Abstract: We propose a new, robust pitch detection algorithm for speech degraded by additive noise. Our algorithm exploits the high correlation between adjacent pitch periods that does not exist for the segment as a whole and performs well in the vicinity of voiced/unvoiced regions where the local SNR is low.

The algorithm works as follows. We first determine an estimate of the pitch period near the short time peak energy where the local SNR is highest. We then adaptively estimate the local pitch period from the peak towards the transition region by using pitch synchronous cross-correlation with an updated waveform. The performance of this new algorithm is compared to the SIFT and CEPSTRUM algorithms.

I INTRODUCTION

Pitch is an important parameter of speech in synthesis, recognition, and speaker verification [1]. A variety of pitch extraction algorithms have been proposed and their performance studied for clean speech [2] as well as for noisy speech [3]. For practical applications, one normally must consider speech in noisy environments.

The acoustic speech waveform can be modeled as the convolution of the impulse response of the vocal tract with quasi-periodic pulses (during voiced speech), or random noise (during unvoiced speech). Accurate and reliable pitch extraction from speech waveform is often difficult for several reasons [2].

1) The glottal excitation is not perfectly periodic but rather a quasi-periodic waveform.
2) The speech waveform itself is time-varying signal as the result of the interaction between the vocal tract and glottal excitation.
3) It is difficult to distinguish between unvoiced speech and low-level voiced speech.
4) Background noise tends to destroy the fine structure of voiced speech.
5) It is difficult to define the exact beginning end of each pitch period during voiced segments.

Two currently popular pitch detectors are cepstrum analysis and the SIFT algorithm. Cepstrum analysis (CEP) uses the property of the logarithm log ab = log a + log b and the periodicity measurement in the quefrency domain, the SIFT algorithm uses the removal of formant structure first and then a periodicity measurement in the time domain.

Both CEP and SIFT techniques are pitch asynchronous (frame-by-frame) analysis algorithms. Because of this fixed frame size, CEP (spectral pitch detector) has difficulty with high pitched speakers and SIFT (hybrid pitch detector) has difficulty with both very low pitched and very high pitched speakers [2]. In addition both the SIFT AND CEP algorithms may perform poorly in classifying voiced/unvoiced (v/uv) segments even for clean speech [2].

It is well known that periodicity measurements in the time domain tend to be robust in noise, and for distortion and spurious transients in the signal [2]. As the result, the SIFT method yields a fairly good performance in noisy speech [3]. Thus, we propose to approach the basic problems in pitch extraction by using a time domain method which will adapt to time varying characteristics of voiced speech.

Because of the slowly time-varying property of voiced speech waveforms and of pitch periods, adjacent pitch periods and waveforms are more highly correlated than those more separated in time.

Our algorithm exploits this high correlation between adjacent pitch periods which does not hold for the segment as a whole. As expected, this approach provides a more reliable estimate of pitch at lower SNR, particularly near the v/uv transition regions.

II PITCH EXTRACTION BY UCC

A) AUTOCORRELATION AND CROSS-CORRELATION [4]

It is well known that the autocorrelation function (ACF) and cross-correlation function (CCF) can be used for pitch extraction.

A frame size of L samples, encompassing several pitch periods is used to calculate ACF

\[ ACF(m) = \sum_{j=1}^{L-1} s(j)s(j+m) \quad m = 0, 1, \ldots, L-1 \]  

where \( s(j) \) is the \( j \)-th sample of input vector. 

CCF is calculated over a portion of the analysis interval \( L' \) where \( L' < L \)

\[ CCF(m) = \sum_{j=1}^{L'-1} s'(j)s(j+m) \quad m = 0, 1, \ldots, L-L' \]  

and \( s'(j) \) is the \( j \)-th sample of input subvector.
The length of $L'$ can be chosen according to the expected pitch period.

Some of the advantages of CCF over ACF are:
1) The relative size of the cross-correlation peak is constant while there is a linear decrease of autocorrelation peaks as a function of the delay.
2) If we choose the length of input vector according to the expected pitch period, CCF can be used in an adaptive pitch synchronous analysis and we can get a more accurate pitch period and avoid averaging over several pitch periods.

Since random noise in speech tends to destroy the fine (quasi-periodic) structure of speech, this adaptive pitch synchronous analysis (updated cross-correlation, UCC) has a relatively good performance as compared to ACF for noisy speech.

B) CORRELATION FUNCTION AS EXTENSION OF MATCHED FILTER

As for any pitch synchronous analysis, the input to the UCC method is the windowed signal whose window size is about 2 pitch periods. In the ideal case, if we consider the maximization of the SNR, we obtain a matched filter [5], with an impulse response which is a delayed, time-reversed version of the input signal, $h(n) = s(L'-n)$.

Fig. 1 shows the input/output relationship of a matched filter with window $w(n)$

$$x(n) \rightarrow [\text{X}] \rightarrow h(n) = s(L'-n) \rightarrow y(n)$$

Its output

$$y(n) = \sum_{j=0}^{L-1} h(n-j)s(j) = \sum_{j=0}^{n-1} s(L'-n+j)s(j) \quad (3)$$

at $n = L'$

$$L'-1 = \sum_{j=0}^{L'-1} s^2(j) = ||s||^2$$

For white input noise of variance $\sigma_n^2$, the variance of the output noise is given by

$$\sigma_o^2 = \sigma_n^2 ||s||^2$$

These classical results are used traditionally to argue that increasing the signal duration in autocorrelation, or cross-correlation, will improve performance because the signal coherence, and thus energy, and noise incoherence, result in an increased signal to noise ratio for increasing duration. We have observed experimentally that, because of fine waveform variations, the cross-correlation for clean speech, will vary within a voiced segment. The maximum cross-correlation of pitch periods at both ends of a voiced segment is as low as 0.78.

Thus, the combined effect of additive noise and of normal changes in speech waveforms lead to a loss of performance towards the ends of voiced segments.

By contrast, as speech is slowly changing signal with time, the waveforms of two consecutive pitch periods are highly correlated in the voiced segment of speech. We thus use the segment of speech corresponding to the previous pitch period as the impulse response of matched filter to determine the next pitch period.

C) UPDATED CROSS-CORRELATION ALGORITHM

An accurate and robust to noise pitch extraction algorithm based on UCC for the range of 50-400Hz works as follows. The speech is sampled at sampling frequency of 10 kHz and the speech samples $s(n)$ are low pass filtered using third order elliptic filter [6].

Accurate Determination of Voiced Segments

The short time energy contour is determined using:

$$E(j) = \frac{1}{N} \sum_{n=0}^{N-1} [s(n*100+n)]^2 \quad j = 0 \ldots J-1$$

where $J = N/100$

It is assumed that during the first 100 ms of recording interval there is no speech present. This initial interval is used to estimate the statistics of background noise. Local energy peaks are found by setting a threshold based on the global peak energy, $ESOM$, and the silence peak energy, $ENOM$. The threshold value is chosen as:

$$THR_1 = 0.035*(ESOM-ENOM) + ENOM$$

The number of local energy peaks is the number of voiced segments in that speech utterance.

Initial Pitch Estimation

We first estimate the pitch period near the first local energy peak. Because the local SNR is highest in that region, the pitch estimates are most accurate and reliable there. Since the maximum anticipated pitch period of speech is 20 msec, the initial pitch estimate is obtained by cross correlating a 20 msec speech segment near the local energy peak with a 40 msec speech segment near the peak using cross correlation equation given in equation (7)

$$C_k(n) = \sum_{j=0}^{L'-1} s_k(j) s_k(j+m) \quad ||s_k|| \neq 0$$

The subscript $k$ is the index of the present pitch period, $L'$ is k-15th the pitch period and should be updated as $k$ changes. The denominator of equation (7) provides normalization.

Tracking the Pitch

We then adaptively estimate the local pitch period from the energy peak toward the transition regions using cross correlation, and a local energy ratio for voiced/unvoiced (v/uv) logic. A local energy ratio is used as subsidiary means to make a v/uv decision and turned out to be useful to distinguish between unvoiced segments and low-level voiced segments in clean speech.
Let $EP(j)$ be the short term energy over an estimated pitch period.

Define a Local Energy Ratio as

$$ER(j) = \frac{EP(j)}{EP(j-1)}$$

We make use of the product of 3 values of $EP(j)$ over successive pitch periods to determine whether the segment is voiced or unvoiced, or

$$ER'(j) = ER(j) \cdot ER(j-1) \cdot ER(j-2) > \text{THRESHOLD (THR2)}$$

To obtain adjacent pitch periods a sliding matched filter (cross correlation) is updated to the speech waveform for the previous pitch period. This process is repeated moving toward the voiced/unvoiced transitions.

We start the process of initial pitch estimation and pitch tracking for each voiced segments, as determined by energy peaks.

D) V/UV CLASSIFICATION USING UCC AND LOCAL ENERGY RATIO

The v/uv decision was made on the basis of the cross correlation coefficients, CCC, and local energy ratios $ER'(j)$ and $ER(j)$ defined before. The v/uv classification diagram is given in Fig. [2].

![V/UV Classification Diagram](image)

**Fig. [2] V/UV Classification Diagram for UCC**

a) CCC: Max. $C_k(m)$ over a pitch period

$$ER'(j) = \frac{EP(j)}{EP(j-1)}$$

b) subsidiary v/uv classification diagram of transition region $T_1$ in diagram a)

$$ER(j) = \frac{EP(j)}{EP(j-1)}$$

Threshold values were determined experimentally and turned out to be very consistent for every speaker.

Two periods of delayed pitch and voicing information are retained for error detection and correction as in the SIFT algorithm [6]. They correct isolated unvoiced decisions in a voiced segment and isolated voiced decisions in an unvoiced segment.

III PERFORMANCE EVALUATION AND EXPERIMENTAL RESULTS

The utterances used in this study included one mono-syllabic word, two double-syllabic words, and two sentences, each of which spoken by a male and a female. They were:

1. nine
2. subtract
3. replace
4. I know when my lawyer is due
5. I was stunned by the beauty of the view

Sentence 4 is all voiced (except for a short unvoiced segment) and sentence 5 contains both voiced and unvoiced speech [2]. The performance of the SIFT and CEPSTRUM algorithm were studied and compared with that of UCC. In order to study the performance of UCC and to compare that of SIFT and CEPSTRUM algorithm in various noisy condition, Gaussian random noise was added to each of utterances so that the peak signal-to-(rms) noise ratio of the resulting signal becomes 40, 35, 30, 25, 23, 21, 19, 17, 15dB.

Errors were evaluated on a pitch period-by-pitch period basis.

Since the SIFT and CEPSTRUM algorithms work on a frame-by-frame basis, the output of these algorithms were converted to pitch periods by interpolation using the LPC SYNTHESIZER program [6].

Reference pitch contours and v/uv decisions for each utterance were made by eye using a graphical display of speech waveforms.

We used the definitions of error parameters of Rabiner et al [2]: fine pitch error, gross pitch error, voiced to unvoiced error and unvoiced to voiced error.

A clean speech record for a female speaker, a noisy speech record (21dB peak signal to rms noise ratio), the reference pitch contour and the results for the 3 algorithms under study are shown in Fig. [3].

![Speech Waveforms and Pitch Contours](image)

**Fig. [3] Speech waveforms and pitch contours for utterance "subtract"**
While algorithms lose substantial portions of voiced segments for increasing amount of noise, gross or fine pitch errors in retained voiced portions do not increase substantially with noise. The rms fine pitch errors are approximately .13 msec for UCC, .17 msec for SIFT, and .15 msec for CEP respectively.

The performances of the UCC, SIFT, and CEP algorithms in noisy environment are tabulated in Table 1. Table 1a gives the gross pitch error, Table 1b the v/uv error, Table 1c, the uv/v error.

Table 1: Performance of three pitch detection algorithms, a) gross pitch error (averaged number of error per utterances) b) voiced-to-unvoiced error rates (%) c) unvoiced-to-voiced error rates (%).

From table 1, we observe that the performance of UCC is substantially superior to both CEP and SIFT for v/uv errors and uv/v errors. Note that the gross pitch error tabulation indicate that the retained voiced segment have few gross pitch errors for any of the algorithms tested.

For the important v/uv error, a 10% error rate corresponds to SNR’s of 32dB for CEP, 26 for SIFT and 20 for UCC. Thus UCC compares in performance to SIFT and CEP for 6 and 12dB higher noise levels respectively.

IV DISCUSSION AND CONCLUSION

Pitch is an important parameter in speech analysis system. While most of currently available algorithms work well for clean speech, their performance degrades when noise is present [3], particularly in the vicinity of voiced/unvoiced regions where the local SNR is low.

Our algorithm (UCC) exploits the high correlation between adjacent pitch periods. Our approach provides a more reliable estimate in lower SNR, particularly near the transition regions. Further an improved smoothed fine pitch estimate could be obtained by fitting the raw data with a low order polynomial over the entire voiced segment.

For the autocorrelation method used in SIFT algorithm it has been shown that the window size must be on the order of several pitch periods to ensure a reliable estimate [6]. By contrast with the SIFT algorithm, just two pitch periods are needed to get a reliable pitch period estimate in the UCC algorithm.

Most pitch asynchronous algorithms, such as SIFT and CEP have a speaker dependent performance. Since UCC, as adaptive pitch synchronous algorithm, makes no assumption on pitch period and no restriction on window size, it has speaker independent performance in clean speech.

REFERENCES