On the Use of Autocorrelation Analysis for Pitch Detection

LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—One of the most time honored methods of detecting pitch is to use some type of autocorrelation analysis on speech which has been appropriately preprocessed. The goal of the speech preprocessing in most systems is to whiten, or spectrally flatten, the signal so as to eliminate the effects of the vocal tract spectrum on the detailed shape of the resulting autocorrelation function. The purpose of this paper is to present some results on several types of (nonlinear) preprocessing which can be used to effectively spectrally flatten the speech signal. The types of nonlinearities which are considered are classified by a nonlinear input-output quantizer characteristic. By appropriate adjustment of the quantizer threshold levels, both the ordinary (linear) autocorrelation analysis, and the center clipping–peak clipping autocorrelation of Dubnowski *et al.* [1] can be obtained. Results are presented to demonstrate the degree of spectrum flattening obtained using these methods. Each of the proposed methods was tested on several of the utterances used in a recent pitch detector comparison study by Rabiner *et al.* [2] Results of this comparison are included in this paper. One final topic which is discussed in this paper is an algorithm for adaptively choosing a frame size for an autocorrelation pitch analysis.

I. Introduction

ALTHOUGH a large number of different methods have been proposed for detecting pitch, the autocorrelation pitch detector is still one of the most robust and reliable of pitch detectors [2]. There are several reasons why autocorrelation methods for pitch detection have generally met with good success. The autocorrelation computation is made directly on the waveform and is a fairly straightforward (albeit time consuming) computation. Although a high processing rate is required, the autocorrelation computation is simply amenable to digital hardware implementation generally requiring only a single multiplier and an accumulator as the computational elements. Finally, the autocorrelation computation is largely phase insensitive.[1] Thus, it is a good method to use to detect pitch of speech which has been transmitted over a telephone line, or has suffered some degree of phase distortion via transmission.

Although an autocorrelation pitch detector has some advantages for pitch detection, there are several problems associated with the use of this method. Although the autocorrelation function of a section of voiced speech generally displays a fairly prominent peak at the pitch period, autocorrelation peaks due to the detailed formant structure of the signal are also often present. Thus, one problem is to decide which of several autocorrelation peaks corresponds to the pitch period. Another problem with the autocorrelation computation is the required use of a window for computing the short time autocorrelation function. The use of a window for analysis leads to at least two difficulties. First there is the problem of choosing an appropriate window. Second there is the problem that (for a stationary analysis),[2] no matter which window is selected, the effect of the window is to taper the autocorrelation function smoothly to 0 as the autocorrelation index increases. This effect tends to compound the difficulties mentioned above in which formant peaks in the autocorrelation function (which occur at lower indices than the pitch period peak) tend to be of greater magnitude than the peak due to the pitch period.

A final difficulty with the autocorrelation computation is the problem of choosing an appropriate analysis frame (window) size. The ideal analysis frame should contain from 2 to 3 complete pitch periods. Thus, for high pitch speakers the analysis frame should be short (5-20 ms), whereas for low pitched speakers it should be long (20-50 ms).

A wide variety of solutions have been proposed to the above problems. To partially eliminate the effects of the higher formant structure on the autocorrelation function, most methods use a sharp cutoff low-pass filter with cutoff around 900 Hz. This will, in general, preserve a sufficient number of pitch harmonics for accurate pitch detection, but will eliminate the second and higher formants. In addition to linear filtering to remove the formant structure, a wide variety of methods have been proposed for directly or indirectly spectrally flattening the speech signal to remove the effects of the first formant [3]-[5], [1]. Included among these techniques are center clipping and spectral equalization by filter bank methods [3], inverse filtering using linear prediction methods [4], spectral flattening by linear prediction and a Newton transformation [5], and spectral flattening by a combination of center and peak clipping methods [1].

Each of these methods has met with some degree of success; however, problems still remain. It is the purpose of this paper to investigate the properties of a class of nonlinearities applied to the speech signal prior to autocorrelation analysis with the purpose of spectrally flattening the signal. Also a solution to the problem of choosing an analysis frame size which adapts to the estimated average pitch of the speaker will be presented.

The organization of this paper is as follows. In Section II we review the theory of short-time autocorrelation analysis and present the various types of nonlinearities to be investigated for spectrally flattening the speech. Examples of signal spectra

Manuscript received April 4, 1976; revised August 16, 1976.
The author is with the Bell Laboratories, Murray Hill, NJ 07974.

[1] In the limit of exactly periodic signals, or for an infinite correlation function it is exactly phase insensitive.

[2] A stationary analysis is one for which the same set of input samples is used in computing all the points of the autocorrelation function. A nonstationary analysis is impractical for pitch detection because of the large number of autocorrelation points involved in the computation.

obtained with the nonlinearities being used will be given in this section. In Section III the results of a limited but formal evaluation of each of the nonlinear autocorrelation analyses are given. Several of the test utterances used in [2] are used in this test for comparison purposes. In Section IV we discuss a simple algorithm for adapting the frame size of the analysis based on the estimated average pitch period for the speaker, and present results on how well it worked on several test examples.

II. Short-Time Autocorrelation Analysis

Given a discrete time signal $x(n)$, defined for all n, the autocorrelation function is generally defined as

$$\phi_x(m) = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m). \tag{1}$$

The autocorrelation function of a signal is basically a (noninvertible) transformation of the signal which is useful for displaying structure in the waveform. Thus, for pitch detection, if we assume $x(n)$ is exactly periodic with period P, i.e., $x(n) = x(n+P)$ for all n, then it is easily shown that

$$\phi_x(m) = \phi_x(m+P), \tag{2}$$

i.e., the autocorrelation is also periodic with the same period. Conversely, periodicity in the autocorrelation function indicates periodicity in the signal.

For a nonstationary signal, such as speech, the concept of a long-time autocorrelation measurement as given by (1) is not really meaningful. Thus, it is reasonable to define a short-time autocorrelation function, which operates on short segments of the signal, as

$$\phi_\ell(m) = \frac{1}{N} \sum_{n=0}^{N'-1} [x(n+\ell)w(n)]\,[x(n+\ell+m)w(n+m)],$$
$$0 \leqslant m \leqslant M_0 - 1 \tag{3}$$

where $w(n)$ is an appropriate window for analysis, N is the section length being analyzed, N' is the number of signal samples used in the computation of $\phi_\ell(m)$, M_0 is the number of autocorrelation points to be computed, and ℓ is the index of the starting sample of the frame. For pitch detection applications N' is generally set to the value

$$N' = N - m \tag{4}$$

so that only the N samples in the analysis frame (i.e., $x(\ell)$, $x(\ell+1), \cdots, x(\ell+N-1)$) are used in the autocorrelation computation. Values of 200 and 300 have generally been used for M_0 and N, respectively, [1] corresponding to a maximum pitch period of 20 ms (200 samples at a 10 kHz sampling rate) and a 30 ms analysis frame size. As will be discussed later a rectangular window (i.e., $w(n) = 1$, $0 \leqslant n \leqslant N-1$, $w(n) = 0$ elsewhere) is used for all the computations to be described in this paper.

To reduce the effects of the formant structure on the detailed shape of the short-time autocorrelation function, two preprocessing functions were used prior to the autocorrelation computation of (3), as discussed in Section I. Fig. 1 shows a block diagram of the processing which was used. The speech signal $s(n)$ is first low-pass filtered by an FIR, linear phase, digital filter with a passband of 0 to 900 Hz, and a stopband beginning at 1700 Hz.[3] The output of the low-pass filter is then used as input to two nonlinear processors, labeled NL1 and NL2 in Fig. 1. The nonlinearities used in each path may or may not be identical. The types of nonlinearities which were investigated were various center clippers, and peak clippers. Based on earlier works [3], [1] it has been shown that such nonlinearities can provide a fairly high degree of spectral flattening, and are computationally quite efficient to implement [1]. Additionally, the capability of correlating two nonlinearly processed versions of the same signal provides a useful degree of flexibility into the system. It has also been argued that such a correlation will be most appropriate in a variety of actual situations in pitch detection.[4]

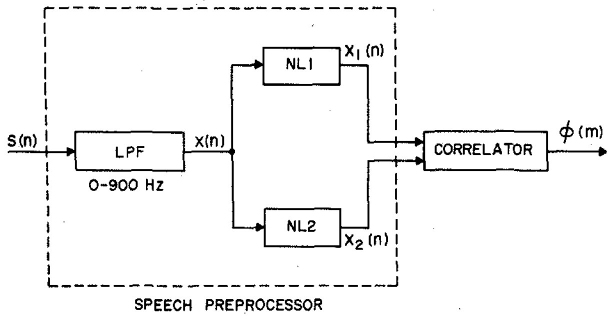


Fig. 1. Block diagram of the nonlinear correlation processing.

Three types of nonlinearity have been considered. They are classified according to their input-output quantization characteristic in the following way. The first type of nonlinearity is a compressed center clipper whose output $y(n)$ obeys the relation (with $x(n)$ as input)[5]

$$\begin{aligned} y(n) = \text{clc}\,[x(n)] &= (x(n) - C_L), && x(n) \geqslant C_L \\ &= 0, && |x(n)| < C_L \\ &= (x(n) + C_L), && x(n) \leqslant -C_L \end{aligned} \tag{5}$$

where C_L is the clipping threshold. The second nonlinearity is a simple center clipper with the input-output relation[6]

$$\begin{aligned} y(n) = \text{clp}\,[x(n)] &= x(n), && x(n) \geqslant C_L \\ &= 0, && |x(n)| < C_L \\ &= -x(n), && x(n) \leqslant -C_L. \end{aligned} \tag{6}$$

Finally, the third nonlinearity is the combination center and peak clipper with the input-output relation[7]

[3] The filter had an impulse response duration of 25 samples. The filter passband was flat to within ±0.03, and the stopband response was down at least 50 dB.

[4] M. Sondhi–personal communication.

[5] The function clc $[x]$ stands for *clip* and *compress* x.

[6] The function clp $[x]$ stands for *clip* x.

[7] The function sgn $[x]$ stands for the *sign* of x.

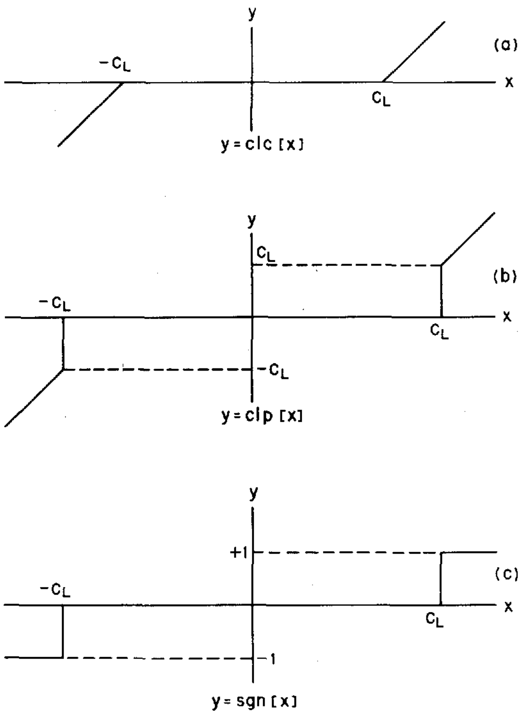


Fig. 2. Input-output characteristics of each of the three nonlinearities used in the investigation.

$$
\begin{aligned}
y(n) = \text{sgn}\,[x(n)] &= 1, & x(n) &\geqslant C_L \\
&= 0, & |x(n)| &< C_L \\
&= -1, & x(n) &\leqslant -C_L.
\end{aligned}
\tag{7}
$$

Fig. 2 illustrates the input-output characteristics for the three nonlinearities of (5)-(7). Allowing a direct path connection between input and output for each of the nonlinearities of Fig. 1 (i.e., $y = x$) it can be seen that there are ten distinct ways[8] in which the signals $x_1(n)$ and $x_2(n)$ can be correlated, depending on which of the nonlinearities is used for NL1 and NL2. Table I summarizes these ten possibilities.

It should be noted that correlation number 1 in Table I corresponds to an ordinary autocorrelation, whereas correlation number 10 corresponds to the combination peak clipping, center clipping correlation discussed in [1]. Also shown in Table I are the required computations needed to implement the combined correlation for each possibility. In the most general case (correlation number 1) a multiply and an add are required for each sample in the computation. For cases 2, 3, 7, 8, and 9, whenever either $x_1(n)$ or $x_2(n+m)$ falls below the clipping level, C_L, no computation is required as indicated by the ϕ in the computation column for these cases. For cases 4, 5, and 6 only an adder/subtractor is required because $\text{sgn}[x(n+m)]$ can only assume the values +1 [addition of $x_1(n)$], 0 (no computation), or -1 [subtraction of $x_1(n)$]. Finally, case 10 only requires an up-down counter to implement as discussed in [1].

TABLE I
COMBINATIONS OF NONLINEARITIES FOR CORRELATON ANALYSIS

Correlation No.	$x_1(n)$	$x_2(n)$	Computation
1	x(n)	x(n)	*,+
2	clc[x(n)]	clc[x(n)]	*,+/Ø
3	clp[x(n)]	clp[x(n)]	*,+/Ø
4	x(n)	sgn[x(n)]	±/Ø
5	clc[x(n)]	sgn[x(n)]	±/Ø
6	clp[x(n)]	sgn[x(n)]	±/Ø
7	x(n)	clc[x(n)]	*,+/Ø
8	x(n)	clp[x(n)]	*,+/Ø
9	clp[x(n)]	clc[x(n)]	*,+/Ø
10	sgn[x(n)]	sgn[x(n)]	counter/Ø

The justification for considering the nonlinearities of Fig. 2 for use in autocorrelation analysis is obtained by examining the effects of the nonlinearities on the waveforms. It can be argued that a center clipper effectively attenuates the effects of first formant structure on the waveform, without seriously affecting the pitch pulse indications. However, it has been argued that the peak clipping of the sgn quantizer [Fig. 2(c)] gives too much weight to signal amplitudes that just exceed the clipping threshold, and too little weight to signal amplitudes that exceed the clipping threshold by a wide margin. Thus, the justification for the clc (*cl*ip and *c*ompress) and the clp (*clip*) quantizers is that they provide a compromise between the extremes of no clipping and infinite peak clipping.

Before proceeding to some examples showing the effects of each of these nonlinearities, it is worth noting that the method used to set the clipping threshold (C_L) for each of these nonlinearities was exactly the method used in [1], i.e., set the clipping as a fixed percentage (68 percent) of the smaller of the maximum absolute signal level over the first and last one-thirds of the analysis frame. This method has proven quite successful in all tests to date [2].

Fig. 3 illustrates the effects of each of the quantizer characteristics of Fig. 1 on a typical frame of voiced speech. The left-hand side of Fig. 3 shows the sequence of signals $x(n)$, $\text{clc}[x(n)]$, $\text{clp}[x(n)]$, and $\text{sgn}[x(n)]$. Superimposed on the plot of $x(n)$ is the clipping level for this frame of speech. The right-hand side of Fig. 3 shows the sequence of autocorrelations corresponding to each of the sequences at the left (i.e., correlations numbers 1, 2, 3, and 10 in Table I).[9] A rectangular window was used in all cases for computing the correlations as no other type of window is reasonable. The effects of the nonlinear clipping are readily evident in this figure. Although there is a sharp peak in the autocorrelation due to the pitch at $m = 80$ for all four correlations, the shape of the correlation function for the unprocessed speech [Fig. 3(a)] is significantly different from the shape of the correlation function for all of the nonlinearly processed signals [Fig. 3(b)-(d)]—especially in the low time part of the correlation (i.e., m going from 20 to 80). Fig. 3(d) also illustrates the problems associated with using the sgn quantizer in that all speech samples which exceed the clipping threshold are weighted equally. Thus in the third

[8] Theoretically there are 16 ways in which $x_1(n)$ and $x_2(n)$ can be correlated. For all practical purposes, however, six pairs of these results are equivalent. Thus, only ten ways of correlating $x_1(n)$ and $x_2(n)$ are considered here.

[9] Each of the signal amplitudes in Figs. 3-7, and 9 is scaled so that the maximum value is set to 1.0 for display purposes. Thus, it is difficult to compare these amplitude sequences against each other.

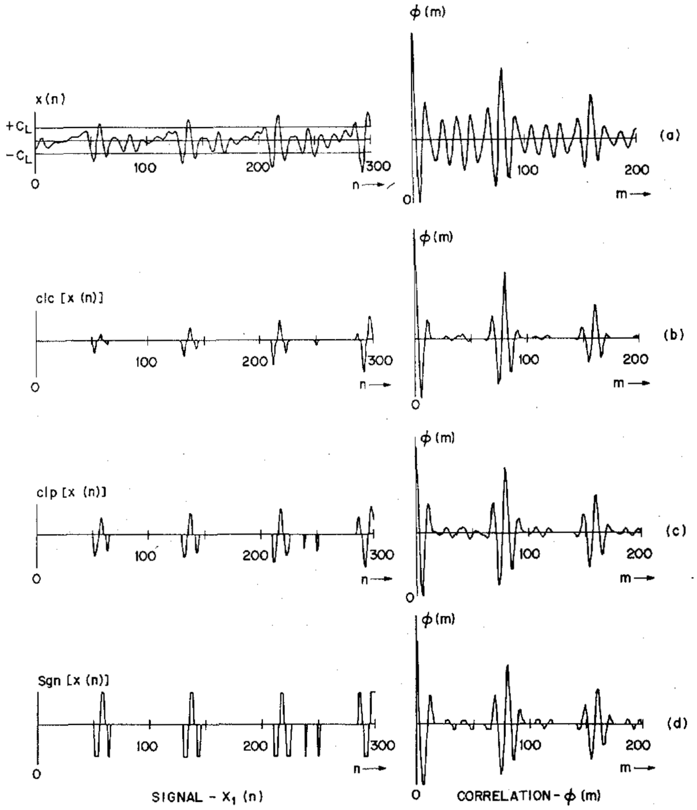


Fig. 3. Each of the processed signals and the resulting correlation function for a section of voiced speech.

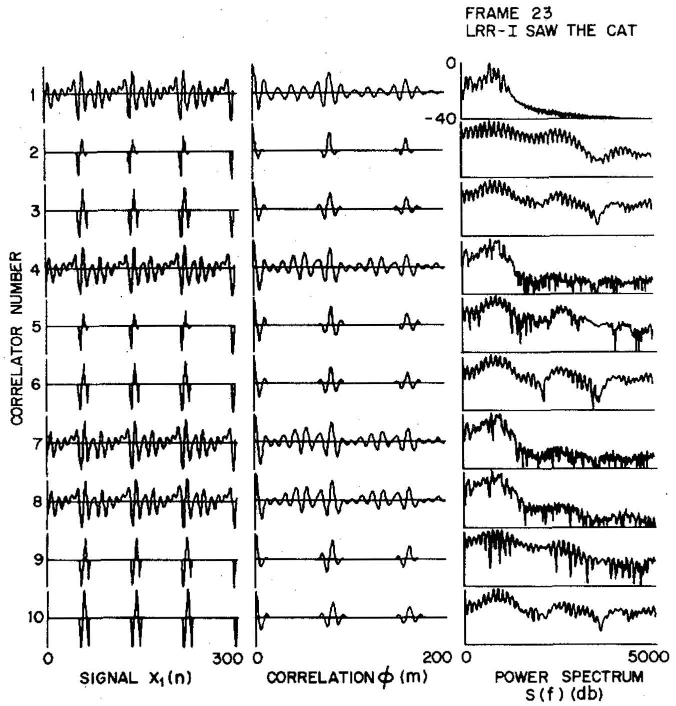


Fig. 4. The signal $x_1(n)$; the resulting correlation and power spectrum for each of the ten correlators of Table I for a section of voiced speech.

period there are five pulses of varying width whereas in the first periods there are only three pulses. Fig. 3(b) shows that such problems are inherently eliminated by the clc quantizer whose output samples are proportional in amplitude to the amount by which they exceed the clipping threshold.

Spectral Flattening from the Quantizers

It has already been argued that the effect of the nonlinear processing preceding the correlation computation is to approximately spectrally flatten the signal spectrum, thereby enhancing the periodicity of the signal. To investigate this, the power spectrum of each of the correlation functions of Table I was computed directly from the correlation function by the Fourier transform relation

$$S(f) = \sum_{M=-(M_0-1)}^{M_0-1} \phi(m) e^{-j2\pi fm}. \tag{8}$$

A 512-point FFT was used to compute $S(f_k)$, $k = 0, 1, \cdots, 511$ where

$$f_k = k \cdot \left(\frac{10\,000}{512}\right), \tag{9}$$

i.e., at 512 points around the unit circle. Since $\phi(m)$ is theoretically infinite, a $(2M_0 + 1)$ point Hamming window was used to taper $\phi(m)$ smoothly to 0. (Note we are assuming $\phi(m)$ is symmetric, i.e., $\phi(m) = \phi(-m)$.)

Figs. 4-7 show plots of the results of processing four different sections of voiced speech. The left-hand column shows the signal $x_1(n)$, the middle column shows the signal $\phi(m) = x_1(n)$ correlated with $x_2(n)$ (where $x_2(n)$ is as specified in Table I), and the right-hand column shows the power spectrum $S(f)$ obtained as described above. The ten rows in each figure correspond to the ten combinations of signals to be correlated as shown in Table I. An examination of Fig. 4 shows that for the unprocessed signal (i.e., the top row) the first several harmonics are seen in the power spectrum. Beyond 1 kHz, the spectrum decays rapidly due to the lowpass filter (the lack of a sharp falloff in the spectrum is due to a combination of the signal and autocorrelation windows). The amplitudes of the harmonics vary with the first formant envelope. It can be seen that the spectrum for the autocorrelations of each of the nonlinear quantizers (i.e., rows 2, 3, and 10) are much flatter than the original signal spectrum. Additionally, the spectra of the nonlinearly processed signals are much broader than the original spectrum. It is interesting to note that the spectra from correlations involving $x(n)$, i.e., correlations numbers 1, 4, 7, and 8, are the least flattened and are generally quite irregular (i.e., the harmonics are not very easy to find).

Fig. 5 shows similar results from a different section of voiced speech. As seen from the spectrum of the unprocessed signal (on the top line) the bandwidth of the first formant is fairly small, causing the correlation function to show a great deal of formant periodicity for small values of m. The effects of the nonlinearities on the signal spectra are quite impressive even for such a difficult case as this one.

Fig. 6 shows results from a section of speech from a female speaker (high pitch). Again the spectrum from the unpro-

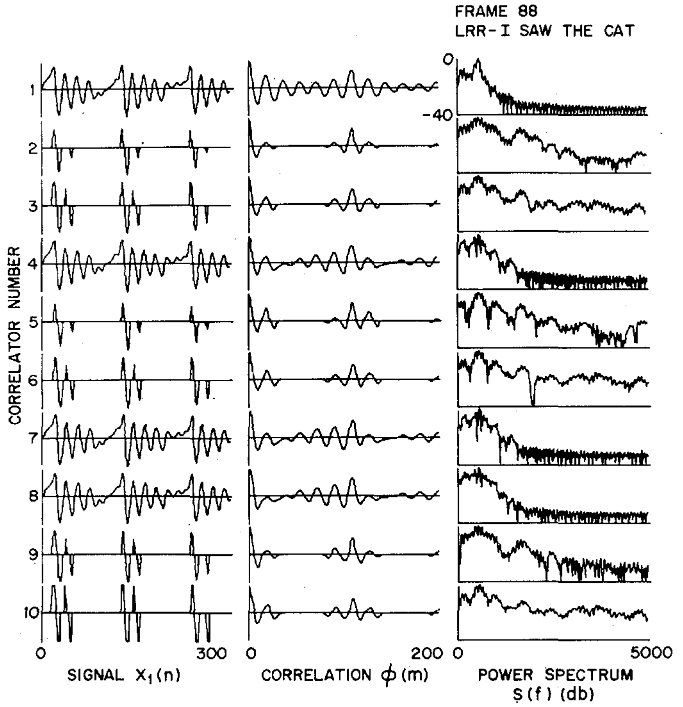


Fig. 5. The signal $x_1(n)$; the resulting correlation and power spectrum for each of the ten correlators of Table I for *another* section of voiced speech.

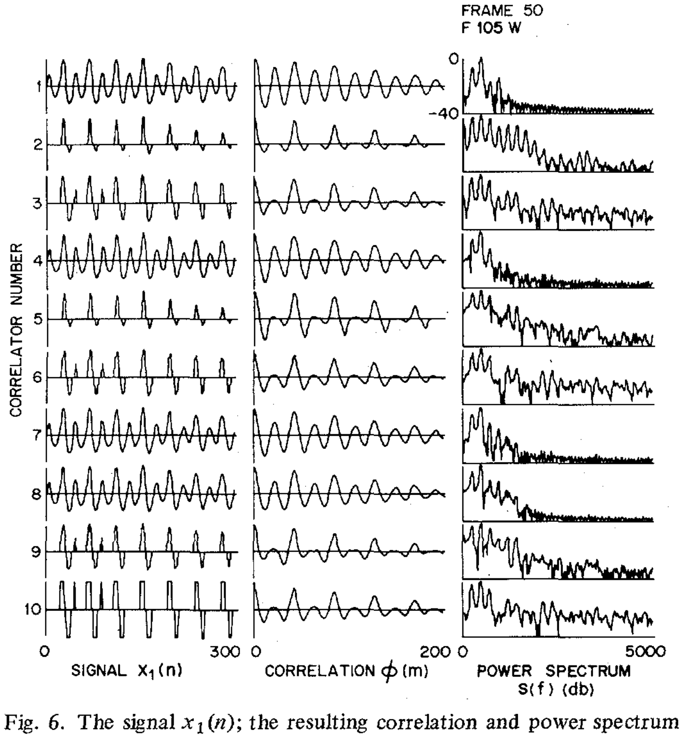


Fig. 6. The signal $x_1(n)$; the resulting correlation and power spectrum for each of the ten correlators of Table I for a section of voiced speech from a female speaker.

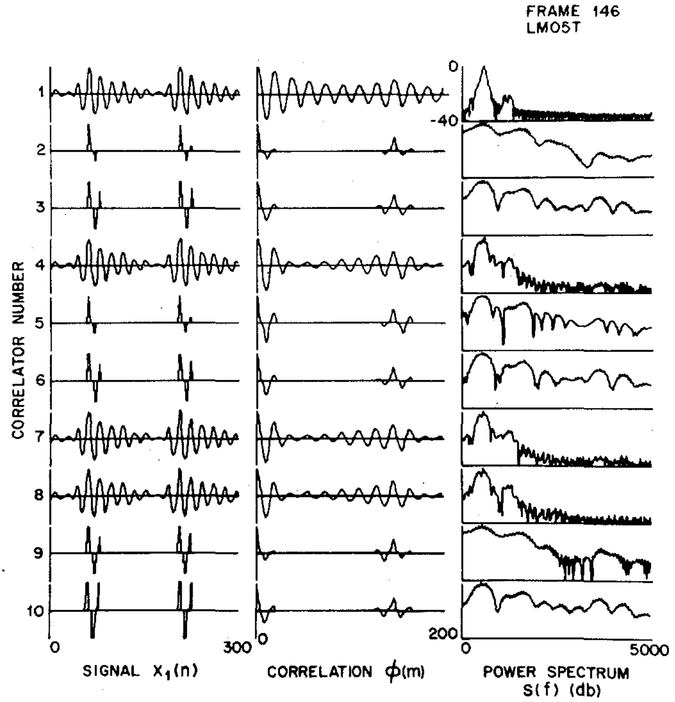


Fig. 7. The signal $x_1(n)$; the resulting correlation and power spectrum for each of the ten correlators of Table I for a section of voiced speech from a low pitched male speaker.

cessed signal shows only a few harmonics whose amplitudes vary with the formant amplitude. The nonlinearly processed samples show various degrees of spectral flattening, as anticipated by the previous discussion.

Finally, Fig. 7 shows the results obtained with a voiced frame from a low pitched (long period) male speaker. In this example the first formant has a very narrow bandwidth as seen in the spectrum at the top of Fig. 7. Pitch detection directly on the autocorrelation of the signal yields incorrect results in this case due to the first formant peak(s) in the autocorrelation function. However, as shown in Fig. 7, almost any of the nonlinearities flatten the spectrum and eliminate the troublesome effects of the sharp first formant in the resulting correlation function.

In summary, we have presented examples which tend to show that, as anticipated, the effect of nonlinearly quantizing the signal amplitudes using the quantizers of Fig. 1 is to effectively flatten and broaden the signal power spectrum, thereby reducing the effects of the first formant on the correlation function, and simplifying the pitch detection problem. In the next section we present results of a comparative test of the performance of the ten correlation pitch detectors discussed in this section on a series of speech utterances.

III. Evaluation of the Ten Nonlinear Correlations

In order to evaluate and compare the performance of the ten nonlinear correlations discussed in the preceding section, a small set of the utterances from the data base in [2] was used as a test set. For each of the utterances a reference pitch contour was available from which an error analysis was made [6]. Since the problem of making a reliable voiced-unvoiced decision was not a concern here, the reference voiced-unvoiced contour was used directly, i.e., each correlator was required to estimate the pitch period, assuming *a priori* that the interval was properly classified as voiced. (No pitch detection was done during unvoiced intervals.) However, if the peak correlation value (normalized) fell below a threshold (0.25), the interval was classified as unvoiced since reliable selection of

TABLE II
STANDARD DEVIATIONS FOR TEN CORRELATORS

Correlation Number	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
1	.53	.60	.80	.97	.54	1.34	1.58	.85	.88	1.23	1.24	1.52	1.23
2	.63	.71	.84	.68	.58	1.59	1.70	1.18	1.19	1.21	1.32	1.39	1.05
3	.44	.64	.72	.70	.56	1.50	1.66	1.10	1.14	1.31	1.17	1.46	1.04
4	.52	.61	.79	.83	.79	1.54	1.82	.97	1.11	1.32	1.24	1.50	1.34
5	.63	.64	.76	.86	.99	1.47	1.75	1.07	1.07	1.24	1.02	1.37	1.27
6	.47	.65	.72	.78	.82	1.55	1.66	1.08	1.19	1.29	1.05	1.43	1.25
7	.40	.63	.80	.78	.54	1.46	1.70	.99	1.04	1.34	1.29	1.45	1.24
8	.57	.65	1.15	.73	.54	1.67	1.88	1.09	1.18	1.29	1.51	1.45	1.24
9	.46	.68	.97	.67	.56	1.56	1.76	1.13	1.17	1.30	1.41	1.46	1.11
10	.45	.79	.69	.75	.75	1.50	1.69	1.13	1.25	1.39	1.28	1.57	1.17

Standard Deviation of Pitch Period - Unsmoothed

Correlation Number	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
1	.61	.40	.51	.50	.50	.84	1.12	1.59	1.52	2.08	1.22	1.92	1.75
2	.62	.48	.61	.50	.59	1.09	1.40	1.63	1.81	1.54	1.25	2.15	1.50
3	.61	.49	.55	.50	.57	.85	1.21	1.49	1.76	1.34	1.28	1.98	1.41
4	.60	.43	.55	.56	.49	1.25	.97	1.48	1.59	1.87	1.33	2.17	1.83
5	.70	.48	.57	.54	.62	1.18	1.02	1.65	1.60	1.36	1.55	2.23	1.53
6	.61	.56	.60	.56	.52	1.06	1.13	1.63	1.68	1.69	1.57	1.93	1.51
7	.59	.46	.54	.71	.51	1.03	1.14	1.55	1.75	1.77	1.21	2.15	2.11
8	.63	.48	.67	.68	.50	1.19	1.43	1.78	1.83	1.78	1.84	2.48	1.92
9	.59	.50	.64	.51	.57	1.04	1.41	1.67	1.92	1.48	1.41	2.39	1.66
10	.59	.59	.54	.56	.58	1.05	1.26	1.63	1.76	1.68	1.56	2.29	1.63

Standard Deviation of Pitch Period - Smoothed

the pitch was not possible with a correlation peak below this threshold.

Thirteen utterances from [2] were used in this comparison. Tables II-V present the results of an error analysis which measured the average and standard deviation of the pitch period, the number of gross pitch period errors, and the number of voiced-to-unvoiced errors [2].[10] For all utterances the average pitch period error was well below 0.5 samples (10 kHz sampling rate) and so the results of this measurement are not presented. Table II presents the standard deviations of the pitch period for the ten correlations. The results are also presented for the errors when the pitch contours were nonlinearly smoothed using a medium smoothing algorithm [7]. From Table II it can be seen that the standard deviations for all correlators were approximately the same for the same utterance. It is also seen that as the average pitch period gets longer (reading from left to right) the standard deviation increases proportionally.

Tables III and IV show the error statistics for gross errors, and voiced-to-unvoiced errors both for the unsmoothed pitch contours (Table III) and for the smoothed pitch contours (Table IV). These tables show that for the high pitched speakers (utterances prefaced by C1, F1, F2), although some differences[11] were present in the error scores for the unsmoothed data, the nonlinear smoother was able to correct most of the errors. Thus the overall performance on the first four utterances was approximately the same for all correlators. For the low pitched speakers (utterances prefaced by LM, M2) there were more significant differences between the correlators. For the category of gross errors, correlators 1 and 8 generally had the largest numbers of errors across the last 6 utterances in the test. However, for the category of voiced-to-unvoiced errors, correlators 2 and 9 had consistently the largest number of errors. Although the smoothing signifi-

TABLE III
ERROR STATISTICS FOR TEN CORRELATORS—UNSMOOTHED

Correlation Number	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
1	6	1	1	2	7	1	4	24	33	40	18	18	25
2	12	8	6	2	8	1	6	8	15	11	6	5	8
3	9	5	2	3	6	1	5	13	22	10	10	7	11
4	6	1	2	4	6	2	3	21	29	24	15	14	20
5	14	6	2	3	6	2	3	14	25	18	18	17	19
6	11	5	2	1	4	2	4	13	24	12	11	14	13
7	8	2	2	4	8	1	4	24	31	27	13	9	16
8	7	2	3	5	10	1	10	28	37	34	16	13	22
9	7	1	3	4	8	1	7	16	25	14	10	8	11
10	14	5	4	1	6	2	6	15	27	15	11	9	13

Number of Gross Voiced Errors (Unsmoothed)

Correlation Number	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
1	0	0	1	2	1	1	0	1	3	0	2	3	2
2	2	1	5	4	3	2	0	10	11	4	10	13	11
3	2	0	2	4	2	2	0	3	6	2	5	4	5
4	0	0	1	0	0	0	0	1	1	0	4	0	2
5	0	0	1	1	0	0	0	2	4	0	4	2	4
6	0	0	2	3	2	0	0	3	5	3	5	2	4
7	0	0	2	1	0	2	0	1	3	0	2	4	4
8	1	0	2	2	1	2	0	5	3	1	3	10	4
9	2	0	3	4	2	2	0	4	7	4	4	11	10
10	0	0	1	5	3	0	1	6	5	2	6	6	7

Number of Voiced - Unvoiced Errors (Unsmoothed)

	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
Total Number of Voiced Intervals	213	133	174	169	118	152	157	170	170	144	105	134	147

TABLE IV
ERROR STATISTICS FOR TEN CORRELATORS—SMOOTHED

Correlation Number	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
1	0	0	0	0	0	0	0	11	17	4	0	2	2
2	0	0	0	0	0	0	0	6	5	6	2	0	3
3	0	0	0	0	0	0	0	7	3	4	1	1	1
4	0	0	0	0	0	0	0	8	8	2	0	2	11
5	0	0	0	0	0	0	0	6	2	4	3	3	5
6	0	0	0	0	0	0	0	3	3	3	2	2	3
7	0	0	0	0	0	0	0	9	8	8	0	2	3
8	0	0	0	0	0	0	1	9	8	8	3	0	7
9	0	0	0	0	0	0	1	4	9	7	2	0	3
10	0	0	0	0	0	0	0	3	4	3	2	2	3

Number of Gross Voiced Errors - (Smoothed)

Correlation Number	C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
1	2	1	0	3	5	2	1	13	13	20	18	11	17
2	3	0	1	4	9	2	1	9	10	7	9	9	2
3	3	0	0	4	8	2	1	3	14	5	9	3	2
4	2	1	0	4	4	1	2	3	8	8	19	8	1
5	6	1	0	3	5	1	2	4	7	11	15	3	0
6	2	0	0	3	6	1	2	9	10	8	10	8	0
7	2	1	0	4	5	2	1	1	6	7	17	3	2
8	2	1	1	5	8	2	1	10	19	8	9	12	6
9	3	0	1	6	9	2	1	7	12	6	7	11	4
10	2	C	0	4	9	1	3	11	10	8	10	10	4

Number of Voiced - Unvciced Errors (Smoothed)

[10] A voiced-to-unvoiced error occurred when a voiced region was improperly classified as an unvoiced region because no peak above the threshold was present in the correlation function.

[11] These differences for the high pitched (short period) speakers were due to pitch period doubling, i.e., the correlation peak at twice the period was somewhat higher than the correlation peak at the true period. This is a common effect when the pitch period is on the order of 30 ms (300 Hz pitch) as was the case for these speakers.

cantly reduced the number of gross errors for many of the correlators, in turn it increased the number of voiced-to-unvoiced errors. Since both errors constitute a pitch error, in this case the most significant error statistic is probably the sum of the gross errors and voiced-to-unvoiced errors, Table V shows these results. Based on this combined error statistic the following conclusions can be drawn about the performance of the ten correlators.

1) For high pitched speakers the differences in performance scores between the different correlators are small and probably insignificant. It is for this class of speakers that *any* type of correlation measurement of pitch period tends to work very well.

2) For low pitched speakers fairly significant differences in the performance scores existed. Correlator number 1 (the normal linear autocorrelation) tended to give the worst performance for all utterances in this class. Correlators numbers 4, 7, 8 (the ones involving an unprocessed $x(n)$ in the computation) were also somewhat poorer in their overall performance based on the sum of gross errors and voiced-to-unvoiced errors.

3) Differences in the performance among the remaining six correlators were not consistent. Thus, any one of these correlators would be appropriate for an autocorrelation pitch detector.

It is interesting to note that (as seen in Tables III-V) the results for utterance M208T were significantly worse than for utterance M208M. These utterances were simultaneously recorded—the difference being that M208T was recorded off a telephone line, whereas M208M was recorded from a close talking microphone. This result is due to the band-limiting effects of the telephone line (300 Hz cutoff frequency) which eliminate the first few harmonics of the pitch, thereby making accurate pitch detection more difficult.

To illustrate the errors made during one of the more difficult utterances, Fig. 8 shows the pitch period contours from three of the correlators for the utterance LMO5T ("we were away a year ago," spoken by a low pitched male over the telephone line). Also shown in this figure is the nonlinearly smoothed pitch contour from correlator number 10. The pitch period contour from correlator number 1 [Fig. 8(a)] shows the large number of gross pitch period errors made during the analysis. It is readily seen that most of the errors involved choosing a low valued correlation peak rather than the one at the pitch period. These errors, although due somewhat to the frame size used for analysis (30 ms or 300 samples), are primarily due to the narrow bandwidth first formant which has a stronger correlation peak than the one due to the pitch period. The results for correlators number 2 [Fig. 8(b)] and 10 [Fig. 8(c)] confirm the fact that the use of the nonlinearities prior to correlation greatly flattens the spectrum, thereby reducing the number of errors of the type discussed above. As shown in Fig. 8(d), the nonlinear smoother is quite capable of correcting most of the gross pitch errors in analysis from the correlators using the nonlinearities; however, the number of errors for correlator number 1 is too large to be adequately corrected by this smoother. The nonlinearly smoothed pitch contour also shows that the only gross pitch period errors which were not corrected by the smoother were those that occurred near an unvoiced boundary. As already mentioned, these gross pitch period errors were often changed into voiced-to-unvoiced errors in the smoothed pitch contour.

TABLE V
TOTAL ERROR STATISTICS FOR TEN CORRELATORS

		Utterance												
		C108T	F107M	F108M	F208T	M107T	M107W	M207W	M208M	M208T	LM05T	LM07W	LM08M	LM08W
Correlation Number	1	6	1	2	4	8	2	4	25	36	40	20	21	27
	2	14	9	11	6	11	3	6	18	26	15	16	18	19
	3	11	5	4	7	8	3	5	16	28	12	15	11	16
	4	6	1	3	4	6	2	3	22	30	24	19	14	22
	5	14	6	3	4	6	2	3	16	29	18	22	19	23
	6	11	5	4	4	6	2	4	16	29	15	16	16	17
	7	8	2	4	5	8	3	4	25	34	27	15	13	20
	8	8	2	5	7	11	3	10	33	40	35	19	23	26
	9	9	1	6	8	10	3	7	20	32	18	14	19	21
	10	14	5	5	6	9	2	7	21	32	17	17	15	20

Total Number of Pitch Errors - Unsmoothed

Correlation Number	1	2	1	0	3	5	2	1	24	30	24	18	13	19
	2	3	0	1	4	9	2	1	15	15	13	11	9	5
	3	3	0	0	4	8	2	1	10	17	9	10	4	3
	4	2	1	0	4	4	1	2	11	16	10	19	10	12
	5	6	1	0	3	5	1	2	10	9	15	18	6	5
	6	2	0	0	3	6	1	2	12	13	11	12	10	3
	7	2	1	0	4	5	2	1	10	14	15	17	5	5
	8	2	1	1	5	8	2	2	19	27	16	12	12	13
	9	3	0	1	6	9	2	2	11	21	13	9	11	7
	10	2	0	0	4	9	1	3	14	14	11	12	12	7

Total Number of Pitch Errors - Smoothed

IV. Adaptive Frame Size for Pitch Analysis

One of the remaining problems in designing an effective correlation pitch detector is to implement an algorithm for making the analysis frame size variable. It is important to note that the variability of frame size for a given speaker is not nearly as important as the variability of frame size from speaker to speaker. The most important feature of the analysis frame size is that it be large enough to encompass at least two complete pitch periods, but not so large that it encompasses a large number of pitch periods. If we consider the range of pitch period variation across speakers [2], then a frame size on the order of 40 samples (4 ms) is required for a high pitched speaker, and a frame size on the order of 400 samples (40 ms) is required for a low pitched speaker. Thus, a single fixed frame size will not be suitable for all speakers.

The question now remains as to a suitable method of adapting the frame size to the pitch of the speaker. We have already argued that adaptation to the detailed pitch variation within an utterance is generally unnecessary—mainly because the range of pitch variation within an utterance is generally 1 octave or less (a factor of 2 to 1) from the average pitch for the utterance. Thus, an instantaneously adapting algorithm for choosing the analysis frame size is not required. This is fortunate in that instantaneously adaptive methods generally do not work well when the pitch estimates include gross pitch errors.

In lieu of an instantaneously adaptive method, a simple but effective method of adapting the frame size is to estimate the average pitch $\overline{P}(m)$ of the speaker using the relation

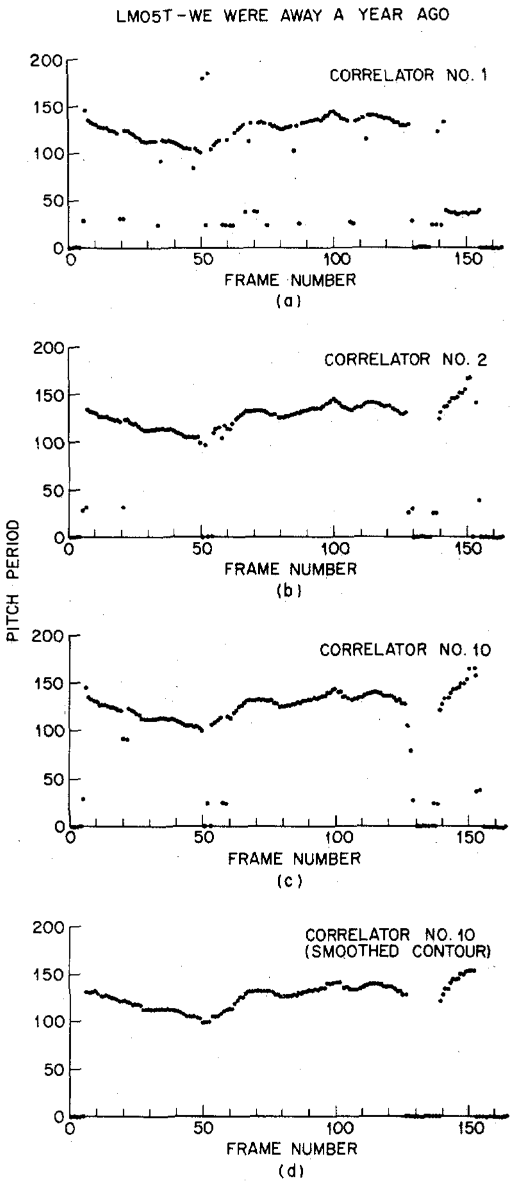
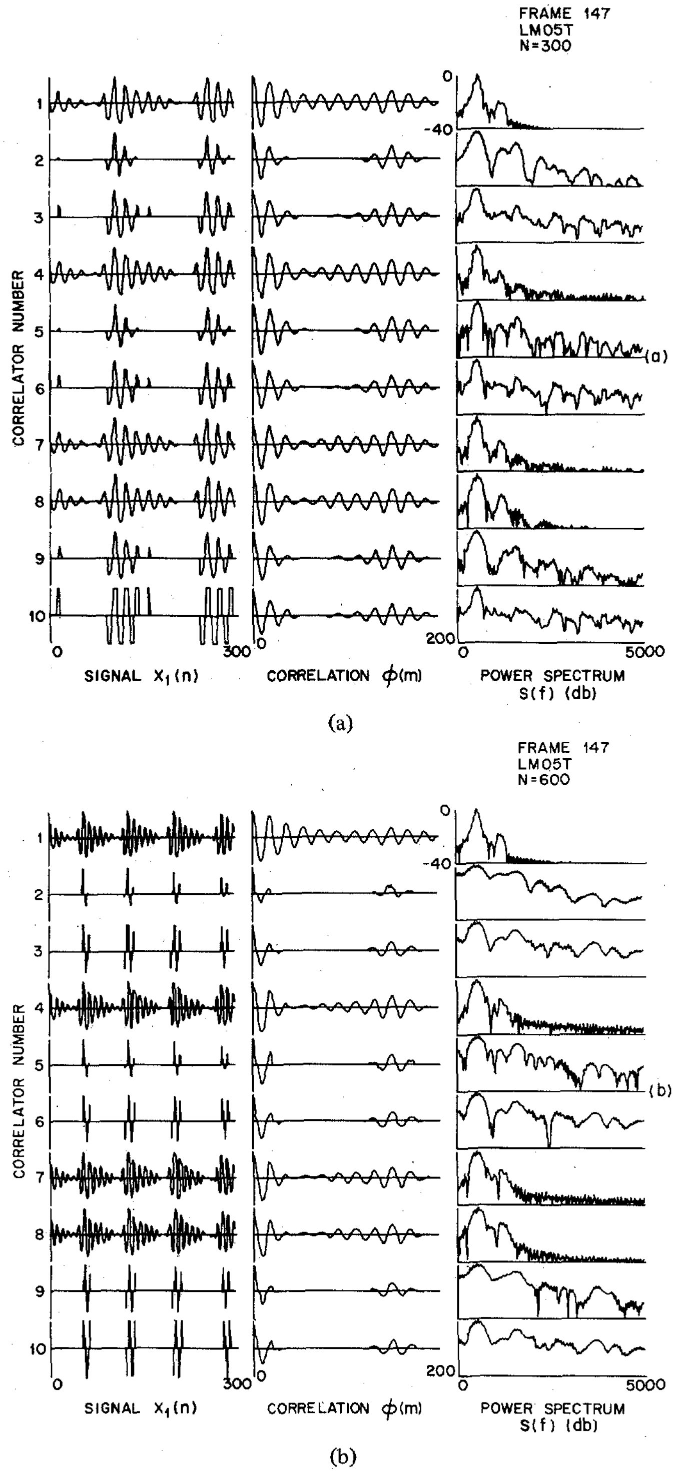


Fig. 8. The pitch period contours for the utterance LMO5T from three of the correlators of Table I and a nonlinearly smoothed pitch contour from correlator number 10.

$$\bar{P}(m) = \frac{1}{N_m} \sum_{i=1}^{N_m} p(i), \qquad N_m \geqslant 10$$
$$= 100, \qquad N_m < 10 \tag{10}$$

where $p(i)$ is the pitch period of the ith voiced frame (i.e., unvoiced frames are not used in the computation), and N_m is the number of voiced frames up to the mth frame.[12] The initial condition $\bar{P}(m) = 100$ for $N_m < 10$ is used to ensure a reasonable "average" pitch period estimate until a sufficient number of voiced frames have been estimated. The frame length $L(m)$ is generated from the simple rule

$$L(m) = 3 \cdot \bar{P}(m). \tag{11}$$



(a)

(b)

Fig. 9. The signal $x_1(n)$, the resulting correlation and power spectrum for each of the 10 correlators of Table I for both a 300 and a 600 sample analysis frame.

The factor of 3 allows up to a 50 percent variation in pitch period from the estimated average pitch period, and still ensures that at least two complete pitch periods are contained within each analysis frame. To prevent the analysis frame from getting too small, or too large, $L(m)$ is restricted to the range

$$100 \leqslant L(m) \leqslant 600, \qquad \text{for all } m. \tag{12}$$

[12] For continuous text, equation 10 should be modified so that $\bar{P}(m)$ is computed over a fixed number of past voiced frames.

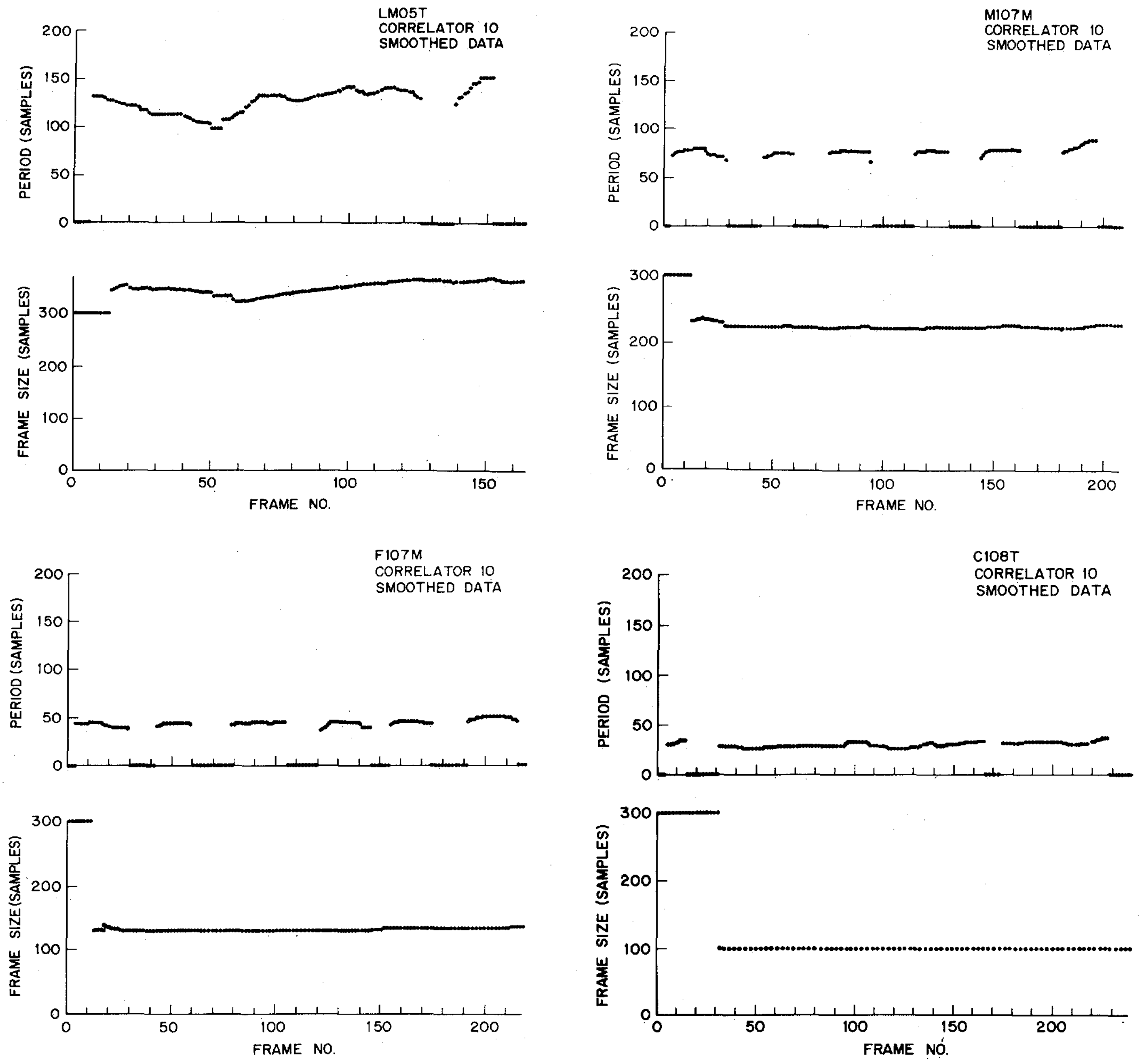


Fig. 10. Plots of the pitch contour and the resulting adaptive frame size for four typical utterances.

To demonstrate the necessity and effectiveness of matching the analysis frame size to the speaker's average pitch, Fig. 9 shows plots of the waveforms, correlation functions, and power spectra for a section of voiced speech from a low pitched male. The pitch during this section was about 150 samples. Fig. 9(a) shows the results for the 10 correlators for an analysis size of 300 samples, Fig. 9(b) shows the results for a 600 sample analysis frame size. By comparing the flatness of the power spectrum for the best correlators (i.e., numbers 2, 3, and 10) it can be readily seen that the longer analysis frame size leads to significantly flatter spectra.

The analysis frame adaptation algorithm discussed above was tested on several utterances used in the study of [2]. Fig. 10 shows plots of both the nonlinearly smoothed pitch period contour, and the analysis frame size as obtained from (11) and (12). Fig. 10(a) shows the results on a low pitched male whose average pitch period was about 140 samples. As discussed above the first 10 voiced frames used a 300 sample frame; after that the frame size adapted slowly to the pitch period, reaching a fairly constant value of about 420 samples. Fig. 10(b) shows the results for a normal pitched male speaker with very little pitch variation throughout the utterance. The algorithm very rapidly converges to an analysis frame size of about 210 samples for this speaker. Fig. 10(c) shows the results for a female speaker. In this case the analysis frame size quickly converged to a length of about 135 samples.

Finally, Fig. 10(d) shows the results for a high pitched child. In this case the frame size reached the lower limit of a 100 sample frame size at the first iteration, and remained at that value throughout the utterance.

Adapting the frame size to the estimated average pitch of the speaker can have advantages other than the ones discussed above. In cases where the resulting frame size is smaller than 300 samples, the computation of the correlation function is speeded up. In cases where the frame size falls below 200 samples, the computation is speeded up even more because fewer than 200 correlations need to be computed. Thus, for example, for a frame size of 300 samples, on the order of $N_1 = 300 \times 200 = 60\,000$ operations (multiply, addition) need to be performed to compute 200 autocorrelation points, whereas for a frame size of 100 samples, on the order of $N_2 = 100 \times 100 = 10\,000$ operations are required providing a 6 to 1 savings in computation. However, in cases where the frame size exceeds 300 samples, the correlation computation time increases, but this increase in computation time is unavoidable if one is to use the proper frame size.

V. Summary

In this paper we have examined several methods for combining nonlinear processing of the speech waveform with a standard correlation analysis to give correlation functions which have sharp peaks at the pitch period. We have shown that the nonlinearities provide some degree of spectral flattening, thereby enhancing the periodicity peaks in the correlation function, and reducing the correlation peaks due to the formant structure of the waveform. A formal evaluation of ten types of nonlinear correlation showed that correlations involving the unprocessed signal were somewhat inferior to correlations involving the nonlinearly processed signal; however, almost all the nonlinearities provided essentially the same performance.

In addition a simple procedure for adapting the analysis frame size of the correlation to the estimated average pitch period of the speaker was proposed and evaluated for several utterances. By basing the adaptation on a running estimate of the pitch period, it was shown that a fairly reliable and robust method of adapting analysis frame size resulted. This method should be appropriate for any frame-by-frame speech analysis system in which pitch is extracted.

References

[1] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-24, pp. 2–8, Feb. 1976.

[2] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-24, pp. 399–418, Oct. 1976.

[3] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262–266, June 1968.

[4] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.

[5] B. S. Atal, unpublished work.

[6] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A semi-automatic pitch detector (SAPD)," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-23, pp. 570–574, Dec. 1975.

[7] 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.