PITCH DETERMINATION OF HUMAN SPEECH BY THE HARMONIC PRODUCT SPECTRUM, THE HARMONIC SUM SPECTRUM, AND A MAXIMUM LIKELIHOOD ESTIMATE

A. Michael Noll
Bell Telephone Laboratories, Inc.
Murray Hill, N. J.

This paper describes three pitch determination methods which were especially designed for noisy speech. The harmonic sum spectrum and harmonic product spectrum developed from intuitive considerations of the harmonic structure of the spectrum of a periodic signal while the maximum likelihood estimate was the result of a purely mathematical treatment of estimating the period of a periodic signal.

The three methods were individually simulated on a digital computer, and the results were compared with the results of a computer simulation of cepstrum pitch determination. All of the new methods performed somewhat better than the cepstrum for very noisy signals, but the cepstrum gave more "accurate" pitch data for non-noisy speech.

INTRODUCTION

During the last few years, there has been considerable interest in talker identification, speech recognition, and voice-privacy systems. Nearly all of the schemes and devices used in such work require accurate information about the fundamental frequency or pitch of the speech signals being analyzed. Also, many years ago there was considerable interest in pitch determination, but this work was primarily concerned with application to speech-bandwidth compression devices called vocoders. The pitch detectors used in these early vocoders were very crude, but they have now become quite sophisticated to such an extent that the vocoder analyzer might almost be regarded primarily as a pitch detector with some less-complicated devices for extracting other information from the speech. Although the virtually unlimited bandwidths that seem to be coming in the future might reduce the usefulness of vocoders as bandwidth-compression devices for long-distance telephony, the vocoder and its associated concepts will still find application in talker identification, speech recognition, processing of helium speech, and voice-privacy systems for band-limited channels. Hence the quest for the "perfect" pitch detector is continuing. This paper reports on some new methods which were reasonably good in their pitch determination abilities for noisy speech. The techniques might also find application in areas other than pitch determination.

Presented at the Symposium on Computer Processing in Communications
Polytechnic Institute of Brooklyn, April 8-10, 1969.

THE PITCH-DETERMINATION PROBLEM

Voiced speech sounds are produced when the quasiperiodic puffs of air produced by the vocal cords excite the resonant modes of the vocal tract. The vocal tract is time varying and can be represented as a time-varying linear filter with a quasiperiodic signal with time-varying period as input. Some simplifying assumptions are possible since the impulse response of the vocal tract does not vary much with respect to time over a few pitch periods, and similarly the period can be assumed constant over a few pitch periods. Actually some deliberate smoothing of the period-to-period fluctuations of the pitch period is perceptually not noticeable, and an average value of the pitch period over a few periods is quite acceptable [12]. Thus, the quasistationary representation of the speech production system becomes as shown in Figure 1. The periodic puffs of air or excitation signal $e(t)$ with period τ_0 is the input to the vocal tract or linear filter with impulse response $h(t)$. The resonant modes of the vocal tract become peaks in the transfer function $H(\omega)$; these resonances are commonly called formants. The pitch-determination problem then is given the speech signal $s(t)$ to determine the period τ of the excitation signal.

Determining the fundamental frequency of a periodic signal should be a relatively simple problem. Unfortunately this is not true for speech signals for a number of reasons. The pitch frequencies can cover a range of from 80 Hz for male speech to 400 Hz for female speech. These frequencies can vary considerably during a single speech utterance for a single person; in fact, sudden doubling of the pitch period is not at all uncommon. If the pitch detector is to work with telephone-quality speech, then it must work with signals for which the fundamental frequency might be absent. Also, phase shifts, poor signal-to-noise ratios, and nonlinear distortions must be expected. The patently simple problem of pitch determination becomes the very difficult and challenging problem of determining the pitch of a signal about which virtually no *a priori* statistics are known and which can be immersed in noise after being passed through nonlinear communication channels.

In a crude categorization, pitch detectors can be divided into two groups depending upon whether the time-domain or the frequency-domain properties of the speech signal are exploited. Many time-domain approaches fail on certain speech sounds because of unpredictable phase effects which drastically affect the actual shape of the speech waveform. Nevertheless, some ingenious methods particularly applicable to computer processing have survived the test of time [3, 11]. Both the amplitude spectrum and power spectrum are insensitive to phase so that any phase problems are obviously eliminated in those pitch detectors which work in the frequency domain [4]. Short-time cepstrum analysis has been found to perform particularly well as a pitch detector [8, 9]. Although the material to be presented in the next section of this paper is a review of previously-published material on the cepstrum, some relevant properties of speech spectra will also be developed which will help in understanding the intuitive motivation of the harmonic-sum spectrum and the harmonic-product spectrum.

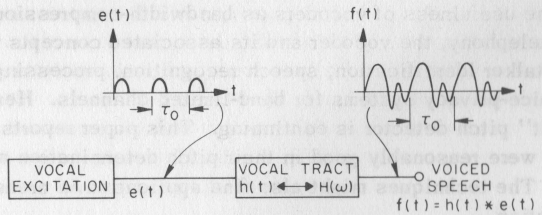


Fig. 1. Basic quasi-stationary system for the production of voiced speech sounds. The impulse response of the vocal tract and its corresponding Fourier transform are represented as $h(t)$ and $H(\omega)$, respectively.

REVIEW OF SHORT-TIME CEPSTRUM PITCH DETERMINATION

As mentioned previously, a voiced periodic speech signal will be assumed to have a constant pitch period over the analysis interval. If the periodic signal is infinitely repeated, a periodic signal with a spectrum consisting of delta functions spaced at the fundamental frequency results. The effect of a finite analysis interval with weights $w(t)$ for $|t| \leq T$ and 0 elsewhere is a convolution of the Fourier transform of $w(t)$ with the impulse train. The result is still a "periodic" spectrum; since this is frequency periodicity, Tukey would call it "repiodicity" with large repiods corresponding to low "quefrency" ripples in the spectrum [1]. This repiodicity is preserved in the power spectrum. Since the usual method of determining the period of a periodic signal is to take its Fourier transform, the Fourier transform of the power spectrum should have a peak corresponding to the repiod of the power spectrum, and this peak should occur at a quefrency equal to the period of the original signal. All of this might sound strange, but it is nothing more than an intuitive development of the autocorrelation function based upon determining repiodicities in the power spectrum.

Unfortunately, simple short-time autocorrelation fails dismally as a pitch detector for speech. The reason for this is inherent in the method of speech production as shown previously in Figure 1. The excitation signal produced by the puffs of air passing through the vocal cords is the source of the periodicity, but this signal passes through the vocal tract so that the final speech signal $f(t)$ is the convolution of the excitation $e(t)$ and the vocal-tract impulse response $h(t)$; $f(t) = e(t) * h(t)$. But then the power spectrum $|F(\omega)|^2$ of $f(t)$ equals the product of the power spectra of $e(t)$ and $h(t)$; $|F(\omega)|^2 = |E(\omega)|^2 \cdot |H(\omega)|^2$. Thus the autocorrelation function is the convolution of the inverse Fourier transforms of $|E(\omega)|^2$ and $|H(\omega)|^2$, and for certain speech signals the vocal-tract transfer function $H(\omega)$ is such that the resulting convolution gives multiple peaks. The convolution becomes, in effect, a confusion! However, if the logarithm of the power spectrum is examined, the two spectra become additive; $\log|F(\omega)| = \log|E(\omega)| + \log|H(\omega)|$. As depicted in Fig. 2, the log spectrum of the vocal-tract causes low quefrency ripples in the speech spectrum corresponding to the resonant peaks or formants of the vocal tract. The log spectrum of the excitation contributes repiodic fine structure to the speech spectrum. The power spec-

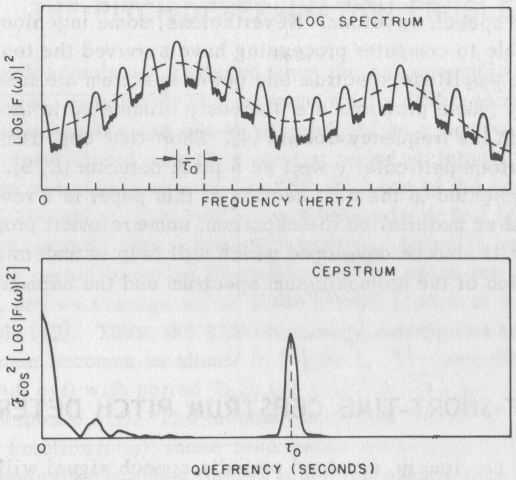


Fig. 2. Logarithm power spectrum (top) of a typical segment of voiced speech showing a spectral periodicity resulting from the original periodicity of the speech. The power spectrum of the log power spectrum, or cepstrum (bottom), therefore has a sharp peak corresponding to this spectral repiodicity. The coarse structure in the log spectrum caused by the formants of the vocal tract produce the low quefrency peaks in the cepstrum.

trum of the speech log spectrum therefore consists of a peak at a quefrency corresponding to the repiodic fine structure and low quefrency information corresponding to the formant structure in the speech log spectrum. These two effects are additive, and since they occupy different quefrency ranges they are separately distinguishable. Thus, simply by taking the logarithm of the power spectrum a new function has been derived. This new function is called the cepstrum and is formally defined as the power spectrum of the log power spectrum;

$$C(\tau) = |\mathcal{F}[\log|F(\omega)|^2]|^2 = |\mathcal{F}_{\cos}[\log|F(\omega)|^2]|^2 \quad , \qquad (1)$$

where $\mathcal{F}$ and $\mathcal{F}_{\cos}$ represent the complex Fourier transformation and Fourier cosine transformation, respectively. $|F(\omega)|^2$ is the power spectrum of the signal in the analysis interval, and the outer-most Fourier transformation becomes the Fourier cosine transform because $|F(\omega)|^2$ is an even function of frequency.

Since speech changes with time, a series of cepstra are calculated for short time segments of the speech signal usually about 40 ms or 400 samples for speech sampled at a 10 kHz rate. The analysis intervals overlap each other to adequately sample the time variation of the short-time spectra. Some computer-calculated spectra and cepstra for female speech are shown in Figure 3. A hamming time window of length 40 ms and moving in jumps of 10 ms was used. Automatic techniques for picking the cepstral peaks have been developed so that actual speech could be synthesized using cepstra-derived pitch signals. The quality of the synthesized speech was excellent thereby attesting to the superb pitch determination properties of the cepstrum [9]. Incidentally, the absence of a cepstral peak is taken to indicate that the speech segment is unvoiced so that

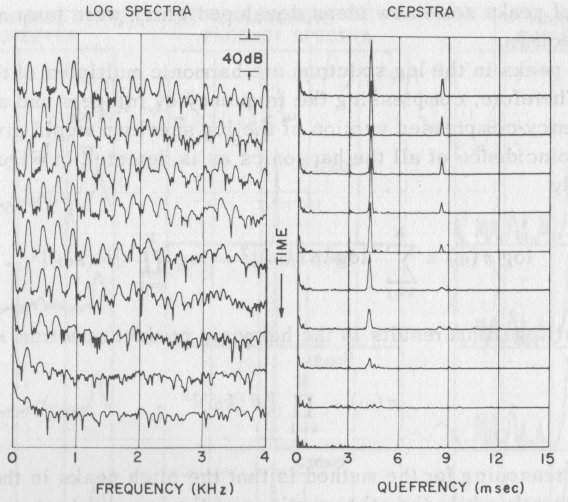


Fig. 3. Computer-calculated short-time log spectra (left) and corresponding short-time cepstra (right) for a segment of female speech recorded with a high-quality condenser microphone. For the short-time log spectra, frequency is plotted along the abscissa, log amplitude along the ordinate, and each successive plot is the log spectrum for a 40 ms speech segment. These speech segments overlap so that time progresses downwards in jumps of 10 ms. A 40 ms long hamming time window was used.

voiced-unvoiced information is obtained as a by-product of cepstrum pitch determination.

It seems that one is never satisfied; it was observed that cepstrum pitch determination failed for very noisy speech signals, and the quest for the "perfect" pitch detector started anew. Such hunts resulted in the clipstrum as an attempt to simplify hardware, and in the harmonic sum spectrum and harmonic product spectrum in an attempt to improve performance with noisy speech signals [10].

THE HARMONIC-PRODUCT SPECTRUM

Although no cepstral pitch peak was obtained for very noisy speech, in many instances it was nevertheless possible to identify visually some of the pitch peaks in the log spectra. This could be done because the shape of the harmonic peaks in the log spectra are nearly identical over a small frequency interval assuming that the magnitude and logarithm operations are linear. Since visual inspection of log spectra of noisy speech indicated some similarity of the pitch peaks, matched "filtering" (really matched "liftering" using Tukey's terminology) of the log spectra might sharpen these peaks. This matched filter concept led to the idea of calculating the mean square difference between the expected peak shape and the actual speech spectrum as a function of frequency. This mean square difference was appropriately called the mean-square-error spectrum and was worthless for pitch determination, but while attempting to determine the

coincidence of peaks some new ideas developed which were later applied directly to the log spectrum.

The pitch peaks in the log spectrum are harmonic multiples of the fundamental frequency. Therefore, compressing the frequency by integers and adding the resulting frequency-compressed version of the log spectrum would give a large peak at the coincidence of all the harmonics as is depicted in Figure 4. Stated mathematically

$$\log \pi(\omega) \equiv \sum_{k=1}^{K} \log |F(k\omega)|^2 = \log \prod_{k=1}^{K} |F(k\omega)|^2 \quad . \tag{2}$$

Taking the antilogarithm results in the harmonic product spectrum $\pi(\omega)$

$$\pi(\omega) \equiv \prod_{k=1}^{K} |F(k\omega)|^2 \quad . \tag{3}$$

The intuitive reasoning for the method is that the pitch peaks in the log spectrum add coherently while the other portions of the log spectrum are uncorrelated and add noncoherently. The frequency compression results in a sharper final peak as is depicted in Figure 4.

Since the low-quefrency structure of the log spectrum of speech only adds confusion when compressed to form the harmonic product spectrum, these low quefrencies are removed. This is done by smoothing the spectrum with a ham-

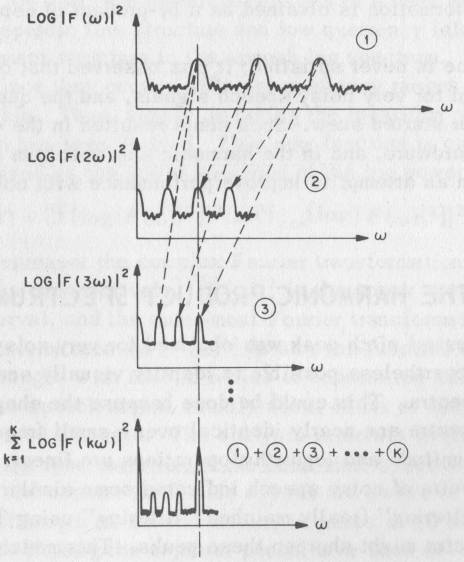


Fig. 4. Intuitive development of the harmonic product spectrum as the antilogarithm of the sum of harmonically-compressed log spectra. The low-quefrency content of the low spectrum is liftered out to prevent too large a low-frequency peak from occurring after the summation.

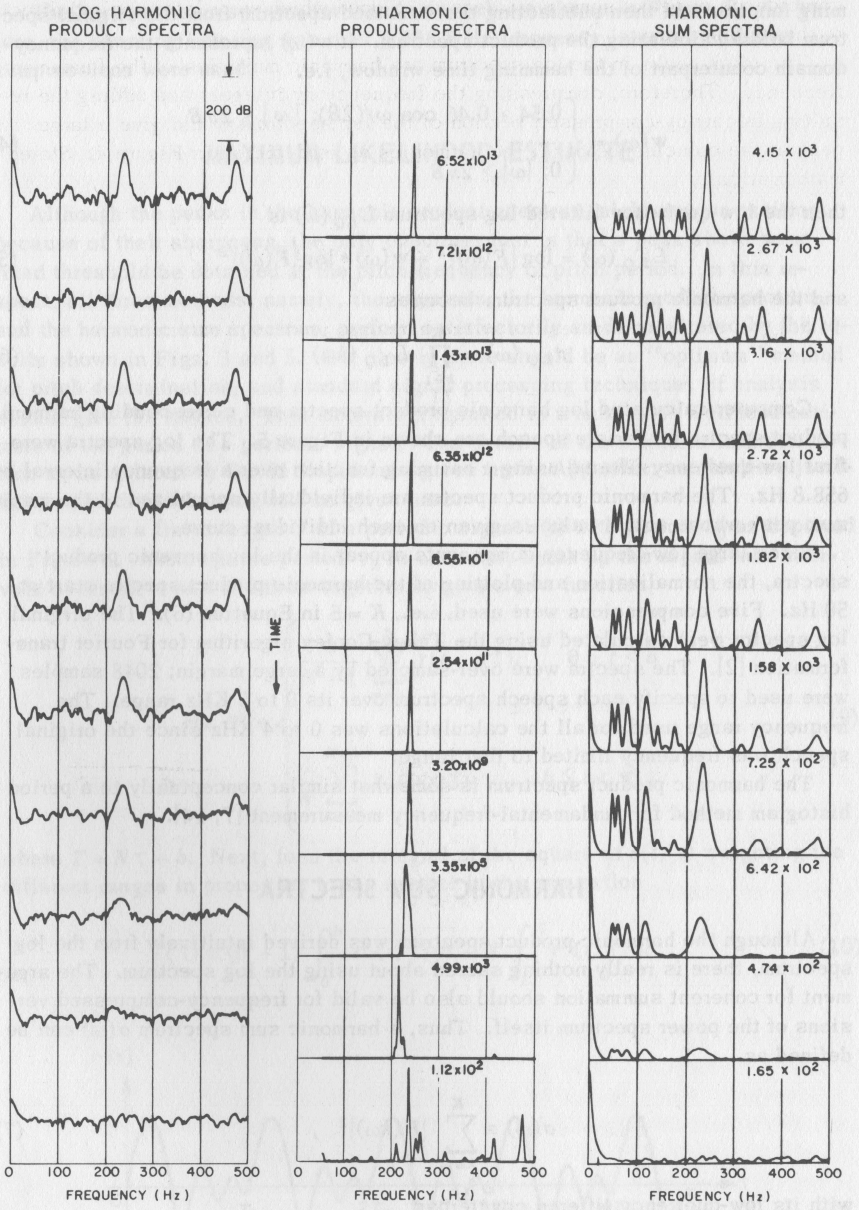


Fig. 5. Computer-calculated log harmonic product spectra (left), harmonic product spectra (center), and harmonic sum spectra (right) for the same segment of female speech used in Figure 3. The log spectra were low-quefrency liftered before calculating the harmonic product spectra and harmonic sum spectra. The harmonic product spectra and sum spectra are all normalized; the actual values of the maximum peaks are given for each individual plot. The normalization and plotting of the harmonic product spectra start at 50 Hz.

ming function and then subtracting the smoothed spectrum from the original spectrum before calculating the product spectrum. If $w(\omega)$ represents the frequency-domain counterpart of the hamming time window, i.e.

$$w(\omega) = \begin{cases} 0.54 + 0.46 \cos \omega/(2B); & |\omega| \leq 2\pi B \\ 0; & |\omega| \geq 2\pi B \end{cases} \tag{4}$$

then the low-quefrency liftered log spectrum $L_{LQ}(\omega)$ is

$$L_{LQ}(\omega) = \log |F(\omega)|^2 - w(\omega) * \log |F(\omega)|^2 \ , \tag{5}$$

and the harmonic product spectrum becomes

$$\pi_{LQ}(\omega) = \prod_{k=1}^{K} L_{LQ}(k\omega) \ . \tag{6}$$

Computer calculated log harmonic product spectra and corresponding harmonic product spectra for female speech are shown in Figure 5. The log spectra were first low-quefrency liftered using a hamming function over a frequency interval of 658.8 Hz. The harmonic product spectra are individually normalized to the maximum peak whose actual value is given on each individual curve.

Since large low-frequency components appear in the log harmonic product spectra, the normalization and plotting of the harmonic product spectra start at 50 Hz. Five compressions were used, i.e., $K = 5$ in Equation (6). The original log spectra were calculated using the Tukey-Cooley algorithm for Fourier transformation [2]. The spectra were over-sampled by a large margin; 2048 samples were used to specify each speech spectrum over its 0 to 5 KHz range. The frequency range used for all the calculations was 0 to 4 KHz since the original speech was frequency limited to that range.

The harmonic product spectrum is somewhat similar conceptually to a period histogram method for fundamental-frequency measurement [7, 14].

HARMONIC SUM SPECTRA

Although the harmonic product spectrum was derived intuitively from the log spectrum, there is really nothing sacred about using the log spectrum. The argument for coherent summation should also be valid for frequency-compressed versions of the power spectrum itself. Thus, a harmonic sum spectrum $\sigma(\omega)$ can be defined as

$$\sigma(\omega) \equiv \sum_{k=1}^{K} |F(k\omega)|^2 \ , \tag{7}$$

with its low-quefrency liftered counterpart

$$\sigma_{LQ}(\omega) \equiv \sum_{k=1}^{K} \log^{-1} [L_{LQ}(k\omega)] \ , \tag{8}$$

where $L_{LQ}(k\omega)$ is given by Equation (5).

The harmonic sum spectra for female speech are shown in Figure 5. As before, the hamming liftering was over a 658.8 Hz frequency interval, and five compressions were used.

MAXIMUM LIKELIHOOD ESTIMATE

Although the peaks in the harmonic product spectrum might appear exciting because of their sharpness, the only important fact is that a peak above some fixed threshold be obtained at the pitch frequency or pitch period. In this respect, all three methods, namely, the cepstrum, the harmonic product spectrum, and the harmonic sum spectrum, perform satisfactorily as demonstrated by the results shown in Figs. 3 and 5. But clearly there should be an "optimum" method for pitch determination, and standard signal processing techniques of analysis should give the method. This motivation resulted in a maximum likelihood estimate of the period of a periodic signal. The details of the mathematical derivation as calculated by David Slepian are given in the Appendix; an intuitive justification of the technique will be given here.

Consider a finite length T of a periodic signal $r(t)$ with period τ_0 as depicted in Figure 6. Although the period τ_0 is unknown, break up the signal into intervals of length τ and add them together to give a new function

$$\mathbf{r}(t, \tau) \equiv \begin{cases} \dfrac{1}{N+1} \displaystyle\sum_{n=0}^{N} r(t + n\tau), & 0 \leq t < b \\ \dfrac{1}{N} \displaystyle\sum_{n=0}^{N-1} r(t + n\tau) \ , & b \leq t < \tau \end{cases} \tag{9}$$

where $T = N\tau + b$. Next, form the integral of the square of $\mathbf{r}(t, \tau)$ weighting the different ranges in proportion to the number in the summation

$$J(\tau) \equiv (N+1) \int_0^b \mathbf{r}^2(t, \tau)\, dt + N \int_b^{\tau} \mathbf{r}^2(t, \tau)\, dt \ . \tag{10}$$

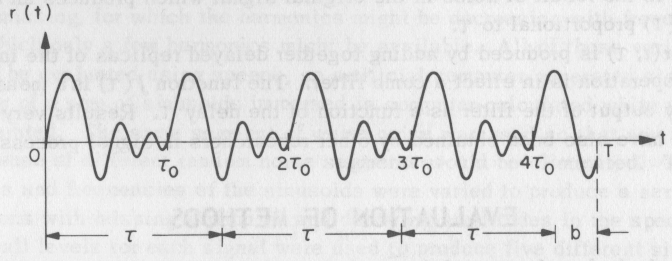


Fig. 6. Finite length T of a periodic speech signal with unknown period τ_0. The signal is broken up into N intervals of length τ so that $T = N\tau + b$.

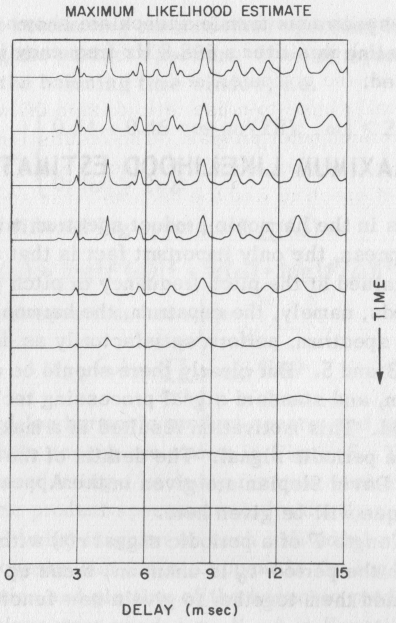


Fig. 7. Computer-calculated maximum-likelihood-estimate function $J(\tau)$ for the same segment of female speech used in Figure 3. The maximum likelihood estimate of the period is that value of delay τ which gives the maximum peak in $J(\tau)$. Unfortunately, multiple peaks with increasing amplitudes are obtained thereby greatly reducing the usefulness of the technique.

All of the small intervals should add coherently when τ is chosen to equal the pitch period τ_0 so that $J(\tau)$ would be maximum at $\tau = \tau_0$. This value of τ which maximizes $J(\tau)$ is the maximum likelihood estimate of the period.

At last, here is an intuitively appealing and mathematically rigorous approach to pitch determination. Unfortunately, there are some problems. Multiple peaks of nearly the same amplitude will occur in $J(\tau)$ thereby somewhat complicating peak picking. However, the method was simulated on a digital computer, and the results for female speech are shown in Figure 7. As expected, multiple peaks are obtained but with increasing amplitude for longer delays. This increasing amplitude is the result of noise in the original signal which produces an additive term in $J(\tau)$ proportional to τ.

Since $\mathbf{r}(t, \tau)$ is produced by adding together delayed replicas of the input signal, this operation is in effect a comb filter. The function $J(\tau)$ is a measure of the energy output of the filter as a function of the delay τ. Results very similar to Eq. (9) have also been obtained by other researchers in signal processing [6].

EVALUATION OF METHODS

One method of evaluation of the pitch-determination properties of all the methods would be to try the cepstrum, the harmonic product spectrum, the harmonic sum spectrum, and the maximum likelihood estimate on noisy speech. The energy of the sentence "High altitude jets whiz past screaming" spoken by a female was calculated on the computer, and white noise with appropriate spectral density to give an overall signal-to-noise ratio of zero dB was added by a random-number-generator computer program. The results for the same segment of female speech used previously are shown in Figs. 8 and 9 for the cepstrum, and for the harmonic product spectrum and the harmonic sum spectrum, respectively.

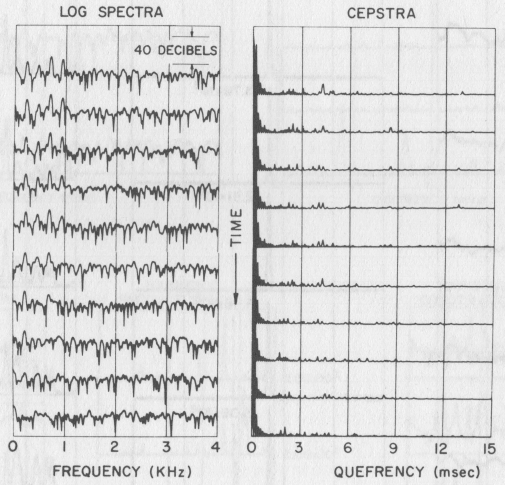


Fig. 8. Computer-calculated log spectra (left) and cepstra (right) for the same segment of female speech used in Fig. 3 but with additive white noise to give an overall signal-to-noise ratio of 0 dB.

The conclusion is that both the harmonic product spectrum and the harmonic sum spectrum are able to follow the pitch slightly better than the cepstrum during the end of voicing that occurs in the lower curves. The same female speech but with a −6 dB signal-to-noise ratio was also tested, but here all three methods were equally poor. Clearly, more evaluation was required to form any final conclusion about the relative merits of the different methods.

A realistic pitch detector must perform with signals for which the fundamental might be missing, for which the harmonics might be decreasing with frequency, and for which only a few harmonics might be available. All of these variations could not be evaluated using speech, so artificial computer-generated signals consisting of a sum of sinusoids immersed in computer-calculated white noise were generated. The same segment of white noise was used so that any fluctuations because of different random noise segments would be eliminated. The amplitudes and frequencies of the sinusoids were varied to produce a series of permutations with missing harmonics and decaying amplitudes in the spectra. Five overall levels for each signal were used to produce five different signal-to-noise ratios. Examples are shown in Figs. 10 and 11 for all the methods including the maximum likelihood estimate. The general conclusion is that the

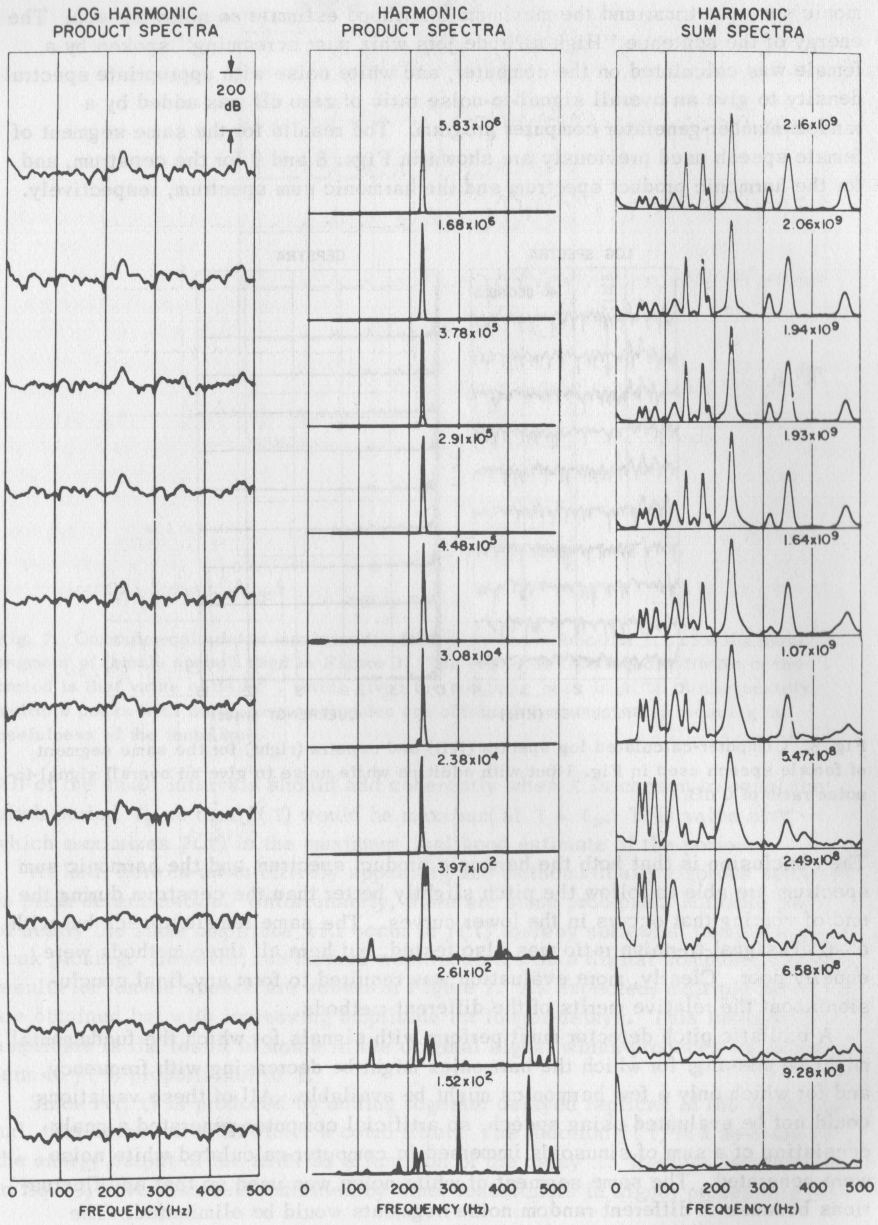


Fig. 9. Computer-calculated log harmonic product spectra, harmonic product spectra, and harmonic sum spectra with liftering for the same segment of noisy female speech used in Figure 8.

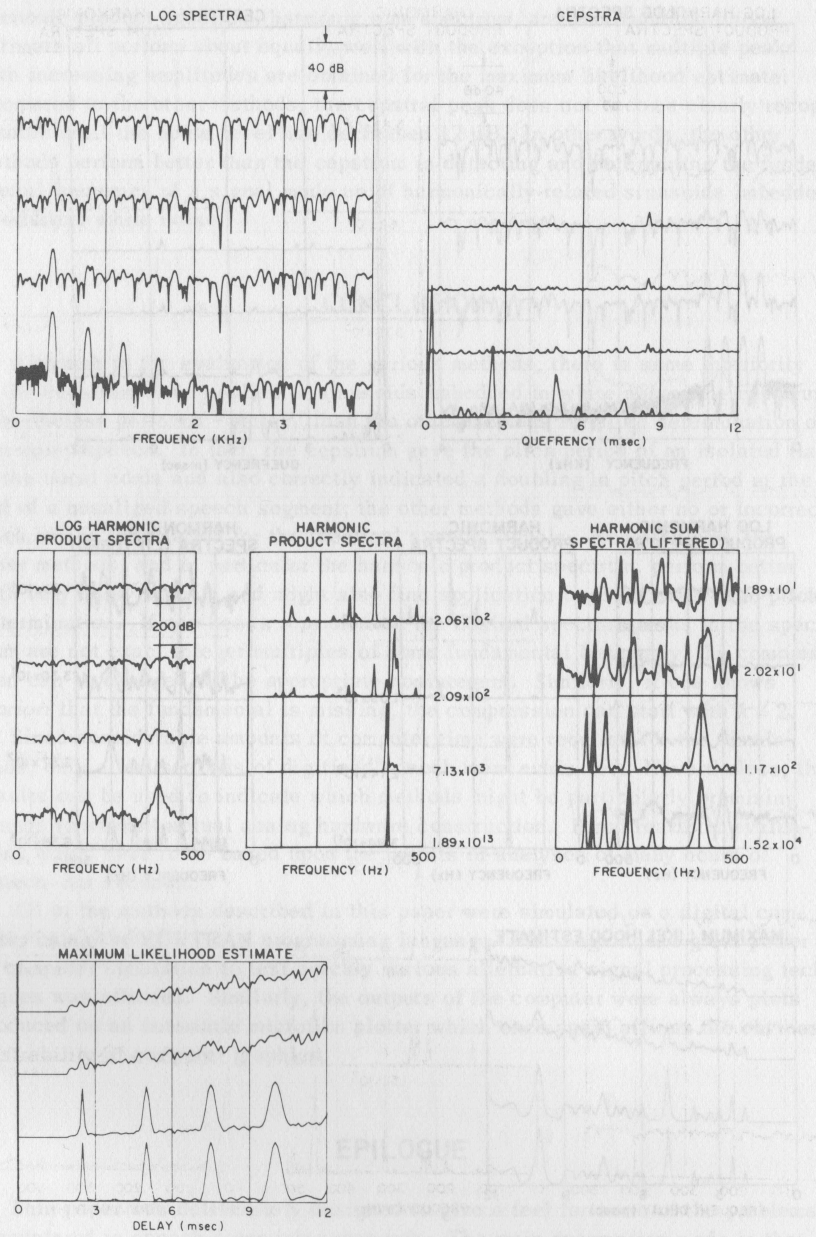


Fig. 10. Comparison of different period-determination techniques. The four signals analyzed by each technique consist of four different levels of the same segment of white noise added to five sinusoids with amplitudes decreasing at 6 dB/octave. The four noise levels are decreasing in steps of 12 dB, and the average spectral level of the noise for the top curve is approximately equal to the spectral peak of the fundamental.

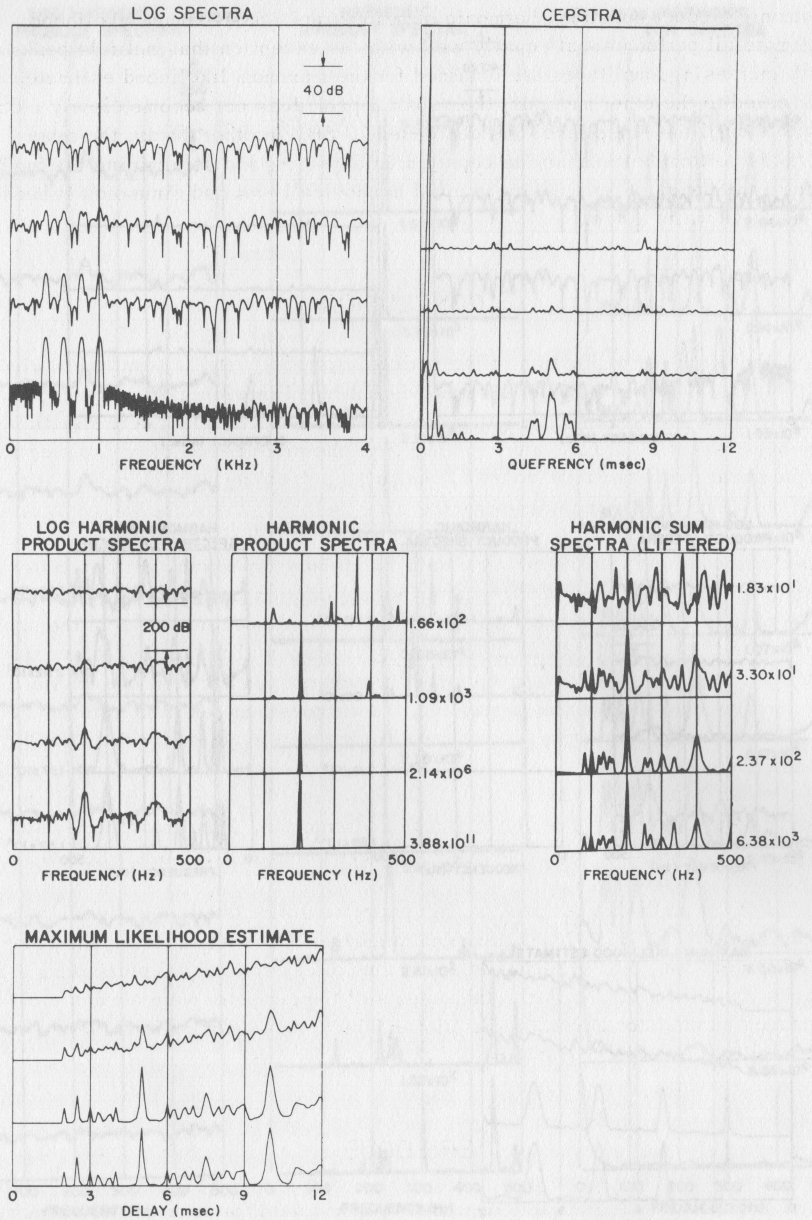


Fig. 11. Comparison of different period-determination techniques. The four signals analyzed by each technique consist of four different levels of the same segment of white noise added to four sinusoids with a missing fundamental. The four noise levels are decreasing in steps of 12 dB, and the average spectral level of the noise for the top curve is approximately equal to the spectral peaks of the harmonics.

harmonic product spectrum, harmonic sum spectrum, and maximum likelihood estimate all perform about equally well with the exception that multiple peaks with increasing amplitudes are obtained for the maximum likelihood estimate. Compared to the other methods, the cepstral peak does not become clearly recognizable until the noise level has decreased 12 dB. In other words, the other methods perform better than the cepstrum in detecting and determining the fundamental frequency of a signal made up of harmonically-related sinusoids imbedded in additive white noise.

CONCLUSION

Although in the evaluation of the various methods, there is some inferiority of the cepstrum when applied to sinusoids imbedded in white noise, the cepstrum nevertheless performs "better" than the other methods for pitch determination of non-noisy speech. In fact, the cepstrum gave the pitch period of an isolated flap of the vocal cords and also correctly indicated a doubling in pitch period at the end of a nasalized speech segment; the other methods gave either no or incorrect pitch period information for the identical speech signals. Nevertheless, the other methods, and in particular the harmonic product spectrum, perform better with very noisy signals and might also find application in areas other than pitch determination. If one knows *a priori* that the desired spectral peaks in the spectrum are not exact integer multiples of some fundamental frequency, the compression can be changed to the appropriate nonintegers. Similarly, if one knows *a priori* that the fundamental is missing, the compression can start with $k = 2$.

Since considerable amounts of computer time were required for the simulations, only a few seconds of digitized speech were evaluated. Nevertheless, the results can be used to indicate which methods might be particularly promising enough to warrant actual analog hardware construction. Final realistic evaluations would have to be based upon the results of analyses of many hours of speech—not seconds.

All of the methods described in this paper were simulated on a digital computer using the FORTRAN programming language. Once again, the great power of computer simulation to test quickly various alternative signal processing techniques was affirmed. Similarly, the outputs of the computer were always plots produced on an automatic microfilm plotter which once again affirms the obvious desirability of computer graphics.

EPILOGUE

This paper was deliberately designed to give a feel for some of the problems encountered in speech-processing research. The main assumption made in the very beginning of the paper was that the vocal tract and the pitch period were not time variant over short time intervals. The whole concept of resonant frequencies or formants of the vocal tract has been arrived at from observations of bars in speech spectrograms. However, the whole speech process is time varying so

that spectra and similar concepts lose any exact meaning. Perhaps now is the time for speech research to drop many of the age old concepts and to attempt to characterize speech as the output of a time varying system. The techniques of adaptive filtering might be particularly applicable here, and in the future perhaps less will be said about such things as speech spectra and formants.

APPENDIX: SLEPIAN'S CALCULATION OF THE MAXIMUM LIKELIHOOD ESTIMATE OF THE PERIOD

The calculations given in this Appendix are essentially similar to those given by David Slepian in an unpublished memorandum.

Maximum likelihood estimation of some signal $s(t)$ imbedded in additive Gaussian noise $n(t)$ and observed in the time interval $0 \le t \le T$ is to be applied to the estimation of the period of a periodic signal. If the Gaussian noise is white, then the maximum likelihood estimate of the signal $s(t)$ belonging to some class C of signals is that signal $\hat{s}(t) \in C$ which minimizes the integral

$$I = \int_0^T [r(t) - \hat{s}(t)]^2 \, dt \quad , \tag{A1}$$

where

$$r(t) = n(t) + s(t), \qquad s(t) \in C \quad . \tag{A2}$$

This particular result can be derived rigorously by manipulations involving the joint density for the received samples of $r(t)$. Intuitively, this result is a logical extension of the maximum likelihood estimate of some parameter of a signal as presented in Helstrom [5].

If the class C is the class of all periodic functions with period $\tau > T$, then

$$T = N\tau + b \quad . \tag{A3}$$

Thus, if τ is the period then N represents the total number of complete periods of $s(t)$ contained in the interval $0 \le t \le T$, and b is the time remaining when T is not an integer multiple of τ. Also,

$$s(t) = s(t + m\tau), \qquad m = \pm 1, \pm 2, \cdots \tag{A4}$$

Equation (A1) can be expanded

$$I = \int_0^T r^2(t)\, dt - 2\int_0^T r(t)\,\hat{s}(t)\, dt + \int_0^T \hat{s}^2(t)\, dt \quad , \tag{A5}$$

and by breaking up the range of integration because of the assumed periodicity of $s(t)$, the second integral can be written

$$\int_0^T r(t)\,\hat{s}(t)\, dt = \sum_{n=0}^{N} \int_{n\tau}^{n\tau+b} r(t)\,\hat{s}(t)\, dt + \sum_{n=0}^{N-1} \int_{n\tau+b}^{(n+1)\tau} r(t)\,\hat{s}(t)\, dt$$

$$= \sum_{n=0}^{N-1} \int_0^b r(u + n\tau)\,\hat{s}(u + n\tau)\, du + \sum_{n=0}^{N-1} \int_b^\tau r(u + n\tau)\,\hat{s}(u + n\tau)\, du$$

$$= \int_0^b \hat{s}(t) \sum_{n=0}^{N} r(t + n\tau)\, dt + \int_b^\tau \hat{s}(t) \sum_{n=0}^{N-1} r(t + n\tau)\, dt \quad . \tag{A6}$$

In a somewhat similar fashion, the third integral of (A5) becomes

$$\int_0^T \hat{s}^2(t)\, dt = (N+1)\int_0^b \hat{s}^2(t)\, dt + N\int_b^\tau \hat{s}^2(t)\, dt \quad . \tag{A7}$$

Combining all the preceding gives

$$I = \int_0^T r^2(t)\, dt + \int_0^b \left[(N+1)\hat{s}^2(t) - 2\hat{s}(t) \sum_{n=0}^{N} r(t + n\tau)\right] dt$$

$$+ \int_b^\tau \left[N\hat{s}^2(t) - 2\hat{s}(t) \sum_{n=0}^{N} r(t + n\tau)\right] dt \quad . \tag{A8}$$

This equation is to be minimized by proper choice of $\hat{s}(t)$ in $0 \le t \le \tau$, but the last two integrals involve $\hat{s}(t)$ over different intervals and both have integrands of the form $\beta s^2 - 2sh$, $\beta > 0$. They are minimized pointwise by choosing $\beta s = 2h$, or calling this optimal estimate $\hat{s}(t, \tau)$

$$\hat{s}(t, \tau) = \begin{cases} \dfrac{1}{N+1} \displaystyle\sum_{n=0}^{N} r(t + n\tau), & 0 \le t < b \\[2ex] \dfrac{1}{N} \displaystyle\sum_{n=0}^{N-1} r(t + n\tau), & b \le t < \tau \end{cases} \tag{A9}$$

$$\hat{s}(t + m\tau, \tau) = \hat{s}(t, \tau), \qquad 0 \le t < \tau, \qquad m = \pm 1, \pm 2, \cdots$$

The corresponding minimum value of I is

$$I_{\min}(\tau) = \int_0^T r^2(t)\, dt - \int_0^T \hat{s}^2(t, \tau)\, dt \quad , \tag{A10}$$

so that the second integral which depends upon τ must be maximized. Thus the maximum likelihood estimate $\hat{\tau}$ of the actual period τ_0 is that value of τ less than a given $\tau_{\max} < T$ for which

$$J(\tau) \equiv \int_0^T \hat{s}^2(t, \tau)\, dt = (N+1)\int_0^b \hat{s}^2(t, \tau)\, dt + N\int_b^\tau \hat{s}^2(t, \tau)\, dt \quad , \tag{A11}$$

achieves its maximum value and where $\hat{s}(t, \tau)$ is given by (A9).

If the restriction that τ be less than T is dropped, then obviously Eq. (A1) is minimized by choosing

$$\begin{aligned} \hat{s}(t) &= r(t), \qquad 0 \le t \le T \\ \hat{s}(t + mT) &= \hat{s}(t), \qquad 0 \le t \le T, \qquad m = \pm 1, \pm 2, \cdots \end{aligned} \tag{A12}$$

The integral I then has the value zero, and the estimate of the period is T. Therefore, the constraint $\tau < \tau_{max} < T$ is necessary to avoid this trivial solution.

ACKNOWLEDGMENTS

The maximum likelihood estimate of the period was derived by David Slepian, and the material in the Appendix closely duplicates his calculations. Manfred R. Schroeder and Man Mohan Sondhi contributed lively enthusiasm and helpful suggestions. Most of the computer programs were written by Mrs. Janice Horn Hartman. The Tukey-Cooley FFT algorithm was adapted for use at Bell Laboratories by Miss Joan E. Miller. The contributions of everyone to the work reported in this paper are most gratefully and respectfully acknowledged.

REFERENCES

[1] B. P. Bogert, M. J. R. Healy, J. W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking," M. Rosenblatt, Ed., *Proceedings of the Symposium on Time Series Analysis*, (New York, N.Y.: John Wiley & Sons, Inc., 1963), 209–243.

[2] J. W. Cooley, J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation*, **19**, 297–301, (April, 1964).

[3] B. Gold, "Computer Program for Pitch Extraction," *J. Acoust. Soc. Am.*, **34**, No. 7, 916–921, (July, 1962).

[4] Cyril M. Harris, Mark R. Weiss, "Pitch Extraction by Computer Processing of High-Resolution Fourier Analysis Data," *J. Acoust. Soc. Am.*, **35**, No. 3, 339–343, (March, 1963).

[5] Carl W. Helstrom, *Statistical Theory of Signal Detection*, (New York, N. Y.: Pergamon Press, 1960).

[6] Thomas G. Kincaid, Henry J. Scudder, III, "Estimation of Periodic Signals in Noise," *J. Acoust. Soc. Am.*, **42**, No. 5, 1166, (Nov., 1967).

[7] R. L. Miller, "Pitch Determination by Measurement of Harmonics," *J. Acoust. Soc. Am.*, **44**, No. 1, 390, (July, 1968).

[8] A. Michael Noll, "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection," *J. Acoust. Soc. Am.*, **36**, No. 2, 296–302 (Feb., 1964).

[9] A. Michael Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.*, **41**, No. 2, 293–309, (Feb., 1967).

[10] A. Michael Noll, "Clipstrum Pitch Determination," *J. Acoust. Soc. Am.*, **44**, No. 6, 1585–1591, (Dec., 1968).

[11] D. R. Reddy, "Pitch Period Determination of Speech Sounds," *Commun. ACM*, **10**, No. 6, 343–348, (June, 1967).

[12] A. E. Rosenberg, "Effect of Pitch Averaging on the Quality of Natural Vowels," *J. Acoust. Soc. Am.*, **44**, No. 6, 1592–1595, (Dec., 1968).

[13] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," *Proc. IEEE*, **54**, No. 5, 720–734, (May, 1966).

[14] M. R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement," *J. Acoust. Soc. Am.*, **43**, No. 4, 829–834, (April, 1968).