by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417-427, Oct. 1973.
[10] F. Itakura and S. Saito, "On the optimum quantization of feature parameters in the PARCOR speech synthesizer," in *Conf. Rec., 1972 Conf. Speech Communications and Processing*, Newton, Mass., Apr. 1972, pp. 434-437.
[11] J. Makhoul and R. Viswanathan, "Adaptive preprocessing for linear predictive speech compression systems," Presented at the 86th meeting of the Acoustical Society of America, Los Angeles, Oct. 30-Nov 2, 1973.
[12] N. Kitawaki and F. Itakura, "Nonlinear coding of PARCOR Coefficients," in *Meeting of the Acoustic Society of Japan* (in Japanese), Oct. 1973, pp. 449-450.
[13] A. E. Bryson, Jr. and Y. C. Ho, *Applied Optimal Control*. Waltham, Mass.: Blaisdell, 1969.
[14] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Trans. Acoust., Speech, Signal Processing*, this issue, pp. 283-296.
[15] A. H. Gray Jr. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207-217, June 1974.
[16] R. Viswanathan and J. Makhoul, "Current issues in linear predictive speech compression," in *Proc. 1974 EASCON Conf.*, Washington, D.C., Oct. 1974, pp. 577-585.

Correspondence

Pitch Extraction by Trigonometric Curve Fitting

K. STEIGLITZ, G. WINHAM, AND J. PETZINGER

Abstract—**An algorithm is proposed for extracting the pitch of voiced speech. The method is based on approximating a given segment of the speech waveform in a least-squares sense by a finite Fourier series. In the approximation the fundamental frequency of the Fourier series, as well as its coefficients, is considered variable.**

I. Least-Squares Formulation and Solution

We consider $(2M + 1)$ consecutive values of a digital speech signal $S(k)$, $k = -M, \cdots, M$. If this signal were a sample of a voiced phoneme, we would have the finite Fourier series representation

$$S(k) = \sum_{i=0}^{N} \alpha_i p_i(k, \omega) + \sum_{j=1}^{N} \beta_j q_j(k, \omega) \tag{1}$$

where $p_i(k)$ is a cosine polynomial of degree i, $q_j(k)$ is a sine polynomial of degree j, ω is the fundamental frequency of repetition, and the α_i and β_j are constants. With this hypothesis, we may estimate the α_i, β_j, and ω in a least-squares sense by minimizing

$$F(\alpha, \beta, \omega) = \sum_{k=-M}^{M} [\alpha' p(k, \omega) + \beta' q(k, \omega) - S(k)]^2 \tag{2}$$

where α, β, $p(k, \omega)$, and $q(k, \omega)$ are appropriately defined vectors.

We can now use a device of Lanczos [1] to break the function F into two pieces. Write

$$S(k) = S^e(k) + S^o(k) \tag{3}$$

where

$$S^e(k) = [S(k) + S(-k)]/2$$

$$S^o(k) = [S(k) - S(-k)]/2$$

are the even and odd parts of $S(k)$, respectively. Equation (2) then becomes

$$F(\alpha, \beta, \omega) = 2 \sum_{k=0}^{M} \epsilon_k [\alpha' p(k, \omega) - S^e(k)]^2 + 2 \sum_{k=0}^{M} [\beta' q(k, \omega) - S^o(k)]^2 \tag{4}$$

where

$$\epsilon_k = \begin{cases} 1 & k > 0 \\ \frac{1}{2} & k = 0. \end{cases} \tag{5}$$

The weighting implied by $\epsilon_0 = \frac{1}{2}$ can be incorporated into what follows, but will have a small effect and complicate the discussion. Hence, we will take $\epsilon_0 = 1$ for convenience, and consider the error

$$F'(\alpha, \beta, \omega) = \sum_{k=0}^{M} [\alpha' p(k, \omega) - S^e(k)]^2 + \sum_{k=0}^{M} [\beta' q(k, \omega) - S^o(k)]^2. \tag{6}$$

For fixed ω, we can minimize F' with respect to α and β. If we choose the p_i and q_j to be orthonormal sets of polynomials over $k = 0, \cdots, M$, the resulting function of ω can be written simply as

$$\min_{\alpha, \beta} F'(\alpha, \beta, \omega) = \sum_{k=0}^{M} [S^e(k)]^2 + \sum_{k=0}^{M} [S^o(k)]^2 - H(\omega) \tag{7}$$

where

$$H(\omega) = \sum_{i=0}^{N} \left[\sum_{k=0}^{M} S^e(k) p_i(k, \omega) \right]^2 + \sum_{j=1}^{N} \left[\sum_{k=0}^{M} S^o(k) q_j(k, \omega) \right]^2. \tag{8}$$

Manuscript received February 1, 1974; revised October 17, 1974. This work was supported in part by the U.S. Army Research Office, Durham, under Contract DAHC04-69-C-0012 and NSF Grant GJ-965.
K. Steiglitz is with the Department of Electrical Engineering, Princeton University, Princeton, N.J. 08540.
G. Winham is with the Department of Music, Princeton University, Princeton, N.J. 08540.
J. Petzinger is with Stevens Institute of Technology, Hoboken, N.J. 07030.

To summarize, we have shown that the least-squares estimate of ω in the Fourier series (1) is a value that maximizes $H(\omega)$ in (8). This function requires for each evaluation at a point ω the computation of $(2N + 1)(M + 2)$ products, where N harmonics are used in the model; assuming the orthonormal trigonometric polynomials are precomputed at the frequency used.

Orthonormal trigonometric polynomials can be generated for fixed ω by recurrence relations given by Newbery and Hunter [2], [3]. As they show, these polynomials can be generalized to arbitrary weighting (corresponding to a window on the data), and to unequally spaced data points. This recursive method of generating orthogonal polynomials [4] and trigonometric polynomials [2], [3] is well known in numerical analysis, and Newbery in fact tested a class of orthogonal trigonometric polynomials by applying them to a periodicity search.

II. A Simple Search Algorithm

We are left with the problem of maximizing the function $H(\omega)$ in (8). The algorithm described here is the simplest one for dealing with this one-dimensional problem: search over a fixed grid of points. The range of possible frequencies is sampled at L points $\omega_1, \cdots, \omega_L$, and the first estimate ω_r computed as

$$H(\omega_r) = \max_i H(\omega_i). \tag{9}$$

This is then refined to a final estimate ω' by passing a quadratic function through the points $H(\omega_{r-1})$, $H(\omega_r)$, and $H(\omega_{r+1})$, and defining ω' to be the value of ω which maximizes this quadratic. (If $r = 1$ or L, we take $\omega' = \omega_r$.)

This grid search method, while expensive computationally, has several advantages. First, it is global and will find the region of highest H even if H has many local maxima (assuming that the grid is sufficiently fine). Second, it does not rely on a previous estimate and hence its accuracy will not be affected by errors made in other frames. Third, it uses the values of the orthonormal trigonometric polynomials at a fixed set of frequency points, so that these can be precomputed and stored once and for all. Last, it is very simple to program. It should be clear that these advantages can be sacrificed with corresponding gains in computation time and space. We concentrate here on showing the effectiveness of the approach without attempting to optimize speed.

III. Implementation Details and Comments on Speed

The implementation for which experimental results are described in the next section was arranged as follows. The original speech (male speaking voice) was sampled at 15 kHz and frames of length 350 points were analyzed for pitch. The digitized signal was low-pass filtered, and every tenth sample taken, so that the effective Nyquist frequency was 750 Hz, and the frame length was $(2M + 1) = 35$, or $M = 17$. Thus the data are smoothed and decimated in a way which insures that the first formant at least is retained.

The frequency range 70 to 210 Hz (this may be too restrictive for some male speakers) was sampled at 50 points, equally spaced, and the number of harmonics chosen to be

$$N = \min\,[5, \text{Nyquist frequency/fundamental frequency}] \tag{10}$$

so that the number of harmonics in the model varied from 5 at the lowest pitch to 3 at the highest. The total number of orthonormal functions, at all frequencies, was then 492, and this should be multiplied by $(M + 2) = 19$ to yield the number of multiplications per frame, ignoring the small amount of computation required for the quadratic interpolation, and the initial generation of the orthonormal polynomials. Thus, the method required about 9348 multiplications per frame (and that many words of storage). The frames were allowed to overlap by 100 points, so that 60 frames were analyzed per second, corresponding to 560 880 multiplications per second of speech.

This amount of computation was found practical for small specimens of speech on the IBM 360/91, where the total analysis, including linear predictive coding and the low-pass filtering, ran in about 6 times real time, using Fortran H. Clearly, the computational and storage requirements need to be reduced if the method is to be practical on smaller and slower computers. One obvious way to reduce computation time is to sample the frequency grid coarsely, narrowing down the range of frequencies considered before sampling finely. Another possibility is using the previous frame's pitch estimate to narrow the search.

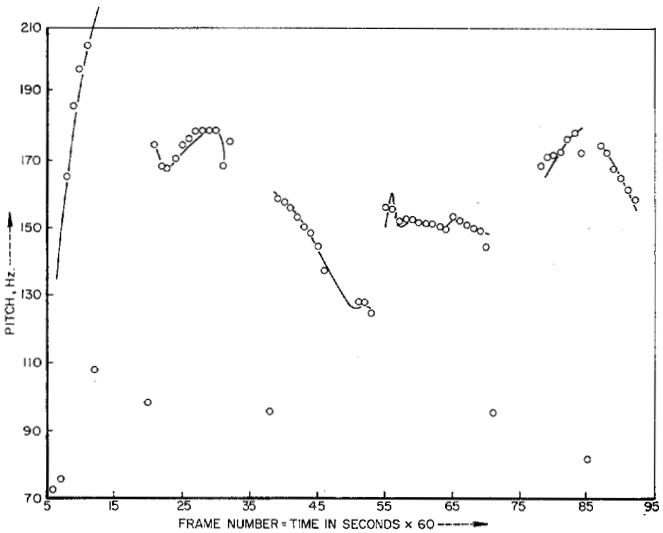

Fig. 1. Pitch contour for the utterance "This work was done twenty (thousand) · · · ." The solid curve is pitch as estimated and smoothed by eye. The open circles are unsmoothed estimates determined by the algorithm.

IV. Experimental Results

Two types of computational experiments were performed; the first type on artificially generated frames, and the second type on actual male speech. The synthetic waveforms were Fourier series of the same form as the model, with varying amounts of additive noise, and with the frequency changing linearly over a frame by varying amounts. The rms error in pitch for 88 equally spaced values of ω, and no noise or frequency modulation, was 0.35 Hz. Note that a systematic error is caused by the fact that the number of harmonics in the model changes from 5 to 4 at 150 Hz, and from 4 to 3 at 187.5 Hz. When the actual pitch is slightly below these points, the error can be as large as 2.3 Hz. This could be compensated for, but the emphasis here is on the simplicity of the algorithm. The method performed well down to an rms signal-to-noise ratio of 2, and frequency modulations of 6 Hz/frame, producing an rms error in pitch of 2.56 Hz in the extreme case of rms signal-to-noise ratio of 2 and a frequency sweep of 6 Hz/frame.

Figs. 1 and 2 show typical unsmoothed pitch contours for the one male speaker tested (GW). The speech was synthesized by analysis, using the pitch estimates from the present method, and linear predictive coding using 250-point frames, the covariance method [5] (non-Toeplitz matrix method), and no window. A fixed threshold of 2^{-7} on the normalized minimum error was used for the voiced–unvoiced decision, and Figs. 1 and 2 show only the frames that were determined in this way to be voiced. Also shown is the hand-tracked pitch, obtained by estimating every pitch period by eye and smoothing the result. It is seen that the gross errors of the algorithm occur at the onset and cessation of voicing, as would be expected by the fact that in these cases the waveform changes drastically over one frame.

V. Conclusions

A method has been proposed for extracting the pitch of voiced speech, based on best fit by trigonometric polynomials. The method has the advantages of being simple to program,

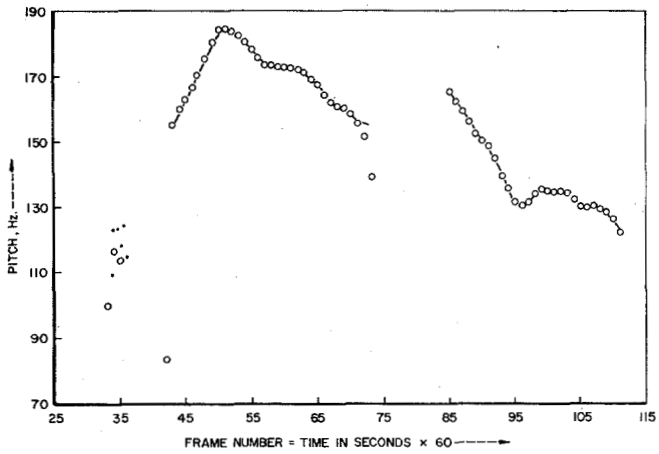

Fig. 2. Same as Fig. 1 for the utterance "The following song was (sung) · · · ." The solid dots in the first word represent eye estimates of individual pitch periods.

easy to justify in terms of maximum likelihood estimation and generally applicable to any quasi-periodic signal in the presence of noise. It has the disadvantage, however, of being computationally expensive, requiring in the simplest implementation about 0.5×10^6 multiplications per second, and about 10^4 words of table storage. However, the computation required is essentially all of the inner product type, and so the algorithm is very well suited to parallel implementation.

Further work is needed to test the method on a variety of subjects, to refine the implementation, and to explore ways of decreasing the computational requirements.

References

[1] C. Lanczos, *Applied Analysis*. Englewood Cliffs, N.J.: Prentice-Hall, 1964, ch. IV, sect. 22.
[2] A. C. R. Newbery, "Trigonometric interpolation and curve-fitting," *Math. Comput.*, vol. 24, no. 112, pp. 869–876, Oct. 1970.
[3] D. B. Hunter, "Algorithm 320: Harmonic analysis for symmetrically distributed data," *Commun. Ass. Comput. Mach.*, vol. 11, no. 2, pp. 114–115, Feb. 1968.
[4] G. E. Forsythe, "Generation and use of orthogonal polynomials for data fitting with a digital computer," *SIAM J. Appl. Math.*, vol. 5, pp. 74–88, 1957.
[5] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pt. 2, pp. 637–655, Aug. 1971.

Book Reviews

Signal Processing—J. W. R. Griffiths, P. L. Stocklin, and C. van Schooneveld, Eds. (New York and London, England: Academic, 1973, 775 pp., $42.00). *Reviewed by Albert H. Nuttall.*

This book contains the papers presented at the North Atlantic Treaty Organization Advanced Study Institute on Signal Processing, with Particular Reference to Underwater Acoustics, held at Loughborough, England, August 21–September 1, 1972. Contributors to the Proceedings were from Austria, France, Germany, Italy, The Netherlands, Norway, Canada, England, and the United States. The representatives summarized the state of the art in their fields of their respective countries, as of mid-1972, in a series of short papers (10–20 pages).

Papers of both tutorial and research character were presented in the fields of spectrum analysis, numerical processing methods, acoustics and propagation, detection and estimation, array processing, adaptive processing and normalization, and display processing. Each paper was followed by a question-and-answer period, the content of which is also presented in the book. The ample list of references afforded by most of the contributors should afford interested readers the opportunity to quickly obtain relevant past work done in foreign countries and thereby avoid some duplication of effort.

The quality of all the articles is very good; among the outstanding ones (in this reviewer's opinion) are those on time-frequency energy distributions, cepstrum analysis, Walsh functions, minimum detectable signals for spectral analyzer systems, modeling of array inputs, and a large section on adaptive array processing.

Since each paper is necessarily confined to a few pages, complete derivations of important results or techniques are naturally not presented. Interested readers must obtain the referenced articles or rederive relations themselves. Nevertheless the final results and conclusions are well documented and give an immediate indication of the state of progress and accomplishment. Considerable background is required of the reader, as these proceedings detail work on the forefront of current research (as of mid-1972). This book should be readable to graduate students and professional scientists and engineers.

The cost of this book, approximately $40, is hard to justify as an individual purchase. The central library of each institution or company could have a loan copy from which selected sections or papers could be copied. This is a text which should be carried by technical libraries.

Albert H. Nuttall (S'63–S'67–M'71) is with the Naval Underwater Systems Center, New London, Conn., where he serves as a Technical Consultant in the fields of signal processing, estimation techniques, and performance evaluation. He has made many contributions in the realm of application of statistical communication theory to practical problems.