Digital Inverse Filtering–A New Tool for Formant Trajectory Estimation

JOHN D. MARKEL
Speech Commun. Res. Lab., Inc.
Santa Barbara, Calif. 93101

Abstract

A new algorithm, based upon a digital inverse filter formulation, is presented and shown to be quite useful for estimating resonance or formant structure of voiced speech. The output of the algorithm is a set of raw data corresponding to peak frequencies versus time which is then used to estimate the first three and sometimes four continuously varying formant trajectories. Although an algorithm for automatically extracting the formants from the raw data is not presented here, for nearly 90 percent of the time an automatic decision algorithm is trivial, namely, the first three peaks of the reciprocal of the inverse filter spectrum define the first three formants.

I. Introduction

In 1966 Saito and Itakura [1] developed a new technique for time domain speech analysis based upon the maximum likelihood estimation method. A discussion in English was published in 1968 [2]. Also in 1968 Atal and Schroeder [3] published a method for linear prediction of the speech wave In 1970 Markel [4] observed that both of the basic analysis equations, independently developed, were of similar form and in fact, were derivable as special cases of Prony's method [5], [6] originally formulated in 1795 and extended to a least squares formulation at least as early as 1924 [7]. Prony's method in z-transform notation has been rediscovered at least twice within the past five years [8], [9]. From the analysis equations, moderate bit-rate speech transmission systems have been developed. The approach had, however, evidently been deemed unworkable as a tool for formant extraction. The purpose of this paper is to show that the basic analysis approach is transformable into a formant extraction algorithm and, moreover, to demonstrate that high quality formant trajectory estimation is possible even for the more difficult problems of closely spaced formants and fast transitions. The algorithm developed is linear, fast, and accurate. The raw data from which the formant trajectories are estimated consists only of a set of peak frequencies versus time. No formant amplitude or allowable formant frequency range information is necessary. On the average, for approximately 90 percent of the analysis frames, the first three formants can be uniquely defined as the first three peaks in the reciprocal of the inverse filter spectrum.

The approach relies heavily upon consideration of a frequency domain point of view, namely, inverse filtering for determining the necessary analysis conditions and parameter values.

Manuscript received August 1971.

This work was supported by the Office of Naval Research under Contract N00014-67-0118

This paper is based largely upon a more detailed description in "Formant trajectory estimation from a linear least-squares inverse filter formulation," SCRL Monograph 7, Speech Commun. Res. Lab., Inc., Santa Barbara, Calif. 93101.

II. Derivation of the Digital Inverse Filter

A digital filter of the form

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}$$

is assumed where $M+1$ defines the filter length. Given an input sequence $\{x_n\}$ of length N we wish to determine the coefficients $\{a_i\}$ such that the total energy as measured at the filter output is minimized. Within a constant factor this formulation is equivalent to the problem of transforming an input sequence into the best least squares estimate of a unit pulse through the filter [10], [11]. Thus, the problem is essentially one of digital Wiener filtering [12].

The form of $A(z)$ is important in our development due to the fact that we are interested only in estimating the resonance structure of spectral data. Defining the leading term as unity instead of some arbitrary a_o has the following effects. 1) If the leading term is arbitrary, the error function must be the difference between a constant and the filter output instead of just the filter output—otherwise the optimum filter would be defined by setting $a_i=0$, $i=0, 1, \cdots, M$; and 2) the order of the equations which must be solved is reduced by one.

The solution to this digital inverse filter formulation is easily obtained as follows. The total energy ϵ of the output filter is calculated from the error function y_n by

$$\epsilon = \sum_{n=0}^{L} y_n^2 \tag{1}$$

where

$$y_n = x_n + \sum_{i=1}^{M} a_i x_{n-i}$$

and

$$L = N + M - 1.$$

The total energy is minimized by taking the partial derivative of ϵ with respect to a_k, $k=1, 2, \cdots, M$, setting to zero, and solving for $\{a_k\}$. The result is

$$\sum_{i=1}^{M} a_i \sum_{n=0}^{L} x_{n-i}x_{n-k} = -\sum_{n=0}^{L} x_{n-k}x_n, \quad k = 1, 2, \cdots, M. \tag{2}$$

By noting that $x_n=0$, $n<0$, and $n\geq N$, (2) can be rewritten as

$$\sum_{i=1}^{M} a_i r_{i-k} = -r_k, \qquad k = 1, 2, \cdots, M \tag{3}$$

where

$$r_k = \sum_{n=0}^{N-1-|k|} x_n x_{n+|k|}. \tag{4}$$

The results are of similar form although not identical to the classical Wiener filter equations in discrete form [13]. These equations define the nucleus of the formant trajectory estimation algorithm which will be developed.

There is certainly nothing difficult about the development of the above equations. In various related forms they have existed for at least 50 years. The nontrivial and necessarily empirical part of the procedure is in determining various relationships between mathematical constants such as N, M, and the system sampling rate and the desired characteristics of the speech wave so that a useful formant analysis technique is developed.

III. Formant Extraction and the Inverse Filter

In essence, the inverse filter attempts to transform the input signal into a constant or white noise spectrum. If $M \rightarrow \infty$ the inverse filter will theoretically predict the exact inverse of the input signal spectrum, resulting in a constant for the output or error spectrum. For a finite M, it is no longer possible to span the input signal and thus the filter can only be designed to approximate the inverse of the signal characteristics.

Our hypothesis was that if M were properly chosen it would be possible to predict the inverse of the gross spectral structure corresponding to the resonance or formant behavior while ignoring the fine spectral structure corresponding to voice fundamental frequency F_o (at least for the case where F_o is in the range of a male voice). The estimate of the resonance behavior would then be determined as the reciprocal of the inverse filter spectrum. Figs. 1 and 2 illustrate the expected results for voiced and unvoiced sounds with proper choice of analysis conditions and parameters. Shown in Fig. 1(A) and (B) is a representative spectrum and its autocorrelation sequence, respectively. With the proper choice of analysis conditions $\log |A(z)|^2$ evaluated along the unit circle gives the appearance shown in Fig. 1(C). The reciprocal of the inverse filter spectrum defines the estimate of the resonance structure of the input spectrum as shown in Fig. 1(D). The spectrum of the inverse filter output corresponding to the error in the representation is shown in Fig. 1(E). The effect of the inverse filter is to transform the input signal into the best estimate of white noise (in the least squares sense). It is seen that the result is roughly a white noise spectrum with a periodic component superimposed upon it. The autocorrelation sequence of the output shown in Fig. 1(F) is an alternate method of illustrating the effect of the inverse filter. The unit pulse at the time origin corresponds to the white noise or constant portion of the spec-

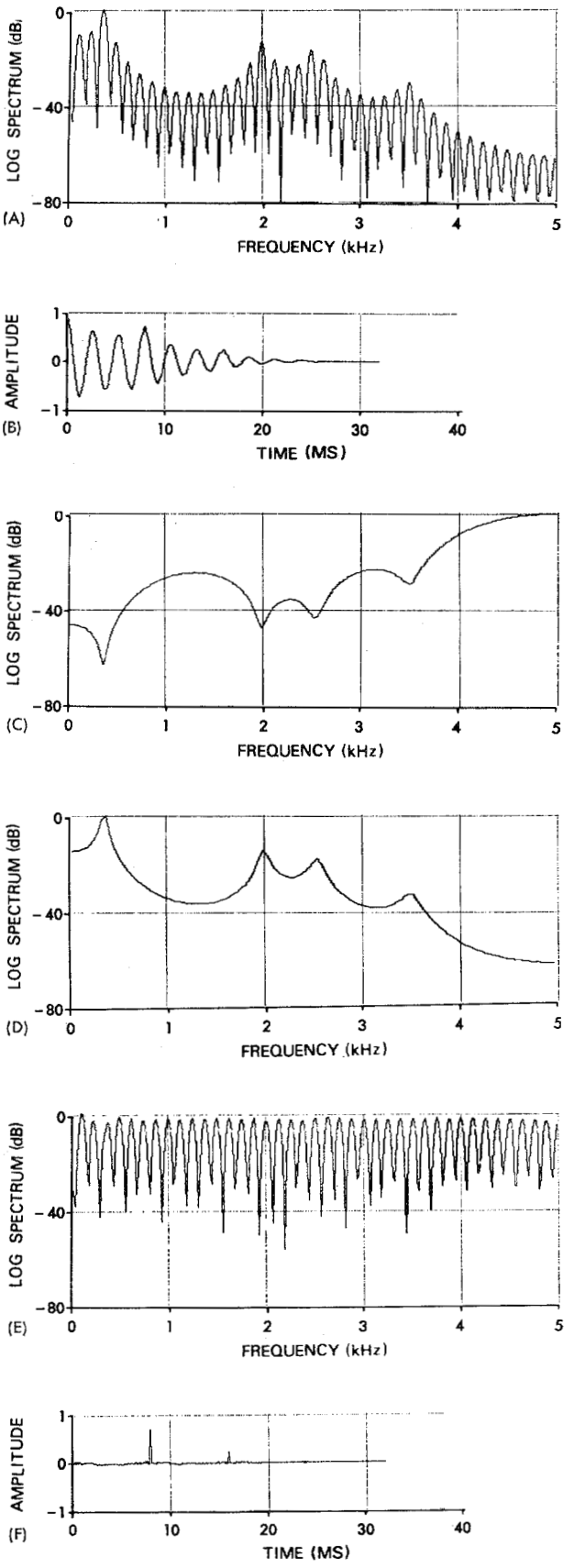

Fig. 1. Representative waveforms from the analysis of a voiced sound with proper choice of analysis conditions.

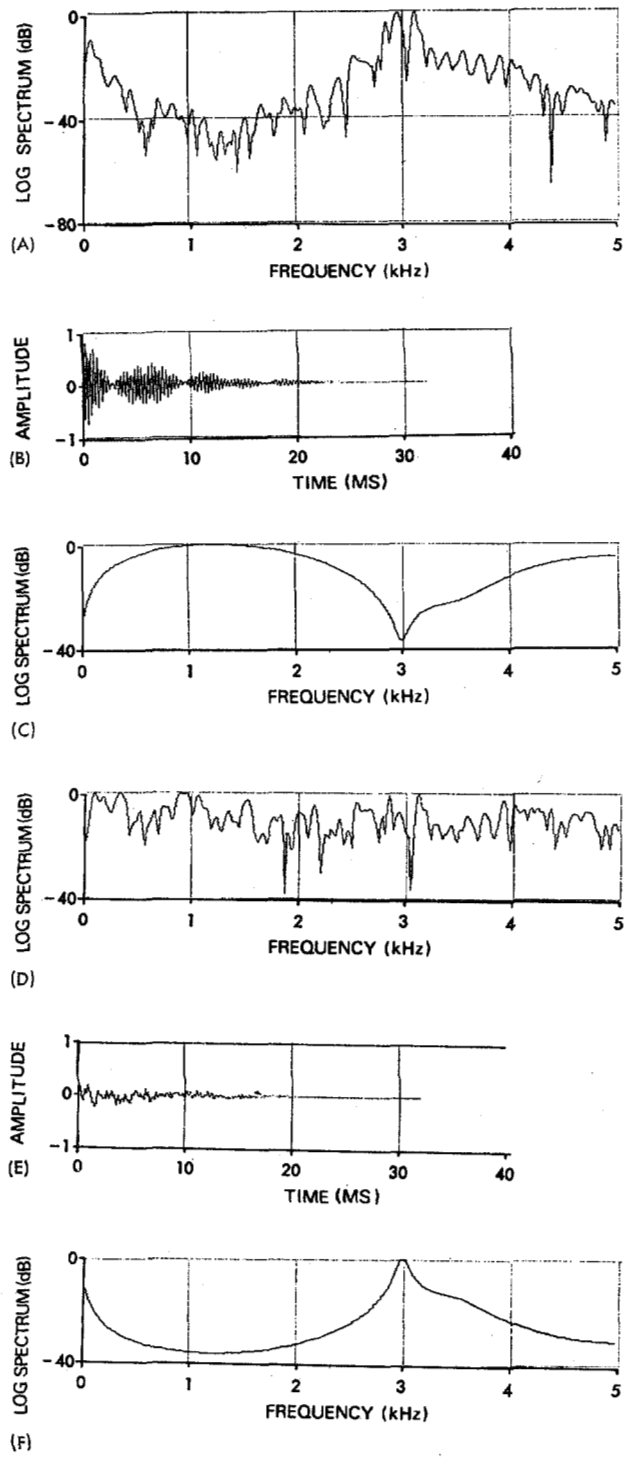

Fig. 2. Representative waveforms from the analysis of an unvoiced sound with proper choice of analysis conditions.

trum while the next highest amplitude spike and its diminished amplitude replications correspond to the periodic component of the spectrum. The application of the autocorrelation equations for obtaining high quality pitch information has previously been recognized by Itakura and Saito [2].

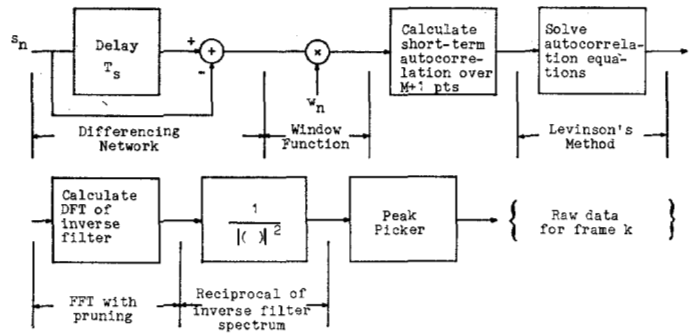

Fig. 3. Block diagram of the inverse filter algorithm for automatically extracting the raw data for use in formant trajectory estimation.

Fig. 2 presents corresponding results for the analysis of an unvoiced segment with proper analysis conditions. It can be seen that a quite reasonable estimate of the noise spectrum is obtained. The output autocorrelation sequence and spectrum illustrate the fact that the predictable (nonwhite) components of the input signal corresponding to the resonance structure have been accurately extracted.

The peak frequency locations in the spectrum shown in Fig. 1(D) (which can be obtained by simple peak peaking) define the "raw data" from which formant frequencies are estimated. For this synthetic example, the peak locations correspond precisely to the formant frequencies. The algorithm used for generating this raw data for continuous real speech will now be considered.

IV. The Inverse Filter Algorithm (IFA)

A block diagram of the IFA is shown in Fig. 3. Let s_k, $k=0, 1, \cdots$, define the sampled data representation of the continuous speech wave $s(t)$, where

$$s_k = s(t)\,|_{t=kT_s}, \qquad k = 0, 1, \cdots,$$

and T_s is the sampling period with corresponding sampling frequency $F_s=1/T_s$. Since speech is a continually time-varying process, short-term analysis of sets of contiguous data samples at some specified frame rate is needed. Each data frame is of duration T_f with corresponding frame rate $F_f=1/T_f$.

For each frame the inverse filter input x_n is defined by $x_n=w_n(s_{k+n+1}-s_{k+n})$, $n=0, 1, \cdots, N-1$ where k denotes the starting point of the particular frame being analyzed. The signal is differenced for two reasons. First, a 6 dB/octave pre-emphasis occurs which with windowing tends to emphasize the higher formants. In addition a troublesome low-frequency nonformant peak occasionally caused by a very strong fundamental frequency component is deemphasized. Secondly, if there is any bias or nonzero mean value in the frame, it will tend to cause the Fourier transform of the window function to appear in the low-frequency region of the discrete spectrum of the input sequence $\{x_n\}$. The

effect is to cause an increase in the filter length so that the peak at zero frequency along with the other resonances can be represented adequately.

If the N differenced speech samples are directly applied to the inverse filter equations, w_n is effectively defined as a rectangular window. Because of the properties of a rectangular window, zeros occur in the DFT of $\{x_n\}$ at the reciprocal of the window width. The effect of these zeros is so severe that often F_2 or F_3 can be completely disguised. A considerably more appropriate window function is the Hamming or Hanning window [14]. Somewhat arbitrarily we have chosen the Hamming window defined by

$$w_n = \begin{cases} 0.54 - 0.46 \cos(2\pi n/(N-1)), \\ \qquad\qquad n = 0, 1, \cdots, N-1 \\ 0, \quad \text{elsewhere.} \end{cases}$$

The necessity for a proper window function is shown in Fig. 4. Fig. 4(A) shows the spectrum of a 32-ms segment taken from a spoken phrase using a rectangular window. The inverse filter analysis with proper choice of conditions results in the spectral estimate shown in Fig. 4(B). In terms of extracting the apparent resonance structure of the spectrum, the analysis is reasonably good. At most, one resonance can be estimated.

However, by observing a spectrogram of the utterance from which this segment was taken, it is quite clear that there are two closely spaced low formants and two additional formants, one at about 2.2 kHz and the other at about 3.1 kHz. By applying a 320-point Hamming window to the input sequence, the spectral representation shown in Fig. 4(C) is obtained. By comparing Fig. 4(A) and (C) one can see the dramatic change in spectral character, due simply to the application of the Hamming window function.

In Fig. 4(A), F_1 is reasonably clear. However, it is not evident whether the single spike at about 0.6 kHz belongs to the first or second formant. In Fig. 4(C) it is clear that two separate peaks occur in the range (0, 1) kHz. Past 1 kHz, the effect of the rectangular window is to mask out the signal components. In Fig. 4(A), the peak at 2.2 kHz is barely perceivable, whereas it is quite distinct in Fig. 4(C). In Fig. 4(A), no resonances are observable past 3 kHz, whereas in Fig. 4(C) an additional resonance is apparent at 3.1 kHz. Since the inverse filter is designed to transform the input spectrum into a white noise or constant spectrum it should not be too surprising that the analysis result as shown in Fig. 4(D) is obtained. The four resonance frequencies can now be extracted by simple peak peaking.

The short term autocorrelation sequence $\{r_k\}$ is computed directly from (4) using the length N sequence $\{x_n\}$. Although r_k, $k=0, 1, \cdots, N-1$, is most efficiently calculated by applying two FFT's of length N (if N is highly composite), for speech analysis $M \ll N$ in general, and direct calculation of the $M+1$ autocorrelation coefficients will be somewhat faster.

Because of the special form of the autocorrelation equations it is possible to recursively solve for the unknown filter coefficients a_i, $i=1, 2, \cdots, M$, in the order of M^2 ($0(M^2)$)

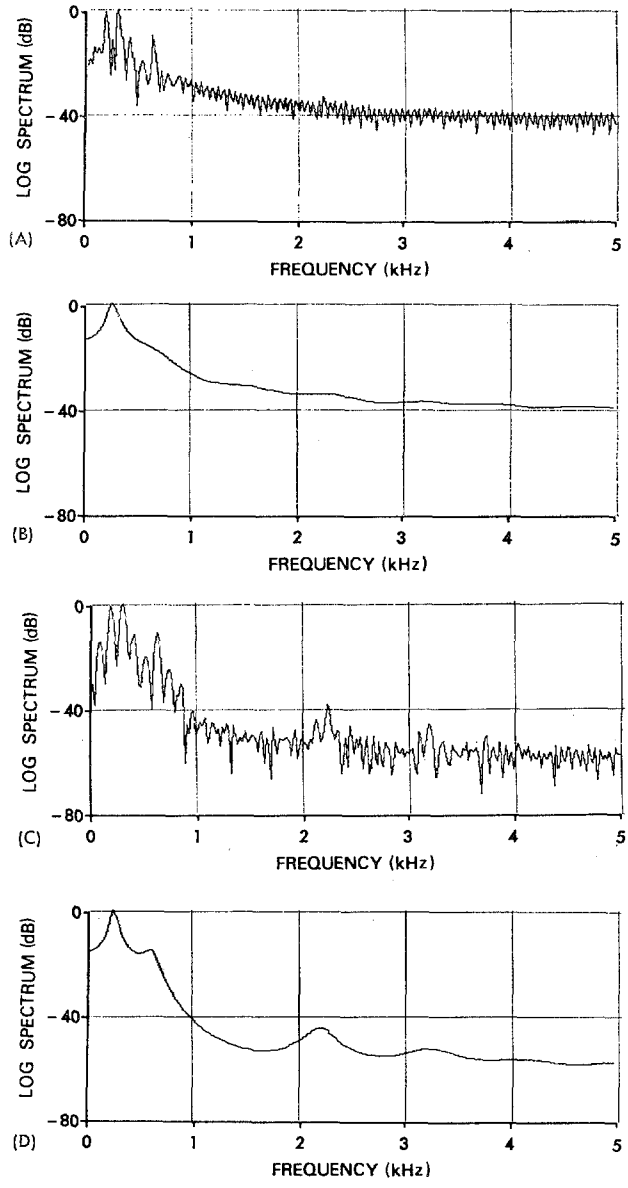

Fig. 4. Illustrations showing the effects of a rectangular window and a Hamming window on resonance estimates from the inverse filter algorithm.

operations as opposed to $0(M^3)$ operations for a general linear simultaneous equation solution. The solution was first developed by Levinson [13] and later derived in a somewhat different form using z transforms by Robinson [10].

From Robinson's development we obtain with only slight modifications the flow chart of Fig. 5 as the most efficient solution to (3). A Fortran program can be written directly from the flow chart by noting that the superscripts correspond to the iteration number (and are thus only implied) while the subscripts define the subscripted variables which must be dimensioned. With respect to (3), $a_n^{(M)} = a_n$ for $n=1, 2, \cdots, M$, as the coefficient solution set. One addi-

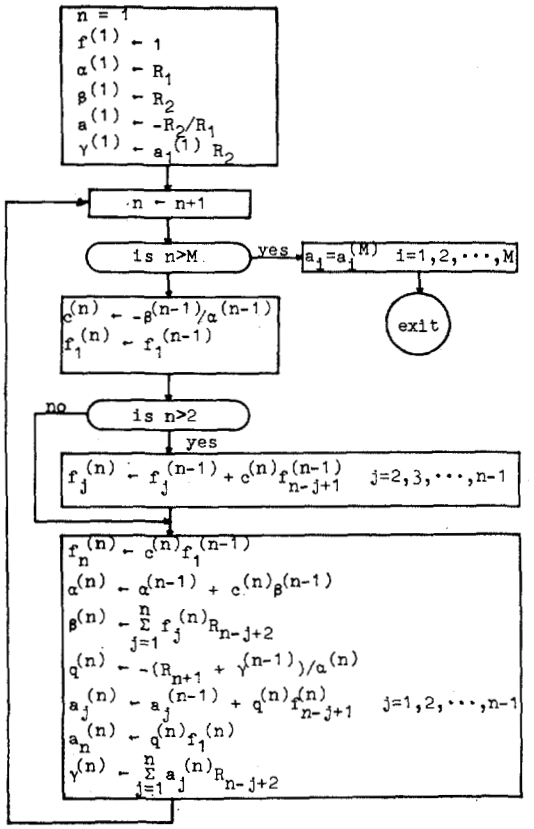

Fig. 5. Flow chart for determining the inverse filter coefficients a_i, $i=1, 2, \cdots, M$ from the short-term autocorrelation coefficients $r_i=R_{i+1}$, $i=0, 1, \cdots, M$ using Levinson's method.

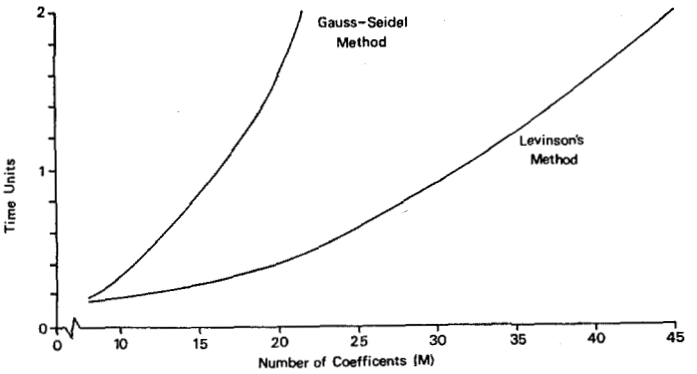

Fig. 6. Relative computational times for the $M \times M$ autocorrelation matrix inversion using Levinson's method and the general-purpose Gauss–Seidel procedure.

tional advantage of Levinson's method is that only $4M$ storage locations are required as opposed to M^2 for general linear simultaneous equation solution. The relative time saving possible from Levinson's method is shown in Fig. 6 where it is compared with the Gauss–Seidel method [15] versus M. The digital filter

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}$$

is now determined for a particular frame. Since $A(z)$ is an inverse filter, $|D(z)| = |1/A(z)|$ must define the estimate of the input spectrum.

After some experimentation using a polynomial root solving program to calculate all the resonances, it became apparent that the roots which were possible candidates for formant parameters were of narrow enough bandwidth that the useful information could also be obtained from a frequency domain representation of the inverse filter.

The discrete frequency spectrum $D(z_k)$, where $z_k = \exp(j2\pi k/N')$, $k=0, 1, \cdots, N'-1$, is most efficiently calculated by pruning the FFT. The input sequence is defined by

$$\{1, a_1, a_2, \cdots, a_M, \underbrace{0, \cdots, 0}_{N'-M-1 \text{ zeros}}\}.$$

An algorithm for pruning a radix-2 algorithm of length $2^{M'}$ where only $2^{L'}$ nonzero input values exist has recently been described [16]. With respect to a nonpruned FFT algorithm a time saving of $M'[L'+2(1-2^{-\Delta})]^{-1}-1$ where $\Delta=M'-L'$ is obtainable. If $M'=9$ and $L'=4$ (which means that $M \leq 16$), a time saving of nearly 60 percent is possible. Since the data are real, the computation time can be further reduced by one half by shuffling odd and even points into the real and imaginary arrays, applying an $N'/2$-point transform and then unshuffling.

The result of the FFT is $A_k=A(z_k)$. The magnitude spectrum corresponding to the estimate of the input spectrum is then computed from

$$|D_k|^2 = 1/[\mathrm{Re}^2(A_k) + \mathrm{Im}^2(A_k)]$$

where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary part, respectively. The magnitude spectrum is scanned and the L_k local maxima $p_k(l)$, $l=1, 2, \cdots, L_k$ in frame k are recorded. The set of all local maxima defines the raw data from which the formant trajectories are to be estimated. (If there is no interest in displaying the resonance structure it should be noted that the most direct approach is to search for the local minima of $|A_k|^2$.)

V. Considerations in the Choice of Analysis Parameters

A study was performed to determine the relationship between the choice of filter length and the accuracy of resonance estimation. For a 10-kHz sampling frequency, typical results are illustrated in Fig. 7. The spectrum of the filter input sequence $\{x_n\}$ for a 32-ms portion of the spoken vowel /I/ is shown in Fig. 7(A). The spectra of the inverse filter are shown in Fig. 7(B)–(F). For $M=6$, a very poor representation is obtained which might be expected since only three resonances are allowed for spanning the entire 5-kHz range. By inspection of Fig. 7(A) one can see that at least four formants are present. For $M=10$ the representation improves somewhat and for $M=14$ quite reasonable estimates of the formant frequencies are possible by simple peak peaking. For $M=16$ the resonances are still accurately predicted. As M is increased still further, better and better ap-

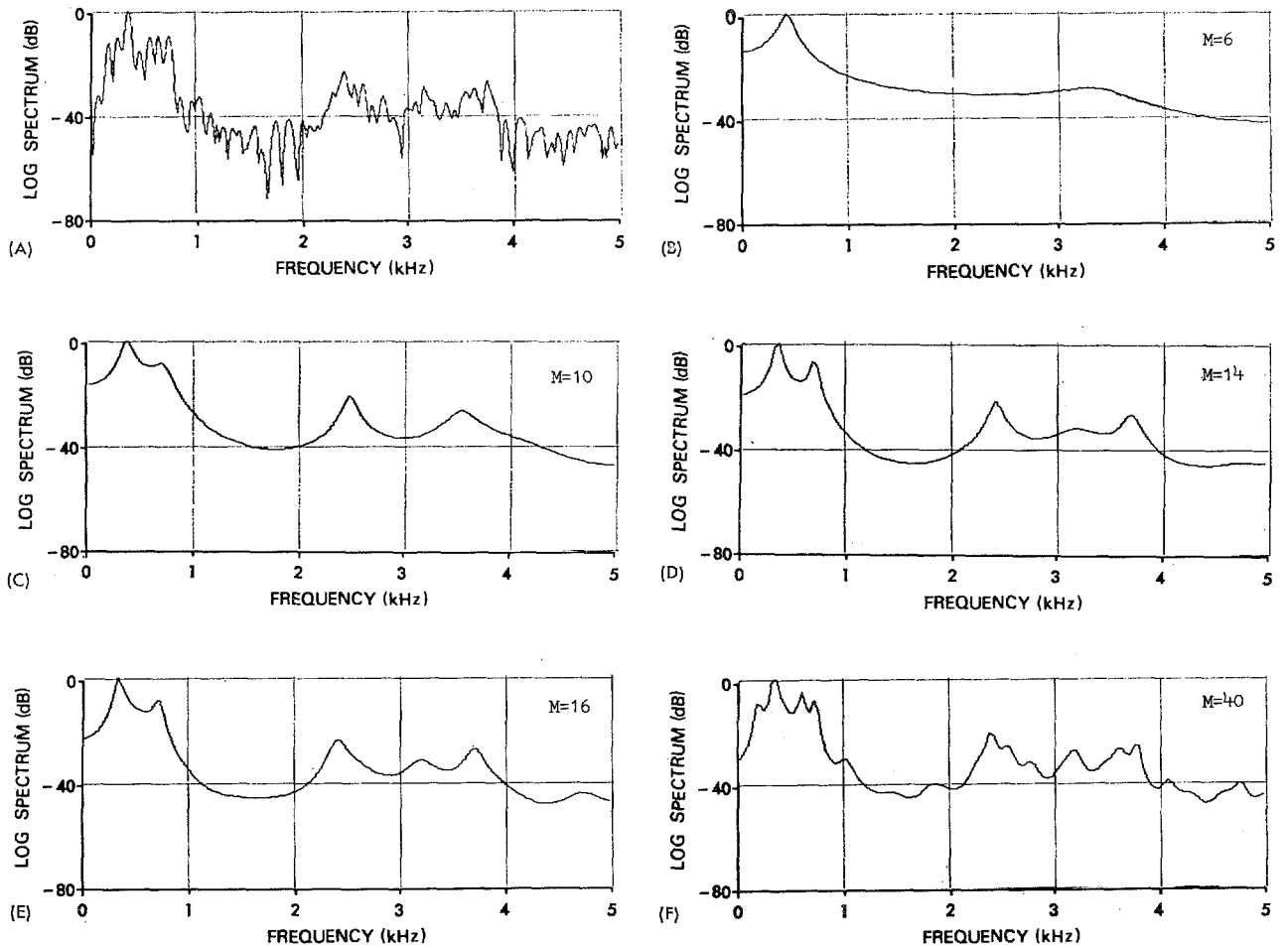

Fig. 7. Estimation of the resonance behavior of the input spectrum (shown at the top) for various filter lengths. A 10-kHz sampling rate was used for this example.

proximation to the input spectrum (as opposed to the resonance behavior) is obtained. This fact is illustrated for $M=40$ which requires the solution of a 40×40 matrix of linear simultaneous equations.

Thus an M in the neighborhood of 14 to 16 would seem the logical choice for this example. Fortuitously it has been discovered that M is not a strong function of the particular speech sound. However, it is a strong function of the system sampling rate. In fact for $6\leq F_s\leq 18$ kHz the equation $M=F_s+\gamma$ where $\gamma=4$ or 5 has been found generally sufficient for the analysis. The physical interpretation of this result is simply that independent of the sampling rate, roughly one complex pole pair is required to span approximately every 700 Hz.

The choice of analysis window length is based upon consideration of two factors: 1) frequency resolution; and 2) spectral averaging. If the window is too short it will not be possible to resolve closely spaced formant structure. On the other hand if it is too long, strong formant peaks will not exist in the input spectrum due to the inherent frequency averaging over the time interval of the window. Again, the window length has been determined to be a rather weak function of the particular speech sound but a strong function of the sampling rate. $N=\delta F_s$ where $\delta=20$ to 35 defines reasonable limits on N.

VI. Experimental Results

In this paper, results will be presented for a constant sampling frequency of 10 kHz since nearly all significant spectral structure of speech (excluding possibly some voiceless fricatives and sibilants or aspiration noise) is contained below 5 kHz. From the Peterson–Barney study [17] it can be seen that the first three formants for vowels in a CVC environment are generally below 3 kHz. The same upper limit is also observed in voiced speech and has been used in the formant analysis algorithm as developed by Schafer and Rabiner [18].

Due to its being previously chosen as representative of a difficult form of voiced speech having both close first and

second formants and close second and third formants and fast transitions, we will consider the voiced phrase "we were away" as a first example [17]. The results of the algorithm are shown in Fig. 8 for the parameter values $M=14$ and $N=256$.

The raw data are plotted on a scale of peak frequency location versus time. With the knowledge that formant trajectories must be continuous and cannot intersect, for voiced nonnasalized speech it is quite easy to estimate the first three formant trajectories by inspection. It can be seen that an occasional omission of a peak is obtained. There were no extraneous insertions (i.e., nonformant peaks) obtained for this example.

How easily the formant trajectories can be estimated is a strong function of the frame rate F_f. As F_f is increased, proportionally more data per unit time are obtained so that the effect of occasional omissions is minimized. In addition, more detailed behavior of the formant trajectories can be observed. To illustrate this point an expanded resolution analysis (ERA) was performed over the 32-ms time period shown in Fig. 8 by tripling F_f in this region. The results in the lower portion of the figure show the fast transitional second formant behavior in considerably more detail. The fastest rate of change of the trajectory is measured as -400 Hz/5 ms $=-80$ Hz/ms.

It is rather informative to consider the spectral displays from which these results were automatically obtained. Shown in Fig. 9 are the spectra corresponding to the inverse filter input on the left and the estimate of the resonance structure (obtained from the inverse of the resulting inverse filter spectra) on the right.

The inherent ability of the IFA to track closely spaced formants and fast transitions is clearly shown. In frame 1, F_2 and F_3 are separated by approximately 300 Hz. From the input spectra (frames 1 through 4) it is quite difficult to ascertain whether one or two formants exist in the region of 2 kHz. However, by invoking continuity it becomes clear that two formants must exist in the 2-kHz region. At frame 4, F_2 begins to break away from F_3 and move toward F_1. These input spectra are particularly good examples of the fact that formant extraction is not quite so simple as defining the three largest peaks of the speech spectrum in ascending order of frequency as the formants. This is necessarily the type of algorithm used in filter bank analysis of formant frequencies. In other words, peak picking of the input spectra will not generally produce correct results. However, peak picking of the inverse filter spectra, in general, will give correct results approximately 90 percent of the time.

From the lower portion of Fig. 8 and the inverse filter spectra it is seen that simple peak picking, namely $F_i(n)=p_i(n)$, $i=1, 2, 3$, $n=1, 2, \cdots, 7$, is 100 percent correct for these frames of data.

Frames 5 and 6 also illustrate the fact that correct results are not generally obtained by solving for the roots of $A(z)$ and then defining the three complex roots with smallest bandwidths as the first three formants. With this approach,

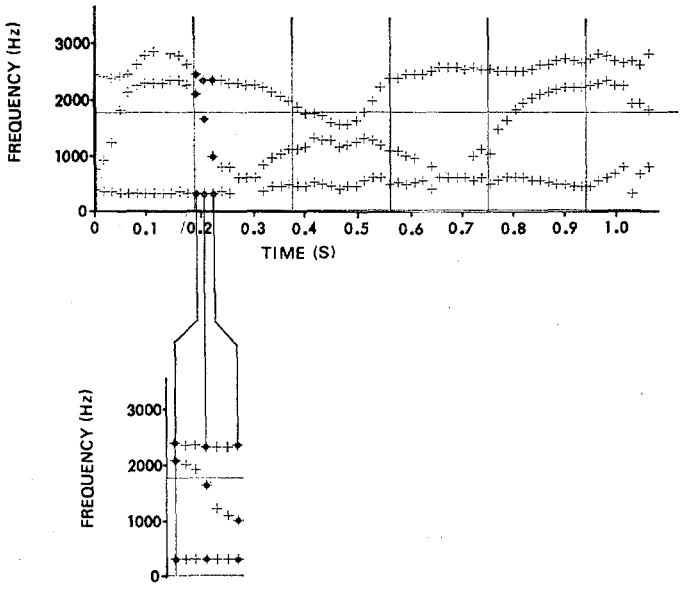

Fig. 8. (Top) The raw data output for the utterance: "We were away." (Bottom) An expanded resolution analysis (ERA) for the interval indicated to illustrate the tracking of a rapid F_2 transition.

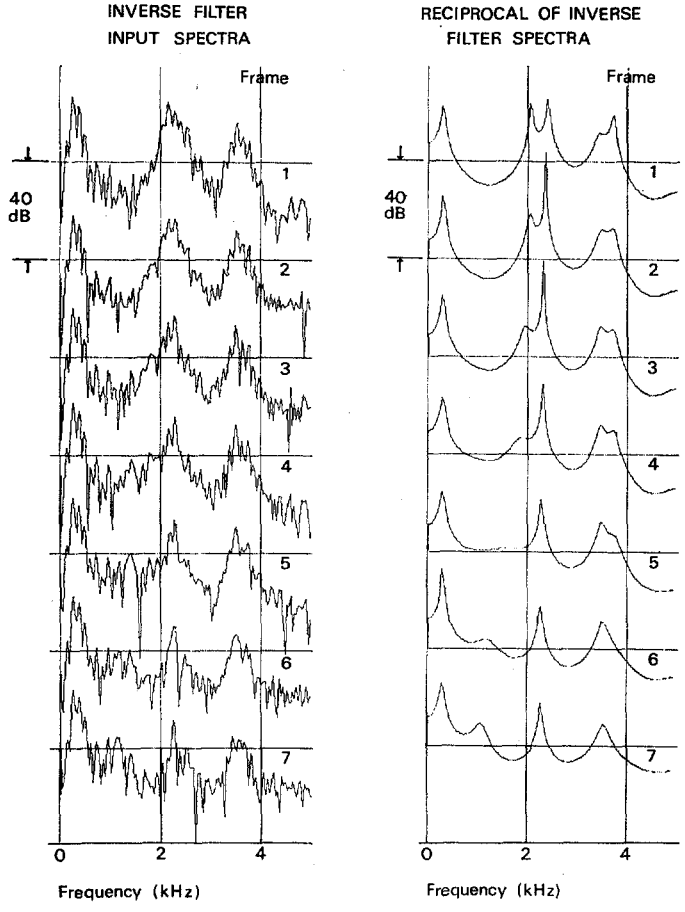

Fig. 9. Spectral representations obtained in the ERA of the transitional segment shown in Fig. 8. (Left) Inverse filter input spectra. (Right) Reciprocal of the resulting inverse filter spectra.

F_2 would be incorrectly defined as F_3, and F_3 would be incorrectly defined as F_4. Thus the IFA has an additional advantage besides computational efficiency over directly solving for the roots of $A(z)$. If a peak can be detected in $|A(e^{j\omega T})|^{-2}$ then it is a candidate for a formant frequency. Generally, only four or five peaks will be detected with peak picking whereas five to seven complex root pairs are usually obtained by directly solving the polynomial $A(z)$. (The extra roots generally have large bandwidths and are thus not detected by moving an analysis contour only along the unit circle in the z plane.)

It is believed that the results as illustrated here are unique in that: 1) the analysis is automatic; 2) the equations are linear, thus giving the results without recursion (in contradistinction to analysis-by-synthesis techniques [19]; 3) the formants are clearly and correctly obtained by simple peak picking of the inverse filter spectra (i.e., no decision criteria are required); and 4) the segment being tracked has formants separated by as little as 300 Hz and an F_2 transition with average slope of greater than -40 Hz/ms.

It should be emphasized that simple peak picking of the inverse filter spectra (more precisely the reciprocal of the inverse filter spectra) will uniquely define the formant trajectories only roughly 90 percent of the time. For the remaining 10 percent of the frames, some kind of decision criteria will be necessary for automatic formant trajectory extraction.

To illustrate the fact that analysis results are not dramatically changed for the suggested ranges of δ and γ, the same phrase was analyzed with $N=320$ and $M=15$. For this case several extraneous insertions are obtained as shown in Fig. 10(A), but the apparent merging of F_1 and F_2 at several locations in Fig. 8 has been eliminated. The first three formant trajectories are still easily definable by inspection using a "connect the dots" procedure. To illustrate this point a subject having no knowledge of acoustic phonetics was asked to draw three smooth continuous curves within the range (0, 3000) Hz that seemed most reasonable without any intersections.

The resulting curves were then superimposed upon a wide-band spectrogram of the utterance for visual comparison as shown in Fig. 10(B). The trajectories are quite realistic with respect to what one experienced with spectrographic analysis would estimate. Note that on the wide-band spectrogram the closely spaced formants appear as a single very thick dark bar, and the fast transition and large portions of the F_3 trajectory appear quite faintly.

If an ERA is performed over the complete phrase with analysis conditions $N=320$ and $M=15$, results over the complete frequency range $(0, F_s/2)$ are obtained as shown in Fig. 11. The first four formant trajectories are easily definable. The tradeoff is simply computation time. The ERA with 5-ms resolution is three times slower than the analysis using a 15-ms frame period.

Although the analysis system was strictly designed for formant trajectory estimation of nonnasalized voiced speech, the system works reasonably well with nasals and unvoiced sounds that contain formant structure.

The phrase: "Hello there, how are you?" was analyzed with parameter values somewhat arbitrarily chosen as

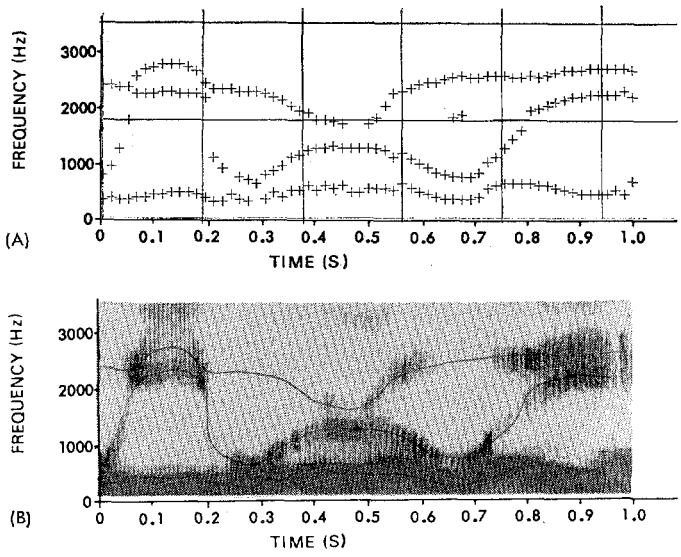

Fig. 10. Analysis of phrase: "We were away." (Top) Raw data from inverse filter algorithm. (Bottom) Spectrographic analysis of phrase with overlay of connected dots from raw data.

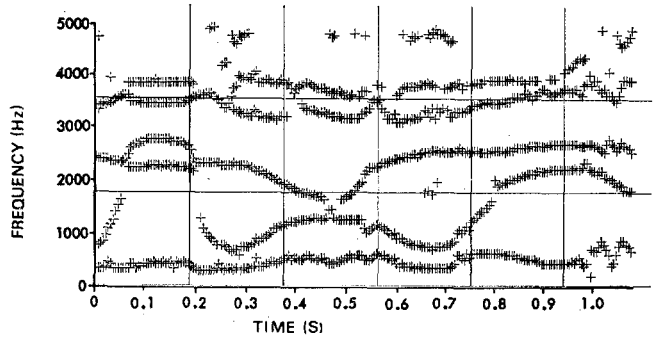

Fig. 11. Analysis of phrase: "We were away" using an expanded resolution analysis (ERA).

$N=256$ and $M=15$. The raw data is shown in Fig. 12(A) and an estimate of three nonintersecting continuous lines made by the same subject (superimposed on a spectrogram of the utterance for comparison) is shown in Fig. 12(B). Although there is a sharp transition in the region (0.24 s, 0.28 s), the raw data shows the formants to be continuously tracked. The spectrogram is used to demonstrate that they are correctly tracked.

The predominantly nasalized phrase: "I am now a man" was analyzed with the parameters $N=256$ and $M=15$. From the raw data as shown in Fig. 13(A) it can be seen that the first four formant trajectories are obtainable. The major difficulty is caused by the initial /m/. The coupling of the vocal tract to the nasal passage introduces zeros in the mathematical model of the physical system. The second formant in the time interval (0.22 s, 0.29 s) appears to have been cancelled by the nasal zero.

A comparison of a spectrogram of the phrase and four smooth continuous nonintersecting curves drawn by the same subject are shown in Fig. 13(B). Once again quite reasonable estimates appear to have been made.

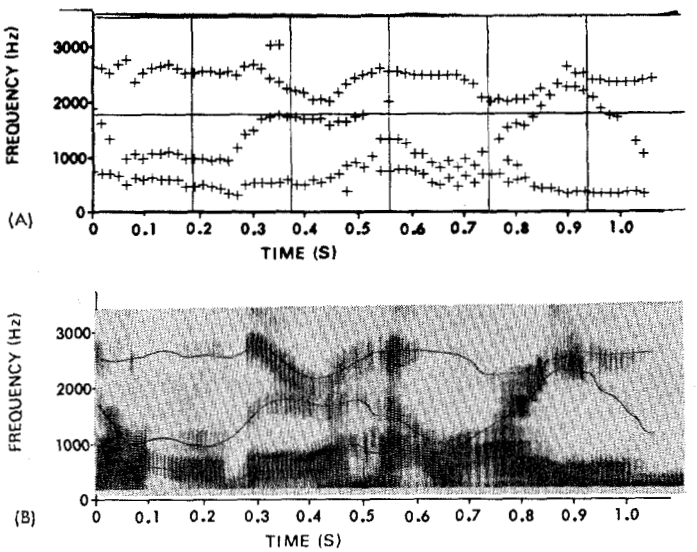

Fig. 12. Analysis of phrase: "Hello there how are you?" (Top) Raw data from inverse filter algorithm. (Bottom) Spectrographic analysis of phrase with overlay of connected dots from raw data.

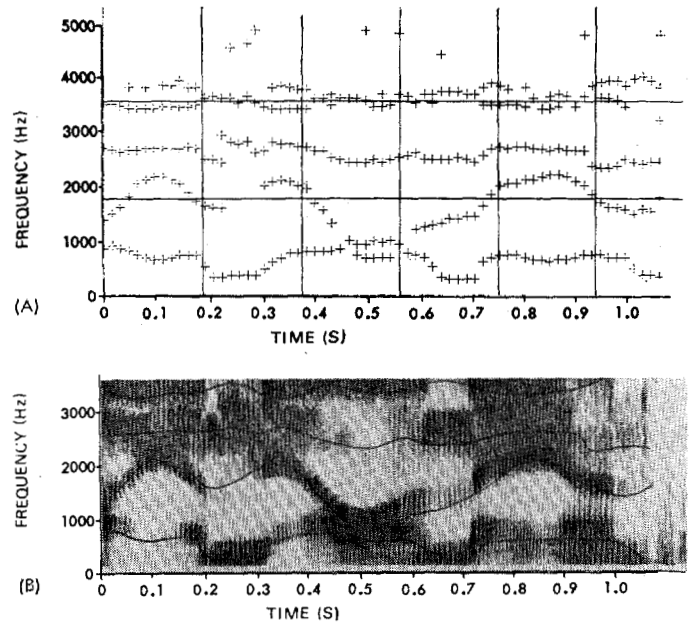

Fig. 13. Analysis of phrase: "I am now a man." (Top) Raw data from inverse filter algorithm. (Bottom) Spectrographic analysis of phrase with overlay of connected dots from raw data.

VII. Conclusion

A new application of the well-known inverse filter concept has been shown—namely, that it is a very useful tool for estimating formant trajectories from speech.

An inverse filter algorithm was presented for efficiently and automatically transforming voiced speech into a set of raw data from which the formant trajectories could then be easily estimated by inspection. Although an algorithm for automatically extracting the formant trajectories from the raw data was not presented, it is believed that the results can still be quite useful for those engaged in basic speech research along with applied research areas such as digital vocoding techniques. The predicted usefulness is based upon the experimental results of the paper and in particular, the fact that the formant trajectories can be easily estimated from the raw data by a naive subject and the fact that simple peak picking of the raw data has been approximately 90 percent effective in estimating the formant trajectories.

Although it is not discussed here, the application of the inverse filter algorithm to the general problem of tracking resonance behavior of time-varying signals should be apparent. Techniques for automatically extracting formant trajectory estimates from the raw data are presently being investigated.

References

[1] S. Saito and F. Itakura, "The theoretical consideration of statistically optimum methods for speech spectral density" (in Japanese), Elec. Commun. Lab., N.T.T., Tokyo, Japan, Rep. 3107, Dec. 20, 1966.

[2] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," *Rep. 6th Int. Congr. Acoust.*, Y. Konasi, Ed., Tokyo, Japan, Rep. C-5-5, Aug. 21–28, 1968.

[3] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," *Rep. 6th Int. Congr. Acoust.*, Y. Konasi, Ed., Tokyo, Japan, Rep. C-5-5, Aug. 21–28, 1968.

[4] J. D. Markel, "The Prony method and its application to speech analysis," *J. Acoust. Soc. Amer.*, vol. 49, pt. 1, p. 105(A), Jan. 1971.

[5] R. Prony, "Essai experimental et analytique sur les lois de la dilatabilite des fluides elastiques et sur celles del la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, a differentes temperatures," *J. Ecole Polytech.*, vol. 1, no. 2, pp. 24–76, 1795.

[6] R. N. McDonough, "Matched exponents for the representation of signals," Ph.D. dissertation, Dep. Elec. Eng., the Johns Hopkins University, Baltimore, Md., Apr. 1963.

[7] C. Runge and H. Konig, *Vorlesungen über Numerisches Rechnen*, vol. 11 of *Die Grundlehren der Mathematischen Wissenschaften*. Berlin: Springer, 1924, p. 231.

[8] C. S. Burrus and T. W. Parks, "Time domain design of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 137–141, June 1970.

[9] J. L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. 32, pp. 33–51, Feb. 1967.

[10] E. A. Robinson, *Statistical Communication and Detection*. New York: Hafner, 1967.

[11] K. L. Peacock and S. Treitel, "Predictive deconvolution: Theory and practice," *Geophysics*, vol. 34, pp. 155–169, Apr. 1969.

[12] E. A. Robinson and S. Treitel, "Principles of digital Wiener filtering," *Geophys. Prospect.*, vol. 15, pp. 311–333, Sept. 1967.

[13] N. Levinson, "The Wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, no. 4, pp. 261–278, 1947; also in N. Wiener, *Extrapolation Interpolation and Smoothing of Stationary Time Series*. Cambridge, Mass.: M.I.T. Press, 1966.

[14] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*. New York: Dover, 1959.

[15] R. H. Pennington, *Introductory Computer Methods and Numerical Analysis*. Toronto, Canada: Macmillan, 1970, pp. 348–355.

[16] J. D. Markel, "FFT pruning," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 305–311, Dec. 1971.

[17] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–184, Mar. 1952.

[18] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, Feb. 1970.

[19] C. G. Bell *et al.*, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725–1736, Dec. 1961.