Predictive Coding in a Homomorphic Vocoder

CLIFFORD J. WEINSTEIN, Member, IEEE
ALAN V. OPPENHEIM, Senior Member, IEEE
Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, Mass. 02173

Abstract

Application of a type of predictive coding to the channel signals of a homomorphic vocoder has produced sizable bit rate reductions. With only slight degradation in speech quality, reduction (for the spectral envelope information) from 7800 to 4000 bits/s was achieved. A technique for obtaining the formant frequencies from the predictive coding parameters is described; this approach promises further bit rate reductions. As a byproduct of this study of predictive coding, direct and cascade form speech synthesizers are compared on the basis of differing quantization effects.

Manuscript received November 21, 1970.

This work was sponsored by the Department of the Air Force.

I. Introduction

Recently, a speech analysis–synthesis system based on homomorphic filtering was described [1]. In the analysis the cepstrum is computed, the low-time portion of which is used to represent the spectral envelope. Excitation information represented by pitch and a voiced–unvoiced decision is obtained either from the cepstrum or by using any of a variety of available pitch detection algorithms. The synthesis consists of converting the low-time cepstral values obtained in the analysis to an impulse response function which is then convolved with an excitation waveform generated from the excitation information to obtain the synthesized speech.

When considered within the context of a bandwidth compression system, the low-time cepstral values were viewed as the channel signals. It was reported that the first 26 cepstral values, updated every 20 ms and each quantized to 6 bits, provided high-quality speech. It was anticipated that more sophisticated coding of the channel signals could lead to a reduction in the bit rate.

As an alternative to more efficient coding of the channel signals, the bit rate required to resynthesize the speech can be reduced by modeling the channel signals, the spectral envelope, or the impulse response function in terms of a smaller number of parameters. An example of a procedure of this type is the use of formant tracking on the transform of the low-time cepstral values [2]. Such procedures seem to be applicable primarily within the context of audio response units because of the complexity of the analysis. In this paper a similar kind of procedure is described. The impulse response function is modeled as the unit sample response of a digital filter. Unlike direct formant tracking, the parameters of the digital filter are obtained by solving a set of linear equations which express the criterion that when the inverse of the recursive filter is excited by the impulse response function to be modeled, the mean-squared difference between the obtained output and an impulse is minimized. This algorithm for obtaining the parameters of a filter to represent a given impulse response function was originally suggested by Atal and Schroeder [3] within the context of a speech compression system based on predictive coding.

The parameters that emerge from the solution of the set of linear equations represent the coefficients of a single direct-form recursive filter. One could quantize and transmit these coefficients or, at a large cost in computation, factor the direct-form system function into first- and second-order sections and quantize and transmit the coefficients of this cascade-form filter. It is generally considered in the implementation of digital filters that a cascade-form implementation is less sensitive to coefficient inaccuracies, and consequently coefficient quantization, than a direct-form implementation [4]. In the present system, however, parameters are initially obtained for a direct-form realization and parameters for the cascade form must be obtained by factoring the direct-form polynomial. Consequently, one aspect of the present investigation was to compare the coefficient word length required for cascade-form and direct-form synthesis.

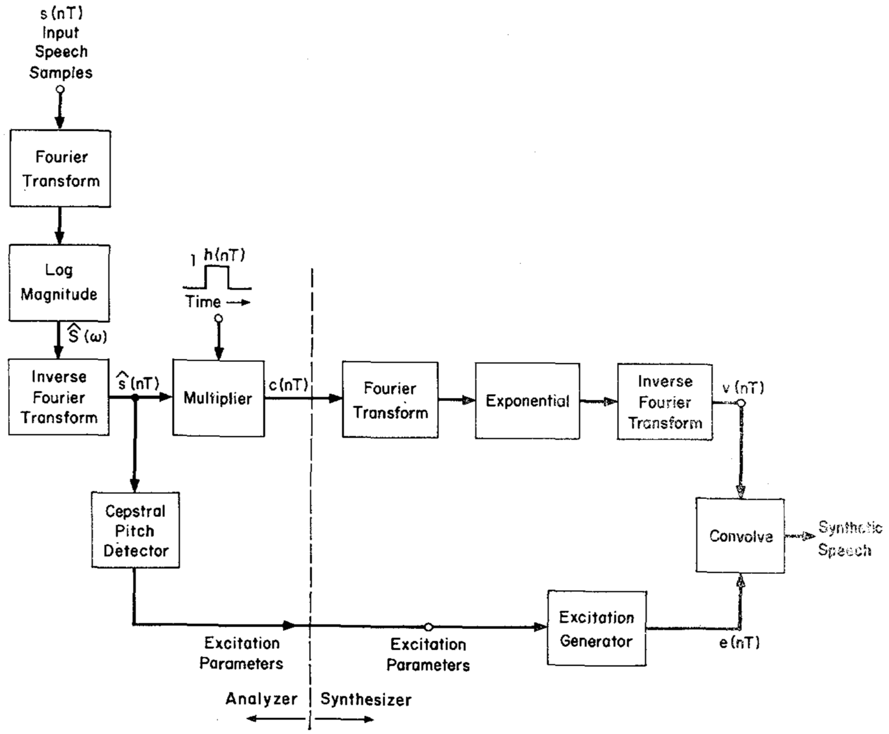

Fig. 1. Block diagram of homomorphic analysis—synthesis system.

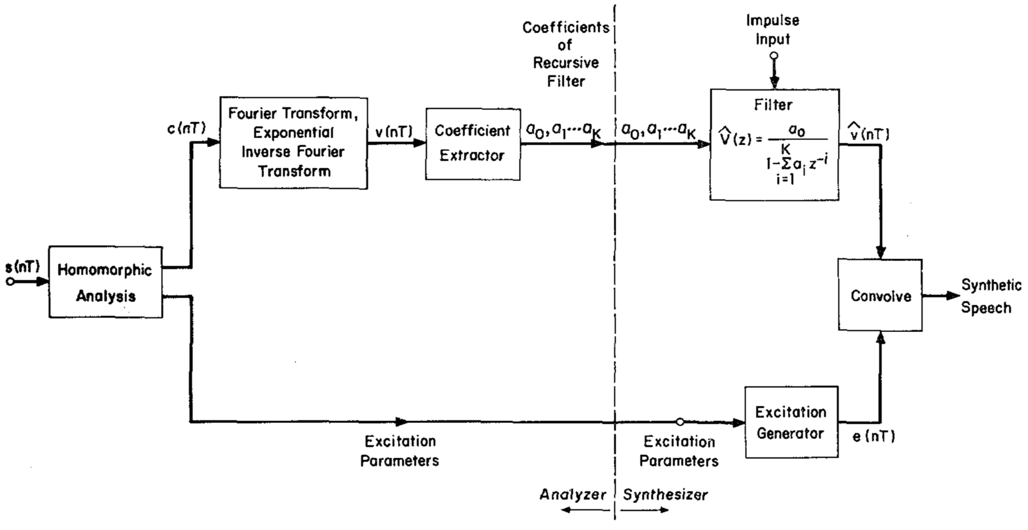

Fig. 2. Block diagram of homomorphic vocoder with predictive coding.

II. System Description

A. General Strategy

A block diagram of the homomorphic analysis–synthesis system is shown in Fig. 1. The spectral envelope information contained in $c(nT)$ (the low-time portion of the cepstrum) is updated and transmitted every 20 ms. The excitation parameters are obtained by cepstral pitch detection. At the receiver the cepstral values $c(nT)$ are transformed into impulse response functions $v(nT)$ and convolved with the excitation function to yield synthesized speech.

In the present system, the entire top path of Fig. 1, through the conversion of $c(nT)$ to $v(nT)$, is included as part of the analyzer. Also, at the analyzer the coefficients of a recursive filter that approximates the given impulse response $v(nT)$ are obtained. These coefficients then represent the spectral envelope information. The excitation function information is represented exactly as in Fig. 1. A general block diagram of our system is shown in Fig. 2. Note that the synthesis is performed by first pulsing the recursive filter to obtain an impulse response function and then convolving this impulse response function with the excitation function.

To obtain the coefficients of the filter used to represent the impulse response function, we seek to approximate the effect of the vocal tract by an all-pole digital filter with transfer function $\hat{V}(z)$ of the form

$$\hat{V}(z) = \frac{a_0}{1 - \sum_{i=1}^{K} a_i z^{-i}} \tag{1}$$

where K is the order of the filter. If the given impulse response $V(nT)$ is played through the *inverse* of $V(z)$, then the output should resemble an impulse. This provides a criterion for selecting the a_i. Specifically, when $\hat{V}^{-1}(z)$ is excited by $v(nT)$, the mean-squared difference between the obtained output and an impulse is to be minimized. This criterion has been selected among many other possible and reasonable criteria chiefly because it leads to a closed procedure for obtaining the a_i. Fig. 3 depicts the situation of interest. (To simplify the notation, we have represented $v(nT)$ by v_n.) We have neglected, for the moment, the factor a_0 since it represents only a gain. Samples of v_n are available for $n=0, 1, \cdots, N-1$. The criterion is that

$$(u_0 - 1)^2 + \sum_{n=1}^{N-1+K-1} u_n^2 \tag{2}$$

is minimum, that is, the mean-squared difference between u_n and an impulse is minimized. We will choose the gain factor a_0 such that $u_0 = 1$ so that the criterion becomes

$$\sum_{n=1}^{M} u_n^2 \tag{3}$$

is minimum, where $M = N+K$ is the maximum length of the sequence u_n.

We now express (3) in terms of the v_n and a_i and carry out the minimization. Since

$$u_n = v_n - \sum_{i=1}^{K} a_i v_{n-i}, \tag{4}$$

then

$$\sum_{n=1}^{M} u_n^2 = \sum_{n=1}^{M} v_n^2 + \sum_{n=1}^{M} \left(\sum_{i=1}^{K} a_i v_{n-i} \right)^2 - 2 \sum_{n=1}^{M} v_n \sum_{i=1}^{K} a_i v_{n-i}. \tag{5}$$

The minimization is carried out by equating the partial derivative of (5) with respect to a_j to zero so that

$$\sum_{n=1}^{M} 2v_{n-j} \sum_{i=1}^{K} a_i v_{n-i} - 2 \sum_{n=j}^{M} v_n v_{n-j} = 0 \tag{6}$$

or

$$\sum_{i=1}^{K} \left(\sum_{n=1}^{M} v_{n-j} v_{n-i} \right) a_i = \sum_{n=1}^{M} v_n v_{n-j}, \quad \text{for } j = 1, 2, \cdots, k. \tag{7}$$

Equation (7) constitutes a set of linear equations that may be solved to determine the a_i. In matrix form,

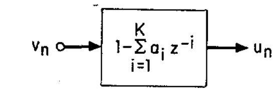

Fig. 3. Criterion for selecting coefficients of vocal tract filter.

$$\phi a = \psi \tag{8a}$$

where ϕ is a symmetric $K \times K$ matrix, a and ψ are $K \times 1$ vectors,

$$\phi_{ij} = \sum_{n=1}^{M} v_{n-i} v_{n-j}, \tag{8b}$$

and

$$\psi_j = \sum_{n=1}^{M} v_n v_{n-j} = \phi_{0j}. \tag{8c}$$

The "coefficient extractor" box of Fig. 2 computes ϕ and ψ from the given v_n, then solves (8a) to obtain the coefficients a_i. Note that the order K of the filter approximating the impulse response function is a parameter that can be varied.

The transfer function $\hat{V}(z)$ has been represented as a single kth-order transfer function, with the coefficients a_i as the coefficients of a direct-form digital filter realization. An algorithm as simple as that described above for determining the coefficients of a cascade-form synthesizer is not available. However, given the coefficients a_i of the direct-form transfer function, one could factor $\hat{V}(z)$ by means of a polynomial root-finding program to determine the coefficients of a cascade realization. From previous results [4] on the sensitivity of different digital filter realizations to coefficient quantization, one might expect that the cascade coefficients could be transmitted with fewer bits of accuracy than the direct-form coefficients. Also, once $\hat{V}(z)$ has been factored, there is the possibility of transmitting only the center frequencies for each cascade section (formant) and using a fixed table to determine the bandwidths at the receiver.

B. Computational Details

In this section we discuss some of the computational details of our system, with particular attention to the formation and solution of the linear equation system derived in (8).

The homomorphic analysis produces a new impulse response function v_n every 20 ms; the duration of v_n is limited to a maximum of 256 points (at the 10-kHz sampling rate). For each impulse response, we must form the equation system described in (8) by computing the K^2 elements of ϕ_{ij} and the K elements of ψ_j. The computation can be carried out efficiently by making use of the relation

$$\phi_{i+1,j+1} = \phi_{ij}. \tag{9}$$

Only one row of the ϕ_{ij} elements needs to be computed and the remaining elements follow directly from this relation with no additional computation.

Once the equation system has been formed, it can be solved by any of the standard methods for solving linear equations. We have used a Gaussian elimination procedure; Atal and Schroeder solved the system of equations by means of "square-rooting" procedure, which is computationally more efficient than the Gaussian elimination procedure for symmetric matrices. The computation involved in solving the system of equations was implemented with 36-bit floating-point arithmetic to insure sufficient accuracy. For the speech data that we processed, the Gaussian elimination procedure always yielded solutions that produced satisfactory speech at the synthesizer. No difficulties with singularities in the ϕ matrix were encountered.

By the procedure just described, a set of K coefficients a_i for a direct form vocal tract filter was obtained for each impulse response. For direct-form synthesis, the a_i were quantized and transmitted to the synthesizer. A new impulse response function was obtained at the synthesizer by pulsing a filter with coefficients a_i. Synthesized speech was then generated by convolving the reconstructed impulse responses with the excitation function.

For cascade synthesis, a root-finding program was used in the analyzer to factor $\hat{V}(z)$ [see (1)] into a product of first- and second-order systems. The coefficients of the first- and second-order sections were then quantized and used at the synthesizer as the coefficients in a cascade-form filter to obtain the impulse response functions.

III. Bandwidth Compression Results

In this section we describe the bandwidth compression results obtained by means of the above procedure, both for direct-form and cascade-form synthesis. In discussing degradation in speech quality, we will consider the speech produced by the original homomorphic system as our standard. Of course, all comments regarding quality of synthesized speech are subjective. The results we discuss are based on processing and listening to five sample sentences for each system bit rate. Three of the sentences were spoken by males and two by females.

The bit rate for the spectral envelope information of the homomorphic system was 7800 bits/s. The corresponding bit rate for the present system is defined by

$$\text{bits/s} = (50)(B)(K) \tag{10}$$

where K is the number of coefficients, B is the number of bits per coefficient, and the factor 50 arises from the fact that the coefficients are updated every 20 ms. Thus, if $K=10$ and $B=8$, we would require 4000 bits/s for the spectral envelope information; adding another 800 bits/s for excitation function information, this corresponds to a 4800-bits/s speech transmission system.

A. Direct-Form Synthesis

1) Number of Coefficients: We first investigated the effect of varying the number of coefficients K while maintaining the accuracy of the individual coefficients at 36 bits (floating point). The largest value of K tried was 14 (corresponding to a 7-formant system). For $K=14$, with 36-bit coefficients, the speech produced by this system was judged to be indistinguishable from the speech originally synthesized by the homomorphic system. For $K=10$, differences from the original speech were barely noticeable, but the degradation in quality was judged to be almost negligible. For $K=8$, there was a noticeable loss of sharpness in the high-frequency components of the speech. The speech quality continued to degrade as K was reduced further, but even for $K=4$ the sentences were intelligible.

Inspection of spectrograms of the synthesized sentences and spectra of individual impulse responses indicated that as K was decreased, the speech spectrum lost its resolution at high frequencies. Said another way, the higher formant resonances were lost as K decreased, although the first and second formant were tracked fairly well even for $K=6$.

To explain why this system matches the vocal tract spectrum more closely at low frequencies than at high frequencies, we refer back to Fig. 3 and (2). Minimization of (2) is equivalent to minimization of

$$\int_{-\pi/T}^{\pi/T} \left| V(e^{j\omega T})\widehat{V^{-1}}(e^{j\omega T}) - 1 \right|^2 d\omega \tag{11}$$

where $V(z)$ is the z transform of v_n and $\widehat{V^{-1}}(z)$ is our approximation to the vocal tract inverse filter. Equation (11) may be rewritten as

$$\int_{-\pi/T}^{\pi/T} \left| V(e^{j\omega T}) \right|^2 \left| \widehat{V^{-1}}(e^{j\omega T}) - V^{-1}(e^{j\omega T}) \right|^2 d\omega. \tag{12}$$

An interpretation of (12) is that the coefficients are chosen to minimize the integrated squared difference between the actual inverse transfer function $V^{-1}(e^{j\omega T})$ to be approximated and the approximation $\widehat{V^{-1}}(e^{j\omega T})$, but that errors are to be weighted nonuniformly over the frequency band, according to the magnitude of the actual transfer function $V(e^{j\omega T})$. Thus, where $|V(e^{j\omega T})|$ is large, we would expect smaller errors in the approximate filter than where $|V(e^{j\omega T})|$ is small. Since the speech spectrum falls off at 6 dB per octave as frequency increases, we would expect our system to match the vocal tract spectrum better at low frequencies than at high frequencies.

2) Quantization of Coefficients: The effect of quantizing the coefficients of the direct-form synthesizer was studied. In a typical experiment, the number of coefficients was kept fixed (e.g., at 14) and the word length of the coefficients was reduced from 36 bits (floating point) to 18 bits (fixed point) and then in steps to 4 bits (fixed point); each of 5 sentences was synthesized with all the various coefficient word lengths.

Degradations due to the quantization became barely audible at 10 bits/coefficient. At 8 bits/coefficient, occasional intrusions (sounding like pops) in the speech were heard, the intelligibility was good, and the quality of speech was fair. At 6 bits/coefficient, the pops and intrusions became quite frequent and disturbing. A probable reason for the intrusions in the speech was that the coefficient errors were causing some of the vocal tract impulse responses to become unstable.

3) Overall Bit Rate: Judging from the results above, it was felt that acceptable (though slightly degraded) speech could be produced with $K=10$ and $B=8$, corresponding to a 4000-bit/s rate. Adding 800 bit/s for excitation function information, this represents a 4800-bit/s speech transmission system.

To obtain speech quality essentially indistinguishable from the speech produced by the homomorphic system, it seemed that $K=10$, $B=10$, corresponding to a spectral rate of 5000 bits/s, would suffice.

B. Cascade-Form Synthesis

1) General Procedure: To obtain the coefficients for a cascade synthesizer, we first solved (8) to obtain a direct-form transfer function, then used a polynomial root-finding program (from the IBM scientific subroutine library) to find the cascade coefficients. The need for root-taking obviously adds significant computational complexity to our system.

At the synthesizer, the cascade filter is pulsed to obtain an estimated vocal tract impulse response, which is then convolved with the excitation function. (The cascade sections were either second-order filters of the form $1/(1-a_1z^{-1}-a_2z^{-2})$ or first-order filters of form $1/(1-az^{-1}.)$ An alternative synthesis procedure that was considered was to play the excitation function directly into a cascade filter with time-varying coefficients. However, as pointed out by Atal [5], this method does not work because of a problem in ordering the cascade sections. One starts out with a particular ordering for the sections, but as the coefficients change with time, one must decide which coefficients to associate with which cascade section. If the order of the sections is not consistent with the filter initial conditions, the output speech degrades badly. For direct-form synthesis, this problem does not appear and therefore the synthesis can be implemented by playing the excitation function into a time-varying filter, instead of carrying out a convolution. However, we chose a convolution in the direct-form case to allow us to change the impulse response every 5 ms by linear interpolation between the synthesized impulse responses, which are computed every 20 ms.

2) Bit Rate Reduction—Straightforward Approach: The bit rate necessary for cascade synthesis could be examined by varying K and B [see (10)] as in the direct-form case and listening to the effects.

The effect of varying the number of coefficients with 36-bit accuracy in the coefficients was essentially identical to the corresponding effect in the direct-form case, which was discussed above. This was to be expected since, except for quantization errors, the two systems would yield identical speech.

The effect of varying B was somewhat different than in the direct-form case. For $B=8$ bits/coefficient, the degradation in the speech was barely noticeable. For $B=6$, spectral distortions in the speech were quite noticeable, but the intrusions and pops that occurred in direct-form synthesis were not present. The speech degraded quite gracefully as B was made smaller, and even for $B=4$ the sentences were intelligible, although quite unnatural sounding.

3) Formant Tracking Approach: In obtaining the coefficients of a cascade synthesizer, we have essentially tracked the center frequencies and bandwidths of the vocal tract formants. In fact, for a cascade section, with transfer function of the form

$$V_c(z) = \frac{1}{1 - a_1z^{-1} - a_2z^{-2}},$$

the center frequency and bandwidth are given approximately by

$$f_c = \frac{1}{2\pi T} \cos^{-1}\left(\frac{a_1}{\sqrt{-a_2}}\right)$$

$$W = \frac{1}{4\pi T}(1 - a_2). \tag{13}$$

In order to reduce bit rate, one could consider transmitting only the center frequency for each formant, and calculating the bandwidth at the synthesizer to fit on a predetermined curve of formant bandwidth versus formant frequency.

This method was tried. The normalized center frequency (angular pole location) $\theta=2\pi f_cT$ was determined for each cascade section and transmitted to the synthesizer, where the radial pole location $r=e^{-2\pi WT}$ was determined according to the formula

$$r(\theta) = 0.982e^{-0.056\theta}. \tag{14}$$

However, the speech produced was wholly unsatisfactory. The problem was that some of the pole pairs obtained by our analysis were well inside the unit circle and corresponded not to formant resonances but to a general shaping of the spectrum. Simple application of (14) brought these low-gain poles out near the unit circle and produced false resonances in the speech. To counteract this effect, some information about r was transmitted to the synthesizer and a rule such as the following replaced (14).

$$r(\theta) = \begin{cases} 0.982e^{-0.056\theta}, & r \overline{>} 0.8 \\ 0.7, & 0.6 \overline{<} r < 0.8 \\ 0.5, & 0.4 \overline{<} r < 0.6 \\ 0.3, & 0 \overline{<} r < 0.4. \end{cases} \tag{15}$$

This corresponded to placing r on one of four possible contours, and required that two bits of information about r (to select the contour) be transmitted.

With this method, reasonably good quality speech was produced, although the quality was definitely inferior to the originally synthesized speech. However, the bit rate for such a system could become attractively low. Typically, one could use 5 second-order systems, with θ quantized to 7 bits, and 2 bits to characterize r, so that the bit rate for spectral information would be (50) (5) (7+2)=2250 bits/s.

IV. Direct-Versus Cascade-Form Synthesis

A specified digital transfer function may be implemented in a number of forms, of which two of the standard forms are the direct and cascade. One of the important criteria in choosing between these forms is the sensitivity of the actual realized transfer function to quantization of filter coefficients.

Kaiser [3] has shown that for a transfer function with closely spaced, high-gain poles, the direct form is quite undesirable because the pole positions are very sensitive to coefficient errors. He also showed that for such a direct-form realization, the pole position sensitivity increased with filter order.

The present study has provided an experimental comparison of the effects of coefficient quantization in direct and cascade filters, in a situation where the filters were of fairly high order (up to 14th), but the pole positions were not too closely spaced. Even for 14th-order vocal tract filters, it was found that the coefficient accuracy needed for acceptable speech output with the direct form was not excessive; with 9-bit coefficients, degradations from the original homomorphic speech were barely noticeable. With a cascade synthesizer of the same order, similar speech quality could be produced with 7-bit coefficients. As the quantization was made much coarser (down to 6 bits), the direct-form speech began to degrade completely, with pops and squeals appearing in the speech, presumably due to instabilities in some of the filters. The cascade-form speech degraded much more gracefully, and even at 4 bits/coefficient was intelligible and not marred by intrusions.

V. Conclusions

Application of a type of predictive coding originally proposed by Atal and Schroeder to the channel signals of a homomorphic vocoder has produced sizable bit rate reductions.

It seems that speech of quality comparable to that of the homomorphic system could be produced at an overall bit rate of 4800 bits/s. To further reduce the bit rate (say down to 2400 bits/s), a technique such as the formant tracking scheme (Section III-B) would have to be utilized; more detailed experimentation with the scheme would be needed.

The computational complexity of the predictive coding algorithm is quite large, especially if a root-finding procedure is included. The system is far from a real-time vocoder. However, most of the complexity is in the analysis, so that a possible application would be for an automatic audio response device, where the analysis could be performed offline.

As a by-product of this study, direct- and cascade-form speech synthesizers were compared on the basis of differing effects of coefficient quantization. The cascade form showed some advantage in necessary coefficient accuracy, although the difference was not as dramatic as we had at first expected.

References

[1] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Accoust. Soc. Amer.*, vol. 45, 1969, pp. 459–462.

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

[3] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, Oct. 1970.

[4] J. F. Kaiser, "Digital filters," in *System Analysis by Digital Computers*, F. Kuo and J. F. Kaiser, Eds. New York: McGraw-Hill, 1959, ch. 7, pp. 218–285.

[5] B. S. Atal, private communication.