SPEECH DIGITIZATION BY LPC ESTIMATION TECHNIQUES

D. T. Magill, et al

Stanford Research Institute

Prepared for:

Advanced Research Projects Agency
Army Research Office-Durham

July 1974

Annual Technical Report

Covering the Period 3 October 1972 through 31 March 1974

SPEECH DIGITIZATION BY LPC ESTIMATION TECHNIQUES

By: D. T. MAGILL, E. J. CRAIGHILL, and D. W. ELLIS

Prepared for:

ADVANCED RESEARCH PROJECTS AGENCY
ARLINGTON, VIRGINIA 22209

CONTRACT DAHC04-72-C-0009
ARPA Order No. 1943
Program Element Code 61101D

Approved for public release; distribution unlimited.

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

Annual Technical Report – Task 2
Covering the Period 3 October 1972 through 31 March 1974

Form Approved
Budget Bureau No. 22-R0293
July 1974

Task 3 in AD-785738

SPEECH DIGITIZATION BY LPC ESTIMATION TECHNIQUES

By: D. T. MAGILL (Task Leader), E. J. CRAIGHILL, and D. W. ELLIS
(415) 326-6200, Ext. 2664

Prepared for:

ADVANCED RESEARCH PROJECTS AGENCY
ARLINGTON, VIRGINIA 22209

CONTRACT DAHC04-72-C-0009
ARPA Order No. 1943
Program Element Code 61101D

SRI Project 1526

Approved by:

ROBERT F. DALY, *Director*
Telecommunications Sciences Center

BONNAR COX, *Executive Director*
Information Science and Engineering Division

CONTENTS

ILLUSTRATIONS

TABLES

1 INTRODUCTION

A. Background

This project was established to study the relevance of linear predictive coding (LPC) estimation techniques for the development of a practical, real-time system for transmitting digitized voice signals. These techniques had been shown to provide excellent quality transmission at modest bit rates when simulated on large-scale digital computers. Our goal was to determine how they can be used in packet communication systems with smaller computers.

During the first year of the project, our perspective on the problem changed in three ways. First, it became apparent that achieving high quality was of paramount importance and that the computational load was not as critical as originally anticipated. The rapid development of high-speed large-scale integrated circuits (LSI) has made it possible to achieve remarkable computational capabilities today, and the projections for future developments are even more promising. In addition, most of the LPC approaches offer roughly comparable computational loads, since the major amount of computation is in the calculation of autocorrelation coefficients and in the synthesizing filter. Thus, we decreased our emphasis on the number of computations per second.

Second, as the program progressed, more literature on the effect of quantization accuracy requirements became available. We were able to adopt the major results and the most promising techniques from this research and, accordingly, to reduce our own efforts in this area.

Third, the importance of accurate pitch for high quality synthesis and the difficulty of the pitch-extraction problem became apparent early

in the program. The high quality of the original LPC-synthesized speech resulted from the accuracy of hand-marked pitch pulses as well as the inherent advan ages of the LPC technique itself. Furthermore, pitch extraction from the residual was far more complex than the original papers implied. As a result, work on pitch extraction was established as Task 3 research under this contract, and major effort was directed toward the study of the excitation function.

B. Summary of Areas Studied During Task 2 Research

1. Asynchronous Operation

This research was directed toward the development of an LPC-speech digitization technique that is compatible with the asynchronous operational mode of packet communication systems. Since previous research on LPC techniques had been concerned exclusively with synchronous systems, a major part of our effort was devoted to the study of asynchronous operation. The result is the adaptive data compression algorithm DELCO, described in detail in Magill (1973), a copy of which is attached as Appendix A to this report.[1]* This algorithm is specifically designed to function with and take advantage of the characteristics of an asynchronous data channel. DELCO offers a data compression factor between 2:1 and 3:1 beyond that achieved by standard LPC approaches, with no degradation in voice quality.† Thus, neglecting the overhead of the packet communication system, we can transmit speech digitally between 1200 and 2400 baud.

---

*References are listed at the end of this report.

†An additional 2:1 data compression is achieved in an asynchronous system, since no channel capacity is allocated for listening as in fixed-channel assignment systems. That is, it is possible to capitalize on the less than 50 percent average duty factor in a two-way conversation.

This excellent performance is obtained by two means. First, the pauses in speech are recognized by a TASI-type speech detector and are not encoded or transmitted. Second, periodic waveforms, such as occur in steady-state vowels, are recognized; LPC coefficients are transmitted only when new values are required--i.e., when the vocal tract configuration changes significantly. The need for coefficient updating is determined from the ratio of the residual energies formed with the previous LPC parameters and the optimum parameters. Note that these two operations do not significantly increase the number of computations, so the ability to achieve real-time operation is not significantly impaired.

2. Error Signal Characterization

In previous studies, two methods have been used to characterize the error--or residual--signal (the difference between the predicted and actual values). In the first method, the error signal is characterized at each time sample by several bits. The quantized error signal is transmitted and used to drive the synthesizer at the receiver. A potential advantage of this approach is that the synthesis procedure should maintain high quality performance even in the presence of audio background noise. The major disadvantage is that the bit rate required to characterize the error signal is high, e.g., nominally at least 7200 baud for a one-bit quantizer.

With the second method, the error signal's features are extracted, so that a much lower bit rate is adequate to represent the error signal. These key features are voiced/unvoiced (V/UV) decision, pitch frequency, and power level. The disadvantage with this method is that, if errors are made in the feature-extraction process, serious degradation of performance will result. Unfortunately, these errors can occur rather easily in the presence of common disturbances, such as audio background noise, phone-line signal distortion, and multiple speakers.

Because of the difficulties in these methods, a major goal in our research was to seek alternative encoding or characterization techniques. Several concepts based on peak-picking, threshold-crossing, and extrema-encoding were proposed; however, a detailed investigation of these techniques was not possible because of the character of the error the signal revealed by experimental observations. The proposed algorithms simply would not function reliably with the observed signals.

This result was not anticipated because some of the foremost researchers in LPC methods had indicated that simple peak-picking was adequate (Atal and Hanauer, 1971).[2] Our experiments, however, showed conclusively that one could not rely on the presence of a readily observable pitch pulse in the residual signal. In fact, the residual frequently was highly oscillatory with multiple peaks per pitch period. This situation destroyed the purpose and the advantages of the proposed algorithms for error-signal characterization.

The difficulty of the encoding problem can best be appreciated by noting that the residual signal is extremely intelligible. In fact, it sounds like differentiated speech. Thus, the problem of encoding the residual signal is virtually as complex as the problem of directly encoding the input speech.

Since this result was so surprising, we made an attempt to determine the cause. First, we tried various forms of analyses (such as pitch-synchronous versus pitch-asynchronous, and Toeplitz versus non-Toeplitz) and varying numbers of coefficients. The most desirable residual signals were found with a pitch-synchronous (over one pitch period), non-Toeplitz analysis or with a preemphasized, Toeplitz analysis over multiple pitch periods. Nevertheless, even in these cases, highly oscillatory residuals were frequently observed. Thus, the proposed algorithms would not function well enough for any of the conventional LPC approaches.

After a literature search and review and after experiments with synthetic speech, we discovered two potential difficulties. First, the use of a stationary model (fixed predictive coefficients for each analysis block) increases the energy of the error signal, especially during speech sounds with changing formant frequencies (vowel glides, transitions from consonant to vowel, and the like). Hence the choice of analysis block size is critical. Second, conventional LPC approaches model the glottal excitation shape as well as the vocal tract. However, the glottal excitation waveshape cannot be modeled accurately by poles (although the spectrum can be approximated quite well). As a result, the residual signal based on these approximate LPC parameters was frequently quite oscillatory and contained a significant amount of formant information.

On the basis of a theoretical model and experiments with synthetic speech, we determined that the true vocal tract parameters could be found by performing an LPC analysis over only the force-free portion of one pitch period. Use of these true vocal tract parameters in the predictor produces the glottal excitation waveshape for the residual signal. This waveshape lends itself naturally to the proposed encoding schemes of peak-picking, threshold-crossing, and extrema-encoding.

Thus, the research indicated that the use of the proposed concepts is possible. First, however, it is necessary to find the force-free period for analysis. This problem is complex but fortunately is not quite as demanding as pitch extraction. Because of the difficulty of the pitch-encoding problem, it was assigned to a separate study of excitation encoding (see the Task 3 report). Meanwhile, we adopted the feature-extraction approach and hand placed the pitch pulses. With this approach, we avoided the problems of algorithmic pitch extraction and could concentrate on the major problem of asynchronous operation.

As mentioned before, the error-signal characterization (or pitch-extraction) problem is extremely difficult. A separate research

effort (Task 3) was devoted to this subject; the reader is referred to the Task 3 final report for more details on error-signal characterization. In this Task 2 report, sections are devoted to time-domain pitch extraction (Section III) and pitch-accuracy requirements (Section IV).

3. Process Modeling

The requirement for zeros in the speech process model was determined to result from the following factors:

- Incorrect analysis time base with respect to the pitch period, i.e., nonminimum phase waveforms during the analysis interval.
- Glottal excitation waveshape.
- Nasals.
- Other sounds with side cavities or branches in the acoustic tract.

We determined that pole approximations to the zeros required for the last two items gave adequate performance with respect to synthesis, provided that the pitch-extraction problem was solved. That is, the ear is relatively insensitive to the phase of the synthesized speech. However, the inability to produce an inverse filter that correctly deconvolves the source zeros greatly hampers pitch extraction based on the residual signal. Thus, for nasals and for other sounds produced with side cavities present, the need for zero modeling is principally associated with the pitch-extraction problem.

To model the excitation waveshape accurately, many zeros--perhaps 50--are required because of the high duty factor of the excitation. The resulting computational problems can be avoided in several ways. First, the residual can be heavily filtered and the sampling rate can be reduced so that fewer zeros suffice. Second, the excitation waveshape can be approximated by a simple waveform--e.g., a triangle--and the

characteristic parameters can simply be encoded so that the problem of zeros is avoided altogether. Third, two LPC analyses can be performed. The first analysis would be based on a selected period to avoid the excitation function. These coefficients would be used to produce a residual that permitted simple pitch extraction. The second LPC analysis would be based on one or more pitch periods and would model the excitation waveshape (by approximating it with poles) as well as the vocal tract transfer function. Thus, this second set of LPC parameters could be used in a synthesizer driven by an impulse function. In this case, no zero modeling is required. Because of the variety in glottal waveshapes, we recommend the use of the third approach.

The major need for zero modeling was determined to be for phonemes in which the acoustic channel has a side branch (nasals included). Here, the major goal of zero modeling is to produce a residual that permits simple pitch extraction.

Preliminary efforts were directed toward methods of zero determination. Methods based on solution of quadratic equations and root-finding of a polynomial were found in the literature (Gersh and Luo, 1972; Hsia and Landgrebe, 1967).[3,4] An adaptive gradient technique that avoids the above complex operations was also found in the literature (Melsa et al., 1973).[5]

We made no attempt to implement any of the zero-finding algorithms in the Task 2 effort. The preliminary need for zero modeling was determined to be for characterization of the excitation function. As a result, further consideration of zero modeling was left for Task 3.

4. Simplification of the Gain Calculation

Adequate synthesized speech quality has been achieved by using a synthesizer excitation power level equal to the residual power level.

Although this approach does not guarantee that the synthesizer output power will match the input signal power perfectly, it does offer a sufficiently good approximation. As a result, the computational load is significantly reduced compared with the original estimation by Atal and Hanauer (1971).[2] Section V of this report presents more details on the gain calculation and the excitation function.

5. Comparison of Toeplitz Versus Non-Toeplitz Form Solutions

Both the Toeplitz form (Markel, 1972; Itakura and Saito, 1972)[6,7] and the non-Toeplitz form (Atal and Hanauer, 1971)[2] have been implemented on the PDP-10 computer. Each can be operated in a variety of modes with a user-selected number of coefficients and block size. Very good performance has been demonstrated with both forms. On the basis of the testing to date, it appears that the Toeplitz form is preferable because it is computationally simpler, particularly with respect to stability determination. However, modest differences in complexity are probably not significant for future systems in light of the great capability of LSI. The non-Toeplitz form appears to produce somewhat more desirable residual signals for pitch extraction; however, it does not solve the pitch-extraction problem (see Task 3 report). The Toeplitz approach is recommended for preliminary real-time demonstrations.

6. Innovations Representation

We concluded that the innovations representation of a random process offers a more generalized viewpoint that may provide useful insight for some speech-processing problems.[8] However, it is much more important to model the physical process accurately--e.g., to include zeros or to use the proper number of coefficients--than to develop sophisticated statistical representations. Consequently, only a modest effort was devoted to the innovations approach.

Specifically, the energy of innovation process, i.e., the residual, was determined to be an extremely useful measure of the quality of the parameter estimation. In fact, this approach led to the DELCO compression algorithm discussed in Section VI and Appendix A and briefly described above. However, no other significant contribution to computational load or data reduction was found by considering the innovations representation.

C. Outline of the Report

The preceding section briefly summarized our research results by study areas. The rest of this report presents the details of our research. Quality considerations of pitch-synchronous analysis and synthesis are considered in Section II. Section III discusses time-domain pitch extraction. Pitch-accuracy requirements are presented in Section IV. The LPC synthesizer excitation function recommendations and results are developed in Section V. The adaptive data compression system, DELCO, is discussed in Section VI. Section VII presents our conclusions.

II PITCH-SYNCHRONOUS ANALYSIS AND SYNTHESIS TECHNIQUES

Conventional linear predictive coding algorithms (Atal and Hanauer and Markel) have concentrated on methods that attempt to characterize not only the vocal tract transfer function but the glottal source itself.[2,6] Thus, the synthesizing filter, when driven by a series of impulse functions at the pitch rate, attempts to reproduce the short-term power spectrum of the speech. Both the excitation spectrum and the vocal tract power transfer functions are represented. This statement holds for both the non-Toeplitz matrix (Atal and Hanauer) and the Toeplitz matrix (Markel) solutions to the problem.

Makhoul has shown that the formulation of the Toeplitz-form matrix equations tends to estimate the peaks of the spectral envelope with great accuracy, while the nulls or dips are estimated less accurately.[9] This performance is well matched to human perception. Thus, it appears that the conventional LPC analysis does what is desired. However, the above result is derived on the assumption of white noise excitation under steady-state circumstances so that it is meaningful to discuss power spectra. In practice, only a short segment of speech, perhaps 30 ms at most, is analyzed. Furthermore, most of the time, the excitation is not white noise but rather is one or two pitch pulses, or possibly several for a high-pitched speaker. Since the analysis is conventionally performed on a pitch-asynchronous basis, different phasings or timings of the excitation with respect to the analysis interval can occur. Thus, depending on this timing, somewhat different estimated short-term power spectrum envelopes may result from the analysis when, in fact, there is no change in the power spectrum.

The best solution to this problem is to increase the analysis period so that more excitation pulses are present. With a sufficient number of pulses, the timing of the pulses with respect to the analysis window is not crucial. Furthermore, the concept of power spectrum becomes more meaningful. Unfortunately, this solution hurts the transient response of the analysis system; i.e., it may not be possible to track rapid transients in the speech spectra with the larger window.

To avoid the sluggish time response of large window analyses and yet avoid timing-induced distortion, SRI has extensively studied pitch-synchronous analysis.* Nominally we used a rectangular window over one pitch period for a Toeplitz-form LPC analysis to derive the LPC coefficients. However, at a later point in our research, we employed a larger Hamming window over three pitch periods. This resulted in an overlapped analysis, since we performed a new analysis each pitch period. A Hamming window was not used unless an overlapped analysis was employed. Otherwise low value nulls caused by the window might have suppressed important data, e.g., when the glottal pulse occurred during a window null.

An advantage of pitch-synchronous analysis is that pitch-synchronous synthesis can be used without the necessity of interpolation. There is considerable debate among the speech community about the necessity for pitch-synchronous synthesis. However, most agree that the synthetic speech quality is not degraded. There is general agreement, too, that if interpolation is required for pitch-synchronous synthesis, one must be very careful about the interpolation technique. A poor interpolation system may do more harm than good. The basic problem is that linear interpolation of LPC parameters, or reflection coefficients, does not

---

*The subroutine EPOCH, which sets up the analysis and synthesis from the pitch marks, is described in Appendix B.

correspond to linear interpolation of the power spectrum. The desired result could be achieved by solving for the poles of the LPC polynomials and linearly interpolating these poles. Unfortunately, this is a messy computational procedure requiring something like a Newton-Raphson root-finding technique. Note that Markel has obtained quite good synthetic speech simply by linearly interpolating the reflection coefficients.[10]

The advantages of the pitch-synchronous analysis approach are that:

- No interpolation of parameters is necessary.
- The calculated LPC parameters will remain constant when the speech process is stationary.

The disadvantages of pitch-synchronous analysis are:

- Variable analysis window size, which causes algorithm complexity.
- Asynchronous rate of generating LPC parameters, which results in an asynchronous data rate.
- Higher transmission rates when parameters are encoded each period, a problem particularly for high-pitched speech.
- Additional analysis system complexity, since pitch marks are necessary before a pitch-synchronous analysis can be performed.
- Incompatibility with many popular pitch-extraction techniques (e.g., autocorrelation) that provide relative--as opposed to absolute--pitch marks.*
- Incompatibility with LPC techniques that do not use pitch extraction, such as RELP (see Task 3 report).

---

*With overlapped analyses, the use of a Hamming window makes the significance of a window of precisely three-pitch periods of dubious value. However, there may be some value in having the window always in the same relative position with respect to the glottal pulse.

We performed extensive pitch-synchronous analysis/synthesis simulations and demonstrated that very high quality synthesis is achievable. The output is virtually indistinguishable from the input speech. However, this high quality was achieved at the price of the disadvantages listed above. If these disadvantages are significant enough (as it now appears), pitch-synchronous analysis will not be used for practical real-time vocoders. Nevertheless, pitch-synchronous analysis/synthesis serves as a useful standard of quality that other more practical systems should strive to achieve. With good pitch extraction and excitation, the only quality degradation is due to the assumptions of the LPC speech model itself, e.g., no zeros appear in the model.

The concept of pitch-synchronous analysis/synthesis is critically dependent on precise, absolute pitch-mark placement. Time-domain pitch extraction is briefly described in Section III of this volume and is described in considerably more detail in the volume devoted to Task 3. The required accuracy of pitch-pulse placement is discussed in Section IV of this volume.

An important point is that pitch marks are placed in unvoiced intervals--during the aspiration after a stop release, for example. Pitch marks, rather than periods, are stored within our computer simulation. Thus, pitch is considered from a time-domain viewpoint (i.e., the excitation required to produce a given waveform), rather than from the prosodic viewpoint of speech analysis systems.

The excitation system is generalized with respect to conventional approaches to include a mixture of noise and pulses. Section V discusses the excitation system in more detail.

III TIME-DOMAIN PITCH EXTRACTION

Pitch extraction has always been a fundamental and difficult research problem of speech analysis, in general, and of vocoder design and implementation in particular. Linear predictive vocoder techniques have yielded significant improvement in vocal tract modeling and, hence, have intensified the need for good pitch extraction. The first sentences on the analog tape accompanying this report demonstrate the good quality achieved by an LPC vocoder when the pitch extraction is done by a human operator using a high resolution CRT interactive display. The sentences were chosen to test a range of difficult speech sounds (such as nasals, vowel glides, and semi-vowels) and are typical of general American conversational speech.

Through numerous experiments performed with interactive hand marking of pitch pulses, we have pinpointed several requirements that a high quality pitch extractor should satisfy. First, pitch marks are desirable for some aperiodic speech signals. Good examples of these transients are (1) stop releases, (2) the first voiced segment in a consonant-vowel transition, and (3) utterance-terminal voiced signals with low amplitude and vocal fry, i.e., erratic pitch. Second, during most significant portions of speech, the pitch estimates should vary smoothly. Based on experience with our data base, the acceptable rms pitch deviation from the true pitch is approximately ± 2 Hz.* "True" pitch is defined by the hand-marked pitch pulses that produce synthetic speech virtually indistinguishable from the original.

---

*The lowest pitch of our data base is approximately 100 Hz. If the data base were expanded to include a speaker with a 50-Hz pitch, the required accuracy is expected to be ± 1 Hz.

Computing the average period over a large window, e.g., by the autocorrelation method, may satisfy the desired smoothness requirements. However, it may not accommodate the required transient situations for high quality synthesis. Indeed, some LPC synthetic speech has a monotone quality when based on a large window autocorrelation-function pitch extractor. The SIFT algorithm of Markel attempts to handle these transient situations by dividing the normally large window into subintervals, each characterized by a particular excitation function type.[11] We believe this artificial approach would not be necessary with the correct representation of pitch pulses.

In contrast to the compromises inherent in correlation pitch extraction, we believe that it is possible to obtain superior performance (at the price of increased bit rate or complexity, or both) by using time-domain pitch extraction. Time-domain techniques are capable, in principle, of yielding smooth pitch and also of marking transient periods accurately. Time-domain pitch marking is described more completely in Sections II, A, 3 and II, E of the Task 3 report. Here we summarize the basic ideas briefly. Time-domain pitch marking is normally done in two stages: first, locate the largest magnitude peak in a 2- to 10-ms window, and second, place the pitch mark at some repeatable feature of the waveform near the large peak. The repeatable feature could be (1) the zero crossing preceding the peak, (2) the peak itself, or (3) the estimated point of transition from a decaying to a growing signal. In general, interactive hand marking of pitch makes use of all these approaches. Each result is tested to see if it meets the smoothness requirement. If none does, it is necessary to use a combination of the above. As one might suspect, the above process is complex, and necessarily so, due to the wide diversity of the possible speech signals. Consequently, our experience indicates that the time-domain approach to pitch extraction is not well suited to implementation as a real-time automatic algorithm.

Nevertheless, it is extremely useful as a laboratory tool and provides a good reference for the best achievable performance with LPC synthesis.

The complexity of time-domain pitch extraction can be simplified by performing preprocessing (filtering) of the speech. Three basic types of filtering may be employed: (1) inverse filtering, (2) low-pass filtering, and (3) formant-isolation filtering. Each of these is described in greater detail in the Task 3 report. Here we simply summarize the results of the research effort.

In general, inverse filtering is an effective method of reproducing the glottal waveshape (and thereby simplifying the time-domain pitch extraction). Analysis over a 20- to 25-ms window on preemphasized speech is necessary. This approach encounters difficulties when significant phase distortion (due to the acoustic environment, for example) exists or when the speech character is rapidly changing so that the window is too large to accurately characterize the speech.

Low-pass filtering the speech, e.g., to a bandwidth of approximately 600 Hz, can significantly simplify the problem of pitch extraction in the time domain. Unfortunately, our experience has been that pitch marking on this baseband signal is not adequate to provide the desired high quality synthesis. Nevertheless, when combined with other information, the results can be useful in estimating the pitch-pulse marks.

Formant-isolation filters can be used with significant performance improvement. However, the complexity of this system is prohibitive for real-time automatic pitch extraction. Formant isolation when combined with low-pass filtering can be used as an effective method of hand marking pitch pulses. It should be noted that at present the process of hand marking pitch pulses can be greatly shortened by using the formant-isolation approach. The reader is referred to Section II, E of the Task 3 report for more details on this subject.

IV TIMING REQUIREMENTS FOR HIGH QUALITY REPRODUCTION

In this section we consider the timing requirements for successful speech reproduction. An accompanying analog tape (see Appendix C for a detailed description) illustrates the effects described here. The output is from our LPC vocoder simulation program residing in the SRI-AI PDP-10 computer system. In all cases, the input speech was band-limited to 4 kHz, sampled at a 10-kHz rate, and preemphasized, i.e., one point differenced, in software. The analysis procedure used 14 coefficients and applied a Hamming window for most data. However, some analysis schemes based on one pitch period used a rectangular window. (All the utterances on the attached tape used overlapped analysis with a Hamming window and pitch-synchronous analysis/synthesis.) The synthesizing filter was of the lattice type described by Itakura.[7] The excitation was determined by the ratio method described in Section V.

The following subsections study the effects of (1) analysis block (window) length and (2) pitch accuracy.

A. Analysis Window Size

The first set of utterances analyzed in the pitch-synchronous analysis/pitch-synchronous synthesis (PSA/PSS) mode used a rectangular data window of one pitch period. A rectangular time window does not have good skirt selectivity in the frequency domain. Consequently, the spectral estimates derived from such an LPC analysis are only approximate. One method of alleviating this problem is to use a Hamming window. However, without overlapping, significant segments of data may be missed due to window nulls. These nulls cannot be avoided unless overlapping and a higher analysis block refresh rate are employed. Of course, this

results in a higher transmission rate unless larger windows are used. Normally, larger windows are used and an increased response time to transient effects results.

Experiments were performed both with a rectangular window of one pitch-period duration and with an overlapped Hamming window of three pitch-period duration, with a new analysis performed each pitch period. These tests were done in the PSA/PSS mode. Very good quality resulted in both cases. However, the overlapped analysis approach appeared to be less sensitive to the precision of pitch-pulse marking. For very low-pitched speakers, the overlapped approach might not yield a sufficiently good transient response to handle very rapidly changing speech segments. For our data base, which had a lowest pitch of approximately 100 Hz, no problems were encountered. Consequently, on the basis of our experiments, we would recommend an analysis window size of 20 to 30 ms, with 25 ms a desired goal.

Use of pitch-asynchronous analysis over a fixed window size may result in slight quality degradation. However, the advantages of a fixed (rather than a pitch variable) window size are significant, in a practical sense. As a result, we recommend a window size of 25 ms, with a new set of coefficients calculated every 10 or 15 ms. The optimum value must be determined on the basis of extensive testing with the adaptive data-compression algorithm DELCO (see Section VI of this report).

B. Pitch Accuracy Requirements

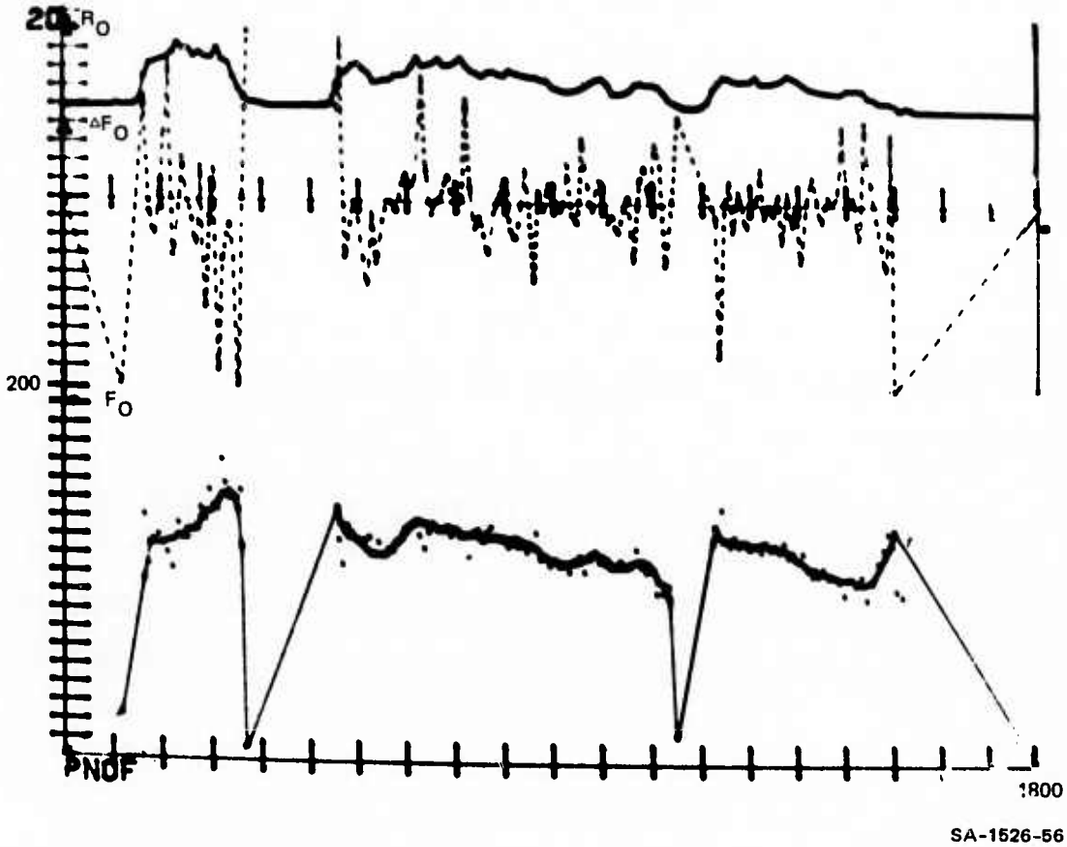
Conflicting estimates of the required pitch accuracy are given in the literature. Gold and Rabiner indicate that pitch marks must be placed within 100 μs of true position.[12] Markel places his requirements in the frequency domain.[11] In describing the SIFT algorithm he concludes that the fundamental frequency estimates must be within 7 Hz of the correct

value. For a nominal pitch of 100 Hz, this corresponds to an accuracy of 655 μs. Thus, a considerable difference exists between these two estimates. Consequently, we performed several experiments on our data base using the LPC synthesizer approach.

Two utterances from our data base were particularly difficult to reproduce without noticeable roughness. We found, by iteratively hand placing pitch marks, that it was possible to produce a very smooth trace for the fundamental frequency. This trace, for utterance number one, is shown as the solid line in the bottom trace of Figure 1. This set of pitch marks (called DTG in our file notation system) was taken to be the true or best estimate of the pitch function. A real-time algorithm would have great difficulty in generating a set of marks as good because of the iterative process used.

Two additional sets of pitch marks were compared with the best or smooth set (file DTG) to determine if it is possible to achieve adequate quality with simpler algorithms. The first set (called DTM) was determined from the unprocessed speech by a simple minimum-phase criterion; that is, the pitch marks were placed so as to make the speech signal appear to be a minimum-phase waveform over the pitch period, i.e., a decaying waveform. This pitch-marking philosophy was adopted since it seemed best suited to the basic assumptions of the LPC approach. The fundamental frequency estimates based on this hand-marked, minimum-phase philosophy are shown in the bottom trace of Figure 1 as the series of dots scattered about the solid line representing the best hand-marked set (file DTG). The middle trace shows the frequency difference between the two sets of pitch contours. The top trace shows the envelope of the sentence.

The standard deviation for the period differences is 400 μs, and the standard deviation of the fundamental frequency differences is 5.3 Hz.



SA-1526-56

FIGURE 1 OSCILLOSCOPE TRACES OF: (A) TOP TRACE – ENVELOPE OF SPEECH SIGNAL, (B) MIDDLE TRACE – FUNDAMENTAL FREQUENCY DIFFERENCE BETWEEN PITCH MARKS IN FILES DTG AND DTM, AND (C) BOTTOM TRACE – SOLID LINE, PITCH CONTOUR FOR THE BEST SET OF HAND-MARKED PITCH PULSES (FILE DTG) AND DOTTED LINE, PITCH CONTOUR FOR PITCH MARKS BASED ON A MINIMUM-PHASE CRITERION (FILE DTM)

The quality of the synthetic speech generated on the basis of the simpler, minimum-phase, nonautomatic, pitch-marking algorithm is substantially worse than that generated on the basis of the best pitch pulses. The degradation is perceived as a roughness in the synthetic speech.

A second set of pitch marks (called DTO) was generated from a bandpass-filtered version of the input speech using formant isolation filters (see Task 3 report). The pitch marks were placed at zero-crossings preceding the largest peak in the waveform in an interval corresponding to the estimated pitch period. An attempt was made to smooth the period estimates but not with the same care and effort as were used for file DTG.

Figure 2 (same format as Figure 1) is a photograph of a CRT display comparing the DTG and DTO files. The standard deviation for the period difference is 200 μs, and the standard deviation of the fundamental frequency difference is 2 Hz. Perceptually, the two sets of pitch marks produce indistinguishable synthetic speech.

As a result of these and other experiments, we conclude that a pitch accuracy of 2 Hz (standard deviation) is adequate for high quality synthesis. Poorer accuracy will result in a perceptual roughness of the synthetic speech. The utterances on the tape compare the three cases described above. The reader (listener) may judge the significance of the roughness effect.

The tape also includes two additional synthetic speech utterances that were generated to determine whether the roughness was caused by poor analysis windows or by poor accuracy excitation. The first synthetic speech utterance used rough (DTM) pitch marks for analysis and smooth (DTG) pitch marks for synthesis. The second utterance used smooth (DTG) pitch marks for analysis and rough (DTM) pitch marks for synthesis. The reader (listener) can readily determine that no quality loss results

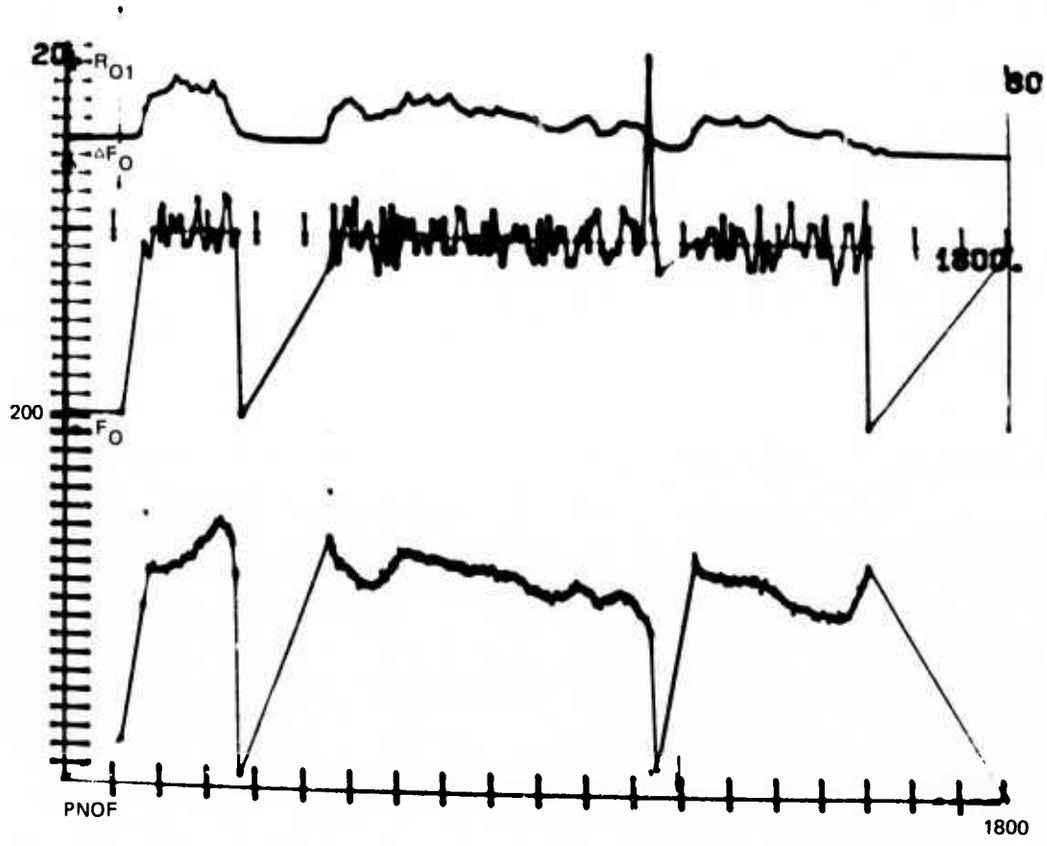


FIGURE 2 OSCILLOSCOPE TRACES OF: (A) TOP TRACE – ENVELOPE OF SPEECH SIGNAL, (B) MIDDLE TRACE – FUNDAMENTAL FREQUENCY DIFFERENCE BETWEEN PITCH MARKS IN FILES DTG AND DTO, AND (C) BOTTOM TRACE – SOLID LINE, PITCH CONTOUR FOR THE BEST SET OF HAND-MARKED PITCH PULSES (FILE DTG) AND DOTTED LINE, PITCH CONTOUR FOR PITCH MARKS DERIVED FROM A SMOOTHED ESTIMATE OF PITCH BASED ON THE OUTPUT OF A LOW-PASS FILTER (FILE DTO)

from the use of rough analysis pitch marks. However, use of the rough pitch marks for the excitation function results in synthetic speech with a rough quality. Thus, we conclude that the roughness results from the excitation function.

V LPC SYNTHESIZER EXCITATION

Conventional channel vocoders use either buzz (pitch pulses) or hiss (random noise) excitation depending on whether voiced or unvoiced synthesis is being performed. This concept has been extended to the original LPC analysis/synthesis systems as well, with reasonably good results.

Part of our research effort was devoted to considering improvements in the excitation function. The most obvious modification is to use a mixture of noise and pulses for the excitation. From a decision-theory point of view, this mixture has the obvious advantage of avoiding catastrophic failures when a V/UV error is made. Instead, the "soft" character of the processing (estimation as opposed to decision) should provide graceful degradation.

Another major advantage is that speech does not consist of solely voiced or solely unvoiced segments. Perhaps the best known example of a different segment is the voiced fricative. Here the excitation is a composite of noise (due to turbulence produced by a constriction) and pulses (due to the action of the vocal cords). Other lesser known cases exist. For example, Fujimura found that many voiced sounds contain unvoiced power in certain portions of the frequency spectrum.[13] Thus, a mixture of noise and pulse excitation appears to provide a better approximation to the true excitation source.

We have developed an excitation function that is just such a mixture of random noise and pulses. The ratio of noise to pulse power is controlled by the normalized error or residual energy, ERRN. The reasoning is that voiced processes are more predictable than unvoiced processes.

Consequently, the normalized (the normalization is required since voiced signals generally have much higher power than unvoiced signals) error energy for voiced signals should be much less. Atal and Hanauer confirm that this is a valid approach.[3]

Many relationships between the ratio of the noise and pulse powers (RATIO) and ERRN have been tried. Through our experiments, we have found that the following characteristic (shown in Figure 3) provides the optimum performance. The ratio of the noise energy, E_n, to the sum of the pitch pulse plus noise energies, $E_p + E_n$, is defined as the variable

$$\text{RATIO} = E_n/(E_n + E_p) \quad .$$

Below a value of ERRN = 0.250,

$$\text{RATIO} = 16\ (\text{ERRN})^2 \quad .$$

For ERRN $\geq$ 0.250,

$$\text{RATIO} = 1.0 \quad .$$

That is, if the normalized error energy exceeds 0.250, only hiss excitation is used. For smaller values of ERRN, the excitation rapidly converges to consist primarily of pulse energy.

The excitation requires the information given by RATIO plus the residual energy, $E = \text{ERRN} \cdot R_o$--where R_o is the input signal power over the analysis window. With this information the proper absolute energy can be applied to each source. Note that our excitation power formula is based on ERRN where

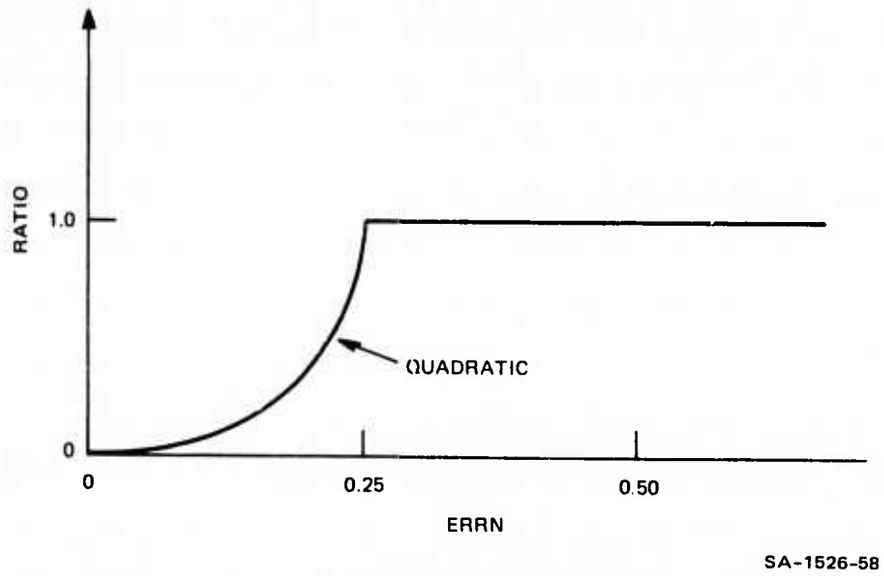


FIGURE 3 RATIO OF NOISE ENERGY TO SUM OF NOISE AND PULSE ENERGIES AS A FUNCTION OF ERRN

$$\text{ERRN} = 1 - \sum_{i=1}^{p} a_i (R_i / R_o) \quad .$$

This value corresponds to the true normalized residual energy for a non-Toeplitz-form LPC analysis. However, for the Toeplitz-form analysis that we conventionally use, ERRN is only an approximation to the correct value. Fortunately, for our size analysis window and for the number of LPC coefficients (14 or fewer), the approximation is quite good and high quality synthesis results.

A more serious approximation exists. The above excitation philosophy is based on the assumption that, if we match the excitation power, the output power will match the power of the input speech. Unfortunately, this result does not hold perfectly because of coherent, transient cancellation effects when the synthesizing filter coefficients are updated. That is, the decay response of the initial conditions left over from the

previous analysis period can coherently add or subtract from the present interval. Fortunately, the magnitude of the initial condition response is normally quite small compared with the impulse response. Nevertheless, the result is that the envelope function of the synthetic speech is considerably more jagged than the input speech. In fact, the dynamic range of the synthetic speech may be four times greater than the input speech.

The above effect leads to a certain harshness (perceptible under ideal listening conditions) in the synthetic speech. However, the procedure for resolving this minor problem is computationally complex; Atal and Hanauer describe this method of guaranteeing a power match between the input and synthetic speech processes.[2] Our conclusion is that the quality improvement is not worth the additional system complexity.

Our experiments with the excitation mixture concept indicate that very high quality synthesis can be achieved. In fact, the resulting synthetic speech is virtually indistinguishable from the input speech. Furthermore, the excitation mixture system appears more robust with respect to other system degradations. For example, some evidence exists that the presence of noise in the excitation signal tends to mask the roughness associated with pitch-asynchronous synthesis. As a result, we recommend the use of the excitation mixture concept.

VI ASYNCHRONOUS TRANSMISSION OF LPC PARAMETERS

A. Introduction

Speech is an inherently asynchronous time-varying process. The properties of the signal vary with the short-term properties of the particular utterance. It is well known that for various reasons speech contains pauses ranging in duration from a few milliseconds to several seconds. Similarly, we find that quasi-stationary portions of voiced speech over several excitation periods, e.g., over approximately 80 ms, are not uncommon. In contrast, we also find significant signal character changes occurring in one or two excitation periods. A characteristic of an adaptive speech compression system designed for asynchronous operation is a nonuniform data transmission rate commensurate with the varying properties of the input signal. An advantage over similar synchronous systems is the retention of a given quality of synthetic speech at a lower average bit transmission rate. An asynchronous system interfaces nicely with asynchronous-transmission circuits, such as those employing packet-switching techniques. The interface to an ordinary synchronous circuit requires data buffering to achieve the uniform transmission rate.

In this section we seek a measure, δ, of the change of signal properties in speech from one analysis frame to another. The transmission strategy is then to transmit new LPC parameters to the synthesizer only when δ (the change between the previously transmitted frame and the current frame) exceeds a predetermined threshold. Four candidate measures (δ_1, δ_2, δ_3, and δ_4) are defined, discussed, and evaluated. Experimental results are presented showing that the adaptive LPC transmission algorithm based on δ_4 yields at 50 percent to 70 percent reduction in bit

rate with negligible loss in speech quality. These results are found for many speakers, utterances, and types of LPC analysis. Statistical results describing the time between coefficient updates and the time between transmission of successive packets in a typical packet communication system are presented and discussed.

Appendix A describes how a particular adaptive compression algorithm (DELCO) that was developed in this research effort can be interfaced to a packet communication system.

B. The LPC Model

In most formulations LPC coefficients are used to model the combined effects of the glottal source, the vocal tract shape, and radiation characteristics. At a particular instant in time a speech sample, s(n), is approximated by a linearly weighted summation of the past p samples. That is,

$$\hat{s}(n) \approx \sum_{i=1}^{p} a(i) \cdot s(n - i) \quad .$$

The prediction error (or residual) is given by

$$e(n) = s(n) - \hat{s}(n)$$

and the linear predictive coefficients are found by minimizing the squared error summed over a given duration. The result is a set of p linear equations in terms of the autocorrelation coefficients. Depending on the precise formulation, the matrix of autocorrelation coefficients may be Toeplitz or non-Toeplitz in form. The impact of this difference is not great. (There are some complexity reductions for the Toeplitz-form case.) In either case the residual energy is given by

$$E = \sum_{n} e^2(n) = R_{oo} - \sum_{n} a(i)R_{oi} \quad .$$

The residual energy is an extremely important measure of system performance. Minimizing the residual error is the basis of the least-mean-square approach.[14] Magill has shown that the residual energy is the key variable in determining the optimum adaptive Kalman filter for all important performance criteria, e.g., minimum mean-square-error and maximum likelihood estimates.[15,16] Consequently, it was immediately recognized as the basis for a very effective adaptive data compression or adaptive sampling scheme for the LPCs.

Both Toeplitz and non-Toeplitz analysis assume that the speech process is stationary over short intervals (approximately 10 to 20 ms). Thus, the LPC model assumes a piece-wise stationary process. In addition, the LPC model assumes that the speech process can be adequately modeled by an all-pole (or autoregressive) source. To date there is no indication that the quality of the reconstructed speech is deteriorated by either of these assumptions.

Atal prefers to view the LPC analysis from the time-domain viewpoint.[17] However one can regard the LPC approach from the frequency domain equally well. In fact, Makhoul has shown that the Toeplitz form of LPC analysis matches the peaks of the envelope of the short-term power spectrum.[9]

For the purpose of evaluating the performance of algorithms for the adaptive transmission of LPC coefficients, we extensively use frequency-domain techniques, such as the short-term power spectra derived from LPCs [see Figures 5 through 10 and the graphs of frequency (formant) peaks, Figures 11 through 18]. Listening tests verify that preserving spectral properties gives good quality reproduction.

C. Description of Adaptive Measures

The problem is to determine a measure of the amount of change in vocal tract parameters from one analysis frame to another and then to use this information as a means of adaptively transmitting the LPC coefficients at a reduced transmission rate. Four measures are examined. For each measure a function, δ, is defined whose value is used to indicate the relative amount of change in coefficients between two analysis frames. A low value of δ should indicate similar vocal tract parameters over the two frames. A high value of δ should indicate that the vocal tract parameters for the two frames are substantially different. The first three measures (δ_1, δ_2, and δ_3) are computed directly from the LPC coefficients. These functions reflect various assumptions about the relationship between changes in vocal tract parameters and the changes in LPC coefficients. The fourth measure δ_4 considers the normalized residual energy over the nth analysis epoch using nonoptimum versus optimum coefficients. Although somewhat more computationally complex than δ_1, δ_2, and δ_3, δ_4 is based on the normalized residual energy and is consistent with the theoretical analysis of Magill.[15,16]

1. Adaptive Measures Based on the LPC Parameters or Transformed Versions of Them

Although the measures about to be described may operate on the coefficients a(i), the same measures may operate on the reflection coefficients or partial correlation (PARCOR) coefficients, k(i), of Itakura and Saito.[7] For reasons that will become evident as the discussion proceeds, we use the reflection coefficients, k(i). Note that in the following definitions, δ_1, δ_2, and δ_3 are not necessarily based on all of the k(i). Only the first few may be used.

Consider a q-dimensional subset of the coefficients k(i) as a discretely varying q-tuple of real numbers, K, on the inner product

space, H, of dimension q. The canonical inner product is defined to be $< U, V > = u(1)v(1) + u(2)v(2) + \ldots + u(q)v(q)$. The length of U is defined as $|U| = \sqrt{< U, U >}$. Let the superscripts m and n respectively denote the mth and nth analysis frames with $m < n$.

Measure 1 is simply the distance between the q-tuples K^n and K^m, i.e., the length of $K^n - K^m$. For Measure 1,

$$\delta_1 = |K^n - K^m| = \left\{ \sum_{i=1}^{q} [k^n(i) - k^m(i)]^2 \right\}^{1/2} .$$

Measure 2 is the length of $K^n - K^m$ normalized (divided) by the length of K^n,

$$\delta_2 = \frac{|K^n - K^m|}{|K^n|} = \left\{ \frac{\sum_{i=1}^{q} [k^n(i) - k^m(i)]^2}{\sum_{i=1}^{q} [k^n(i)]^2} \right\}^{1/2} .$$

Measure 3 is the length of $K^n - K^m$ where each component of $K^n - K^m$ is scaled by a factor inversely proportional to the magnitude of each component of K^n,

$$\delta_3 = \left\{ \sum_{i=1}^{q} \frac{[k^n(i) - k^m(i)]^2}{|k^n(i)|} \right\}^{1/2} .$$

2. Theoretically Optimum Adaptive Measure

Let the superscripts m and n respectively denote the mth and nth analysis frames with $m < n$, as in the previous section.

Measure 4 derives the function δ from a comparison of the normalized error energy over the nth analysis frame with the normalized error energy over the same analysis frame using the coefficients from the mth frame,* denoted by the vector A^m,

$$\delta_4 = 1 - E^n(A^n)/E^n(A^m) \quad .$$

$E^n(A^n)$ is the error energy using optimum coefficients for the nth frame,

$$E^n(A^n) = R^n(0,0) - \sum_{i=1}^{p} a^n(i) \cdot R^n(0,i)$$

and $E^n(A^m)$ is the error energy using nonoptimum coefficients for the nth frame,

$$E^n(A^m) = R^n(0,0) - 2\sum_{i=1}^{p} a^m(i) \cdot R^n(0,i)$$

$$+ \sum_{i=1}^{p} \sum_{\ell=1}^{p} a^m(i) \cdot a^m(\ell) \cdot R^n(i,\ell) \quad .$$

D. Transmission Strategy

For each measure, we hypothesize that a low value of δ will indicate similar vocal tract parameters over both the mth and the nth frames and that a higher value of δ will indicate different vocal tract parameters for the nth frame, compared with the mth frame. The transmission strategy is to send coefficients when δ exceeds a given threshold, γ, where the

---

* Note that this measure is simply a transformed version of the measure used in Appendix A. Here δ_4 is constrained to lie between 0 and 1, whereas in Appendix A the equivalent parameter DEL lies between 1 and ∞.

mth frame corresponds to the last transmitted coefficients and the nth frame is the current frame.

A typical synthetic speech waveform resulting from the above transmission strategy is shown in Figure 4. This figure presents three separate waveforms. The top trace is the envelope function associated with the utterance, "Add the sum to the product of these three." The two marks corresponding to the speech segment (duct) represent the interval that is shown in greater detail in the lower traces. The middle trace is the synthetic speech for this interval; the lower trace is the input speech during this interval. Note how the peaks of the input speech vary smoothly with time. By contrast, the synthetic speech peaks tend to follow a step-function-like contour. This is the case since the excitation power level is updated only when the LPC parameters are updated. Thus, by observing the middle trace one can see that new LPC parameters were transmitted at approximately 1.522, 1.535, 1.556, and 1.582 s. The step-like character of the envelope of the synthetic speech appears to the eye to be a significant distortion. Fortunately, it is virtually imperceptible to the human ear. Consequently, no attempt has been made to update the excitation power levels more frequently.

E. Empirical Evaluation of Coefficient Measures

A typical test case is shown in Figures 5 through 10. In all cases the lower trace, which shows the input speech, is the same. The upper trace, which differs from figure to figure, shows the power spectrum computed from different speech segments. The speech sample (lower trace) is the syllable "Pete," minus the initial stop release, taken from the utterance, "Pete Cooper's dog toyed with Dick Todd's cat." The LPC power spectra (upper trace) were computed by taking the reciprocal of the log magnitude spectrum of the inverse filter. The basic technique is described by Markel.[6] For the test cases, LPC coefficients were computed

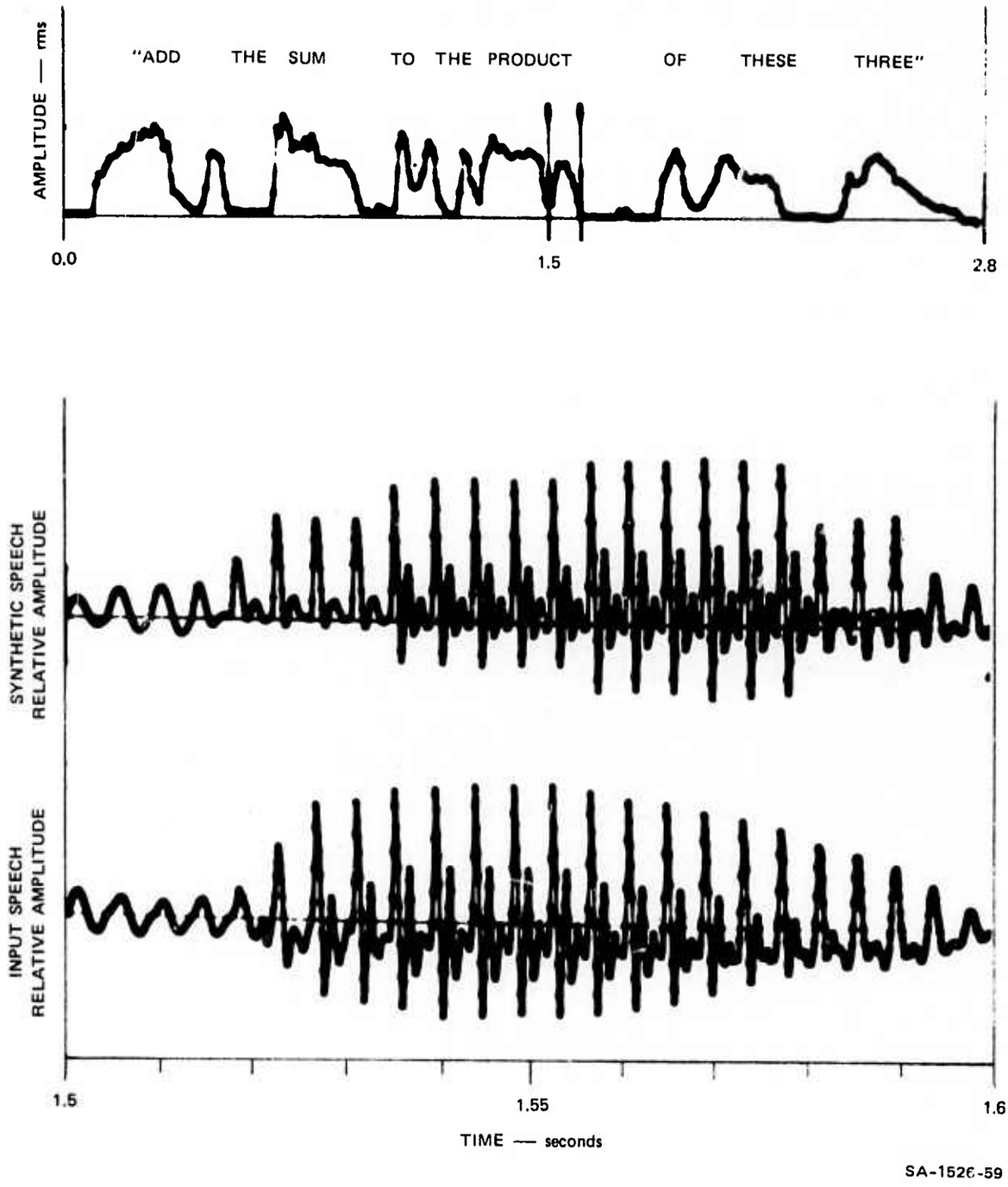


FIGURE 4 COMPARISON OF INPUT SPEECH AMPLITUDE WITH AMPLITUDE OF DELCO GENERATED SYNTHETIC SPEECH, MEASURE 4, Y = 0.3

Analysis type: PTOVR (pitch synchronous overlapped).

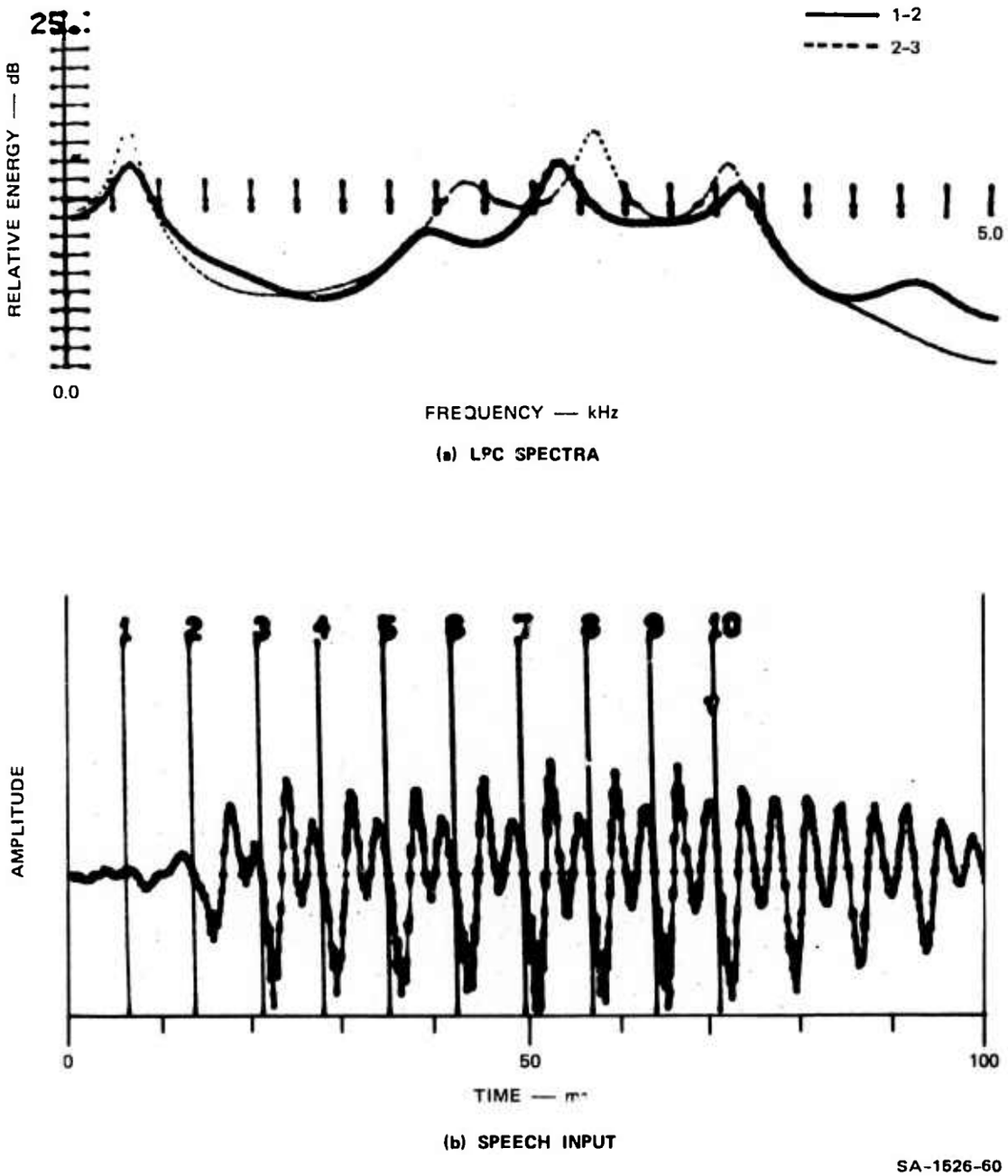


FIGURE 5 LPC SPECTRA OVER THE SYLLABLE "PETE"

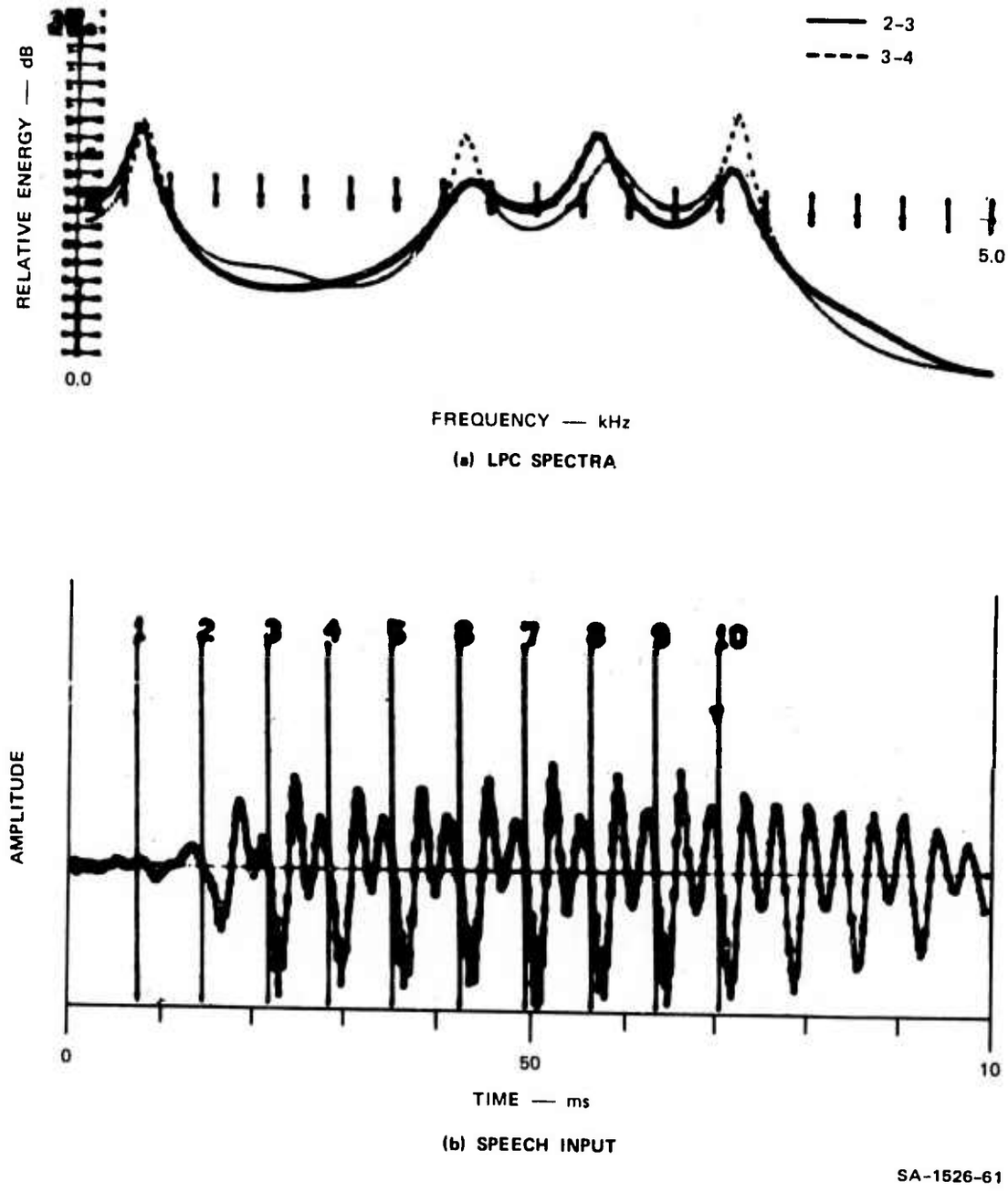


FIGURE 6 LPC SPECTRA OVER THE SYLLABLE "PETE"

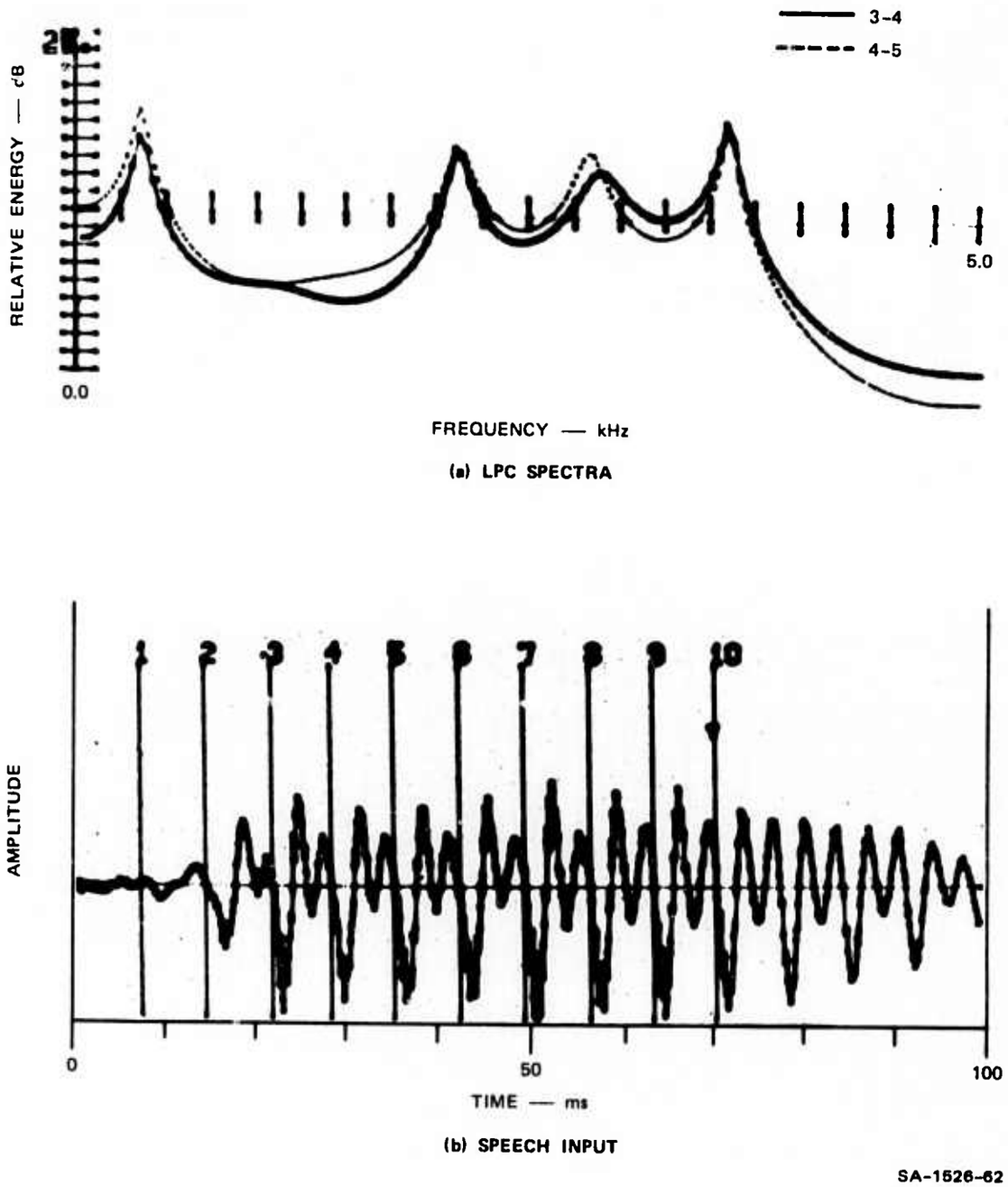


FIGURE 7 LPC SPECTRA OVER THE SYLLABLE "PETE"

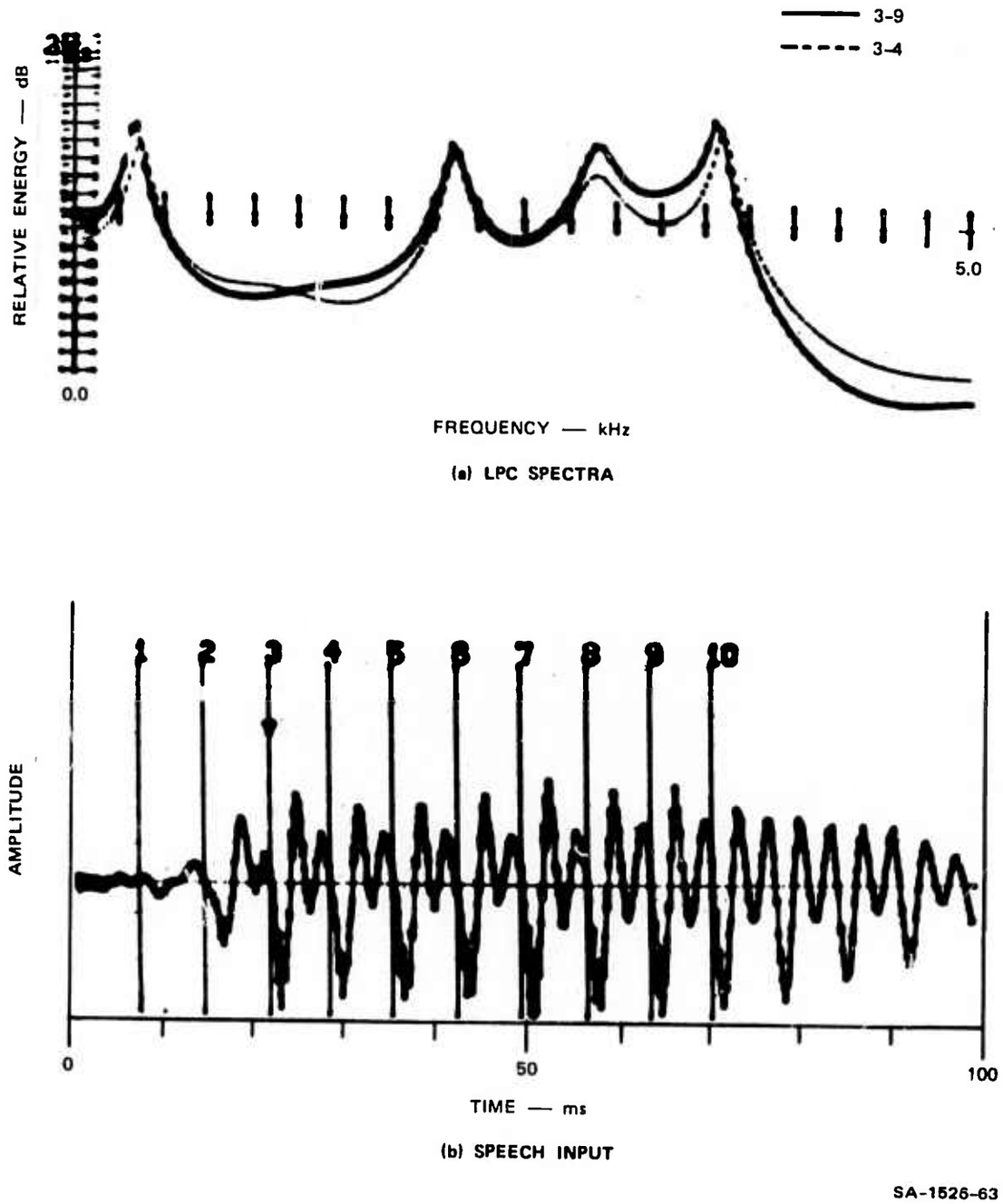


FIGURE 8 LPC SPECTRA OVER THE SYLLABLE "PETE"

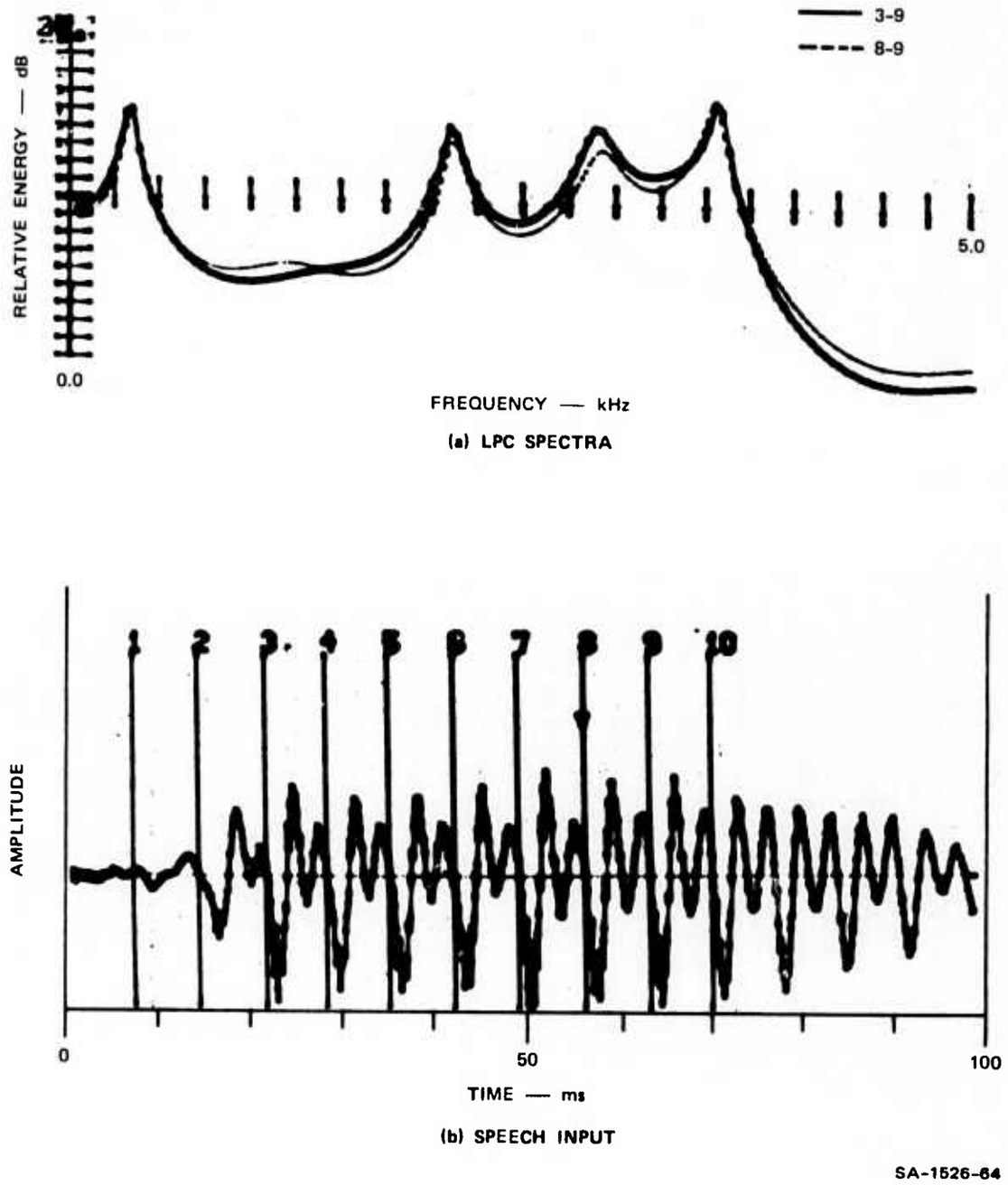


FIGURE 9 LPC SPECTRA OVER THE SYLLABLE "PETE"

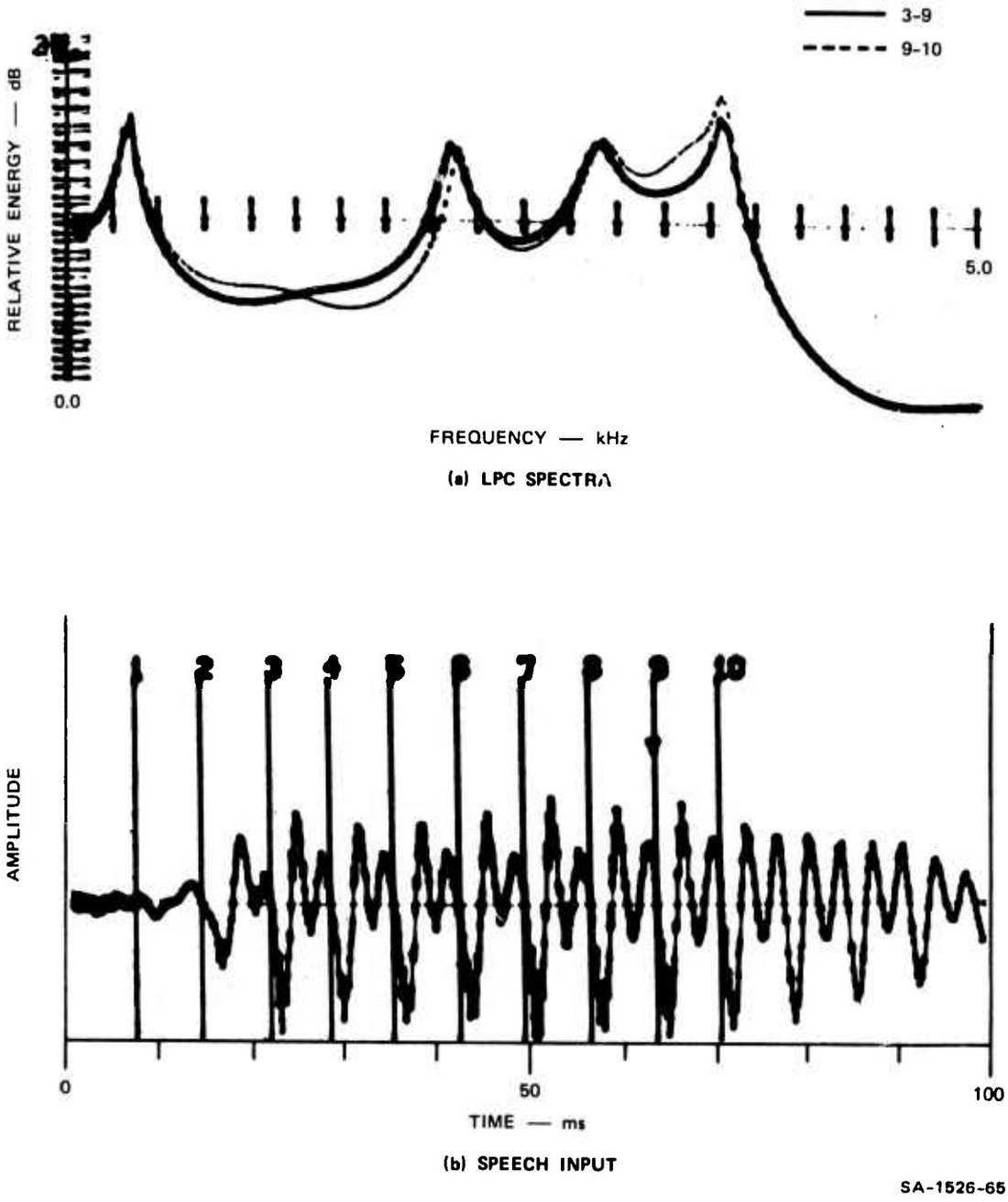


FIGURE 10 LPC SPECTRA OVER THE SYLLABLE "PETE"

using analysis systems denoted as either PTOVR (pitch-synchronous analysis using overlapped analysis frames, p = 14, with a Hamming window applied) or PTSYN (pitch-synchronous analysis, no overlap of analysis frames, p = 14, no Hamming window applied). The spectra were computed for 128 spectral points over the indicated marks with the vertical scale in decibels and the horizontal scale going to 5 kHz.

Note that the spectra of intervals 1-2 versus 2-3 (Figure 5) show transitions both in frequency (about 5 kHz/s for second and third formants) and in amplitude. For intervals 2-3 and 3-4 (Figure 6) the same is true but to a lesser extent. Commencing with interval 3-4, we experience a relatively slowly changing spectrum. Figures 7 through 10 show that a significant portion of the speech sample, say 3-9 or perhaps 3-10, may be approximated as being quasi-stationary. Aided by this information, we may compare results obtained using the various coefficient measures.

Tables 1 and 2 briefly summarize results obtained for the syllable "Pete" using PTSYN and PTOVR analyses for various threshold values. In each case the nth frame is the current frame and the mth frame is the frame at which time the coefficients were last transmitted. If the threshold is exceeded, we transmit coefficients; otherwise the previous coefficients are used. For example, the δ_4 measure with PTSYN analysis, $\gamma = 0.4$, transmitted a new set of coefficients at 187.9 ms (Index 1) and at 195.0 ms (Index 2). The set at 195.0 ms was used over the next four periods and then subsequently refreshed by a new set at 230.3 ms. The table dramatically points out the poor performance of some of the measures, e.g., Measure 2, PTSYN, $q = 4$, $\gamma = 0.4$ and Measure 1, PTSYN, $q = 4$, $\gamma = 0.3$. These measures transmit coefficients during the quasi-stationary portion of the signal when it is unnecessary.

Another performance evaluation is obtained by comparing the function δ given γ and the resultant transmission decisions with the graphs of

Table 1

SUMMARY OF TRANSMISSION DECISIONS FOR LPC COEFFICIENTS OVER THE SYLLABLE "PETE": PTSYN (Pitch Synchronous) ANALYSIS

Index Shown in Figures 5 Through 10	Time (ms)	Measure 1		Measure 2		Measure 3		Measure 4	
		q=4 γ=0.3	q=12 γ=0.5	q=4 γ=0.4	q=12 γ=0.4	q=4 γ=0.6	q=12 γ=1.5	γ=0.25	γ=0.4
1	187.9	1*	1	1	1	1	1	1	1
2	195.0	0†	1	0	1	1	1	1	1
3	202.4	1	0	1	0	1	1	0	0
4	209.1	1	0	1	0	1	0	0	0
5	216.1	1	1	1	1	1	0	1	0
6	223.3	1	1	1	1	1	0	1	0
7	230.3	0	0	0	0	0	0	0	1
8	237.5	0	0	0	0	0	0	0	0
9	244.5	0	0	0	0	0	0	0	0
10	251.4	1	1	1	1	1	0	1	0
-	258.4	0	1	0	1	1	1	1	0
-	265.5	0	0	0	0	0	0	0	1
-	272.4	0	0	0	1	0	1	1	0
-	279.7	1	1	1	1	1	0	1	1

*1--Transmit new coefficient set.

†0--Use last transmitted coefficient set.

Table 2

SUMMARY OF TRANSMISSION DECISIONS FOR LPC COEFFICIENTS OVER THE SYLLABLE "PETE": PTOVR (Pitch Synchronous Overlapped) ANALYSIS

Index Shown in Figures 5 Through 10	Time (ms)	Measure 1		Measure 2		Measure 3		Measure 4	
		q=4 γ=0.25	q=12 γ=0.3	q=4 γ=0.25	q=12 γ=0.3	q=4 γ=0.6	q=12 γ=0.75	γ=0.2	γ=0.25
1	187.9	1*	1	1	1	1	1	1	1
2	195.0	0†	0	1	0	1	1	0	0
3	202.4	1	1	0	1	0	0	0	0
4	209.1	1	1	1	0	0	0	0	0
5	216.1	0	0	0	1	0	1	0	0
6	223.3	1	1	1	1	1	1	1	1
7	230.3	0	0	0	0	0	0	0	0
8	237.5	0	1	0	0	0	1	0	0
9	244.5	0	0	0	0	0	0	0	0
10	251.4	1	1	1	1	1	1	0	0
11	258.4	0	0	0	0	0	0	0	0
-	265.5	0	1	1	0	0	1	1	0
-	272.4	0	1	0	1	0	1	1	1
-	279.7	1	1	1	1	1	1	0	0

*1--Transmit new coefficient set.

†0--Use last transmitted coefficient set.

frequency (formant) peaks. For visual representation of transmission decisions, the function δ^* is defined:

$\delta^* = \delta$ if $\delta \geq \gamma$ (transmit coefficients)

$\delta^* = \delta - 1.0$ if $\delta < \gamma$ (do not transmit coefficients) (biased downward)

Figures 11 through 18 show the decision process for two different utterances by different speakers using various analysis techniques. The uppermost solid trace is the rms power envelope of the speech signal. Immediately below it are three formant traces, frequency (each division corresponds to 250 Hz) versus time, which mark the location of the frequency peaks in the LPC power spectra computed at 10-ms intervals over an analysis window of 15 ms (overlapped analysis). The lower trace is a plot of δ^* versus time as computed for the particular run. When the formant traces remain constant, we expect to see a small number of occurrences where the decision is to change coefficients (δ^* biased downward). When the formant traces are changing, we expect to see a large number of occurrences where the decision is to change coefficients (δ^* remains unbiased).

F. Results

Of the four coefficient measures, best results are obtained using Measure 4 (the measure based on the residual signal energy). Extensive listening tests verify that high performance is maintained for male and female speakers over several utterances using both pitch-synchronous and pitch-asynchronous analysis. Using overlapping analysis frames introduces redundant data in the overlap period and produces a smoother δ_4 function. This allows for more accurate extraction of the changes in the vocal tract parameters than is obtained using nonoverlapped analysis frames. However, with the transmission strategy previously described, the

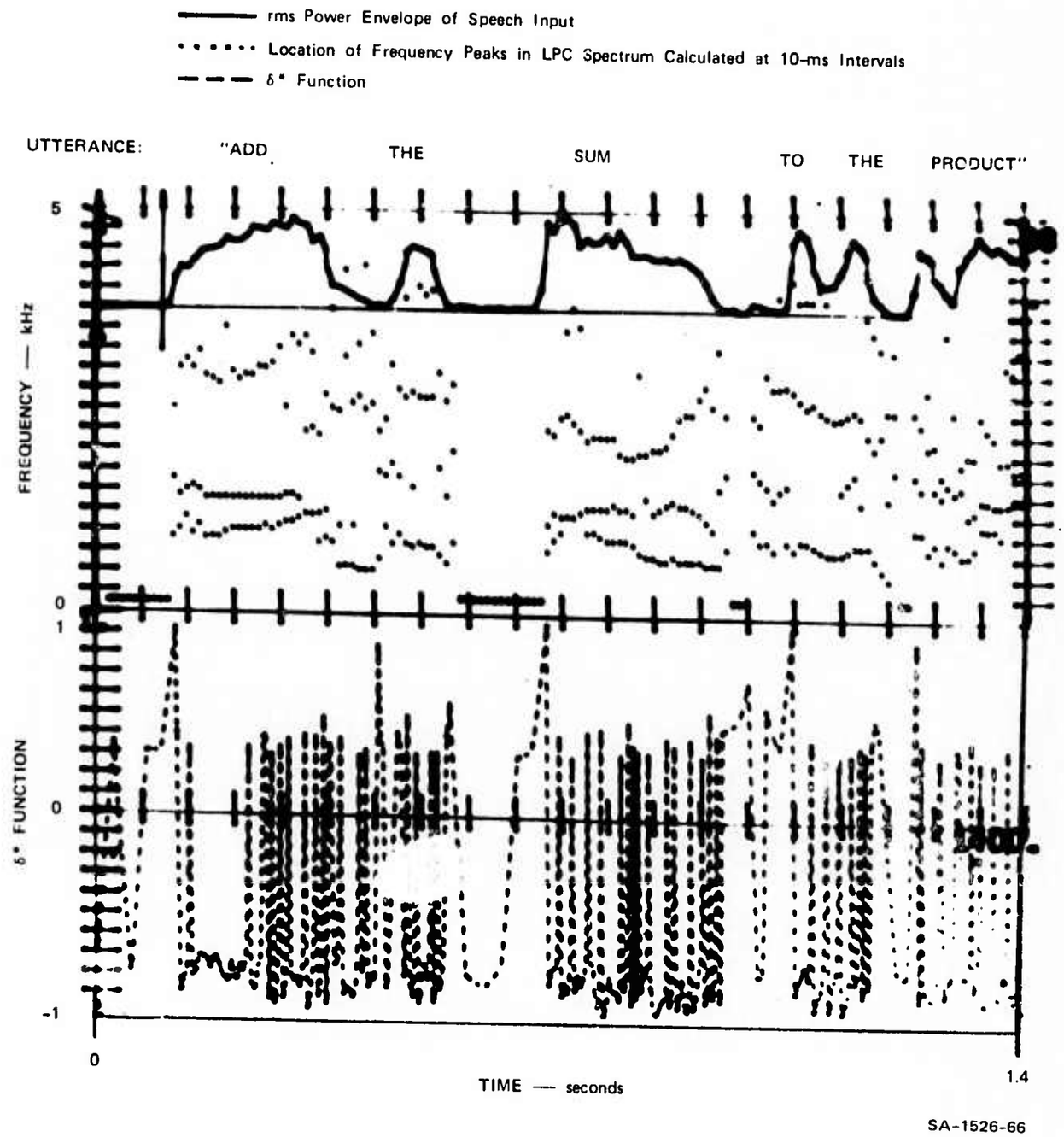


FIGURE 11 TRANSMISSION DECISION FUNCTION δ^* FOR MEASURE 4, Y = 0.3
Analysis type: PTOVR (pitch synchronous overlapped)

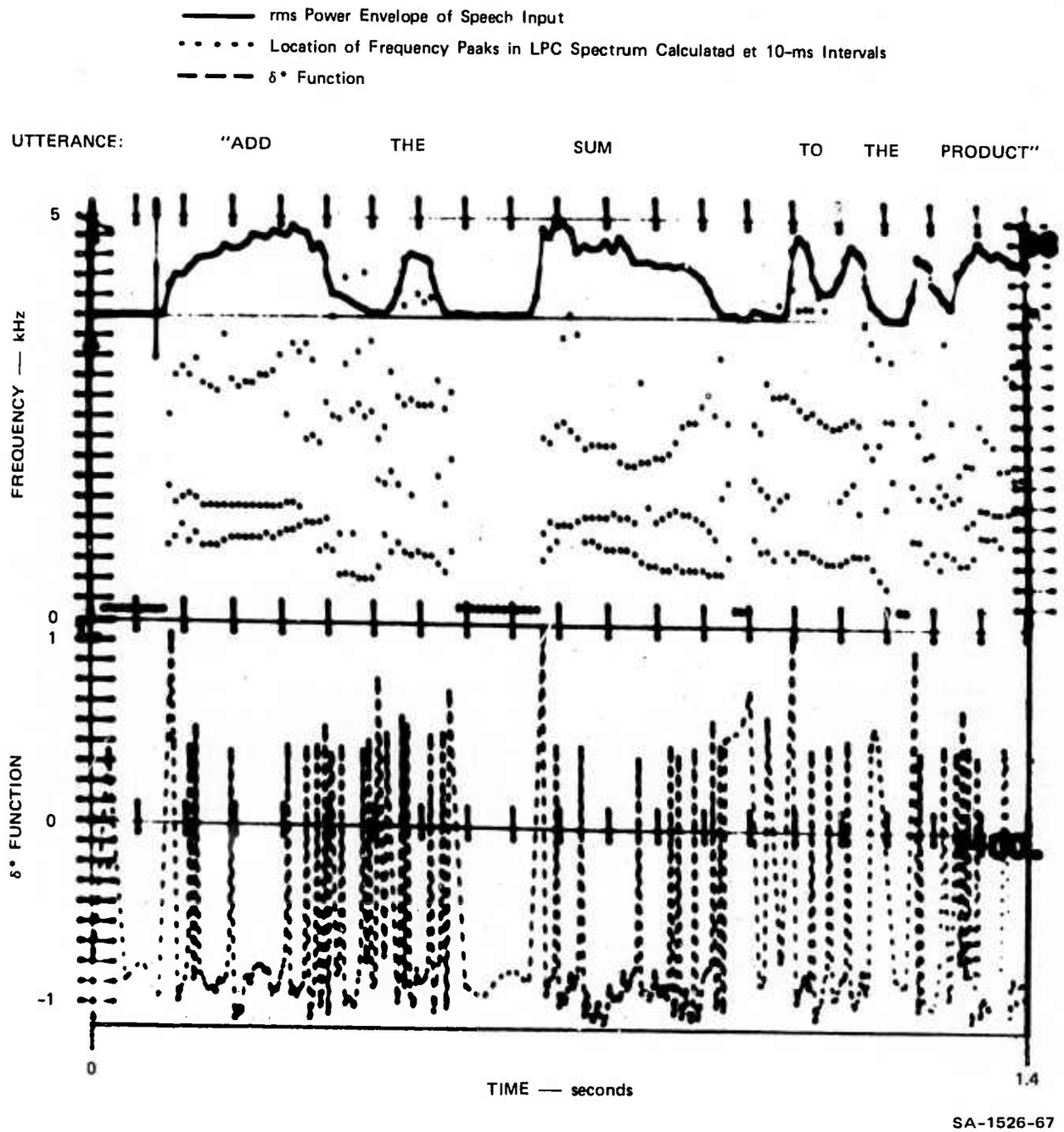


FIGURE 12 TRANSMISSION DECISION FUNCTION δ* FOR MEASURE 4, Y = 0.35
Analysis type: PTOVR (pitch synchronous overlapped)

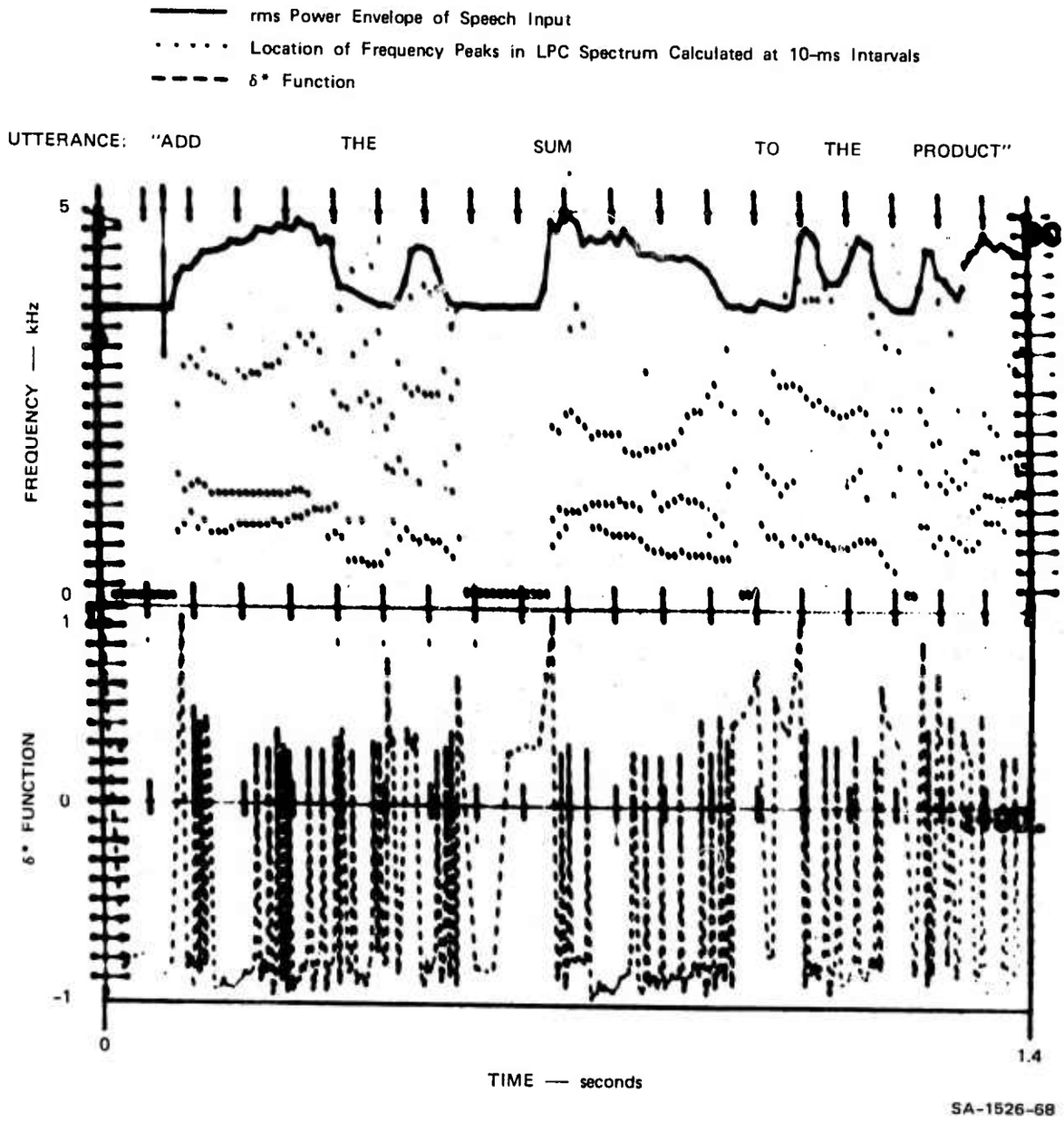


FIGURE 13 TRANSMISSION DECISION FUNCTION δ^* FOR MEASURE 4, Y = 0.25
Analysis type: PTOVR without pre-emphasis

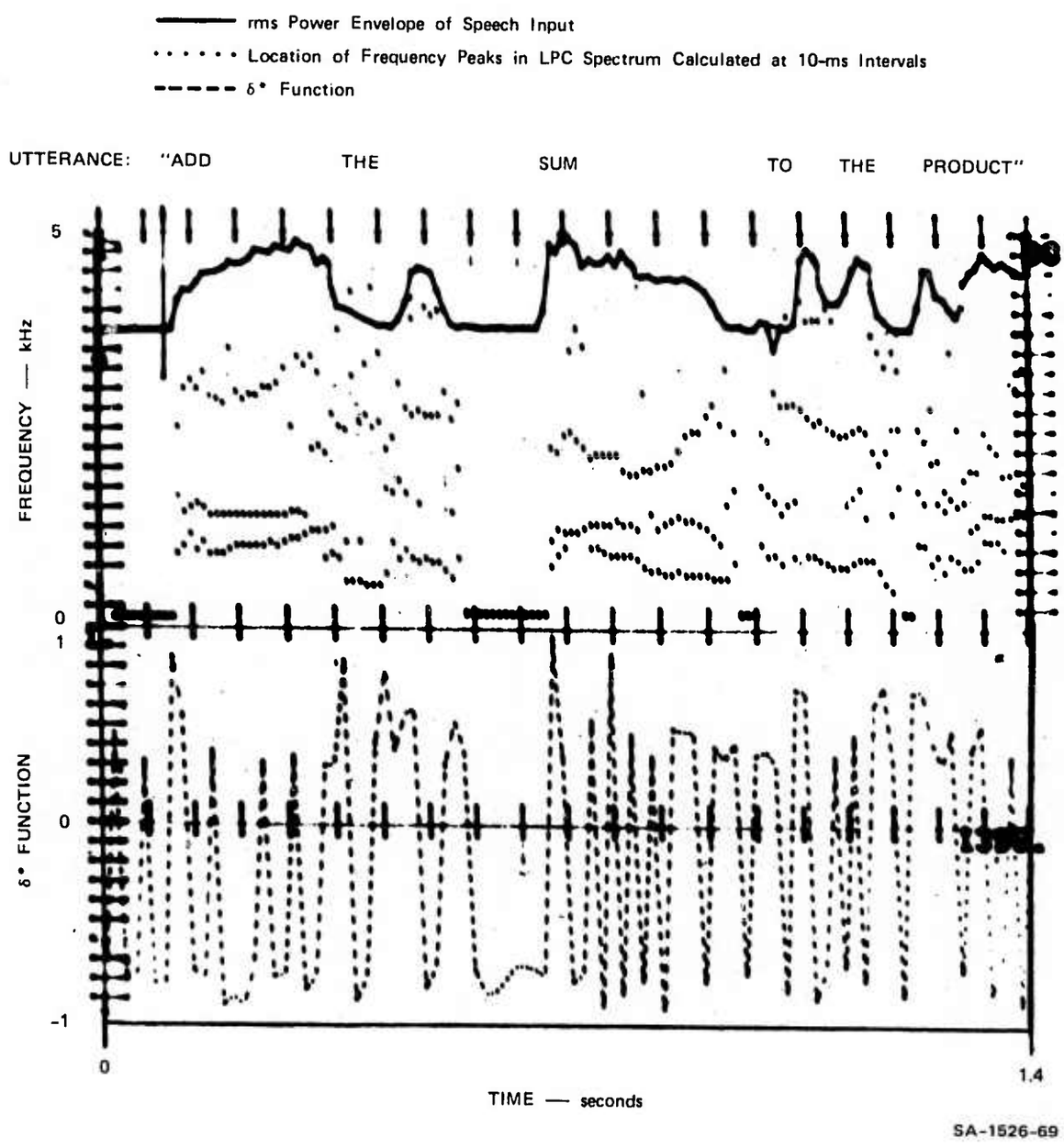


FIGURE 14 TRANSMISSION DECISION FUNCTION δ^* FOR MEASURE 4, Y = 0.3

Analysis type: Block synchronous using overlapped 25-ms analysis frames shifted at 15-ms

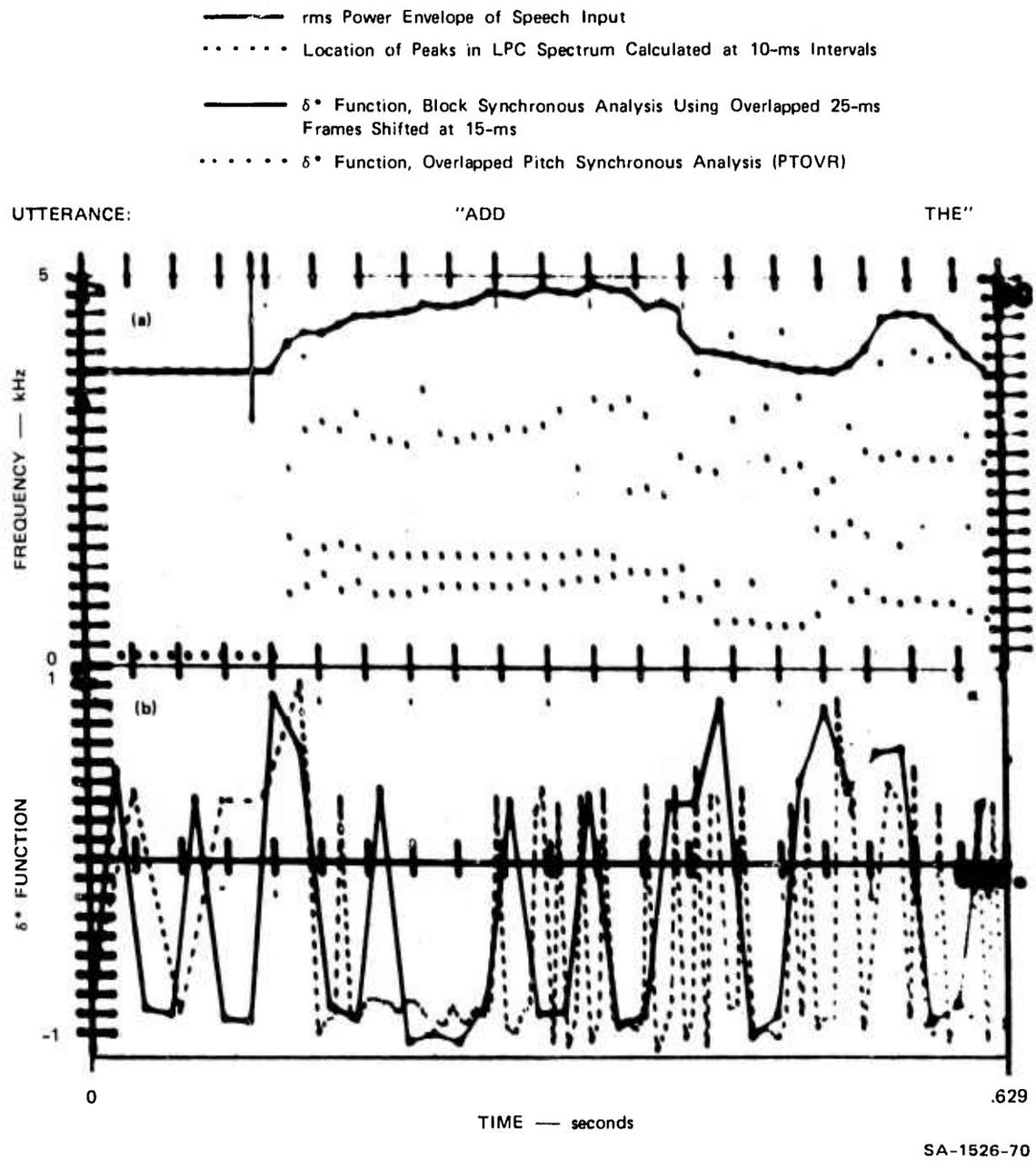


FIGURE 15 COMPARISON OF TRANSMISSION DECISION FUNCTION δ* FOR PITCH SYNCHRONOUS AND BLOCK SYNCHRONOUS ANALYSIS TYPES USING MEASURE 4, Y = 0.3

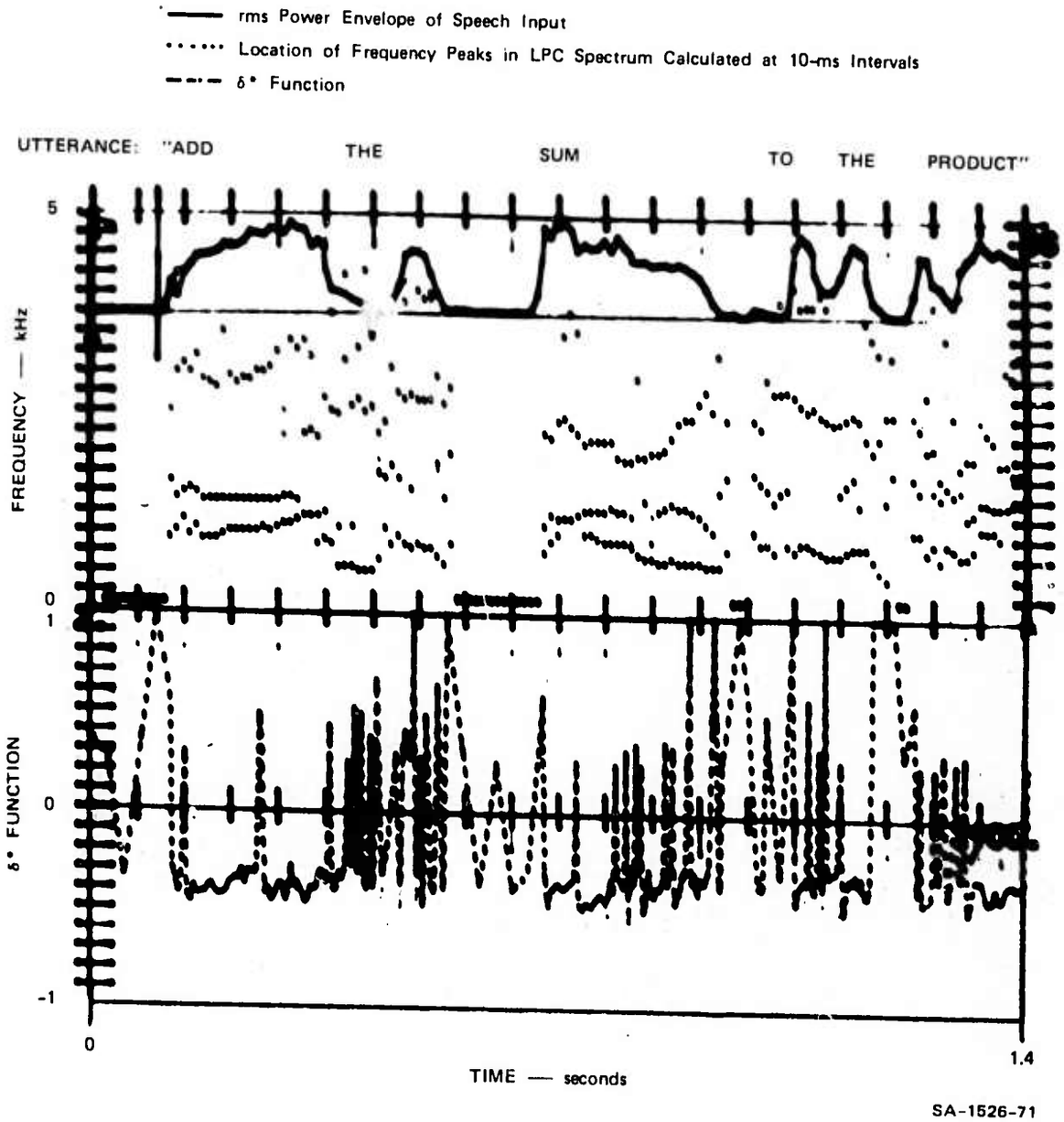


FIGURE 16 TRANSMISSION DECISION FUNCTION δ* FOR MEASURE 3, Y = 0.5
Analysis type: PTOVR (pitch synchronous overlapped)

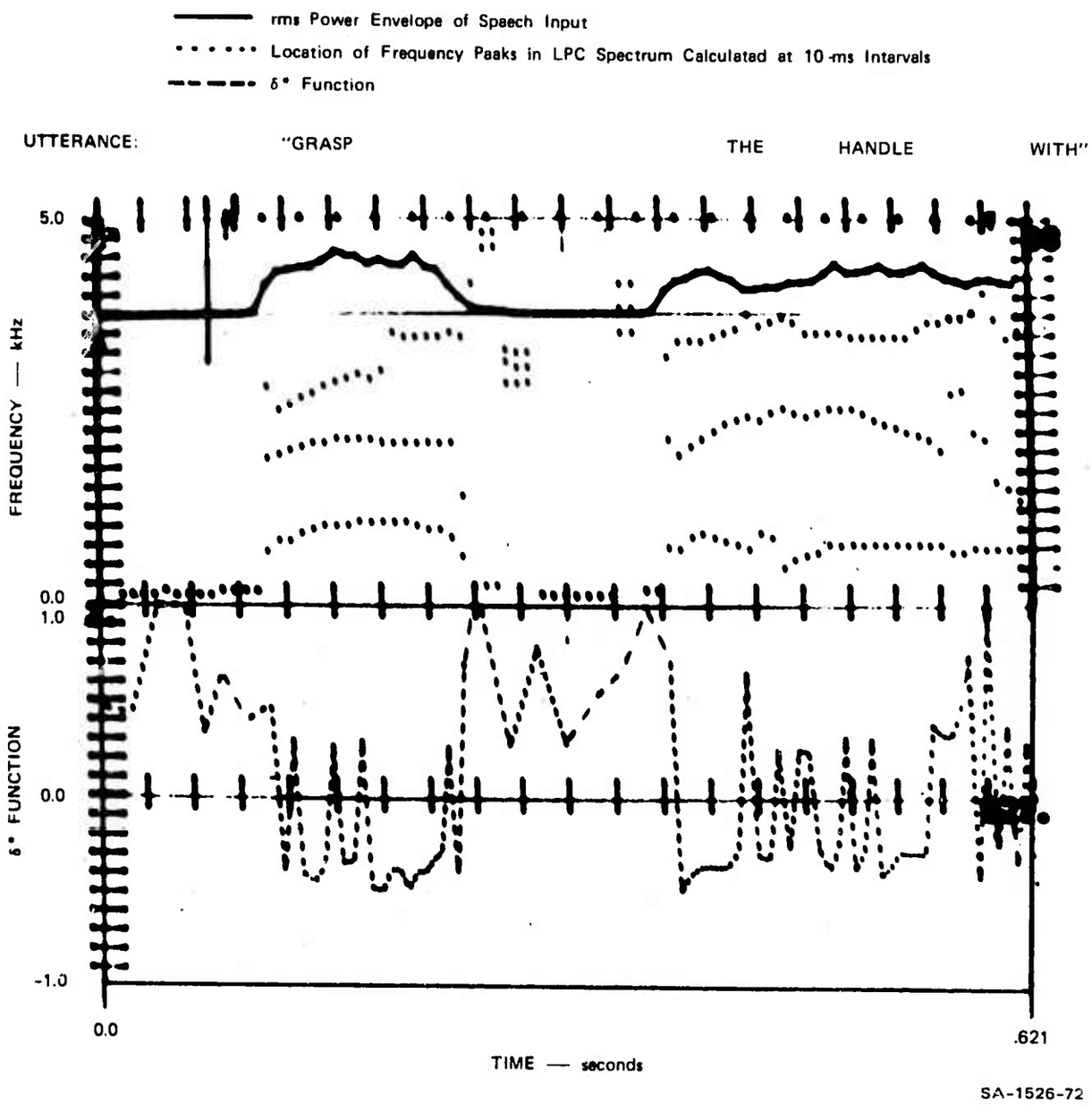


FIGURE 17 TRANSMISSION DECISION FUNCTION δ* FOR MEASURE 3, Y = 0.5
Analysis type; PTOVR (pitch synchronous overlapped)

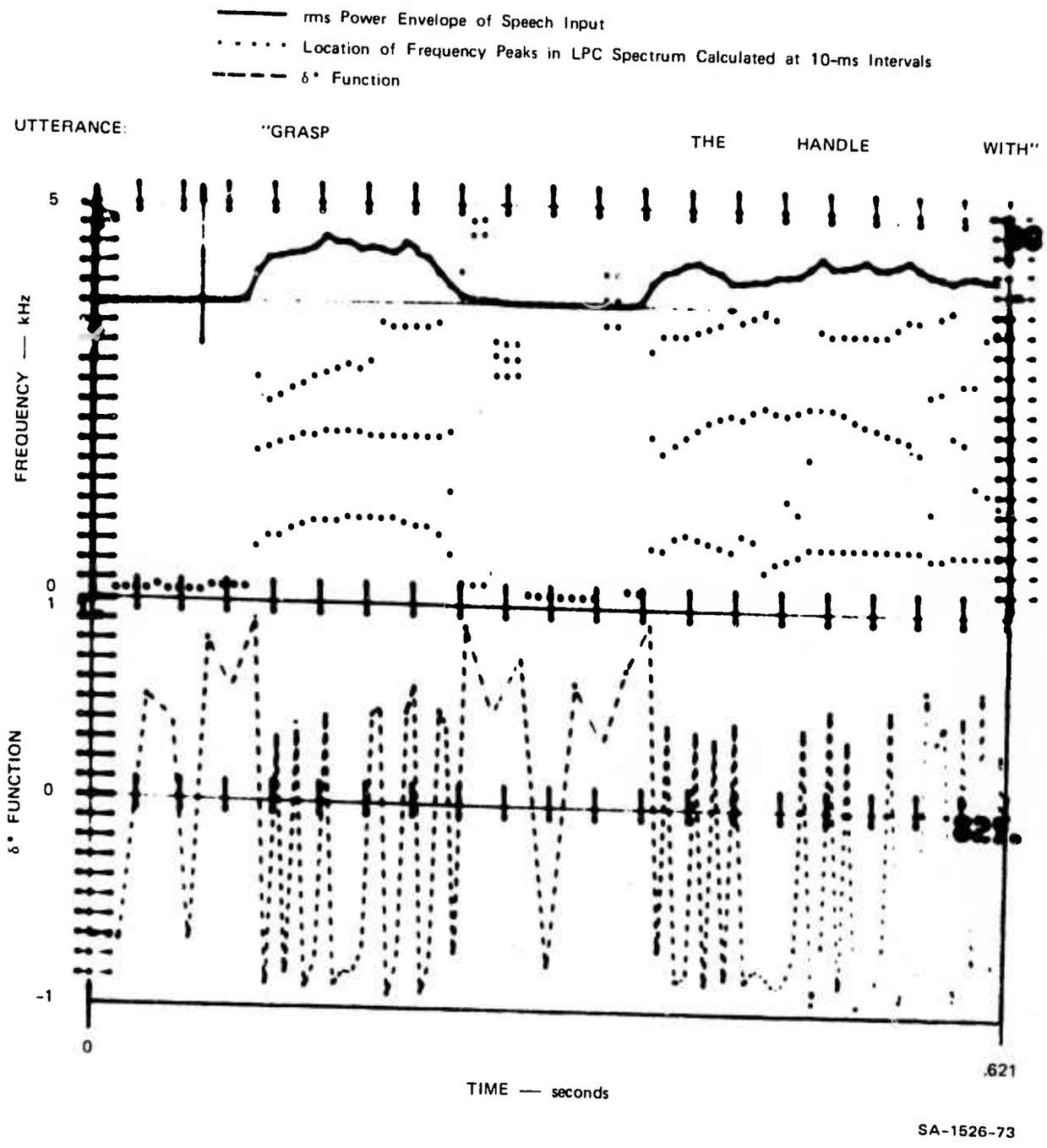


FIGURE 18 TRANSMISSION DECISION FUNCTION δ* FOR MEASURE 4, Y = 0.3
Analysis type: PTOVR (pitch synchronous overlapped)

discernible difference in performance for overlapped versus nonoverlapped analysis is small. Generally, better speech quality is obtained for pitch-synchronous than for pitch-asynchronous analysis. However, this same result was observed without adaptive compression. Comparison of δ_4^* functions for pitch-synchronous analysis is not important for adaptive compression.*

Measure 3 [based directly on the coefficients k(i)] produces a slightly higher average bit rate for a given quality of synthetic speech than that obtained using Measure 4. This measure (and Measures 1 and 2) performs best when applied over only the first few k(i), using overlapped analysis frames. The results indicating better performance with Measures 1, 2, and 3 applied over only a few of the k(i), e.g., q=4, are not expected. However, letting q=4 eliminates the "noise" introduced into the computation of δ by the higher-order terms, which are not as accurate as the lower-order terms. By contrast, Measure 4 does not require overlapped analysis for acceptable performance. For this reason, Measures 1 and 2 are dropped entirely. Although Measure 4 is more robust and theoretically more justifiable than Measure 3, the latter demonstrates that tolerable transmission decisions may be extracted directly from the k(i).

Typical rates of compression obtained over and above the data transmission rates obtained from synchronous LPC systems are summarized in Tables 3, 4 and 5. The baud rates are computed on the basis of 72 bits per transmitted frame with 14 coefficients quantized at 6, 6, 4, 4, 4 ... bits respectively for a total of 60 bits. Twelve additional bits are provided for excitation amplitude, pulse/noise ratio, and pitch information. Since the pitch frequency for these tests is derived from

---

*We hypothesize that the pitch-asynchronous degradation is associated with imperfect gain settings in the excitation function and not with the LPC parameters themselves.

Table 3

SUMMARY OF COMPRESSION RATES FOR LPC COEFFICIENTS.
UTTERANCE--EAIF.DTG ("Add the Sum to the Product of These Three.")*

Run	Analysis Type	No. of Analysis Frames	Initial Baud Rate†	DELCO Measure	DELCO Threshold (%)	No. of Coefficient Sets Transmitted	Coefficient Sets Transmitted (%)	Final Baud Rate†
1	PTOVR	465	11917	4	0.30	137	0.295%	3523
2	PTOVR	465	11917	4	0.35	106	0.228	2726
3	Over 10/25	278	7149	4	0.30	170	0.612	4371
4	Over 15/25	186	4783	4	0.30	92	0.495	2365
5	PTOVR (no preemphasis)	465	11917	4	0.25	141	0.303	3626
6	PTOVR (no preemphasis)	465	11917	4	0.30	107	0.230	2751
7	PTSYN	465	11917	4	0.20	164	0.353	4217
8	PTSYN	465	11917	4	0.30	107	0.230	2751
9	PTOVR	465	11917	3	0.50	122	0.262	3137

*For all runs, approximate speaker pitch frequency was 250 Hz, length was 2.8 s, and sample rate was 10 kHz.

†Estimated (the pitch frequency is unquantized); see text for details.

Table 1

SUMMARY OF COMPRESSION RATES FOR LPC COEFFICIENTS,
UTTERANCE F0021.DTM ("Pete Cooper's Dog Toyed with Dick Todd's Cat.")*

Run	Analysis Type	No. of Analysis Frames	Initial Baud Rate†	DELCO Measure	DELCO Threshold (v)	No. of Coefficient Sets Transmitted	Coefficient Sets Transmitted (%)	Final Baud Rate†
1	PTOVR	228	5863	1	0.3	78	0.342	2006
2	PTOVR	228	5863	1	0.35	72	0.316	1851
3	Over 10/25	278	7148	1	0.3	103	0.371	2649

*For all runs, approximate speaker pitch frequency was 125 Hz, length was 2.8 s, and sample rate was 10 kHz.

†Estimated (the pitch frequency is unquantized); see text for details.

Table 5

SUMMARY OF COMPRESSION RATES FOR LPC COEFFICIENTS,
UTTERANCE BN-IF.DTM ("Grasp the Handle with the Hole in It.")*

Run	Analysis Type	No. of Analysis Frames	Initial Baud Rate†	DELCO Measure	DELCO Threshold (V)	No. of Coefficient Sets Transmitted	Coefficient Sets Transmitted (%)	Final Baud Rate†
1	PTOVR	161	5040	4	0.30	84	0.522%	2630
2	PTOVR	161	5040	4	0.35	79	0.491	2473
3	Over 10/25	228	7137	4	0.30	113	0.496	3537

*For all runs, approximate speaker pitch frequency was 120 Hz, length was 2.3 s, and sample rate was 10 kHz.

†Estimated (the pitch frequency is unquantized); see text for details.

hand-marked pitch pulses, it is unquantized; therefore, the given baud rates are estimated baud rates. The results show that adaptive transmission of LPC parameters allows an impressive reduction in average bit transmission rate.

All four of the transmission measures described above attempt to respond only to spectral changes and, of course, are derived from values that are normalized with respect to signal power.* Since, in general, the gross spectral properties cannot be expected to remain constant during pauses in the speech, some unnecessary transmission of LPC parameters may occur during pauses. This problem is clearly evident in Figures 17 and 18. The algorithm should therefore be augmented with a signal present/absent detector.

G. Transmission Statistics

The time between coefficient updates using the asynchronous strategy (pitch-synchronous analysis) varies from one to several pitch periods. The minimum, maximum, average, and standard deviation of the time between coefficient updates for several speech utterances by a variety of speakers are given in Table 6. The table shows that, for a typical speaker with $\gamma = 0.3$, an average time between coefficient updates of approximately 30 ms can be expected, with a standard deviation of about 20 ms--although minimum and maximum times between coefficients can be expected to range from 3 to 200 ms.†

---

*Theoretically, only Measure 4 can be clearly tied to spectral changes, In general, measures based on the reflection coefficients or the LPC parameters are not reliable since the transformation between them and the power spectrum is not metric preserving.

†The minimum value of 3 ms results for pitch-synchronous analysis with a female speaker of approximately 300 Hz pitch. More realistically for PAA, the minimum value is 10 ms.

Table 6

TIME BETWEEN COEFFICIENT UPDATES

DELCO Threshold (Y)	File	No. of Analysis Frames	No. of Transmitted Frames	Minimum TBT* (ms)	Maximum TBT* (ms)	Average TBT* (ms)	Standard Deviation (ms)
0.35	PNOF	160	52	6.9	85.6	32.4	19.6
0.35	BHOF	134	59		140.0	31.5	21.9
0.35	F0021	228	72	5.7	194.6	37.5	28.4
0.35	BN4F	161	79		100.0	27.5	16.1
0.35	EA1F	465	106	3.4	144.4	26.0	25.4
0.30	PNOF	160	58	6.9	75.0	29.0	15.6
0.30	BHOF	134	66		140.0	28.2	19.2
0.30	F0021	228	78	5.7	194.6	34.6	26.7
0.30	BN4F	161	84		100.0	26.5	15.6
0.30	EA1F	465	137	3.4	100.0	20.0	17.9

*Time between packet transmissions.

If we assume that no special buffering or smoothing of the data takes place, the effect of the asynchronous data rate on a packet transmission system will be to produce a corresponding asynchronous packet transmission rate. Table 7, which uses the same speech utterances and speakers as Table 6, presents statistical data with respect to the time between packet transmissions. For 360 data bits/packet and a typical speaker with $\gamma = 0.3$, an average time between packet transmissions of approximately 160 ms can be expected. The standard deviation is about 40 ms. Minimum and maximum times between packet transmissions can be expected to range from 40 to slightly over 250 ms. Similar conclusions may be derived from the 7[illegible] data bits/packet statistics. It is worth noting that a practical packet transmission system will require the maximum time between packet transmission to be limited.

Table 7

TIME BETWEEN PACKET TRANSMISSIONS (TBT)
360 Data Bits/Packet (Lines 1-5) and 792 Data Bits/Packet (Lines 6-9)

DELCO Threshold (Y)	File	No. of Analysis Frames	No. of Transmitted Frames	No. of Transmitted Packets	Minimum TBT (ms)	Maximum TBT (ms)	Average TBT (ms)	Standard Deviation (ms)
0.30	PNOF	160	58	11	74.2	207.5	141.7	39.7
0.30	BHOF	134	66	13	92.3	216.3	141.0	45.3
0.30	F0021	228	78	15	87.2	252.1	170.0	52.3
0.30	BN4F	161	84	16	86.6	169.5	126.6	25.8
0.30	EA1F	465	137	27	45.3	177.3	100.8	38.5
0.30	PNOF	160	58	5	233.1	377.4	288.5	48.2
0.30	F0021	228	78	7	252.9	481.3	382.1	89.7
0.30	BN4F	161	84	7	214.4	345.1	278.7	37.2
0.30	EA1F	465	137	12	150.8	286.0	219.1	39.0

VII CONCLUSIONS

Based on our simulation results, reconstructed speech quality appears not to depend on whether the LPC analysis is of the Toeplitz or the non-Toeplitz type. Other factors, such as pitch extraction, have a much greater bearing on the speech quality. The advantage of the Toeplitz analysis is that the computed reflection coefficients are guaranteed to produce a stable synthesizing filter. Consequently, our major research effort concentrated on Toeplitz-form LPC analysis/synthesis systems.

Our research demonstrated that the best quality synthetic speech resulted when pitch-synchronous analysis and synthesis were performed. The degradation with pitch-asynchronous synthesis was much greater than that associated with pitch-asynchronous analysis. Of course, significant pitch-pulse location errors in the synthesizer excitation function are far more noticeable than either of the above degradations. A major difficulty with pitch-synchronous analysis is that the analysis window varies in size with the speaker's pitch.

Since better performance was achieved with pitch-synchronous analysis, investigation of time-domain (i.e., absolute pitch-pulse placement) pitch extraction was performed. The difficulty of constructing a good, reliable time-domain pitch extractor is great. The reader is referred to the Task 3 report for further details. Here, it suffices to say that we developed an algorithm that greatly simplified the job of hand placing pitch marks. A human operator (needed to correct occasional pitch errors) using this algorithm can generate a set of absolute-time pitch-pulse marks that, when used with pitch-synchronous LPC analysis and synthesis, produces synthetic speech virtually indistinguishable from the input speech. These absolute pitch marks serve as a useful reference set for

comparison with the outputs of more practical pitch extractors. A computer program has been developed that computes the standard deviation between two sets of pitch marks, making it convenient to compare any absolute-time pitch extractor with the best possible pitch marks.

Based on our simulations with inferior pitch extractors, we determined that the required accuracy (on a pitch of 100 Hz) is approximately 2 Hz rms. That is, a set of pitch marks with a standard deviation of 2 Hz, with respect to the best set of hand-marked pitch pulses, produced acceptable quality synthetic speech. However, when the standard deviation was increased to 4 Hz, a definite roughness was perceptible in the synthetic speech. The required pitch accuracy scales with frequency so that 1-Hz and 4-Hz standard deviations are acceptable at pitches of 50 and 200 Hz, respectively.

Use of an excitation function that consists of a mixture of pulses and random noise produces very high quality synthetic speech. No quality degradation was found with this concept when the proper combination rule was used. In fact, the mixture concept seemed to offer an unexpected degree of robustness with respect to a variety of system degradations. For example, the use of the noise mixture concept, rather than a hard buzz-hiss decision, improved the quality of the synthetic speech with pitch-asynchronous synthesis. Furthermore, the mixture concept is clearly better suited to handling signals such as the voiced fricatives. The major question is whether the improvement is worth the effort of transmitting two or three extra bits each analysis block to convey this information. For the first systems developed, it is clearly an unnecessary luxury. However, future systems may find this structure desirable.

The major contribution of our research has been the development of an adaptive data compression algorithm for the linear predictive coefficients. The algorithm (known as DELCO) recognizes steady-state segments

of speech and transmits new LPC parameters only when there are significant changes in the parameter values from the previously transmitted values. Thus, an adaptive sampling system is used between the LPC analysis system and the transmission system. The DELCO algorithm is preferable to a fixed, low-rate LPC analysis system, since DELCO can respond to rapid changes in signal structure when necessary. By contrast, the fixed, lower-rate LPC system (with the same average transmitted bit rate) will miss or will not accurately represent these rapid changes.

The result of this data compression is a reduction in the required average data rate by a factor in excess of two, with no discernible quality loss. The exact compression factor depends on the speaker and the utterance. Frequently, the compression is significantly greater than two to one. DELCO produces a nonuniform data rate since it is based on adaptive sampling of a fixed-rate system. Data compression systems that produce nonuniform data rates require rate-smoothing buffers to interface with synchronous communication systems. However, DELCO can be interfaced with an asynchronous communication system, such as a packet-switching transmission system, without requiring rate-smoothing buffers. Thus, DELCO is ideally suited for operation with packet-switching systems.

In summary, the major contribution of our research has been the development of the adaptive data compression algorithm DELCO. DELCO reduces the average data rate of an LPC vocoder by a factor of two or more while maintaining excellent speech quality. DELCO is a proved concept that can be readily interfaced with packet-switching systems and other asynchronous communication systems.

Appendix A

ADAPTIVE SPEECH COMPRESSION FOR PACKET COMMUNICATION SYSTEMS

Appendix A

ADAPTIVE SPEECH COMPRESSION FOR PACKET COMMUNICATION SYSTEMS*

D. T. Magill
Stanford Research Institute
Menlo Park, California 94025

Packet communication systems offer many significant advantages for low duty factor user populations. These advantages can be applied to voice communication. Additional data compression beyond that achievable with the new linear predictive encoding techniques can be obtained by exploiting the asynchronous character of the packet communication channel. The adaptive data compression algorithm DELCO achieves a compression factor greater than two while maintaining high quality.

I INTRODUCTION

The conventional approach to joint utilization of a common communication resource among multiple users is frequency-division multiple access (FDMA).† Each user is assigned a separate frequency channel (and in some cases, such as satellite communication, a fraction of the available power) on a dedicated basis. This traditional approach is efficient for static user populations and has been used with great success for analog communication systems.

---

*This work was supported by the Advanced Research Projects Agency of the Department of Defense (DAHC04-72-C-0009).

†In this appendix the term multiple access is used generally and includes, as a special case, multiplexing, i.e., the case when all users are, effectively, collocated.

The advent of the digital computer and its associated digital technology has had a tremendous impact on both the concepts and the hardware of communication systems. In particular, digital modulation has rapidly grown in prominence due to theoretical and hardware advantages. Digital signaling has been employed successfully with the conventional FDMA approach. However, it has been recognized that time-division multiple access (TDMA) offers significant advantages. With TDMA the communication resource, a single wideband channel, is shared on a time-division basis. Thus, for example, the problem of frequency stability for many narrowband channels is greatly alleviated. In many parts of the TDMA communication system, a single piece of time-shared equipment replaces multiple units in the conventional FDMA system. This is possible due to the inherent high speed of present day digital circuits. There are other advantages of digital TDMA systems, which are not listed in the interest of brevity. The important point is that, so far, the discussion refers to a synchronous TDM or TDMA system with a relatively static user population, each user receiving a dedicated link. Such a system might be reconfigured relatively infrequently, perhaps once a day or once a month.

In practice there are many communication environments in which the user population possesses far different characteristics. For example, a communication system may consist of very many remote data terminals accessing a central computer. In this case, these data terminals might have a very low duty factor and have independent statistics. These messages might be very short in duration and occur randomly. For such a system the conventional FDMA or the relatively recent digital TDMA system might be quite inefficient. The basic problem is that these systems have been designed on the basis of the dedicated circuit concept. This concept simply is not suited to a very large user population that has a very low duty factor. For example, there simply may not be enough bandwidth

to allocate each of the many system users a dedicated circuit. Even if it were possible, inefficient use of the communication resource would result.

One efficient method of operating with such a user population is known as packet communication.[1,2,3]* With this system all users share a common wideband channel in a random, asynchronous mode. Each user transmits its information in packets or short bursts. These packets consist of the data plus preamble bits that carry source and sink (destination) information. Parity bits are also attached for error detection and correction.

Many forms and variations of packet communication are possible. However, the following example suffices to illustrate the major concepts. If the message is received correctly at the intended destination, i.e., no parity errors are detected, then an appropriate acknowledgment is transmitted back to the sender. If the acknowledgment is correctly received, then the message is removed from the sender's buffer storage and the sender is ready to progress to its next message. However, if an incorrect message is received at the destination, a repeat request is generated. When this is correctly received at the source the original message is repeated and the process continues as described above. Most often, the necessity for a repeat transmission is generated by the simultaneous transmission of two or more messages from random sources. However, receiver noise may occasionally cause such a repeat request.

Clearly as the system usage factor becomes higher, more frequent repeat requests will become necessary. Thus, the effective system usage will increase, resulting in further repeat requests. Such a system has a "snowballing" effect if the system usage becomes excessively high.

---

*References are listed at the end of this appendix.

With a well-designed system the usage factor can be kept appropriately low and this problem avoided. The net result is that for a sufficiently large and low-duty factor user population, packet communication can offer significant advantages over the conventional dedicated circuit approach. Furthermore, since packet communication can be regarded as a form of asynchronous TDMA, it possesses most of the advantages of TDMA with respect to FDMA. Consequently, packet communication offers many important advantages. A very significant characteristic of such systems is their asynchronous, random signal flow.

To date, the advantages of packet communication have been described with respect to data systems. However, voice communication systems frequently have user populations with similar characteristics, e.g., low-duty factor. Thus, voice communication systems need to be considered from the packet communication system viewpoint. Furthermore, in many cases, it is desirable to mix voice and data within a common system. In addition, the security advantages of digitized speech are well recognized. Consequently, the performance and capabilities of digitized voice in packet communication systems were investigated.

II SPEECH COMPRESSION

It is well known that digital transmission of speech is a difficult problem with many trade-offs between data rate, system complexity, and voice quality. Simple systems such as delta modulation (and its numerous variations) offer high quality, i.e., input and output virtually indistinguishable,* only for high data rates. Complex systems such as vocoders operate at modest signaling rates, i.e., 2400 to 9600 baud, but are prone to providing inconsistent quality. That is, while high intelligibility

*By high quality we refer to the quality obtainable in a standard 4-kHz phone channel.

may be maintained, loss of speaker identification, emotional content, and naturalness may result under certain circumstances. Thus, with conventional approaches it does not appear possible to obtain the desired high quality with a data rate that can readily be transmitted through a 4-kHz phone circuit.

At present the most promising new technique for speech digitization is based on linear predictive encoding.[4-7] With linear predictive encoding, short-term properties of the speech process S(t) are deduced by posing the linear one-step prediction problem. That is, it is desired to select a set of p coefficients $\{a_i\}$ such that the error

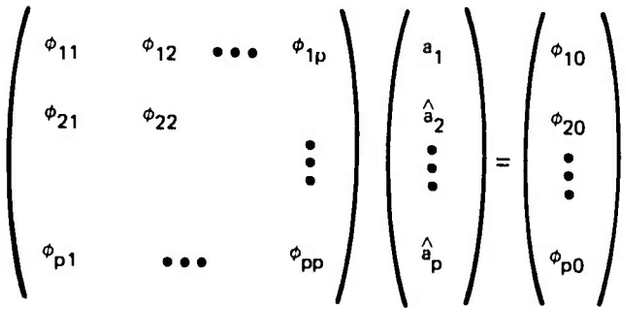
$$E(t) = S(t) - \sum_{i=1}^{p} a_i S(t - i)$$

is minimized in a mean-square error sense over some interval. While there are several formulations of the problem, it is convenient to choose the following example. The mean-square error is minimized over a finite block size of 100 samples or a pitch period, depending on whether the speech signal is voiced or unvoiced.*

Posing the above problem leads to a set of p simultaneous equations in the autocorrelation coefficients and the unknown linear predictive coefficients (LPC), i.e., the $\{a_i\}$, which may be solved for the latter. Figure A-1 illustrates these equations in matrix form. The LPC partially characterize the speech process on a short-term basis and, in fact, can

---

*Here we assume a sampling rate of approximately 10 kHz so that an analysis block of 100 samples corresponds to a 100-Hz refresh rate on the analysis. This appears to be sufficiently often to track the changes in the vocal tract configuration.



Where $\phi_{ij} = \sum_{t=0}^{N-1} S(t-i)\, S(t-j)$

Non-toeplitz Form – Cholesky's Method

Toeplitz Form – Levinson's Method

SA-1526-83

FIGURE A-1 MATRIX FORMULATION OF LPC EQUATIONS

be readily related to the short-term power spectral density. This power spectral density is known to be an adequate characterization of the speech process when used in conjunction with other important parameters such as the voiced-unvoiced (V/UV) decision, the pitch, and the overall signal power.

If the LPC (or a suitably transformed version of them) and the V/UV, pitch, and power parameters are encoded and transmitted, the receiver can synthesize a signal that accurately models the input speech short-term power spectral density.[7] In this case satisfactory quality will be obtained. The LPC parameters are used in a recursive (all-pole filter) that is excited by an appropriate source. For unvoiced segments an independent noise generator is used. For voiced segments an impulse generator (the frequency is controlled by the pitch parameter) is employed. In both cases the excitation level is controlled by the power parameter. Figures A-2 and A-3 are block diagrams of the transmitter and receiver, respectively.

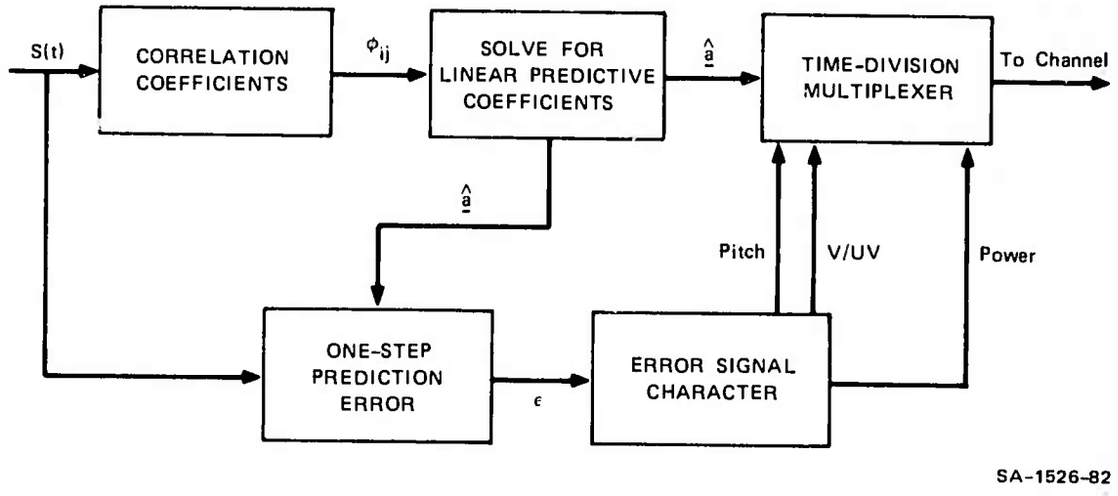


FIGURE A-2 BLOCK DIAGRAM OF LPC ANALYZER

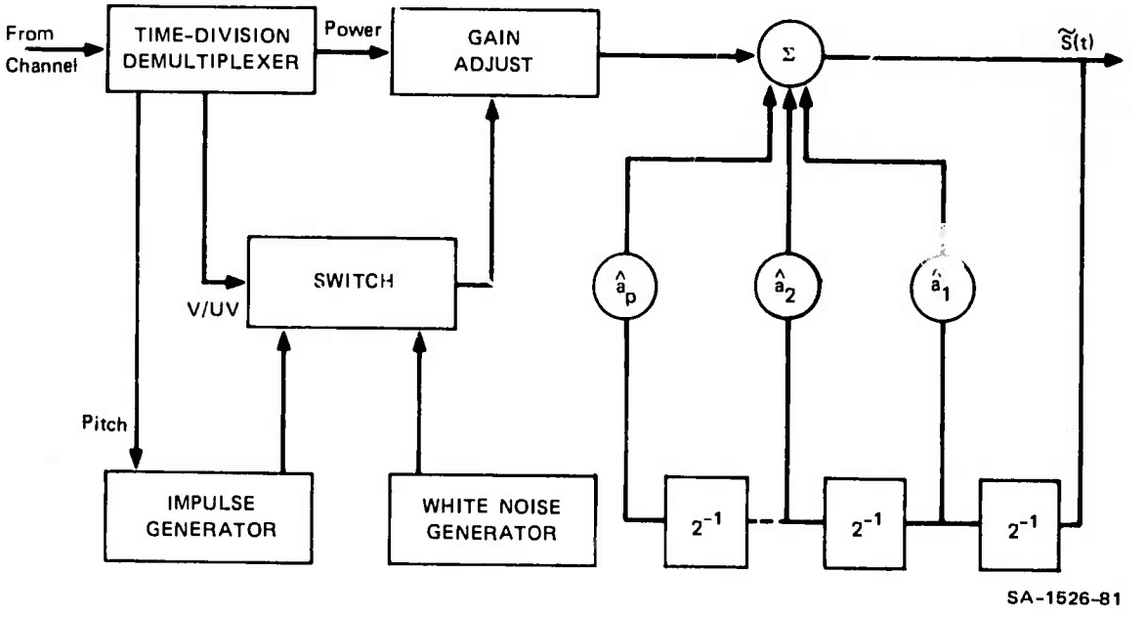


FIGURE A-3 BLOCK DIAGRAM OF LPC SYNTHESIZER

At this point one can note the obvious similarities with the conventional channel vocoder approach. It is reasonable to ask what advantages the LPC approach offers over the conventional approach. Basically the higher quality of the former approach can be related to the greater flexibility of the recursive synthesizing filter as compared with the relatively fixed capabilities of the channel vocoder synthesizing filter. In addition, the LPC technique is directly suitable for computer processing and digital implementation. Note that poor quality in the synthesized speech due to errors in the V/UV decision and pitch extraction is not avoided by adopting the LPC approach. Since the LPC approach has proved more successful (on the basis of preliminary research) than any other speech compression technique, it has been investigated for application with packet communication systems.

III DELCO ALGORITHM

To date all LPC algorithms and systems have been designed for operation with a synchronous, dedicated circuit. Thus, both active speech and speech pauses are transmitted. Since typical conversational speech has a duty factor of less than 50 percent, it should be possible to reduce the bit rate of a typical LPC speech compression digitizer by a factor of two. With a normal synchronous communication system this would result in a buffering problem since the achievable compression is a nonuniform function of time. Fortunately with packet communication, an asynchronous or burst-type transmission is acceptable and no rate smoothing buffer is required.

The compression algorithm developed modified the basic LPC algorithms to permit adaptive operation appropriate to the input speech. Data compression beyond that obtainable by the LPC algorithms is obtained in two ways. First, pauses in speech are eliminated by a TASI-type speech

detector that determines the presence or absence of a speech signal.[8,9] Second, steady-state portions of speech are recognized and only the new information is encoded. The synthesizer maintains the previous parameter values unless new values are transmitted. Consequently, the proposed scheme transmits no unnecessary speech information.

The necessity of transmitting new LPC parameters is established by considering the energy in the one-step prediction error or residual signal. This error energy is determined assuming that the last transmitted LPC parameter vector is used to form the prediction. Rather than computing the residual energy in the obvious but lengthy fashion, one can use the formula

$$E^{(k)}(\underline{a}^{(j)}) = \varphi_{00}^{(k)} - 2\sum_{i=1}^{p} a_i^{(j)} \varphi_{0i}^{(k)}$$

$$+\sum_{i=1}^{p}\sum_{\ell=1}^{p} a_i^{(j)} a_\ell^{(j)} \varphi_{i\ell}^{(k)} \qquad \text{(A-1)}$$

where the superscript denotes the analysis block $\varphi_{i\ell}^{(k)}$ are the auto-correlation coefficients in the kth (the present) analysis block, and $a_i^{(j)}$ are the LPC parameters from the jth (previous) analysis block. This energy is compared with the residual energy that would be obtained if the optimized LPC parameters were used to form the predicted signal.

$$E^{(k)}(\underline{a}^{(k)}) = \varphi_{00}^{(k)} - \sum_{i=1}^{p} a_i^{(k)} a_i^{(k)} \varphi_{0i}^{(k)} \qquad \text{(A-2)}$$

If

$$DEL = E^{(k)}(\underline{a}^{(j)})/E^{(k)}(\underline{a}^{(k)}) \qquad (A-3)$$

is less than a threshold value γ then parameter vector $\underline{a}^{(j)}$ is judged to be sufficiently accurate that it can be used for the present analysis/synthesis block.

Experience with the DELCO algorithm indicates that a threshold value of a 40 percent increase, i.e., $\gamma = 1.4$, yields a compression factor of approximately two without producing noticeable degradation. Thresholds as high as $\gamma = 2$, i.e., 100 percent increase in residual energy, have been employed yielding compression factors of approximately five. While the resulting speech is intelligible, it is noticeably distorted--primarily with an echo effect. Consequently, a conservative estimate of the compression factor (while maintaining high quality) is two to one. Table A-1 presents the simulation results for speaker Number 2 with the sentence, "Pete Cooper's dog toyed with Dick Todd's cat."

The results of Table A-1 are based on this sentence, which has only very short pauses, and on the DELCO algorithm without using the TASI-tape signal presence detector. With the speech detector installed in the voice digitizer, it should be possible to obtain an overall compression factor of four to one or better since a user's average duty factor is less than 50 percent.

Atal has demonstrated high quality speech at transmission rates in the range of 2400 to 9600 baud.[5] Thus, one might expect that the DELCO algorithm with packet communication might yield data rates as low as 600 to 2400 baud. Such is not the case for several reasons. First, the lowest rate of 2400 baud is achieved by using the low frame rate of 33-1/3

Table A-1

DELCO COMPRESSION FACTOR

Threshold Value (γ)	Number of Blocks Transmitted Out of 288 Analysis Blocks	Compression Factor	Quality
1.0	288	1.0	High
1.15	210	1.37	High
1.40	126	2.28	High
2.0	62	4.65	Distorted with an echo effect

Hz rather than the 100 Hz previously described. With such long analysis blocks, it is less likely that the subsequent analysis blocks will pass the DELCO threshold test than when shorter blocks are used. Second, the packet communication system concept has overhead bits associated with it and these will increase the average baud rate to convey a speech channel. At this point it is desirable to consider further this expansion factor.

IV VOICE PACKET FORMAT

Each packet must convey appropriate routing information such as source and destination identification. Since these bits are a type of fixed overhead, it is desirable to make each packet as large as possible to minimize the inefficiency due to the overhead bits. However, an increased packet length increases the average propagation delay through the network.

The minimum cycle time for the sink to acknowledge to the source that the packet was properly received is

$$T_{cycle} = 2T_P + T_m + T_a + T_r \qquad \text{(A-4)}$$

where T_p is the physical propagation path delay and T_m is the message or packet duration. T_a is the duration of the acknowledgment message, and T_r is the processing delay in the receiver. If the message or the acknowledgment are incorrectly received, then it is necessary to repeat the cycle. In this case the network propagation delay is significantly increased.

Use of excessively long packets can result in network propagation delays that are unacceptable. Nominally, it is desirable to maintain the network delay below 0.3 s to avoid conversation difficulties, such as simultaneous speech. However, it has been reported that users can tolerate delays as large as 1.2 s.[10]

In addition to the maximum tolerable delay effect, which limits packet sizes, there is a random variation in the propagation delay. The magnitude of this effect depends on the variables of cycle time equation and on the system usage factor, i.e., the likelihood of cycle repeats. For most reasonably designed systems the variation in the network delay will significantly distort the time base. As a result it is necessary to append additional overhead bits that identify the proper time placement for the information bits describing the speech process. Many formats are possible but it is clearly advantageous to use relative timing information rather than absolute values since the former procedure results in a significant data rate reduction.

At present a variable packet structure is envisioned. The data are arranged in the following sequence: (1) destination, (2) source,

(3) parity, (4) power level and signal presence/absence, (5) voiced/unvoiced ratio, (6) pitch, (7) LPC parameters, and (8) relative time. If no signal is present, the packet could be truncated after the fourth position. Otherwise, the full duration packet would be transmitted. The synthesizer at the receiver continues to employ the previous values until it is signaled to change to new values. Nominally one might expect some 60 or so overhead bits for source, destination, and parity bits. Thus, if the speech information requires 60 or more bits, the packet efficiency should exceed 50 percent. Atal has shown that 72 bits per analysis block are adequate to provide high quality synthetic speech.[5] Thus, so long as the packet describes one or more analysis blocks, then the packet efficiency should exceed 50 percent.

The above arguments neglect the loss due to the necessity of transmitting timing information. The number of bits required depends on the range of the relative time measurement and the required resolution. The range can be reduced by periodically transmitting fixed time references even when it is unnecessary to transmit new speech coefficients. A time resolution of 10 ms should be adequate for the speech process parameters. As a result, ten bits should be more than sufficient for timing information. Thus, the requirement for timing bits does not significantly affect the packet efficiency.

V SIMULATION

The existing system used to generate the demonstration tape has been implemented on a large, general purpose, time-shared computer--a PDP-10 that is part of the ARPANET. Input/output and display are handled through an auxiliary PDP-15 computer that permits interactive operation. The analysis can be performed either on a Toeplitz or non-Toeplitz basis.[6,5] The synthesizing filter can be either of the direct or ladder forms.[11,7]

The latter is preferred from a coefficient accuracy point of view. At present, with no significant effort on algorithm speed the program runs about 60 times slower than real time.

The simulated performance is based on pitch-synchronous analysis using hand-placed pitch pulses. This was done as the initial stage since the major effort of this study was to explore the interaction between the LPC approach and the packet communication system--rather than to develop pitch extractors. The excitation function driving the synthesizing filter uses these pitch pulses for pitch-synchronous synthesis. The excitation power is divided between random noise and pulse power, depending on the energy in the residual signal normalized by the signal power. If the normalized residual energy exceeds a threshold, all of the excitation energy is noise-like. Otherwise the ratio of noise power to total power is a quadratic function of the normalized residual energy. The threshold value has been selected on the basis of providing high quality synthesis for the speech data base.

Typically, the existing DELCO algorithm has been run with a non-windowed Toeplitz analysis, a ladder synthesizer, 14 coefficients, and a pitch-synchronous analysis/synthesis structure. However, many other modes are possible. To date the data compression algorithm has been applied to the excitation energy, the V/UV ratio, and the LPC parameters. No attempt has been made to adaptively encode the pitch parameters. There are several reasons for this. First, the data rate required to transmit pitch information is only about one-tenth of the rate required to characterize the complete speech process. Thus, the requirement to continually update pitch is not burdensome. Second, the quality of synthetic speech is critically dependent on the pitch signal. Thus, it is important to accurately transmit pitch information. Third, the normalized residual energy is not a good measure of changes in pitch. However, in

the future it may be desirable to develop a good method for compressing the pitch information, e.g., perhaps DPCM is such a method. At present this problem area has been reserved for future study.

VI CONCLUSIONS

The adaptive data compression technique DELCO works very well, yielding significant data compression without degrading voice quality. It is estimated that data rates about 1200 to 4800 baud permit high quality voice transmission with packet communication systems. Such systems avoid the wasteful practice of dedicating circuits to low duty-factor users. Thus, based on this initial research effort, the concept of a voice packet communication system appears very promising. Much work remains to develop the full capabilities of such a system.

VII ACKNOWLEDGMENTS

The author is indebted to Dr. E. J. Craighill for his development of the excitation function concept and his general support in the field of speech processing. Mr. D. W. Ellis has been instrumental in developing the DELCO algorithm code and has provided much of the important programming.

REFERENCES FOR APPENDIX A

1. H. Frank, R. Kahn, and L. Kleinrock, "Computer Communication Network Design--Experience with Theory and Practice," Spring Joint Computer Record, pp. 255-270 (16 May 1972).

2. N. Abramson, "The ALOHA System--Another Alternative for Computer Communications," 1970 Fall Joint Computer Conference Record, pp. 281-285 (November 1970).

3. L. Roberts and B. Wessler, "The ARPA Computer Network," Computer Communication Networks, Abramson and Kuo, eds. (Prentice-Hall, Englewood Cliffs, New Jersey, 1972).

4. B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell Sys. Tech. J., Vol. 49, No. 8, pp. 1973-1986 (October 1970).

5. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acous. Soc. Am., Vol. 50, No. 2 (Part 2), pp. 637-655 (August 1971).

6. J. D. Markel, "Digital Inverse Filtering--A New Tool for Formant Trajectory Estimation," IEEE Trans. Audio and Electroacous., pp. 129-137 (June 1972).

7. F. Itakura and S. Saito, "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Conference Record, pp. 434-437, 1973 International Conference on Speech Communication and Processing, Boston, Massachusetts (April 1972).

8. H. Miedema and M. G. Schachtman, "TASI Quality-Effect of Speech Detectors and Interpolation," Bell Sys. Tech. J., pp. 1455-1972 (July 1962).

9. G. R. Leopold, "TASI-B: A System for Restoration and Expansion of Overseas Circuits," Bell Laboratories Record, pp. 299-306 (November 1970).

10. P. T. Brady, "Effects of Transmission Delay on Conversational Behavior on Echo-Free Telephone Circuits," Bell Sys. Tech. J., pp. 115-134 (January 1972).

11. B. Gold and C. M. Rader, Digital Processing of Signals, p. 214 (McGraw-Hill, Inc., New York, New York, 1969).

APPENDIX B

DESCRIPTION OF SUBROUTINE EPOCH

Appendix B

DESCRIPTION OF SUBROUTINE EPOCH

The subroutine EPOCH sets up the analysis and synthesis intervals based on a set of (pitch) marks. These marks can be the output of a pitch extractor (of the absolute-time type) or the result of human pitch-mark placement. The linear predictive analyses and syntheses are then performed over these epochs. Figure B-1 is a listing of the subroutine EPOCH.

```
      SUBROUTINE EPOCH (NPTMX,MOVPTS,IMHER,MARK,NMK,IWSW,
     X           KAN,NPTAN,KSYN,NPTSYN)
      DIMENSION MARK(NMK)
C---THIS ROUTINE SETS UP EPOCHS FOR ANALYSIS AND SYNTHESIS.
C
C---     INPUTS
C       NPTMX    - MAXIMUM LENGTH OF EPOCH FOR BLOCK ANALYSIS
C                  AND FOR PITCH SYNCH IN UNVOICED INTERVALS.
C       MOVPTS   - #PTS TO MOVE ANALYSIS EPOCH FOR 'OVER',
C                  ALSO LENGTH OF SYNTHESIS EPOCH.
C       IMHER    - ABSOLUTE TIME INDEX FOR PITCH MARKS.
C       MARK     - ARRAY OF PITCH MARKS, STORED AS ABSOLUTE TIME
C                  INDICIES OF THE DATA ARRAY (1 SAMPLE / COUNT).
C       NMK      - NO. OF MARKS.
C       IWSW     - ALPHA SWITCH THAT SELECTS THE EPOCH OPTION.
C         'NO'   - BLOCK ANALYSIS WITH LENGTH NPTMX.
C         'OVER'- BLOCK ANALYSIS WITH LENGTH NPTMX
C              OVERLAPPED BY MOVPTS.
C              SYNTHESIS EPOCH IS MOVPTS.
C         'PTSYN'- PITCH SYNCHRONOUS ANALYSIS & SYNTHESIS,
C              ANALYSIS & SYNTHESIS EPOCHS THE SAME.
C              EPOCH = DISTANCE BETWEEN 2 MARKS,
C              UNLESS DISTANCE > NPTSMX
C              THEN,
C              IF DISTANCE >2*NPTSMX, EPOCH = NPTSMX
C              IF DISTANCE <NPTSMX, EPOCH = DISTANCE/2.
C         'PTOVR'- PITCH SYNCH ANALYSIS & SYNTHESIS OVERLAPPED.
C              SAME RULES FOR SYNTHESIS EPOCH AS 'PTSYN'.
C              ANALYSIS EPOCH IS 3 OVERLAPPING PERIODS DURING
C              VOICED PORTIONS AND THE SAME FOR UNVOICED.
C              ONLY 2 PERIODS ARE USED AT BEGINNING OF VOICED
C              INTERVAL AND AT END.
C         'SLO'- BLOCK SYNCHRONOUS ANALYSIS WITH OVERLAPPING
C               SYNTHESIZER COEFFICIENTS ARE CHANGED WITH PITCH
C
C---     OUTPUTS
C       KAN      - RELATIVE INDEX WHERE ANALYSIS EPOCH STARTS.
C       NPTAN    - NO. OF POINTS IN ANALYSIS EPOCH.
C       KSYN     - RELATIVE INDEX WHERE SYNTHESIS EPOCH STARTS.
C       NPTSYN   - NO. OF POINTS IN SYNTHESIS EPOCH.
C
C
      LOGICAL OLDMRK,SWITCH,PT
```

FIGURE B-1 LISTING OF SUBROUTINE EPOCH

```
^L
;  <CRAIGHILL>ANSYN.F4;306    TUE 6-AUG-74 3:31PM       PAGE 15:1

C     ---KSYN=0 MEANS INITIALIZE
          IF (KSYN.GE.0)  GO TO 100
          NEXT = 2
          PT = 'PT'.AND."777760000000
          SWITCH = IWSW.AND."777760000000
          ASSIGN 999 TO IT
          IF (SWITCH.EQ.PT) ASSIGN 300 TO IT
          NPTAN = NPTMX
          NPTSYN = NPTMX
          IZERO = IMHER
          KAN = - NPTSYN
          KSYN= - NPTSYN
          NADD = 0
          IF (.NOT.(IWSW.EQ.'OVER'.OR.IWSW.EQ.'SLO')) GO TO 100
C---      OVERLAPPING BLOCK SYNCH. ANALYSIS        ++++++++++
          NPTSYN = MOVPTS
          KAN = (NPTSYN-NPTAN)/2 - NPTSYN
          KSYN = -NPTSYN
C
C
C---      BLOCK SYNCH. ANALYSIS                    ++++++++++
100       KAN = KAN + NPTSYN
          KSYN = KSYN + NPTSYN
          IF (IWSW.NE.'SLO')       GO TO IT
C
C---      MODIFIED OVERLAPPING BLOCK SYNCH. ANALYSIS  ++++++++
C         COEFFICIENTS SWITCHED WHEN A PULSE OCCURS.
C
          KAN = KAN - NADD
C---      FIND NEXT 2 MARKS
410       NDIS1 = 0
          NDIS2 = 0
420       IF (NEXT.GT.NMK)  GO TO 490
          NDIS2 = MARK(NEXT) - IMHER -1
          IF (NDIS2.GE.MOVPTS-NADD)  GO TO 425
          NDIS1 = NDIS2
          NEXT = NEXT +1
          GO TO 420
425       IF (NDIS1.LE.0.AND.NDIS2.LE.0)  GO TO 490
          IF (NDIS1.LE.0)  NDIS1 = -30000
          IF (NDIS2.LE.0)  NDIS2 = 30000
C---      PICK THE CLOSEST MARK
          NDIS = NDIS2
          IF (NDIS2+NADD-MOVPTS.LT.MOVPTS-NDIS1-NADD)  GO TO 427
          NDIS = NDIS1
          NEXT = NEXT-1
C---      BR IF ROOM FOR MORE THAN ONE EPOCH
427       IF (NDIS.GT.1.5*MOVPTS-NADD)  GO TO 490
          NPTSYN = NDIS
          NADD = NADD + NPTSYN - MOVPTS
          GO TO 999
```

FIGURE B-1 LISTING OF SUBROUTINE EPOCH (Continued)

```
^L
;   <CRAIGHILL>ANSYN.F4;306    TUE 6-AUG-74 3:31PM        PAGE 15:2


C---      NO MARKS FOR A WHILE, GO OVERLAPPED BLOCK SYNC.
490       NPTSYN = MOVPTS - NADD
          NADD = 0
          GO TO 999
C
C---      PITCH SYNCH. ANALYSIS                          ♦♦♦♦♦♦♦♦♦♦♦
C                   IMHER                          MARK(NEXT)
C                     ^------< NDIS >---------^
300       OLDMRK = .TRUE.
310       NDIS = MARK(NEXT) - IMHER - 1
          NPTSYN = MIN0 (NDIS,NPTMX)
          IF (NPTSYN)320,330,340
C
C---      NPTSYN LT 0 MEANS -1, IGNORE IT
320       IF (NEXT.LT.NMK) GO TO 330
          NPTSYN = NPTMX
          NPTAN = NPTSYN
          GO TO 999
C
C---      NPTSYN = 0 MEANS WE NEED A NEW MARK
330       NEXT = NEXT + 1
          OLDMRK = .FALSE.
          GO TO 310
C
C---      NPTSYN GT 0 MEANS WE GOT A GOODIE
340       NPTAN = NPTSYN
          IF (NPTSYN.LT.NPTMX) GO TO 360
          KAN = KSYN
C---      IS THERE ENUF FOR TWO FULL BLOCKS?
          IF (NDIS.GE.2*NPTMX) GO TO 999
C---      SPLIT IT
          IF (NDIS.EQ.NPTMX) GO TO 999
          NPTSYN = NDIS/2
          NPTAN = NPTSYN
          GO TO 999
```

FIGURE B-1 LISTING OF SUBROUTINE EPOCH (Continued)

```
C
360     IF (IWSW.NE.'PTOVR'.OR.OLDMRK) GO TO 999
C---    OVERLAPPING PITCH SYNCH. ANALYSIS        ◆◆◆◆◆◆◆◆◆◆◆
C       OBJECTIVE◆◆◆ TO USE PROCEEDING  AND FOLLOWING
C       PITCH PERIODS IN ADDITION TO PRESENT ONE IN ANALYSIS
C       EPOCH DURING VOICED (PITCH MARKED) INTERVALS.
C
C---    GENERAL CASE
        KAN = MARK(NEXT-2) - IZERO
        NPTAN = MARK(NEXT+1) - IZERO - KAN
        IF (KSYN-KAN.LT.NPTMX) GO TO 362
C---    FIRST PERIOD
        KAN = KSYN
        NPTAN = MARK(NEXT+1) - IZERO - KAN
        GO TO 999
^L
;  <CRAIGHILL>ANSYN.F4;306   TUE 6-AUG-74 3:31PM      PAGE 15:3


362     IF (MARK(NEXT+1)-MARK(NEXT).LT.NPTMX.AND.NEXT.LT.NMK) GO TO 999
C---    LAST PERIOD
        NPTAN = MARK(NEXT) - IZERO -KAN
        GO TO 999
C
999     CONTINUE
        RETURN
        END
```

FIGURE B-1 LISTING OF SUBROUTINE EPOCH (Concluded)

APPENDIX C

DESCRIPTION OF SIMULATION TAPE DEMONSTRATING THE EFFECT OF TIMING ACCURACY ON SYNTHETIC SPEECH QUALITY

Appendix C

DESCRIPTION OF SIMULATION TAPE DEMONSTRATING THE EFFECT OF TIMING ACCURACY ON SYNTHETIC SPEECH QUALITY

The accompanying tape is restricted to the particularly difficult utterance, "Grasp the handle with the hole in it," by a male speaker. This utterance is low-pass filtered to a 4-kHz passband and is sampled at 10 kHz. In all cases 14 LPC parameters, preemphasis, a Hamming window, pitch-synchronous analysis overlapped over three pitch periods, and ratio excitation (see Section V) were used.

Five groupings of three utterances are presented on the tape. In the first grouping one hears (1) the input (original), (2) the synthetic, and (3) the input utterances. The synthetic utterance is based on the best set of hand-marked pitch pulses (file DTG). Note the high quality of the synthetic speech.

In the second grouping one hears (4) the input, (5) the synthetic speech (file DTO), and (6) the synthetic speech (file DTG). The synthetic speech (file DTO) is based on hand-marked pitch pulses on the output of a formant-isolation filter; less care in iteratively placing the pitch pulses was taken than for file DTG. Roughly comparable quality is perceived for both synthetic files.

In the third grouping one hears (7) the input, (8) the synthetic speech (file DTM), and (9) the synthetic speech (file DTG). File DTM is created from pitch marks based on the minimum-phase philosophy. Note the rough quality of file DTM.

In the fourth grouping one hears (10) the input, (11) the synthetic speech (file DTM/DTG), and (12) the synthetic speech (file DTG). File DTM/DTG uses inaccurate and accurate pitch marks for analysis and synthesis, respectively. Note that inaccurate analysis pitch marks have very little effect on the quality of the synthetic speech.

In the fifth grouping one hears (13) the input, (14) the synthetic speech (file DTG/DTM), and (15) the synthetic speech (file DTG). File DTG/DTM uses accurate and inaccurate pitch marks for analysis and synthesis, respectively. Note the very significant quality loss due to the use of inaccurate pitch marks for exciting the synthesizing filter.

Based on a comparison between the fourth and fifth groupings, one can say that accurate excitation pitch marks are much more important than accurate analysis marks. Furthermore, one can say that an rms pitch accuracy of 2 Hz (file DTO) provides excellent speech quality. In addition, it is clear from these recordings that it is possible to produce outstanding speech quality with the LPC method.

REFERENCES

1. D. T. Magill, "Adaptive Speech Compression for Packet Communication Systems," Conference Record, Vol. 1.2, pp. 29D-1-5, National Telecommunications Conference, Atlanta, Georgia (26-28 November 1973).

2. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acous. Soc. Am., Vol. 50, No. 2 (Part 2), pp. 637-655 (August 1971).

3. W. Gersch and S. Luo, "Discrete Time Series Synthesis of Randomly Excited Structural System Response," J. Acous. Soc. Am., Vol. 51, No. 1 (Part 2) (August 1971).

4. T. C. Hsia and D. A. Landgrebe, "On a Method for Estimating Power Spectra," IEEE Trans. on Instrumentation and Measurement, Vol. IM-16, No. 3, pp. 255-257 (September 1967).

5. J. L. Melsa, J. D. Gibson, and S. V. Jones, "Vocal Tract Parameter Identification Using Sequential Estimation Techniques," Conference Record, Vol. 2, pp. 29A-1-5, National Telecommunications Conference, Atlanta, Georgia (26-28 November 1973).

6. J. D. Markel, "Digital Inverse Filtering--A New Tool for Formant Trajectory Estimation," IEEE Trans. Audio and Electroacous., Vol. AU-20, No. 2, pp. 129-137 (June 1972). Also, SCRL Monograph 7, SCRL, Santa Barbara, California, (1971).

7. F. Itakura and S. Saito, "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Conference Record, pp. 434-437, 1972 Conference on Speech Communication and Processing, Newton, Massachusetts, (April 1972).

8. T. Kailath, "The Innovations Approach to Detection and Estimation Theory," Proc. IEEE, Vol. 58, No. 5, pp. 680-695 (May 1970).

9. J. Makhoul, "Spectral Analysis of Speech by Linear Prediction," IEEE Trans. Audio and Electroacous., Vol. AU-21, No. 3, pp. 140-148 (June 1973).

10. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-22, No. 2, pp. 124-134 (April 1974).

11. J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio and Electroacous., Vol AU-20, No. 5, pp. 367-377 (December 1972).

12. B. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain, " J. Acous. Soc. Am., Vol. 46, No. 2 (Part 2), pp. 442-448 (1969).

13. O. Fujimura, "An Approximation to Voice Aperiodicity," IEEE Trans. Audio and Electroacous., Vol. AU-16, No. 1, pp. 68-72 (March 1968).

14. A. Papoulis, Probability Random Variables, and Stochastic Processes, pp. 241-244 (McGraw-Hill, New York 1965).

15. D T. Magill, "Optimal Adaptive Estimation of Sampled Stochastic Processes," IEEE Trans. on Automatic Control, Vol. AC-10, No. 4, pp. 434-439 (October 1965).

16. D. T. Magill, "A Sufficient Statistic Approach to Adaptive Estimation," Conference Record, International Federation on Automatic Control, London, England (June 1966).

17. B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell Sys. Tech. J., Vol. 49, No. 8, pp. 1973-1986 (October 1970).