[redacted]

NRL Report 8295

UNCLASSIFIED

Multirate Processor (MRP) for Digital Voice Communications

G. S. Kang, L. J. Fransen and E. L. Kline

Systems Integration and Instrumentation Branch
Communications Sciences Division

March 21, 1979

NAVAL RESEARCH LABORATORY
Washington, D.C.

Approved for public release; distribution unlimited.

84 02 08 032

CONTENTS

MULTIRATE PROCESSOR (MRP) FOR DIGITAL VOICE COMMUNICATIONS

INTRODUCTION

The multirate processor (MRP) is a design option for use in upgraded Department of Defense (DOD) voice communications. As the name implies, the MRP is a voice processor in which a single voice processing algorithm provides both low and high data rates, 2.4 and 9.6 kilobits per second (kb/s). The high data rate is for the transmission of high-quality speech over wideband channels (line of sight radio links). The low data rate is for the transmission of lower quality (but highly intelligible) speech to those users who do not have access to wideband links or rely exclusively on narrowband links, such as high-frequency (HF) channels.

The most significant characteristic of the MRP is that the bit stream of the high data rate mode contains the bit stream of the low data rate mode as a subset. This embedded data structure makes it possible to interconnect, without user intervention, a wideband and narrowband system via a rate converter located somewhere along the link, thus maintaining end-to-end security. (In this report, a 2.4 kb/s system is referred to as a narrowband system, and a 9.6 kb/s system is regarded as a wideband system. However, some may prefer to also call a 9.6 kb/s system a narrowband system, because 9.6 kb/s speech can be transmitted over a 3 kHz telephone channel, which is a narrowband channel.)

In the framework of the MRP both narrowband and wideband communication resources are regarded as a single, integrated capability. Thus, most of the narrowband communication plant of DOD or non-DOD can be employed for wideband communications to:

- Increase system survivability through rerouting,
- Provide secure connectivity between wideband and narrowband users, and
- Increase channel capacity during peak loading by lowering the data rate.

This MRP concept was conceived by the authors in 1975. It took 3 years of effort to devise a voice processing algorithm in which the wideband mode (9.6 kb/s) was equivalent to or better than the presently deployed 16 kb/s CVSD family and in which the embedded narrowband mode (2.4 kb/s) was compatible with other narrowband voice terminals in development under DOD sponsorship. (The acronym CVSD and the acronyms in the next paragraph are explained in Appendix A.)

On September 26, 1976, the Director, Telecommunications and Command and Control Systems (DTACCS) (the predecessor of (C^3I) tasked the Navy to undertake MRP technology and system tasks. The objective of these tasks was to determine through system analysis

Manuscript submitted November 1, 1978.

and demonstration, by DSARC III, the validity of including the MRP as an element of the AUTOSEVOCOM II program. These tasks were conducted on a noninterfering, nonduplicative basis with AUTOSEVOCOM II and ANDVT programs. The MRP program, however, would coordinate activities with ANDVT, ESVN, and AUTOSEVOCOM II programs, the Office of World Wide Secure Voice Architect, the Narrowband Digital Voice Processor Consortium, and other agencies pursuing related work.

Accordingly the Navy, assisted by IBM, carried out a preliminary system study of the MRP. Additional study tasks related to system analysis and component technology are to begin shortly. In addition a demonstration model of the MRP is being fabricated.

The consequence of an integrated approach to wideband/narrowband systems is that it directly addresses the transition from narrowband to wideband communication. The communication architect is given freedom to expand the secure voice user base using the present analog plant while upgrading the digital wideband service at a rate consistent with budget availability, instead of being forced to choose between increased near-term digital upgrade costs and fewer secure-voice subscribers.

This report documents the in-house R&D efforts related to the formation of the MRP concept, synopsizes operational usage, and describes the voice-processing algorithm including samples of MRP-processed speech.

MRP CONCEPT

The unique characteristic of the MRP is that the bit stream representing voice encoded at 9.6 kb/s also contains the bit stream representing voice encoded at 2.4 kb/s. Figure 1 depicts the data structure of the MRP for each frame of speech data (22.5 ms, 180 speech samples). The 54 bits represent the set of data required for the generation of 2.4-kb/s speech: one bit for synchronization, 41 bits for synthesis filter weights, and 12 bits for excitation signal. The 162 bits represent the supplementary data required for the generation of 9.6-kb/s speech: three bits for synchronization and 159 bits for improved excitation signal.

The data are structured in a building-block form, in which the wideband mode shares the data belonging to the narrowband mode. The use of a single voice-processing principle, linear predictive coding (LPC), makes it possible to form the embedded data structure.

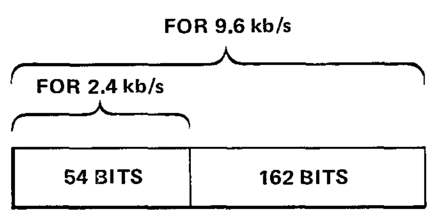


Fig. 1 — MRP data structure of each frame

Because of this embedded data structure, 9.6 kb/s can be converted to 2.4 kb/s by simply deleting 162 bits. On the other hand 2.4 kb/s can be converted to 9.6 kb/s by inserting 162 bits (the message indicator which tells the receiver that speech is originally encoded at 2.4 kb/s) in each frame. The rate conversion can be made anywhere within the link without user intervention while synchronization and encryption are maintained.

OPERATIONAL BENEFITS OF THE MRP SYSTEM

As a result of the unified voice processing technique for the wideband and narrowband operational modes and of flexibility of data rate conversion, the MRP can provide numerous operational benefits, described in the following paragraphs.

Elimination of Wideband/Narrowband Tandem Loss

The importance of DOD secure voice communications, including conferencing, is evidenced in virtually every master plan for C^3 existing today. The uses of secure voice are not limited to internal DOD application but extend to the State Department, civil agencies, NATO, and allied forces as well.

Presently the data rate for each user is selected by the speech quality required, the availability of transmission channel bandwidth, and such considerations as leased line costs. Wideband systems operating at 32 kb/s (or greater) have virtually no degradation in speech quality (Appendix B) but require large transmission channel bandwidth. The availability of these wideband trunks and access lines is limited, thus some users must rely primarily on narrowband channels. In the face of this dilemma the DOD solution is to use two voice processors, one for wideband at 16 kb/s or 32 kb/s and the other for narrowband at 2.4 kb/s. Therefore a requirement exists for tandem operation between wideband and narrowband users.

The incompatibility of these voice processors results in present and planned wideband and narrowband systems being tandemed only via digital to analog and analog-to digital conversions. These conversions degrade speech intelligibility and quality below that of a single narrowband communication system. Degradation is even more severe if speech is contaminated by background noise (Appendix B). During the past 2 years a number of researchers have tried to improve analog tandem performance [1-4]. The conclusion is that no significant improvement presently seems feasible. More important, however, is that an analog interface prevents end-to-end security, since speech data must be decrypted at the tandem junction.

The implementation of the MRP system eliminates the tandem loss, because the rate conversion is carried out on the digital data stream, and provides end-to-end security, because the digital stream is never decrypted at the tandem junction.

Improved Connectivity and Survivability

The DOD voice system must support operations ranging from peacetime to national emergency, with a user priority structure that ensures availability to priority users. The MRP system satisfies the priority-user requirement by providing rerouting options in the event of network outages for wideband users. Calls placed on a wideband system or by priority users can be rerouted on government, public, or foreign analog telephone networks. Likewise, lower priority wideband calls can be assigned to lower-data-rate channels rather than being blocked or preempted. The system controller will sense the operating environment and provide the logical capability to automatically reroute calls or reduce the data rate required.

Increased Channel Capacity

Under peak loads greater channel capacity can effectively be created by the MRP system by reducing data rates and rerouting wideband communications to less expensive analog trunks at the expense of reduced speech quality or by multiplexing a number of lower rate channels into a single wideband link.

Unified Communication-Security System

The MRP makes it possible to employ a single communication-security (COMSEC) principle that provides intrasystem end-to-end encryption regardless of data rate. The embedded data structure will permit passage of electronic keying material across system interfaces.

Ease of Conferencing

An outstanding merit of the MRP is its capability to address both narrowband and wideband users simultaneously and receive messages from a narrowband or wideband user. Since there is no analog tandeming, the minimum speech quality attainable in a conference situation is equivalent to the quality obtainable from 2.4-kb/s LPC.

MRP VOICE ALGORITHM

The MRP voice algorithm is an extension of the LPC technique. In this technique a speech sample is represented by a weighted sum of past speech samples (one-step-forward prediction model). The speech signal is transformed into a set of weighting factors (prediction coefficients) and the prediction error (the residual).

Grossly stated, the LPC technique can be used in two ways for speech digitization. One way it can be used is as a means of *inverse filtering* the voice coloration induced by the human vocal tract [5,6]. A sufficient number of prediction coefficients estimated from the given speech waveform removes much of its resonance-frequency components, leaving considerably whitened prediction residual. This is an approach (similar to H. Dudley's approach [7,8]) that separates the vocal excitation function from the vocal-tract transmission function. The prediction residual can be remodeled with an overall data rate of 2.4 kb/s or greater. This approach performs an open-loop analysis, in which there is no interaction between the given speech waveform and the reconstructed speech waveform.

The LPC technique can also be used as a means of *differential coding* for an efficient waveform transmission of speech [9]. A limited number of prediction coefficients are derived to minimize the mean-square difference between the given speech waveform and the reconstructed waveform as seen by the receiver. This closed-loop LPC analysis does not make a rigid distinction between the vocal excitation function and the vocal tract transmission function. The overall data rate is 6.4 kb/s or greater.

All current narrowband LPC systems operating at 2.4 kb/s are implemented in the inverse filtering method, although wideband LPC systems operating at a data rate of 6.4 kb/s or greater may be implemented by either method.

The MRP narrowband algorithm is presented in Appendix C. As indicated previously, the implementation is identical to the current 2.4 kb/s LPC used for the ANDVT program. Justification of the MRP wideband algorithm, however, requires that the following topics be discussed:

- Review of LPC analysis and synthesis.
- Summary of previously known wideband LPCs.
- Motivation behind the MRP wideband algorithm.
- Description of the MRP wideband algorithm.
- Experimental results.

These topics will be discussed in the following subsections.

Review of LPC Analysis and Synthesis

LPC based on the inverse filtering principle performs an open loop (feed-forward) analysis in which the present speech sample is represented by a weighted sum of past speech samples (an autoregressive model). The prediction error, $F_o(z)$, in terms of the speech signal, $S(z)$, and the predictor, $P_o(z)$, is expressed

$$F_o(z) = [1 - P_o(z)]S(z), \tag{1}$$

where

$$P_o(z) = \sum_k \alpha_o(k) z^{-k}, \tag{2}$$

in which $\alpha_o(k)$ is the kth prediction coefficient estimated in an open-loop LPC analysis. The quantized prediction residual, $X_o(z)$, is

$$\begin{aligned} X_o(z) &= F_o(z) + Q_o(z) \\ &= [1 - P_o(z)]S(z) + Q_o(z), \end{aligned} \tag{3}$$

where $Q_o(z)$ is open loop residual quantization noise.

The synthesis filter output, $Y_o(z)$, in terms of prediction coefficients and quantized prediction residual is

$$Y_o(z) = \frac{X_o(z)}{1 - P_o(z)}$$

$$= S(z) + \frac{Q_o(z)}{1 - P_o(z)}. \tag{4}$$

A block diagram of an open-loop residual transmission system is shown in Fig. 2.

On the other hand LPC based on the differential coding principle performs a closed-loop (feedback) analysis in which the present speech sample is represented by a weighted sum of past speech samples as seen by the receiver (if no transmission error is present). The LPC analyzer output, X_c, in terms of the predictor, $P_c(z)$, and closed-loop quantization noise, $Q_c(z)$, is obtained from

$$X_c(z) = S(z) - \frac{P_c(z)}{1 - P_c(z)} X_c(z) + Q_c(z),$$

where

$$P_c(z) = \Sigma\, \alpha_c(k) z^{-k}. \tag{5}$$

Thus

$$X_c(z) = [1 - P_c(z)]\,[S(z) + Q_c(z)], \tag{6}$$

in which $\alpha_c(k)$ is the kth prediction coefficient.

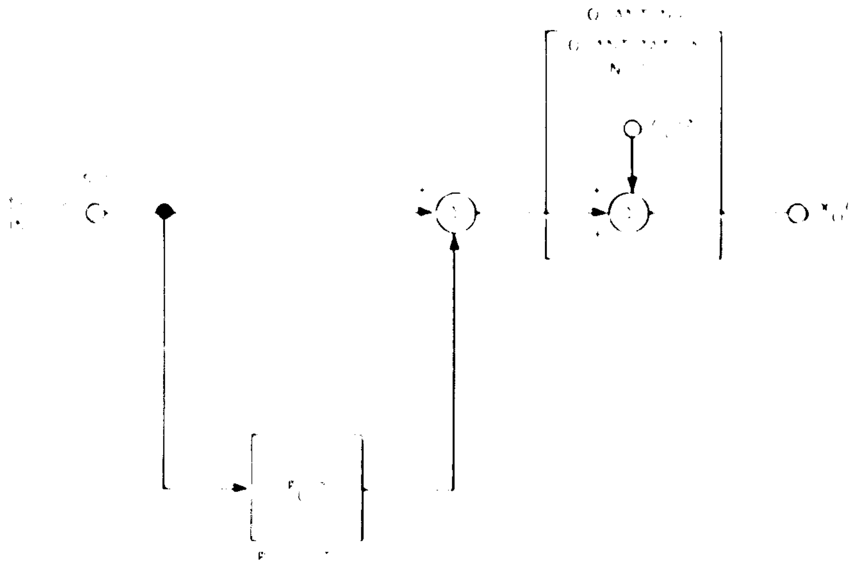
The synthesis filter output, $Y_c(z)$, in terms of prediction coefficients and the analyzer output, is

$$Y_c(z) = \frac{X_c(z)}{1 - P_c(z)}$$

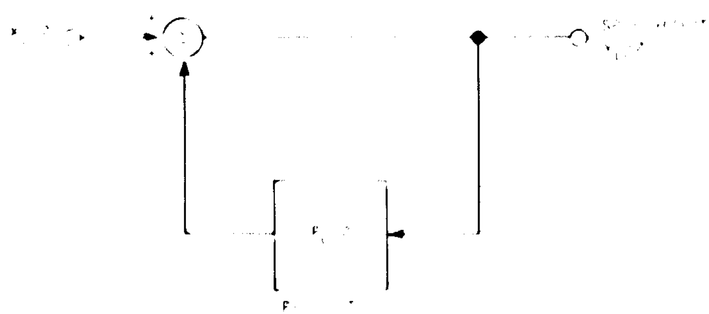
$$= S(z) + Q_c(z). \tag{7}$$

A block diagram of a closed-loop residual transmission system is shown in Fig. 3. This method is commonly known as adaptive predictive coding (APC).

The open-loop residual quantization noise as seen by the receiver output has a spectral shape similar to that of the speech signal as noted by equation 4, whereas closed-loop residual quantization noise as seen by the receiver output has a flat spectrum, as noted by equation 7. In the absence of quantization noise $P_o(z) = P_c(z)$; both systems are identical.

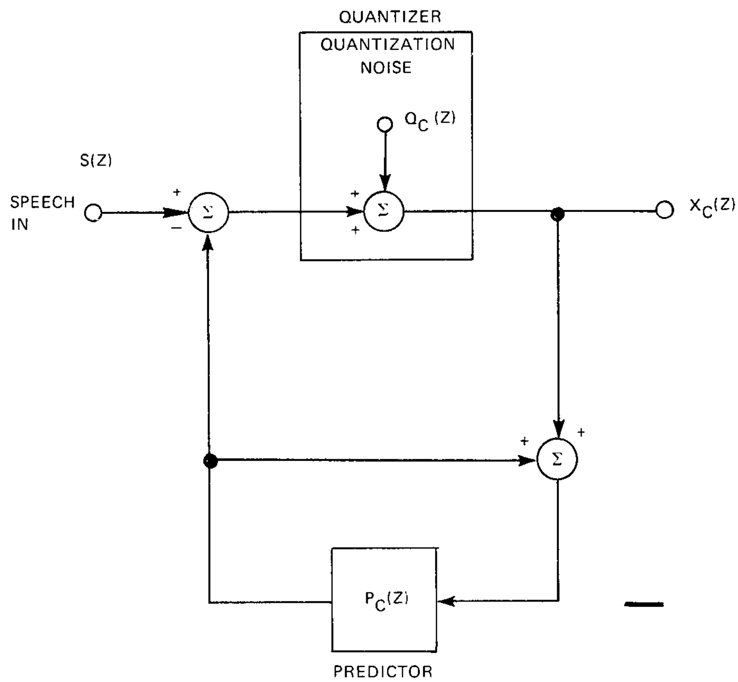


(a) Transmitter

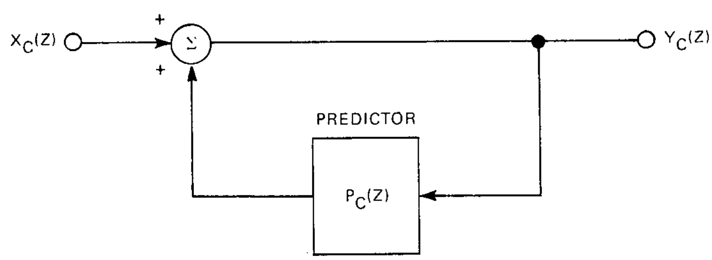


(b) Receiver

Fig. 2 — Wideband LPC system with open loop residual quantization



(a) Transmitter



(b) Receiver

Fig. 3 — Wideband LPC system with closed-loop residual quantization

In either the open-loop or closed-loop case prediction coefficients are estimated through the minimization of the chosen errors. Open-loop prediction coefficients may be estimated by a sample-by-sample form of computation [10] or alternatively by a frame-by-frame form of computation (Appendix C). Closed-loop prediction coefficients are estimated by an iterative procedure, but they are often replaced by open-loop prediction coefficients. In this case the solution is suboptimum, but the computations become simpler. As usual a set of prediction coefficients are transformed to an equal number of reflection coefficients. An advantage of transmitting reflection coefficients is that if the magnitude of each coefficient is less than unity, the synthesis filter stability is assured [10].

The prediction error is a byproduct of the linear predictive analysis, because a speech sample is not always perfectly represented by a weighted sum of past samples. To validate a linear prediction model of speech, the prediction error is allowed (but minimized in the mean square sense) in the estimation of the prediction coefficients. In the framework of linear prediction analysis and synthesis the prediction error behaves as an ideal excitation signal.

The LPC performance is further enhanced by the removal of long-term correlated speech components at the pitch period (if speech is voiced). Thus, the analyzer contains a notch filter, $H_n(z)$, and the synthesizer has a comb filter, $H_c(z)$:

$$H_n(z) = 1 - \beta z^{-T} \tag{8}$$

and

$$H_c(z) = \frac{1}{1 - \beta z^{-T}}, \tag{9}$$

where T is the pitch period and β is a correlation coefficient between the delayed and undelayed signals.

Recently, additional LPC improvements were achieved by feeding filtered residual quantization noise back to the quantizer input. This method may be applied to either a closed-loop configuration [11] or an open-loop configuration [12]. The purpose of this arrangement is to shape residual quantization noise in such a way that it becomes perceptually more tolerable. In effect a quantization noise shaper accomplishes one or both of the following:

- Shifts quantization noise more in the frequency regions where speech energy is greater and less where speech energy is weaker.

- Reduces quantization noise in the lower end of the frequency spectrum.

In closed-loop LPC, quantized residual samples are customarily transmitted sample by sample. Therefore a wideband LPC system operating at a data rate of 6.4 or 9.6 kb/s would transmit the residual at 1 bit per sample with a suitable selected speech sampling rate and a suitable number of bits assigned to each prediction coefficient.

On the other hand open-loop LPC operating at this data rate often favors the transmission of the baseband residual (typically below 1 kHz) in lieu of the entire residual. Although there are some pitfalls (as will be discussed later), it is possible to recreate *spectrally compatible* upperband residual from the baseband residual on account of its relatively flat spectral characteristic. In this approach the preservation of the time waveform is no longer desired.

Previously known wideband LPC systems will be briefly described to establish the commonality and originality of the MRP voice algorithm. For convenience open-loop LPC devices and closed-loop LPC devices will be described in separate subsections.

Summary of Previous Wideband Open-Loop LPC Devices

The following brief descriptions are of wideband LPC devices which transmit open-loop quantized prediction residuals. Data rates range between 6.4 kb/s and 16 kb/s.

ITT 9.6-kb/s LPC Device (J. G. Dunn, 1971)

This device is a completely hardware implemented voice digitizer operable in a full duplex mode [13]. It is remarkable that this device was fabricated as early as 1970, when LPC was relatively unknown. In 1972 the authors of this report talked over this device. It was far superior to existing CVSD equipments operating at the same data rate of 9.6 kb/s. Because only two prediction coefficients are used, it may not be categorized into the inverse-filtering mode of wideband LPC. However, it processed the open-loop prediction residual and transmitted it in the form of delta modulation.

NRL Embedded 16/4.8/2.4-kb/s LPC Device (G. S. Kang and L. Fransen, 1975)

This device, which demonstrates real-time 16-kb/s LPC with embedded 4.8-kb/s and 2.4-kb/s modes, is now superseded by the 9.6/2.4-kb/s MRP discussed in this report. The analyzer derives ten reflection coefficients and a set of prediction residuals from a ten-section cascaded lattice analysis filter. The pitch information for the narrowband mode is obtained from the autocorrelation function of the prediction residual. The information for the 2.4-kb/s LPC mode is transmitted by a frame rate of 44.444 Hz (as in current 2.4-kb/s LPC), whereas the information for 4.8-kb/s mode is transmitted at a double rate. The 16-kb/s mode uses the same synthesis filter as the 4.8-kb/s mode, but the excitation signal is derived from the residual as follows: the prediction residual is down-sampled by a factor of 3 to 1, transmitted at four bits per sample with block companding, and spectrally flattened by the use of an inverse filter at the receiver. Although the intelligibility score for the wideband mode is virtually equal to that for the 4.8-kb/s mode (93), the synthesized speech at 16 kb/s is much more robust (understandable) in the presence of background noise.

SRL 6- to 9.6-kb/s LPC Device (D. T. Magill, C. K. Un, and S. E. Canon, 1975)

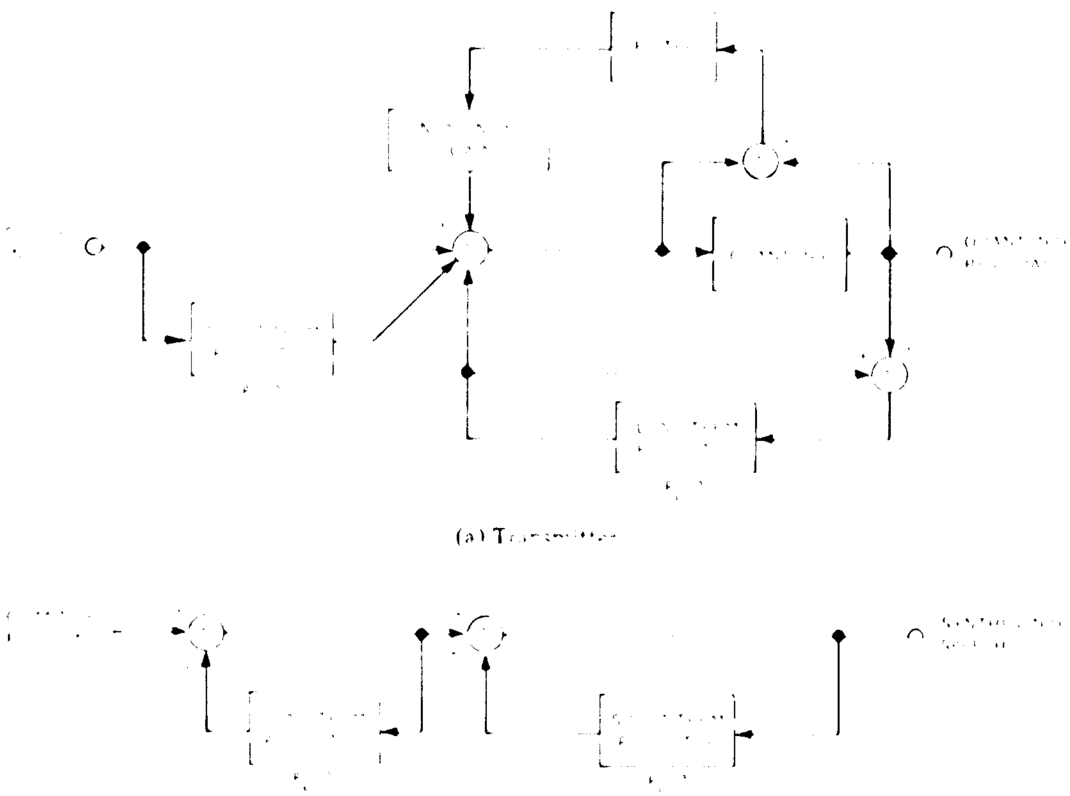
This Stanford Research Laboratory wideband LPC device is the result of a study contract entitled "Speech Digitization Excitation Study" for the Defense Communication Agency [14]. Ten filter coefficients are derived via the autocorrelation method of LPC analysis. Subsequently the prediction residual is generated, down-sampled by a factor of 4 to 1, and encoded by an adaptive delta modulator with hybrid (syllabic and instantaneous) companding. The received residual is linearly interpolated, it is spectrally flattened by a full-wave rectifier, and random noise is added prior to the excitation of the synthesis filter. The data rate can be set between 6 and 9.6 kb/s. No intelligibility scores are available.

French IBM 9.6- or 7.2-kb/s LPC Device (D. Esteban et al., 1978)

In this device the baseband residual (300 to 1000 Hz) is transmitted in the form of splitband filter outputs processed by quadrature mirror image filters [15]. Six of eight filter outputs, each down sampled by a factor of 8 to 1, are transmitted along with eight filter coefficients and the upperband residual amplitude information. At the receiver, six split band filter outputs are combined to form the baseband residual. The upperband residual is regenerated from the baseband residual by rectification, random noise addition, bandpass filtering, and energy calibration. The synthesized speech at 9.6 kb/s appears to be as good as that processed by 9.6 kb/s APC, according to independent testing.

BTL Wideband Coder (B. S. Atal and M. R. Schroeder, 1978)

The Bell Telephone Laboratories wideband coder (of an unspecified data rate) combines two approaches in one: inverse filtering to process open loop prediction coefficients and residual, and differential coding to remove long term correlated pitch redundancies [12]. Furthermore quantization noise is shaped by feeding the quantization error back to the quantizer input as indicated in Fig. 4. The pitch loop reduces noise during a sustained



(a) Transmitter

(b) Receiver

Fig. 4 — Wideband coder (after Atal and Schroeder [12])

voice state. The noise-shaping filter reduces noise in the interformant region and lower frequency range. The data rate must be approximately 16 kb/s, because the residual is transmitted in three levels. The synthesized speech is almost toll quality.

Philco-Ford 16-kb/s LPC Device (J. R. Welsh and C. Teacher, 1976)

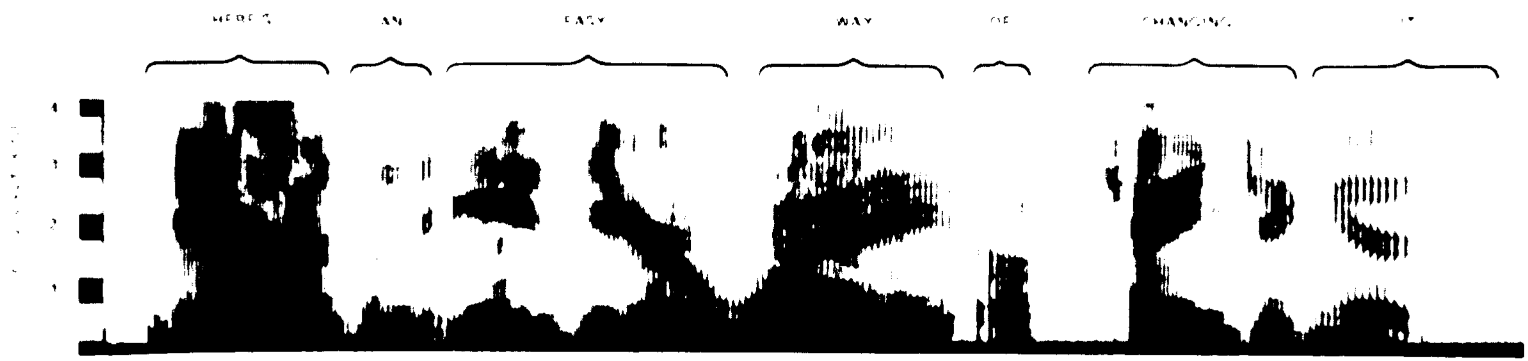
This device is a splitband voice digitizer in which the upperband is high-passed LPC synthesized data at 3.2 kb/s and the baseband is a 1-kHz low-passed speech signal transmitted at 12.8 kb/s [16]. The upperband speech is generated by the use of ten prediction coefficients and the pitch excitation signal (a pulse train or random noise). The baseband is transmitted in six-bit logarithmic PCM at a down-sampled rate of 3 to 1. As expected, there is no coherency between the upperband and baseband in terms of pitch epoch and pitch period. The human ear, however, is surprisingly insensitive to this sort of misalignment in the presence of a well-reproduced baseband. In fact a clean pitch-excitation signal for the upperband can be more desirable than poorly regenerated upperband from the baseband by the use of a spectral flattener, a common technique employed by a baseband excitation method. This method has been programmed on the Philco-Ford signal processor for a real-time demonstration. Some of those who conversed over this device were favorably impressed by its responsiveness to casual conversations. Although not stated by the authors of Ref. 16, this method can be made to embed a 3.2-kb/s mode in the 16-kb/s mode.

NRL Embedded 12.6/2.4-kb/s LPC Device (G. S. Kang and L. Fransen, 1976)

This device is an extension of the Philco-Ford 16-kb/s LPC device in which the baseband speech signal is replaced by the baseband residual and 3.2-kb/s LPC is replaced by 2.4-kb/s LPC. The baseband residual is low-pass filtered at 1 kHz, down-sampled by a factor of 4 to 1, and encoded in five bits per sample after block companding. The upperband excitation signal is a high-passed narrowband excitation signal in which the excitation power is derived from the transmitted baseband residual samples. If speech is voiced, the upperband excitation power is pitch-synchronously updated by the quantity proportional to the baseband residual power over one pitch period. If speech is unvoiced, the upperband excitation is updated four times per frame. Furthermore the voiced excitation is initiated somewhere near the peak amplitude of the baseband residual when going from an unvoiced to voiced state, in an attempt to provide an initial phasing between the baseband and the upperband.

Figure 5 shows spectrograms of original speech and synthesized speech at 12.6 kb/s. The discontinuities of pitch lines near 1 kHz are not very distinguishable by the human hearing mechanism because of the masking due to strong first-formant frequency. The synthesized speech is quite natural, and there seems to be no audible quantization noise.

This method may not be practical with high-level noise-contaminated speech, because the upperband is pitch excited. However, this method will provide a wideband capability to a narrowband voice processor with a small number of additional computations. No efforts have been made to reduce the high data rate from 12.6 kb/s to 9.6 kb/s.



(a) Original speech

(b) Synthesized speech at 12.6 kb/s (mixed excitation). The labels A and B at the right indicate (A) 2.4-kb/s LPC excitation and (B) 10.2-kb/s baseband residual excitation

Fig. 5 — Spectrum of original speech and of synthesized speech (mixed excitation modes)

Summary of Previous Wideband Closed-Loop LPCs

The following are brief descriptions of wideband LPC devices which transmit closed-loop quantized prediction residual. Data rates vary from 6.4 kb/s to 16 kb/s.

BTL 10-kb/s LPC Device (B. S. Atal and M. R. Schroeder, 1970)

This earlier Bell Telephone Laboratories wideband coder [9] is a forerunner of virtually all current wideband coders with closed-loop residual quantization and a pitch loop, in which both prediction coefficients and quantization step size are updated periodically. The prediction residual is quantized in two levels, and the binary difference signal is transmitted at a rate of 6.67 kb/s with overhead data at approximately 3 kb/s. The overhead data include the pitch period, the quantization step size, and eight prediction coefficients, which are all updated every 5 ms. Subjective tests have indicated that the quality of the synthesized speech was better than that of logarithmic PCM speech with five bits per sample. As was stated earlier, this method is commonly known as adaptive predictive coding (APC).

GTE Sylvania 6.4-kb/s Wideband Coder (1973)

This device is part of a secure voice terminal developed for the U.S. Government, in which the principle of APC is used to code speech at 6.4 kb/s. The speech is sampled at 5120 Hz, from which four reflection coefficients are derived once per frame via the inversion of the autocorrelation matrix. As a byproduct a parameter related to the rms value of the error is also obtained. The pitch period is obtained, once per frame, by the average magnitude difference function [17]. The prediction residual is transmitted in one bit per sample. Despite plainly audible quantization noise in the synthesized speech background the device is much more responsive to casual conversations, unlike any 2.4-kb/s LPC. Yet (from a test under the auspices of the DOD Narrowband Digital Voice Processor Consortium in 1975) the intelligibility and quality scores fall below those attained by the current (and best) 2.4-kb/s LPC device under all operational conditions except with severely noise contaminated speech (speech with helicopter noise).

GTE Sylvania 9.6-kb/s Wideband Coder (1975)

This device is another wideband coder designed by GTE Sylvania [18,19] and uses adaptive residual coding (ARC). Although the functional block diagram of ARC is identical to the previously mentioned APC, both methods are substantially different in the parameter estimation process. The prediction coefficients are not calculated by matrix inversion as in the APC; rather they are computed by stochastic estimation or Kalman filtering, and they are updated many times per frame, normally after each sample of the residual is computed and quantized. The ARC has been programmed to operate in real time and has been tested for intelligibility and quality. The intelligibility scores (again from the test under the auspices of the DOD Narrowband Digital Voice Processor Consortium in 1975) indicate no decided advantage of the 9.6-kb/s ARC over the 6.4-kb/s APC.

GTE Sylvania 16-kb/s Wideband Coder (A. J. Goldberg et al. 1976)

GTE Sylvania developed [illegible] for 16-kb/s APC with an adaptive multilevel quantizer (referred to as APCQ) [illegible] a full-duplex [illegible] on two GTE Sylvania programmable [illegible] [21]. [illegible] placed on the investigation of various quantizers which provide [illegible] in the presence of transmission errors. GTE investigated [illegible] fixed (but updated frame by frame) quantizer and [illegible] fixed quantizer, because it outperformed other[s] in a high-error [illegible] that the APCQ outperforms a 16-kb/s CVSD at a low error rate [illegible] that a CVSD outperforms the APCQ at a higher error rate.

GTE Sylvania [illegible] 16 [illegible] (A. J. Goldberg 1978)

In this [illegible] 2.4-kb/s LPC and 13.6-kb/s APCQ were combined [22]. [illegible] of 3400 Hz and 2867 Hz the quantization levels are four and five [illegible]. No [illegible] or quality test scores are available to date.

BBN 16-kb/s Wideband Coder (M. Berouti and J. Makhoul 1978)

The Bolt, Beranek and Newman 16-kb/s voice digitizer is basically an APC device with a noise-shaping filter [11]. By the [illegible] method, eight log area ratios (in 33 bits) and a gain factor (in six bits) are [illegible] for each frame every 25 ms. The remaining 14 kb/s are assigned for the transmission of the residual. To achieve the best speech quality, the residual is transmitted [illegible] coding. Sample recordings of speech synthesized by this system are [illegible].

Rationale of the MRP Algorithm

The voice processor [illegible] in the MRP provides the following capabilities:

- It generates two data rates: 2.4 kb/s and 9.6 kb/s.
- The 2.4-kb/s data stream is [illegible] in the 9.6-kb/s data stream.
- The 2.4-kb/s [illegible] with current 2.4-kb/s LPC devices.
- The 9.6-kb/s mode produces speech intelligibility and quality equal to or better than that produced by a 16-kb/s CVSD.

The MRP can be simply a combination of the current 2.4-kb/s LPC and 7.2-kb/s wideband LPC implemented by any one of the previously mentioned wideband coding principles. Unfortunately none of the previously mentioned wideband coders operating at 7.2 kb/s can compete against a 16-kb/s CVSD in performance. Therefore it was necessary to devise a new wideband coding technique. In the generation of this new coding principle, the following areas were considered:

A Common Synthesis Filter for Both Narrowband and Wideband Modes

A system advantage if the MRP wideband mode uses the same synthesis filter as the narrowband mode is the saving of transmission bits, because the filter weights are transmitted as part of the narrowband data. Since the filter weights are transmitted at a rate of 1.822 kb/s, a common filter represents a data-rate savings of 19%. A second advantage is a simpler implementation, because only one set of filter weights is derived at the transmitter and a unified speech synthesis method is implemented at the receiver.

A synthesis filter with ten prediction coefficients provides a fairly good spectral envelope, as indicated in the spectrum shown in Fig. 6. However, a good synthesis filter does not necessarily produce good synthesized speech unless it is excited by a quality excitation signal. This is because the synthesized speech spectrum is a product of the filter frequency response and the excitation spectrum (in the form of a spectral envelope modulated by a carrier spectrum). The quality of narrowband speech is relatively poor mainly because its excitation signal lacks details of an ideal excitation signal. The wideband mode of the MRP eliminates this problem by using the prediction residual as an excitation signal.

In the framework of LPC the prediction residual is an ideal excitation signal because its spectrum is complementary to the filter frequency response for a perfect synthesis of the original speech (in the absence of quantization). Thus certain deficiences of the synthesis filter (such as a lack of zeros in the transfer function) are no longer a problem with the residual excitation, because zeros (if any) are reflected in the residual.

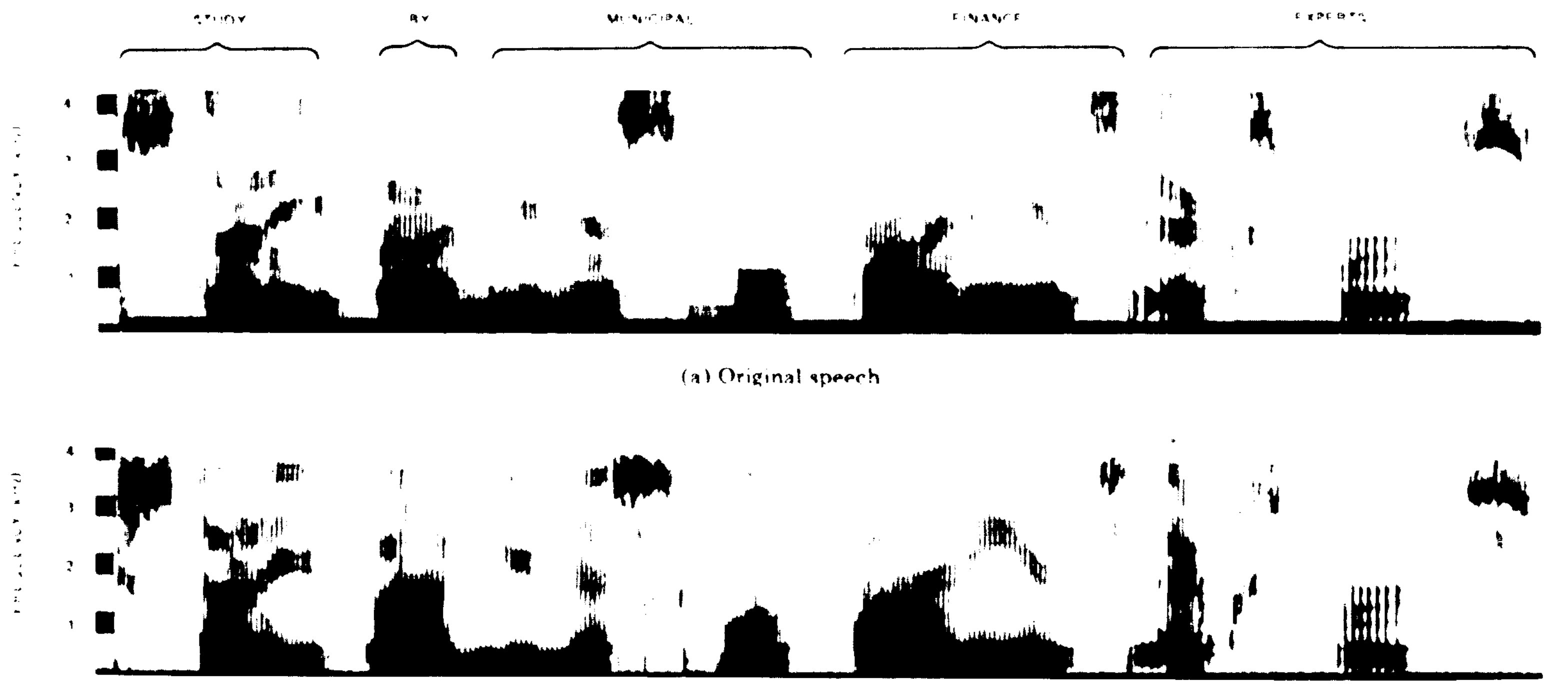
Likewise a slow update rate of filter weights (44.444 Hz) does not create sluggishness in synthesized speech with residual excitation (as often noted in narrowband speech). In fact filter weights can be time invariant as long as the corresponding residual is used for excitation. Such a wideband coder operating at either 9.6 kb/s or 16 kb/s has been programmed by Lincoln Laboratory to run in real time. In terms of the intelligibility and quality it was nearly equivalent with other wideband coders operating at the same data rates.

For the voice algorithm in the MRP, a synthesis filter with ten adaptive prediction coefficients was selected, because it can be shared by both narrowband and wideband modes. The major issue of the MRP wideband mode was how to encode the residual so that synthesized speech at 9.6 kb/s had intelligibility and quality comparable to 16-kb/s CVSD.

More Resolution for Residual Samples

Some of the previously mentioned wideband coders, operating at data rates between 6.4 kb/s and 9.6 kb/s, encoded the residual in one bit per sample and transmitted at the speech sampling rate. Synthesized speech generated by this residual have been described as rough, raspy, gruffy, or gritty. It is inconceivable how speech quality can be improved as long as the residual is transmitted in this form. In fact under certain favorable conditions (with a high-quality microphone and an articulate speaker in a noise-free environment) 2.4-kb/s LPC can produce a more intelligible and pleasing synthesized voice than wideband coders using a one-bit per sample residual transmission.

The human ear is quite tolerant of speech sounds that have spectral envelope distortions (such as CB radio sounds) as long as the distortions remain quasi-stationary. What is



(a) Original speech

(b) 2.4-kb/s speech (processed in real time)

Fig. 6 — Spectrum of original speech and of 2.4-kb/s speech

objectional to most is nonstationary distortions (similar to excitation distortions) which make speech waveforms vary substantially from one pitch period to the next. The previously mentioned wideband coders (and current 9.6-kb/s or 16-kb/s CVSDs) generate speech sounds which fit this description.

Another undesirable feature in a wideband coder that transmits the residual in one bit per sample is the use of a reduced speech bandwidth. If the residual is transmitted at 7.2 kb/s (at an overall data rate of 9.6 kb/s with an embedded 2.4-kb/s data rate), the speech sampling rate must be below 7.2 kHz in order to transmit the overhead data (such as an adaptive quantization level, the pitch period, and the pitch-loop gain). Thus the speech input cutoff frequency must be below 3.6 kHz — possibly 3.2 kHz. A speech bandwidth of 3.2 kHz is adequate for unprocessed speech, but extensive tests have shown that a bandwidth reduced to 3.2 kHz results in lower intelligibility for synthetic speech. This is due to the high-frequency-related speech attributes being obscured by the presence of quantization noise.

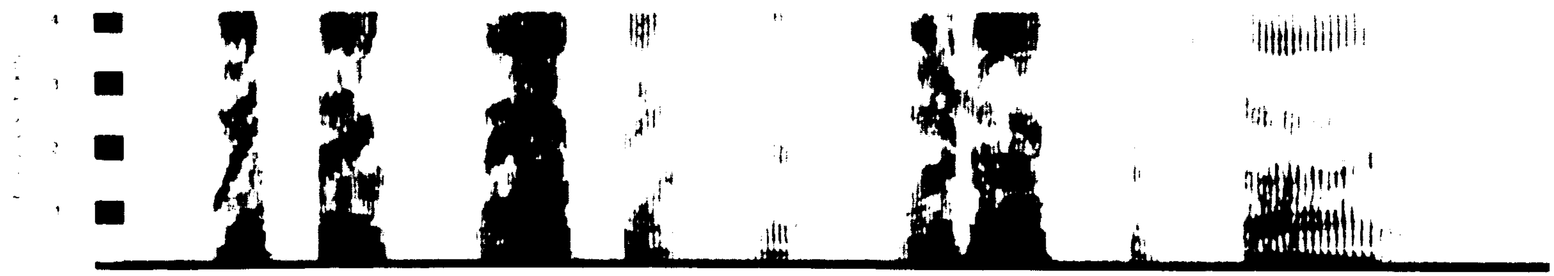
These two observations indicate that a 9.6-kb/s wideband coder should transmit the baseband residual (at a down-sampled rate) in a finer quantization, rather than transmitting the entire residual (at a speech sampling rate) in a coarser quantization. Then the baseband residual should be employed to derive the upperband excitation signal up to 4 kHz. This approach is feasible because the prediction residual processed by ten adaptive prediction coefficients has a relatively flat spectral envelope, as illustrated by the spectrum in Fig. 7. As noted, the prediction residual spectrum has a strong correlation along the frequency axis (depicted by vertical lines in the figure), thus making it possible to regenerate the upperband excitation from the baseband residual.

The majority of previous baseband excitation methods employed some form of spectral flattening to spread the baseband residual into the upperband region. The most commonly used spectral flattener was a nonlinear device, such as a full-wave or half-wave rectifier. Although these nonlinear elements spread the baseband residual spectrum (at an increased attenuation with frequency), they also generated undesirable side effects. One such effect was the crossmodulation of remnant formant frequencies present in the prediction residual (Fig. 7), which made synthesized speech sound fuzzy and unfocused. Furthermore additive quantization noise in the baseband residual becomes multiplicative noise in the resulting upperband excitation signal. Neither of these effects have been evaluated to determine techniques to minimize them.

The MRP voice processor, detailed in the next section, uses a method of regeneration in which the entire baseband is frequency-shifted to create the upperband excitation signal. Therefore this method does not create crossmodulation among remnant formant frequencies or quantization noise in the baseband. The regenerated upperband residual is no longer a replica of the original upperband residual time waveform; instead its short-term spectrum approximates that of the original upperband residual. Comparison of the original upperband residual and the regenerated upperband residual indicates that they are almost acoustically indistinguishable. This new method of generating the upperband excitation signal is partially responsible for the good quality of the MRP wideband mode.

The following discussion concerns the baseband bandwidth. If there is no data-rate restriction, a larger bandwidth is more desirable. This is because the prediction residual still

(a) Original speech



(b) Prediction residual

Fig. 7 — Spectrum of original speech and of the prediction residual

contains nonflat spectral information that is not mapped into prediction coefficients. This includes mostly antiresonance components (zeros), but some weak formant components (poles) are also included. In practice, however, the baseband bandwidth is chosen as a compromise between the available bits for the residual coding (162 bits per frame) and the desired quality for the residual. At a minimum the baseband residual must carry the information on the fundamental pitch frequency to be useful as an excitation signal. Thus the minimum baseband bandwidth is twice the fundamental pitch frequency, if the fundamental pitch component is present in the given speech. Since it may not be present, as in telephone analog communication, the upper cutoff frequency of the baseband must be high enough to include at least two pitch harmonics. Then the human ear will sense the fundamental pitch frequency through its own peculiar mechanism. Since the fundamental pitch frequency for a female can be as high as 400 Hz, the upper cutoff frequency of 1200 Hz is recommended as the implementation for the MRP.

Elimination of Speech Information Below 250 Hz

Speech information below 250 Hz should be eliminated from encoding. One reason is that these frequency components are not essential to speech communications. In fact the presence of these frequency components tends to mask the perception of higher frequency components, which are vital to speech intelligibility. A second reason is that certain speech waveforms do not contain any significant amount of these frequency components. An example would be telephone speech. In certain cases these frequency components are intentionally suppressed in order to minimize interferences caused by 60-Hz-related hums or rumbling noise present in the speaker background (such as shipboard environments). Thus, encoding of virtually nonexistent speech components results in wasting transmission data bits.

If speech components below 250 Hz are eliminated from encoding, a bit saving of 20% is realized for the system in which the baseband residual, up to 1.25 kHz, is transmitted. In practice, however, the actual saving would be more than 20%, because lower frequency components are normally quantized with a finer resolution than for higher frequency components. This is because the human ear is more sensitive to quantization noise contained in a lower frequency range than in a higher frequency range.

Importance of Quantization Noise Shaping

The spectral distribution of quantization noise with respect to the speech spectrum has a significant impact on how noise is perceived by the human ear [11,12]. The human ear apparently prefers quantization noise whose spectrum is congruent with the speech spectrum. The degree of preference increases with diminished noise densities everywhere in the frequency band, but the effect is more pronounced if lower frequency noise is further decreased. An open-loop residual quantization produces quantization noise which has a spectral shape similar to the speech signal and is the first step toward noise shaping for the MRP. In additional the open-loop residual quantization in the MRP will be such that lower frequency components are quantized in a finer resolution than higher frequency components.

Transmission of Spectral Information

The residual may be transmitted in either its time waveform (as in practically all previous wideband coders) or by its spectral waveform. The transformation from one waveform to the other is reversible, and no information is lost in the process. The transmission of spectral information is advantageous to the MRP, because all design features previously discussed are in some way involved with the spectral components of the residual:

- Transmission of the baseband residual only.
- Elimination of frequency components below 250 Hz.
- Finer quantization for lower frequency components.
- Generation of spectrally compatible upperband excitation from baseband residual.

Although the transformation from the time waveform to the spectral waveform requires additional computations, the inherent advantages outweigh increased computations, as indicated in the following discussions.

- Transmitting spectral information does not require low-pass filtering to extract base baseband residual samples. Low-pass filtering implies a flow-form computation of the product sums, where the number of terms can be anywhere between 20 and 50.

- The use of spectral information does not require a nonlinear operation for spectral flattening. Although computations required for nonlinear operation (rectification) are trivial, the generated upperband is often flawed by the presence of crossmodulated remnant formant frequencies, as discussed previously. In the MRP the upperband residual is simply a frequency-shifted baseband residual, brought about by the substitution of the baseband spectra for the upperband spectra. The inverse Fourier transform of the resulting spectrum yields the total excitation signal.

- The use of spectral information does not require high-pass filtering, which was often required in previous baseband excitation methods. High-pass filtering implies a flow-form computation of the product sums with the number of terms varying between 20 and 50, as in low-pass filtering.

- Transmitting spectral information permits encoding of the amplitude and phase spectra. The amplitude spectrum is a necessary part of the total excitation information, and the phase spectrum describes how various frequency components are phased with respect to a fixed time reference (such as the beginning of each analysis frame). Thus the phase spectrum contains more timing information than the amplitude spectrum [23]. A well-preserved phase spectrum reduces the unnatural pitch-to-pitch waveform variations that are so offensive to the human ear. Accordingly the phase spectrum is quantized in a finer step size than is the amplitude spectrum.

- If so desired, the data rate can be increased by transmitting more spectrum information. Since each frequency component is transmitted in five bits, the data-rate increment can be as small as 222 b/s.

- Packet-system transmission can be applied to MRP. At a minimum 54 bits per frame of data are transmitted for 2.4-kb/s speech quality. As network conditions permit, an additional 162 bits per frame of data may be transmitted for 9.6-kb/s speech quality. Beyond this point, additional speech data in five-bit-per-frame multiples may be transmitted for improved speech quality. (The application of the MRP embedded data concept to a packet system was originally considered by Ted Bially of Lincoln Laboratory.)

Problems Associated with the Use of Narrowband Excitation Parameters for Wideband LPCs

In essence the narrowband excitation signal is a drastically simplified model of the prediction residual. It is represented by a quasi-periodic broadband signal if speech is voiced or by random noise if speech is unvoiced. The narrowband excitation signal is parameterized into the pitch period (six bits), the voicing decision (one bit), and the amplitude information (five bits). At a frame rate of 44.444 Hz these parameters are transmitted at a rate of only 533.3 b/s.

The primitive nature of the excitation signal is primarily responsible for the poor quality of narrowband synthesized speech. The extraction of the excitation parameters is somewhat unreliable, even under favorable operational conditions (a person speaking in a clear, articulated manner and a good microphone in a noise-free speaker site). Furthermore, certain sound elements (such as /z/) are neither entirely voiced nor unvoiced, making mapping into a binary voicing decision difficult.

Some previous wideband coders used the voicing decision and the pitch period for the generation of the upperband excitation signal from the baseband residual (or speech signal). (The use of voicing decision is mandatory in a voice-excited wideband digitizer, which transmits the baseband speech signal in place of the residual. The synthesized speech is a combination of the transmitted baseband speech signal and the regenerated upperband excitation signal. Since the baseband speech signal does not contain the information on certain high-frequency unvoiced sounds (/sh/ or /ch/), these sounds are generated from a random-noise generator controlled by the voicing decision.)

In addition most wideband LPC devices have extensively used the pitch period as a long-term prediction parameter to suppress pitch-related frequency components. Although this is effective with clean speech and in the absence of transmission errors, it actually diminishes overall performance when it is used under adverse conditions [21].

As stated previously, anomalies in the excitation signal have a great impact on the quality of synthesized speech. Since the narrowband excitation parameters are somewhat unreliable, they should not be used as control parameters in the generation of the wideband excitation signal. An advantage of having a wideband coder is its robust performance under adverse operating conditions. The MRP wideband mode does not use the narrowband excitation parameters in any form.

Description of the MRP Algorithm

Based on the preceding discussion the general approach to the MRP algorithm is as follows:

- Ten prediction coefficients are extracted from the speech waveform by the same procedures employed in 2.4 kb/s LPC (Appendix C).
- The same filter is used for the speech synthesis at 2.4 and 9.6 kb/s;
- The excitation signal for the narrowband mode is either a quasi periodic broadband signal (if speech is voiced) or random noise (if speech is unvoiced) as is used with the current 2.4 kb/s LPC (Appendix C).
- The excitation signal for the wideband mode is derived from the prediction residual;
- The residual is transformed into a set of amplitude and phase spectrum components and transmitted for the baseband only from 250 to 1250 Hz.
- The residual information is quantized on an open loop basis;
- The phase information is quantized with a finer resolution than is the amplitude information;
- The lower frequency components are quantized with a finer resolution than are the high frequency components.
- The upperband excitation signal is generated by frequency shifting of the baseband signal.

In essence the MRP is a 2.4 kb/s LPC device with an add-on processor for a 9.6 kb/s capability. The add-on processing is nothing more than a means to transmit the residual information. Figure 8 is a block diagram of the MRP. The two blocks that are hatched imply the add-on processing to 2.4 kb/s LPC to produce the MRP. The equipment listed in these two blocks performs the following operations:

- Residual transformation,
- Baseband residual encoding/decoding,
- Excitation signal generation.

Each item is discussed in the following subsections.

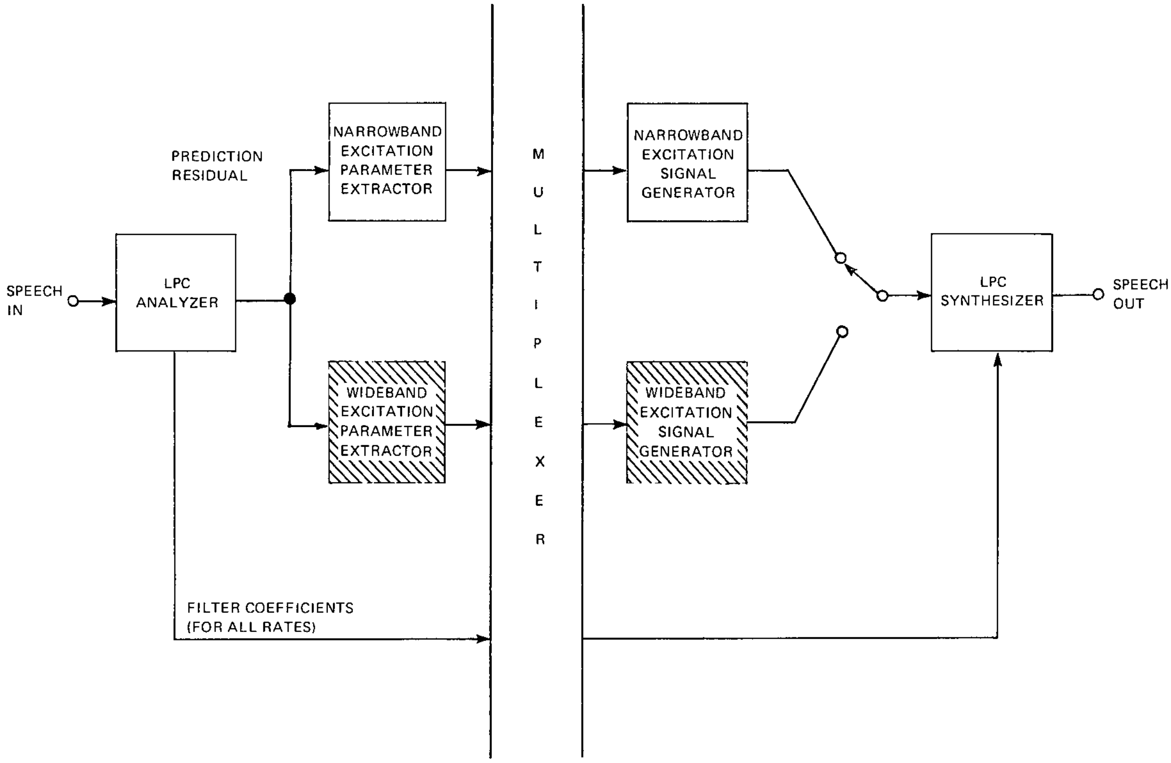


Fig. 8 — MRP voice processor. The hatched blocks are additions to the current 2.4-kb/s LPC device.

Residual Transformation

The residual transformation begins with the transformation of time samples into Fourier components. Let the residual samples be denoted by $x_0, x_1, \ldots, x_{N-1}$, where N is even. Then the real and imaginery components are expressed by

$$a_m = \frac{1}{N} \sum_{n=0}^{N-1} x_n \cos\left(\frac{2\pi mn}{N}\right), \quad m = 0, 1, \ldots, \frac{N}{2}, \tag{10a}$$

and

$$b_m = \frac{1}{N} \sum_{n=0}^{N-1} x_n \sin\left(\frac{2\pi mn}{N}\right), \quad m = 1, 2, \ldots, \frac{N}{2} - 1, \tag{10b}$$

where a_m is the mth real Fourier component and b_m if the mth imaginary Fourier component.

Conversely, time samples can be obtained from a_m and b_m by the following expression:

$$x_n = \left[a_0 + (-1)^n a_{N/2}\right] + 2\sum_{m=1}^{\frac{N}{2}-1}\left[a_m \cos\left(\frac{2\pi mn}{N}\right) + b_m \sin\left(\frac{2\pi mn}{N}\right)\right] \tag{11}$$

$$= \left[a_0 + (-1)^n a_{N/2}\right] + 2\sum_{m=1}^{\frac{N}{2}-1}\left[|c_m| \cos\left(\frac{2\pi mn}{N} + \phi_m\right)\right], \quad n = 0, 1, \ldots, N-1, \tag{12}$$

with

$$c_m = \sqrt{a_m^2 + b_m^2} \tag{13}$$

and

$$\phi_m = \tan^{-1}\left(\frac{b_m}{a_m}\right). \tag{14}$$

where $|c_m|$ is the mth amplitude spectral component and ϕ_m is the mth phase spectral component.

The transform pairs expressed by equations 10 and 11 are reversible. Thus N time samples may be transformed to N spectral components and vice versa. No information will be lost by the transformation process. If the spectral components are unquantized, a back-to-back transformation gives an output identical to the input (at least in theory, if numerical and rounding off errors are neglected). In the presence of quantization, however, the reconstructed time sequence tends to have discontinuities at analysis frame boundaries. Hence it is desirable to have the analysis frames overlap at the expense of a reduction of data transmission efficiency. (Efficiency is reduced because more transmission bits are required for the same amount of residual information.)

For convenience, the analysis frame size was chosen as $N = 128$, of which eight samples are overlapped with the eight samples of the previous analysis frame. This frame size was chosen because the fast Fourier transform algorithm can be exploited for the spectral analysis. On the other hand the overlap sample size of eight is chosen because 120 samples (the analysis frame size less the overlap sample size) is a submultiple of the total number of residual samples contained in two LPC frames. This means that the spectral analysis may be performed three times for every two LPC frames. Figure 9 illustrates how spectral analysis frames are placed in reference to LPC frames.

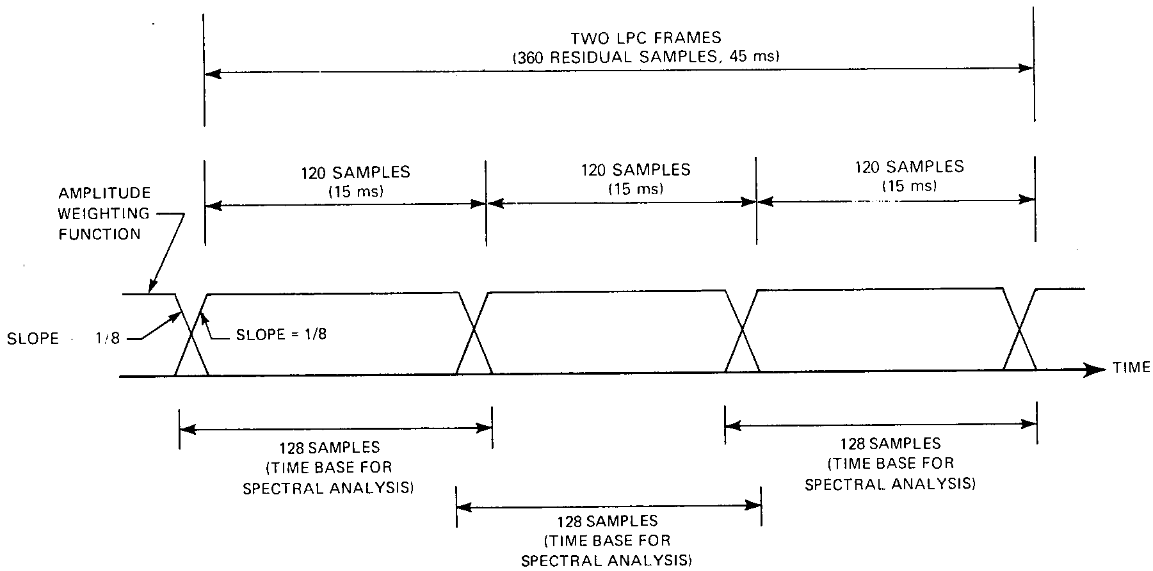


Fig. 9 — Overlap of time samples for spectral analysis

Baseband Residual Encoding/Decoding

For a 9.6-kb/s mode, 216 bits are transmitted for every LPC frame. Of the 216 bits, 54 are embedded data which are used by both the 2.4- and 9.6-kb/s modes, and the remaining 162 bits are for the 9.6-kb/s mode exclusively. The 162 bits are further partitioned as follows:

- LPC frame synchronization: three bits (with the embedded 54 bits containing a fourth LPC frame synchronization bit),
- Error protection of the first four filter coefficients: 12 bits,
- Baseband residual transmission: 147 bits.

The 147 bits in an LPC frame (22.5 ms) correspond to 98 bits in a spectral analysis frame (15 ms). These 98 bits are allocated as follows:

- Amplitude spectrum normalization factor: six bits,
- Error protection of the normalization factor: three bits,
- Seventeen amplitude spectral components (with no error protection): 34 bits,
- Seventeen phase spectral components (with no error protection): 55 bits.

Within the total bandwidth of 4000 Hz the baseband bandwidth ranges from 250 to 1250 Hz. Since spectral components are separated by 62.5 Hz (4000/64 Hz), the baseband contains the fourth through the twentieth component. Table 1 lists the spectral-component frequency locations and the number of bits assigned to each component.

Table 1 — Frequency Location and Resolution of Spectral Components

Index (m)	Frequency (Hz)	Number of Bits: Amplitude (c_m)	Number of Bits: Phase (ϕ_m)
4	250.0	2	4
5	312.5	2	4
6	375.0	2	4
7	437.5	2	4
8	500.0	2	3
9	562.5	2	3
10	625.0	2	3
11	687.5	2	3
12	750.0	2	3
13	812.5	2	3
14	875.0	2	3
15	937.5	2	3
16	1000.0	2	3
17	1062.5	2	3
18	1125.0	2	3
19	1187.5	2	3
20	1250.0	2	3
Total		34	55

To use the full input range of the quantizer, the 17 amplitude spectral components are normalized by the maximum input value. The amplitude normalization factor is encoded into six bits based on the coding table (Table 2). The coding is linear for lower amplitude values and logarithmic for larger amplitude values. The three most significant bits of the amplitude normalization factor are error protected by three additional bits.

Phase spectral components are uniformly distributed from $-\pi$ to $+\pi$ as illustrated in Fig. 10a. The probability density function was obtained from 1,600,000 phase spectral components from the prediction residual generated from both male and female speech samples. Since the distribution is uniform, an equal step symmetric quantizer may be used for the quantization of phase spectral components. The coding may be effected by the algebraic expression

$$\left|PC_m\right| = \mathrm{Int}\left(\frac{|\phi_m|}{2\pi} 2^J\right), \tag{15}$$

where PC_m is the code for the mth phase spectral component, J is the bit number assigned for ϕ_m (3 or 4, as given in Table 1), Int $(\cdot)$ implies an integerizing operation with truncation, and the sign of ϕ_m is extended to the sign of PC_m.

Table 2 — Coding/Decoding Table for the Normalization Factor

Original Value	Code	Decoded Value	Original Value	Code	Decoded Value
0,1	0	1	72-79	32	75
2	1	2	80-87	33	83
3	2	3	88-96	34	92
4	3	4	97-106	35	101
5	4	5	107-118	36	112
6	5	6	119-130	37	124
7	6	7	131-144	38	137
8	7	8	145-158	39	151
9	8	9	159-176	40	167
10	9	10	177-194	41	185
11	10	11	195-215	42	205
12	11	12	216-237	43	226
13	12	13	238-263	44	250
14	13	14	264-289	45	277
15	14	15	290-322	46	306
16	15	16	323-356	47	339
17	16	17	357-393	48	374
18	17	18	394-435	49	414
19	18	19	436-481	50	458
20	19	20	482-532	51	506
21,22	20	22	533-589	52	560
23-25	21	24	590-648	53	619
26-28	22	27	649-720	54	685
29-31	23	30	721-796	55	757
32-35	24	33	797-880	56	837
36-39	25	37	881-974	57	929
40-43	26	41	975-1077	58	1024
44-47	27	45	1078-1190	59	1132
48-52	28	50	1191-1316	60	1252
53-58	29	55	1317-1455	61	1384
59-64	30	61	1456-1609	62	1530
65-71	31	68	1610 or more	63	1692

The phase spectral component is decoded by the expression

$$\left|\phi_m\right| = \left(\left|PC_m\right| + \frac{1}{2}\right)(2\pi)2^{-J}, \tag{16}$$

where ϕ_m is the mth decoded (and quantized) phase spectral component. Again the sign of PC_m is extended to the sign of ϕ_m.

The probability density function of amplitude spectral components (normalized by the maximum spectral component in each analysis frame) is shown in Fig. 10b. The quantizer is designed to minimize the mean-square values of quantization error. Since all normalized amplitude spectral components are quantized to two bits (Table 1), a four-level quantizer is desired. The amplitude transfer characteristic of such a quantizer is expressed by

$$\begin{aligned} y(x) &= y_1, \text{ if } x \leq x_1, \\ &= y_2, \text{ if } x_1 < x \leq x_2, \\ &= y_3, \text{ if } x_2 < x \leq x_3, \\ &= y_4, \text{ if } x_3 < x \leq 1, \end{aligned} \tag{17}$$

where x is the input amplitude, $y(x)$ is the output amplitude, y_1, y_2, y_3, and y_4 are four discrete output amplitude levels, and x_1, x_2, and x_3 are the input amplitude break points.

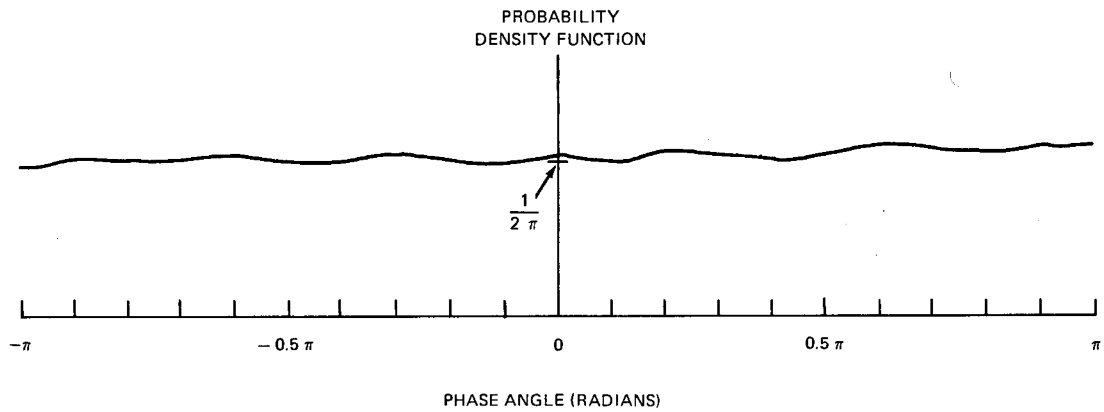
The quantization error is defined as the difference between the actual and the ideal amplitude transfer characteristics:

$$\epsilon(x) = y(x) - x. \tag{18}$$

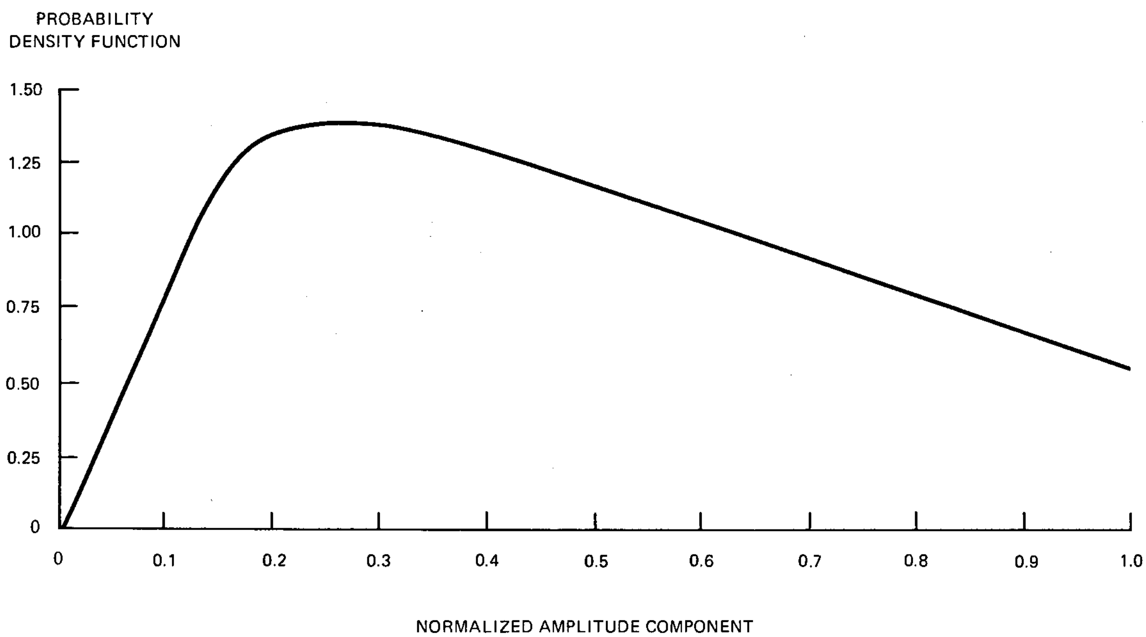
The mean-square value of the quantization error is

$$\begin{aligned} \overline{\epsilon^2} &= \int_0^1 \epsilon^2(x)p(x)\,dx \\ &= \int_0^{x_1} (y_1 - x)^2 p(x)\,dx + \int_{x_1}^{x_2} (y_2 - x)^2 p(x)\,dx \\ &+ \int_{x_2}^{x_3} (y_3 - x)^2 p(x)\,dx + \int_{x_3}^{1} (y_4 - x)^2 p(x)\,dx. \end{aligned} \tag{19}$$

With use of the probability density shown in Fig. 10b the quantizer parameters which minimize the mean-square error are computed as $x_1 = 0.25$, $x_2 = 0.469$, $x_3 = 0.719$, $y_1 = 0.125$, $y_2 = 0.375$, $y_3 = 0.594$, and $y_4 = 0.859$. The amplitude transfer characteristic of the quantizer is shown in Fig. 11. The coding is as detailed in Table 3.



(a) Probability density function of phase components



(b) Probability density function of normalized amplitude components

Fig. 10 — Probability density functions of residual spectra

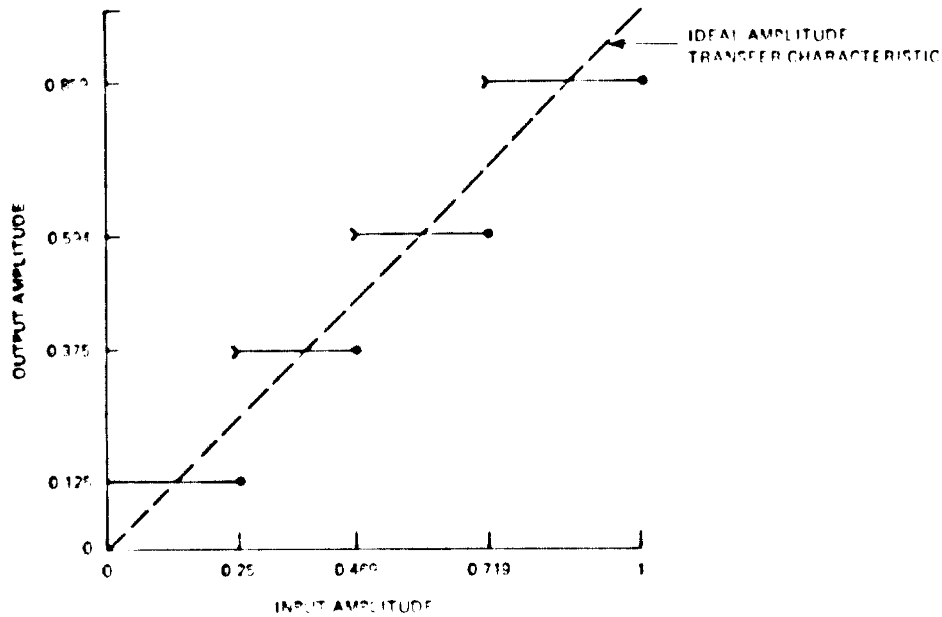


Fig. 11 — Quantizer for amplitude components

Table 3 — Encoding and Decoding Table for Normalized Amplitude Spectral Components

Original Value (x)	Code	Decoded Value
$0 \leq x \leq 0.25$	0	0.125
$0.25 < x \leq 0.469$	1	0.375
$0.469 < x \leq 0.719$	2	0.591
$0.719 < x \leq 1.000$	3	0.859

Excitation Signal Generation

The excitation signal is obtained by the inverse Fourier transform of the composite spectrum (the baseband and the upperband) through the use of equation 12. The baseband spectrum consists of 17 amplitude and phase spectral components that have been transmitted: ϕ_m and $|c_m|$, where $m = 4, 5, \ldots, 20$. The upperband consists of three subbands, each containing the same spectral information as the baseband. Thus,

$$\left.\begin{aligned} \phi_{m+17} &= \phi_m \\ |c_{m+17}| &= |c_m| \end{aligned}\right\}, \quad m = 4, 5, \ldots, 20, \tag{20}$$

$$\left.\begin{aligned} \phi_{m+34} &= \phi_m \\ |c_{m+34}| &= |c_m| \end{aligned}\right\}, \quad m = 4, 5, \ldots, 20, \tag{21}$$

$$\left.\begin{aligned} \phi_{m+51} &= \phi_m \\ |c_{m+51}| &= |c_m| \end{aligned}\right\}, \quad m = 4, 5, \ldots, 20. \tag{22}$$

For the synthesis of unvoiced sounds, the regenerated excitation signal is completely satisfactory, because it is as random as the ideal excitation signal (the prediction residual). Even a burst of noise by a stop consonant is satisfactorily synthesized by the use of the regenerated excitation signal, as is illustrated in Fig. 12.

The regenerated excitation signal is adequate for the synthesis of voiced sounds, because each subband provides a time sequence which undulates pitch-synchronously and has a flat spectral envelope (as flat as the baseband information). In other words it is similar to the prediction residual in the same band. Figure 12 shows the time sequences of the prediction residual and regenerated excitation signal, both within the first upper subband (a frequency range of 1312.5 to 2562.5 Hz and frequency indices from $m = 21$ to $m = 37$). Figure 13 shows the autocorrelation functions of the same time sequences. The regenerated excitation signal is not expected to agree with the original prediction residual in every detail, but the most important pitch information is satisfactorily preserved in the regenerated excitation signal.

Experimental Results

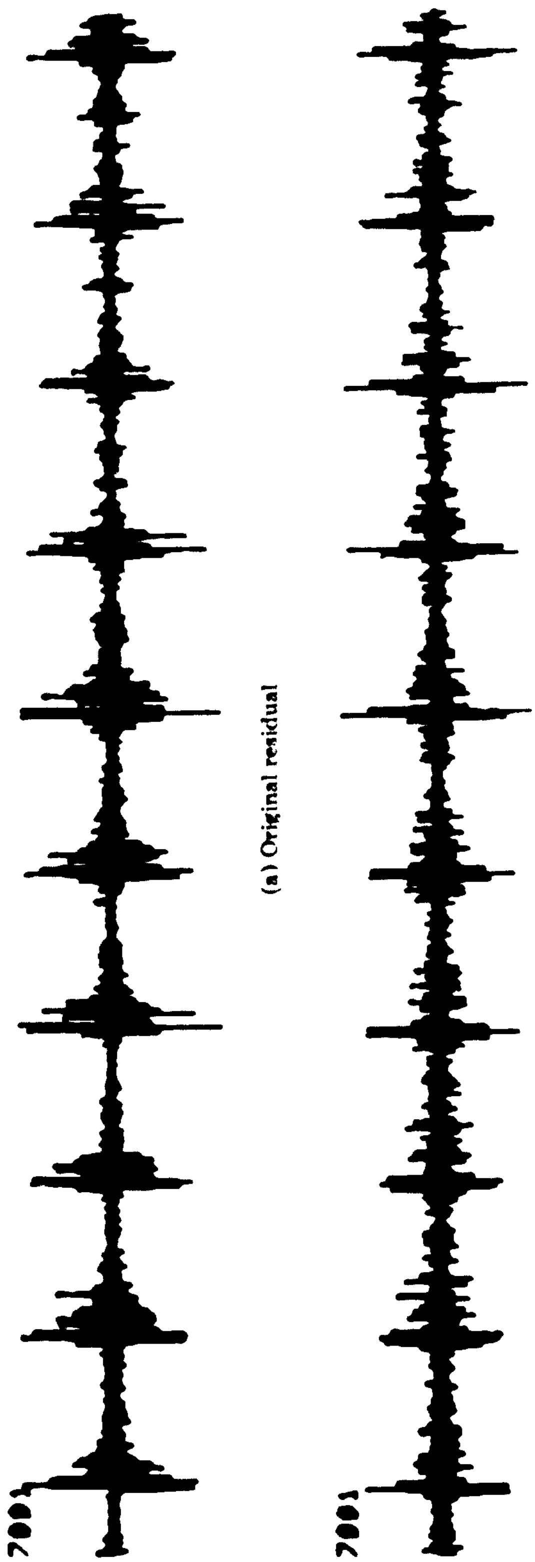
Speech samples synthesized at 9.6 kb/s will be presented in three ways to demonstrate their characteristics:

- Spectrographic analysis,
- Intelligibility testing, and
- Audio demonstration.

Speech samples synthesized at 2.4 kb/s are excluded from this demonstration, because they have been extensively tested by the DOD Narrowband Digital Voice Processor Consortium in 1975 and also by NRL in 1977 (Appendix B).

The MRP has a 9.6-kb/s mode because the quality of 2.4-kb/s speech may not satisfy some communication users. A lack of naturalness is a major problem, but a lack of robustness can also be a problem under conditions such as the following:

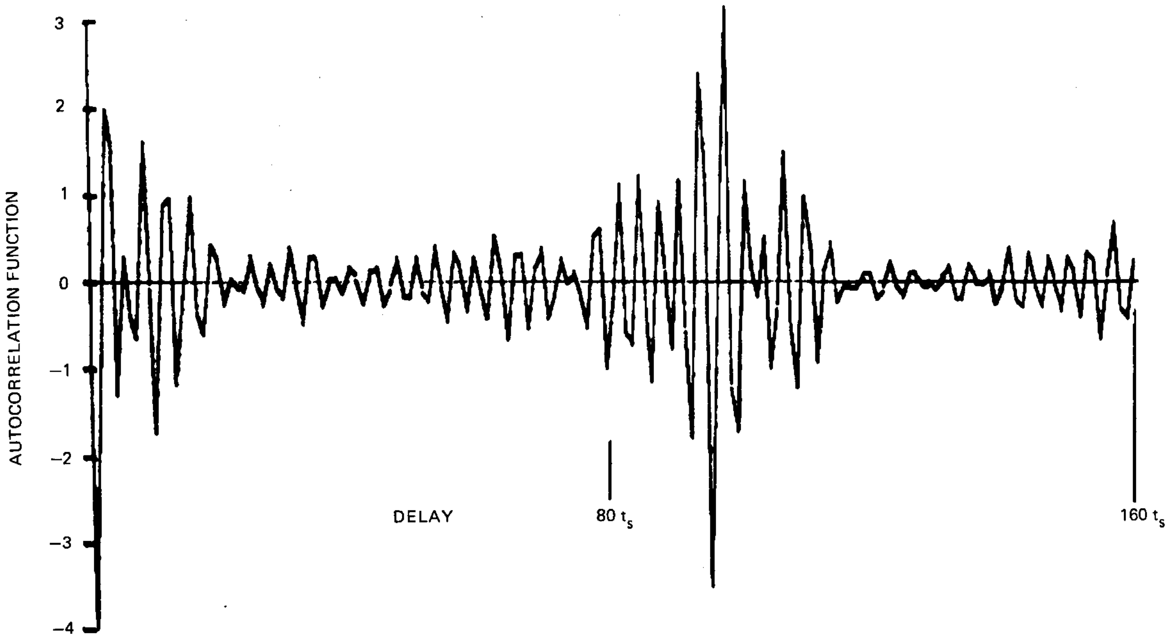
- Background acoustic noise interference,
- A poor microphone with ripples in the frequency response which generates amplitude distortions,
- Users who do not articulate words or who talk too fast.



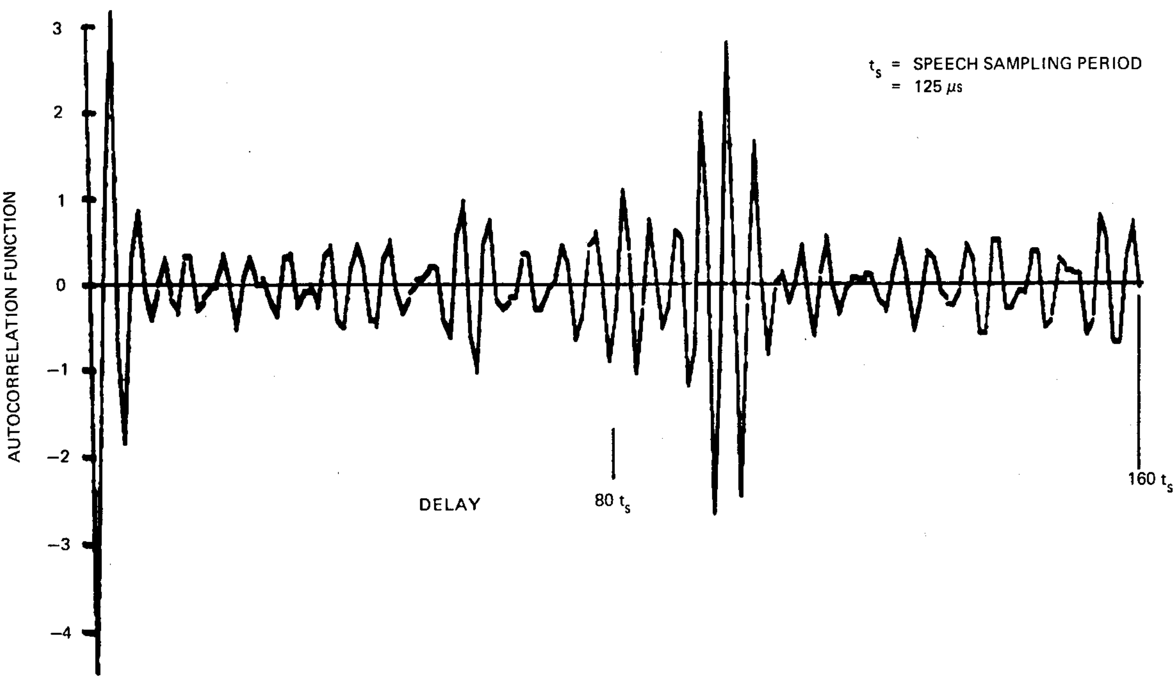
(a) Original residual

(b) Regenerated residual

Fig. 12 — Time sequences of the original and regenerated residual in the first upper subband



(a) Original residual



(b) Regenerated residual

Fig. 13 — Autocorrelation functions of the original and regenerated excitation signals in the first upper subband (t_s = speech sampling period = 125μs)

Speech material which examplifies these conditions will be used for this demonstration.

Spectogram of 9.6-kb/s Speech

Figure 14 is a spectrogram of the original speech and of 9.6-kb/s speech. The following details are often missing in 2.4-kb/s speech, but they are observable in this figure:

- A sharp noise burst for the consonant /t/ in "weight,"
- The formant transition during an unvoiced segment of /s/ in "has,"
- A distinctive gap between "and" and "volume,"
- Irregular (but natural) pitch excitation at onsets of "air," "has," and "volume."

The presence of these details make 9.6-kb/s speech more natural and responsive to a casual mode of conversation.

Intelligibility Test

Quantitative evaluations of synthesized speech can be made by means of the diagnostic rhyme test (DRT). The DRT word list comprises 448 monosyllable rhyming word pairs in which initial consonants differ by only a single feature. An important objective of the DRT [24] is determining speech perception as influenced by process parameters (the parameter update rate, the number of bits for each parameter, and the choice of parameters). The test provides a measure of intelligibility and allows one to evaluate the discriminability of six distinctive features: voicing, nasality, sustention, sibilation, graveness, and compactness.

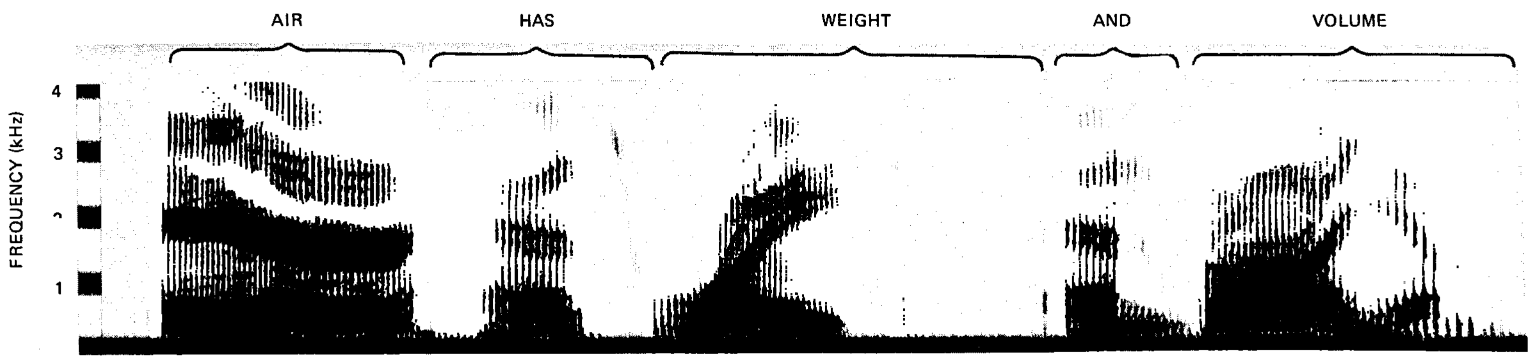
Table 4 lists the DRT scores for the 9.6-kb/s MRP. For comparison the scores for a 16-kb/s CVSD are also listed. The 9.6-kb/s MRP compares favorably with a 16-kb/s CVSD. In particular the improvement in "sustention" is substantial (+17 points).

Audio Demonstration

The attached cassette* contains selected samples of 9.6-kb/s synthesized speech processed by the algorithm discussed in this report, with speech from a 16-kb/s CVSD being recorded first for each sample. The input material was selected to include diversified speech with varying characteristics. The characteristics for each of the selections on the tape are as follows:

- Selection 1: male speech with a high-quality microphone,
- Selection 2: male and female dialog with a high-quality microphone,
- Selection 3: telephone (high-passed) voices with a carbon microphone,

*If a cassette is not included with this report, one can be requested from the Naval Research Laboratory, Code 7526, Washington, DC 20375



(a) Original speech

(b) MRP-processed speech at 9.6 kb/s

Fig. 14 — Spectrum of original speech and of MRP-processed speech at 9.6 kb/s

Table 4 — Diagnostic-Rhyme-Test (DRT) Scores for the 9.6-kb/s MRP and, for Comparison, for a 16-kb/s CVSD

Feature	Perception	DRT Score	
		9.6-kb/s MRP	16-kb/s CVSD
Voicing	Distinguishes /b/ from /p/, /d/ from /t/, /v/ from /f/, etc.	92.6	97.9
Nasality	Distinguishes /n/ from /d/, /m/ from /b/, etc.	94.9	99.2
Sustention	Distinguishes /f/ from /p/, /b/ from /v/, /t/ from /θ/, etc.	91.0	73.7
Sibilation	Distinguishes /s/ from /θ/, /ʃ/ from /d/, etc.	96.5	91.4
Graveness	Distinguishes /p/ from /t/, /b/ from /r/, /m/ from /n/, etc.	88.7	84.6
Compactness	Distinguishes /y/ from /w/, /g/ from /d/, /k/ from /t/, /ʃ/ from /s/, etc.	97.3	95.8
Average		93.5	90.5

- Selection 4: noisy speech taken from an actual tactical helicopter platform,
- Selection 5: fast talking from an AM broadcast with some static noise,
- Selection: casual talking by President Kennedy at a White House press conference (with interfering laughter).

REAL-TIME SIMULATION

Real-time simulations of multirate processing have been performed on two microprogrammable speech processor systems (or terminals) built for the Navy by TRW, Inc. Each system consists of three dual-arithmetic (2-AU) processors, each with its own dedicated data memory and program memory. The three 2-AU processors are interfaced through a common multiplexer for exchange of information. The initial software for the 2-AU terminals was written to simulate a full-duplex very-low-data-rate (VLDR) voice communication system at 600 b/s and a full-duplex narrowband LPC terminal operating at 2.4 kb/s. Software was later generated to simulate an MRP terminal with 4.8 kb/s embedded in a 16-kb/s data stream and 2.4 kb/s embedded in the 4.8-kb/s data stream (the NRL embedded 16/4.8/2.4-kb/s LPC device briefly described in an earlier subsection). Software was also written to demonstrate 2.4 kb/s embedded in a 9.6-kb/s data stream, which are the data rates of the MRP that is the subject of this report.

Processor Description

The three 2-AU units are identical except for different offline and input/output (I/O) capabilities. For instance the first 2-AU can perform an offline normalized division or logarithm in parallel with computations. Each 2-AU processor consists of two identical arithmetic logic units (ALUs) (Fig. 15) and a controller (Fig. 16). The processor has a 150-ns microinstruction cycle time, 2048 16-bit words of data memory, 2048 69-bit words of program memory, and a 32-bit double-precision processing capability. Each 2-AU can perform the following arithmetic functions:

- Sixteen-bit-times-16-bit 2's complement multiplication with a 32-bit product and optional accumulation of products,
- Sixteen-bit-times-16-bit 2's complement multiplication with a 16-bit rounded or truncated product,
- Scaler operations,
- Sixteen- or 32-bit ALU functions with a hard-limit option (no overflow).

The two ALUs (ALU A and ALU B) that are in each 2-AU processor can perform a number of 16- or 32-bit 2's complement arithmetic operations. The operations are the same for either ALU A (Fig. 17) or ALU B in program software. Consider the input to ALU A shown in the block diagram. The input to ALU A is from MUX 3A and registers AAM/AAL. Register AAM stores the 16 most significant bits of a data word, and register AAL stores the 16 least significant bits. Thus AAM/AAL stores a 32-bit word. A description of some of the available ALU A functions together with their appropriate symbols are shown in Table 5.

Software Description

The first 2-AU processor, known as the master unit, has an executive routine that controls sample-by-sample and frame-by-frame processing as well as the other two slave 2-AU processors. The executive, time permitting, also performs confidence checks on some critical constants stored in data memory. As part of the sample-by-sample processing the master 2-AU serves as the LPC analyzer and synthesizer. The frame-by-frame processing done in the master includes encoding and decoding of the LPC parameters and the interpolation of receiver parameters.

The second 2-AU processor, known as the excitation processor, has an executive routine that is slave to the executive in the master 2-AU. As part of the sample-by-sample processing, the excitation processor extracts pitch and voicing information from the residual provided by the master 2-AU.

For the MRP simulation the third 2-AU performs a 128-point FFT and inverse FFT. The input to the FFT is the residual from the first 2-AU. Three FFTs and inverse FFTs are performed every two frames. A block diagram depicting the working arrangement of the three 2-AU processors is shown in Fig. 18.

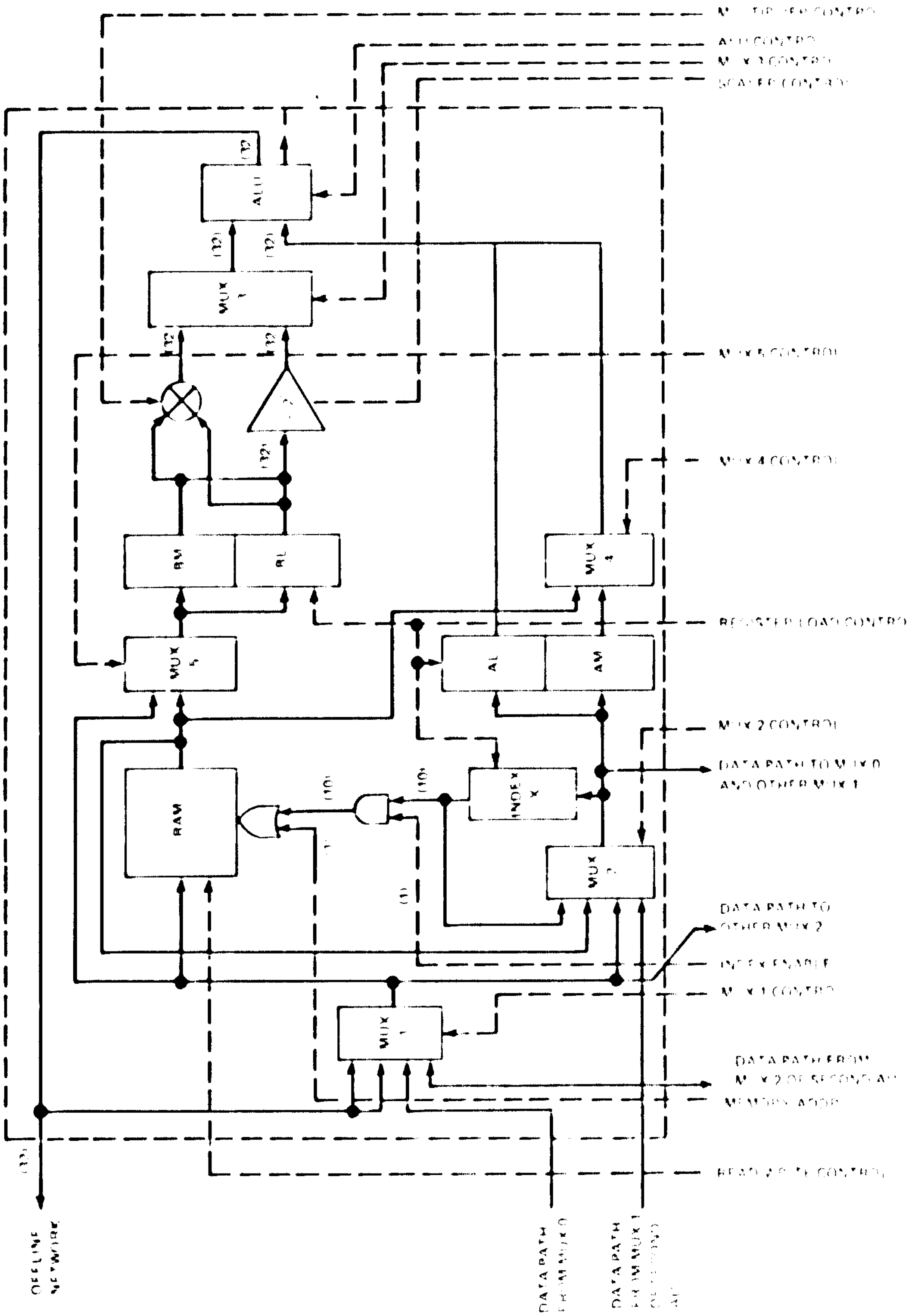
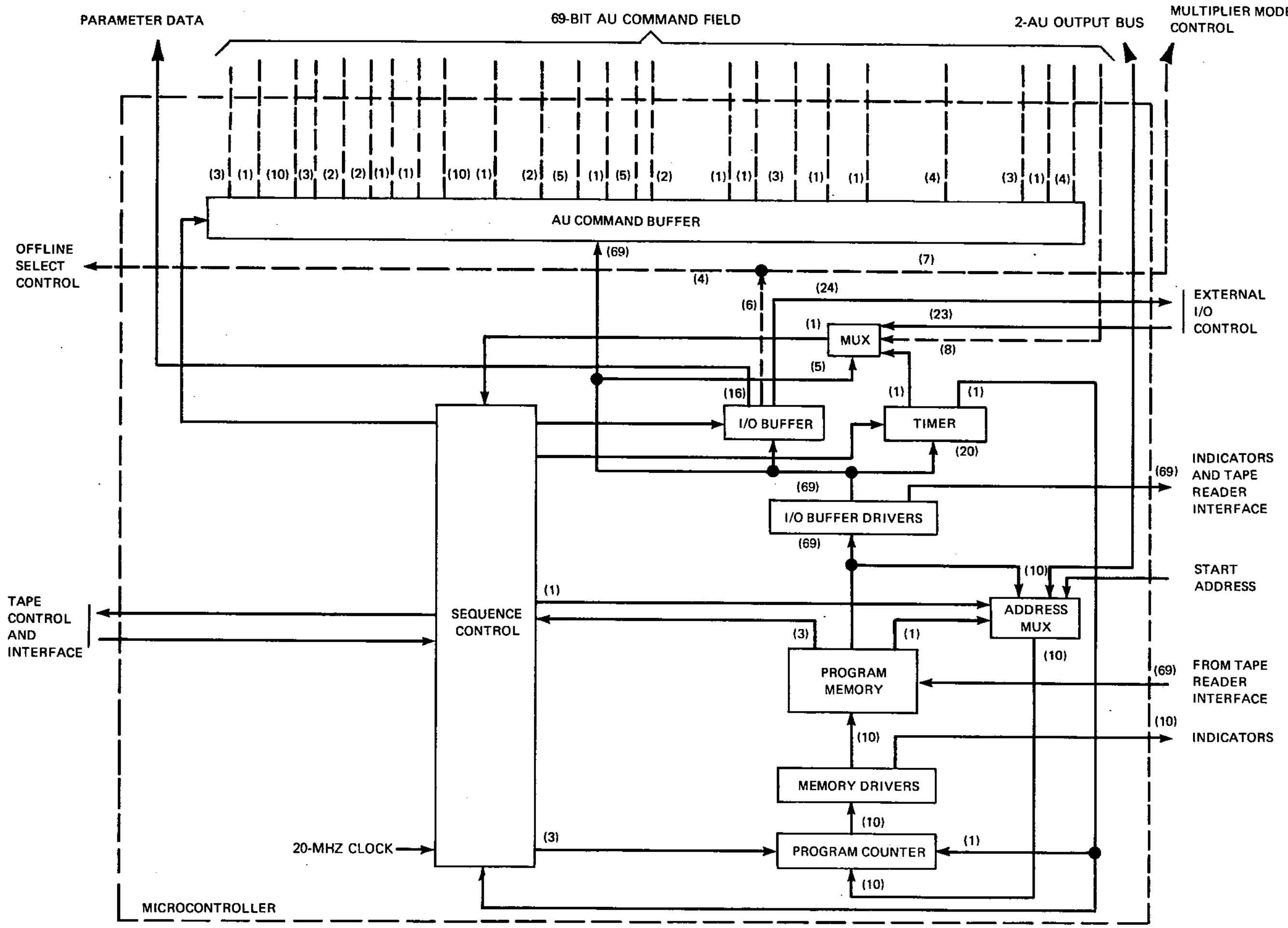


Fig. 15 — Arithmetic unit and memory of the 2-AU processor. If the number of bits associated with a data path is other than 16, the number of bits is labeled in parentheses.



NOTE: ALL DATA PATHS ARE 16 BITS UNLESS OTHERWISE SPECIFIED.
( ) INDICATES NUMBER OF BITS ASSOCIATED WITH A PARTICULAR DATA PATH.

Fig. 16 — Microprocessor controller in the 2-AU processor. Again a number of bits other than 16 associated with a data path is labeled in parentheses.

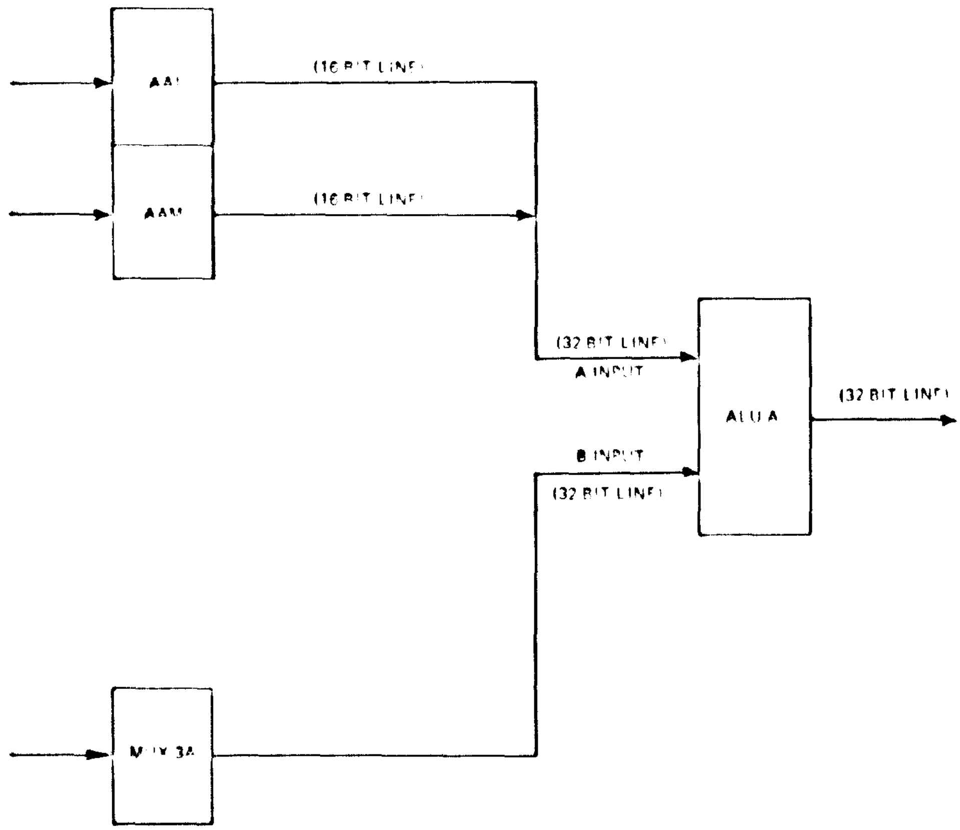


Fig. 17 — ALU A input and output lines

Table 5 — Some of the ALU A Functions

Symbol	ALU A Output	Description
ADD	<AAM,AAL>+<MUX 3A>	Addition
ADL	<AAM,AAL>+<MUX 3A>	Addition with hard limit
SUB	<AAM,AAL>-<MUX 3A>	Subtraction
SBL	<AAM,AAL>-<MUX 3A>	Subtraction with hard limit
ADC	<AAM,AAL>+~<MUX 3A>	Add complement
DEC	<AAM,AAL>-1	Decrement
DBL	<AAM,AAL>*2	Double
MOVA	<AAM,AAL>	Move A
MOVB	<MUX 3A>	Move B
COMA	~<AAM, AAL>	Complement A
COMB	~<MUX 3A>	Complement B

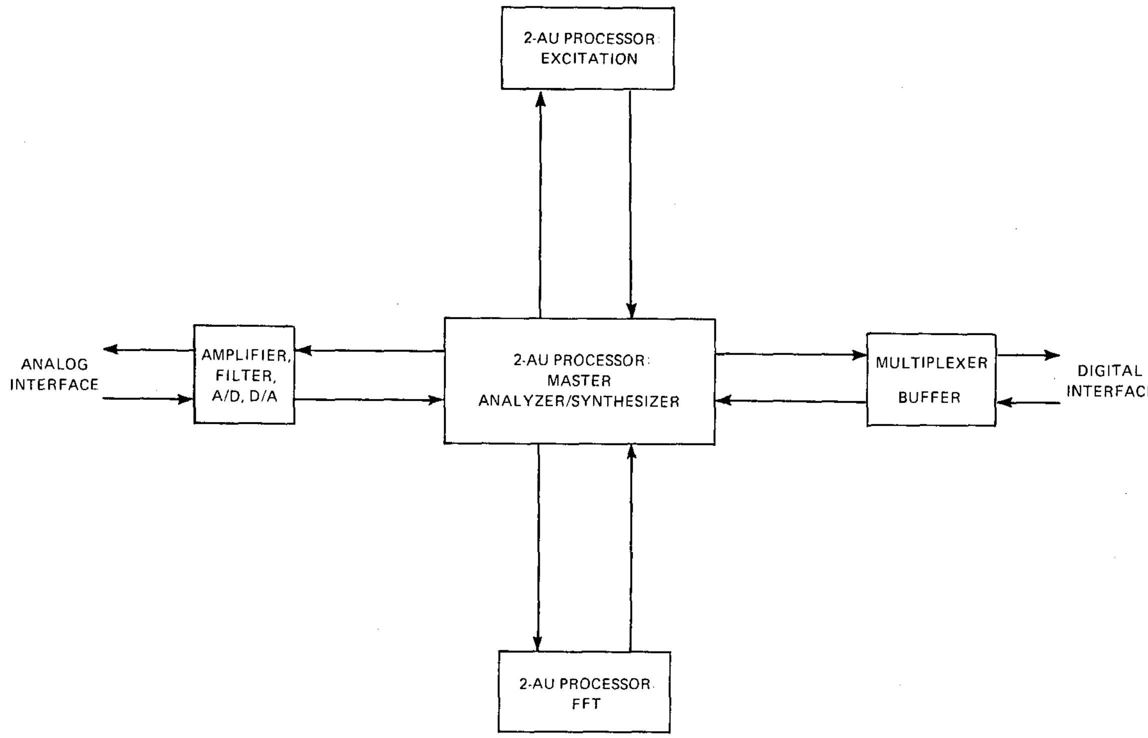


Fig. 18 — 2-AU terminal

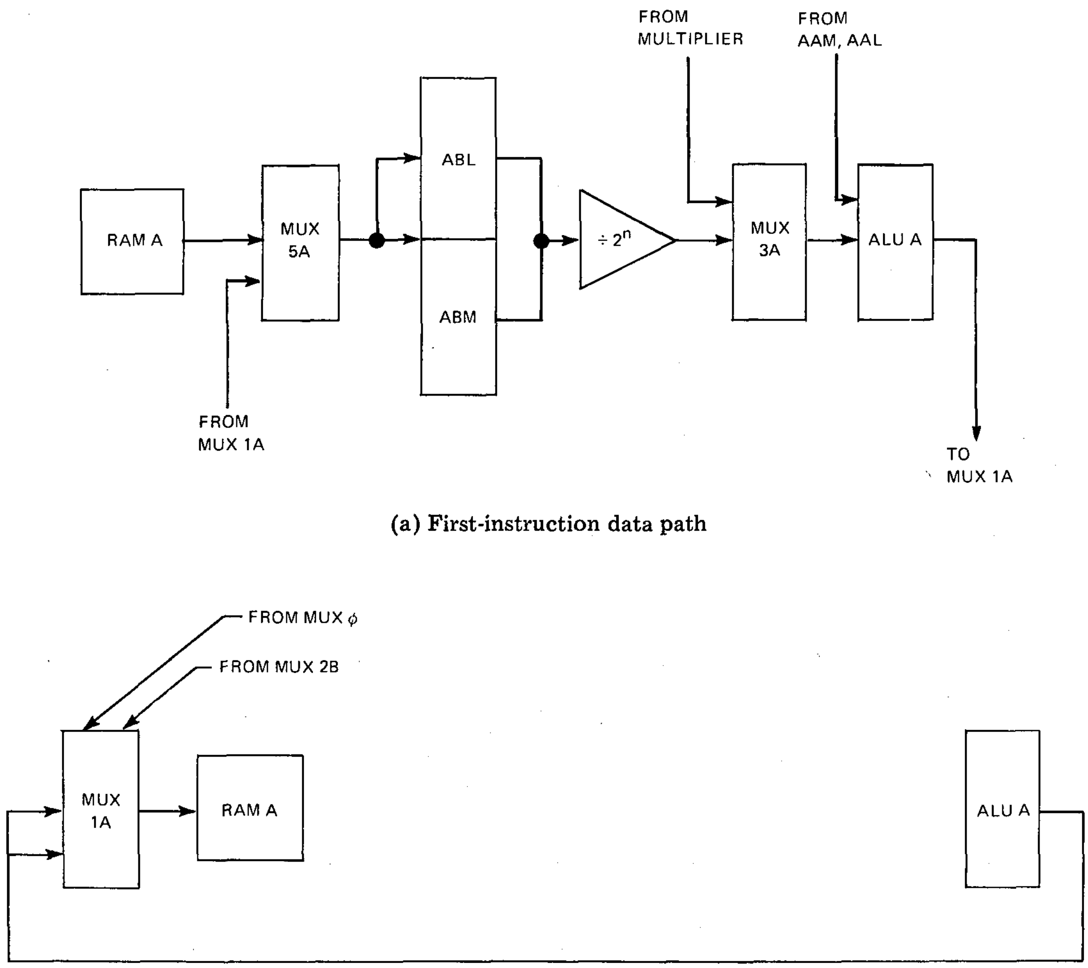
A crossassembler, called VOCASM, runs on a CDC-6600 computer and is used to generate microcode for the 2-AU program memories. A separate set of microinstructions is maintained for each of the three-2-AU processors on the CDC-6600 and can be accessed through time-sharing terminals. After crossassembly in the CDC-6600, the set of 69-bit microcoded instructions are transferred by means of a modem and general-purpose interface to a disk file on a PDP-11/45 minicomputer. These instructions provide an input to a program running on the PDP-11/45 that punches a paper tape to be read into the 2-AU terminals.

VOCASM converts the symbolic representation of the 2-AU multiplexer settings, ALU mode selection, offline control, AU register control, and internal program transfers to an organized sequence of bits for each microcoded instruction. A number of symbolic micro-operation source lines are needed to produce a line of microcode. The load command, LDCMD, collects the current value of each control field and assembles them into a line of code. Most control fields maintain their last value until redefined.

To illustrate the programming of the 2-AU terminals, consider the microprogramming necessary to fetch a value from data memory, scale the value, and store the scaled value in its original location. It takes two 69-bit microcoded instructions to perform this scaling operation. Since the instruction cycle time is 150 ns, the entire operation will take 300 ns. The instructions needed for this example listed in Table 6. The first and second instruction data paths are shown in block-diagram form in Fig. 19.

Table 6 — Instructions Needed to Fetch a Value From Memory, Scale the Value, and Store the Value in Its Original Location

Symbolic Representation	Interpretation of the Symbolic Representation
	First Instruction
RAMR A, VALUE	Read data in location VALUE from RAM A.
MUX5 A, RAMA	Select the RAM A input to MUX 5A.
LOAD ABM	Load the output of MUX 5A into ABM, which is the most significant part of the 32-bit ABM/ABL register.
SCAL A, 1	Set up the scaler to divide the input by 2.
MUX3 A, SCALER	Select the scaler as input to MUX 3A.
ALU A, MOVB	Select the MUX 3A output as the output of ALU A.
LDCMD	Delineates the end of the first instruction.
	Second Instruction
MUX1 A, ALUM	Select the most significant 16 bits coming out of ALU A as the input to MUX 1A.
RAMW A, VALUE	Write the scaled-down data in location VALUE.
LDCMD	Delineates the end of the second instruction.



(a) First-instruction data path

(b) Second-instruction data path

Fig. 19 — Data paths for the instructions used in the example of table 6

The hardware architecture of the 2-AU allows for the efficient computation of the equation $Z = X + WY$, normally a time-consuming calculation in most computers. This type of computation is needed in the analysis and synthesis filter. Assuming 16-bit accuracy, registers AAM, ABL, and ABM can be loaded from three locations in random accces memory (RAM) called X, W, and Y respectively, and the 16-bit resultant Z can be stored in RAM in 600 ns (four instruction cycles).

CONCLUSIONS

Security, connectivity, and survivability are the most desirable operational features for military communications, particularly under stress conditions. Intelligibility, quality, and robustness are also desirable for user acceptability.

The MRP is a first step toward attaining all of these features. The MRP generates two data rates such that the 9.6-kb/s data stream contains the 2.4-kb/s data stream as a subset, making the following possible:

- Encrypted (secure) digitized voice data can be transmitted from one terminal to another, even if data rates are different;
- Network connection can be made even under overloaded channel conditions by appropriate rate change or rerouting.
- Voice communications can continue even under disrupted channel conditions by rerouting through existing government or civilian telephone circuits;
- Persons having no access to wideband channels can use the narrowband mode for voice communication;
- Persons having access to wideband channels or operating in a severe background noise environment can use the wideband mode to transmit high-quality speech.

This report described a method of digitizing speech at 9.6 kb/s with an embedded 2.4 kb/s. The design goal was to make the performance of the 9.6-kb/s mode as good as or better than performances of the present 16-kb/s CVSDs and to make the 2.4-kb/s mode compatible with 2.4-kb/s LPC devices now under DOD development. This goal has been reached.

In the past, the most successful 16-kb/s voice digitizers used the pinciple of waveform preservation, whereas all 2.4-kb/s voice digitizers have used the principle of spectral-envelope preservation. Previously, several 6.4-kb/s or 9.6-kb/s voice digitizers have also been implemented using the principle of waveform preservation, and they were not too successful.

In principle the MRP 9.6-kb/s mode combines both approaches: The speech spectral envelope is preserved by ten prediction coefficients extracted by the same procedures used by a 2.4 kb/s LPC, and the excitation signal (the baseband only) is transmitted in the form of amplitude and phase components. Since phase components carry more information about the excitation, they are quantized with a finer resolution than are the amplitude components.

In conclusion the MRP voice terminal combines narrowband and wideband communication resources. It gives security, operational flexibility (connectivity and survivability), and preferred data rates based on the quality requirement of each communication user.

ACKNOWLEDGMENTS

Bruce Wald, Superintendent of the NRL Communication Sciences Division, and Don I. Himes, Head of the Systems Integration and Instrumentation Branch, showed keen interest in the MRP from the beginning. Their consultation and encouragement have been most welcome.

This work could not have been completed without the help of Jim West and Ruth Phillips of NRL, who maintained our extensive computer facilities in good working order.

The authors also thank Marc Rubenstein of Booz, Allen, and Hamilton, Inc., who performed the various tests required for this report.

Finally, the authors gratefully acknowledge the support of Irv Smietan of the Naval Electronic Systems Command, John Boyd of Naval Communication Architect Group, and Dr. Robert Conley of CNO, who supported and promoted this R&D effort.

REFERENCES

1. T.P. Barnwell and R.W. Schafer, "Improving the Performance of LPC-CVSD Tandem Connections by Phase Modifications," pp 433-436 in Proceeding of 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing.
2. R.S. Cheung, "Application of CVSD with Delayed Decision to Narrowband/Wideband Tandem," pp 437-439 in Proceedings of 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing.
3. L.R. Rabiner, M.R. Sambur, R.E. Crochiere, and D.J. Goodman, "Analysis of the Performance of a Tandem Connection of a 2.4-kb/s LPC Vocoder to a 16-kb/s CVSD Coder" (Abstract), J. Acoustic Soc. Am. **61** (Suppl. 1), S89 (Spring 1977)
4. L.R. Rabiner, M.R. Sambur, R.E. Crochiere, and D.J. Goodman, "Analysis of the Performance of a Tandem Connection of a 16-kb/s CVSD Coder to a 2.4-kb/s LPC Vocoder" (Abstract), J. Acoust. Soc. Am. **61** (Suppl. 1), S89 (Spring 1977).
5. F. Itakura and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," Electron. Commun. Jap. **53A** (No. 1), 36-43 (1970).
6. J.D. Markel, "Digital Inverse Filtering—A New Tool for Formant Trajectory Estimation," IEEE Trans. Audio Electroacoust. **AU-20** (No. 2), 129-137 (1972).
7. H. Dudley, "Remaking Speech," J. Acoust. Soc. Am. **11**, 165 (1939).
8. H. Dudley, "Signal Transmission," U.S. Patent 2,151,091, Mar. 12, 1939.
9. B.S. Atal and M.R. Schroeder, "Adaptive Predictive Coding for Speech Signals," Bell System Tech. 48, 1973-1986 (Oct. 1970).
10. G.S. Kang, "Application of Linear Prediction Encoding to a Narrowband Voice Digitizer," NRL Report 7774, Oct. 31, 1974.

UNCLASSIFIED

11. M. Berouti and J. Makhoul, "High Quality Adaptive Predictive Coding of Speech," pp 303-306 in Proceedings of 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, Apr. 10-12.

12. B.S. Atal and M.R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," pp 573-576 in Proceedings of 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 573-576, Tulsa, Apr. 10-12.

13. J.G. Dunn, "An Experimental 9600-bit/s Voice Digitizer Employing Adaptive Prediction," IEEE Trans. of Comm Tech. *COM-19* (No. 6) 1021-1032 (1971).

14. D.T. Magill, C.K. Un, and S.E. Cannon, "Speech Digitization Study," final report prepared for Defense Communication Agency by Stanford Research Institute, 1975.

15. D. Esteban, C. Galand, D. Mauduit, and J. Menez, "9.6/7.2 kb/s Voice Excited Predictive Coder (VEPC)," pp 307-311 in Proceedings of 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, Oklahoma, Apr. 10-12.

16. J.R. Welsh and C.F. Teacher, "A Split-Band Predictive Coding System at 16 kb/s," pp 241-243 in Proceedings of 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Apr. 12-14.

17. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoust., Speech, Signal Processing *ASSP-22*, 353-362 (Oct. 1974).

18. A.J. Goldberg et al., "Kalman Predictive Encoder," final report prepared for Defense Communication Agency, 1975.

19. A.J. Goldberg and A. Arcese, "A Quantitative Comparison of Residual Encoding Techniques," pp 28A-28F in Proceedings of EASCON 1975.

20. A.J. Goldberg et al., "Adaptive Multilevel 16-kb/s Speech Coder," final report prepared for Defense Communication Agency, 1976.

21. A.J. Goldberg, R.L. Freudberg, and R.S. Cheung, "High Quality 16-kb/s Voice Transmission," pp 244-246 in Proceedings of 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Apr. 12-14.

22. A. Goldberg, "2400/16,000-b/s Multirate Voice Processor," pp 299-302 in Proceedings of 1978 IEEE International Conference on Acoustics, Speech and Signal Processing, Tulsa, Apr. 10-12.

23. J.L. Flanagan and R.M. Golden, "Phase Vocoder," Bell System Tech. J. 45, 1493-1509 (1966).

24. W.D. Voiers, A.D. Sharpley, and C.J. Hemsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," AFCRL-72-0694, prepared by TRACOR, Inc., for the Air Force Cambridge Research Laboratories, Air Force Systems Command, 1973.

APPENDIX A

GLOSSARY OF ACRONYMS

ANDVT — Advanced Narrowband Digital Voice Terminal. This is a digitized narrowband voice terminal applicable to Defense Communication System (DCS) and military requirements that will fulfill the needs of a single interoperable system which is global. The terminal uses linear prediction coding as the voice-processing algorithm.

AUTODIN — Automatic Digital Network. This is the principal, long-haul, DOD digital network for transmitting administrative, logistical, special category intelligence, and command and control record/data traffic on an automatic switched basis. It provides rapid digital data communications among organizational elements. No special survivability provisions are included in AUTODIN other than the redundancy of circuit paths which are also inherent in AUTOVON.

AUTODIN II — Automatic Digital Network II. The first phase of AUTODIN II is within its early stage of initiation. It will provide a standalone capability within the Continental United States (CONUS) for interactive data service, man-machine and machine-machine message transfer. It is contemplated that this network will expand to meet an overseas demand and will integrate the functions of AUTODIN I.

AUTOSECVOCOM I — Automatic Secure Voice Communications I. This network provides rapid secure communications among organization elements. No special provisions are included in the AUTOSEVOCOM network for survivability other than the redundancy of circuit paths inherent in Defense Communication System (DCS) facilities.

AUTOSECVOCOM II — Automatic Secure Voice Communications II. The inability of AUTOSEVCOM I to meet DOD needs for secure voice communications resulted in the design of a new secure voice network. Full-scale development was approved in May 1976. This network is being developed as a digital overlay onto the AUTOVON network. Subscribers' terminals will digitally encode and encrypt speech using a 16-kb/s data rate. Interconnection through the AUTOVON network will not be possible. Although a substantial number of secure voice subscribers will be added to the voice network, about 98% of the current AUTOVON subscriber population will still be without end-to-end security protection.

AUTOVON — Automatic Voice Network. This is a worldwide DOD switched telephone service via common user trunks. It provides voice communications among organizational elements. The system has a goal to complete at least 95 of every 100 call attempts during the peak busy hour. It provides the trunks for AUTOSEVOCOM communications. The survivability feature of AUTOVON is its grid routing of interswitch trunks.

CVSD — continuously-variable-slope delta modulation (data rate from 9.6 kb/s to 32 kb/s). This is the voice processing algorithm implemented in the digital-secure-voice-terminal (DSVT) and Vinson families of digital wideband secure voice equipments. In this implementation the adaptive delta-modulation technique is used, with the quantizer step size being varied over a continuous range of values.

DSARC III — Defense System Acquisition Review Council. This council determines whether the system under review requires further development, or if initial production should be started, or if the program should be eliminated.

ESVN — Executive Secure Voice Network. This is a planned secure network overlay on the commercial telephone system used by U.S. civil agencies and certain high-ranking executives of the government. It uses subscriber telephone units (STU-II) and employs 2.4-kb/s LPC-10 or 9.6-kb/s adaptive predictive coding (APC) as the voice-processing algorithm.

APPENDIX B

INTELLIGIBILITY OF NARROWBAND AND WIDEBAND VOICE PROCESSORS: BACK-TO-BACK AND TANDEM OPERATION

This appendix lists intelligibility test scores of narrowband and wideband voice processors in back-to-back and tandem operation. The tests were conducted at NRL during April and May 1978. They include male and female voices from ALTEC high-fidelity, carbon, and/or noise-cancellation (NC) microphones. The effects of additive noise were also examined using airborne-command-post (ABCP), shipboard (SB), office (Off.), and helicopter (Hel.) noise.

TEST SCORES FOR A NARROWBAND VOICE DIGITIZER

The narrowband voice digitizer tested was the latest, real-time, linear predictive encoder operating at 2.4 kb/s in a back-to-back mode. The results are tabulated in Table B1.

Although there were a limited number of test samples, the following observations are worth noting:

- In a noise-free back-to-back mode with a dynamic microphone the score for all attributes was distributed over a narrow range. Previously the majority of 2.4-kb/s linear-predictive-coding (LPC) devices gave low scores for one or two of the tested attributes (such as 50 for "sustention"). A narrower distribution of the individual attribute scores with a higher total score (88.6 for six male speakers and 85.9 for three female speakers) represents a subtle improvement in the latest LPC device.

- The DRT scores were lower if a carbon microphone was used (—2.8 for six male speakers and —5.2 for three female speakers) rather than a high-quality high-fidelity microphone. The microphone degradation occurs in "voicing" and "sibilation" with male speakers and in "sustention" and "compactness" with female speakers.

- In the presence of airborne-command-post noise or shipboard noise, the use of a noise-cancellation microphone was better (+5.7 and +7.3, respectively) than the use of a dynamic microphone. On the other hand, in the office environment, the use of a dynamic nonnoise-canceling microphone is preferred (+6.2 for three male speakers and +10.0 for one female speaker). If a microphone is optimized for the particular operational environment, the DRT scores would be in the middle to lower 80s — acceptable levels with respect to those set by the Narrowband Digital Voice Processor Consortium.

- The DRT score with helicopter noise was still lower than desired (54.9). A noise-suppressing preprocessor is definitely needed for this environment.

Table B1 — Diagnostic-Rhyme-Test (DRT) Scores for 2.4-kb/s Linear-Predictive-Coding (LPC(2.4)) Devices

No. of Male and Female Speakers	Microphone	Noise	DRT Scores for LPC(2.4)						
			Voicing	Nasality	Sustention	Sibilation	Graveness	Compactness	Total
6M	ALTEC	–	87.1	91.7	82.3	96.1	77.6	91.0	88.6
3F	ALTEC	–	93.5	93.5	81.2	79.9	74.2	93.2	85.9
6M	Carbon	–	79.3	93.7	80.3	90.5	78.0	93.0	85.8
3F	Carbon	–	87.8	93.7	71.1	82.8	61.3	81.6	80.7
3M	ALTEC	ABCP	65.4	66.7	67.7	79.9	77.6	93.7	75.2
3M	NC	ABCP	77.9	78.9	73.2	90.1	77.1	91.7	81.5
3M	ALTEC	SB	79.7	81.2	62.2	85.4	73.2	83.9	77.6
3M	NC	SB	83.1	89.1	78.9	89.6	76.0	92.7	84.9
3M	M78	Hel	63.3	56.8	40.6	54.2	51.3	63.5	54.9
3M	ALTEC	Off	85.4	90.4	75.8	92.4	81.3	91.0	86.5
1F	ALTEC	Off	98.4	93.0	76.6	88.3	76.6	89.8	87.1
3M	NC	Off	71.1	83.6	79.9	88.0	65.6	93.7	80.3
1F	NC	Off	78.1	96.9	61.7	83.6	60.9	80.5	77.0

TEST SCORES FOR WIDEBAND VOICE DIGITIZERS

Three CVSDs were selected for the wideband voice digitizer: DSVT(32), DSVT(16), and Vinson(16), with the numeral denoting the data rate in kb/s. The DRT scores, summarized in Table B2, indicate the following:

- The weakest attribute of wideband voice digitizers is "graveness," which scores several points below the other attributes. The presence of quantization noise seems to lower the score the "gravness." This is also exhibited in an NRL experimental residual-excited 16-kb/s LPC device.

- The scores for "sustention" were also low for female voices, particularly at 16 kb/s. Since "sustention" is related to voice onset rise characteristics, slope overloading of the CVSD could be the source for this degradation.

- The use of a carbon microphone does not limit wideband performance, contrary to the narrowband cases. In fact the scores for "sibilation" for the Vinson(16) is higher with a carbon microphone than with a dynamic microphone. This is due to the inherent preemphasis characteristics of a carbon microphone. The score for "nasality" is surprisingly high despite a lack of low-frequency response typical of a carbon microphone.

- In the presence of airborne-command-post noise the Vinson(16) behaved as poorly as an LPC(2.4) device — scoring in the lower 80s.

Table B2 — DRT Scores for Wideband Voice Digitizers

No. of Male and Female Speakers	Micro-phone	Noise	DRT Scores						
			Voicing	Nasality	Sustention	Sibilation	Graveness	Com-pactness	Total
DSVT(32)									
3M	ALTEC	—	99.0	99.7	96.9	99.5	88.0	99.0	97.0
3M	Carbon	—	99.5	99.5	94.0	98.7	89.8	99.5	96.8
1F	ALTEC	—	98.4	97.7	93.0	96.9	89.8	100.0	96.0
1F	Carbon	—	98.4	100.0	82.0	97.7	85.9	99.2	93.9
DSVT(16)									
3M	ALTEC	—	99.7	99.2	94.3	97.4	85.4	99.2	95.9
3M	Carbon	—	98.4	99.0	94.5	97.7	87.2	99.5	96.1
1F	ALTEC	—	98.4	100.0	86.7	95.3	83.6	100.0	94.0
1F	Carbon	—	99.2	99.2	82.8	95.3	85.2	100.0	93.6
Vinson(16)									
3M	ALTEC	—	97.9	99.2	73.7	91.4	84.6	95.8	90.5
3M	Carbon	—	94.0	97.9	79.4	97.4	85.7	96.9	91.9
1F	ALTEC	—	94.5	99.2	64.1	71.9	68.8	98.4	82.8
1F	Carbon	—	96.1	98.4	78.9	86.7	70.3	97.7	88.0
3M	ALTEC	ABCP	87.2	84.9	70.6	76.6	76.6	94.3	81.7
3M	NC	ABCP	94.0	92.2	72.1	71.9	75.0	89.8	82.5

- The use of a noise-cancellation microphone provided marked improvement in the presence of acoustic background noise for the narrowband digitizer (LPC device), whereas improvement in the wideband case was limited.

TEST SCORES FOR LPC(2.4) DEVICES TANDEMED INTO WIDEBAND VOICE DIGITIZERS

The DRT scores for tandem LPC(2.4)-to-wideband links are tabulated in Table B3. As expected, a DSVT(32) is almost transparent in a tandem link. The most encouraging result was the tandem performance of an LPC(2.4) device. It was only slightly degraded from a back-to-back configuration when tandemed with a Vinson(16), if the speech originated from a noise-free environment. For narrowband-to-wideband tandem links, the results indicate the following:

- In the absence of acoustic noise, the DRT scores are in the 80s, which are acceptable.

- The deficiencies of a wideband voice digitizer (for example, low scores for "graveness") are more exaggerated in the tandem links with an LPC(2.4) device.

Table B3 — DRT Scores for LPC(2.4) Devices Tandemed into Wideband Voice Digitizers

No. of Male and Female Speakers	Microphone	Noise	DRT Scores						
			Voicing	Nasality	Sustention	Sibilation	Graveness	Compactness	Total
LPC(2.4) Tandemed into DSVT(32)									
3M	ALTEC	—	95.6	97.1	87.5	95.1	80.2	93.0	91.4
3M	Carbon	—	89.8	91.8	67.0	90.4	78.4	95.8	89.4
1F	ALTEC	—	97.7	97.7	82.0	85.9	67.2	88.3	86.5
1F	Carbon	—	95.3	98.4	77.3	91.4	63.3	93.0	86.5
LPC(2.4) Tandemed into DSVT(16)									
3M	ALTEC	—	91.5	96.4	82.0	91.7	69.8	93.7	88.0
3M	Carbon	—	85.9	95.3	80.7	82.0	75.0	93.0	85.3
1F	ALTEC	—	91.5	97.7	73.4	82.8	59.4	82.8	81.8
1F	Carbon	—	91.4	96.1	62.5	76.6	56.3	85.2	78.0
LPC(2.4) Tandemed into Vinson(16)									
3M	ALTEC	—	91.8	96.6	85.7	87.5	78.1	90.0	88.9
3M	Carbon	—	88.5	91.9	79.9	86.5	77.6	91.1	85.9
1F	ALTEC	—	92.2	96.9	81.3	78.9	64.8	88.3	83.7
1F	Carbon	—	88.3	99.2	65.6	92.2	61.7	88.3	82.6
3M	ALTEC	ABCP	57.8	65.6	55.5	62.5	70.8	81.6	66.1
3M	NC	ABCP	81.8	76.3	69.3	74.5	71.6	86.2	76.6

- In the presence of airborne-command-post noise the use of a noise-cancellation microphone produces better DRT results, similar to the results for an LPC(2.4) device in the back-to-back configuration.

TEST SCORES FOR WIDEBAND VOICE DIGITIZERS TANDEMED INTO LPC(2.4) DEVICES

The DRT scores for wideband-to-LPC(2.4) links are listed in Table B4. The DRT scores are quite similar to those of the reverse link (Table B3) except in the case of tandeming with a Vinson(16). There was some difficulty in setting up the tandem link with this particular Vinson unit. Thus the results indicated in Table B4 may not be representative of the Vinson family of equipment. However, since the DSVT(16) and Vinson(16) are both of the CVSD family, the DSVT(16) tandem performance with an LPC(2.4) device should provide the necessary data to determine future Vinson operation in this tandem configuration.

Table B-4 — DRT Scores for Wideband Voice Digitizers Tandemed into LPC(2.4) Devices

No. of Male and Female Speakers	Micro-phone	Noise	DRT Scores						
			Voicing	Nasality	Sustention	Sibilation	Graveness	Com-pactness	Total
DSVT(32) Tandemed into LPC(2.4)									
3M	ALTEC	—	97.7	97.1	87.8	93.7	75.3	92.2	90.6
3M	Carbon	—	90.9	94.8	85.4	93.0	82.6	96.4	90.5
1F	ALTEC	—	97.7	97.7	80.5	93.0	55.5	100.0	97.4
1F	Carbon	—	85.5	95.3	71.1	80.5	67.2	96.1	82.7
DSVT(16) Tandemed into LPC(2.4)									
3M	ALTEC	—	96.1	97.4	77.3	87.5	70.8	93.5	87.1
3M	Carbon	—	90.6	95.6	75.8	78.9	76.6	91.4	84.8
1F	ALTEC	—	96.9	93.8	79.7	84.4	51.6	82.8	81.5
1F	Carbon	—	84.4	96.1	58.6	82.8	63.3	80.5	77.6
Vinson(16) Tandemed into LPC(2.4)									
3M	ALTEC	—	93.2	96.1	77.1	86.2	71.6	84.9	84.9
3M	Carbon	—	75.3	93.5	70.3	84.6	66.9	86.2	79.5
1F	ALTEC	—	71.9	87.6	52.3	68.0	54.7	93.0	71.1
1F	Carbon	—	81.3	97.7	59.4	80.5	39.1	87.5	74.2
3M	ALTEC	ABCP	54.4	57.8	45.1	62.2	70.3	84.4	62.4
3M	NC	ABCP	76.6	75.5	54.9	63.8	55.2	73.7	66.6

TEST SCORES FOR LPC(2.4) DEVICES TANDEMED INTO LPC(2.4) DEVICES

The DRT scores for self-tandeming LPC(2.4) devices are listed in Table B5. The performance for an LPC(2.4) device tandemed into an LPC(2.4) device (Table B5) falls slightly (a few points) below that of an LPC(2.4) device in a back-to-back mode (Table B1).

Table B5 — DRT Scores for LPC(2.4) Devices Tandemed into LPC(2.4) Devices

No. of Male and Female Speakers	Micro-phone	Noise	DRT Scores						
			Voicing	Nasality	Sustention	Sibilation	Graveness	Com-pactness	Total
3M	ALTEC	—	93.5	97.1	79.9	93.2	72.9	96.4	88.8
3M	Carbon	—	84.9	90.4	76.6	86.7	75.3	93.0	84.5
1F	ALTEC	—	93.8	85.9	65.6	89.8	67.2	96.9	83.2
1F	Carbon	—	80.5	89.8	59.4	75.0	62.5	80.5	74.6

UNCLASSIFIED

APPENDIX C

ANDVT PROCESSOR ALGORITHM

On the following pages of this appendix the authors of the report describe the latest 2.4-kbs LPC algorithm. These pages are duplicated in the form used as section 10 of the ANDVT specifications. To achieve system interoperability between the ANDVT and the ESVN, this LPC algorithm was compiled in close coordination with the National Security Agency, who is responsible for implementing the ESVN. The authors express their apprecia-tion to Thomas Tremain and Jesse Fussell for their invaluable assistance.

10.1 LPC Processor

The voice digitizer shall be based on the analysis/synthesis principle of speech encod-ing. In particular, the theory of linear prediction shall be employed as the analysis technique. The speech shall be synthesized by a linear time-varying filter with an excitation signal con-trolled by pitch and voicing decision information. The voice digitizer shall have the follow-ing overall characteristics:

Sampling Rate:	8 kHz ± .1%
Predictor Order:	10
Data Rate:	2400 b/s
Frame Length:	22.5 ms (54 bits per frame)
Excitation Analysis:	
Pitch:	Average magnitude difference function (AMDF) with Dynamic Pitch Tracker, 51.3-400 Hz, semilogarithmic coding (60 values)
Voicing:	Based on lowband speech energy, background noise energy, zero crossing, and first two reflection coefficients
Amplitude:	Speech root-mean-square (rms) value, semilogarithmic coding (32 values)
Filter Analysis:	Semi pitch synchronous
Preemphasis:	Filter transfer function: $1-0.9375z^{-1}$
Matrix Load:	Covariance method
Matrix Inversion:	Cholesky decomposition
Coding:	Log area ratio for first two coefficients and linear for the remainder

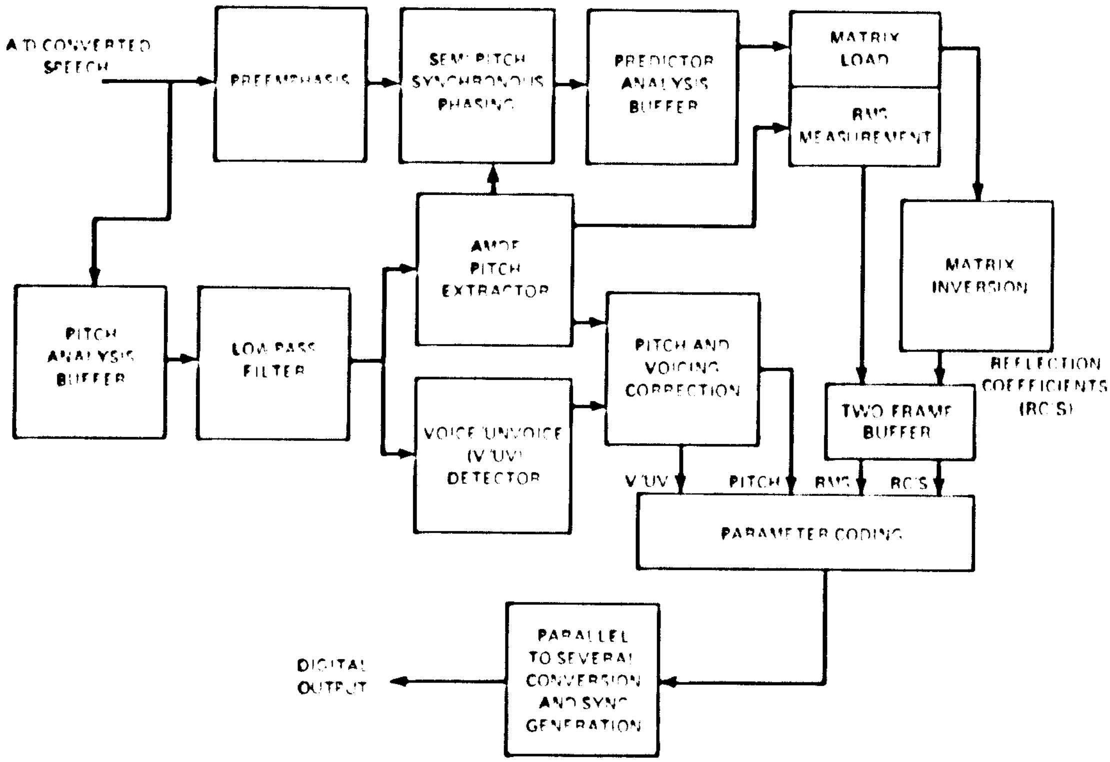


Fig 10 1 – LPC Transmitter

10.2.1 Excitation Analysis

The excitation signal, shall be either a broadband signal that repeats quasi-periodically at the rate of the pitch frequency for voiced sounds or random noise for unvoiced sounds. The excitation analysis shall include the following five computational and logic processes: (1) signal conditioning, (2) average magnitude difference function, (3) pitch tracking, (4) voicing decision, and (5) pitch/voicing decision coding.

10.2.1.1 Signal Conditioning

The speech signal shall be low pass filtered at approximately 1000 Hz to suppress higher formant frequency interferences for the pitch estimation process. It shall be either a 3-4-5 sum filter with the following transfer function:

$$H(z) = \frac{1}{60} \prod_{j=1}^{3} \left(1 + \sum_{i=1}^{5-j} z^{-i}\right)$$

or a fourth order Butterworth filter expressed as

$$H(z) = \frac{1}{1 - 2.455z^{-1} + 2.455z^{-2} - 1.152z^{-3} + 0.210z^{-4}} \cdot$$

10.2.1.2 Average Magnitude Difference Function

The average magnitude difference function (AMDF) shall be derived from a low-pass filtered speech signal. It shall be computed for 60 values of delay, and they shall be transferred to the pitch tracker to generate the pitch trajectory. The definition of the AMDF shall be

$$\text{AMDF}(i) = \sum_{n=n_1}^{n_2} |s(n) - s(n+\tau_i)| \quad i = 1,2,\ldots,60$$

where $s(n)$ is the n^{th} low-passed sample

τ_i is the i^{th} delay equal to the i^{th} pitch period.

The pitch period shall be quantized semilogarithmically into 60 values from 51 Hz to 400 Hz, as listed in Table 10-I.

The AMDF shall be computed with respect to the following rules:

1. The computations shall be in the form of block analysis.
2. The summation shall be computed on every 4^{th} term of the absolute different (i.e., $n = n_1, n_1 + 4, n_1 + 8, \ldots$), in order to reduce computations.
3. The total number of summations shall be fixed to 32. Thus, the lower and the upper limits of the summation shall be related by

$$n_2 = n_1 + 127.$$

Table 10-I. Pitch Values

Index	Pitch Period*	Pitch Frequency (Hz)	Index	Pitch Period	Pitch Frequency (Hz)
1	20	400.000	31	60	133.333
2	21	380.952	32	62	129.032
3	22	363.636	33	64	125.000
4	23	347.826	34	66	121.212
5	24	333.333	35	68	117.647
6	25	320.000	36	70	114.286
7	26	307.692	37	72	111.111
8	27	296.296	38	74	108.108
9	28	285.714	39	76	105.263
10	29	275.862	40	78	102.564
11	30	266.667	41	80	100.000
12	31	258.065	42	84	95.238
13	32	250.000	43	88	90.909
14	33	242.424	44	92	86.957
15	34	235.294	45	96	83.333
16	35	228.571	46	100	80.000
17	36	222.222	47	104	76.923
18	37	216.216	48	108	74.074
19	38	210.526	49	112	71.429
20	39	205.128	50	116	68.966
21	40	200.000	51	120	66.667
22	42	190.476	52	124	64.516
23	44	181.818	53	128	62.500
24	46	173.913	54	132	60.606
25	48	166.667	55	136	58.824
26	50	160.000	56	140	57.143
27	52	153.846	57	144	55.555
28	54	148.148	58	148	54.054
29	56	142.857	59	152	52.632
30	58	137.931	60	156	51.282

*In multiples of the speech sampling period.

4. The computations shall be centered at the midpoint of the data sequence (the point of minimum variance). The center point (n_c), as determined by the maximum pitch period and the initial summation point of $n_1 = 1$, shall be

$$n_c = \text{INT}\left[\frac{n_1 + n_2}{2} + \frac{\tau_i}{2}\right] + 1$$

$$= \text{INT}\left[\frac{129}{2} + \frac{156}{2}\right] + 1$$

$$= 143.$$

Accordingly, the initial point of summation (n_1) for a given delay (τ_i) shall be

$$n_1 = \text{INT}\left[\frac{157 - \tau_i}{2}\right] + 1.$$

10.2.1.3 Pitch Tracker

The pitch tracker shall determine the pitch trajectory based on the following information:

1. AMDF indicating the most probable pitch periods
2. History of past pitch periods.

The pitch trajectory shall be chosen to provide smooth pitch transitions for a given AMDF. It shall be derived through a joint consideration of the two cost functions: the intrinsic cost of choosing a pitch period τ_i at the j^{th} frame (i.e., AMDF) and the transition cost of choosing τ_i at the $(j-1)^{th}$ frame to τ_k at the j^{th} frame (i.e., previous pitch periods).

The intrinsic cost, a measure of the "unlikelihood" of the pitch period τ_i at the j^{th} frame, shall be the AMDF function:

$$A(j,i) = \text{AMDF}(j,i).$$

The transition cost, a measure of perceptual change of pitch, shall be proportional to the logarithmic difference between the two pitch periods. Since pitch periods are semilogarithmically quantized, the transition cost shall be proportional to the absolute difference of the pitch indices:

$$f(j,k-i) = \alpha_j |k - i|_j$$

where α_j is a measure of the confidence level of the AMDF and behaves as a pitch tracker time constant (i.e., the smaller the value of α_j the more reliance the pitch tracker puts on the present AMDF values).

The confidence level α_j shall be directly related to the minimum value of the AMDF, and it shall be updated frame by frame based on the following expressions:

$$\alpha_{j+1} = \begin{cases} \alpha_j & \text{for unvoiced frames} \\ \frac{1}{4}\left[\frac{m(j)}{2} + 3\,\alpha_j\right] & \text{for sustained voiced frames} \end{cases}$$

where m(j) is the minimum value of the AMDF at the j^{th} frame

α_j is the confidence level of the j^{th} frame.

The cost of the cheapest sequence of pitch values that end with τ_i at the j^{th} frame (i.e., a winner function denoted by W(j,i)) and the index of the immediately preceding pitch period in the cheapest sequence (a pointer function denoted by P(j,i)) shall be determined recursively by the following expressions:

$$W(1,i) = A(1,i)$$

$$P(1,i) = \text{not needed and undefined}$$

$$W(j+1,i) = A(j,i) + \min_k \{W(j,i) + f(j,k-1)\}$$

$$= \text{AMDF}(j,i) + \min_k \{W(j,i) + \alpha_j\, |k - i|_j\}$$

$$P(j+1,i) = \text{value of k that satisfies the above expression.}$$

The term $\min_k \{W(j,i) + \alpha_j\, |k - i|_j\}$, an intermediate winner function, shall be evaluated by the comparison between the old winner function W(j,i) and $\alpha_j\, |k - i|_j$, which is a pair of rays of slopes $\pm\,\alpha_j$ emanating from a point at height W(j,i). The smaller value between these two quantities shall be added to the current AMDF, A(j+1,i). In order to guard against overflow, a bias (the minimum value of the winner function) shall be subtracted from the winner function. The value of the index corresponding to the pitch period for the source of the ray shall be stored in the pointer array P(j+1,i).

The best choice of past pitch values shall be obtained by backward tracing of pointer functions, starting with the location of the minimum point (viz., i=I) in the present winner function. This shall be used as an index for the location of the current point function P(j,i=I). The content of this pointer function shall be used as the index to the location of the previous pointer function P(j−1,P(j,i=I)), and so on.

Table 10-II presents a FORTRAN program of this particular pitch tracker algorithm. It shall be referened during the development of the pitch tracker software.

10.2.1.4 Voicing Decision

The voicing decision shall be based on the information derived from the low-pass filtered speech signal. In order to generate the voicing decision, the following quantities shall be required:

1. Low-passed Speech Energy (ISTU): This quantity shall be approximated by the absolute value summation of the dc-removed, low-passed speech signal:

$$\text{ISTU} = \frac{1}{4} \sum_{i=m_1}^{m_2} |s(i) - \overline{s}|$$

where s(i) is the i^{th} low-passed speech sample

$\frac{1}{4}$ is a scale factor

the summation shall be commuted for every other sample

and

$$\overline{s} = \frac{2}{m_2 - m_1 + 1} \sum_{i=m_1}^{m_2} s(i)$$

in which the midpoint of the summation range (m_1, m_2) shall be at n_c (the pitch computation center, defined in 10.2.1.2) and $m_2 - m_1 + 1$ shall be 180 samples (Fig. 10-2). This summation shall be computed for every other sample. ISTU shall be computed twice per frame.

2. Background Noise Energy (ALO): The background noise energy shall be computed from the low-passed speech energy, updated only when the raw voicing decision is "unvoiced":

$$\text{ALO} \leftarrow \left(\frac{15}{16}\,\text{ALO} + \frac{1}{16}\,\text{ISTU}\right).$$

Table 10-II. FORTRAN Program of Pitch Tracker

```
      SUBROUTINE PTRACK(AMDF,MIN,IV,IPOINT,ITMIDX)
C     PITCH TRACKER USING AMDF VALUES,ITS MINIMUM AND VOICING DECISION
C
      IN[illegible]
      IN[illegible]
C
C     ----- [illegible] (IPTR IS CURRENT FRAME) -----
C
      IPTR=IPOINT+1
C
C     ----- UPDATE CONFIDENCE LEVEL -----
C
      J[illegible](IPTR,3)+1
      IF(IV.EQ.0) GO TO 98
      IA[illegible]+3*IALPHA)/4
C
C     ----- FIND POINTER ARRAY AND INTERMEDIATE WINNER FUNCTION -----
C
C     *** LEFT TO RIGHT PASS ***
C
98    P(1,IPTR)=1
      I=1
      FSAV=1
      SSAV=S(1)
100   I=I+1
      IF(I.GT.60) GO TO 105
      SSAV=SSAV+IALPHA
      IF(SSAV-S(I)) 101,102,102
101   S(I)=SSAV
      P(I,IPTR)=FSAV
      GO TO 100
102   SSAV=S(I)
      P(I,IPTR)=I
      FSAV=I
      GO TO 100
C
C     *** RIGHT TO LEFT PASS ***
C
105   I=FSAV
      SSAV=S(I)
110   I=I-1
      IF(I.LT.1) GO TO 119
      SSAV=SSAV+IALPHA
      IF(SSAV-S(I)) 111,112,112
111   S(I)=SSAV
      P(I,IPTR)=FSAV
      GO TO 110
112   FSAV=P(I,IPTR)
      I=FSAV
      SSAV=S(I)
      GO TO 110
```

Legend

IPOINT = frame index
IPTR = current frame index
IV2F = voicing decision
IALPHA = confidence level
AMDF(·) = AMDF
MIN = minimum of AMDF
P(·,·) = pointer function
S(·) = intermediate winner function
MIDX = minimum location of current winner function
PATH(·) = pitch trajectory
ITMIDX = pitch value with two-frame delay decision
All variables are internal variables

Table continues.

Table 10-II. FORTRAN Program of Pitch Tracker (Concluded)

```
C
C       *** UPDATE S( ), USING AMDF AND INTERMEDIATE WINNER FUNCTION ***
C           FIND MAXIMUM AND MINIMUM VALUES OF S( ) AND INDEX OF MINIMUM
C
119     S(1)=S(1)+AMDF(1)
        MINSC=S(1)
        MAXSC=MINSC
        MIDX=1
        DO 10 I=2,60
        S(I)=S(I)+AMDF(I)
        IF(S(I).GT.MAXSC) MAXSC=S(I)
        IF(S(I).LT.MINSC) MIDX=I
        IF(S(I).LT.MINSC) MINSC=S(I)
10      CONTINUE
C
C       *** SUBTRACT MINSC FROM S( ) TO PREVENT OVERFLOW ***
C
        DO 15 I=1,60
        S(I)=S(I)-MINSC
15      CONTINUE
        MAXSC=MAXSC-MINSC
C
C       ----- TRACE POINTER -----
C
        J0=IPTR
        PATH(J0)=MIDX
        DO 20 J=1,2
        J1=MOD(J0,5)+1
        PATH(J1)=P(PATH(J0),J0)
        J0=J1
20      CONTINUE
        ITMIDX=PATH(J1)
C
        RETURN
        END
```

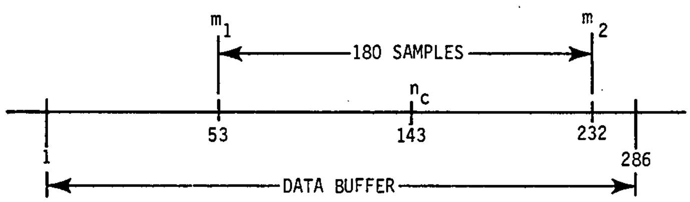


Fig. 10-2. Time Window for Low-passed Speech Energy (ISTU)

The maximum allowable range for the ALO shall be between 50 and 1024. However, the ALO shall have a minimum threshold (CLAMPI) defined as

$$\text{CLAMPI} \leftarrow \frac{1}{32}\,\text{AVENG}$$

where

$$\text{AVENG} \leftarrow \left(\frac{63}{64}\,\text{AVENG} + \frac{1}{64}\,\text{ISTU}\right)$$

which shall be updated when the raw voicing decision is "voiced." ALO shall be computed twice per frame.

3. Zero Crossing Count (NOZ): The number of zero crossings shall be counted on the input speech signal in which its dc value has been removed and a zero crossing bias (ITHZ) has been added. The zero crossing bias shall be defined as

$$\text{ITHZ} = 4 + \frac{\text{ALO}}{32}\,.$$

The count shall begin with $s(m_1 - 1)$ and continue through the succeeding 180 samples. NOZ shall be obtained twice per frame.

4. Two Reflection Coefficients (RK1 and RK2): These are simply the first two reflection coefficients derived as vocal tract filter parameters, available once per frame.

The voicing decision shall have two frames of delay. The initial raw voicing decision (IV2F) shall be refined to IV1F with one frame of delay, which is further refined to IVD with another frame of delay. The raw voicing decision shall have three possible states: IV2F = 2 for definitely voiced, IV2F = 1 for tentatively voiced, and IV2F = 0 for definitely unvoiced. The voicing decision shall proceed in the following four steps:

a. A voicing decision shall be made based on the following energy related parameters:

(1) The low passed speech energy of the present frame (ISTU)

(2) Low passed speech energy of the previous frame (OSTU)

(3) The updated background noise energy (ALO).

The voicing decision shall follow these independent paths depending upon the value of IV2F (i.e., 0, 1, or 2) as illustrated in Fig. 10-3.

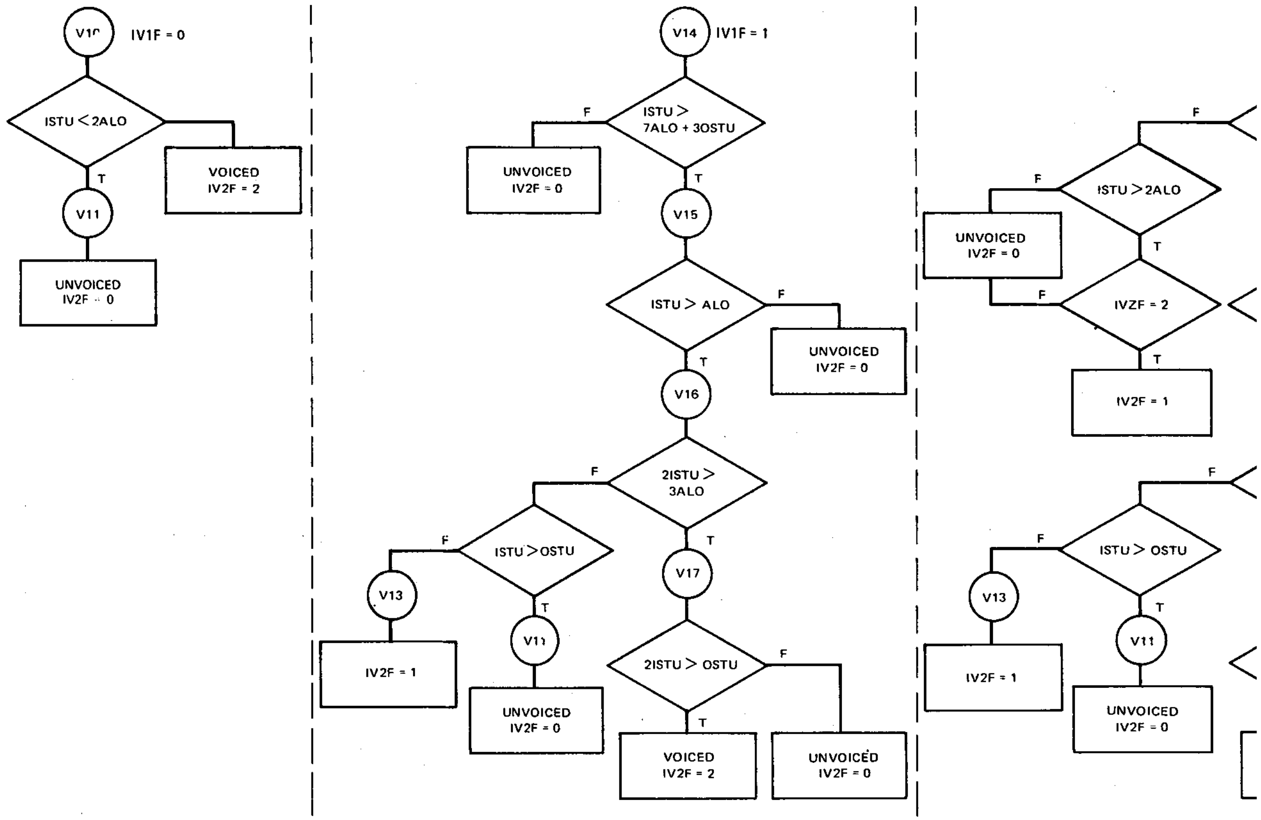


Fig. 10-3 — Raw voicing decision

b. The zero crossing count (NOZ) and the two reflection coefficients from the input speech (RK1 and RK2) shall be used to override the voicing decision made above if

$NOZ > 52$:	change to "unvoiced"
or	
$100(RK1+RK2) < -80$:	change to "unvoiced."

This logical decision shall be delayed by one frame because RK1 and RK2 are not available at this time.

c. IV1F = 1 shall revert to IV1F = 0 if the conditions indicated in Fig. 10-4 are satisfied.

d. The voicing patterns of IV2F, IV1F, and IVD shall be compared to prohibit a single frame of voiced speech.

10.2.1.5 Pitch/Voicing Decision Coding

The pitch period (60 possible candidate values requires six bits for resolution) and the voicing decision (one bit) shall be combined to form a seven bit code. These parameters shall be encoded so that the presence of a one bit error does not produce a decoded voicing error. Likewise, the encoding of the pitch information shall prohibit large pitch differences between the originally encoded pitch and the decoded pitch with single errors. Specifically, the full-frame unvoiced decision shall be encoded as a seven bit word having a Hamming weight of zero (seven zeros). The half-frame voicing transition shall be encoded as seven-bit word having a Hamming weight of 7 (seven ones). The pitch values shall be Gray coded. Table 10-III shall be the encoding table for a combination of the pitch period and the voicing decision.

10.2.2 Vocal Tract Filter Analysis (Reflection Coefficient Estimation)

The vocal tract filter analysis shall produce ten reflection coefficients via linear predictive analysis. Coefficient extraction shall begin with the selection of 130 speech samples from the current frame, which are separated from the previous sample set by an exact multiple of the current pitch period. The selected samples shall be preemphasized. The covariance method of linear prediction analysis shall generate a ten-by-ten and a ten-by-one covariance matrix. Cholesky decomposition shall be used to derive the ten reflective coefficients from these matrices.

10.2.2.1 Semi-Pitch Synchronous Phasing

In order to minimize pitch modulation effects, the analysis shall be made semi-pitch synchronous. This shall be accomplished by placing the present analysis window at a location which is an exact multiple of the current pitch period from the previous window. The

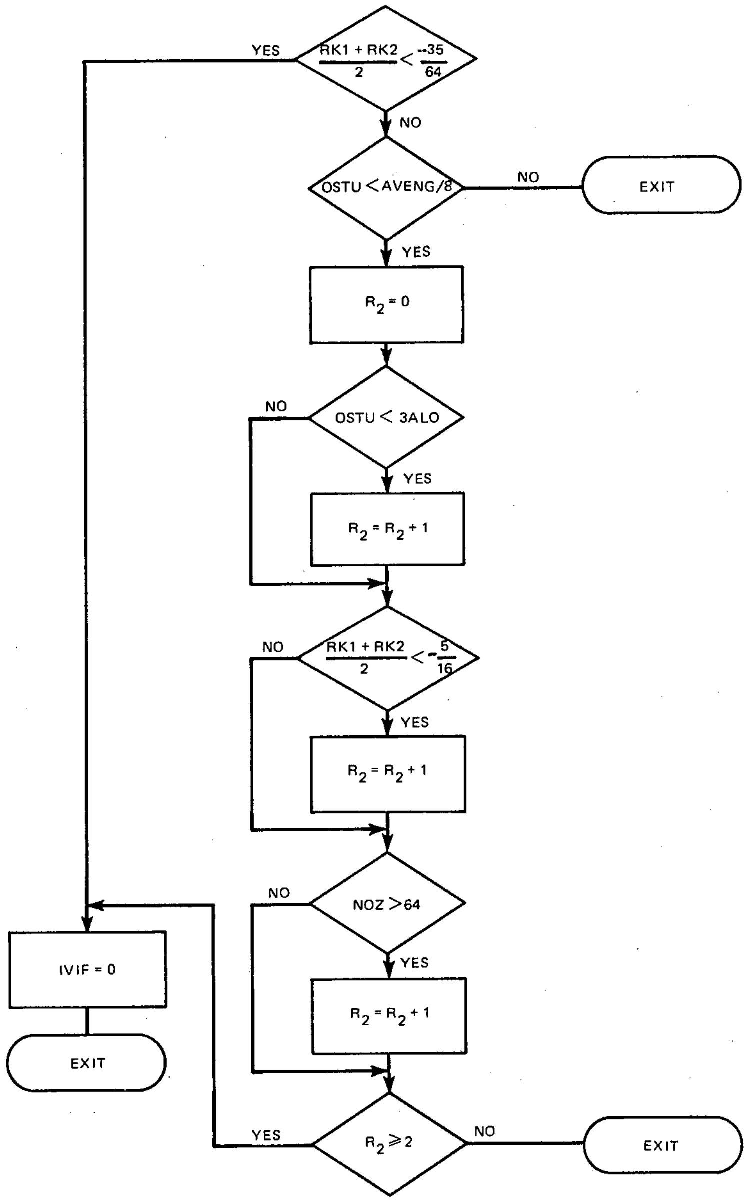


Fig. 10-4 — Voicing decision override logic

Table 10-III. Pitch/Voicing Decision Coding Table

Pitch Period	Pitch Code	Pitch Period	Pitch Code
Unvoiced	0		
20	19	60	90
21	11	62	88
22	27	64	92
23	25	66	84
24	29	68	86
25	21	70	82
26	23	72	83
27	22	74	81
28	30	76	85
29	14	78	69
30	15	80	77
31	7	84	73
32	39	88	75
33	38	92	74
34	46	96	78
35	42	100	70
36	43	104	71
37	41	108	67
38	45	112	99
39	37	116	97
40	53	120	113
42	49	124	112
44	51	128	114
46	50	132	98
48	54	136	106
50	52	140	104
52	60	144	108
54	56	148	100
56	58	152	101
58	26	156	76
		Voicing Transition	127

beginning and end points of the j^{th} analysis window, denoted by $p_1(j)$ and $p_2(j)$, respectively, shall be determined by the following expressions:

$$N(j) = \mathrm{INT}\left[\frac{360 - p_2(j-1)}{\tau(j)}\right]$$

$$p_2(j) = \mathrm{MOD}\left\{[p_2(j-1) + \tau(j)N(j)],\ 180\right\}$$

(if the calculated value of $p_2(j)$ is 0, then replace $p_2(j)$ with the value of 180) and

$$p_1(j) = p_2(j) - 129$$

where $N(j)$ is the analysis window separation in terms of $\tau(j)$

$\tau(j)$ is j^{th} frame pitch period.

For pitch periods greater than or equal to 52 (excluding pitch periods of 60 and 90), the negative values of $p_1(j)$ shall be interpreted as the present analysis window containing speech samples belonging to the previous frame, as illustrated in Fig. 10-5.

10.2.2.2 Preemphasis

A simple and fixed preemphasis (a single zero at $z = 0.9375$) shall be used. The preemphasized speech shall be processed from the expression

$$x(i) = 2s(i) - 1.875\, s(i-1) \qquad i = 1, 2, 3, \ldots, 130$$

where $x(i)$ is the i^{th} preemphasized speech sample

$s(i)$ is the i^{th} speech sample.

10.2.2.3 Matrix Load

A set of simultaneous equations shall be generated by applying the principle of linear prediction to preemphasized speech samples. A speech sample shall be represented by a linear combination of the past P samples:

$$x(i) = \sum_{k=1}^{P} \alpha(k)x(i-k) + \epsilon(i) \qquad i = P+1, P+2, \ldots, M$$

where $\epsilon(i)$ is the i^{th} prediction error sample

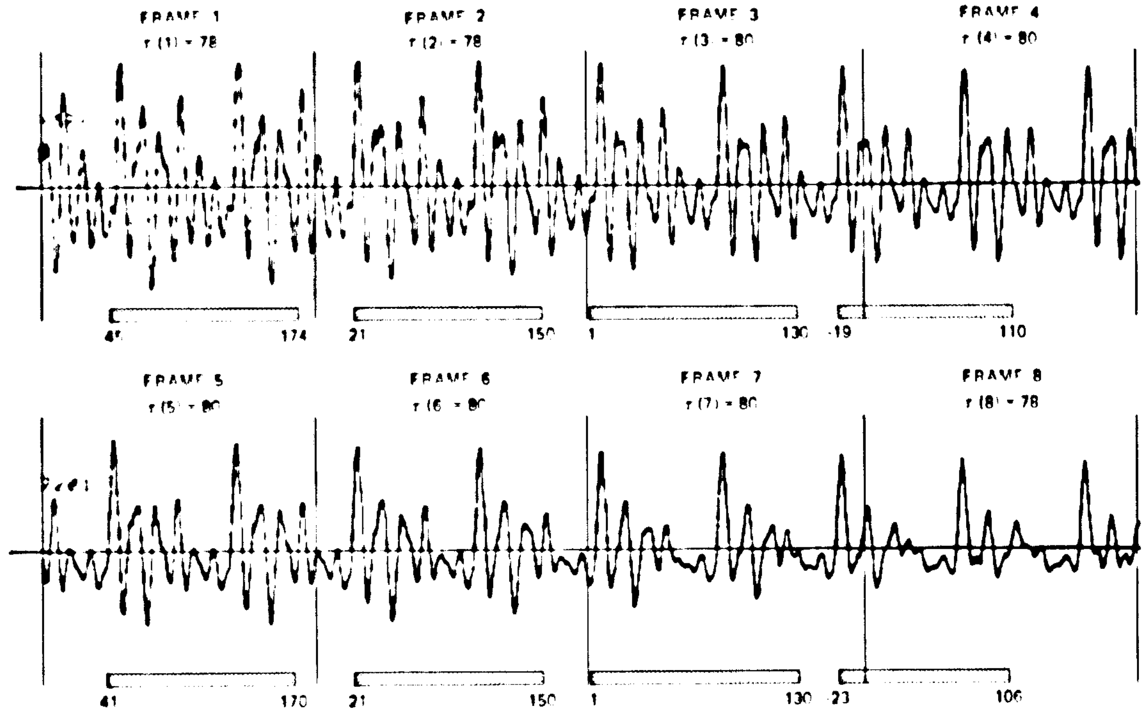


Fig. 10.5 – Example of semi-pitch synchronous phasing

$\alpha(k)$ is the k^{th} prediction coefficient

P is the order of prediction

M is the number of speech samples needed for the analysis.

An implicit solution for prediction coefficients obtained from the above expression shall be in the form of

$$R\alpha = Q$$

where α is the ten by one prediction coefficient matrix

R is the ten by ten covariance matrix (see Fig. 10.6)

Q is the ten by one covariance matrix (see Fig. 10.6).

$$
\underline{R} = \left|\begin{array}{c:c:c:c:c}
\sum_{i=P}^{M-1} x_i^2 & \sum_{i=P}^{M-1} x_i x_{i-1} & \sum_{i=P}^{M-1} x_i x_{i-2} & \cdots & \sum_{i=P}^{M-1} x_i x_{i-P+1} \\ \hdashline
\sum_{i=P-1}^{M-2} x_i x_{i+1} & \sum_{i=P-1}^{M-2} x_i^2 & \sum_{i=P-1}^{M-2} x_i x_{i-1} & \cdots & \sum_{i=P-1}^{M-2} x_i x_{i-P} \\ \hdashline
\sum_{i=P-2}^{M-3} x_i x_{i+2} & \sum_{i=P-2}^{M-3} x_i x_{i+1} & \sum_{i=P-2}^{M-3} x_i^2 & \cdots & \sum_{i=P-2}^{M-3} x_i x_{i-P-1} \\ \hdashline
\vdots & \vdots & \vdots & \cdots & \vdots \\ \hdashline
\sum_{i=1}^{M-P} x_i x_{i+P-1} & \sum_{i=1}^{M-P} x_i x_{i+P-2} & \sum_{i=1}^{M-P} x_i x_{i+P-3} & \cdots & \sum_{i=1}^{M-P} x_i^2
\end{array}\right|
$$

$$
\underline{Q} = \left|\begin{array}{c}
\sum_{i=P}^{M-1} x_i x_{i+1} \\ \hdashline
\sum_{i=P-1}^{M-2} x_i x_{i+2} \\ \hdashline
\sum_{i=P-2}^{M-3} x_i x_{i+3} \\ \hdashline
\vdots \\ \hdashline
\sum_{i=1}^{M-P} x_i x_{i+P}
\end{array}\right|
$$

M = 130 and P = 10. The numbers of terms in the product summation is M − P = 120.

Fig. 10-6 — R and Q covariance matrices

The following properties ofthe R and Q matrices shall be used for the reduction of computations during matrix loading:

1. The first column (or the first row) of the R matrix shall be computed first. The remaining elements along the principle diagonal axis and the axes parallel to it shall be computed from the expression

$$R(i,j) = R(i-1, j-1) - x(M+1-i) \times (M+1-j) + x(P+1-i) \times (P+1-j)$$

where i = 2, 3, . . . , P and j = 2, 3, . . . , i.

2. R(j,i) shall be made equal to R(i,j).

3. The Q-matrix elements shall be obtained from the expression

$$Q(i) = R(i+1, 1) - x(P) \times (P-1) + x(M) \times (M-i)$$

where $i = 1, 2, \ldots, P-1$.

The matrix loading shall use 130 speech samples, and the number of product terms in each matrix element shall be 120. Computations shall include the following scaling operations:

1. The calculations of each matrix element shall produce double precision sums of products (i.e., each speech sample is represented by sign plus 11 bits; when multiplied against another speech sample the result is sign plus 22 bits. The 120 product terms in each matrix element consist of sign plus 29 bits).

2. The double precision values of the main diagonal elements of the R matrix shall be computed first Then the main diagonal elements shall be block scaled to maintain 16 significant bits in the largest matrix element.

The FORTRAN program listed in Table 10-IV shall be referenced during the development of the matrix load software.

10.2.2.4 Matrix Inversion

The reflection coefficients shall be derived through Cholesky decomposition of matrix R:

$$\begin{aligned} R &= L\,D\,L^T \\ &= S\,L^T \end{aligned}$$

where L is the lower triangular matrix

D is the diagonal matrix

L^T is the transposed L matrix

where S matrix elements are obtained from

$$S(i,1) = R(i,1) \qquad i = 1, 2, \ldots, P$$

and

$$S(i,j) = R(i,j) - \sum_{k=1}^{j-1} \frac{S(i,k)S(j,k)}{S(k,k)} \qquad \begin{array}{l} i = 2, 3, \ldots, P \\ j = 2, 3, \ldots, i. \end{array}$$

Table 10-IV. FORTRAN Program of Matrix Load

```
      SUBROUTINE MLOAD(X,R,Q)
C     MATRIX LOAD (NEEDS 130 SPEECH SAMPLES)
C
      DIMENSION X(130),R(10,10),Q(10)
      INTEGER P
      ROF(W)=AINT(W+SIGN(.5,W))
C
C     ----- FIXED CONSTANTS -----
C
C     P-ORDER OF PREDICTOR; M-ANALYSIS WINDOW SIZE
      P=10
      M=130
      BIG=32767.
C
C     ----- COMPUTE FIRST COLUMN ELEMENTS OF R MATRIX -----
C
      IP=P+1
      DO 50 I=1,P
      R(I,1)=0.
      DO 50 K=IP,M
      X1=X(K-1)*X(K-I)/BIG
      R(I,1)=R(I,1)+ROF(X1)
50    CONTINUE
C
C     ----- COMPUTE LAST ELEMENT OF Q MATRIX -----
C
      Q(P)=0.
      DO 60 K=IP,M
      X1=X(K)*X(K-P)/BIG
      Q(P)=Q(P)+ROF(X1)
60    CONTINUE
C
C     ----- COMPUTE REMAINING R MATRIX ELEMENTS -----
C
      DO 70 I=2,P
      DO 70 J=2,I
      X1=X(M+1-I)*X(M+1-J)/BIG
      X2=X(P+1-I)*X(P+1-J)/BIG
      R(I,J)=R(I-1,J-1)-ROF(X1)+ROF(X2)
70    CONTINUE
C
C     ----- COMPUTE REMAINING Q MATRIX ELEMENTS -----
C
      IPP=P-1
      DO 80 I=1,IPP
      X1=X(P)*X(P-I)/BIG
      X2=X(M)*X(M-I)/BIG
      Q(I)=R(I+1,1)-ROF(X1)+ROF(X2)
80    CONTINUE
C
C     ----- FIND MAXIMUM OF DIAGONAL ELEMENTS OF R MATRIX -----
C
      AMAX=0.
      DO 90 I=1,P
```

Legend

ROF = rounding-off operation
BIG = 15-bit number
X = preemphasized speech samples
R = ten-by-ten covariance matrix
Q = ten-by-one covariance matrix.

Table continues.

Table 10-IV. FORTRAN Program of Matrix Load (Concluded)

```
      IF(AMAX.GE.R(I.I))GO TO 90
      AMAX=R(I.I)
90    CONTINUE
      IF(AMAX.LE.0.)AMAX=BIG/4.
C
C     ----- SCALE R AND Q MATRICES -----
C
      DO 100 I=1.P
      Q1=Q(I)*BIG/AMAX
      Q(I)=FIX(Q1)
      DO 100 J=1.I
      R1=R(I.J)*BIG/AMAX
      R(I.J)=FIX(R1)
      R(J.I)=R(I.J)
100   CONTINUE
      DO 110 I=1.P
      IF(R(I.I).GE.BIG)R(I.I)=BIG-1.
110   CONTINUE
C
      RETURN
      END
```

The reflection coefficients (denoted by r) shall be computed from

$$r = S^{-1} Q$$

or

$$r(i) = \frac{Q(i)}{S(i,i)} - \sum_{j=1}^{i-1} \frac{r(j)S(i,j)}{S(i,i)} \qquad i = 1, 2, \ldots, P.$$

Computations shall include block scaling to provide high accuracy with single-precision arithmetic. Multiplication and accumulation operations shall be carried out in double precision followed by rounding to single-precision format. The following computational procedures shall be used:

1. Compute the j^{th} column of the S matrix.

2. Find the maximum of the j^{th} column of the S matrix and store a shift factor y(j) which will shift the MSB of the maximum entry to sign-plus-15-bit format.

3. Scale the j^{th} column of the S matrix by y(j).

4. Find a shift factor which will make S(j,j) to a sign-plus-15-bit format.

5. Compute r(j) and clamp to a sign-plus-15-bit format.

If any of the following conditions occur, computations shall be halted and the reflection coefficients of the previous frame shall be repeated.

1. Any column of the S matrix is too small.
2. Any diagonal element of the S matrix is less than or equal to 0.
3. Any reflection coefficient overflows.

The FORTRAN program listed as Table 10-V shall be referenced during the development of the matrix inversion software.

10.2.2.5 Reflection Coefficient Coding

The first through tenth reflection coefficients shall be coded to five, five, five, five, four, four, four, four, three, and two bits, respectively. The first two reflection coefficients shall be quantized nonlinearly (or linearly on their log area ratios), as indicated in Table 10-VI. The sign bit of each coefficient shall be extended as the sign bit of the code.

The third through tenth reflection coefficients shall be quantized linearly via the arithmetic process indicated below:

1. Expand r_i to a sign plus 14 bits: $-2^{14} < r_i < 2^{14}$ $(i = 3, 4, \ldots, 10)$.

2. Remove coefficient offset bias from each r_i: $r_i \leftarrow (r_i - b_i)$.

3. Multiply r_i by a scale factor and clamp to a sign plus seven bits: $-2^7 < r_i < 2^7$.

4. Shift bits by preassigned amount: $-2^{(7 - L_i)} < r_i < 2^{(7 - L_i)}$.

The constants required for these transformations shall be those listed in Table 10-VII.

For unvoiced frames, only the first four reflection coefficients shall be encoded and transmitted. The 21 available bits (originally allocated for the fifth through tenth reflection coefficients) shall be utilized to protect the amplitude information and the first four reflection coefficients by Hamming (8,4) codes. Table 10-VIII lists the parameters and the number of bits to be protected.

10.2.3 Amplitude Analysis

The amplitude information shall be the speech rms value. The synthesized speech amplitude shall be calibrated so that its rms value matches the transmitted rms value.

UNCLASSIFIED

Table 10-V. FORTRAN Program of Matrix Inversion

```
      SUBROUTINE INVERT(R,Q,RC)
C     (MATRIX INVERSION USING R AND Q MATRICES)
C
      DIMENSION R(10,10),S(10,10),Q(10),RC(10),X(10),Y(10)
      INTEGER P
      [illegible]
C
      P=10
      [illegible]
      SMALL=0.
      Y(1)=1.
C
C     --- --------- -------------------------
C
      DO 120 J=1,P
      JJ=J-1
C
C                   ****************
C                   FORMATION OF S MATRIX
C                   ****************
C
C     ----- TRANSFER R MATRIX COLUMN TO S MATRIX COLUMN -----
C
      DO 5 I=J,P
      S(I,J)=R(I,J)
5     CONTINUE
      IF(J.EQ.1)GO TO 70
C
C     ----- COMPUTE PRODUCT TERM & CHECK OVERFLOW -------
C
      DO 10 K=1,JJ
      SHIFT=X(K)*Y(K)
      SAVE=RTF(S(J,K)*S(K,K)/BIG)
      DO 10 I=J,P
      AII=RTF(S(I,K)*SAVE/BIG)
      AII=AINT(AII*BIG/SHIFT)
      IF(ABS(AII).GE.BIG)GO TO 140
      S(I,J)=S(I,J)-AII
10    CONTINUE
C
C     ----- OVERFLOW & UNDERFLOW CHECKS FOR S MATRIX -----
C
      DO 20 I=1,P
      IF(ABS(S(I,J)).GE.BIG)GO TO 140
20    CONTINUE
      IF(ABS(S(J,J)).LE.SMALL)GO TO 130
C
C     ----- FIND MAX OF S MATRIX COLUMN -----
C
      COLMAX=0.
      DO 30 I=J,P
      IF(ABS(S(I,J)).LE.COLMAX)GO TO 30
      COLMAX=ABS(S(I,J))
30    CONTINUE
```

Legend

R = ten-by-ten covariance matrix
Q = ten-by-one covariance matrix
RC = reflection coefficients

Table continues.

Table 10-V. FORTRAN Program of Matrix Inversion (Continued)

```
C
C       ----- FIND BIT SHIFT FACTOR Y(J) -----
C
        Y(J)=1.
40      IF(COLMAX.GE.BIG/2.)GO TO 50
        COLMAX=COLMAX*2.
        Y(J)=Y(J)*2.
        GO TO 40
C
C       ----- SCALE S MATRIX COLUMN BY Y(J) --------
C
50      DO 60 I=J,P
        S(I,J)=S(I,J)*Y(J)
60      CONTINUE
C
C       ----- UNDERFLOW CHECK FOR DIAGONAL ELEMENTS OF S MATRIX -----
C
        IF(S(J,J).EQ.0)GO TO 150
C
C       ----- FIND BIT SHIFT FACTOR FOR DIAGONAL ELEMENTS, X(J) -----
C
70      X(J)=BIG/2.
80      IF(ABS(S(J,J)).GT.X(J))GO TO 90
        X(J)=X(J)/2.
        GO TO 80
C
C       ----- SCALE DIAGONAL ELEMENTS OF S MATRIX BY X(J) -----
C
90      S(J,J)=ROF(BIG*X(J)/S(J,J))
C
C               ****************************************
C               REFLECTION COEFFICIENT COMPUTATIONS
C               ****************************************
C
        RC(J)=Q(J)
        IF(J.EQ.1)GO TO 110
        DO 100 K=1,JJ
        ADD=ROF(RC(K)*S(J,K)/BIG)
        ADD=AINT(ADD/Y(K))
        RC(J)=RC(J)-ADD
100     CONTINUE
C
C       ----- OVERFLOW CHECK FOR RC(J) -----
C
        IF(ABS(RC(J)).GE.BIG)GO TO 140
C
C       ----- STORE RC'S IN S+15 BIT FORMAT -----
C
110     RC(J)=AINT(RC(J)*BIG/X(J)*Y(J)/2.)
        IF(RC(J).LT.-BIG)RC(J)=-BIG
        IF(RC(J).GE.BIG)RC(J)=BIG-1.
        RC(J)=ROF(RC(J)*S(J,J)/BIG)*2.
        IF(RC(J).LT.-BIG)RC(J)=-BIG
        IF(RC(J).GE.BIG)RC(J)=BIG-1.
```

Table continues.

Table 10 V. FORTRAN Program of Matrix Inversion (Concluded)

```
120     CONTINUE
        GO TO 170
C
C       ----- END OF LOOP -----
C
C       *** UNDERFLOW CASES (PRINT OUT, IF YOU WISH) ***
C
130     PRINT 900,I
900     FORMAT(' UNDERFLOW ON ',I2,'-TH COEFFICIENT')
        GO TO 150
C
C       *** OVERFLOW CASES (PRINT OUT, IF YOU WISH) ***
C
140     PRINT 910,I
910     FORMAT(' OVERFLOW ON ',I2,'-TH COEFFICIENT')
C
C       ----- ZEROIZE R0'S (AND REPEAT PREVIOUS FRAME R0'S) ----
C
150     DO 160 I=1,M
        R0(I)=0.
160     CONTINUE
C
C       ----- CONVERT R0'S TO FRACTIONAL NUMBERS (IF NEEDED) -----
C
170     DO 200 I=1,M
        R(I)=R0(I)/BIT
200     CONTINUE
C
C       -----------------------
C
        RETURN
        END
```

Table 10-VI. Coding Table for First Two Reflection Coefficients

Coefficient Magnitude	Code
0-5	0
6-12	1
13-19	2
20-26	3
27-33	4
34-38	5
39-43	6
44-48	7
49-52	8
53-55	9
56-57	10
58-59	11
60	12
61	13
62	14
63	15

Table 10-VII. Coding Constants for Third Through Tenth Reflection Coefficients

Coefficient Index (i)	Bias (b_i)	Scale Factor (g_i)	Bit-Shift Factor (L_i)
3	1152	0.0112	3
4	– 2816	0.0125	3
5	– 1536	0.0135	4
6	– 3584	0.0143	4
7	– 1280	0.0147	4
8	– 2432	0.0145	4
9	768	0.0167	5
10	– 1920	0.0204	6

Table 10-VIII. Parameter Protection for Unvoiced Frames

Parameter	No. of Bits	No. of Bits Protected	Available Code Bits
Amplitude	5	4	r(8)
r(1)	5	4	r(5)
r(2)	5	4	r(6)
r(3)	5	4	r(7)
r(4)	5	4	r(9) & r(10)*

*One bit of r(10) is not used.

10.2.3.1 Amplitude Computations

The speech power value shall be obtained from the first diagonal element of the R matrix, as a byproduct of the matrix loading:

$$\text{Speech power} = R(1,1) = \sum_{i=P}^{M-1} x_i^2 .$$

The speech rms value shall be obtained by

$$\text{Speech rms value} = \sqrt{\frac{R(1,1)}{M-P}} \quad \text{(for unvoiced)}$$

$$= \sqrt{\frac{R(1,1)}{N}} \quad \text{(for voiced)}$$

where M = 130, P = 10, and

$$N = \text{INT}\left[\frac{120}{\text{current pitch period}}\right].$$

If the current pitch period is greater than 120, then N = 1.

10.2.3.2 Amplitude Coding

The speech rms value shall be encoded by five bits, semilogarithmically, as indicated in Table 10-IX.

Table 10-IX. Amplitude Coding Table

Speech rms Value	Code	Speech rms Value	Code
838 or more	31	34-41	15
686 – 837	30	28-33	14
561 – 685	29	23-27	13
459 – 560	28	19-22	12
375 – 458	27	16-18	11
307 – 374	26	13-15	10
251 – 306	25	11,12	9
206 – 250	24	9,10	8
168 – 205	23	8	7
138 – 167	22	6,7	6
113 – 137	21	5	5
93 – 112	20	4	4
76 – 92	19	3	3
62 – 75	18	2	2
51 – 61	17	1	1
42 – 50	16	0	0

10.2.4 Transmission Data Format (Multiplexing)

Data conditioning shall be provided to reconfigure the 2400-b/s data stream into 16-bit blocks. The sequential order of the data multiplexing shall be as illustrated in Tables 10-X and 10-XI. The sync bit shall be an alternating "1" and "0" from frame to frame.

10.3 Synthesis (LPC Receiver)

The LPC receiver shall perform frame synchronization and data demultiplexing to provide the synthesizer with pitch, voicing decision, amplitude information, and the reflection coefficients. Parameter errors due to transmission anomalies shall be detected and corrected.

The synthesizer shall generate one pitch period at a time using interpolated parameters. The synthesizer shall be a direct-form recursive filter with the prediction coefficients as its filter weights. It shall be driven by a constant excitation signal, and the output shall be matched to the input speech level of the transmitter by scaling the synthesizer output on a pitch-period basis. A functional block diagram of the LPC receiver is illustrated in Fig. 10-7.

Table 10-X. Transmission Data Format for Voiced Frames

TRANSMISSION SEQUENCE

16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
r (4)	P	A	r (3)	r (4)	r (1)	A	P	r (3)	r (2)	r (1)	A	P	r (3)	r (2)	r (1)
1	2	2	2	0	2	1	1	1	1	1	0	0	0	0	0

BIT SIGNIFICANCE

32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17
r (4)	P	r (8)	r (7)	r (2)	P	A	r (4)	r (3)	r (2)	r (1)	A	r (4)	r (3)	r (2)	r (1)
4	4	0	0	4	3	4	3	4	3	4	3	2	3	2	3

48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33
r (9)	P	r (8)	r (10)	r (6)	r (5)	P	r (9)	r (7)	r (6)	r (5)	r (8)	r (10)	r (7)	r (6)	r (5)
1	6	2	1	2	2	5	0	2	1	1	1	0	1	0	0

54	53	52	51	50	49
SYNC	r (8)	r (9)	r (7)	r (6)	r (5)
	3	2	3	3	3

BIT SIGNIFICANCE—"0" MEANS LSB
A: AMPLITUDE PARAMETER
P: PITCH/VOICING PARAMETER
R: REFLECTION COEFFICIENT

Table 10-XI. Transmission Data Format for Unvoiced Frames

TRANSMISSION SEQUENCE

16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
r (4)	P	A	r (3)	r (4)	r (1)	A	P	r (3)	r (2)	r (1)	A	P	r (3)	r (2)	r (1)
1	2	2	2	0	2	1	1	1	1	1	0	0	0	0	0

BIT SIGNIFICANCE

32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17
r (4)	P	A	r (3)	r (2)	P	A	r (4)	r (3)	r (2)	r (1)	A	r (4)	r (3)	r (2)	r (1)
4	4	5	5	4	3	4	3	4	3	4	3	2	3	2	3

48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33
r (4)	P	A	D/C	r (2)	r (1)	P	r (4)	r (3)	r (2)	r (1)	A	r (4)	r (3)	r (2)	r (1)
7	6	7		7	7	5	6	7	6	6	6	5	6	5	5

54	53	52	51	50	49
SYNC	A	r (4)	r (3)	r (2)	r (1)
	8	8	8	8	8

BIT SIGNIFICANCE—"0" MEANS LSB
A: AMPLITUDE PARAMETER
P: PITCH/VOICING PARAMETER
R: REFLECTION COEFFICIENT

ERROR CORRECTION BITS ARE
IDENTIFIED BY THICK LINED BOXES

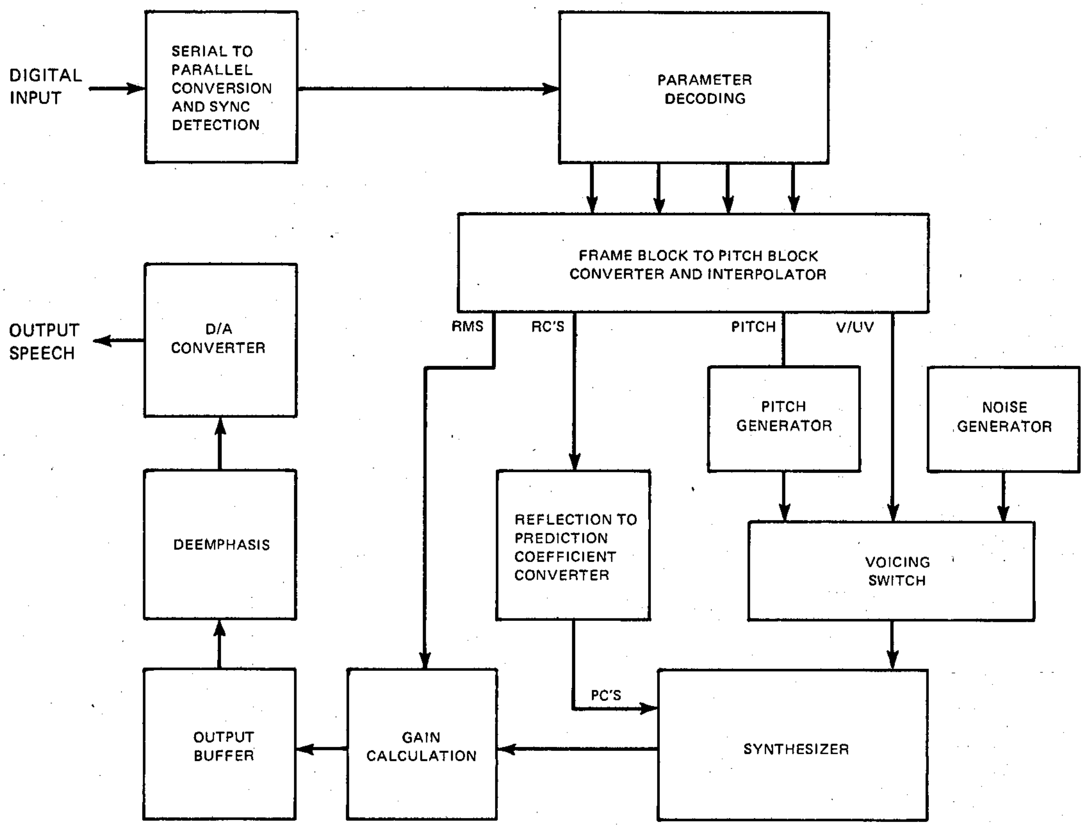


Fig. 10-7 — LPC Receiver

10.3.1 Frame Synchronization

The LPC receiver shall be capable of extracting frame sync (and crossvalidating the modem frame sync) from the 54-bit serial frame sequences generated by the LPC transmitter. Frame synchronization shall consist of two operational modes: sync acquisition and sync maintenance.

10.3.1.1 Sync Acquisition

Frames of raw channel data shall be used to update 54 counters. Each counter shall correspond to a bit location in the raw frame and shall be incremented or reset depending on the level of correlation with the sync sequence at each location. When a counter reaches the preestablished sync threshold of eight, the corresponding bit location shall be taken as the sync location. The pointer specifying the bit location shall be returned to the frame synchronizer with an indication that sync has been achieved.

10.3.1.2 Sync Maintenance

Once sync has been acquired, the sync-maintenance mode shall check the assumed sync bit once per frame for sync verification. It shall maintain a counter indicating the sync confidence level. This counter, with an initial value of eight, shall be incremented by one count for every correct sync bit and decremented by two counts for each incorrect sync bit. The maximum count shall be limited to 15. If the counter is decremented to a value of zero, the sync-acquisition mode shall be reinitiated.

10.3.2 Demultiplexing

The demultiplexer shall accept the sync pointer and new frame of data (16-bit words, right-justified format). It shall be implemented as a software routine driven by a table that indicates the number of parameters and the number of bits for each parameter and their location within the data frame.

10.3.3 Decoding

The 12 unpacked words shall be decoded into 13 parameters (pitch, voicing decision, amplitude information, and ten reflection coefficients). The pitch and voicing parameters shall be demultiplexed from a seven-bit word. The amplitude information shall remain coded and left-shifted one bit before being placed into the frame buffer. Likewise, the first two reflection coefficients shall be left-shifted one bit for interpolation and decoding.

10.3.3.1 Pitch and Voicing

The pitch period and voicing decision shall be decoded as indicated in Table 10-XII. A decoded number greater than or equal to 20 shall indicate a "voiced" decision with a pitch period identical to the decoded value. The decoded value of "0" shall indicate an "unvoiced" decision. The decoded value of "3" implies an invalid pitch period, for which the pitch period of the previous frame shall be repeated. The decoded value of "1" shall imply a voicing transition.

10.3.3.2 Reflection Coefficients

The first two reflection coefficients shall be decoded directly by the values indicated in Table 10-XIII. The sign of each reflection coefficient shall be extended from the sign of the respective code.

The third through tenth reflection coefficients shall be decoded by the inverse arithmetic process of encoding in the following manner:

1. Expand r_i to a sign + 14 bits by shifting bits by $7+L_i$: $-2^{14} < r_i < 2^{14}$ $(i = 3, 4, \ldots, 10)$.

Table 10-XII. Decoding Table for Pitch and Voicing Decision

Received Pitch Code	Pitch Period/ Voicing	Received Pitch Code	Pitch Period/ Voicing	Received Pitch Code	Pitch Period/ Voicing
0	0	46	34	92	64
1	0	47	3	93	3
2	0	48	3	94	3
3	3	49	42	95	1
4	0	50	46	96	3
5	3	51	44	97	116
6	3	52	50	98	132
7	31	53	40	99	112
8	0	54	48	100	148
9	3	55	3	101	152
10	3	56	54	102	3
11	21	57	3	103	3
12	3	58	56	104	140
13	3	59	3	105	3
14	29	60	52	106	136
15	30	61	3	107	3
16	0	62	3	108	144
17	3	63	1	109	3
18	3	64	0	110	3
19	20	65	3	111	1
20	3	66	3	112	124
21	25	67	108	113	120
22	27	68	3	114	128
23	26	69	78	115	3
24	3	70	100	116	3
25	23	71	104	117	3
26	58	72	3	118	3
27	22	73	84	119	1
28	3	74	92	120	3
29	24	75	88	121	3
30	28	76	156	122	3
31	3	77	80	123	1
32	0	78	96	124	3
33	3	79	3	125	1
34	3	80	3	126	1
35	3	81	74	127	1
36	3	82	70		
37	39	83	72		
38	33	84	66		
39	32	85	76		
40	3	86	68		
41	37	87	3		
42	35	88	62		
43	36	89	3		
44	3	90	60		
45	38	91	3		

Table 10-XIII. Decoding Table for First Two Reflection Coefficients

Code	Coefficient Magnitude	Code	Coefficient Magnitude
0	2	16	50
1	6	17	52
2	9	18	54
3	13	19	55
4	16	20	57
5	19	21	58
6	23	22	59
7	27	23	59
8	30	24	60
9	33	25	60
10	36	26	61
11	39	27	61
12	41	28	62
13	43	29	62
14	46	30	63
15	48	31	63

2. Add quantization bias (q_i): $r_i \leftarrow r_i + q_i$.

3. Compress r_i by scale factor (g_i'): $r_i \leftarrow r_i g_i'$.

4. Add coefficient offset bias (b_i): $r_i \leftarrow r_i + b_i$.

The constants required shall be those presented in Table 10-XIV.

10.3.3.3 Amplitude Information

The amplitude information shall be interpolated based on its code (i.e., log rms of speech) prior to decoding. The interpolated amplitude (in a six-bit format) shall be decoded by utilizing the values indicated in Table 10-XV.

10.3.4 Error Corrections

For unvoiced frames, if two errors are detected in a parameter (a single error is corrected by the Hamming (8,4) code), its value shall be replaced by the value from the previous frame.

Table 10-XIV. Decoding Constants for Third Through Tenth Reflection Coefficients

Coefficient Index (i)	Coefficient Offset Bias (b_i)	Quantization Bias (q_i)	Scale Factor (g_i')	Bit Shift Factor (L_i)
3	1152	511	0.6953	3
4	−2816	511	0.6250	3
5	−1536	1023	0.5781	4
6	−3584	1023	0.5469	4
7	−1280	1023	0.5312	4
8	−2332	1023	0.5391	4
9	768	2047	0.4688	5
10	−1920	4095	0.3828	6

Table 10-XV. Decoding Table for Amplitude Information

Code	Speech rms Value	Code	Speech rms Value
0	0	32	50
1	1	33	55
2	1	34	61
3	2	35	68
4	2	36	75
5	3	37	83
6	3	38	92
7	4	39	101
8	4	40	112
9	5	41	124
10	5	42	137
11	6	43	151
12	7	44	167
13	7	45	185
14	8	46	205
15	9	47	226
16	10	48	250
17	11	49	277
18	12	50	306
19	14	51	339
20	15	52	374
21	17	53	414
22	18	54	458
23	20	55	506
24	22	56	560
25	25	57	619
26	27	58	685
27	30	59	757
28	33	60	837
29	37	61	926
30	41	62	1024
31	45		

For voiced frames, gross errors in the pitch, amplitude, and first six reflection coefficients shall be corrected by the following rule:

$$\text{If } |w(j-1) - w(j)| > \text{threshold}$$

and if

$$|w(j+1) - w(j)| > \text{threshold}$$

then let

$$w(j) = \text{median}\,[w(j-1), w(j), w(j+1)]$$

where $w(j-1)$, $w(j)$, and $w(j+1)$ are parameter values of the previous, present, and future frames, respectively. The threshold level is a function of the average error rate as indicated by the Hamming codes during unvoiced frames (as the error rate increases, the threshold decreases). The threshold level shall be updated every frame.

10.3.5 Reflection-to-Prediction-Coefficient Conversion

Reflection coefficients shall be converted to prediction coefficients by the following recursive expressions:

$$\alpha(1,1) = r(1)$$

and

$$\alpha(n,j) = \alpha(n\text{-}1,j) - r(n) \quad \alpha(n\text{-}1,n\text{-}j)$$

$$n = 2, 3, \ldots, P$$
$$j = 1, 2, \ldots, n$$

where n is the index for the recursion cycle

$\alpha(P,j)$ is the j^{th} prediction coefficient.

The first few iterations shall generate the following:

n=1: $\alpha(1,1) = r(1)$

n=2: $\alpha(2,1) = \alpha(1,1) - r(2)\alpha(1,1)$
$\alpha(2,2) = r(2)$

n=3: $\alpha(3,1) = \alpha(2,1) - r(3)\alpha(2,2)$
$\alpha(3,2) = \alpha(2,2) - r(3)\alpha(2,1)$
$\alpha(3,3) = r(3)$

n=4: $\alpha(4,1) = \alpha(3,1) - r(4)\alpha(3,3)$
$\alpha(4,2) = \alpha(3,2) - r(4)\alpha(3,2)$
$\alpha(4,3) = \alpha(3,3) - r(4)\alpha(3,1)$
$\alpha(4,4) = r(4)$

and so on.

The conversion shall utilize a sign-plus-15-bit format for each reflection coefficient and each prediction coefficient. A scale factor shall be stored indicating the number of bits by which the prediction coefficients were right-shifted to avoid overflow. The FORTRAN program, listed in Table 10-XVI, shall be referenced during the development of the coefficient conversion software.

10.3.6 Parameter Interpolation

The decoded parameter values shall be converted from frame block format to pitch-period block format by interpolation of the pitch period, amplitude, and prediction coefficients.

Table 10-XVI. FORTRAN Program for Coefficient Conversion

```
      SUBROUTINE RPCONV(RC.PC.SCALE)
C     REFLECTION TO PREDICTION COEFFICIENT CONVERSION (NEEDS 10 RC'S)
C
      DIMENSION RC(10).PC(10).DUMY(10)
      INTEGER P
      ROF(W)=AINT(W+SIGN(.5.W))
C
      P=10
      BIG=32767.
C
      DO 1 I=1.P
      RC(I)=RC(I)*BIG
1     CONTINUE
C
      SCALE =1.
      PC(1)=RC(1)
      DO 50 I=2.P
      II=I-1
C
      DO 30 J=1.II
      IJ=I-J
10    HOLD=PC(J)-ROF(RC(I)*PC(IJ)/BIG)
      IF(ABS(HOLD).LT.BIG)GO TO 30
      DO 20 L=1.II
      DUMY(L)=AINT(DUMY(L)/2.)
      SCALE=SCALE*2.
20    CONTINUE
      GO TO 10
30    DUMY(J)=HOLG
      DO 40 J=1.II
      PC(J)=DUMY(J)
40    CONTINUE
C
      PC(I)=AINT(RC(I)/SCALE)
50    CONTINUE
      END
```

Legend

RC = reflection coefficients
PC = prediction coefficients
SCALE = scale factor.

The pitch period shall be interpolated linearly at each pitch epoch (i.e., the beginning of each pitch cycle). In order to perform pitch period interpolation, the intraframe pitch trajectory shall be determined (see Figure 10-8) by the equation

$$PP(t) = Mt + B$$

with

$$M = \frac{NP - OP}{L + \Delta L}$$

and

$$B = OP + M(\Delta L + 1)$$

where OP is the old pitch period at the last pitch epoch of the previous frame (at $t = -\Delta L - 1$)

NP is the new decoded pitch period for the present frame (at $t = L$).

The above computations shall be performed once per frame. A new pitch epoch location (t_j) in terms of the previous location (t_{j-1}), the slope (M), and the intercept (B) of the intraframe pitch period trajectory shall be determined by

$$PP(t_j) = \frac{Mt_{j-1} + B}{1 - M}.$$

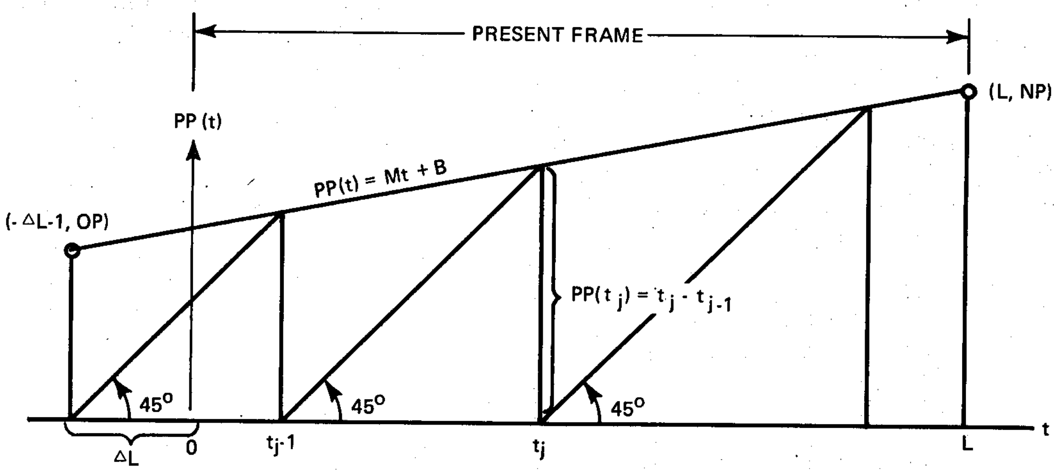


Fig. 10-8 — Interpolation of pitch period and

$$t_j = t_{j-1} + PP(t_j).$$

Prediction coefficients shall be selected (the new, the old, or the average) depending on the location of a pitch epoch referenced to the frame. Table 10-XVII shall be the prediction-coefficient selection rules.

The amplitude information shall be interpolated linearly on a log rms basis. It shall have two forms:

1. Normal, as illustrated in Fig. 10-9A
2. Delayed, as shown in Fig. 10-9B.

Depending on the voicing states between two adjacent half frames (note that the voicing decision is computed twice per frame), the following four rules shall be applied:

1. Voiced to voiced:

   a. Compute pitch for each epoch
   b. Select prediction coefficients
   c. Do normal log rms interpolation and table lookup.

2. Unvoiced to unvoiced:

   a. Let the pitch period equal 45
   b. Select prediction coefficients
   c. Do normal log rms interpolation and table lookup.

Table 10-XVII. Prediction Coefficient Selection Rule

Epoch Location in Terms of Frame Reference	Coefficient Selection Rule
$t_j < 45$	Old value
$46 < t_j < 135$	Mean of old and new values
$t_j > 136$	New value

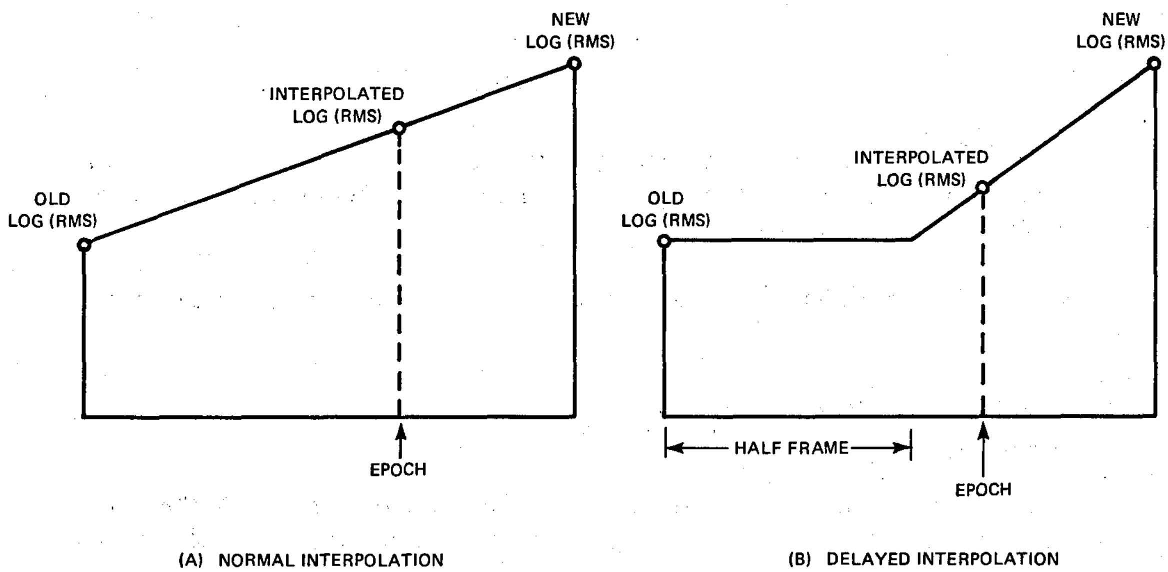


Fig. 10-9 — Amplitude interpolation rule

3. Voiced to unvoiced:

   a. Generate pitch epoch using the pitch of last frame
   b. Select old prediction coefficients
   c. Do normal log rms interpolation and table lookup.

4. Unvoiced to voiced:

   a. Generate one unvoiced pitch epoch (90) using old prediction coefficients
   b. Generate voiced pitch epoch using new prediction coefficients and new pitch
   c. Do delayed log rms interpolation and table lookup.

10.3.7 Excitation Signal Generator

The input to the excitation signal generator shall be the:

1. Voicing indicator
2. Scale factor
3. Pitch period.

If the voicing indicator is 0 (i.e., unvoiced), a random-number generator shall provide one interpolated pitch period of random numbers, uniformly distributed in the range of ±256. Random numbers shall be generated from a table with a random starting point and a random indexing increment or any other equivalent method.

If the voicing indicator is 1, the excitation signal shall be as indicated in Table 10-XVIII. Depending on the current interpolated pitch period, one of the following loading procedures shall be used:

1. If the current pitch period is greater than or equal to 40, the 40 excitation signal samples shall be loaded into the synthesis buffer. Zeros shall be entered for the remainder of the pitch period.

2. If the current pitch period is less than 40, the excitation signal samples shall be loaded only until a new pitch epoch (and a new pitch period) has been generated. If the previous pitch period (ℓ) was less than 40, the last $(40 - \ell)$ samples in the excitation signal shall be added to the first $(40 - \ell)$ excitation points generated.

10.3.8 Vocal-Tract Filter

The vocal-tract filter shall be a direct-form recursive filter with prediction coefficients as its filter weights. The speech synthesis buffer shall contain an excitation signal preceded by ten (=P) samples from the tail of the previous synthesized speech. The operation performed by the speech synthesizer shall be

$$y(n) = \sum_{j=1}^{P} \alpha(j) y(n-j) + v(n)$$

$$= 2^{8} \sum_{j=1}^{P} \left[\alpha(j) 2^{-8}\right] y(n-j) + v(n)$$

where $\alpha(j)$ is the j^{th} prediction coefficient

$y(n)$ is the n^{th} sample of synthesized speech

2^{8} is a scale factor by which the predictive coefficients have been shifted down

$v(n)$ is the n^{th} excitation signal sample. The above operation shall be performed for increasing values of n from P+1 to P plus the interpolated pitch period.

If an overflow occurs, the value of $y(n-1)$ shall be entered in the location for $y(n)$. The maximum absolute sample value generated during the pitch cycle shall be detected and saved for the output amplitude calbration.

Table 10-XVIII. Excitation Signal for Voice Sounds

Index	Amplitude
1	249
2	– 262
3	363
4	– 362
5	100
6	367
7	79
8	78
9	10
10	– 277
11	– 82
12	376
13	288
14	– 65
15	– 20
16	138
17	– 62
18	– 315
19	– 247
20	– 78
21	– 82
22	– 123
23	– 39
24	65
25	64
26	19
27	16
28	32
29	18
30	– 15
31	– 29
32	– 21
33	– 18
34	– 27
35	– 31
36	– 22
37	– 12
38	– 10
39	– 10
40	– 4

10.3.9 Amplitude Calibration

The output amplitude calibration shall utilize the following:

1. The rms value (A)
2. The maximum absolute value of synthesized samples ($\hat{Y}$)
3. The pitch period length (ℓ)
4. The number of remaining samples needed for output for current frame
5. The location of the last sample entered in the speech output buffer.

The output amplitude calibration shall begin with scaling of the speech samples currently in the synthesis buffer:

$$y(n) \leftarrow y(n)G \qquad n = P+1, P+2, \ldots, P+\ell$$

where

$$G = \frac{32767}{\hat{Y}}.$$

Then the 16 LSB's of y(n) shall be taken to compute the total power by

$$\text{power} = \sum_{n=P+1}^{P+\ell} y^2(n)$$

where the power is in double precision and only 16 MSB's of $y^2(n)$ shall be accumulated. Finally, speech samples shall be calibrated by

$$y(n) \leftarrow y(n)G' \qquad n = P+1, P+2, \ldots, P+\ell$$

where

$$G' = \frac{8A}{\sqrt{\frac{\text{power}}{\ell}}}.$$

10.3.10 Deemphasis

The deemphasis characteristic shall be the inverse of the preemphasis. The deemphasis output shall be the operation

$$z(n) = y(n) + 0.9375z(n-1)$$

where y(n) is the input

z(n) is the output.

For each selection on the attached cassette a 16-kb/s CVSD is played and then the 9.6-kb/s MRP is played:

1. Male speech with a high-quality microphone,
2. Male and female dialog with a high-quality microphone,
3. Telephone (high-passed) speech with a carbon microphone,
4. Noisy speech taken from a tactical helicopter,
5. Fast talking from AM broadcast,
6. Casual talking by President Kennedy at a White House press conference (with interfering laughter).

DEPARTMENT OF THE NAVY

NAVAL RESEARCH LABORATORY
Washington, D.C. 20375

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
DEPARTMENT OF THE NAVY
DoD-316

THIRD CLASS MAIL

U.S.MAIL