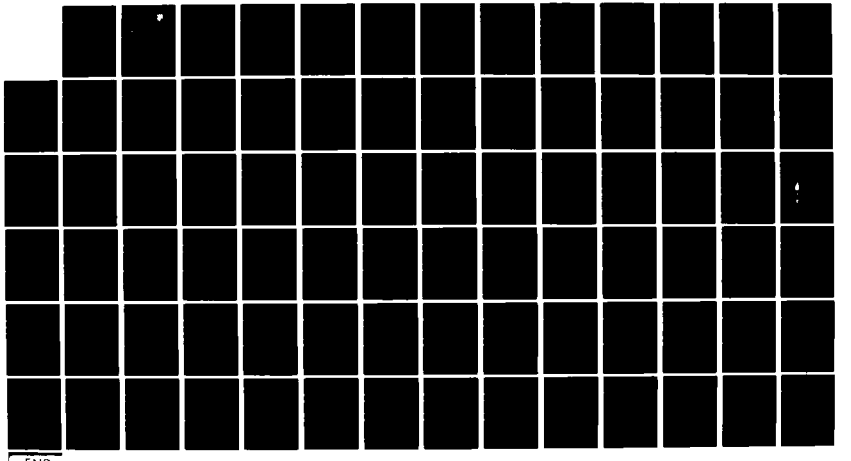
AD-A122 519 VECTOR/MATRIX QUANTIZATION FOR NARROW-BANDWIDTH DIGITAL SPEECH COMPRESSION(U) SIGNAL TECHNOLOGY INC GOLETA CA D Y WONG SEP 82 RADC-TR-82-246 F30602-81-C-0054 1/1

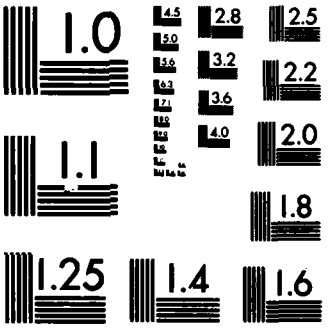
UNCLASSIFIED F/G 17/2 NL



END

FILMED

DTIC



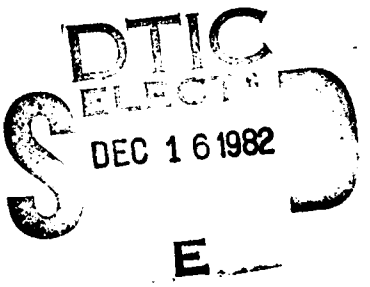
(12)

RADC-TR-82-246
Final Technical Report
September 1982

VECTOR/MATRIX QUANTIZATION FOR NARROW-BANDWIDTH DIGITAL SPEECH COMPRESSION

Signal Technology, Inc.

David Y. Wong



ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-82-246 has been reviewed and is approved for publication.

APPROVED: Silvi K. Steigerwald

SILVI STEIGERWALD, 2/Lt, USAF
Project Engineer

APPROVED: John N. Entzminger

JOHN N. ENTZMINGER, JR.
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: RADC-TR-82-246	2. GOVT ACCESSION NO.: AD-A122519
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): VECTOR/MATRIX QUANTIZATION FOR NARROW-BANDWIDTH DIGITAL SPEECH COMPRESSION	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report 13 Jan 81 - 13 May 82
	6. PERFORMING ORG. REPORT NUMBER: N/A
7. AUTHOR(s): David Y. Wong	8. CONTRACT OR GRANT NUMBER(s): F30602-81-C-0054
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Signal Technology, Inc. 5951 Encina Road Goleta CA 93117	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62702F 45941580
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAA) Griffiss AFB NY 13441	12. REPORT DATE: September 1982
	13. NUMBER OF PAGES: 76
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Silvi K. Steigerwald, 2/Lt, USAF (IRAA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

speech compression, linear predictive coding (LPC), speech coding, vocoders, switched source model, rate distortion, vector quantization, codebook generation, clustering, spectral distortion, formant frequencies, allophones, frame predictive coding, M-L search, scrambling function, diagnostic rhyme (DRT)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Speech compression techniques for very low data rate compression are studied. The techniques are based on a standard LPC analysis/synthesis (vocoder) system. Significant advances are made in the quantization algorithms to achieve bit rates of 200 to 400 bps.

Frame predictive vector quantization is developed to compress the bit rate for the LPC model filter to under 250 bps. The vector quantization

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

technique developed applies to continuous speech and is independent of both speaker and vocabulary.

An innovative LPC compression technique, matrix quantization, is also developed to compress the LPC model filter to a rate under 150 bps. The design is applicable to continuous speech and unlimited vocabulary. At this stage of development, it is adapted to a single speaker, but theoretically it can be generalized to a selection of speakers or even the general population.

In comparison, the LPC filter in a standard 2400 bps LPC-10 system is encoded at a rage of 1820 bps.

Fake process trellis coding algorithms are developed for compressing the vocoder excitation parameters. The results show that if these parameters are compressed independent of the LPC model, an overall bit rate for under 125 bps can be obtained while preserving the prosodic information and natural quality of the speech. In comparison, the bit rate of encoding these parameters in a 2400 bps LPC-10 system is 533 bps.

By combining frame predictive LPC vector quantization with trellis coding of the excitation parameters, the overall vocoder bit rate is reduced to under 400 bps. The bit rate is reduced to about 200 bps by combining LPC matrix quantization with trellis coding of the pitch and gain parameters.

Subjective evaluation of both the vector and matrix LPC quantization approaches using the diagnostic rhyme test (DRT) has been performed and the test scores are analyzed in detail. The results indicate that the proposed techniques are feasible for intelligible speech transmission at bit rates of 400 bps and 200 bps. Recommendations for improvements to the algorithm for better quality and lower complexity are also presented.

Accession For	
NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

TABLE OF CONTENTS

Summary of Technical Effort

Speech compression techiques for very low data rate compression are studied. The techniques are based on a standard LPC analysis/synthesis (vocoder) system. Significant advances are made in the quantization algorithms to achieve bit rates of 200 to 400 bps.

Frame predictive vector quantization is developed to compress the bit rate for the LPC model filter tc under 250 bps. The vector quantization technique developed applies to continuous speech and is independent of both speaker and vocabulary.

An innovative LPC compression technique, matrix quantization, is also developed to compress the LPC model filter to a rate under 150 bps. The design is applicable to continuous speech and unlimited vocabulary. At this stage of development it is adapted to a single speaker, but theoretically it can be generalized to a selection of speakers or even the general population.

In comparison, the LPC filter in a standard 2400 bps LPC-10 system is encoded at a rate of 1820 bps.

Fake process trellis coding algorithms are developed for compressing the vocoder excitation parameters. The results show that if these parameters are compressed independent of the LPC model, an overall bit rate for under

125 bps can be obtained while preserving the prosodic information and natural quality of the speech. In comparison, the bit rate of encoding these parameters in a 2400 bps LPC-10 system is 533 bps.

By combining frame predictive LPC vector quantization with trellis coding of the excitation parameters, the overall vocoder bit rate is reduced to under 400 bps. The bit rate is reduced to about 200 bps by combining LPC matrix quantization with trellis coding of the pitch and gain parameters.

Subjective evaluation of both the vector and matrix LPC quantization approaches using the diagnostic rhyme test (DRT) has been performed and the test scores are analyzed in detail. The results indicate that the proposed techniques are feasible for intelligible speech tranmission at bit rates of 400 bps and 200 bps. Recommendations for improvements to the algorithm for better quality and lower complexity are also presented.

1.0 INTRODUCTION

1.1 Background

For about a decade since the introduction of LPC techniques [1-3] the bit rate of 2400 bps has become a recognized lower bound for practical good quality speech coding. A number of LPC vocoders have already been built and some commercial models are already in use with reported success.

A number of speech coders at bit rates of 1200 bps to 600 bps have also been developed [4-6] and are implementable in real time. These systems are inferior to the 2400 bps in quality but appear to be acceptable for communication purposes. Their acceptability has yet to be demonstrated through more tests and actual usage.

In the last few years, Oshika [7] and Schwartz et al. [8] have reported the development of systems that operate at 200 bps or lower. These approaches are similar in that they both exploit existing techniques in automatic speech recognition. The belief is that there is no graceful degradation from 2400 bps LPC to a 100-200 bps system. Therefore, in these systems, the speech signal is compressed to the phonemic level along with some prosodic information such as pitch, gain and duration.

The research on very low rate speech compression discussed in this report is based on the recent development of an optimal rate distortion vector quantization technique [9,10]. With the vector quantization approach, an 800 bps LPC system has recently been implemented [6]. Trained to a specific speaker, this design is equal in quality to existing 2400 bps LPC systems. For a general population, the quality of the present version 800 bps system is found to be slightly inferior, but the degradation is graceful.

Conceptually, the basic theory involved in the development of an 800 bps vector quantization coder points to the existence of various speech coder designs below 800 bps. Of particular interest are the predictive vector coding and the matrix coding techniques. A qualitative review of vector quantization is presented below to motivate the frame predictive vector quantization and matrix quantization coding techniques.

Human speech perception can be thought of as an information processing structure involving (i) acoustic analysis, (ii) phonological analysis, and (iii) higher level linguistic analysis such as syntactic and semantic analysis (Fig. 1.1) [11]. The phonemic and diphone approaches essentially try to substitute the first two levels of the processing structure with machine recognition, reducing speech to the phonemic level. With the vector/matrix approach, human phonological analysis is not replaced by

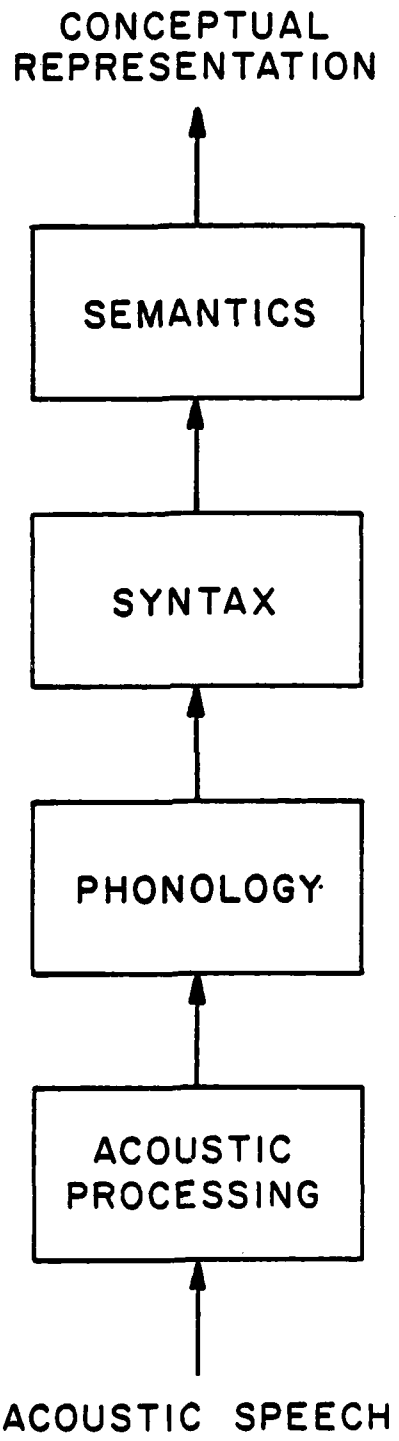


Figure 1.1 Information Processing Structure of Human Speech Perception.

machine processing. Instead, a more fundamental approach of efficient parametric coding based on minimizing a spectral distortion measure is taken.

Vector quantization, or block quantization, has been studied for several decades by information theorists and communication engineers [12,13]. When applied to LPC speech compression, vector quantization is the more appropriate terminology because a vector here refers to a set of LPC filter coefficients representing a particular spectral model. Later in this discussion, the term matrix is used to refer to a sequence of several time consecutive vectors.

To motivate the technique of vector quantization, we first model the speech production process as a switched source as shown in Fig. 1.2. It consists of a composite source and a switch. In the composite source resides a finite (but large) number of different short term speech models. Each unit in the source corresponds to T sec (e.g. 10 msec) of speech. A speech signal is produced by switching from one of the sources to another at T sec intervals. This model is based on the common knowledge that a speaker generally produces only a finite number of perceptually distinct speech sounds, each lasting a short duration of time, typically under 100 msec.

In the traditional technique of LPC speech compression, each T sec of speech, f, is replaced by an LPC model $\tilde{f}$, a gain, and a pitch value. The LPC coefficients are then

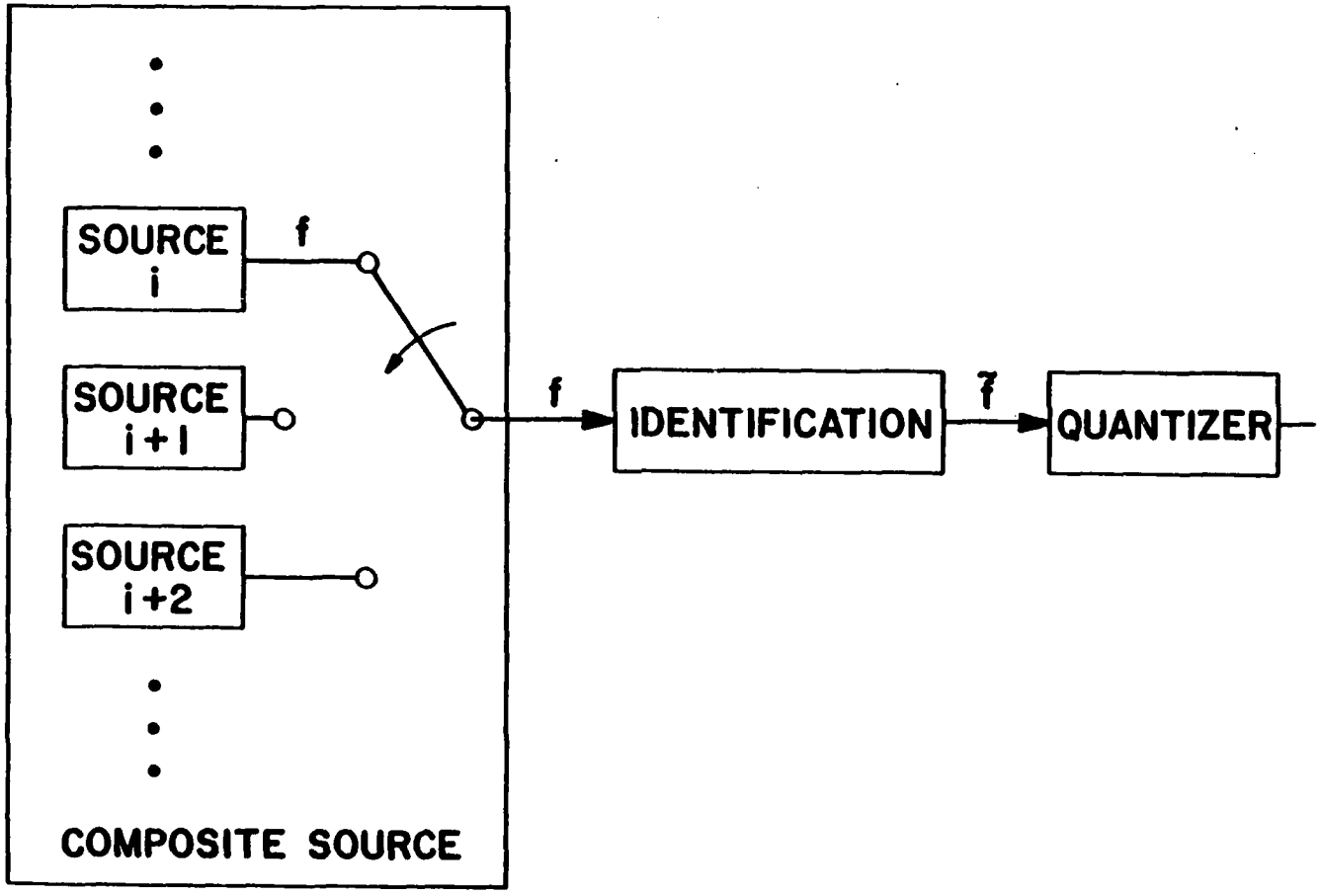


Figure 1.2 A Switched Source Model of Speech Production.

quantized and transmitted. For a 2400 bps system, a set of 10 LPC coefficients typically requires 40-50 bits for transmission. There are a number of serious inefficiencies to such an approach for quantization. They are as follows.

(1) The LPC model $\tilde{f}$ is extracted to minimize some error criterion such as the likelihood ratio or the Itakura-Saito distortion measure [14,17], but the LPC coefficients are quantized according to an error criterion on the coefficients. There is an inconsistency in the criterion, since minimizing the coefficient error does not lead to minimization of the spectral error criterion, and it certainly does not minimize the overall distortion between the speech spectrum f and the quantizer output $\hat{f}$.

(2) Adopting the switched source model, 40-50 bits can theoretically encode from 2^{40} to 2^{50} different spectral models! In reality, we can safely assume nobody produces more than several hundred perceptually distinct speech sounds. There is obviously great inefficiency in the scalar quantization of LPC coefficients. The causes are found to be the following:

(i) A vast majority of the different LPC coefficient vectors allowed by the scalar quantization tables never occur in encoding actual LPC models. From the viewpoint of vocal tract modelling, most of the vocal tract configurations allowed by the scalar quantization tables are not realized by human speakers.

(ii) LPC filter with different coefficient values are coded as different with scalar quantization. However, quite often two LPC filters with different coefficients correspond to very similar spectra. Such LPC filters should be consolidated into one LPC model in the quantizer.

The logical approach to eliminating these inefficiencies is to quantize the LPC coefficients one vector at a time and to quantize them according to the same error criterion used in LPC analysis. Thus an optimal vector quantization LPC coding system has been developed [9]. Such a system consists of a codebook of LPC vectors and a search algorithm (Fig. 1.3). Each incoming vector is compared to each codeword (prestored LPC vector) in the codebook until the best match according to a distortion measure criterion is found.

The codebook is obtained by a clustering procedure which minimizes the average distortion for a large training data base of LPC vectors obtained from real speech. For the given training data base, the codebook achieves a local minimum in average distortion. The minimum is local because the clustering process depends on the initial conditions of the clustering process.

If the training data base is adequately large, the codebook generated from it will perform equally well for any input speech. Based on the vector quantization technique,

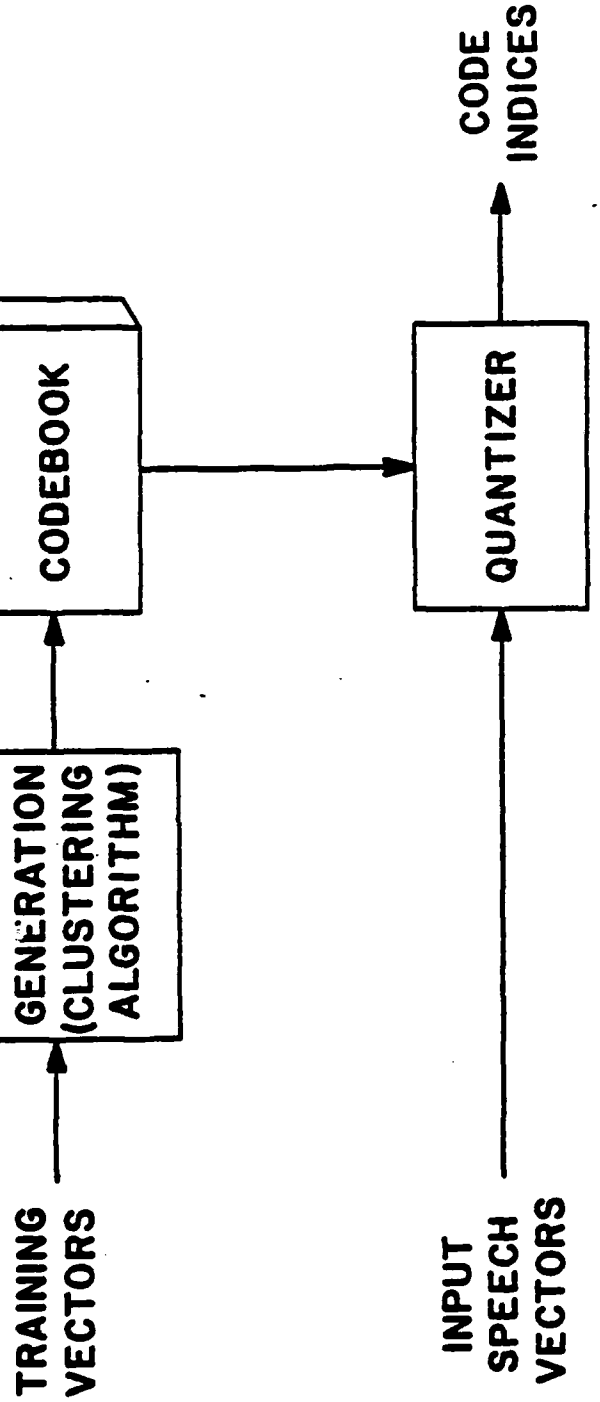


Figure 1.3 Block Diagram of a Vector Quantization System

an 800 bps LPC vocoder has been implemented and fully demonstrated to be feasible for very low rate speech coding [6]. At the frame rate of 44.4 frames/sec, 10 bits/frame are used to quantize the LPC model, thus allowing 2^{10} (1024) different spectral models to be transmitted. The average distortion performance of a 10-bit vector quantizer is found to be comparable to a 27 bits/frame optimized scalar quantizer. The perceived quality is considerably better as discussed in [10].

The code words in the vector quantization code book have been found to be very similar in function to the allophones (variations of phonemes) used in phonemic synthesis. Fig. 1.4 is a plot of the first two formants (F_1,F_2) of a 5-bit (32 code words) vector code book. The dots are the F_1 and F_2 values of the 32 code words, and the ellipse-like cells correspond to Peterson-Barney phonemic spaces for the standard American English vowels [15]. An important point to note here is that while vector quantization leads to an allophonic-like classification of speech spectra, the coding is performed entirely at the acoustic level based on a spectral distortion criterion. The system, therefore, does not attempt to replace the phonological process of the human listener. This is why under a number of channel and ambient noise conditions, the 800 bps system has been found to be just as robust as the 2400 bps LPC approach [6].

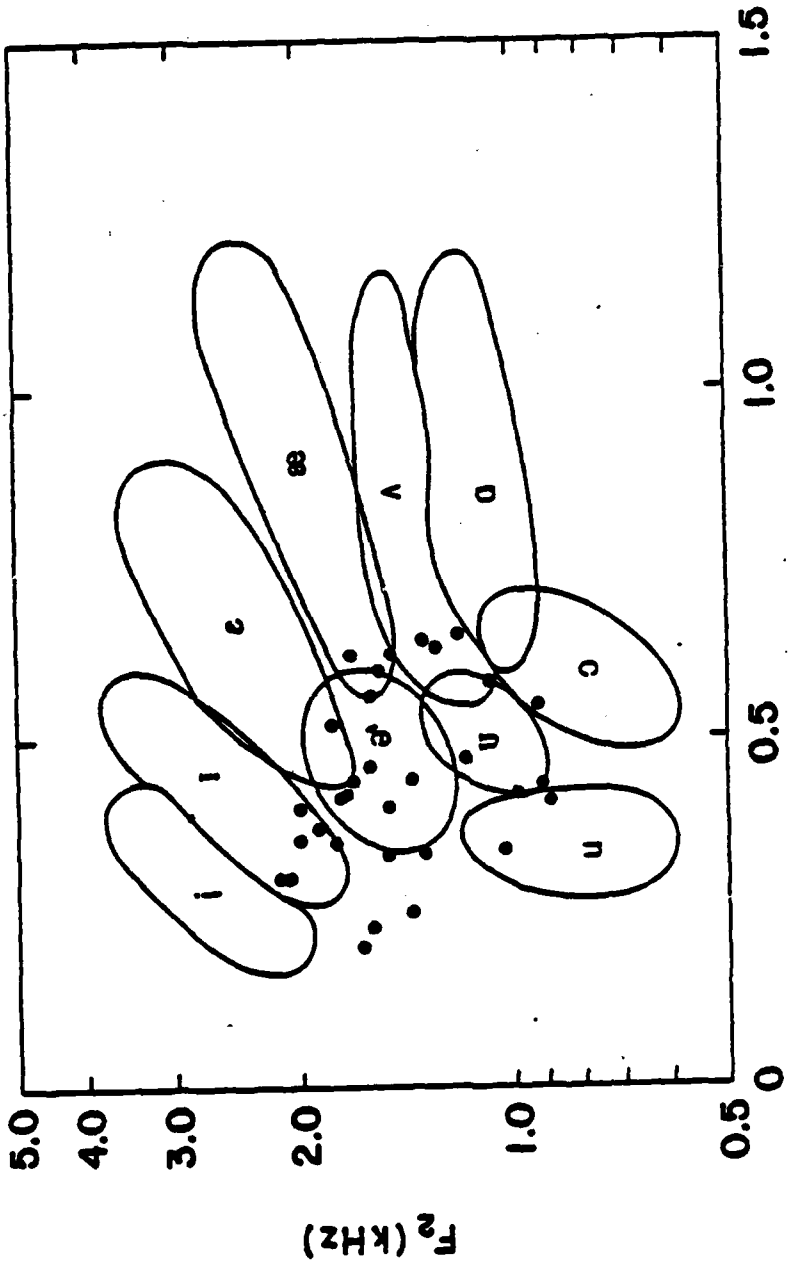


Figure 1.4 First and Second Formants of a 5-bit LPC Vector Codebook.

While vector quantization has reduced the LPC vocoder bit rate by 2/3, it has not removed all of the obvious redundancies in the code. Techniques can therefore be developed to achieve even lower rate speech compression using vector quantization.

Further reduction in the bit rate for coding the LPC coefficient vector exploits the following remaining areas of redundancy or inefficiency in the vector quantization approach to the 800 bps system.

(1) It has been shown that the quantization codebook for unvoiced frames need not be as large as that for voiced speech [6]. Therefore, if variable rate transmission is applied, the present bit rate can be further reduced.

(2) Natural pauses exist even within a very short duration (less than 1 second) of speech. Such pauses can be identified and encoded with the gain and voicing codes, and no LPC vector code needs to be transmitted. With variable rate coding, significant bit reduction can be obtained.

(3) In the present vector quantization approach, frame to frame redundancy in the LPC model has not been exploited. Two techniques which can reduce the bit rate by a factor of 2 to 4 are:

(i) frame predictive coding and

(ii) matrix quantization.

A matrix here refers to a sequence of time consecutive LPC coefficient vectors.

It is relatively straightforward to reduce the first two types of redundancy in the vector code. However the achievable bit rate reduction is minor compared to that from reducing the frame-to-frame redundancy of the LPC vector code. It is also interesting to note that all of these techniques lead to either variable rate transmission or synchronous transmission with increased delay. Such a consequence is inevitable because speech is a variable rate information source, so any efficient coding technique must also be variable rate.

1.2 Report Outline

Two techniques for reducing the time redundancy of the LPC code have been studied. The frame predictive vector quantization approach is discussed in section 2.0. The matrix quantization approach is discussed in 3.0. With these compression techniques, the bit rate for encoding the LPC spectral model is reduced by 50% or more.

To preserve the prosodics and natural quality of the speech, excitation parameters for the synthesizer, namely pitch, gain, and voicing must also be transmitted. The fake process trellis coding technique has been studied for very low rate compression of these parameters. The theory and results are discussed in 4.0.

Formal subjective evaluation of the frame predictive

vector quantization and matrix quantization techniques have been conducted to verify the intelligibility of the speech output of these systems. The results are discussed in 5.0. Conclusions and recommendations for efficient speech coding based on the techniques developed in this study are discussed in 6.0.

2.0 FRAME PREDICTIVE LPC VECTOR QUANTIZATION

It has been shown that vector quantization is near optimal in its distortion performance for encoding LPC coefficients. The vector code is also nearly optimal in its discrete memoryless source code entropy [10]. Higher code efficiency is thus attainable only by exploiting frame to frame redundancy in the LPC coefficient vectors.

It is well known that speech is a variable rate information source and that some phonetically stationary sounds may be sustained for over a hundred msec. Several techniques, such as frame repeat or frame fill coding, have been proposed [4,16] to take advantage of this fact. It has been reported that significant bit rate reduction from the standard memoryless design, sometimes as much as 50% for scalar quantization, can be achieved. A frame repeat coding system for vector quantization is developed to study its effectiveness for bit rate reduction.

For efficient coding, significant bit reduction can also be made on the coding of pitch and gain information. This topic will be discussed in Section 4.0.

2.1 Frame Repeat Coding Design

A vector quantizer maps each input LPC model vector x onto an index j. The index designates the codeword y_j in a codebook $C=\{y_i\}$ which best matches the input vector. In short, $\hat{x}=y_j$. The frame repeat operation for such a vector quantizer can be described as follows. Let x(n), x(n-1), $\hat{x}(n)$, and $\hat{x}(n-1)$ denote the current input vector, the previous input vector, the current quantizer output vector (to be chosen), and the previous quantizer output vector, respectively. The quantizer output vector is always one of the prestored codebook entries. Let the quantizer output for x(n-1) be denoted by $\hat{x}(n-1)=y_j$. With a given distortion threshold t, frame repetition occurs (i.e. x(n) is mapped into j, so that $\hat{x}(n) = \hat{x}(n-1) = y_j$) when either one (or both) of the following conditions is met:

i) $d[x(n), x(n-1)] < t$;

ii) $d[x(n), y_j] = \min d[x(n), y_i]$

where d(.,.) denotes the distortion measure for the vector quantizer. When frame repetition is performed, no codeword index has to be sent for the new frame, thereby reducing the overall bit rate. However, a 1-bit/frame repetition flag (repeat/no-repeat) must be transmitted.

2.2 Experimental Results

A test speech sample of 400 frames was processed to investigate the performance of the above frame repeat vector quantizer. The likelihood ratio measure [14,17] with full search vector quantization [9] was employed. To investigate the relationship between performance and codebook size, three sets of codebooks, 6-bit, 8-bit, and 10-bit in size, were tested. Each set consists of two codebooks, one for voiced speech and one for unvoiced speech. For each codebook size, threshold (t) values ranging from 0.0 to 0.8 in 0.1 increments were tested. It is important to note that these codebooks were obtained using a full scale multi-speaker training speech sequence (consisting of over thirty thousand LPC vectors). The average distortion would be lower than those reported here if the codebook is specifically trained for a single speaker.

The percentage of frames repeated plotted against the repetition threshold for three codebook sizes is shown in Fig. 2.1. As expected, all three plots are monotonically increasing functions of the threshold t. It is observed that in terms of incremental effect on the repetition percentage, threshold values of $t \geq 0.4$ is desirable for all codebook sizes. For smaller threshold values ($t < 0.4$), the percentage of repetition is higher for smaller codebooks, which is primarily due to repetition condition (ii). That

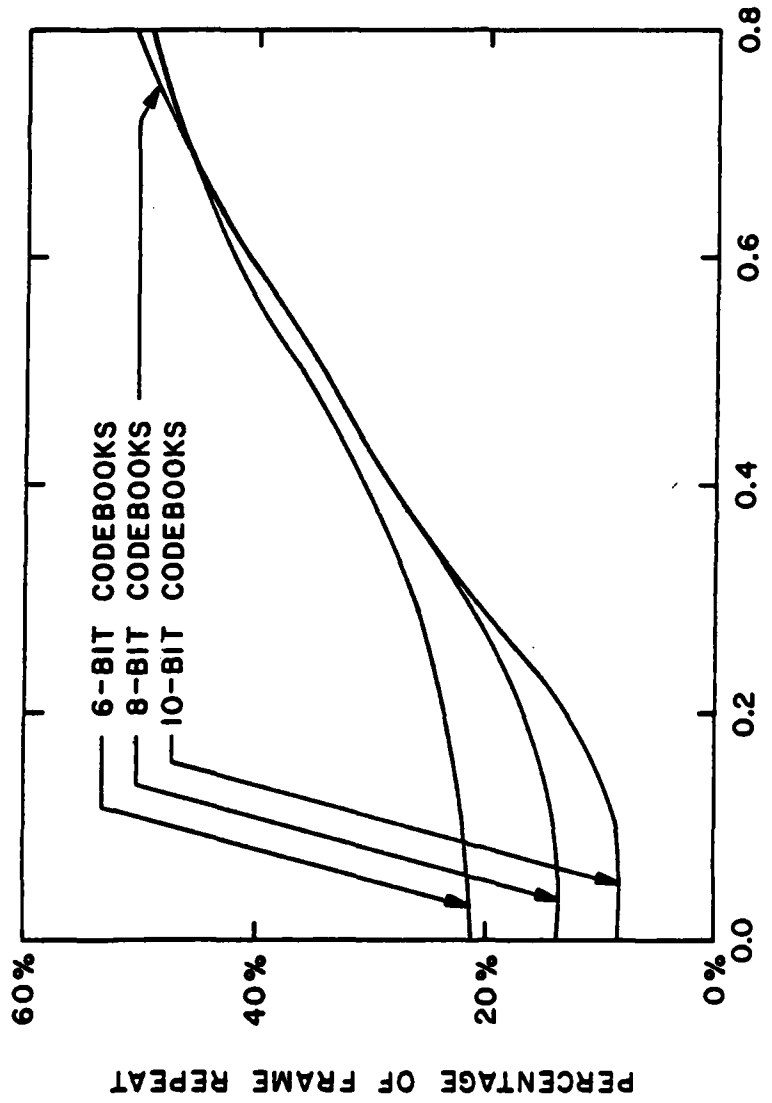


Figure 2.1 Number of frames repeated (percentage) as a function of the distortion threshold.

is, even though the spectral change for the input vector from, for example, frame n-1 to frame n has exceeded the threshold, no better matched codeword can be found in the codebook, so the same codeword is assigned to frame n. In other words, the codebook resolution is not fine enough to capture the change.

In Fig. 2.2, the average distortion for all three codebook sizes is plotted against the average number of bits (inclusive of the repetition flag.) The average distortion as expected, is a monotonically decreasing function of the average number of bits per frame for a fixed codebook size.

The trade-off between codebook size and the threshold for repetition is analogous to that between frequency and time resolution. To achieve a given bit rate, a large codebook will require a higher percentage of frame repetition and thus a higher repeat threshold. The output of such a quantizer will have accurate spectral features when a new vector code is sent, but due to more frequent repetitions time resolution will be compromised. The reverse is true for a smaller codebook, where fewer frames are repeated, but the quantizer output spectrum is not as well matched to the input even when a vector code is sent.

From the three curves in Fig. 2.2, it is seen that for a fixed codebook size, the incremental performance gain (i.e. drop in average distortion) decreases with the bit rate. To achieve an average distortion of about 0.3 or

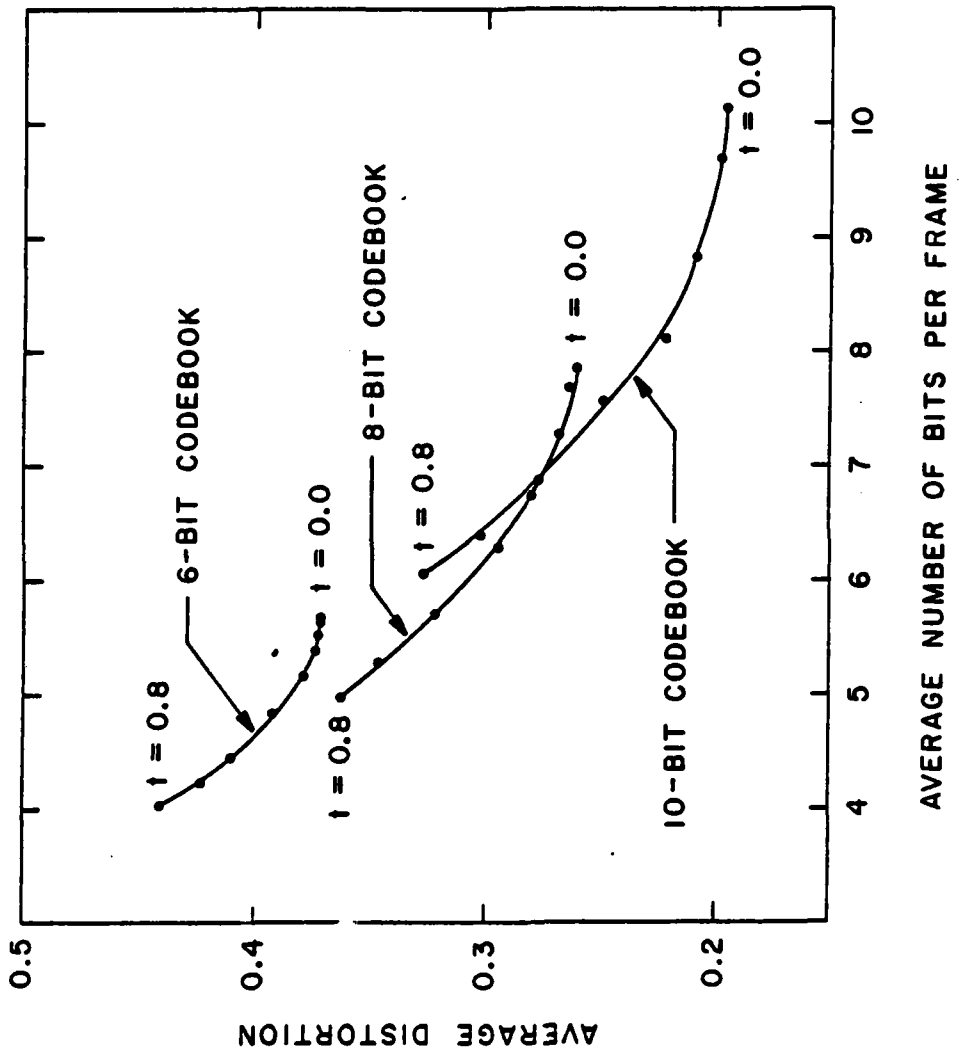


Figure 2.2 Average distortion as a function of the average number of bits per frame.

less, a 6-bit codebook is simply not adequate; an 8-bit or 10-bit codebook would have to be used. Intersection of the curves for the 8-bit and 10-bit codebooks occur at the average rate of 7 bits/vector; for a lower bit rate an 8-bit codebook yields better performance (i.e. a lower average distortion), but for a higher bit rate, the 10-bit codebook performs better.

In achieving an average rate of about 250 bps for the LPC vector code (i.e. 5.6 bits/frame at 44.4 frames/sec), several configurations appear to be possible. These configurations are listed in Table 2.1, together with their expected distortion performance and average bit rate for the test speech sample. The average bit rate is computed based upon a standard frame rate of 44.44 frames/sec.

A demonstration of frame repeat coding is included in the audio tape accompanying this report. The speech sample is not the same as that used to obtain the results of Table 2.1. For the demonstration speech sample, a likelihood ratio threshold value of 0.6 is used, yielding an average bit rate of 228 bps for encoding the LPC coefficients. The output speech quality is informally judged to be very close to the 800 bps vector quantized LPC synthesis [6]. See Appendix A for the tape list.

Based on the results presented above, it is concluded that vector quantization with frame predictive coding can achieve bit rates below 250 bps. A formal subjective evaluation of the system will be presented in Section 5.0.

Codebook Size	Threshold	Average Distortion Performance	Average Bit Rate (bps)
6	0.3	0.374	238
6	0.4	0.38	230
8	0.6	0.32	253
8	0.7	0.348	234
8	0.8	0.363	220

Table 2.1 Several Configurations for Quantizing Spectral Coefficients with Frame Repeat Vector Quantization.

3.0 LPC MATRIX QUANTIZATION

In the switched source model for speech production presented in Section 1.0 (Fig. 1.2), the switch changes state randomly every T sec.. In a more realistic model of speech production, the switch must not change from any state to the next arbitrarily. Given that only a small set of target phonemes are intended by the speaker and that the articulatory transition from one phoneme to the next must follow a certain path, it is postulated that a finite number of transition vector sequences are adequate to construct all of the speech sounds produced.

A natural extension to the vector quantization technique is, therefore, to assemble the LPC vectors into NxM matrices $\underline{X}(n)$, where

$\underline{X}(n) = (x(n-N+1), x(n-N+2)..., x(n))$,

$x(i)$ = An Mx1 vector of LPC coefficients for frame i,

(M=order of the LPC filter),

and N = number of vectors in the matrix,

so that the time duration of the matrix = NT sec. Each matrix is treated as the basic unit for quantization.

A block diagram of the matrix quantization system is shown in Fig. 3.1. There are three key components in the design. An analysis and segmentation algorithm transforms the speech signal into LPC matrices. A data base of actual

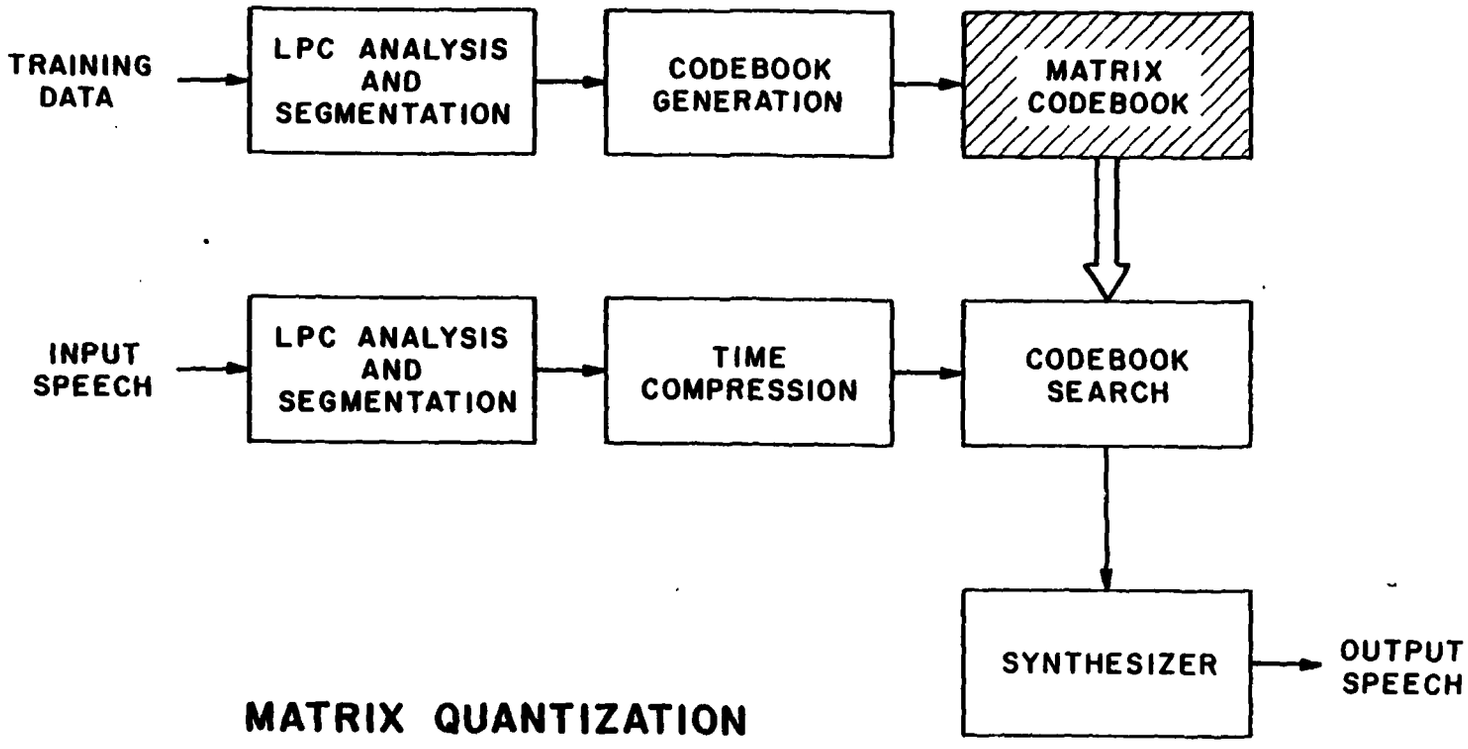


Figure 3.1 A Matrix Quantization Speech Coding System.

the speech signal into LPC matrices. A data base of actual speech data is segmented and transformed into a large set of matrices obtained from real speech, which is labeled as training data in Fig. 3.1. The codebook consists of codeword matrices which represent all possible naturally occurring transition segments in speech. In actual operation, the input speech is analyzed and segmented into matrices with the same algorithm used during codebook generation. Codebook search is then carried out to find the codeword matrix which best matches the input matrix. The matching is performed according to a well defined spectral distortion measure between two matrices. It is the index of the best match codeword that is transmitted. At the receiver, this index is used to retrieve the same codeword matrix for synthesis. The matrix is then used to synthesize a segment of speech with a standard LPC synthesizer.

In addition to the matrix code, timing information on the matrix must be transmitted. The synthesizer excitation parameters, pitch, gain, and voicing information, must also be extracted and transmitted. This section will concentrate on matrix quantization of the LPC filter information. Encoding of the pitch and gain information will be discussed in 4.0.

In developing a matrix quantizer, a segmentation algorithm must be developed to assemble the LPC vectors into matrices. Then a distortion measure must be defined for

comparing two matrices. Finally, a codebook generation procedure must be developed.

A fundamental issue that is critical to the feasibility of the matrix quantization approach is whether a codebook of reasonable size can in fact produce good quality synthesis. We postulate that for a large but limited vocabulary (500-1000 words and phrases), a codebook of 1000-5000 codewords should be adequate for producing intelligible speech.

Given that a 10-bit vector quantization codebook (1024 vector codewords) can produce very intelligible vocoder speech, the number of vector codewords representing only steady state sounds must be significantly smaller than 1024. The lower bound for this number is the total number of sustained vowels and consonants (or sustained sounds within a consonant, such as the aspiration for a plosive) in English which is below 40. If we assume M such codewords are adequate (M>40), then the total number of transitional sounds which connect one steady state sound to the next must be reasonably close to M^2 for a general vocabulary. In fact, since not all such transitions occur naturally, the lower bound is below M^2. Based on these estimates, the lower bound on the matrix quantization codebook for a general vocabulary is estimated at about 1600. For a limited vocabulary, the lower bound is estimated to be close to 1000.

Having argued that the matrix quantization approach is feasible for speech coding, such a system for coding LPC models is designed and implemented. The details are presented below.

3.1 Matrix Quantization Design

Analysis and Segmentation

While it is helpful to think of speech information in terms of phonemes, they are very difficult to segment and identify acoustically because they are not articulated independently, but are articulated in groups to form syllables, words, or phrases. It is thus easier to define speech segments according to acoustically observable events such as speech onset, speech offset, steady states, and transitions. By defining onsets, offsets, and the centers of steady states as segment boundaries, both isolated or connected speech can be segmented into transition matrices. Each matrix is made up of a sequence of vectors beginning at a speech onset point or steady state center and ending at the next steady state center or speech offset point. The duration of such matrices may vary from 50 msec to over 300 msec.

The above definition of segment boundaries may also be argued from the viewpoint that speech is a variable rate information source, and the information resides mostly

within transitions in the speech signal. Therefore, the basic units for coding should correspond to speech transitions.

To transform the speech signal into transition matrices, tenth order LPC analysis at a frame rate of 100 frames/sec is first performed. The analysis window length is 16 msec; pre-emphasis with a factor of 0.9 and Hamming windowing are applied to the signal before autocorrelation computation. The autocorrelation terms are transformed into a set of reflection coefficients with the Levinson recursion algorithm [3]. A set of excitation parameters, namely pitch, voicing, and residual energy values, is also extracted for each frame of speech. The speech signal is thus transformed into a sequence of LPC vectors and excitation parameter vectors. The next step in the process is to assemble the LPC vectors into transition matrices with a segmentation algorithm.

Segmentation is based on the discrimination between speech and non-speech signals (pauses) and between steady-state and transition speech sounds. The discrimination algorithm is based on a subset of the parameters extracted by standard LPC, namely the filter reflection coefficients, the voicing decision, and the speech rms (gain) value.

Speech/pause discrimination is primarily based on the gain and voicing features. A voiced frame is automatically

defined as speech. A maximum likelihood pattern classifier based on gain is used to discriminate pauses from speech. For such a single parameter case, the classifier reduces to a simple threshold test. If the gain value exceeds a threshold, the signal is classified as speech. The gain threshold should be adaptively adjusted so that the algorithm can operate under different noise environments. It may also be desirable to include acoustic features such as zero-crossing count, and the first one-to-two LPC coefficients into the pattern classifier. However, in our initial design, a fixed gain threshold is used. The decision made by the threshold test is then processed by a smoothing algorithm which eliminates speech or pause segments that are under 50 msec. in length. This smoothing procedure eliminates fluctuations in the speech/pause decision during transitions or due to background noise and voicing decision errors. Based on the above speech/pause classification results, decisions on speech onset and offset points are made.

Steady-state/transition classification is primarily based on a spectral variance measure defined as follows. Denote the LPC vectors extracted from the signal at every T sec by x(n), then the spectral variance at index n is given by

$$\sigma_d(n) = \frac{1}{2L} \sum_{j=n-L}^{n+L} d[x(j), x(n)] \qquad (3.1)$$

where L=3, and d[,] is the COSH distortion measure [17]. The COSH measure is defined as

$$d[A_1,A_2] = \int_{-\pi}^{\pi} \cosh[v(\theta)]-1 \frac{d\theta}{2\pi} \qquad (3.2)$$

where

$$v(\theta) = \ln [1/1\ A_1(e^{j\theta})|^2]$$

$$\cosh [v(\theta)] = \frac{e^{jv(\theta)} + e^{-jv(\theta)}}{2}$$

and $A_1(z)$, $A_2(z)$ are the linear prediction all-zero fiters in z-transform notation. Note that the arguments x(j) and x(n) in (3.1) may denote any one-to-one transformations of the LPC filter coefficients (such as reflection coefficients or the predictive filter A(z) coefficients), but the COSH measure will always be defined in terms of the all-zero predictive filter A(Z) as shown in (3.2).

Heuristics for detecting dips and valleys in the σ_d contour are applied to locate steady state sounds. The algorithm is as follows:

(1) A fixed COSH threshold value of 0.45 is set to detect strong steady state sounds. In general, two LPC filters with a COSH measure under 0.3 are perceptually indistinguishable. Over seven frames, a spectral distortion variance of $\sigma_d(n)<0.45$ indicates a highly stationary speech segment of 70 msec (for L=3) and n is situated at the center of such a segment. Frame n is thus labeled steady state. All other frames for which $\sigma_d(n)>0.45$ are tentatively labeled as transition until detected otherwise by a number

of other criteria.

(2) If $\sigma_d(n)$ consistently stays below 0.45 for over seven frames (corresponding to a steady state segment of about 130 msec), then a search for a local minimum is performed. Such a minimum very likely corresponds to a transition. The reasoning is that long (>100 msec) segments of slow spectral change often correspond to slow phonemic transitions such as those found in diphthongs and final vowels. If $\sigma_d(n)$ is detected to be a local minimum, frames n-1, n, n+1 are all labeled transition. Such a transition segment will not be eliminated by post-processing.

(3) While a steady state sound usually corresponds to a dip in the σ_d contour, there may not exist a true minimum. A "soft minimum" criterion is thus established to detect such dips. The criterion is as follows: Frame n is defined as a "soft minimum" if

$$\sigma_d(i) \geq 2\,\sigma_d(n) \quad \text{for} \quad i = n\pm2,\ n\pm3$$

$$\sigma_d(i) \geq 0.9\,\sigma_d(n) \quad \text{for} \quad i = n\pm1$$

Such a soft minimum is labeled as steady state.

(4) An abrupt drop in the spectral variance contour is detected at location n when

$$\frac{\sum_{i=n-4}^{n-1} \sigma_d(n)}{\sum_{i=n}^{n+2} \sigma_d(n)} \geq C,$$

where C is set to 4.0. Frames n, n+1, and n+2 are labeled steady state. Similarly an abrupt rise in the spectral variance contour is detected when

$$\frac{\sum_{i=n-2}^{n} \sigma_d(n)}{\sum_{i=n+1}^{n+4} \sigma_d(n)} \leq 1/C$$

Frames n-2, n-1, and n are labeled as steady state.

(5) After the decision process of (1) through (4), long transition segments that are over KT sec long are further processed to locate possible steady state segments within. In this study, T = .01 sec and K is set to 18.

A running average $m_\sigma(n)$

$$m_\sigma(n) = \frac{1}{K} \sum_{i=n-K}^{n-1} \sigma_d(i)$$

is computed, and frame n is labeled steady state if

$$\sigma_d(n) < 0.5\ M_\sigma(n)$$

After the detection of speech onset/offset and steady state center points, a segment is defined as any speech interval beginning at an onset point or a steady state center, and ending at an offset point or a steady state center. Final smoothing algorithms are then applied to alter segments that are too short (<50 msec), too long (>300

msec), or contain too many (>3) voicing transitions.

An example of the segmentation results is shown in Fig. 3.2. The onset and offset points are labeled in the bottom plot of the figure. The unlabeled vertical bars in the plot mark the end and beginning of segments at steady state centers.

Codebook Generation

A database of speech is first collected. Such a data base must contain at least several occurrences of each word in the vocabulary being considered. For continuous speech processing, each word should occur under different syntactical and contextual environments.

Standard LPC analysis is peformed at a rate of 100 frames/sec, so that each vector represents the short term spectral model for 10 msec. of speech. The segmentation algorithm described above is applied to collect the vectors into transition matrices. Codebook generation is performed using a minimax criterion.

Denote each transition matrix by b(i), and the collection forming the training database by {b(1),...,b(N)}. The first training sample b(1) forms the first initial code word w(1). A new training sample b(2) is then compared to w(1), if the spectral distance d [w(1), b(2)] is less than some threshold t, no new codeword is created. Otherwise b(2) becomes a new word, i.e. w(2) = b(2). This process is continued for i=3,..., N. At each stage, the new matrix

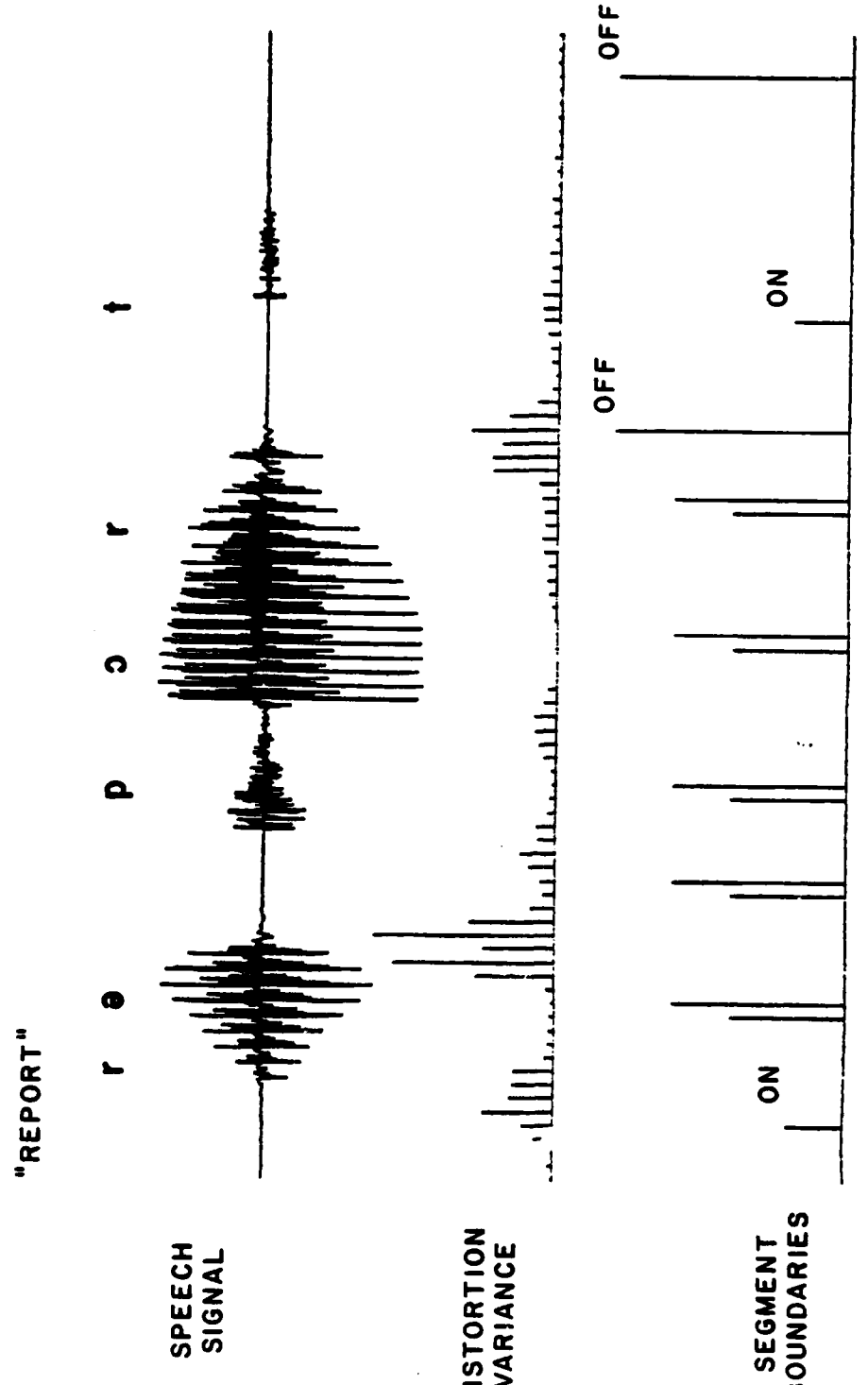


Figure 3.2 Segmentation of Speech based on the Spectral Distortion Variance Measure.

b(i) is matched to all codewords stored prior to processing it. The maximum distortion for encoding any matrix b(i) in the data base is thus less than t. By varying t, the codebook size can be controlled accordingly. Initially a large value is selected for t to prevent overflowing memory. Then t is reduced until the desired codebook size is obtained.

Time Warping

Note that b(i) and w(j) may be of different lengths. To compute the distortion d[w(j), b(i)], dynamic (non-linear) time warping is applied so that d [,] is accumulated over the optimal warping path for some prespecified continuity and range constraints. The topic of dynamic time warping is very well covered in the literature [18,19,20] and will not be discussed here. The optimal dynamic programming approach is adopted in this study. The continuity condition (slope constraint) selected is the simplified path with slope intensity P=1 as defined in [18]. No global range constraint [20] other than that implied by the continuity condition is applied. The distance measure used is the COSH measure defined in equation (3.2).

The non-linear time warping algorithm (with optimal dynamic programming) is very computationally intensive. Therefore, each input matrix is pre-compressed to a minimum length by eliminating any vector which does not vary significantly from the vector preceding it. Subjective

listening experiments have shown that a COSH threshold value of 0.4 will eliminate most redundant frames and still preserve all perceptually significant information. With this threshold, about 50% of the speech frames (at 100 frames/sec.) are eliminated, i.e. a segment is in general reduced by half, and the computation approximately by 3/4. To satisfy the time warping continuity and range constraint conditions, codewords that are too long (> twice the length of the input matrix) or too short (<1/2 of the input matrix) are not compared to the input. This further reduces the computation considerably. Other techniques such as aborting unlikely warping paths or discarding unlikely candidates before completing the optimal path search [19], may also be applied to reduce computation time.

Quantizer Simulation

In a matrix quantization speech coding system, a copy of the codebook is stored at both the transmitter and the receiver. At the transmitter, for each input matrix b(i) of LPC vectors, the code word w(j), which minimizes the time warped spectral distortion d [w(j), b(i)], is found. The code word w(j) is then assigned to b(i) and the index j is transmitted for b(i).

In addition to the matrix code index, timing information must be transmitted so that the codeword w(j) can be warped to the right length at the receiver. At a minimum, the duration of the input matrix (50 msec to 300

msec) must be transmitted. This would require 4 bits (rounding to the nearest 20 msec) or 5 bits (rounding to the nearest 10 msec) for each matrix. At an average rate of about 8 matrices/sec (with no pauses) the bit rate for duration information is 32 to 40 bits/sec.

If the timing information is to be exactly encoded, the dynamic time warping path and the pre-compression timing information must be combined to yield one of three options for each input frame: repeat the last codeword frame, advance one codeword frame, or advance two codeword frames (i.e. skip one codeword frame). Since for the continuity condition selected the skip option cannot occur successively for two frames, eight possible timing patterns are possible for every two frames (corresponding to 20 msec), requiring 3 bits for encoding. If no pauses occur, then a rate of 150 bits/sec is needed to exactly encode the time warping path. Such a high bit rate for transmitting timing information is clearly unnecessary. It is estimated that simple coding techniques can be applied to reduce timing information to a rate of 50-75 bps. The combination of matrix and timing code will then be under 150 bps.

At the receiver, the matrix code j is used to retrieve w(j). An output $\hat{b}(i)$ is obtained by time warping w(j) according to the timing information. The quantized LPC matrix b(i) is fed to the synthesizer to produce the output speech.

3.2 Experimental Results

An LPC matrix quantization system as described in 3.1 has been fully simulated in Fortran. All computation is done in floating point and no attempt is made at this phase of the study to compromise performance for speed or simplicity. The intention of this study is to verify the validity of the matrix quantization concept.

A data base of about 16 minutes of speech from a single male talker recording is used as the training data for generating the codebook. The data base includes single words (of one to many syllables), short phrases, and complete sentences that are typical of cockpit communication. The vocabulary consists of approximately 450 words. The speech is digitized at 8 KHZ, and after analysis and segmentation processing, 3478 transition matrices are obtained with an average length of 130 msec/segment (or 13 LPC vectors). For this recording, there are only about 3.6 segments per second because the recording contains long pauses between utterances.

The transition matrices are used to generate a codebook with the minimax criterion. The COSH distance measure, and an optimal dynamic time warping algorithm (with symmetric distortion, and a simplified path for P=1 [18]) are selected in computing the distance between two matrices. The minimax procedure described in 3.1 is used to generate the codebook.

The training data is collected into groups of monosyllabic words, disyllabic words, trisyllabic words, multi-syllabic words, short phrases, and sentences, so that they may be processed by the codebook generation program in that order. In Fig. 3.3 the ratio (in percentage)

$$\frac{\text{number of codewords generated}}{\text{total number of training matrices processed}}$$

is plotted against the total number of matrices processed. It is seen that as more and more training data is processed, the percentage decreases, i.e. fewer new codewords are created. It is not clear how much training data is needed before the percentage will fall below an acceptable convergence threshold (say <10%). Extrapolation of the curve in Fig. 3.3 suggests that such a threshold may never be reached. However, extrapolation on the last few points of this curve may not be justified because the acoustic characteristic of the training material changed very drastically from single words to rapidly spoken sentences. If only single words are processed (corresponding to the first five data points) the plot is almost linear. Extrapolation on this linear part of the plot suggests that no more than 5000 training matrices will be needed for the percentage to drop below 10%.

The codebook size is also inversely related to the distortion threshold value chosen. A codebook of 1185 codewords is generated from the database with a COSH

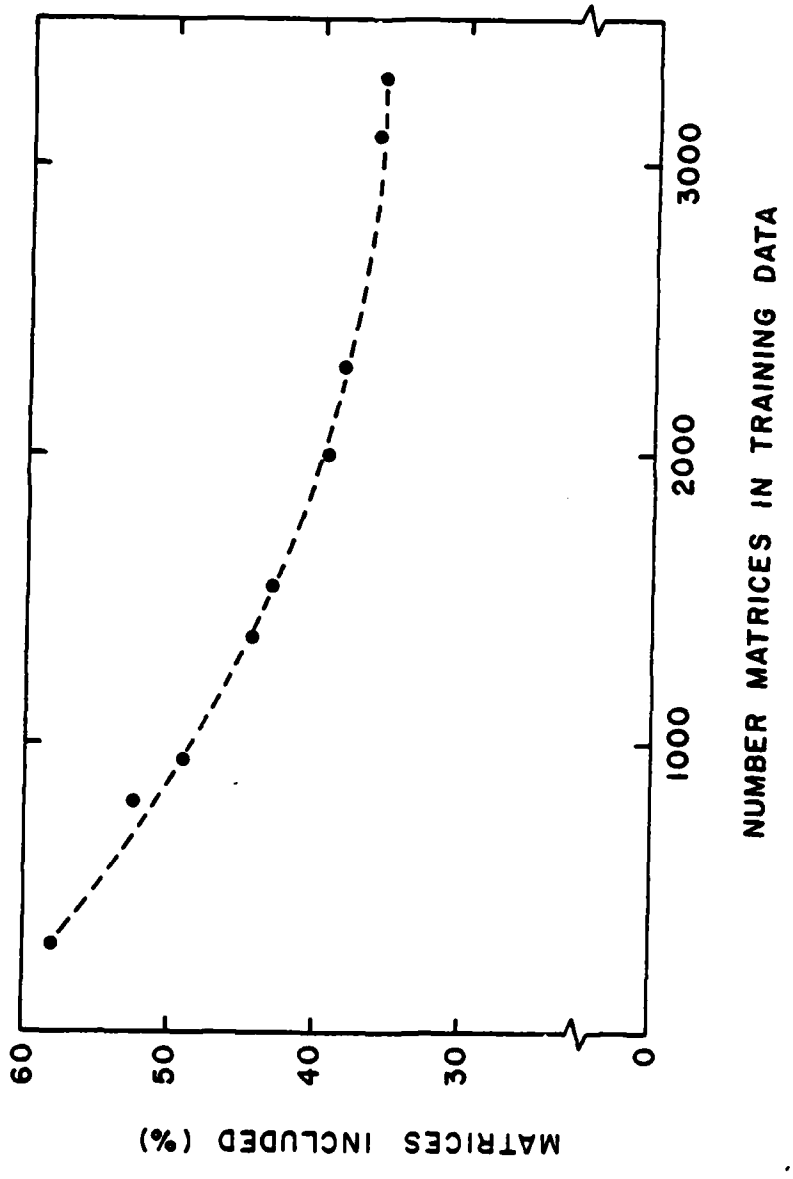


Figure 3.3 Percentage of Training Matrices Included as Codewords as a function of the Training Data Size.

threshold value of 1.0. The average length of a codeword is 7.64 LPC vectors. Given that about 10 bits is required for coding one matrix, and an average matrix is 130 msec long, the average bit rate for the matrix code is 77 bits/sec when no pauses are present.

The quantizer has also been simulated and a number of speech samples have been used to test the codebook in a preliminary experiment. The speech samples include single words, phrases, and sentences from the training data, words outside the training data (and vocabulary) by the same speaker, and speech by different speakers. The results are very encouraging in every case. The speech is very intelligible in most cases and the quality is surprisingly well preserved. One of these results is demonstrated in the audio tape accompanying this proposal. The tape content is listed in Appendix A. Using the codebook obtained in this experiment, a full DRT word list has been used to test the intelligibility of the matrix quantizer. The DRT results are presented in 5.0.

4.0 EXCITATION PARAMETER COMPRESSION

Two different approaches for encoding the LPC spectral model information, namely frame predictive vector quantization and matrix quantization, have been presented in sections 2.0 and 3.0. While most (but not all) of the phonetically important information is contained in the LPC spectral model, the vocoder excitation parameters, namely gain, pitch, and voicing, are also crucial for preserving prosodic information and natural quality as well as phonetic information. Gain and voicing contours are in fact vital for indentifying stop consonants and separating voiced and unvoiced consonants.

The excitation parameters may be treated as separate waveforms so that any coding techniques can be applied as long as the decoded parameter contours are time synchronized with the LPC vectors before synthesis. In 4.1 the theory of fake process trellis coding is discussed. Application of this coding technique for compressing the excitation parameters are presented in 4.2. It should be noted that the excitation parameter coding algorithm may be used with either frame predictive vector quantization or matrix quantization of the LPC coefficients.

4.1 Fake Process Trellis Coder

A fake process trellis coding [21] is a special case of a trellis coder. The basic structure of a trellis encoder is shown in Fig. 4.1. It consists of a search algorithm and a copy of the decoder. The decoder is a time invariant filter (denoted f in the figure) which transforms the contents of the shift register into the decoded output $\hat{x}_n$. The search algorithm determines what values for the M-ary code $\{u_n\}$ would minimize the expected distance, $Ed(x_n,\hat{x}_n)$, between the input sequence x_n and the decoded output $\hat{x}_n$. The search algorithm may be any one of many proposed algorithms such as the Viterbi algorithm [22, 23] or the M-L algorithm [24-26]. The output of the encoder is the M-ary code $\{u_n\}$. In this discussion we will consider only the binary case, ie. u_n=0 or 1. While a good search algorithm can lead to lower expected error $Ed(x_n,\hat{x}_n)$, more important perhaps is the decoder design. In fact, the uniqueness of the fake process trellis coder is in the decoder design. Detailed theoretical discussion of the system can be found in [21]. A brief discussion is provided below.

Decoder Design

In order that $\hat{x}_n$ is closely matched to x_n, it is necessary that for an independent identically distributed

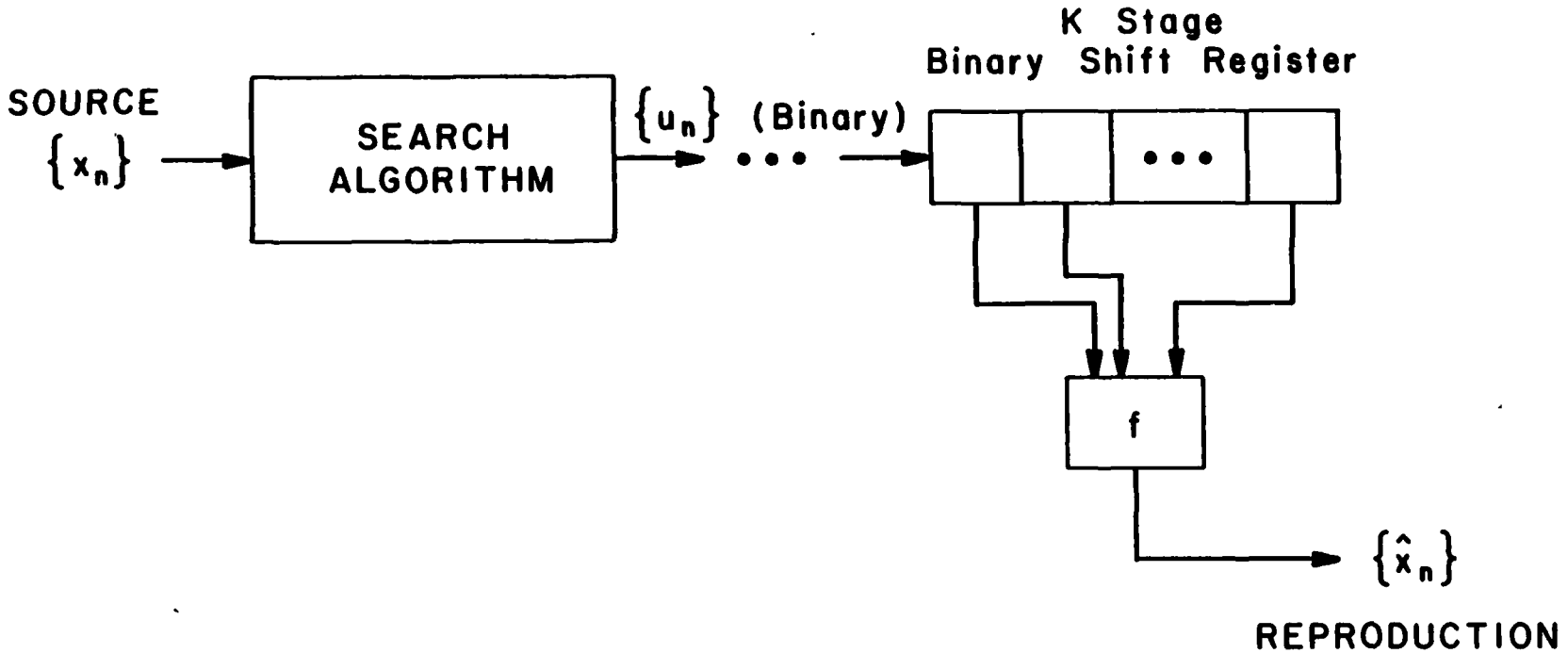


Figure 4.1 A Trellis Coding System.

(i.i.d.) input sequence to the filter f, $\{u_n\}$, the characteristics of the output $\hat{x}_n$ closely match those of x_n. Specifically their density function and spectral density, must match. The reasoning is that

$$\int_0^1 [F_x^{-1}(u) - F_{\hat{x}}^{-1}(u)]^2 \, du \leq \bar{\rho}(x,\hat{x})$$

$$\text{and } \frac{1}{2\pi}\int_{-\pi}^{\pi} [\sqrt{S_x(f)} - \sqrt{S_{\hat{x}}(f)}]^2 \, df \leq \bar{\rho}(x,\hat{x}) \tag{4.1}$$

where $\bar{\rho}(x,x)$ is the generalized Orstein distance which can be made arbitrarily close to the rate distortion function (i.e. the lowest achievable average distortion for a given rate) if a good coder is designed [21]. Although the conditions of (4.1) are only necessary and not sufficient for approaching the rate distortion bound, in practice they have been found to yield near optimal performance.

To achieve these conditions, two different operations are required. First, the content of the register (length K), $\underline{u}_n=(u_n,\ldots,u_{n-k+1})$ must be transformed into z_n such that z_n has the same cumulative distribution as x_n. This requires that $\underline{u}_n$ (in theory) be first transformed into a scalar v_n, where

$$v_n = \sum_{i=1}^{K} u_{n-i+1} 2^{-i} + 2^{-K-1} .$$

The term 2^{-K-1} is added to avoid zero values for v_n. In general, for an i.i.d. input process $\{u_n\}$,$\{v_n\}$ is correlated. However, a scrambling function g(.) can be applied to decorrelate v_n (even though $\{v_n\}$ will always be

statistically dependent). A class of functions which decorrelates v_n is given by

$$g(t) + g(t+\tfrac{1}{2}) = 1 \quad o \leq t \leq \tfrac{1}{2} \tag{4.2}$$

For any g(.) which satisfies (4.2), the $z_n = g(v_n)$ is decorrelated, i.e. the autocorrelation terms satisfy the condition

$$R_z(i) = 0 \quad \text{for all } i \neq 0 .$$

It can also be shown that if $\{u_n\}$ is a symmetric Bernoulli process, then u_n approaches uniform distribution as $K \to \infty$. If g(.) is properly selected, the uniform distribution of v_n can be preserved. Thus with proper scrambling, the output process $\{z_n\}$ is uniformly distributed and uncorrelated. To fake the distribution of $\{x_n\}$, z_n is transformed by the inverse function $F_x^{-1}(.)$ where $F_x(.)$ is the cumulative distribution of $\{x_n\}$.

So far only white signals for $\{x_n\}$ are considered. If $\{x_n\}$ is not white and has a known power spectral density $S_x(f)$, then $\{X_n\}$ can be modeled as the output of a linear time invariant filter B(z) with a white input process $\{e_n\}$. In this case, the scrambler output is transformed by $F_e^{-1}(.)$ where $F_e(\cdot)$ is the cumulative distribution of the innovation process $\{e_n\}$. The output is then filtered by B(z) to produce a fake process $\{\hat{x}_n\}$ which fakes both the probability density function and power spectral density of $\{x_n\}$. The block diagram of a fake process trellis decoder for a correlated process is illustrated in Fig. 4.2.

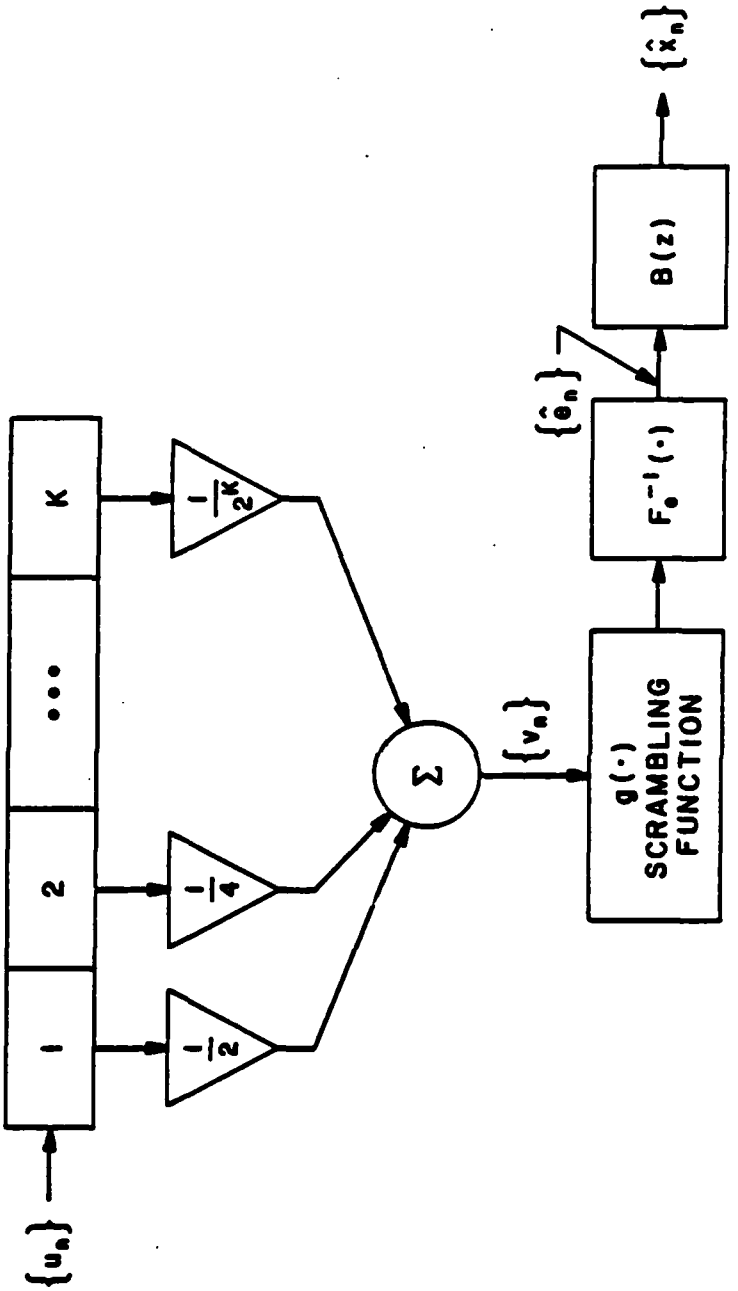


Figure 4.2 A Fake Process Decoder for a Correlated Process.

Encoder Design

The encoder consists of a copy of the fake process decoder and a search algorithm which selects a sequence $\{u_n\}$ which produces a good estimate $\{\hat{x}_n\}$ of the input sequence $\{x_n\}$. The search algorithm considered in this study is that of Look-Ahead-Delta-Modulation (LADM) [25, 26]. Given a decoder, the LADM search algorithm is determined only by a parameter M, the search depth. In LADM, the state of the shift register will be advanced one step to the next optimal state after each search of 2^M possibilities. The algorithm is illustrated in Fig. 4.3 for a 3-stage shift register (K=3 in Fig. 4.1) and a search depth of M=4. Suppose the shift register is at state (01) at the moment n; that is, $\underline{u} = [u_n\ 0\ 1]$ and we are to choose 0 or 1 for u_n. Suppose that after LADM search, the path A-B-C-D is chosen because along this path the corresponding decoder outputs, $\hat{x}_{n+1},\ldots,\hat{x}_{n+4}$, have the minimum distortion from $x_{n+1},\ldots,x_{n+4}$ among all 16 paths. Then, the encoder symbol for time n is 0 and the shift register advances to the state (10). The LADM search is equivalent to the M-L algorithm [24] in the case M=L, and there is only one encoder symbol output after every search of 2^M possibilities.

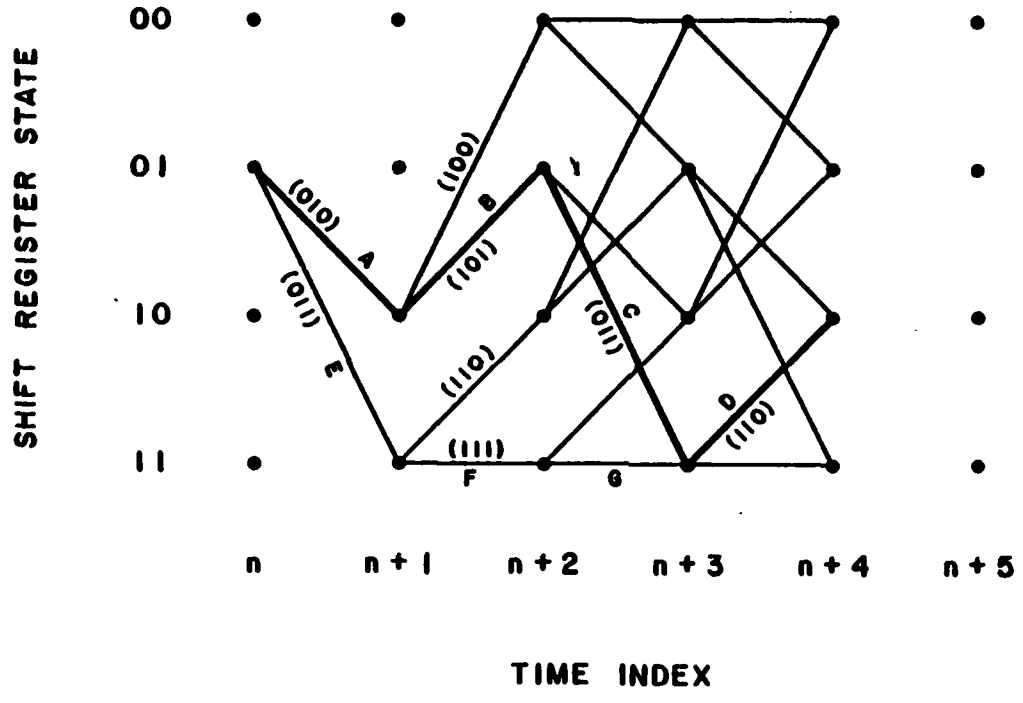


Figure 4.3 Illustration of a LADM Search with a Shift Register Length of 3 and a Search Depth of 4.

4.2 Experimental Results

A trellis coder is implemented to encode the gain parameter at 1 bit/sample. The squared error $(\hat{x}_n - x_n)^2$ is defined as the distortion $d(x_n, \hat{x}_n)$. No linear filtering operations are incorporated in the decoder in this experiment. Its cumulative probability density is quite similar to the output of the linear product operation $\{v_n\}$. Assuming $\{u_n\}$ is i.i.d., the autocorrelation of $\{v_n\}$ is given by [21]

$$R_v(i) = \begin{cases} \sigma_v^2 \frac{K-|i|}{K} & |i| \leq K \\ 0 & |i| > K \end{cases}$$

where σ_v^2 is the variance of $\{v_n\}$. The autocorrelation terms of the gain parameter are also found to be somewhat similar to R .

A shift register of K=5, and a LADM search depth of 5 (i.e. 32 searches per sample), are selected for the trellis coder. A speech sample of 30.4 seconds of speech is analyzed at 44.4 frames/sec. to generate 1350 frames of gain parameters. The results are illustrated in Fig. 4.4. The discrete lines correspond to the encoder input and the connected lines correspond to the decoded gain contour. The signal-to-noise ratio obtained is 10.53 dB, which is significantly better than the single sample optimal quantizer [27].

Informal listening comparison which compared LPC

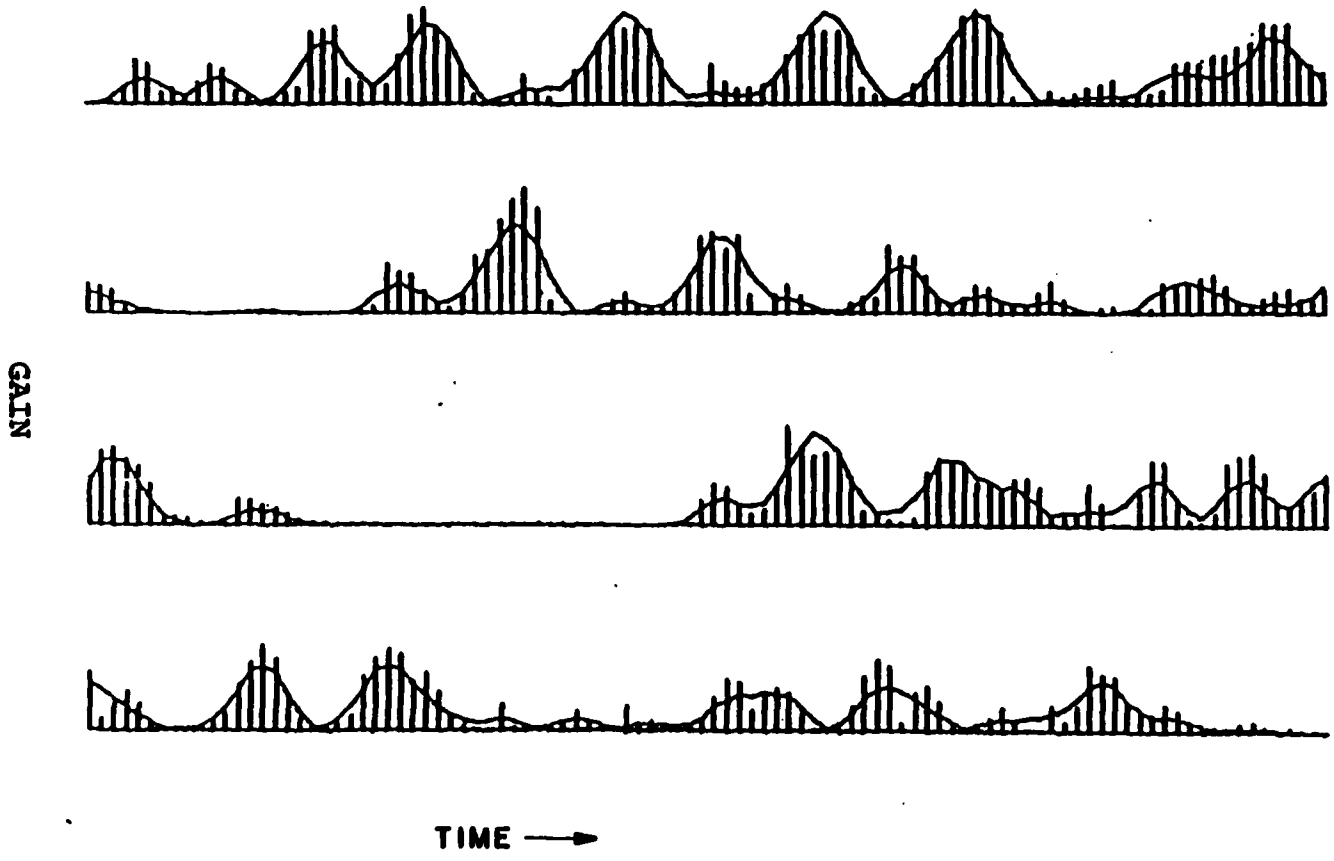


Figure 4.4 Gain Contour Before (Discrete Lines) and after (connected lines) Trellis Coding.

synthesis using gain parameters before and after trellis coding was done. The difference between the two was found to be minimal. It is thus concluded that except for very rapid changes in the gain contour, such as during stop consonants, 44.4 bit/sec trellis coding will preserve the gain contour for natural quality speech synthesis.

A trellis coder that attempts to compress both pitch and voicing to 1 bit/sample has also been studied. Instead of encoding the pitch period value, its inverse (fundamental frequency) is defined as the input. The distortion function adopted is

$$d(x_n, \hat{x}_n) = [(x_n/\hat{x}_n) - (\hat{x}_n/x_n)]^2$$

which can be expressed as

$$d(x_n+ \hat{x}_n) = [(x_n+ \hat{x}_n)/x_n\hat{x}_n]^2 \quad (x_n- \hat{x}_n)^2 \ .$$

The distortion is thus the product of a squared error term and a scaling factor. For high pitch values (x_n and $\hat{x}_n$ large), the squared error is scaled down so that the distortion is approximately normalized by the square of the pitch frequency.

To combine pitch and voicing into the same parameter contour, the unvoiced decision is imbedded in the pitch code as a zero pitch value. Such a pitch contour will typically change very slowly except at voicing transitions. To compress such a signal, more than one state of the K-length

shift register is assigned to the zero pitch value (unvoiced). Thus the inner product operation of the decoder is preceded by a voicing check. In addition, the shift register values corresponding to the unvoiced state are arranged so that the decoded output can change into a high pitch value in a single register shift. This is equivalent to a form of scrambling. The inverse probability scrambling algorithm (i.e. the g(.) and F (.) operations) are not adopted in the decoder. However, the decoded output is post processed to smooth over any output errors due to voicing transitions. Thresholding is applied to convert pitch frequencies which are much lower than the majority of values in a voiced segment. Linear low pass filtering is also applied over a voiced segment to smooth out any abrupt pitch changes.

The same speech data used for the gain compression experiment was used to test the pitch and voicing coder. The results are illustrated in Fig. 4.5. It can be seen that the voicing contour is accurately reproduced and in a majority of the cases the pitch contour is reproduced. The SNR obtained is 16.44 dB. This result is remarkable from the viewpoint that the voicing decision is in itself a 1 bit/sample code. By trellis coding, both pitch and voicing information have been compressed into a 1 bit/frame combined code. While quantitatively the trellis coding results are remarkable, the perceptual tolerance for incorrect pitch

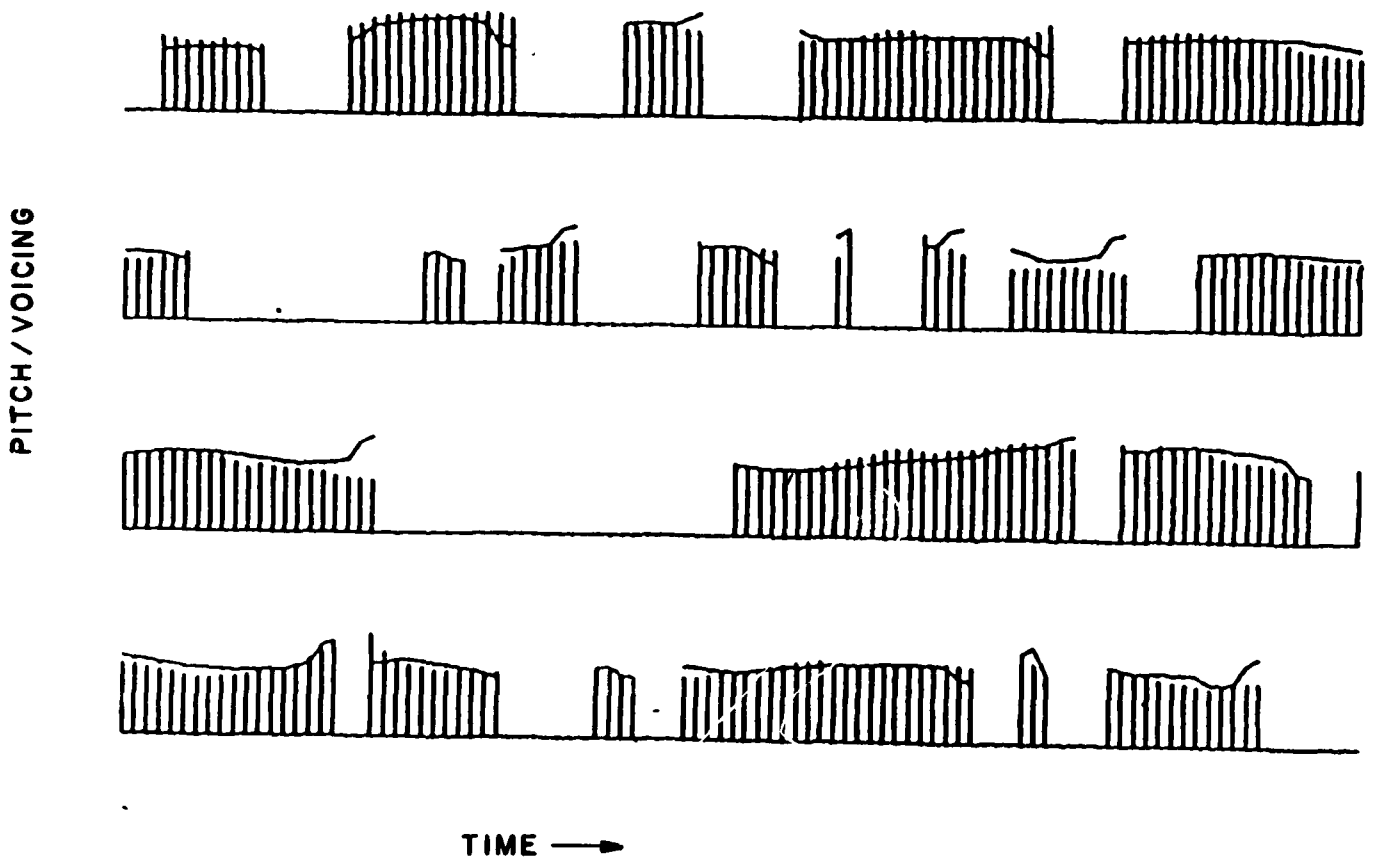


Figure 4.5 Pitch and Voicing Contours Before (Discrete Lines) and after Trellis Coding (Connected Lines).

contour is extremely low. Just the few errors in the pitch contour as seen in the figure produce a sing-song quality to the speech synthesis. It is thus concluded that pitch and voicing will have to be encoded separately for natural quality speech output. The pitch frequency can be compressed with a tree or trellis coder at a rate of 1 bit/frame or lower.

5.0 SUBJECTIVE EVALUATION

A single speaker diagnostic rhyme test (DRT) procedure [28] is used to evaluate the intelligibility of both the frame repeat vector quantization and matrix quantization coding techniques as discussed in 2.0 and 3.0. In these tests, the excitation parameters, pitch, voicing, and gain, are not quantized or coded. Only the LPC filter coefficients are quantized. The total score for the DRT results therefore reflects only the effects of quantization on the LPC filters. This allows the testing to be focused on LPC quantization, which is a much more important problem than excitation parameter quantization.

5.1 Frame Predictive LPC Vector Quantization

The basic frame rate for LPC analysis is 44.4 frames/sec. The analysis window length is 16 msec, the filter order is 10, and the pre-emphasis factor is 0.94. These analysis conditions are identical to those of the ANDVT LPC-10 system [29]. However, autocorrelation analysis preceded by Hamming windowing [3], instead of LPC-10 covariance analysis, is used. The pitch algorithm is based on a modified cepstral detection scheme [30] and the voicing algorithm is based on the cepstral peak value, gain, the

first two reflection coefficients, and zero-crossing count.

Two vector quantization codebooks are used, one for voiced and one for unvoiced speech. Even though different codebooks are used, frame repetition across a boundary is allowed, although this feature does not have significant impact on either the speech quality or bit rate. The vector quantization codebooks are generated from 30 minutes of conversational speech collected from ten talkers (3 females, 7 males, at 3 minutes/talker). The details of the codebook generation procedure are described in [10].

The DRT word list is spoken by a talker outside the training data. Furthermore, most of the DRT words do not occur in the training speech data. The test is truly an open test.

The scores for this single speaker (8 listeners) DRT are tabulated in Table 5.1. The individual feature scores are also plotted in Fig. 5.1. The total score of 78.9 compares favorably with that of a fixed frame rate (44.4 frames/sec) vector quantization (without pitch and gain coding) based on the ANDVT LPC-10. The latter system attained a score of 82.5. The score difference due to frame predictive coding is -3.6 points.

Of the six DRT features tested, four of them are primarily dependent on the spectral features, and are most directly affected by LPC filter coefficient quantization [31]. They are nasality, graveness, compactness, and

CONTRACTOR : SIGNAL TECHNOLOGY TEST CONDITION : 1ST SESSION DATE TESTED : 05/07/82

	PRESENT	S.E.*	ABSENT	S.E.*	BIAS	S.E.*	TOTAL	S.E.*
VOICING	100.0	0.00	85.9	3.69	14.1	3.69	93.0	1.84
FRICTIONAL	100.0	0.00	87.5	4.72	12.5	4.72	93.7	2.36
NONFRICTIONAL	100.0	0.00	84.4	4.57	15.6	4.57	92.2	2.29
NASALITY	93.7	3.34	84.4	5.66	9.4	6.14	89.1	3.49
GRAVE	96.9	3.12	84.4	4.57	12.5	4.72	90.6	3.12
ACUTE	90.6	6.58	84.4	9.37	6.2	11.33	87.5	5.79
SUSTENTION	65.6	5.66	73.4	6.44	-7.8	10.28	69.5	3.22
VOICED	90.6	6.58	50.0	11.57	40.6	13.31	70.3	6.66
UNVOICED	40.6	9.37	96.9	3.12	-56.2	10.30	68.7	4.72
SIBILATION	65.6	6.99	85.9	5.99	-20.3	8.82	75.8	4.79
VOICED	34.4	12.44	78.1	9.95	-43.7	13.15	56.2	9.15
UNVOICED	96.9	3.12	93.7	4.09	3.1	5.66	95.3	2.29
GRAVENESS	68.7	7.83	62.5	5.79	6.2	12.27	65.6	3.12
VOICED	71.9	11.02	68.7	11.33	3.1	20.29	70.3	4.69
UNVOICED	65.6	6.58	56.2	6.25	9.4	9.37	60.9	4.38
PLOSIVE	71.9	11.99	90.6	4.57	-18.7	14.75	81.2	5.28
NONPLOSIVE	65.6	9.37	34.4	8.10	31.2	14.75	50.0	4.72
COMPACTNESS	84.4	6.14	76.6	2.83	7.8	5.76	80.5	3.82
VOICED	78.1	8.76	96.9	3.12	-18.7	6.25	87.5	5.79
UNVOICED	90.6	4.57	56.2	4.09	34.4	6.58	73.4	2.83
SUSTAINED	93.7	4.09	84.4	4.57	9.4	4.57	89.1	3.69
INTERRUPTED	75.0	9.45	68.7	4.09	6.2	7.83	71.9	6.14
BK/MD	78.1	7.38	96.9	3.12	-18.7	7.83	87.5	4.09
BK/FR	90.6	6.58	56.2	6.25	34.4	4.57	73.4	5.99
EXPERIMENTAL **	96.9	2.05	90.6	6.14	6.2	7.09	93.7	2.89

SPEAKER LIST # CH 307A
DRT SCORE 78.9
S.E.* 1.79

8 LISTENERS, CREW (02), 192 TOTAL WORDS
1 SPEAKER(S), 192 WORDS PER SPEAKER

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X DRT TOTAL SCORE = 78.9 X
X STANDARD ERROR* = 1.79 X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

TOTAL VOICED SCORE = 71.1
TOTAL UNVOICED SCORE = 74.6

* STANDARD ERRORS BASED ON LISTENER MEANS.
** EXPERIMENTAL ITEMS ARE NOT INCLUDED IN ANY SUMMARY SCORES.

Table 5.1 DRT Results for the Frame Predictive Vector Quantization System.

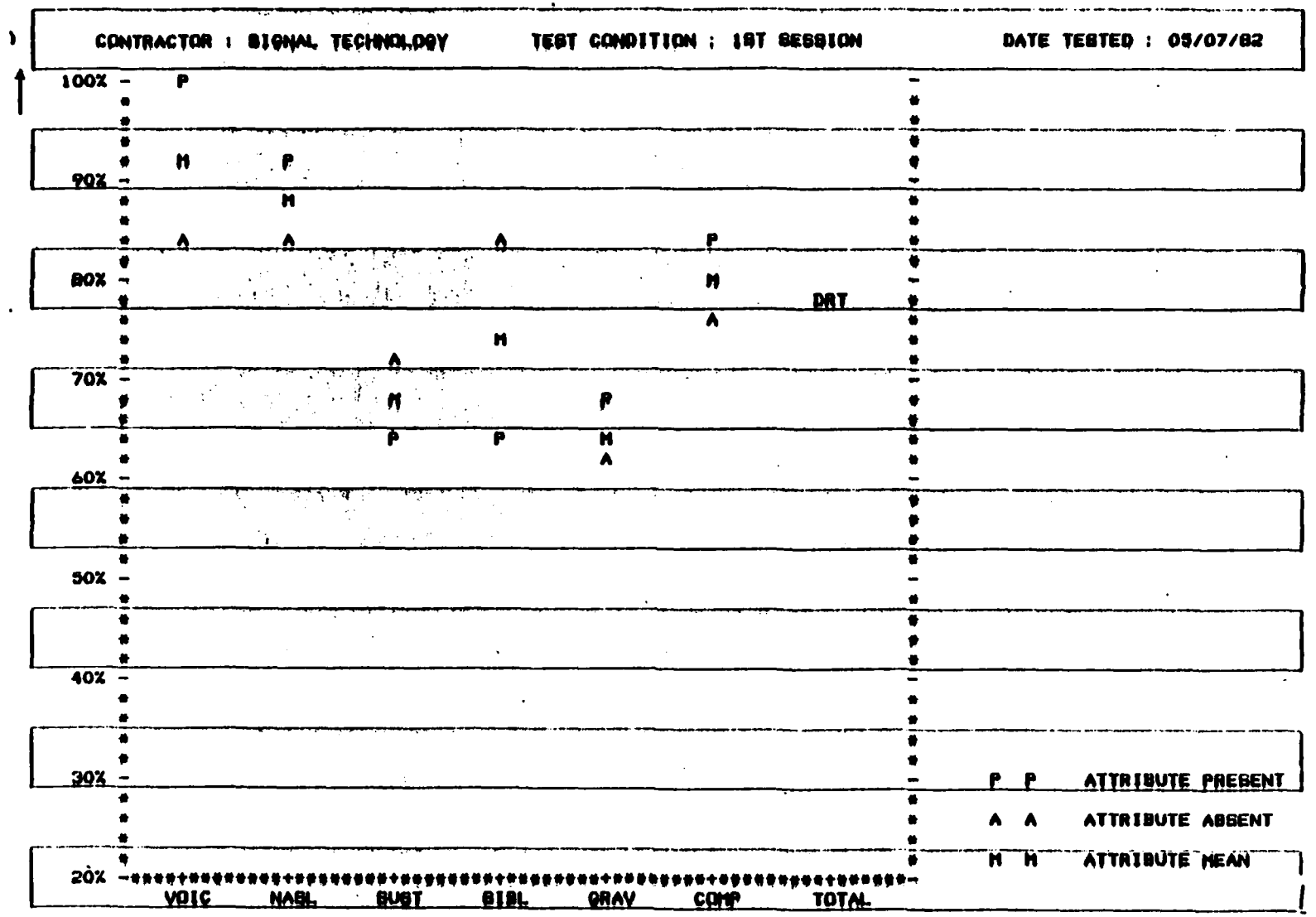


Figure 5.1 Individual DRT Feature Scores for the Frame Predictive Vector Quantization System.

sibilation. The total score for this subset of features (called the spectral scores) is 77.8. In comparison, the spectral score for the unquantized ANDVT LPC-10 system is 90.2, and that for LPC-10 with fixed rate vector quantization is 83.4 [6]. Therefore, frame predictive vector quantization leads to a drop of 5.6 in the spectral score compared to fixed frame rate vector quantization. It may be concluded that the degradation from fixed rate vector quantization to frame predictive vector quantization is limited to the spectral features as expected. While these score comparisons lead to useful interpretations, it must be cautioned that the LPC analysis and synthesis algorithms for the frame predictive experiment is not identical to LPC-10, so that some of the score differences may be due to analysis/synthesis algorithm differences.

5.2 LPC Matrix Quantization

The same analysis/synthesis system as the frame predictive system is used. The frame rate for LPC analysis is changed to 100 frames/sec. All other analysis conditions remain the same. Only the LPC filter coefficients are quantized using the matrix quantization codebook obtained in the experiment discussed in Section 3.2. The training data used for generating the matrix codebook does not contain the DRT word list. In fact, at best only a few of the DRT words

may be present in the training data vocabulary. The speaker for the DRT word list is the same. Based on the informal test discussed in 3.2, the same codebook seems to also work well for other male voices. The DRT results here are thus for a semi-open (same-speaker), unlimited vocabulary test situation.

The DRT scores are tabulated in Table 5.2. The individual feature scores are also plotted in Fig. 5.2. The DRT score of 67.7 is a great improvement over the score of 42.8 reported by Oshika [7]. It must be stressed that matrix quantization is completely automatic and very likely speaker independent. Its performance must be compared only with other fully automatic systems. The system reported by Oshika [7] achieved a DRT score of 83.5 with hand edited phonemic analysis, which is essentially a dyadic phoneme to speech synthesizer. With automatic analysis the DRT score drops to 42.8. Based on informal listening, the matrix quantization output speech is judged more natural than a dyad speech synthesis with hand edited phonemic input.

The spectral features for matrix quantization is 66, which is 11.8 points lower than the frame predictive vector quantization system. There is also a significant drop of 18.4 in the sustention score. This is due to the fact that while the gain parameter (residual energy) has not been quantized, it is affected by LPC filter quantization. If we denote the excitation signal energy by α_M, the synthesis

CONTRACTOR : SIGNAL TECHNOLOGY TEST CONDITION : 2ND SESSION DATE TESTED : 05/07/82

	PRESENT	S.E.*	ABSENT	S.E.*	BIAS	S.E.*	TOTAL	S.E.*
VOICING	85.9	4.38	92.2	3.29	-6.2	6.25	89.1	2.29
FRICTIONAL	100.0	0.00	87.5	4.72	12.5	4.72	93.7	2.36
NONFRICTIONAL	71.9	8.76	96.9	3.12	-25.0	9.45	84.4	4.57
NASALITY	79.7	8.16	65.6	6.58	14.1	10.41	72.7	5.27
GRAVE	69.6	10.50	75.0	11.57	-9.4	15.62	70.3	7.81
ACUTE	93.7	6.25	56.2	4.09	37.5	6.68	75.0	4.09
SUSTENTION	40.6	8.44	65.6	9.07	-25.0	15.85	53.1	3.74
VOICED	81.2	6.25	37.5	15.67	43.7	19.34	59.4	6.99
UNVOICED	0.0	16.37	93.7	4.09	-93.7	16.87	46.9	8.44
SIBILATION	48.4	5.51	90.6	3.92	-42.2	8.16	69.5	2.49
VOICED	21.9	12.88	90.6	4.97	-68.7	15.49	56.2	5.79
UNVOICED	75.0	4.72	90.6	4.57	-15.6	4.57	82.8	4.05
GRAVENESS	48.4	8.98	40.6	8.44	7.8	15.48	44.5	3.99
VOICED	62.5	6.68	21.9	12.88	40.6	16.99	42.2	5.76
UNVOICED	34.4	14.12	59.4	9.37	-25.0	18.90	46.9	7.38
PLOSIVE	71.9	9.95	78.1	9.95	-6.2	15.49	75.0	6.25
NONPLOSIVE	25.0	12.50	3.1	11.02	21.9	21.36	14.1	4.98
COMPACTNESS	82.8	3.29	71.9	4.57	10.9	6.86	77.3	2.02
VOICED	75.0	0.00	90.6	6.58	-15.6	6.58	82.8	3.29
UNVOICED	90.6	6.58	53.1	5.66	37.5	10.56	71.9	3.12
SUSTAINED	75.0	4.72	96.9	3.12	-21.9	5.66	85.9	2.83
INTERRUPTED	90.6	4.57	46.9	8.76	43.7	10.30	68.7	4.72
BK/MD	90.6	4.57	75.0	8.18	15.6	10.50	82.8	4.05
BK/FR	75.0	4.72	68.7	4.09	6.2	7.83	71.9	2.05
EXPERIMENTAL **	92.2	3.29	70.3	3.29	21.9	5.15	81.2	2.05

SPEAKER CH
LIST # 307A
DRT SCORE 67.7
S E * 1.90

8 LISTENERS, CREW (02), 192 TOTAL WORDS
1 SPEAKER(S), 192 WORDS PER SPEAKER

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X DRT TOTAL SCORE = 67.7 X
X STANDARD ERROR* = 1.90 X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

TOTAL VOICED SCORE = 60.2
TOTAL UNVOICED SCORE = 62.1

* STANDARD ERRORS BASED ON LISTENER MEANS.
** EXPERIMENTAL ITEMS ARE NOT INCLUDED IN ANY SUMMARY SCORES.

Table 5.2 DRT Results for the Matrix Quantization System.

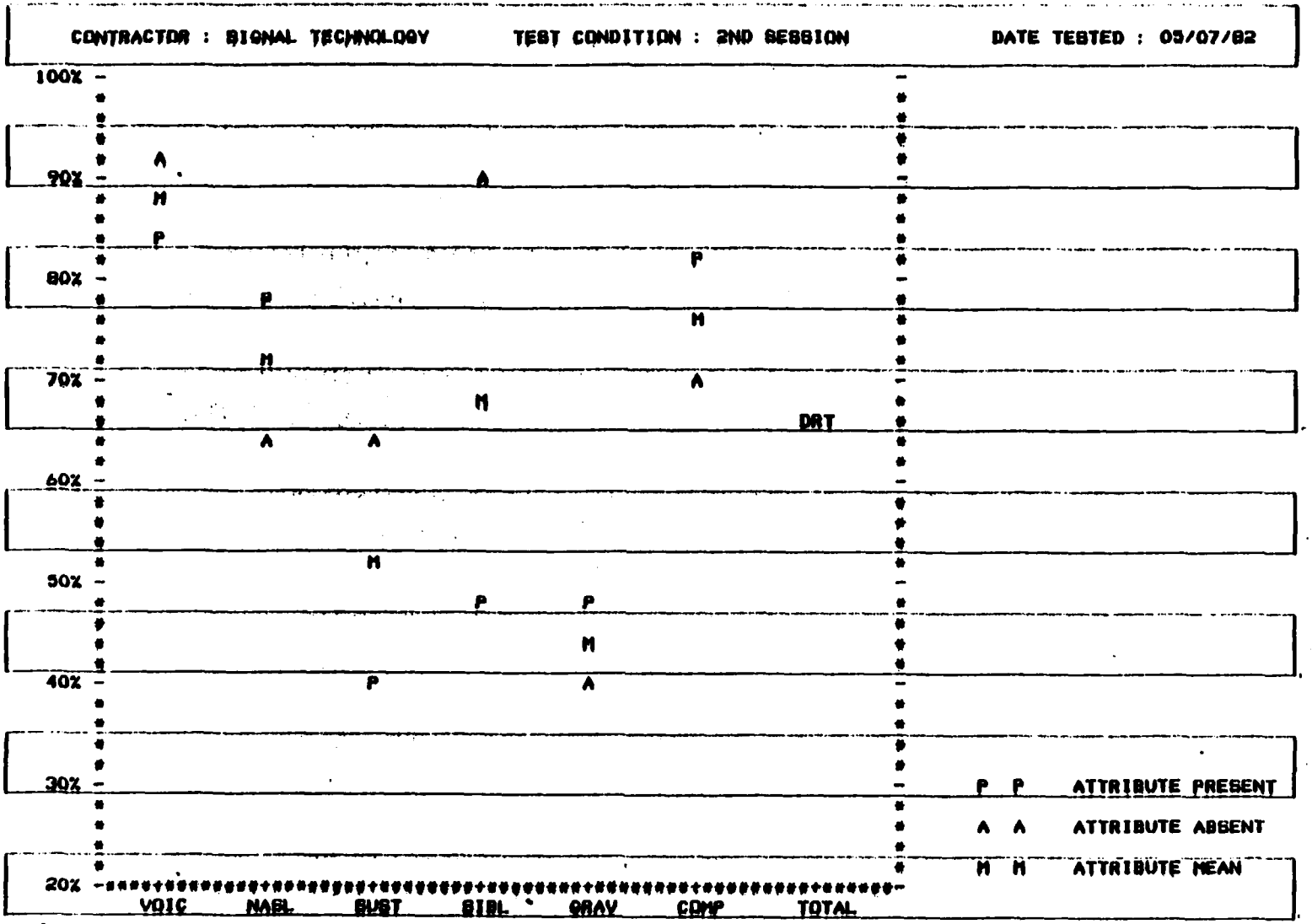


Figure 5.2 Individual DRT Feature Scores for the Matrix Quantization System.

signal energy by α_o, and the reflection coefficients of the M order LPC synthesis filter by $\{k_i\}$, then α_M and α_o are related by the equation

$$\alpha_o = \alpha_M / \prod_{i=1}^{M} (1-k_i^2)$$

It is clear from the equation that if the LPC filter (the k_i's) are changed, the same excitation gain (α_M) contour will produce a very different synthesis gain (α_o) contour.

Another aspect of the DRT score we find useful in understanding the matrix quantization system is the scores of the ten word-pairs which produced the lowest scores. The scores for these ten word-pairs are tabulated in Table 5.3. The average score for these ten words is -10. Excluding these ten words from the total DRT score would result in an overall score of 76.7, a nine point improvement. A check through the vocabulary of the training data finds that of the 20 initial consonant-vowel (CV) combinations in these DRT word-pairs, only nine may be phonemically matched to some word in the training data, and only in the word pairs "shad/chad" and "weed/reed" do both CV combinations exist (phonemically) in the training data. For the other eight word-pairs, either the same matrix codeword is used to quantize the minimally distinct CV pairs or some other poorly matched matrix codeword is introduced.

It is clear from the results above that the intelligibility of the matrix quantization system would

CONTRACTOR : SIGNAL TECHNOLOGY TEST CONDITION : 2ND SESSION DATE TESTED : 09/07/82

FOR THE PURPOSES OF FURTHER RESEARCH DESIGNED TO IMPROVE YOUR SYSTEM OR DEVICE, YOU WILL FIND IT ADVANTAGEOUS TO GIVE SPECIAL ATTENTION TO THE DISTINGUISHABILITY OF THE FOLLOWING WORD PAIRS :

	WORD PAIRS		P(C)
54	FORE/THOR		-75.0
40	MET/NET		-25.0
102	SOLE/THOLE		-12.5
67	ZEE/THEE		-12.5
103	FIN/THIN	**	0.0
101	VON/BON	**	0.0
34	COOP/POOP		0.0
19	WEED/REED		0.0
100	SHAD/CHAD		12.5
97	YALL/WALL		12.5

** THE CONTRASTS: FAD/THAD, FIN/THIN, FOUGHT/THOUGHT, VON/BON, VOX/BOX, VEE/BEE, VILL/BILL, VAULT/FAULT ARE GENERALLY AMONG THE MOST DIFFICULT TO DISTINGUISH. THEIR PRESENCE ON THE FOREGOING LIST DOES NOT, THEREFORE, REFLECT UNIQUELY UPON THE PERFORMANCE OF YOUR SYSTEM OR DEVICE.

Table 5.3 Scores for the most difficult word-pairs processed by the Matrix Quantization System.

improve if a larger training data base is used, so that all commonly occuring transition sounds are included in the codebook. More specifically, if the DRT word list is included in the training data vocabulary, better scores will be obtained.

Overall, the matrix quantization technique has been found to be highly promising for very low rate (efficient) coding of speech for large to unlimited vocabulary, isolated word or continuous speech input.

6.0 CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

Two different techniques for very low rate compression of the LPC spectral model have been developed and tested. With frame predictive vector quantization, the average bit rate for the LPC model can be reduced to about 230 bps. The DRT score with such a vector quantization approach is 78.9. Such an approach is fully implementable in real time with existing VLSI signal processors and is speaker and vocabulary independent. Speech quality may be better for a limited vocabulary or a single speaker. Also by using a better tuned analysis/synthesis system, the DRT score is expected to improve by 5 points [6]. The newly advanced matrix quantization technique is capable of reducing the bit rate for the LPC model to under 150 bps. The DRT score for the matrix quantized LPC models is 67.7. This DRT score is a considerable improvement over past results for automatic and unlimited vocabulary systems at a similar bit rate. Significantly better scores, estimated to be about 76.7, would be obtained for a limited vocabulary, which is estimated to be about 76.7.

A very efficient fake process trellis coding approach to compressing the vocoder excitation parameters (i.e.

gain, pitch, and voicing) has also been implemented. The results indicate that the gain parameter can be compressed to under 50 bps for a quantization SNR of 10.53 dB. While pitch and voicing combined can be compressed to under 50 bps with a quantization SNR of 16.44 dB, the perceived quality is unnatural. By compressing voicing and pitch parameters separately, it is expected that a bit rate of about 75 bps can be attained. The fake process trellis coding approach treats the excitation parameters as totally independent waveforms. It can thus be combined with any compression scheme for coding the LPC model filter. Combined with frame predictive vector quantization, an overall bit rate of under 400 bps is achieved.

While the excitation parameters may be coded completely independent of the LPC model, there is strong correlation between them for the matrix quantization approach. Recent results in a study of a similar system [32] suggest that the voicing parameter is highly correlated to the LPC matrix. Thus by combining the LPC matrix code (<150 bps) with a trellis coder for the gain (<50 bps) and pitch (<25 bps), an overall bit rate close to 200 bps can be achieved.

6.2 Recommendations

While the results of this study have validated the capability of the vector quantization and matrix

quantization to reduce the bit rate of an LPC vocoder to 400 bps and 200 bps respectively, further research in the following areas are recommended.

(1) Frame Predictive LPC Vector Quantization

One drawback to the frame predictive approach is that the bit rate is variable. For fixed rate transmission, buffering is needed, leading to significant time delay. A trade-off study between bit rate and delay is needed.

The vector quantization system developed in this study is for a general population and unlimited vocabulary. Quality improvement and/or bit rate reduction by tuning the system to a limited vocabulary and/or a specific speaker (or a small group of speakers) should be studied. It is clear from the design of the vector quantization system that adaptive training (i.e. the vector quantization codebook is automatically trained to the speaker's voice while in use) can also be implemented.

(2) Matrix Quantization

Overall, the goal of the present study on matrix quantization is to demonstrate the validity of its underlying concepts. The emphasis was not to simplify the algorithms for real time implementation nor to fine tune it for actual use. Considerable research thus remains to be done before the algorithms can be ready for real-time

The quality of the matrix quantization can be improved with a larger training data base and proper smoothing between matrices. The effects of limiting the vocabulary manner of speech input must also be studied. Improvements to the segmentation algorithm should also be studied.

The computation of the system can be reduced dramatically through judicious simplification of the time warping or codebook search algorithms. This will allow cost effective real time implementation of the system in the near future (before 1985).

The possibility of imbedding the voicing code in the matrix should be studied. This could lead to a bit rate reduction of 50 bps. The matrix quantization approach also results in a variable rate code. The time delay and buffering requirements must also be studied.

(3) Pitch and Gain Coding

A trellis coder for just the pitch parameter needs to be developed. Other approaches to pitch and gain coding, such as block coding with syllabic update may be more effective and should be studied.

(4) Integration and Evaluation:

The LPC vector and matrix quantization coders will have to be fully integrated with the excitation parameter coders. A study on the trade-offs in bit rate, quality, and

complexity should then be performed.

(5) Acoustic and Channel Noise Effects

The vector code has been found to perform relatively well with channel error rates of 1% to 2% [6], and is also robust in environments with mild to medium levels of noise [6]. However, the frame predictive vector and matrix coding systems will be slightly more vulnerable to acoustic noise and transmission errors due to the lower redundancy of the code. A study on their performance under different acoustic and channel noise environments is recommended.

Appendix A

Demonstration Tape List

Set 1: Frame Repeat LPC Vector Coding Male Speaker,

"Rainbow Passage"

Codebook is for general population, not trained to the speaker.

Only LPC vectors are quantized, pitch and gain are not quantized.

Quantized Synthesis	twice
Unquantized LPC (~ 2400 bps)	twice
Original 8 KHZ PCM (96 Kbps)	twice

Set 2: Matrix Coding

Male Speaker, Cockpit Communication

Codebook is trained to the speaker

Only LPC vectors are quantized

Quantized Synthesis	twice
Unquantized Synthesis (~ 2400 bps)	twice
Original 8 KHZ PCM (96 Kbps)	twice

References

1. B. S. Atal and S. L. Hanauer, Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, JASA Vol. 50, pp. 637-655, April 1971.

2. I. Itakura and S. Saito, Analysis Synthesis Telephony Based Upon the Maximum Likelihood Method, Reports of 6th Int. Cong. Acoust., Ed. Y. Kohasi, Paper C-5-5, pp. 17-20, 1968.

3. J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, Springer Verlag, 1976.

4. P. E. Blankenship and M. L. Malpass, Frame Fill Techniques for Reducing Vocoder Data Rate, Technical Report 556, Lincoln Lab, February 1981.

5. G. S. Kang and D. C. Coulter, 600 BPS Voice Digitizer (Linear Predictive Formant Vocoder), Naval Research Lab. Report 8043, Nov. 1976.

6. D. Y. Wong, B. H. Juang, and A. H. Gray, Jr., An 800 bps Vector Quantization LPC Vocoder, to be published by IEEE Trans. ASSP.

7. B. T. Oshika, FACP Speech Recognition/Transmission System, System Development Corp., RADC-TR-78-193, August 1978. (A060115)

8. R. Schwartz et al., A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model, Conf. Rec. of 1980 ICASSP, pp. 32-35, April 1980.

9. A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, Speech Coding Based on Vector Quantization, IEEE Trans. ASSP, pp. 562-574, October 1980.

10. B. H. Juang, D. Y. Wong, and A. H. Gray, Distortion Performance of Vector Quantization for LPC Voice Coding, to be published by IEEE Trans. ASSP.

11. D. B. Pisoni and J. R. Sawusch, Some Stages of Processing in Speech Perception, Structure and Process in Speech Perception, Ed. A. Cohen and S. G. Nooteboom, Springer Verlag, 1976.

12. L. O. Davisson and R. M. Gray, Ed., Data Compression, Dowden, Hutchingson and Ross, Inc., 1976.

13. T. Berger, Rate Distortion Theory, Englewood Cliffs, NJ, Prentice Hall, 1971.

14. R. M. Gray, A. Buzo, A. H. Gray, Jr., and J. D. Markel, Distortion Measures for Speech Processing, IEEE Trans., Vol. ASSP-28, August 1980.

15. G. E. Peterson and H. L. Barney, Control Methods Used in a Study of Vowels, JASA, Vol. 24, pp. 175-184, March 1952.

16. J. Makhoul, Speech Compression Research at BBN, BBN Report No. 2976, December 1974.

17. A. H. Gray, Jr. and J. D. Markel, Distance Measures for Speech Processing, IEEE Trans. ASSP Vol. ASSP-24, pp. 380-391, 1976.

18. H. Sakoe and S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Trans. Vol. ASSP-26, pp. 43-49, Feb. 1978.

19. F. Itakura, Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. Vol. ASSP-23, pp. 67-72, February 1975.

20. C. Myers, L. R. Rabiner, and A. E. Rosenberg, Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition, IEEE Trans. ASSP, Vol. ASSP-28, pp. 622-635, December 1980.

21. Y. Linde and R. M. Gray, A Fake Process Approach To Data Compression, IEEE Trans. Comm., Vol. COM-26, No. 6, June 1978.

22. G. D. Forney, Jr., The Viterbi Algorithm, IEEE Proc., Vol. 61, pp. 268-278, March 1973.

23. F. Jelinek and J. B. Anderson, Instrumentable Tree Encoding of Information Sources, IEEE Trans. Info. Theory, Vol. IT-17, pp. 118-119, January 1971.

24. J. B. Anderson and J. B. Bodie, Tree Encoding of Speech, IEEE Trans. Inform. Theory, Vol. IT-21, pp. 379-381, 1975.

25. J. Uddenfeld and L. H. Zettenberg, Algorithms for Delayed Encoding in Delta Modulation with Speech-like Signals, IEEE Trans. Comm., Vol. COM-24, pp. 652-658, June 1976.

26. L. H. Zettenberg and J. Uddenfeld, Adaptive Delta Modulation With Delayed Decision, IEEE Trans. Comm., Vol. COM-22, pp. 1195-1198, September 1974.

27. J. Max, Quantizing for Minimum Distortion, IRE Trans. Inform. Theory, Vol. IT-6, pp. 7-12, 1960. (Also reprinted in 12.)

28. W. D. Voiers, Diagnostic Evaluation of Speech Intelligibility, Speech Intelligibility and Speaker Recognition, ed. M. E. Hawley, Dowden, Hutchingson, and Ross, 1977.

29. G. S. Kang, Application of Linear Prediction Encoding to a Narrowband Voice Digitizer, NRL Report 7774, October 1974.

30. B. H. Juang and J. D. Markel, Cepstrally Based Pitch and Voicing Estimation with Statistical Assistance, ARPA NSC Note No. 140, unpublished technical memorandum, Signal Technology, Inc., Santa Barbara, CA. October 1979.

31. D. Y. Wong and J. D. Markel, An Intelligibility Evaluation of Several Linear Prediction Vocoder Modifications, IEEE Trans. ASSP, Vol. ASSP-26, pp. 424-435, October 1978.

32. Roucous et al., Segment Quantization for Very Low Rate Speech Coding, Conference Record ICASSP 82, pp. 1565-1568, May 1982.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.