OPTIMUM QUANTIZERS IN LINEAR PREDICTIVE ENCODING OF SPEECH

by

Marc L. Belleau, B.Sc. (Hons. Physics)

Department of Electrical Engineering
McGill University
Montreal, Canada

A thesis submitted to the Faculty
of Graduate Studies and Research in partial fulfillment
of the requirements for the degree of Master of Engineering

March, 1978

Electrical Engineering M.Eng.

QUANTIZERS IN LINEAR PREDICTIVE CODING OF SPEECH

Marc L. Belleau

Abstract

There have been many attempts in the past to reduce the transmission rate for a digital representation of a speech waveform. One technique for achieving this goal is a parametric representation using linear prediction, in which the parameters of that model are quantized before being transmitted. The purpose of this thesis is to study the effects of quantization. First, linear prediction methods in analysis, pitch extraction and synthesis are reviewed. Different distance measures and fidelity criteria are introduced. Then, for the reflection coefficients of linear prediction, schemes like inverse sine quantization and one which minimizes the expected spectral deviation bound, are discussed in detail. Finally, because these coefficients are mutually dependent, a decorrelation procedure is applied, and for the set of parameters obtained in this way, a quantization method which minimizes the expected spectral deviation bound is then derived and compared to the above mentioned schemes.

Génie Electrique Maîtrise

QUANTIFICATEURS OPTIMAUX DANS LE CODAGE DE LA PAROLE UTILISANT LA PREDICTION LINEAIRE

Marc L. Belleau

Résumé

Afin de diminuer la vitesse de transmission dans la représentation digitale de la parole, la prédiction linéaire est utilisée, et les coefficients de réflexion, implicite dans la solution aux équations de cette méthode, sont quantifiés. Tout d'abord, une revue est faite des méthodes de la prédiction linéaire dans l'extraction de la fréquence fondamentale, l'analyse et la synthèse de la parole. Ensuite, différentes mesures de distorsion et différents critères de fidélité sont considérés. Pour les coefficients de réflexion, des méthodes telles que la quantification arcsinus et celle qui minimise la borne supérieure de la déviation spectrale moyenne, sont examinées. Etant donnée l'interdépendance des coefficients de réflexion, ces derniers sont transformés en d'autres paramètres, pour éliminer cette corrélation. Finalement, la méthode de quantification, minimisant la borne supérieure de la déviation spectrale moyenne de ces nouveaux paramètres, est comparée aux méthodes mentionnées ci-dessus.

ACKNOWLEDGEMENTS

I would especially like to thank Dr. P. Kabal under whose supervision this research was conducted.

I am also grateful to the staff of INRS-Telecommunication for helping out in the experimental set-up and in the solution of numerous technical problems.

Thanks must also be given to Miss Cerrone and Miss Gotts whose skillful typing allowed the completion of the thesis.

TABLE OF CONTENTS

PAGE

I: INTRODUCTION

Over the past ten years much effort has been spent trying to reduce the bit rate of digitized speech subject to a fidelity criterion. Bit rate reduction is necessary in the transmission of speech signals over noisy communication channels. Conventional sampling and quantizing of a speech waveform requires 36,000 bits/sec if no difference between the original and output waveform is to be perceived by the ear. However the entropy of the written information of a spoken language in terms of the relative frequencies of occurrence of independent letters is about 50 bits/sec [1]. If the dependencies introduced by the constraints in the language are introduced, the entropy is even smaller. Furthermore, as stated in [1], experiments have shown that human subjects probably cannot process information at a rate above 50 bits/sec. Hence, if a subject is to perceive all the particular characteristics of a speaker, such as vocal inflections, timbre, nasality, the written version of the spoken utterance must contain redundant information. In view of these facts, the speech waveform is seen to be highly redundant. Therefore a scheme is sought that will extract as few parameters as possible and will permit reproduction of the original speech waveform as well as possible in some perceptual sense. Many such methods have been proposed and in particular the method of linear prediction has been quite successful in achieving that goal.

The following is a brief list of quantization methods based upon linear prediction, that have been found useful in the reduction of bit rate in speech:

- equal area coding of the reflection coefficients by Seneff [17, December 1974].
- uniform quantization of the reflection coefficients by Markel and Gray [10, 1974] and also by Chandra and Lin [16, August 1977]. In connection with this method there is also the dynamic programming bit allocation of Itakura and Saito mentioned in [10] (1972).
- the log area quantization of the log area parameters by Viswanathan and Makhoul [15, June 1975]. The Huffman coding of these parameters by Makhoul (1974) is also described in detail in [2].
- the inverse sine quantization of the reflection coefficients and the two parameter quantization scheme by Markel and Gray [14, December 1976].
- the minimum expected spectral deviation bound quantization of the reflection coefficients by Markel and Gray [12, February 1977].
- the decorrelation and DPCM approach of Sambur [18, December 1975].

All of the above methods will be discussed in the following chapters. First, an overall introduction to the thesis will be given.

The first section of Chapter II is essentially a review of linear prediction analysis as covered by Markel and Gray in [2]. The solution parameters of the linear prediction equations are the basic building blocks of all later work in this thesis. Section 2.2 then expounds on the physical models of the vocal tract, in order to obtain some insight into how well the above linear prediction model applies to it. Most of this work is covered by Flanagan in [1] and the relation with linear prediction is the subject of Chapter 4 in [2]. As the model is deficient in many respects, the efforts of Strube [5], Steiglitz [6] and Kopec [7] in improving it are briefly discussed in Section 2.3. With a better model, it is then shown that the poles and zeroes of the vocal tract are in closer agreement with actual values.

Chapter III first presents a short review on the results of a subjective comparison between various pitch extractors by McGonegal, Rabiner and Rosenfeld in [22]. The SIFT algorithm, as developed by Markel and Gray in [1], [9], is then discussed in some detail since it was the pitch tracker used in the present studies.

Chapter IV then reviews the particular analysis conditions used on speech when performing linear prediction analysis, and the type of synthesis structures and driving function to the speech synthesizer. The latter discussion culminates in the synthesizer program of Section 4.5. This pitch-synchronous synthesizer will be used to obtain the results of Chapter VI. All the above material is covered by Markel and Gray in [2]. Chapter IV is then concluded by the review of Markel and Gray on autocorrelation linear prediction vocoders [10].

In reducing the total bit rate some suitable quantization schemes are needed. This is the subject of Chapter V.

To this end, a spectral deviation measure is introduced and two fidelity criteria based on this measure are applied to quantization of the linear prediction parameters. Section 5.1 is essentially the work of Markel and Gray on distance measures, [11], and on optimal quantization using the expected spectral deviation bound, [12]. There is also a mention of another distance measure and of a proof concerning the maximum deviation bound criterion which is taken from Viswanathan and Makhoul in [15]. The material of Section 5.2 on the use of various sets of parameters in quantization is also to be found in [15].

Section 5.3 then describes the efforts of researchers in trying to reduce the bit rate using reflection coefficient quantization. First the maximum entropy coding scheme of Seneff [17] is discussed for comparison. An average bit rate of 1450 bits/sec was achieved when variable frame rate transmission is used in conjunction with equal area quantization. Then more details are given about the theoretical and experimental results of Viswanathan and Makhoul on two distance measures [15]. It is mentioned in the article, that speech quality is better using the $\sqrt[p]{D^p}$ distance measure of Markel and Gray [11] in the case of p = 1. The rest of Section 5.3 then expounds on the theoretical and experimental results of Markel and Gray on minimum max $\bar{D}$ and two parameter quantization [14] and minimum $E(\bar{D})$ quantization [12]. Using an optimum bit allocation procedure, they find that the total bit rate for direct, inverse sine and log area ratio quantization is about 3500 bits/sec for max $\bar{D}$ = 3dB as opposed to 2800 bits/sec in the two parameter scheme. The speech quality is the same in both cases. [14] is a theoretical study giving only the number of bits allocated to the first and tenth reflection coefficient for a fixed $E(\bar{D})$ = .3dB each. It is then mentioned that the reflection coefficients are dependent on past values and also on each other, and that further bit rate reduction would be possible if this dependence could somehow be extracted. In [14], Sambur's work on decorrelation of data and DPCM is pointed out. This scheme-[18] and decorrelation especially, is discussed at the beginning of Section 5.4. In conjunction with DPCM, decorrelation can reduce the bit rate to 600 bps and for some utterances the quality will still be acceptable. The purpose of this research is then to test whether or not decorrelation of the reflection coefficients, as done in [21], will reduce the total bit rate when the minimum expected spectral deviation bound quantization scheme of [12] is applied to the decorrelated parameters. Only dependence within a frame is treated in this study (no DPCM). In order to decorrelate the data, a Jacobi diagonalization of the covariance matrix of the reflection coefficients is performed, [19]. A summary of the basic ideas behind this diagonalization is presented. In the remainder of Section 5.4, the relation between the sensitivity function of the new parameters and the sensitivity function of the reflection coefficients is then derived. The new parameter is a known linear combination of the reflection coefficients (from the Jacobi diagonalization) and if these relations are used in conjunction with the equations of [14], then the desired result is obtained. Then, a few assumptions will be made on what the probability density function and average sensitivity function of the new parameters should be. These results are then substituted into the equations of [12], to yield the optimum quantizer curves and the number of levels. An alternative scheme which was developed is to compute these functions using time averages.

In order to establish a comparison with other schemes, experimental results on quantization of the reflection coefficients themselves using the $E(\bar{D})$ fidelity criteria are also computed. These will at the same time complement the theoretical study of [12]. For this study, two quantizer functions are selected: the inverse sine quantization which optimizes the fidelity criteria max $\bar{D}$ of [14] and min $E(\bar{D})$ quantization. A time average of the sensitivity function will be computed as was done above for the decorrelated parameters.

Experimental results appear in Chapter VI. The set-up procedure is first described, and then the logarithmic quantization of the pitch and gain [10], is discussed. With a fidelity criteria $E(\bar{D})_{tot} = 3.5$ dB, it is found that inverse sine and min $E(\bar{D})$ quantization of the reflection coefficients, and min $E(\bar{D})$ quantization of the decorrelated parameters result in a total bit rate of 3070, 2750, 2884 bits/sec respectively. Moreover, the subjective quality of speech processed under these three conditions is the same.

The conclusion and suggestion for further research appear at the end.

II: THE LINEAR PREDICTION MODEL OF SPEECH

In section 2.1, the method of solution to the covariance and autocorrelation equations is presented. Using the correlation matching criterion, the energy in the output signal from the linear prediction analysis is then shown to be equal to the gain of the linear prediction filter. Section 2.2 then describes the physics of the vocal tract and its excitation sources. A simplified model consisting of a cascade of transmission lines is then developed. If further assumptions are made, then the model is found to be mathematically equivalent to the solution to the autocorrelation equations. Finally, section 2.3 gives a brief discussion about more accurate methods of obtaining the parameters of the speech waveform, in the case where the above assumptions are not made.

2.1. The Basic Equations of Linear Prediction

Linear prediction attempts to achieve bit rate reduction by, as the name implies, approximating a speech sample value using a linear combination of a certain number M (to be specified later) of past speech samples. Namely,

$$s(n) = \sum_{K=1}^{M} -a_K s(n-K) + e(n) \qquad (2.1.1)$$

where the error signal $e(n)$ is small. The parameters to be extracted are the $-a_K$'s and they are chosen to be those which minimize $\alpha = \sum_{n=n_0}^{n_1} e^2(n)$ where the interval (n_0, n_1) to be used will also be specified. Extrema can be obtained by setting the derivative with respect to each a_K to zero. Let

$$c_{ij} = \sum_{n=n_0}^{n_1} s(n-i)s(n-j) \tag{2.1.2}$$

and

$$\alpha = \sum_{n=n_0}^{n_1} e^2(n) = \sum_{i=o}^{M} \sum_{i=o}^{M} a_i c_{ij} a_j \tag{2.1.3}$$

$$\frac{\partial \alpha}{\partial a_K} = 0 = 2 \sum_{i=o}^{M} a_i c_{ik} \tag{2.1.4}$$

yields

$$\sum_{i=1}^{M} a_i c_{ik} = -c_{ok} \tag{2.1.5}$$

and

$$\sum_{i=o}^{M} a_i c_{oi} = \alpha \tag{2.1.6}$$

In linear prediction, the calculation of (2.1.2) involves a finite number N of samples. Let them be denoted by $s(o)$, $s(1)$, ..., $s(N-1)$. As will be seen later, N depends on the region of validity of (2.1.1). Two methods [2, p. 14-15, Chapter I] are used for solving the system of simultaneous linear equations $A\underline{b} = \underline{c}$ as represented by (2.1.5). They differ in the way the N samples are used to obtain the a_k's.

Autocorrelation method

Here, $n_0 = -\infty$ and $n_1 = \infty$. Hence, because only N samples are used, this is equivalent to windowing the speech waveform over the N samples. Note that $c_{ij} = c_{ji}$ and

$$c_{ij} = \sum_{n=-\infty}^{\infty} s(n-i)s(n-j) = \sum_{n=-\infty}^{\infty} s(n)s(n-|j-i|) = c_{o,|j-i|}$$

Hence, $c_{i+1,j+1} = c_{o,|j+1-(i+1)|} = c_{ij}$ and the matrix $[c_{ij}]$ is Toeplitz. $c_{o,|j-1|}$ is then an autocorrelation coefficient and is denoted by $r(|j-i|)$. Also from (2.1.1), e(n) is defined for n = 0,1,....,N+M-1.

Covariance method

Here, $n_0 = M$ and $n_1 = N-1$. The symmetric matrix $[c_{ij}]$ is no longer Toeplitz because

$$c_{i+1,j+1} = c_{ij} + s(M-i-1)s(M-j-1) - s(N-1-i)s(N-1-j).$$

An attractive scheme for the numerical solution to (2.1.5) and 2.1.6) is now discussed.

<u>The inner product formulation</u> [2, p. 35-38, Chapter 2]

For any two arbitrary polynomials in z^{-1}, of degree M

$F(z) = \sum_{i=0}^{M} f_i z^{-i}$ and $G(z) = \sum_{i=0}^{M} g_i z^{-i}$ where f_i, $g_i \varepsilon R$, define an operation $(F(z), G(z)) = (G(z), F(z)) = \sum_{i=o}^{M} \sum_{j=o}^{M} f_i (z^{-i}, z^{-j}) g_j$.

From its form, it is seen to satisfy some of the properties of the inner product. Equation (2.1.3) together with $A(z) = \sum_{i=o}^{M} a_i z^{-i}$ is seen to be an inner product $(A(z), A(z))$ with $(z^{-i}, z^{-j}) = c_{ij}$. Similarly, (2.1.5) is $\sum_{i=o}^{M} \sum_{\ell=o}^{M} a_i c_{ij} \delta_{\ell j}$

$$= \left(\sum_{i=o}^{M} a_i z^{-i}, \sum_{\ell=o}^{M} \delta_{\ell j} z^{-\ell} \right) = (A(z), z^{-\ell}) = 0$$

$$\ell = 1, 2, \ldots M .$$

This orthogonal relationship is the basis for a recursive scheme used to calculate the a_k's. The idea is to solve the problem for m = 1, 2, ... M successively.

Let $e_m^+(n) = \sum_{i=o}^{m} a_{mi} s(n-i) \qquad a_{mo} = 1$

It is called the forward prediction. Similarly, let the backward predictor be

$$e_m^-(n) = \sum_{i=1}^{m+1} b_{mi} s(n-i) \qquad b_{m,m+1} = 1$$

As before, the extremum α_m of $\sum_{n=n_o}^{n_1} [e_m^+(n)]^2$ and the extremum β_m of $\sum_{n=n_o}^{n_1} [e_m^-(n)]^2$ are obtained by setting the derivatives with respect to a_{mi}, b_{mi} to zero. In inner product notation the solution is

$$(A_m(z), z^{-\ell}) = 0 \quad \ell = 1,2,\ldots m \quad \text{where } A_m(z) = \sum_{i=o}^{m} a_{mi} z^{-i}$$

$$(B_m(z), z^{-\ell}) = 0 \qquad \text{and} \quad B_m(z) = \sum_{i=1}^{m+1} b_{mi} z^{-i}$$

$$\alpha_m = (A_m(z), A_m(z)) \qquad \beta_m = (B_m(z), B_m(z)).$$

Now it is shown that these extrema are indeed minima.

Proof: Let $F(z)$ be a polynomial minimizing $(F(z),F(z))$.

Then,

$$(F(z)+Cz^{-j}, F(z)+Cz^{-j}) \geq (F(z),F(z)) \qquad \forall C \in R$$

$$j = 1,2,\ldots,\deg F .$$

Then $$2C(F,z^{-j}) + C^2(z^{-j},z^{-j}) \geq 0 \tag{2.1.7}$$

Since it is true for any C, choose C to be $-(F(z),z^{-j})/z^{-j},z^{-j})$.

(2.1.7) then implies $(F(z),z^{-j})^2 \le 0$. However, if $(z^{-j},z^{-j}) = 0$ let $C = -(F(z),z^{-j})$. In both cases then, $(F(z),z^{-j})^2 \le 0$. All speech samples $s(n)$ are real $\rightarrow (z^{-i},z^{-j}) = c_{ij} \varepsilon R$ and coefficients of any polynomial are real. Hence $(F(z),F(z))$ is a minimum implies $(F(z),z^{-j}) = 0$. Conversely, given $(F(z),z^{-\ell}) = 0$ for any $Q(z) = \sum_{j=o}^{M} q_j z^{-j}$, $(F(z)+Q(z),\ F(z)+Q(z)) = (F(z),\ F(z))+(Q(z),Q(z))$. But $(Q(z),\ Q(z)) =$

$$\sum_{ij}\sum q_i \sum_{n=n_0}^{n_1} s(n-i)s(n-j)q_j = \sum_{n=n_0}^{n_1} \{\Sigma q_i s(n-i)\}^2 \ge 0 .$$

Consequently, $(F(z),\ z^{-\ell}) = 0$ implies $(F(z),\ F(z))$ is a minimum. Hence the necessary and sufficient conditions have been proven. From the orthogonality properties of A_m and B_m and the linearity of the scalar product,

$$(A_m(z),\ \sum_{i=1}^{m} r_i z^{-i}) = \sum_{i=1}^{m} r_i (A_m, z^{-i}) = 0 = (B_m(z),\ \sum_{i=1}^{m} d_i z^{-i})$$

Notice then that $(A_m(z),\ A_n(z)) = (A_m, z^{\circ}) \neq 0$.

However, $(B_m,B_i) = \delta_{mi}\beta_m$.

Proof: The case $i = m$ is from the definition: $(B_m,B_m) = \beta_m$. Since the inner product is symmetric, the case $m < i$ is the same as $m > i$ and $(B_m,B_i) = \sum_{j=1}^{m+1} b_{mj}(B_i,z^{-j}) = 0$ because

$1 \le j \le m+1 \le i$ satisfies $1 \le j \le i$, Q.E.D.

Going back to the problem of finding A(z) which satisfies $(A(z), z^{-\ell}) = 0$, the recursive procedure to be followed is to find an $A_m(z)$ orthogonal to the basis $z^{-\ell}$ given that orthogonal polynomials $A_{m-1}(z)$ and $B_{m-1}(z)$ are already known [2, p.48-56, Chapter 3]. Since $A_m(z) = \deg B_{m-1}(z)$,

$$A_m(z) = A_{m-1}(z) + k_m B_{m-1}(z) \tag{2.1.8}$$

is easily seen to be orthogonalized by letting

$$\begin{aligned} k_m &= -(A_{m-1}(z), z^{-m})/(B_{m-1}(z), B_{m-1}(z)) \\ &= -(A_{m-1}(z), z^{-m})/\beta_{m-1} \end{aligned} \tag{2.1.9}$$

From (2.1.8)

$$A_m(z) = 1 + \sum_{i=1}^{m} k_i B_{i-1}$$

$$(A_m(z)-1, A_m(z)-1) = \sum_{i=1}^{m} \sum_{j=1}^{m} k_i k_j (B_{i-1}, B_{j-1}) \tag{2.1.10}$$

$$= \sum_{i=1}^{m} k_i^2 \beta_{i-1} \tag{2.1.11}$$

$$(A_m(z), A_m(z)) - 2(A_m(z), 1) + (1,1) = \sum_{i=1}^{m} k_i^2 \beta_{i-1} \tag{2.1.12}$$

$$\alpha_m = (1,1) - \sum_{i=1}^{m} k_i^2 \beta_{i-1} \tag{2.1.13}$$

$$\alpha_{m+1} = \alpha_m - k_{m+1}{}^2\beta_m \tag{2.1.14}$$

$$A_o(z) = 1$$

$$B_o(z) = z^{-1}$$

$$\alpha_o = (1,1) = c_{00}$$

$$\beta_o = (z^{-1},z^{-1}) = c_{11}$$

$$k_1 = -\frac{(A_o(z),z^{-1})}{\beta_o} = -\frac{c_{01}}{c_{11}}$$

$$A_1(z) = A_o(z) + k_1B_o(z) = 1 + -\frac{c_{01}}{c_{11}}z^{-1}$$

Hence, $a_{10} = 1$; $a_{11} = -c_{01}/c_{11}$ which completes the initialization. Notice that $B_1(z)$ has not been found. In fact at any step $m-1$, $B_m(z)$ is not obtained by the above procedure. The solution is to use a Gram-Schmidt orthogonalization

$$B_m(z) = z^{-(m+1)} + \sum_{i=1}^{m} b_{mi}z^{-i} = z^{-m+1} - \sum_{i=0}^{m-1} \gamma_{mi}B_i(z) \tag{2.1.15}$$

Because the $B_m(z)$ are orthogonal to each other,

$$(z^{-m+1},B_j(z)) = \sum_{i=o}^{m-1}\gamma_{mi}(B_i(z),B_j(z))$$

$$\gamma_{mj} = \frac{(z^{-(m+1)}, B_j(z))}{(B_j(z), B_j(z))} = \frac{(z^{-m+1}, B_j(z))}{\beta_j} \qquad 0 \leq j \leq m-1$$

$$\beta_j \neq 0$$

If $\beta_j = 0$, γ_{mj} is arbitrary. Then,

$$B_m(z) = z^{-m+1} - \sum_{i=o}^{m-1} \gamma_{mi} \sum_{j=1}^{i+1} b_{ij} z^{-j} = z^{-m+1} - \sum_{i=o}^{m-1} \sum_{j=1}^{i+1} \gamma_{mi} b_{ij}\, z^{-j}$$

or,

$$b_{mj} = - \sum_{i=j-1}^{m-1} \gamma_{mi} b_{ij}$$

Now that $B_m(z)$ is known, β_m and k_{m+1} are easily calculated. Rewriting (2.1.8) for step m as

$$a_{m+1,o} = 1$$

$$a_{m+1,i} = a_{mi} + k_{m+1} b_{mi} \tag{2.1.16}$$

$$a_{m+1,m+1} = k_{m+1} \tag{2.1.17}$$

and substituting the above values of b_{mi}, β_m, k_{m+1} in (2.1.14), (2.1.16), and (2.1.17), step m is therefore completed and applies to both the covariance and autocorrelation method. In the latter, the relation $c_{ij} = r(j-i)$ simplifies the algorithm even more. For, let $j = m+1-\ell$ in $\sum_{i=o}^{m} a_{mi} r(i-j) = 0 \qquad a_{mo} = 1$

$i = m+1-k$

Then

$$\sum_{k=1}^{m+1} a_{m,m+1-k} r(m+1-k-(m+1-\ell)) = \sum_{k=1}^{m+1} a_{m,m+1-k} r(\ell-k) = 0$$

Let $b_{mk} = a_{m,m+1-k} \qquad k = 1, 2, \ldots\ldots m+1$

Then, $b_{m,m+1} = a_{mo} = 1$ as required for $B_m(z)$ and $(B_m, z^{-\ell}) = 0$.

Furthermore,

$$B_m(z) = \sum_{i=1}^{m+1} b_{mi} z^{-i} = \sum_{i=1}^{m+1} a_{m,m+1-i} z^{m+1-i} \; z^{-(m-1)} = +z^{-(m+1)} A_m(1/z) \tag{2.1.18}$$

Substituting (2.1.18) in (2.1.8) gives

$$z^{-m+1} B_m(1/z) = z^{-m} B_{m-1}(1/z) + k_m z^{-m} A_{m-1}(1/z) \qquad z \rightarrow 1/z$$

$$z^{m+1} B_m(z) = z^{+m} B_{m-1}(z) + k_m z^m A_{m-1}(z)$$

$$zB_m(z) - k_m A_{m-1}(z) = B_{m-1}(z) \tag{2.1.19}$$

This autocorrelation algorithm has been implemented as a FORTRAN subroutine program in [2], and will be used in analysis and pitch extraction of speech as described in Chapter VI.

Correlation matching - calculation of gain

In z transform notation (2.1.1) may be expressed as $S(z) = E(z)/A(z)$. In the autocorrelation method it is desired to match the autocorrelation $\rho(j)$ of the unit sample response of the vocal apparatus to that of the input speech signal $s(n)$ within the window used: $\rho(j)=r(j)$, $j=0,1,\ldots,M$ [2, p. 31-32, chapter 2]. Assume the transfer function of this unit sample response to be a causal all-pole filter $H(z) = \sigma/A(z)$ and rewrite this as

$$\sum_{i=o}^{m} a_i h_{n-i} = \sigma\delta_{no}$$

Then

$$\sum_{i=o}^{m} a_i \rho(i-j) = \sigma\sum_{n} \delta_{no} h_{n-j} = \sigma h_{-j} = 0 \qquad j > 0 \qquad (2.1.20)$$

because of the causality. From (2.1.19) $n=0$ gives $a_o h_o = h_o = \sigma$. Consequently,

$$\sum_{i=o}^{m} a_i \rho(i) = \sigma^2 \qquad (2.1.21)$$

But since $\rho(j) = r(j)$ for $j = 0,1,\ldots M$, the solution to (2.1.20) is the a_k's obtained in the previous section and $\sigma^2 = \alpha = (A,A)$ is the minimum energy. In particular since $\rho(o) = r(o) =$ the energy of the input signal, by Parseval's theorem then, σ matches the average value of $|S(e^{j\theta})|^2$ to the average value of $\sigma^2/|A(e^{j\theta})|^2$.

Now, as $M \to \infty$, $\rho(j) = r(j)$, $j \in I$ and since the spectrum equals the transform of the autocorrelation sequence, $|S(e^{j\theta})|^2 = \sigma^2/|A(e^{j\theta})|^2$. The autocorrelation method then gives a perfect fit to the magnitude of the speech spectra. Consider the log spectra of $|A(e^{j\theta})|^2$

$$= \int_{-\pi}^{\pi} \ln|A^*(e^{j\theta})|^2 \frac{d\theta}{2\pi} = \int_{\pi}^{\pi} |\ln|A(e^{-j\theta})|^2 \frac{d\theta}{2\pi} \quad \text{since } a_i \in R$$

$$= 2\mathrm{Re} \oint_{|z|=1} \ln A(1/z) \frac{dz}{2\pi j z} . \quad \text{But } A(z) \text{ is causal}$$

and therefore the roots of $A(1/z)$ are all outside the unit circle. The residue is simply

$$2\ \mathrm{Re}\ell n A(\infty) = 2\ \mathrm{Re}\ell n 1 = 0 \qquad \text{since } a_o = 1.$$

Consequently,

$$\int \ln\left|\frac{\sigma}{A(e^{j\theta})}\right|^2 \frac{d\theta}{2\pi} = \ln\sigma^2 = \ln\alpha$$

Experimentally it is found [2, Chapter 6] that the log spectrum of the speech signal tends to lie below the model log spectrum and also the latter tends to fit the peaks more accurately than the dips. Actually this observation is desired because the peaks represent the resonance frequencies of the vocal tract and these play a dominant role in the perception of voiced speech. In the covariance formulation a power spectrum cannot really be defined since c_{ij} is not an autocorrelation.

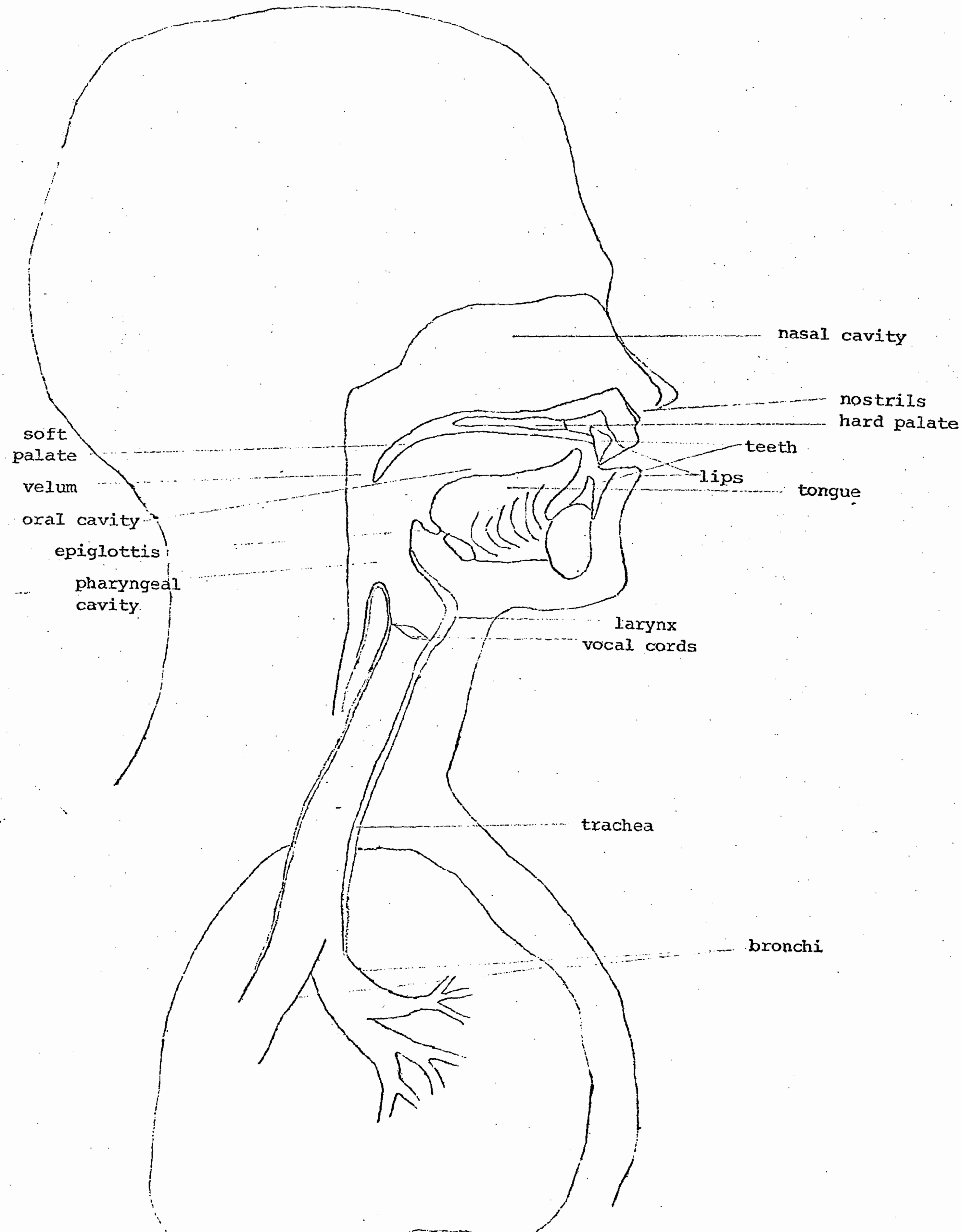
Nevertheless, if A(z) is causal and $\ln|1/A(e^{j\theta})|^2$ is compared to $\ln|S(e^{j\theta})|^2$ the same observations are made [2, Chapter 6].

2.2 The Speech Production Model and its Relation to Linear Prediction

Vocal tract apparatus [1, p. 9-15, Chapter 2].

The complex sound which is perceived as speech is the result of a pressure wave generated by our vocal apparatus. The major components of the system are shown on diagram (2.2.1). The source of power for the expiration of air is the contraction of the lungs by the rib muscles. The sources of excitation for modulating this mass air flow are (1) vocal cord vibrations and (2) any constriction at an arbitrary location in the vocal apparatus. The first gives rise to speech classified as voiced. By voluntarily tightening the vocal cords which are attached to the arytenoid cartilages in the glottis, the subglottal pressure will force them apart to allow the air to be expired. But by the Bernoulli principle, which is a form of energy conservation, a moving

Figure 2.2.1 Vocal Apparatus



fluid exerts less pressure on the walls than a stationary enclosed one. Hence, the pressure in the glottal region drops and the vocal cords are brought closer together reducing the air flow and building up the pressure again. This vibratory behavior of the vocal cord then results in quasi periodic variations in the output airflow. The tension in the vocal cords and the subglottal pressure determine respectively the pitch and intensity of the resultant pressure wave. The duty cycle of the waveform is also proportional to the pitch and intensity. The second excitation can be subdivided into two categories. If a pressure is built up behind a closure point constriction and is suddenly released by opening the latter, then a plosive unvoicedsound is produced. If a constriction creates local turbulence in the air stream, the resulting random pressure wave is called a fricative sound. It is possible to have sounds characterized as voiced and unvoiced. When the velum is open the air passes through both the nasal and oral cavity giving rise to nasal sounds.

Models of the vocal apparatus [1, Chapter 3]

Consider a stationary vocal tract configuration (with the velum closed) and a pressure wave emanating from it. For the range of frequencies involved in the production of audible sounds, the length of the vocal tract from the glottis to the lips is of the same order of magnitude as the sound wavelengths. Consequently a wave analysis of sound production is required. Moreover, if the transverse dimensions to the tract are small compared to a wavelength then the analysis is one-dimensional and reduces to solving the classical Webster-Horn equation subject to the given boundary conditions at the lips and the glottis. However, the analysis does not lead to tractable mathematics because the vocal tract's cross section is a function of the distance from the glottis (a non uniform tube). An approximate solution to the problem is to represent the vocal tract by a finite number of series interconnections of uniform tubes each of which has a short length compared to the range of wavelengths of interest. The solution to a one-dimensional wave analysis of a uniform tube is analogous to that of a uniform electrical transmission line. Here the inertia of the air particles, the compressibility of the air volume and the viscous and heat conduction losses at the walls are playing the role of inductance, capacity and resistance respectively. These losses are even more important when modelling the nasal tract (velum open) because of its convoluted surface area. In addition the walls themselves are not smooth and rigid and this is another contribution to the net impedance of the tube. A cascade connection of tubes of different cross sections is then analogous to a cascade connection of transmission lines of different lengths and impedances per unit length.

Models of excitation sources [1]

First, consider voiced excitation. The subglottal pressure P_s is almost equal to the lung pressure P_L because of the negligible drop across the bronchi and trachea. P_s is also constant over many pitch periods because the rib muscles contract the lungs in proportion to the quantity of air expelled. Consequently, the lung capacitance and inductance are variable. It was already pointed out that the vocal cords vibrate under tension. Consequently their inertia can be represented by an inductance and the damping of their motion due to the viscous fluid flow by a resistance. However, during their vibratory cycle, the cords' inertia and damping are time varying. The model of the glottis assumes that the glottal output volume velocity of air is not perturbed at all by the presence of the vocal tract. This is obviously not true, especially when a tight constriction exists, because the pressure wave is partially reflected back into the glottis.

The model for the source when a constriction occurs is a random impedance and generator whose mean values depend on the volume velocity and the area of the constriction in a nonlinear way. The spectrum of this noise source has been determined to be relatively uniform at the point of the constriction. It can then be modelled as white noise. A similar model can be used for plosive sounds.

Termination at the lips

Since a pressure wave is radiated from the lips, there is a non-zero output impedance. It varies with the size of the mouth opening and for wavelengths long compared to the mouth opening, it behaves as a resistance proportional to w^2 in series with an inductance proportional to w, where w is the frequency of a sinusoidal input. The model used to compute the impedance is even more suited to the nasal tract because the nostril opening is even smaller. Also because the distance from the nostrils to the mouth is short compared to the wavelength, the phase difference between the mouth and nostril pressure waveforms is small and to a good approximation the output speech is the sum of the two contributions.

Relation to the linear prediction all-pole model

Since a computer simulation was used, the speech had to be digitized. Procedures in recording speech on discs and playing it back will be discussed in Chapter VI. If no aliasing is desired, then one sets the sampling frequency of the converter to at least twice the cutoff frequency. To show this, let $F_a(s)$ be the Fourier transform of $f_a(t)$ and let $f(n) = f_a(nT)$ be the equally spaced samples of $f_a(t)$. Then

$$f(n) = \frac{1}{2\pi j} \oint F(z) z^{n-1} dz$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{jw}) e^{jwn} dw \qquad \text{where } F(z) = \sum_{n=-\infty}^{\infty} f(n) z^{-n}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} F_a(jw) e^{jwnT} dw$$

Combining these, results in [3, p.26-29 (Chapter 1]

$$F(e^{jw}) = \frac{1}{T} \sum_{m=-\infty}^{\infty} F_a(j(\frac{w}{T} - \frac{2\pi m}{T})) \tag{2.2.1}$$

Consequently, if no aliasing is desired, it is necessary that $|F_a(\frac{w}{T})| = 0$ for $|w| > \pi$. The speech must then be bandlimited prior to sampling.

In the language of sequences let $T[e(n)] = s'(n)$ be a transformation from glottal input, to the output speech waveform. Now,

$$e(n) = \sum_{K=-\infty}^{\infty} e(K)\delta(n-K) \qquad \text{since } \delta(m) = \delta_{mo} .$$

T will be assumed to be linear. This requires among other things that the glottis be uncoupled from the vocal tract.

$$T[e(n)] = \sum_{K=-\infty}^{\infty} e(K)T[\delta(n-K)] = \sum_{K=-\infty}^{\infty} e(K)h(n-K) \tag{2.2.2}$$

(letting $T[\delta(m)] = h(m)$). Up to now it was assumed that the vocal tract configuration did not change in time. This condition will now be relaxed. (2.2.2) will not be true for all n (i.e. $h(K) = h(K;n)\ \forall\ K$). However, it will be assumed that h(K) does not depend on n for a certain range of n say, from 0 to N-1. Therefore let $s(n) = w(n)s'(n)$ where $w(n) = 0$ for $n \notin (0,N-1)$. Then by the convolution theorem

$$S(e^{jw}) = \int S'(e^{j\theta})W(e^{j(w-\theta)})d\theta$$

Details can be found in [3, section 5.5] about the type of windows used to approximate $S'(e^{jw})$ by $S(e^{jw})$. For example a Hamming window will be used before performing autocorrelation analysis. Notice that even if the system was time invariant, only an approximation to a spectral computation $S(z) = \sum_{n=-\infty}^{\infty} s(n)z^{-n}$ is possible because of the infinite limits of summation.

With the series connection of uniform tubes model of the vocal tract (i.e. velum closed), it can be shown, from [1] and also from the further use of (2.2.1) that to a good approximation, the transfer function from the glottal output, to the lips is of the form

$$V(z) = \frac{A}{\prod_{k=1}^{L}(1-c_k z^{-1})(1-c^*_k z^{-1})} \qquad |c_k| < 1 \tag{2.2.3}$$

in the case of voiced excitation. In the case of excitation at a constriction in the vocal tract, there is also generation of zeroes and to a good approximation, the transfer function is

$$H_u(z) = \frac{A \prod_{k=1}^{m} (1-\alpha_k z^{-1})(1-\alpha^*_k z^{-1})}{\prod_{k=1}^{p} (1-c_k z^{-1})(1-c^*_k z^{-1})} \tag{2.2.4}$$

It can also be shown from [1], that the poles and zeroes will be perturbed by the lip radiation model's poles and zeroes contribution. However the contributions due to this model can be simplified by an additional factor $1-z^{-1}$ in the numerator [2, section 1.3]. The z transform for the noise source is a constant as it is represented by white noise. Since the output of the glottis is a periodic pulse, the input to the glottis can be modelled by an infinite train of unit pulses equally spaced by an amount equal to the pitch period. The transfer function of the glottis will modify the pulses. Since it is uncoupled from the rest of the vocal tract its poles and zeroes contribution will not perturb those of the vocal tract.

This glottal transfer function is often approximated by a 2 pole filter $(1/(1-\alpha z^{-1})^2$ [2, section 1.3]. One of these factors can then cancel the numerator $1-z^{-1}$ due to the lips because α is close to 1 in this model. Hence for the voiced situation the net transfer function $1/A(z)$ is all-pole. Using (2.2.2), $s(n) = w(n)(e(n)*h(n))$. If $h(n)$ varies slowly with respect to $w(n)$ [3, p. 514, chapter 10] then

$$s(n) \sim [w(n)e(n)]*h(n) \tag{2.2.5}$$

since $1/A(z) = \sum_{n=-\infty}^{\infty} h(n)z^{-n}$,

$$S(z) \sim \frac{E_w(z)}{A(z)}$$

where $E_w(z)$ is an all zero input because it is of finite duration. This last equation is the z transform of (2.1.1).

Next using the mass continuity, momentum and the Webster-Horn equations (the latter being easily derived from the first two) and the continuity equations for volume velocity and pressure at the boundary between two uniform tubes, it is shown in [2] that in the case of no pressure wave leaving the lips (i.e., the output impedance at the lips is zero), equations entirely analogous to the autocorrelation equations

$$A_o(z) = 1 \tag{2.2.6}$$

$$B_o(z) = z^{-1} \tag{2.2.7}$$

$$A_m(z) = A_{m-1}(z) + k_m B_{m-1}(z)$$

$$zB_m(z) = k_m A_{m-1}(z) + B_{m-1}(z)$$

are obtained. In the present situation, m is the index denoting a uniform tube. m=0 stands for the tube terminated on one side, at the lips and m=M for the tube terminated on one side at the glottis. Here $k_m = 1-A_m/A_{m-1} \;/\; 1 + A_m/A_{m-1}$

where A_m is the cross-section of uniform tube m and it represents the fraction of the energy which is reflected back into the tube. This is the reason for calling the M parameters k_m in autocorrelation linear prediction, reflection coefficients.

2.3 Improved Parameter Representation of Speech

The error signal e(n) which is the output of the linear prediction filter A(z) exhibits the following properties [4, page 11].

(1) It is quasi-periodic due to the vibratory motion of the vocal cords.

(2) No interval can be found within a period, which will possess a flat amplitude spectrum like that of silence or white noise.

(3) A jitter from one pulse to the next in the instantaneous period of the waveform is observed because of instabilities in the vocal cord motion.

In addition, the glottal transfer function is time-varying within a pitch period (Section 2.2). A(z) and e(n) as obtained from an interval covering several periods might then not accurately represent the vocal tract transfer function and the input to it. For example, as pointed out in

[5], there is no clear cut one-to-one correspondence between two adjacent peaks of e(n) and the points of strong excitation in pre-emphasized speech. However, as will be done in the next chapter, e(n) can still be used to provide an estimate of the pitch. Once having obtained such an estimate, it is then proposed in [5], to perform linear prediction over intervals short compared to this calculated pitch period. Then, assuming hard glottal closure, it is then expected that $\sum_{n=n_0}^{n_1} e^2(n)$ would fall to zero as the segment of constant length is shifted to an interval lying between two points of glottal closure. In practice it should not fall exactly to zero even if glottal closure is quite sharp, because of the slow rise of the next glottal pulse. However this is not a practical scheme to be implemented in a speech transmission system because once an initial pitch estimate is obtained for an analysis frame (10-30 ms in length), the computation involved in the search of just one excitation-free interval is to be done on all such intervals within that analysis frame if correct information about the excitation signal is to be transmitted. The method might also not be accurate if the assumption of hard, glottal closure does not hold.

Nevertheless, returning to the error signal e(n) obtained from the original analysis frame, it is found in [6], that linear prediction applied to an interval of speech lying between two finite duration pulses, will result in a spectral plot $\alpha/|A(e^{j\theta})|^2$ which averages the peaks of $|S(e^{j\theta})|^2$ better than the previous analysis. Letting E(z) be the z transform of the new error signal, it is then suggested to obtain the zeroes of the spectrum by performing linear prediction on the z^{-1} transform of 1/E(z) or by solving for the roots of $\sum_{n \in J} e(n) z^{-n}$ where J is an interval lying within one of the finite duration pulses. It is then observed in [6] that approximately the same zeroes are obtained if the interval J is shifted to a region between pulses. The zeroes are then more likely to be due to an opening of the velum than to the presence of a glottal pulse.

Up to now, methods of obtaining the error signal e(n) and the vocal tract transfer function in the presence of a voiced excitation, have been briefly described. However there is a method which avoids the difficulties arising from the existence of such an error signal. It is called homomorphic deconvolution and in some cases [3, Chapter 10] is useful in separating a signal into its basic components. It involves finding the z^{-1} transform $\hat{x}(n)$ of log X(z) where $X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}$.

Now from (2.2.5) $S(z) = E_w(z)H(z)$.

Therefore $\log S(z) = \log E_w(z) + \log H(z)$

or $\hat{s}(n) = \hat{e}_w(n) + \hat{h}(n)$

It is then shown in [3] that for large pitch periods, $\hat{h}(n)$

does not overlap $\hat{e}_w(n)$ appreciably because of its rapid decay ($\hat{h}(n) \leq C^n/n$, where C is a bound). Consequently it is then possible to separate $\hat{h}(n)$ from $\hat{e}_w(n)$ and hence $h(n)$ from $e_w(n)$. Writing the vocal tract transfer function H(z) as

$$\frac{\sum_{i=1}^{N_z} a_i z^{-i}}{\sum_{j=o}^{N_p} b_j z^{-j}}$$

the problem then becomes that of solving for the a_i's and b_i's simultaneously. As it is a highly non-linear problem, its solutions are approximated by those solutions to modified linearized problems. Methods of solution to two such simplified problems have been proposed by Kalman and Shank [8]. The original non-linear problem can only be solved iteratively, and even then, there is no guarantee that the algorithm will converge. One such scheme, called iterative prefiltering, is discussed in [8], where it was shown that it actually results in a more accurate representation of the vocal tract than Shank's method. However the two main disadvantages are increased complexity and execution time of the algorithm.

In conclusion, this section was basically concerned with the limitations of the linear prediction algorithm. Further problems arise in including zeroes as parameters. First there is the difficulty in locating them in any real system due to ever-present interfering signals. Also recall that if aliasing is avoided, then a cutoff frequency $f_c \leq f_s/2$ is necessary. But then f_c must be as close to $f_s/2$ as possible if zeroes in the spectrum are also to be avoided. Also since a windowed frame contains a finite number of samples only, the z transform is then a polynomial (an all zero transform). Zeroes in the transmission are in addition masked by these artificially created zeroes. Conventional linear prediction will from now on be used. Also the input to the glottis will from now on be approximated by a train of equally spaced input samples.

III: PITCH EXTRACTORS

One parameter of great importance in the perception of voiced speech is the fundamental frequency of the glottal excitation, [2], more commonly called the pitch. Therefore the conception of a very accurate pitch tracker would allow a great reduction in transmission bit rate at little loss of fidelity. Several pitch detectors have already been proposed. In section 3.1, the subjective results [22] of speech synthesized using different pitch detectors are summarized and section 3.2 describes in more detail one particular detector which was used in obtaining the results of Chapter VI.

3.1 Comparison of Various Pitch Extractors

In [22] a subjective comparison of linear prediction synthesized speech in which only the method of pitch extraction is allowed to vary, was carried out. In all, eight such methods were studied and are listed below:

(1) SAPD (semi automatic pitch contour)

(2) LPC (spectral equalization LPC method)

(3) AMDF (average magnitude difference function)

(4) PPROC (parallel processing method)

(5) AUTOC (modified autocorrelation method)

(6) SIFT (simplified inverse filtering method)

(7) CEP (cepstrum method)

(8) DARD (data reduction method)

Details on the theory of operation of each of these algorithms are provided in the references listed in [22]. The original unprocessed utterance was also included in the study of [22], for a total of nine versions of an utterance. For each of these versions, the speaker, listener, sentence uttered and recording conditions were varied. To remove as much as possible any bias on the part of a listener, the utterances were randomly selected among all values of the above parameters. This preference ranking method is described in detail in [22]. Denoting a preference of method A over method B by A > B it is seen from a plot of the average of the preference over all parameters (keeping the detection method fixed) versus the detection method that

original utterance>SAPD>LPC>AMDF>PPROC>AUTOC>SIFT>CEP>DARD .

Also, with respect to this average, the original utterance scores considerably better than any of the eight LPC synthesized utterances, and the variation of the average preference among these eight methods is not as great. Moreover, the standard deviation in preference scores is much larger for the eight detection methods than for the natural utterance. Plots of the average preference score versus detection method used, keeping not only the detection method but also either of the listener, speaker, recording conditions, fixed, are also shown in [22]. Variations in preference scores among speakers are seen to be larger than variations among recording conditions and these are in turn larger than those among either listeners or sentence uttered.

Another comparison experiment, in which the mean preference for utterances synthesized with smoothed pitch contours over those synthesized with unsmoothed pitch contours is plotted versus the pitch detection method, was carried out in [22]. The same general trend concerning the preference scores keeping the sentence uttered, listener, speaker and recording conditions fixed, respectively, is observed in this experiment. Generally speaking, the higher an utterance scores in the previous experiment, the lower is its need for pitch smoothing in order to improve its subjective quality.

In conclusion, the fact that no LPC synthesized utterance comes close in quality to the original utterance should not be surprising in view of the discussion in section 2.3 on the limitations of linear prediction. Further work on pitch extraction algorithms is also necessary in view of the fact that on the average, the semi-automatic pitch contour method scores higher than the seven pitch detectors.

3.2 The SIFT Algorithm

From the previous discussion of section 3.1 on subjective testing, it is clear that SIFT is not a particularly good algorithm for pitch extraction. However, as the quantization properties of the reflection coefficients and some of their transformations is the subject of this thesis, the particular pitch extraction algorithm to be chosen is not of prime concern. Besides, implementations of SIFT by two FORTRAN subroutine programs were readily available for use in [2, Chapter 8]. Therefore, this algorithm will now be discussed in some detail.

First, it is observed that direct extraction of the pitch from the speech signal $s(n)$ can be done manually and is quite accurate. However for the purpose of implementing an automatic procedure of pitch extraction, the logical step to follow is to compute the autocorrelation

$$R(j) = \sum_{n=0}^{N-1-|j|} s(n)s(n+j)$$

where the interval $(0, N-1)$ includes many pitch periods. Obviously, $R(0) \geq R(j)$. Suppose there is a priori knowledge of the interval $J \subset (0, N-1)$ in which the pitch value should lie. Then compute $R(j)$ for all $j \in J$ and assign the value ℓ to the pitch where ℓ satisfies

$$R(\ell) = \max_{\substack{j \in J, \\ j \neq 0}} R(j)$$

Notice that if the gain R(0) changes by a constant factor α then so does any R(j). Because $R(0) \geq R(j)$ the normalization R(j)/R(0) can then always be compared with a fixed threshold function D(j) independent of gain. Unfortunately, the poles of the vocal tract transfer function have narrow bandwidths (especially those of low frequency). Therefore components of the speech waveform at those frequencies will not decay considerably within a pitch period. High amplitude correlation peaks due to those components could result in false pitch detection [9].

Inverse filtering [9]

This is simply linear prediction and ensuing inverse filtering of the speech signal s(n). Autocorrelation is then performed on the error signal. Gain normalization is then applied and a simple voiced-unvoiced decision based upon a fixed threshold function D(j) can be defined. In this way, most of the source vocal tract interaction is eliminated. Refinements of the method have led to the simplified inverse filter technique (SIFT) [9].

SIFT

Preliminaries [10]. Before performing linear prediction analysis the mean of the input signal within the analysis frame is extracted and subtracted from each sample value. If this was not done, the bias in the windowed frame would contribute to $R(j)$, a linear term monotonically decreasing in j. By its presence it is possible that a peak which would otherwise be below the threshold $D(j)$, could cross it and have an amplitude greater than a peak to its right corresponding to the actual pitch value. It is also possible that the threshold $D(j)$ is exceeded for a value of j smaller than the highest fundamental frequency of interest.

If the speech energy in the frame is less than some number called the lower dynamic range, then the frame is defined as silence. This allows the number of computations involved in linear prediction analysis and pitch extraction to be greatly reduced because of the substantial fraction of silence frames even in continuous speech. The same lower bound is used in gain quantization (see Chapter VI).

Finally, if the zero crossing density exceeds 2/ms, the frame is defined as unvoiced. This is because in unvoiced frames, the source of excitation has higher frequency components than for voiced frames, corresponding to a zero crossing density of at least 2/ms.

Human pitch for the average male or female speaker ranges from 50 to 250 Hz. The input speech can then safely be bandlimited (prior to the above preliminaries) to 1 KHz without any loss of pitch information. As will become clearer in Chapter IV, a sampling frequency f_s of 2 KHz and a filter order M=4 is sufficient for the linear prediction analysis. The advantage of this approach lies in the great reduction in the total number of necessary operations in the analysis. This scheme does not work well in the case of nasal or voiced plosive sounds because the speech signal contains zeroes around the frequencies of human pitch. To cancel this zero spectrum a pre-emphasis filter $1-z^{-1}$ is used before performing linear prediction [2, p. 193-197]. To get the filter coefficients, the input speech is also windowed using a Hamming window in order to obtain a more accurate representation of the speech spectrum. Then the error signal is obtained by inverse filtering the unwindowed and nonpre-emphasized speech signal. If the filter order M had been chosen to be much larger for such a bandlimited signal then the output would have been a unit sample $(e(n) = \delta(n))$ because

$$\frac{\alpha_M}{|A(e^{j\theta})|} \rightarrow |S(e^{j\theta})|$$

as $M \rightarrow \infty$ for autocorrelation linear prediction. The length of the analysis frame should encompass several pitch periods yet be small enough to ensure that the vocal tract does not change shape appreciably within the frame, and that pitch variation from pulse to pulse is insignificant. At $f_s = 2$ KHz 80 samples are used. The autocorrelation sequence is then twice that long but is symmetrical $R(j) = R(-j)$.

Interpolation

The sampling period T is .5 ms. Taking a typical pitch period P to be of the order of 6 ms [9] the quantization error in Hertz is

$$\left|\frac{1}{P+T/2} - \frac{1}{P}\right| = \frac{1}{P}\left|\frac{1}{1+T/2P} - 1\right| \sim \frac{T}{2P^2} = \frac{.25\text{x}1000}{6\text{x}6} = 7 \text{ Hz}$$

which is large enough to be noticeable. Since increasing the sampling frequency is undesirable a more accurate peak value and location is obtained from a simple parabolic interpolation of the maximum autocorrelation $R(\ell)$ and its two adjacent samples [9].

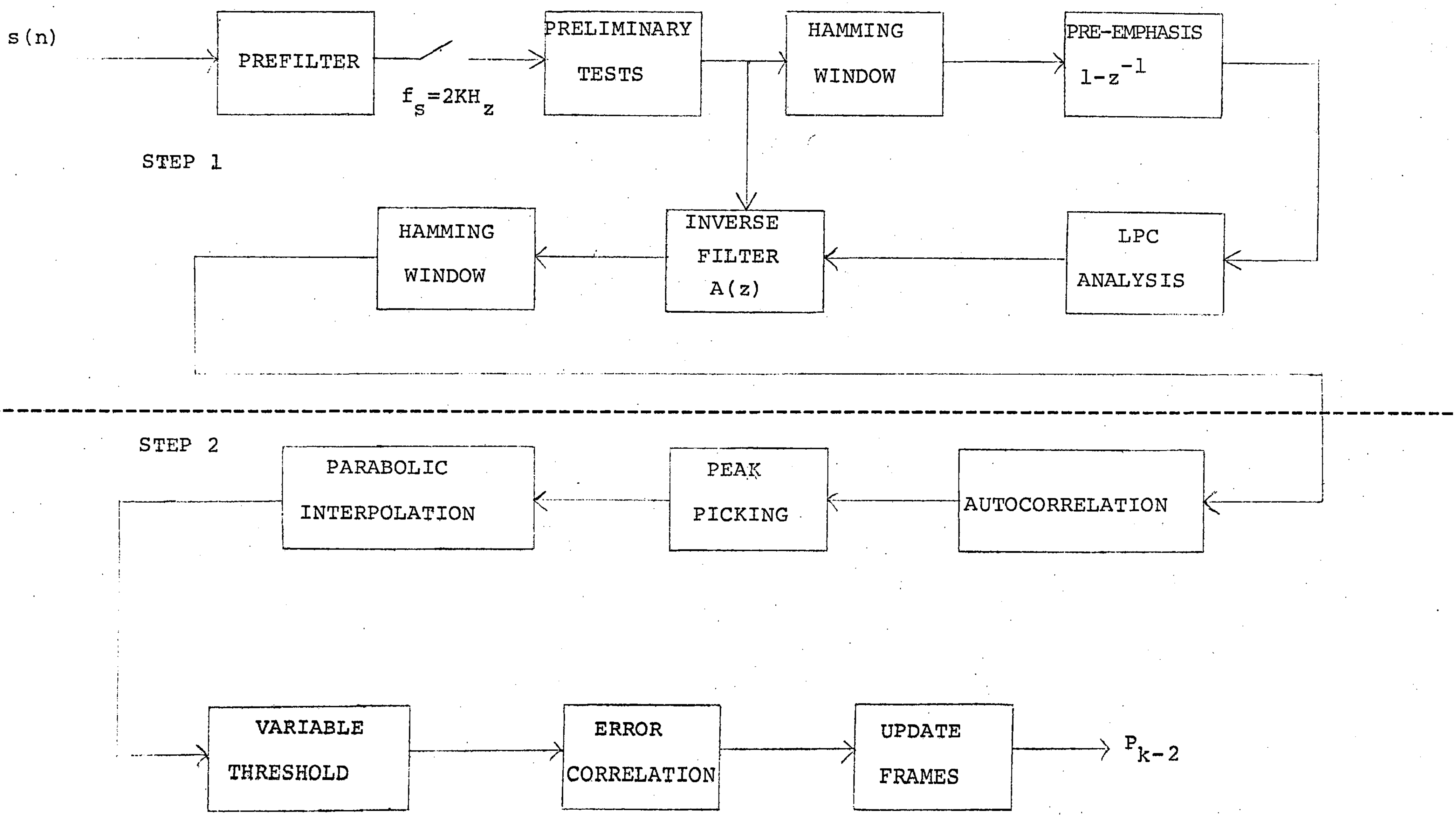
A block diagram of the SIFT algorithm is shown in Figure 3.2.1.

The variable threshold $D(j)$ and the error detection and correction logic are discussed in more detail in [2, Chapter 8]. In addition STEP 1 and STEP 2 of Figure 3.2.1 are implemented as two FORTRAN subroutine programs.

As a tradeoff between complexity and accuracy, SIFT uses only two frames of delayed pitch information for the

SIFT

Figure 3.2.1



detection and correction of errors. To further reduce the amount of computation involved, SIFT only searches pitch values over the range (50,250) Hz even though human pitch can go as high as 500 Hz.

Because linear prediction results are very sensitive to recording conditions [10], any type of background noise (including more than one speaker) must be kept to a minimum. Otherwise the performance of the SIFT algorithm will be considerably degraded. For the same reason, because of the binary voiced-unvoiced classification of each frame, implicit in linear prediction, voiced plosive and fricative sounds cannot be well reconstructed.

It should be pointed out that a single parameter extraction from the error signal, as is done above, accounts for the largest reduction in the transmission bit rate of speech.

IV: ANALYSIS AND SYNTHESIS USING PITCH EXCITATION

In this chapter, the basic building blocks of a pitch-excited vocoder are reviewed. Section 4.1 essentially deals with preprocessing and input variables to either a covariance or autocorrelation analyzer: sampling frequency, filterorder, analysis frame length, frame rate, windowing and pre-emphasis of the input speech. In Section 4.2 the stabilizing property of the reflection coefficients is briefly discussed. In the next section, two important synthesis structures are described. One of them, the two-multiplier lattice structure becomes part of the pitch synchronous synthesizer briefly discussed in Section 4.5. The driving function to this synthesizer uses the gain matching criterion discussed in the previous section. Finally, in view of the fact that quantization properties of various transformations of the reflection coefficients will be the main topic of Chapters V and VI, this synthesizer program is adopted and Section 4.5 concludes by enumerating some characteristics of autocorrelation vocoders.

4.1 Analysis Conditions [2, sections 6.5.2-6.5.6]

In order to account for the most important formant structure of speech, a sampling frequency f_s of at least 6 KHz is necessary. If low intensity and high frequency fricatives sounds were to be represented, a high SNR and f_s = 20 KHz would be required unless the technique of selective linear prediction [2, chapter 6] was employed. As discussed earlier, to prevent any aliasing, the speech must be bandlimited to $|f| < f_s/2$. However, since the introduction of artificial zeroes in the spectrum is undesirable, a variable filter with a very sharp cutoff at $f = f_s/2$ is required.

A figure of merit for the filter order M is f_s(KHz) + 4. This can be accounted for in the following way. In relating linear prediction to the speech production model, an equation of the form

$$y_m^-(t+T) = \mu_m y_{m-1}^+(t) + y_{m-1}^-(t)$$

is derived in [2, Chapter 4]. $T = 2\ell/c$ where ℓ is the length of a uniform tube and c is the speed of sound. T represents the time it takes for a wave to traverse the length of a uniform tube and be reflected back to its starting point. However, in digital representation of speech the samples are spaced $1/f_s$ apart. In order to be aware of the existence of such a tube a resolution $1/f_s \leq 2\ell/c$ is required. Let the number of tubes be M. Then $M\ell = L$ is the distance from the glottis to the lips. For humans, $2L/c \sim 1$ ms. Hence $M \leq f_s$(KHz). In other words it is useless to use $M > f_s$ because no additional formants are present in the range $(0, f_s/2)$. The best that can be done is $M = f_s$(KHz). However there are 4 or 5 additional poles which are observed in the input speech spectrum and these are due to the glottal transfer function and lip radiation model. Therefore to represent these poles a filter order value of at least f_s(KHz) + 4 is used. For unvoiced speech the vocal tract formant structure does not stand out as clearly in the input speech spectrum. If unvoiced frames of speech are analysed, then a smaller value for M than the one above could be used to accurately represent speech. Also there might not be a contribution from the glottis.

The analysis frame length N is limited by the time varying nature of the vocal tract. For most speech sounds it should not exceed (15-20) f_s(KHz) [2, Chapter 6]. However it would be preferable for some voiced and especially plosive sounds to use a value of N/f_s (KHz) of only a few msec if accurate representation of these sounds is desired. As these values of N cover many pitch periods, absolute placement of the interval is unnecessary in both the covariance and autocorrelation methods. To accurately represent the continuous nature of speech, a frame rate f_r of at least

50 Hz is recommended. Hence for a typical f_s of 10 KHz, $f_s/f_r = 200$ and with the above values of N, shifted intervals do not overlap. This is to be contrasted with the SIFT algorithm in which the overlap ratio is 1/2 (N=80 and $f_s/f_r = 2\text{KHz}/50\text{Hz} = 40$).

As was previously mentioned, windowing of input speech reduces the distortion between the actual and truncated speech spectra. Specific details about these distortions depend on the shape and length of the windows. For analysis lengths of order of magnitude as stated above, non-rectangular windowing of the speech is desirable.

Recall that an approximate way to account for the effect of glottal transfer function and lip radiation model on the output speech is to divide the all pole filter $1/A(z)$ of a vocal tract transfer function with zero lip impedance and infinite glottal impedance by the term $1-z^{-1}$. Since performing linear prediction to obtain the original all pole filter $1/A(z)$ is desirable the input speech is then preemphasized by a factor $1-z^{-1}$. This will lower the energy of the low frequency part of the spectrum. However, most unvoiced sounds contribute energy mostly to the high frequency part of the spectrum. For most of these sounds, the glottis does not contribute an all pole filter $1/(1-e^{-cT}z^{-1})^2$. There is then no reason to preemphasize the speech. Therefore, prior to the autocorrelation analysis an adaptive preemphasis filter $1-uz^{-1}$ where $u = r(1)/r(0)$, is used. $r(0)$ is the energy of the input speech in the analysis interval. For unvoiced sounds, the autocorrelation r(1) is much less than r(0) because there is practically no correlation among samples. There is then no preemphasis. For voiced sounds preemphasis is greatest because $r(1) \lesssim r(0)$. [2, Chapter 6].

4.2 Stability Problems and Comparison of Autocorrelation and Covariance Analyses

Recall from Section 2.2 that the parameters k_m involved in the solution to the autocorrelation linear prediction equations are termed reflection coefficients because they represent the fraction of the energy which is reflected at a boundary between two uniform tubes. More precisely it was found in [2, Chapter 4] that

$$k_m = \frac{1-A_m/A_{m-1}}{1+A_m/A_{m-1}}$$

where A_m is the cross-section of the mth uniform tube. An area is a positive quantity and therefore from simple inspection of the above equation, $|k_m|<1$, as is required from physical grounds since apart from the glottal input, there is no additional source of energy. This result can also be seen from (2.1.14) since $\alpha_m = \beta_m$ in the autocorrelation analysis and therefore the equation reduces to

$$\alpha_{m+1} = \alpha_m(1 - k_{m+1}^2) \tag{4.2.1}$$

But α_m is a sum of squares and is always positive. Hence $|k_m| < 1$ for all m and consequently stability is ensured. (A more rigorous proof relating the condition $|k_m| < 1$ to the requirement that the roots of A(z) lie inside the unit circle $|z| < 1$ for stability of 1/A(z), can be found in [2, Chapter 5].) This result does not generally hold for the covariance method since α_m is not necessarily equal to β_m in (2.1.14). However, combining (2.1.8) and (2.1.19) yields

$$A_{m-1}(z) = \frac{A_m(z) - zk_m B_m(z)}{1-k_m^2} \tag{4.2.2}$$

and using (2.1.18), (4.2.2) can be rewritten as

$$A_{m-1}(z) = \frac{A_m(z) - z^{-m} k_m A_m(1/z)}{1-k_m^2} \tag{4.2.3}$$

or in time-domain notation

$$a_{m-1,i} = \frac{a_{mi} - k_m a_{m,m-i}}{1-k_m^2} \tag{4.2.4}$$

for m = M, M-1, ... 1 and i = 0, 1, ..., m-1. Therefore all $A_m(z)$ can be found given A(z). But from (2.1.17), $a_{mm} = k_m$. Therefore if a filter A(z) is obtained by the covariance method, the step down recursion 4.2.4 can be used to test for a possible occurrence of at least one

$|k_m| > 1$. If there is one, A(z) is expanded in product form, and for the roots z_i which lie outside the unit circle, let $z_i = 1/z_i$. Then reconstruct the new polynomial A(z). If reflection coefficients k_m are to be used in transmission apply (4.2.4) once more to find all k_m for the new A(z). It must be noted that this new A(z) does not satisfy the original minimization criterion. The above procedure is called the step down-step up method. The advantages of the autocorrelation over the covariance method are therefore (1) the filter is assured to be stable, (2) a useful gain matching is easily computed and (3) for the same analysis frame length, it requires less calculations. However, the quality of the synthesized speech is often lower than that of the pitch synchronous covariance analysis. i.e., a frame of duration less than a pitch period [2, section 10.3.3]. However the gain calculation in the covariance analysis may require a larger frame of data [2, Section 6.5.1]. Notice that both methods should give similar results as the frame length increases because then c_{ij} differs from r(i-j) only in the end terms in the summation over (n_0, n_1).

4.3 Synthesis Structures [2, sections 5.4,5.5]

Up to now, analysis has been discussed. However, many of the ideas involved in linear prediction can be used in the inverse problem of synthesizing speech. First assume that an all-pole linear prediction filter $1/A(z)$ and an arbitrary input signal $E(z)$ to this filter are given. Then the output is

$$D(z) = \frac{E(z)}{A(z)} = \frac{E(z)}{1+\sum_{i=1}^{M} a_i z^{-i}} \tag{4.3.1}$$

or in time domain notation

$$d(n) = e(n) - \sum_{i=1}^{M} a_i d(n-i) \tag{4.3.2}$$

The idea in synthesis is to compute $d(n)$ consecutively for a certain range of n, given the input $e(n)$ and the filter coefficients a_i, $i = 2, 2, \ldots M$, and updating the a_i's at the first n outside the above range. Notice that, by the above computation 4.3.2, the memory $d(n-1), \ldots, d(n-M)$ is updated for every new input $e(n)$. In the SIFT algorithm there is such a filter memory used in the computation of the error signal:

$$E(z) = S(z)\ A(z) \tag{4.3.3}$$

or

$$e(n) = s(n) + \sum_{i=1}^{M} a_i s(n-i) \tag{4.3.4}$$

For every analysis frame of length N there are N new samples s(1) .... s(N) but for n=1 it must be decided which values should be assigned to the memory s(0), s(-1) .... s(-M). These are chosen to be zero at the initiation of every frame.

The computation scheme (4.3.2) is called the DIRECT FORM synthesis structure. Now the parameters which are often transmitted to the receiver are the reflection coefficients k_i. As can be understood from the previous discussion, this is because stability is guaranteed under quantization of the k_i's in the open interval (-1,1). Therefore a scheme which computes the output speech samples directly from the k_i's should be sought. Such a method is presented below and is called the TWO-MULTIPLIER LATTICE structure. First rewrite (2.1.8) and (2.1.19) as

$$A_{m-1}(z) = A_m(z) - k_m B_{m-1}(z) \qquad (4.3.5)$$

and

$$zB_m(z) = k_m A_{m-1}(z) + B_{m-1}(z) \qquad (4.3.6)$$

Combining (2.2.6) and (2.2.7) gives

$$A_o(z) = zB_o(z) = 1 \qquad (4.3.7)$$

Multiply (4.3.5-4.3.7) by E(z)/A(z) and let

$$E_m^+(z) = A_m(z)E(z)/A(z) \quad (4.3.8)$$

$$E_m^-(z) = B_m(z)E(z)/A(z) \quad (4.3.9)$$

(4.3.8) and (4.3.9) are the z transform of the forward and backward predictor respectively. Equations (4.3.5-4.3.7) then become

$$E_{m-1}^+(z) = E_m^-(z) - k_m E_{m-1}^+(z) \quad (4.3.10)$$

for m=M, M-1,...1

$$zE_m^-(z) = k_m E_{m-1}^+(z) + E_{m-1}^-(z) \quad (4.3.11)$$

$$E_\circ^+(z) = zE_\circ^-(z) \quad (4.3.12)$$

In z^{-1} transform notation it is written

$$e_{m-1}^+(n) = e_m^+(n) - k_m e_{m-1}^-(n) \quad (4.3.13)$$

for m=M, M-1,...1

$$e_m^-(n+1) = k_m e_{m-1}^+(n) + e_{m-1}^-(n) \quad (4.3.14)$$

$$e_\circ^+(n) = e_\circ^-(n+1) \quad (4.3.15)$$

The k_m are only updated after a certain value of n. The input to the synthesis structure is $e_M^+(n)$ and the memory is

$e^-_{M-1}(n)$, ..., $e^-_\circ(n)$. The output $e^+_\circ(n)$ can be calculated recursively in the order of decreasing m by the sole use of equation (4.3.13). Equation (4.3.14) and (4.3.15) compute the new memory $e^-_{M-1}(n+1)$,..., $e^-_\circ(n+1)$ to be used with the next input $e^+_M(n+1)$. The two-multiplier lattice structure was implemented in [2, Chapter 5] as a Fortran subroutine program and will be used in Chapter VI for the reconstruction of speech which was analyzed by the autocorrelation linear prediction method. Other practical structures exist which are simple modifications of the above two-multiplier lattice. [2, Chapter 5]

4.4 The Driving Function to the Synthesizer [2, section 10.2.4]

For the purpose of speech transmission it would be possible to use the error signal itself as input to the synthesizer. Figure 4.4.1.

$$S(z) \longrightarrow \boxed{A(z)} \xrightarrow{E(z)} \boxed{1/A(z)} \longrightarrow S(z)$$

However the quantization of the error signal for its subsequent transmission would result in an excessive bit rate. To obtain a relatively low bit rate, pitch extraction from the input $\acute{s}(n)$ is suggested. The pitch estimate as obtained, say, by the SIFT algorithm is transmitted along with the reflection coefficients and the gain information through the channel. At the receiver a sequence $e(n)$ is constructed from the pitch and gain information. Basically if the frame is unvoiced a randomly generated sequence e(n) is chosen as input to the synthesizer and if it is voiced it will consist of fixed amplitude samples equally spaced by the pitch value $P(ms)f_s(KHz)$ where P(ms) is obtained from the pitch extractor and f_s is the sampling frequency of the output speech. The gain of the output speech is to be calculated subject to some matching criterion. One suggestion is to match the energy of the input speech to the analysis filter, to that of the output speech at the receiver within each consecutive interval of length equal to a pitch period [2, Chapter 10]. Transient contributions to the gain from one previous pitch period are taken into account. The disadvantage of the approach is that there is no guarantee that the gain will not vary discontinuously from one pitch period to the next. Also notice that in addition to the filter coefficients and pitch period information, the transmitter has to send the gain information for all pitch periods encompassed by the length of the frame. If the frame is unvoiced then the situation is simpler in that the pitch period can be assigned the value of the synthesis frame length since there is no memory involved. In the synthesizer program of [2, Chapter 10] the synthesis frame length is f_s/f_r (the same number as used for the elapsed time before an analysis frame is updated). It employs a different gain matching method based on the error signal energy per analysis frame, namely α. If the frame is unvoiced, the excitation is provided by randomly generated samples g(n). The mean Eg(n) is set to zero and a uniform probability distribution over a range (-b,b) is utilized for g(n):

$$\int_{-b}^{b} p(x)\,dx = A.2b = 1 \quad \text{or } p(x) = 1/2b$$

$$Eg^2(n) = \sigma_g^{\ 2} = \int_{-b}^{b} x^2.\ \ 1/2b\ dx = 1/2b\ \frac{x^3}{3} = \frac{b^2}{3}$$

The gain of the excitation e(n) is then matched by

$$\frac{e(n)}{\sqrt{\alpha/N}} = \frac{g(n)}{\sigma_g} \tag{4.4.1}$$

where N is the frame length in the analysis. If the frame is voiced, an excitation consisting only of fixed amplitude samples equally spaced by a pitch period, will not have a zero mean. To force it to have a zero mean, a fixed amplitude of opposite sign is assigned to the remaining samples. More quantitatively, let C_1 and C_2 be these two respective amplitudes. With an analysis frame length N and a pitch period I there are then N/I samples of amplitude C_1 and N-N/I samples of amplitude C_2. Then with the same gain matching criterion as used for unvoiced speech, plus the zero mean requirement, there are two constraint equations in C_1 and C_2:

$$C_1^2(N/I) + C_2^2(N-N/I) = \alpha \tag{4.4.2}$$

$$C_1(N/I) + C_2(N-N/I) = 0 \tag{4.4.3}$$

Solving these two equations yields

$$C_1 = \sqrt{\alpha/N}\ \sqrt{I-1} \tag{4.4.4}$$

and

$$C_2 = -\sqrt{\alpha/N}\ /\sqrt{I-1} \tag{4.4.5}$$

4.5 A Pitch Synchronous Synthesizer

A synthesizer has been implemented as a FORTRAN program in [2, Chapter 10]. It performs pitch-synchronous linear interpolation of the gain, pitch and reflection coefficients from the present and previous frames. The idea behind this is that speech of better quality can be obtained by smoothening out discontinuities in going from one frame to the next. Because reflection coefficients are inputted, the two-multiplier lattice synthesis structure implemented as a subroutine program is utilized. A constant postemphasis value of .9, an analysis frame length N of 128 and a synthesis frame length of 64, are used. For unvoiced excitations, gain matching criterion (4.4.1) is employed while for voiced excitation, the constants C_1 and C_2 are obtained by solving the equations

$$C_1^2 N/I = \alpha$$

and (4.4.3) simultaneously. (This is only slightly less accurate than solving 4.4.2 and 4.4.3 since $C_1 >> C_2$). To obtain the results of Chapter VI, the above program was used, with only slight modifications. The value of f_s/f_r in both the analysis and synthesis, is 200. Also, if an analysis frame was pre-emphasized by a factor μ, then the corresponding frame in the synthesis will be post-emphasized by the same factor:

$$x(n) = y(n) - \mu y(n-1) \quad y(n) \text{ is pre-emphasized}$$

$$y(n) = x(n) + \mu y(n-1) \quad x(n) \text{ is post-emphasized}$$

If a frame is voiced, then the constants C_1 and C_2 are obtained by solving 4.4.2 and 4.4.3. There is no Hamming window $w(n)$ in the above synthesizer program. However, it was introduced in the analysis, for better spectral representation of speech, and this reduces the gain of the input speech by a factor $\sum_{n=1}^{N-1} w^2(n) \sim 1.58$ for the range of N under consideration. Taking this into account, the gain of the output speech was increased by a factor of 1.58.

4.6 Some Characteristics of Autocorrelation Vocoders

Fig 4.6.1 is a block diagram of a basic pitch excited vocoder. Either covariance or autocorrelation analysis could be performed. The parameters are then quantized before being transmitted through the channel. More details on the transformations and quantization of parameters will be given in Chapter V.

Markel and Gray have used autocorrelation analysis and the SIFT algorithm as the pitch extractor [10]. A summary of the results in [10] is now presented. The sampling frequency, preemphasis and windowing considerations already mentioned were taken into account. From the analysis, the reflection coefficients are obtained and are linear quantized while the pitch and gain are logarithmically quantized [see Chapter V]. After quantizing, the speech was synthesized as described under Section 4.5. Even though interpolation is important for speech quality it can cause blurring of fast transitions from one class of sound to another. Fixed frame analysis can cause errors in the timing and gain of some plosive sounds. Fricative sounds are more difficult to represent in view of their voiced-unvoiced character. As will be seen in Chapter V spectral distortion of speech is important in its perception. Consequently, it is more important to have an accurate spectral representation of the original speech utterance rather than to have an accurate

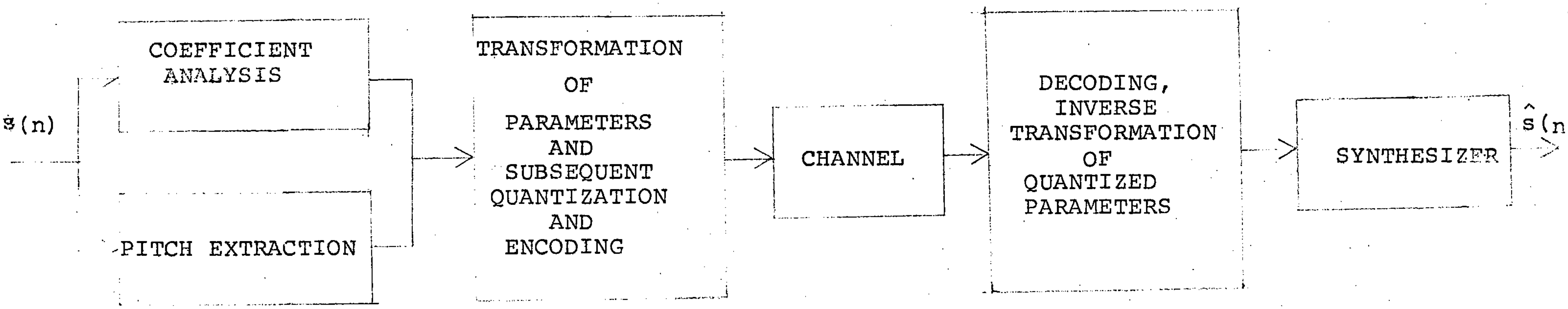


Figure 4.6.1. Pitch Excited Vocoder

temporal structure. Part of this distortion in the temporal domain is due to using too simplified a gain matching criterion [Equations 4.4.1, 4.4.2] in the synthesizer program. This can be remedied by replacing it with a more accurate criterion [2, Chapter 10]. As the error signal contains most of the information in speech it is important that the artificial excitation input e(n) to the synthesizer matches it as closely as possible if the output speech is to be almost indistinguishable from the original utterance. In cases where the match between the error signal peaks and those of e(n) is good (voiced speech) it is observed that perceivable differences are much smaller [10]. It is concluded that for bit rates as low as 3300 bits/sec, the quality of synthesized speech is good in general. Between 1400 and 3300 bits/sec the degradation in the quality depends on the particular speaker and also on the speech content. Unless variable, bit rate analysis is used, synthesized speech is unintelligible at bit rates under 1400 bits/sec. It is possible to use variable bit rate transmission because of the large number of silence and unvoiced intervals requiring less spectral information, even in continuous speech (See equal area coding Chapter V). A particular variable bit rate scheme [2, section 10.3.2] was used involving a maximum likelihood distance measure which will also be discussed in Chapter V. The filter order M is also variable and Huffman coding is performed on the quantized parameters in order for the average bit rate to approach their entropy [13]. An average bit rate of 1500 bits/sec was then achieved although the analysis frame rate was as high as 100 Hz. The quality of the output speech was even better by using time synchronous instead of pitch synchronous, interpolation of the parameters.

V: QUANTIZATION*

In section 5.1 the basic properties of the log spectral deviation measure, are reviewed, in view of their application to speech parameter quantization. The emphasis is on their behavior in fine quantization. After a sensitivity function and deviation bound are defined for single parameter quantization, two fidelity criteria, the maximum and expected spectral deviation bound, are introduced [12,14]. Non-asymptotic and asymptotic results involving these criteria are then derived. Section 5.2 then briefly enumerates the properties of different sets of parameters that have found use in quantization. One of these, the set of reflection coefficients, is then the subject of section 5.3. Several quantization schemes are discussed. First, there is uniform and equal area quantization. Then inverse sine and log area ratio quantization [14] are shown to be optimal in the sense of minimizing the maximum spectral deviation bound criterion. After an alternative scheme, the two-parameter quantization method [14], is presented, overall deviation bounds in terms of the above single parameter deviation bounds are derived in order to determine the optimum bit allocation among the parameters. Two parameter quantization is then shown to be superior to log area or inverse sine quantization, in terms of bit rate, for the same quality of speech. The

*In the following, except where specifically mentioned, autocorrelation linear prediction is assumed.

bit rate results of [12], where the fidelity criterion is the expected spectral deviation bound, are then summarized. As entropy coding does not reduce the bit rate substantially, decorrelation of the reflection coefficients is suggested. Section 5.4 first describes the eigenvector analysis method of [18] for decorrelating the reflection coefficients within a frame. The DPCM technique is briefly mentioned. The theoretical development which led to the experimental results of Chapter VI is then introduced. Using the quantization scheme which minimizes the expected spectral deviation bound in the asymptotic limit, on the decorrelated parameters resulting from the eigenvector analysis of [18], it is hoped that a lower total bit rate can be achieved. The eigenvector analysis will be carried out by the Jacobi method for diagonalizing a matrix. The sensitivity function of the decorrelated parameters is then derived. Next, assumptions involving the probability density function and also the average sensitivity function of these parameters are made. One difficulty concerning the average sensitivity is then resolved, and an alternative, more accurate method of obtaining the density and average sensitivity function is proposed, based on time averages. These results are then substituted in the already derived formulae for the quantizer curve function and the number of levels. These time averages are also computed for the reflection coefficients themselves as these will also be quantized for a comparison of their performance with that of the decorrelated parameters.

5.1 Introduction to Distortion Measures and Fidelity Criteria

It is desired to greatly reduce the bit rate in transmission of speech, subject to the requirement that no difference between the original and synthetic speech shall be perceived. Unfortunately, the perception mechanism is extremely complicated and far from being understood. It will therefore be necessary to work with empirical distortion measures which describe some aspects of the hearing mechanisms. Many of these distortion or distance measures find use in both quantization and variable frame rate transmission. A few of the most commonly used ones will now be discussed.

Consider a set $\{a_i\}$ of filter coefficients or any transformation of them. These will be discussed later. One distortion measure is based on the difference $a_i - a_i'$. For example in variable frame rate transmission the fidelity criterion could be $\sum_{i=1}^{M} (a_i - a_i')^2$ where a_i belongs to one frame and a_i' belongs to the adjacent one [2, section 10.2.3]. If this quantity is smaller than some prescribed number, then no information is sent to the receiver and the synthesizer reconstructs the speech using the previous frame's parameters. It has been shown experimentally that poor results are obtained unless the parameters used are the cepstral coefficients [11]. As will be shown, this is because of their relationship to a spectral deviation which has been successful in bit rate reduction. In single parameter quantization (letting x stand for the parameter a_i) the following fidelity criterion has been used [12]:

$$M_p = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} |x-\hat{x}_n|^p p_X(x)\,dx$$

where N, x_n, $\hat{x}_n$, p_X stand for the number of levels, the boundary values, the levels and the probability density function of x respectively and p is an arbitrary integer. Subject to this constraint, it can be shown that as $N \to \infty$, uniform quantization of x will minimize the entropy defined by

$$H = -\sum_{n=0}^{N-1} P_n \log P_n \qquad \text{where } P_n = \int_{x_n}^{x_{n+1}} p_X(x)\,dx$$

Now, $H \leq \log N$, with equality iff $P_n = 1/N$, [13] and in cases where it is considerably less than log N, it becomes advantageous to reduce the bit rate to as close to H as possible by an appropriate scheme such as Huffman coding [13]. Uniform quantization has been applied to reflection coefficients and this will be described in more detail later.

Though the approach to be followed should be to minimize the entropy subject to a fidelity criterion [13], it is possible that a scheme which maximizes output entropy is successful in reproducing speech. In such a case, $P_n = \int_{x_n}^{x_{n+1}} p_X dx = 1/N$ and hence the scheme is also called equal area quantization. This has been applied on reflection coefficients and will be described later. The distance measure $|x-\hat{x}_n|^p$ is however not appropriate for speech reflection coefficients because it does not take into account gross spectral errors as $|x| = |k_i| \to 1$. (The condition $|k_i| < 1$ is required for stability). Hence a distance measure which takes into account the filter A(z) should be sought [12].

Spectral deviations [11]

Letting unprimed and primed variables correspond to different values of the same set of parameters in

$$V(\theta) \triangleq \ln \sigma^2/|A(e^{j\theta})|^2 - \ln \sigma'^2/|A'(e^{j\theta})|^2 \qquad (5.1.1)$$

a particular distance measure D is defined by the following relation:

$$D^p = \int_{\pi}^{\pi} |V(\theta)|^p \, d\theta/2\pi \qquad (5.1.2)$$

It is stated in [11] that a deviation D in the speech spectrum of at least 3 to 4dB is required in order to be able to perceive any difference between the original and synthetic speech. Now, as $p \rightarrow \infty$, $\sqrt[p]{D^p} \rightarrow |\Delta V(\theta)|_{max}$ [11]. This quantity is plotted in [11] versus $\sqrt[2]{D^2}$ for every 2 successive frames in an utterance with the following analysis conditions $f_s = 6.5$ KHz, $M = 10$, $N = 120$, $1/f_r = 20$ ms. The cross correlation coefficient was then measured to be .84. It was concluded that the choice of p is not significant. Hence $p = 2$ was selected because known properties of analytic functions can be used to evaluate D^2 in terms of an infinite summation instead of having to use an approximation for the integral in (5.1.2). (This would involve the use of two FFT's for the evaluation of $A(e^{j\theta})$ and $A'(e^{j\theta})$. Since

$$A(z) = \sum_{i=0}^{M} a_i z^{-i}, \qquad a_i \in R$$

$$\log A(z) = - \sum_{K=1}^{\infty} \hat{a}_K z^{-K}$$

where $\hat{a}_K$ is a cepstral coefficient.

It is then shown in [11] that,

$$\log \sigma^2/|A(e^{j\theta})|^2 = \log \sigma^2 - \log A(e^{j\theta}) - \log A(e^{-j\theta})$$

$$= \sum_{K=-\infty}^{\infty} \hat{a}_K e^{-jK\theta} \qquad \text{where } \hat{a}_K = \hat{a}_{-K}$$

$$\text{and } \hat{a}_o = \log \sigma^2$$

Consequently,

$$D^2 = \int_{-\pi}^{\pi} (\Sigma \hat{a}_K e^{-jK\theta} - \Sigma \hat{a}_K' e^{-jK\theta})^2 \, d\theta/2\pi$$

$$= \sum_K \sum_m \int_{-\pi}^{\pi} (\hat{a}_K - \hat{a}_K')(\hat{a}_m - \hat{a}_m') e^{-j(K+m)\theta} d\theta/2\pi$$

$$= \sum_{K=-\infty}^{\infty} (\hat{a}_K - \hat{a}_K')^2 \tag{5.1.3}$$

Since a computer only sums a finite number of terms, only the most important contributions are summed over. As already mentioned in a previous chapter the $\hat{a}_n$'s decay as C^n/n. Since D^2 is an infinite sum of squares, such a finite approximation is a lower bound to D^2. This representation of D^2 is used in variable frame rate transmission. In quantizing speech, however, the main interest is in the behavior of D in the limit of small perturbations in the values of the parameters. Going back to (5.1.2), assume $\sigma^2 = \sigma^2(\underline{\lambda})$ and $A = A(e^{j\theta}; \underline{\lambda})$ where $\underline{\lambda}$ is a vector of parameters $(\lambda_1, \lambda_2 \ldots \lambda_L)$ [14].

Next, rewrite $V(\theta)$ as

$$\ln \frac{\sigma^2(\underline{\lambda}+\Delta\underline{\lambda})}{\sigma^2(\underline{\lambda})} - \ln \frac{A(e^{j\theta};\underline{\lambda}+\Delta\underline{\lambda})}{A(e^{j\theta};\underline{\lambda})} - \ln \frac{A^*(e^{j\theta};\underline{\lambda}+\Delta\underline{\lambda})}{A^*(e^{j\theta};\underline{\lambda})}$$

In the calculation of the gain using the correlation matching of section 2.1, it was shown that

$$\oint_{|z|=1} \ln A(1/z)\; dz/2\pi jz = 0 \qquad (5.1.4)$$

From these analyticity properties and again in the case $p = 2$, rewrite (5.1.2) as

$$D^2 = 4\left[\ln \frac{\sigma(\underline{\lambda}+\Delta\underline{\lambda})}{\sigma(\underline{\lambda})}\right]^2 + 2\int_{-\pi}^{\pi} \left| \ln \left(\frac{A(e^{j\theta};\underline{\lambda}+\Delta\underline{\lambda})}{A(e^{j\theta};\underline{\lambda})}\right) \right|^2 \frac{d\theta}{2\pi} \qquad (5.1.5)$$

A conventional method which quantizes the gain independently will be discussed in Chapter VI, and since its contribution to D^2 is additive, $\sigma(\underline{\lambda})$ is normalized to 1. Then, writing $A(e^{j\theta};\underline{\lambda}+\Delta\underline{\lambda})$ as $A(e^{j\theta};\underline{\lambda}) + \Delta A(e^{j\theta})$, using the approximation $\ln(1+x) \sim x$ for small x since $\Delta\lambda$ is infinitesimal

$$D^2 = 2\int_{-\pi}^{\pi} \left| \frac{\Delta A(e^{j\theta})}{A(e^{j\theta};\underline{\lambda})} \right|^2 d\theta/2\pi \qquad (5.1.6)$$

This expression is also involved in another distance measure discussed in [11]. (5.1.2) with p = 1, and the following distance measure (where α is the minimum energy of the error signal)

$$D = \ln \left[\frac{1}{2\pi} \frac{\alpha(\underline{\lambda})/\alpha_o(\underline{\lambda})}{\alpha(\underline{\lambda}+\Delta\underline{\lambda})/(\alpha_o(\underline{\lambda}+\Delta\underline{\lambda})} \int_{-\pi}^{\pi} \left|\frac{A(e^{j\theta};\underline{\lambda}+\Delta\underline{\lambda}}{A(e^{j\theta};\underline{\lambda})}\right|^2 \frac{d\theta}{2\pi}\right] \tag{5.1.7}$$

have been used in the quantization studies of [15]. It will be discussed later in connection with reflection coefficient quantization. Denote D explicitly as D(.,.) where the two arguments will refer respectively to different values of the same set of parameters. Then it can be shown that (5.1.2) satisfies the following properties:

$$D(\underline{\lambda},\underline{\lambda}) = 0 \tag{5.1.8a}$$

$$D(\underline{\lambda},\underline{\lambda}') = D(\underline{\lambda}',\underline{\lambda}) \tag{5.1.8b}$$

$$D(\underline{\lambda},\underline{\lambda}') > 0 \qquad \underline{\lambda} \neq \underline{\lambda}' \tag{5.1.9}$$

$$D(\underline{\lambda},\underline{\lambda}'') \leq D(\underline{\lambda},\underline{\lambda}') + D(\underline{\lambda}',\underline{\lambda}'') \tag{5.1.10}$$

Properties (5.1.8) - (5.1.9) are almost self-evident from the form of (5.1.2). (5.1.10) is the continuous analog of the triangle inequality, whose proof can be found in [20].

Independent parameter quantization [12]

As it is much easier to obtain quantizer curves in the asymptotic limit of a large number of levels, (5.1.6) can be the starting point of the analysis instead of (5.1.2). In terms of a single parameter variation, the following sensitivity function is then defined

$$s_{\Lambda}(\lambda) = \lim_{\Delta\lambda \to 0} \frac{D(\lambda,\lambda+\Delta\lambda)}{|\Delta\lambda|} \tag{5.1.11}$$

in which λ stands for a single parameter. Also define

$$\bar{D}(x,y) = \left| \int_x^y s_{\Lambda}(\lambda)\, d\lambda \right|$$

The following proof is from [12]. Let D stand for any measure like (5.1.2) which obeys properties (5.1.8) - (5.1.10).

Show that $\quad D(x,y) \leq \bar{D}(x,y) \qquad (5.1.12)$

Proof: $\quad D(x,\lambda+\Delta\lambda) \leq D(x,\lambda) + D(\lambda,\lambda+\Delta\lambda) \qquad (5.1.13)$

$$D(x,\lambda) \leq D(x,\lambda+\Delta\lambda) + D(\lambda+\Delta\lambda,\lambda) \tag{5.1.14}$$

Taken together, (5.1.13) and (5.1.14) imply

$$|D(x,\lambda+\Delta\lambda) - D(x,\lambda)| \leq D(\lambda,\lambda+\Delta\lambda)$$

or
$$\left|\frac{dD(x,\lambda)}{d\lambda}\right| \leq s_\Lambda(\lambda)$$

Using (5.1.8), (5.1.12) is then obtained.

Hence $\bar{D}(x,y)$ is an upper bound to $D(x,y)$. Also for $x \sim y$, $D(x,y) \sim \bar{D}(x,y)$. Recall that the fidelity criterion M_p used an inappropriate distance measure. Replacing it with $\bar{D}(x,y)$ the new fidelity criteria is then

$$E(\bar{D}) = \int_a^b \bar{D}(x, q(x))p_X(x)dx$$

$$= \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \bar{D}(x,\hat{x}_n)\ p_X(x)dx \qquad (5.1.15)$$

where x_n and $\hat{x}_n$ are the quantization boundaries and levels respectively, and (a,b) lies in the allowed range for X. The values to be chosen for x_n, $\hat{x}_n$, a, and b will be discussed later. As stated in [12] it is not clear as to whether or not the value of E(D) is close to that of its upper bound $E(\bar{D})$. Now as mentioned in [12], the minimization of

$E(\bar{D})$ with respect to all x_n, $\hat{x}_n$, keeping N fixed results in equations which require an iterative numerical technique for their solution. To avoid this procedure, the asymptotic case of large N is treated in [12] in order to get a closed form solution for the quantizer curve. Let $z = U(x)$ such that z is uniformly quantized in the range $U(a) = 0$ to $U(b) = 1$. Hence $z_n = n/N$ and $\hat{z}_n = (n+1/2)/N$ where z_n and $\hat{z}_n$ are the boundaries and levels respectively. Since the probability and sensitivity measures should not depend on the coordinates used, it is required that

$$p_Z(z)dz = p_X(x)dx \tag{5.1.16}$$

$$s_Z(z)dz = s_X(x)dx \tag{5.1.17}$$

Let $u(x) = dz/dx$. In the new coordinates, using (5.1.16) and (5.1.17), $E(\bar{D})$ becomes

$$\sum_{n=0}^{N-1} \int_{z_n}^{z_{n+1}} \left| \int_{z}^{\hat{z}_n} s_Z(\lambda)d\lambda \right| p_Z(z)dz \tag{5.1.18}$$

It is then shown in Appendix B of [12] that, in the asymptotic limit of large N, (after transforming back to the old coordinates)

$$E(\bar{D}) \sim \frac{1}{4N} \int_a^b s_X(x) \frac{p_X(x)dx}{u(x)} \qquad (5.1.19)$$

and

$$H \sim -\int_a^b p_X(x)\log[p_X(x)/u(x)]dx + \log N \qquad (5.1.20)$$

Using the Schwartz Inequality

$$\left|\int X(t)Y(t)dt\right|^2 \leq \int|X(t)|^2dt\int|Y(s)|^2ds$$

with equality iff $X = dY$, where d is a constant, and substituting $u(x)$ in $|Y|^2$, and $s_X(x)p_X(x)/u(x)$ in $|X|^2$ gives

$$\int_a^b \frac{s_X(x)p_X(x)dx}{u(x)} \geq \left|\int_a^b \sqrt{\frac{s_X(x)p_X(x)}{u(x)}} \cdot \sqrt{u(x)}\, dx\right|^2 \qquad (5.1.21)$$

Hence, for fixed N, $E(\bar{D})$ is smallest iff

$$\sqrt{\frac{s_X(x)p_X(x)}{u(x)}} = d\sqrt{u(x)}$$

Using the normalization $\int_a^b u(x)dx = U(b) = 1$,

$$u(x) = \frac{\sqrt{s_X(x)p_X(x)}}{\int_a^b \sqrt{s_X(\lambda)p_X(\lambda)d\lambda}} \qquad (5.1.22)$$

then achieves the global minimum of $E(\bar{D})$ [12]. Introducing another fidelity criterion, $\max\limits_{a \le x \le b} \bar{D}(x,q(x))$ it can be shown that

$$u(x) = \frac{s_X(x)}{\int_a^b s_X(\lambda)\,d\lambda}$$

minimizes the above criterion even for finite N. A proof of this is given in [15] and can also be found in Appendix A. This criterion will be discussed later in connection with reflection coefficients. This u(x) can also be shown to minimize the entropy H for fixed $E(\bar{D})$ in the asymptotic limit of large N. A proof of this is given in [12] and is also included in Appendix A.

The asymptotic results (5.1.19) and (5.1.20) together can be used to find H and N given a fixed value for $E(\bar{D})$ and a general quantizer curve U(x).

5.2 Characteristics of Various Parameters under Uniform Quantization [15]

The filter coefficients a_i or some transformation of them are then quantized before being transmitted through a channel. Using distance measure (5.1.2) with p = 1, results have been obtained and compared for commonly used transformations [15].

Some of these results will now be summarized:

(1) If the filter coefficients are themselves quantized, then the reconstructed filter at the receiver might very well be unstable. (The roots of a filter with quantized coefficients do not necessarily have to be within the unit circle). If such a method is employed, then very fine quantization would have to be used and thus the bit rate would be too high for transmission purposes.

(2) Similarly, the quantization of the auto-correlation coefficients of $a_i/\sqrt{\alpha}$ might result in an unstable filter.

(3) The DFT of the sequence in (2) once quantized gives superior results which are comparable to method (6) below.

(4) The cepstral coefficients, obtained from the a_i's, are then quantized and the inverse transformation is applied to give the modified a_i's. Instabilities are still possible although results are also (like (3) above) superior to (1) and (2).

(5) If the roots of A(z) are quantized to values within the unit circle, the instability problem is solved. Bandwidths are not as important as frequencies and so the

quantization of the absolute magnitude of the roots does not have to be as fine as for the frequencies. Unfortunately, the set of roots $\{z_i\}$ is not an ordered set like the other transformations which have been mentioned. In fact it is difficult to associate a root with a particular peak in a spectrum. In addition the computation of the roots of a high degree polynomial like A(z) is not easy.

(6) An alternative set of parameters [2, section 10.2.1] would be the autocorrelation sequence, r(n), of the input speech to the LPC analysis itself. Autocorrelation linear prediction would be performed at the receiver instead of at the transmitter. Stability of the all-pole filter A(z) is ensured if quantization is performed in such a way that the transformed autocorrelation coefficient matrix remains positive definite.

(7) The best set for transmission purposes is the set of reflection coefficients. In addition, this set is ordered and from Chapter IV, it was mentioned that the condition $|k_i| < 1$, for all i, always results in a stable filter A(z). Hence, the k_i's can be quantized to the range (-1,1) without any stability problem. Of course any function which maps (-1,1) to another interval in a one-to-one correspondence is equally acceptable. Examples of these, mentioned in [2, section 10.2.1] are the area ratios $A_m/A_{m-1} = 1-k_m/1+k_m$, the log

area ratios, and the areas A_m themselves. To conclude, the roots and any such function of the k_i's will produce stable filters at the output, after quantization at the input. Also for exactly the same reason as quantization, linear interpolation of parameters whose values lie within the region of stability will also be in the same region provided the region is convex. The unit circle and the straight line are convex regions, so that there is no stability problem with regard to the above sets of parameters. Linear interpolation was used in the synthesizer in Chapter IV.

It must be further noted that all transformations considered have a unique inverse. A computer program could then be developed that would produce any set of parameters given any other set as input. Also if covariance analysis had been applied, the step down-step up method of Chapter IV could be used in order to be assured of starting with a stable filter A(z) from which any set of parameters could be transmitted after quantization.

5.3 Reflection Coefficient Quantization

Uniform and nonuniform quantization of the k_i's subject to various fidelity criteria will now be discussed.

Uniform quantization

As will be seen later, this scheme is suboptimal because of the non-uniform spectral sensitivity $s_{k_i}(k_i)$ when the distance measure is (5.1.2). This is especially so when $k_i \sim 1$ and is even more pronounced for k_1 and k_2. Hence k_1 and k_2 are most important parameters in accurate representation of speech. Unfortunately, as was observed by many researchers, the probability distribution of k_1 and k_2 are highly skewed (especially k_1) towards -1 and +1 respectively. The probability distribution of the other less important k_i's look more or less like truncated Gaussian densities with mean zero and range $\pm.7$ [2, section 10.2.2]. The skewness property of k_1 and k_2 was derived in [10] using an approximation to the autocorrelation $r(n)$ valid for high sampling frequencies. The k_i's for all $i > 2$ were then uniformly quantized to the interval (-.7, .7) [10].

The same was done for $i = 1,2$ except that k_1 and k_2 are linearly shifted by .3 and -.3 respectively, because of their skewness. For $i > 2$, fewer bits are necessary because the singularity of $s_{k_i}(k_i)$ becomes less pronounced as mentioned above. More quantitatively, it is stated in [10] that dynamic programming has been used to allocate bits in the optimum fashion for this uniform quantization. As expected the optimum allocation is non-uniform. Another study, [16], drawing on the fact that for all k_i where i is even, the probability distribution is less symmetrically distributed than that for odd i, avoided uniform quantization throughout the range $((k_i)_{min}, (k_i)_{max})$. (The limits $(k_i)_{min}$ and $(k_i)_{max}$ are here defined as the values at which the probability function is truncated and depend on i). In conclusion, when using distance measure (5.1.2), uniform quantization comes close to being optimal except for k_1 and k_2. Moreover in the limit of fine quantization, it minimizes entropy subject to the fidelity criterion M_p [12].

Equal area quantization

This scheme has been applied in [17]. As was shown previously, it maximizes output entropy. The results of the study in [17] will now be summarized. Histograms of the relative frequency distribution of the reflection coefficients were collected for silence, voiced and unvoiced intervals, separately. For this scheme, $u(x) = dz/dx = p_X(x)$. Since z is uniformly quantized in (0, 1) the corresponding levels and boundaries for x (where x is a reflection coefficient) can be found. The bit allocation was determined empirically from listening tests. It was found that for unvoiced speech, the total number of bits used is only slightly over half the number of bits used for voiced speech, and they are distributed among the first 5 reflection coefficients only. As the probability distributions obtained depend much more upon the recording conditions (background noise) than on the speaker or speech content, the quantization tables for silence, unvoiced and voiced speech were kept fixed under fixed recording conditions. Because k_1 is important as far as minimizing spectral deviation, adaptive pre-emphasis of input speech is suggested, to match the distribution of k_1 to the one resulting in the fixed quantization table for k_1. However, there is no guarantee that the other k_i will be simultaneously matched. The speech, must then be post-emphasized by the same factor at the receiver. By processing speech with these fixed quantization tables, it was found that 25% of the analysis frames were silence, 30% unvoiced speech and 45% voiced speech. The relatively high percentage of silence intervals is due to stop gaps and short pauses unavoidable even in continuous speech. Only 2 bits are needed in order to distinguish between the 3 above classes of intervals. Two bits represent 4 levels. One of the levels could then be used to inform the receiver that the present frame belongs to the same class as the previous one if variable frame rate transmission is used. In conclusion, because of the above relative percentages, plus the fact that no additional information needs to be sent if the frame is silent, and that unvoiced frames require a much smaller number of bits than voiced frames, Seneff was able to achieve a variable rate vocoder with an average bit rate of 1450 bits/sec.

Spectral deviation quantization

The derivations are taken mostly from [14]. The first step is to use the distance measure (5.1.6) as an approximation to equation (5.1.2) with p = 2. Let

$$\Delta A = \sum_{i=1}^{M} [a_i(\underline{\lambda}+\Delta\underline{\lambda}) - a_i(\underline{\lambda})]z^{-i}$$

The inverse Fourier transform of $|\Delta A|^2$ is then

$$r_\Delta(n) = \sum_{i=1}^{M} [a_i(\underline{\lambda}+\Delta\underline{\lambda}) - a_i(\underline{\lambda})][a_{i+n}(\underline{\lambda}+\Delta\underline{\lambda}) - a_{i+n}(\underline{\lambda})] \tag{5.3.1}$$

But $r_\Delta(n) = 0$ for $|n| > M - 1$ because

$$a_o(\underline{\lambda}) = 1 \qquad \forall\ \underline{\lambda}$$

and

$$a_i(\underline{\lambda}) = 0 \qquad i \notin \{0,1,2,\ldots.,M\}$$

Also $r_\Delta(n) = r_\Delta(-n)$. Hence, by Parseval's theorem

$$D^2 = \frac{2}{\alpha}\,[r(0)r_\Delta(0) + 2\sum_{n=1}^{M-1} r(n)r_\Delta(n)] \tag{5.3.2}$$

where r(n) is the autocorrelation sequence of the input speech. Assume that $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots., \lambda_L)$ reduces to $\underline{\lambda} = \lambda$, i.e. consider single parameter variation only. First, let the parameter be a filter coefficient a_ℓ. Then

$$\Delta A = \frac{\partial A}{\partial a_\ell} \Delta a_\ell = \frac{\partial}{\partial a_\ell} \sum_m a_m z^{-m} \Delta a_\ell$$

$$= \Delta a_\ell z^{-\ell} \tag{5.3.3}$$

and

$$r_\Delta(n) = z^{-1} \text{ transform } [\Delta A(z) \Delta A(1/z)]$$

$$= z^{-1} \text{ transform } [(\Delta a_\ell)^2]$$

$$= (\Delta a_\ell)^2 \delta_{no} \tag{5.3.4}$$

Hence by (5.3.2)

$$D^2 = \frac{2}{\alpha} r(0) (\Delta a_\ell)^2 = 2(\Delta a_\ell)^2 (\frac{\alpha_o}{\alpha})$$

$$= \frac{2(\Delta a_\ell)^2}{\prod_{i=1}^{M} (1-k_i^2)}$$

So, as $|k_i| \to 1$, D^2 becomes unbounded. Therefore apart from the stability problem that arises in quantizing the filter coefficients, using these as parameters to be quantized is to be avoided.

Consider using as the single parameter, an arbitrary transformation of a single reflection coefficient. Namely,

$$\Delta A \sim \Delta\lambda \frac{\partial k_\ell}{\partial\lambda} \frac{\partial A(z;\lambda)}{\partial k_\ell} \tag{5.3.5}$$

As was shown in Chapter II

$$A_m(z) = A_{m-1}(z) + k_m B_{m-1}(z)$$

$$B_m(z) = z^{-1}[k_m A_{m-1}(z) + B_{m-1}(z)]$$

From the form of these equations, it is seen that $A(z;\lambda)$ is a linear function of every k_ℓ. Consequently,

$$\frac{\partial A(z;\lambda)}{\partial k_\ell} = \frac{\Delta A(z;\lambda)}{\Delta k_\ell}$$

Let $k_\ell = k_\ell(\lambda)$ and λ' be such that $k_\ell(\lambda') - k_\ell(\lambda) = 1$.

Then
$$\frac{\partial A(z;\lambda)}{\partial k_\ell} = A(z;\lambda') - A(z;\lambda)$$

or
$$\frac{\partial a_i(\lambda)}{\partial k_\ell} = a_i(\lambda') - a_i(\lambda) \tag{5.3.6}$$

To get D, (5.3.6) is first computed for all i and the results are substituted in (5.3.1) and (5.3.2). Note that $\lim_{\Delta\lambda\to 0} \frac{D}{\Delta\lambda}$ is the sensitivity $s_\Lambda(\lambda)$. Now it can be shown that if $|k_\ell| < 1$, then $r(n)$ and $r_\Delta(n)$ are bounded and therefore so is $\Sigma r(n) r_\Delta(n)$ [14]. But,

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

Therefore $s^2_\Lambda(\lambda)$ can be written in a form

$$\left(\frac{\partial k_\ell}{\partial \lambda}\right)^2 \frac{f_\ell(k_1,\ldots\ldots,k_M)}{1-k_\ell^2} \tag{5.3.7}$$

where the only singular contribution of k_ℓ to $s_\Lambda(\lambda)$ is due to the denominator $(1-k_\ell^2)$.

This singularity can then be cancelled by the transformation

$$k_\ell = \sin \lambda/c_\ell$$

or
$$\lambda = c_\ell \sin^{-1} k_\ell \tag{5.3.8}$$

as it is easily seen to satisfy $(\partial k_\ell/\partial\lambda)^2 = 1-k_\ell^2$. Now recall that the choice $dz/dx = u(x) = s_X(x)$ minimizes

$$\max_{a \le x \le b} \bar{D}(x,q(x)) \tag{5.3.9}$$

where z is uniformly quantized. Therefore, if this fidelity criterion is to be satisfied $s_\Lambda(\lambda)$ in (5.3.7) must be equal to a constant, which implies that $\partial\lambda/\partial k_\ell$ is proportional to $\sqrt{f_\ell(k_1,\ldots\ldots,k_M)/1-k_\ell^2}$. In cases where the function f_ℓ can be represented by a constant when compared to $1-k_\ell^2$ it is seen that the inverse sine quantization (5.3.8), to a good approximation, satisfies the minimization of the above fidelity criterion. Now (5.1.6) is the result of using $\sigma(\underline{\lambda}) = 1$, $\forall\underline{\lambda}$. The following normalization will now be used:

$$\sigma^2(\underline{\lambda}) = \alpha = \alpha_o \prod_{i=1}^{M} (1-k_i^2) \tag{5.3.10}$$

The input energy α_o is independent of all k_i's. The first term on the right hand side of (5.1.5) can be written as,

$$4\left\{\ln\left[\frac{\sigma(\underline{\lambda}+\Delta\underline{\lambda})}{\sigma(\underline{\lambda})}\right]\right\}^2 = \left\{\ln\left[\frac{\sigma(\underline{\lambda}+\Delta\underline{\lambda})}{\sigma(\underline{\lambda})}\right]^2\right\}^2$$

In the one parameter variation this becomes

$$\left\{ \frac{\Delta\lambda[\ln\sigma^2(\lambda+\Delta\lambda)-\ln\sigma^2(\lambda)]}{\Delta\lambda} \right\}^2$$

$$\sim (\Delta\lambda)^2 \left\{ \frac{\partial[\ln\sigma^2(\lambda)]}{\partial\lambda} \right\}^2$$

Substituting $\sigma(\lambda)$ as given by (5.3.10) results in

$$(\Delta\lambda)^2 \ (\frac{\partial k_\ell}{\partial\lambda})^2 \ \frac{4k_\ell^2}{(1-k_\ell^2)^2}$$

Adding this to (5.3.7),

$$s_\Lambda^2(\lambda) = (\frac{\partial k_\ell}{\partial\lambda})^2 \ \frac{g_\ell(k_1,k_2,\ldots.,k_M)}{(1-k_\ell^2)^2}$$

where the only singularity due to k_ℓ appears in the denominator as $(1-k_\ell^2)^2$. Straightforward differentiation will show that $\lambda = c_\ell \ln (1+k_\ell)/(1-k_\ell)$ satisfies $\partial k_\ell/\partial\lambda = 1-k_\ell^2$. But $\ln (1+k_\ell)/(1-k_\ell)$ is a log area ratio and this parameter has already been mentioned a few times. Hence there are two quantization schemes which minimize the fidelity criterion max $\bar{D}(x,q(x))$ in an approximate manner. Inverse sine and log area ratio quantization. Log area ratio quantization has also been empirically arrived at in [15] using the same gain normalization (5.3.10), but a value of p equal to 1 in distance measure (5.1.2). Hence it can be concluded that a different quantizing scheme is arrived at solely because of the use of a different gain normalization and not because of the choice of p in (5.1.2).

Now distance measure (5.1.7) has gain normalization (5.3.10), with the input gain α_o being a function of the parameter vector $\underline{\lambda}$. Consider the single parameter variation where $\lambda = k_\ell$. Then the gain normalization is exactly like (5.3.10) with α_o independent of all k_i. (5.1.7) then becomes

$$D = \ln\left[\frac{\alpha(k_\ell)}{\alpha(k_\ell+\Delta k_\ell)}\right] + \ln\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{A(e^{j\theta};k_\ell+\Delta k_\ell)}{A(e^{j\theta};k_\ell)}\right|^2 d\theta\right] \qquad (5.3.11)$$

It is proved in Appendix B that the second term on the right is simply

$$\ln\left[\frac{(A(e^{j\theta};k_\ell+\Delta k_\ell),A(e^{j\theta};k_\ell+\Delta k_\ell))}{(A(e^{j\theta};k_\ell),A(e^{j\theta};k_\ell))}\right]$$

in the inner product notation of section 2.1.

Denoting $A(e^{j\theta};k_\ell)$, $A(e^{j\theta};k_\ell+\Delta k_\ell)$ by A and A' respectively,

$$\lim_{\Delta k_\ell \to 0} \left\{ \frac{\ln(A',A') - \ln(A,A)}{\Delta k_\ell} \right\}$$

$$= \frac{d}{dk_\ell} \ln(A(e^{j\theta};k_\ell), A(e^{j\theta};k_\ell))$$

$$= 0$$

because (A,A) is minimum since it is the error signal energy of the linear prediction analysis.

Therefore,

$$s_{k_\ell}(k_\ell) = \left| \lim_{\Delta k_\ell \to 0} \frac{D}{\Delta k_\ell} \right| \tag{5.3.12}$$

In combination with (5.3.10) this simplifies to

$$\left| \lim_{\Delta k_\ell \to 0} \frac{\ln(1-k_\ell^2) - \ln(1-(k_\ell+\Delta k_\ell)^2)}{\Delta k_\ell} \right| = \left| \frac{d}{dk_\ell} \ln(1-k_\ell^2) \right| = \frac{2|k_\ell|}{1-k_\ell^2} \tag{5.3.13}$$

With respect to (5.1.7) this is an exact result for k_ℓ which is independent of the values of all other k_i's [15]. The requirement for λ to have a constant sensitivity measure is that

$$\frac{\partial \lambda}{\partial k_\ell} = s_{k_\ell}(k_\ell) \tag{5.3.14}$$

Integration of (5.3.13) evidently results in

$$\lambda(k_\ell) \;\alpha\; \mathrm{sgn}(k_\ell)\; \ln \frac{1}{1-k_\ell^{\,2}} \tag{5.3.15}$$

$$= \mathrm{sgn}(k_\ell)\; \ln \frac{\alpha_{\ell-1}}{\alpha_\ell} \tag{5.3.16}$$

where $\alpha_\ell = \alpha_o \prod_{i=1}^{\ell} (1-k_i^{\,2})$ is the forward prediction residual energy of Chapter II. 5.3.16 is called log error ratio quantization and it is pointed out in [15] that speech quality is better using log area ratio quantization rather than log error ratio quantization. From this fact, it is concluded that distance measure (5.1.2) describes the speech perception mechanism better than distance measure (5.1.7). An additional reason for preferring (5.1.2) over (5.1.5) is also included in Appendix B.

<u>Two parameter quantization [14]</u>

In this method the roots of the filter $A(z) = \sum_{i=0}^{M} a_i z^{-i}$ are computed. $A(z)$ is then factored into $\lfloor M/2 \rfloor$ quadratic polynomials. If $M/2 \neq \lfloor M/2 \rfloor$, then there is a leftover linear term $z-z_M$ where z_M is a real root. Which real root z_M is chosen to be the leftover, and which real root is to be associated with which real root in the formation of a quadratic with real coefficients will be considered later when quantization is discussed in more detail. For the moment assume a polynomial $A(z;\underline{\lambda}) = 1+a_1z^{-1}+a_2z^{-2}$ has been formed. Then treating it as a linear prediction filter, (2.1.16) and (2.1.17) yield

$$a_2 = k_2 \tag{5.3.17}$$

$$a_1 = k_1 + k_2 . k_1 \tag{5.3.18}$$

If both k_1 and k_2 are quantized simultaneously, then

$$A(z;\underline{\lambda}) \sim \frac{\partial A}{\partial k_1} \Delta k_1 + \frac{\partial A}{\partial k_2} \Delta k_2 \tag{5.3.19}$$

$$= \Delta k_1 (1+k_2) z^{-1} + \Delta k_2 (k_1 z^{-1} + z^{-2})$$

after substituting (5.3.17) and (5.3.18) in $A(z;\underline{\lambda})$. Now in scalar product notation, distance measure (5.1.6) is $2/\alpha\ (\Delta A, \Delta A)$. Then take the ΔA of (5.3.19) but first write it as a linear combination of the orthogonal polynomials

$B_m(z)$ of Chapter II. If a polynomial $P_m(z) = \sum_{i=1}^{m+1} p_{mi} z^{-i}$ is constructed using the recursive formula $P_{m-1}(z) = P_m(z) - p_{m,m+1} B_m(z)$ for m = M-1, M-2, ..., 0 then, starting from the initial conditions $\Delta A(z) = P_{m-1}(z)$, $\Delta A(z)$ is represented as $\sum_{m=0}^{M-1} p_{m,m+1} B_m(z)$. Therefore by the orthogonality of the B_m's, i.e., $(B_m, B_n) = \delta_{mn} \alpha_m$,

$$D^2 = 2 \sum_{m=0}^{M-1} p^2_{m,m+1} \frac{\alpha_m}{\alpha} \tag{5.3.20}$$

Here M=2 and using (2.1.16)-(2.1.18)

$$B_0(z) = z^{-1}$$

$$B_1(z) = k_1 z^{-1} + z^{-2}$$

and $$\Delta A = (1+k_2) B_0(z) \Delta k_1 + B_1(z) \Delta k_2$$

or $$p_{01} = (1+k_2) \Delta k_1$$

$$p_{12} = \Delta k_2$$

Consequently $$D^2 = 2(1+k_2)^2 \Delta k_1{}^2 \frac{\alpha_0}{\alpha} + 2(\Delta k_2)^2 \frac{\alpha_1}{\alpha}$$

$$D^2 = 2\,\frac{1+k_2}{1-k_1}\,\frac{(\Delta k_1)^2}{1-k_1{}^2} + \frac{2(\Delta k_2)^2}{1-k_2{}^2} \tag{5.3.21}$$

Relation between number of bits and D

The asymptotic result relating N or the entropy H to the fidelity criterion $E(\bar{D})$ has already been derived in the case of single parameter variation. This has been applied in [12] when the parameter is a reflection coefficient. Details will be described later. If the fidelity criterion is $\max \bar{D}(x,q(x))$, then it was proven in Appendix A, that this quantity is minimized by transforming k_i to a constant sensitivity parameter that is uniformly quantized. For such a parameter, the maximum quantization error is equal to half the distance between levels. Let $\underline{k}_\ell$ and $\bar{k}_\ell$ define the range of the truncated probability distribution of k_ℓ. Then define $\underline{\lambda}_\ell$ and $\bar{\lambda}_\ell$ to be the transformed values of these two numbers. If the number of levels is N_ℓ the following relation is obtained

$$\max \bar{D}(k_\ell,q(k_\ell)) = \frac{\bar{\lambda}_\ell-\underline{\lambda}_\ell}{2N_\ell}\,\frac{\partial k_\ell}{\partial \lambda}\,s_{k_\ell}(k_\ell)$$

a constant independent of k_ℓ. The form of $s_{k_\ell}(k_\ell)$ will depend upon the choice of p in (5.1.2) and the gain normalization $\sigma(\underline{\lambda})$.

Recall that in the case of fine quantization $D(k_\ell, q(k_\ell)) \sim \bar{D}(k_\ell, q(k_\ell))$ and the above holds approximately if $\bar{D}$ is replaced by D. Furthermore, if k_ℓ is transformed to a uniformly quantized variable λ with a non constant sensitivity, then the above result is valid for large N if $\partial k_\ell/\partial\lambda \; s_{k_\ell}(k_\ell)$ is maximized over k_ℓ. For the 2 parameter variation described above, there is no one-dimensional sensitivity function $s_\Lambda(\lambda)$ defined as (5.1.11). Hence a bound $\bar{D}$ will not be defined either. The smallest number of levels (in the asymptotic limit) that are required if a fixed spectral deviation D is not to be exceeded will now be computed. From (5.3.21) with the change of variables $\psi_i = \sin^{-1} k_i$, $i = 1,2$, (5.3.21) becomes

$$D^2 = 2\,\frac{(1+\sin\psi_2)}{1-\sin\psi_2}\,(\Delta\psi_1)^2 + 2(\Delta\psi_2)^2 \qquad (5.3.22)$$

which is the equation of an incremental ellipse. For simplicity, rectangular boundaries would be desired when quantizing ψ_1 and ψ_2. To minimize the number of levels the area of a rectangle inscribed in the incremental ellipse with center (ψ_1, ψ_2) must then be maximized if D is not to be exceeded. The area is $4\Delta\psi_1\,\Delta\psi_2$ and differentiation with respect to $\Delta\psi_1$, with the value of $\Delta\psi_2$ given by (5.3.22), will yield a maximum when the derivative is set to zero. In this way, the height of the rectangle is found to be $\Delta\psi_2 = D$ and its width $\Delta\psi_1 = D\sqrt{1-\sin\psi_2/1+\sin\psi_2}$. If k_1 and k_2 satisfy $-1 \leq k_1 \leq 1$ and $\underline{k}_2 \leq k_2 \leq \bar{k}_2$ where $\underline{k}_2$ and $\bar{k}_2$ are determined empirically, then $-\pi/2 \leq \psi_1 \leq \pi/_2$ and $\sin^{-1}\underline{k}_2 \leq \psi_2 \leq \sin^{-1}\bar{k}_2$. Therefore a necessary and sufficient condition for a spectral deviation not to exceed D is that (ψ_2 axis is vertical) the number of horizontal strips N_s is to be at least

$$\frac{\sin^{-1}\bar{k}_2 - \sin^{-1}\underline{k}_2}{D} \tag{5.3.23}$$

Let the boundaries values be $\psi_2(n)$, $n=1,2,\ldots\ldots,N_s$. Similarly for a fixed $\psi_2(n)$, the number of vertical strips is

$$N(n) = \frac{\pi}{D}\left[\frac{1+\sin\psi_2(n)}{1-\sin\psi_2(n)}\right]^{1/2} \tag{5.3.24}$$

Obviously ψ_2 is uniformly quantized and so is ψ_1 for fixed ψ_2. Therefore the total required number of quantization levels is

$$N = \sum_{n=1}^{N_s} N(n) \tag{5.3.25}$$

Define $\Delta\psi_2 \stackrel{\Delta}{=} D = \psi_2(n+1) - \psi_2(n)$. (5.3.25) can then be rewritten for small $\Delta\psi_2$ as

$$N \sim \frac{\pi}{D^2} \int_{\sin^{-1}\underline{k}_2}^{\sin^{-1}\bar{k}_2} \left(\frac{1+\sin\psi}{1-\sin\psi}\right)^{1/2} d\psi = \frac{\pi}{D^2} \ln\left(\frac{1-\underline{k}_2}{1-\bar{k}_2}\right) \tag{5.3.26}$$

This is the minimum number required if D is not to be exceeded. If a pair of uniformly quantized parameters is desired, ψ_1 is then multiplied by $[(1+\sin\psi_2)/(1-\sin\psi_2)]^{1/2}$ and the new transformation of k_1 and k_2 is given by

$$\lambda_2 = \sin^{-1}k_2$$

$$\lambda_1 = \left(\frac{1+k_2}{1-k_2}\right)^{1/2} \sin^{-1}k_1$$

<u>Bounds and bit allocation</u>

If the single parameter analysis is applied to each of the reflection coefficients, it must be decided on how the total number of bits B should be allocated among each k_i in order that the threshold of a certain fidelity criterion shall not be exceeded.

To find this optimum allocation, it is first necessary to get a bound on the overall spectral deviation when all parameters are simultaneously quantized. As the derivations of the results were rather lengthy and would have interfered with the continuity of the subject, a third appendix, devoted to these proofs, was added. Only the final results are summarized below.

$$\max D(\underline{\lambda},\underline{\lambda}'') \leq \sum_{i=1}^{M} \frac{\bar{\lambda}_i-\underline{\lambda}_i}{2N_i} \max_{k_1,k_2,..k_M} \frac{\partial k_i}{\partial \lambda_i} s_{k_i}(k_i) \tag{5.3.27}$$

$$E\, D(\underline{\lambda},\underline{\lambda}'') \leq \sum_{m=1}^{M} \frac{1}{4N_m} \left[\int_{\underline{\lambda}_m}^{\bar{\lambda}_m} \frac{E s_{\Lambda_m}(\lambda_m) p_{\Lambda_m}(\lambda_m) d\lambda_m}{u_m(\lambda_m)}\right]^2 \tag{5.3.28}$$

where $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_M)$ is a vector of the M parameters to be quantized. $\underline{\lambda}'' = (\lambda_1'', \lambda_2'', \ldots, \lambda_M'')$ where λ_j'' is a quantized value of λ_j. Therefore, the maximum of the spectral deviation over all values of $\underline{\lambda}$, and its expected value where the average is taken over all $\underline{\lambda}$, are respectively bounded by the sum of the M single parameter maximum and expected spectral deviation bounds [12,14].

A similar result is proven for the case of two parameter quantization:

$$D_{tot} \leq \sum_{j=1}^{\lfloor M/2 \rfloor} D_j + \lceil M/2 - \lfloor M/2 \rfloor \rceil D_o \tag{5.3.29}$$

where $D_j^{\,2} = \frac{\pi}{N_j} \ln \frac{1-\underline{k}_{2,j}}{1-\bar{k}_{2,j}}$ as in (5.3.26) and $\lfloor x \rfloor$, $\lceil x \rceil$ stand respectively for the integer smaller and greater than x, which are closest to x. If there is a leftover root (i.e. $M/2 \neq \lfloor M/2 \rfloor$) then an additional bound

$$D_o = \frac{\bar{\lambda}_1 - \underline{\lambda}_1}{2N_1} \frac{\partial k_1}{\partial \lambda_1} \sqrt{\frac{2}{1-k_1^{\,2}}}$$

is present (see Appendix C). Denoting the overall bounds in (5.3.27) and (5.3.28) by max $\bar{D}_{tot}$ and $E\bar{D}_{tot}$ respectively, it is then shown using Lagrangian multipliers that minimization of the total bit rate subject to a fixed max $\bar{D}_{tot}$ (or $E\bar{D}_{tot}$) is achieved by setting all individual single parameter bounds to the same value, namely, (max $\bar{D}_{tot}$)/M (or ($E\bar{D}_{tot}$)/M), [12,14].

For the two parameter quantization scheme , a similar result holds. Denoting the overall bound of (5.3.29) by D_b, $D_j = \frac{2D_b}{M}$ and $D_o = \frac{D_b}{M}$ minimize the total bit rate subject to a fixed D_b [14]. (For details, refer to Appendix C.)

The results of [14] will now be summarized. By assigning arbitrary values A to max $\bar{D}_{tot}$ and B_i to max $[(\partial k_i)/(\partial \lambda_i)]\, s_{k_i}(k_i)$ (for all i), a number N_i can be found for which the single parameter deviation bound does not exceed A/M except for those points $(k_1, k_2, \ldots, k_M)$ whose corresponding value of $[(\partial k_i)/(\partial \lambda_i)]\, s_{k_i}(k_i)$ exceeds B_i. In terms of this number N_i, it is then experimentally determined for k_i where $i > 2$, that uniform reflection coefficient quantization is slightly superior to log area ratio and inverse sine quantization of the k_i's. In spite of the gain normalization $\sigma(\underline{\lambda}) = 1$, inverse sine quantization is only slightly superior to log area ratio quantization. For $i = 1,2$, however, inverse sine quantization is significantly superior to uniform quantization. In terms of overall bit rate, the 3 schemes are almost equivalent.

For the 2 parameter quantization schemes [14], it is easy to derive by direct substitution, that the roots of the quadratic polynomial $A(z;\underline{\lambda}) = 1 + k_1(1+k_2)z^{-1} + k_2 z^{-2}$ are related to k_1 and k_2 by

$$k_2 = (x_i)^2 + (y_i)^2 \tag{5.3.30}$$

$$k_1 = -2x_i/(1+k_2) \tag{5.3.31}$$

if the roots are $z_i = x_i \pm jy_i$, and by

$$k_2 = x_i x_j \tag{5.3.32}$$

$$k_1 = -(x_i+x_j)/(1+k_2) \tag{5.3.33}$$

if x_i and x_j are the 2 real roots.

In order to find the $\underline{k}_2$ and $\bar{k}_2$ values to be substituted in (5.3.26), a histogram approach must be used. However there is a k_2 associated with each $\lfloor M/2 \rfloor$ polynomials. Therefore, to obtain statistics about each k_2, an ordering scheme must be developed. It is observed from (5.3.30) that k_2 is the magnitude of the root z_i which is inversely proportional to the exponential of the bandwidth. The $[M/2]$ k_2's are then ranked in order of increasing bandwidth. To find the largest k_2, the two largest real roots or, complex root with largest magnitude, are chosen at any step in the procedure, depending on which yields the largest k_2. This procedure ensures that the leftover term, if there is one, is associated with the smallest real root. If this scheme is repeated for every analysis frame, $\lfloor M/2 \rfloor$ scatter plots of (k_1,k_2) planes are obtained. By inspection, $\underline{k}_2$ and $\bar{k}_2$ are found for each ordered k_2. The numbers $\underline{k}_2$ and $\bar{k}_2$ of course decrease with decreasing k_2. It is observed that, for each plot, the range $(\underline{k}_2,\bar{k}_2)$ is small compared to the allowed range $(-1,1)$ for a reflection coefficient. (in fact much smaller than the observed range for k_2 in single reflection coefficient quantization). This is one of the reasons for the experimental fact that with a frame rate f_r of 50 Hz and 5 bits per frame for pitch and gain respectively, any one of the above three single parameter quantization requires 3500 bits/sec given a fixed value of 3dB for max $\bar{D}_{tot}$ as compared with 2800 bits/sec for D_b = 3dB in the two parameter quantization scheme [14]. The quality of speech is the same in both cases and bit rate reduction has been achieved for the two parameter method at the expense of more computation involved in polynomial root solving.

In [12] results on the first and tenth reflection coefficients using the min $E(\bar{D})$ fidelity criterion are presented. Let the variable stand for k_1. Then it was found that even in the case of only 4 quantization levels, the distribution of the points x_n, $\hat{x}_n$ obtained by using the quantizer curve which minimizes $E(\bar{D})$ asymptotically (5.1.22), is almost identical to that obtained using the quantizer curve which minimizes $E(\bar{D})$ non-asymptotically (the latter being found iteratively starting from 5.1.15). Then, still using 4 quantization levels, $E(\bar{D})$ is compared as obtained both asymptotically (5.1.19) and non-asymptotically (5.1.15) for the following 5 quantization schemes:

(1) $u(x) \propto 1$

(2) $u(x) \propto s_X(x)$

(3) $u(x) \propto p_X(x)$

(4) $u(x)$ which minimizes $E(\bar{D})$

$$(5)\quad u(x)\ \alpha\ \sqrt{s_X(x)p_X(x)}$$

For non-asymptotic cases, x_n, $\hat{x}_n$ are known from $x = U^{-1}(z)$ and are then substituted in (5.1.15) while for the asymptotic cases, u(x) is directly substituted into (5.1.19). In general, it is found that for any particular u(x), the asymptotical result for $E(\bar{D})$ is surprisingly close to the actual non-asymptotic result. Next, the asymptotic results for the minimum number of bits and entropy are obtained for $E(\bar{D})$ set at .3dB. Recall that, in the asymptotic limit, over all choices of u(x), the above scheme (2) minimizes entropy while scheme (5) minimizes log N. Unfortunately, it is experimentally determined that the difference between those values of log N and H is only .25 and .28 bits for k_1 and k_{10} respectively. For such small differences, it is not worthwhile to use variable bit rate coding which achieves rates close to entropy.

If further bit rate reduction is desired, then some other scheme which may involve an hitherto unexploited property of speech must be sought. Such a property exists and is stated in [12]. It has been experimentally verified that for voiced speech, reflection coefficients are dependent of each other and also from frame to frame. The dependence within a frame is greatest between k_1 and k_2.

The frame to frame dependence is felt to be even more significant. If this total dependency could somehow be extracted before transmission, a means for further reducing the bit rate without diminishing the quality of the output speech would have been achieved.

5.4 Orthogonal Parameter Quantization

To achieve a certain measure of independence among the reflection coefficients within a frame, a technique found in [18] is used to decorrelate them. Basically, the covariance matrix $R = [R_{ij}]$ is first obtained

$$R_{ij} = E[(k_i - Ek_i)(k_j - Ek_j)] \qquad (5.4.1)$$

In practice, using the law of large numbers and stationarity the mean of all k_i's should be computed using a time average over N frames and then the cross-correlation obtained by a time average over N-1 frames. The equation $|R-\lambda I| = 0$ (for the M eigenvalues λ_i of the matrix R) is then solved, where I is the identity matrix and $|..|$ is the notation for the determinant of a matrix. Then solve the simultaneous equations

$$R\underline{\phi}_i = \lambda_i \underline{\phi}_i$$

where $\underline{\phi}_i$ is the eigenvector corresponding to eigenvalue λ_i ($\underline{\phi}_i = (\phi_{1i},\phi_{2i},\ldots,\phi_{Mi})^T$). Now let Λ be a diagonal matrix $[\lambda_i\delta_{ij}]$ and U be the M x M matrix $[\underline{\phi}_1,\underline{\phi}_2,\ldots,\underline{\phi}_M]$. Then the previous equation can be rewritten as

$$RU = U\Lambda$$

or
$$U^{-1}RU = \Lambda$$

But $R = R^T$ and $\lambda = \lambda^T$ and consequently,

$$\Lambda^T = (U^{-1}RU)^T = U^T R^T (U^{-1})^T$$

$$= U^T R (U^{-1})^T = \Lambda$$

Therefore $U^{-1} = U^T$ (U is orthogonal).

Claim: The covariance matrix of the M parameters θ_i's, where

$$\theta_i \stackrel{\Delta}{=} \sum_{\ell=1}^{M} \phi_{\ell i} k_\ell \qquad \text{is } \Lambda$$

Proof: $E[(\theta_i - E\theta_i)(\theta_j - E\theta_j)] = E\sum_{\ell m}\sum \phi_{\ell i}\phi_{mj}(k_\ell - Ek_\ell)(k_m - Ek_m)$

$$= \sum_{\ell m}\sum \phi_{\ell i}\phi_{mj}R_{m\ell} = (U^T RU)_{ij} = (U^{-1}RU)_{ij} = (\Lambda)_{ij}$$

There is then no correlation between different θ_i's and in this sense they are termed orthogonal parameters. In addition, the total variance $\sum_{i=1}^{M} R_{ii}$ will be reallocated among the orthogonal parameters in such a way, that few of these will possess a large variance λ_i. This can be seen from the following observation.

Note that from the unitary property of U that k_j is $\sum_{i=1}^{M} \phi_{ji}\,\theta_i$. The variance of k_j can then be expressed as

$$R_{jj} = E(k_j - Ek_j)^2 = E\Big(\sum_{i=1}^{M} \phi_{ji}(\theta_i - E\theta_i)\Big)^2$$

$$= \sum_{i=1}^{M} \sum_{\ell=1}^{M} \phi_{ji}\phi_{j\ell} E[(\theta_i - E\theta_i)(\theta_\ell - E\theta_\ell)]$$

But the θ_i's are decorrelated, so that $R_{jj} = \sum_{i=1}^{M} \phi_{ji}{}^2 \lambda_i$. Again, from the unitary property of U

$$\sum_{j=1}^{M} \phi_{ji}\phi_{j\ell} = \sum_{j=1}^{M} \phi_{ij}\phi_{\ell j} = \delta_{i\ell}$$

and consequently $\sum_{j=1}^{M} R_{jj} = \sum_{i=1}^{M} \lambda_i$. This is true in general: the trace of a matrix is equal to the sum of its eigenvalues.

Now apply Holder's inequality [20]:

$$\sum_{i=1}^{M} |x_i y_i| \leq \left(\sum_{i=1}^{M} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{M} |y_i|^q \right)^{1/q}$$

for $1/q = 1 - 1/p$ and $p > 1$. In the case of $p = 2$, this reduces to the Cauchy-Schwartz inequality:

$$\sum_{i=1}^{M} |x_i y_i| \leq \left(\sum_{i=1}^{M} |x_i|^2 \sum_{i=1}^{M} |y_i|^2 \right)^{1/2}$$

Consequently,

$$\sqrt{R_{jj}} = \left(\sum_{i=1}^{M} (\phi_{ji} \sqrt{\lambda_i})^2 \right)^{1/2} .1 = \left(\sum_{i=1}^{M} (\phi_{ji} \sqrt{\lambda_i})^2 . \left(\sum_{i=1}^{M} (\phi_{ji})^2 \right) \right)^{1/2}$$

and therefore by the above inequality

$$\sqrt{R_{jj}} \geq \sum_{i=1}^{M} \phi_{ji}{}^2 \sqrt{\lambda_i}$$

and

$$\sum_{j=1}^{M} \sqrt{R_{jj}} \geq \sum_{i=1}^{M} \sqrt{\lambda_i}$$

Hence, by decorrelating the data, the sum of the square root of the variances is minimized. The problem then reduces to finding the λ_i's which minimize $\sum_{i=1}^{M} \sqrt{\lambda_i}$ subject to a known constraint $P = \sum_{i=1}^{M} \lambda_i$ and $\sqrt{\lambda_i} > 0$, $i = 1,2,...M$. The inverse problem, that of maximizing $\sum_{i=1}^{M} \sqrt{\lambda_i}$ subject to $P = \sum_{i=1}^{M} \lambda_i$ is easily solved by the Lagrangian function

$$F = \sum_{i=1}^{M} \sqrt{\lambda_i} + \alpha \sum_{i=1}^{M} \lambda_i$$

to yield $\lambda_i = 1/4\alpha^2$ which, substituted in the constraint gives $\alpha^2 = M/4P$ or $\lambda_i = P/M$. In other words, the total variance P is distributed equally among each of the M parameters. Therefore, following the decorrelating scheme it is expected that the total variance will be redistributed among the parameters in an uneven way. In section 6.1, a tabulation of R_{ii} and λ_i will demonstrate this fact. Sambur has applied decorrelation on the log area ratio parameters as well as on the k_i's [18,21]. (It was already seen that as far as stability is concerned, log A_m and k_m are equally good representations.) Using a filter order M = 12, he obtained statistics over N frames about individual utterances. From Table VIII, [18], with the 12 eigenvectors ordered in terms of decreasing variance, it is observed "that 90% of the total statistical variance is contained in the first 5 or 6 eigenvectors". This redundancy can then be exploited in a DPCM scheme, [18] resulting in further bit rate reduction, by sending the 5 or 6 parameters with largest variance. DPCM is basically a scheme where linear prediction is performed on data and the difference between the data and its linear prediction estimate is quantized before being sent to the receiver. Good results will be obtained if the original data is correlated in time. This is the case for speech where the solution to the linear prediction minimization criterion is consistent with the simplified model of the vocal tract (Chapter II). However, quantization of the error signal itself will not lead to substantial bit rate reduction. But it is mentioned at the end of the last section, that the k_i's are themselves dependent on their own past values. It is then proposed in [18], to apply linear prediction on the k_i's, the gain and pitch information. The linear prediction coefficients which can also be variable in an adaptive scheme, are then known to the receiver, and after probability distributions in the linear prediction errors in the pitch, gain and k_i's are obtained, optimal quantization levels and boundaries are calculated for each of these differences. The quantized values of these differences are then ready to be transmitted. To achieve further reduction in bit rate, dependence among the k_i's within a frame is taken into account. Linear prediction analysis is then performed on the θ_i's instead of the k_i's. Because 6 of these θ_i's have a very small variance, they do not vary much across an utterance. In the DPCM scheme, these parameters can be considered as constant and only their average values need to be sent. The number of bits is then allocated to the linear prediction errors of the remaining θ_i's with greater variance λ_i. It must be emphasized that once a number of bits, N_i, is determined that optimal quantizer curves must still be calculated for each of the linear prediction errors. This requires a knowledge of the probability distribution of these errors, which is not necessarily equal to the distributions of the original θ_i's. Sambur then maintains that it is possible to achieve a total bit rate from 600 bps to 1000 bps, "and still yield acceptable quality speech". The degradation is as mentioned before, dependent on the content and particularly on the speaker. The drawback to using this method is the amount of computation involved in the eigenvector-eigenvalue analysis. Moreover, if the gathering of statistics to obtain R, and the subsequent computation, is done for every consecutive N frame utterance in continuous speech, then the system could not be operated in real time. However the probability distribution of the k_i's are not very speaker and content dependent. In fact it was stated under the discussion on equal area coding that they are much more dependent on the amount of background noise.

Keeping this to a minimum, and assuming that the correlation among different k_i's is also speaker and content independent, the computation can be done prior to any transmission of orthogonal parameters θ_i's, if the speech data is first processed for the sole purpose of obtaining the necessary statistics, once and for all.

Introduction to the present study: theory

From now on, the dependence among k_i's within a frame only, will be taken into account and the necessary analysis leading to a comparison of results under the min $E(\bar{D})$ quantization scheme obtained using on the one hand, k_i's, and on the other θ_i's as the parameters will be described. Following the inequality $\sum_{j=1}^{M} \sqrt{R_{jj}} \geq \sum_{i=1}^{M} \sqrt{\lambda_i}$, it is hoped that not only will N_i increase as λ_i increases, but that $\sum_{i=1}^{M} \log N_i$ will be greater for the reflection coefficients than for the orthogonal parameters.

Diagonalization of the covariance matrix

Since the R matrix is symmetric, the Jacobi method for diagonalizing a matrix will be used to get both the eigenvalues and eigenvectors. The basic idea is as follows. Starting with a matrix $A = [a_{ij}]$, let

$$A_{k+1} = [a_{ij}^{(k+1)}] = (U_1U_2\ldots\ldots U_{k+1})^{-1}A(U_1U_2\ldots U_{k+1}) \tag{5.4.2}$$

$$= U_{k+1}^{-1}U_k^{-1}\ldots U_1^{-1}A\; U_1U_2\ldots\ldots U_{k+1} \tag{5.4.3}$$

A_{k+1} has obviously the same eigenvalues as A and is also symmetric if U_i is orthogonal (i.e. $U_i^{-1} = U_i^T$ for all i). Notice that it is possible to diagonalize a matrix A where $\Lambda = S^TAS$ for some S. But if $S^T \neq S^{-1}$, Λ is not the matrix of eigenvalues of A.

Furthermore, the trace of A.A being the sum of the diagonal elements

$$= \sum_i^M \sum_j^M a_{ij}\, a_{ji} = \sum_i^M \sum_j^M a_{ij}^2 \qquad \text{(because A is symmetric)}$$

= the sum of the eigenvalues of A.A.

Now let T^{-1} be any nonsingular matrix. Then, $T^{-1}(A.A)T = (T^{-1}AT)(T^{-1}AT)$ has the same eigenvalues as A.A. Let $T = U_1U_2\ldots U_{k+1}$ where all U_i are orthogonal. Then by (5.4.2) and the resulting symmetry of A_{k+1}

$$\sum_{i=1}^M \sum_{j=1}^M (a_{ij}^{(k)})^2 = \text{sum of the eigenvalues of A.A.} = \sum_{i=1}^M \sum_{i=1}^M a_{ij}^2$$

If the U_k's are such that

(1) $\sum_{i=1}^{M} (a_{ii}^{(k+1)})^2 > \sum_{i=1}^{M} (a_{ii}^{(k)})^2$ for all integer k

and

(2) $\sum_{i=1}^{M} \sum_{j=1}^{M} (a_{ij}^{(k)})^2 - \sum_{i=1}^{M} (a_{ii}^{(k)})^2 \to 0$ as $k \to \infty$

then the $a_{ij}^{(k)}$, $i \neq j$, converge to zero as $k \to \infty$ and A has been diagonalized

$$\lim_{k\to\infty} A_k = \Lambda = [\lambda_i \delta_{ij}] \tag{5.4.4}$$

and letting $U = \lim_{k\to\infty} (U_1 U_2 \ldots U_k)$, from (5.4.2),

$$AU = U\Lambda$$

It is proved in [19] that there exists a sequence $\{U_k\}$ that will result in (5.4.4). At step k, the largest $a_{ij}^{(k)}$ which is denoted by $a_{\ell m}^{(k)}$, is to be zeroed out. U_k is of the form

$$
I + \begin{array}{cc}
 & \begin{array}{ccccc} \quad & \ell & \quad & m & \quad \end{array} \\
\begin{array}{c} \\ \ell \\ \\ m \\ \\ \end{array} & \begin{bmatrix} \bigcirc & \vdots & \bigcirc & \vdots & \bigcirc \\ \cdots & \cos\alpha & \cdots & \sin\alpha & \cdots \\ \bigcirc & \vdots & \bigcirc & \vdots & \bigcirc \\ \cdots & \sin\alpha & \cdots & \cos\alpha & \cdots \\ \bigcirc & \vdots & \bigcirc & \vdots & \bigcirc \end{bmatrix}
\end{array}
$$

where α has to be properly chosen in terms of $a_{\ell m}{}^{(k)}$, $a_{\ell\ell}{}^{(k)}$ and $a_{mm}{}^{(k)}$. For details, see [19].

The eigenvectors and eigenvalues of the matrix R have therefore been obtained. Autocorrelation analysis will now be performed to obtain the k_i's as usual, and by transforming them to the set of uncorrelated parameters θ_i given by $\theta_i = \sum_{j=1}^{M} \phi_{ji} k_j$ a certain measure of independence has been achieved. The parameter λ_m in (5.3.28) then becomes θ_m instead of k_m. $\sum_{i=1}^{M} \log N_i$ is then minimized by letting each term in this asymptotic formula for $E(\bar{D}_{tot})$ be equal to $E(\bar{D}_{tot})/M$. Substituting (5.1.22) into (5.1.19) with this value for the individual bounds results in

$$N_i \sim \frac{1}{\frac{4E(\bar{D}_{tot})}{M}} \int_a^b \frac{Es_{\theta_i}(x)p_{\theta_i}(x)dx}{\left(\sqrt{Es_{\theta_i}(x)p_{\theta_i}(x)} \Big/ \int_a^b \sqrt{Es_{\theta_i}(x)p_{\theta_i}(x)}dx\right)} \tag{5.4.5}$$

$$= \frac{M}{4E(\bar{D}_{tot})} \left[\int_a^b \sqrt{Es_{\theta_i}(x)p_{\theta_i}(x)}\ dx\right]^2 \tag{5.4.6}$$

The range (a,b) will be discussed shortly. As was stated previously 3 to 4dB is the smallest distortion that can be perceived when using distance measure (5.1.2). In the theoretical study of [12], $E(\bar{D}_{tot})$ is set to 3dB. As a compromise, I set it to 3.5dB. Thereafter (5.4.6) is used to obtain the number of bits N_i for all the parameters. If θ_i has a small variance λ_i, it is hoped the the outcome of the computation (5.4.6) will be a small number. One other remark is in order: (5.4.6) is the asymptotic formula valid for large N_i. However, the interest lies in obtaining a small N_i. It is assumed that (5.4.6) is still accurate for small N_i as is demonstrated for k_1 in [12].

Notice also that only λ_i and ϕ_i are evaluated. The quantizer curve

$$\frac{\int_a^{\theta_i} \sqrt{Es_{\theta_i}(x)p_{\theta_i}(x)}\, dx}{\int_a^{b} \sqrt{Es_{\theta_i}(x)p_{\theta_i}(x)}\, dx}$$

and the number of bits as given by (5.4.6) cannot be computed, until $Es_{\theta_i}(x)$ and $p_{\theta_i}(x)$ are specified. Two methods will be proposed for the computation of the quantizer curve and hence of (5.4.6).

METHOD I - This method assumes that θ_i has a Gaussian p.d.f. of the form:

$$p_{\theta_i}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-Ex)^2/2\sigma^2} \qquad (5.4.7)$$

where σ^2 is the eigenvalue λ_i and Ex is the mean $\sum_{j=1}^{M} \phi_{ji} Ek_j$. This is easily obtained since the k_j's were computed in getting the covariance matrix R. Notice in passing, that if the k_i's were all normally distributed, that θ_i, being a linear combination of the k_i's would also be Gaussian. Since the θ_i's are uncorrelated, this would imply their independence and this is exactly what is desired. The assumption is of course false since $\text{Prob}\{|k_i| > 1\} = 0$ and therefore the range of θ_i is $\sum_{j=1}^{M} |\theta_{ji}|$ for all i. Consequently, the θ_i's are not Gaussian variables and it does not follow that they are independent. The best that can be said is that they are uncorrelated. However, for the convenience of representing $p_{\theta_i}(x)$ by an analytic function, (5.4.7) can be used because it is a good fit to experimentally obtained relative frequency histograms of θ_i (see Chapter VI). The problem now is to get an expression for the average over all $\theta_{m\neq i}$, $Es_{\theta_i}(x)$, as a function of θ_i. In its derivation it is required to know the sensitivity as a function of θ_i, for fixed but arbitrary $\theta_{m\neq i}$. Now in terms of a single parameter variation (where θ_i is the parameter), $A(e^{j\theta})$ in (5.1.6) becomes

$$\Delta A(e^{j\theta}) \sim \Delta\theta_i \frac{\partial A}{\partial \theta_i}(z;\theta_i) \quad \text{and since}$$

$$k_j = \sum_{i=1}^{M} \phi_{ji}\theta_i \qquad j=1,2,\ldots.M \tag{5.4.8}$$

$$\Delta A(e^{j\theta}) \sim \Delta\theta_i \sum_{j=1}^{M} \frac{\partial k_j}{\partial \theta_i} \frac{\partial A}{\partial k_j}(z;\theta_i)$$

$$= \Delta\theta_i \sum_{j=1}^{M} \phi_{ji} \frac{\partial A(z;\theta_i)}{\partial k_j} \tag{5.4.9}$$

But from (5.3.6)

$$\frac{\partial A}{\partial k_j} = \sum_{i=1}^{M} z^{-i}[a_i(k_j+1)-a_i(k_j)]$$

The inverse Fourier transform of $|\Delta A|^2$ is then

$$r_{\Delta\theta_\ell}(n) = (\Delta\theta_\ell)^2 \sum_{i=1}^{M} \left[\sum_{j=1}^{M} \phi_{j\ell}(a_i(k_j+1)-a_i(k_j)) \right] \left[\sum_{j=1}^{M} \phi_{j\ell}(a_{i+n}(k_j+1)-a_{i+n}(k_j)) \right] \tag{5.4.10}$$

This equation is to be substituted in (5.3.2) to yield D^2. But from (5.4.10), $D^2/\Delta\theta_i^{\ 2}$ depends on all k_i's, or transforming to $\theta_i = \sum_{j=1}^{M} \phi_{ji} k_j$, depends on all θ_i's. For example, in [12], a one parameter sensitivity function $D^2/\Delta k_i^{\ 2}$ is desired for the computation of N_i. Some sort of averaging procedure is required.

The following is the approach used in [12]. Since all $s_{k_i}(x)$ have, as only singularity, the factor $(1-x^2)^{1/2}$ in the denominator (recall gain normalization $\sigma(\underline{\lambda}) = 1$), $s_{k_i}(x)$ is multiplied by $\sqrt{1-x^2}$. This is performed for each point in the scatter plot. Then a histogram of the relative frequency of occurrence of points is obtained over the whole range of x.

A mean value B for $s_{k_i}(x)\sqrt{1-x^2}$ is then extracted from the histogram and the one-dimensional function $s_{k_i}(x)$ to be used in the quantization schemes of [12] is then $B/\sqrt{1-x^2}$.

Following the discussion that led to (5.3.28) the representative one dimensional function that will be used for $s_{\theta_i}(x)$ is the average value $Es_{\theta_i}(x)$ where the average is taken over all possible values $\theta_{m\neq i}$. Using the maximum range $R_i = \sum_{j=1}^{M} |\phi_{ij}|$ for θ_i, and $p_{\theta_i}(x)$ as given by (5.4.7), (5.4.6) is then evaluated as

$$N_i = \frac{M}{4x3.5dB} \left[\int_{-R_i}^{R_i} \sqrt{Es_{\theta_i}(x) \frac{1}{\sqrt{2\pi\lambda_i}} e^{-(x-E\theta_i)^2/2\lambda_i}} \, dx \right]^2 \tag{5.4.11}$$

This integration is carried on using the approximation by Simpson's Rule with 200 subdivisions. This number was found to be sufficient in depicting the shape of the quantizer curve. Once N_i is known from (5.4.11) the quantization levels and boundaries are then obtained from the quantizer curve. One technical remark is in order: The D^p measure (5.1.2) is derived using a natural logarithm whereas values for $E(\bar{D})$ and max $\bar{D}$ were always quoted in decibels. If a variable x has units of power (e.g. $\alpha/|A(e^{j\theta})|^2$), then the definition of x in dB is 10 $\log_{10} x$, and using the conversion $\log_e x = \log_{10}x/\log_{10}e$, the sensitivity function must then be multiplied by a factor 10 $\log_{10}e \sim 4.3492$. Now, $s_{\theta_i}(x)$ is a very complicated formula involving all θ_m and moreover the actual formula for $\text{Prob}\{\theta_m/\theta_i$, all $m\neq i\}$ is unknown. Even if a multidimensional Gaussian density function was used, the calculation of $Es_{\theta_i}(x)$ would be prohibitively difficult. The easiest way to obtain $Es_{\theta_i}(x)$ is through a time average of $s_{\theta_i}(x)$. By the law of large numbers, the sum of the $s_{\theta_i}(x)$ that occur in the scatter plot for a given x, divided by the number of these occurrences should be a good approximation to $Es_{\theta_i}(x)$. This will be the approach to be followed in METHOD II. In the present method, $Es_{\theta_i}(x)$ is assumed to be the $s_{\theta_i}(x)$ given by $\theta_m = E\theta_m$ for $m\neq i$. It is in general not true that $\sum_x f(x)\text{Prob}\{x\} = f(\sum_x x\text{Prob}\{x\})$. But the quantizer curves that are plotted using the two different methods, turn out to be similar in shape (see Chapter VI).

There is still one inconsistency which must be resolved. There is no guarantee that the necessary conditions $|k_i| < 1$ for stable filters will be satisfied with the set of orthogonal parameters consisting of an arbitrary θ_i and $\theta_m = E\theta_m$ for $m\neq i$. Indeed, from computer printouts, values of θ_i outside a certain range that will be denoted by (f_{i1}, f_{i2}) for convenience, always results in absolute values of k_ℓ slightly greater than 1, for a few index values of ℓ. In fact, $|k_\ell|-1$

increases monotonically as θ_i goes from $E\theta_i$ to $\pm R_i$. The scheme that was adopted then, was to alter $E\theta_m$ to new values $\tilde{\theta}_m$, $m \neq i$, in such a way that all k_i satisfy $|k_i| < 1$ for any particular θ_i. This cannot be said to be a single parameter variation. However as will be seen in Chapter VI, $\sqrt{\lambda_i}$ is relatively small in comparison with the range R_i, and also the actual probability density function of θ_i is truncated to an interval $(t_{i1},t_{i2}) \subset (-R_i,R_i)$, which is approximately the above interval (f_{i1},f_{i2}). Also it happens that $\sqrt{\lambda_i} < \min(|f_{i2}-E\theta_i|,|f_{i1}-E\theta_i|)$. From an inspection of (5.4.11) it is therefore seen that, because of the Gaussian density term, under the condition that $Es_{\theta_i}(x)$ is not too singular, the complement of either (f_{i1},f_{i2}) or (t_{i1},t_{i2}) does not contribute very much to the number of bits N_i, and the quantizer curve will be flat outside the truncated range. This is substantiated by the results of Chapter VI. Notice that because θ_i has a truncated density function, only the interval (t_{i1},t_{i2}) needs to be quantized instead of the whole interval $(-R_i,R_i)$. This was done in inverse sine quantization of the k_i's as their p.d.f. are truncated also. But in the min $E(\bar{D}_{tot})$ quantization scheme as discussed above, because the quantizer curve is flat outside the truncated range (t_{i1}, t_{i2}), it makes no difference whether that range or $(-R_i,R_i)$ is chosen for quantization. The latter is chosen because initially it is desired to prove that the integral over the complement of (t_{i1}, t_{i2}) was close to zero.

Letting θ_i run from $-R_i$ to R_i, it is first tested if

$$k_j = \phi_{ji}\theta_i + \sum_{\substack{\ell=1 \\ \ell\neq i}}^{M} \phi_{j\ell}E\theta_\ell \tag{5.4.12}$$

results in just one $|k_j| > 1$ for some j. If there is one such k_j, other values $\tilde{\theta}_\ell$ have to be used instead of $E\theta_\ell$. The basic assumption is to let

$$\tilde{\theta}_\ell - E\theta_\ell = \beta\sqrt{\lambda_\ell}/\phi_{j\ell} \tag{5.4.13}$$

where β is a constant of proportionality which is to be sought. The reason behind this assumption is that the bigger the variance λ_ℓ, the more likely it is that θ_ℓ departs from its mean value $E\theta_\ell$ and if $\phi_{j\ell}$ is small in (5.4.12), then in order for a change $\tilde{\theta}_\ell - E\theta_\ell$ to make its presence felt, a corresponding factor $\phi_{j\ell}$ must appear in the denominator. A value for β must now be found. In order to minimize the change $\tilde{\theta}_\ell - E\theta_\ell$, $|k_j| < 1$ can be made as close to 1 as is desired. An arbitrary value $|K| = .99$ is chosen. Then

$$\phi_{ji}\theta_i = k_j - \sum_{\substack{\ell=1\\ \ell\neq i}}^{M} \phi_{j\ell}E\theta_\ell = K - \sum_{\substack{\ell=1\\ \ell\neq i}}^{M} \phi_{j\ell}\tilde{\theta}_\ell \tag{5.4.14}$$

from which,

$$K-k_j = \sum_{\ell\neq i} \phi_{j\ell}(\tilde{\theta}_\ell - E\theta_\ell)$$

$$= \beta \sum_{\ell\neq i} \phi_{j\ell} \sqrt{\lambda_\ell}/\phi_{j\ell} \tag{5.4.15}$$

Consequently,

$$\tilde{\theta}_\ell - E\theta_\ell = \frac{K-k_j}{\sum_{n\neq i}\sqrt{\lambda_n}} \frac{\sqrt{\lambda_\ell}}{\phi_{j\ell}} \qquad \ell\neq i \tag{5.4.16}$$

Therefore, the values $\tilde{\theta}_\ell$ for which k_j becomes $\pm$ .99 have been found. Now $\tilde{\theta}_\ell$ must lie between $\pm\ R_\ell$ and it would be preferable that $|\tilde{\theta}_\ell - E\theta_\ell|$ does not exceed $\sqrt{\lambda_\ell}$, i.e. if in (5.4.16) it turns out that for some ℓ

$$\sum_{n\neq i} \sqrt{\lambda_n} \ \frac{\phi_{j\ell}}{K-k_j} < \sqrt{\lambda_\ell}$$

then the value of θ_ℓ is kept at $E\theta_\ell$ and (5.4.14) is changed to

$$\phi_{ji}\theta_i + \phi_{j\ell}E\theta_\ell = k_j - \sum_{\substack{n\neq i\\ n\neq \ell}} \phi_{jn} E\theta_n = K - \sum_{\substack{n\neq i\\ n\neq \ell}} \phi_{jn}\tilde{\theta}_n \tag{5.4.17}$$

But this would result in

$$\tilde{\theta}_m - E\theta_m = \frac{K-k_j}{\sum\limits_{\substack{n\neq i\\ n\neq \ell}} \sqrt{\lambda_n}} \; \frac{\sqrt{\lambda_m}}{\phi_{jm}} \qquad \text{for} \quad \begin{matrix} m\neq \ell \\ \neq i \end{matrix} \tag{5.4.18}$$

If for some m, $|\tilde{\theta}_m - E\theta_m| > \sqrt{\lambda_m}$, the same procedure is repeated until all the remaining $|\tilde{\theta}_m - E\theta_m|$ are less than $\sqrt{\lambda_m}$. There might be none remaining in which case the method failed. After all, the subscripts m, n run over a decreasing set of values from $\{1,2,....m\}$ and as the number of differences $\tilde{\theta}_m - E\theta_m$ decreases, their value tends to increase because the denominator $\Sigma\sqrt{\lambda_n}$ becomes smaller as the sum is over fewer elements. If the procedure fails, then an alternative simpler scheme is developed and described below.

But first supposing the method does not fail, then given this new set of orthogonal parameters, a check is made from (5.4.8) for the first occurrence of a $|k_i| > 1$. Recall that the above method, guarantees the inequality $|k_j| < 1$ for one j only. If all $|k_i| < 1$, then $s_{\theta_i}(x)$ is computed using this set of orthogonal parameters. If there is just one $|k_i| > 1$, the above procedure must be repeated in order to find a new set that will satisfy $|k_i| < 1$. If the procedure itself should fail at some point (no $\tilde{\theta}_m - E\theta_m$ remain which satisfy $|\tilde{\theta}_m - E\theta_m| < \sqrt{\lambda_m}$) or if after repeating the procedure a certain number of times, no set of orthogonal parameters have been found to yield $|k_i| < 1$ for all i, then the following step is taken. (From computer printouts, it was seen that the following scheme was forced upon, even for values of θ_i relatively close to (t_{i1}, t_{i2}). Only for values even closer to that interval is the above procedure successful.) Let

$$k_j = \frac{\theta_i \operatorname{sgn}\phi_{ji}}{\sum_{\ell} |\phi_{\ell i}|} = \frac{\theta_i \operatorname{sgn}\phi_{ji}}{R_i} \qquad (5.4.19)$$

for all j.

It must then be shown that these are a consistent set of equations. This certainly satisfies $|k_j| < 1$, since $|\theta_i| < R_i$ as it runs over $(-R_i, R_i)$. Also

$$\theta_n = \sum_{m=1}^{M} \phi_{mn} k_m = \frac{\theta_i}{R_i} \sum_{m=1}^{M} \phi_{mn} \operatorname{sgn}\phi_{mi} \qquad (5.4.20)$$

If n = i, the R.H.S. of (5.4.20) becomes $(\theta_i/R_i) \sum_{m=1}^{M} |\phi_{mi}| = \theta_i$ as required by equation (5.4.20). If $n \neq i$, the R.H.S. of (5.4.20) is less or equal to $(\theta_i/R_i)\, R_n \leq R_n$ also required of θ_n for all n. The set of equations (5.4.19), therefore satisfies the necessary constraints. $s_{\theta_i}(x)$ is then computed from (5.3.2) and (5.4.10) using this set of k_i's. Results using METHOD I will be shown at the end. It is desired to compare the quality and the bit rate of the speech generated by the above "optimal quantizer" with that obtained by using other quantization schemes. As the fidelity criterion in the above is $E(\bar{D}_{tot})$, for easier comparison, this is the fidelity criterion that will also be used to find the necessary number of levels in the other quantization schemes. Furthermore, the asymptotic formula in the limit of large N_ℓ, relating $E(\bar{D}_\ell)$ to N_ℓ

$$N_\ell \sim \frac{1}{4E(\bar{D}_\ell)} \int_a^b \frac{Es_{\Lambda_m}(\lambda_m)p_{\Lambda_m}(\lambda_m)d\lambda_m}{u(\lambda_m)}$$

will be used. The first quantization scheme that will be studied is the one that minimizes max $\bar{D}(\lambda_\ell, q(\lambda_\ell))$ for the reflection coefficients. This quantizer curve was derived in section 5.3:

$$\lambda_\ell = c_\ell \sin^{-1} k_\ell \equiv U(k_\ell)$$

Let the range of k_ℓ be $(\underline{k}_\ell, \bar{k}_\ell)$. Then normalization of U requires

$$c = \frac{1}{\sin^{-1}\bar{k}_\ell - \sin \underline{k}_\ell}$$

Then $u(\lambda_\ell) = dU/d\lambda_\ell = c_\ell/\sqrt{1-\lambda_\ell^2}$ is substituted in the above asymptotic formula. Once these quantizer curves are assigned to each reflection coefficients, the total number of bits $B = \sum_\ell \log N_\ell$ is minimized subject to the constraint $E(\bar{D}_{tot}) =$ 3.5dB (as was shown previously) by letting $E(\bar{D}_\ell) = E(\bar{D}_{tot})/M$. The second quantization scheme that is next considered is asymptotic min $E(\bar{D})$ on the k_i's.

$$U(k_\ell) \;\alpha \int_{-1}^{k_\ell} \sqrt{Es_{k_\ell}(\tau) p_{k_\ell}(\tau)}\; d\tau$$

and

$$N_\ell \sim \frac{1}{4E(\bar{D}_\ell)} \left[\int_{-1}^{1} \sqrt{Es_{k_\ell}(\tau) p_{k_\ell}(\tau)}\; d\tau \right]^2$$

where $E(\bar{D}_\ell) = E[\bar{D}_{tot}]/M$. This will then be an experimental result following the theoretical development in [12]. Minimum deviation orthogonal parameter quantization will then be compared with minimum deviation and inverse sine reflection coefficient quantization. As was already mentioned, the k_1 and k_2 distributions are skewed and hence do not look like symmetric truncated Gaussian densities. Although analytical functions which approximate their empirical distributions are derived in [10], the following empirical method to obtain the probabilities and the sensitivities will be used in comparing the 3 quantization schemes.

METHOD II - Histograms of the relative frequency of occurrences of the θ_i's and k_i's are obtained. The full range of the parameter (θ_i or k_i) is subdivided into 200 intervals. The counts in any given interval are added. For this particular interval this value is then divided by the sum of the counts over all intervals, and this number is assigned to the probability of the parameter lying in that interval. Since a probability density function

$$\lim_{\Delta x \to 0} \frac{\mathrm{Prob}\{x \leq \lambda \leq x+\Delta x\}}{\Delta x}$$

is desired, the probability of the interval just computed is divided by its length and this number is assigned to the probability of the parameter at the value halfway between the ends of the interval. As was previously stated in the section on METHOD I, $Es_{\theta_i}(x)$ is obtained empirically by again subdividing the range into 200 subdivisions, then the sum of all values that occur in a given interval divided by the number of occurrences in that interval is computed and that number is assigned to $Es_{\theta_i}(x)$ where x is a point midway in that interval. Notice that in the 3 quantization schemes, Es_{Λ_m} and p_{Λ_m} appear only as a product $Es_{\Lambda_m} p_{\Lambda_m}$ in the asymptotic formula for N_ℓ. Since p_{Λ_m} is the number of occurrences in a given interval divided by the sum of counts over all intervals, $Es_{\Lambda_m} p_{\Lambda_m}$ does not explicitly depend on the number of occurrences within that particular interval.

VI: EXPERIMENTAL RESULTS

The experimental setup will first be briefly described. It was mentioned in the last chapter that gain quantization is often done independently of the vocal tract parameters' quantization. In the present study, logarithm, of the gain and also pitch, quantization as used in [10] is adopted. The range for quantization of the gain is also chosen to be the range in one of the preliminary tests to the SIFT algorithm. More details about the SIFT algorithm and the subsequent autocorrelation linear prediction analysis, are then given. In order to study the dependence of the reflection coefficients on the text and speaker, statistics about 1 file and 14 files of speech were separately compiled. The dependence was found to be rather small. The Jacobi diagonalization procedure is then carried out, and the results using METHOD I and II are then tabulated. In terms of bit rate reduction, it is then seen that min $E(\bar{D}_{tot})$ quantization of the orthogonal parameters performs better than inverse sine quantization of the reflection coefficients but not as well as min $E(\bar{D}_{tot})$ quantization of the reflection coefficients. Plots of the relative frequency of occurrence histograms, averaged sensitivity functions and quantizer curves for the orthogonal parameters using METHOD I and II, are then compared. Then, plots of the histograms and sensitivity functions for the reflection coefficients are compared favorably with those of [12,14]. To obtain the quantizer levels and boundaries, linear interpolation on the quantizer curves, is then performed. Finally, a subjective comparison is established. It is found that the quality of synthesized speech using pitch extraction is very much the same for all quantization methods, and only slightly worse than that of speech synthesized with no quantization of the parameters. When the input to the synthesizer is the unquantized error signal, the quality of the output speech is somewhat more dependent on which of the three quantization schemes is used but is better than that of any speech obtained using the pitch-synchronous synthesizer.

Procedures in recording and playing back speech

The original speech utterances were recorded on analog magnetic tape using a high impedance microphone at INRS-Telecom, Montreal. The input gain to the tape was set by observing the peaks in the utterance. Then a converter was set in A/D mode. To prevent aliasing, the input speech is first passed through a variable analog filter with a value for the cutoff frequency, less or equal to half the sampling frequency of the converter. This filter allows frequency settings from 0 to 100 KHz in steps of 10 Hz, the selection of high pass versus low pass characteristics and also flat amplitude versus delay characteristics. The sampling frequency of the converter is then set at 10 KHz, thereby assuming that the amount of energy of the input speech in the range 5 to 10 KHz can be neglected. There is an implicit quantization of every speech sample because of the finite memory of the computer: a sample is stored as an integer in the range $(-2^{14}, 2^{14}-1)$. Overload lights indicate whether the input utterance exceeds this range. To avoid overloads, the input gain to the tape must be reduced. Once the speech is stored on computer disk as a file, a FORTRAN program which can further filter and down-sample the file is also available. The file can then be played back, by putting the converter in D/A mode. Since the D/A creates an analog signal by a sampled-and-hold method, the above mentioned variable analog filter is used as a low-pass filter in order to smoothen out the discontinuities introduced by that method. Before listing the experimental conditions, the conventional approach to quantizing the pitch and gain will now be described. This quantization, done independently, of the vocal tract parameters, is the reason behind preferring the gain normalization $\sigma(\underline{\lambda})$ as unity in the spectral distance measures.

Quantization of the pitch and gain

Pitch

As discussed in Chapter III the SIFT algorithm determines as estimate of the pitch P in the range 2.5 to 20 ms. The sampling frequency f_s of the input speech was 2 KHz. In dimensionless units then the pitch P' is Pf_s. The question is how the interval should be quantized. Evidence pointed out in [10] suggests that the ear is sensitive to relative fundamental frequency error $\Delta f/f$. Since $\Delta \ln f \sim \Delta f/f$, uniform quantization of lnf is necessary if a relative error independent of frequency is desired. Let

$$f_{min} = 1/P'_{max} \text{ and } f_{max} = 1/P'_{min}$$

stand for the range of frequencies of interest in the SIFT algorithm. If β_P is the number of bits used, then lnP' is quantized to the value

$$2^{\beta_P} \cdot \frac{\ln(1/P') - \ln(1/P'_{max})}{\ln(1/P'_{min}) - \ln(1/P'_{max})}$$

unless the speech is unvoiced, in which case, $P' = 0$. The inverse operation $\ln P' \to P'$ is then carried out at the receiver.

<u>Gain of the Error Signal [10]</u>. Experiments have shown that the probability density function of the gain can be roughly represented by an exponential [10]. It follows that if the logarithm of the gain is uniformly quantized, then the probability of occurrence of an interval is approximately uniformly distributed over all intervals. If β_G bits are used, then as for the pitch, the quantized value of $\ln G$ is

$$2^{\beta_G} \cdot \frac{\ln G - \ln G_{min}}{\ln G_{max} - \ln G_{min}}$$

As for pitch the inverse operation $\ln G \to G$ is carried out at the receiver. $G = 0$ is a problem but since there is always some background noise G_{min} is selected to be just above the upper cutoff for the noise gain. Adopting the figure in [10] this is set at $G_{max}/300$. G_{max} must now be found. Recall that in the autocorrelation method

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

For small α_o as in low amplitude fricative noise, α_M is not much less than α_o and for large α_o as in some voiced sounds, α_M is usually $<< \alpha_o$. Consequently,

$$\frac{(\alpha_M)_{max}}{(\alpha_M)_{min}} < \frac{(\alpha_o)_{max}}{(\alpha_o)_{min}}$$

and a dynamic range greater than that of the input speech is not needed. In [10], $(\alpha_M)_{max}$ is set to .3 $(\alpha_o)_{max}$. Since there are N samples in a frame, this would then correspond to an average amplitude $\sqrt{.3(\alpha_o)_{max}/N}$. This is the adopted value for G_{max} in [10]. α_o is obtained from the autocorrelation analysis of the input speech. In the present study, the pitch extraction is performed before the analysis. This is described in more detail in the next subsection. In one of the preliminary tests (prior to the pitch extraction) the value of G_{min} and hence of G_{max} is required. Since α_o is as yet undetermined, the value of G_{max} will be set at $\sqrt{.3}$ A where A is the maximum amplitude over all speech samples in an utterance. Recall that speech samples are quantized to 2^{15} levels when stored on computer disk. Only integers ranging from -2^{14} to $2^{14}-1$ are then possible for representing speech. The input gain to the converter (in A/D mode) is then kept at a constant value. This value must not be too large, as overloads, which are indicated by the A/D overload light, are to be avoided. Table 6.1.1 lists a few characteristics of 14 utterances which are described below. The value of A is set at the maximum over the most positive amplitude and the absolute value of the most negative amplitude. Values for β_G and β_P of 5 bits each were allocated to the pitch and gain. According to [10] these should result in reasonably good quality speech. Indeed it was observed that with only pitch and gain quantization, the output speech is almost indistinguishable from that synthesized with no quantization at all.

In all, 14 speech files of approximately 2 to 3 seconds in duration, were recorded and stored on computer disk, as described earlier. The data were chosen from a selection of well-known phonetically balanced utterances, [4]:

(1) OAK IS STRONG AND ALSO GIVES SHADE
(2) CATS AND DOGS EACH HATE THE OTHER
(3) ADD THE SUM TO THE PRODUCT OF THESE THREE
(4) THIEVES WHO ROB FRIENDS DESERVE JAIL
(5) THE PIPE BEGAN TO RUST WHILE NEW
(6) OPEN THE CRATE BUT DON'T BREAK THE GLASS

There were 3 adult male speakers and 2 adult female speakers. The first male uttered sentences (1), (3) and (4); the second male, sentences (2) and (3) and the third, (1), (3) and (4). The first female uttered (2), (3), (5) and the second (1), (3) and (6). A file will be denoted by a-b-c, where a stands for the sex of the speaker (M or F), b which of the speakers of the same sex and c, which of the above 6 sentences. A speech sample is denoted by s(n) and the speech characteristics in Table 6.1.1, are taken over the whole speech file.

Analysis conditions

The variable filter cutoff frequency was set at 5 KHz with a low pass flat amplitude characteristic. The sampling frequency of the converter in A/D mode was set at 10 KHz. The cutoff is abrupt enough to make the contribution to the spectrum, of aliasing and zeroes introduced in this way, negligible. The SIFT algorithm is then applied to produce 14 pitch files, one corresponding to each input speech file. SIFT uses an elliptic filter of third order, in prefiltering the speech file down to 1 KHz. The file is then downsampled to 2 KHz. (This is a computer simulation: all these operations were carried out with FORTRAN programs). The frame rate was 50 Hz, the analysis length N, 80 and the linear prediction filter order M, was 4. The preliminary test lower gain value was set to $G_{max}/300$ where G_{max} is obtained from Table 6.1.1 as was discussed previously. This same value of G_{max} was also used in quantization studies. Then, the 33 autocorrelation values R(1), R(2),....,

Table 6.1.1

File	Min s(n)	Max s(n)	E s(n)	$\sqrt{\text{Var s(n)}}$
M-1-1	-14077	11395	-22	1748.5
M-1-3	-14419	10222	-22	1545.4
M-1-4	-13967	12700	-22	1703.8
M-2-2	-15466	9549	-23	1800.8
M-2-3	-15503	10715	-23	1697.7
M-3-1	-14199	11579	-24	1976.0
M-3-3	-16384	11772	-23	2394.2
M-3-4	-13403	9631	-21	1980.0
F-1-2	-15010	9160	-22	1762.0
F-1-3	-12015	9179	-22	1759.8
F-1-5	-12782	8366	-22	1847.8
F-2-1	- 7685	6407	-21	919.0
F-2-3	- 6083	5230	-22	749.9
F-2-6	- 5173	4196	-23	640.4

R(33) are obtained from the last 76 samples in the 80 samples analysis frame. Following Figure 3.2.1, the procedure up to now is called STEP 1 and the further processing of the autocorrelation values R(n) is called STEP 2. For additional details, see Section 3.2 and [9]. The pitch decision of the SIFT algorithm, for each analysis frame in the speech file, is then stored in a pitch file. Recall that, because of the error detection and correction performed in STEP 2, there is a delay of 2 frames in the computation of the pitch.

Autocorrelation analysis is then performed on the 10 KHz speech file. The frame rate $f_r = 50$ Hz and the filter order $M = f_s(\text{KHz}) + 4 = 14$. For the m^{th} frame, the analysis frame length N is chosen to be $.01\ f_s = 100$ or $.02\ f_s = 200$, depending on whether the decision in the corresponding $(m-2)^{th}$ pitch frame is unvoiced or voiced respectively. Adaptive pre-emphasis using a factor $\mu = r(1)/r(0)$, and windowing using a Hamming window with a scale factor of .54, is done prior to this linear prediction analysis. The pitch, gain and reflection coefficient information for each analysis frame is then stored in a speech parameter file. Statistics necessary in the evaluation of the covariance matrix R are then gathered about the k_i's. Statistics, about the 1 file of reflection coefficients corresponding to speech file M-1-3 and about the 14 files of reflection coefficients were separately compiled in order to study their dependence on the text and speaker. For the purpose of calculating R and $E\theta_i$ as required in METHOD I, Ek_i must first be obtained. The values of the Ek_i and $Vark_i$ are shown in Table 6.1.2 for the 1 file and 14 file statistics. Other data on the k_i's will be presented when results on METHOD II are discussed. Table 6.1.3 and 6.1.4 are computer printouts from the Jacobi diagonalization Fortran program using 1 file and 14 file statistics respectively.

N is the filter order and thus is the rank of the covariance matrix. In this program, this matrix is denoted by A instead of R and because of its symmetry, only its upper triangular form is stored. ITER counts the number of times the whole procedure is repeated, and ITMAX is the maximum number of these iterations allowed in the program. SIGMA 1 and SIGMA 2 are respectively

$$\sum_{i=1}^{N} (a_i^{(k)})^2 \quad \text{and} \quad \sum_{i=1}^{N} (a_i^{(k+1)})^2$$

of the previous discussion on Jacobi diagonalization leading to (5.4.4). EPS1 and EPS2 are arbitrary threshold values used in the zeroeing of some elements $a_{lm}^{(k)}$ and in the selection of the value of α in orthogonal matrix U_k, respectively. Approximate convergence is achieved when

Table 6.1.2

1 FILE STATISTICS

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ek_i	-.359	.067	-.212	.157	-.169	.173	.025	.163	.135	.202	.009	-.044	-.107	-.019
$Vark_i$	.133	.138	.055	.072	.082	.039	.037	.038	.035	.056	.028	.022	.015	.012

14 FILE STATISTICS

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ek_i	-.430	.228	-.198	.089	-.071	.125	.038	.157	.070	.108	.008	.014	-.004	.033
$Vark_i$	.152	.172	.091	.070	.077	.060	.037	.044	.056	.048	.028	.021	.017	.013

Table 6.1.3

1 file statistics

DETERMINATION OF EIGENVALUES BY JACOBI METHOD,WITH

N = 14
ITMAX = 100
EPS1 = 0.10000E-09

EPS2 = 0.10000E-09
EPS3 = 0.10000E-04

THE STARTING MATRIX A(1,1)...A(N,N) IS

0.15215	-0.05960	-0.02159	-0.03892	-0.02702	-0.02589	-0.00796	-0.01445	-0.01032	-0.00444	-0.00183	-0.00592	-0.00512	-0.00124
0.00000	0.17244	-0.00491	0.02124	-0.02448	-0.00033	-0.00185	0.01396	0.02329	0.00671	0.00121	0.00823	-0.00156	-0.00147
0.00000	0.00000	0.09060	0.00371	-0.00458	-0.01778	-0.01758	-0.01404	-0.00559	-0.00269	0.00038	-0.00156	0.00305	0.00149
0.00000	0.00000	0.00000	0.07046	-0.00168	0.01328	-0.00378	-0.00030	-0.00064	0.00703	0.00482	0.00100	-0.00268	0.00033
0.00000	0.00000	0.00000	0.00000	0.07720	0.00142	0.00641	-0.00556	-0.01738	-0.01775	-0.00112	0.00142	0.00330	-0.00131
0.00000	0.00000	0.00000	0.00000	0.00000	0.05951	0.00407	0.01520	-0.00594	-0.00064	-0.00244	0.00133	-0.00323	-0.00034
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03733	0.00443	-0.00161	-0.00549	-0.00562	0.00007	-0.00181	0.00029
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04397	0.00474	0.00176	-0.00459	0.02169	-0.02223	0.00751
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.05586	0.01619	-0.02415	-0.00422	-0.00219	-0.00289
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04780	0.00399	-0.00552	-0.00436	-0.00128
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02819	0.00152	-0.00195	-0.00157
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02101	0.00318	0.00755
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01454	-0.00394
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01319

ITER = 1 SIGMA1 = 0.85976E-01 SIGMA2 = 0.11132

ITER = 2 SIGMA1 = 0.11132 SIGMA2 = 0.11267

ITER = 3 SIGMA1 = 0.11267 SIGMA2 = 0.11269

CONVERGENCE HAS OCCURRED,WITH

ITER = 4 SIGMA2 = 0.11269 S = 0.11269 N = 14

EIGENVALUES EIGEN(1)...EIGEN(N) ARE

0.13791 0.25120 0.11069 0.01937 0.07962 0.06167 0.03280 0.03050 0.05555 0.03694 0.02621 0.01810 0.01456 0.01132

THE FINAL TRANSFORMED MATRIX A IS

0.13791	-0.00000	0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.00000	-0.00000	0.00000
0.00000	0.25120	-0.00000	-0.00000	-0.00000	0.00000	0.00000	-0.00000	0.00000	0.00000	-0.00000	-0.00000	-0.00000	0.00000
0.00000	0.00000	0.11069	-0.00000	-0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000	0.00000	-0.00000	-0.00000	0.00000
0.00000	0.00000	0.00000	0.01937	-0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.00000	0.00000	-0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.07962	-0.00000	-0.00000	-0.00000	-0.00000	-0.00000	0.00000	-0.00000	0.00000	-0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.06157	0.00000	-0.00000	-0.00000	0.00000	-0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03280	-0.00000	-0.00000	0.00000	0.00000	0.00000	-0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03050	-0.00000	-0.00000	-0.00000	-0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.05555	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03694	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02601	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01810	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01456	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01132

EIGENVECTORS ARE IN CORRESPONDING COLUMNS OFTHE FOLLOWING T MATRIX

0.51470	-0.64550	-0.03071	0.23205	0.10689	-0.20301	-0.19747	0.03579	0.02919	-0.00167	-0.20086	0.12803	-0.18171	0.18705
0.43598	0.68038	0.02601	0.12004	0.42159	-0.21037	-0.03817	-0.01856	-0.00562	0.09673	-0.18196	0.00964	-0.05127	0.08555
-0.20240	0.03965	0.02447	0.28250	-0.06129	0.01020	-0.04041	0.01316	-0.32119	-0.06070	-0.14796	0.13959	-0.14915	0.12[illegible]
-0.14945	0.23077	-0.02896	0.07537	-0.48397	-0.55392	-0.17491	0.19358	0.40204	-0.35986	-0.05695	0.06410	-0.08167	0.04456
-0.54762	-0.02034	-0.15309	0.23298	0.45200	0.09335	-0.28596	0.09269	0.37171	0.24131	-0.25227	0.15460	-0.10000	0.15027
-0.17665	0.12317	-0.39093	0.22628	-0.32938	-0.13455	-0.17609	-0.49160	-0.49382	0.20547	-0.19104	0.01455	-0.18511	0.14274
-0.06154	0.01150	-0.25182	0.33449	0.12935	0.15394	0.61953	-0.03469	-0.02263	-0.53001	-0.00000	0.13288	-0.23954	0.17641
0.04357	0.12053	-0.26740	0.11475	-0.09392	0.18223	-0.19357	0.79098	-0.41109	0.03106	0.09261	-0.02148	-0.11463	0.07730
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.15428	0.05503	0.03624	-0.00916	-0.43986	0.25996	0.15955	0.12934	0.27446	0.40515	[illegible]	[illegible]	-0.11[illegible]	0.11167
-0.07235	0.00978	0.04652	0.40290	-0.05598	-0.16582	-0.21976	-0.01047	0.16418	0.41702	0.68185	-0.13858	-0.12750	0.23410
-0.02426	0.04408	-0.03637	-0.35229	0.07843	-0.11429	-0.01200	-0.05892	-0.10847	0.03740	0.31969	0.02964	-0.21705	0.[illegible]
-0.05764	0.03500	0.03513	-0.52425	0.06009	0.00490	-0.04995	-0.04609	-0.00216	-0.05934	0.01159	-0.38575	-0.55725	0.43527
-0.01295	-0.00105	0.01387	-0.09525	-0.00725	-0.00226	0.01531	0.02960	-0.03812	-0.07542	0.04412	0.12452	0.66117	0.72779

Table 6.1.4

14 file statistics

```
DETERMINATION OF EIGENVALUES BY JACOBI METHOD,WITH

N       = 14
ITMAX  = 100
EPS1     = 0.10000E-09

 EPS2    = 0.10000E-09
EPS3     = 0.10000E-04

THE STARTING MATRIX A(1,1)...A(N,N) IS

0.13272 -0.03728 -0.03155 -0.04729 -0.01348 -0.02177  0.00076  0.00108 -0.00629 -0.02096 -0.00215  0.00108 -0.00517  0.00194
0.00000  0.13807 -0.01771  0.01957 -0.04985  0.02142  0.00121  0.01026  0.01177  0.01968  0.01013  0.01191  0.00225 -0.00735
0.00000  0.00000  0.05540  0.00567  0.01465 -0.00167 -0.00500 -0.01089 -0.00966  0.00059  0.00088 -0.00970  0.00149  0.00046
0.00000  0.00000  0.00000  0.07213 -0.00269  0.01274 -0.01651  0.00735  0.00239  0.02414  0.00853 -0.00565  0.00024 -0.00145
0.00000  0.00000  0.00000  0.00000  0.08216 -0.00924  0.00235 -0.02511 -0.01293 -0.02888 -0.00337 -0.00230  0.00671  0.00350
0.00000  0.00000  0.00000  0.00000  0.00000  0.03976 -0.00044  0.00259  0.00528  0.01757  0.00374  0.00284 -0.00241 -0.00371
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.03749 -0.00397 -0.00358 -0.01495 -0.00070 -0.00350 -0.00311  0.00021
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.03247  0.01632  0.01858 -0.00310 -0.00024 -0.00552 -0.00316
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.03450  0.01850 -0.00205 -0.00296 -0.00416 -0.00255
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.05577  0.00314 -0.00613 -0.00759 -0.00442
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.02772  0.00192 -0.00223 -0.00410
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.02210  0.00510  0.00147
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.01515 -0.00738
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.01238

 ITER    -    1          SIGMA1 - 0.62037E-01          SIGMA2 - 0.92494E-01

 ITER    -    2          SIGMA1 - 0.92494E-01          SIGMA2 - 0.93911E-01

 ITER    -    3          SIGMA1 - 0.93911E-01          SIGMA2 - 0.93939E-01

CONVERGENCE HAS OCCURRED,WITH

 ITER    -    4      SIGMA2 - 0.93940E-01      S  - 0.93940E-01      N   -   14

EIGENVALUES EIGEN(1)...EIGEN(N) ARE

0.15963  0.22398  0.01887  0.09087  0.05492  0.04911  0.03307  0.03607  0.00950  0.01594  0.02319  0.02733  0.02743  0.01383

 THE FINAL TRANSFORMED MATRIX A IS

0.15963 -0.00000 -0.00000 -0.00000  0.00000  0.00000 -0.00000 -0.00000 -0.00000 -0.00000 -0.00000  0.00000 -0.00000 -0.00000
0.00000  0.22398 -0.00000 -0.00000 -0.00000 -0.00000  0.00000  0.00000  0.00000 -0.00000 -0.00000  0.00000 -0.00000 -0.00000
0.00000  0.00000  0.01887 -0.00000 -0.00000 -0.00000 -0.00000 -0.00000  0.00000  0.00000  0.00000 -0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.09087  0.00000  0.00000  0.00000  0.00000 -0.00000 -0.00000 -0.00000  0.00000 -0.00000 -0.00000
0.00000  0.00000  0.00000  0.00000  0.05492  0.00000 -0.00000 -0.00000 -0.00000 -0.00000  0.00000 -0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.04911 -0.00000 -0.00000 -0.00000 -0.00000  0.00000 -0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.03307  0.00000  0.00000 -0.00000 -0.00000  0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.03607  0.00000 -0.00000 -0.00000  0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00950  0.00000  0.00000 -0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.01594  0.00000 -0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.02319 -0.00000  0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.02733 -0.00000  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.02743  0.00000
0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.01383

EIGENVECTORS ARE IN CORRESPONDING COLUMNS OFTHE FOLLOWING T MATRIX

0.62943 -0.50395  0.22316  0.12171  0.32931 -0.12622  0.11678 -0.08148  0.16429  0.00432 -0.16073 -0.27771  0.09697  0.04623


0.33259  0.62385  0.12425 -0.54974  0.13885 -0.09179 -0.14926 -0.01211 -0.04082  0.12987 -0.11616 -0.21023  0.02215  0.06762
-0.35650 -0.00101  0.38272  0.05421 -0.20080 -0.73565 -0.06828 -0.04435  0.25469 -0.04649 -0.06296 -0.17926  0.09775  0.12779
-0.25463  0.32340  0.14206  0.30306  0.56585  0.15310  0.30950  0.41527  0.23363 -0.15744 -0.08498 -0.10709 -0.07608  0.09247
-0.45824 -0.28918  0.19784 -0.21948  0.21213  0.40494 -0.21867 -0.25525  0.02266  0.33524 -0.05245 -0.37644  0.10335  0.02840
-0.01607  0.20421  0.40729  0.04904 -0.04734  0.14756  0.48495 -0.64094 -0.16885 -0.18551 -0.14151  0.16213  0.01855 -0.05675
0.03315 -0.06244 -0.03246 -0.28400 -0.45199  0.20369  0.55224  0.21762  0.38593  0.08407  0.18014 -0.19688  0.29045  0.02523
0.18149  0.12877  0.50046  0.32507 -0.21159  0.19162 -0.23074  0.28078 -0.14444  0.34585  0.20050  0.14711  0.22920 -0.12586
0.10376  0.12519  0.02973  0.22724 -0.25317  0.31081 -0.39292 -0.15017  0.24043 -0.59766  0.11545 -0.32580  0.02007  0.19116
0.05492  0.29885 -0.41289  0.59543 -0.09014 -0.04075  0.01694 -0.30739  0.19635  0.46795 -0.21316 -0.18840  0.20761 -0.03823
0.01433  0.06595 -0.10582  0.00541  0.34167 -0.17763  0.02198 -0.25410 -0.02400 -0.01704  0.81596  0.00472  0.32303  0.00520
0.07063  0.02176  0.05741 -0.16444  0.12409  0.10173 -0.19916 -0.15219  0.63685  0.14220 -0.04270  0.60944 -0.01264  0.27415
-0.05717 -0.02851 -0.06703 -0.13000  0.10379 -0.01308 -0.16613  0.03904  0.10424 -0.28394 -0.27437  0.14563  0.50146 -0.52286
-0.02475 -0.04678 -0.09971 -0.01488 -0.00122  0.01432  0.00229  0.08699 -0.36346 -0.07334 -0.22059  0.12856  0.56617  0.57098
```

$$\sum_{i=1}^{N} (a_{ii}^{(k+1)})^2 - \sum_{i=1}^{N} (a_{ii}^{(k)})^2 < EPS3$$

With the values of EPS1, EPS2 and EPS3 as listed in the printouts, it can be seen that the matrix has for all practical purposes been diagonalized, after only 4 iterations. The diagonal elements are the eigenvalues of A and the eigenvectors $(\phi_{1i}, \phi_{2i}, \ldots, \phi_{Ni})$ corresponding to each eigenvalue λ_i appear in the columns of matrix T. For additional details concerning the flowchart and the program listing, see [19].

Straightforward calculation yields

$$\sum_{i=1}^{14} \sqrt{\text{Vark}_i} = 3.085 \quad > \quad \sum_{i=1}^{14} \sqrt{\lambda_i} = 2.881, \text{ for 1 file statistics}$$

and

$$\sum_{i=1}^{14} \sqrt{\text{Vark}_i} = 3.308 \qquad \sum_{i=1}^{14} \sqrt{\lambda_i} = 3.181, \text{ for 14 file statistics}$$

$\Sigma\sqrt{\text{Vark}_i} - \Sigma\sqrt{\lambda_i} = .204$ for 1 file as opposed to .127 for 14 file statistics. This is to be expected since statistics on some data should yield larger correlation values, than when other less correlated data are added to the previous data.

Table 6.1.5 lists characteristics of the orthogonal parameters θ_i and also the number of levels N_i for METHOD I.

Note that the θ_i's are listed in order of decreasing variance λ_i. The range $\sum_{j=1}^{M} |\phi_{ji}|$ is denoted by R_i.

For 1 file statistics, the total variance is $\sum_{i=1}^{14} \lambda_i = .783$ whereas for the 14 file statistics, it is .888, which is larger as expected. Also the variance is allocated among the parameters in the same way for both statistics. Notice that the range is always much larger than $\sqrt{\mathrm{Var}\theta_i}$. For the smallest λ_i, it is in fact 31 and 26 times larger than R_i for the 1 and 14 file statistics respectively.

The probability distribution of the k_i's does not depend on the filter order M for all $i < M$, i.e. taking two arbitrary filter orders M_1 and M_2, the distributions are the same for $1 \leq i \leq \min(M_1, M_2)$. In [21] a filter order $M = 12$ is used as opposed to $M = 14$ in the present study. Similarly it is expected that the probability distributions of the θ_i's do not depend very much on the value of M if the latter is large because the variance and the cross-correlation of the k_i's decreases as i increases. Comparing the 12 eigenvalues from Table 1 in [21], it is found that the sum of the 12 variances is roughly the same and is also distributed in the same way.

From a previous discussion, the expected spectral deviation for each parameter is $E(\bar{D}_{tot})/M = 3.5\text{dB}/14$. The optimum allocation of levels N_i to each orthogonal parameter θ_i is listed in Table 6.1.5 for METHOD I. N_i is first computed in floating point notation. The values obtained

Table 6.1.5

1 file statistics				14 file statistics			
$E\theta_i$	λ_i	R_i	N_i	$E\theta_i$	λ_i	R_i	N_i
.451	.224	2.695	36	.495	.251	2.085	50
-.033	.160	2.619	22	-.027	.138	2.705	18
.190	.091	2.954	12	-.231	.111	2.158	16
-.124	.055	3.072	9	-.130	.080	2.912	10
.243	.049	2.722	9	.104	.062	2.800	10
.008	.036	2.947	7	-.022	.056	2.904	8
.070	.033	2.930	7	.032	.037	2.839	7
.103	.027	3.171	7	.069	.033	2.844	7
.055	.023	2.696	5	.039	.031	2.233	6
-.081	.019	2.892	6	.054	.026	2.848	6
-.027	.016	2.871	5	-.078	.019	3.276	6
-0.30	.013	2.364	4	-.302	.018	2.414	5
-.090	.010	2.994	4	.061	.015	2.836	5
.025	.007	2.598	3	-.012	.011	2.688	4

are then rounded off to the next greater integer. From inspection of Table 6.1.5, it is seen that with one minor exception under 1 file statistics, N_i decreases as λ_i decreases. Converting levels to bits and allocating β_P bits to pitch, β_G bits to gain, with a frame rate f_r, the total bit rate is

$$\left(\sum_{i=1}^{M} \log_2 N_i + \beta_G + \beta_P\right) f_r$$

In the present study, f_r = 50 Hz, $\beta_G = \beta_P = 5$. In [10], an extra bit per frame is allocated to the variable pre-emphasis $\mu = r(1)/r(0)$. The levels are $\hat{\mu}_1 = 0$, $\hat{\mu}_2 = .9$ and the boundaries are $\mu_1 = 0$, $\mu_2 = .6$, $\mu_3 = 1.0$. But as will be seen under the results of METHOD II, the absence or presence of pre-emphasis quantization is insignificant perceptually. Then, using the above formula for total bit rate, 2539 bits/sec and 2674 bits/sec are required for the 1 file and 14 file statistics respectively, if $E(\bar{D}_{tot})$ in the asymptotic minimum deviation scheme, is not to exceed 3.5 dB.

Table 6.1.6 lists results for the orthogonal parameters and reflection coefficients, using METHOD II. Only the 14 file statistics results of the Jacobi diagonalization will be utilized, because in order to obtain a good representative time average of the sensitivity and relative frequency of occurrence of the parameter, a large number of frames encompassing all 14 files is required. Table 6.1.6a then lists the variance λ_i, the range R_i (both also found in Table 6.1.5), the values $\underline{\theta}_i$ and $\bar{\theta}_i$ at which the probability distribution of θ_i is truncated and the number of levels N_i under the min $E(\bar{D}_{tot})$ scheme, for each of the orthogonal parameters θ_i.

Table 6.1.6b then lists the values $\underline{k}_i$ and $\bar{k}_i$ at which the probability distribution of the k_i's is truncated, the number of levels, N_{i1} using inverse sine quantization, and the number of levels, N_{i2} using the min $E(\bar{D}_{tot})$ quantization scheme, for all k_i's. The number of levels have been calculated using the bound $E(\bar{D}_{tot})/M = 3.5\text{dB}/14$ for all parameters in all 3 of the quantization schemes.

With $\beta_G = \beta_P = 5$, $f_r = 50$ Hz, as in METHOD I, the total number of bits required if a bound $E(\bar{D}_{tot}) = 3.5\text{dB}$ is not to be exceeded, is 3070 bits/sec for inverse sine quantization of the k_i's, 2750 bits/sec for min $E(\bar{D}_{tot})$ quantization of the k_i's and 2884 bits/sec for min $E(\bar{D}_{tot})$ quantization of the θ_i's. Min $E(\bar{D}_{tot})$ quantization of the k_i's is therefore slightly superior to inverse sine quantization of the k_i's as predicted in the theoretical study of [12]. Unfortunately, even though $\sum_{i=1}^{M} \sqrt{R_{ii}} \geq \sum_{i=1}^{M} \sqrt{\lambda_i}$ as was already derived using

Table 6.1.6a

METHOD II

λ_i	N_i	R_i	$\underline{\theta}_i$	$\overline{\theta}_i$
.251	35	2.085	-0.856	1.376
.138	25	2.075	-1.055	1.001
.111	18	2.158	-1.122	.734
.080	14	2.912	-1.048	.728
.062	14	2.800	- .756	.952
.056	11	2.904	- .871	.813
.037	8	2.839	- .653	.823
.033	9	2.844	- .512	.796
.031	8	2.233	- .580	.603
.026	8	2.848	- .427	.655
.019	8	3.276	- .524	.360
.018	6	2.414	- .507	.531
.015	6	2.836	- .425	.454
.011	5	2.688	- .376	.430

Table 6.1.6b

METHOD II

i	N_{i1}	N_{i2}	$\underline{k}_i$	$\bar{k}_i$
1	33	27	-.98	.72
2	26	22	-.85	.97
3	21	16	-.86	.76
4	18	13	-.70	.85
5	17	13	-.79	.73
6	13	10	-.53	.77
7	13	8	-.57	.73
8	11	8	-.50	.75
9	12	9	-.60	.80
10	11	8	-.52	.84
11	9	6	-.53	.65
12	7	5	-.48	.52
13	7	4	-.57	.38
14	5	4	-.36	.41

Hölder's inequality, min $E(\bar{D}_{tot})$ quantization of the orthogonal parameters is not an improvement over min $E(\bar{D}_{tot})$ quantization of the reflection coefficients as far as the bit rate is concerned, given a fixed bound $E(\bar{D}_{tot})$. The final conclusion must however be based on perception tests since the actual hearing mechanism is far from being understood. But first, before quantizing the input parameters, the quantization levels and boundaries must be known. A few approximations will be made in both METHOD I and II. So the graphical results obtained in both cases will first be compared. Figure 6.1.1a and Figure 6.1.2a represent the 14 file statistics Gaussian probability density function of the first and second largest variance θ_i's, as used in METHOD I. Figure 6.1.1b and 6.1.2b are the corresponding 14 file statistics relative frequency of occurrence histograms as used in METHOD II. The corresponding diagrams are to the same scale and a quick inspection will show that they are quite similar. The Gaussian assumption is then not a bad one. For the largest variance θ_i, Figure 6.1.3 is the average sensitivity of METHOD I using 1 file statistics, Figure 6.1.4, using 14 file statistics and Figure 6.1.5 the time averaged sensitivity of METHOD II. All 3 graphs are to the same vertical scales. For Figure 6.1.5, the value of the sensitivity will depend on the number of occurrences at a particular value of θ_i and consequently

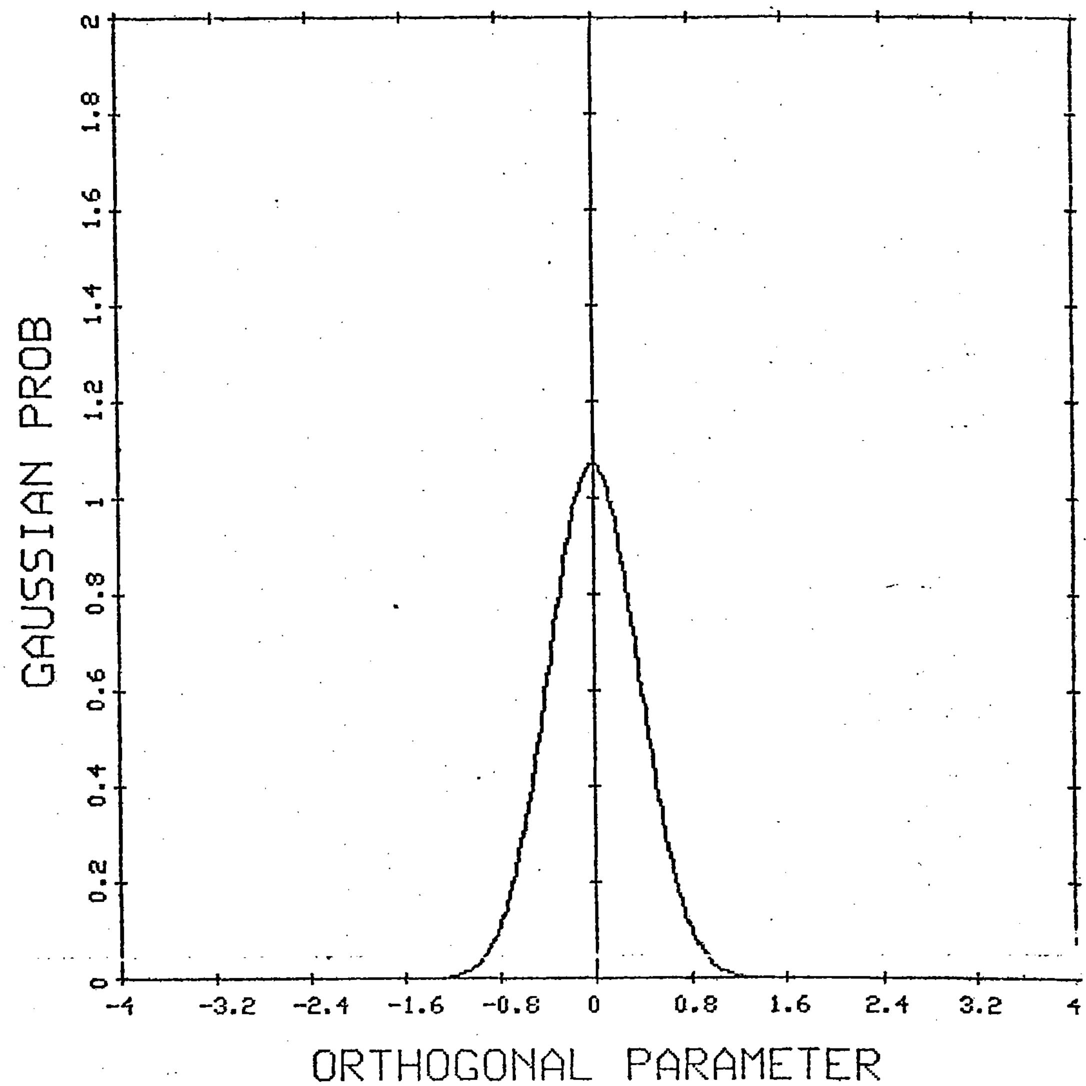


Figure 6.1.1a: Gaussian probability density function of the second largest variance orthogonal parameter.

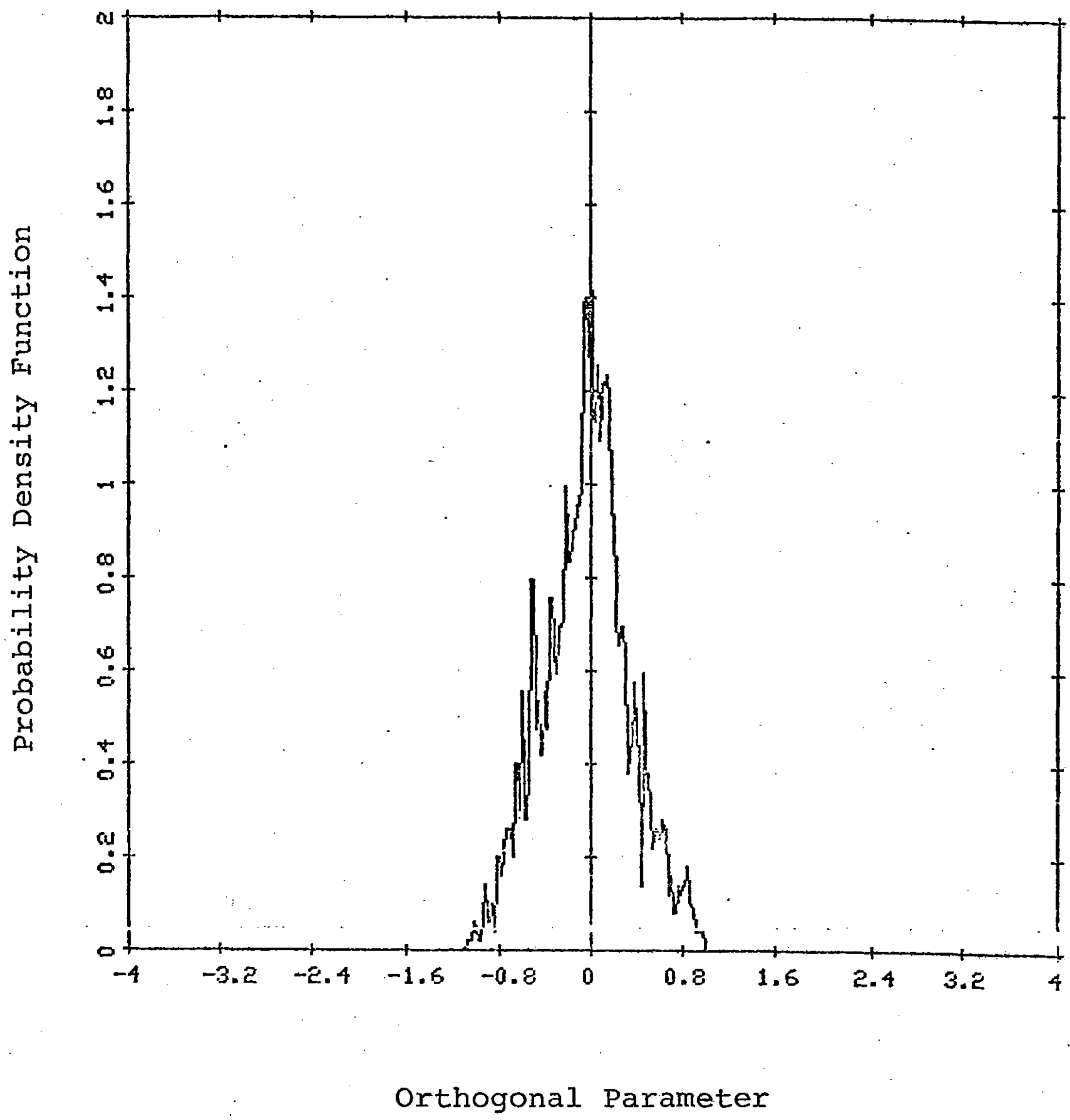


Figure 6.1.1b: Relative frequency of occurrences histogram of the second largest variance orthogonal parameter.

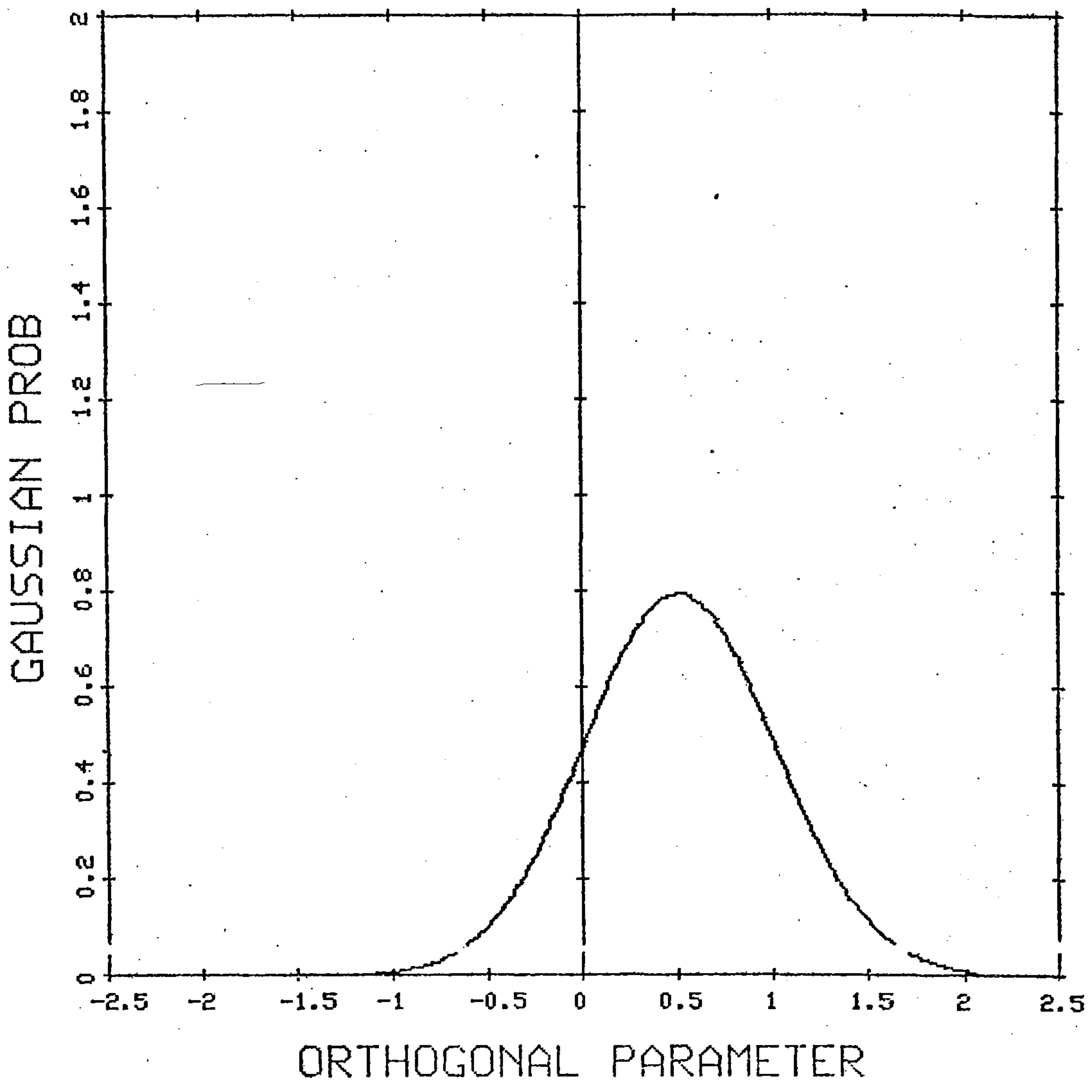


Figure 6.1.2a: Gaussian probability density function of the largest variance orthogonal parameter.

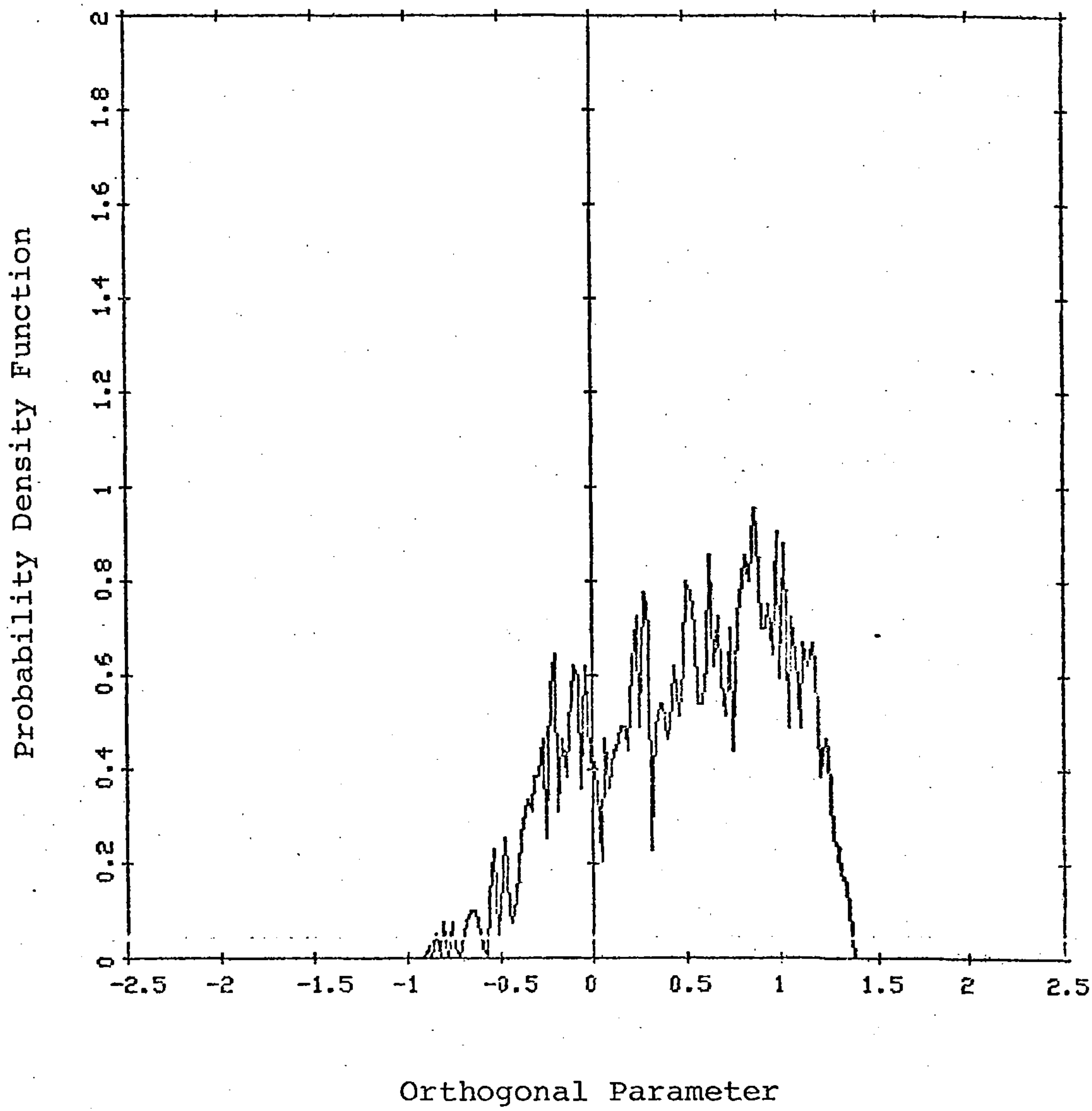


Figure 6.1.2b: Relative frequency of occurrences histogram of the largest variance orthogonal parameter.

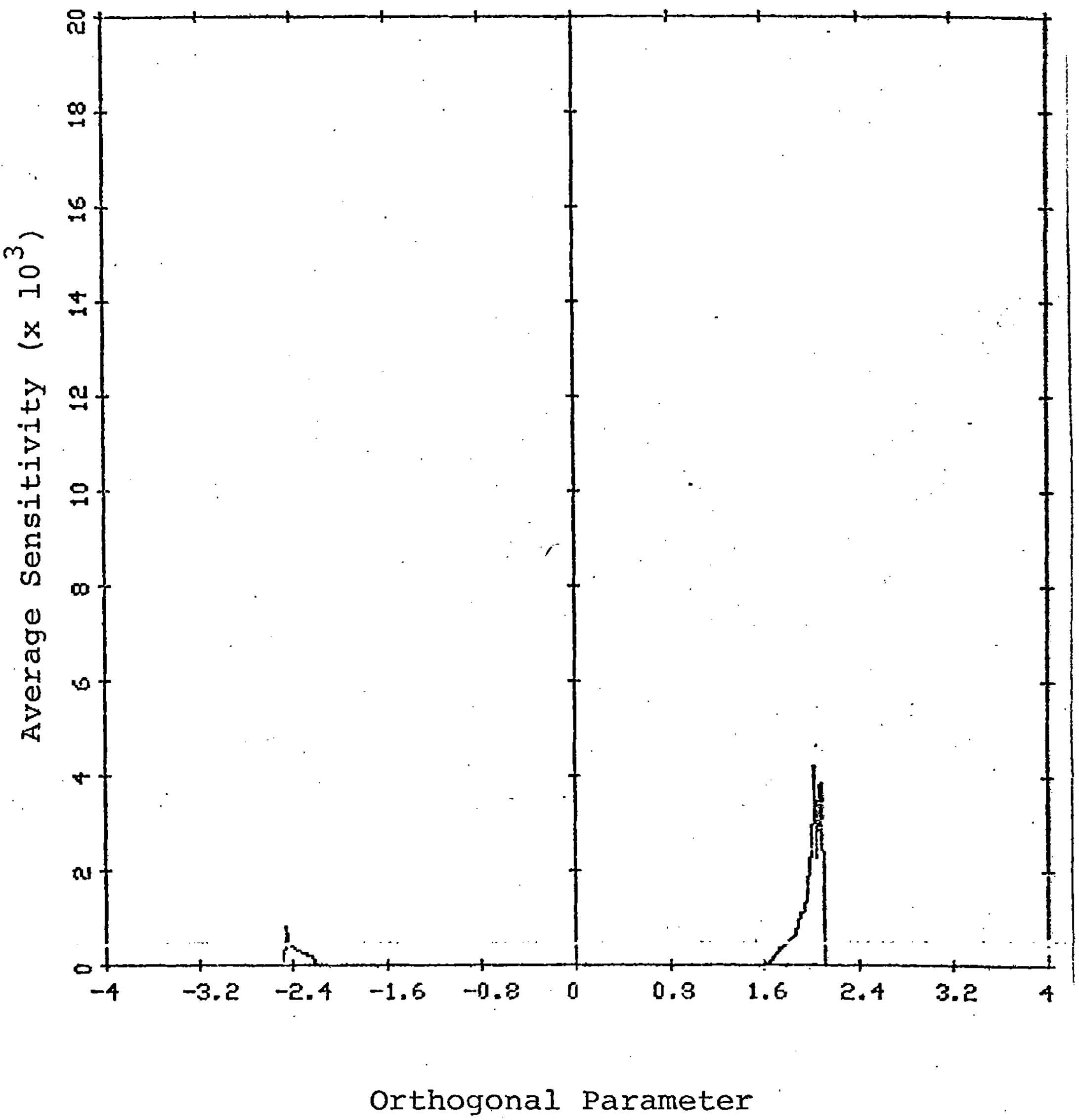


Figure 6.1.3: The average sensitivity function of the largest variance orthogonal parameter, using METHOD I with 1 file statistics.

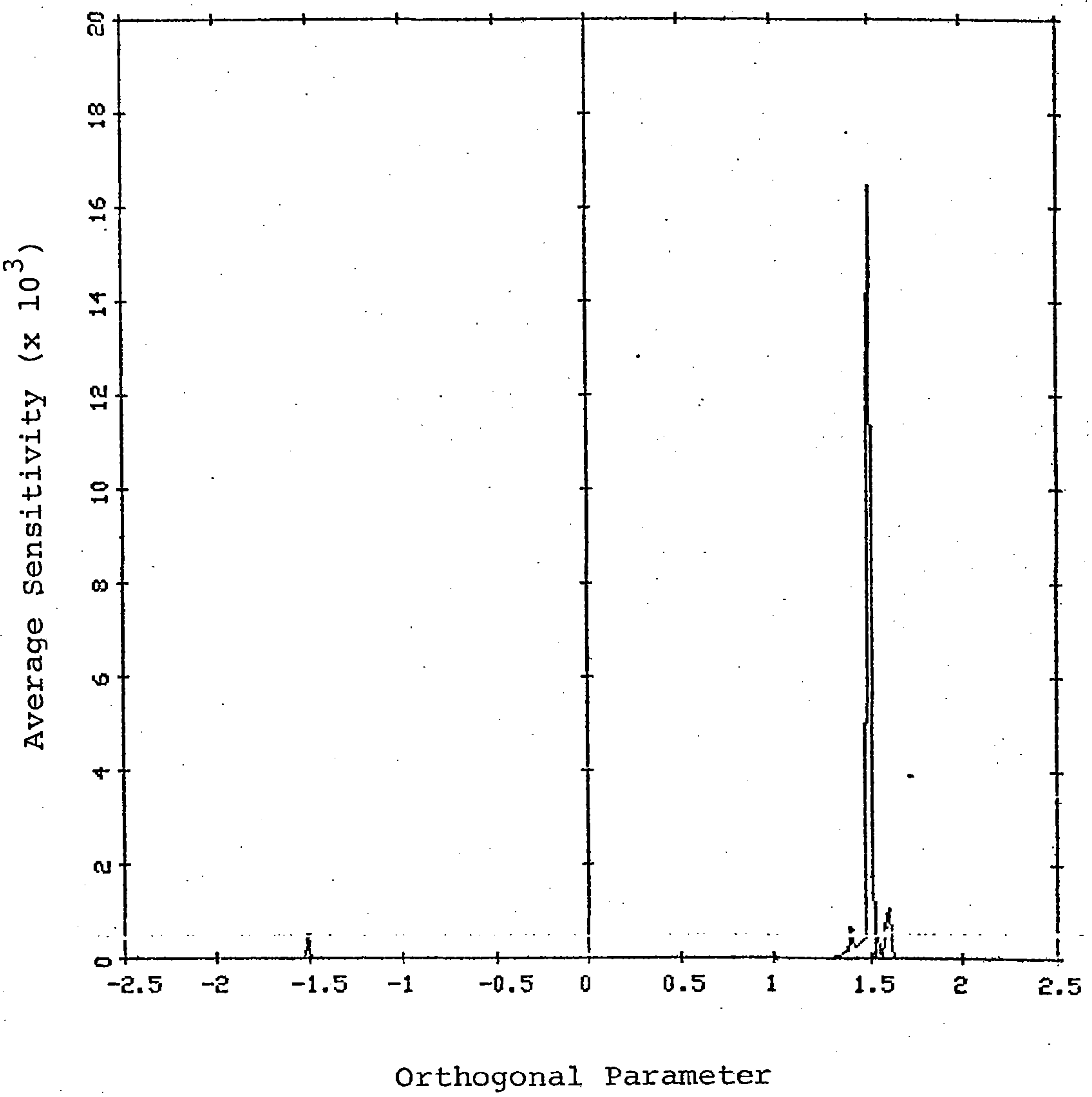


Figure 6.1.4: The average sensitivity function of the largest variance orthogonal parameter, using METHOD I with 14 file statistics.

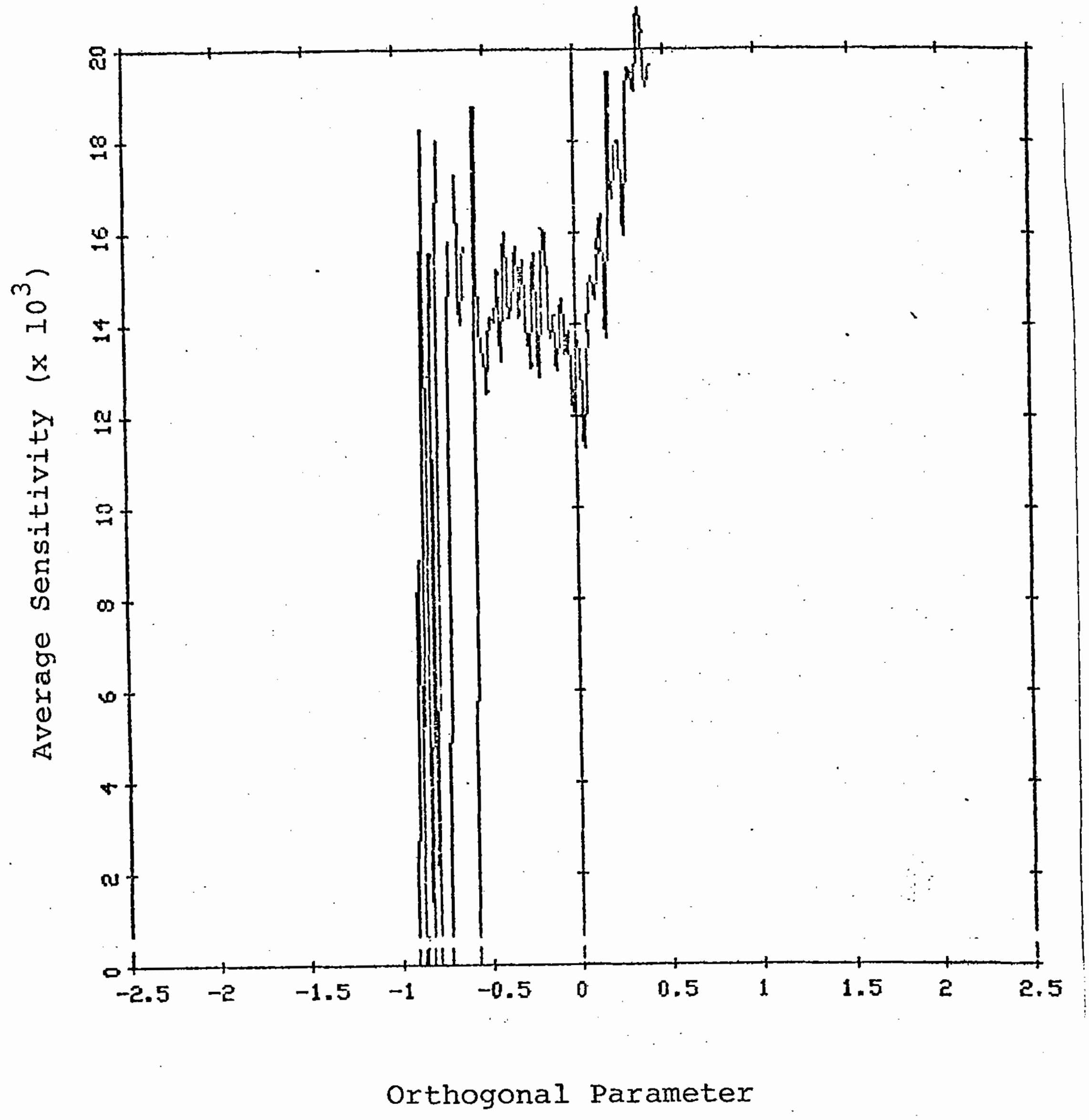


Figure 6.1.5: The time-averaged sensitivity function of the largest variance orthogonal parameter.

the graph is truncated because the p.d.f. of the orthogonal parameter is truncated. Extrapolation of these results outside this truncated range would give the indication that the sensitivity might be unbounded as $\theta_i \to \pm R_i$. This would not be surprising in view of the fact that s_{θ_i} is a linear combination of s_{k_i}'s each of which becomes unbounded as $\theta_i \to \pm R_i$ because then all $|k_j| \to 1$. The truncated p.d.f. will however be responsible for flattening out the quantizer curve U(x) as θ_i moves away from $E\theta_i$. Figure 6.1.3 and 6.1.4 show clear spikes outside the above truncated interval, a region where the values $E\theta_m$ had to be changed to (5.4.20) or to $\tilde{\theta}_m$ as explained earlier. It is therefore seen that as far as the sensitivity is concerned, METHOD I and II give quite different results. There was no guarantee that the outcome should be similar under the assumption that the average of the sensitivity for a fixed θ_i is given by the sensitivity at, the average values of θ_m, or $\tilde{\theta}_m$, or by (5.4.20) for all $m \neq i$. Nevertheless taking this sensitivity function in conjunction with the Gaussian density seems to give comparable results for the number of levels and as will also be seen below, for the shape of the quantizer curves.

Similar sets of 3 sensitivity graphs are obtained for all smaller variance orthogonal parameters. Figures 6.1.6, 6.1.7, 6.1.8 are the min $E(\bar{D}_{tot})$ quantizer curves of the

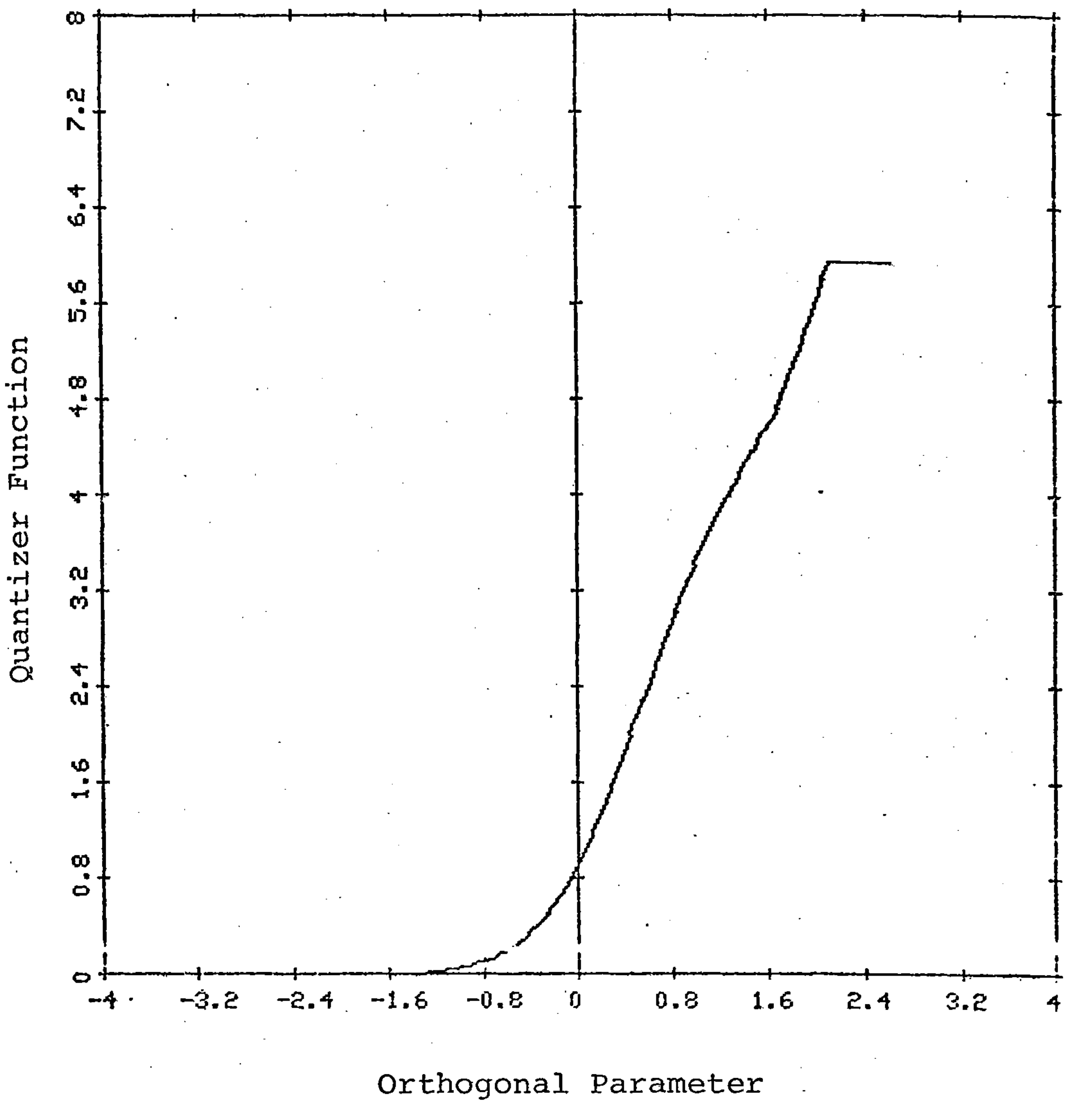
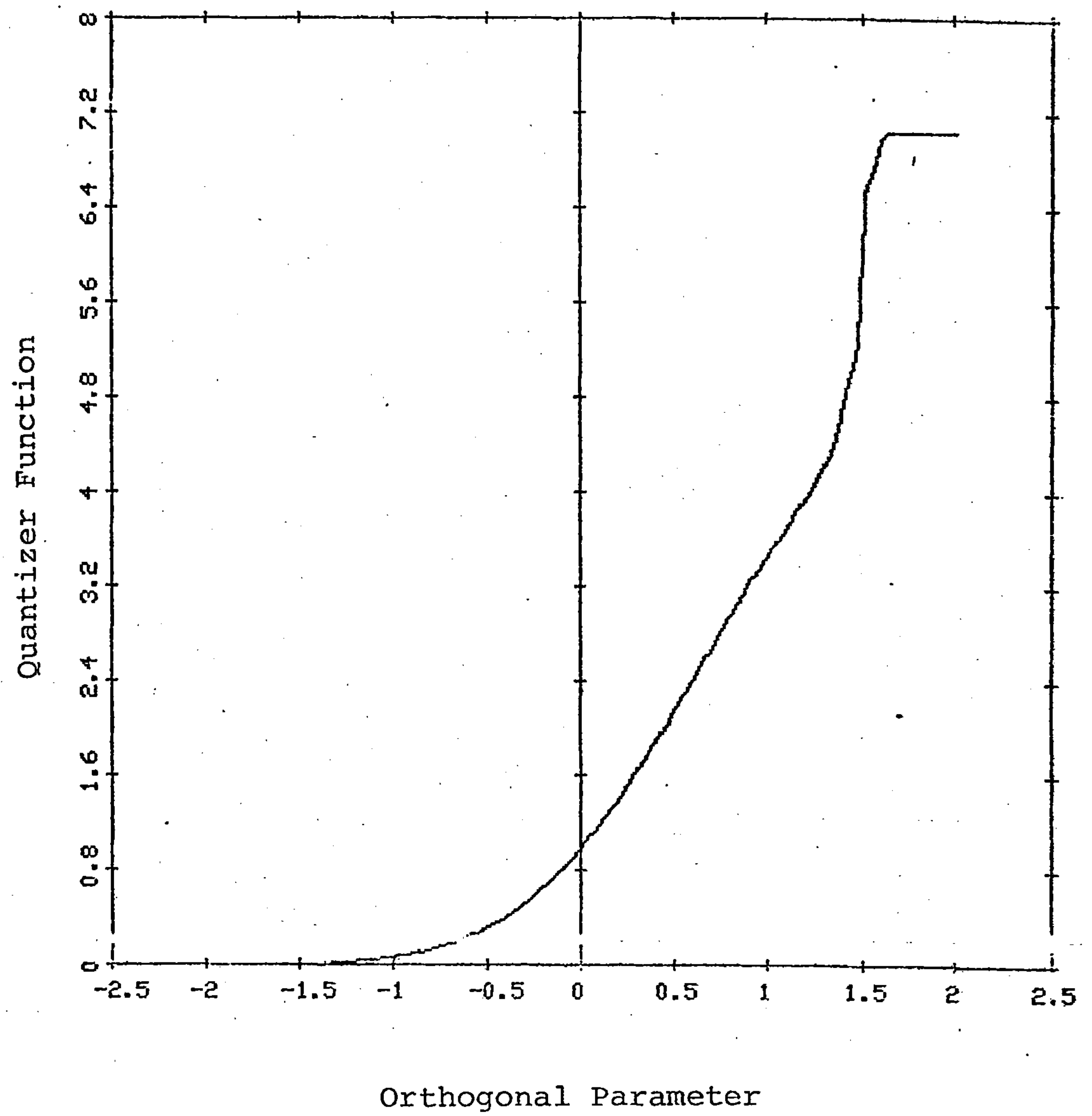
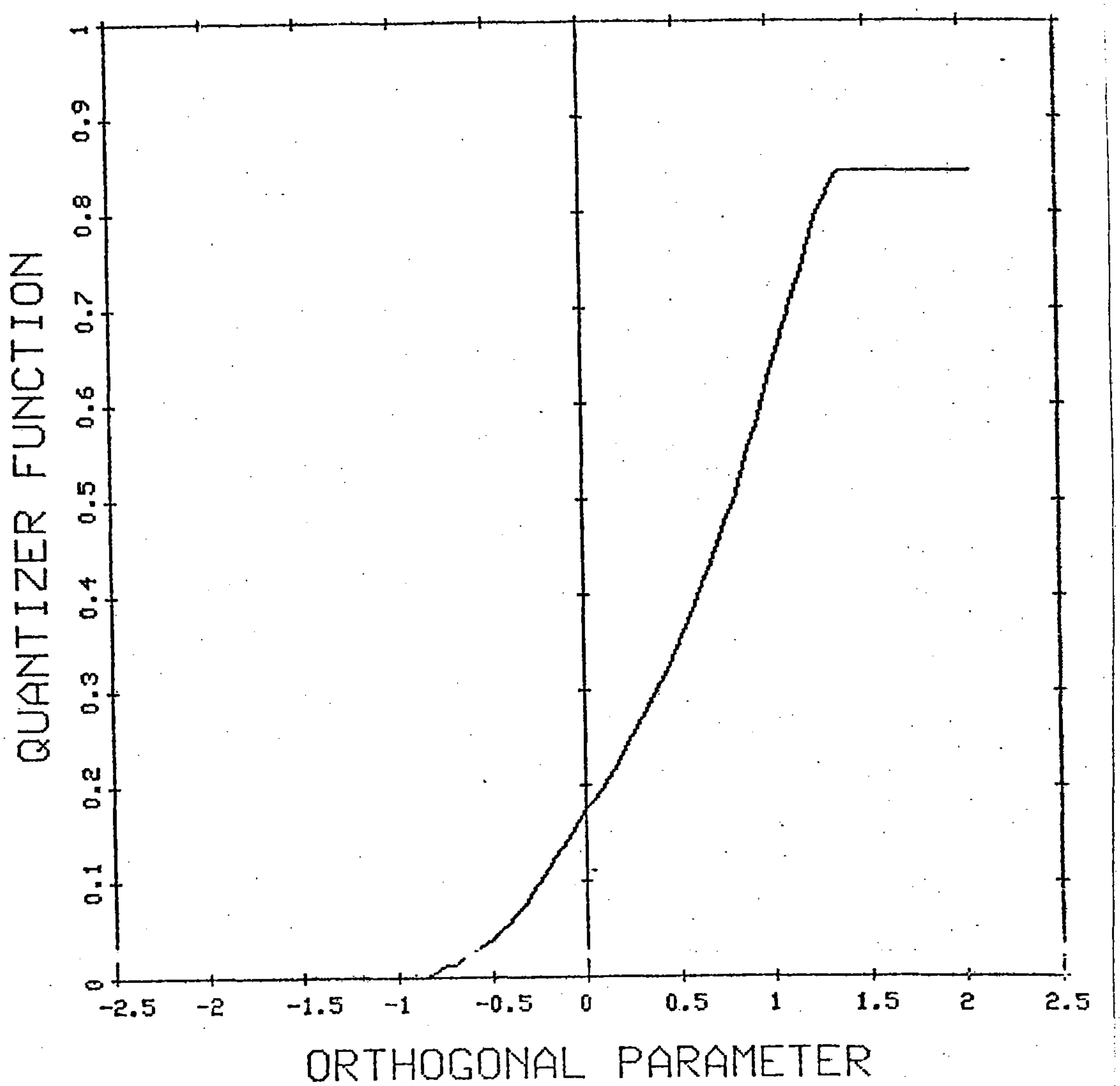


Figure 6.1.6: The unnormalized quantizer curve for the largest variance orthogonal parameter using METHOD I with 1 file statistics.



<u>Figure 6.1.7:</u> The unnormalized quantizer curve for the largest variance orthogonal parameter using METHOD I with 14 file statistics.



<u>Figure 6.1.8:</u> The unnormalized quantizer curve for the largest variance orthogonal parameter using METHOD II.

largest variance θ_i for METHOD I using 1 file statistics, METHOD I using 14 file statistics and METHOD II using 14 file statistics respectively. The third graph is somewhat different from the first two and is not to the same scale either. As far as finding the levels and boundaries it is only necessary to know the shape of the quantizer curve although its correct normalization is required in computing the number of levels. It is seen from Table 6.1.6a or from Figure 6.1.1b that the quantizer curve of Figure 6.1.8 is flat outside the range defined by the values at which the probability density function of the parameter is truncated. This transition is less abrupt in Figure 6.1.7 since a true Gaussian density is used as the p.d.f. It was judged superfluous to include the corresponding graphs of the smaller variance parameters as they were even more comparable and symmetrical about a vertical line close to $E\theta_i$.

Figure 6.1.9a, 6.1.10a, 6.1.11a are respectively the relative frequency of occurrence histogram, the time averaged sensitivty function and the min $E(\bar{D}_{tot})$ quantizer curve for the first reflection coefficient. Figure 6.1.9b, 6.1.10b, and 6.1.11b are the corresponding graphs for the second reflection coefficient. Of course, the time averaged sensitivity function will depend on the number of occurrences at any given value of the reflection coefficient and consequently the graphs are truncated at the values at which

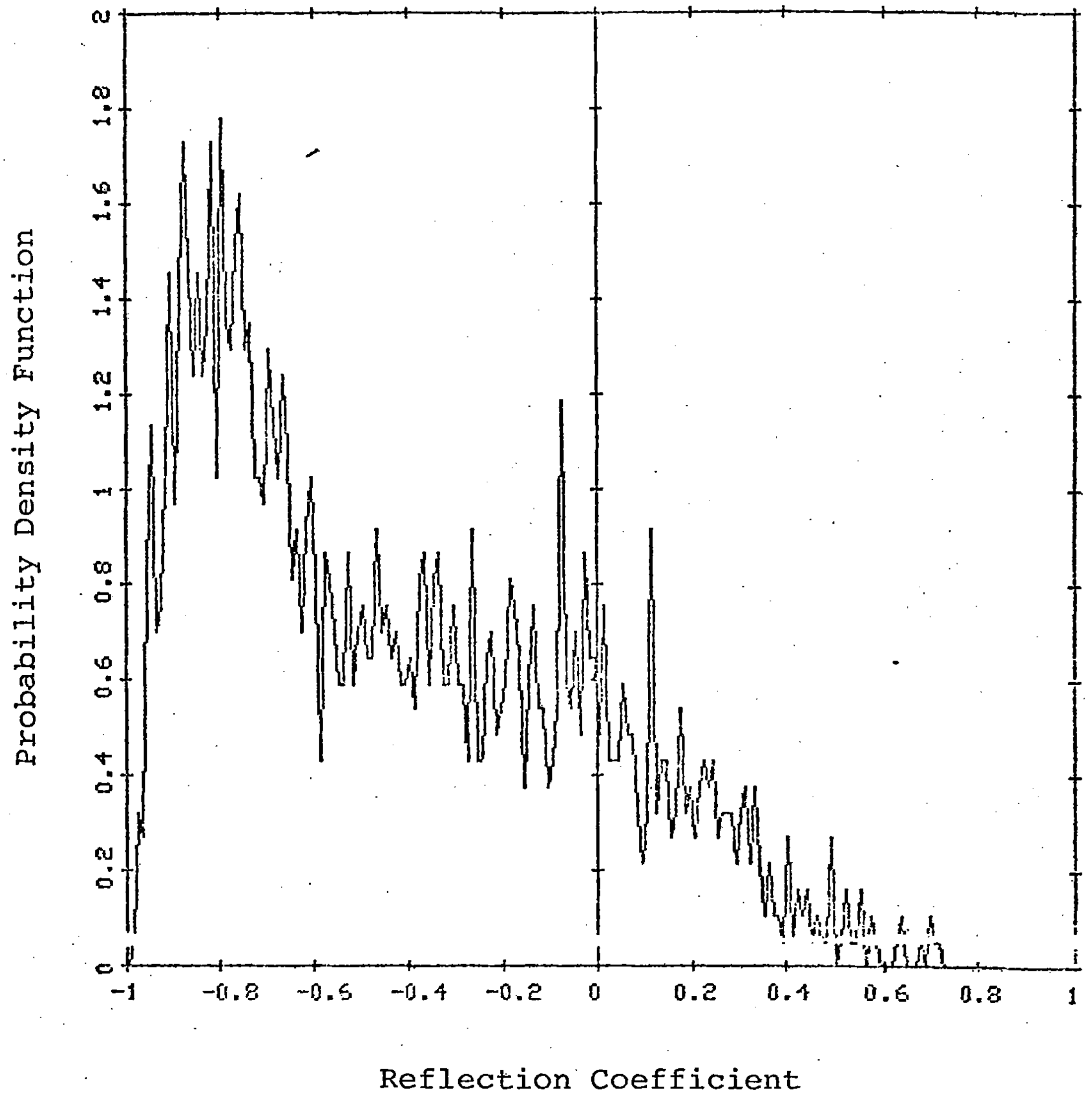


Figure 6.1.9a: Relative frequency of occurrences histogram of the first reflection coefficient.

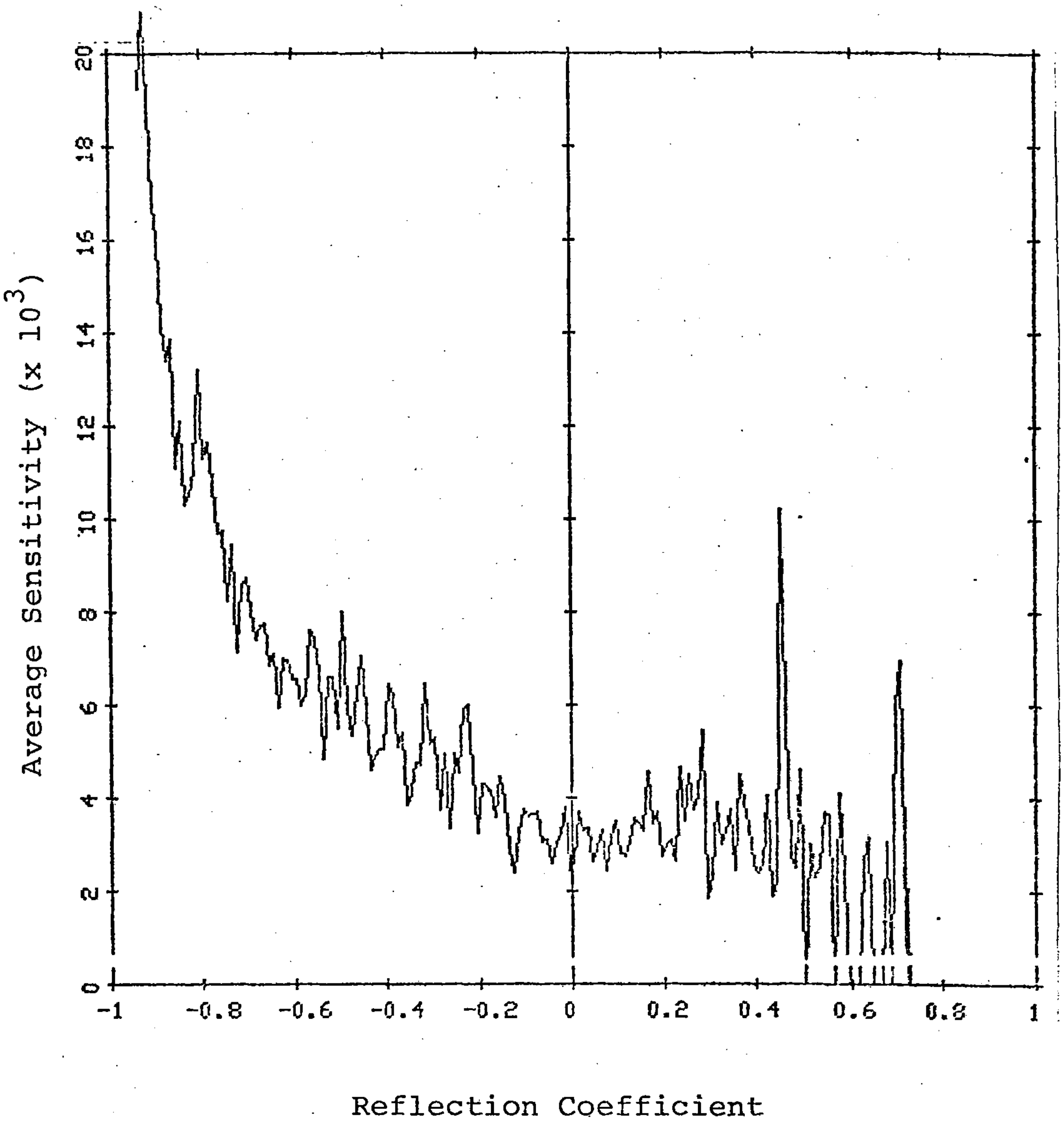


Figure 6.1.10a: The time-averaged sensitivity function of the first reflection coefficient.

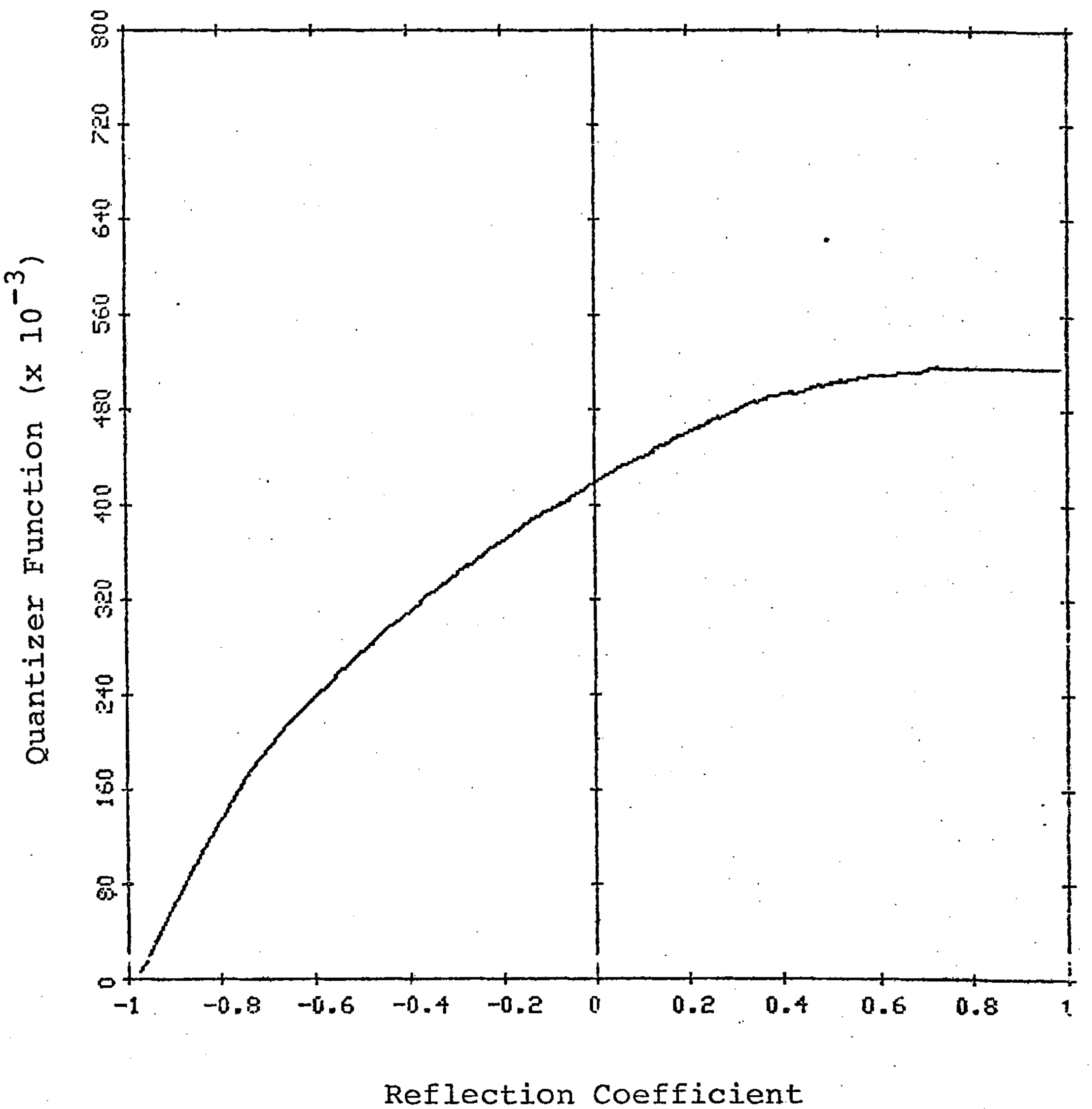


Figure 6.1.11a: The unnormalized quantizer curve for the first reflection coefficient.

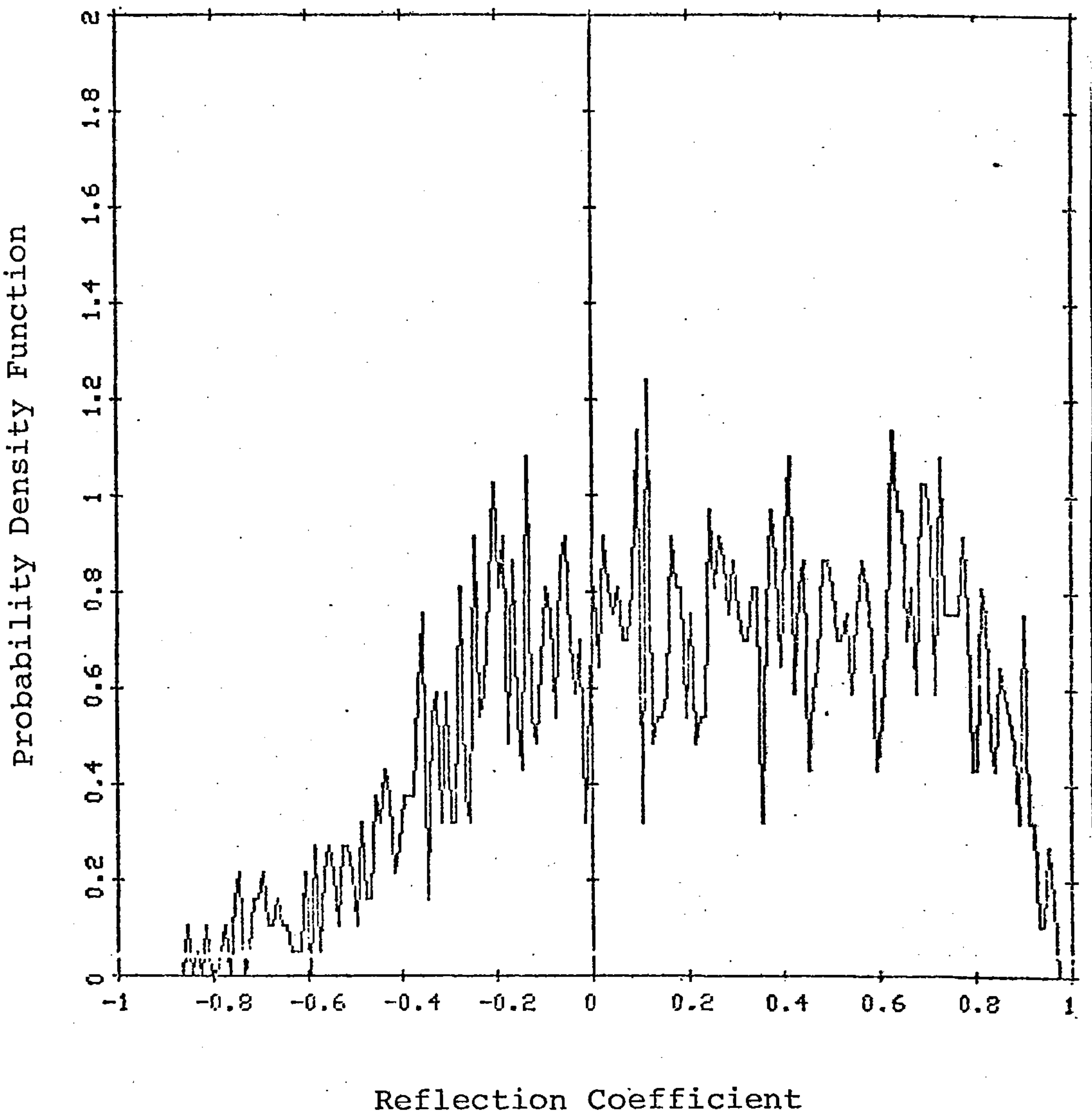


Figure 6.1.9b: Relative frequency of occurrences histogram of the second reflection coefficient.

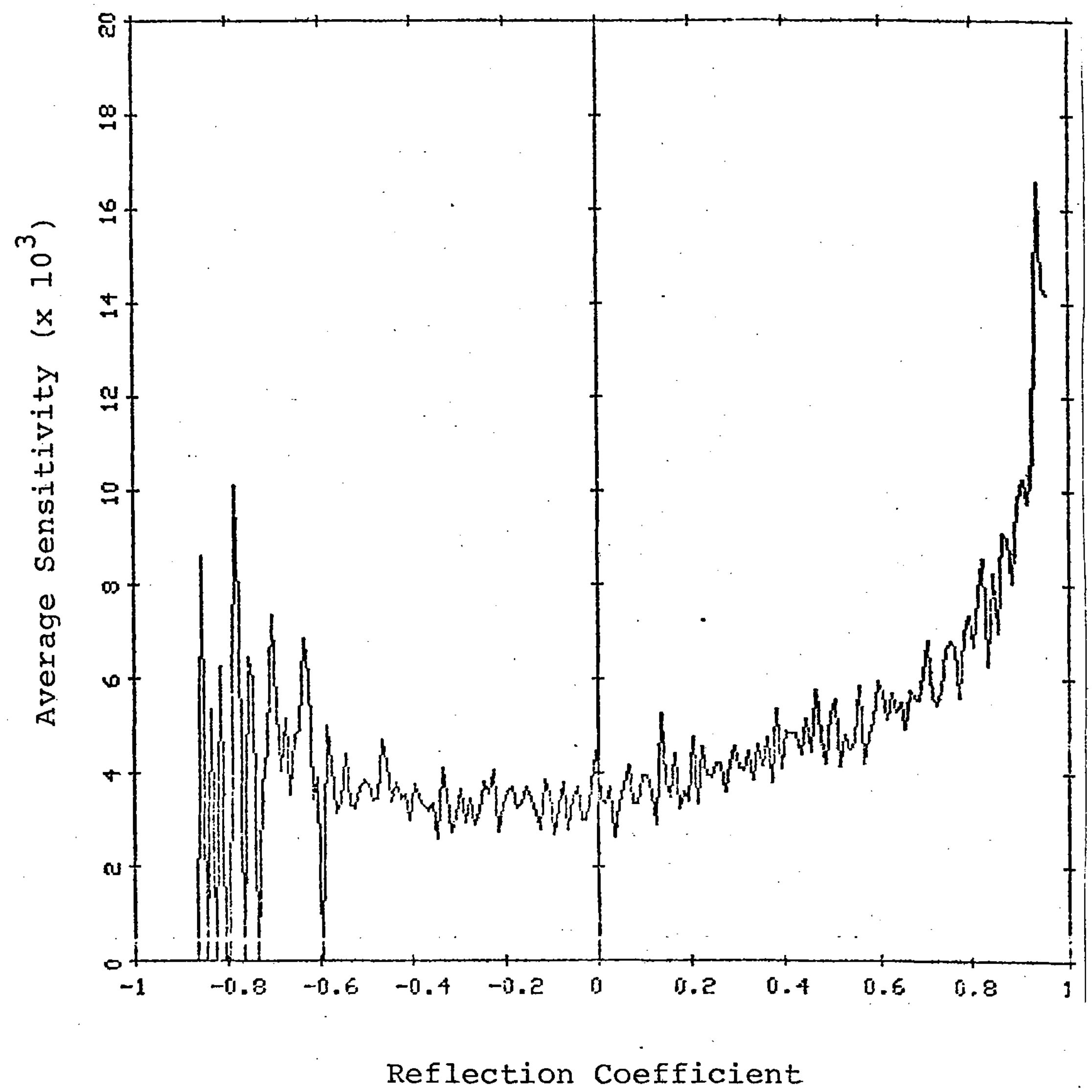


Figure 6.1.10b: The time-averaged sensitivity function of the second reflection coefficient.

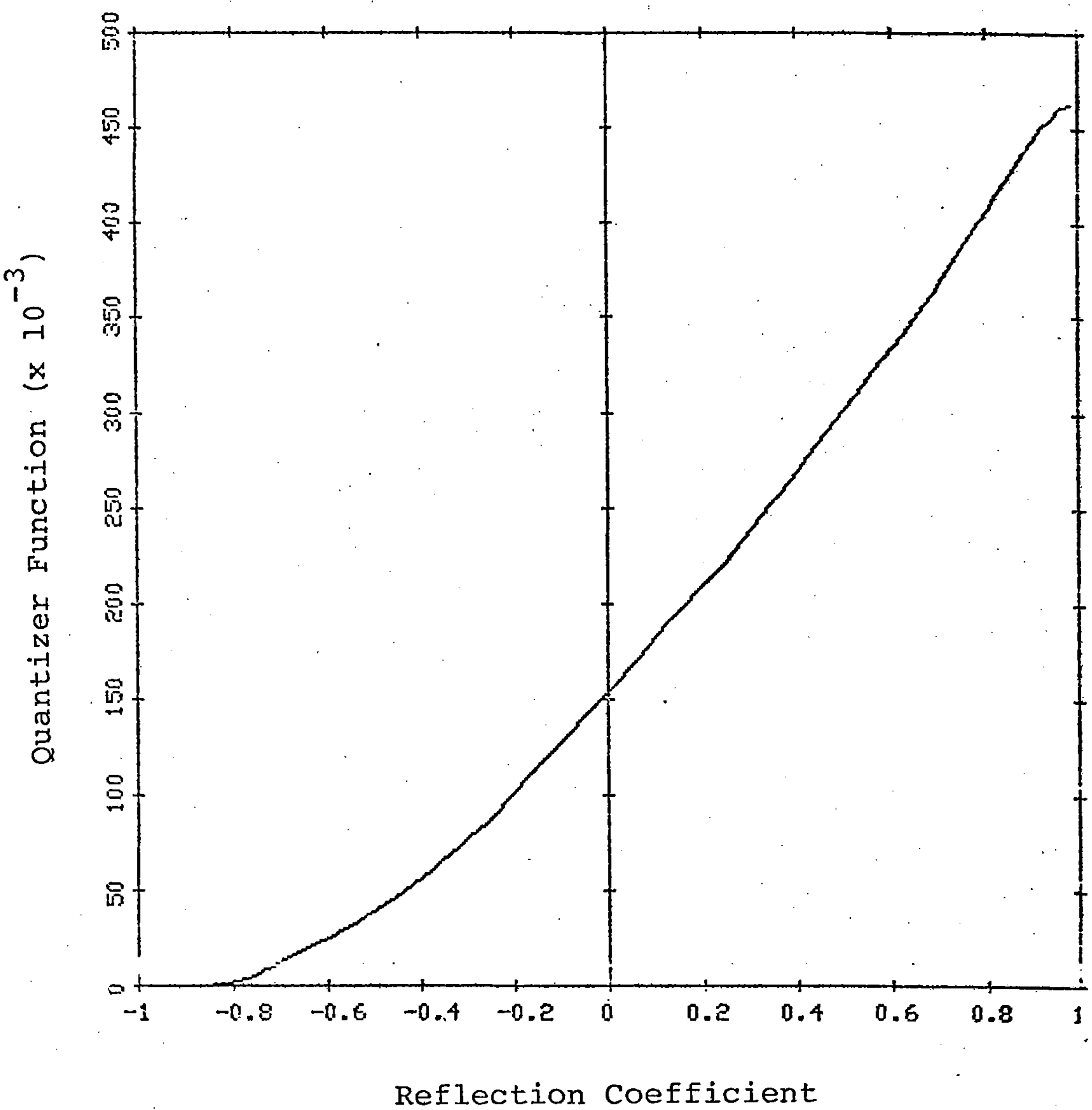


Figure 6.1.11b: The unnormalized quantizer curve for the second reflection coefficient.

the probability density function of the k_i's is truncated. It can be seen that the general shape of these histograms is in agreement with the histograms and scatter plots of [12], [14]. Of course, the quantizer curve is flat outside the truncated range. It was also found unnecessary to include the graphs for the other k_i's because as i increases, the quantizer curves of the k_i's become more symmetrical about a vertical line close to Ek_i and in fact their shape is more reminiscent of that of the orthogonal parameters' quantizer curves.

In the calculation of the number of levels and shape of the quantizer curves in the min $E(\bar{D}_{tot})$ quantization scheme, it was mentioned already, that the integrals are approximated by Simpson's Rule with 200 subdivisions. Therefore 200 values of sensitivities and probabilities are computed and assigned to the point midway between the ends of each subdivision, and then 99 values of the unnormalized quantizer curve

$$U(x) = \int_a^x \sqrt{s_\Lambda(\lambda) p_\Lambda(\lambda)}\, d\lambda$$

are obtained for the corresponding values x which are equally spaced by twice the original subdivision length. Denoting the range by (a,b), the number of levels is then

$$N = [U(b)]^2 \big/ [4E(\bar{D}_{tot})]/M$$

Let W_1 and W_2 be respectively the closest values of x to a and b. Then $z = U(x)$ can be uniformly quantized in the range $(U(W_1), U(W_2))$ because since the number of subdivisions is large, W_1 and W_2 will be respectively close enough to a and b to ensure that the quantizer curve $U(x)$ will be flat outside a truncated range $(t_1,t_2) \subset (W_1,W_2) \subset (a,b)$.

It is then easy to compute all levels and boundaries $\hat{z}_n$, z_n if the number of levels is known. The problem is then to find which of the 99 values $U(x)$ is closest to one of the computed $\hat{z}_n$ (or z_n). Since $U(x)$ is obviously monotonic in x, a Fortran program is easily implemented with a few DO-LOOP's, that will search for those values x_i, x_{i+1}, x_j, x_{j+1} which satisfy

$$W_1 = x_1 < x_2 \;\; \ldots\ldots < x_{99} = W_2$$

and $$U(x_i) \leq \hat{z}_n \leq U(x_{i+1}) \leq U(x_j) \leq z_{n+1} \leq U(x_{j+1})$$

for all n.

The problem is then to find values $\hat{x}_n$, x_{n+1} which satisfy

$$x_i \leq \hat{x}_n \leq x_{i+1}$$

and $$x_j \leq x_{n+1} \leq x_{j+1}$$

such that $U(\hat{x}_n) = \hat{z}_n$ and $U(x_{n+1}) = z_{n+1}$. Since z was not computed as an analytic function of x, but is rather found empirically, the inverse function U^{-1} is unknown. However because the number of subdivisions is large, the function U in the interval (x_i, x_{i+1}) can be approximated by a straight line and thus, linear interpolation can then be performed. Consequently, $\hat{x}_n$ is solved for, by using

$$\frac{\hat{x}_n - x_i}{x_{i+1} - x_i} = \frac{\hat{z}_n - U(x_i)}{U(x_{i+1}) - U(x_i)}$$

This idea was applied in the min $E(\bar{D}_{tot})$ scheme of both METHOD I and II. For inverse sine quantization of the k_i's however, it is only necessary to uniformly quantize $z = \sin^{-1} k_i$ in the interval $(\sin^{-1} \underline{k}_i, \sin^{-1} \bar{k}_i)$ and to apply the inverse transformation to get $\hat{x}_n = \sin \hat{z}_n$ and $x_{n+1} = \sin z_{n+1}$. The values $\underline{k}_i$ and $\bar{k}_i$ are taken from Table 6.1.6b.

Subjective results and Conclusion

It is first checked that the original file M-1-3 is perfectly reconstructed when played back through the converter in D/A mode. Figure 6.1.12a shows the time domain representation of the file covering 2.432 seconds of speech, sampled at 10 KHz and bandlimited to 5 KHz. Figure 6.1.12b is a corresponding low time resolution spectrogram of the

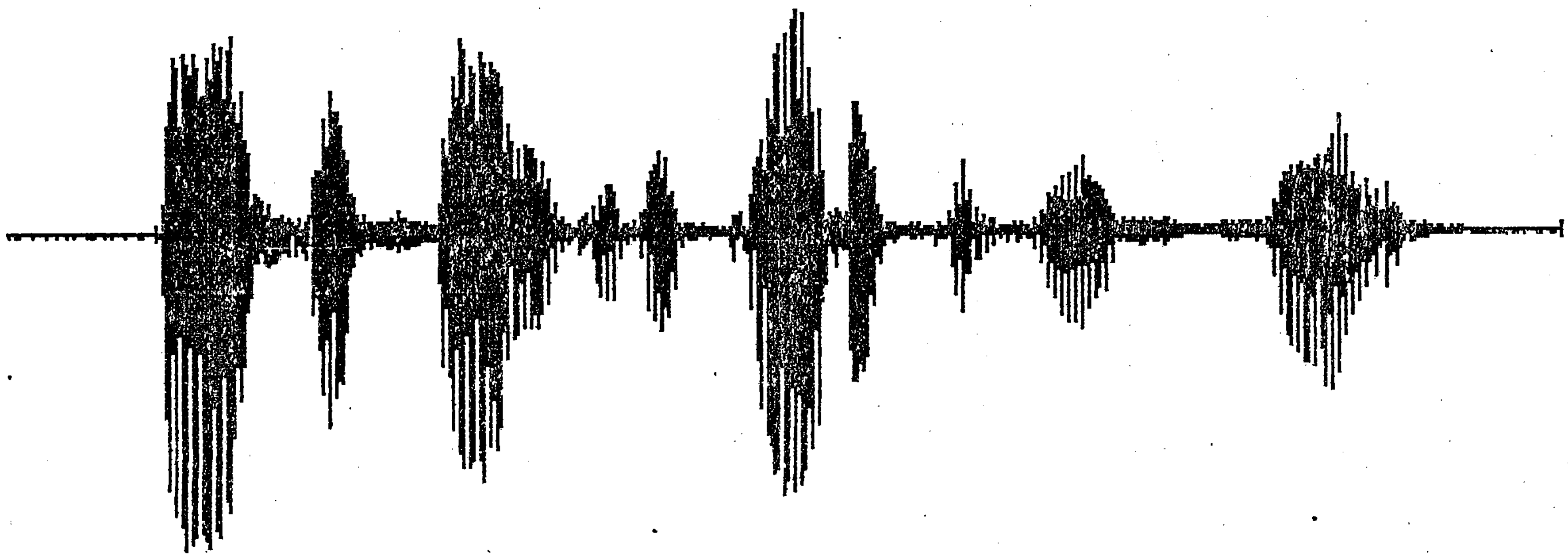
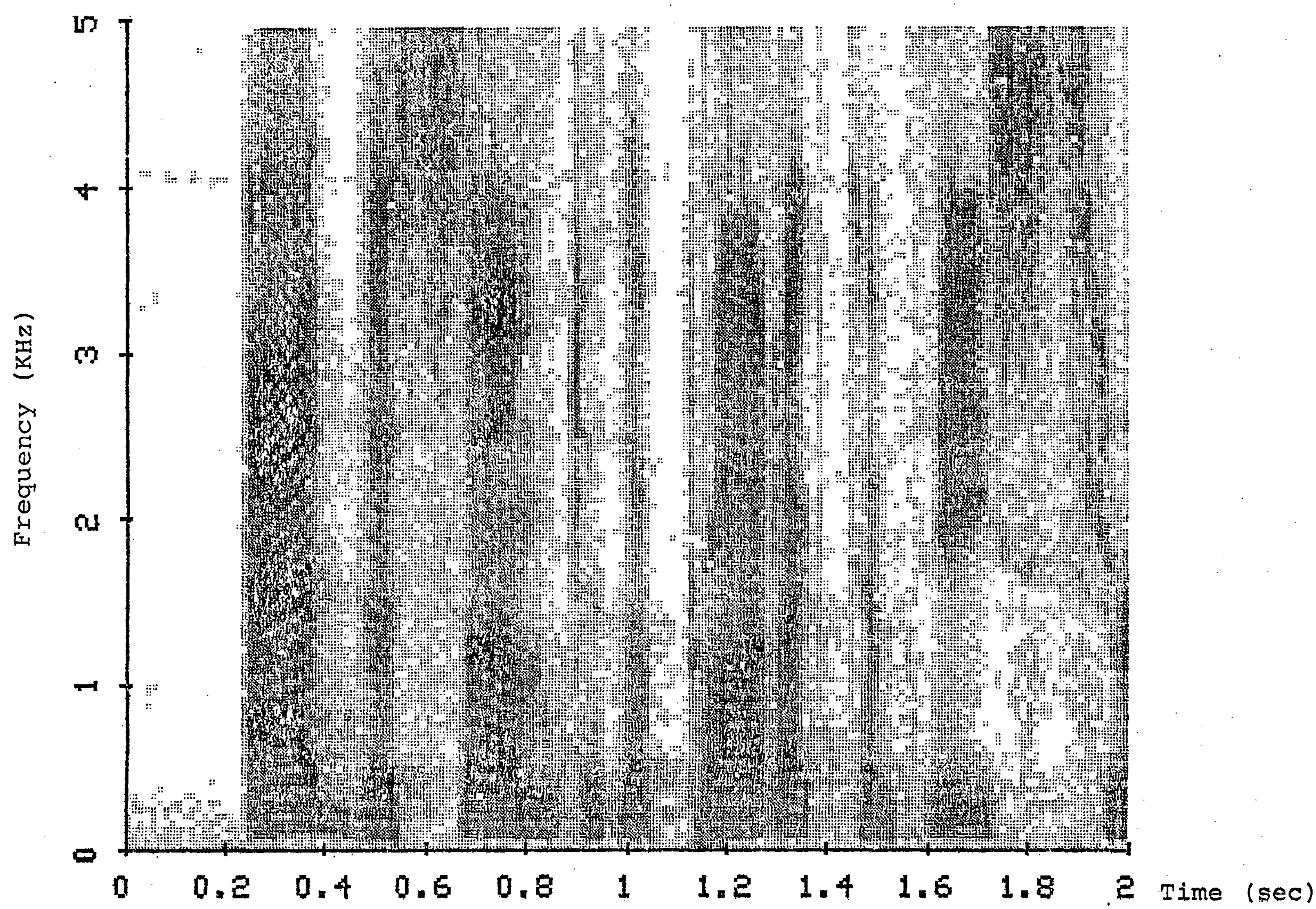


Figure 6.1.12a: The original utterance 'add the sum to the product of these three', (also denoted by file M-1-3).



<u>Figure 6.1.12b:</u> Spectrogram corresponding to the first two seconds in Figure 6.1.12a.

first 2 seconds of speech. (An FFT of length N=128 is used. The 128 speech samples are first windowed using a Hamming window with a scale factor of .54). The darker areas indicate larger concentration of energy. The horizontal striations represent harmonics of the pitch period. If, in a particular interval, these are absent and there is a non-zero concentration of energy, then this interval of time corresponds to unvoiced speech. The frequency axis extends only up to 5 KHz since the speech is bandlimited.

The speech parameter file obtained in autocorrelation analysis is then inputted to the synthesizer program discussed in section 4.5.

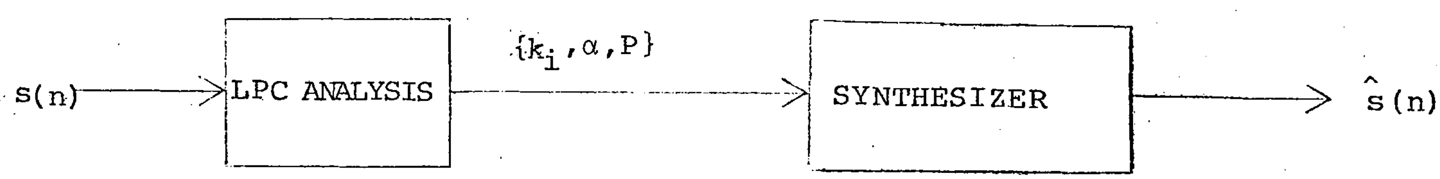


Figure 6.1.13

Since there is no quantization involved (except for the negligible quantization implicit in the integer storage of the speech samples) the reconstructed speech utterance should be the one most similar to the original one. Other utterances by the 3 male speakers were also analyzed in this way. The output speech is of acceptable quality and nothing peculiar was discerned that was not already discussed in section 4.6.

Figure 6.1.14a and 6.1.14b represent the time domain and corresponding spectrogram respectively. Figure 6.1.15a and 6.1.15b demonstrate the fact that quantization of pre-emphasis to 2 levels results in output speech virtually indistinguishable from non-quantized synthesized speech. Figure 6.1.16a and 6.1.16b demonstrate results when, in addition to pre-emphasis, pitch and gain are both logarithmically quantized to 5 bits. The only noticeable change is the repression of a few consecutive peaks in the middle of the time domain diagram.

Figure 6.1.17 shows the sequence of steps that was followed in obtaining synthesized speech using inverse sine quantization of the k_i's.

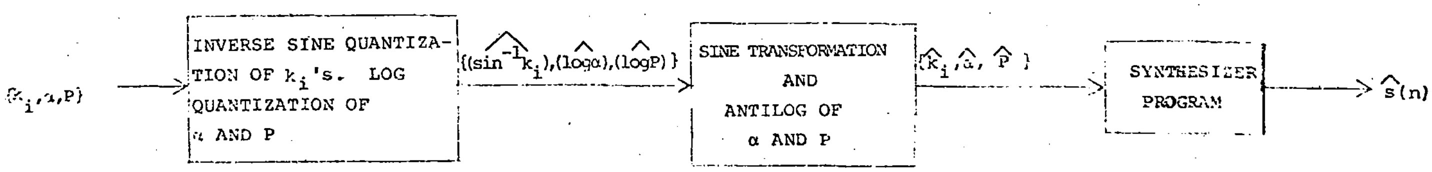


Figure 6.1.17

Figure 6.1.18a and 6.1.18b are the time domain and spectrogram respectively of the output speech, at a total bit rate of 3070 bits/sec. A slight degradation in quality is now perceived when the speech is compared with non-quantized synthesized speech.

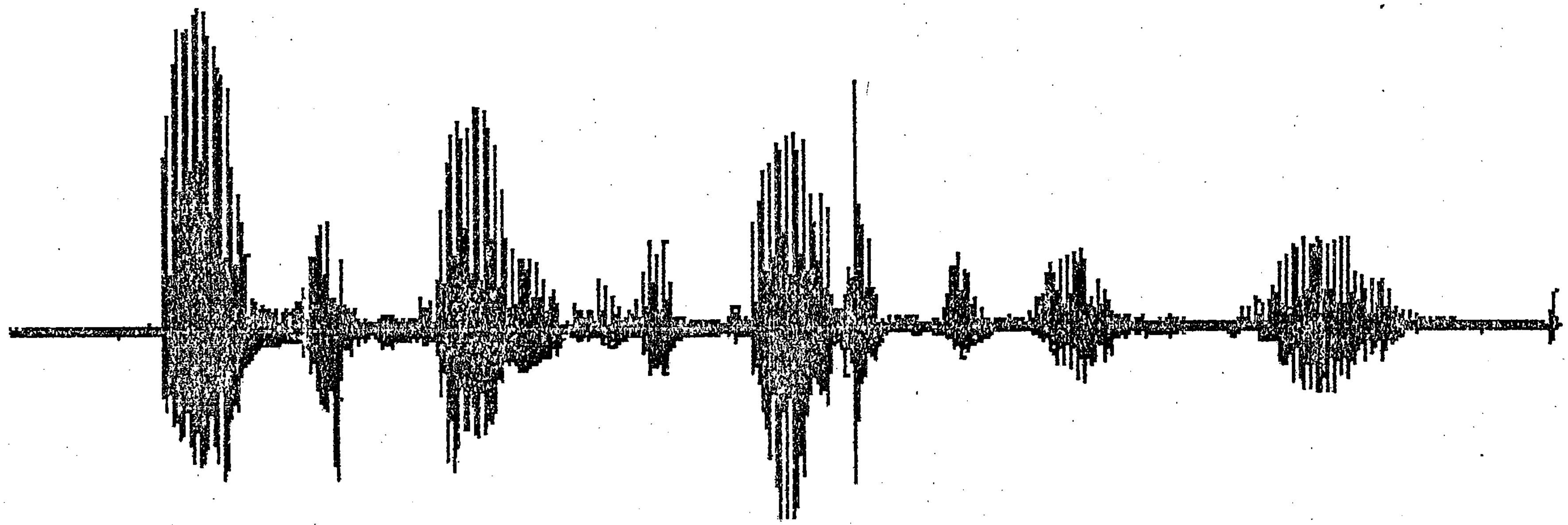


Figure 6.1.14a: The synthesized version of file M-1-3 with no quantization at all.

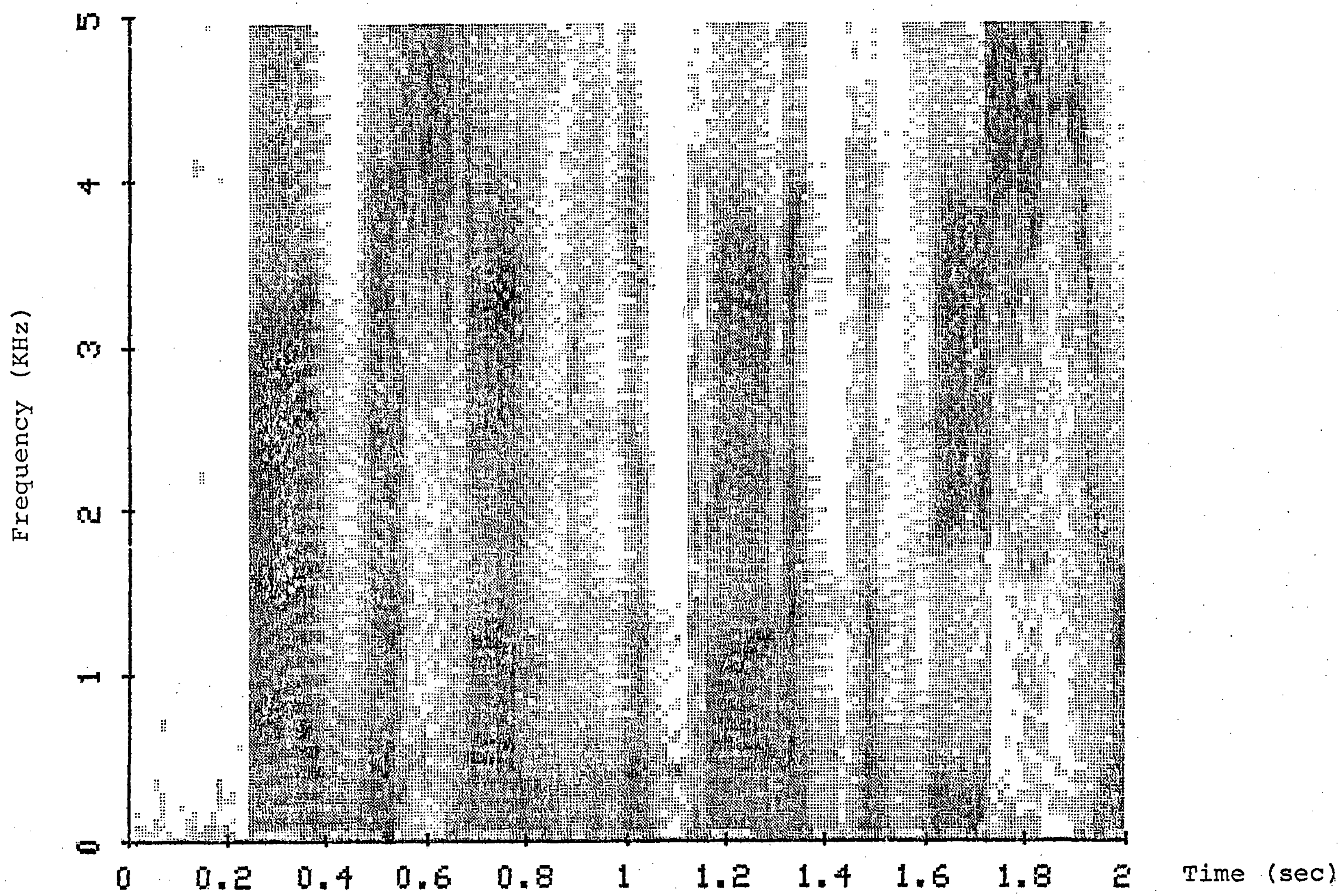
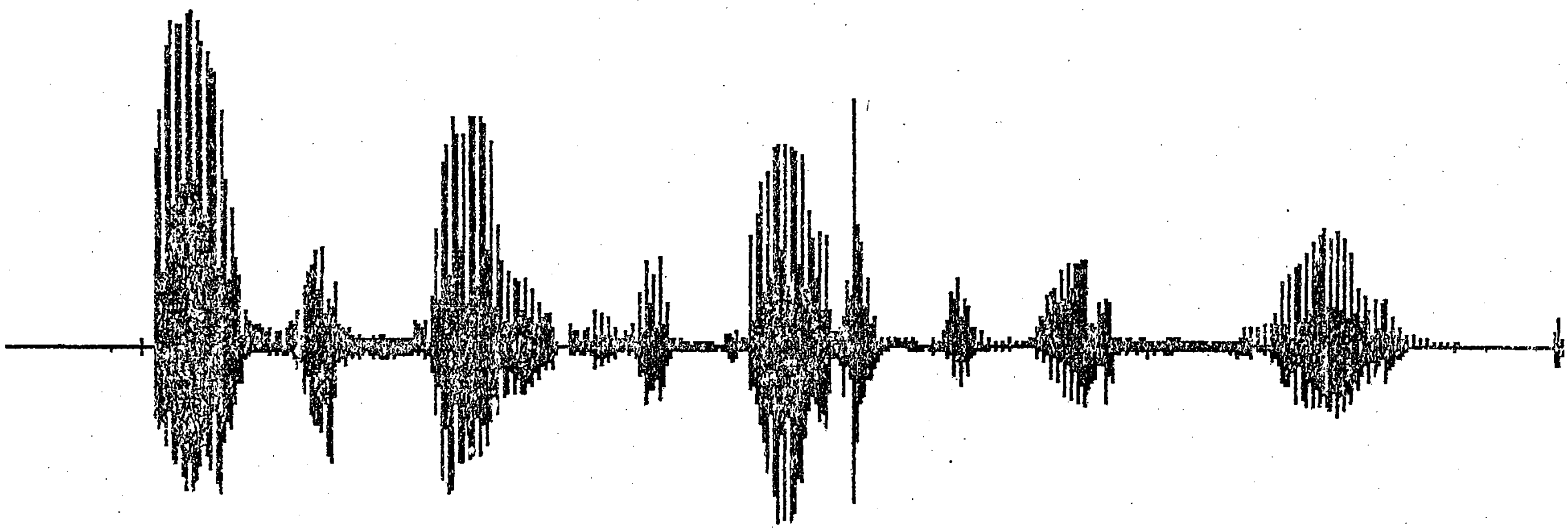


Figure 6.1.14b: Spectrogram corresponding to the first two seconds in Figure 6.1.14a.



<u>Figure 6.1.15a:</u> The synthesized version of file M-1-3, with pre-emphasis quantization only.

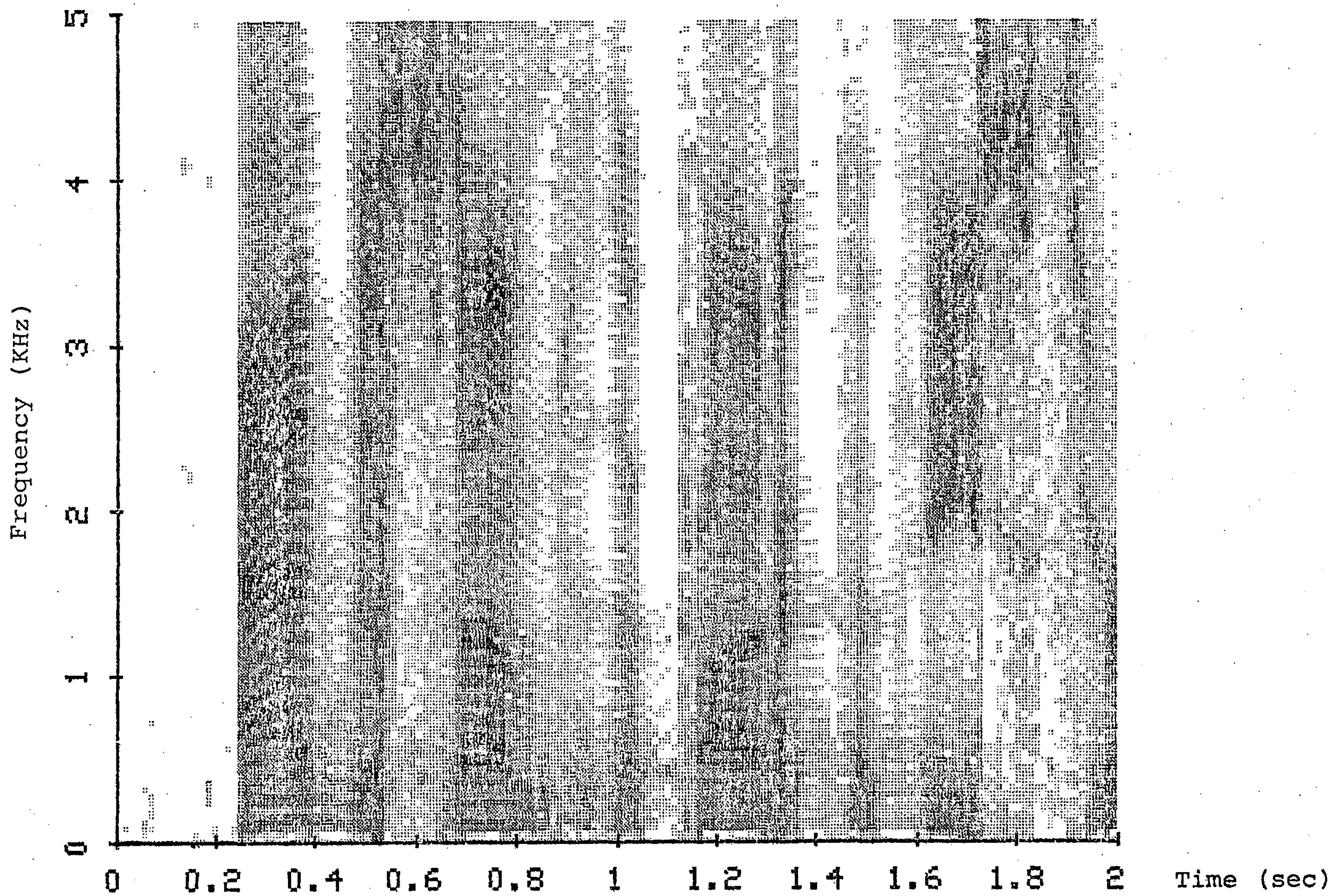


Figure 6.1.15b: Spectrogram corresponding to the first two seconds in Figure 6.1.15a.

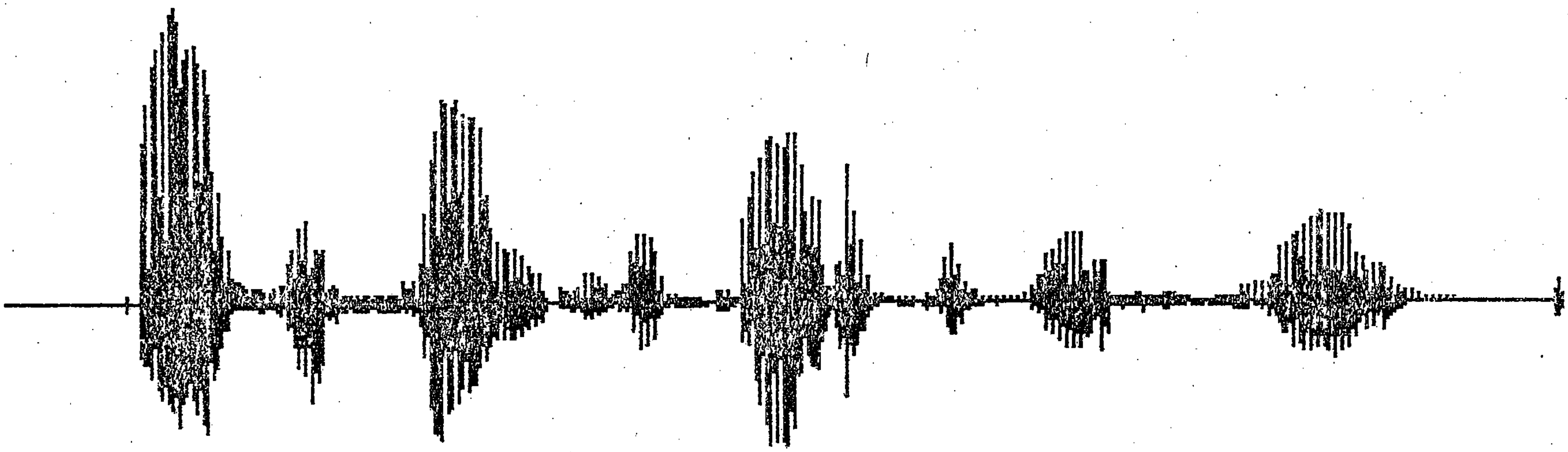


Figure 6.1.16a: The synthesized version of file M-1-3 with pre-emphasis, gain and pitch quantization only.

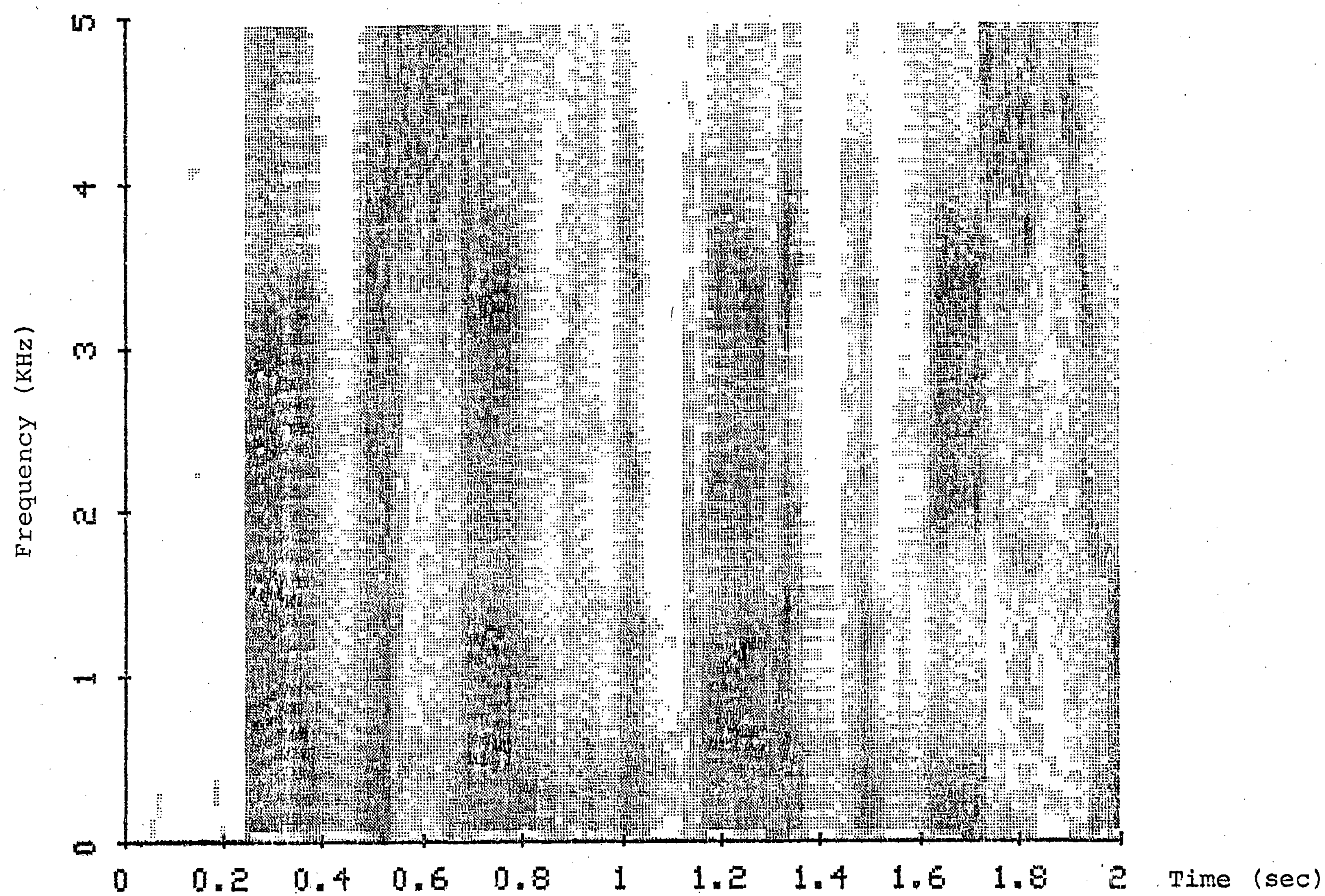


Figure 6.1.16b: Spectrogram corresponding to the first two seconds in Figure 6.1.16a.

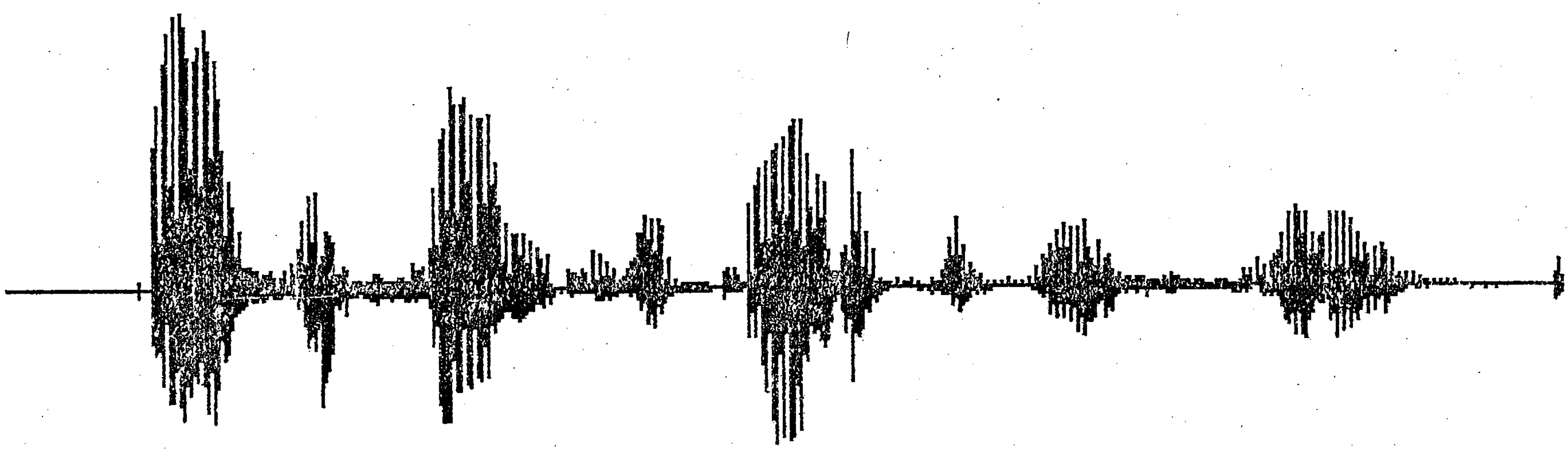
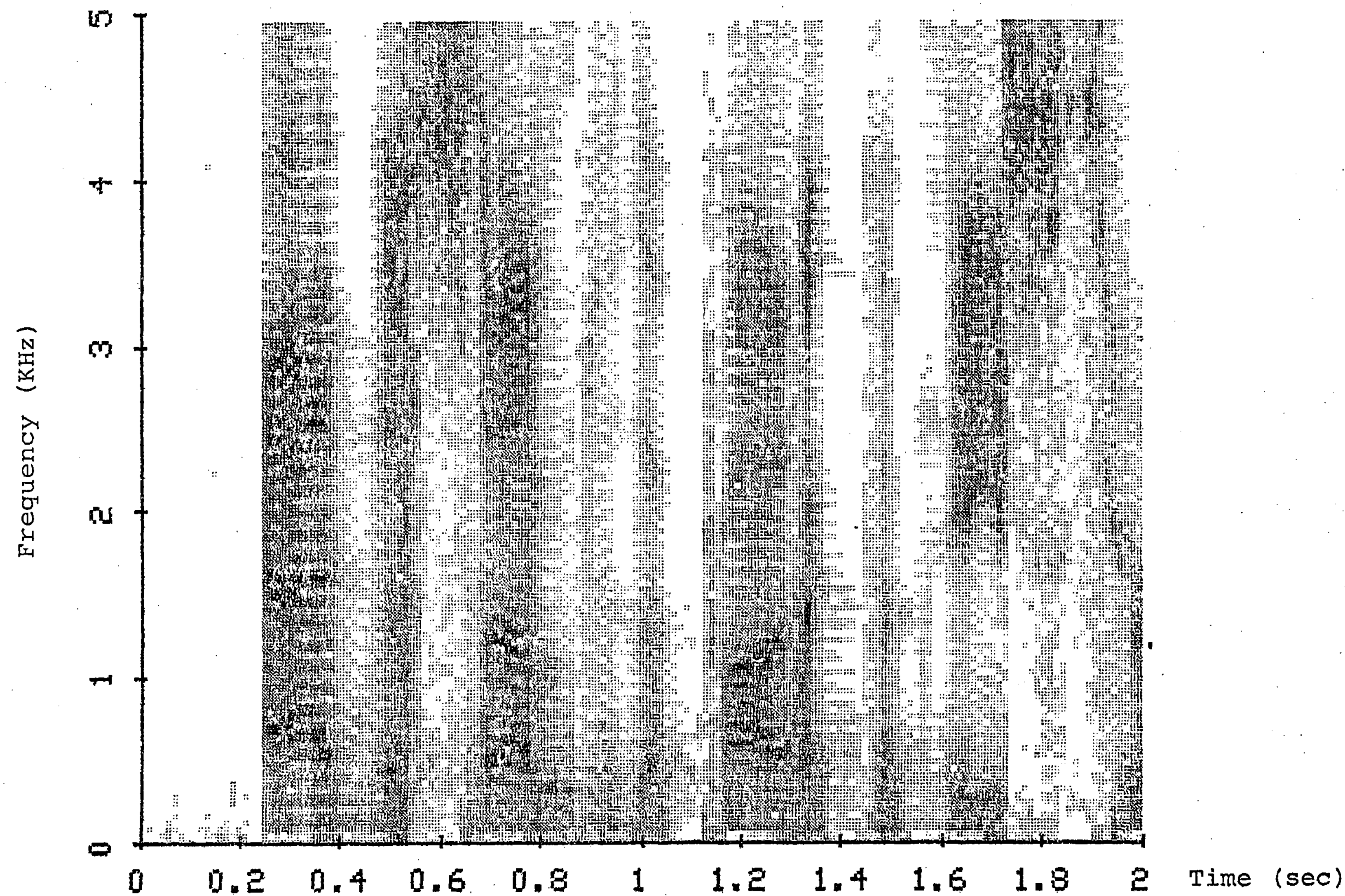


Figure 6.1.18a: Synthesized version of file M-1-3, with inverse sine quantization of the reflection coefficients but no pre-emphasis quantization.



<u>Figure 6.1.18b</u>: Spectrogram corresponding to the first two seconds in Figure 6.1.18a.

If inverse sine quantization is replaced in Figure 6.1.17 by min $E(\bar{D}_{tot})$ quantization of the k_i's, and $E(\bar{D}_{tot})$ is fixed at 3.5 dB, there results Figure 6.1.19a and b, representing quantized speech transmitted at a total bit rate of 2750 bits/sec. It was not possible to discern any difference in quality when compared to speech processed using inverse sine quantization.

Figure 6.1.20 then represents the sequence of steps followed in min $E(\bar{D}_{tot})$ quantization of the θ_i's. Figure 6.1.21a and 6.1.21b and, Figure 6.1.22a and 6.1.22b represent 1 file and 14 file statistics results respectively, using METHOD I. At $E(\bar{D}_{tot})$ = 3.5 dB, the total bit rate is 2539 bits/sec and 2674 bits/sec respectively. Finally, in the case of METHOD II on the θ_i's, Figure 6.1.23a and 6.1.23b and, Figure 6.1.24a and 6.1.24b are the results for pre-emphasis quantization but no pitch and gain quantization, and pitch and gain quantization, but no pre-emphasis quantization respectively. The total bit rate of the quantized parameters is in each case 2884, and 2934 bits/sec respectively, at $E(\bar{D}_{tot})$ = 3.5 dB. Again the only major difference when pitch and gain are quantized is the repression of the same peaks as discussed earlier. The quality of speech produced by min $E(\bar{D}_{tot})$ quantization on the orthogonal parameters is very comparable to that of reflection coefficient quantization. If one method happens to perform better than

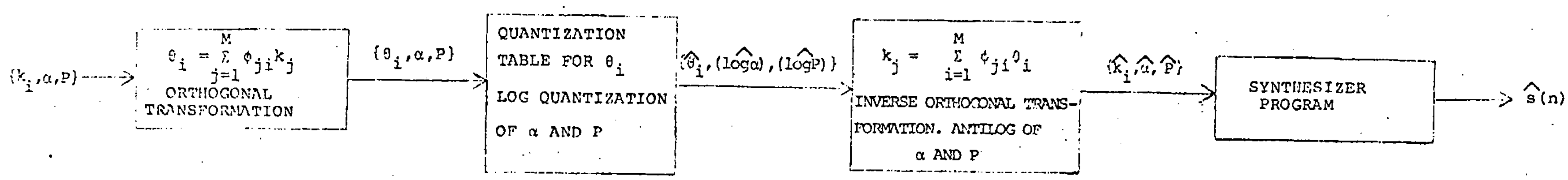
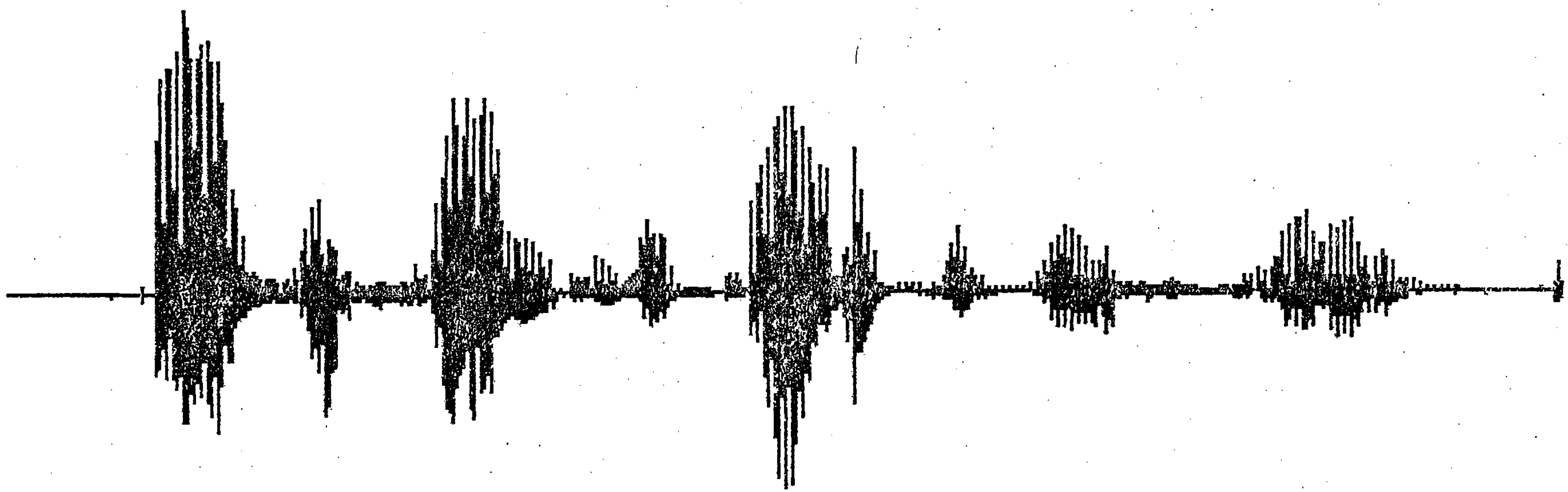


Figure 6.1.20: Vocoder Using Orthogonal Parameter Quantization



<u>Figure 6.1.19a</u>: Synthesized version of file M-1-3 with min $E(\bar{D})$ quantization of the reflection coefficients but no pre-emphasis quantization.

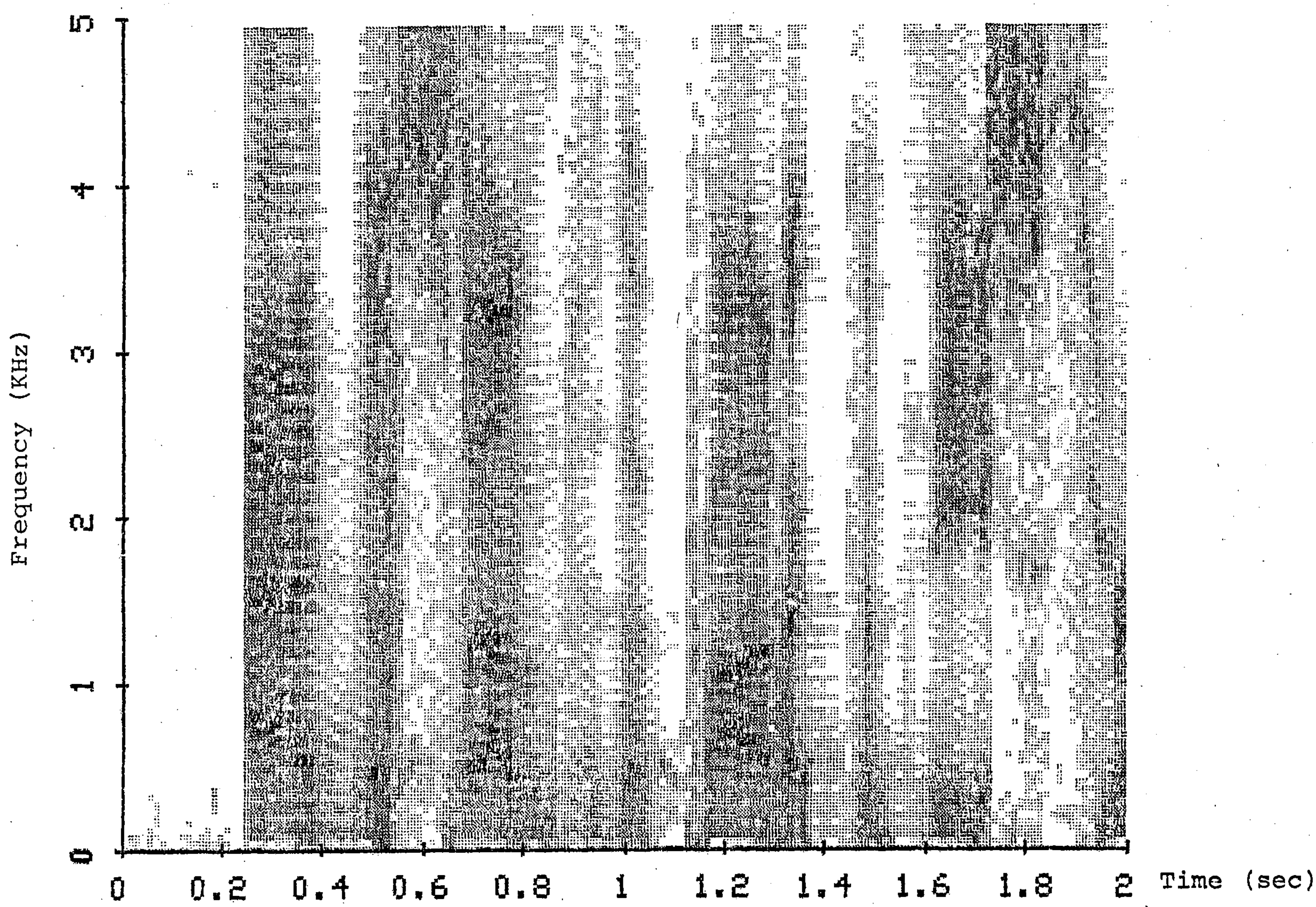


Figure 6.1.19b: Spectrogram corresponding to the first two seconds in Figure 6.1.19a.

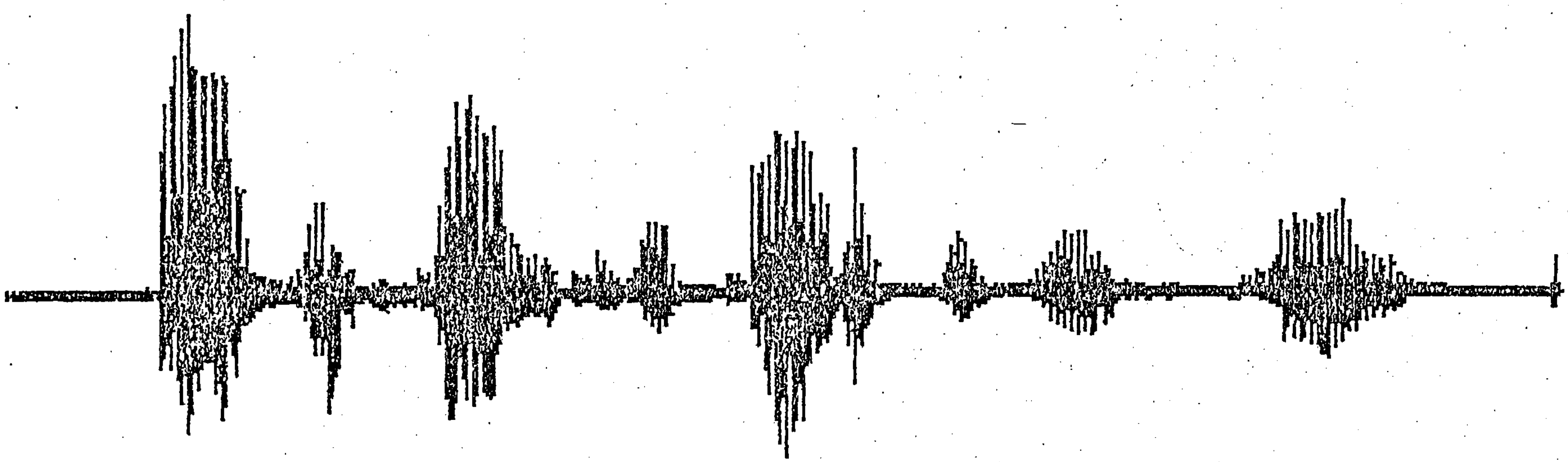


Figure 6.1.21a: The synthesized version of file M-1-3 using METHOD I with 1 file statistics but no pre-emphasis quantization.

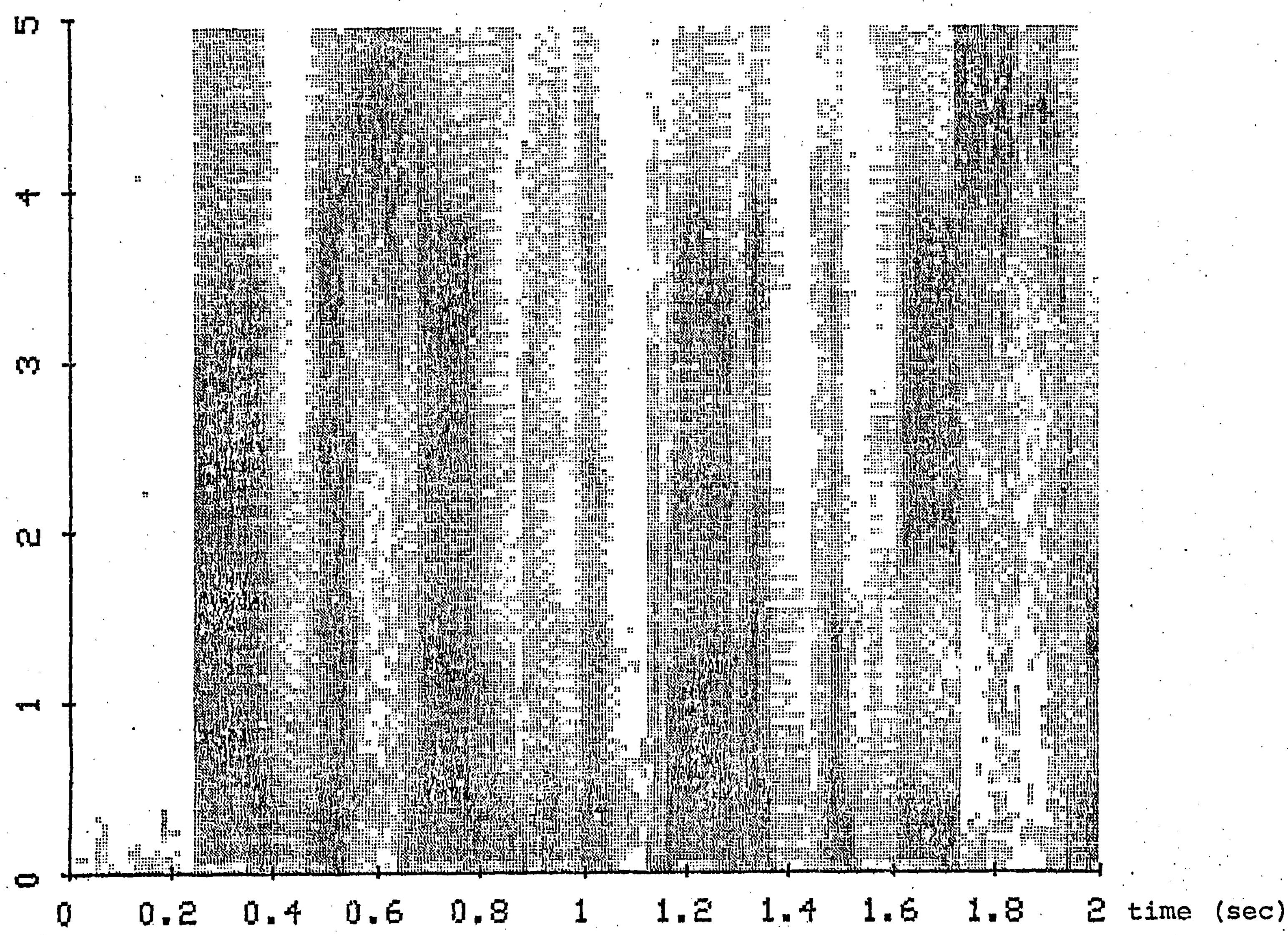


Figure 6.1.21b: Spectrogram corresponding to the first two seconds in Figure 6.1.21a

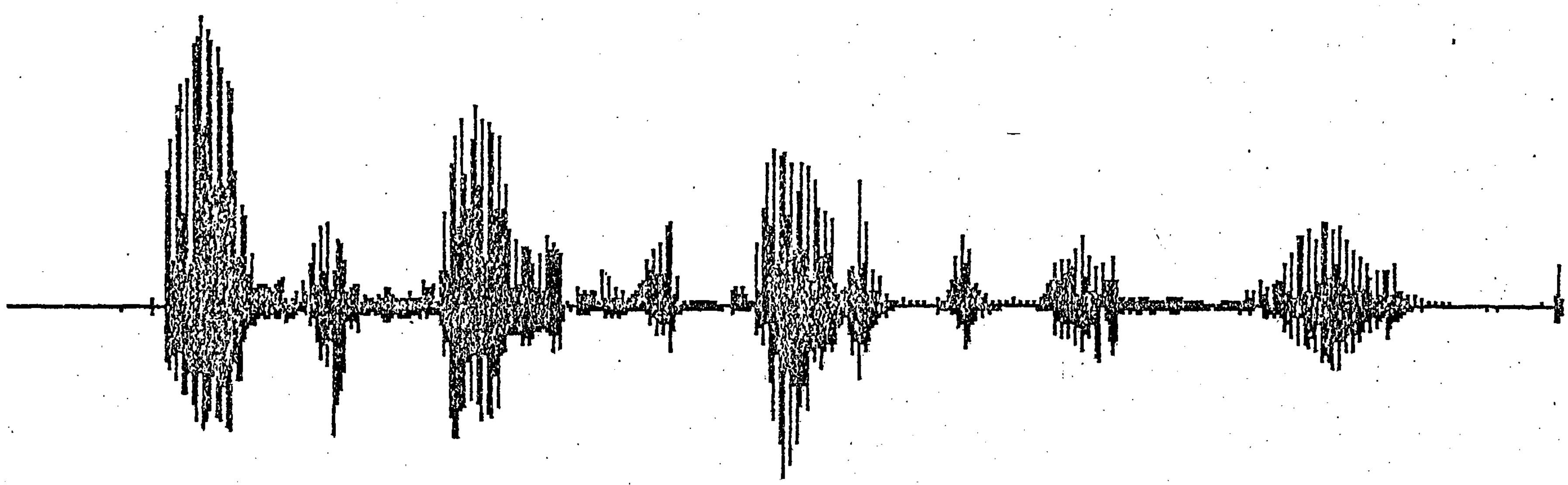


Figure 6.1.22a: The synthesized version of file M-1-3 using METHOD I with 14 file statistics but no pre-emphasis quantization.

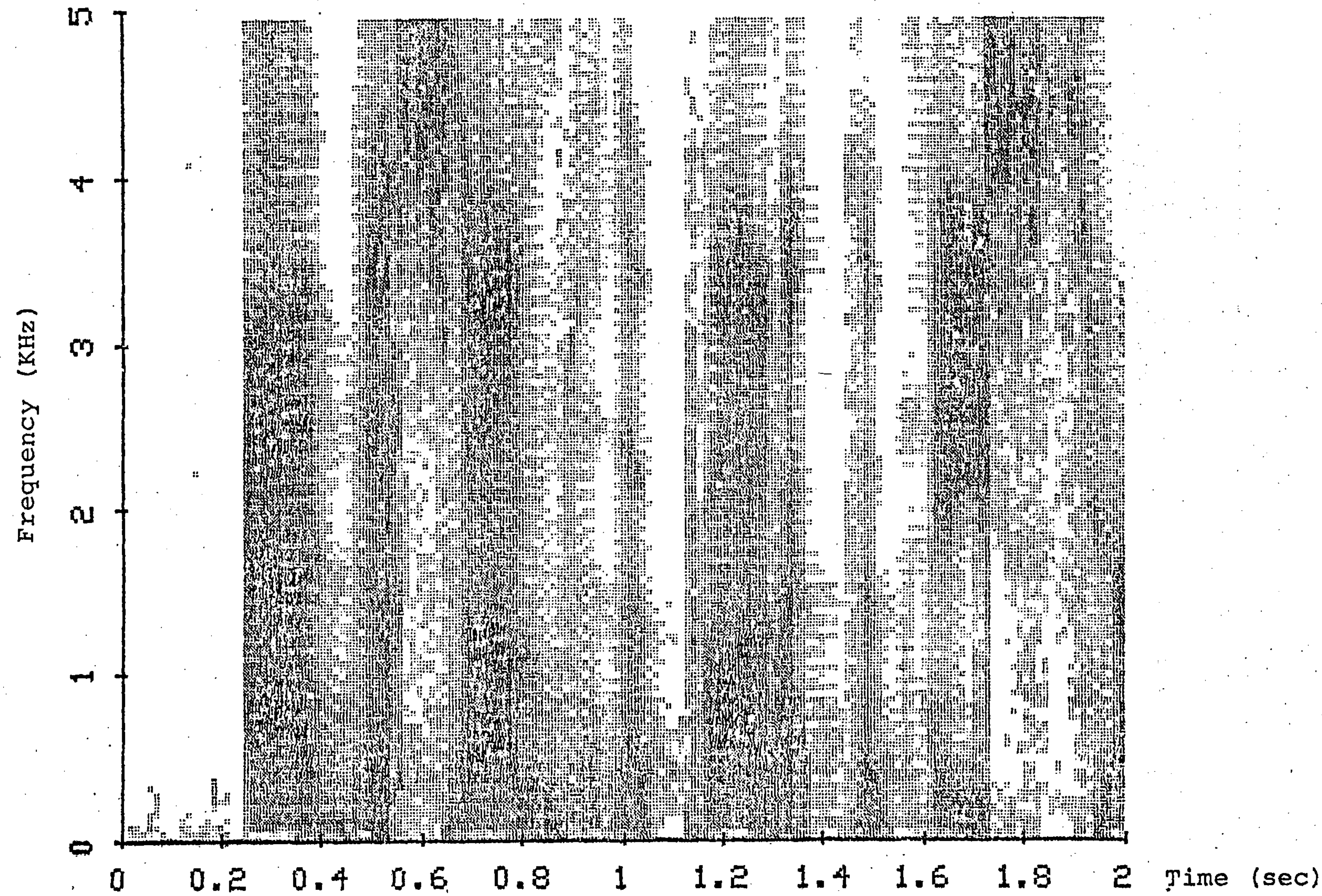


Figure 6.1.22b: Spectrogram corresponding to the first two seconds in Figure 6.1.22a.

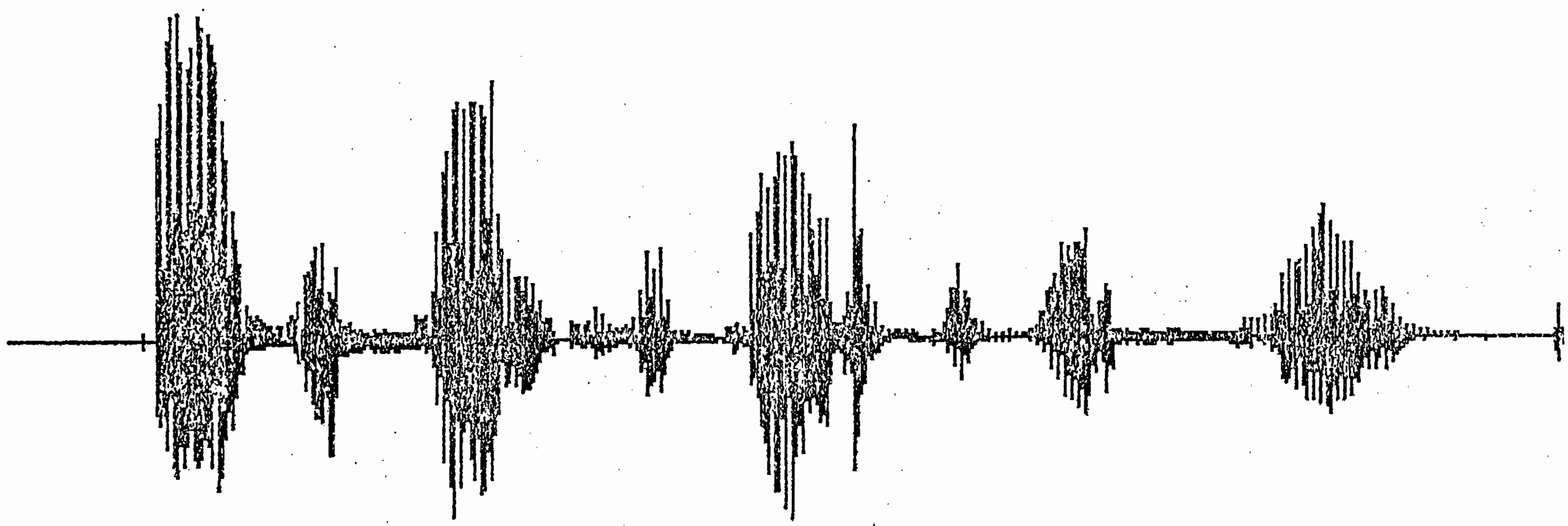


Figure 6.1.23a: Synthesized version of file M-1-3 using METHOD II with min $E(\bar{D})$ quantization of the orthogonal parameters but no pitch and gain quantization.

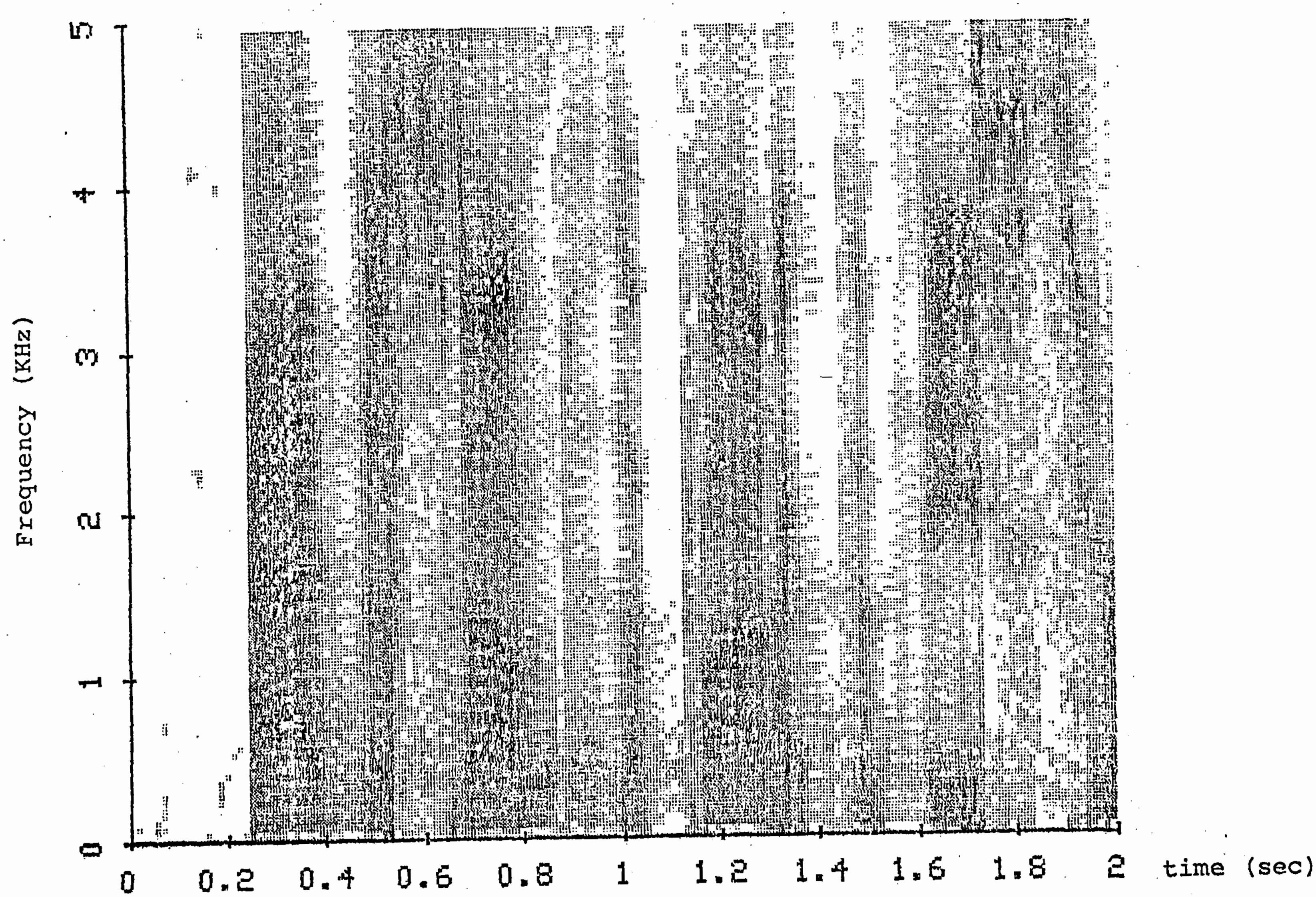


Figure 6.1.23b: Spectrogram corresponding to the first two seconds in Figure 6.1.23a.

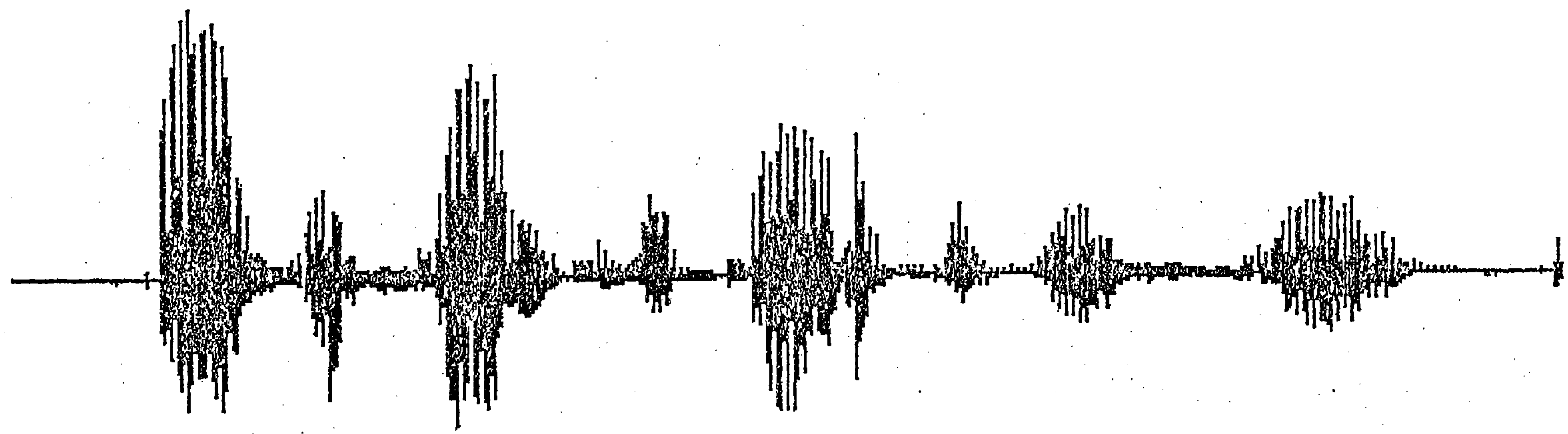
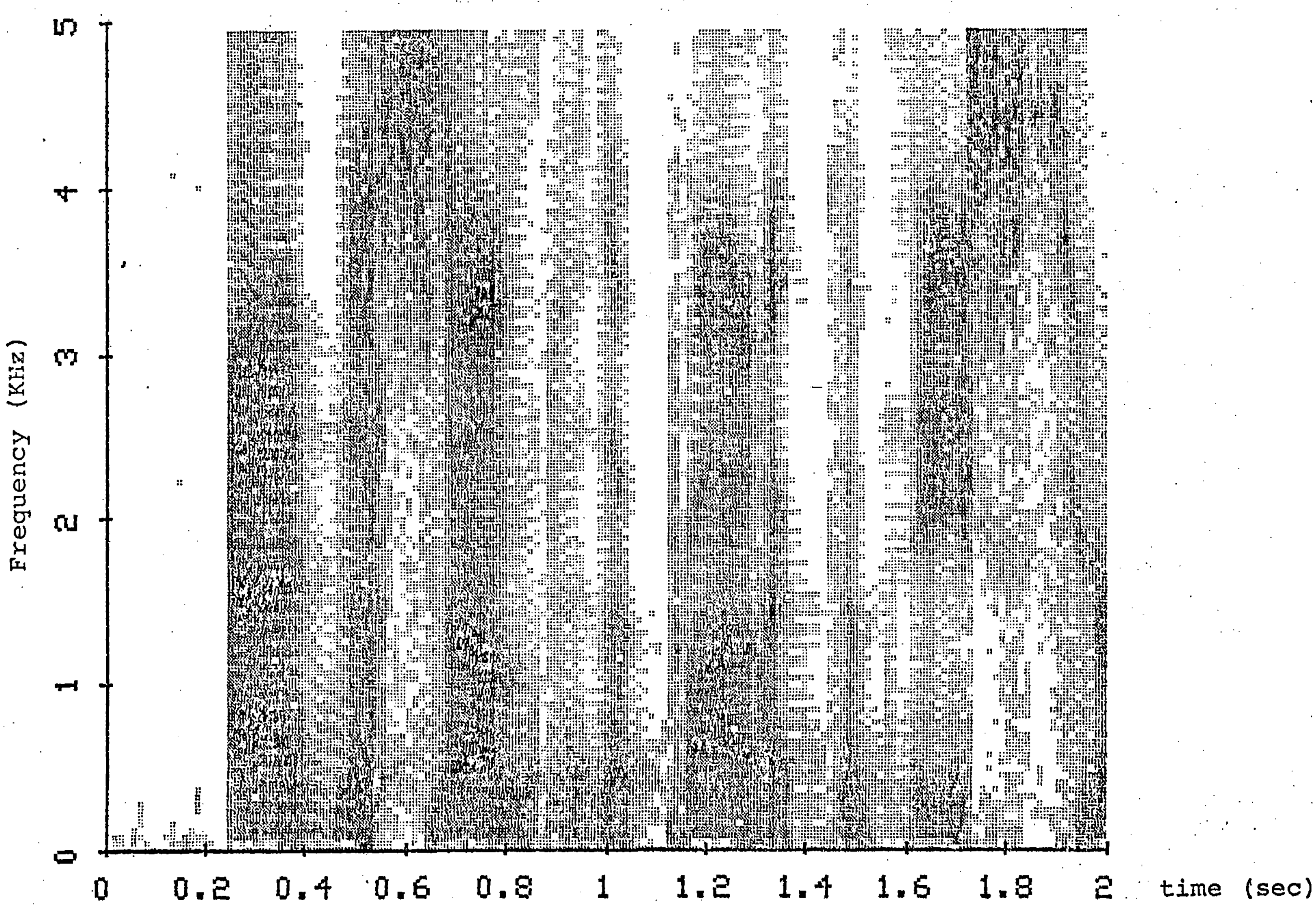


Figure 6.1.24a: Synthesized version of the file M-1-3 using METHOD II with min $E(\bar{D})$ quantization of the orthogonal parameters but no pre-emphasis quantization.



<u>Figure 6.1.24b:</u> Spectrogram corresponding to the first two seconds in Figure 6.1.24a.

another in some portion of the utterance, the other method will be found to produce speech of better quality in another segment. Now, the following experiment was also carried out. The error signal of file M-1-3 was used as input to the two-multiplier lattice synthesis structure. (The basic block diagram of the procedure is simply Figure 4.4.1). The error signal is obtained by passing the nonpre-emphasized and unwindowed version of the original file M-1-3 through the inverse filter A(z). The pre-emphasis factor and the reflection coefficients being already stored in a speech parameter file, it is only necessary to apply a step-up procedure on the k_i's in order to obtain the filter coefficients of the inverse filter A(z). However, the k_i's used in the synthesizer are those from the quantized reflection coefficient files. This experiment then permits a subjective comparison of processed speech files in which only the reflection coefficients are varied. The important degradation due to pitch extraction is therefore eliminated. Figures 6.1.25-6.1.27 represent synthesized speech in which inverse sine and min $E(\bar{D})$ quantization of the reflection coefficients, and min $E(\bar{D})$ quantization on the orthogonal parameters was applied, respectively. Subjectively speaking, all 3 files were almost indistinguishable from the original file M-1-3.

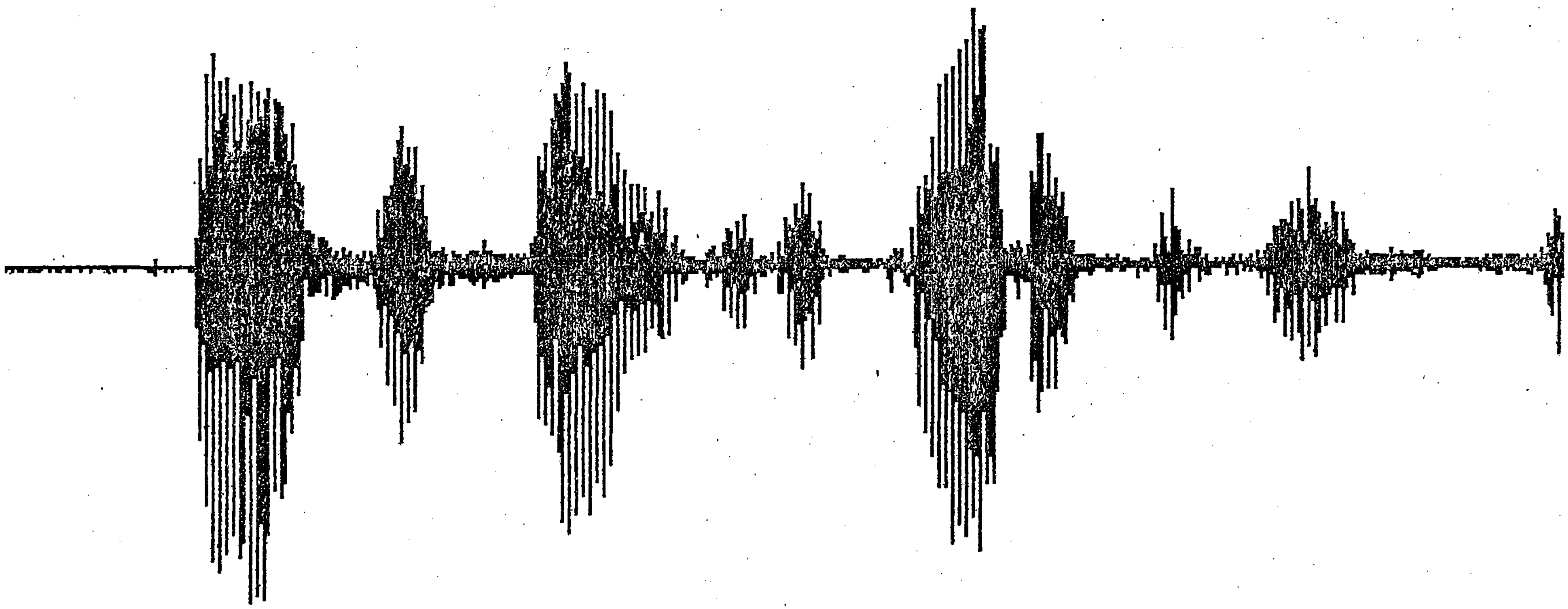


Figure 6.1.25: Synthesized version of file M-1-3 using inverse sine quantization of the reflection coefficients and the unquantized error signal as the driving function.

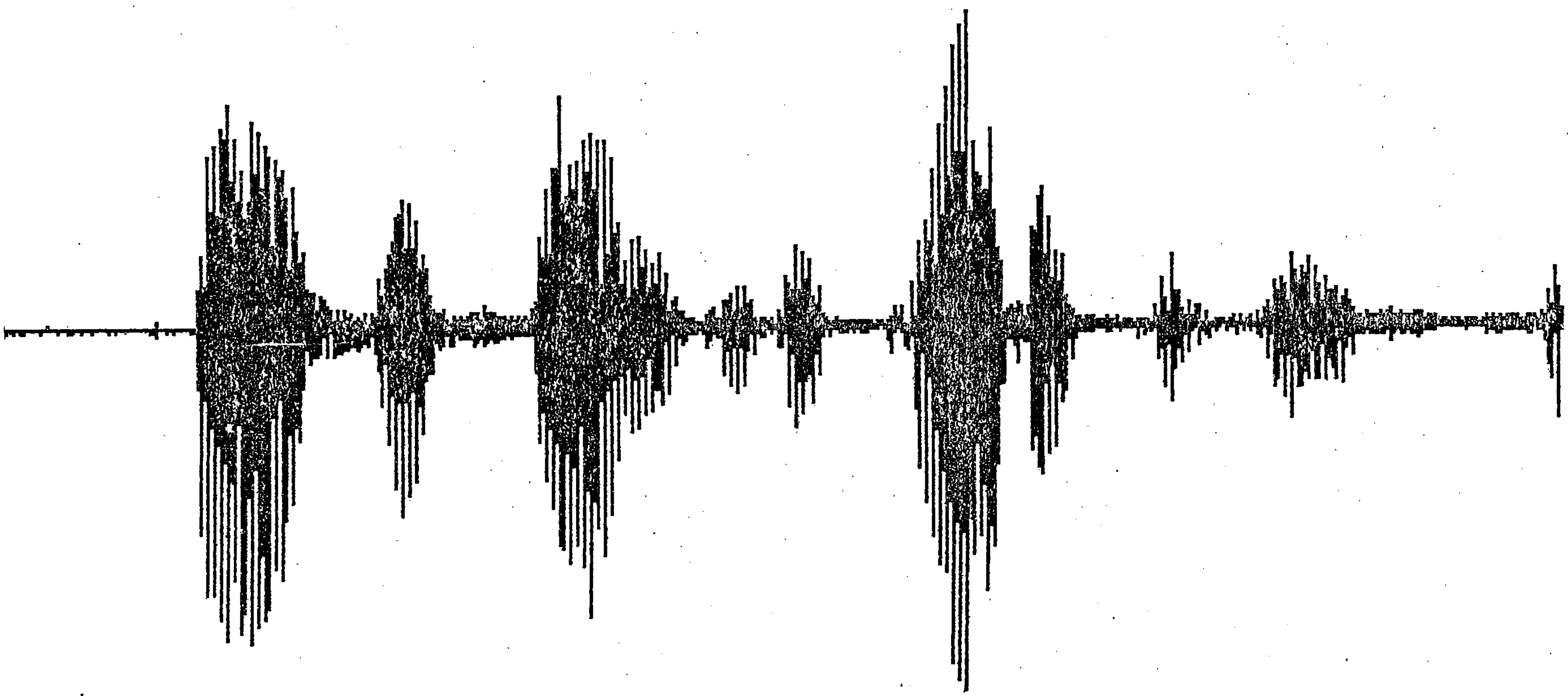


Figure 6.1.26: The synthesized version of file M-1-3 using min $E(\bar{D})$ quantization of the reflection coefficients and the unquantized error signal as the driving function.

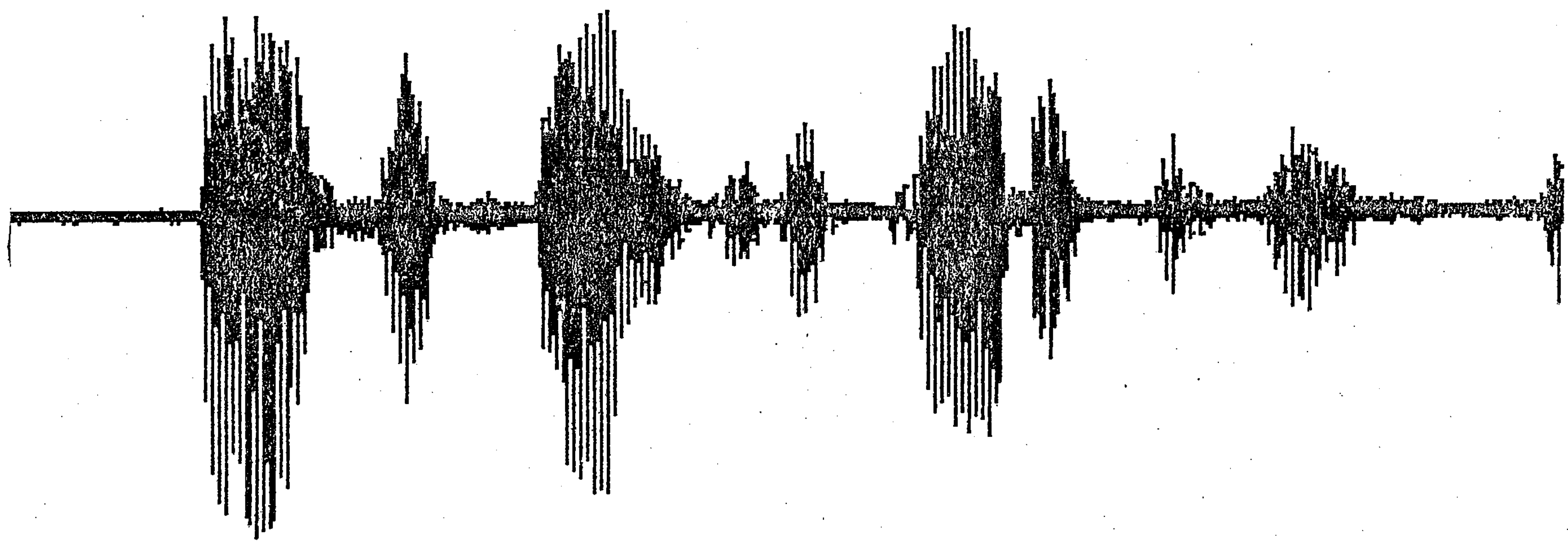


Figure 6.1.27: The synthesized version of file M-1-3 using min $E(\bar{D})$ quantization of the orthogonal parameter and the unquantized error signal as the driving function.

However when the original utterances were processed, it was found that, on the average, inverse sine quantization produces speech of quality, close to that of the original, and better than that using min $E(\bar{D}_{tot})$ quantization on the θ_i's, while min $\bar{E}(D_{tot})$ quantization on the k_i's results in the most discernable degradation. It must be emphasized that for this synthesizer with the error signal as the driving function, 14 file statistics were used on all files including M-1-3. File M-1-3 performs better than other files and this was first thought to be due to the fact that its statistics are similar to the statistics obtained using 14 files. For example, file M-1-4, whose performance is the worst, has statistics less comparable with the 14 file statistics (see Table 6.1.7). However, tests using METHOD II with its statistics instead of the usual 14 file statistics seem to indicate that the statistics are not the major reason for the poor performance since the latter does not improve at all under 1 file statistics.

DETERMINATION OF EIGENVALUES BY JACOBI METHOD, WITH

N = 14
ITMAX = 100
EPS1 = 0.10000E-09

EPS2 = 0.10000E-09
EPS3 = 0.10000E-04

Table 6.1.7: 1 File Statistics of M-1-4.

THE STARTING MATRIX A(1,1)...A(N,N) IS

0.17941	-0.05904	-0.07869	-0.08671	0.01398	-0.02251	0.01877	0.01674	-0.00191	-0.01862	-0.02463	-0.00218	-0.01152	-0.00403
0.00000	0.16921	0.04465	0.02627	-0.06794	0.00400	-0.01713	0.00961	0.02265	0.02298	0.00603	0.00145	0.00912	-0.00299
0.00000	0.00000	0.09034	0.04425	-0.01306	-0.01669	-0.02464	-0.03759	-0.00503	-0.00041	0.01141	-0.00218	0.01317	-0.00061
0.00000	0.00000	0.00000	0.08596	-0.01230	0.02697	-0.02579	-0.01122	-0.00262	0.02075	0.03092	0.00638	0.00619	0.00417
0.00000	0.00000	0.00000	0.00000	0.08863	-0.01767	0.00367	-0.01983	-0.01289	-0.02891	-0.00859	-0.00335	0.00053	-0.00564
0.00000	0.00000	0.00000	0.00000	0.00000	0.05917	0.00305	0.01659	0.00917	0.01660	0.00936	0.00063	-0.00571	0.00262
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03722	0.00786	-0.00436	-0.00853	-0.01183	-0.00272	-0.00965	0.00134
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04592	0.01845	0.01556	-0.01165	0.00159	-0.00709	0.00083
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03537	0.00943	-0.00869	-0.00982	-0.00387	0.00131
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04083	0.00531	-0.00266	-0.00378	-0.00075
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04148	0.00527	0.00148	-0.00111
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02067	0.00285	-0.00120
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01269	-0.00160
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01194

ITER = 1 SIGMA1 = 0.96597E-01 SIGMA2 = 0.17457

ITER = 2 SIGMA1 = 0.17457 SIGMA2 = 0.17762

ITER = 3 SIGMA1 = 0.17762 SIGMA2 = 0.17777

CONVERGENCE HAS OCCURRED WITH

ITER = 4 SIGMA2 = 0.17777 S = 0.17777 N = 14

EIGENVALUES EIGEN(1)...EIGEN(N) ARE

0.34949 0.18015 0.00973 0.06476 0.04408 0.11587 0.01167 0.01603 0.04286 0.03009 0.02154 0.02377 0.00417 0.00668

THE FINAL TRANSFORMED MATRIX A IS

0.34949	-0.00000	0.00000	-0.00000	0.00000	-0.00000	0.00000	0.00000	0.00000	0.00000	-0.00000	-0.00000	-0.00000	-0.00000
0.00000	0.18015	-0.00000	0.00000	-0.00000	0.00000	-0.00000	-0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00973	-0.00000	0.00000	-0.00000	0.00000	0.00000	-0.00000	0.00000	-0.00000	-0.00000	-0.00000	-0.00000
0.00000	0.00000	0.00000	0.06476	-0.00000	0.00000	-0.00000	-0.00000	0.00000	-0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.04408	-0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000	-0.00000	-0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.11587	0.00000	0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000	-0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01167	-0.00000	0.00000	-0.00000	0.00000	0.00000	-0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01603	0.00000	-0.00000	0.00000	-0.00000	-0.00000	-0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.04286	0.00000	-0.00000	-0.00000	-0.00000	-0.00000
0.00000	[illegible]	[illegible]	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03009	-0.00000	-0.00000	-0.00000	-0.00000
0.00000	[illegible]	[illegible]	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02154	0.00000	0.00000	0.00000
0.00000	[illegible]	[illegible]	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.02377	0.00000	-0.00000
0.00000	[illegible]	[illegible]	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00417	0.00000
0.00000	[illegible]	[illegible]	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00668

EIGENVECTORS ARE IN CORRESPONDING COLUMNS OFTHE FOLLOWING T MATRIX

0.61944	0.33519	[illegible]	0.49430	0.20677	-0.19346	0.12907	0.20727	0.07305	-0.23395	-0.17924	0.08903	0.12682	-0.00705
-0.47[illegible]	[illegible]	[illegible]	-0.12719	0.33487	-0.37149	-0.06862	0.07572	-0.35015	-0.07350	-0.13502	-0.06454	0.02056	0.05350
-0.34975	[illegible]	[illegible]	[illegible]	-0.22587	-0.40872	0.40585	0.00874	0.32670	-0.24253	-0.10613	0.14172	0.35558	0.27997
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.30994	0.06053	0.60909	0.03296	-0.00993	-0.17555	0.15111	-0.19846	-0.25286
[illegible]	[illegible]	[illegible]	[illegible]	0.61473	-0.09746	0.06822	0.05167	-0.14826	-0.01579	-0.20212	-0.12742	0.26543	0.13163
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.57496	0.03039	-0.34378	-0.19235	-0.58963	-0.18090	0.26899	0.13425	0.15374
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.07167	0.28397	0.35975	-0.37347	-0.14903	-0.15973	-0.25110	0.26491	-0.32541
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.31900	0.25340	0.14717	0.11439	0.41143	0.17554	0.17634	0.22065	0.37645
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.08676	0.23594	0.08702	0.46574	-0.27193	0.46550	-0.04230	0.08950	-0.36760
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.26947	-0.02173	-0.15596	0.36560	0.29358	-0.53369	-0.44061	0.29188	-0.07923
[illegible]	[illegible]	[illegible]	[illegible]	0.11807	0.13096	0.12654	-0.07622	-0.31895	-0.01122	0.50203	-0.50832	0.26594	0.09278
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.02594	0.33459	-0.04113	-0.31182	0.40639	0.02389	0.46960	0.18052	-0.09612
[illegible]	[illegible]	[illegible]	[illegible]	0.03239	-0.11717	-0.25715	-0.29816	-0.01179	0.10603	0.04072	0.26877	0.45117	-0.58917
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.05556	-0.53998	0.42427	0.05501	-0.06995	0.14793	0.12320	0.45854	0.24453

CHAPTER VII: CONCLUSION

Using the $E(\bar{D}_{tot})$ fidelity criterion, it has therefore been verified that asymptotic min $E(\bar{D}_{tot})$ quantization of the k_i's results in a slightly lower bit rate than inverse sine quantization, as is expected from the results of [12]. Next decorrelation of the k_i's results in a total bit rate which is also lower than that using inverse sine quantization but unfortunately, is higher than that using min $E(\bar{D}_{tot})$ quantization on the k_i's. Recall from page 138, that the difference $\Sigma\sqrt{\text{Var } k_i} - \Sigma\sqrt{\lambda_i}$ is not substantial for either 1 file or 14 file statistics. As can be seen from Table 6.1.3 and Table 6.1.4, this is because the cross-correlation in the original covariance matrices is not pronounced. Now as was already mentioned under equal area quantization (Chapter V) a great percentage of speech consists of silence and unvoiced intervals. Also, from page 102, section 5.3, it is stated that the frame to frame dependence of the k_i's is felt to be even more significant than the above cross-correlation within a frame. Afterall, the variable frame rate approaches of Makhoul (section 4.6) and Seneff (section 5.3) and the DPCM approach of Sambur all result in an average bit rate of about 1500 bits/sec. Hence, if decorrelation is to be performed, it should be followed by variable frame rate transmission and/or DPCM on the orthogonal parameters themselves.

As was shown in [18] this can further reduce the bit rate in DPCM by about 500 bits/sec.

Notice that if the spectral deviation D is an adequate representation of the hearing mechanism, then as discussed previously a value of D in the range 3 to 4dB is required if a difference is to be perceived. As the gain quantization is done independently, the distance measure D depends only on the k_i's. As the degradation due to the use of pitch in the construction of the driving function to the synthesizer masks the differences in quality among the 3 reflection coefficient quantization methods studied, it was decided in the end to use the error signal as driving function to the synthesizer. In Chapter V two fidelity criteria were introduced: the maximum spectral deviation bound, max $(\bar{D}_{tot})$ and the expected spectral deviation bound, $E(\bar{D}_{tot})$. The $E(\bar{D}_{tot})$ criterion was then chosen for study. It is then found that min $E(\bar{D}_{tot})$ quantization on the θ_i's results in speech quality slightly superior to that using min $E(\bar{D}_{tot})$ quantization on the k_i's. However the performance under these two methods is noticeably worse than that under inverse sine quantization on the k_i's. In fact, the latter method results in speech quality fairly close to that of the original utterance. But, from Chapter V, it is observed that inverse sine quantization does not minimize the $E(\bar{D}_{tot})$ criteria, but instead, minimizes the max $(\bar{D}_{tot})$ criteria. The fact that under the $E(\bar{D}_{tot})$ criterion inverse sine quantization is subjectively a better scheme than min $E(\bar{D}_{tot})$ quantization, seems to suggest that, as far as the minimization of criteria is concerned, the max $(\bar{D}_{tot})$ criterion is a better approximation to some aspect of the hearing mechanism than the $E(\bar{D}_{tot})$ criterion.

For this error signal synthesis, the degradation in quality (which on the average is especially apparent when using min $E(\bar{D}_{tot})$ quantization on the k_i's, shows itself in the introduction of discontinuous dips and peaks fairly well distributed throughout the whole speech file (see Figures 6.1.28-6.1.31). However, the difference in quality between the original utterance and the unquantized linear prediction synthesized utterance, is even greater. The reason for this was discussed before: linear prediction is only an incomplete description of the speech production mechanism and among other things, the actual pitch values for each frame are not necessarily extracted. It is possible that these errors are larger than those resulting from quantization (as is the case here). The natural quality of the speech is also degraded because of the difficulty in reproducing speech when dealing with nasal and fricative sounds and, fast transitions from one class of sounds to another. Additional problems also arise because of the use of a fixed frame analysis.

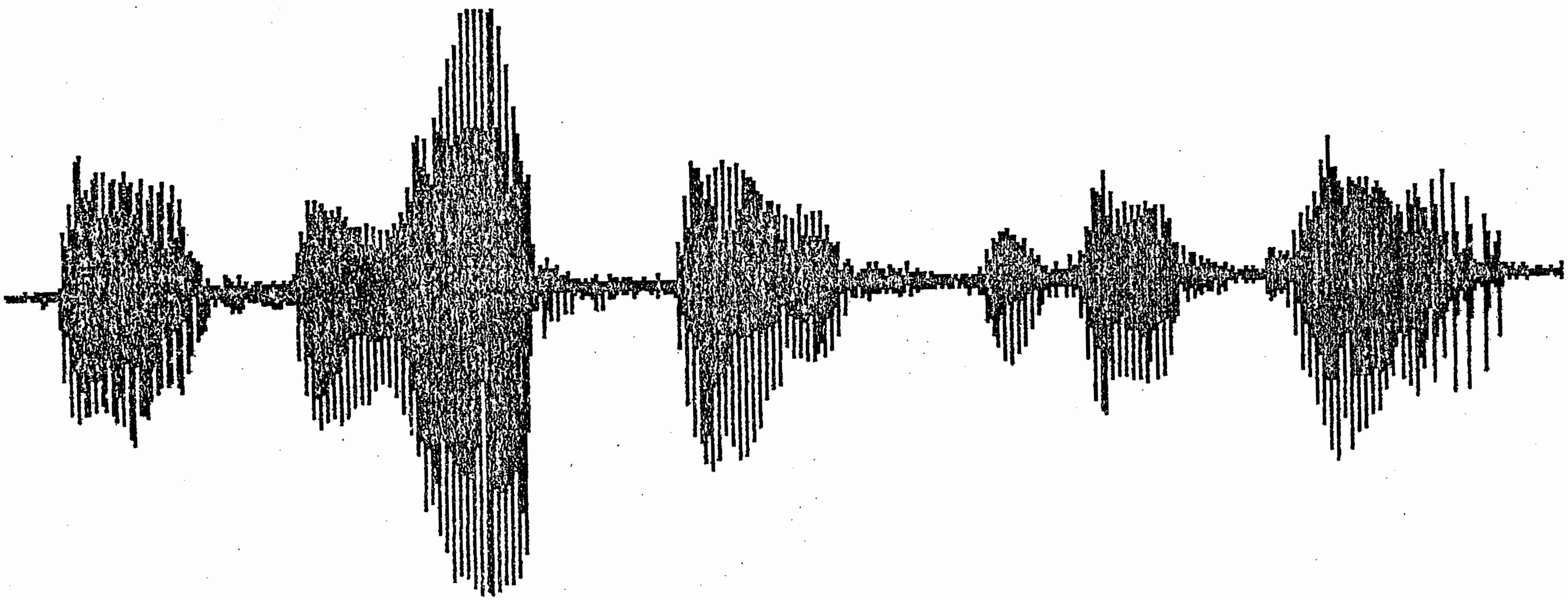


Figure 6.1.28: The original utterance 'thieves who rob friends deserve jail' (also denoted by file M-1-4).

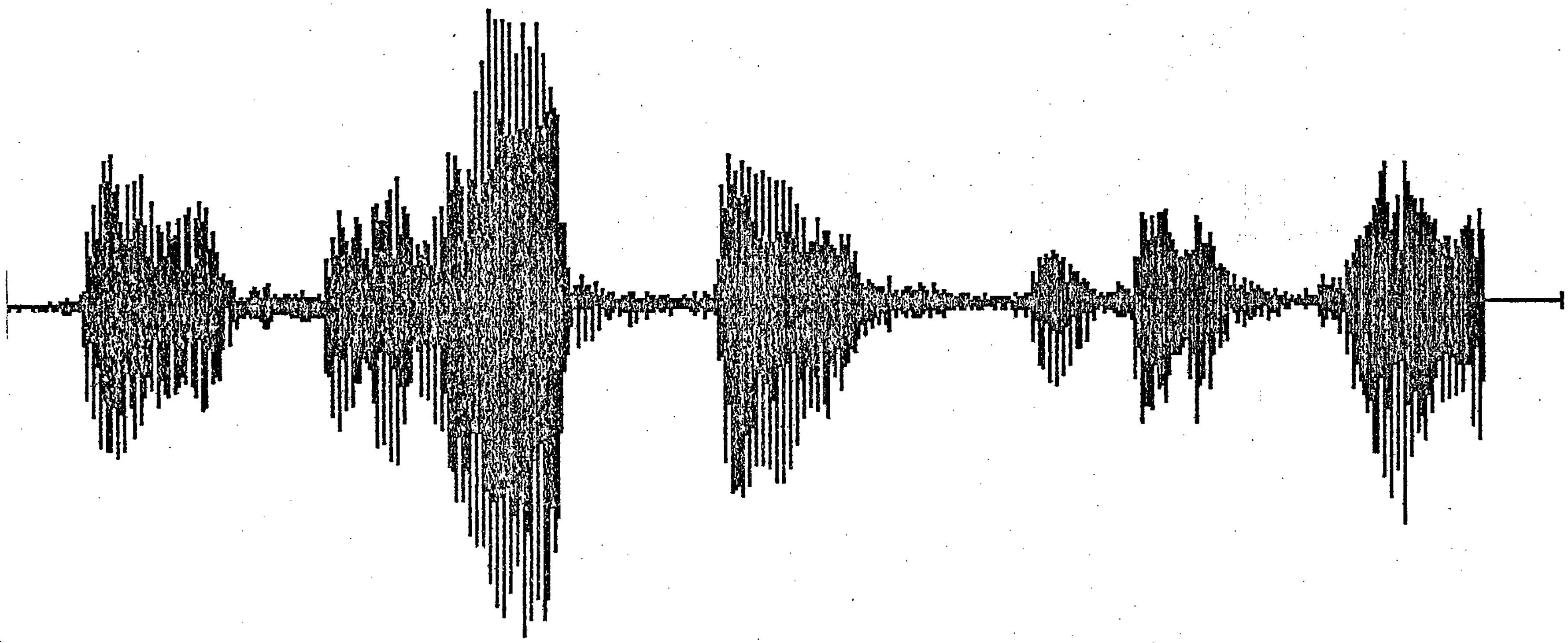


Figure 6.1.29: Synthesized version of file M-1-4 using inverse sine quantization of the reflection coefficients and the unquantized error signal as the driving function (1 file statistics).

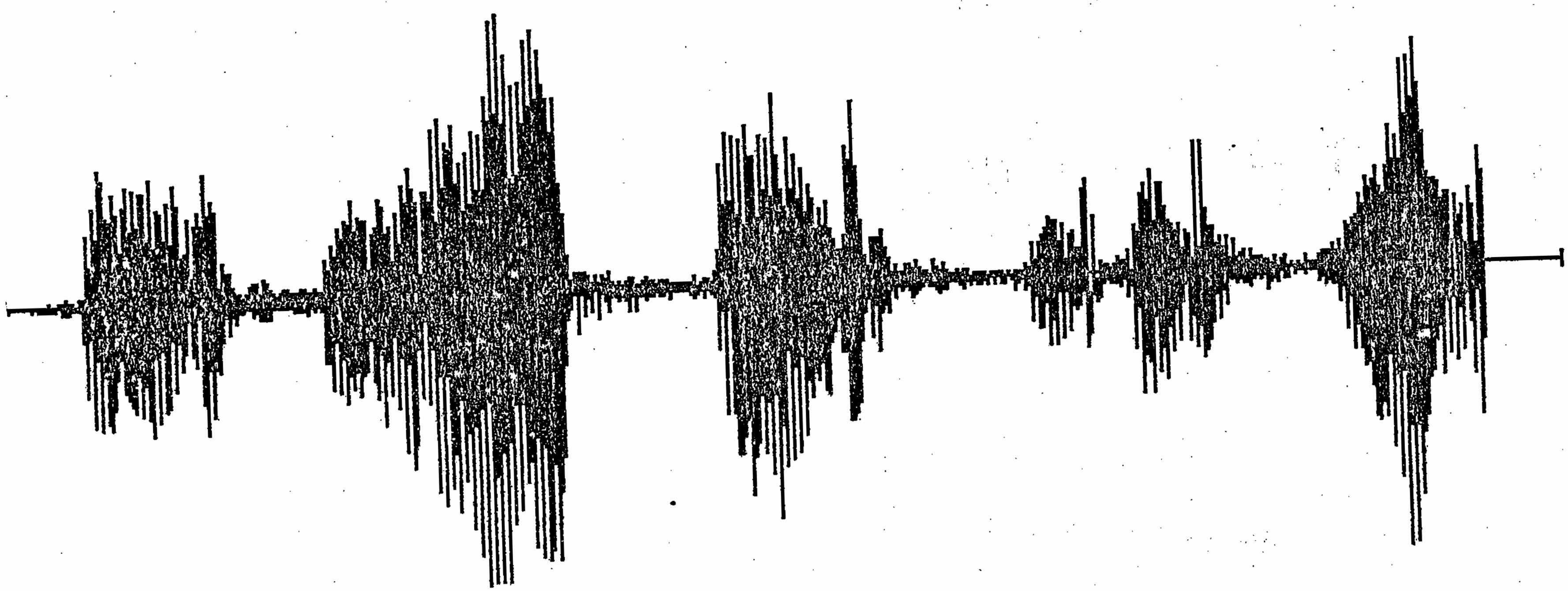


Figure 6.1.30: The synthesized version of file M-1-4 using min $E(\bar{D})$ quantization of the reflection coefficients and the unquantized error signal as the driving function (1 file statistics).

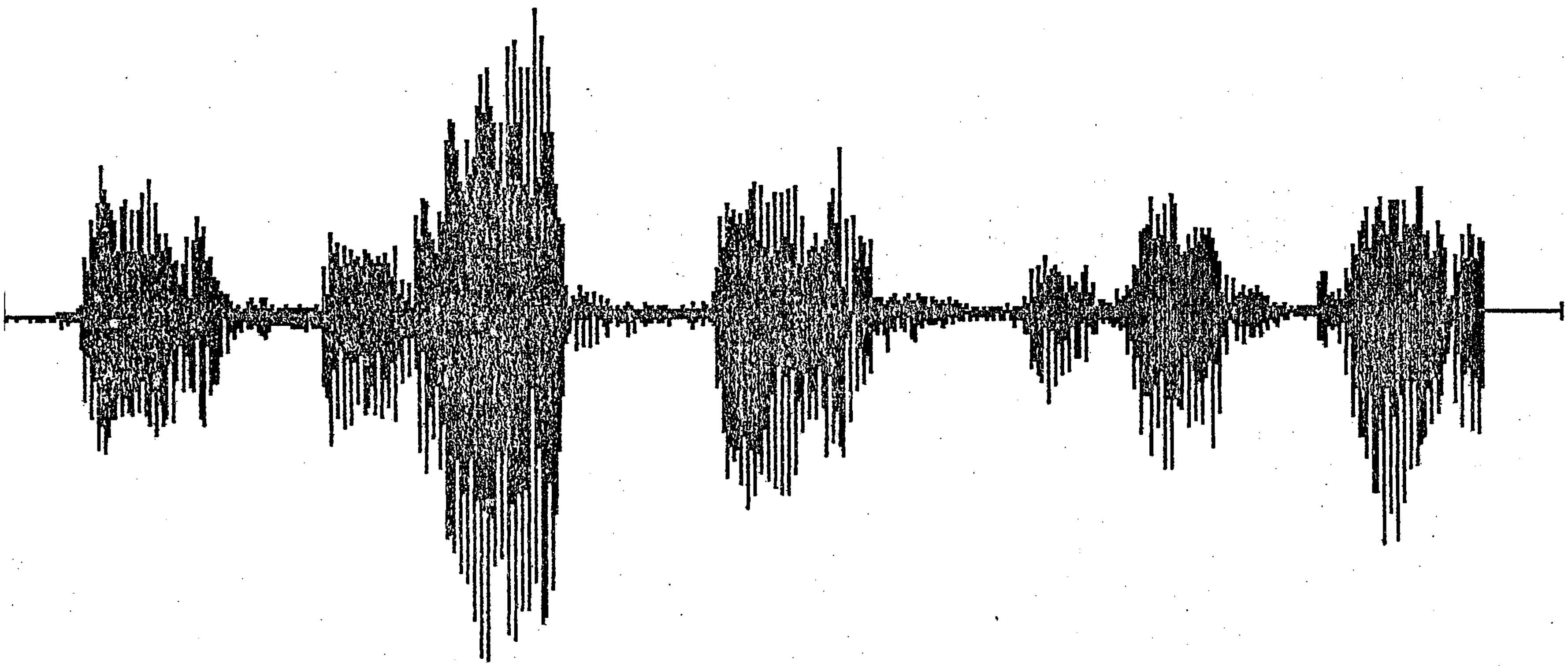


Figure 6.1.31: The synthesized version of file M-1-4 using min $E(\bar{D})$ quantization of the orthogonal parameter and the unquantized error signal as the driving function (1 file statistics).

If the actual hearing mechanism was understood, then which parameters should be extracted from the speech waveform and how they should be quantized would then be known. Only further basic research into speech production and hearing mechanisms and the construction of efficient algorithms will permit the reduction of the total bit rate by a great factor and at no price in speech quality.

Appendix A

It is required to show that $u(x) = \dfrac{\hat{s}_X(x)}{\int_a^b s_X(\lambda)\,d\lambda}$

minimizes $\max\limits_{a\le x\le b} \bar{D}(x,q(x))$

Proof: (from [15]) Transform coordinates to $z = U(x)$.
Then, using (5.1.17)

$$\bar{D}(x,q(x)) = \left|\int_x^{q(x)} s_X(\lambda)\,d\lambda\right| = \int_{U(x)}^{U(q(x))} s_Z(z)\,dz = \int_{U(x)}^{U(q(x))} \frac{s_X(x)\,dz}{u(x)}$$

Hence if $u(x) = \dfrac{s_X(x)}{\int_a^b s_X(\lambda)\,d\lambda}$ then $s_Z(z)$ is a constant

sensitivity measure. The problem reduces to proving that a constant $s_Z(z)$ minimizes $\max \bar{D}$ iff z is uniformly quantized.

Necessary condition: let $s_Z(z) = C$, a constant.
Then if z is uniformly quantized into N levels, $\max \bar{D} = \frac{C}{2N}$.
However, if it is not uniformly quantized $\max \bar{D} > \frac{C}{2N}$.
Consequently, if $s_Z(z) = C$, then uniform quantization of z is required.
Sufficient condition: let z be uniformly quantized.
Then if $s_Z(z)$ is not constant it is obvious that non uniform quantization of z will decrease $\max \bar{D}$. Consequently if uniform

quantization of z is to be optimal, $s_Z(z)$ must be constant.

Next, it is shown that the same choice of u(x) also minimizes the entropy H for fixed $E(\bar{D})$ in the asymptotic limit of large N.

Proof: (from [12])

Substituting (5.1.19) in (5.1.20) yields

$$H = -\log 4E(\bar{D}) + \log E \frac{s_X(x)}{u(x)} - E \log \frac{\bar{p}_X(x)}{u(x)} .$$

Recall from Chapter V, that the integral of u(x) over (a,b) is normalized to 1. Now, using the following inequality (stated in [12])

$$\log E \frac{s_X(x)}{u(x)} \geq E \log \frac{s_X(x)}{u(x)}$$

satisfied with equality iff

$$u(x) = Cs_X(x) = \frac{s_X(x)}{\int_a^b s_X(\lambda)\, d\lambda}$$

yields

$$H \geq -\log 4E(\bar{D}) + E \log \frac{s_X(x)}{p_X(x)}$$

the lower bound being attained by the above choice of u(x).

Appendix B

To show that $$\psi \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{A'(e^{j\theta})}{A(e^{j\theta})} \right|^2 d\theta \qquad \text{(B-1)}$$

is equivalent to $\frac{(A',A')}{(A,A)}$, where A is the linear prediction analysis filter.

Proof: (from [11]) $|A'|^2$ is the inverse Fourier transform of the autocorrelation $r_a'(n)$ of the sequence $\{a_i'\}$. But $a_i = 0$ for $i \not\subset (0,M)$ which implies that $r_a'(n)$ is zero for $|n|>M$. Let the autocorrelation of $\alpha/|A|^2$ be $\rho(n)$. (B-1) can then be written as

$$\psi = \frac{1}{\alpha} \sum_{n=-M}^{M} r_a'(n)\rho(n).$$

But by the correlation matching of section 2.1, $\rho(n) = r(n)$ for $|n|\leq M$. Substituting this in the above summation, (B-1) is seen to be equivalent to

$$\psi = \frac{1}{\alpha} \int_{-\pi}^{\pi} |A'(e^{j\theta})|^2 |S(e^{j\theta})|^2 \frac{d\theta}{2\pi}$$

Let $E' = A'S$. Then by Parseval's theorem,

$$\psi = \frac{1}{\alpha} \sum_{n=-\infty}^{\infty} [e'(n)]^2$$

$$= \frac{1}{\alpha}(A',A') = \frac{(A',A')}{(A,A)}$$

Note that (A',A') is greater than the minimum value α since α = (A,A) is the error signal energy of the linear prediction analysis.

Also recall from Chapter II that

$$(A(z),z^{-i}) = 0 \qquad \text{for} \qquad i = 1,2,....,M.$$

Since A(z)-A'(z) does not contain z°, A(z) is orthogonal to it and consequently

$$\begin{aligned}(A',A') &= (A'-A+A,A'-A+A)\\ &= (A,A)+(A'-A,A'-A)\end{aligned}$$

or

$$\psi = 1 + \frac{(A'-A,A'-A)}{(A,A)},$$

Therefore, the right hand term in (5.3.11) can be written as

$$\ln\ [1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{A(e^{j\theta};\lambda+\Delta\lambda)-A(e^{j\theta};\lambda)}{A(e^{j\theta};\lambda)}\right|^2 d\theta]$$

which is

$$\sim \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{A(e^{j\theta};\lambda+\Delta\lambda)-A(e^{j\theta};\lambda)}{A(e^{j\theta};\lambda)}\right|^2 d\theta] \qquad \text{(B-2)}$$

in the limit of small $\Delta\lambda$.

However consider

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln\left|\frac{A(e^{j\theta};\lambda+\Delta\lambda)}{A(e^{j\theta};\lambda)}\right|^2 d\theta \tag{B-3}$$

Let $x = \dfrac{A(e^{j\theta};\lambda+\Delta\lambda)-A(e^{j\theta};\lambda)}{A(e^{j\theta};\lambda)} = \dfrac{\Delta A(e^{j\theta})}{A(e^{j\theta};\lambda)}$

then the integrand in (B-3) becomes

$$\begin{aligned}\ln|1+x|^2 &= \ln[(1+x)(1+x^*)] \\ &= \ln[1+2\mathrm{Re}x+|x|^2] \\ &\sim 2\mathrm{Re}x+|x|^2 \quad \text{for small } x.\end{aligned}$$

However

$$\int_{-\pi}^{\pi} \frac{\Delta A(e^{j\theta})}{A(e^{j\theta};\lambda)} \frac{d\theta}{2\pi} = \int_{-\pi}^{\pi} \frac{\Delta A(e^{j\theta})A^*(e^{j\theta};\lambda)}{|A(e^{j\theta};\lambda)|^2} \frac{d\theta}{2\pi}$$

$$= \frac{(A,A'-A)}{(A,A)} = 0$$

Consequently (B-3) is approximately (B-2) and therefore

$$\ln \int_{-\pi}^{\pi} \left|\frac{A(e^{j\theta};\lambda+\Delta\lambda)}{A(e^{j\theta};\lambda)}\right|^2 \frac{d\theta}{2\pi} \sim \int_{-\pi}^{\pi} \ln\left|\frac{A(e^{j\theta};\lambda+\Delta\lambda)}{A(e^{j\theta};\lambda)}\right|^2 \frac{d\theta}{2\pi}$$

for small $\Delta\lambda$. But notice that, after the gain contribution is substracted, as was done for (5.3.11), distance measure (5.1.2) with p=1 is

$$\int_{-\pi}^{\pi} \left| \ln \left| \frac{A(e^{j\theta};\lambda+\Delta\lambda)}{A(e^{j\theta};\lambda)} \right|^2 \right| \frac{d\theta}{2\pi} \quad .$$

It is similar to (B-3) except that the absolute value of the log term is taken before integrating. This is an additional reason for preferring distance measure (5.1.2) to (5.1.5), because the absolute value prevents contributions with

$$|A(e^{j\theta};\lambda+\Delta\lambda)| < |A(e^{j\theta};\lambda)|$$

to cancel those with

$$|A(e^{j\theta};\lambda+\Delta\lambda)| > |A(e^{j\theta};\lambda)|$$

as can happen in (B-3) [15].

Appendix C

It is desired to obtain bounds concerning spectral deviations, for three different quantization schemes. The optimum bit allocation procedure for these three methods will then be discussed. It is first necessary to get a bound on the overall spectral deviation when all parameters are simultaneously quantized. Distance measure (5.1.2) will be used throughout. From the triangle inequality (5.1.10), it follows, inductively, that

$$D(\underline{\lambda},\underline{\lambda}'') \leq \sum_{i=1}^{T} D(\underline{\xi_i},\underline{\xi_{i+1}})$$

where $\underline{\xi_1} = \underline{\lambda}$, $\underline{\xi}_{T+1} = \underline{\lambda}''$ and all $\underline{\xi_i}$ are L-vectors with components $(\underline{\xi_i})_j$, $j = 1,2,\ldots,L$. Let $T\to\infty$ and $\underline{\xi_i} \to \underline{\xi_{i+1}}$. Expand $D(\underline{\xi_i},\underline{\xi_{i+1}})$ in a Taylor series about $\underline{\xi_i}$.

$$D(\underline{\xi_i},\underline{\xi_{i+1}}) \sim D(\underline{\xi_i},\underline{\xi_i}) + \sum_{j=1}^{L} \left.\frac{\partial D(\underline{\xi_i},\underline{\xi_{i+1}})}{\partial(\underline{\xi_{i+1}})_j}\right|_{\underline{\xi_{i+1}}=\underline{\xi_i}} ((\underline{\xi_{i+1}})_j-(\underline{\xi_i})_j)$$

But $D(\underline{\xi_i},\underline{\xi_i}) = 0$.

Therefore replacing the sum over index i by an integral over a continuous variable $(\underline{\xi})_j$

$$D(\underline{\lambda},\underline{\lambda}") \leq \sum_{j=1}^{L} \int_{(\underline{\lambda})_j}^{(\underline{\lambda}")_j} \frac{\partial D(\underline{\xi},\underline{y})}{\partial(\underline{y})_j} \Bigg|_{\underline{y}=\underline{\xi}} d(\underline{\xi})_j \qquad \text{(C-1)}$$

Defining $\Delta\underline{y} = (0,0,\ldots,0,(\Delta\underline{y})_j,0,\ldots,0)$ the integrand could have been written as

$$\lim_{(\Delta\underline{y})_j\to 0} \frac{D(\underline{\xi},\underline{y}+\Delta\underline{y})-D(\underline{\xi},\underline{y})}{(\Delta\underline{y})_j}\Bigg|_{\underline{y}=\underline{\xi}}$$

$$= \lim_{(\Delta\underline{\xi})_j\to 0} \frac{D(\underline{\xi},\underline{\xi}+\Delta\underline{\xi})}{(\Delta\underline{\xi})_j} = s_{(\underline{\xi})_j}(\underline{\xi}) \qquad \text{from (5.1.11),}$$

since only one parameter $(\underline{\xi})_j$ is varied by the definition of a partial derivative. However the integrand is a function of $\underline{\xi}$ and in going from $\underline{\lambda}$ to $\underline{\lambda}"$, variations have not been restricted to any particular subset of parameters. Therefore choose a path

$$\underline{\lambda} \triangleq (\lambda_1,\lambda_2,\ldots,\lambda_L) = \underline{\phi}_1 \to \underline{\phi}_2 \to \cdots \to \underline{\phi}_L \to \underline{\phi}_{L+1} = \underline{\lambda}" \triangleq (\lambda_1",\lambda_2",\ldots,\lambda_L")$$

such that

$$\underline{\phi}_m = (\lambda_1",\lambda_2",\ldots,\lambda_{m-1}",\lambda_m,\lambda_{m+1},\ldots,\lambda_L)$$

$$m = 1,2,\ldots,L+1$$

and only $\lambda_m = (\underline{\phi}_m)_m$ varies in going from pt $\underline{\phi}_m$ to pt $\underline{\phi}_{m+1}$. Using this path,

$$D(\underline{\lambda},\underline{\lambda}") \leq \sum_{m=1}^{L} \int_{(\underline{\phi}_m)_m}^{(\underline{\phi}_{m+1})_m} s_{(\underline{\xi})_m}((\underline{\xi})_m)\, d(\underline{\xi})_m$$

where $s_{(\underline{\xi})_m}(\underline{\xi})$ is written as $s_{(\underline{\xi})_m}((\underline{\xi})_m)$ to emphasize the fact that only the parameter $(\underline{\xi})_m$ varies in going from $\underline{\phi}_m$ to $\underline{\phi}_{m+1}$. As a result of this restriction, the definition of $\overline{D}$ can now be used to obtain

$$D(\underline{\lambda},\underline{\lambda}") \leq \sum_{m=1}^{L} \overline{D}(\lambda_1",\lambda_2",\ldots\lambda_{m-1}",\lambda_m,\lambda_{m+1},\ldots\lambda_L;\lambda_1",\lambda_{m-1}",\lambda_m",\lambda_{m+1},\ldots,\lambda_L)$$

(C-2)

The $\lambda_m"$ are to be interpreted as arbitrary but fixed quantized values of λ_m. Then choose the L parameters λ_m which will maximize $D(\underline{\lambda},\underline{\lambda}")$. Since (C-2) is true for any values of the parameters λ_m

$$\max_{\lambda_1\lambda_2\ldots\lambda_L} D(\underline{\lambda},\underline{\lambda}") \leq \max_{\lambda_1,\lambda_2\ldots\lambda_L} \sum_{m=1}^{L} \overline{D}(\underline{\phi}_m;\underline{\phi}_{m+1}) \leq \sum_{m=1}^{L} \max_{\lambda_1,\lambda_2,\ldots,\lambda_L} \overline{D}(\underline{\phi}_m;\ \underline{\phi}_{m+1})$$

$$\leq \sum_{m=1}^{L} \max_{\lambda_1,\lambda_2,\ldots\lambda_L} \overline{D}(\lambda_1,\lambda_2,\ldots\lambda_{m-1},\lambda_m,\ldots\lambda_L;\lambda_1,\lambda_2,\ldots\lambda_{m-1},\lambda_m",\ldots\lambda_L)$$

Let λ_m be uniformly and finely quantized into N_m levels (λ_m is an arbitrary transformation of a reflection coefficient k_m). Then

$$\max D(\underline{\lambda},\underline{\lambda}") \leq \sum_{i=1}^{L} \frac{\overline{\lambda_i}-\underline{\lambda_i}}{2N_i} \max_{k_1,k_2,\ldots k_m} \frac{\partial k_i}{\partial \lambda_i} s_{k_i}(k_i) \qquad \text{(C-3)}$$

Now consider the E $D(\underline{\lambda},\underline{\lambda}")$ where the average is over the random variables $\lambda_1,\lambda_2,\ldots\lambda_L$. Denote the probability of this set of parameters by $p(\lambda_1,\lambda_2,\ldots,\lambda_L)$. This can be rewritten as $p(\lambda_1,\lambda_2,\ldots\lambda_L/\lambda_m)p(\lambda_m)$. Hence

$E\ D(\underline{\lambda},\underline{\lambda}") \leq \sum_{m=1}^{L} E\overline{D}(\underline{\phi}_m;\underline{\phi}_{m+1})$. First average $D(\underline{\phi}_m;\underline{\phi}_{m+1})$ over $\lambda_1,\lambda_2\ \ldots,\lambda_L$ for fixed λ_m. Since m-1 parameters are already quantized in $\underline{\phi}_m$, integrating $p(\lambda_1,\lambda_2\ldots\lambda_L/\lambda_m)\overline{D}(\underline{\phi}_m;\underline{\phi}_{m+1})$ over any one of these m-1 parameters (say the jth one), yields

$$\sum_{n=1}^{N_j} \overline{D}(\lambda_1",\lambda_2",\ldots\lambda_j"(n),\ldots,\lambda_m"\ldots\lambda_L;\lambda_1",\lambda_2",\ldots,\lambda_j"(n),\ldots\lambda_m",\ldots\lambda_L)$$

$$\int_{\lambda_j(n)}^{\lambda_j(n+1)} p(\lambda_1,\lambda_2,\ldots,\lambda_L/\lambda_m)\,d\lambda_j$$

where $\lambda_j"(n)$ is the quantized value of the jth parameter if that parameter lies in $(\lambda_j(n),\ \lambda_j(n+1))$. For fine quantization of all L parameters, replace $\underline{\phi}_m$ and $\underline{\phi}_{m+1}$, by $(\lambda_1,\lambda_2,\ldots,\lambda_{m-1}\lambda_m,\lambda_{m+1}\ldots,\lambda_L)$ and $(\lambda_1,\lambda_2\ldots\lambda_{m-1}\lambda_m'',\ldots\lambda_L)$ respectively and $\bar{D}$ can appear inside the integral in the above expression which reduces then to:

$$\int_{\underline{\lambda}_j}^{\bar{\lambda}_j} p(\lambda_1\lambda_2\ldots\lambda_L/\lambda_m)\bar{D}(\lambda_1\ldots\lambda_m\ldots\lambda_L;\lambda_1\lambda_2\ldots\lambda_m''\ldots\lambda_L)\,d\lambda_j$$

further integrating this over all parameters $\lambda_{i\neq m}$, an average denoted by $\bar{D}_m$ is obtained. Hence, for fine quantization,

$$ED(\underline{\lambda},\underline{\lambda}'')\leq\sum_{m=1}^{L}E\bar{D}_m \equiv \sum_{m=1}^{L}\sum_{n=1}^{N_m}\int_{\lambda_m(n)}^{\lambda_m(n+1)} p_{\Lambda_m}(\lambda_m)\bar{D}_m(\lambda_m,\lambda_m''(n))\,d\lambda_m$$

which by the previous asymptotic result in the single parameter case, equals

$$\sum_{m=1}^{L}\frac{1}{4N_m}\int_{\underline{\lambda}_m}^{\bar{\lambda}_m}\frac{Es_{\Lambda_m}(\lambda_m)p_{\Lambda_m}(\lambda_m)}{u_m(\lambda_m)}\,d\lambda_m \qquad \text{(C-4)}$$

where $Es_{\Lambda_m}(\lambda_m)$ is the average over all other parameters $\lambda_{i\neq m}$.

A bound on the total spectral deviation must now be found when $A(z) = \sum_{i=0}^{M} a_i z^{-i}$ is factored into a product of quadratic polynomials and 2 parameter quantization is applied on each of these polynomials. First, factor A(z) into q polynomials: $A_1 A_2 \ldots A_q$. Denote the corresponding quantized polynomial A'(z) by $A_1' A_2' \ldots A_q'$. Substituting in (5.1.2) yields (gain normalization $\sigma(\underline{\lambda}) = 1$)

$$D_{tot} = \left[\int \left| \sum_{j=1}^{q} \left(\ln \frac{1}{|A_j|^2} - \ln \frac{1}{|A_j'|^2} \right) \right|^p \frac{d\theta}{2\pi} \right]^{1/p}$$

Now by the Minkowski inequality [20]

$$\left[\sum_{i=1}^{n} |x_i + y_i|^p \right]^{1/p} \leq \left[\sum_{i=1}^{n} |x_i|^p \right]^{1/p} + \left[\sum_{i=1}^{n} |y_i|^p \right]^{1/p}$$

This can be generalized if $x_i + y_i$ is replaced by $\sum_{j=1}^{q} x_{ji}$ to yield

$$\left[\sum_{i=1}^{n} \left| \sum_{j=1}^{q} x_{ji} \right|^p \right]^{1/p} \leq \sum_{j=1}^{q} \left[\sum_{i=1}^{n} |x_{ij}|^p \right]^{1/p}$$

Replacing the summation by an integral gives

$$\left[\int \left| \sum_{j=1}^{q} x_j(t) \right|^p dt \right]^{1/p} \leq \sum_{j=1}^{q} \left[\int |x_j(t)|^p \right]^{1/p} \qquad \text{(C-5)}$$

let $dt = d\theta$ and

$$x_j = \ln \frac{1}{|A_j|^2} - \ln \frac{1}{|A'_j|^2}$$

if $\lfloor M/2 \rfloor = M/2$, then $q = \lfloor M/2 \rfloor$ and each A_j will be a quadratic polynomial. From (5.3.26), the jth term in (C-5) becomes

$$D_j = \left[\frac{\pi}{N_j} \ln \frac{1-\underline{k}_{2,j}}{1-\overline{k}_{2,j}} \right]^{1/2}$$

If $M/2 \neq \lfloor M/2 \rfloor$, then there is a leftover linear term $1 + a_1 z^{-1}$. Treating it as a linear prediction filter of order M=1, a single parameter analysis is applied since there is only one parameter, namely $a_1 = k_1$. Recall that a general filter $A(z;\lambda)$ is a linear function of each k_i and therefore, using the recursion formulae developed in Chapter II, $A(z;\lambda) = A_{M-1}(z)+k_M B_{M-1}(z)$. But k_M does not appear in any A_m, B_m where $m<M$. As a result

$$\frac{\partial A}{\partial k_M}(z;\lambda) = B_{M-1}(z) \qquad \text{so that}$$

$$D^2 = \frac{2}{\alpha}(\Delta A, \Delta A) = \frac{2}{\alpha}(\Delta k_M)^2 \left(\frac{\partial A}{\partial k_M}, \frac{\partial A}{\partial k_M}\right)$$

$$= \frac{2(\Delta k_M)^2}{\alpha} \alpha_{M-1} = \frac{2(\Delta k_M)^2}{1-k_M^2} \qquad \text{(C-7)}$$

Therefore in the 2 parameter quantization scheme, [14], the filter of order M=1 will contribute a term

$$D = \frac{\overline{\lambda}_1 - \underline{\lambda}_1}{2N} \frac{\partial k_1}{\partial \lambda_1} \cdot \sqrt{\frac{2}{1-k_1^{\,2}}} \qquad \text{(C-8)}$$

There remains to determine the optimum allocation of bits which minimizes the total bit rate $B = \sum_i \log N_i$, subject to equality constraints on the total bounds. Denoting bounds (C-3) and (C-4) by max $\overline{D}_{tot}$ and $E\overline{D}_{tot}$ respectively, it is seen that their dependence on the N_i's are both of the form

$$\sum_{i=1}^{M} T_i/N_i \quad \text{where } T_i \text{ does not depend on } N_i. \quad \text{(C-9)}$$

This constraint problem is then solved by introducing a Lagrangian multiplier γ and a function F defined by

$$F = \gamma \sum_{i=1}^{M} \frac{T_i}{N_i} + \sum_{i=1}^{M} \log N_i$$

The solution is given by

$$\frac{\partial F}{\partial N_i} = -\frac{\gamma T_i}{N_i^{\,2}} + \frac{1}{N_i} = 0 \text{ or } T_i/N_i = 1/\gamma \quad \text{a constant.}$$

The value of this constant γ^{-1}, after substituting in (C-9) is found to be either $\frac{\max\overline{D}_{tot}}{M}$ or $\frac{E\overline{D}_{tot}}{M}$ depending upon which criterion is utilized. It is therefore seen that minimization of total bit rate is achieved by setting all individual single parameter bounds to the same value.

If however, a parameter quantization is performed, then the overall bound in (C-5) is used. The Lagrangian multiplier solution of the constrained minimum is derived from using

$$F = \gamma \frac{T_o}{N_o} + \gamma \sum_{i=1}^{\lfloor M/2 \rfloor} \frac{T_i}{\sqrt{N_i}} + \{\log N_o + \sum_{i=1}^{\lfloor M/2 \rfloor} \log N_i\} \qquad \text{(C-10)}$$

where the first term represents (C-8) in the case there is a leftover term ($M/2 \neq \lfloor M/2 \rfloor$), the second term the bound (C-5) with $D_i = \frac{T_i}{\sqrt{N_i}}$ as defined by (C-6) and the third term in brackets is the total bit rate. If $M/2 = \lfloor M/2 \rfloor$ then T_o is set to 0 and N_o is set to 1 in order that $\log N_o$ equals 0. If $T_o \neq 0$, the solution to the leftover term is $\frac{T_o}{N_o} = \frac{1}{\gamma}$ as in the single parameter analysis.

For $i = 1,2,\ldots, \lfloor M/2 \rfloor$

$$\frac{\partial F}{\partial N_i} = -\frac{1}{2} \frac{\gamma T_i}{N_i^{3/2}} + \frac{1}{N_i} = 0 \quad \text{or} \quad \frac{T_i}{\sqrt{N_i}} = \frac{2}{\gamma}.$$

Therefore, if $T_o \neq 0$, (denoting the overall bound on the right-hand side of C-5 by D_b),

$$D_b = \frac{1}{\gamma} + \lfloor M/2 \rfloor \frac{2}{\gamma} = \frac{1}{\gamma} + \frac{M-1}{\gamma} = \frac{M}{\gamma}$$

Therefore $\frac{T_o}{N_o} = \frac{D_b}{M}$ and $\frac{T_i}{\sqrt{N_i}} = \frac{D_b}{M/2}$

if $M/2 = \lfloor M/2 \rfloor$, then $T_o = 0$ and $D_b = (M/2)\frac{2}{\gamma}$

or $\frac{T_i}{\sqrt{N_i}} = \frac{D_b}{M/2}$ as before.

REFERENCES

1. Flanagan, J.L., Speech Analysis, Synthesis and Perception, second expanded edition, Springer-Verlag (1972).
2. Markel, J.D., Gray, A.H., Jr., Linear Prediction of Speech, Springer-Verlag (1976).
3. Oppenheim, A.V., Schafer, R.W., Digital Signal Processing, Prentice-Hall, Inc. (1975).
4. Kang, G.S., Application of linear prediction encoding to a narrowband voice digitizer, NRL Report #7774, October 1974, Naval Research Lab.
5. Strube, H.W., Determination of the instant of glottal closure from the speech wave. J. Acoust. Soc. Am., 56, No. 5, November 1974, pp. 1625-1629.
6. Steiglitz, K., Dickinson, B., The use of time-domain selection for improved linear prediction, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 25, No. 1, February 1977, pp. 34-39.
7. Kopec, G.E., Oppenheim, A.V., Tribolet, J.M., Speech analysis by homomorphic prediction, IEEE Transactions on ASSP, Vol. 25, No. 1, February 1977, pp. 40-49.
8. Steiglitz, K., On the simultaneous estimation of poles and zeros in speech analysis, IEEE Transactions on ASSP, Vol. 25, No. 3, June 1977, pp. 229-234.

9. Markel, J.D., The SIFT algorithm for fundamental frequency estimation, IEEE Transactions on Audio and Electroacoustics, Vol. 20, No. 5, December 1972, p. 367-377.

10. Markel, J.D., Gray, A.H., Jr., A linear prediction vocoder simulation based upon the autocorrelation method, IEEE Transactions on ASSP, Vol. 22, No. 2, April 1974, pp. 124-124.

11. Markel, J.D., Gray, A.H., Jr., Distance measures for speech processing, IEEE Transactions on ASSP, Vol. 24, No. 5, October 1976, pp. 380-390.

12. Gray, A.H., Jr., Gray, R.M., Markel, J.D., Comparison of optimal quantizations of speech reflection coefficients, IEEE Transactions on ASSP, Vol. 25, No. 1, February 1977, pp. 9-22.

13. Gallager, R.G., Information Theory and Reliable Communication, New York: Wiley,1968.

14. Gray, A.H., Jr., Markel, J.D., Quantization and bit allocation in speech processing, IEEE Transactions on ASSP, Vol. 24, No. 6, December 1976, pp. 459-473.

15. Viswanathan, R., Makhoul, J., Quantization properties of transmission parameters in linear predictive systems, IEEE Transactions on ASSP, Vol. 23, No. 3, June 1975, pp. 309-321.

16. Chandra, S., Lin, W.C., Linear prediction with a variable analysis frame size, IEEE Transactions on ASSP, Vol. 25, No. 4, August 1977, pp. 322-330.

17. McCandless, S.S., A new encoding technique for the k-parameters: A statistical approach, December 1974, NSC Note # 53.

18. Sambur, M.R., An efficient linear prediction vocoder, Bell Syst. Tech. J., Vol. 54, pp. 1693-1723, December 1975.

19. Carnahan, B., Luther, H.A., Wilkes, J.O., Applied Numerical Methods, New York: Wiley, 1969.

20. Goffman, C., Pedrick, G., First Course in Functional Analysis, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1965.

21. Sambur, M.R., Speaker recognition using orthogonal linear prediction, IEEE Transactions on ASSP, Vol. 24, No. 4, August 1976, pp. 283-289.

22. McGonegal, C.A., Rabiner, L.R., Rosenberg, A.E., A subjective evaluation of pitch detection methods using LPC synthesized speech, IEEE Transactions on ASSP, Vol. 25, No. 3, June 1977, pp. 221-229.