

University of Windsor

Scholarship at UWindor

Electronic Theses and Dissertations

Theses, Dissertations, and Major Papers

1973

Synthesis and coding of voice signals.

Bohsko. Cirjanic
University of Windsor

Follow this and additional works at: <https://scholar.uwindsor.ca/etd>

Recommended Citation

Cirjanic, Bohsko., "Synthesis and coding of voice signals." (1973). *Electronic Theses and Dissertations*. 3689.

<https://scholar.uwindsor.ca/etd/3689>

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000ext. 3208.

SYNTHESIS AND CODING OF VOICE SIGNALS

by

BOSKO CIRJANIC

A DISSERTATION

Submitted to the Faculty of Graduate Studies
through the Department of Electrical Engineering
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario,

Canada

1973

© Boshko Cirjanic 1973

447639

ABSTRACT

The thesis describes new techniques for speech synthesis and coding. The linear predictor proposed by Atal[13] is analyzed from the point of view of practical realization and an adaptively adjusted parallel form realization has been developed to overcome the distortion due to parameter quantization and instability due to finite word length of data. The new scheme is also shown to result in a significant improvement in the quality of reconstructed speech.

A new coding scheme that incorporates a primary stage (this is the adaptive DPCM scheme developed by Atal and Schroeder) and a secondary stage incorporating an optimally designed quantizer is shown to result in a higher quality speech for bit rates that are comparable to those in recently proposed schemes.

The use of the FFT algorithm in conjunction with a spectral shaping technique is shown to result in a computationally efficient scheme for synthesizing speech whose quality is superior to those synthesized by the use of linear predictors. Also the high computational speeds attainable with the use of the FFT algorithm is shown to result in a new coding scheme that permits analytic determination of optimum quantizer levels.

ACKNOWLEDGEMENT

The advice, guidance and constant encouragement of my supervisor Dr. M. Shridhar is acknowledged with gratitude. In addition, the many stimulating discussions with other faculty members of the Electrical Engineering Department are gratefully acknowledged.

I am also indebted to my wife, Mirjana, and my daughter, Jasminka, for their moral support and understanding during the course of this work. Thanks are due to Mrs. S. Ouellette who did an excellent job of typing this dissertation.

I would like to acknowledge the financial support afforded me in the form of a Scholarship by the National Research Council of Canada.

FIGURE. ILLUSTRATIONS

<u>No.</u>	<u>Title</u>	<u>Page No.</u>
1.	Speech production apparatus	3.
2.	The auditory system	6.
3.	Formant method of speech synthesis	9.
4.	Homomorphic speech analysis	11.
5.	Homomorphic speech synthesis	11.
6.	Maximum likelihood vocoder synthesis scheme	13.
7.	Linear predictor speech synthesis	16.
8.	Open loop predictive coding	19.
9.	Differential pulse code modulation (DPCM)	20.
10.	2-stage adaptive coding	21.
11.	Adaptive predictive coding	22.
12.	Linear predictor frequency	34.
(a)	Square root method solution	34.
(b)	Gaussian elimination method solution	34.
13.	Operations involved in obtaining the Cepstrum	36.
14.	Simplified model of the vocal tract impulse response	37.
15.	Cepstrum and normalized correlation plots	40.
16.	Direct realization form of the linear predictor	42.
17.	Approximation to the glottal pulse shape	45.
18.	Triangular excitation pulse	45.
19.(a)	Original waveform of the word - "NOON"	48.
(b)	Original waveform of the word - "DAY"	49.
(c)	Original waveform of the word - "SLEEPY"	50.
(d)	Original waveform of the word - "HOW ARE YOU"	51.

20.(a)	Reconstructed waveform using impulse excitation - "NOON"	52.
(b)	Reconstructed waveform using impulse excitation - "DAY"	53.
(c)	Reconstructed waveform using impulse excitation - "SLEEPY"	54.
(d)	Reconstructed waveform using impulse excitation - "HOW ARE YOU"	55.
21.(a)	Reconstructed waveform using triangular excitation - "NOON"	56.
(b)	Reconstructed waveform using triangular excitation - "DAY"	57.
(c)	Reconstructed waveform using triangular excitation - "SLEEPY"	58.
(d)	Reconstructed waveform using triangular excitation - "HOW ARE YOU"	59.
22.(a)	Pole-zero model reconstruction of the word - "NOON"	62.
(b)	Pole-zero model reconstruction of the word - "DAY"	63.
(c)	Pole-zero model reconstruction of the word - "SLEEPY"	64.
(d)	Pole-zero model reconstruction of the word - "HOW ARE YOU"	65.
23.(a)	Initial condition model reconstruction of the word - "NOON"	67.
(b)	Initial condition model reconstruction of the word - "DAY"	68.
(c)	Initial condition model reconstruction of the word - "SLEEPY"	69.
(d)	Initial condition model reconstruction of the word - "HOW ARE YOU"	70.
24.	Parallel realization form of the linear predictor	73.
25.(a)	Parallel realization model reconstruction of the word - "NOON"	75.
(b)	Parallel realization model reconstruction of the word - "DAY"	76.
(c)	Parallel realization model reconstruction of the word - "SLEEPY"	77.
(d)	Parallel realization model reconstruction of the word - "HOW ARE YOU"	78.
26.(a)	Closed form model with adaptive order of system reconstruction of the word - "NOON"	82.
(b)	Closed form model with adaptive order of system reconstruction of the word - "DAY"	83.

26.(c)	Closed form model with adaptive order of system reconstruction of the word - "SLEEPY"	84.
(d)	Closed form model with adaptive order of system reconstruction of the word - "HOW ARE YOU"	85.
27.	Input/output relationship of a 4 level quantizer	90.
28.	Odd and even number of quantizer levels	90.
29.	Open loop predictive coding	93.
30.	Proposed 2 stage predictive coding scheme	96.
31.	Equivalent system of the DPCM	99.
32.(a)	Primary coding stage reconstruction of the word - "NOON"	104.
(b)	Primary coding stage reconstruction of the word - "DAY"	105.
(c)	Primary coding stage reconstruction of the word - "SLEEPY"	106.
(d)	Primary coding stage reconstruction of the word - "HOW ARE YOU"	107.
33.(a)	Secondary coding stage reconstruction of the word - "NOON"	108.
(b)	Secondary coding stage reconstruction of the word - "DAY"	109.
(c)	Secondary coding stage reconstruction of the word - "SLEEPY"	110.
(d)	Secondary coding stage reconstruction of the word - "HOW ARE YOU"	111.
34.	FFT model for speech analysis	116.
35.	FFT model for speech synthesis	117.
36.	Proposed FFT coding scheme	121.
37.(a)	FFT model speech synthesis using 6 bands - "NOON"	125.
(b)	FFT model speech synthesis using 6 bands - "DAY"	126.
(c)	FFT model speech synthesis using 6 bands - "SLEEPY"	127.
(d)	FFT model speech synthesis using 6 bands - "HOW ARE YOU"	128.
38.(a)	FFT model speech synthesis using 12 bands - "NOON"	129.

38.(b)	FFT model speech synthesis using 12 bands - "DAY"	130.
(c)	FFT model speech synthesis using 12 bands - "SLEEPY"	131.
(d)	FFT model speech synthesis using 12 bands - "HOW ARE YOU"	132.
39.(a)	FFT model speech coding using 6 bands - "NOON"	133.
(b)	FFT model speech coding using 6 bands - "DAY"	134.
(c)	FFT model speech coding using 6 bands - "SLEEPY"	135.
(d)	FFT model speech coding using 6 bands - "HOW ARE YOU"	136.
40.(a)	FFT model speech coding using 12 bands - "NOON"	137.
(b)	FFT model speech coding using 12 bands - "DAY"	138.
(c)	FFT model speech coding using 12 bands - "SLEEPY"	139.
(d)	FFT model speech coding using 12 bands - "HOW ARE YOU"	140.
41.	Laboratory computing facility for speech processing	144.
42.	Flow chart describing the operation of programme PART 1	146.
43.	Flow chart describing the operation of programme OVERLAY 1	148.
44.	Flow chart describing the operation of programme OVERLAY 2	150.
45.	Original waveform of the sentence - "WE WERE AWAY A YEAR AGO"	152.
46.	Initial condition model reconstruction of the sentence - "WE WERE AWAY A YEAR AGO"	153.
47.	DPCM fixed quantizer height reconstruction of the sentence - "WE WERE AWAY A YEAR AGO"	155.
48.	DPCM variable quantizer height reconstruction of the sentence - "WE WERE AWAY A YEAR AGO"	157.
49.	Phomeme synthesis of the word - "NOON"	159.
50.	Phomeme synthesis of the words - "HOW ARE YOU"	160.

LIST OF TABLES

<u>Number</u>	<u>Title</u>	<u>Page No.</u>
TABLE 1	Number of Multiplications Required to Form $\phi a = \psi$	30.
TABLE 2	Predictor Parameters Obtained By Using Two Different Techniques	31.
TABLE 3	Pole Locations of the Linear Predictor Obtained By The Two Different Methods.	32.
TABLE 4	Speech Signal Frequency Spectrum Divided Into 6 Bands For the FFT Model	114.
TABLE 5	Speech Signal Frequency Spectrum Divided Into 12 Bands For The FFT Model	120.

TABLE OF CONTENTS

	<u>Page No.</u>
TITLE PAGE.	i
APPROVAL PAGE	ii
ABSTRACT.	iii
ACKNOWLEDGEMENTS.	iv
FIGURE ILLUSTRATIONS.	v
LIST OF TABLES.	ix
TABLE OF CONTENTS	x
Chapter	
I. INTRODUCTION	1
1.1. Historical Background.....	1
1.2.i. Mechanism of Speech Production.....	2
1.2.ii. Ear and the Hearing Mechanism.....	5
1.3. Current State of the Art in Synthesis and Coding..	7
1.3.i. Formant Analysis-Synthesis of Speech.....	8
1.3.ii. The Homomorphic Vocoder.....	10
1.3.iii. Maximum Likelihood Vocoder.....	12
1.3.iv. Linear Predictor.....	15
1.3.v. Digital Inverse Filtering.....	16
1.4. Problem Statement.....	23
1.5. Thesis Organization... ..	24
II. TIME DOMAIN SPEECH SYNTHESIS.	26
2.1. Linear Predictor.....	26
2.1.i. Evaluation of Predictor Parameters.....	28
2.2. Pitch Period Extraction.....	33
2.2.i. Zero Crossing Rate.....	33
2.2.ii. The Cepstrum.....	33
2.2.iii. Normalized Correlation.....	38
2.3. Linear Predictor Realization Schemes.....	41
2.3.i. Direct Form.....	41
2.3.i(a) Linear Predictor Excitation Function.....	43
2.3.i(b) Linear Predictor With Zeros.....	47
2.3.i(c) Linear Predictor With Initial Conditions.....	61
2.3.ii. Parallel Realization.....	71
2.4. Closed Form.....	74
2.4.i. Exponential Curve Fitting.....	74
2.4.ii. Adaptive Adjustment of System Order.....	80
2.5. Summary.....	81

III. CODING OF SPEECH SIGNALS.	88
3. Coding of Speech Signals.....	88
3.1. Predictive Coding of Speech.....	88
3.1.1. Quantization Techniques.....	89
3.2. Open-Loop Predictive Coding.....	92
3.3. Differential Pulse Code Modulation.....	94
3.4. Two Stage Adaptive Predictive Coding.....	95
3.4.1. Primary Stage.....	97
3.4.1i. Secondary Stage.....	98
3.5. Further Observations.....	99
3.6. Bit-Rate Requirements.....	101
3.7. Simulation Results.....	103
3.8. Summary.....	103
IV. FREQUENCY DOMAIN SPEECH SYNTHESIS AND CODING	112
4. Frequency Domain Speech Synthesis and Coding.....	112
4.1. FFT Model for Speech Synthesis.....	112
4.2. Digital Implementation.....	118
4.3. Techniques For Improving Speech Quality.....	119
4.4. FFT Model Coding.....	121
4.5. Bit-Rate Requirements.....	122
4.5.1. FFT Model Speech Synthesis.....	122
4.5.1i. FFT Model Coding.....	123
4.6. Simulation Results.....	124
4.7. Summary.....	124
V. MINI-COMPUTER IMPLEMENTATION OF SPEECH SYNTHESIS & CODING	142
5. Mini-Computer Implementation of Speech	
Synthesis and Coding.....	142
5.1. Recording of Spoken Material.....	142
5.2. Aid Conversion.....	143
5.3. Software Package For Speech Synthesis.....	143
5.3.1. PART 1.....	145
5.3.1i. OVERLAY 1.....	147
5.3.1ii. OVERLAY 2.....	147
5.4. Implementation and Results.....	149
5.4.1. Initial Condition Model.....	149
5.4.1i. Coding of Speech Signals.....	151
5.4.1i(a) DPCM-Fixed Quantizer Height.....	154
5.4.1i(b) DPCM - Variable Quantizer Height.....	154
5.5. Phoneme Synthesis.....	156
5.6. Summary.....	158

~~SECRET~~

VI. CONCLUSIONS.	162
6.1. Linear Predictor.	162
6.2. Coding of Speech Signals.	162
6.3. Speech Synthesis and Coding by Use of an FFT Model.	163
REFERENCES.	164
VITA AUCTORIS	168

CHAPTER I

INTRODUCTION

1.1. Historical Background

Speech production by machines or instruments has been the goal of many researchers for the past two centuries. One of the earliest successful attempts at mechanical production of speech, using accoustical resonators, was demonstrated by Krantzenstein in 1774. These resonators could only produce the five vowels, a, e, i, o, and u. Soon after Krantzenstein's success, Von Kempelen demonstrated a speaking machine which was not taken seriously. Sir Charles Wheatstone (a maker of musical instruments) built a replica of Von Kempelen's speaking machine, which was capable of generating connected speech. Other peoples interests were aroused, one of these being Alexander Graham Bell.

The first electrical speech synthesizer, called the Voder[1], was demonstrated in 1939 at the New York world fair. The synthesizer, developed by Dudley et al, at the Bell Telephone Laboratories, was capable of producing connected speech. It was played in much the same way as a piano, and the operators required extensive training to become proficient. Soon after the introduction of the Voder, a new invention was demonstrated by H. Dudley of the Bell Telephone Laboratories. It was called the Vocoder [2], deriving its name from Voice Coder.

The Vocoder consisted of essentially three separate units, the analyzer, the transmission path and the synthesizer. The speech signal was applied to a bank of 10 band-pass filters, each 300 Hz wide and covering the 3 kHz speech spectrum in adjacent bands. The output of each filter is rectified and smoothed before being sampled and sent

along the transmission line. At the receiving end, there are ten identical band-pass filters which are excited either by random noise or by quasi periodic pulses. The output of each band-pass filter is then modulated by the spectrum defining signal. The signals are then summed and applied to an equalizer to produce speech. Its main use was seen to be in speech transmission, where a bandwidth compression was possible. Many variations of the original Vocoder have been built.

In 1946, Koenig, Dun and Lacey[3], demonstrated a machine that was capable of converting audio signals into visual patterns that describe the characteristics of a particular message. This machine was called the spectrograph, and it was seen as a means of teaching the deaf to communicate with other people by observing the different patterns (spectrograms) produced by the spectrograph. The training process was long, especially since the analyzing time of the spectrograph was of the order of 100 times real time. The resulting plot from a spectrograph is one of time-intensity-frequency. These plots can now be generated on a digital computer[4], especially with the discovery of the Fast Fourier Transform algorithm[5].

1.2.1 Mechanism of Speech Production

Speech begins as an idea expressible in some language. The idea starts a sequence of neural and muscular operations which will eventually result in an audible message being generated, by habit or previous experience[6]. The major parts involved in the generation of speech are shown in Figure 1.

Speech is produced by the actions of the nose, mouth, throat, etc.

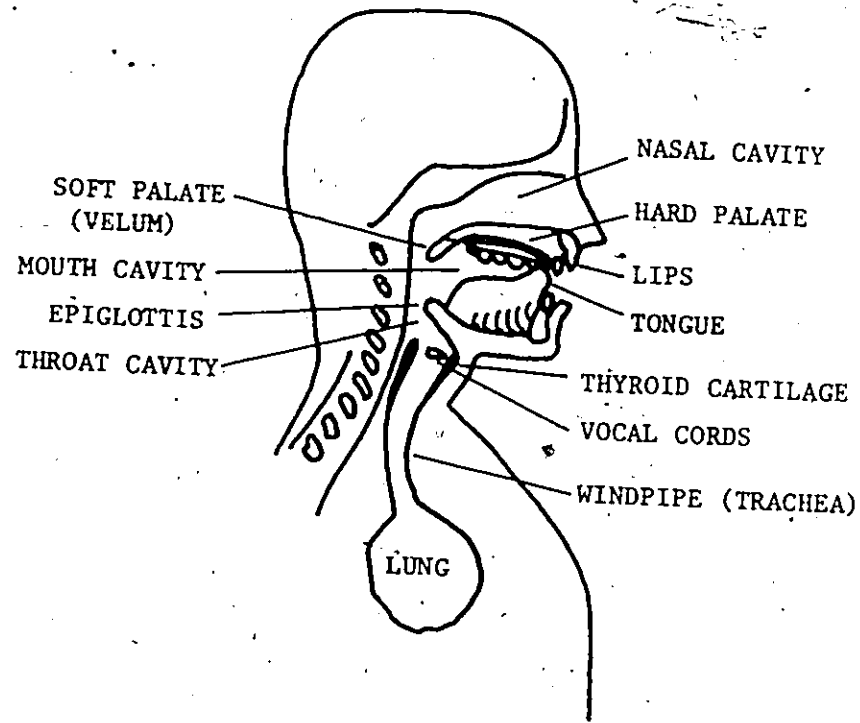


FIGURE 1

Speech Production Apparatus

upon the breath stream as it passes from the glottal orifice to the air surrounding the lips. For most speech sounds, air is exhaled from the lungs, except for a class of sounds known as clicks, which are produced during inhalation. The windpipe connects the lungs at the one end and the throat cavity at the other end. At the top of the windpipe is the larynx, which houses the vocal cords. The vocal cords are shaped like two lips and lie at either side of the larynx. These cords play an important role in speech production of voiced sounds. For voiced sounds, the vocal cords are brought together and then set into vibration by forcing air between them. The mass, compliance and the flow of air through the glottal orifice, determine the frequency of vibration of the cords.

In normal speech, the vocal cord vibration rate, or the pitch frequency covers the range from about 80 to 350 Hz [7]. For voiced speech the vocal cords are made to vibrate at a particular frequency, thus allowing bursts of air to enter the throat cavity. The air then enters the vocal tract, whose shape determines the type of sound to be produced. The shape of the vocal tract is determined by the position of the lips, jaw, tongue and the velum. In connected speech all of these are moving continuously to produce the desired sounds. The velum, when in the open position, allows air to enter the nasal cavity. Sounds produced by the nasal cavity are characterized as nasals, and for these sounds, the teeth and lips are brought close together to effectively block off the passage of air through them.

Sounds are basically generated in three different ways. Voiced sounds are produced by forcing air to flow through the vocal cord orifice,

causing the cords to vibrate. The opening and closing of the glottis produces quasi periodic pulses of air which excite the vocal tract.

Nasal sounds are normally excited by the vocal cords and hence are voiced sounds. Typical examples of voiced sounds are the vowels and vowel like sounds.

Fricative sounds are produced by forming a constriction at some point in the vocal tract, anywhere from the glottis to the lips. Air is forced through these constrictions, causing random eddies to be produced which are heard as hissing noise. The point of noise excitation can, therefore be situated anywhere along the vocal tract. These sounds are normally of a very low intensity, having a wide frequency spectrum. Typical examples of fricative sounds are s, f and sh.

The third type, the plosive sounds, are created by making a complete closure of the vocal tract, building up a pressure behind the closure and abruptly realizing it. The abrupt release of pressure provides an impulsive excitation of the vocal tract. Examples of the plosive sounds are, p, t and k.

1.2.11 Ear and the Hearing Mechanism

The auditory system, shown in Figure 2, can be divided into three regions, the outer ear, the middle ear and the inner ear. The external and middle ears together amplify the pressure variation of auditory sounds and transfer them to fluids within the inner ear. The pressure variations are sorted out by the frequency-analyzer action of the cochlea, and encoded into pulses of electro-chemical activity in the cochlear nerve fibres. These pulses are transmitted through the neural system

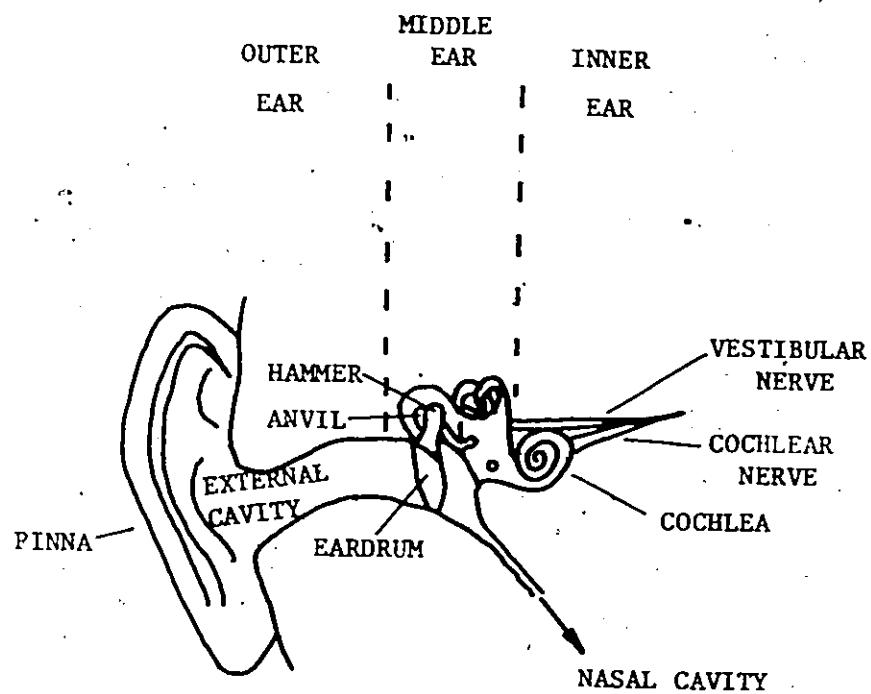


FIGURE 2

The Auditory System

which extends to the brain.

Any pressure variation acting upon the eardrum, must be of a sufficiently high amplitude to overcome the auditory system background noise. The background noise is made up from the circulation of the blood, breathing, muscle tremors and movements of the head. However, the ear is constructed to respond far less efficiently to such vibration than it is to air-borne sounds.

The outer ear consists of the pinna which serves two purposes:

- i) that of protecting the external canal -and-
- ii. to direct the surrounding sounds into the external canal.

The middle ear consists of the eardrum and the ossicular bones.

The hammer is fixed to the eardrum and makes contact with the anvil.

The anvil connects to the stirrup, and the foot plate of the stirrup seats in a port, the oval window.

The operation of the ear is as follows. A sound wave entering the external canal sets the eardrum into vibration. This vibration is transmitted, with a slight amplification, via the ossicular bones to the inner ear. When the stapes are set into vibration, they produce a displacement in the cochlear fluid. The motion of the fluid has the effect of polarizing the hair cells found inside the cochlea. These hair cells produce electrical impulses which travel through the neural system to the brain. The eardrum, itself, is not a stretched diaphragm but resembles a shallow rigid cone with a fold in the lower part allowing the eardrum to oscillate as a whole unit.

1.3 Current State of the Art in Synthesis and Coding

The techniques used in the analysis and synthesis of speech can be

broadly classified into two groups. The first group makes use of the frequency domain information and essentially tries to reproduce the vocal tract resonances. The second group uses time domain methods and can be thought of as some form of curve fitting. In the past, the frequency domain techniques have attracted much more attention, because certain features are more easily recognized using frequency domain information. It is normally easier to reproduce the frequency spectrum envelope, than it is to reproduce the actual time domain signal.

1.3.1 Formant Analysis-Synthesis of Speech[8]

The analysis and synthesis of speech by the formant method, has been used by Bell Telephone Laboratories for some time. Essentially, the speech signal is analyzed and stored in terms of the vocal tract resonances. These resonances, called, formants, are changing continuously as the shape of the vocal tract changes. The rate of change of these formants is much smaller as compared to the rate of change of the actual speech signal. Typically, if the range of human voice is limited to lie between 0 and 3 kHz, for voiced sounds, there will be approximately 3 formants.

The pitch period is estimated using the cepstrum, and the formant frequencies are obtained from the location of peaks in the spectrum envelope. For voiced sounds, three time varying formants are used, whereas for the unvoiced sounds a pair of complex poles and zeros are sufficient to describe the vocal tract characteristics. A typical formant method of speech synthesis is shown in Figure 3. The formant technique can present difficulties in cases where two formants are

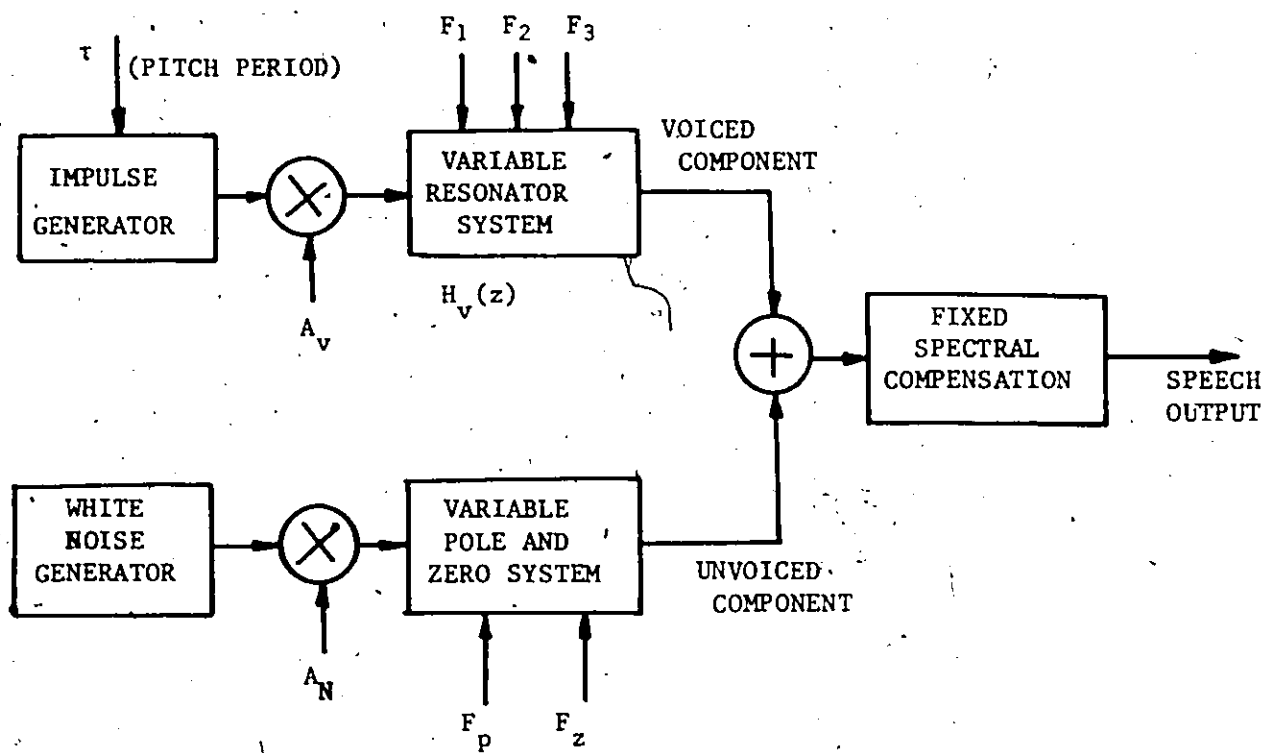


FIGURE 3 - Formant Method of Speech Synthesis

close together, usually the second and third. One way of overcoming this problem is to use the chirp z-transform[9,10], which evaluates the vocal tract transfer function in a manner that sharpens the resonances.

With efficient coding of the pitch and formant data, it is possible for speech to be represented by approximately 600 bits/sec. Another feature of this technique is, that, it is possible to connect different segments of the formant data to form a new message. However, it is not a straight forward matter to connect discrete formants together, since in connected speech the word or phoneme boundaries are not abrupt, but more of a transient nature. Furthermore, there is an interaction between adjacent phonemes, which must be preserved in high quality speech synthesis.

1.3.11 The Homomorphic Vocoder[11]

Homomorphic filtering is a technique used in separating signals combined by convolution and multiplication. In the case of speech signals, it is used to separate out the excitation and vocal tract components. The speech signal is essentially the result of the glottal pulse excitation convolved with the transfer function of the vocal tract. A homomorphic system for estimating the speech signal is shown in Figure 4.

The speech signal at A is essentially a discrete convolution of the excitation pulse and the vocal tract impulse response. At point B, we have the product of the two DFT's. By taking the logarithm of the transforms of the excitation and the vocal tract impulse response, the two have now become additive. The inverse DFT is a linear operation

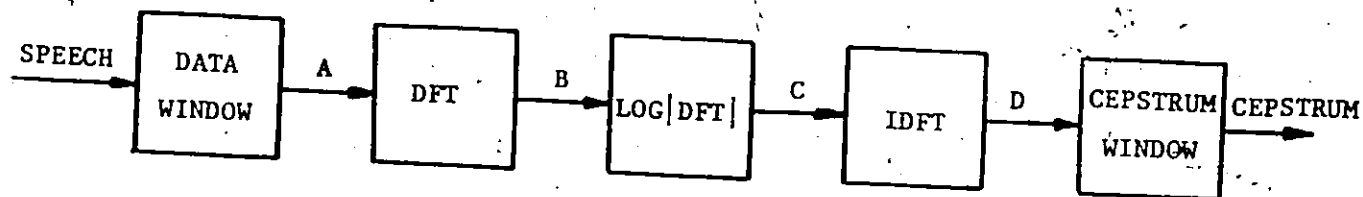


FIGURE 4 - Homomorphic Speech Analysis

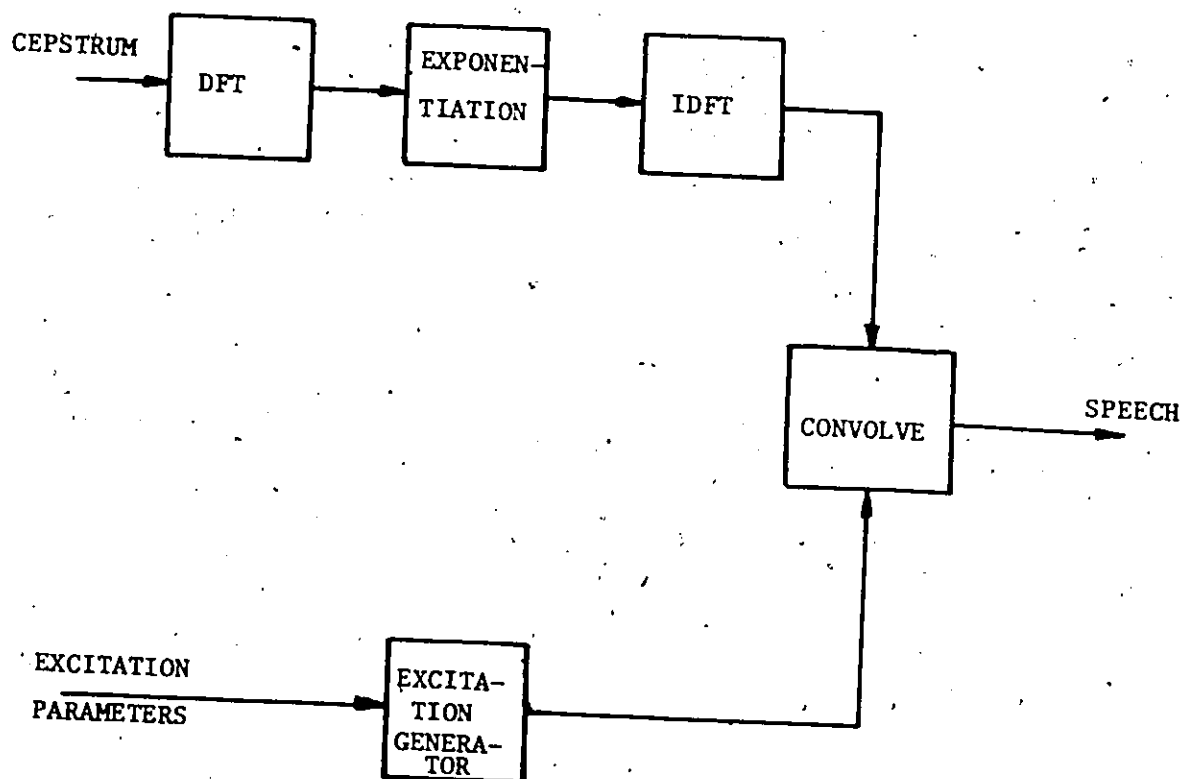


FIGURE 5 - Homomorphic Speech Synthesis

and hence the result at D is the additive cepstra of the excitation and the vocal tract components. The important feature of the cepstrum is that it consists of two additive components in which the vocal tract and the excitation components do not normally overlap. At D it is possible to extract both the pitch period for voiced sounds and also the vocal tract parameters. The synthesis part is achieved by convolving the vocal tract impulse response with the excitation function. The impulse response function is obtained by computing the DFT of the cepstrum sequence followed by an exponential transformation and the inverse DFT. This is shown in Figure 5.

1.3.iii Maximum Likelihood Vocoder[12].

In this method, an all pole model of the power spectrum of the speech signal is assumed. Zeros are omitted in the model, because their effect can be represented to any desired accuracy by a suitable number of poles[12]. The synthesis method for the maximum likelihood vocoder is shown in Figure 6. The synthesizer transfer function in the z-transform notation is

$$\begin{aligned}
 T(z) &= \frac{1}{1 + H(z)} \\
 &= \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_p z^{-p}}
 \end{aligned}
 \tag{1.1}$$

where $z^{-1} = e^{-ST}$

T is the sampling interval

a_k are the digital filter coefficients.

The maximum likelihood estimate of the a_k 's is obtained by minimizing the function of the logarithmic difference between the power

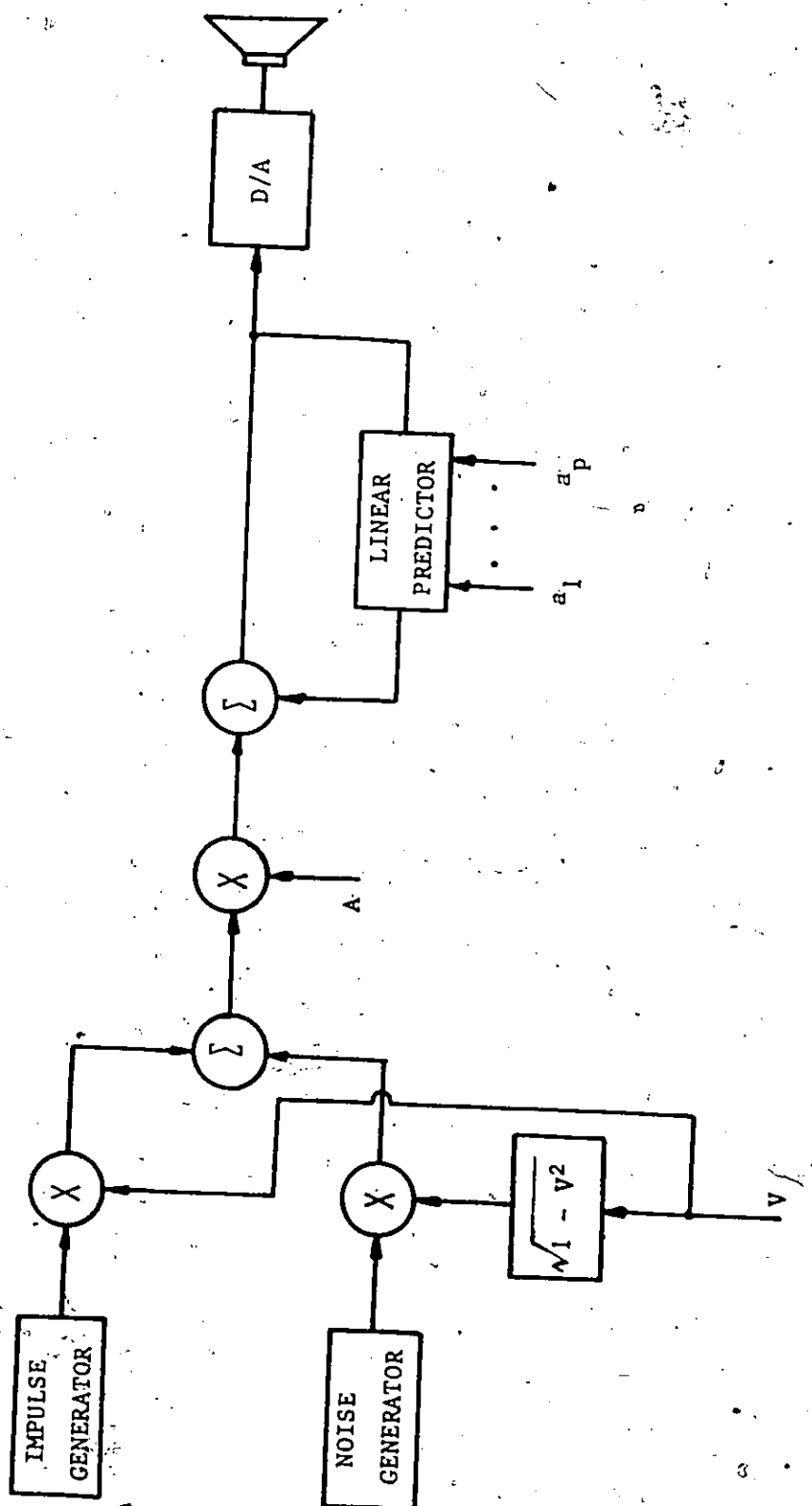


FIGURE 6 - Maximum Likelihood Vocoder Synthesis Scheme

spectrum of the filter $|T(\omega)|^2$ and the short-time power spectrum of the signal samples,

$$S(\omega) = \frac{1}{2\pi N} \left| \sum_{n=1}^N S_n e^{-jn\omega T} \right|^2 \quad (1.2)$$

where S_n is the speech sample sequence

$S(\omega)$ is the short-time power spectrum of the speech signal

T is the sampling interval.

The minimization will result in a fit which is more sensitive to spectral peaks than it is to the valleys in between the formants. The maximum likelihood of the filter coefficients is obtained from the short-time correlation function,

$$\phi_i = \frac{1}{N} \sum_{j=1}^{N-1} S_j S_{j+i} \quad i = 0, 1, \dots, p \quad (1.3)$$

by solving the set of linear equations

$$\sum_{i=1}^p \phi_{|i-j|} a_i = -\phi_j \quad j = 1, 2, \dots, p \quad (1.4)$$

The amplitude scale factor for matching the synthesized speech signal is given by

$$A^2 = \sum_{i=-p}^p A_i \phi_i \quad (1.5)$$

where $A_i = \sum_{j=0}^p a_j a_{j+|i|}$

and $a_0 = a_k = 0 \quad k > p$

The complex roots $1 + H(z)$ give the real and imaginary parts of the formant frequencies.

1.3.iv Linear Predictor[†] [13]

Speech analysis and synthesis by linear prediction, utilizes an all pole digital filter to represent the characteristics of the speech signal. The speech signal is analyzed by estimating the present speech sample as a linear combination of the previous P samples. The predicted speech sample is given by

$$\hat{s}_n = \sum_{k=1}^P a_k s_{n-k} \quad (1.6)$$

where \hat{s}_n is the predicted speech sample
 a_k are the predictor coefficients.

The predictor coefficients a_k , are obtained by minimizing the mean squared error between the actual and the predicted value of the speech signal. These coefficients are the time varying digital filter parameters whose transfer function is given by

$$H(z) = \frac{1}{1 - \sum_{k=1}^P a_k z^{-k}} \quad (1.7)$$

The excitation parameters are obtained from the actual speech waveform.

The formant frequencies can be obtained by factoring the denominator of $H(z)$. Also, the speech signal spectrum can be determined by evaluating $H(z)$ on the unit circle in the z -plane. The speech synthesis by linear prediction scheme is shown in Figure 7.

The excitation is either a quasi periodic pulse train or white noise, depending upon whether the speech is voiced or unvoiced. The gain A is made proportional to the energy level of the signal in a given analysis interval.

[†]The term "linear predictor" implies a discrete filter, whose impulse response is a least-squares approximation to the vocal tract response.

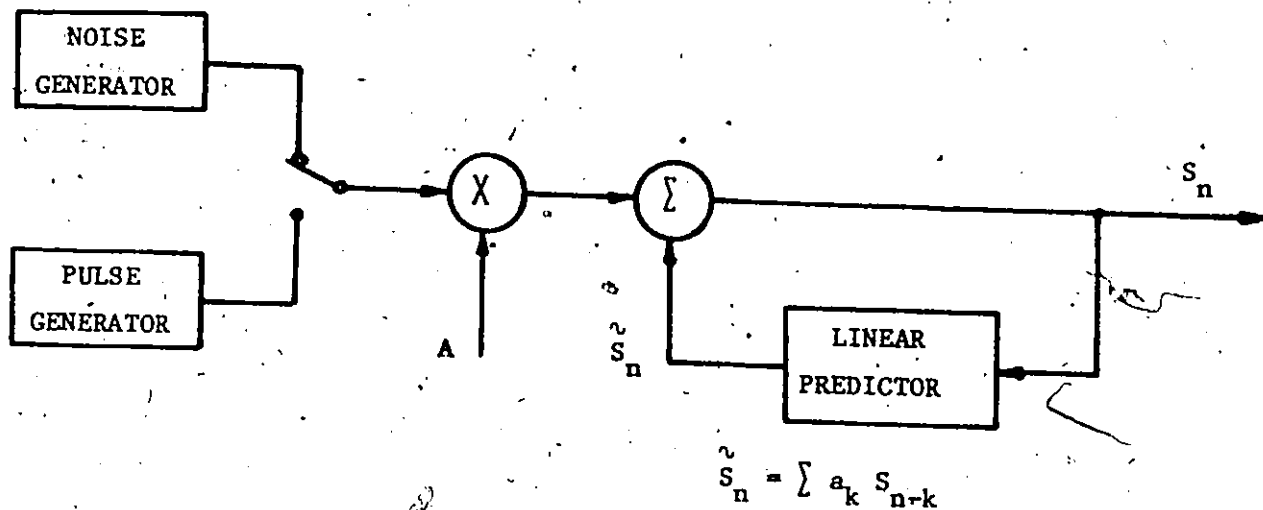


FIGURE 7 - Linear Predictor Speech Synthesis

1.3.v Digital Inverse Filtering[14]

This method was proposed as an efficient means of formant extraction from a speech signal, using the discrete linear least-squares inverse filter. For speech signal representation, the most efficient basis set is one which most closely matches the desired characteristics of the signal to be approximated. In speech analysis-synthesis, we are interested in finding the best representation to the resonances or formants of the vocal tract, impulse response. The inverse filter formulation is as follows.

Given a digital inverse filter

$$A(z) = 1 + \sum_{i=1}^M a_i z^{-i} \quad (1.8)$$

where M is specified, find the coefficients a_i , $i = 1, 2, \dots, M$, such that the energy measured at the filter output $\{y_n\}$ is minimized. If $\{x_n\}$ is the input sequence of speech data, then the filter output in the transform domain is given by

$$Y(z) = X(z) A(z) \quad (1.9)$$

If $A(z)$ could be designed so that it is precisely the inverse of $X(z)$, i.e. $A(z) = \frac{1}{X(z)}$, then $Y(z) = 1$. Thus it is seen that the inverse filter attempts to transform the input signal into white noise. When the output of the filter is as close to a constant, in the transform domain, as possible, the frequency response of the filter will be the exact inverse of the input signal. The inverse filter coefficients are obtained in the following way.

Let the approximation of S_n be given by

$$\hat{S}_n = \sum_{i=1}^M a_i S_{n-i} \quad i = 1, 2, \dots, p \quad (1.10)$$

The error in the estimate is e_n , where

$$e_n = S_n - \hat{S}_n = S_n - \sum_{i=1}^M a_i S_{n-i} \quad (1.11)$$

The portion of the signal to be analyzed is multiplied by a finite window of length N (the window can be other than rectangular). In this case we have

$$S_n = \begin{cases} \text{some sample sequence} & 0 \leq n \leq N-1 \\ 0 & n < 0 \text{ and } n \geq N \end{cases} \quad (1.12)$$

To obtain the filter coefficients we form

$$E = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} (S_n - \hat{S}_n)^2$$

$$= S_0^2 + \sum_{n=1}^{N-1+M} \left(S_n - \sum_{i=1}^M a_i S_{n-i} \right)^2 \quad (1.13)$$

Differentiating E with respect to a_j and setting the result to zero, we get

$$\frac{\partial E}{\partial a_j} = 0 = -2 \sum_{n=1}^{N-1+M} \left(S_n - \sum_{i=1}^M a_i S_{n-i} \right) S_{n-j} \quad (1.14)$$

This equation reduces to

$$\sum_{i=1}^p a_i R_{|j-i|} = R_j \quad 1 \leq j \leq p \quad (1.15)$$

where $R_j = \sum_{n=0}^{N-1+|j|} S_n S_{n+|j|}$

1.3.vii Predictive Coding of Speech Signals [13]

A means of reducing the redundancy in speech signals is by the use of predictive coding. In predictive coding, the transmitted signal is made up of the error between the actual message and its predicted value. Since the variance of the error signal is much smaller as compared to the variance of the speech signal, less bits are required for transmission purposes.

One of the earliest attempts at predictive coding was by Elias [15], who used an open loop coding scheme as shown in Figure 8. The predictor used was of the non-adaptive nature, and its parameters were obtained from a knowledge of the statistics of the set of messages to be transmitted. This type of coding scheme is not used in practice because if any noise is introduced in the system, it will be perpetuated as error in all future values of the message. Eventually the error will accumu-

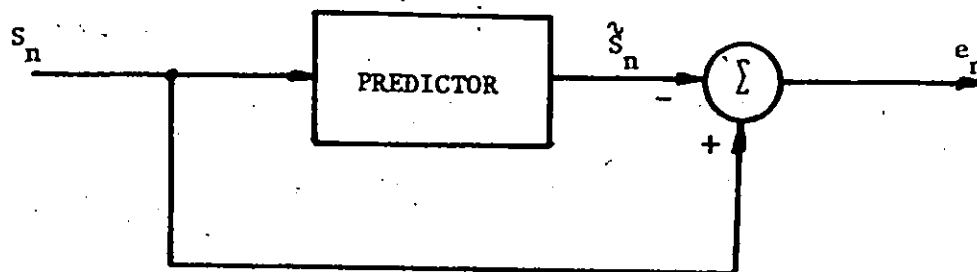


FIGURE 8 - Open Loop Predictive Coding

late to such an extent that the message will be completely lost.

A coding technique which was proposed by McDonald[16] that overcomes the error build up, is the differential pulse code modulation scheme, see Figure 9.

An error signal is formed by subtracting the estimate from the actual speech sample. The error signal is then quantized and coded before being transmitted. The predictor used by McDonald[16] was also of the non-adaptive type.

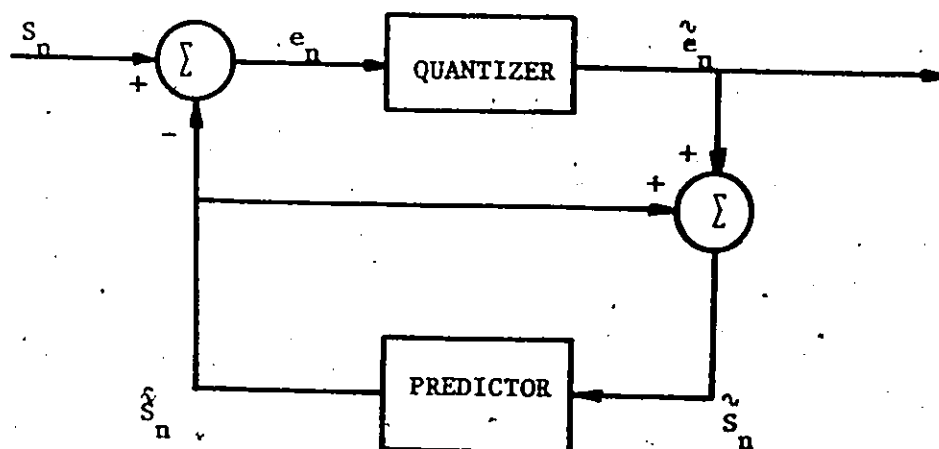


FIGURE 9 - Differential Pulse Code Modulation. (DPCM)

More recently, Atal and Schroeder [17], proposed an adaptive predictive coding scheme, which had the advantages of the DPCM, in which the predictor model was updated periodically. The signal redundancy was reduced in two stages: first, by a predictor that removes the quasi periodic nature of the signal, and second by a predictor that removes formant information from the spectral envelope. The first predictor is simply a gain and a delay adjustment, and the second is a linear combination of the past p values of the speech signal. The predictor coding scheme is shown in Figure 10. The overall transfer

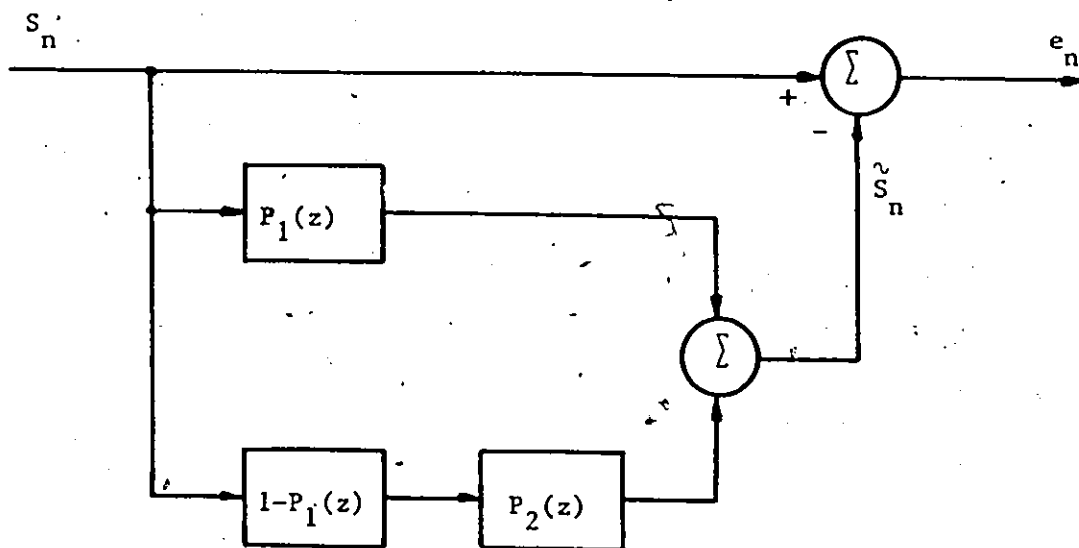


FIGURE 10 - 2-Stage Adaptive Coding

$$P_1(z) = \alpha z^{-k}$$

$$P_2(z) = \sum_{j=1}^N a_j z^{-j}$$

where α is a constant

N is the number of predictor parameters

a_j are the predictor parameters.

function of the adaptive predictor is given by

$$P(z) = [P_1(z) + P_2(z)][1 - P_1(z)] \quad (1.16)$$

The predictor given by equation (16) was used in the differential pulse code modulation scheme, using a 1 bit quantizer for the error signal. This scheme is shown in Figure 11. The quantizer step size Q , is chosen according to the following criterion,

$$Q = \frac{1}{N} \sum_{n=1}^N |e_n| \quad (1.17)$$

where e_n is the error in the estimate at the n th sampling instant and N is the number of samples over which Q is constant.

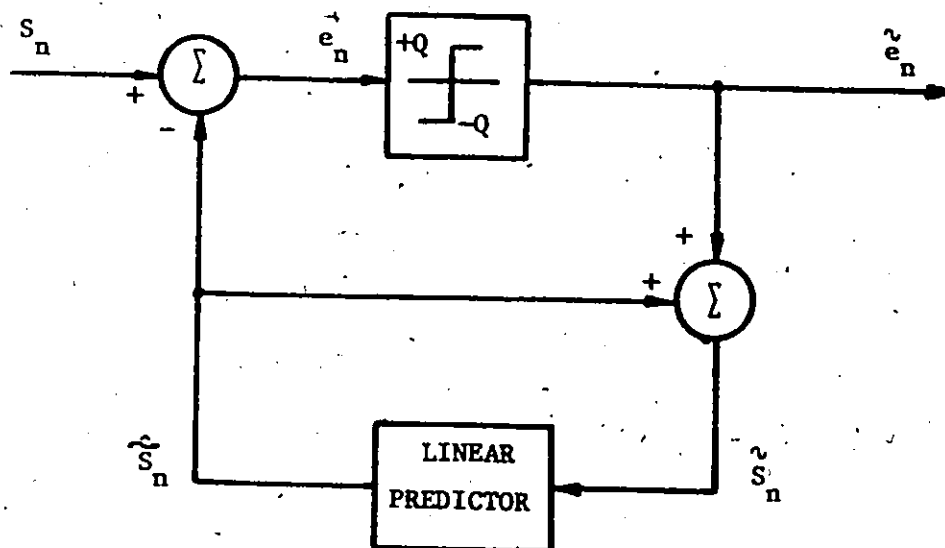


FIGURE 11 - Adaptive Predictive Coding

The results quoted for this coding scheme show that a bit rate of 9600 bits/sec. is possible for speech sampled at 6.667 kHz rate.

The criteria for obtaining the optimum quantizer step size as shown in equation (1.17) will not work for the coding scheme of Figure 11. The error signal e_n is a function of the quantizer level because of the feedback loop around the quantizer. The determination of the optimum quantizer step size is not possible, a sub-optimal solution may be obtained. Since the error is a function of the quantizer level, this will result in a multi-modal optimization problem. Even with the use of optimizing algorithms, such as Rosenbrok's steepest descent[18], or Fletcher and Powell's conjugate gradient method[19], there is no guarantee that a global optimum will be found. Often these techniques will get trapped in a local minima, that may or may not give acceptable results. Both of the above mentioned optimizing algorithms were tried, in each case, the resulting minima was approached very slowly. The slow convergence combined with the uncertainty of the solution render the scheme of Atal and Schroeder unfeasible.

1.4. Problem Statement

The aims of this project were,

1. To critically analyze the existing speech analysis-synthesis methods and recommend suitable techniques to improve the quality of synthetic speech. In particular the investigation was to cover
 - i) the performance of the linear predictor and the different realization schemes.
 - ii) Determine a suitable excitation function.

- iii) To examine the effects of adding zeros to the model of the linear predictor.
2. To develop a suitable criteria for the choice of quantizer levels in adaptive predictive coding method, and to develop new coding schemes for efficient voice transmission.
3. To examine the application of Fast Fourier Transform for on-line synthesis and coding of speech.

In particular, the speed of operation was to be critically analyzed.

1.5. Thesis Organization

The introduction covers a brief review of some past and present techniques in speech synthesis and coding, especially those which use the linear predictor.

In Chapter II, a number of time domain techniques are analyzed and their implementation is discussed in detail. In particular, the different ways of implementing the linear predictor and the problems associated with each scheme is indicated.

Chapter III describes the various coding schemes that have been proposed recently, one of these is the differential pulse code modulation. The implementation of this scheme is shown to be impractical due to the problem of obtaining acceptable results under all conditions.

Chapter IV describes a frequency domain technique for speech synthesis and coding, which is computationally very efficient, with a bandwidth compression of about 5:1 over equivalent PCM. When used in a 1 bit coding scheme, the bit rate requirement for speech transmission is about 14,200 bits/sec.

Chapter V describes the implementation of speech synthesis and coding schemes using the linear predictor on the Digital Equipment Corporation PDP-8/I computer.

Chapter VI contains the conclusions of the speech synthesis and coding schemes that were implemented on the IBM-360/65 and the PDP-8/I computers.

CHAPTER II

TIME DOMAIN SPEECH SYNTHESIS

2.1 Linear Predictor

The linear predictor is assumed to be represented by an all pole digital filter, whose transfer function is given by

$$H(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (2.1)$$

where a_k , $k = 1, 2, \dots, p$ are the predictor parameters and a_0 has been normalized to unity.

The above model has been used for speech production in conjunction with an excitation source, which is either a pulse generator with adjustable amplitude and period or a white noise generator depending upon whether the speech sound is voiced or unvoiced. The speech sample at the n th sampling instant is estimated as

$$\hat{S}_n = \sum_{k=1}^p a_k S_{n-k} + y_n \quad (2.2)$$

where a_k are the predictor parameters

y_n is the excitation function.

The predictor parameters are obtained by minimizing the mean squared error between the actual speech sample S_n and its estimate \hat{S}_n as,

$$\begin{aligned} e_n &= S_n - \hat{S}_n \\ &= S_n - \sum_{k=1}^p a_k S_{n-k} \end{aligned} \quad (2.3)$$

The sum of the error squared in a given analysis interval is given by

$$E = \sum_n e_n^2 = \sum_n \left(S_n - \sum_{k=1}^p a_k S_{n-k} \right)^2 \quad (2.4)$$

The optimum predictor parameters a_k , are obtained by differentiating E , with respect to a_j and setting the result to zero, one then obtains,

$$\frac{\partial E}{\partial a_j} = -2 \sum_n \left(S_n - \sum_{k=1}^p a_k S_{n-k} \right) S_{n-j} = 0 \quad (2.5)$$

rearranging

$$\sum_n \sum_{k=1}^p a_k S_{n-j} S_{n-k} = \sum_n S_n S_{n-j}, \quad j = 1, 2, \dots, p \quad (2.6)$$

or in matrix notation

$$\Phi a = \psi$$

where $\phi_{jk} = \sum_n S_{n-j} S_{n-k}$

$$\psi_j = \sum_n S_n S_{n-j} \quad (2.7)$$

The solution of the system of linear equations (2.7) yields the optimum (in the mean squared error sense) parameters a_k , $k=1,2,\dots,p$. Typically the number of parameters p is chosen to be 12 for speech filtered by a low pass filter with a 5 kHz cut-off, and sampled at 10 kHz rate.

The above method, while basically simple and elegant, has however, the following limitations.

1. The mean squared error criterion may result in the poles of the predictor going outside the unit circle, causing instability problems.
2. The finite precision of the speech samples can also cause the model to be unstable.
3. The assumption that the voiced excitation is of an impulse nature is questionable.

4. The iterative method of speech reconstruction in a finite-precision machine is subject to quantization and truncation errors that can seriously degrade the overall quality of the reconstructed speech.

2.1.1. Evaluation of Predictor Parameters

The predictor parameters are evaluated by solving the system of equations (2.7). The ϕ matrix is a $p \times p$ symmetric matrix since $\phi_{jk} = \phi_{kj}$. Thus in setting up the elements of this matrix it is sufficient to form the upper triangle and the leading diagonal, the lower triangle follows directly. The number of multiplications required to set up the ϕ matrix is $Np \left(\frac{p}{2} + 1 \right)$, where N is the number of samples over which the model is being evaluated. Typically $N = 80$ and $p = 12$, which requires 6720 multiplications.

The number of multiplications can be reduced if one makes the assumption that the speech signal is stationary in the given analysis interval. In such a case

$$\phi_{jk} = \phi_{|j-k|,0} \quad (2.8)$$

$$\text{and } \phi_{j+1,k+1} = \phi_{jk} \quad (2.9)$$

By making the above assumption, the number of multiplications required to set up the ϕ matrix now becomes Np , which is only 960 for the same example as before. By making use of equation (2.8) only the elements of the first row need be evaluated, the rest follow directly from equation (2.9).

One need not make the assumption that the speech sample is stationary to reduce the number of multiplications. Let

$$\phi_{jk} = \sum_{n=L}^N S_{n-j} S_{n-k} \quad (2.10)$$

where $N-L$ is the number of samples in the analysis interval

and $j, k = 1, 2, \dots, p$

Now let $L = p + 1$ (2.11)

and $S_n = 0 \quad n < 0, \quad N > 0$

$$\text{Now } \phi_{j+1, k+1} = \sum_{n=L}^N S_{n-j-1} S_{n-k-1} \quad (2.12)$$

$$= \sum_{n=L-1}^{N-1} S_{n-j} S_{n-k}$$

Equation (2.12) can be rearranged in the following way

$$\phi_{j+1, k+1} = \phi_{jk} + S_{L-j} S_{L-k} - S_{N+1-j} S_{N+1-k} \quad (2.13)$$

Equation (2.13) shows that to set up the ϕ matrix we need to form the first row using Np multiplications and the other elements of the upper triangle and the leading diagonal only require two multiplications each. Thus the total number of multiplications required in this case is $Np + \frac{p}{2}(p-1)$. Using the same example as before we would require 1026 multiplications. The number of multiplications required to set up the system of equations (2.7) for each of the three methods is given below.

$$\text{DIRECT } Np \left(\frac{p}{2} + 1 \right) + Np = Np \left[\frac{p}{2} + 2 \right]$$

$$\text{CYCLIC } Np + Np = 2Np$$

$$\text{SYMMETRY } Np + \frac{p}{2}(p-1) + Np = 2Np + \frac{p}{2}(p-1)$$

The possible savings in setting up the system of equations (2.7)

for the case $N = 80$ samples and $P = 8$ to 16 is shown in Table 1.

TABLE 1

	p=8	p=10	p=12	p=14	p=16
DIRECT	3840	5600	7680	10080	12800
CYCLIC	1280	1600	1920	2240	2560
SYMMETRY	1308	1645	1986	2331	2680

Number of Multiplications Required to Form $\phi a = \psi$, for
 $N = 80$ samples, using the three different methods.

The solution of the system of equation (2.7) was carried out using Gaussian elimination with maximum pivot strategy. Once the system of equations is formed, the actual solution for a 12th order system, takes 4 to 5 msec. on the IBM 360/65 computer. The square root method[20] of solving linear systems of equations was also implemented, giving computational times of about 5 msec.

However, the square root method gave more accurate results over those obtained by Gaussian elimination. A typical set of results obtained by the two methods is shown in Table 2. The error in the solution was defined as $\phi a - \psi = e$.

TABLE 2

SQUARE ROOT METHOD a_k 's	ERROR IN SQUARE ROOT SOLUTION	GAUSSIAN ELIMINATION METHOD a_k 's	ERROR IN GAUSSIAN SOLUTION
1.493688	- 0.000003	1.493676	- 0.000008
- 0.363467	- 0.000004	- 0.363458	- 0.000007
- 0.037537	- 0.000003	- 0.037541	- 0.000008
0.028311	- 0.000002	0.028322	- 0.000006
- 0.078749	- 0.000001	- 0.078761	- 0.000005
- 0.165538	- 0.000001	- 0.165553	- 0.000005
0.016301	- 0.000001	0.016334	- 0.000005
0.0180350	- 0.000000	0.180345	- 0.000004
0.017452	- 0.000000	0.017426	- 0.000003
- 0.174168	- 0.000000	- 0.174116	- 0.000002
- 0.077913	- 0.000000	- 0.077980	- 0.000002
0.139857	- 0.000000	0.139889	- 0.000001

Predictor Parameters Obtained By Using Two Different Techniques
For Solving $\phi a = \psi$. The Error In The Solution $E \Delta \phi a - \psi$.

The error in the evaluation of the predictor parameters was relatively small, but its effect on the pole locations movements was checked, to ensure that the shift, if any, was minimal. The roots of the polynomial

$$1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_p z^{-p} = 0$$

were found using the Newton-Raphson algorithm, and are shown in Table 3.

TABLE 3

SQUARE ROOT METHOD POLES		GAUSSIAN ELIMINATION METHOD POLES	
R_e	I_m	R_e	I_m
0.951236	0.153709	0.951241	0.153719
0.951236	- 0.153709	0.951241	- 0.153719
- 0.759653	0.0	- 0.759713	0.0
- 0.377018	- 0.708345	- 0.376988	- 0.708356
- 0.377018	0.708345	- 0.376988	0.708356
0.906886	0.0	0.906880	0.0
0.610830	- 0.560510	0.610841	0.560515
0.610830	0.560510	0.610841	- 0.560515
- 0.655227	0.399102	- 0.655248	- 0.399155
- 0.655227	- 0.399102	- 0.655248	0.399155
0.143407	0.904894	0.143409	- 0.904899
0.143407	- 0.904894	0.143409	0.904899

Pole Locations Of The Linear Predictor Obtained By The Two Different Methods.

The results shown in Table 3 indicate that the poles are not highly sensitive to small errors in the predictor parameters.

Based on this, both methods, the Gaussian elimination and the square root method can be used for solving the predictor parameters.

The frequency response of the linear predictor was evaluated for the two solutions, and is shown in Figure 12. Examination of Figure 12, will reveal that the frequency response for the two solutions are identical, and hence there is very little to choose between two techniques of solving linear systems of equations.

2.2. Pitch Period Extraction

A number of techniques [21-25], have been proposed for the determination of the pitch period of a segment of speech waveform. Most speech synthesis schemes make use of the pitch period, and a great deal of work has been done in obtaining more accurate estimates of the pitch period. Some of the more common techniques of pitch determination will be discussed together with their advantages and disadvantages, in the following sections.

2.2.1. Zero Crossing Rate

This was one of the earliest attempts at pitch period determination. The speech signal was passed through a low pass filter and the rate or the distance between zero crossings was measured. This scheme will work satisfactorily in some cases, but in cases where the fundamental component of the speech signal is weak, the process will break down [26]. Also, in some cases it is the formant structure in the speech signal that causes the speech waveform to cross zero. When this happens it is easy to estimate the pitch period to be twice its true value, or some integer multiple.

2.1.11. The Cepstrum

A computationally efficient technique of pitch period extraction

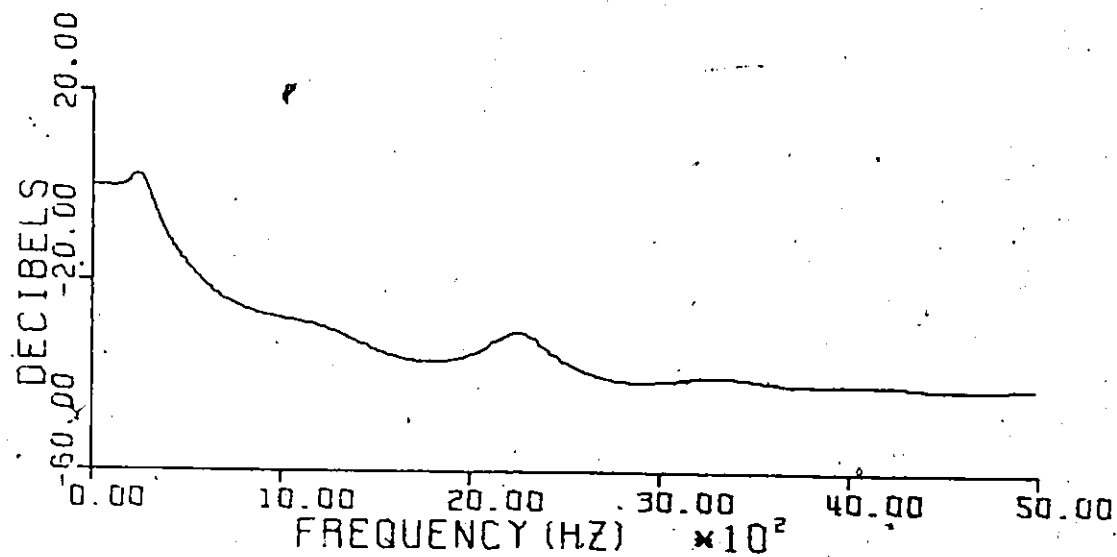


FIGURE 12 - Linear Predictor Frequency Response

(a) Square root method solution

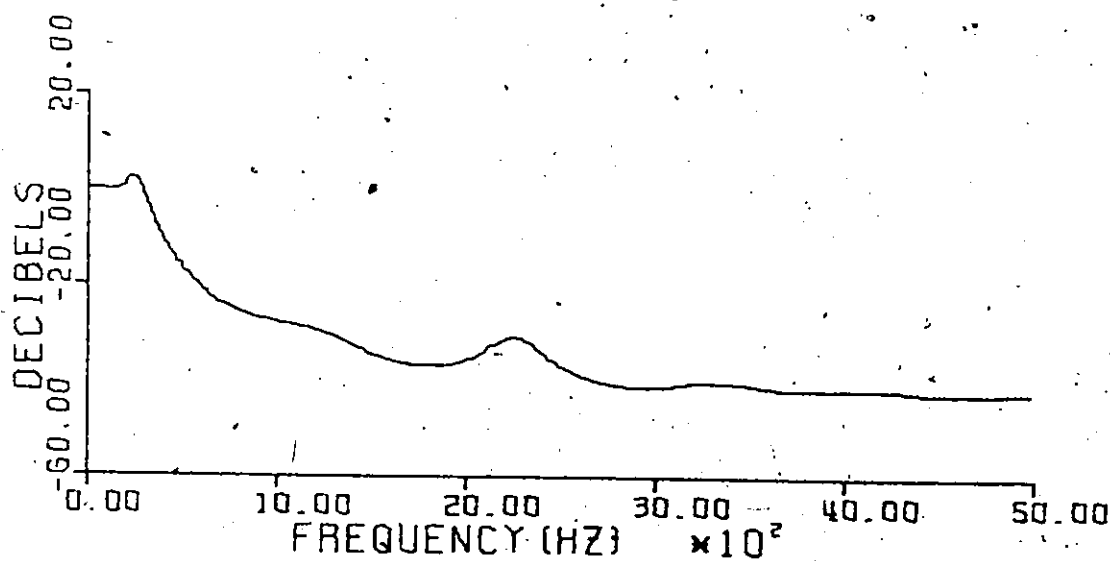


FIGURE 12 - Linear Predictor Frequency Response

(b) Gaussian elimination method solution

is by the use of the cepstrum [22]. This technique was made possible by the discovery of the Fast Fourier Transform algorithm of Cooley-Tukey[5]. The basic operations involved in obtaining the cepstrum are shown in Figure 13.

The cepstrum produces two peaks in the transformed signal. The first peak is at the origin and the second peak is usually at about 8 msec. from the origin for voiced speech. It is the second peak that is of most importance and is an indication of the pitch period of voiced sounds. The lack of it, indicates that the speech sound is unvoiced.

The cepstrum method can best be explained by considering the following simplified model of speech production. For voiced sounds, the vocal tract is effectively excited by a quasi-periodic pulse train. The resultant speech signal can be thought of as a convolution of the pulse train with the impulse response of the vocal tract, see Figure 14.

The output of this system is given by

$$S(t) = \int_{-\infty}^t h(\tau) \times (t - \tau) d\tau \quad (2.8)$$

where $h(\tau)$ is the impulse response of the vocal tract

$x(t)$ is the excitation function.

The convolution integral of equation (2.8) is equivalent to multiplication in the frequency domain,

$$S(\omega) = H(\omega) \cdot X(\omega) \quad (2.9)$$

The problem now is to separate out the effects of the excitation function in the overall speech signal spectrum. By taking the logarithm

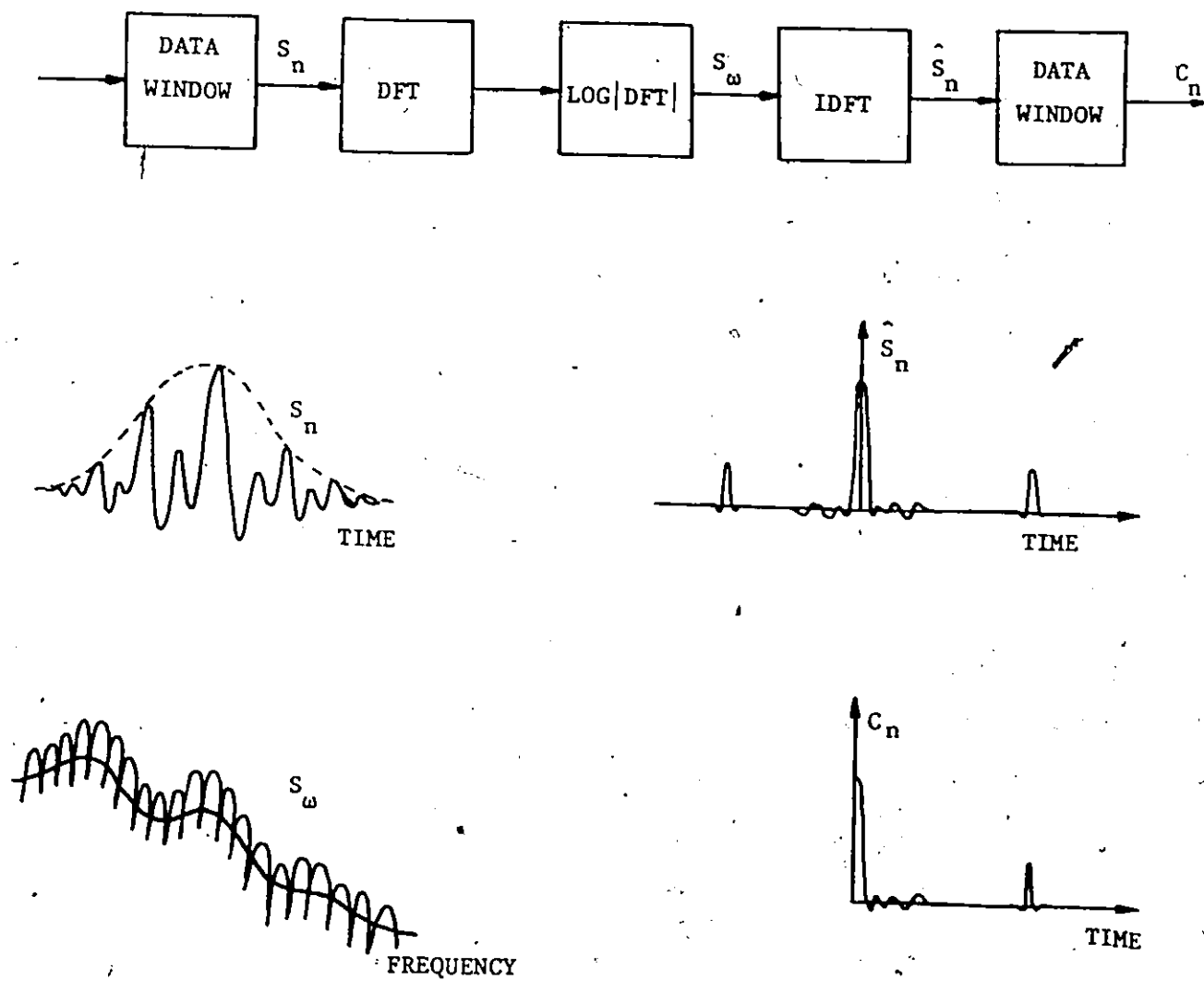
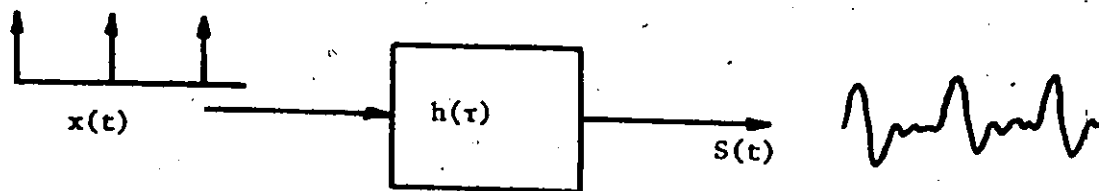


FIGURE 13 - Operations Involved In Obtaining The Cepstrum



where $h(\tau)$ is the vocal tract impulse response
 $x(t)$ is the glottal excitation
 $S(t)$ is the resultant speech signal

FIGURE 14 - Simplified Model Of The Vocal Tract Impulse Response

of both sides of equation (2.9), we obtain,

$$\log[S(\omega)] = \log[H(\omega)] + \log[X(\omega)] \quad (2.10)$$

The effects of the vocal cords and that of the vocal tract are now additive, but still not separated. By examining the log-spectrum plot of Figure 13, the fundamental frequency (pitch frequency) is represented by many relatively closely spaced peaks. The formants, on the other hand, are distinguished by the slowly varying broad peaks. By performing an inverse FFT on the log-spectrum, the resulting cepstrum has a peak at the higher quefrequency (time), which corresponds to the fundamental period. The formant structure is limited to low

quefrequencies because of their much slower rate of change in actual speech.

In the implementation of the cepstrum, the original speech signal was multiplied by a Hamming data window. Other type of windows can be used, their operation and the relative merits are discussed in detail elsewhere[27]. The sample sequence which corresponds to the cepstrum, shows that the cepstral peaks decrease with increasing quefrehuecy (time). To compensate for the decrease, a weighting function[22], was applied over the range 0-12 msec. The cepstrum is multiplied by the linear weighting of 1 at 0 msec. and 5 at 12 msec. Any increase in the weighting, may result in picking a wrong peak at around 12 msec. due to the weighting. A typical plot of the cepstrum is shown in Figure 15.

As shown in Figure 15, the peak corresponding to the pitch period is not very pronounced. It is for this reason that the cepstrum was not used for pitch extraction, and computationally, the normalized correlation was found to be as efficient.

2.2.iii. Normalized Correlation

The pitch period of a segment of speech data is obtained from the normalized short time autocorrelation function as follows [21],

Let S_n represent a speech sample which is periodic with period k . Then,

$$S_{n+k} = \alpha S_n \quad (2.11)$$

where α and k are parameters that have to be determined by minimizing the mean squared error criterion, which is given by

$$E = \sum_n \left(S_{n+k} - \alpha S_n \right)^2 \quad (2.12)$$

Differentiating equation (2.12) with respect to α and equating to zero, we obtain

$$\frac{\partial E}{\partial \alpha} = -2 \sum_n (S_{n+k} - \alpha S_n) S_n = 0 \quad (2.13)$$

rearranging

$$\alpha_{\text{opt}} = \frac{\sum_n S_n S_{n+k}}{\sum_n S_n^2} \bigg|_{k=k_{\text{opt}}} \quad (2.14)$$

The optimum value of k is found by determining the value of k that maximizes the normalized correlation coefficient $\rho(k)$ where,

$$\rho(k) = \frac{\sum_n S_n S_{n+k}}{\left(\sum_n S_n^2 \sum_n S_{n+k}^2 \right)^{1/2}} \quad (2.15)$$

The maximization of $\rho(k)$ is performed by a linear search of the computed values of $\rho(k)$ for $k = 1, 2, \dots, N/2$. The length of the sample sequence N is usually about 2 to 3 pitch periods. A typical plot of the normalized correlation is shown in Figure 15.

Of the three techniques described the cepstrum and the normalized correlation were both implemented. However, the results indicate that the normalized correlation gave slightly better results. Also the peak picking is made easier in the case of the normalized correlation due to the very pronounced peak. It was found that more reliable results could be obtained with the normalized correlation, if the speech sequence was passed through a filter prior to the pitch extraction. The smoothing filter used is a cosine filter [36], whose input-output relationship is given by

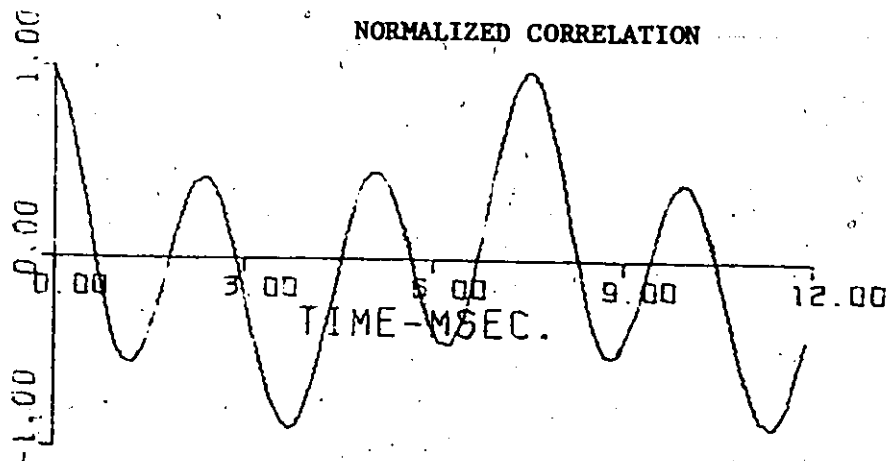
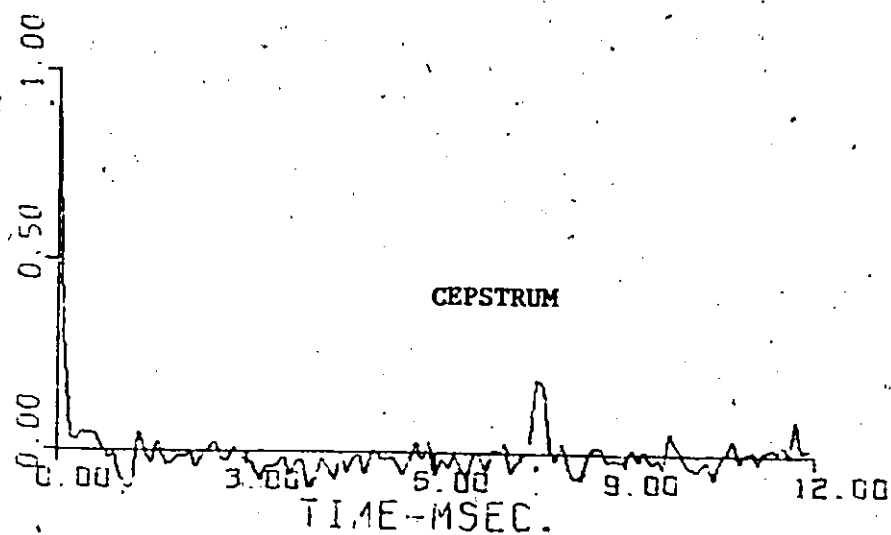


FIGURE 15 - Cepstrum And Normalized Correlation Plots

$$y_n = (S_{n-1} + 4S_n + S_{n+1})/6$$

(2.16)

where y_n is the output of the filter
and S_n is the input to the filter.

The normalized correlation coefficient is then given by

$$\rho(k) = \frac{\sum_{n=1}^{N/2} y_n y_{n+k}}{\left[\sum_{n=1}^{N/2} y_n^2 \sum_{n=1}^{N/2} y_{n+k}^2 \right]^{1/2}} \quad k = 1, 2, \dots, N/2 \quad (2.17)$$

The normalized correlation is evaluated using only $N/2$ multiplications for each summation in equation (2.17). The search for the optimum value of k was constrained to lie between $k = 50$ and $k = 120$. This corresponds to 5 and 12 msec respectively. Of the speech samples tested, the imposed constraint did not affect the accuracy of the pitch detection.

2.3. Linear Predictor Realization Schemes

The linear predictor can be thought of as an all pole digital filter. The three basic forms for realizing linear digital filters are the direct[†], the cascade and parallel. The advantages and the disadvantages of each realization scheme, have been treated in some detail [28-30]. The direct and the parallel realization scheme were implemented, the results for each case are discussed below.

2.3.1. Direct Form

The realization of the linear predictor in the direct form is shown in Figure 16.

The direct form of realization is sensitive to the truncation and roundoff errors [28-30] in the predictor coefficients, and in some cases

[†] The direct form is equivalent to the canonic form for an all pole filter.

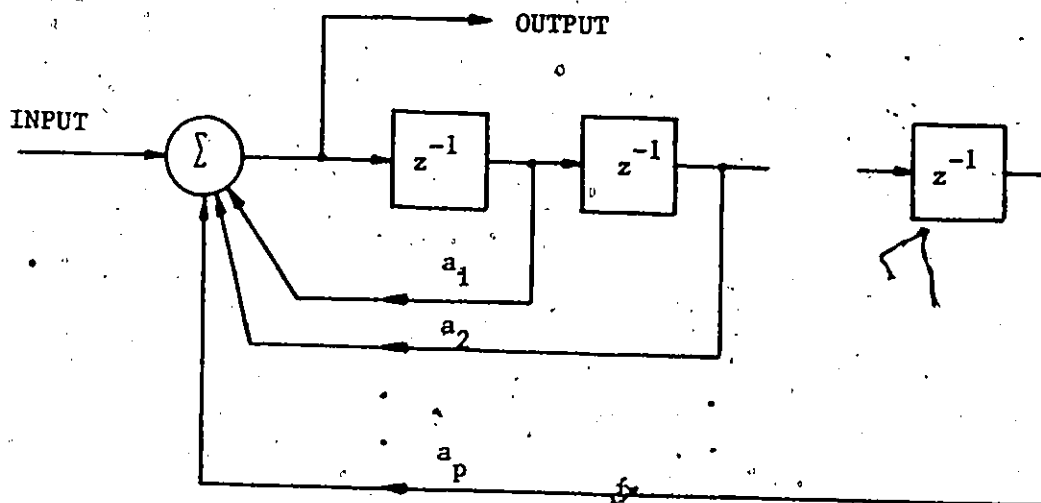


FIGURE 16 - Direct Realization Form Of The Linear Predictor

it can cause stability problems. The error build up[†] in the predictor estimate is very rapid especially when realized in the direct form. The error build up is caused by the following.

1. Errors in the assumed model
2. Finite precision of the data
3. Truncation and roundoff error due to the finite word length in the calculations.

The linear predictor estimates the speech samples recursively, and any error in a previous estimate will affect the subsequent estimates.

[†] The error build up is mainly due to the effects of parameter quantization.

This can be best observed by considering the following example.

Let the error in the predictor parameters be Δa_i , and the error in the sample estimate be ΔS_n . Under these conditions the predictor model is given by

$$\hat{S}_n = \sum_{i=1}^P (a_i + \Delta a_i) \cdot (S_{n-1} + \Delta S_{n-1}) \quad (2.18)$$

multiplying and rearranging,

$$\hat{S}_n = \sum_{i=1}^P a_i S_{n-1} + \sum_{i=1}^P \{a_i \Delta S_{n-1} + \Delta a_i S_{n-1} + \Delta a_i \Delta S_{n-1}\} \quad (2.19)$$

The second summation on the RHS is the cause of the error build up. Increasing the order of the predictor will not improve the sample estimate by a large amount. The greatest improvement will most likely come from higher precision data, since this will also increase the accuracy of the model. A significant improvement can be expected from using a larger word length in the evaluation of the model and also in the estimation process.

2.3.1.(a) Linear Predictor Excitation Function

The use of an impulse for the excitation of the linear predictor may not be the best approximation to the actual excitation of the vocal tract. For voiced sounds, an air pressure is built up within the lungs and is allowed to escape through the glottal orifice into the throat cavity. The glottal volume air flow has been studied extensively [24,32], and their findings indicate that the excitation could be considered to be triangular. One of these studies has reported [32], that when a triangle is used as the excitation function, best results are obtained if the opening time is about 40% - 50% of the total open

time, see Figure 17.

The duration of the pulse τ is variable with a total open time between about 2 to 8 msec. However, when using the linear predictor with a triangular pulse as the excitation, the equivalent open time or the time over which an input is applied is less than or equal to p times the number of sampling intervals (p is the order of the predictor). It was found that a longer open time produced a heavily smoothed output. To prevent this, the open time was constrained to be less than 1 msec. The effect of using a triangle as the excitation pulse is to place zeros at $\pm j 2\pi/\tau_0$ Hz in the s -plane [21].

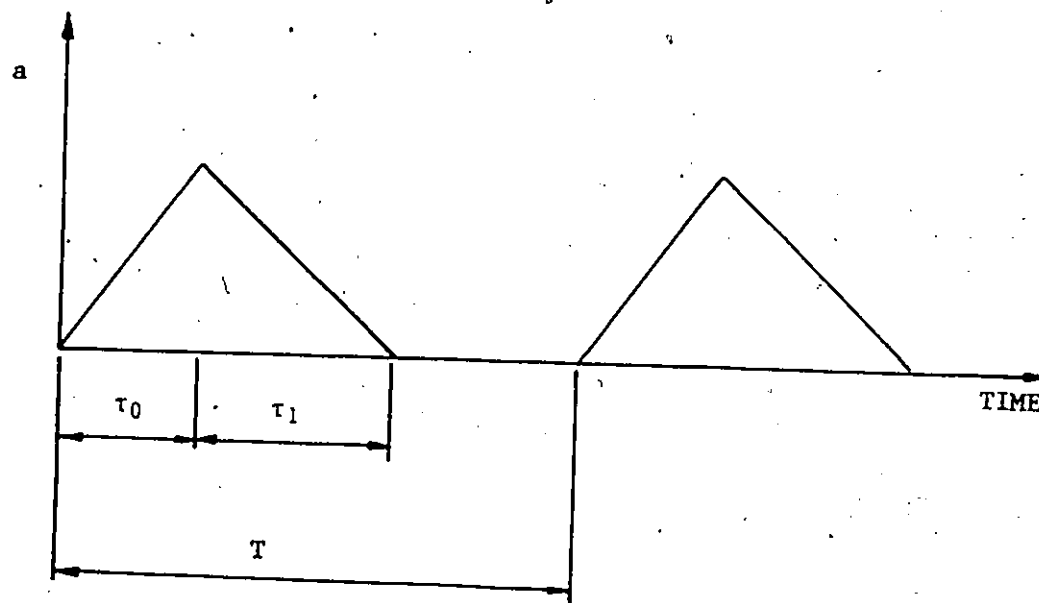
Theoretically, the input to the linear predictor could be made an impulse, and the order of the predictor would be increased to compensate for the actual pulse shape. However, it is not apparent as to the increase in the order required to adequately describe the pulse shape. Also, increasing the order of the predictor will not guarantee that the increase in the model will get assigned to the input pulse. It may well be that it will instead create an extra formant.

If the linear predictor has both poles and zeros, then the model for the speech production process is given by

$$H(z) = \frac{\sum_{j=0}^N b_j z^{-j}}{1 - \sum_{i=1}^p a_i z^{-i}} = \frac{Y(z)}{X(z)} \quad (2.21)$$

rearranging

$$Y(z) = \sum_{i=1}^p a_i z^{-i} Y(z) + \sum_{j=0}^N b_j z^{-j} X(z) \quad (2.22)$$



where T = pitch period

τ_0 = opening time

τ_1 = closing time

$\tau = \tau_0 + \tau_1$ = total open time

FIGURE 17 - Approximation To The Glottal Pulse Shape

Taking the inverse z-transform of equation (2.22) we get,

$$y(nT) = \sum_{i=1}^P a_i y(n-i)T + \sum_{j=0}^N b_j x(n-j)T \quad (2.23)$$

If we now let the input be an impulse $\delta(nT)$, then the predictor zeros b_j 's must describe the shape of the desired excitation function. Thus, if we desire a symmetrical triangle to be the excitation, which is of 4 sampling intervals in duration, see Figure 18, then

$$\begin{aligned} b_0 &= b_4 = 0 \\ b_1 &= b_3 = 1/2 \\ b_2 &= 1 \end{aligned} \quad (2.24)$$

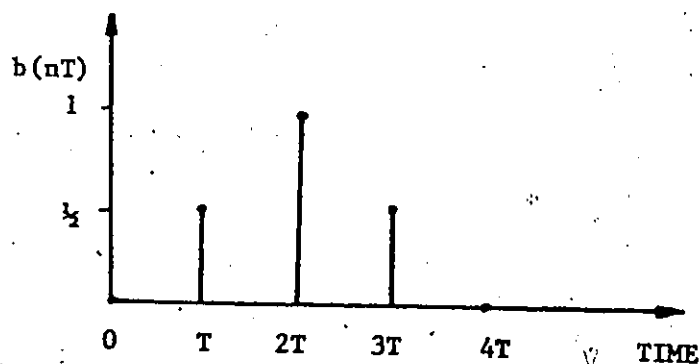


FIGURE 18 - Triangular Excitation Pulse

Impulse and triangular pulses were tried with varying pulse duration. It was found that a symmetric triangle of duration of about 10 sampling intervals gave the best results. Also, the pulse height was made propor-

tional to the energy of the signal in a given analysis interval. To ensure that the energy of the predicted signal was identical to that of the original, the following scaling procedure was found to give good results.

Let the predictor estimate be given by

$$\hat{S}_n = \sum_{i=1}^P a_i S_{n-i} \quad (2.25)$$

$$\text{and let } S_n = k \hat{S}_n \quad (2.26)$$

where k is the scaling factor to be determined.

Now equating the energies of the two waveforms, we obtain,

$$\sum_n S_n^2 = \sum_n k^2 \hat{S}_n^2 \quad (2.27)$$

rearranging

$$k = \sqrt{\frac{\sum_n S_n^2}{\sum_n \hat{S}_n^2}} \quad (2.28)$$

Figure 19 shows plots of the original waveforms for the words "NOON", "DAY", "SLEEPY", and "HOW ARE YOU". Figures 20 and 21 show the results of the impulse and triangle excitation respectively for the same four sets of data.

2.3.1.(b). Linear Predictor With Zeros

The use of the linear predictor as a model for speech production process is only an approximation. In the previous sections, the linear predictor was assumed to be an all pole digital filter. The inclusion of both poles and zeros in the predictor model is expected to improve the model performance. If the model has both poles and zeros,

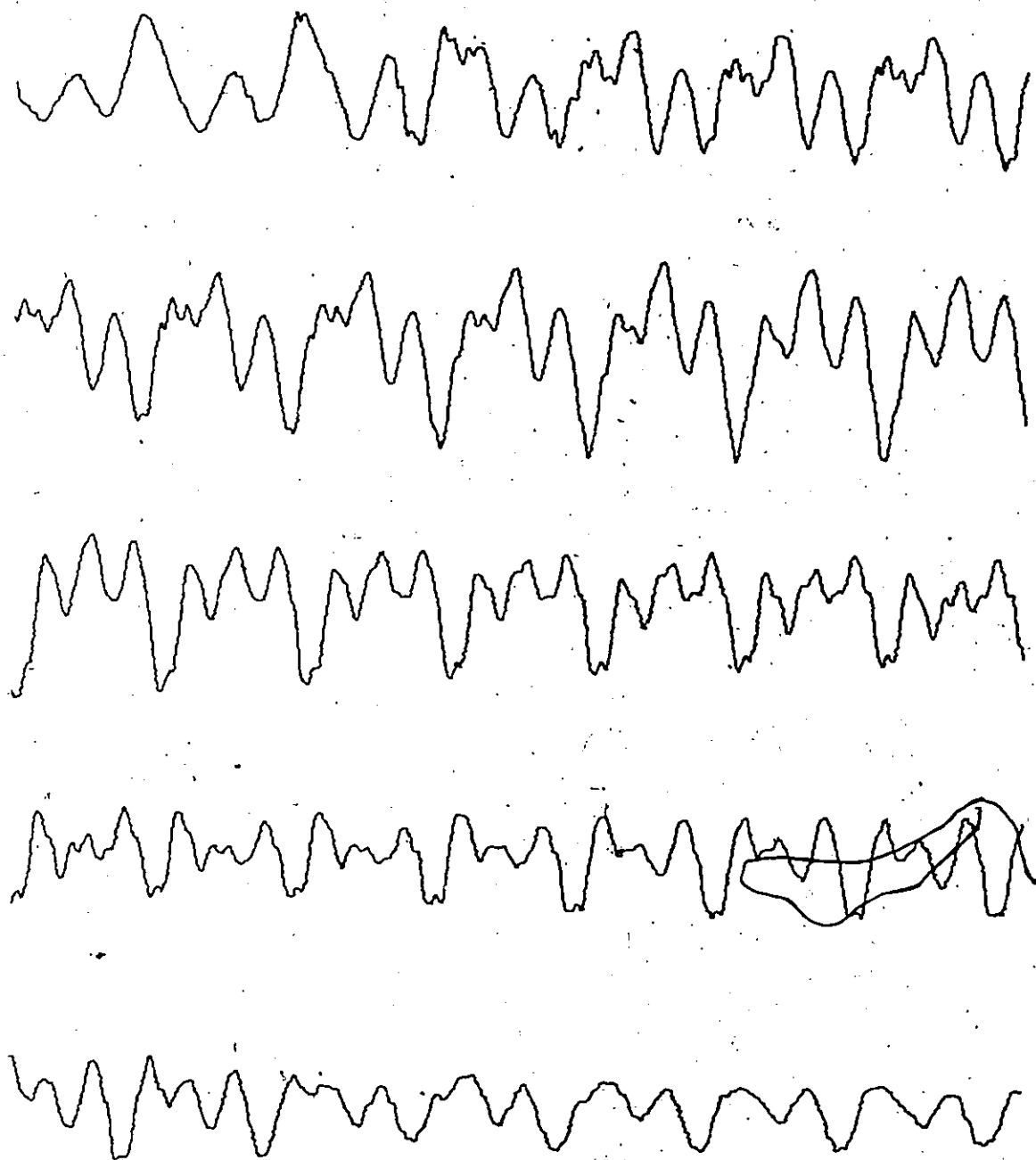


FIGURE 19 (a) - Original Waveform Of The Word - "NOON"

SCALE 0. 10 msec

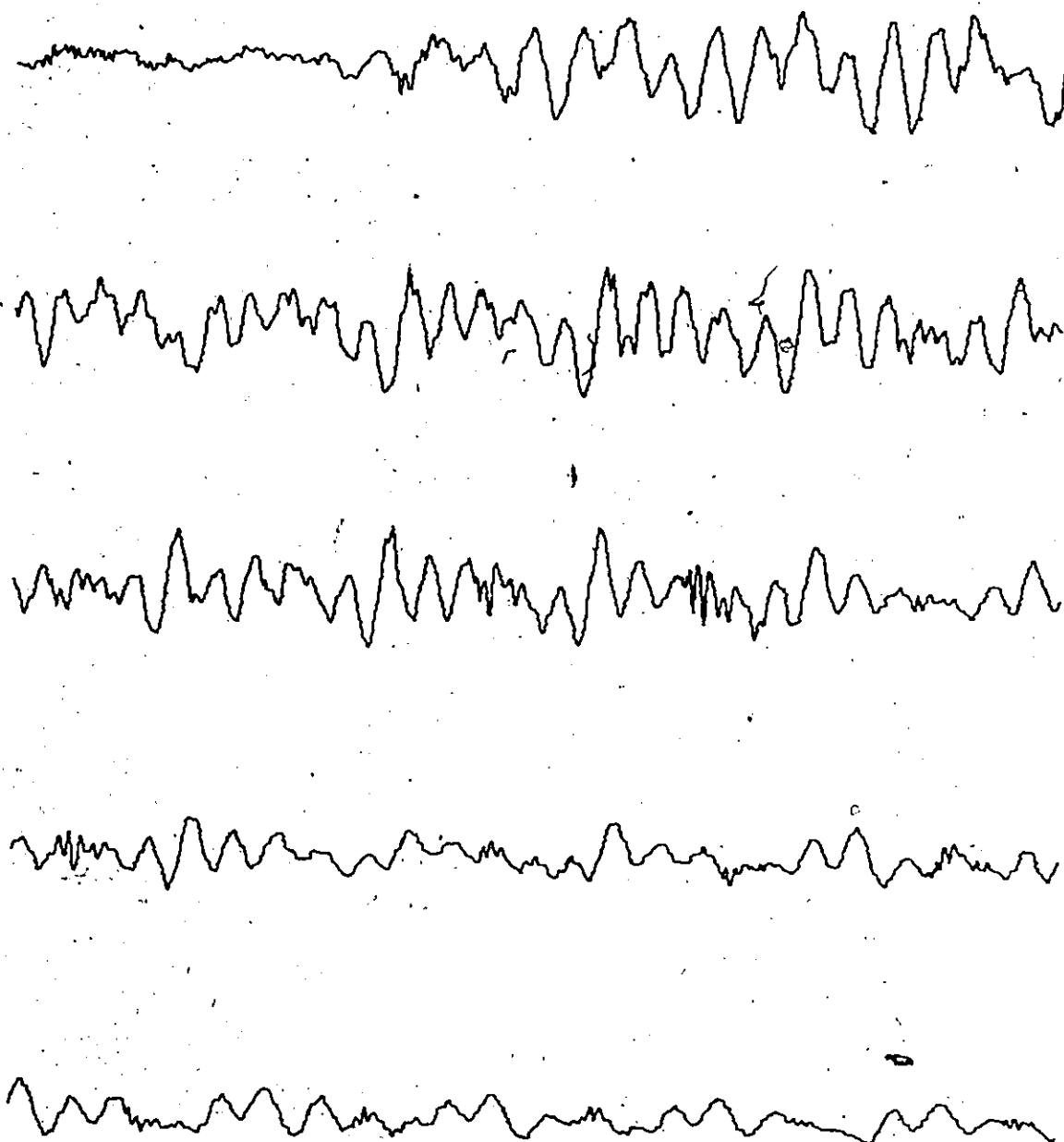


FIGURE 19 (b) - Original Waveform Of The Word - "DAY"

SCALE 0 10 msec

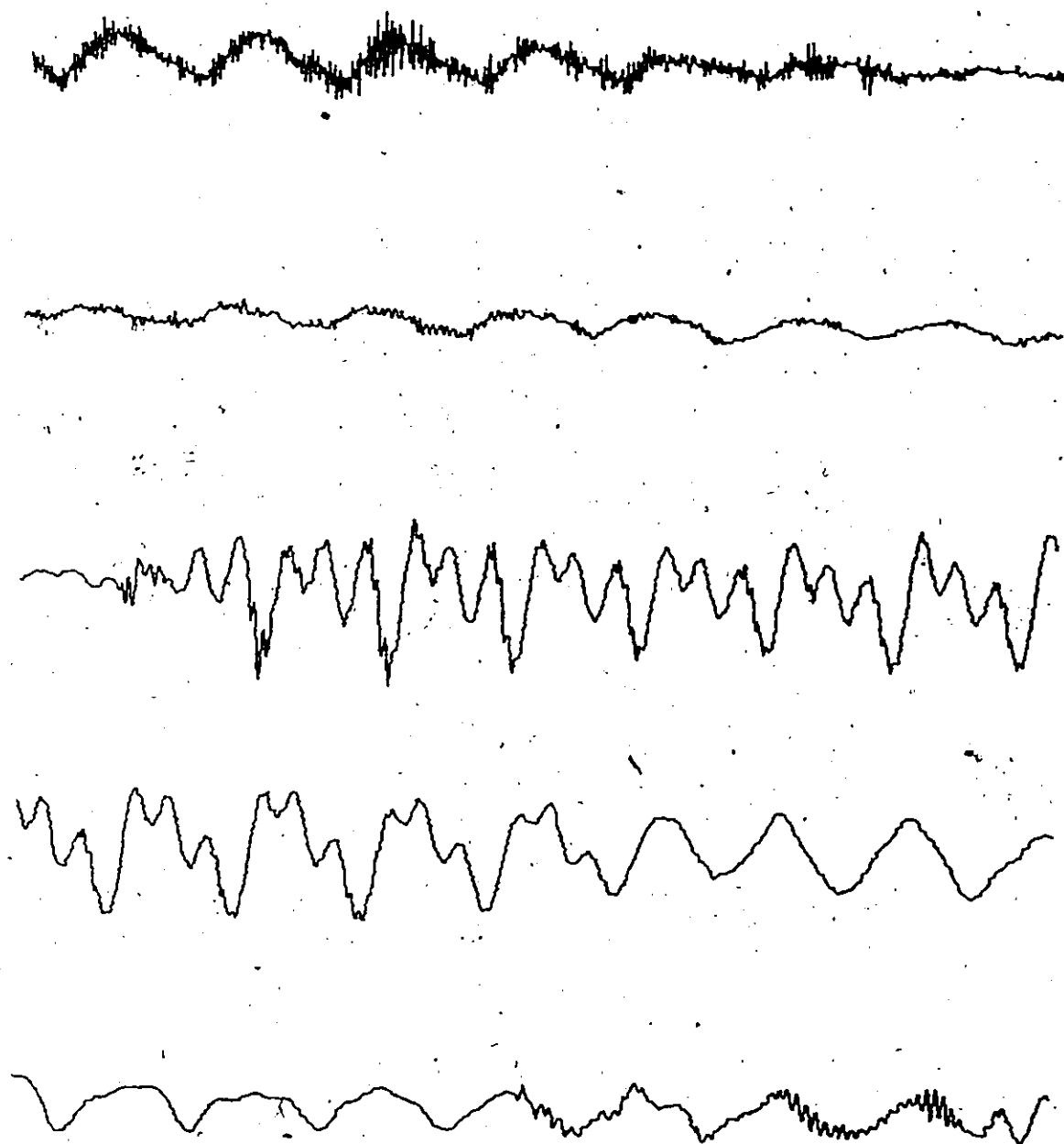


FIGURE 19 (c) - Original Waveform Of The Word - "SLEEPY"

SCALE 0  10 msec

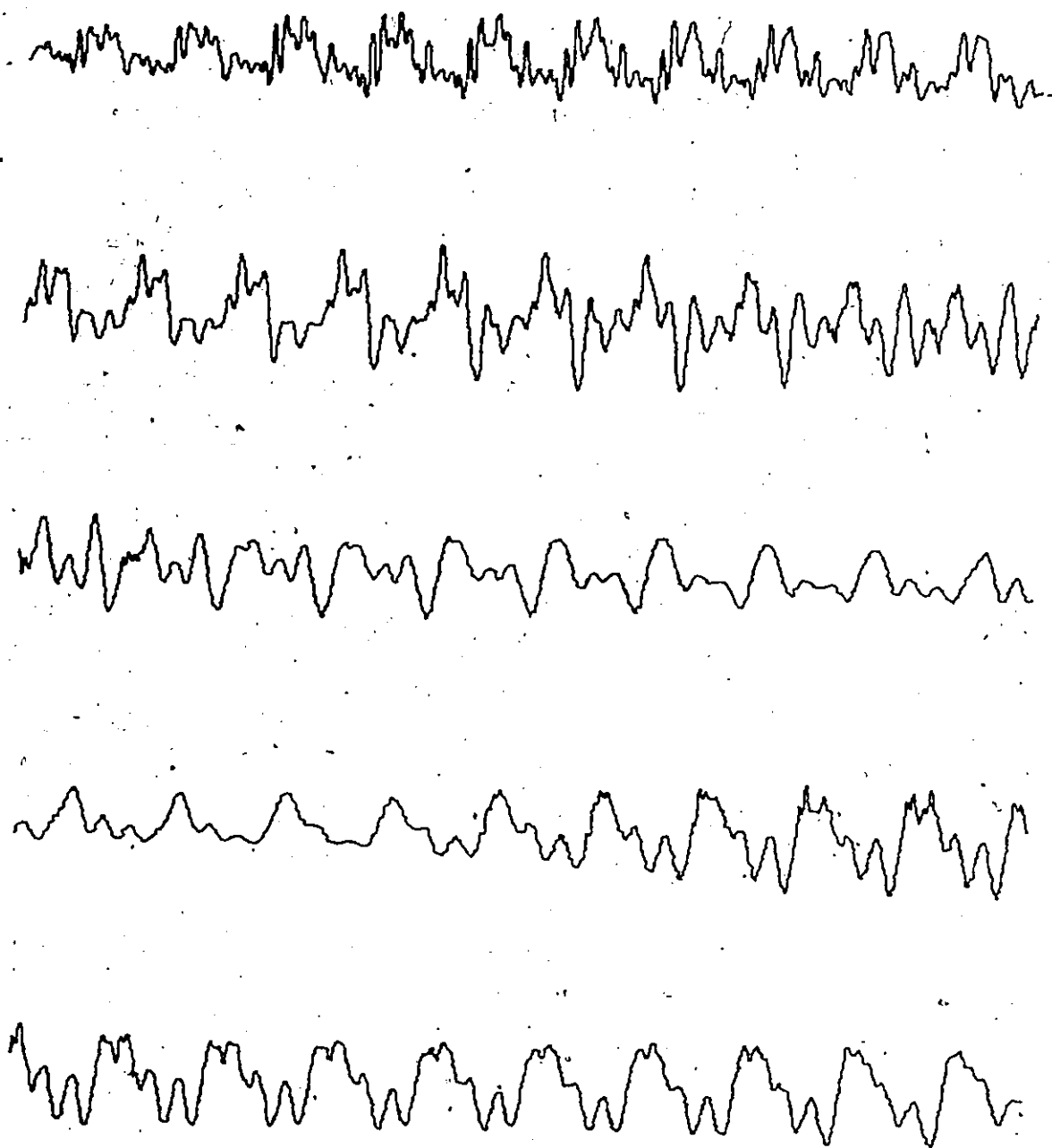



FIGURE 19 (d) - Original Waveform Of The Word - "HOW ARE YOU"

SCALE 0  10 msec

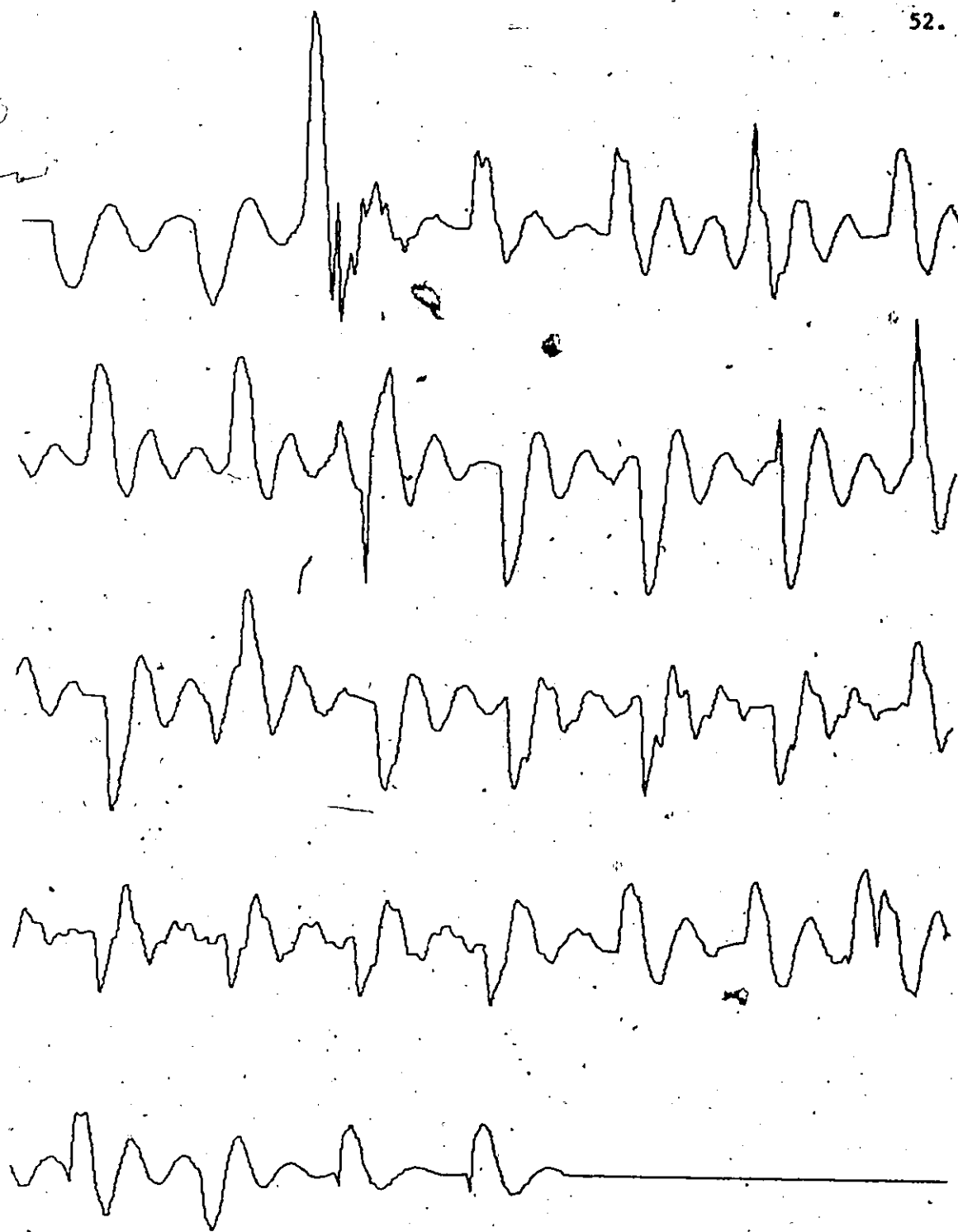


FIGURE 20 (a) - Reconstructed Waveform Using Impulse Excitation - "NOON"



FIGURE 20 (b) - Reconstructed Waveform Using Impulse Excitation - "DAY"

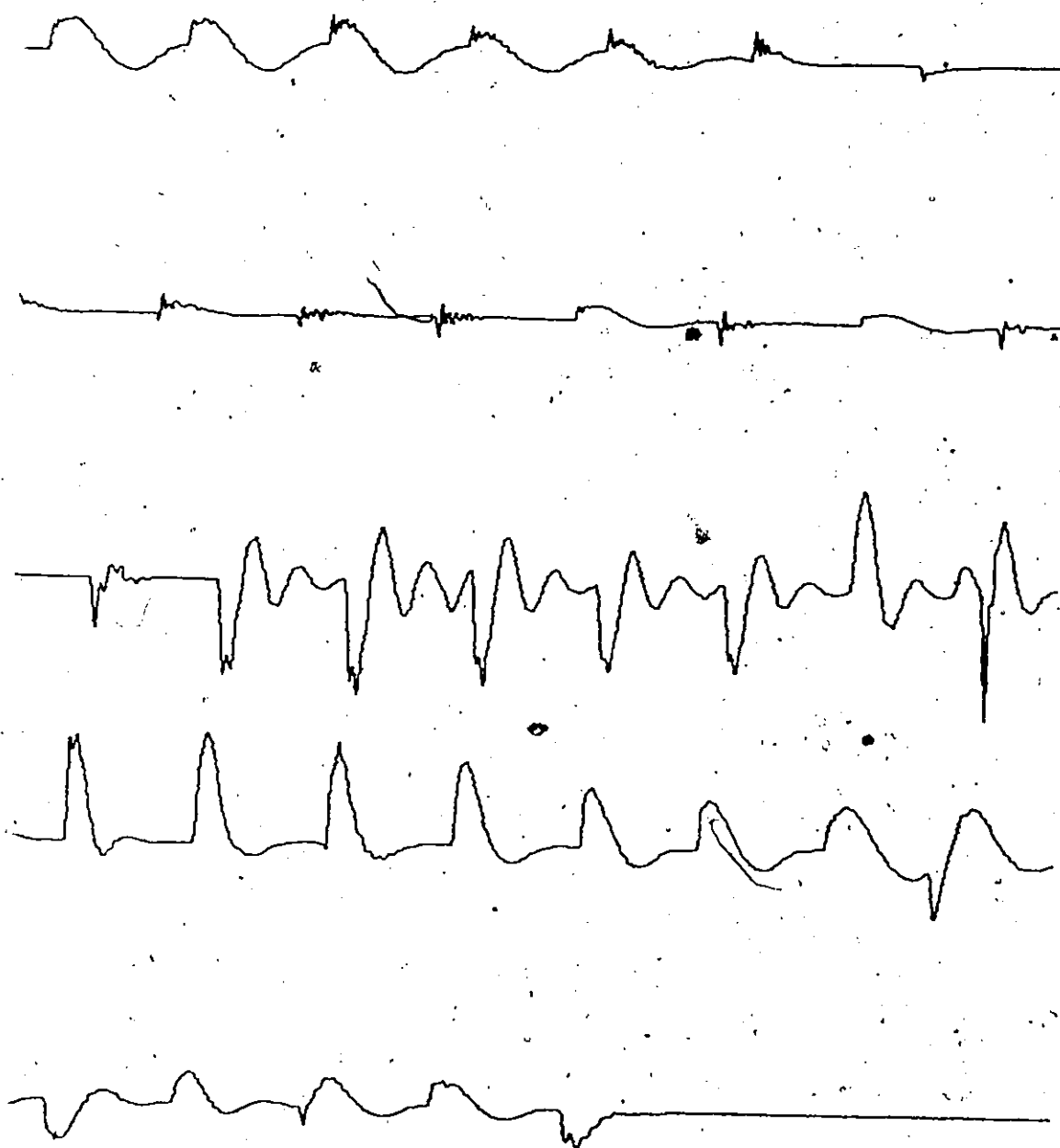


FIGURE 20 (c) - Reconstructed Waveform Using Impulse Excitation - "SLEEPY"

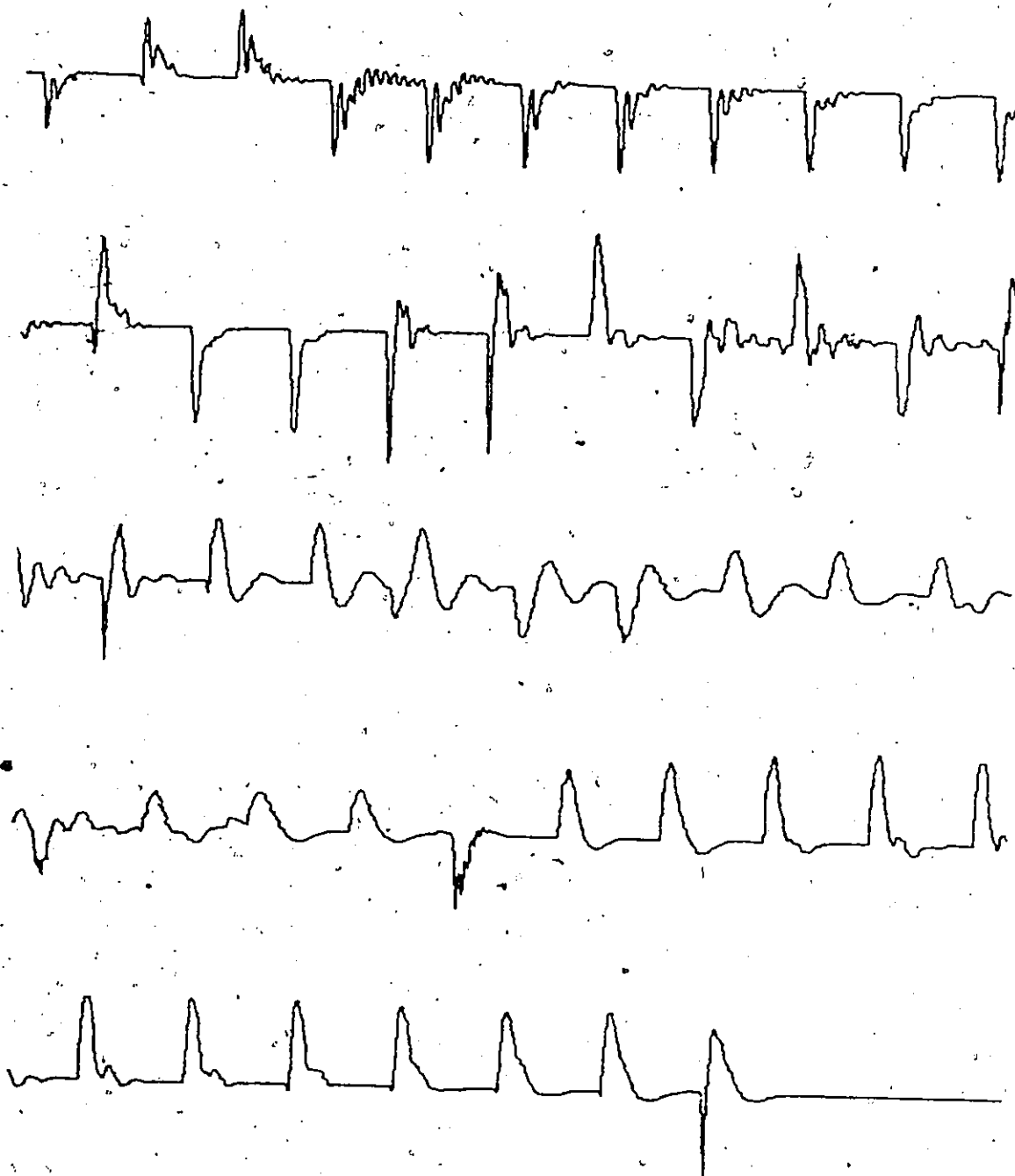


FIGURE 20 (d) - Reconstructed Waveform Using Impulse Excitation - "HOW ARE YOU"

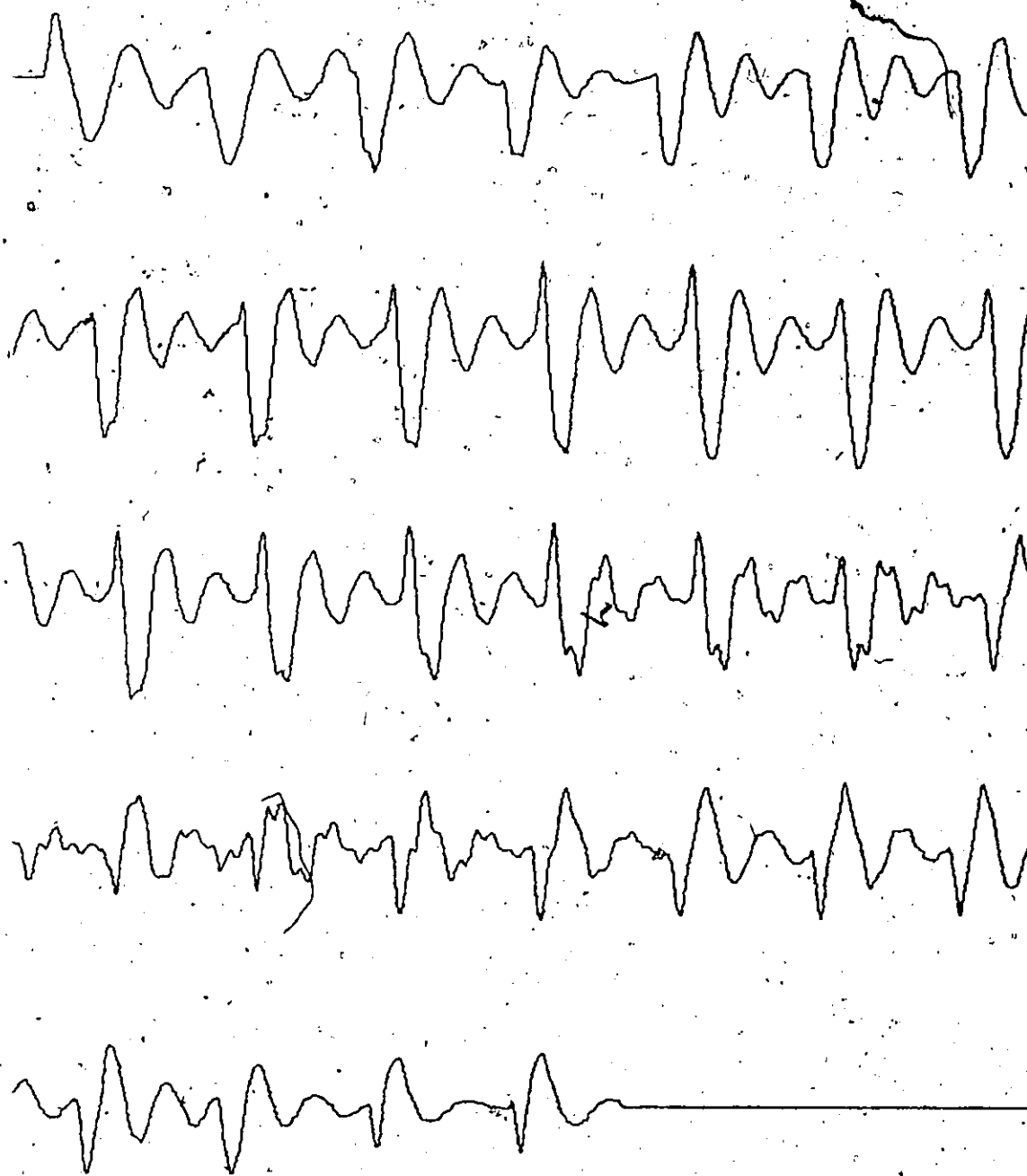


FIGURE 21 (a) - Reconstructed Waveform Using Triangular Excitation - "NOON"

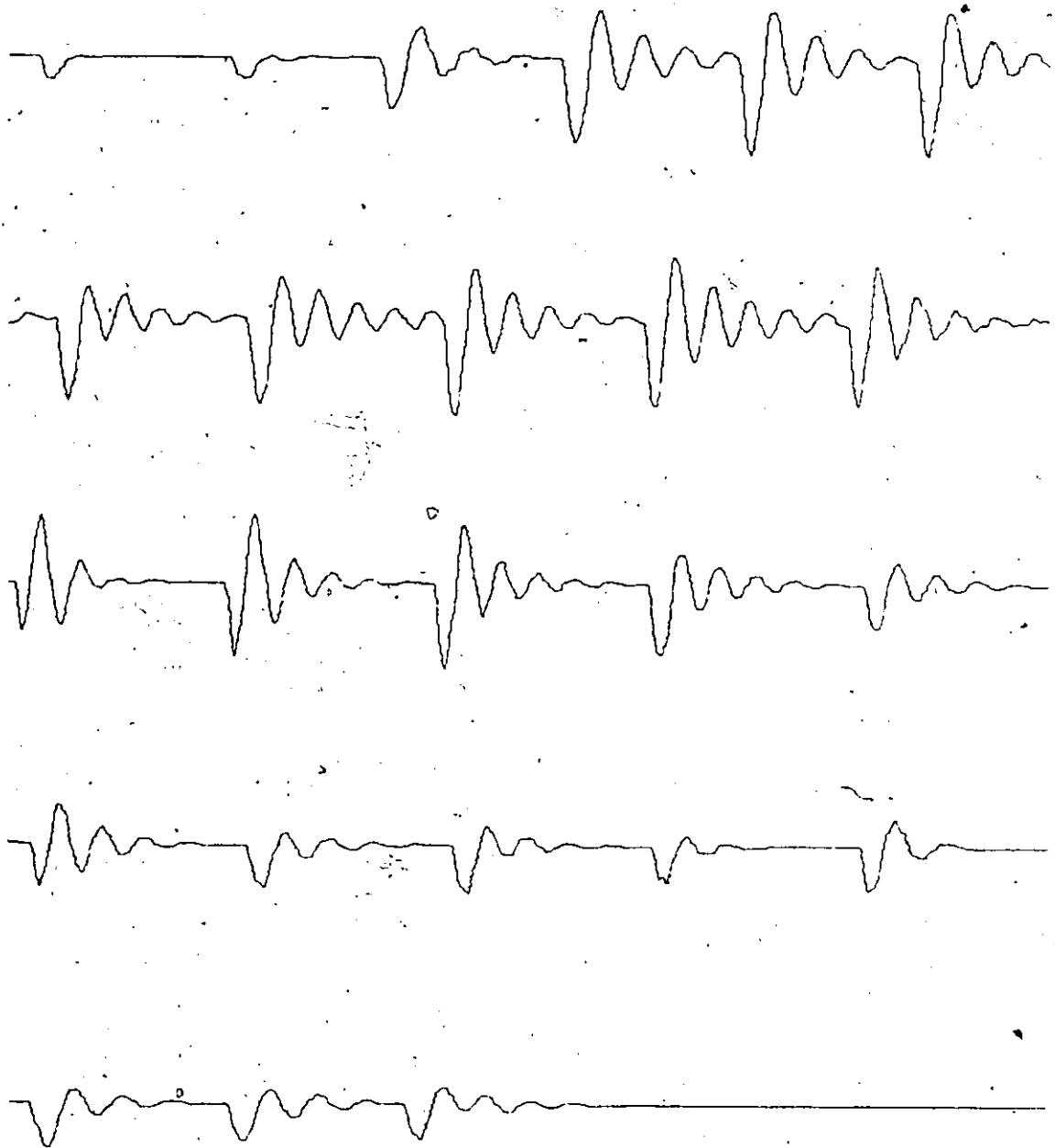


FIGURE 21 (b) - Reconstructed Waveform Using Triangular Excitation - "DAY"



7
FIGURE 21 (c) - Reconstructed Waveform Using Triangular Excitation - "SLEEPY"

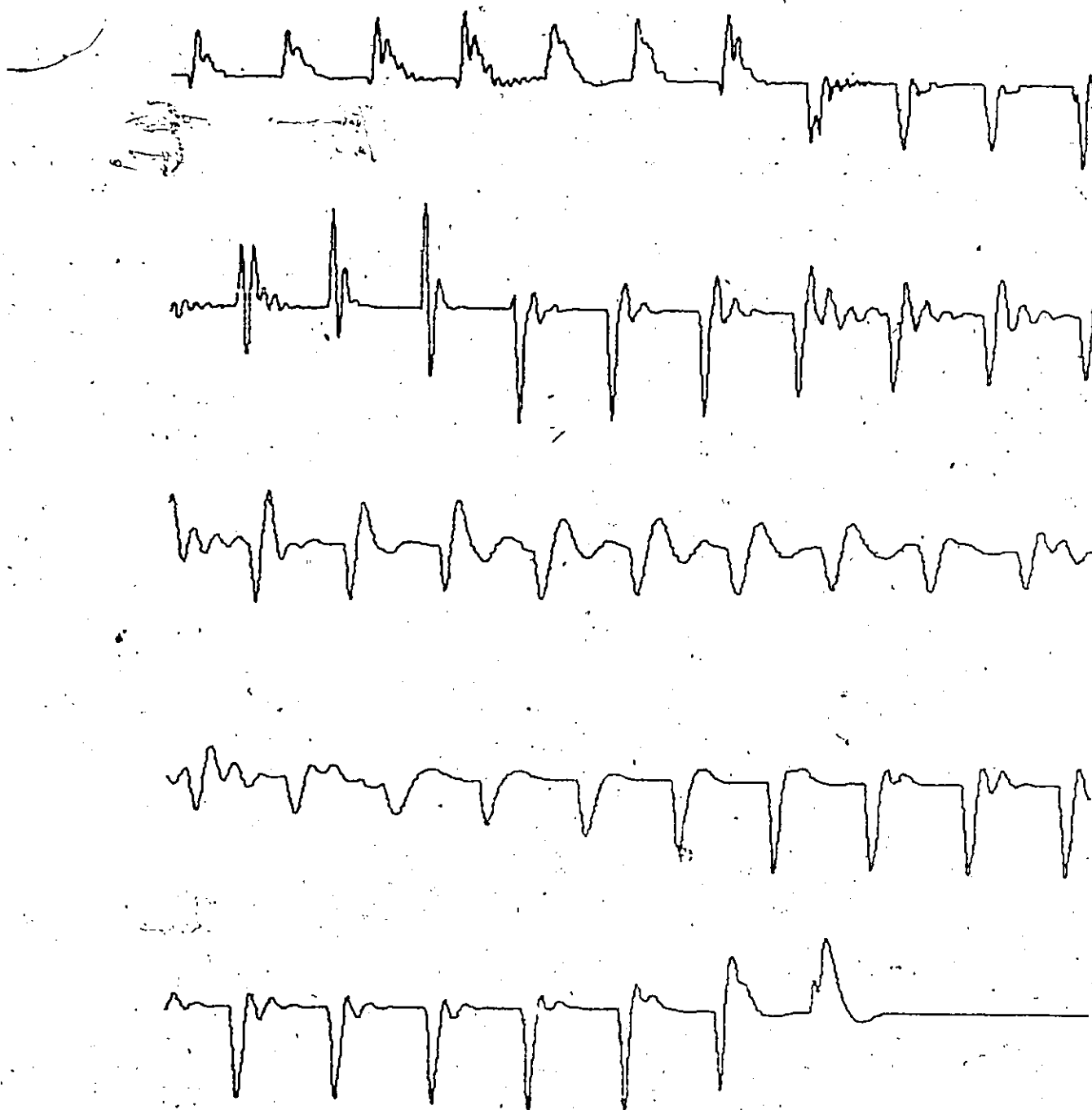


FIGURE 21 (d) - Reconstructed Waveform Using Triangular Excitation - "HOW ARE YOU"

the predictor transfer function is given by

$$H(z) = \frac{\sum_{j=0}^N b_j z^{-j}}{1 - \sum_{i=1}^p a_i z^{-i}} \quad N \leq p \quad (2.29)$$

where a_i and b_j are the predictor coefficients.

Let $X(z)$ and $Y(z)$ be the input and the output of the predictor respectively, then

$$\frac{Y(z)}{X(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_N z^{-N}}{1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_p z^{-p}} \quad (2.30)$$

If we now let $X(z) = 1$, that is, an impulse, then

$$(b_0 + b_1 z^{-1} + \dots + b_N z^{-N}) = (1 - a_1 z^{-1} - \dots - a_p z^{-p}) \cdot Y(z) \quad (2.31)$$

Collecting terms in equation (2.31), we have

$$b_0 = y_0$$

$$b_1 = y_1 - a_1 y_0$$

$$b_2 = y_2 - a_1 y_1 - a_2 y_0$$

(2.32)

$$b_N = y_N - a_1 y_{N-1} - \dots - a_N y_0$$

The predictor zeros can be obtained by solving the set of equations (2.32). This procedure has been suggested [33,34] as a means of obtaining the digital filter zeros, and a similar technique was described by Freeman [35].

The above technique will not work efficiently because any error in the predictor coefficients a_i 's will also affect the b_j 's. Round off and

truncation errors in solving equation (2.32) will further degrade the predictor zeros. Furthermore, the zeros obtained by this method do not represent the true speech signal zeros, since a small time shift in the signal will give totally different results. This procedure was implemented and results are shown for the words "NOON", "DAY", "SLEEPY", and "HOW ARE YOU" respectively, in Figure 22.

2.3.1.(c) Linear Predictor with Initial Conditions

The best approximation to the actual speech production process is by a model which has both poles and zeros. Let the difference equation describing the speech production process be given by

$$\hat{s}_n = \sum_{i=1}^p a_i s_{n-i} + \sum_{j=0}^N b_j e_{n-j} \quad (2.33)$$

where a's and b's are the predictor coefficients and e_n is the excitation function.

Now let $\hat{s}_1 = s_1$

$$\hat{s}_2 = s_2$$

...

$$\hat{s}_p = s_p$$

and $e_n = \delta(n)$

For $p < n \leq N$, the equation (2.33) reduces to

$$\hat{s}_n = \sum_{i=1}^p a_i s_{n-i} \quad (2.35)$$

From equation (2.35) we see that, if the first p speech samples are used as initial conditions in equation (2.33), then the summation

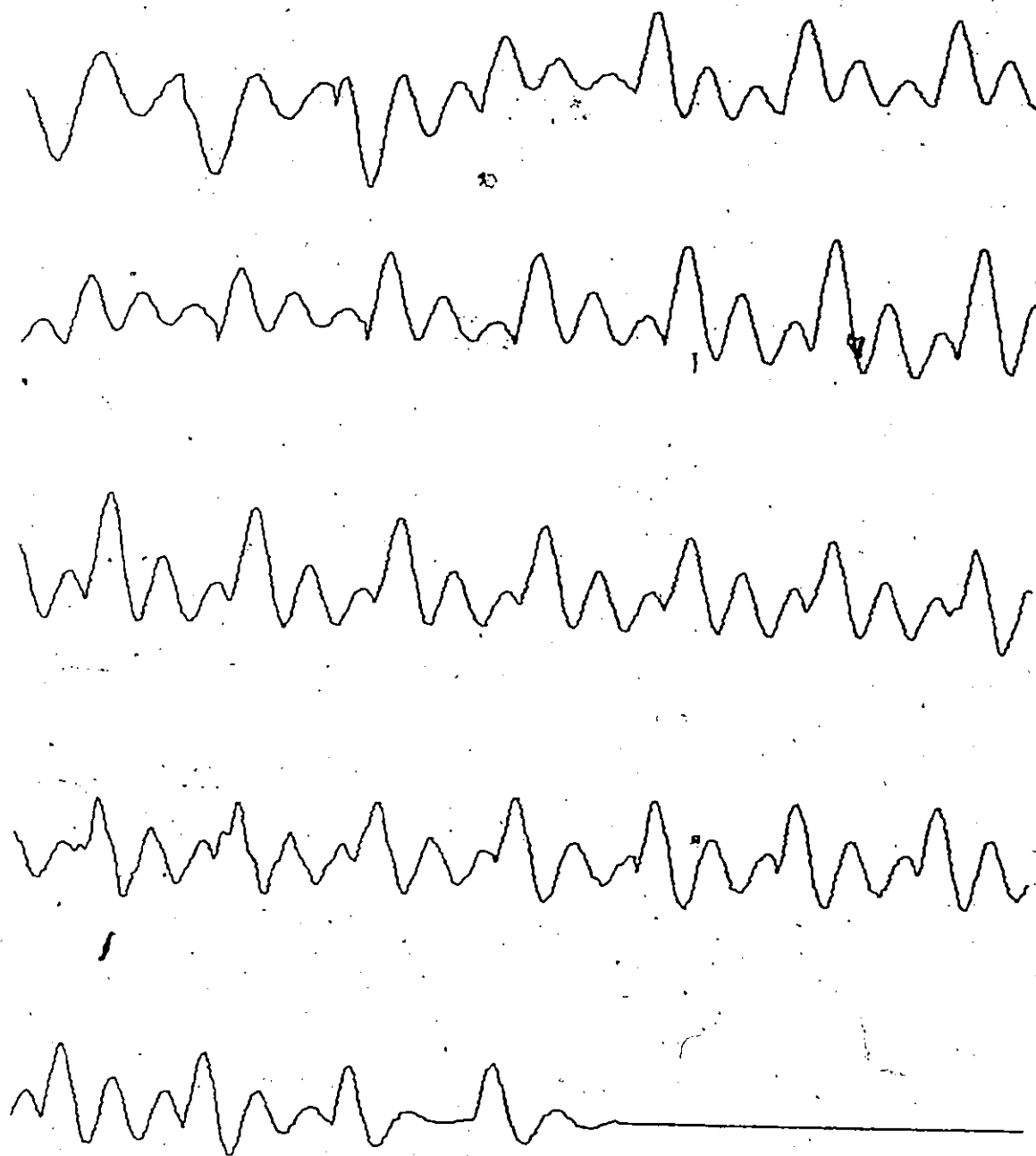


FIGURE 22 (a) - Pole-Zero Model Reconstruction Of The Word - "NOON"

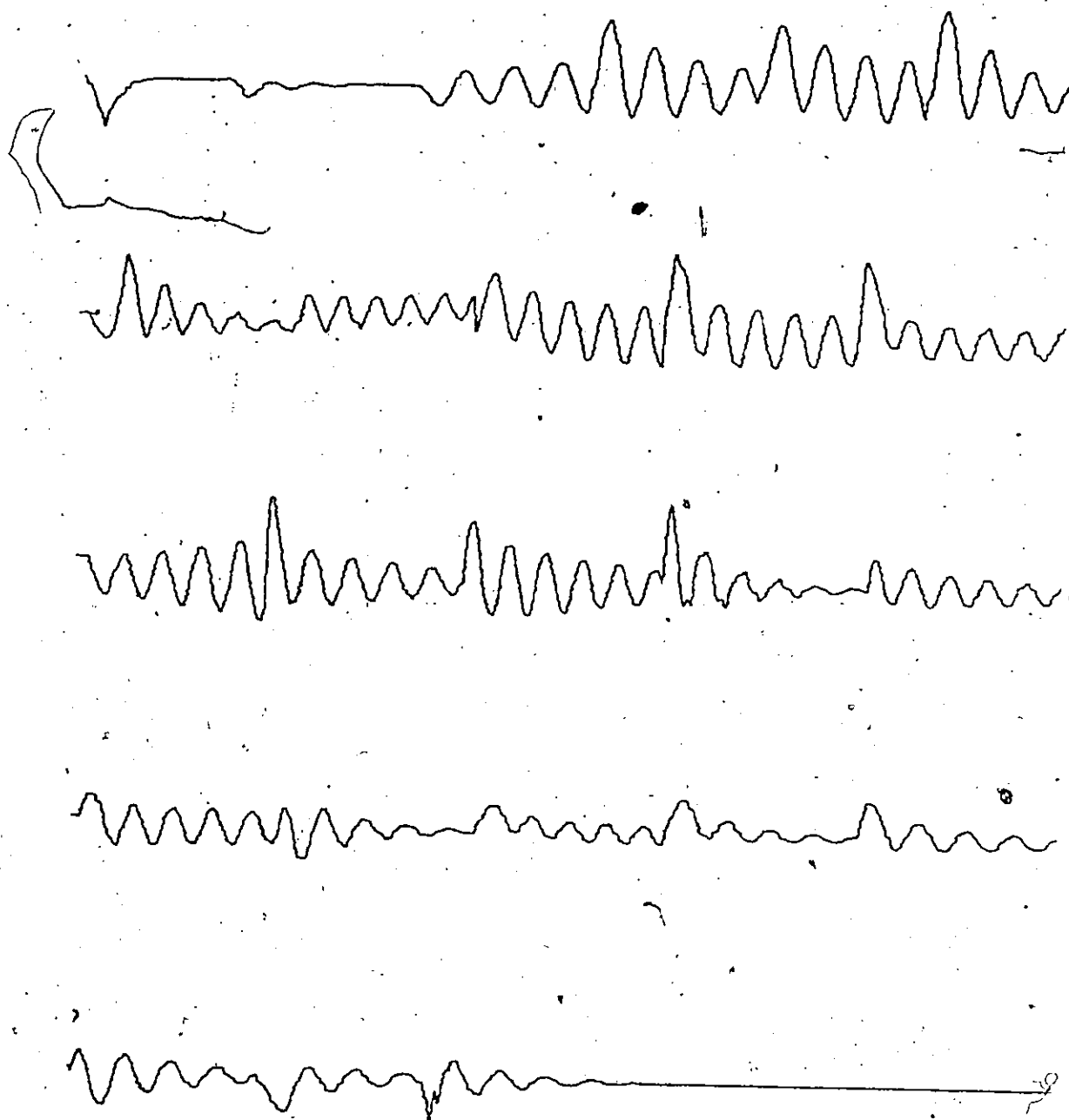


FIGURE 22 (b) - Pole-Zero Model Reconstruction Of The Word - "DAY"

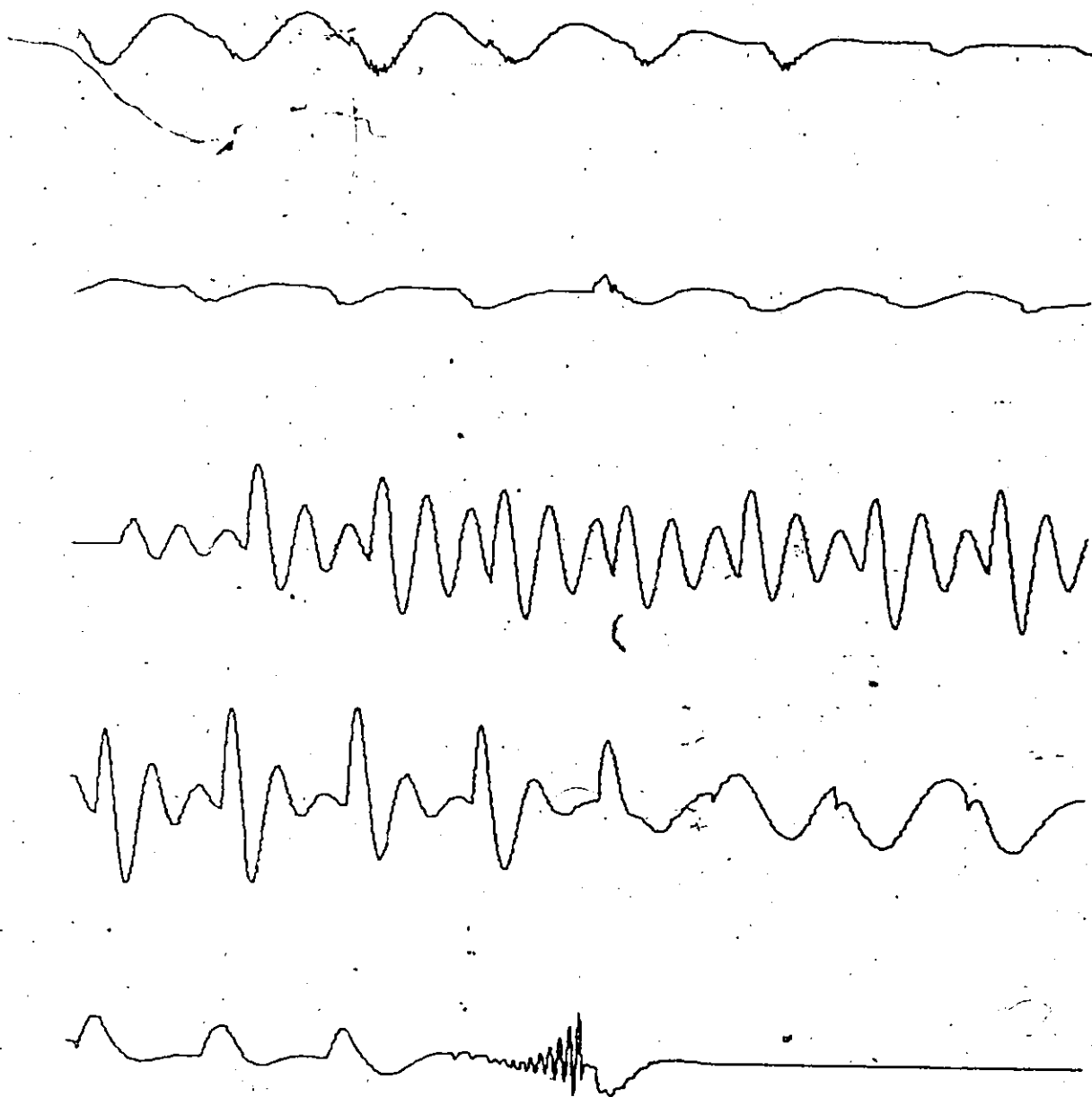


FIGURE 22 (c) - Pole-Zero Model Reconstruction Of The Word - "SLEEPY"

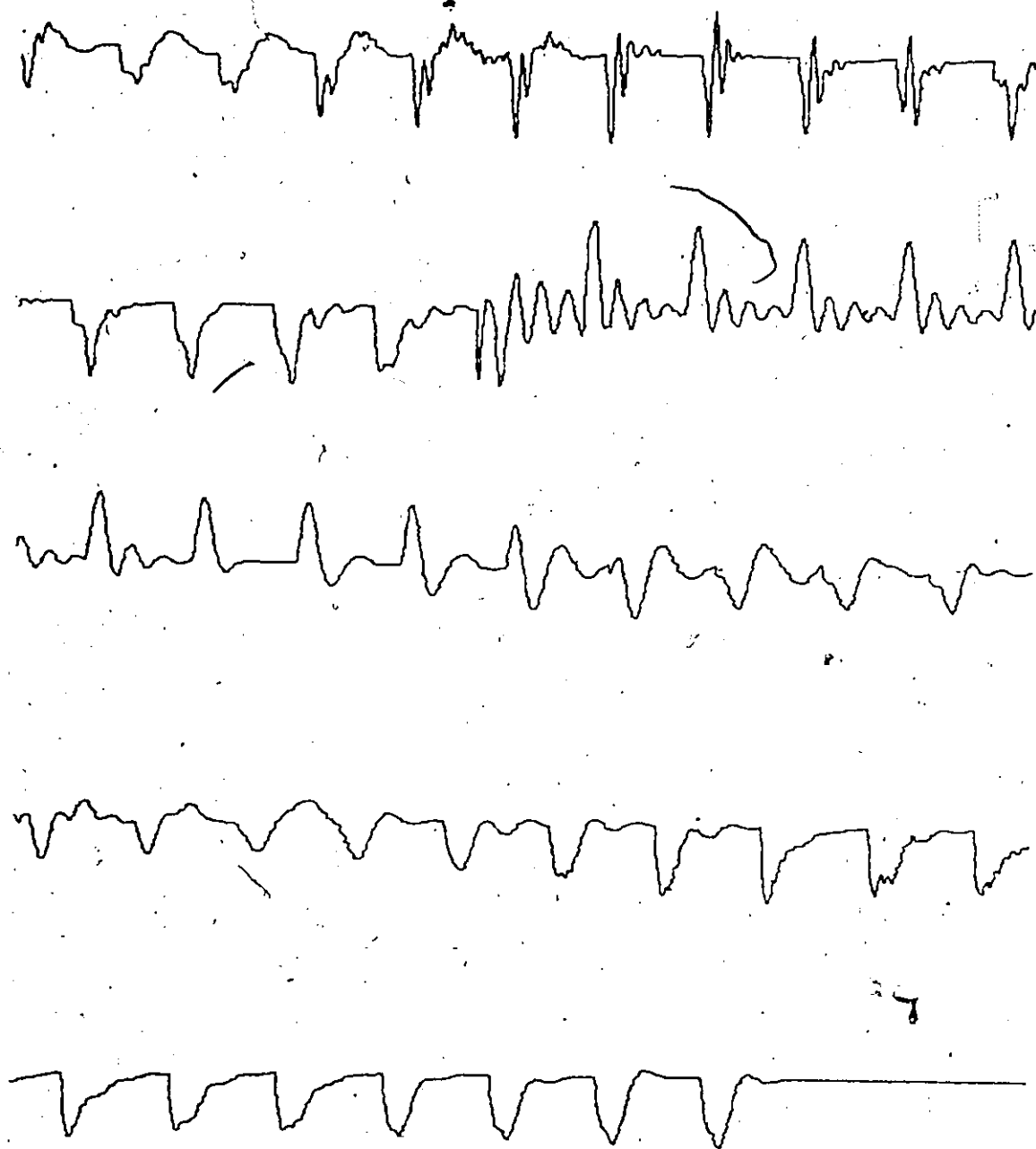


FIGURE 22 (d) - Pole-Zero Model Reconstruction Of The Word - "HOW ARE YOU"

involving the excitation does not contribute to the speech estimate, as shown in Equation (2.35). The use of initial conditions (initial conditions are the first p speech samples in a given analysis interval at which time, the time parameter is reset to zero) in equation (2.33) has removed the excitation function from the difference equation. This is an added advantage, since we are uncertain as to the actual shape of the excitation pulse or of the location of zeros which will adequately define the speech waveform.

Implementation of equation (2.35) for speech synthesis by using the linear predictor will run into the following problems.

1. The evaluation of the predictor coefficients a_1 's requires very accurate data if direct-realization is desired.
2. The iterative nature of equation (2.35) is very sensitive to truncation and roundoff errors. Any error in a previous estimate will affect the subsequent estimates.

The initial conditions model will give good results if the predictor is initialized at the start of a pitch period. This is not always possible due to the difficulty in estimating the actual location of the pitch excitation. One way to overcome this problem, is to re-initialize the predictor at an extremum point close to the location of the pitch excitation. When this is done, the initial condition model will give good results. However, if the predictor is re-initialized at some time where the speech signal is low in amplitude, the resulting reconstruction will be poor. Figure 23 shows the results of the initial conditions model reconstruction for the words "NOON", "DAY", "SLEEPY", and "HOW ARE YOU".

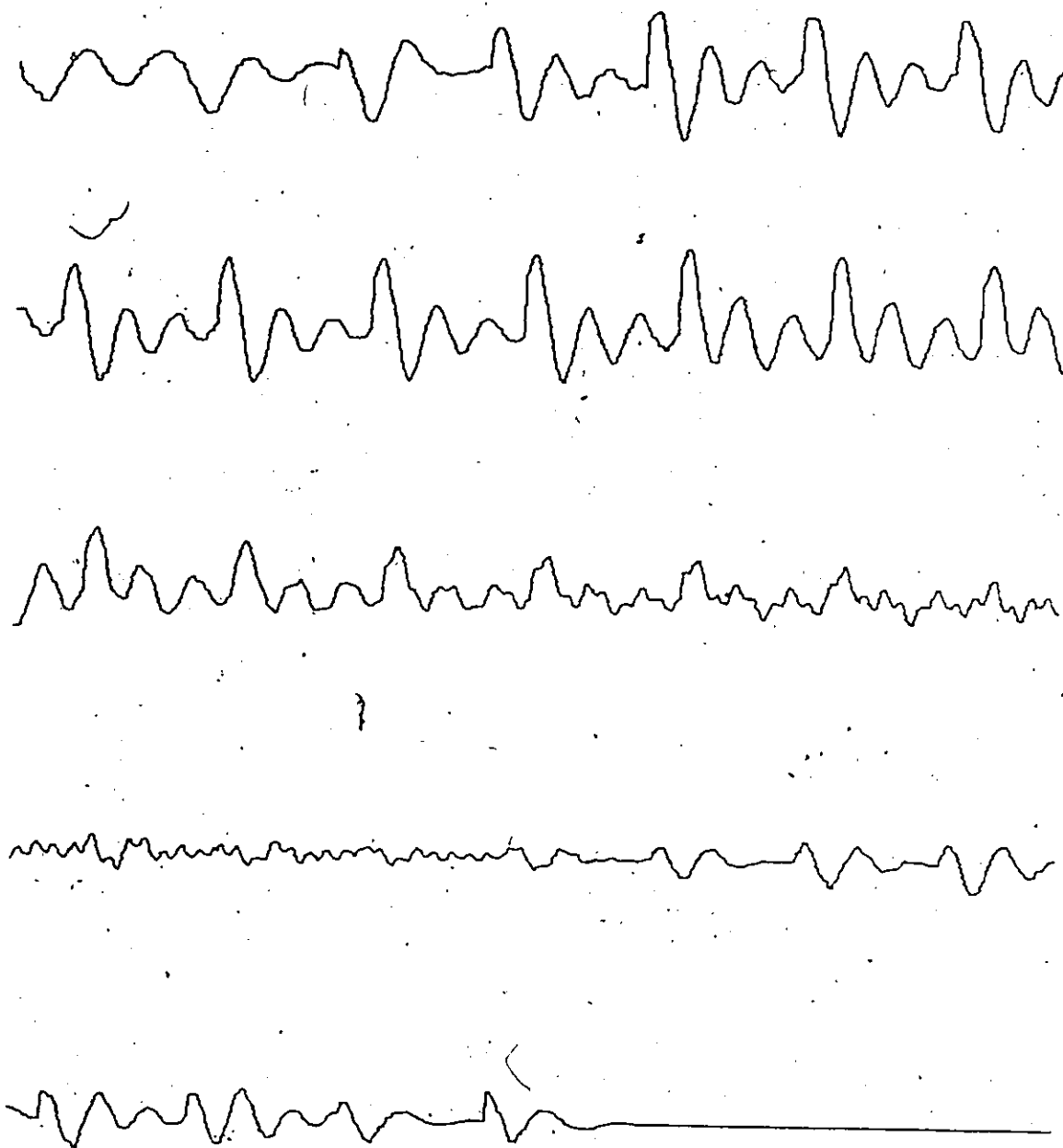


FIGURE 23 (a) - Initial Condition Model Reconstruction Of The Word - "NOON"

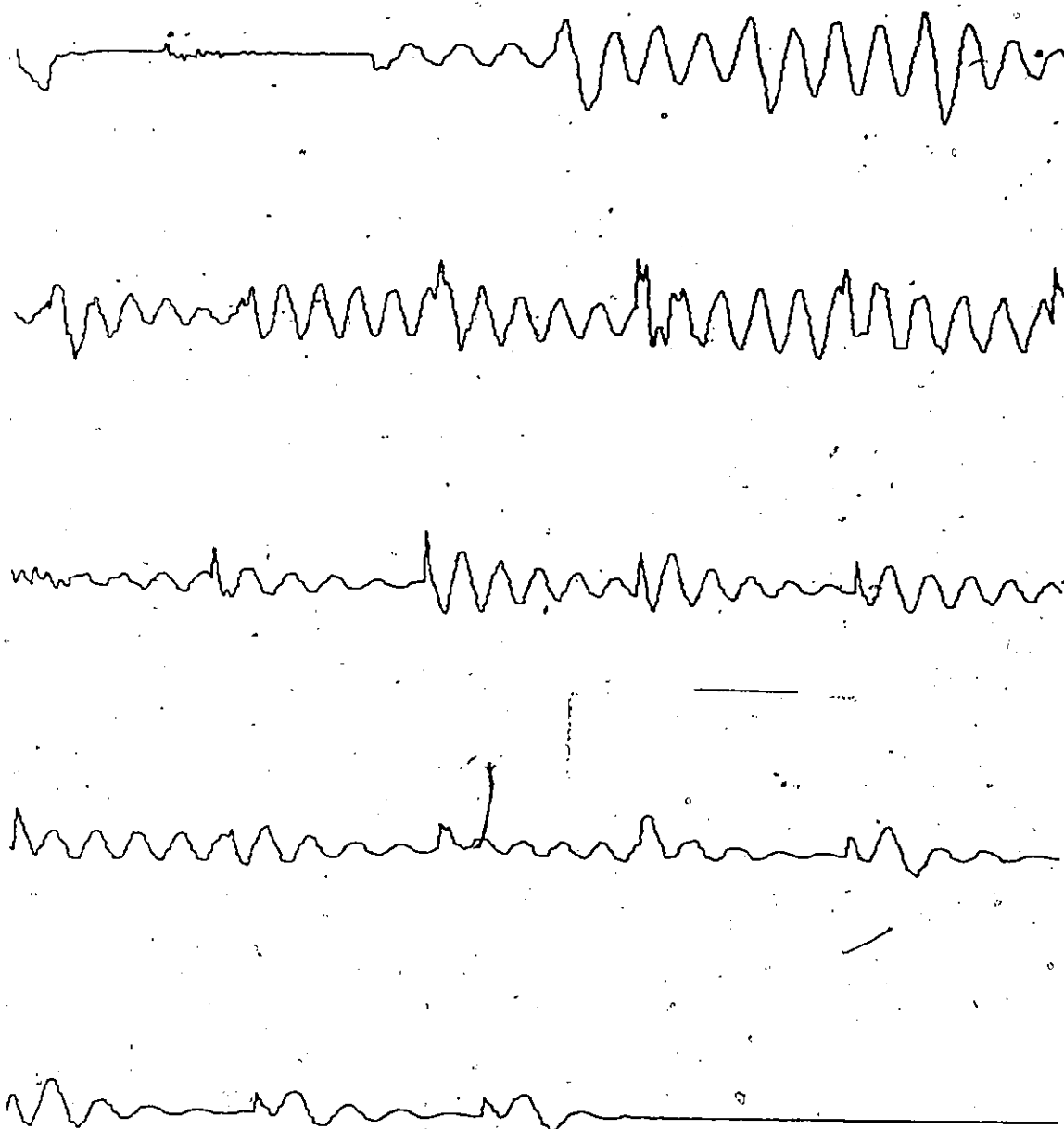


FIGURE 23 (b) - Initial Condition Model Reconstruction Of The Word - "DAY"

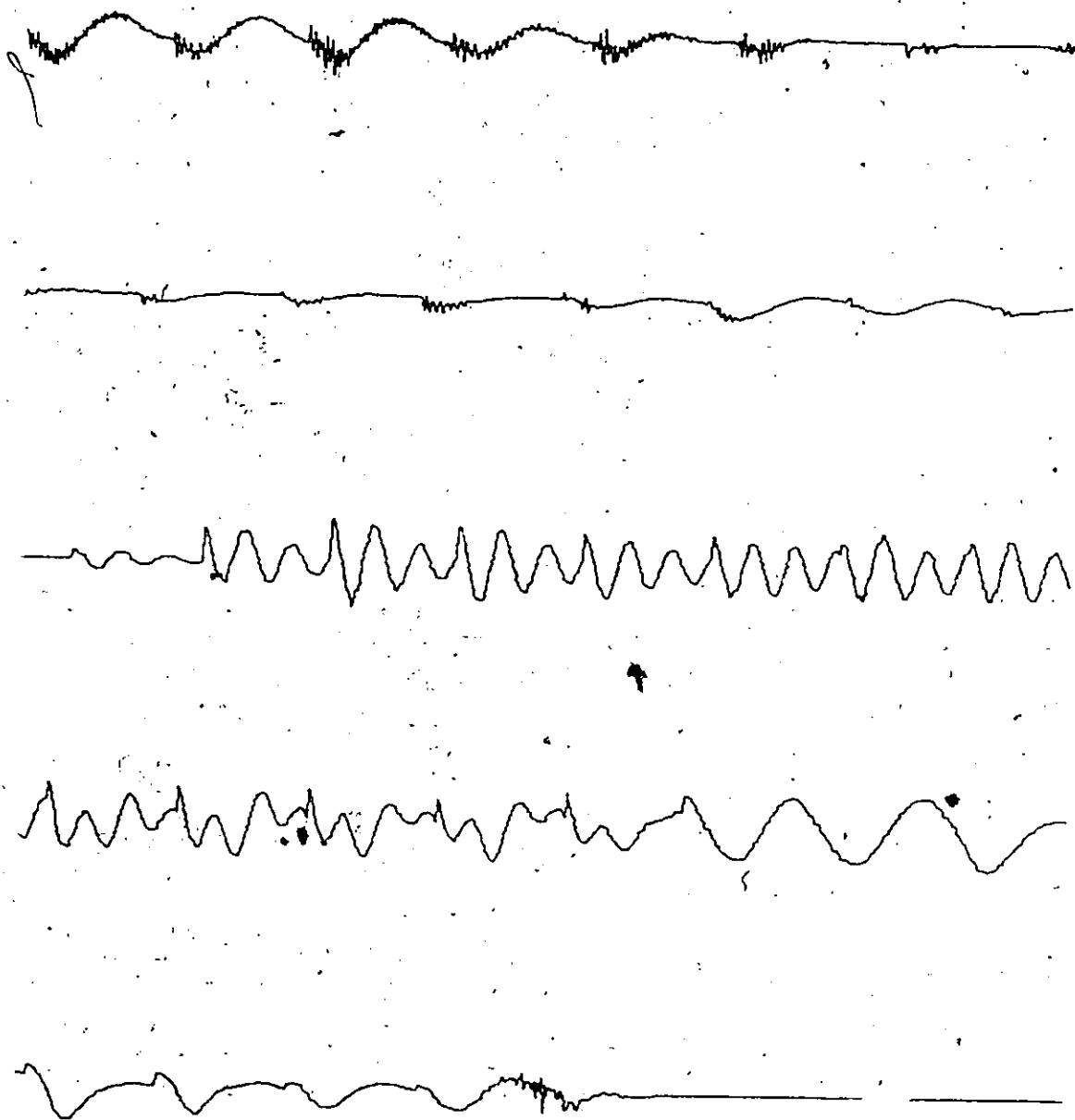


FIGURE 23 (c) - Initial Condition Model Reconstruction Of The Word - "SLEEPY"

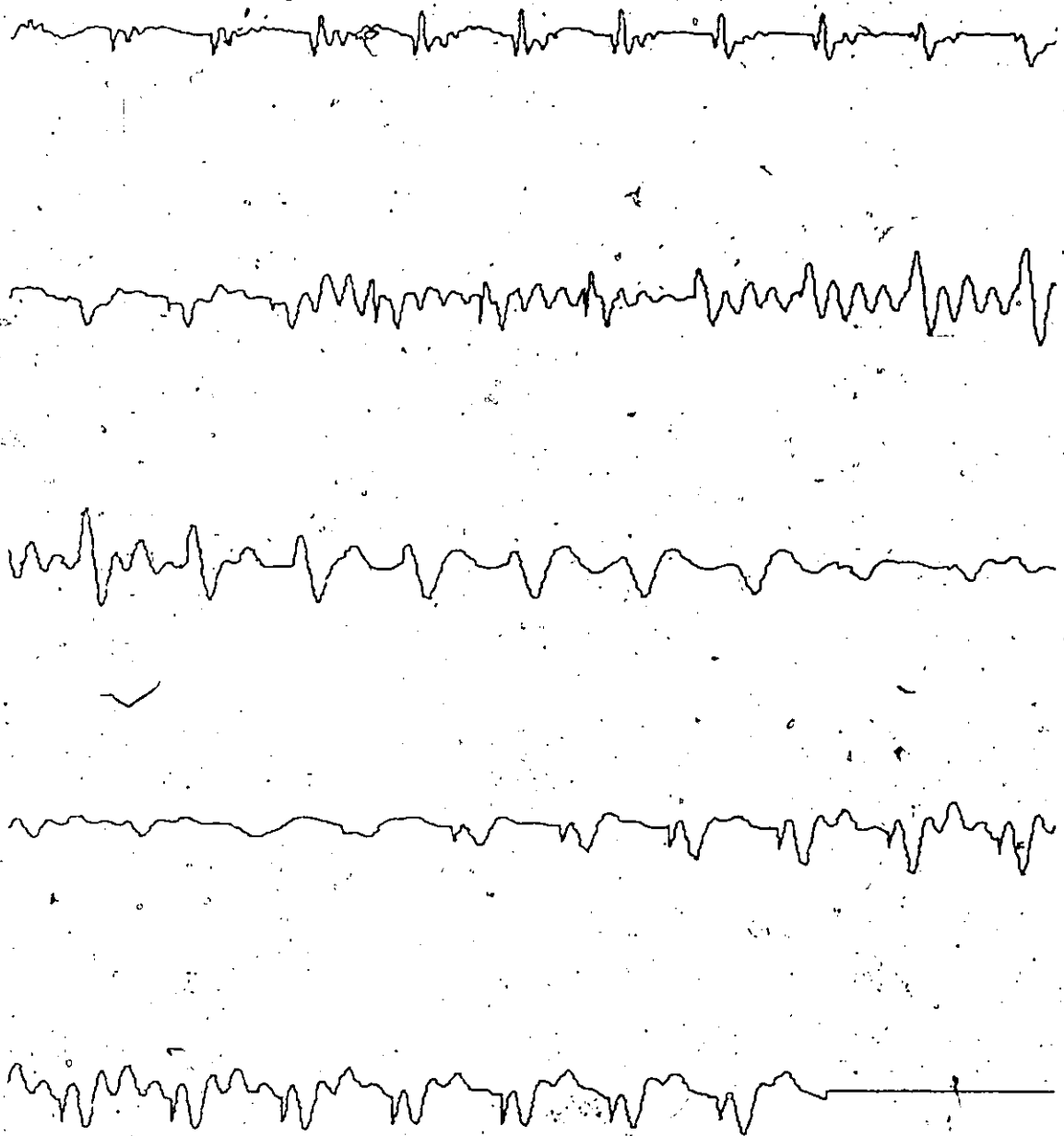


FIGURE 23 (d) - Initial Condition Model Reconstruction Of The Word - "HOW ARE YOU"

2.3.11 Parallel Realization

In the parallel realization scheme, the rapid error build up which was associated with the direct realization is avoided. Although the parallel realization is also iterative or recursive, the error build up is much slower. In this type of realization the model for the speech production process is not constrained to be an all pole digital filter. The effect of the speech signal zeros is included in the residues.

Let the transfer function of the speech production model be given by

$$H(z) = \frac{\sum_{j=0}^N b_j z^{-j}}{1 - \sum_{i=1}^p a_i z^{-i}} \quad N \leq p \quad (2.36)$$

$$= \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_N z^{-N}}{1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_p z^{-p}}$$

let $N = p$ and multiply top and bottom by z^p

$$H(z) = \frac{z^p b_0 + b_1 z^{p-1} + \dots + b_N}{z^p - a_1 z^{p-1} - \dots - a_p} \quad (2.37)$$

Now the roots of numerator and the denominator

of equation (2.37) can be found and $H(z)$ can be expressed as,

$$H(z) = \frac{k(1 - \beta_0 z^{-1})(1 - \beta_1 z^{-1}) \dots (1 - \beta_N z^{-1})}{(1 - \alpha_0 z^{-1})(1 - \alpha_1 z^{-1}) \dots (1 - \alpha_p z^{-1})} \quad (2.38)$$

Equation (2.38) can be factored using partial fractions into the following form.

$$H(z) = \frac{A_1}{1 - \alpha_1 z^{-1}} + \frac{A_2}{1 - \alpha_2 z^{-1}} + \dots + \frac{A_p}{1 - \alpha_p z^{-1}} \quad (2.39)$$

where A_i is the residue of the i th pole
and α_i , $i = 1, 2, \dots, p$ are the poles.

Thus the p th order difference equation can be reduced to p first order difference equations. Alternatively one could combine the complex conjugate pole pairs to form a number of second order systems. The realization of the linear predictor in the parallel form is shown in Figure 24. The inputs to the system shown in Figure 24, are the residues A_i and the poles α_i , $i = 1, 2, \dots, p$. The poles are found from the predictor parameters a_i , and one technique of finding the residues will be discussed in the next section. The output of Figure 24, $y(nT)$ is a real function, but any of the intermediate outputs may be complex depending upon the respective pole. When the intermediate outputs are summed, all the imaginary components should cancel leaving a real result. The transfer function of any subfilter is given by

$$H(z) = \frac{A}{1 - \alpha z^{-1}} = \frac{Y(z)}{X(z)} \quad (2.40)$$

rearranging

$$Y(z) = \alpha z^{-1} Y(z) + A X(z) \quad (2.41)$$

taking the inverse z -transform of both sides we obtain

$$y_n = \alpha y_{n-1} + A x_n \quad (2.42)$$

There will be p difference equations of this type instead of one p th order difference equation. The error build up in this model will be much less severe, since each subfilter requires only one multiplication and one addition. Also, the n th estimate, now only depends on the $(n-1)$ th estimate, instead of the previous p estimates in the direct realization scheme.

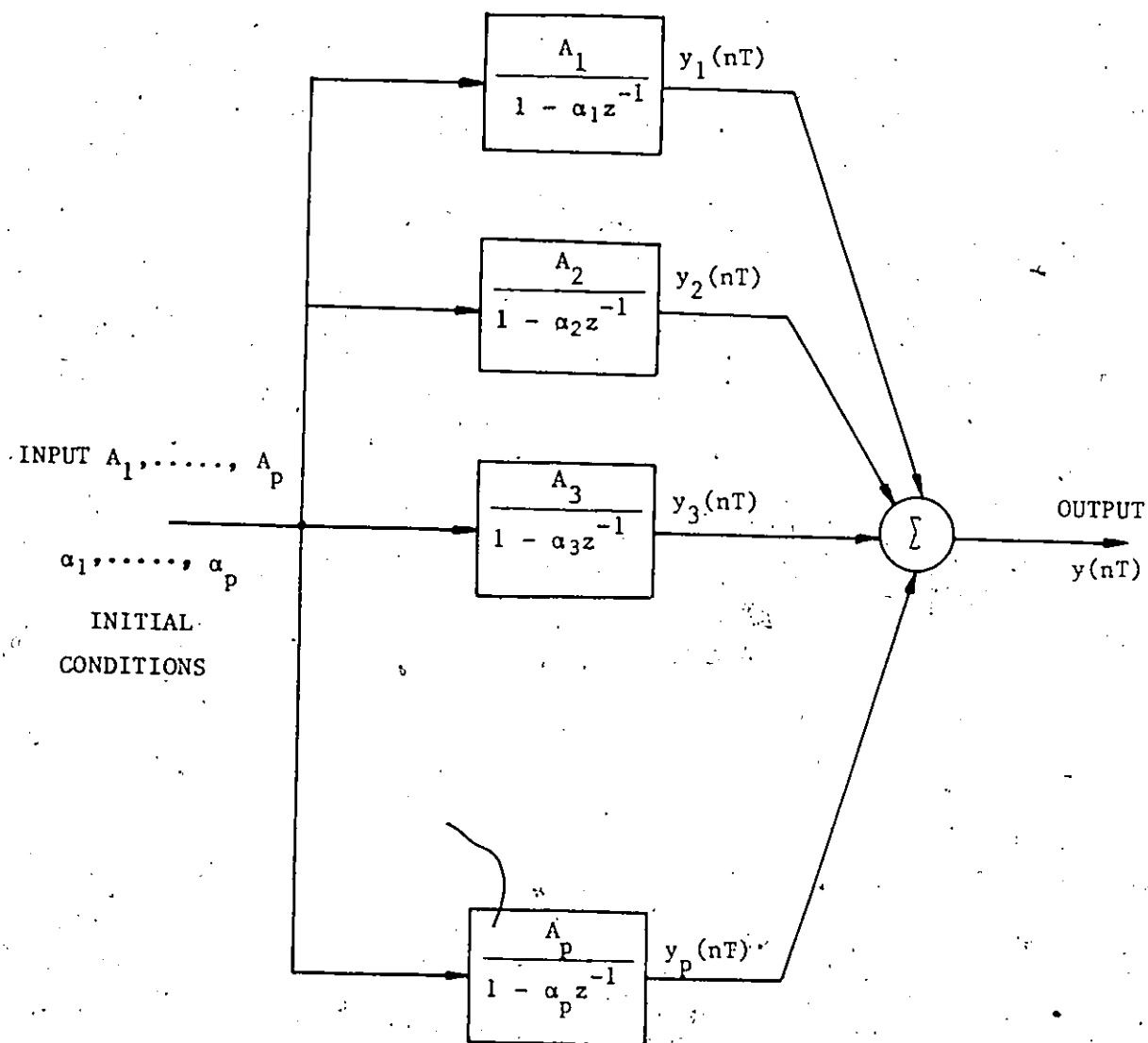


FIGURE 24 - Parallel Realization Form Of The Linear Predictor

The results of the parallel realization are shown in Figure 25 for the words "NOON", "DAY", "SLEEPY", and "HOW ARE YOU", respectively.

2.4. Closed Form Reconstruction

This form of realization makes use of the assumption that the speech signal can be expressed as a sum of exponentials. It differs from the iterative cases, in that the estimate of the n th speech sample does not depend on any past estimates. Because of this, it is less sensitive to round off errors associated with the finite word length in digital computers.

2.4.1. Exponential Curve Fitting

In this method, the speech signal $S(t)$ is estimated as a sum of exponential functions in a given analysis interval. Thus

$$S(t) = \sum_{k=1}^P A_k \exp[(\alpha_k + j\beta_k)t], \quad j = \sqrt{-1} \quad (2.43)$$

where α_k , β_k and A_k are parameters that characterize the speech wave.

The derivation of the parameters is based on the minimization of the mean squared error between the speech sample and its estimate. A function that satisfies equation (2.43) may be shown [36] to satisfy a p th order difference equation. Then

$$S[(n+p)T] + a_1 S[(n+p-1)T] + a_2 S[(n+p-2)T] + \dots + a_p S(nT) = 0 \quad (2.44)$$

where T is the sampling interval

a_k , $k = 1, 2, \dots, p$ are the predictor parameters

$n = 1, 2, \dots, N$.

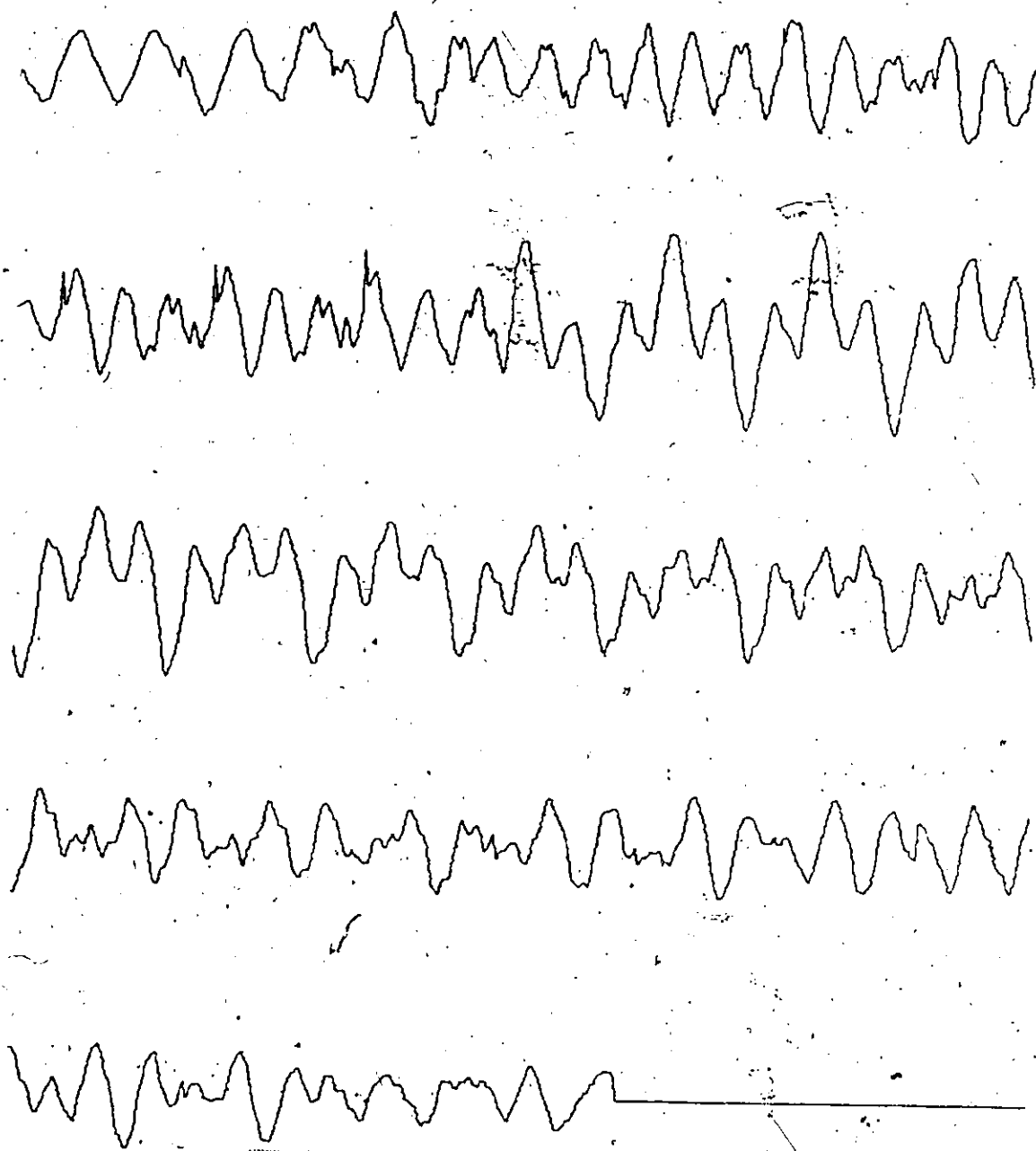


FIGURE 25 (a) - Parallel Realization Model Reconstruction Of The Word - "NOON"

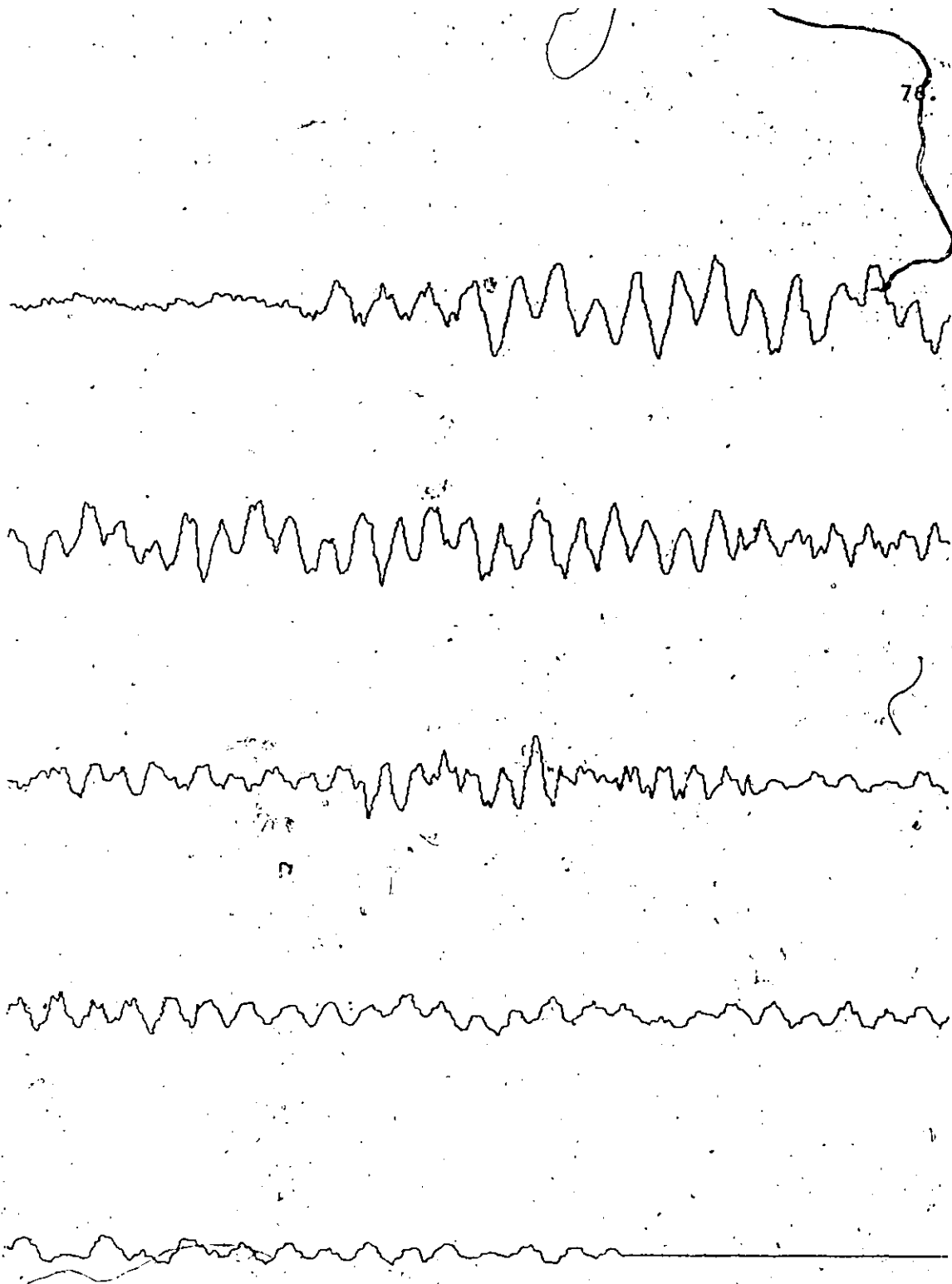


FIGURE 25 (b) - Parallel Realization Model Reconstruction Of The Word - "DAY"

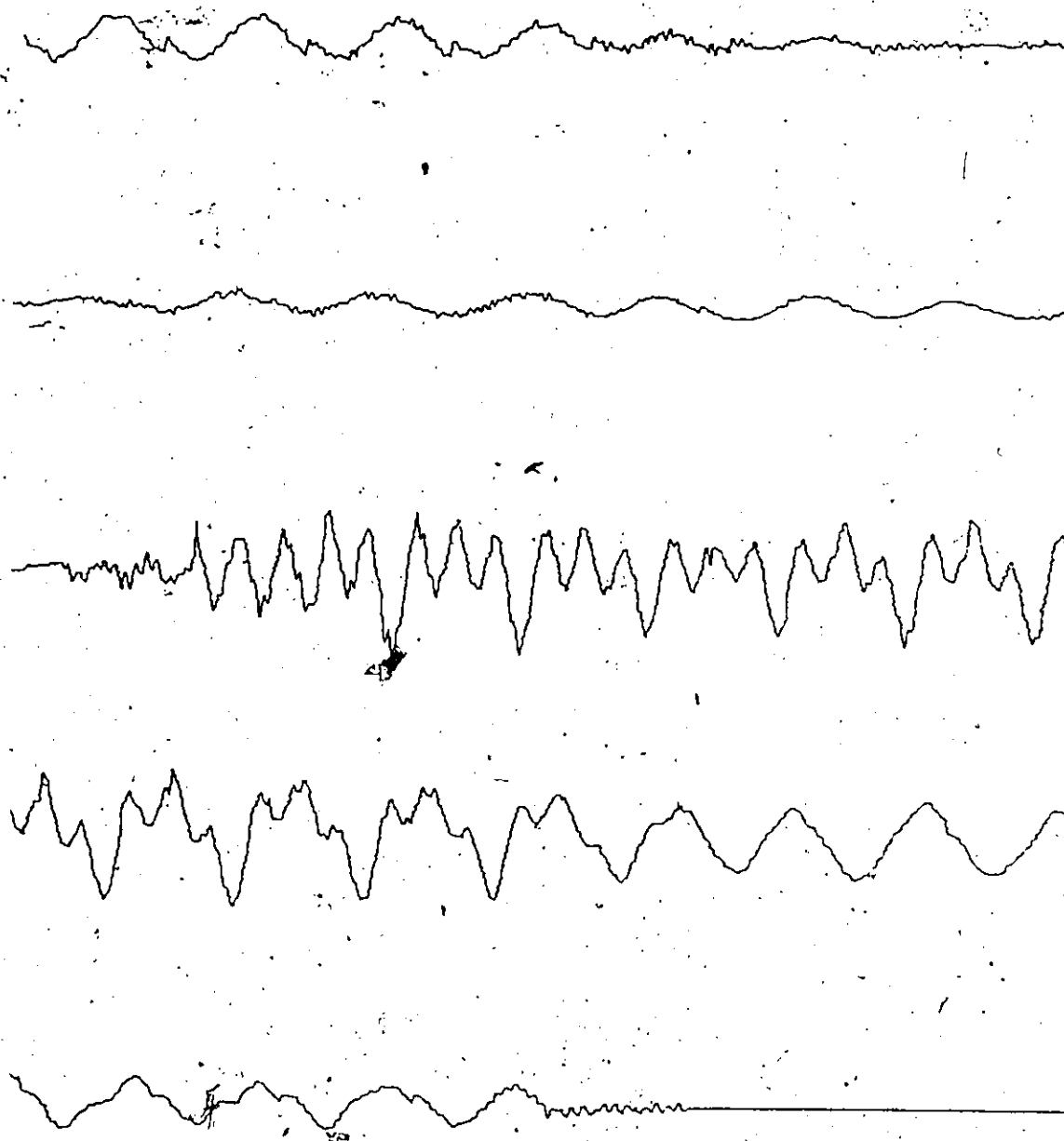


FIGURE 25 (c) - Parallel Realization Model Reconstruction Of The Word - "SLEEPY"

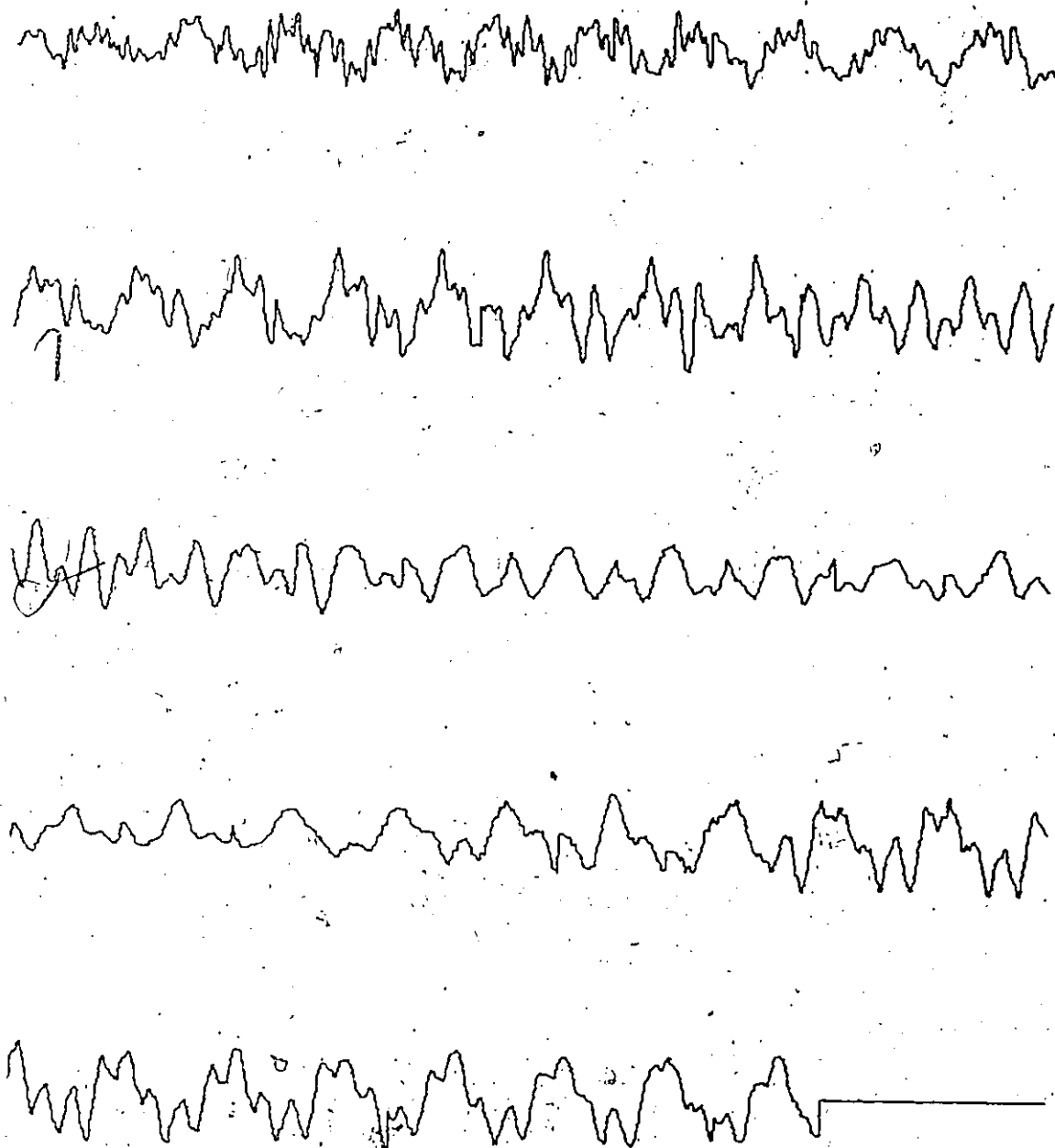


FIGURE 25 (d) - Parallel Realization Model Reconstruction Of The Word - "HOW ARE YOU"

The system of equations given by (2.44) can be solved for the unknown parameters a_k , by least squares techniques. With,

$$X_k = \exp[(\alpha_k + j\beta_k)T] \quad (2.45)$$

it can be shown [36], that the X_k 's satisfy the following polynomial equation,

$$a_p + a_{p-1}x + a_{p-2}x^2 + \dots + a_1x^{p-1} + x^p = 0 \quad (2.46)$$

Equation (2.46) may be solved numerically for X_k 's (the poles of the predictor in the z-plane) by the use of Newton Raphson algorithm. The parameters α_k and β_k are then obtained as follows:

$$\left. \begin{aligned} \text{Let } X_k &= c_k + jd_k \\ \text{then } \alpha_k &= \frac{1}{2} \ln [c_k^2 + d_k^2] \\ \beta_k &= \tan^{-1} [d_k/c_k] \\ k &= 1, 2, \dots, p \end{aligned} \right\} \quad (2.47)$$

α_k and β_k are the real and imaginary parts of the s-plane poles of the predictor, can be obtained by the method presented above.

The evaluation of the residues A_k is based on the following analysis.

Using equations (2.45) in (2.47) one obtains

$$\hat{S}(nT) = \sum_{k=1}^p A_k X_k^n \quad n = 1, 2, \dots, N \quad (2.48)$$

The A_k 's are obtained by minimizing the mean squared error between $S(nT)$ and its estimate $\hat{S}(nT)$ given by equation (2.48). The mean squared error is given by

$$\begin{aligned}
 E &= \frac{1}{N} \sum_{n=1}^N [S(nT) - \hat{S}(nT)]^2 \\
 &= \frac{1}{N} \sum_{n=1}^N \left[S(nT) - \sum_{k=1}^p A_k x_k^n \right]^2
 \end{aligned} \tag{2.49}$$

Differentiating equation (2.49) with respect to A_j and setting the result to zero, one obtains

$$\frac{\partial E}{\partial A_j} = -\frac{2}{N} \sum_{n=1}^N \left[S(nT) - \sum_{k=1}^p A_k x_k^n \right] x_j^n = 0 \tag{2.50}$$

$$j = 1, 2, \dots, p$$

rearranging

$$\sum_{n=1}^N \sum_{k=1}^p x_k^n x_j^n A_k = \sum_{n=1}^N S(nT) x_j^n \tag{2.51}$$

The solution of equation (2.51) yields the residues A_k .

The advantage of this method over the methods discussed in the previous sections, is due to the closed-form analytic expression that is obtained for estimating speech samples. Thus the quantization effects that occur in the iterative methods are avoided. The parallel realization and the closed form realization are identical except in the reconstruction.

2.4.ii. Adaptive Adjustment of System Order

Due to the inaccuracies of numerical computation and the finite precision of the data sequence, it is possible to have the poles of the closed form model in the right half S-plane indicating instability. In such cases, the poles are simply shifted to the imaginary axis. Another difficulty that might be encountered is the overestimation of the system.

Overestimation of the system order results in large values for the residues A_k (equation 2.48), resulting in excessive overshoots in the reconstructed data. Another symptom of the overestimation is the location of the poles of the model in the far left half plane. The procedure here is to ignore any poles with large real part before evaluating the residues A_k . Poles whose time constants are less than 0.1 msec., may simply be ignored as having insignificant effect on the speech reconstruction. With the above procedure an adaptive adjustment of system order is permitted. Such a feature would not be possible with the iterative model.

The results of the closed form model with adaptive adjustment of system order are shown in Figure 26, for the words, "NOON", "DAY", "SLEEPY", and "HOW ARE YOU".

2.5. Summary

The evaluation of the predictor parameters by either the square root method or the Gaussian elimination is fairly straightforward. The computational time can be considerably reduced by using the techniques indicated in section 2.1.(1). The problem of singular solutions did not arise with either the square root or the Gaussian elimination method. The computational times for the two methods of solving a set of linear system of equations were approximately 5 msec. for a 12th order system. As shown in Figure 12, the frequency response of the linear predictor, obtained by the two different methods, is identical. Thus, there is no real advantage with any one method for solving a linear system of equations.

The impulse excitation of the linear predictor on the whole did

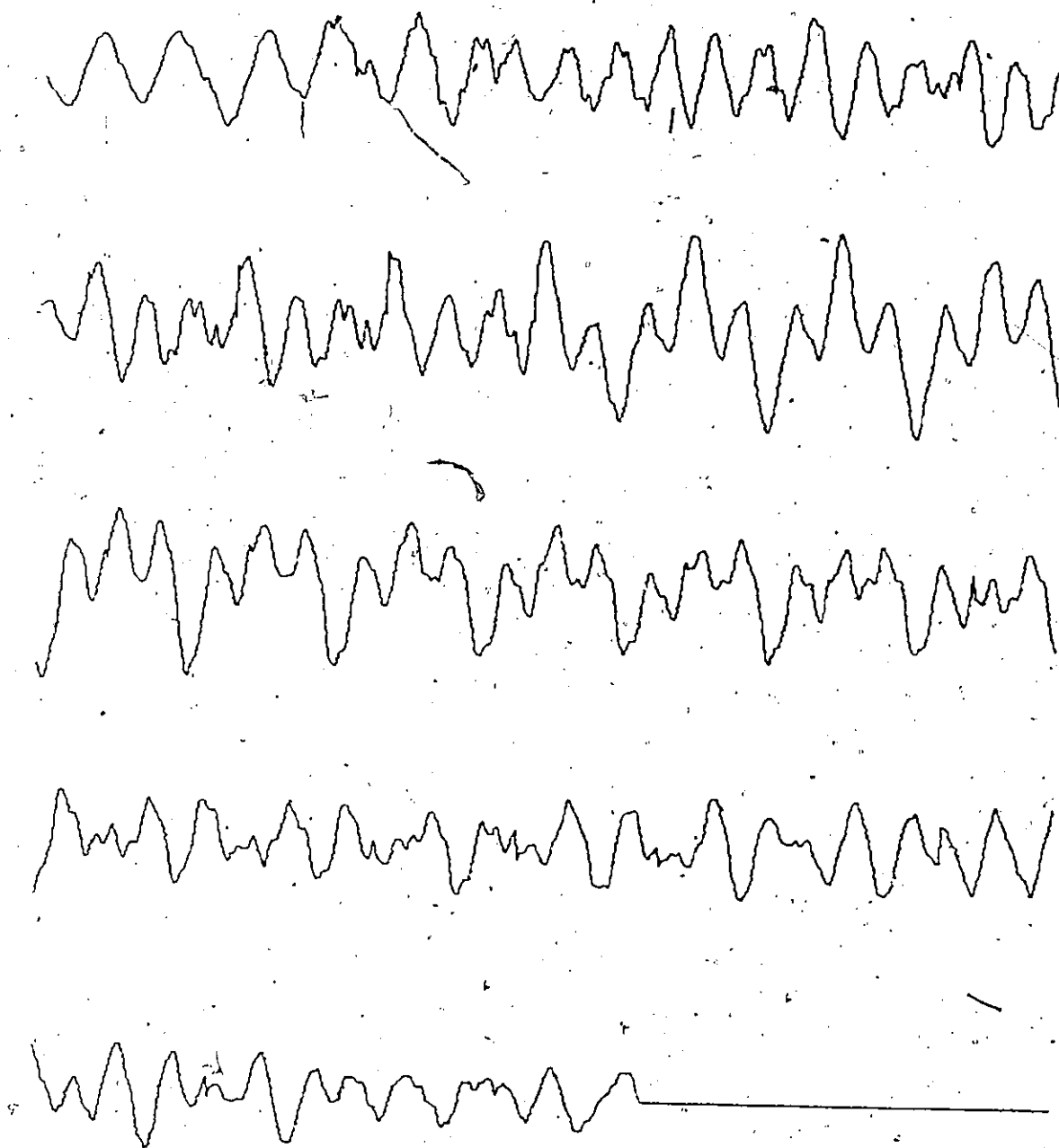


FIGURE 26 (a) - Closed Form Model With Adaptive Order Of System

Reconstruction Of The Word - "NOON"

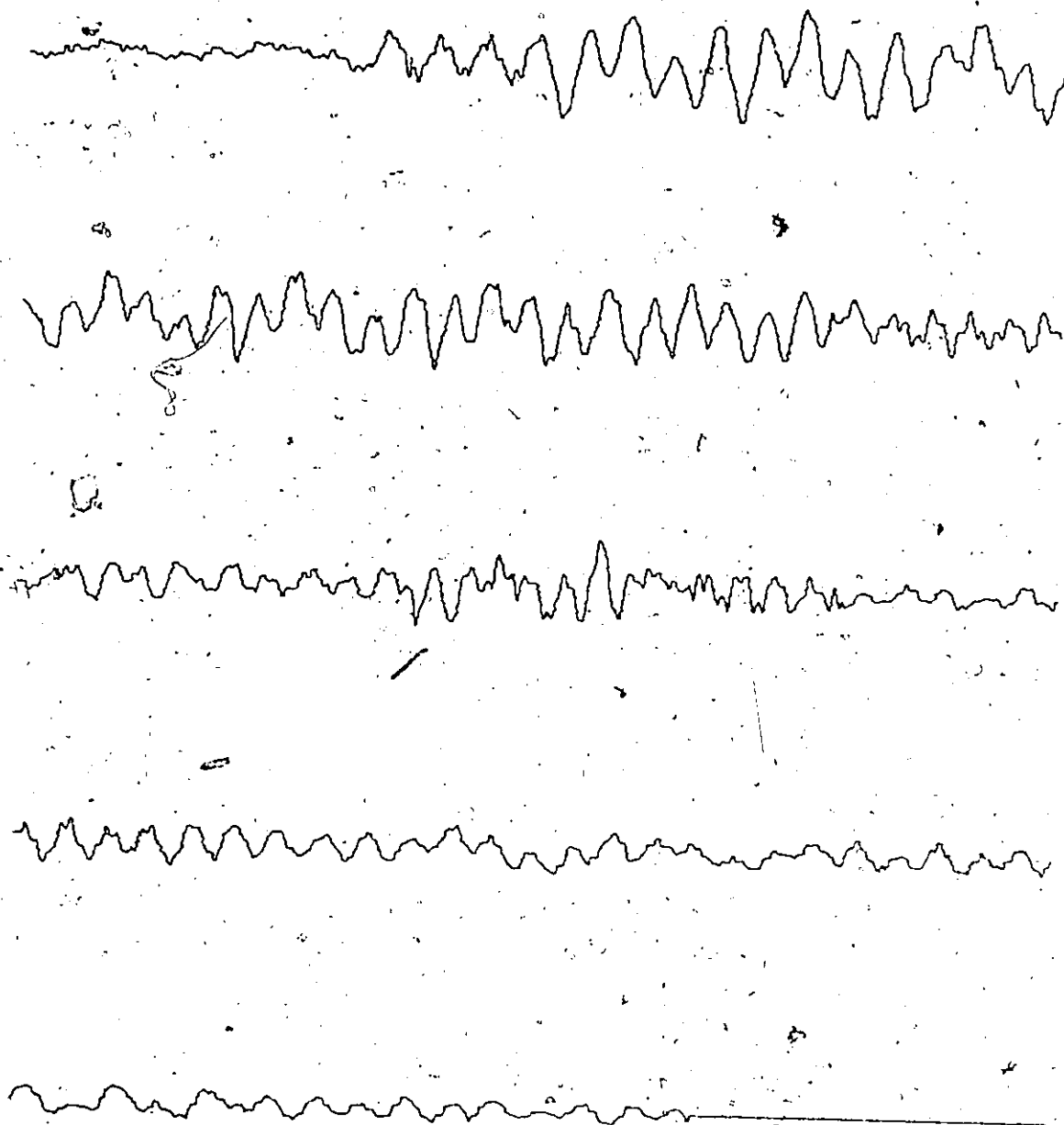


FIGURE 26 (b) - Closed Form Model With Adaptive Order Of System

Reconstruction Of The Word - "DAY"

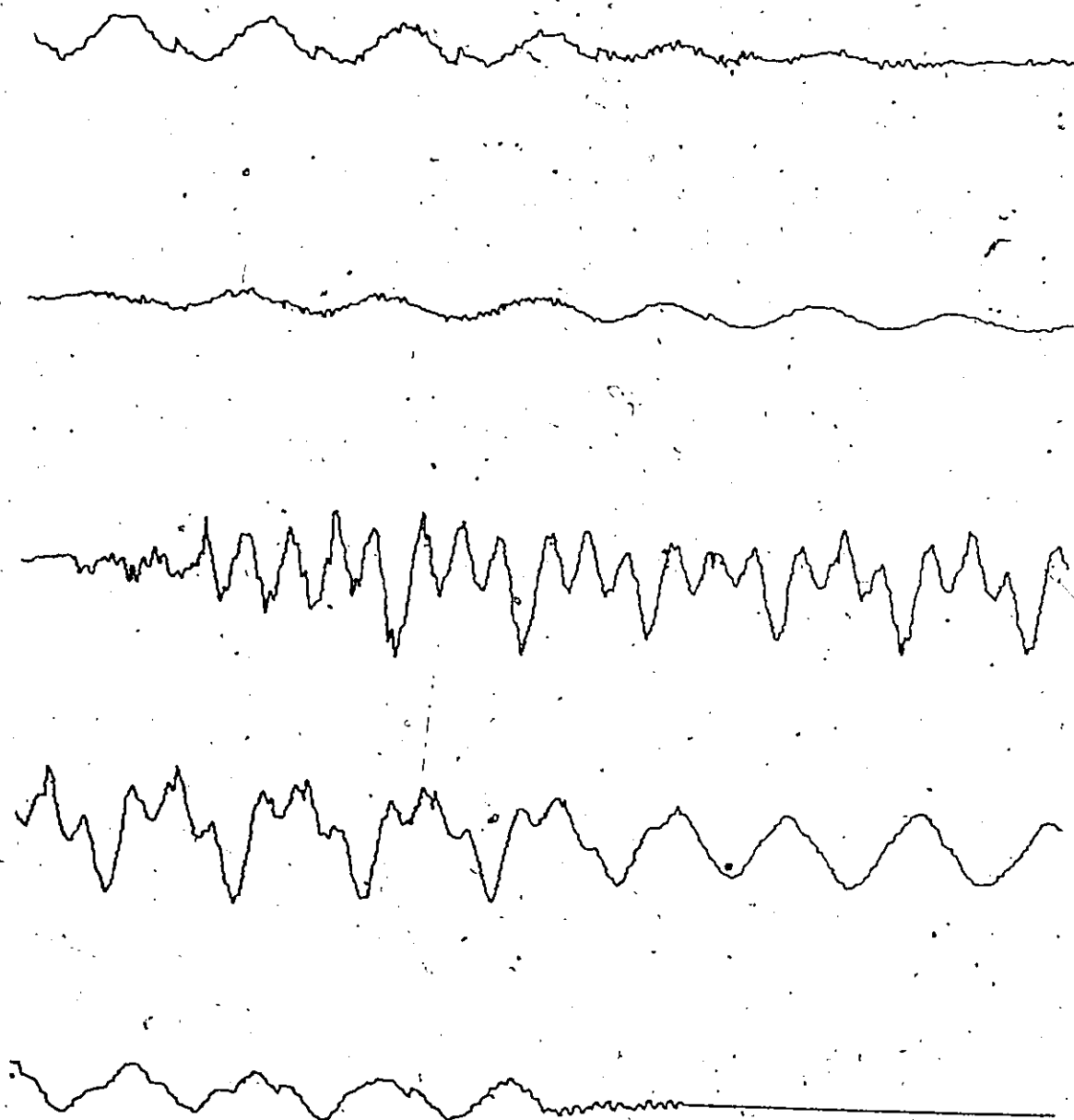


FIGURE 26 (c) - Closed Form Model With Adaptive Order Of System

Reconstruction Of The Word - "SLEEPY"

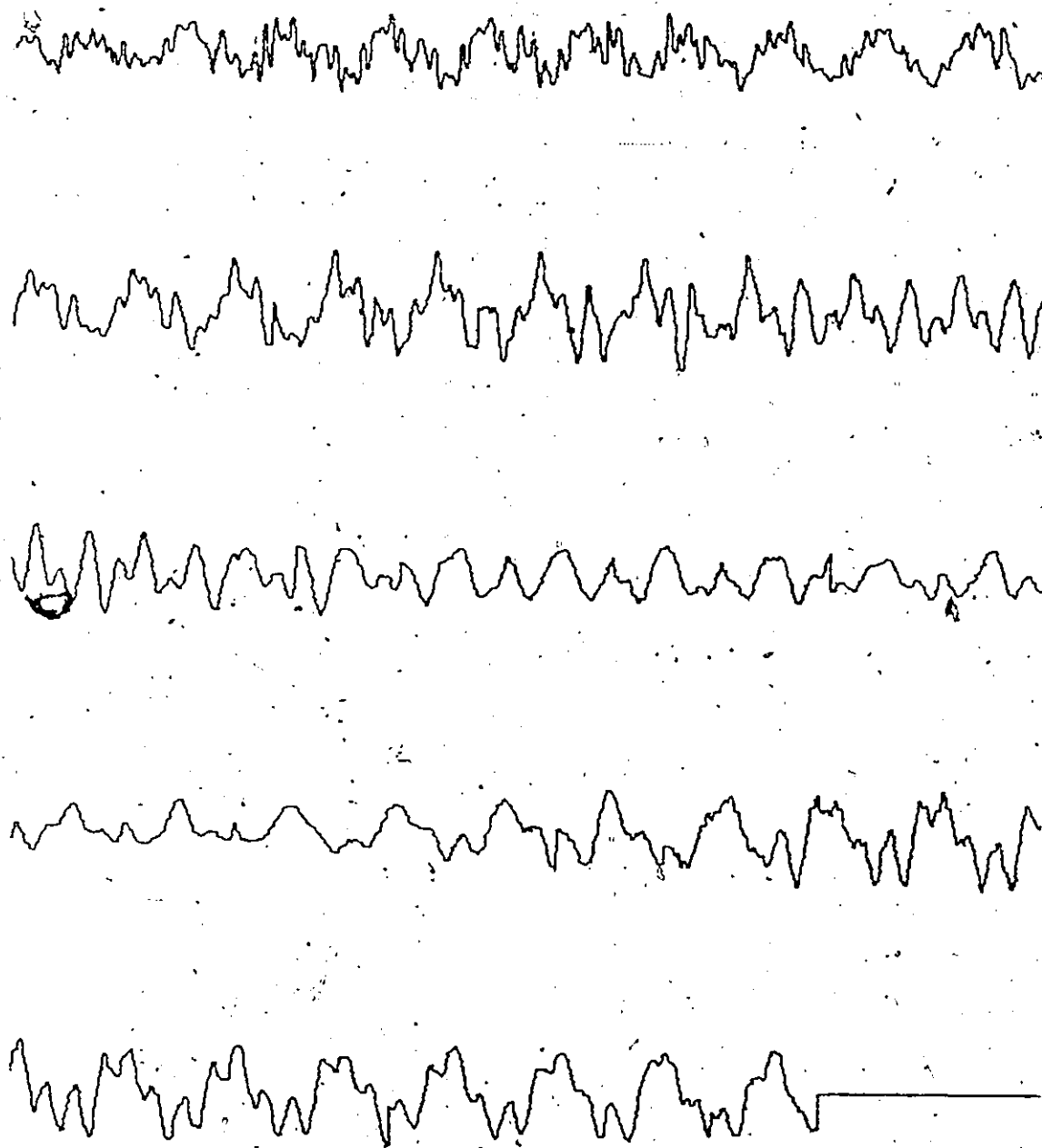


FIGURE 26 (d) - Closed Form Model With Adaptive Order Of System

Reconstruction Of The Word - "HOW ARE YOU"

not produce very good results. In some portions of the word "NOON", Figure 20(a), the reconstruction seems very promising. The use of an excitation function other than an impulse is recommended for high quality speech synthesis.

The use of a triangular pulse for the excitation is an improvement over the impulse, as can be seen by comparing the results of Figures 20 and 21. On the whole, triangular excitation produced better results in all the samples tested. One noticeable feature with both types of excitation is that the reconstructed signal was much smoother than the original waveform.

The inclusion of zeros in the predictor model tended to produce an oscillatory reconstructed signal. This is very much apparent in the reconstructed waveform of the word "SLEEPY", Figure 22(c), in the last block. The inclusion of zeros in the linear predictor model did not produce a significant improvement over that obtained with the impulse excitation which uses an all-pole model.

The use of an all-pole model with the appropriate initial conditions, gave good results, see Figure 23. As mentioned previously, if the initializing is done at a point where the speech signal is low in amplitude, the resulting reconstruction will be poor.

The other noticeable feature in the reconstructed signal is a slight loss in the signal amplitude. This can be compensated without too much difficulty.

By far the best results were obtained with the parallel realization and the closed-form models (the two are essentially identical, except

in the reconstruction). This is evident by comparing the reconstructed signals of Figures 25 and 26. The improvement in the reconstructed signal is due to,

- i) the reconstruction is no longer recursive (closed form model)
- ii) the effect of the speech signal zeros has been included in the residues.

The parallel realization form, while being recursive, does not have the same error build up as the direct realization models. The estimate of the n th sample only depends on the n th-1 sample, instead of the previous p samples. Because of this, quantization effects are not as severe.

The use of an adaptive system order has been shown to be feasible. This feature makes it possible to prevent overestimation of the system order, which could otherwise create problems.

With the triangular pulse excitation, the following model was found to give the best results.

$$\phi_{jk} = \sum_{n=1}^{N-j} s_n s_{n+j} \quad (2.52)$$

Similarly, for the other techniques, the pole-zero, initial conditions, the parallel and the closed form realization, the following model gave the best results.

$$\phi_{jk} = \sum_{n=p+1}^N s_{n-j} s_{n-k} \quad (2.53)$$

CHAPTER III

CODING OF SPEECH SIGNALS

3. Coding of Speech Signals

In the transmission of speech signals, there is a large amount of redundant information, which can be removed before the signal is transmitted. The removal of any redundancy in a speech waveform will result in bandwidth compression. Coding of speech signals is a means of achieving bandwidth compression, and in doing so, more efficient channel utilization is effected. A means of transmitting the same information but using less channel space has been the aim of numerous coding schemes. Two of the more recent coding schemes are analyzed in this chapter, and means of improving their performance are indicated.

3.1. Predictive Coding of Speech

For many classes of information signals, the value of the signal at a given instant can be expressed as a linear combination of its immediate past values. This represents redundant information, which if eliminated results in data compression. Several digital techniques [16, 17, 37], for data compression have recently been proposed, that utilize a time varying linear predictor for forming estimates of information samples.

An error signal, e_i , between the samples of an input sequence S_i and linear estimates of these samples \hat{S}_i is obtained. If the energy of the error signal sequence is small compared to the energy of the input sequence S_i , digital encoding of error samples would require a fewer number of bits, compared to 8 bits per sample for an input sequence of speech signals. In a typical coding scheme, the error signal is transmitted at the same rate as the input sample sequence (10 kHz), while the parameters

of the linear predictor, are transmitted at a much slower rate (40 Hz). Using a scheme similar to the above, digital encoding of speech signals at 15,000 bits/sec. have been reported [17]. However, the choice of optimum quantization levels for maximum signal-to-quantization noise ratio has not been established, owing to the multimodal and non-linear characteristics of the system.

In this chapter, a technique is presented, for digital encoding of the signals, which besides keeping quantization noise level low, enables the determination of optimum quantization levels.

3.1.1. Quantization Techniques

The choice of a suitable quantizer for digital encoding of speech signals is important in the realization of maximum signal to quantization noise ratio. A brief discussion of the different types of quantizer will be presented here. A quantizer, basically, transforms a continuous signal into predetermined discrete levels. A typical quantizer with 4 discrete levels is shown in Figure 27:

The input is a sinusoidal signal and the output is seen to be a staircase function.

Typically two types of quantizer are used for coding. The first type has a zero level and the number of levels is odd. The second type has no zero level and the number of levels is even. These are shown in Figure 28.

Besides the quantizers are also characterized as linear, non-linear, odd symmetric and non-symmetric. In this discussion only linear odd symmetric quantizers will be considered. For a quantizer with odd number

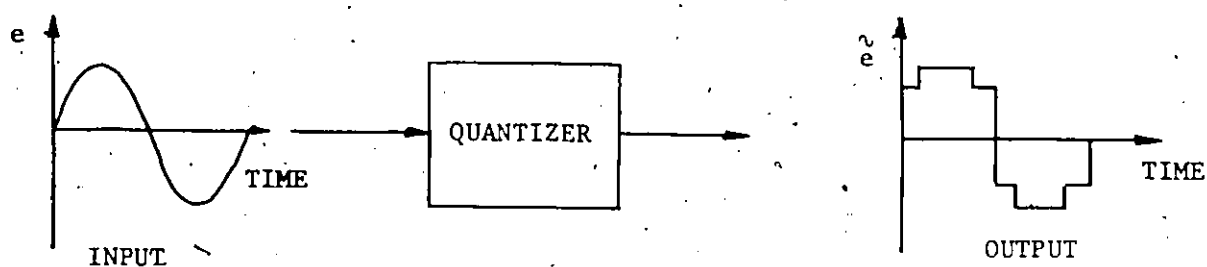


FIGURE 27 - Input/Output Relationship Of A 4-Level Quantizer

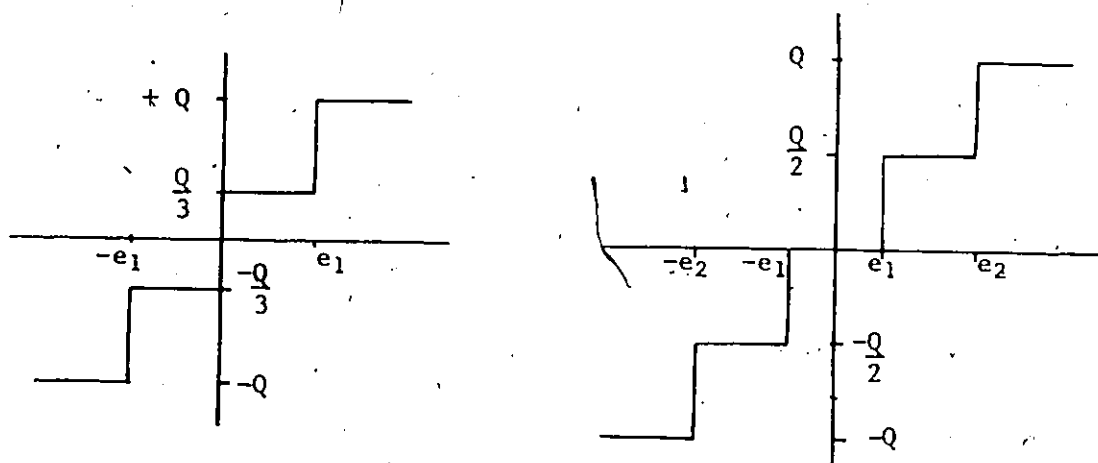


FIGURE 28 - Odd And Even Number Of Quantizer Levels

of levels the output is given by

$$\hat{e}_n = \frac{Q}{N-1} \left[\sum_{i=1}^{\left\lfloor \frac{N-1}{2} \right\rfloor} \{ \text{sgn}(e_n - e_i) + \text{sgn}(e_n + e_i) \} \right] \quad (3.1)$$

where $\text{sgn}(e) = 1$, $e \geq 0$

$= -1$ $e < 0$

Q is the maximum output of quantizer

$\pm e_i$ are the transition points ($i = 1, \dots, \left\lfloor \frac{N-1}{2} \right\rfloor$)

N is the number of levels.

For a quantizer with even number of levels

$$\hat{e}_n = \frac{Q}{N} \left[2 \text{sgn}(e_n) + \sum_{i=1}^{\left(\frac{N}{2} - 1 \right)} \{ \text{sgn}(e_n + e_i) + \text{sgn}(e_n - e_i) \} \right] \quad (3.2)$$

In the general case of a linear odd symmetric quantizer, the output \hat{e}_n may be expressed as

$$\hat{e}_n = Q \cdot A_n \quad (3.3)$$

where A_n is obtained from equation (3.1) or (3.2)

The optimum value of Q is determined by minimizing

$$E = \frac{1}{M} \sum_{n=1}^M (e_n - \hat{e}_n)^2 \quad (3.4)$$

Differentiating equation (3.4) with respect to Q and equating to zero, one obtains

$$\frac{\partial E}{\partial Q} = \frac{-2}{M} \sum_{n=1}^M (e_n - Q A_n) A_n = 0 \quad (3.5)$$

rearranging

$$Q = \frac{\sum_{n=1}^N A_n e_n}{\sum_{n=1}^M A_n^2} \quad (3.6)$$

Example:

$$N = 2$$

$$\tilde{e}_n = Q \operatorname{sgn}(e_n) \quad (3.7)$$

where $Q = \frac{1}{M} \sum_{n=1}^M |e_n|$ is optimum.

Thus Q optimum may be easily determined if the transition points are predetermined. The choice of transition points is dependent on the nature of the application; for speech coding, the transition points may be chosen to be a linear function of the RMS value or the mean of the absolute value of the signal, that is

$$e_n = K \sqrt{\frac{1}{M} \sum_{n=1}^M S_n^2} \quad (3.8)$$

$$\text{or } e_n = \frac{K}{M} \sum_{n=1}^M |S_n| \quad (3.9)$$

where K is a constant that would have to be determined empirically.

3.2. Open Loop Predictive Coding

Figure 29 shows a digital implementation of a predictive coding scheme for voice signals. The linear predictor forms an estimate \hat{S}_n of the input signal S_n , by a linear combination of p past samples,

$$\hat{S}_n = \sum_{k=1}^p a_k S_{n-k} \quad (3.10)$$

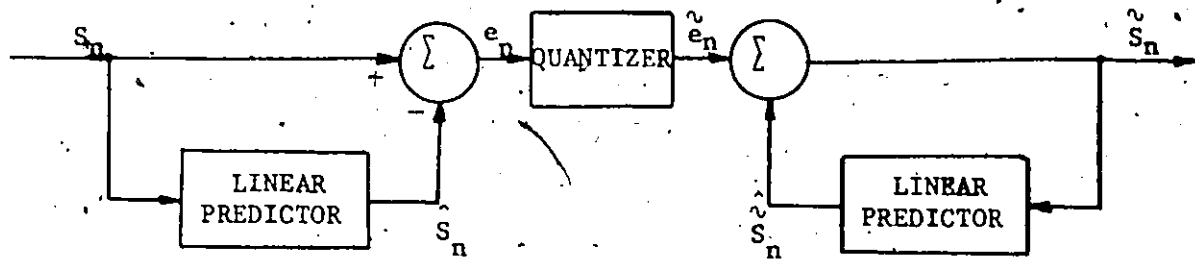


FIGURE 29 - Open Loop Predictive Coding

Since the voice signal is non-stationary, the weighting coefficients a_k are updated periodically. Typically a value for p is 10, while the coefficients a_k are updated every 24 msec. for speech signals sampled at 10 kHz. The coefficients a_k are determined by minimizing the energy of the error signal $e_1 = S_1 - \hat{S}_1$.

In digital transmission, the predictor coefficients and the error samples are quantized to achieve data compression. At the receiver, the transmitted data is used to reconstruct estimates of the speech samples. Referring to the scheme shown in Figure 29, the local estimate of input samples is given by

$$\hat{S}_n = \sum_{k=1}^P a_k S_{n-k} \quad (3.11)$$

The error in the estimate is

$$e_n = S_n - \hat{S}_n \quad (3.12)$$

The transmitted error signal

$$\tilde{e}_n = [e_n]_{\text{quantized}} \quad (3.13)$$

The error in the reconstructed signal at the receiver, neglecting the quantizing effects on a_k , is given by

$$S_n - \hat{S}_n = e_n - \tilde{e}_n + \hat{S}_n - \hat{\hat{S}}_n \quad (3.14)$$

The error term contains not only the quantizing noise of the error signal, but also includes the quantizing error in the estimate. This noise may accumulate as the sequence length increases, resulting in poor quality of the reconstructed signal.

3.3. Differential Pulse Code Modulation

In order to overcome the cumulative effects of quantization error in the estimate, a differential pulse code modulation scheme proposed by McDonald[16] was studied. This scheme is shown in Figure 9. It can be shown that the error in the reconstructed signal at the receiver for the above scheme is the same as that in the error signal.

At the receiver

$$\hat{\hat{S}}_n = \hat{S}_n + \tilde{e}_n \quad (3.15)$$

$$\hat{\hat{S}}_n = \sum_{k=1}^P a_k \hat{\hat{S}}_{n-k} \quad (3.16)$$

now
$$e_n = S_n - \hat{\hat{S}}_n \quad (3.17)$$

and $\tilde{e}_n = [e_n]_{\text{quantized}}$ (3.18)

The error in the reconstructed signal is given by

$$S_n - \hat{S}_n = S_n - \hat{S}_n + \hat{S}_n - \tilde{S}_n \quad (3.19)$$

$$= e_n - \tilde{e}_n \quad (3.20)$$

Equations (3.20), shows that the quantization noise at the receiver does not accumulate. However, this advantage is offset by the difficulty in predetermining the optimum quantization levels for maximum signal-to-quantizing noise ratio. The differential pulse code modulation scheme is highly sensitive to the quantization level and the problem of determining the global minimum of the mean squared error between e_n and \tilde{e}_n is still to be solved. An arbitrary choice for the quantizer level has been found to also lead to instability in the coding scheme[40].

3.4. Two Stage Adaptive Predictive Coding

It is seen from the above discussion that both the open-loop and the differential pulse code modulation have certain disadvantages that render the practical implementation of data compression techniques difficult. It is further observed that the open-loop quantizer enables the predetermination of optimum quantizer levels, while the differential pulse code modulation avoids the cumulative effects of quantizing noise due to error in estimate at the receiver.

In order to combine the advantages of both open-loop and differential pulse code modulation, the scheme shown in Figure 30 is proposed. In this scheme, differential pulse code modulation using a two-level quantizer is used to produce the sequence \tilde{S}_n . The sequence \tilde{S}_n is compared with S_n and

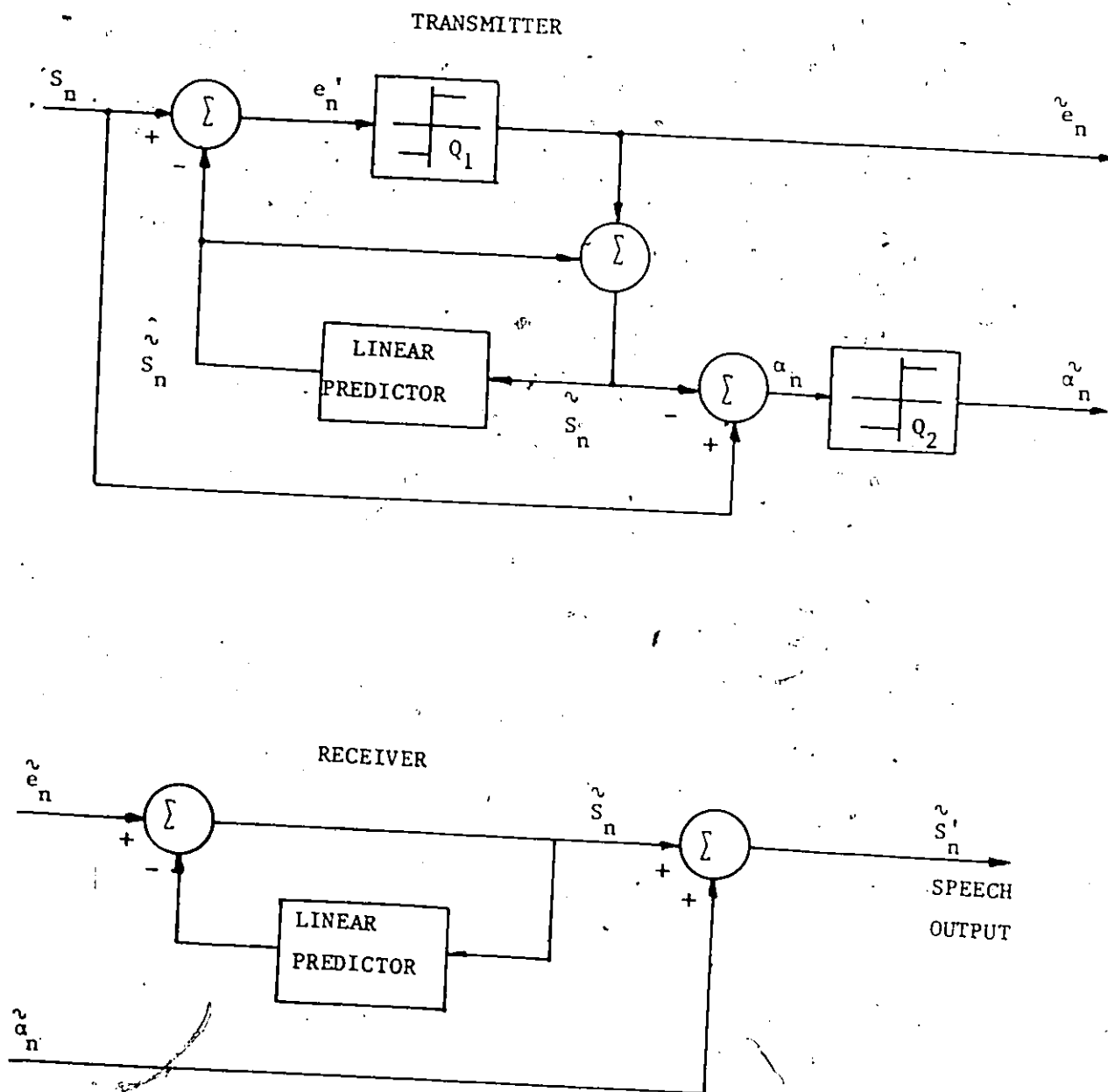


FIGURE 30 - Proposed Two Stage Predictive Coding Scheme

the resulting error is optimally quantized into two levels. Thus the overall scheme results in a two bit quantization of the sequence S_n . The analysis of this scheme is described in the next section.

3.4.1. Primary Stage

In this stage, a differential pulse code modulation of S_n is performed. Referring to Figure 30, the quantizer level Q_1 is chosen to be

$$Q_1 = \frac{1}{N_1} \sum_{n=1}^{N_1} \left| S_n - \sum_{k=1}^P a_k S_{n-k} \right| \quad (3.21)$$

where N_1 is the interval over which Q_1 is constant. For this interval, the sequence \tilde{S}_n is given by

$$\tilde{S}_n = \sum_{k=1}^P a_k \tilde{S}_{n-k} + Q_1 \operatorname{sgn} e_n \quad (3.22)$$

where $e_n = S_n - \sum_{k=1}^P a_k \tilde{S}_{n-k}$

and $\tilde{e}_n = [e_n]_{\text{quantized}} = Q_1 \operatorname{sgn} e_n$

The choice of Q_1 as given by equation (3.21) is based on the following reasons,

1. The optimum value of Q_1 for maximum signal-to-quantizing noise ratio is very difficult to determine, owing to the non-linear characteristics of the quantizer and existence of a closed loop around the quantizer.
2. The original sequence S_n is given by

$$S_n = \sum_{k=1}^P a_k S_{n-k} + \left(S_n - \sum_{k=1}^P a_k S_{n-k} \right) \quad (3.23)$$

The reconstructed sequence \tilde{S}_n is

$$\tilde{S}_n = \sum_{k=1}^p a_k \tilde{S}_{n-k} + Q_1 \operatorname{sgn} e_n \quad (3.24)$$

$$\text{where } e_n = S_n - \sum_{k=1}^p a_k \tilde{S}_{n-k}$$

A comparison of equations (3.23) and (3.24) reveals that if $Q_1 \operatorname{sgn} e_n$ is chosen to be a good approximation to $S_n - \sum_{k=1}^p a_k \tilde{S}_{n-k}$, then the reconstructed sequence \tilde{S}_n will be close to S_n . Since $\operatorname{sgn} e_n = \pm 1$, the choice of Q_1 is chosen to be an optimum estimate of the magnitude of $S_n - \sum_{k=1}^p a_k \tilde{S}_{n-k}$ in the analysis interval. The value of Q_1 is obtained from equation (3.21).

3.4.11. Secondary Stage

In this stage, the reconstructed sequence \tilde{S}_n of the primary stage is compared with S_n and the resulting error is quantized by a 1 bit quantizer. Since this stage represents an open loop coding scheme, the quantizer levels can be optimized for maximum signal-to-quantizing noise ratio. Thus for the secondary stage

$$\alpha_n = S_n - \tilde{S}_n \quad (3.25)$$

If α_n is quantized by a 1 bit quantizer, then, the quantizer output is given by

$$\tilde{\alpha}_n = Q_2 \operatorname{sgn} \alpha_n \quad (3.26)$$

The optimum value of Q_2 is shown to be

$$Q_2 = \frac{1}{N_2} \sum_{n=1}^{N_2} |\alpha_n| \quad (3.27)$$

where N_2 is the interval of optimization.

The output of the secondary stage is,

$$\hat{S}_n' = \hat{S}_n + \hat{\alpha}_n \quad (3.28)$$

The overall error at the output of the second stage is

$$S_n - \hat{S}_n' = S_n - \hat{S}_n + \hat{S}_n - \hat{S}_n' = \alpha_n - \hat{\alpha}_n \quad (3.29)$$

Thus there is no build up of quantization error. The secondary stage has been found to yield on an average a three fold improvement in the signal-to-noise ratio over the primary stage.

3.5. Further Observations

The differential pulse code modulation scheme shown in Figure 9 can be redrawn as shown in Figure 31. This is seen to be a closed-loop

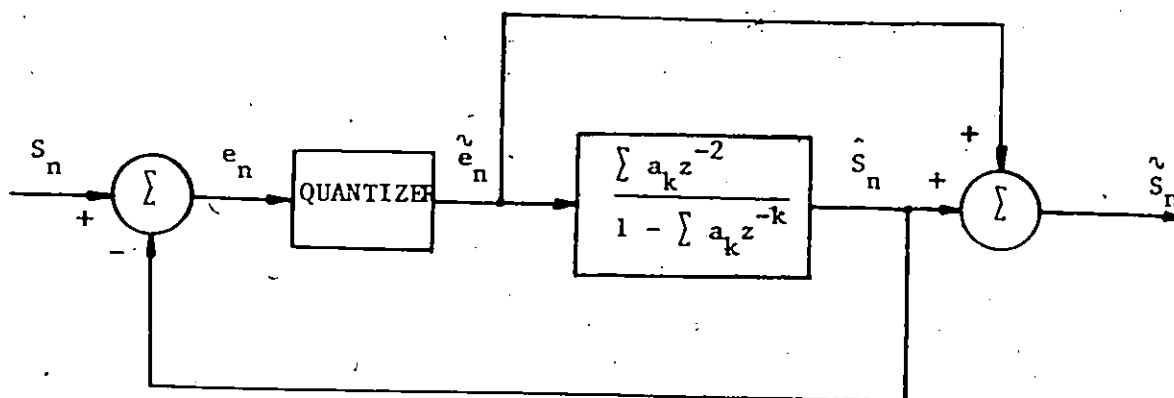


FIGURE 31 - Equivalent System Of The DPCM

feedback system with a non-linear element in the forward loop. This system can go unstable if the quantizer is chosen arbitrarily. A second observation concerns the derivation of the model parameters. If a model derived from a sequence of say 80 points is used to reconstruct a sequence of 240 points, instabilities might occur, owing to the inaccuracies of numerical computation and the finite precision of the data sequence. It is therefore recommended that a new model be derived for each segment of the speech sequence to ensure stability of the model.

If the number of samples in a given analysis interval is more than 128 points the determination of the predictor parameters by the covariance matrix may become time consuming. In such cases an alternate approach could be used which makes use of the power spectrum[11]. The procedure for this technique is as follows:

A segment of speech sequence (typically 80 samples at 10 kHz) is transformed by the FFT algorithm. The power spectrum of the speech samples is then obtained by squaring the magnitude of the FFT sequence. Thus

$$P(k\Omega) = \left| \sum_{n=0}^{127} S_n e^{-jk\Omega T} \right|^2, \quad k = 0, 1, \dots, 127 \quad (3.30)$$

where $S_n = 0 \quad n > 80$

$$P(k\Omega) = S(j\omega) \cdot S(-j\omega) \quad (3.31)$$

The inverse FFT of $P(k\Omega)$ is then given by

$$\phi(l) = \sum_{n=0}^{N-l} S_n S_{n+l} \quad (3.32)$$

where N is typically 80.

The parameters a_i of the linear predictor are then obtained by solving

$$\sum_{i=1}^p \phi(i-j) a_i = \phi(j) \quad (3.33)$$

where $S = 1, 2, \dots, p$.

Such a model using 80 samples, may be used to reconstruct 240 speech samples (at 10 kHz). Further improvement is possible if a new model is evaluated every pitch period. In cases where the analysis interval is taken as 240 speech samples, the evaluation of the linear predictor parameters by the power spectrum method will be slightly faster. The extraction of the pitch period is achieved by the use of cepstrum or the normalized autocorrelation methods.

3.6. Bit-Rate Requirements

The bit-rate requirements for a practical coding scheme are discussed in this section. For speech signal passed through a 5 kHz low pass filter and sampled at 10 kHz, the following bit rate is required:

- 1.(a) Model adjusted every 25 msec.
- (b) Number of parameters for model = 10
- (c) Number of bits per parameters = 12

Bit rate for model adjustment	4800 bits/sec.
Bit rate for primary stage	10000 bits/sec.
Bit rate for secondary stage	10000 bits/sec.
Bit rate for adjustment of Q_1 and Q_2	<u>2000 bits/sec.</u>
Total bit rate	26800 bits/sec.

The equivalent PCM requirement at 8 bits/word = 80,000 bits/sec.

The above analysis indicates that a 3:1 reduction is feasible.

An improvement in the quality of the reconstructed speech is attainable, at the cost of a higher bit rate, by the use of appropriate initial conditions, everytime a new model is evaluated, or by initializing the reconstructed signal every pitch period. For a coding scheme incorporating the above, the bit rate is evaluated as follows:

1. Model adjusted every 24msec.

(a) Number of parameters per model = 10

(b) Number of bits per parameter = 12

Bit rate for model adjustment	4800 bits/sec.
Bit rate for primary stage	10000 bits/sec.
Bit rate for secondary stage	10000 bits/sec.
Bit rate for adjustment of Q_1 and Q_2	2000 bits/sec.
Bit rate for initializing every 8 msec.	<u>3200 bits/sec.</u>
Total bit rate	30000 bits/sec.

The quality of the reconstructed speech can be further improved by using a higher order model for the linear predictor. The improvement in quality will be at the expense of a higher bit rate requirement. If the quality of reconstructed speech is not of utmost importance (as in ham radio or army field communication) the bit rate can be further reduced by using a 6 kHz sampling rate, the bit rate required for the two cases discussed will be 18,800 bits/sec. and 22,000 bits/sec. respectively. Further reductions are possible if one is willing to sacrifice quality.

3.7. Simulation Results

The criterion for the selection of the quantizer level in the first stage coding scheme is seen to give acceptable results. This can be verified by observing the reconstructed speech waveform of the primary coding stage, for the words "NOON", "DAY", "SLEEPY" and "HOW ARE YOU". These are shown in Figure 32.

With the inclusion of the secondary coding stage, the results are further improved. The results of the secondary stage are shown in Figure 33. On average a three fold increase in the signal-to-noise ratio can be expected by the addition of the secondary coding stage.

3.8. Summary

The proposed coding scheme is capable of reproducing the speech waveform extremely well. The use of a multi-level quantizer in the primary stage is not recommended due to the sub-optimal operation of this stage. The need for the secondary stage is evident by comparing the reconstructed waveforms of Figures 32(d) and 33(d) with the original waveform of Figure 19(d).

The use of the same linear predictor model for more than one analysis interval is possible if the system of equations (3.33) is used. The results of Figures 32 and 33 are shown with the model kept constant for three consecutive pitch periods. This procedure is not recommended for other models since it can lead to instabilities in both the primary and the secondary coding stages.

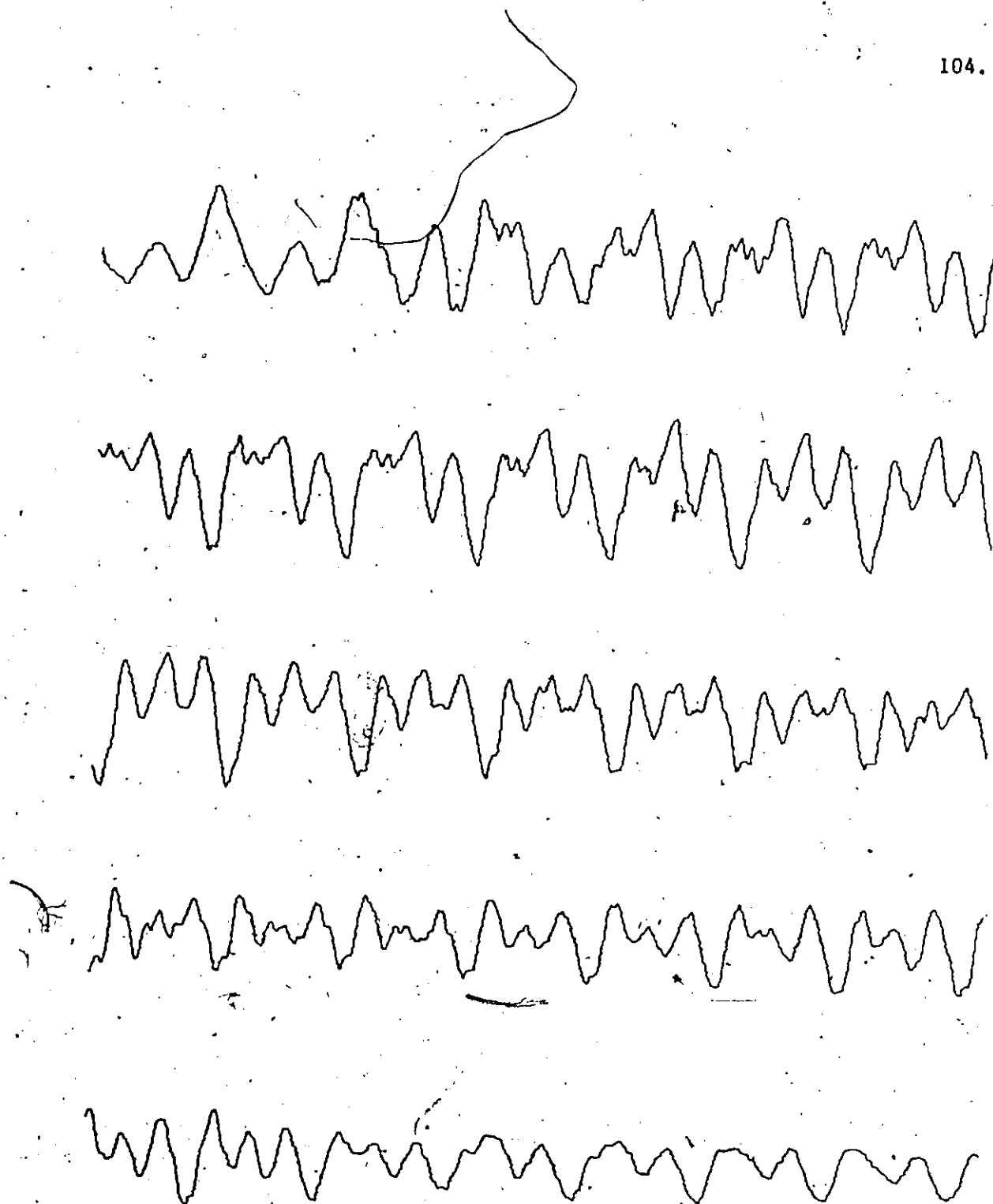


FIGURE 32 (a) - Primary Coding Stage Reconstruction Of The Word "NOON"

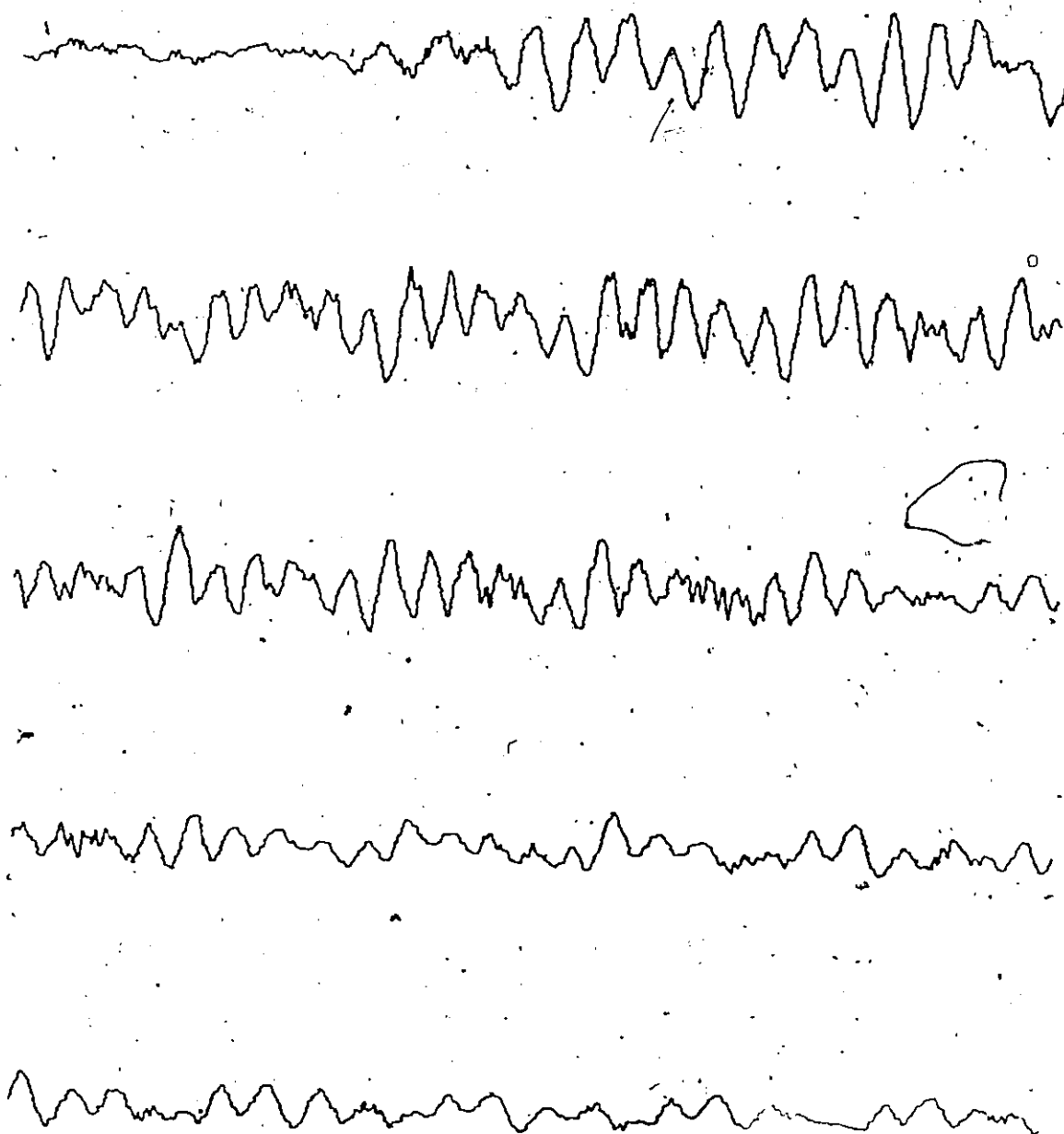


FIGURE 32 (b) - Primary Coding Stage Reconstruction Of The Word - "DAY"

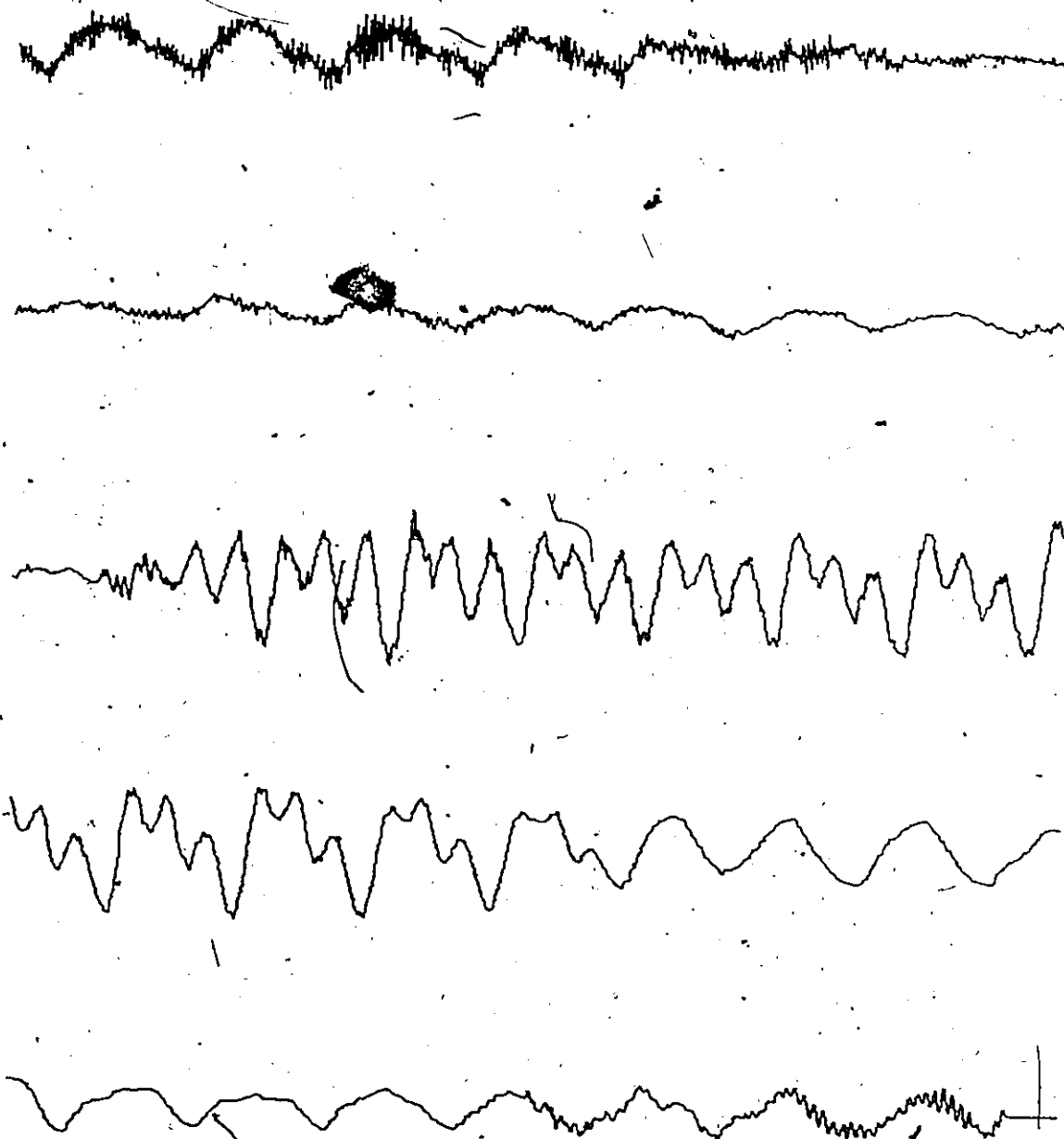


FIGURE 32 (c) - Primary Coding Stage Reconstruction Of The Word - "SLEEPY".

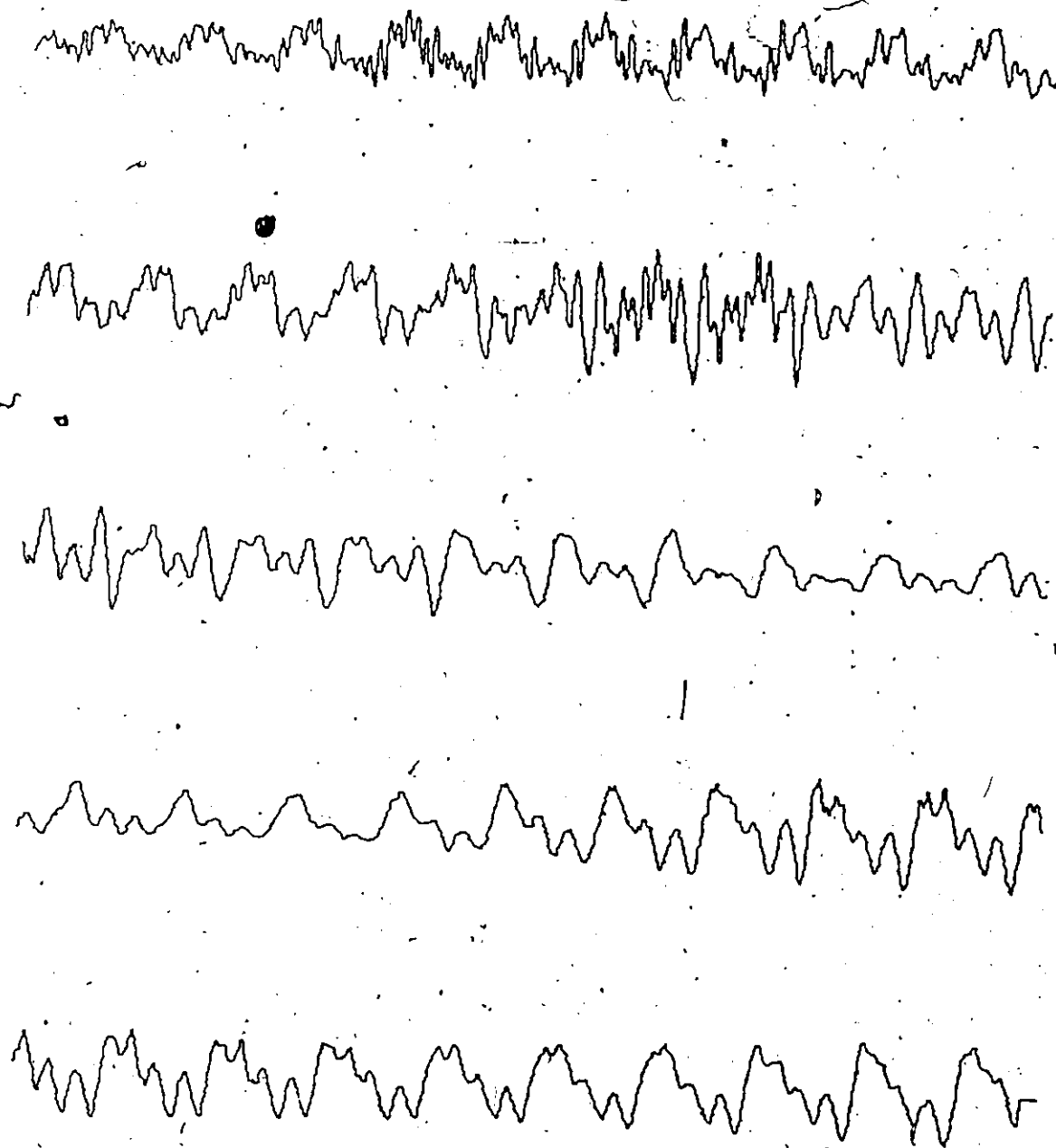


FIGURE 32 (d) - Primary Coding Stage Reconstruction OF The Word - "HOW ARE YOU"

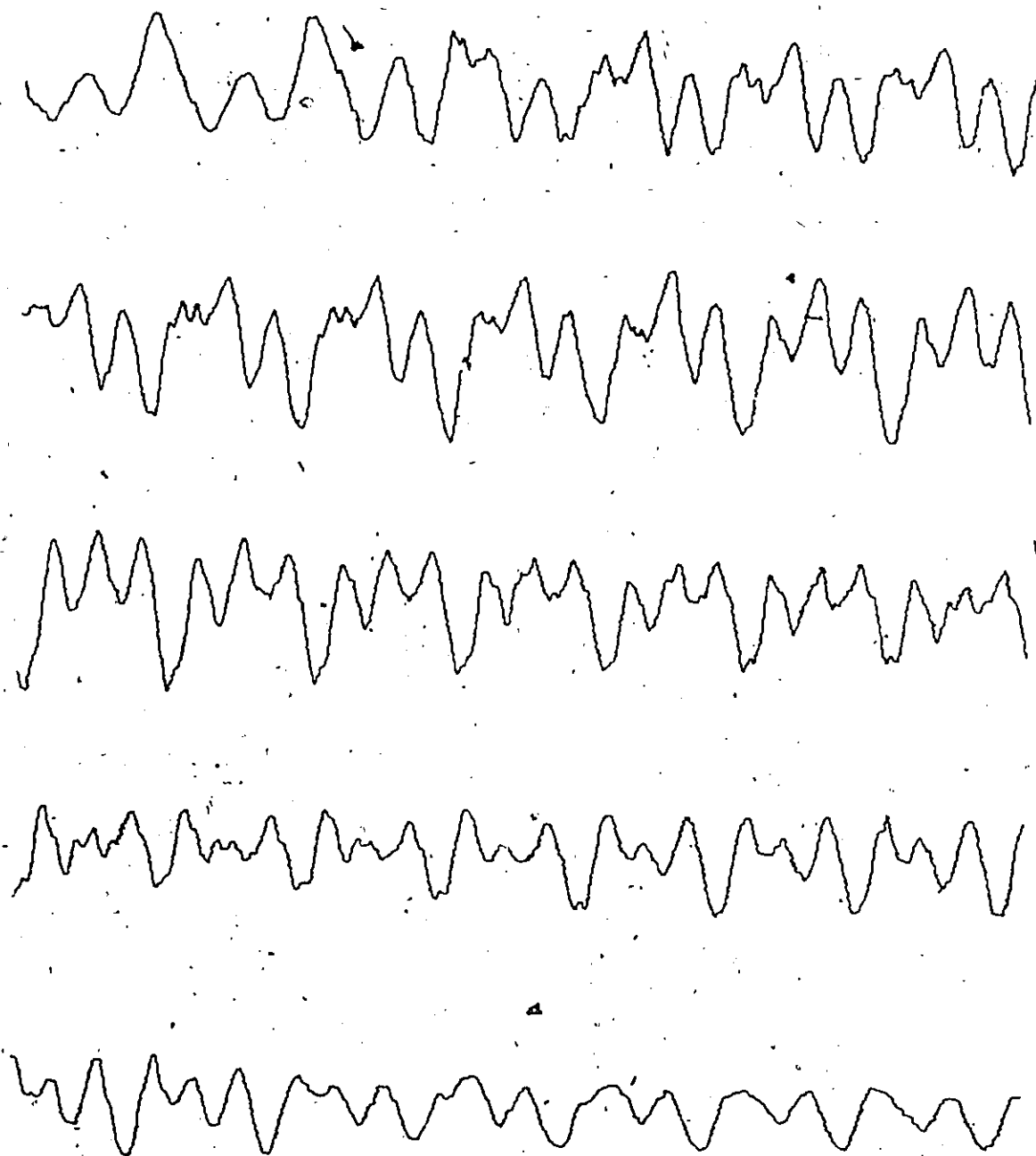


FIGURE 33 (a) - Secondary Coding Stage Reconstruction Of The Word - "NOON"

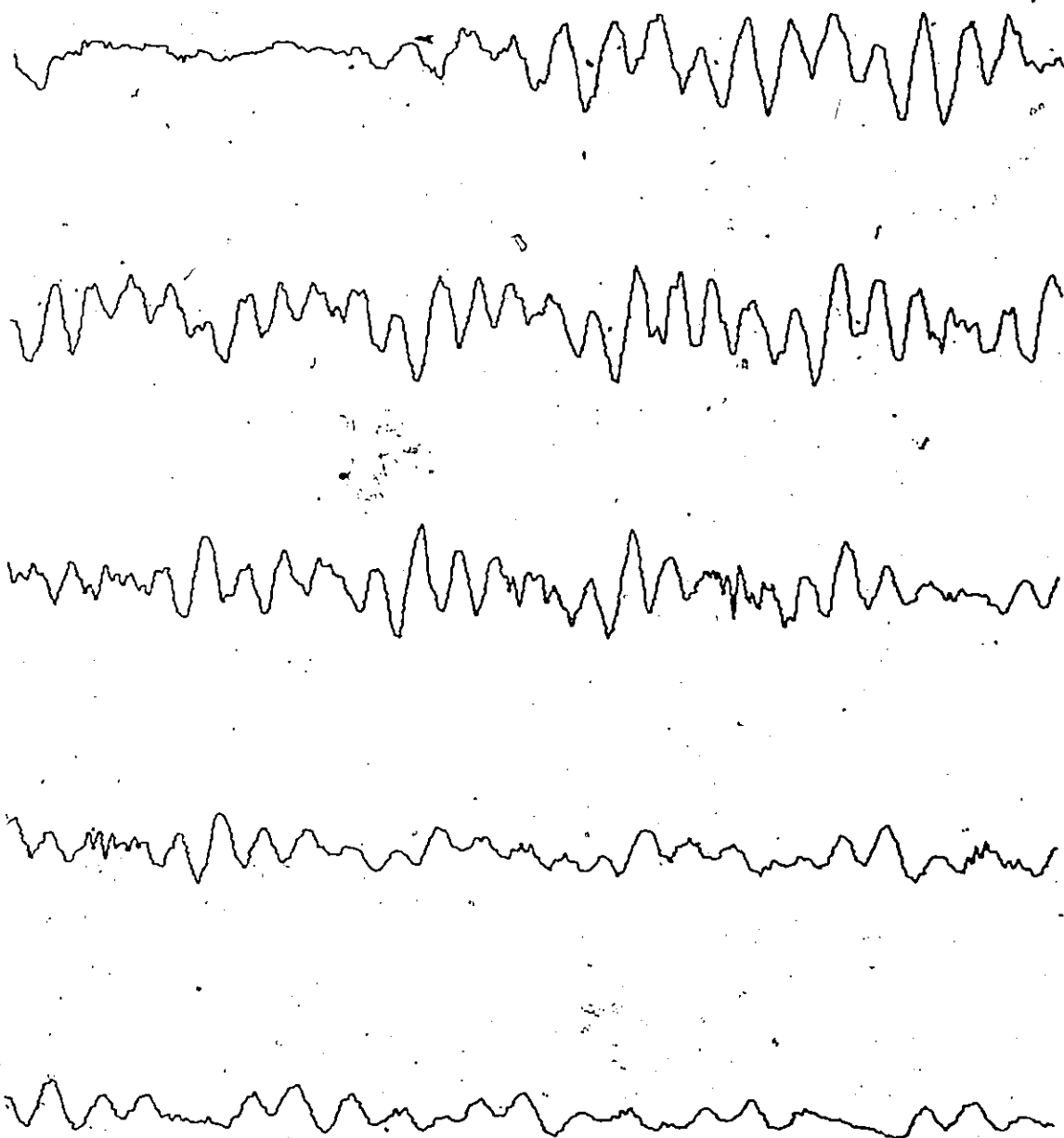


FIGURE 33 (b) - Secondary Coding Stage Reconstruction Of The Word - "DAY"

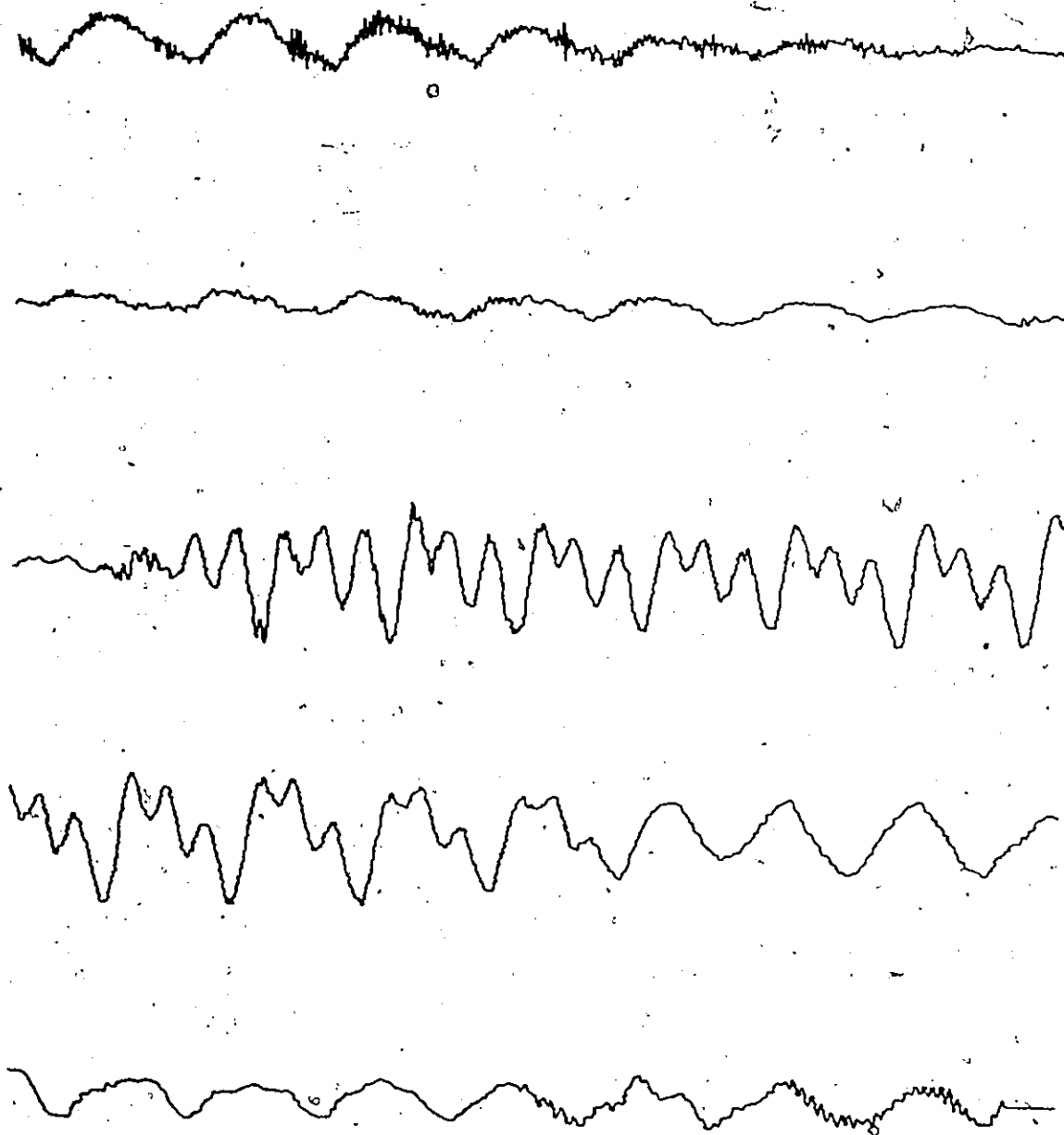


FIGURE 33 (c) - Secondary Coding Stage Reconstruction Of The Word - "SLEEPY"

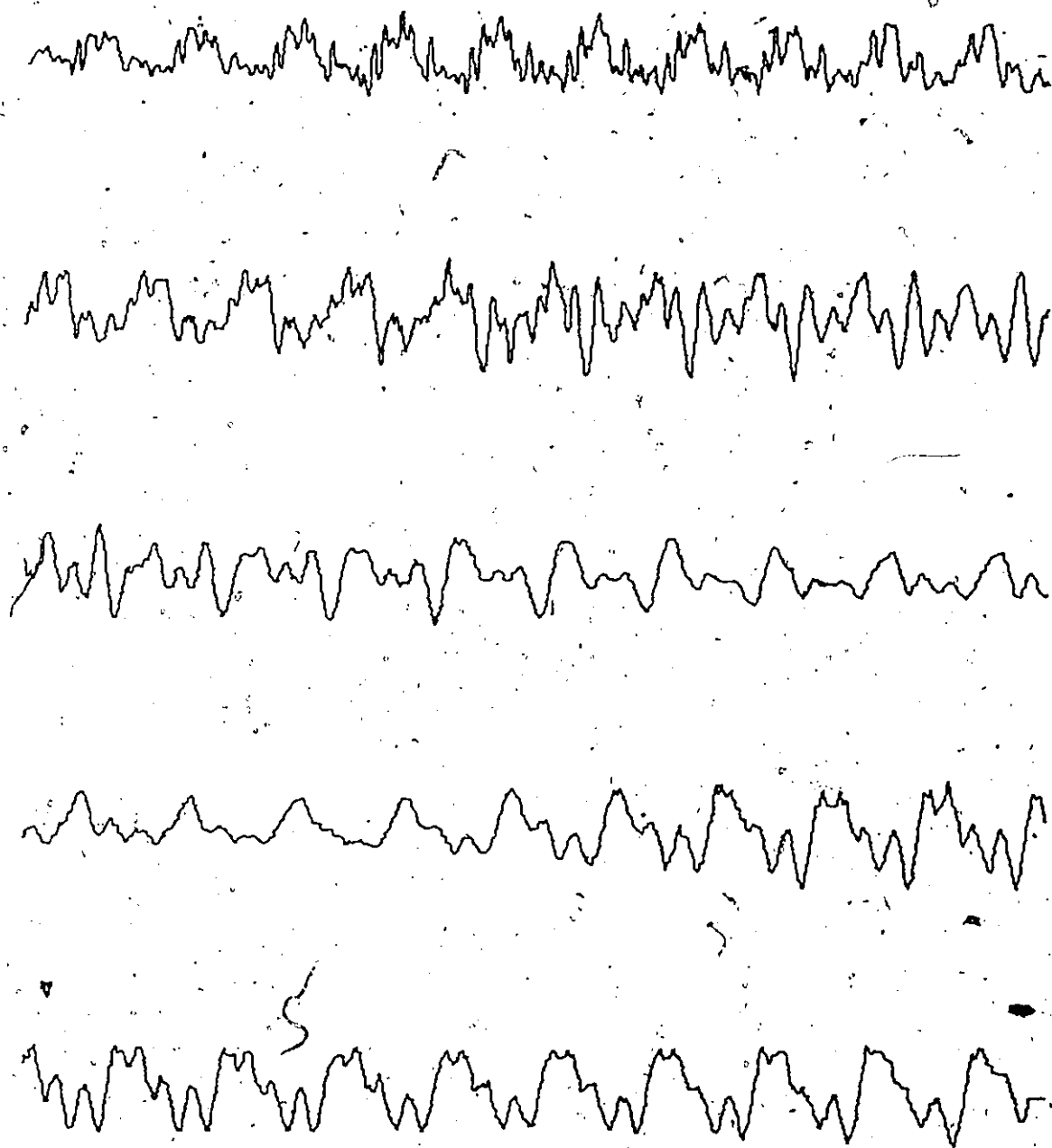


FIGURE 33 (d) ~ Secondary Coding Stage Reconstruction Of The Word - "HOW ARE YOU"

CHAPTER IV

FREQUENCY DOMAIN SPEECH SYNTHESIS AND CODING

4. Frequency Domain Speech Synthesis

One of the problems associated with most coding and synthesis techniques is the development of an accurate model for the speech production process. Contributing to this problem is also the voiced-unvoiced decision, and the determination of the pitch frequency of voiced sounds. A number of techniques using spectral and time domain information have been recently proposed that utilize a linear predictor for the speech production process. The limitations of the linear predictor have been discussed in Chapter II in some detail.

As pointed out in Chapter II, the linear predictor is an all-pole model and does not include the zeros of the vocal tract. Also, it is essentially a recursive digital filter which is prone to error build up because of its dependence on the past values of the speech signal. Any error in an estimate will eventually build up causing severe degradation in the reconstructed speech signal.

The technique discussed in this Chapter makes use of the speech signal frequency spectrum information. A speech production model is then formed which takes into account the effect of the poles and zeros on the speech sample. It is computationally very efficient and real time operation is possible with the appropriate hardware.

4.1. FFT Model For Speech Synthesis

The important parameters to be considered in speech production are,

1. Pitch period
2. Glottal pulse shape

3. Vocal tract resonances
4. Voiced-unvoiced excitation

The vocal tract resonances (called formants) are determined either from the poles of the linear prediction or the peaks in the short-time spectrum of the speech wave. The pitch period and the voiced-unvoiced decision are obtained from the short-time autocorrelation or cepstrum techniques. The glottal pulse shape is characterized by the presence of zeros in the overall model for speech production.

Thus quality speech synthesis requires the determination of all the above parameters accurately and this renders real-time implementation of speech synthesis and transmission practically unfeasible. In the proposed method that uses the Fast Fourier Transform algorithm, all the important parameters are determined as spectral peaks in predetermined frequency bands. Such a technique obviates the need for separate determination of the pitch period, vocal tract resonances, glottal zeros and the voiced-unvoiced decision.

The determination of the frequency bands is based on the following reasons:

1. The pitch frequency is usually in the range 50-200 Hz.
Hence a spectral peak in this range indicates the fundamental component of pitch excitation.
2. The vocal tract resonances occur in the range of 200-600 Hz., 600-2400 Hz and 2400-3200 Hz.
3. The glottal pulse shape is accounted for by the inclusion of additional bands in the 200-800 Hz to account for harmonics in the excitation sources.

4. For the reasons mentioned above in (3), voiced-unvoiced decision is not needed.

Based on the above reasoning, a typical arrangement of the bands is shown in Table 4.

TABLE 4

BAND	FREQUENCY RANGE Hz
1	50 - 100
2	101 - 200
3	201 - 400
4	401 - 800
5	801 - 1600
6	1601 - 3200

Speech Signal Frequency Spectrum Divided Into 6 Bands

For The FFT Model

Since a high computational speed is the main objective, the bands were kept fixed for the analysis of connected speech.

The speech signal frequency spectrum is divided into bands in a similar manner to those used in the calculation of the articulation index[38], except that in the latter case 20 bands are used. In both cases the bands are placed in the frequency range where the ear is the most sensitive. The articulation index is a measure of the quality of synthesized speech, and hence a larger number of bands are necessary. However, in speech synthesis, the number of bands can be reduced, while

still maintaining good quality reconstruction.

In this technique a short time spectrum of a segment of speech is obtained by the use of FFT algorithm. The resulting spectrum is scanned for peaks in predetermined bands. The amplitude and the frequency at which the peak occurs, form the parameters of the model for speech production. This procedure is illustrated in Figure 34.

The following analysis describes the derivation of the FFT model.

Let S_n , $n = 0, 1, \dots, N-1$ be the sequence corresponding to a segment of speech. The short-time discrete frequency spectrum of the speech sequence is given by,

$$S(k\Omega) = \sum_{n=0}^{M-1} S(nT) e^{-j\Omega nk} \quad (4.1)$$

where $k = 0, 1, \dots, M-1$

$$\Omega = \frac{2\pi}{M}$$

T = sampling interval

and $S_n = 0$, $n \geq M$

The frequency spectrum may be obtained efficiently on a digital computer by the use of FFT algorithm. Typically T is 10^{-4} seconds and $M = 2048$. For these values a resolution in the frequency spectrum of about 5 Hz is obtained. The parameters F_j and A_j are given by

$$A_j = S(k\Omega) \quad (4.2)$$

where k is found from $\text{Max}_k (|S(k\Omega)|)$

and $M_j\Omega < k\Omega < M_{j+1}\Omega$, $j = 1, 2, \dots, p$

also $M_j\Omega$ and $M_{j+1}\Omega$ are the end frequencies of the j th band.

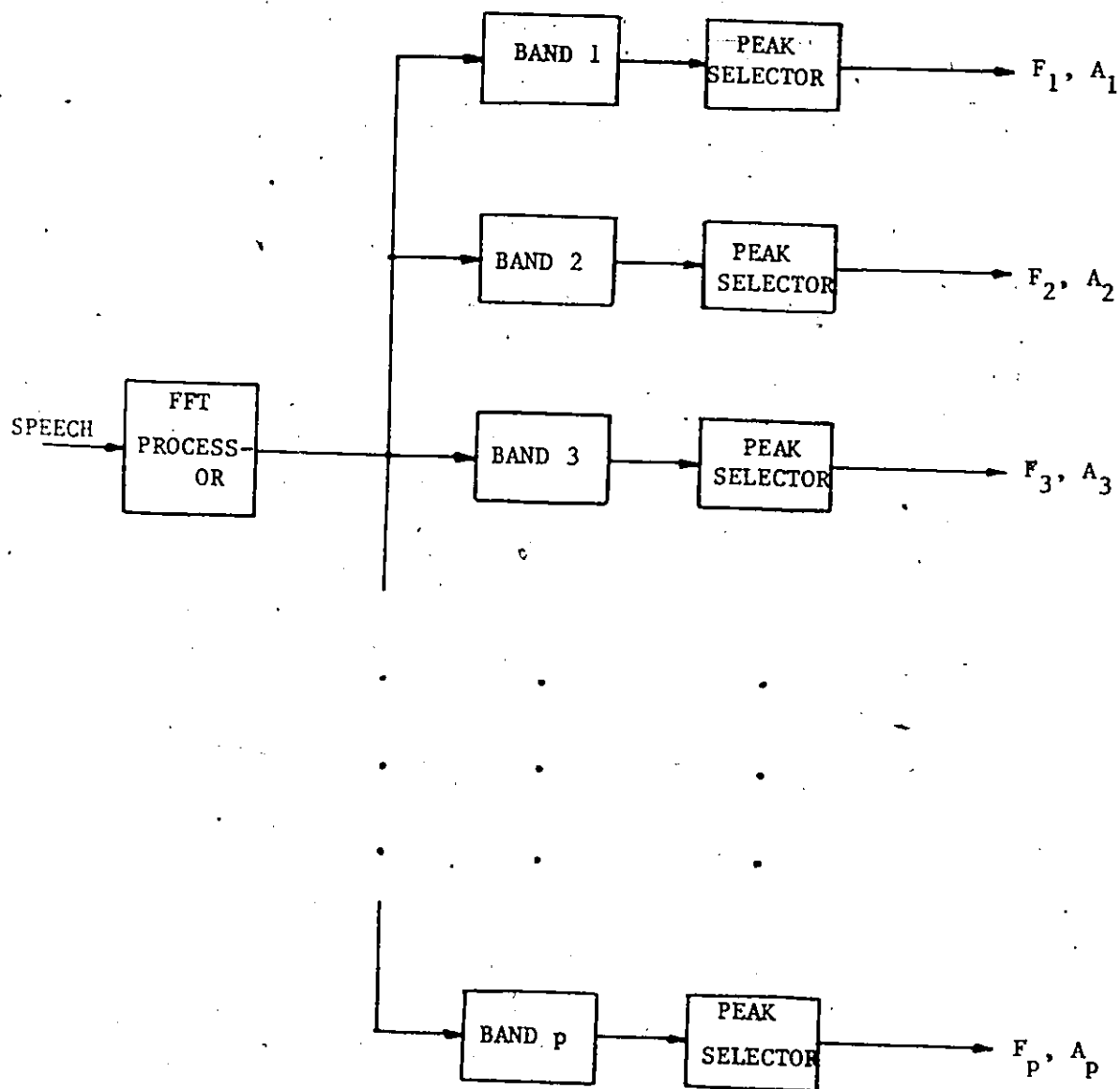


FIGURE 34 - FFT Model For Speech Analysis

$$F_j = \ell_j \Omega, \quad j = 1, 2, \dots, p \quad (4.3)$$

where ℓ_j is the value of k at which the peak occurs.

The real and imaginary parts of A_j are given by

$$\begin{aligned} R_e[A_j] &= R_e[S(\ell_j)] \quad j = 1, 2, \dots, p \\ I_m[A_j] &= I_m[S(\ell_j)] \end{aligned} \quad (4.4)$$

Once the parameters A_j and F_j are determined, the reconstructed speech is obtained as

$$\hat{S}_n = \sum_{j=1}^p R_e(A_j) \cdot \cos(F_j nT) + \sum_{j=1}^p I_m(F_j) \sin(F_j nT) \quad (4.5)$$

The reconstruction scheme is illustrated in Figure 35.

4.2. Digital Implementation

The length of the speech sequence is typically 256, and the number of points in the FFT computation is 2048. With a sampling frequency of 10 kHz, the resolution in the frequency spectrum

$$\Delta f = \frac{1}{NT} \approx 5 \text{ Hz.} \quad (4.6)$$

where $N = 2048$ (number of points in FFT computation)

$T = 10^{-4}$ sec. (sampling interval)

If the spectrum is divided into 6 bands for the peak picking algorithm, the band structure shown in Table 4 could be used.

The reconstruction of speech is achieved by the use of inverse FFT algorithm. An improvement in the reconstructed speech is possible by passing the signal through a digital filter whose input-output relationship is given by

$$y_n = (\hat{S}_{n-1} + 4\hat{S}_n + \hat{S}_{n+1}) / 6. \quad (4.7)$$

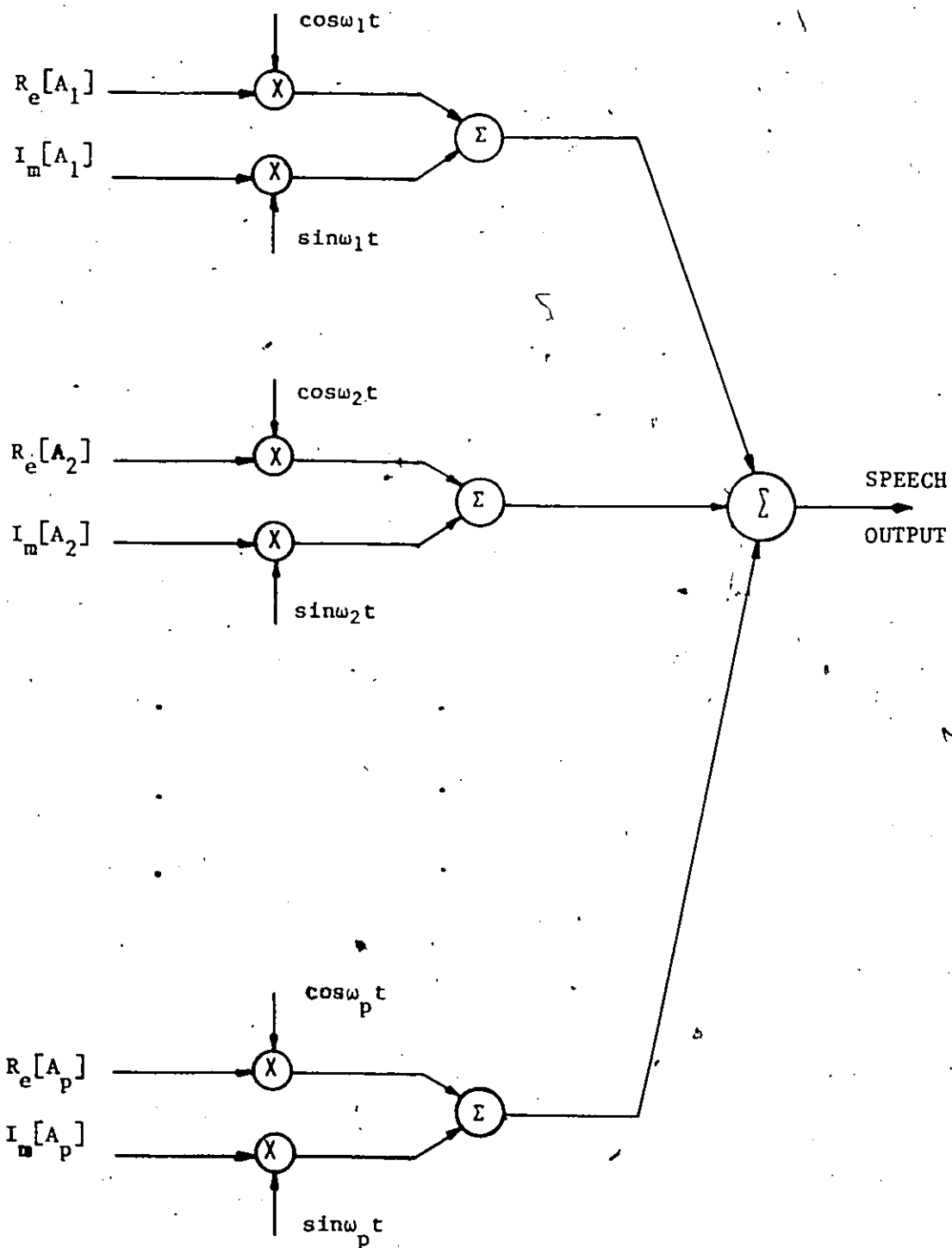


FIGURE 35 - FFT Model For Speech Synthesis

Computationally, the derivation of the parameters of this model is simple and very efficient. With software implementation, the use of Markels[39] FFT algorithm with pruning will decrease the computational time. The computational time can be made extremely small with use of FFT hardware pressures. Typically, the HP Fourier Analyzer[41] takes 10 msec. to compute an FFT of 4096 samples. With hardware FFT processors real time operation is feasible.

The advantages of this model in speech synthesis are,

1. Very high computational speeds
2. Since the model derivation is performed in the frequency domain, problems of stability of the model do not arise.
3. The poles and zeros of the speech production process are included in this scheme.
4. No error accumulation results, owing to the manner in which the reconstruction is performed.

4.3. Techniques For Improving Speech Quality

The speech quality can be improved by a proper selection of the number of bands and the band frequencies. For speech synthesis the frequency range 0 to 3 kHz is adequate and only a marginal improvement is obtained by increasing the upper limit. Also, there is very little information in the 0-100 Hz band. This band could be eliminated and assigned to a different frequency. Typically, if the frequency spectrum is divided into 12 bands much better results can be obtained. For this case, the band frequencies could be, as shown in Table 5.

TABLE 5

BAND	FREQUENCY RANGE Hz
1	50 - 150
2	151 - 250
3	251 - 350
4	351 - 450
5	451 - 550
6	551 - 750
7	751 - 950
8	951 - 1150
9	1151 - 1550
10	1551 - 1950
11	1951 - 2350
12	2351 - 3150

Speech Signal Frequency Spectrum Divided
Into 12 Bands For The FFT Model

With the 12 bands, there are now more bands situated at around 1000 Hz because the ear is most sensitive at this frequency. Also, of the three formants, the first formant carries the most signal energy. For most speech sounds, the first formant will lie between about 200 and 800 Hz.

A further improvement in the speech quality is possible using the same 12 bands, but to scan the real and imaginary parts of the FFT separately, for the selection of the peaks and the frequencies.

4.4. FFT Model Coding

For high quality speech transmission the results of the FFT model can further be improved by using a 1 bit coding scheme. The proposed coding scheme is shown in Figure 36.

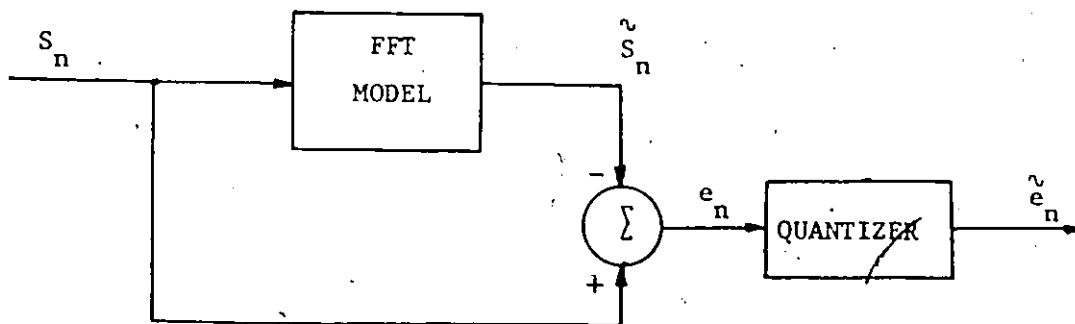


FIGURE 36 - 1-Bit Coding Scheme Using The FFT Model

In the coding scheme shown in Figure 36, the estimate of the speech signal \hat{S}_n is compared to the actual value S_n . The error signal $e_n = S_n - \hat{S}_n$ is quantized, using a variable height 2 level quantizer. The quantized error \hat{e}_n would then be transmitted at the signal sampling rate (10 kHz or may be as low as 6 kHz for speech signals) together with the FFT model parameters. The quantizer is operating in the open loop mode and hence, optimization of the quantizer level is possible. The optimum quantizer level is given by

$$Q = \frac{1}{N} \sum_{n=1}^N |S_n - \hat{S}_n| \quad (4.8)$$

where S_n is the actual value of the speech sample

\hat{S}_n is the estimate of the speech sample

N is the length of the optimization interval (256 samples).

The FFT coding scheme could work in real time allowing for an initial delay of 25.6 msec. This delay is not serious at all, since some existing telephone lines exhibit a delay of as much as 100 msec. The implementation of the coding scheme would be fairly straightforward and allowing a sizeable bandwidth compression. The bit rate requirements for the FFT model are discussed in the following section.

4.5. Bit Rate Requirements

With the 6-band model the bit rate for speech synthesis and coding is of the order of 7000 bits/sec. and 17000 bits/sec. respectively. If higher quality speech is required the number of frequency bands can be increased until the desired quality is achieved. It is unlikely that more than 12 bands would be required for normal speech quality transmission.

4.5.1. FFT Model-Speech Synthesis

(a) 6 bands - Model adjusted every 25 msec.

Number of parameters per model = 3×6

Number of bits per parameter = 10

∴ Total bit rate = $3 \times 6 \times 10 \times 40$
= 7200 bits/sec.

NOTE: - There are 6 bands, and in each band we require the following information.

(i) $R_e[A_j]$

(ii) $I_m[A_j]$

(iii) Position of the peak in each band, ℓ_j

(b) 12-bands - Model adjusted every 25 msec.

Number of parameters per model = 3×12

Number of bits per parameter = 10

Total bit rate = $3 \times 12 \times 10 \times 40$
= 14,400 bits/sec.

4.5.11. FFT Model-Coding

(a) 6 bands - Model adjusted every 25 msec.

10 kHz sampling rate.

Number of parameters per model = 3×6

Number of bits per parameters = 10

Number of bits for the error signal = 10,000

Total bit rate = $10,000 + 3 \times 6 \times 10 \times 40$
= 17,200 bits/sec.

(b) 12 bands - Model adjusted every 25 msec.

10 kHz sampling rate.

Number of parameters per model = 3×12

Number of bits per model = 10

Number of bits for the error signal = 10,000

Total bit rate = $10,000 + 3 \times 12 \times 10 \times 40$
= 24,400 bits/sec.

The bit rate requirement compares favourably with other recently reported schemes[42]. However, in the proposed scheme, the real advantage lies in its speed and also a stable model is guaranteed under all conditions. The bandwidth compression with the proposed coding scheme is of the order of 3:1 at a 10 kHz sampling rate. For equivalent PCM, the bit rate would be of the order of 80,000 bits/sec. (10 kHz sampling rate, and 8 bits per sample). The bit rate requirement can be decreased if the sampling frequency were reduced to 6 kHz.

4.6. Simulation Results

The results of the FFT model for speech synthesis are shown in Figures 37 and 38. Figure 37 shows the reconstructed waveforms for the words "NOON", "DAY", "SLEEPY" and "HOW ARE YOU", using the 6 band structure. The results for the 12 band synthesis scheme are shown in Figure 38.

The results of the FFT model when used in conjunction with a 1 bit open-loop coding scheme, are shown in Figures 39 and 40, using the 6 and the 12 bands respectively. The reconstructed speech waveform is of a very good quality, as can be seen by comparing the results of Figures 37 to 40 with those of the original waveforms shown in Figure 19. Computationally the proposed method for speech synthesis and coding is very efficient and of a high quality.

4.7. Summary

A new fast and accurate technique for speech synthesis and coding has been derived. With the present hardware FFT processors, real time operation is feasible, with a bandwidth compression of about 3:1. The bit rate requirement for the proposed coding scheme is of the same order

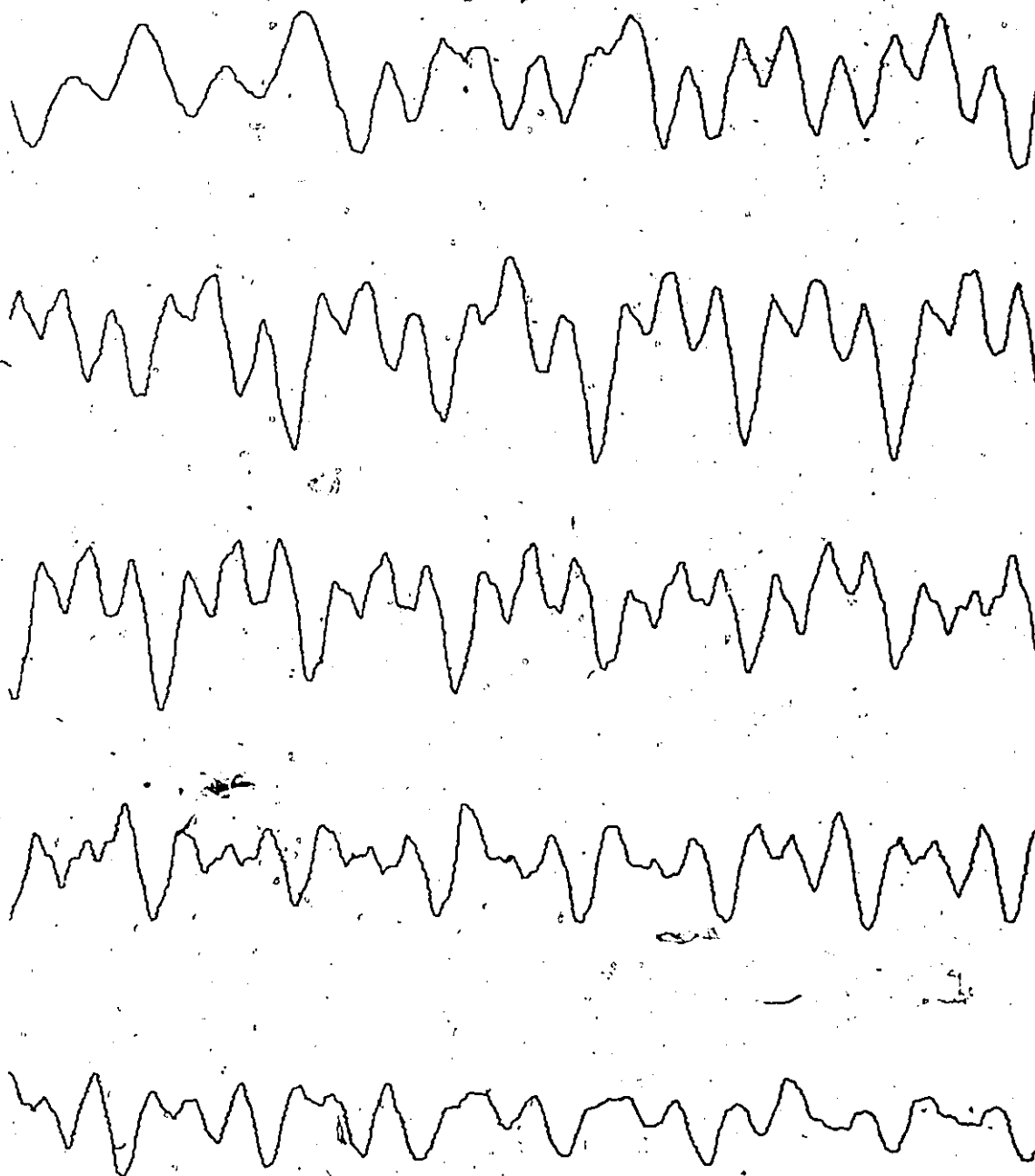


FIGURE 37 (a) - FFT Model Speech Synthesis Using 6 Bands - "NOON"

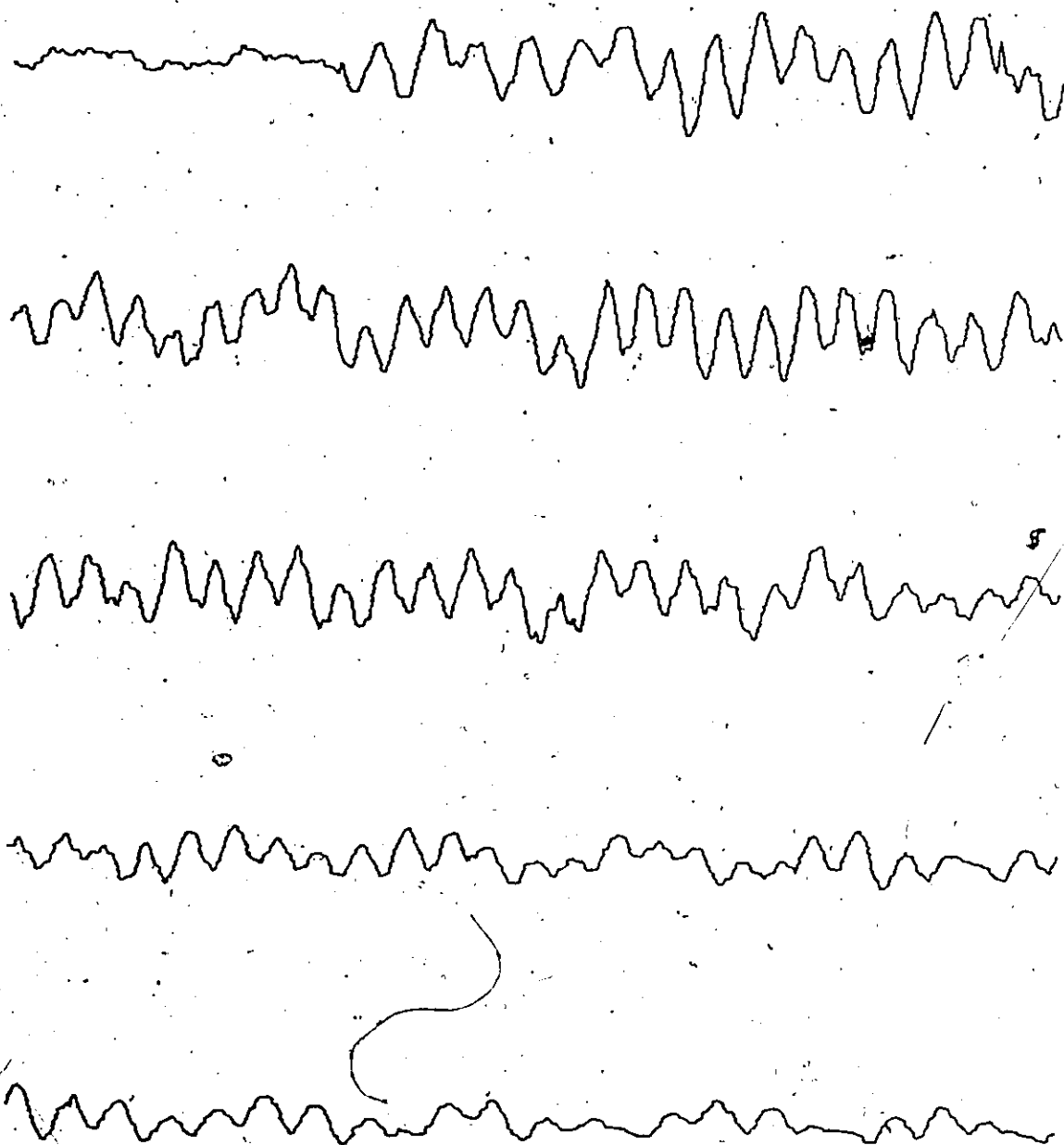


FIGURE 37 (b) - FFT Model Speech Synthesis Using 6 Bands - "DAY"

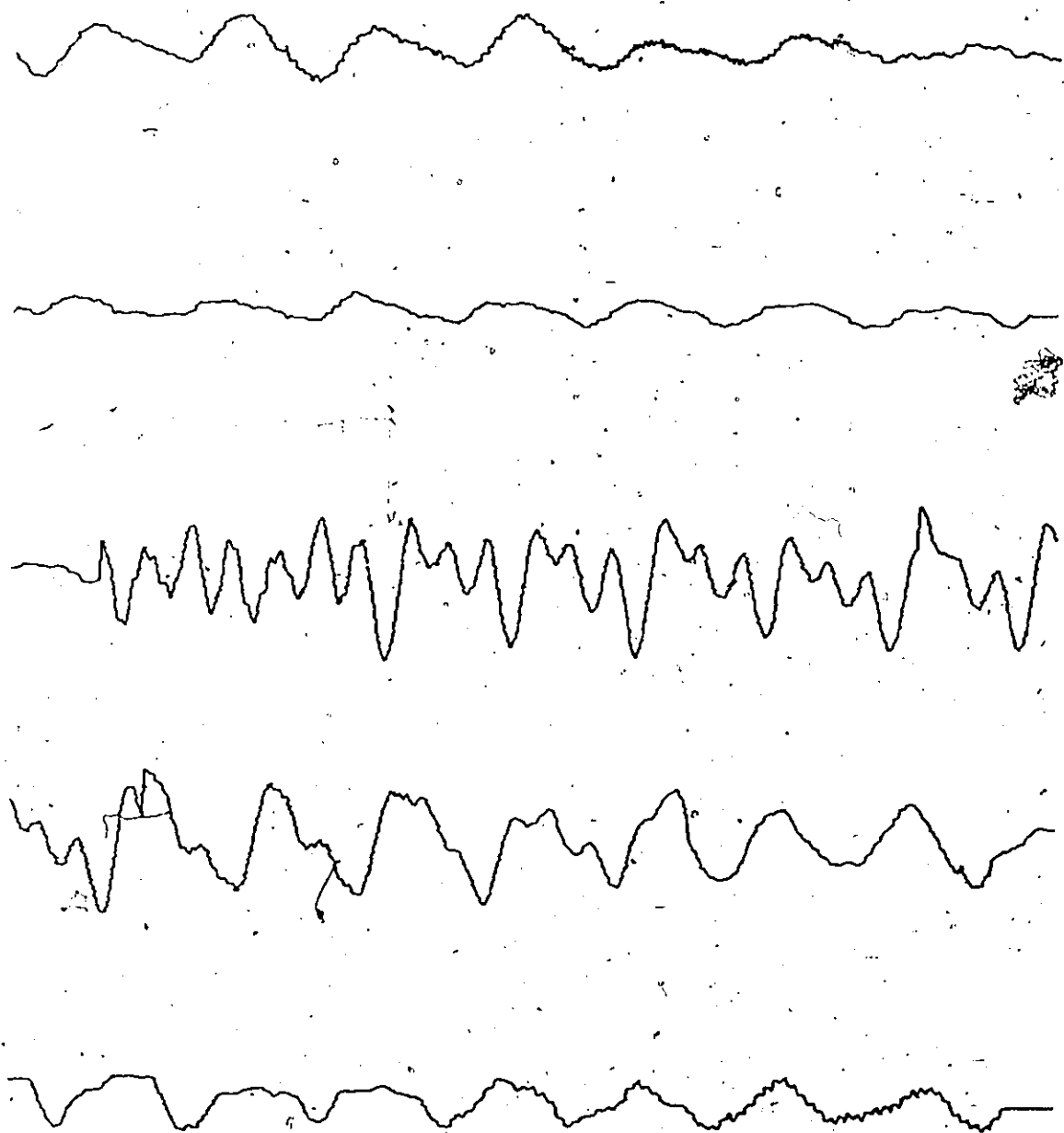


FIGURE 37 (c) - FFT Model Speech Synthesis Using 6 Bands - "SLEEPY"

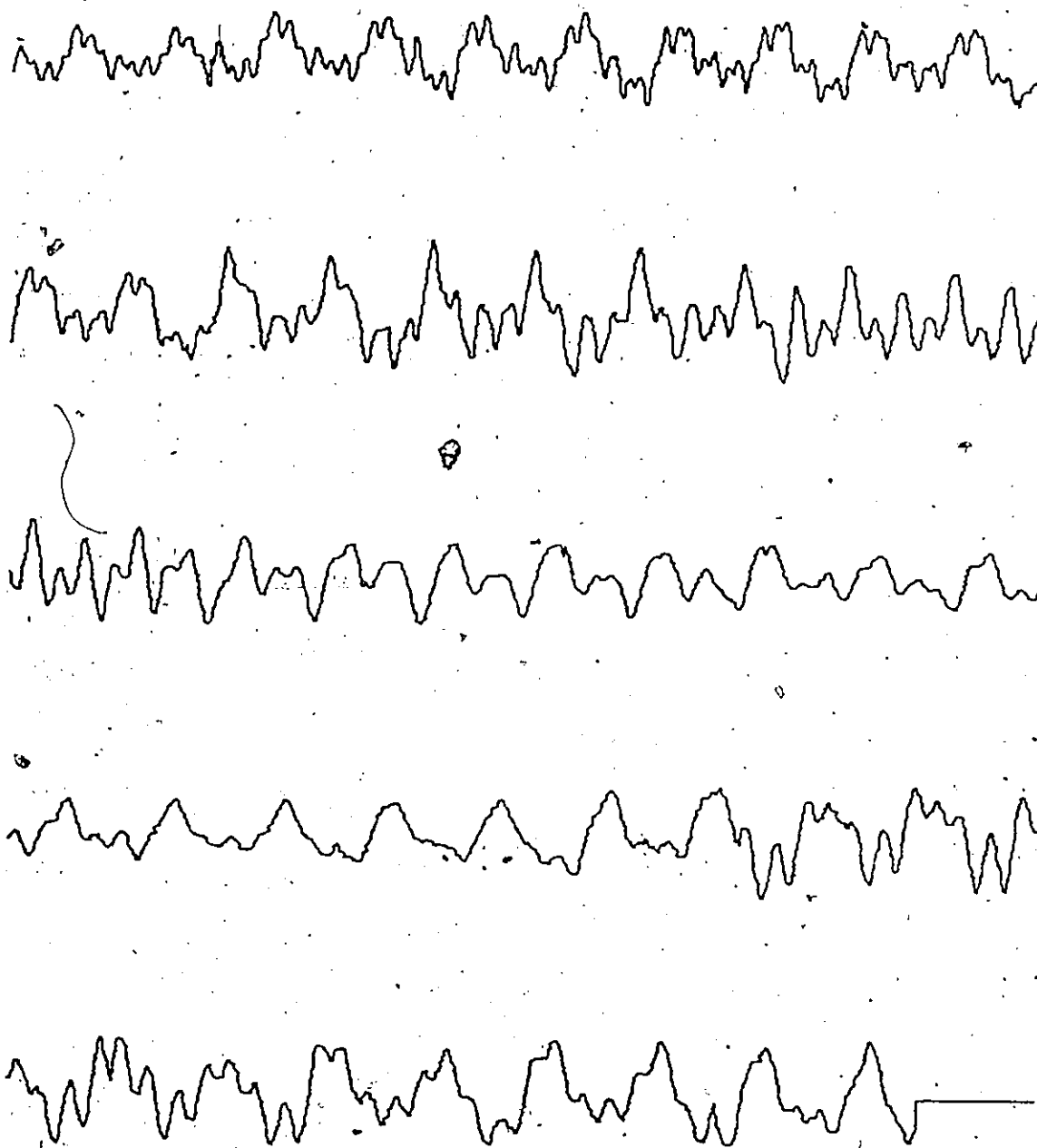


FIGURE 37 (d) - FFT Model Speech Synthesis Using 6 Bands - "HOW ARE YOU"

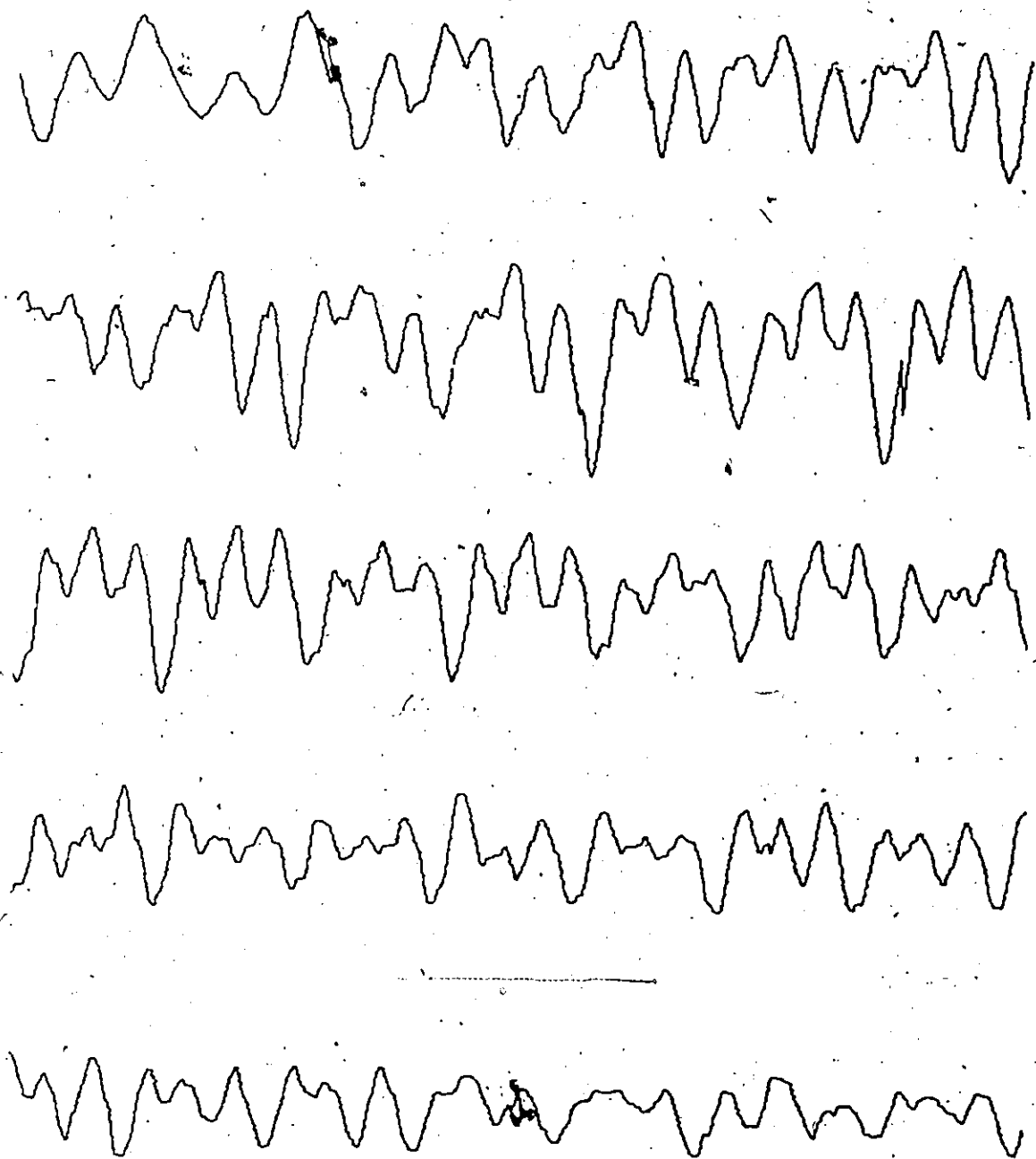


FIGURE 38 (a) - FFT Model Speech Synthesis Using 12 Bands - "NOON"

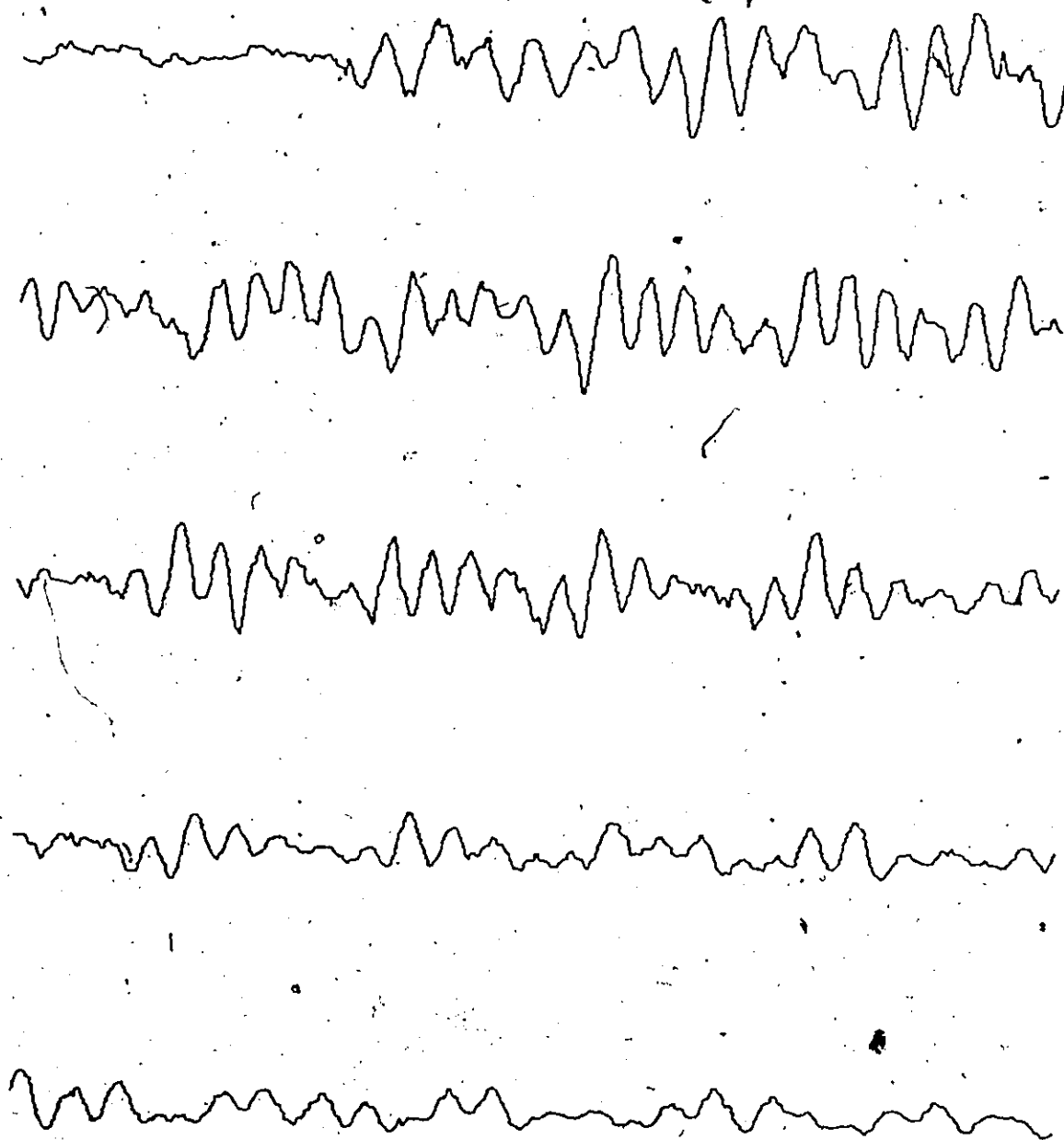


FIGURE 38 (b) - FFT Model Speech Synthesis Using 12 Bands 0 "DAY"

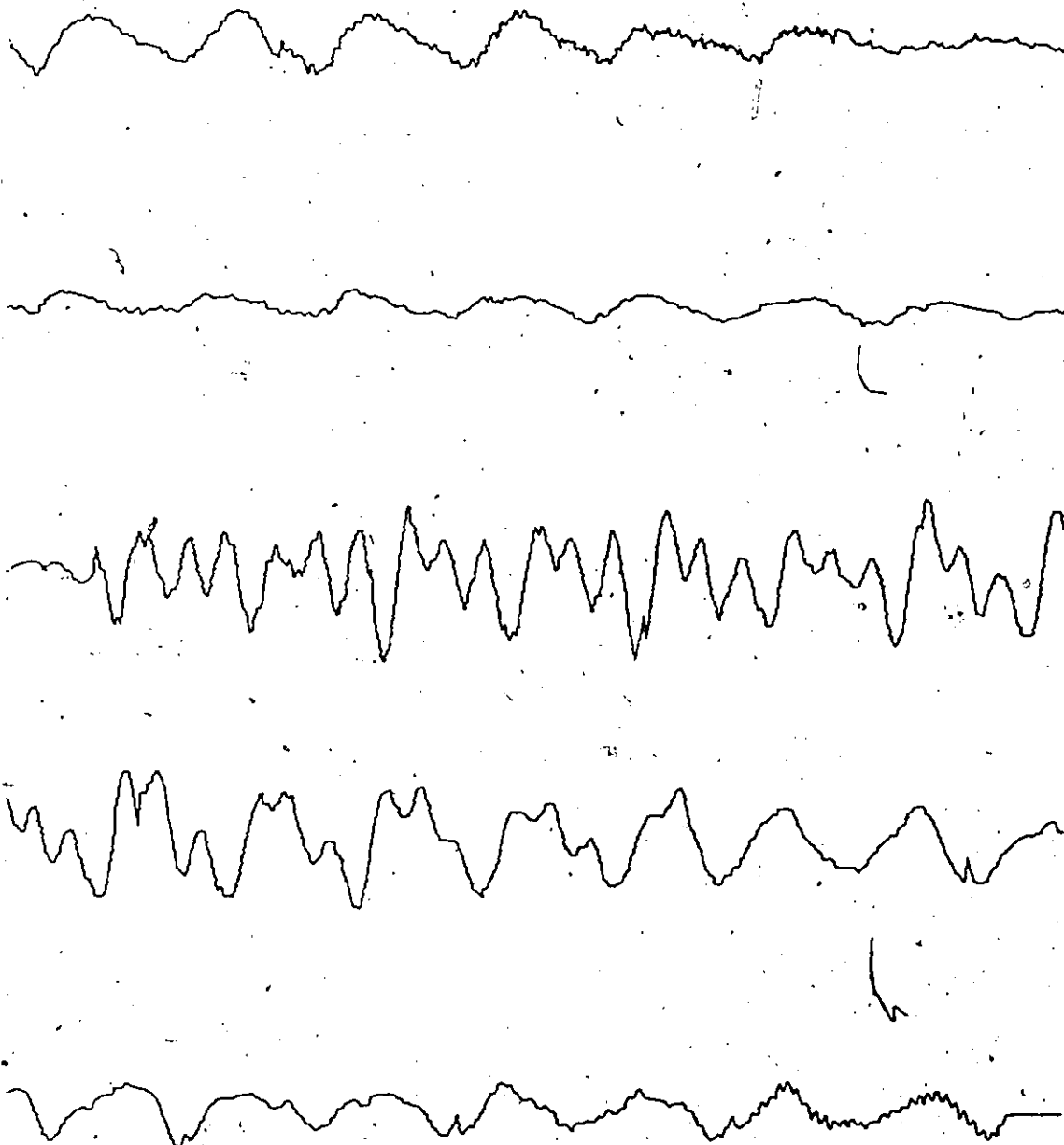


FIGURE 38 (c) - FFT Model Speech Synthesis Using 12 Bands - "SLEEPY"

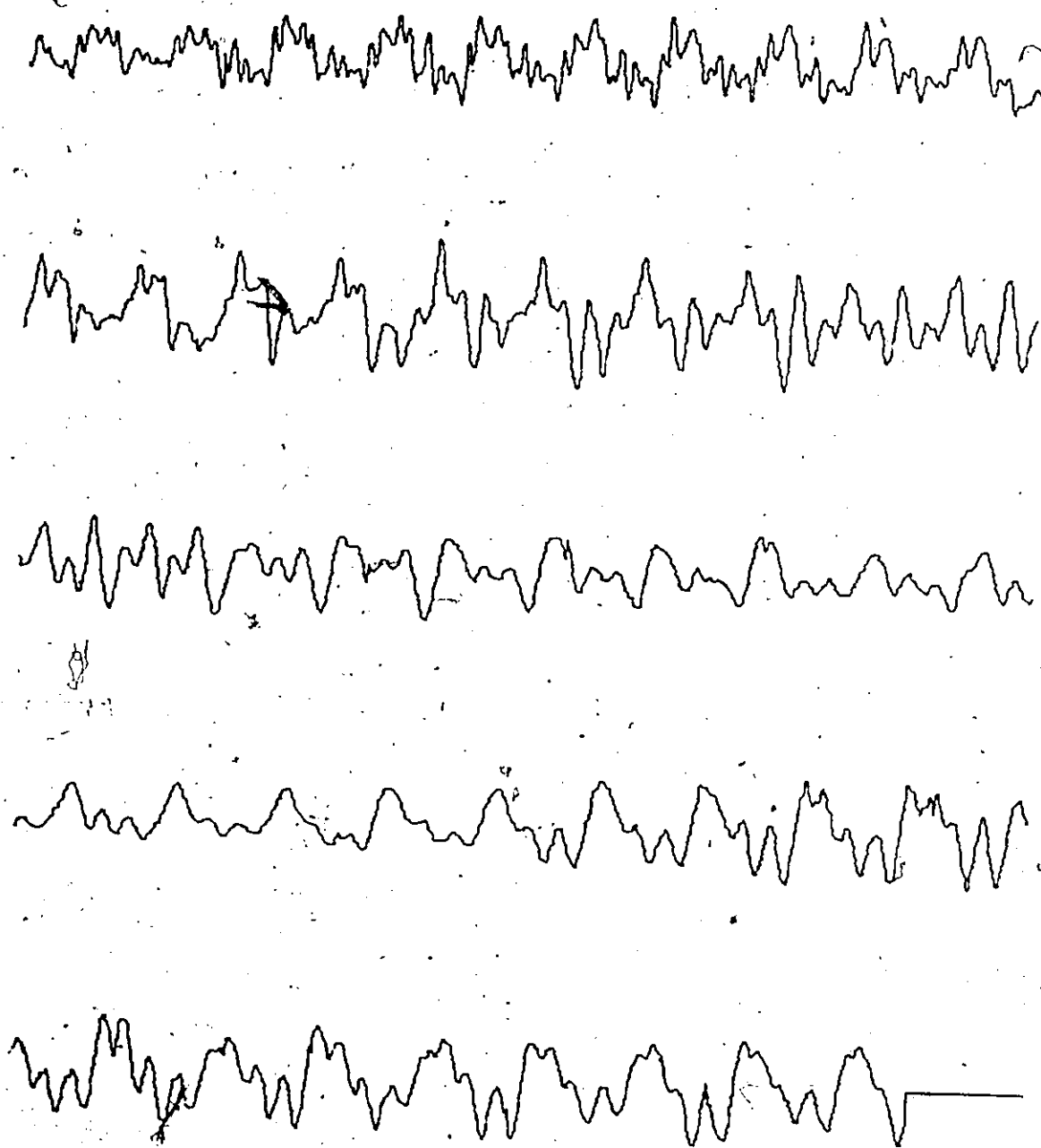


FIGURE 38 (d) - FFT Model Speech Synthesis Using 12 Bands - "HOW ARE YOU"

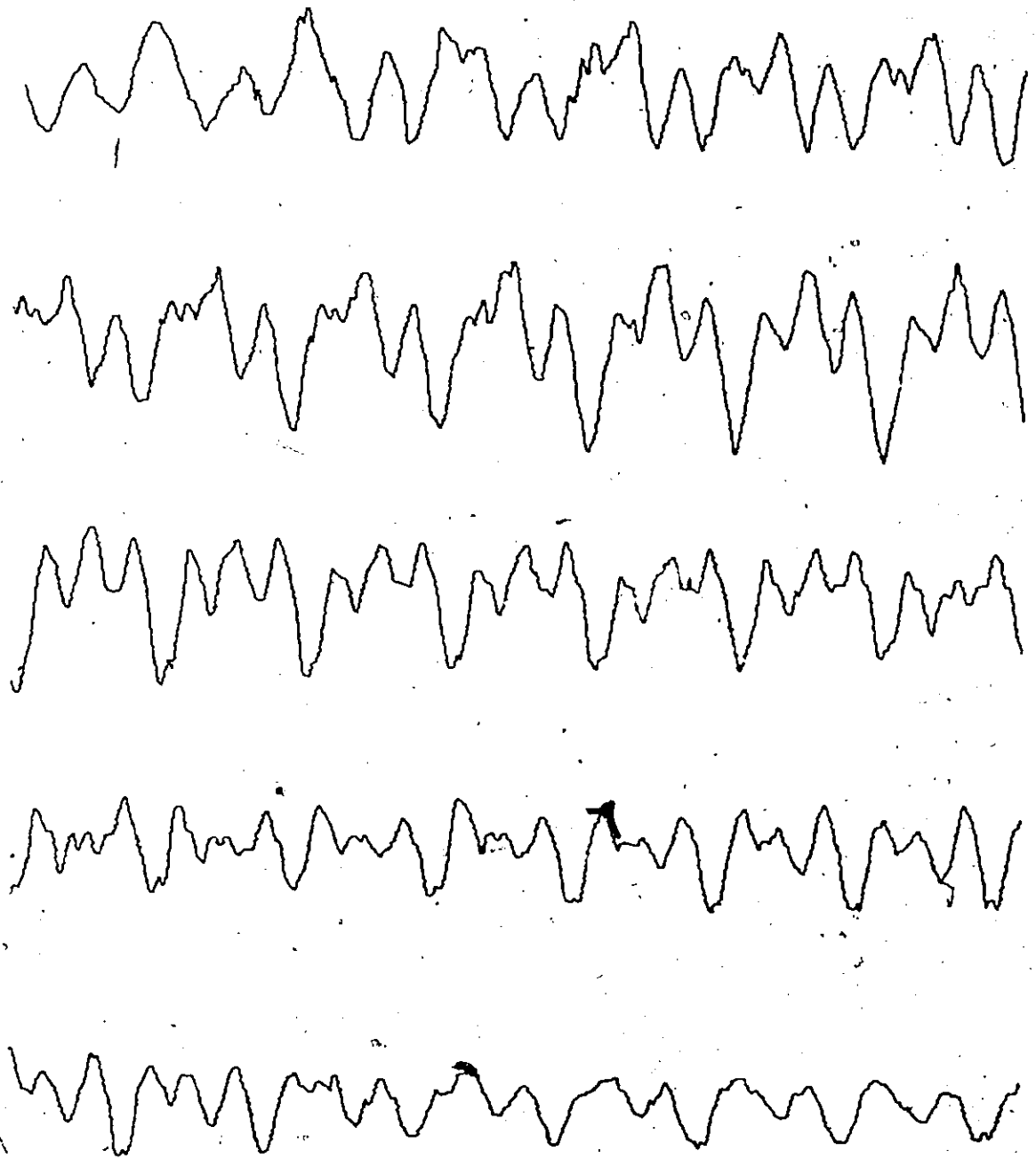


FIGURE 39(a) - FFT Model Speech Coding Using 6 Bands - "NOON"

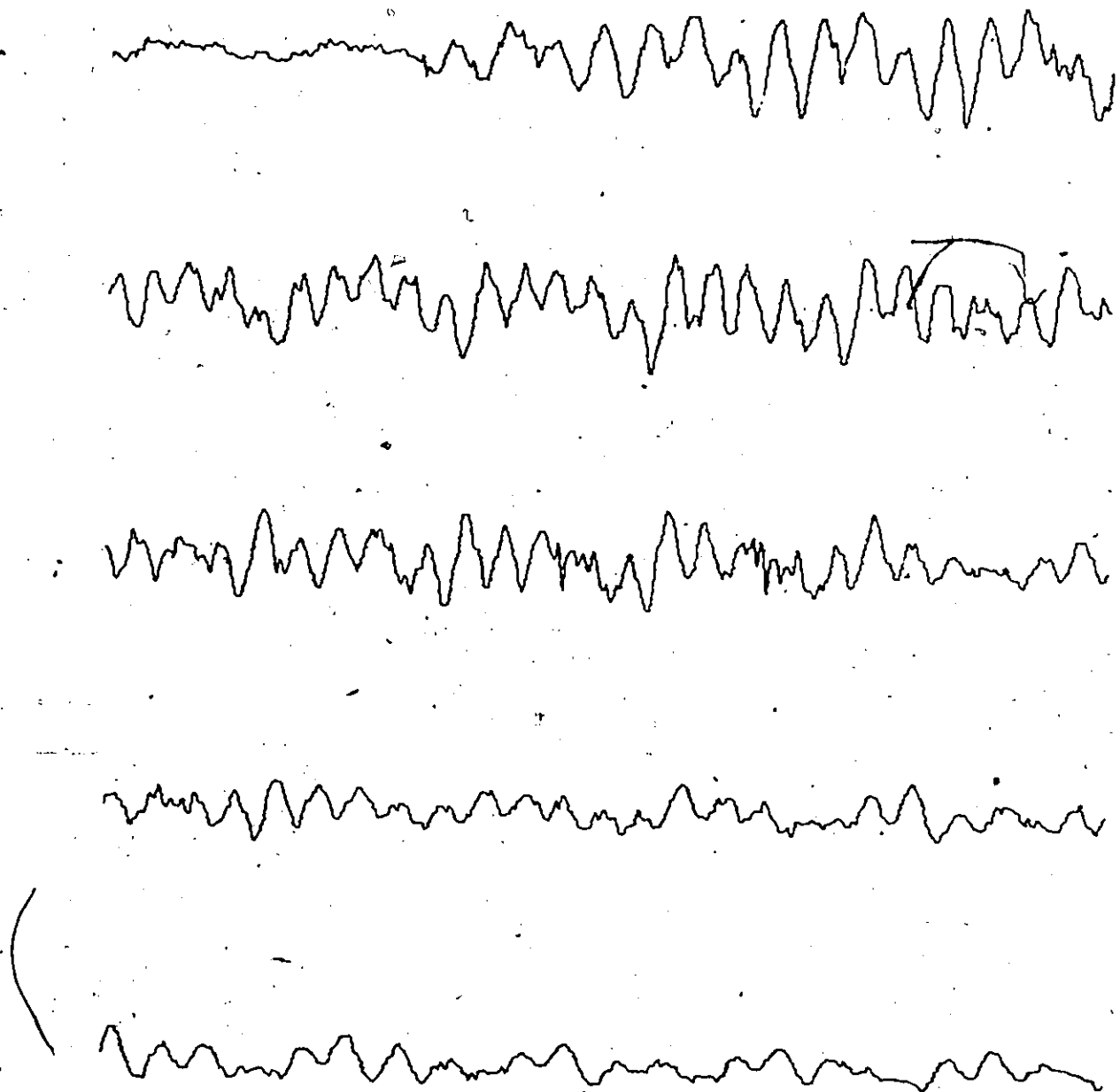


FIGURE 39 (b) - FFT Model Speech Coding Using 6 bands - "DAY"

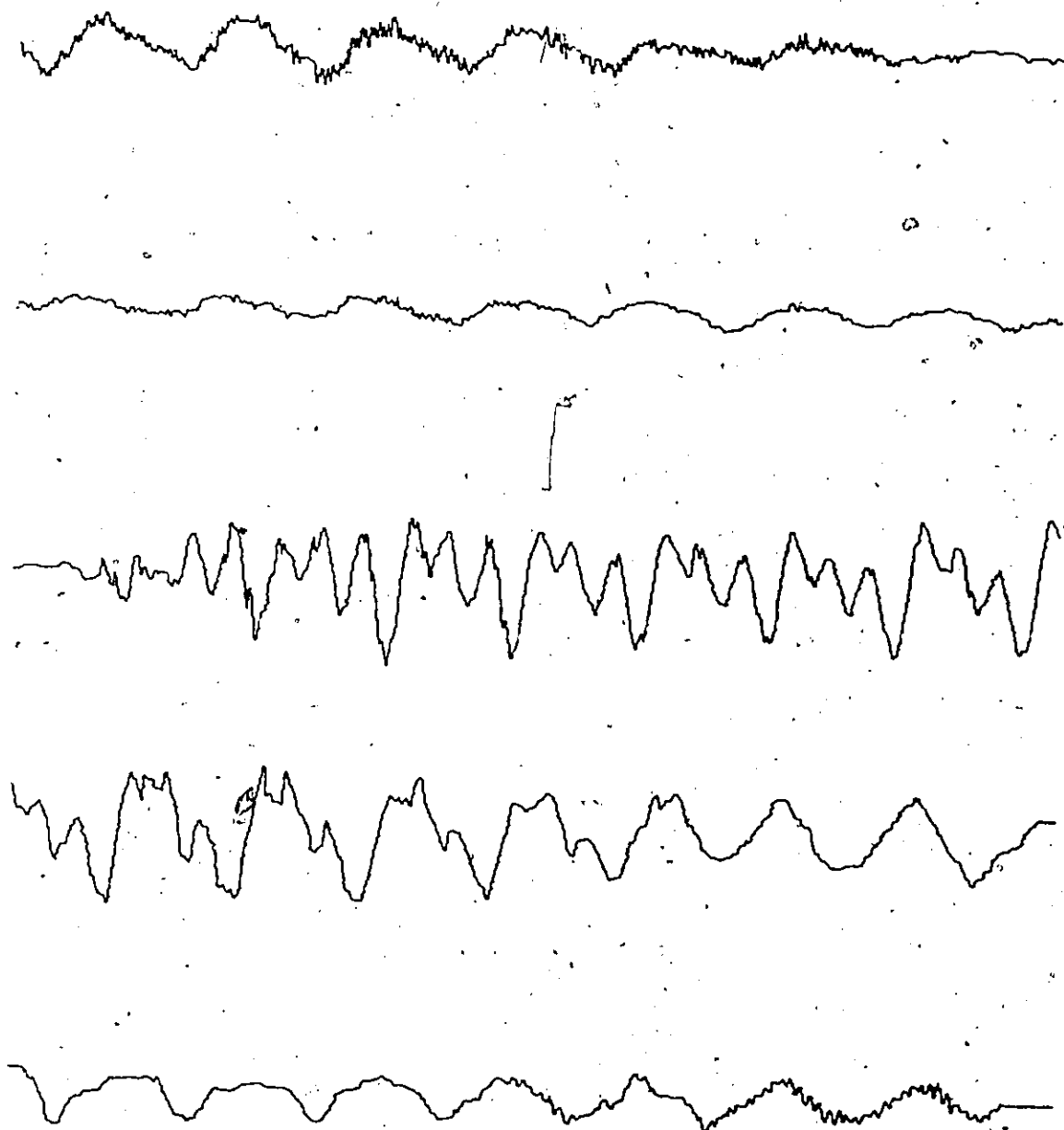


FIGURE 39 (c) - FFT Model Speech Coding Using 6 Bands - "SLEEPY"

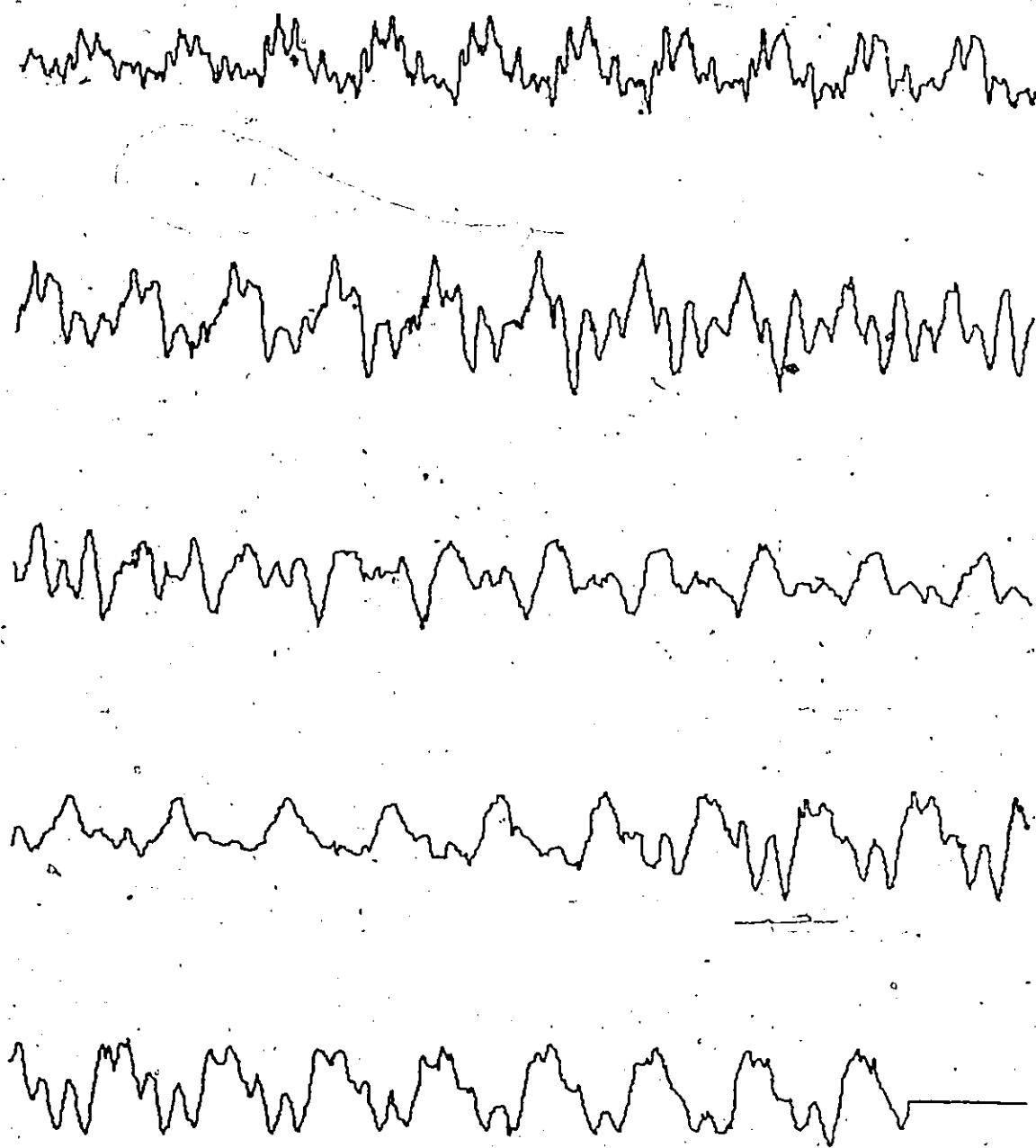


FIGURE 39 (d) - FFT Model Speech Coding Using 6 Bands - "HOW ARE YOU"

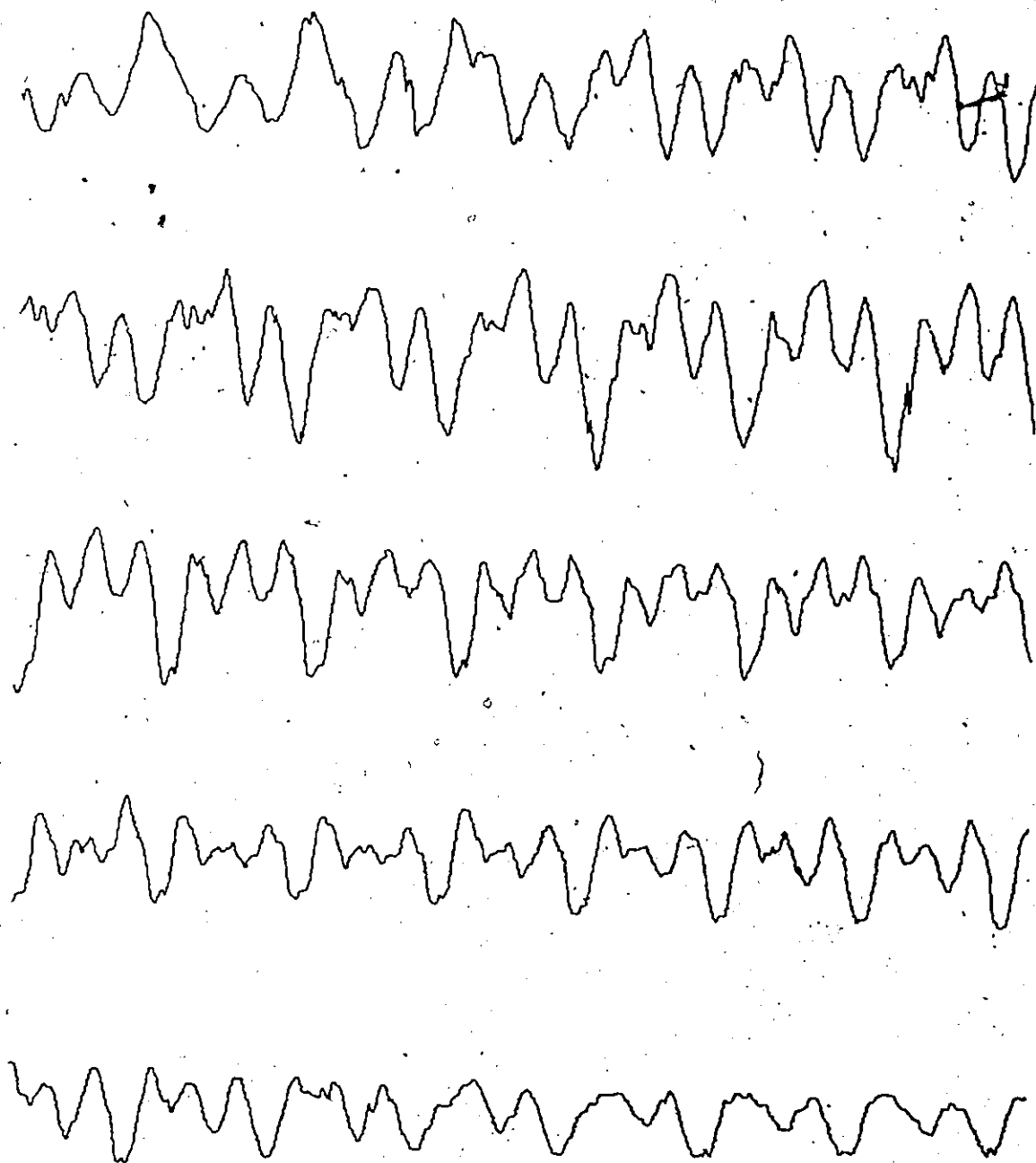


FIGURE 40 (a) - FFT Model Speech Coding Using 12 Bands - "NOON"

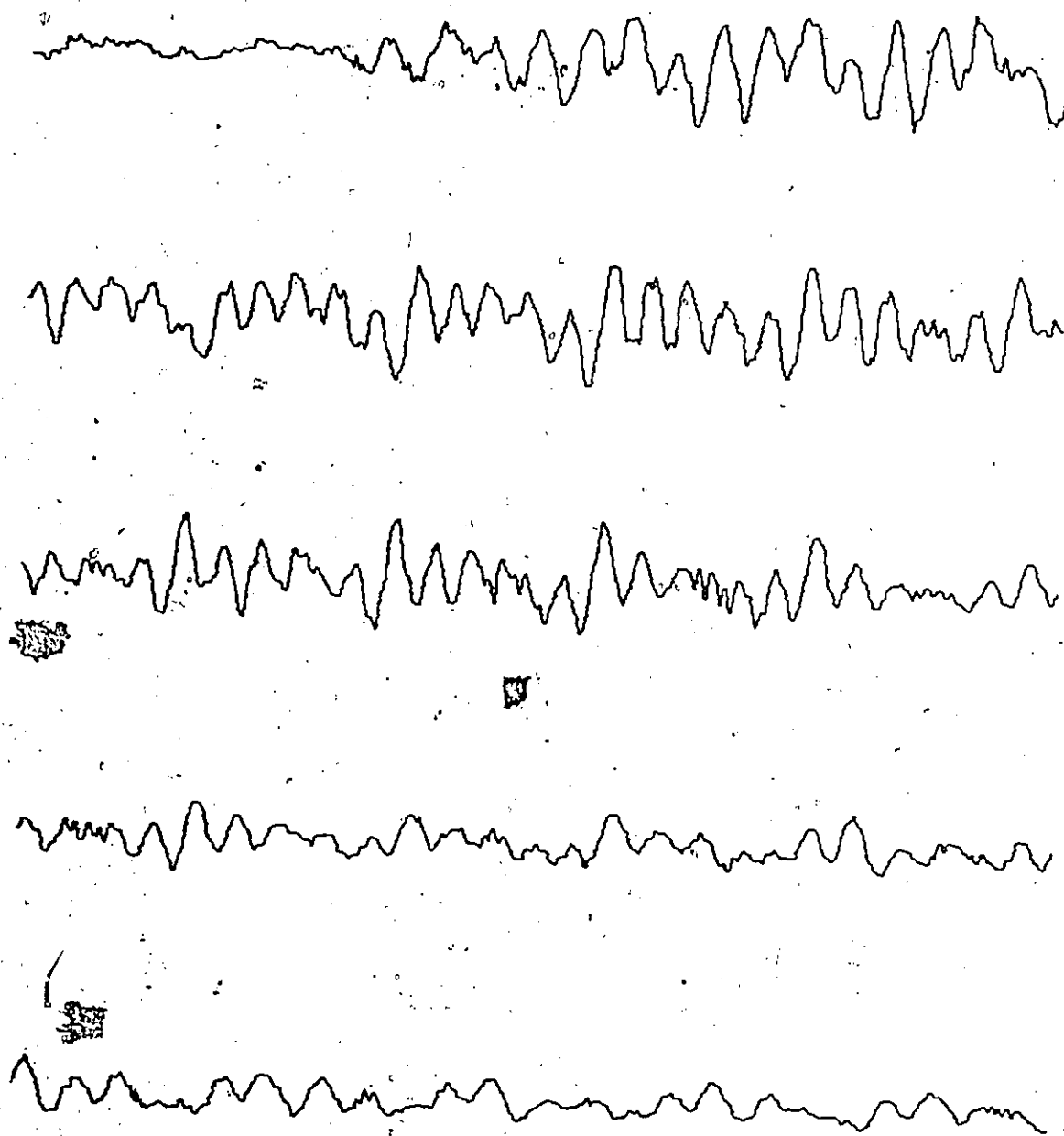


FIGURE 40 (b) - FFT Model Speech Coding Using 12 Bands - "DAY"

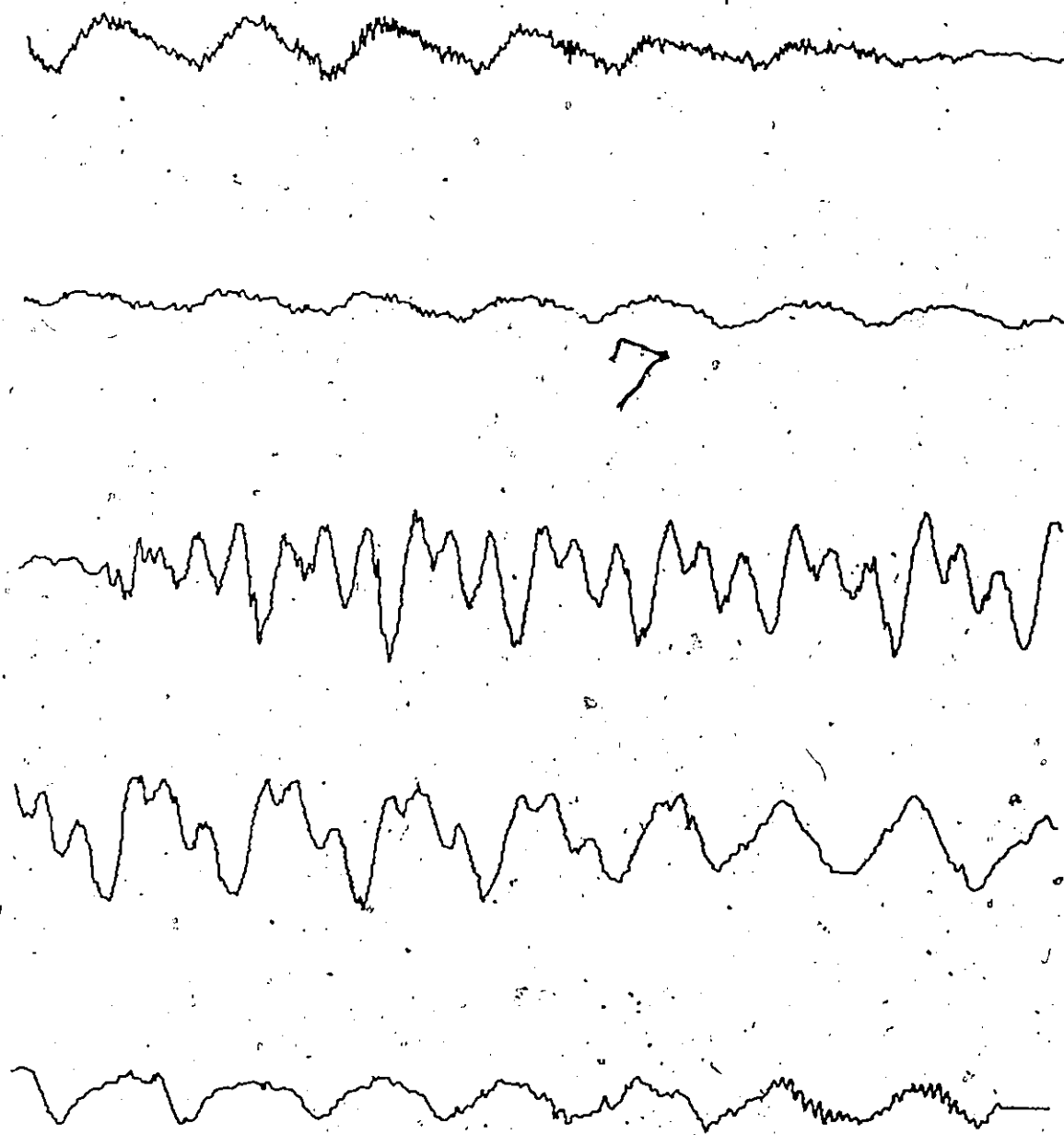


FIGURE 40 (c) - FFT Model Speech Coding Using 12 Bands - "SLEEPY"

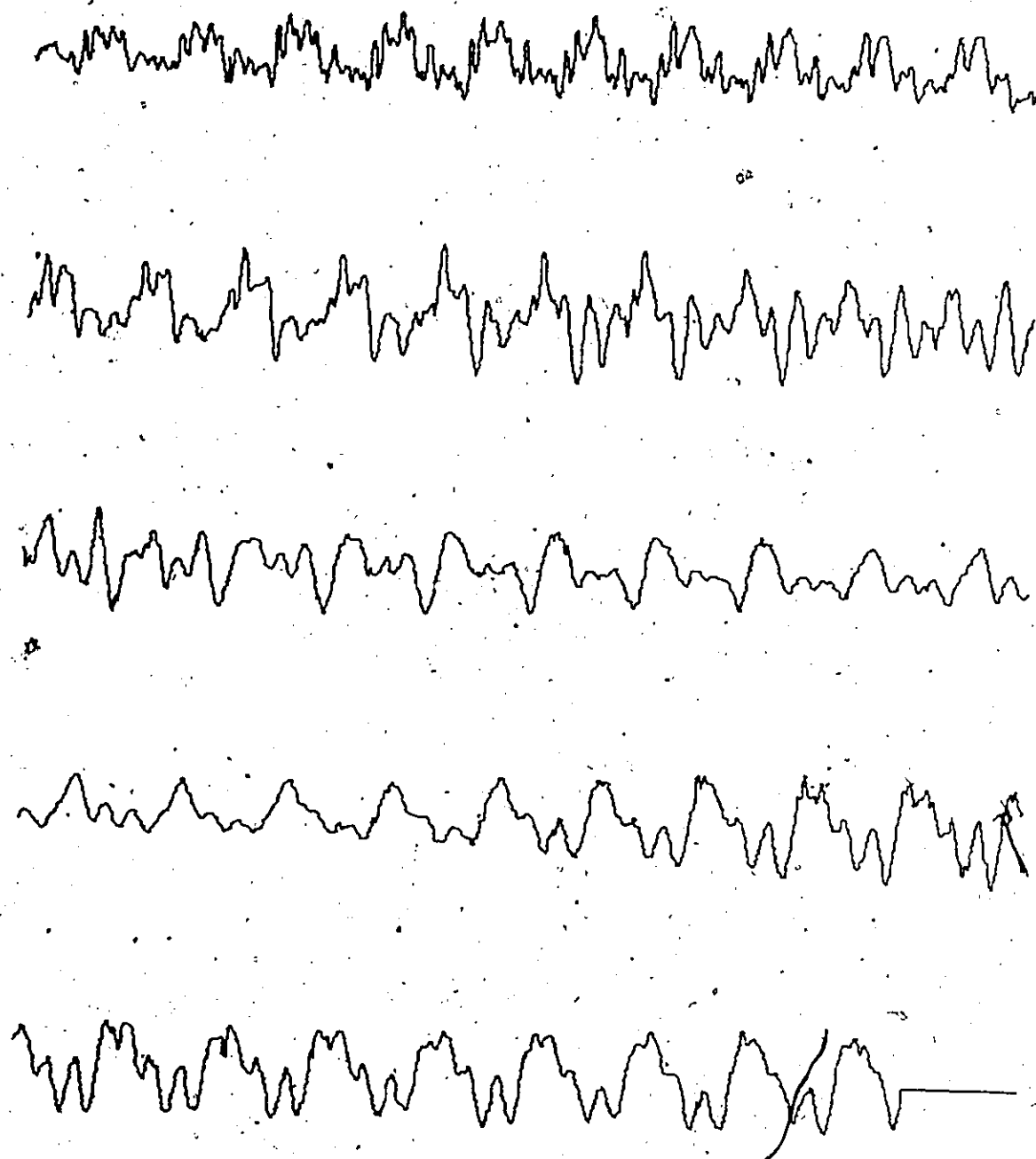


FIGURE 40 (d) - FFT Model Speech Coding Using 12 Bands - "HOW ARE YOU"

as that reported by Flanagan et al [42].

The use of spectral information is shown to result in a stable model. This is not always the case, with some time domain techniques. Also, by the manner in which speech is reconstructed, quantization effects have been eliminated. By a proper selection of the number and the frequencies of the bands it has been shown that high quality speech can be synthesized.

The bandwidth of the peaks in the speech signal spectrum could be fixed or alternatively, they could be found by locating the half power frequencies around the peaks. Both of these techniques were tried with hardly any improvement in the reconstructed speech signal.

CHAPTER V

MINI-COMPUTER IMPLEMENTATION OF SPEECH

SYNTHESIS AND CODING

The recording procedure and the implementation of the speech synthesis and coding schemes on the Digital Equipment Corporation PDP-8/I computer are described in this chapter. Because of its limited memory, all of the software was written in assembler language with overlaying of programmes. The minimum equipment required to operate the software is, an A/D and D/A converter, 4k word memory and one 32k word disk. The addition of an extra 4k word memory will allow a data scan routine to be incorporated for displaying the sampled data.

5.1. Recording of Spoken Material

The recording of the speech data was made in an anechoic chamber using an AKG model D224E dynamic microphone held in a desk top stand. The output of the microphone was recorded on one track of a Crown model S x 800 four track tape recorder, operating at a speed of $7\frac{1}{2}$ inches per second (maximum speed of the particular machine). The tape recorder was placed outside the anechoic chamber to minimize background noise. The recordings were made on Scotch 202 magnetic tape, which was the recommended tape for the particular machine.

The speaker used for the recording, was a male in his late forties having a distinct English accent. The recording material consisted of a number of phonemes and five sentences. The material to be recorded was read twice in order to obtain the most natural recording. No indication was given to the speaker as to which of the spoken material was to be used.

5.2. A/D Conversion

The A/D converter used, was the Digital Equipment Corporation (DEC), Laboratory Peripheral AX08, which converts a continuous signal into a sequence of 9-bit binary numbers. The speech was passed through a low pass filter (with the -3db point at 4k Hz.), and sampled at a rate of 10,000 samples per second. The sampling frequency was above the Nyquist rate, virtually eliminating any foldover, since the attenuation of the low pass filter was 40db outside the pass band. The laboratory computing facility is shown in Figure 41.

5.3. Software Package For Speech Synthesis

At the beginning of this work, the PDP-8/I computer had a 4k word core memory and one 32k word disk. An additional 4k word core memory was added at a later date, after most of the software had been written. The disk was used extensively for storing the original and the reconstructed speech samples. For the 10k Hz sampling rate, the maximum duration of connected speech that could be stored on the disk is 1.6 seconds. The 4k word core memory was used for processing the data and for any transfer between the core and the disk. Due to the limited memory of the PDP-8/I, the software package was made modular to permit overlaying of programmes. The complete software package consists of three programmes,

PART 1

OVERLAY 1

OVERLAY 2

To change from speech synthesis to speech coding, requires only substituting one programme, OVERLAY 1. The other programmes PART 1 and OVERLAY 2

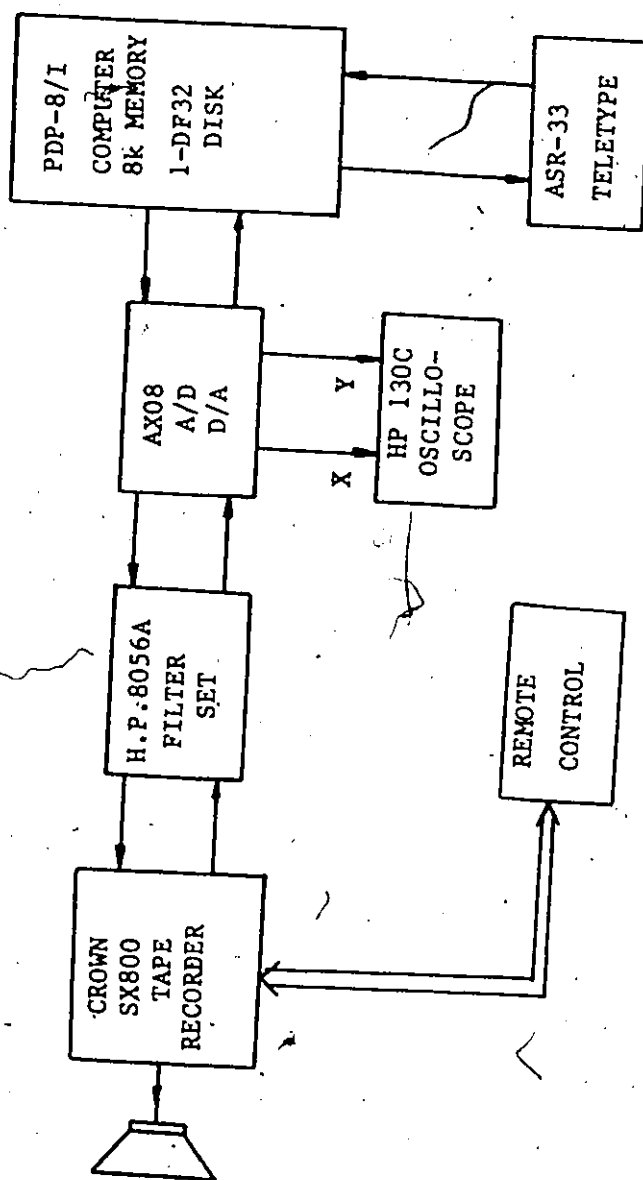


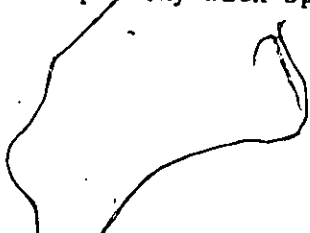
FIGURE 41 - Laboratory Computing Facility For Speech Processing

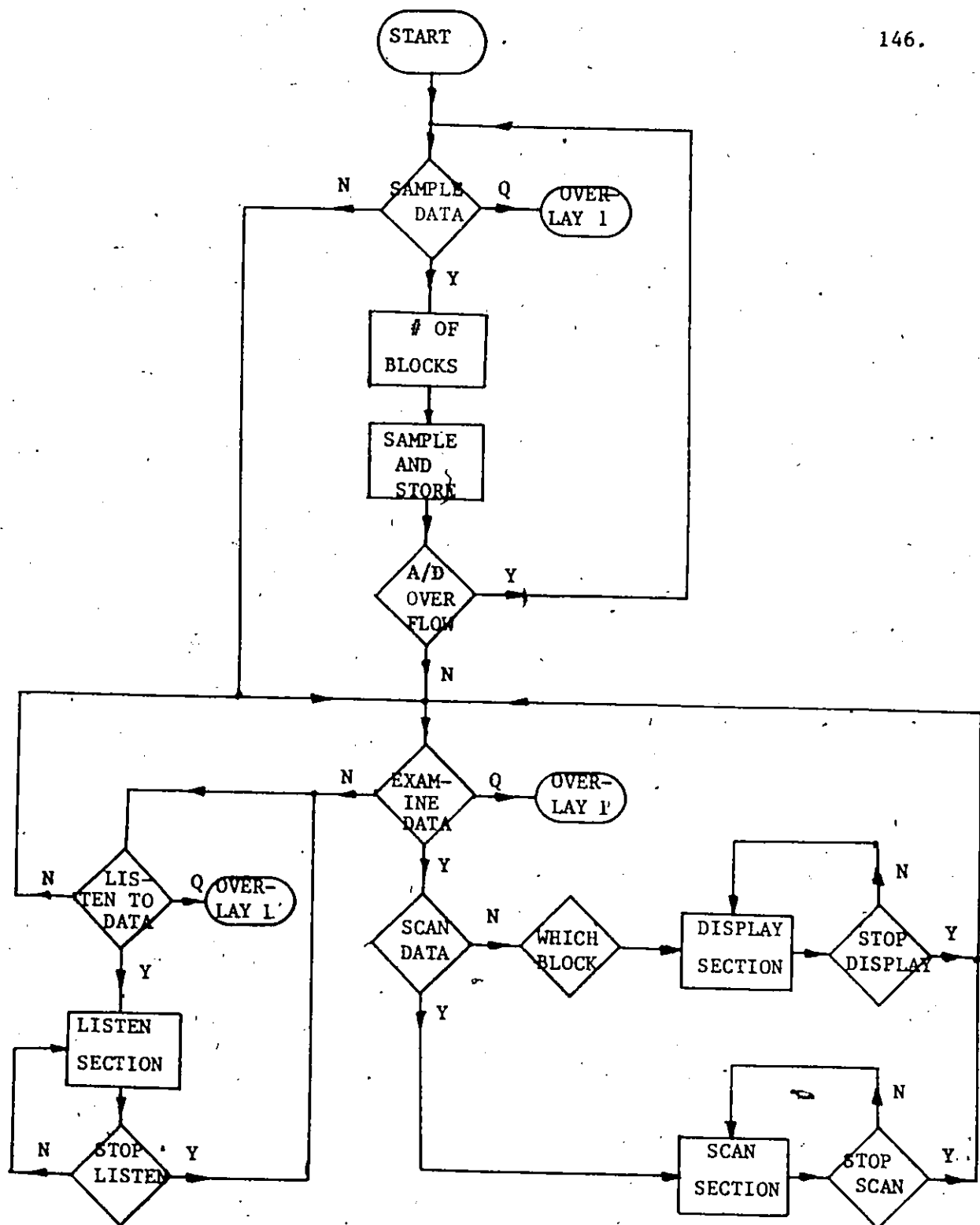
are common to all schemes.

5.3.1. PART 1

The first programme, called PART 1, performs the sampling, display of sampled data on the oscilloscope and provides an audio output of the sampled speech signal. After the sampling process is complete and the speech sample has been stored on the disk, the programme will examine the stored data for A/D converter overflow. This can arise if the sampled signal exceeds 1.02 volts in amplitude. If an overflow has occurred, the sampled signal will be folded back at the position of the overflow with a sign change, resulting in a distorted signal being stored. For maximum accuracy in the sampling process, the full range of the A/D converter should be used, without causing overflow. There is also a data scan option, which requires an additional 4k word memory. However, if the data scan option is not requested, all of the software will operate using the 4k word memory and the 32k word disk. The data scan is useful for checking the signal for any irregularities in the sampled waveform. A flow chart describing the operation of the PART 1 software is shown in Figure 42.

The correct response to the questions, sample data, examine data, scan data and listen to data, are Y-YES, N-NO, and Q-QUIT. Once the teletype has accepted a Q as the response, the second programme OVERLAY 1 will be read in under programme control. The number of blocks must be a 2 digit number which refers to the number of 1k octal blocks to be sampled. This can be any number from 01 to 40, the upper limit corresponds to filling up one half of the disk completely with speech samples.





EXAMINE: - Bit 11 = 1 Stops Display

LISTEN: - Bit 10 = 1 Repeated Listen

0 = 1 Stops Listen

SCAN: - Bits 6 to 11 Control Scanning Speed

0 = 1 Stops Scan

FIGURE 42 -Flow Chart Describing Operation of Programme - Part 1

Programme control is initiated either from the teletype or from the computer panel switch register. The switch register bit settings required to control the programme are shown in Figure 42.

5.3.ii. OVERLAY 1

When the teletype has accepted a Q as a valid response in PART 1, the second programme will be read in the OVERLAY 1. After this programme has been read in, the floating point package should be placed in the high speed reader and the continue switch depressed. OVERLAY 1 performs the evaluation of the predictor parameters and the reconstruction of the speech signal. The evaluation of the predictor parameters is carried out using Gaussian elimination with maximum pivot strategy.

The two responses required in this programme are,

OF EQUATIONS =

ANALYSIS INTERVAL =

For the first response, the number of equations must be a 2 digit octal number which can be from 01 to 16. The analysis interval must be a 3 digit octal number less than 156. These limits were imposed by the initial 4k word memory. The flow chart of OVERLAY 1 is shown in Figure 43 for speech synthesis by linear prediction using the initial conditions model.

5.3.iii. OVERLAY 2

OVERLAY 2 performs the display on the oscilloscope and also provides an audio output of the original and the reconstructed speech waveforms. In the display section, it is possible to select any 1k block of data, either from the original or the reconstructed speech signal. Here also,

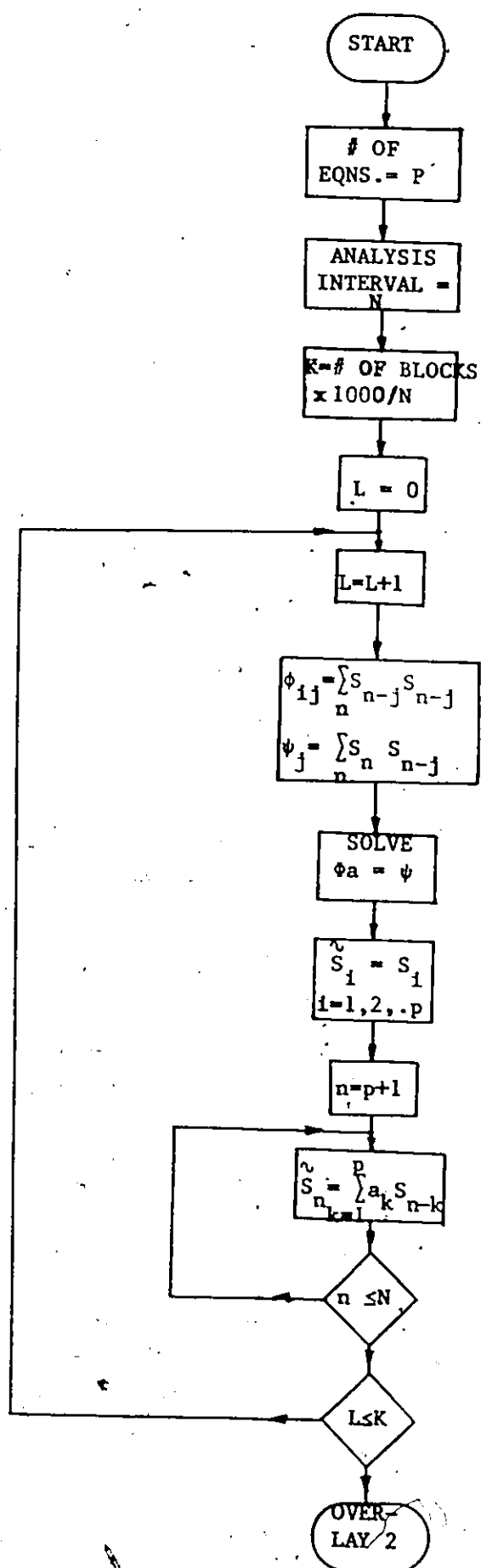


FIGURE 43 - Flow Chart Describing
Operation Of Programme Overlay 2
Using The Initial Conditions Model.

the data scan can be selected providing an 8k word memory is available.

Another feature which was added was that of scaling the reconstructed data to prevent possible overflow of the 9 bit D/A converter. The scaling does not alter the actual reconstructed speech sample that is stored on the disk. The scaling is carried out prior to loading the D/A converter buffer. The flow chart of OVERLAY 2 is shown in Figure 44.

When the teletype accepts Q as a valid response, the programme will evaluate the signal to noise ratio of the reconstruction. This quantity is given by,

$$S/N = \frac{\sum_n S_n^2}{\sum_n (S_n - \hat{S}_n)^2} \quad (5.1)$$

where S_n is the original speech signal

\hat{S}_n is the reconstructed speech signal

and n is summed over all the speech samples.

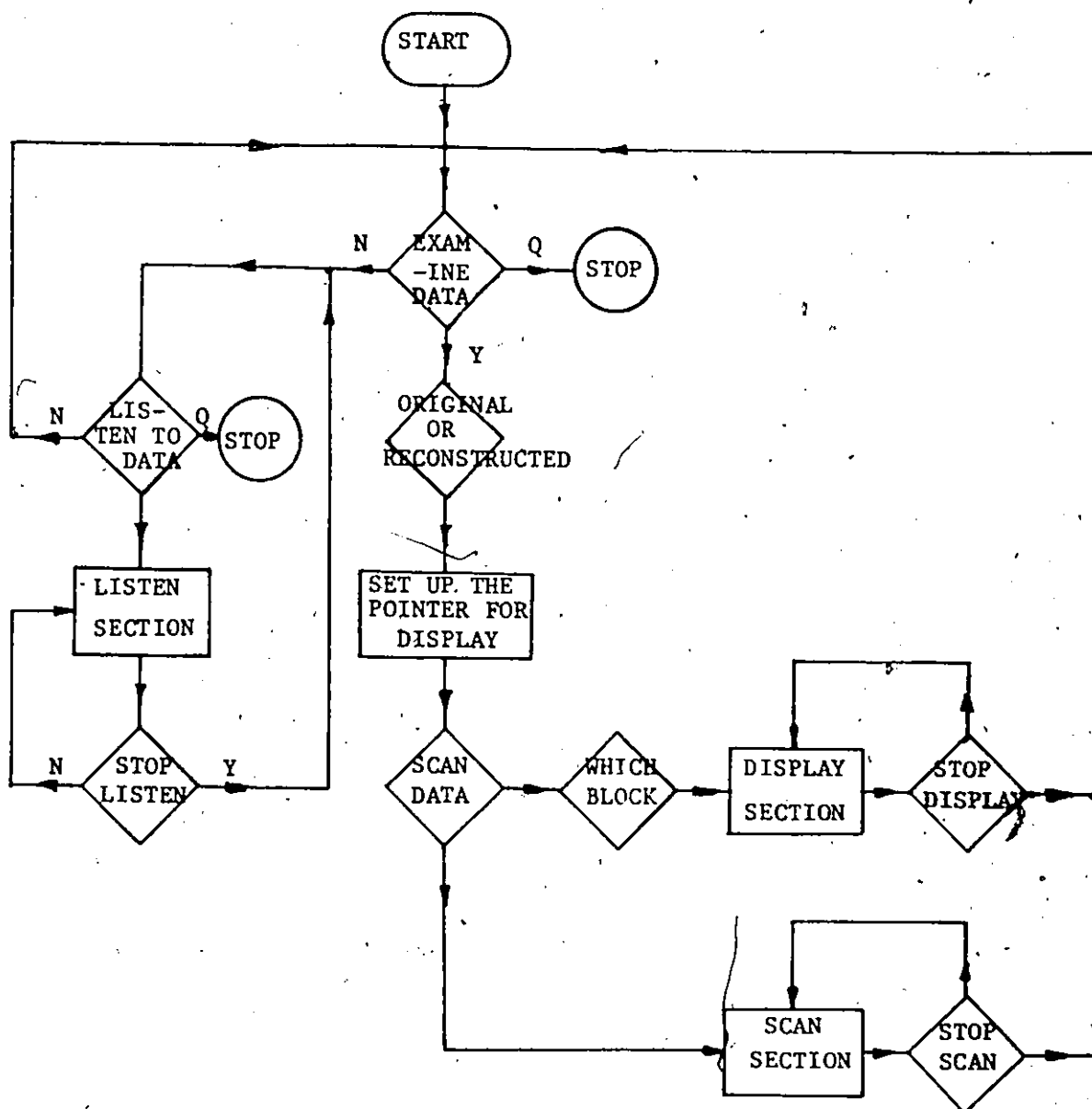
After the signal to noise ratio has been evaluated and typed out on the teletype, the programme can either terminate or reset certain locations so that PART 1 can be read in again and the whole operation repeated on the same speech sample but using different parameters.

5.4. Implementation and Results

This section describes the speech synthesis and coding schemes as implemented on the PDP-8/I computer.

5.4.1. Initial Condition Model

The linear predictor speech synthesis, using the initial condition model, namely



EXAMINE: - Bit 11 = 1 Stops Display

SCAN: - Bits 6 to 11 Control Scanning Speed

0 = 1 Stops Scan

LISTEN: - Bit 10 = 1 Listen to Original and Reconstructed

Bits 10 = 11 = 1 Listen to Reconstructed Only

Bit 0 = 1 Stops Listen

FIGURE 44 - Flow Chart Describing Operation of Programme Overlay 2

$$\hat{s}_1 = s_1$$

$$\hat{s}_2 = s_2$$

$$\hat{s}_p = s_p$$

(5.2)

is given by

$$\hat{s}_n = \sum_{k=1}^p a_k \hat{s}_{n-k} \quad p+1 \leq n \leq N \quad (5.3)$$

where s_n is the original speech sample

\hat{s}_n is the predicted speech sample

a_1 are the predictor parameters

and N is the number of samples over which the model is evaluated.

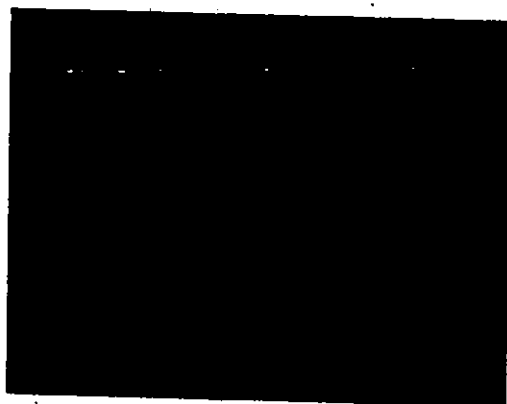
The analysis interval was kept constant throughout the simulation and at the beginning of each interval, the predictor was re-initialized.

The remaining samples are evaluated recursively using equation (5.3).

Figure 45 shows the original waveform for the sentence, "WE WERE AWAY A YEAR AGO". The results of the initial conditions model for the same sentence are shown in Figure 46.

5.4.ii. Coding of Speech Signals

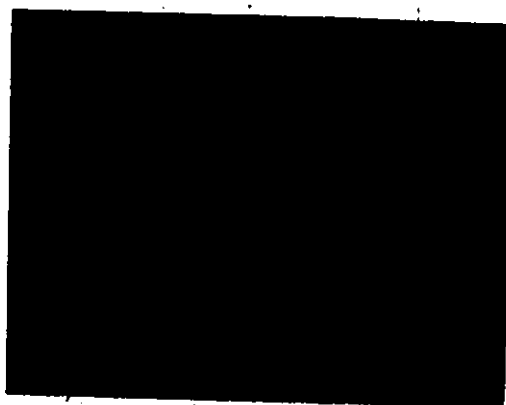
In this section the implementation of coding using the differential pulse code modulation scheme are shown for two cases. In the first case, the quantizer level is kept fixed throughout the simulation, whereas in the second case, the quantizer level is evaluated according to the following criterion,



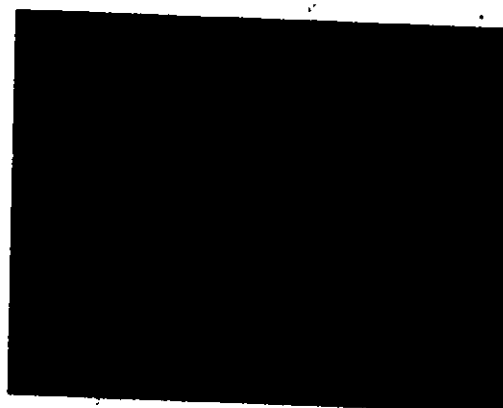
0 - 300 msec.



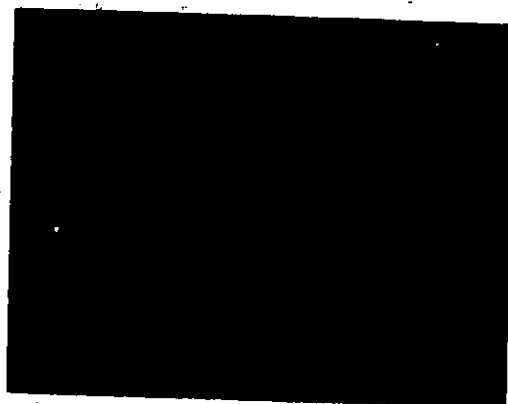
300 - 600 msec.



600 - 900 msec.



900 - 1200 msec.

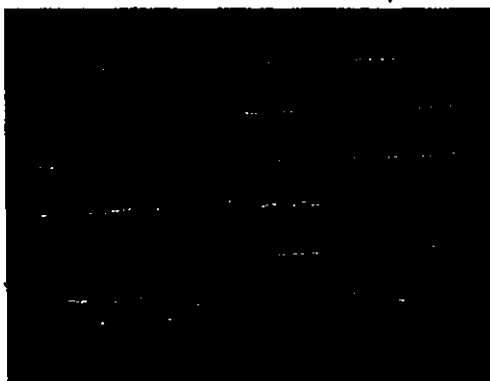


1200 - 1500 msec.

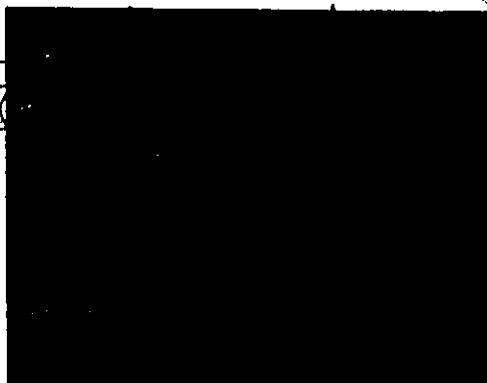
FIGURE 45

Original Waveform of the Sentence

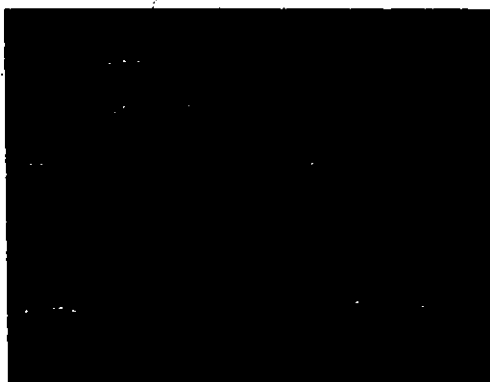
"WE WERE AWAY A YEAR AGO"



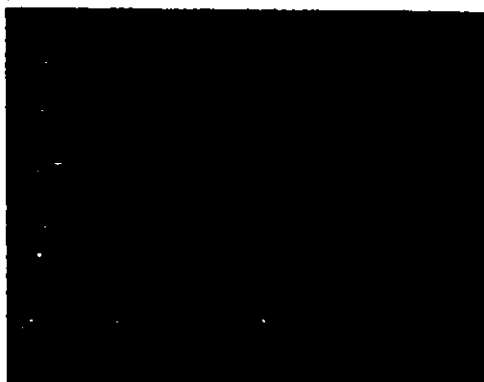
0 - 300 msec.



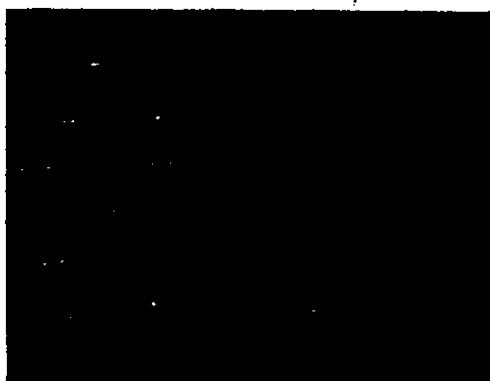
300 - 600 msec.



600 - 900 msec.



900 - 1200 msec.



1200 - 1500 msec.

FIGURE 46

Initial Condition Model Reconstruction
of the Sentence

"WE WERE AWAY A YEAR AGO"

$$Q = \frac{1}{N} \sum_{n=1}^N \left| S_n - \sum_{k=1}^p a_k S_{n-k} \right| \quad (5.4)$$

where S_n is the original speech sample

a_k are the predictor parameters $k = 1, 2, \dots, p$

and N is the number of samples over which Q is kept constant.

5.4.11(a) DPCM-Fixed Quantizer Height

In this simulation, the number of quantizer levels and the step size was controlled from the keyboard. Although, the aim was to use a 2 level quantizer, the software was modified to include multi levels for maximum flexibility.

Two additional responses are required to those described in section 5.3.11, namely

OF QUANTIZER LEVELS =

QUANTIZER HEIGHT =

The number of quantizer levels must be a single digit number from 1 to 7, and represents the number of positive levels. Thus if the response were 2, the number of actual levels would be 4, half of which are positive and the other half negative. The quantizer height must be a 2 digit octal number and be from 01 to 77.

The results of the two level, fixed step size quantizer scheme are shown in Figure 47 for the sentence, "WE WERE AWAY A YEAR AGO".

5.4.11(b) DPCM Variable Quantizer Height

In this scheme the differential pulse code modulation was used with a 2 level variable height quantizer. This programme only requires the number of predictor parameters and the analysis interval to be



0 - 300 msec.



300 - 600 msec.



600 - 900 msec.



900 - 1200 msec.



1200 - 1500 msec.

FIGURE 47

DPCM Fixed Quantizer Height Reconstruction
of the Sentence

"WE WERE AWAY A YEAR AGO"

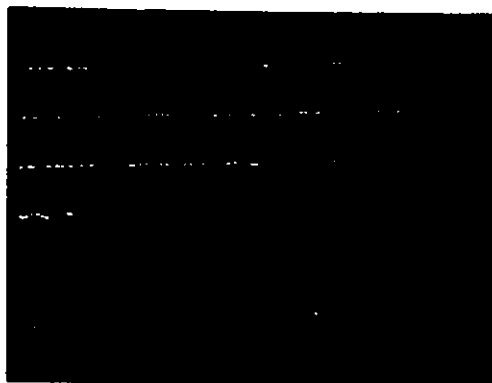
supplied via the keyboard of the teletype. The results of this simulation are shown in Figure 48, for the sentence, "WE WERE AWAY A YEAR AGO".

5.5. Phoneme Synthesis

It is conceivable that all the speech could be built up using basic sounds, called phonemes. These basic sounds would be connected together in a certain order to produce the required word or sentence. In the English language there are 45 phonemes that are recognized by the International Phonetic Alphabet (IPA). Thus, it would seem, that, for the purpose of speech synthesis, all that is required are the 45 phonemes to be able to produce connected speech.

The use of phonemes as building blocks for speech is somewhat more complicated than indicated above. Words and sentences can be made up from distinct phonemes, however, the resulting speech will often sound unnatural. The duration of the same phoneme will be different, depending upon the actual position in a word, and also the position where the stress occurs. Most phoneme synthesis schemes lack the transient nature of adjacent phonemes, which is important for high quality speech synthesis.

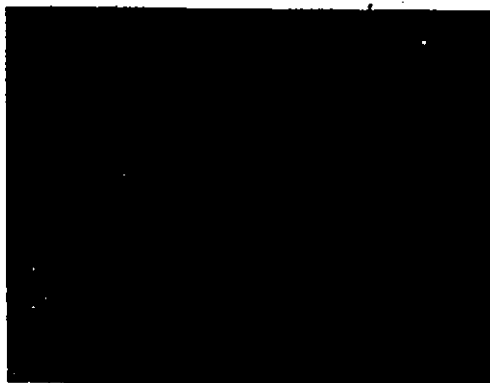
Even with the above mentioned limitations, it is still possible to obtain intelligent speech using phoneme synthesis. Phoneme synthesis does offer a tremendous saving in storage over straight PCM. For the word "NOON", which is approximately 0.35 seconds in duration, PCM would require 3500 samples to be stored (assuming a 10 kHz sampling rate). For PCM, the number of bits required to store the data = $3500 \times 8 = 28,000$.



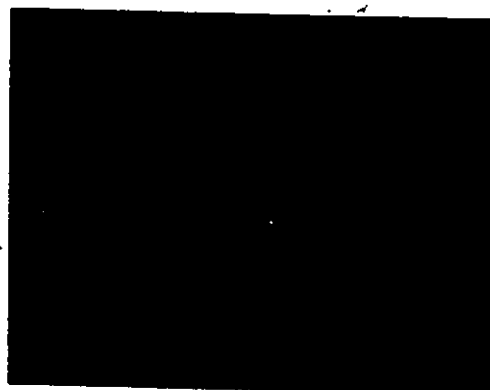
0 - 300 msec.



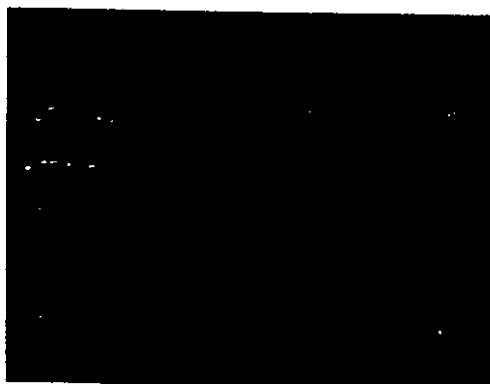
300 - 600 msec.



600 - 900 msec.



900 - 1200 msec.



1200 - 1500 msec.

FIGURE 48

DPCM Variable Quantizer Height
Reconstruction of the Sentence

"WE WERE AWAY A YEAR AGO"

The same word when synthesized using phonemes, would require

$$2(p \times n + n_1 + n_2 + n_3) \text{ bits}$$

where p is the number of predictor parameters (= 12)

n is the number of bits for each phoneme (10)

n_1 is the number of bits representing the phoneme pitch period

n_2 is the amplitude of the excitation

and n_3 is the duration of each phoneme

the 2 represents 2 phonemes in the word "NOON".

$$\therefore \# \text{ of bits} = 2(12 \times 10 + 8 + 8 + 4)$$

$$= 280 \text{ bits.}$$

In this example, the saving would be of the order of 100 times over straight PCM.

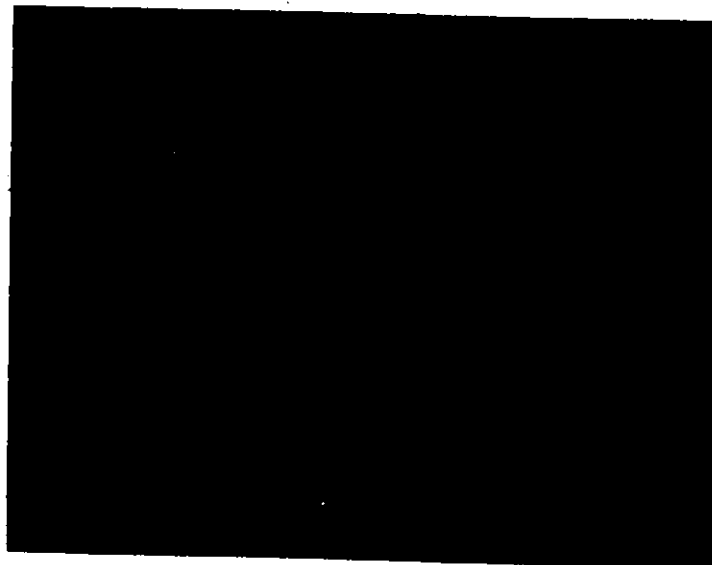
The words "NOON" and "HOW ARE YOU", were synthesized using phoneme synthesis. The predictor parameters, the pitch period and the excitation were obtained from an actual speech sample. Figures 49 and 50 show the phonemes synthesis of the words "NOON", and "HOW ARE YOU", respectively. The synthesis procedure is identical to that described in Chapter II using the linear predictor with triangular excitation, except that the same model is kept for the duration of each phoneme.

5.6. Summary

The rapid decay of the reconstructed signal which was observed on the IBM-360/65 computer was even more severe on the PDP-8/I. This shows that great care must be exercised when implementing speech synthesis and coding schemes on a machine such as the PDP-8/I. With some of the coding schemes, it was possible to keep the same model for two or three

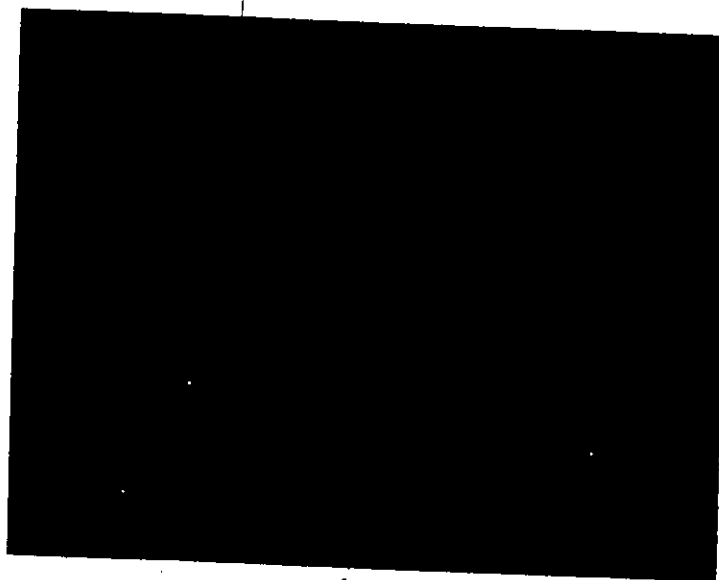


0 - 200 msec.

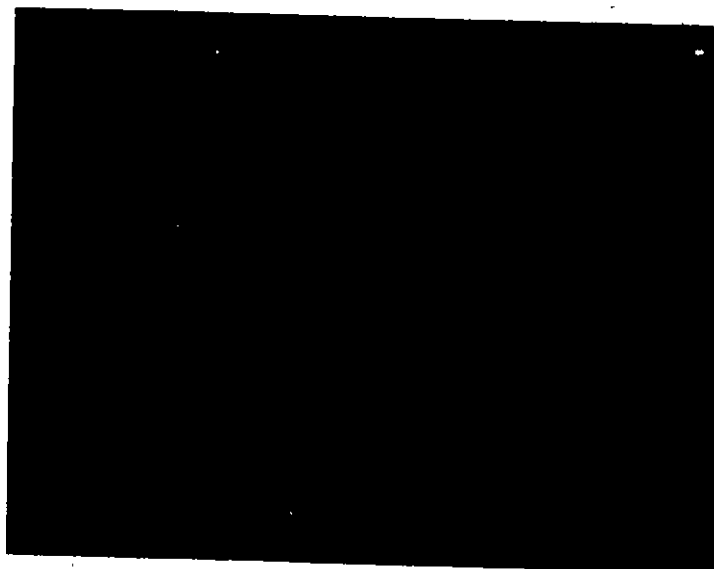


200 - 400 msec.

FIGURE 49 - Phoneme Synthesis Of The Word - "NOON"



0 - 200 msec.



200 - 400 msec.

FIGURE 50 - Phoneme Synthesis Of The Words - "HOW ARE YOU"

pitch periods when using the IBM 360/65 computer, without running into stability problems. This could not be achieved on the PDP-8/I, since any extrapolation of the predictor model invariably resulted in instabilities in the reconstructed speech signal.

The results of the initial conditions model show the rapid decay of the reconstructed speech signal. The coding schemes are able to follow the variations in the speech signal due to the constant correction that is applied at each sampling interval. In all three cases, there were no real signs of instabilities in the reconstructed speech signal. The results of the speech synthesis and coding schemes described in this chapter were recorded on magnetic tape to provide an audio output.

CHAPTER VI

CONCLUSIONS

The major conclusions and contributions of the research described in this dissertation may be summarized as follows.

6.1. Linear Predictor

1. It has been shown that by the proper use of initial conditions, the need to include zeros in the discrete model and the need to determine the nature of excitation were eliminated. This procedure has been shown to yield results comparable to the results obtainable from models that incorporate both poles and zeros as well as those that used special excitation functions.
2. The linear predictor when realized in the parallel or closed form has been shown to result in a significant improvement in the quality of reconstructed speech. This conclusion has been arrived at after subjective listening tests that are not readily expressible in numerical terms.
3. By an algorithm that incorporates adaptive adjustment of system order, it has been shown that a stable model is always realized.

6.2. Coding of Speech Signals

1. It has been shown that the determination of optimum quantizer levels in differential pulse code modulation schemes is not feasible in practice, as there exists no efficient algorithm for determination of the global minimum.
2. A new coding scheme that combines the advantages of DPCM and the open loop coding schemes that yields consistently superior results over the currently used techniques has been developed. A criterion for determining the quantizer level in the primary coding scheme, has

been shown to eliminate the need for determining optimum quantizer levels. With this scheme a bit rate of about 25,000 bits/sec that results in an articulation index of 0.65 for isolated words has been realized.

6.3. Speech Synthesis and Coding by Use of an FFT Model

1. A conceptually new model that utilizes the FFT algorithm has been developed for speech synthesis.
2. It has been shown that a spectral shaping technique, using FFT algorithm in conjunction with a peak-picking algorithm, yields a model that is superior to the linear predictor.
3. It has also been shown that, with the high computational speeds attainable by the use of FFT algorithm, efficient coding schemes, superior in performance to the DPCM scheme are realizable. A bit rate of 24,400 bits/sec with an articulation index of 0.7 for isolated words has been realized.

REFERENCES

- [1] DUDLEY, H., REISZ, R.R., WATKINS, S.A., "A Synthetic Speaker",
J. Acoust. Soc. Amer. vol.11, pp.169-177, 1939 a.
- [2] DUDLEY, H., "The Vocoder", Bell Lab. Record 17, pp.122-126, 1939.
- [3] KOEINING, W., et al, "The Sound Spectrograph", J. Acoust. Soc. Amer.
vol.18, pp.19-49, 1946.
- [4] OPPENHEIM, A.V., "Speech Spectrograms Using the Fast Fourier
Transform", IEEE Spectrum, pp.57-60, August 1970.
- [5] COOLEY, J.W., TUKEY, J.W., "An Algorithm for the Machine Calcula-
tion of Complex Fourier Series", Math. of Comp. vol. 19, no.90,
pp.297-301, 1965.
- [6] RICHARDSON, E.G., "Technical Aspects of Sound", vol.1,
Elsevier Publishing Co., 1953.
- [7] DAVID Jr., E.E., "Signal Theory in Speech Transmission",
I.R.E. Trans. Circuit Theory, pp.232-244, December 1956.
- [8] FLANAGAN, J.L., "Synthetic Voices for Computers", IEEE Spectrum,
vol.7, no.10, pp.22-45, October 1970.
- [9] RABINER, L.R., SCHAFFER, R.W., RADER, C.M., "The Chirp Z-Transform
Algorithm and its Applications", Bell System Tech. J. vol.48,
no.5, pp.1249-1292, May-June 1969.
- [10] SCHAFFER, R.W., RABINER, L.R., "System for Automatic Formant
Analysis of Voiced Speech", J. Acoust. Soc. Amer., vol.45,
pp.634-648, February 1970.
- [11] OPPENHEIM, A.V., "A Speech Analysis-Synthesis System Based on
Homomorphic Filtering", J. Acoust. Soc. Amer., vol.45,
pp.458-465, February 1969.

- [12] ITAKURA, F., SAITO, S., "Analysis Synthesis Telephony Based on the Maximum Likelihood Method", The 6th International Congress on Acoustics, Tokyo, Japan, C-5-5, August 21-28, 1968.
- [13] ATAL, B.S., HANAUER, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Amer., vol.50, part 2, pp.637-655, August 1971.
- [14] MARKEL, J.D., "Formant Trajectory Estimation from a Linear Least-Squares Inverse Filter Formulation", Speech Communications Research Laboratory Inc., Monograph No.7, October 1971.
- [15] ELIAS, P., "Predictive Coding - Parts I and II", IRE Trans. Information Theory, pp.16-33, March 1955.
- [16] McDONALD, R.A., "Signal to Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems - Particular Application to Voice Signals", Bell System Tech. J., vol.45, pp.1123-1151, 1966.
- [17] ATAL, B.S., SCHROEDER, M.R., "Predictive Coding of Speech Signals", Proc. Int. Congress on Acoust. C-5-4, Tokyo, Japan, August 1968.
- [18] ROSENBROCK, H.H., "An Automatic Method of Finding the Greatest or Least Value of a Function", Computer Journal, vol.3, pp.175-184, 1960.
- [19] FLETCHER, R., POWELL, M.D., "A Rapid Descent Method for Minimization", Computer Journal, vol.6, ISS 2, pp.163-168, 1963.
- [20] FADDEEVA, V.N., "Computational Methods in Linear Algebra", Dover Publications Inc., New York.
- [21] FLANAGAN, J.L., "Speech Analysis, Synthesis and Perception", Springer Verlag, New York (2 edition) 1972.

- [22] NOLL, A.M., "Cepstrum Pitch Determination", J. Acoust. Soc. Amer., vol.41, pp.293-309, February 1967.
- [23] NOLL, A.M., "Pitch Determination of Human Speech", Proceedings of the Symposium on Computer Processing in Communications, New York, April 1969, (Polytechnic Press).
- [24] FANO, R.M., "Short-time Autocorrelation Functions and Power Spectra", J. Acoust. Soc. Amer., vol.22, no.5, pp.546-550, September 1950.
- [25] GOLD, B., "Computer Program for Pitch Extraction", J. Acoust. Soc. Amer., vol.34, no. 3, pp.916-921, July 1962.
- [26] SCHROEDER, M.R., "Parameter Estimation in Speech", Proc. IEEE, vol.58, no.5, pp.707-712, May 1970.
- [27] BLACKMAN, R.W., TUKEY, J.W., "Measurement of Power Spectre", Dover Publications Inc., New York, 1958.
- [28] KUO, F.F., KAISER, J.F., "System Analysis by Digital Computer", John Wiley and Sons Inc., 1966.
- [29] GOLD, B., RADER, C.M., "Digital Processing of Signals", McGraw-Hill Book Co., New York, 1969.
- [30] RABINER, L.R., RADER, C.M., "Digital Signal Processing", IEEE Press, 1972.
- [31] FLANAGAN, J.L., "Some Properties of the Glottal Sound Source", J. Speech Hear. Res., vol.1, pp.99-116, 1958.
- [32] ROSENBERG, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", J. Acoust. Soc. Amer., vol.49, pp.583-590, 1971.
- [33] BURRUS, C.S., PARKS, T.W., "Time Domain design of Recursive Digital

- Filters", IEEE Trans. Audio and Electroacoustics, vol.AU-18, pp.137-141, June 1970.
- [34] EVANS, A.G., FISCHER, R., "Optimal Least Squares Time-Domain Synthesis of Recursive Digital Filters", IEEE Trans. Audio and Electroacoustics, vol.AU-21, no.1, pp.61-65, February 1973.
- [35] FREEMAN, H., "Discrete Time Systems", An introduction to the theory, John Wiley and Sons Inc., 1965.
- [36] HAMMING, R.W., "Numerical Methods For Scientists and Engineers", McGraw-Hill Book Co., Inc., New York, 1962.
- [37] KELLY, J.M., et al, "Final Report on Predictive Coding of Speech Signals", Bell Laboratories, June 1970.
- [38] BERANEK, L., "Signal Processing Methods for Voice Communications To Aircraft and Evaluation in Terms of Test Time Signal-To-Noise Ratio and Calculated Articulation Index", C.C.I.R. Special Joint Meeting, Geneva, 1971.
- [39] MARKEL, J.D., "FFT Pruning", IEEE Trans. Audio and Electroacoustics, vol.AU-19, pp.305-311, December 1971.
- [40] SHRIDHAR, M., CIRJANIC, B., "Digital Encoding of Speech Signals", Sixth Biennial Symposium on Communication Theory and Signal Processing, Queen's University, Kingston, Ontario, August 28-30, 1972.
- [41] HEWLETT PACKARD FOURIER ANALYZER, Model 5470A.
- [42] CUMMISKEY, P., JAYANT, N.S., FLANAGAN, J.L., "Adaptive Quantization in Differential PCM Coding of Speech", Proc. IEEE Int. Communications Conference, Seattle, Washington, June 1973.

VITA AUCTORIS

- 1943 Born on November 11th, in Kadina Luka, Ljig, Yugoslavia.
- 1959 Completed secondary education at Turves Green Boys School, Birmingham, England.
- 1965 Graduated from the South Birmingham Technical College, Birmingham, England, with the Higher National Certificate in Electrical Engineering.
- 1969 Graduated from the University of Aston in Birmingham, England, with the degree of B.Sc. in Electrical Engineering.
- 1970 Graduated from the University of Windsor, with the degree of M.A.Sc. in Electrical Engineering.
- 1973 Candidate for the degree of Doctor of Philosophy in Electrical Engineering at the University of Windsor.