A STUDY OF SAMPLING, SMOOTHING AND SEGMENTATION OF SPEECH FOR RECOGNITION BY COMPUTERS

A Thesis
Presented to
the Faculty of Graduate Studies and Research
The University of Manitoba

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in the Faculty of Science

by
A. R. Bibik
October 1970

TITLE: A STUDY OF SAMPLING, SMOOTHING AND SEGMENTATION OF SPEECH FOR RECOGNITION BY COMPUTERS

AUTHOR: A. R. BIBIK

ABSTRACT

This thesis studies the different aspects of speech recognition by computers. The work is divided into two parts: (a) expository part and (b) research part.

In the first part an historical introduction is presented followed by a study of the complicated process of speech production. Included in this part is a description of a spectrograph, vocoder, low-pass filter, high-pass filter and band-pass filter.

In Part Two, simulation of an ideal-filter on the IBM 360/65 is developed together with a study of different simple smoothing routines. In the final stages of this part a recognition algorithm which enabled us to recognize five out of six words for three different speakers is discussed.

ACKNOWLEDGEMENTS

I would like to extend sincere thanks to Profressor J. C. Muzio, my Thesis Supervisor for his extremely valuable assistance throughout this investigation.

Furthermore, I would like to thank Don Costin for his original ideas and encouragements in this field, as well as Miss Irene Rourke and Allan Yost, who, together with Don Costin, volunteered their voices for the research part in this thesis.

Secondly, I would also like to thank Profressors P. Dirksen and R. Collens for their time spent in reading this thesis.

Finally I would like to express my gratitude to the Operation Staff at the University of Manitoba Computer Centre for their co-operation in the production of the many necessary graphs and simulations, as well as Mrs. Christine Schneider for the many hours spent typing this thesis.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF FIGURES

LIST OF TABLES

EQUIPMENT USED FOR SPEECH PROCESSING

Analog Digital (A/D) Converter

Conversion of the analog voice data to IBM digital format is carried out using the Radiation Inc. A/D converter. This device samples all input data at a rate of 7000 cps. The data samples are subsequently multiplexed and written on seven channel, IBM, one inch magnetic tape. The digitized voltage levels range from 2047 to -2047, corresponding to analog signal voltages with a full scale range of plus two volts to minus two volts. The output from the A/D converter can be written on one to twelve channels.

IBM 360/65 Computer

System 360/65 is a general-purpose system which employs solid-logic integrated components. System 360 is designed to accommodate large quantities of addressable storage. Increased capacities are provided by the combined use of high-speed storage of medium size and large-capacity storage of medium speed.

Although the description of system 360/65 could be very long and complicated, it should be noted that the only parts which are of real interest and importance to this thesis is the disk which is used for partial storage and the magnetic tape which is used for storing the digitized data. The machine has been used extensively for running simulation programs.

CHAPTER I

GENERAL INTRODUCTION

Modern computers are capable of processing information at speeds considerably faster than man is capable of supplying it. A great loss of efficiency occurs at this man-machine interface chiefly because man must presently encode this information into machine language to communicate with the system. A great need therefore arises for devices with which man will be able to communicate directly in his own human speech.

The development of the computer has increased the need for man to answer the questions which have puzzled him for countless years: What is the nature of speech?, and, how is it perceived and classified by the ear? At this stage the investigations conducted by many men like Reddy [1], Helmholtz [2] and Lindgren [3] have provided us with an extensive list of facts about the speech signal and the organs by which it is produced.

It is known, for example, that the acoustic properties of the vocal tract cause selective transmission of the frequency components of the harmonically rich sounds generated by the glottis. The glottis is an aperture between the vocal cords. With sufficient pressure from a puff of air coming from lungs the cords, which are approximately one inch long for man and three eights of an inch long for women, are forced apart briefly allowing the puff of air to escape. The sounds are transformed into recognizable speech sounds by such articulators as lips, tongue, vocal cavities or the combination of two or more of these.

The spectrographic manifestations of this selective transmission process are the formants which indicate around which frequencies the excitatory energy has been concentrated by the vocal tract.

In this thesis by direct processing of the speech wave form, frequency and amplitude changes which combine to produce an utterance have been derived. All major computing work has been done on the IBM 360/65 and has been written and organized to be complete and self-contained account for a reader with a background in physics, computer studies and electrical engineering. However, properties of the speech signal, as well as the structure of the vocal tract will not be described in any great detail. (For a detailed treatment see "Visible Speech", Potter, Kopp, and Green, 1947.)

To obtain these results, six words were recorded on a tape recorder by three different speakers. The speakers were chosen by the pitch of their voice, i.e. low-pitch voice (man), medium-pitch voice (man), and high-pitch voice (woman). The recorded signal was then digitized at a frequency of 7000 samples/sec. and finally stored on a disk. By simulating a filter bank on an IBM 360/65 (Chapter III) an ideal filter comprised of forty band pass filters from zero to 7000 in steps of 175 cycles per second was produced. The filtered data was sampled using different time intervals, with both frequency and amplitude being recorded and stored. The parameter determining the sampling time was made variable so that a best sampling time could be obtained. By experimentation it was determined that a sample time of $175^{-\frac{1}{2}}$ sec. produced best results. The filtered results were then passed through a filtering routine (Chapter IV) which eliminated all noise produced by organs such as teeth, tongue and nasal cavities which are of different size and shape

for each speaker. The final result was smooth frequency and amplitude data which was stored on disk and plotted on the Calcomp plotter. The data was then scanned by a second routine to detect voice onset and end. All routines and filter simulations have been written using simple logic in order that hardware could easily be constructed to perform these processes.

Vocoding devices which are instruments that pass a speech signal to synthesizers along a narrow bandwidth channel, and machines which transduce artificial speech from spectrograms have demonstrated that speech sounds can be adequately characterized by specification of the frequencies and amplitudes of the first three formants and by specification of the type of excitation. (See Chapter II)

In Chapter II we give an historical introduction to the study of speech and a comprehensive study of what is presently known about speech signals and their production in the vocal tract.

Chapter III contains detailed descriptions for the construction of hardware filters in addition to filter simulation program written by the author.

Chapter IV discusses different smoothing routines used by the author. The recognition algorithm is given in Chapter V and the final conclusion in Chapter VI.

CHAPTER II

REVIEW OF PROBLEMS IN SPEECH ANALYSIS AND SPEECH SYNTHESIS

Introduction

Study of human speech has a long history. Some of the highlights of this research are specified in Section 2.1

Section 2.2 provides a basic introduction to the theory of speech. It describes the vocal apparatus, the spectrograph and the vocoder, thus serving as the basis for the remaining chapters.

2.1 Historical Introduction

The earliest known characterization of the sounds of human speech was made by the Hindu grammarians about 300 B.C. [4]. Their work consisted simply of describing the positions of the articulators necessary for the production of any speech sound. The very fact that their method is being frequently used by phoneticians and language teachers up to the present day attests that their research was most effective.

In 1679 a German scholar, Samuel Reyherr, published the "Characteristic Pitches" of French and German vowels [5].

The first frequency standard to be used was a brass wheel with equally spaced teeth demonstrated by British physicist, Robert Hooke in 1681 [6]. The wheel was rotated with the teeth striking a reed which produced a tone. The frequency of this tone could be specified in terms of the number of revolutions per second and the number of teeth on the wheel. A more widely used frequency standard was the

tuning fork developed by a British musician, John Shore, in 1711 [7]. A Frenchman by the name of Mical constructed several speaking machines between the years of 1750 and 1780 [8]. Unfortunately no details of these machines have survived.

In 1779 the Imperial Academy of Sciences of St. Petersburg announced a contest for the answer to two questions: What is the nature of the vowel sounds?, and, Is it possible to construct a machine to successfully synthesize these sounds? The first prize award was won by Dr. Christian Gottlieb Kratzenstein who used a set of organ pipes to produce the vowel sounds [9]. In 1790 Wolfgang Von Kempelen constructed the most elaborate and most documented machine of these times [10]. The important thing about his work is the fact that he was the first investigator who concentrated his attention on the consonants as well as vowels. His speaking machine allowed control of all the following factors:

i) The frequency of vibration of the vocal chords.

ii) The role of the nasal cavity.

iii) The position of the tongue.

iv) The role of the teeth.

v) The position of the lips.

vi) The manner of articulation.

vii) The volume of the resonant cavities.

He was the first to talk of modulation of the sound from the glottis by the transmission characteristics of the vocal cavity.

In 1835 a modified version of Von Kempelen's machine was constructed by Wheatstone. This machine was capable of synthesizing the back vowels and a few consonants [11]. Willis also became interested

in Von Kempelen's work and was able to synthesize vowel sounds using a reed and a funnel-shaped pipe. The most important observation made by Willis was that the vowel quality was dependent only on the length of the pipe and not on the frequency of the vibration of the reed. He also described the vowel sounds as a succession of damped vibrations [12].

In his "Die Lehre von den Tonempfinoungen" in 1862, Helmholz pointed out that the vocal cavity is a resonator whose resonances alone determine vowel quality [2]. He was able to separate the German vowels into two groups according to whether they exhibited one or two resonances. He then specified these resonances by comparison with some standard frequency source. His results are shown in Table 2.1

TABLE 2.1

Vowel Resonances as Perceived by Helmholz

Single Resonance	Double Resonance
ü - 175	i - 175, 2349
ö - 466	e - 349, 1976
ä - 932	u - 175, 1468
	o - 349, 1109
	a - 587, 1568

Modern investigations have shown that for every vowel there are three or more resonances in the vocal tract, some more predominant than others.

In 1870's detailed investigations of the consonants were carried out by Grassman, Michaelis and Trautmann. In these investigations they have distinguished two features of the consonants, the type of "noise" and the "characteristic pitch". Although they were unable to make any

scientific characterization of the types of noise due to the fact that the analyzing equipment was not in existence, they did however publish data on the characteristic pitches of the consonant sounds [13]. Their work has illustrated that the characteristic pitch for a particular consonant varied greatly upon which vowel or consonant sound precedes and followed that consonant. There is therefore, a continuous transition of the vocal resonances of the vowel into those of the consonant and vice versa. Recent investigations have shown that these transitions are the most important perceptual clues in the recognition of consonant sounds.

As time progressed, electronic measuring devices made the older mechanical methods obsolete. I. B. Crandall used the vacuum tube devices at Bell Telephone Laboratories to perform Fourier analysis of many speech sounds [14]. Fletcher investigated the perception of speech by using various forms of filters to distort the frequency spectrum of the speech signal [15].

In 1920's early forms of the spectrograph first made their appearance [16]. It was about this time that the term formant was used to describe energy bands in the speech spectrum. The term was first used by Hermann and has been in general use ever since. The first results of spectrographic investigations of the speech signal were published by Steinberg of Bell Telephone Laboratories in 1934 [17].

The work of Chiba and Kajiyama (1941) has greatly increased the knowledge of the means by which speech is produced in the vocal tract [18]. These men used X-rays to obtain a better picture of the articulators and the vocal tract.

In closing this historical section, it should be noted that although research in this field has been conducted for hundreds of years, many of the basic questions such as what is the necessary part and what is the redundant part of speech signal or how to determine the beginning and end of a word still remain unanswered. It is hoped that the present day, fast, large storage computers will simplify greatly the complicated work and research, and that this thesis will be of some help in this fascinating but intricate field.

2.2.1 The Speech Process

The various speech sounds will first be classified and their acoustic properties described. All classification is done on an articulatory basis.

2.2.2 Articulation

Speech sounds are produced on exhalation. The air stream passes through an opening between the vocal cords making them vibrate in voiced sounds, periodically modulating the stream in such a way as to produce a harmonically rich spectrum. This moving stream of air is acted upon by some parts of the vocal system, i.e. the throat, mouth, or nasal cavities (see Fig. 2.1) to create various acoustic disturbances from which a listener extracts linguistic information. To produce complex patterns of shifting resonances one must modify the size and shape of the vocal cavities through time-varying tongue and lip positions. The oral and throat cavities may or may not be coupled to the nasal cavities by the action of the valve at the rear of the mouth called the velum.

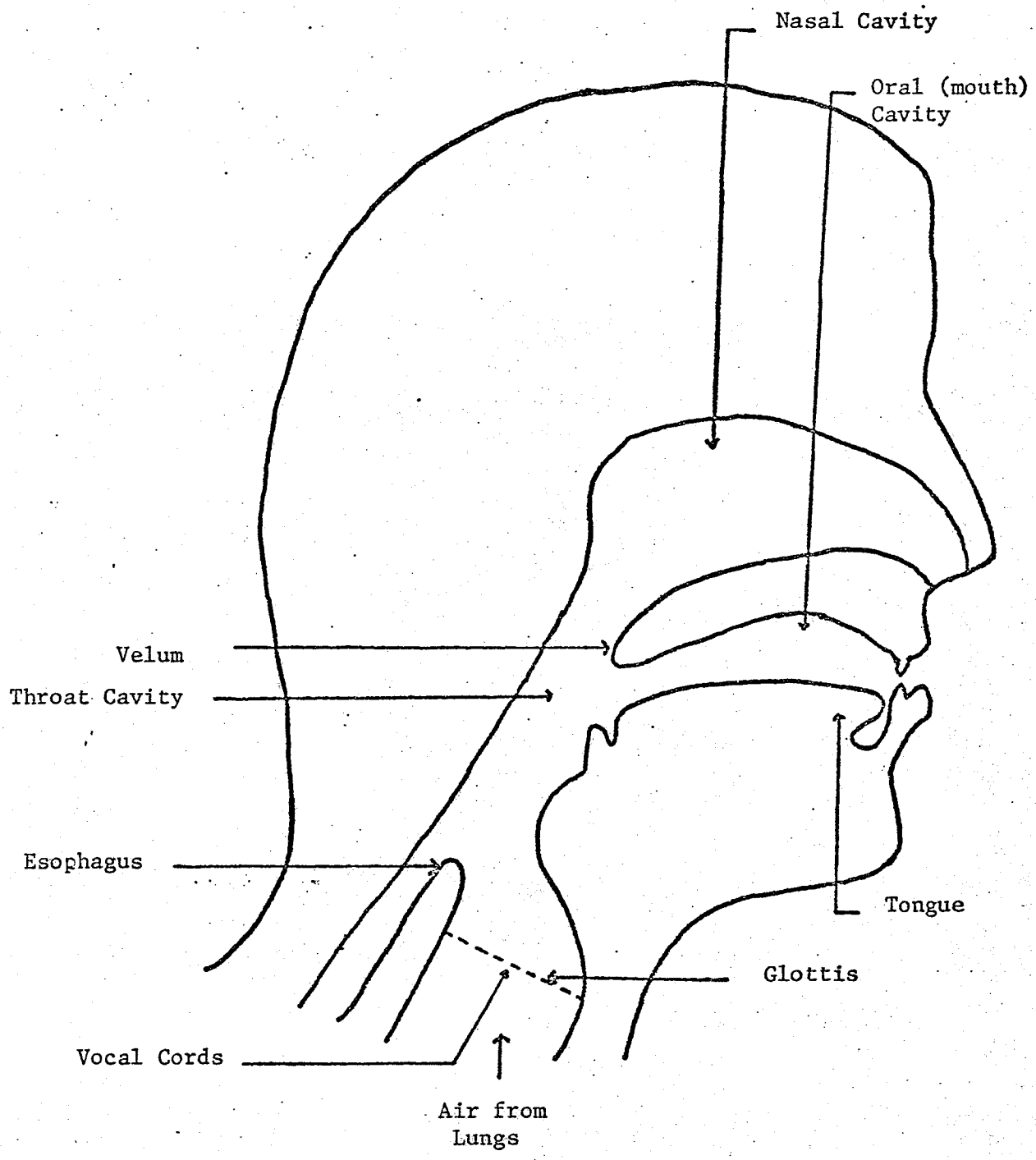


Fig. 2.1 The speech organs.

Turbulence (noiselike sound) is produced by the movement of the air across the edges of the teeth, and by partial closure of the vocal cords. In actual speech, these physical articulators are rarely stationary, but are enacting complex programs of gestures which have their analogs in the modifications of the acoustic output. Changes in output frequencies are perceived subjectively as length. By coupling the throat, oral and other nasal cavities one produces changing patterns of resonant frequencies. Excitation harmonics in the neighborhood of a cavity resonance are strongly transmitted, forming fairly narrow frequency regions of energy concentration (the formants), the first three of which are the most important for speech. The general range of formant frequencies produced by the vocal tract also depends to some extent upon the relative size of the cavities. This is the main reason why men with larger cavities often produce a louder range of such frequencies, and women a higher range. In addition, male voices, with their lower fundamental frequencies and closer harmonic spacing often show more clearly defined formants than those found in female voices.

The linguistic output possible from this acoustic system is a lexicon of thousands of words. These words in turn are composed of syllables which in turn are composed of roughly 40 distinct elementary sounds called phonemes.

2.3.1 Elementary Speech Sounds

The articulatory processes are traditionally classified into two groups, those associated with vowels and those associated with consonants. Vowels and consonants combine in speech to form syllables and the syllables combine to form words. Throughout the history of

speech recognition, the vowels have been studied rather more thoroughly than the other speech sounds.

2.3.2 Vowels

Vowels are voiced i.e. vocal cords are in vibration. The vocal tract is relatively open, i.e. there is an open passage between the vocal cords and the outside atmosphere. Different vowels are characterized by different tongue tip and hump (the back part of the tongue) positions, and by the degree of rounding of the lips. In the production of vowels, the breath stream excites coupled vocal cavities. If the vowel is voiced, the breath stream excitation consists of impulsive puffs which are repeated at the fundamental frequency of the vocal cords, and consequently has a spectrum in which the harmonic amplitudes decrease with frequency. Since the coupled vocal cavities are resonators, excitation harmonics which are in the neighborhood of a resonance will be strongly transmitted and will form regions of energy concentration for the particular vowel. If the vowel is whispered, the excitation of the vocal cavities is noiselike in character and the frequency spectrum will be continuous. The formant regions will still be present, but the relative formant amplitudes will be modified.

In speech certain vowel pairs often occur which are called diphthongs. In this case the formant frequency positions change smoothly between one vowel and the other of the pair. It is quite clear that diphthong cannot be sustained.

For a particular sustained vowel, the formant frequency positions vary between speakers and depend on the speaker's sex.

2.3.3 Consonants

The consonants are classified as vowel-like sounds, fricatives, and plosives (stops). Furthermore, the vowel-like sounds can be subdivided into glides and semi-vowels. For the vowel-like sounds as for vowels the vocal cords are in vibration. The glides are transitory in nature mainly because they are formed by rapid articulatory changes at the beginning of a vowel. Due to the fact that the size and shape of the vocal cavities are changing, the frequency resonances which are analogous to vocal formants are altering in position. As a result of these changes, there is a considerable steepness variation in the time track of resonances. This steepness depends greatly on the glide articulation speed. The four glides usually considered are |w|, |j|, |l|, and |r| as in we, you, let, and read.

Semi-vowels in contrast to glides may be sustained. These sounds are produced with closed mouth and open nasal cavities, i.e. they are nasalized. The frequency resonances of the semi-vowels are analogous to vowel formants. The semi-vowels are |m|, |n|, and |ŋ| as in me, no, and sing.

For fricatives which may be sustained, the air flow is usually predominantly turbulent in character. In their production, the air is usually passed through the narrow openings in the vocal tract or over the edge of the teeth. In addition, the vocal cords may or may not be in vibration. For example |s| in see is an unvoiced fricative, while |z| in zoo is voiced. The fricatives have a low acoustic power. The spectrum of these sounds may exhibit broad noise-like frequency bands, and certain frequency regions may be accentuated.

The plosives or stops are always of a transient nature. In most cases, a practically silent interval is followed by the removal of a block formed with the lips and tongue. If air is then released to the outer atmosphere, an acoustic pulse or burst will be produced. The burst is of much shorter duration than the acoustic co-relates of the other speech sounds. The plosives are usually low in acoustic power and may be voiced or unvoiced. The |p| and |t| in pat are unvoiced plosives, whereas |b| in bit is voiced. An example of a plosive which does not involve a burst is |k| in act.

Perceptually the fricatives are often hiss-like, but a sonorous quality is sometimes evident as in the |z| in zoo. The vowels and vowel-like sounds are noticeable sonorous in quality. The plosives on the other hand have more the quality of pops.

2.4 The Spectrograph

In order to get an exact understanding of voiced sounds, man has constructed many machines which break down the speech signal into its frequency components. These machines, known as spectrographs, have now become standard equipment in all laboratories involved in the research of voice analysis.

The spectrograph is constructed essentially of a set of band-pass filters (see Fig. 2.2) with considerable overlapping between filter frequency ranges. A simple filter is the RLC circuit shown in Fig. 2.3. This type of circuit has a very steady state and transient behaviour. The inputs of all the filters are connected in parallel to the source of the speech signal. The output of each filter is rectified and smoothed to eliminate all ripple or noise which occurs at the glottal rate of

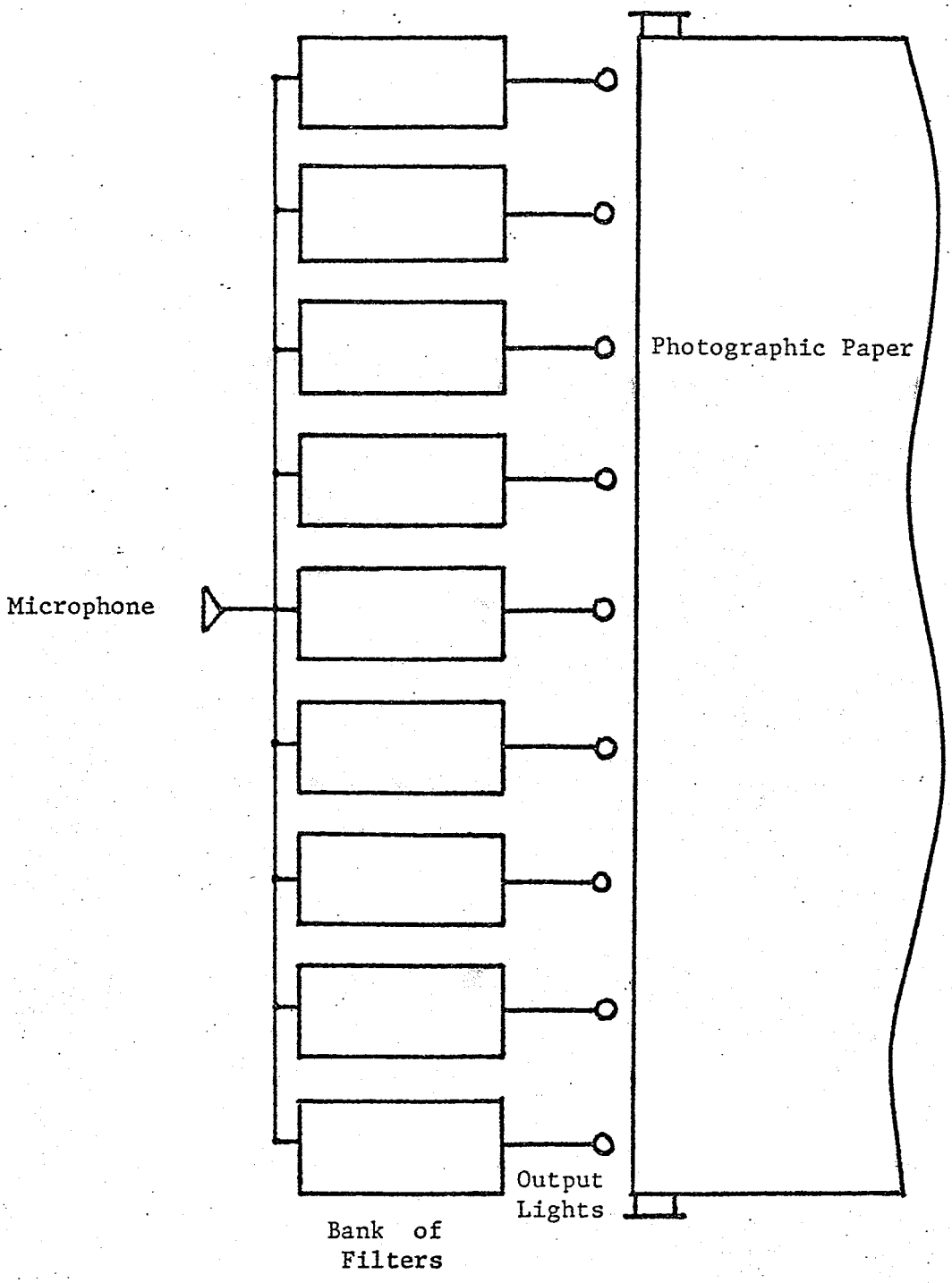


Fig. 2.2 Simple Diagram of a Spectrograph

approximately 100 cps. The presence of frequency in a specific band triggers off the light bulb connected to that band exposing the film which is constantly in motion. The final output of the spectrograph is usually presented in a photographic form (see Fig. 2.4) with time as the abscissa and frequency (number of cycles per second) as the ordinate. The amplitude or intensity of the output signal is represented by the degree of exposure of the photographic paper.

Fig. 2.4A(Potter, Kopp and Green, 1947) is an example of such a film from a spectrograph with filters of 45 cps bandwidth. Fig. 2.4B (Potter, Kopp and Green, 1947) gives the analysis of the same wave-form at 300 cps bandwidth. It is clear from Fig. 2.4A that sound consists mainly of harmonic components of the voiced or laryngeal frequency. In Fig. 2.4B these harmonics are not pronounced and the most important feature of this graph is the fact that the energy is concentrated about two or three main frequencies which vary in time. These bands of concentration of energy are caused mainly by the resonances of the vocal tract and cavities corresponding to the formants.

The degree of accuracy of a spectrograph has been demonstrated more recently with the invention of a machine which performs the spectrograph process in reverse order. A glass slide on which is painted an idealized representation of the band structure of the spectrogram is passed into this machine at exactly the same rate as the scale rate at which a similar drawing would have been produced. This machine then synthesizes a signal which it transduces to sound which is found to correspond very closely to the original word whose spectrum was copied.

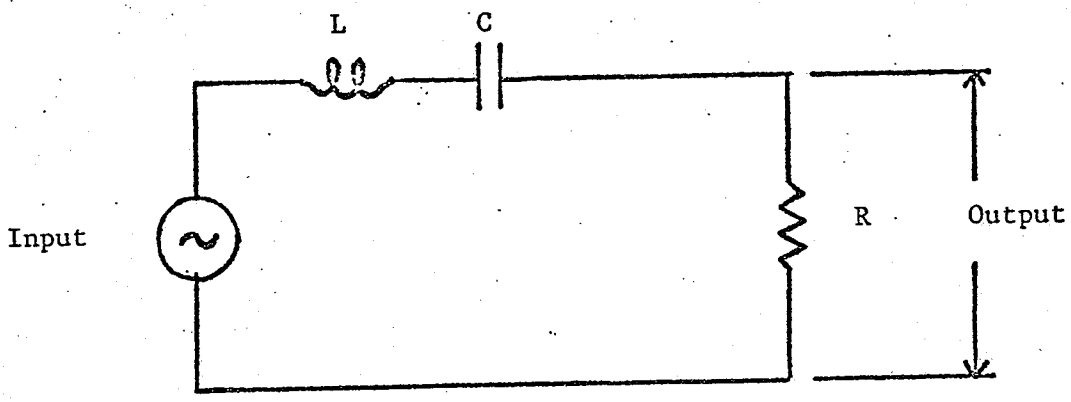


Fig. 2.3 Simple RLC Filter

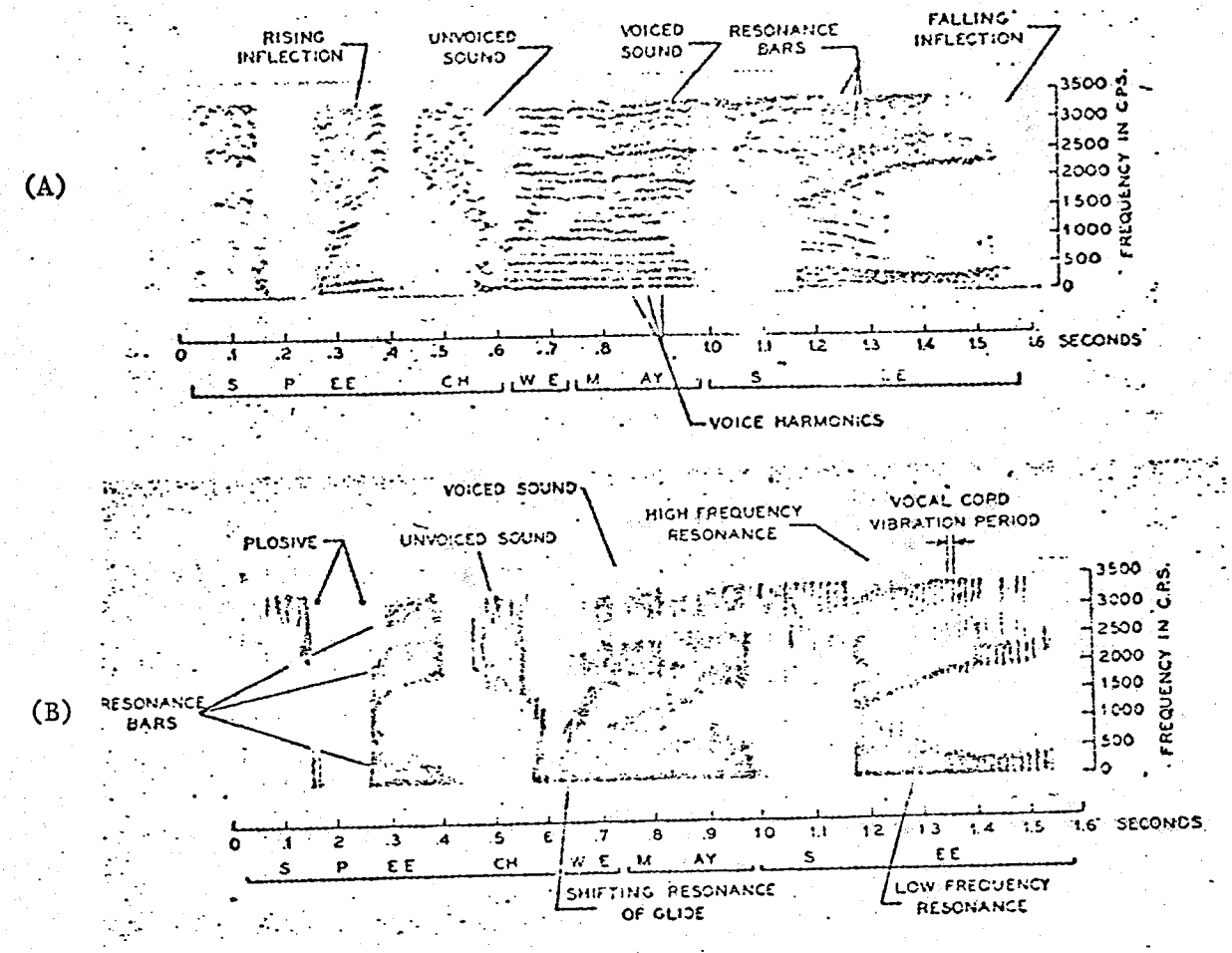


Fig. 2.4 (A) Sound spectrum from an array of narrow - band filters (45 cps)

(B) Sound spectrum from an array of wider - band filters (300 cps)

2.5 Vocoders

The human speech signal contains frequency components extending from tens of cycles per second to about 10 kilocycles per second. However, it is sufficient for recognition of the speech signal to limit transmission to a channel extending from approximately 300 cps to 3000 cps as is the case with the telephone system. It has been proved that a bandwidth of less than 10 cps is sufficient to transmit a signal corresponding to one formant frequency where the channel has a signal-to-noise ratio * of 40 db** or greater [19]. It should be noted that noise is often introduced in practical circuits by extraneous voltage signals coupled into the circuit from the surroundings. The most common noise is the 60 cps power lines.

A vocoder is a device which achieves bandwidth compression of the speech signal. In the vocoders, signals corresponding to the frequencies of the first three formants and their respective amplitudes are extracted from the speech signal and passed to a synthesizer along a narrow bandwidth channel. The most important requirement for a vocoder is to ensure that the synthesized speech be indistinguishable from the original speech signal. It has been shown by Flanagan that the amplitudes of formants are functions of the formant frequencies thus making their transmission redundant.

---

* Signal-to-noise ratio (s/n): The energy of a desired event divided by all the remaining energy (noise) at that time.

** Decibel (db): A unit used in expressing power or intensity ratios. $20 \log_{10}$ of the amplitude ratio or $10 \log_{10}$ of the power ratio.

CHAPTER III

3.1 Introduction

In any speech recognition apparatus, the most important component must be the filter bank that breaks down the speech signal into its frequency components.

In this chapter filters and filter banks will be discussed (3.1, 3.2, 3.3). Their construction and response characteristics will be stated as well as the characteristics of an ideal filter. In Section 4 the author's ideal filter will be presented and the complete program given in Appendix A. In all calculations standard notation is used, i.e. C = capacitance, L = inductance, ω = angular frequency, Z = impedance and X = reactance.

3.2 Low-pass Filter

A low-pass filter is a two-port filter which passes all frequencies less than a certain cutoff frequency ω_c and attenuates all frequencies greater than ω_c.

Figure 3.1 shows a filter which is called T section, and 3.2 a π section filter. For each of these filters, $X_a = \omega L$ and $X_b = -1/\omega C$. The name of the class indicates that X_a and X_b vary in an inverse manner with frequency, one being directly and the other inversely proportional to frequency. Another name for this class is constant-k filter. If we define k as:

$$k^2 = Z_a Z_b$$

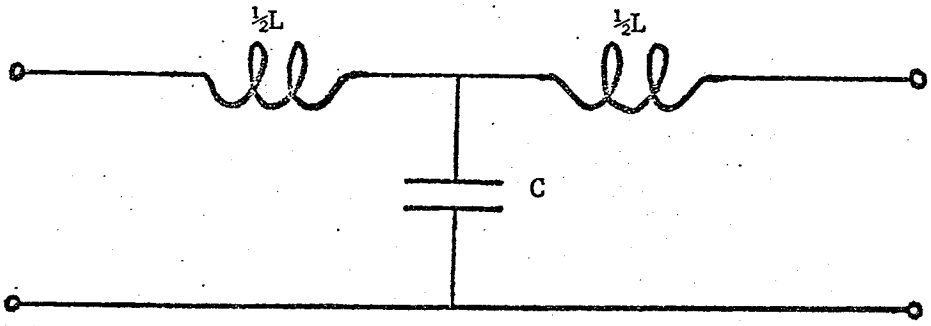


Fig. 3.1 Low-pass T Section Filter

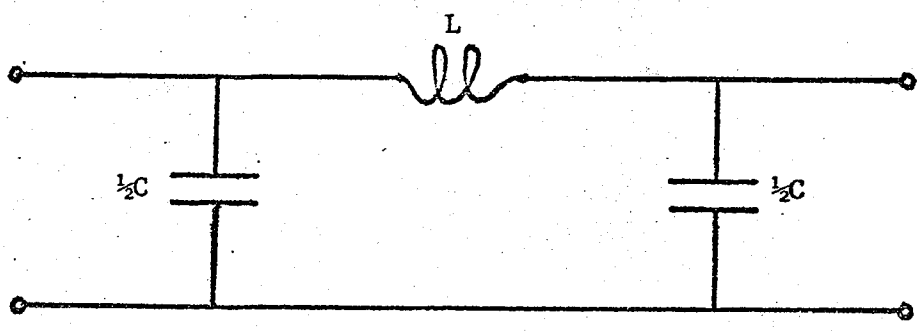


Fig. 3.2 Low-pass π Section Filter

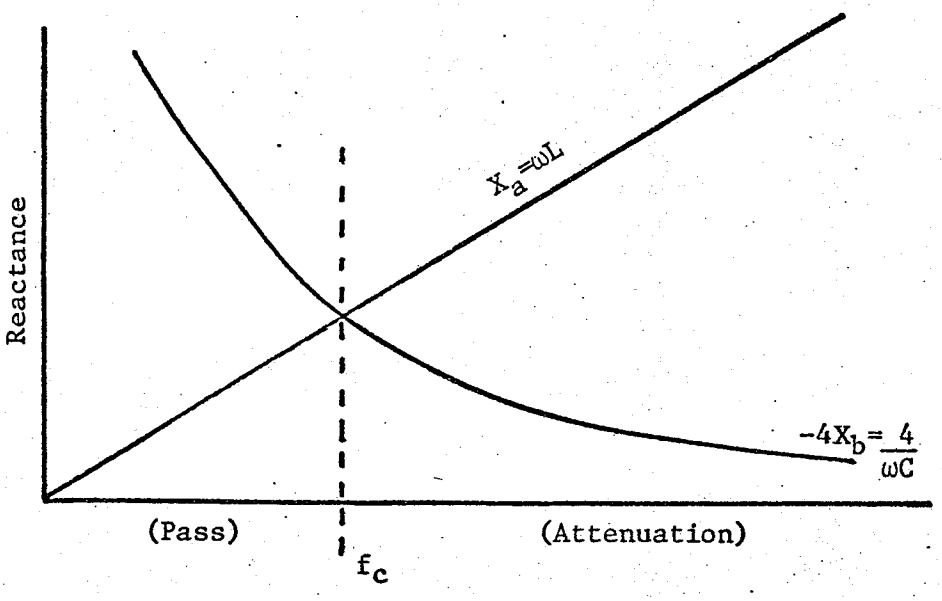


Fig. 3.3

then we can compute k for the sections of Figures 3.1 and 3.2 and find it to be constant, i.e. independent of frequency:

$$k^2 = Z_a Z_b = -X_a X_b = (\omega L)\left(\frac{1}{\omega C}\right) = L/C$$

In Figure 3.3 X_a and $-4X_b$ are plotted as functions of frequency. This kind of diagram is highly desirable for showing pass and attenuation bands. Due to the fact that the cutoff frequency is that at which $X_a = -4X_b$, the curves intersect at exactly this frequency. Furthermore, because the pass band is the range in which X_a lies between $-4X_b$ and zero, the diagram shows clearly that the filter is a low pass filter. To find f_c the cut-off frequency we write:

$$X_a = -4X_b$$

$$\omega_c L = 4\,\frac{1}{\omega_c C} \qquad \text{or} \qquad \omega_c^2 = \frac{4}{LC}$$

$$\omega_c = 2\pi f_c = \frac{2}{(LC)^{1/2}} \qquad \text{or} \qquad f_c = \frac{1}{\pi (LC)^{1/2}}$$

This is extremely useful for designing a low-pass filter, since the formulas can be used in reversed order, i.e. knowing the desired cutoff frequency L and C for the filter can be derived.

3.3 High-pass Filter

The high-pass filter is a two-port filter which passes all frequencies greater than a set cutoff frequency ω_c and attenuates all frequencies less than ω_c.

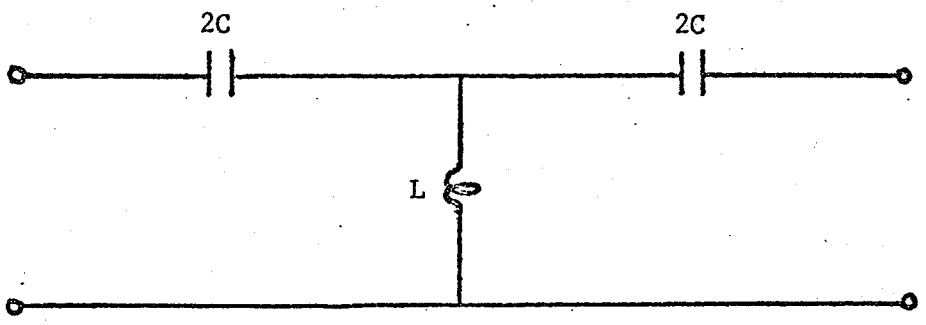


Fig. 3.4 High-pass T Section Filter

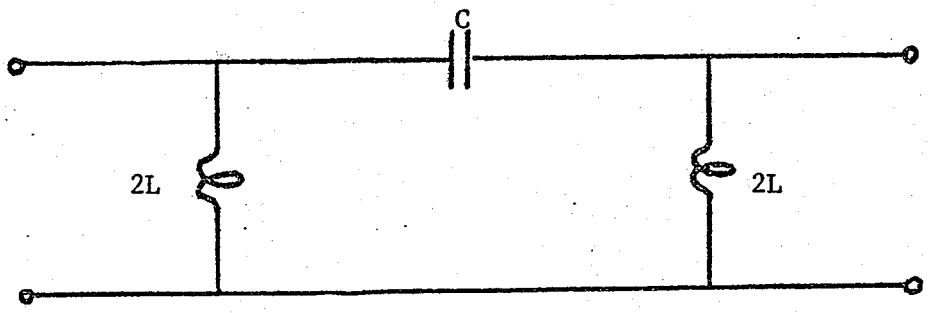


Fig. 3.5 High-pass π Section Filter

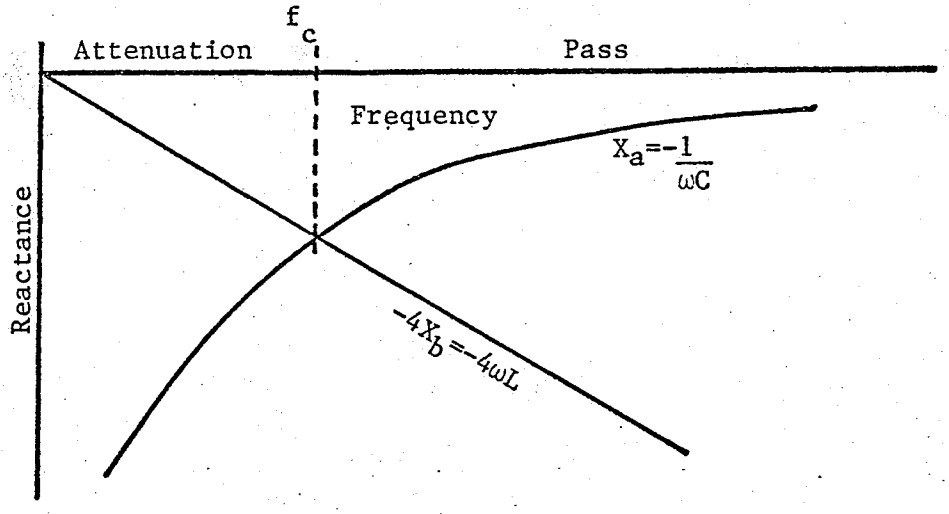


Fig. 3.6

By interchanging inductance and capacitance in the filters of Figures 3.1 and 3.2, one obtains a pair of high-pass filters shown in Figures 3.4 and 3.5. As before, one is a T section and the other a π section. To prove that these are inverse-network filters we write:

$$Z_a Z_b = -X_a X_b = \frac{L}{C} = k^2$$

again showing that k is independent of frequency. To prove that they are high-pass filters, we plot X_a and $-4X_b$ as in Figure 3.6. If we apply the usual criterion that X_a lies between $-4X_b$ and zero in the pass band, we find that this condition is met at frequencies higher than f_c or the point of intersection. To find the cutoff frequency we write:

$$X_a = -4X_b$$

$$-\frac{1}{\omega_c C} = -4\,\omega_c L \qquad \text{or} \qquad \omega_c^2 = \frac{1}{4LC}$$

$$\omega_c = \frac{1}{2(LC)^{1/2}} \qquad \text{or} \qquad f_c = \frac{1}{4\pi(LC)^{1/2}}$$

3.4 Band-pass Filters

A band-pass filter is usually used when frequencies between a lower cutoff frequency f_1 and higher cutoff frequency f_2 are to be transmitted freely, while other frequencies, both higher and lower are to be attenuated. Figures 3.7 and 3.8 show the inverse or constant-k type of band-pass network for both T and π sections and Figure 3.9 shows the reactance curves that apply to either. The pass band is the frequency range in which X_a lies between $-4X_b$ and the axis, extending

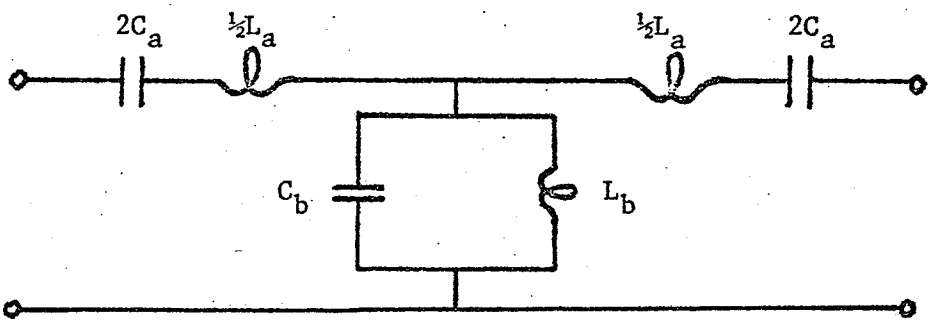


Fig. 3.7 T Section Band-pass Filter

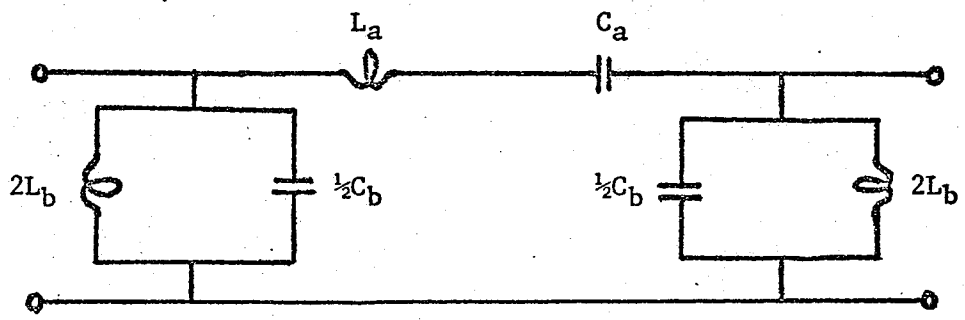


Fig. 3.8 π Section Band-pass Filter

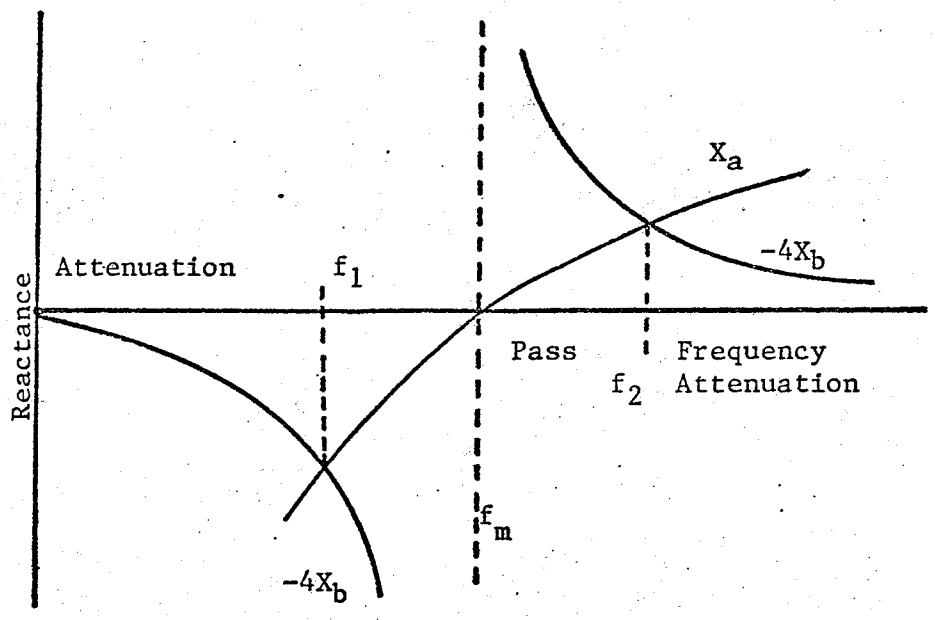


Fig. 3.9

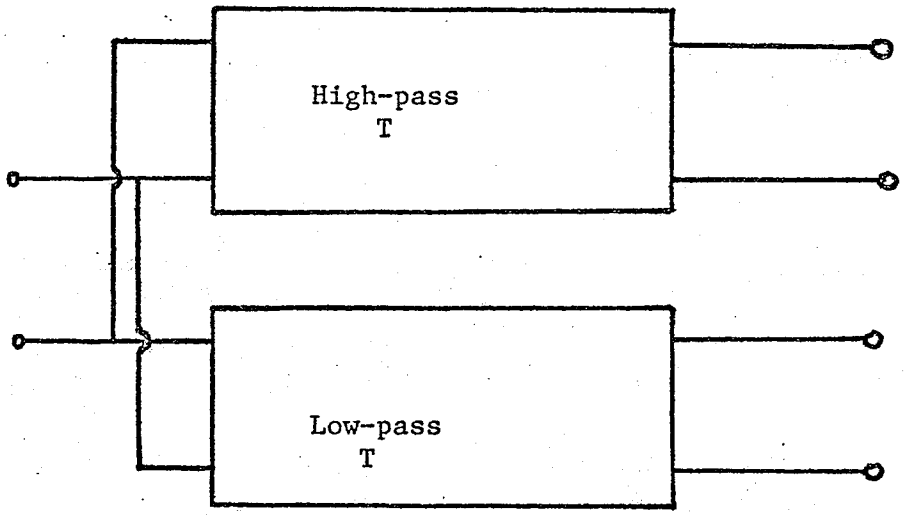


Fig. 3.10 Filters in parallel

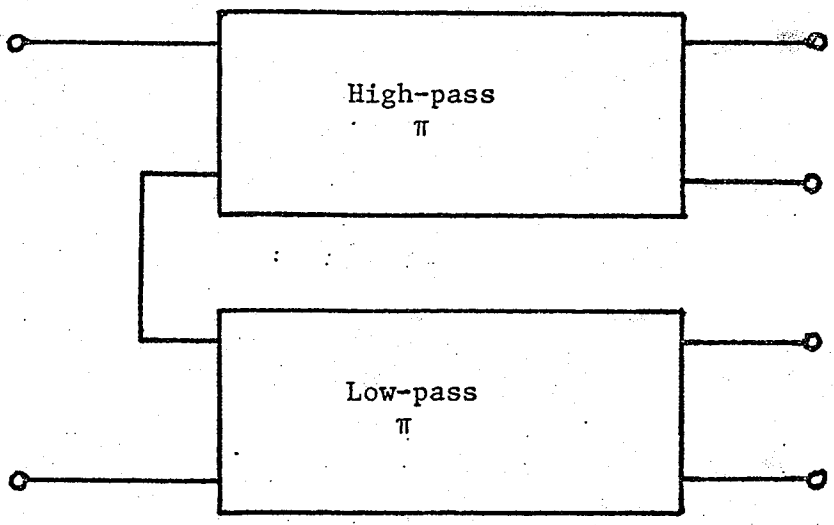


Fig. 3.11 Filters in series

between the two points of intersection of the reactance curves.

One should note that X_a and X_b are both resonant at the same frequency near the middle of the pass band, i.e. $L_aC_a = L_bC_b$. If this was not the case then there would be an attenuation band in the middle of the pass band.

When several frequency channels are to be separated by band-pass filters, several filters may be connected in parallel or in series, as in Figures 3.10 and 3.11. In voice analysis many filters are connected to form a filter bank. The author in his experimentation used a bank consisting of forty filters separating the range into steps of 175 cycles per second. However, it was discovered that twenty filters with the same frequency bands would have been sufficient.

3.5 Ideal Filters

An ideal band-pass filter transmits, without any distortion, all of the signals of frequencies between a certain frequencies f_1 and f_2. The signals of frequencies below f_1 and above f_2 are completely attenuated (see Fig. 3.12).

Unfortunately it is not possible in practice to build such a circuit with crystal ball properties. To obtain the same results the author performed a simulation using the IBM 360/65. Hence the hardware which will replace the ideal filter, i.e. give the same results as an ideal filter, will have to be a small special purpose computer.

This special purpose computer can be easily constructed with the aid of an A/D converter, register, shift register, comparitor and a counter. The number coming from the A/D converter would be compared with the number in the register (originally zero). The number of changes

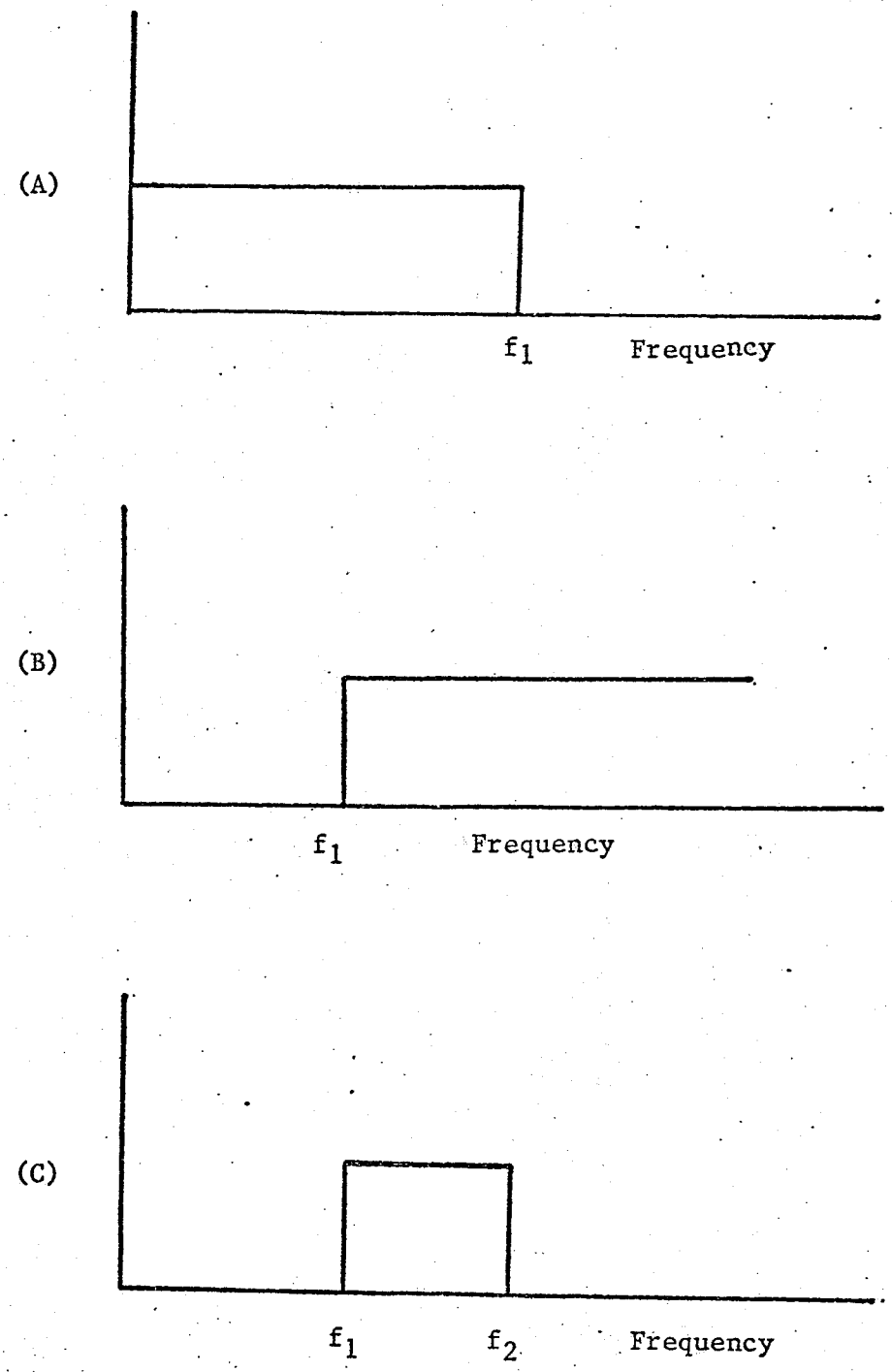


Fig. 3.12 Ideal Filters (A) Low-pass
(B) High-pass
(C) Band-pass

would be counted by the counter. After forty points the number in the counter would be placed in a shift register, shifted right by one, i.e. divided by two giving the number of vibration cycles per that specific time sample. As one can see, there is no real problem in the construction of such equipment.

The program which performs this simulation is called "FILTER" (see Appendix A). The input data for this program is the speech signal which is recorded on a tape recorder and digitized using the Radiation Inc. A/D converter at the rate of 7000 samples/sec. The result is a set of integers between plus and minus 2047, which are stored on disk in blocks of 1200 numbers. Each block is labelled with a number from one to 255. The program which performs this blocking is called "VOICE" (Appendix A).

There are two variables which control the input to "FILTER". The variable "BLOCK" contains the value or label of the block where the desired word starts, whereas "NO" contains the number of blocks which comprise the given word. The program samples the data at a specified time rate controlled by the variable "IN". The rate used by the author is forty which is 40 x 1/7000 sec. First of all the first two integers are compared. If $A_2^* > A_1$, control is passed to the part of the program which passes all integers for which $A_i > A_{i-1}$, until the condition is not satisfied. At this point variable "FREQ" is incremented by one and control is passed to the part of the program which passes all integers for which the condition $A_i < A_{i-1}$ is satisfied. As before, when this test fails "FREQ" is incremented by one. At the completion of examining forty points "FREQ" is divided by two and multiplied by 7000/40 giving the number of cycles per second for that specific time interval. This number is then passed to the subroutine called "IPRINT" which fits it into

---

* Note- A_1 and A_2 are the elements of the array "A" used in "FILTER"

a specific band-pass range and stores it on disk. The final result stored is an integer between one and forty which is the number of the band-pass filter passing that specific frequency.

In addition to the frequency analysis, "FILTER" also performs amplitude analysis. This is accomplished by comparing the absolute value of each integer and recording the largest one as the amplitude of the given time interval.

The final result for this program is plotted by the subroutine "IPLOT" which produces two curves, i.e. frequency vs time sample (thick curve) and amplitude vs time sample (thin curve). For each graph a heading is printed specifying the word represented by the plotted frequency combination, and the speaker pronouncing the word. The difference between two horizontal bars is 175 cps. The frequency range is from zero to 7000 cps. The X axis represents the time sample with each sample being 40/7000 sec. The output of "IPLOT" is presented in Table 3.1. The top graph on each page in this Table represents the raw spectra of the word specified by the heading. It should be noted that in the case of A.Y., the spectra is sometimes flat at the top and bottom. This characteristic is produced by the input into A/D converter reaching the saturation point, i.e. $\pm 2V$. The reason for allowing the input to go into saturation is the fact that under normal conditions the signal produced by I.R. is so small that it resembles a straight line. Furthermore, by observing the two graphs on each page it can be seen that the thin curve, i.e. amplitude trace in graph two is an enlarged shape of the overall positive shape of graph one. It should be noted that Table 3.1 has been divided into two parts with Part A on page 29 and Part B in Appendix B.

TABLE 3.1 (Part A)

Note - The area indicated by * represents solid black spectra eliminated by the reducing instrument.

THAT BY A.Y.

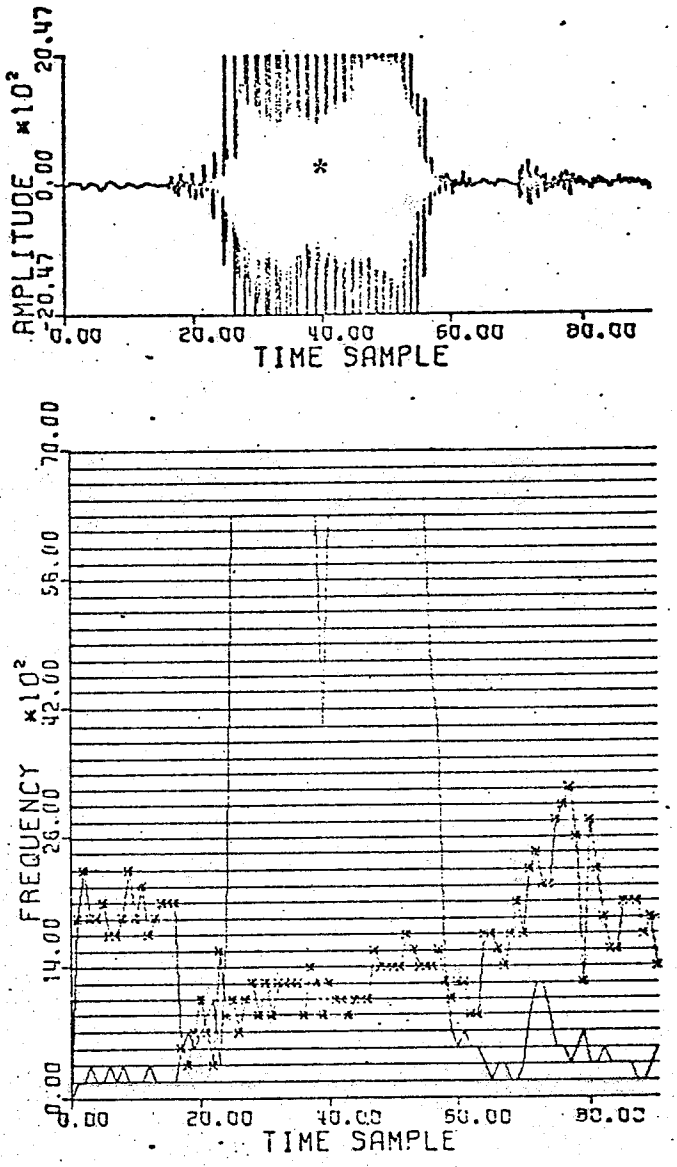
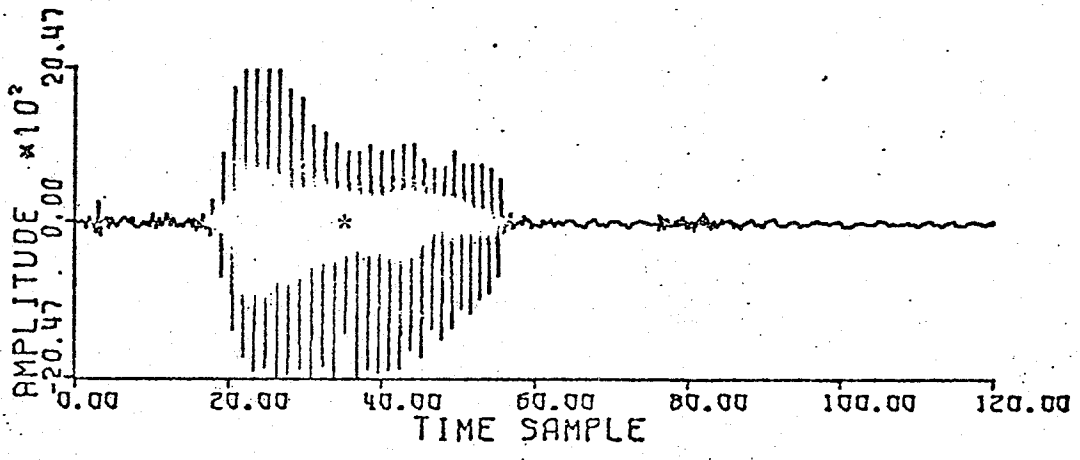


TABLE 3.1 (continued)

THAT BY D.C.



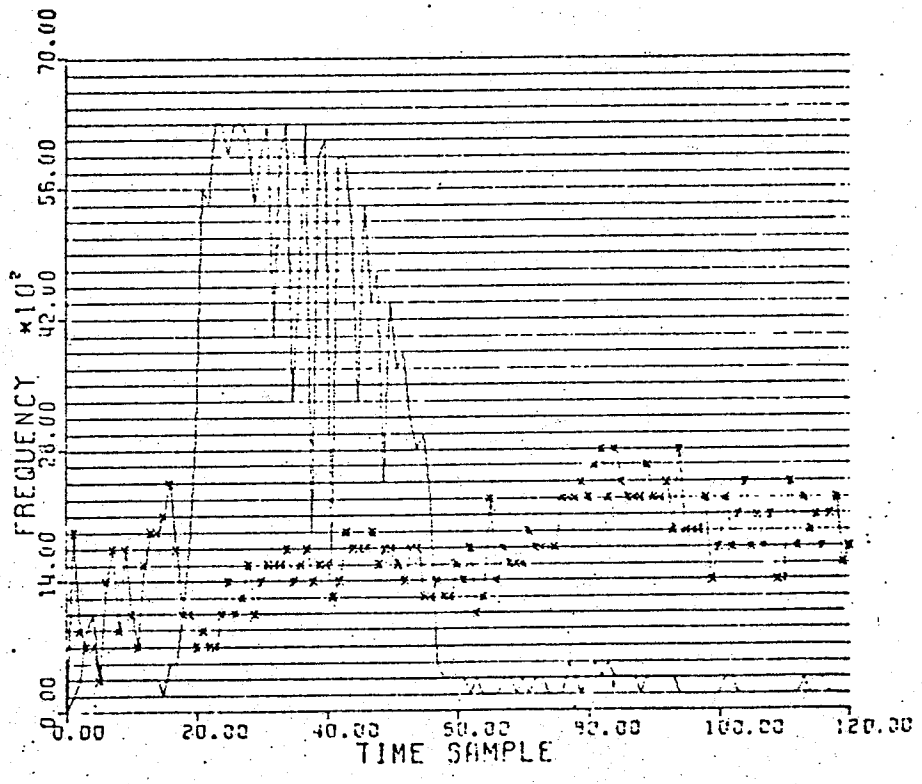
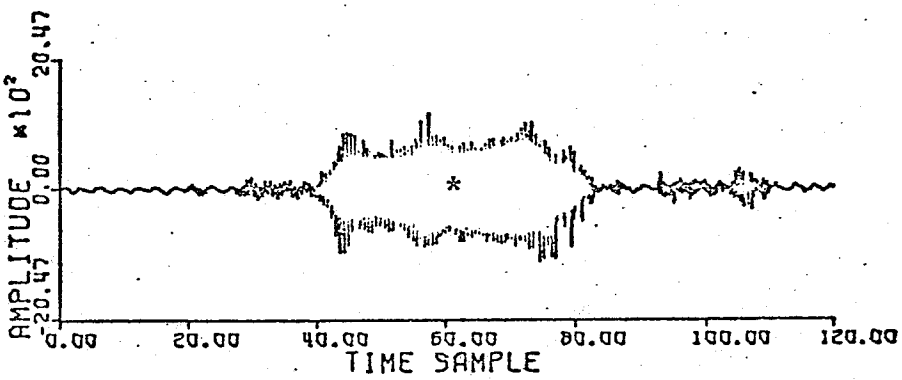
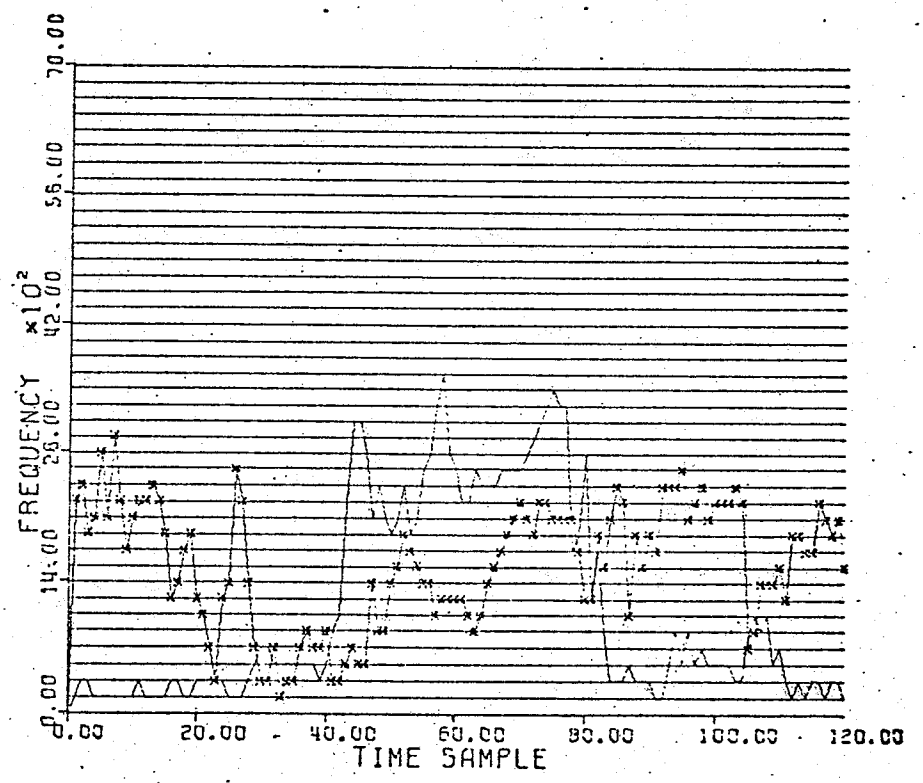


TABLE 3.1 (continued)

THAT BY I.R.





Continued on page 32

CHAPTER IV

SMOOTHING AND SEGMENTATION OF FILTERED DATA

4.1 Introduction

As can be seen from Table 3.1 the general shape of the fundamental wave for each word is basically similar. The only difference is the position of the wave in the frequency range as well as the harmonics of each wave. In this chapter a smoothing technique developed by the author will be presented as well as a plotting routine which eliminates inbetween word noise.

4.2 Smoothing Routine

When writing the smoothing routine, two very important objectives were kept in mind: (a) the routine had to serve the purpose, (b) it had to be as simple as possible. The purpose of the routine was to produce as unique as possible an output for each of the six words regardless of the speaker. The reason for simplicity was the fact that in the final stage the author's objective is to be able to replace software with hardware.

By studying the graphs it was observed that the waves for each word, regardless of the speaker, were comprised of almost identical fundamental wave and harmonics, which were different for each speaker. The reason for the great differences in harmonics for each speaker was assumed to be the fact that no two people have the same vocal tract

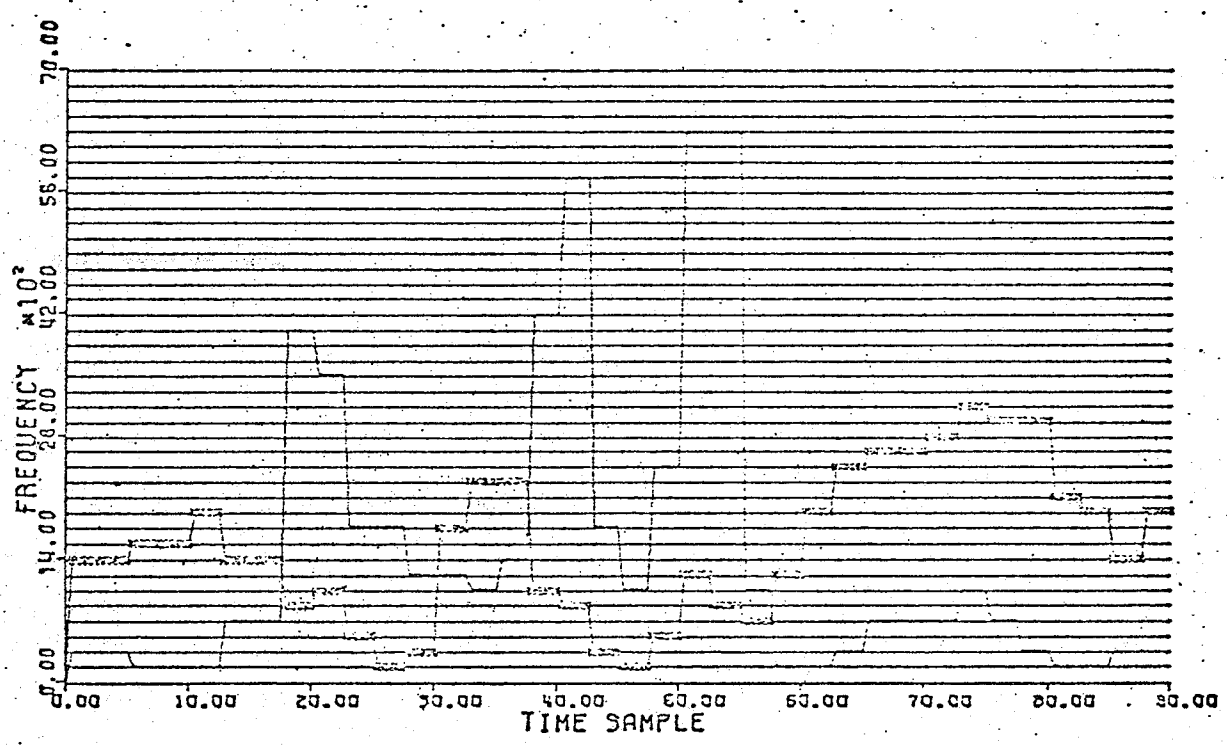
structure. Again assuming that the harmonics are produced by such parts of the vocal tract as spaces between teeth or different vibrations of lips, tongue or cheeks, one can easily see how these differences are bound to arise. It is worth noting that in a system whose main purpose was to recognize speakers, it would be the harmonics which would play the main function in the recognition process. However, our purpose is the opposite one, i.e. identification of the words irrespective of the speaker. Furthermore, it will be noticed that the average length of a phoneme is approximately 600/7000 sec. keeping in mind that phonemes are the basic sounds which form recognizable words. A fluctuation in frequency lasting for only 40/7000 sec. would seem to be too small an interval to be detected by the ordinary human ear. It is the author's belief that what the human ear detects is the changes in the average frequency, and interprets the words from that information. A more detailed consideration of these questions will be given in Chapter V. Taking all the assumptions into consideration, the author decided to eliminate all the harmonics leaving as the final result only the fundamental wave. The drawback of such a process is the fact that the smoothing technique if carried too far, can eliminate together with the redundant parts or meaningless changes for our purposes, parts of the wave which are necessary to distinguish between words. This of course raises the primary question which is how to determine which are the necessary and which the redundant parts of each graph. The author's method was to test the routines on all words for each speaker and see which produced the best results.

Many routines have been tried. The first was to take an interval of 5 x 40/7000 sec., find the average frequency throughout the interval, and set this average as the frequency of the whole of that

specific time sample. This routine was called "Straight Five Averaging". As can be seen from Table 4.1, the routine did not produce a unique representation of the word "exhibits". It should be noted that although the routine was tested on all words for each speaker, the word "exhibits" best demonstrates the general failure of the routine. The reason for the failure of this routine was attributed to the fact that a variation in frequency between two filtering time intervals increased the average of only the smoothing interval in which the variation occurred. Furthermore, the method is somewhat arbitrary in the sense that a comparatively small change in the starting position could cause considerable variations in the smoothed graph. To rectify this problem another routine was designed. In this routine, a variation between two points reflected on the position of the average preceding that interval as well as the one following it. The routine "Six Point Overlap" was tested on all words. However, as in the previous case, only the results for the word "exhibits" which demonstrates the improvements over the last method are given (see Table 4.2). In this routine, the first six points represented the first smoothing time interval. From that point on the interval was increased by the next three points which replaced the first three points. Hence, if the original interval I_1 consisted of points P_1 to P_6, the next interval I_2 would be comprised of points P_4 to P_9 (see Fig. 4.1). As can be seen, each interval overlaps the previous interval by three points. The resultant wave produced was much smoother than the one produced by "Straight Five Averaging", i.e. the sharp discrepancies were reduced considerably. Nevertheless, the wave was far from unique for the word "exhibits". The extra features imposed on the fundamental wave

TABLE 4.1

EXHIBITS BY A.Y.



EXHIBITS BY D.C.

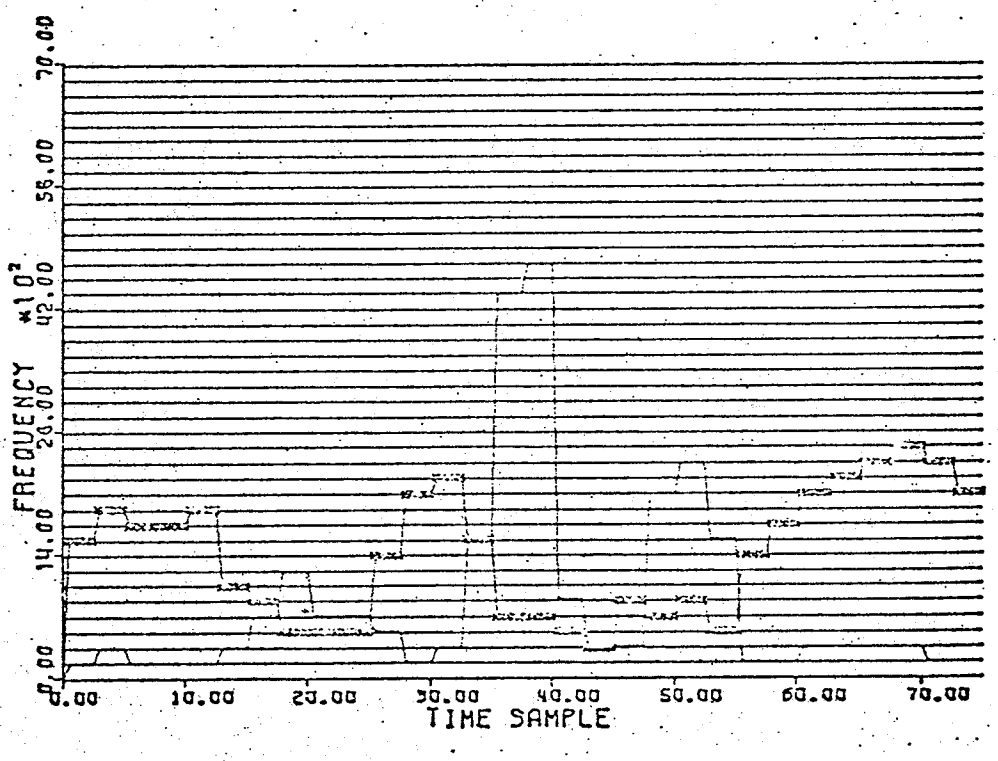
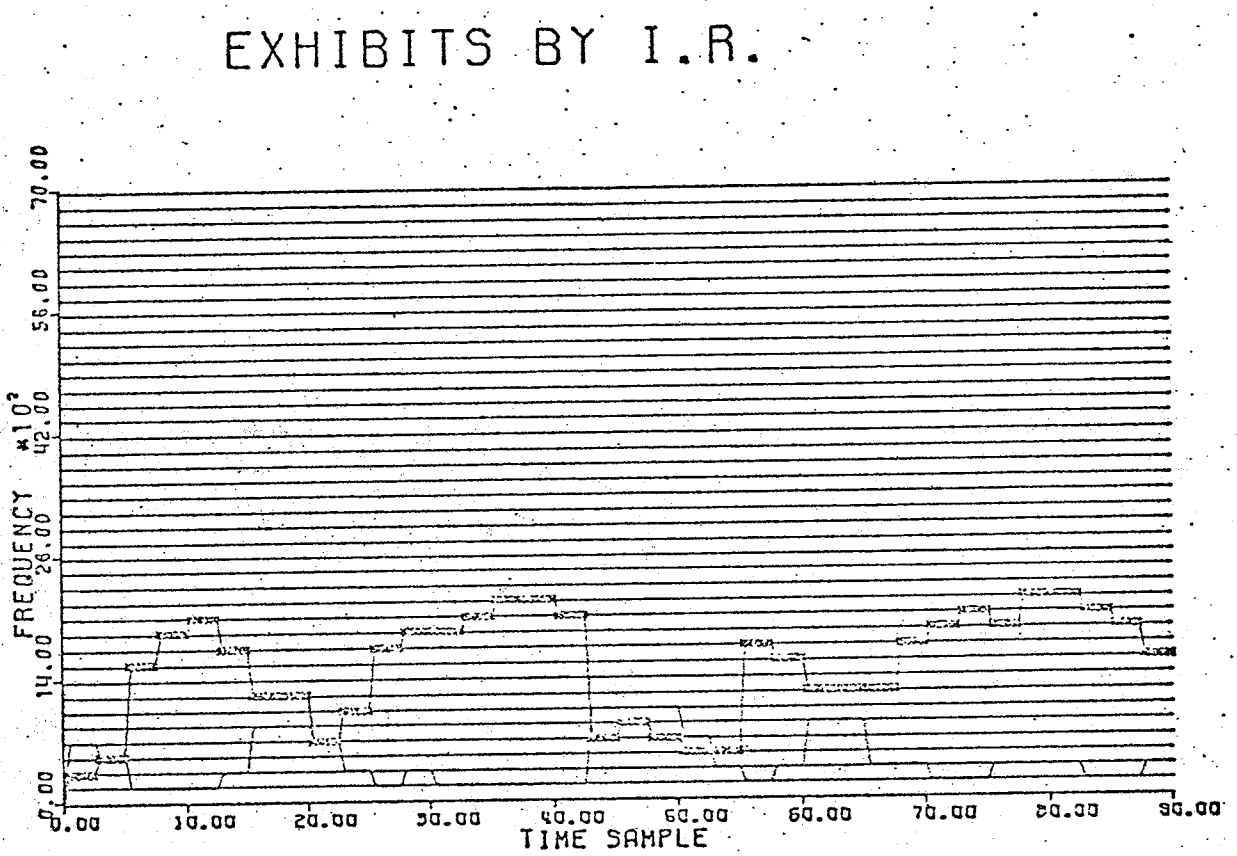


TABLE 4.1 (continued)



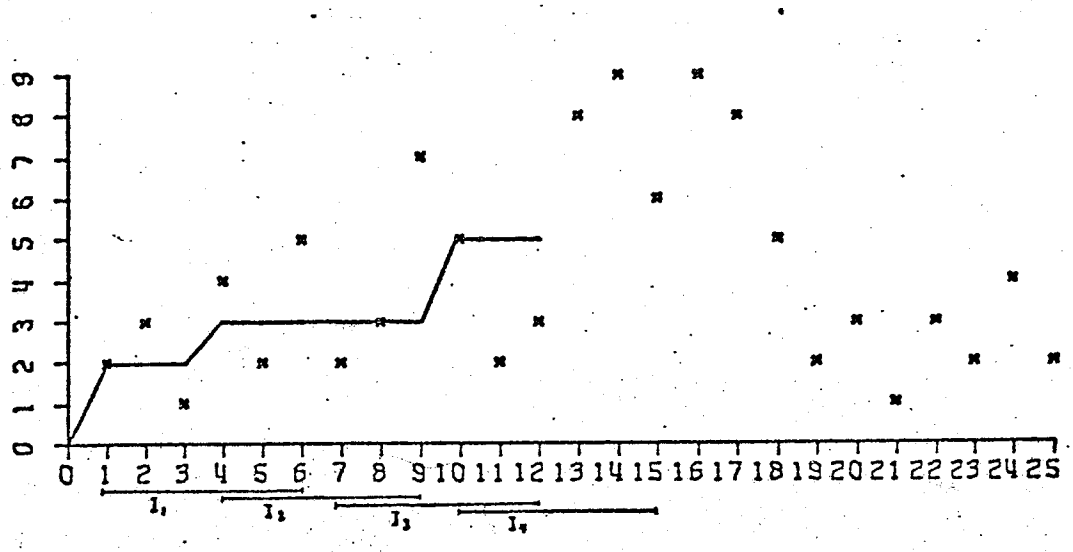
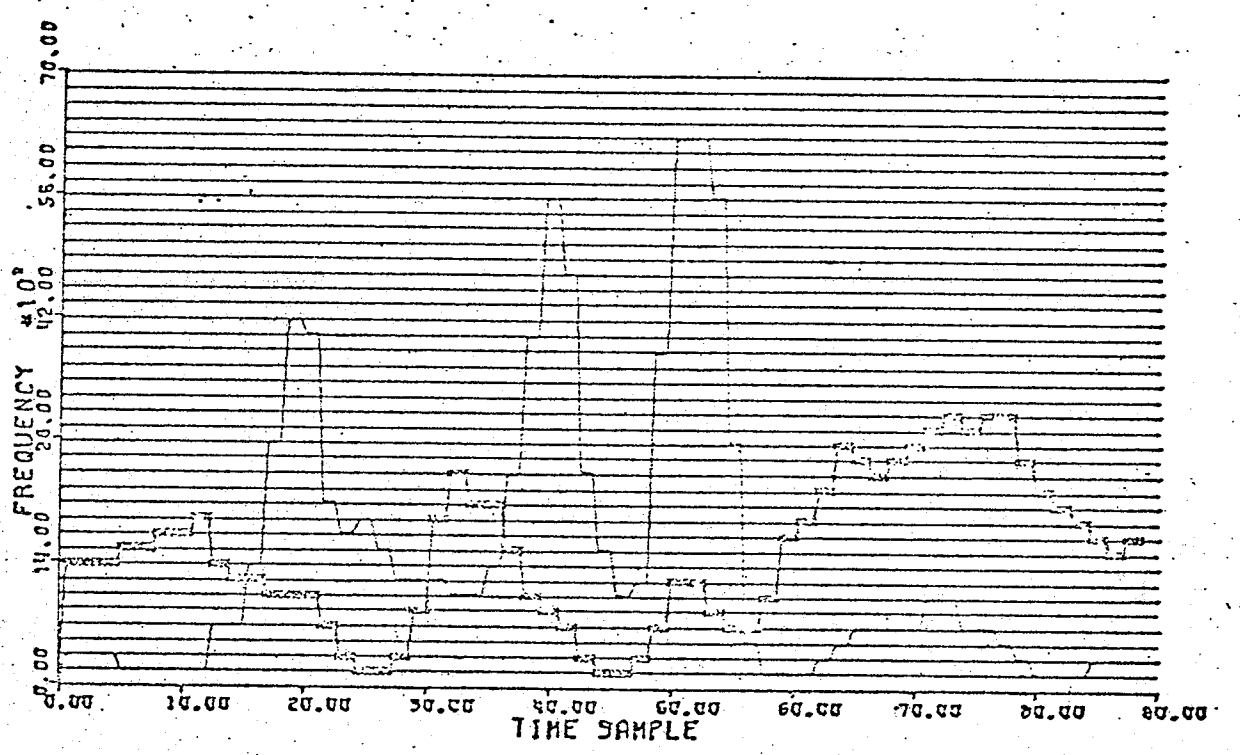


Fig. 4.1

TABLE 4.2

EXHIBITS BY A.Y.



EXHIBITS BY D.C.

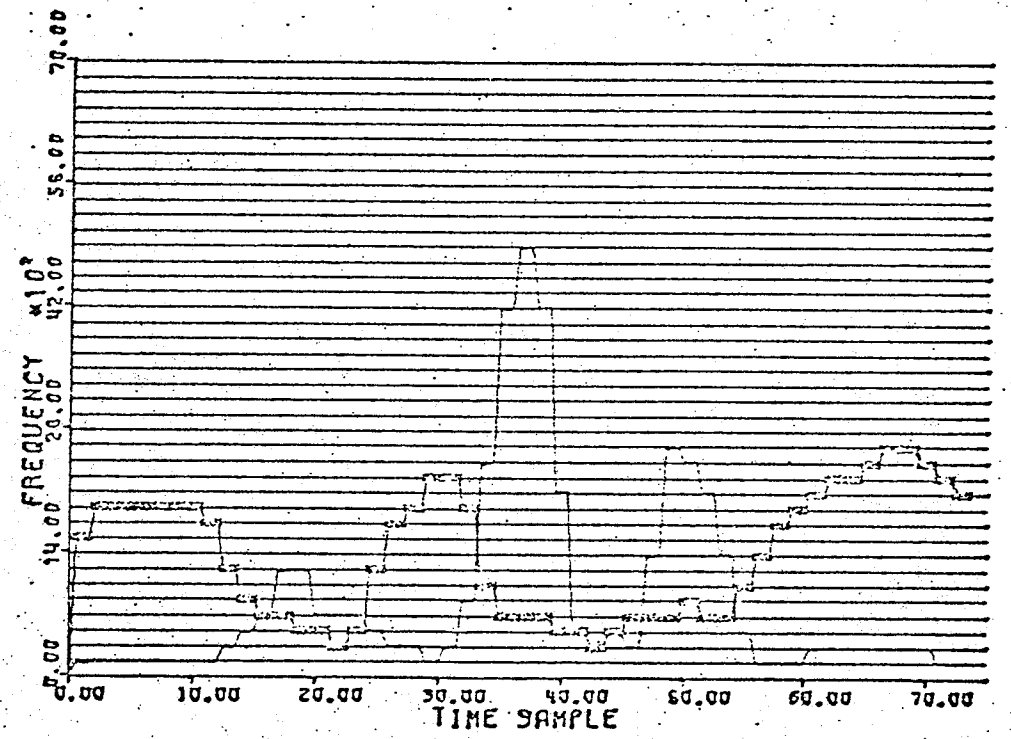
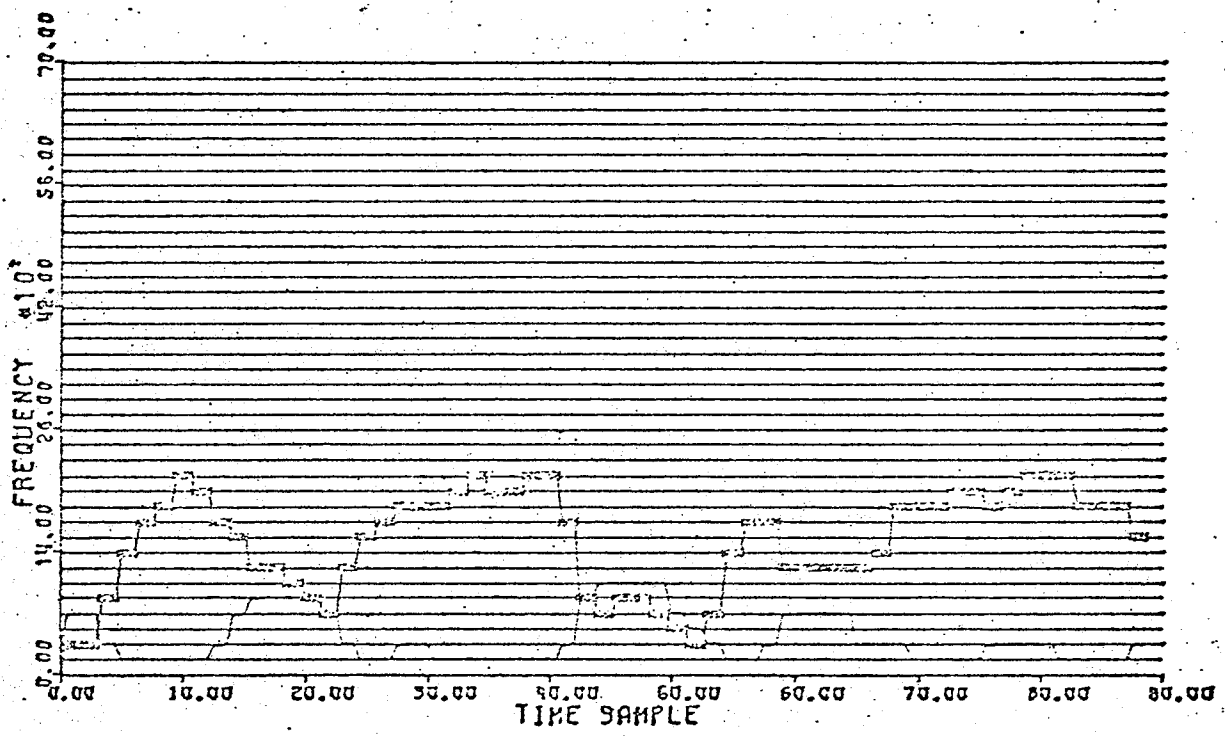


TABLE 4.2 (continued)

EXHIBITS BY I.R.



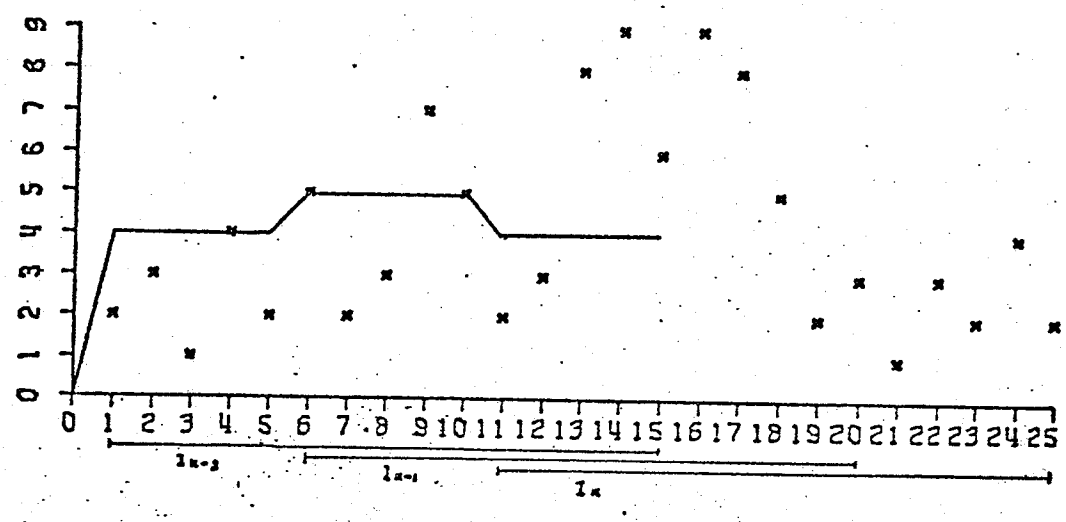


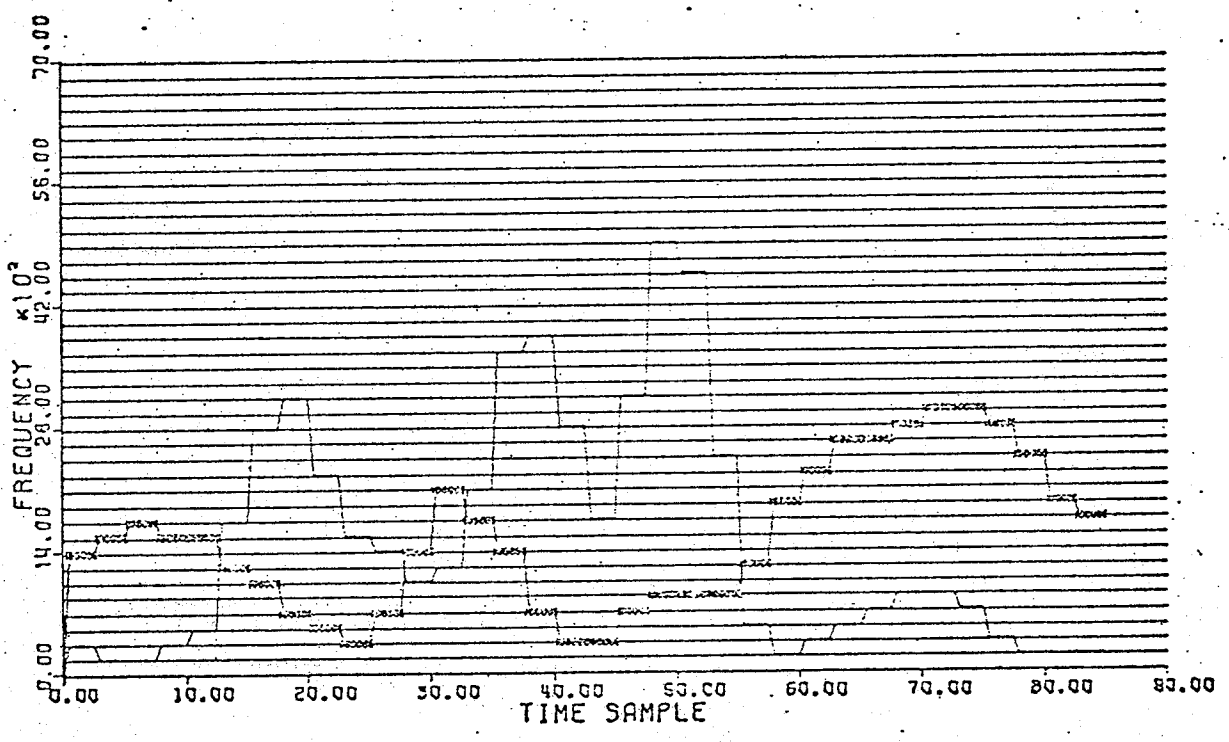
Fig. 4.2

were attributed to the fact that an increment of three points was much too small to eliminate the harmonics, keeping in mind that average phoneme is comprised of approximately 20 points.

To rectify this problem the interval I_k was increased to 15 points with an overlap of ten points with the previous interval I_{k-1} (see Fig. 4.2). The results were quite good for the word "exhibits" (Table 4.3), however, in "resonance" (Table 4.3) the frequency change of "nan" was eliminated by A.Y. and D.C. Furthermore, if one eliminates the time dependance, then "resonance" for A.Y. and D.C. and "exhibits" for the same two speakers possess very similar fundamental waves failing to enable us to distinguish them. Obviously in speech recognition, it is more desirable to have two recognition patterns for one word, than having two words with the same pattern. Although all our methods so far have been time-dependent, this is not a criterion that can be relied on since each speaker has his own way of pronouncing words and phonemes, both with respect to stress as well as time. Hence for one speaker the length of a phoneme may be as short as 400/7000 sec. whereas for another speaker it could be as long as 1000/7000 sec. Continuing this empirical method, the author decreased the interval of the last smoothing routine to nine points, with increments of three points, and overlap of six points with interval I_{j-1} and three points with interval I_{j-2} (see Fig. 4.3). Looking at Table 4.4 one can see that the results are very much similar to those of "Six Point Overlap". The final routine which produced the best results was similar to "Six Point Overlap", however, the interval was increased to ten points with five point overlap (see Fig. 4.4). As can be seen from Table 4.5, the results are quite good. It should be noted that the author judged the smoothing routines on the basis of

TABLE 4.3

EXHIBITS BY A.Y.



EXHIBITS BY D.C.

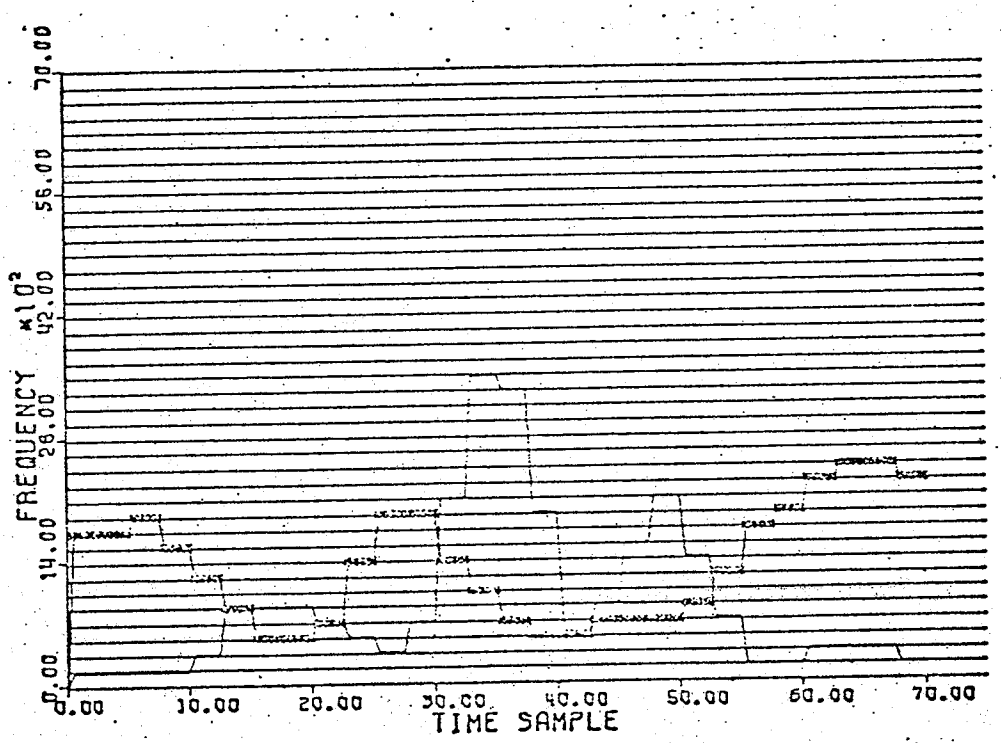
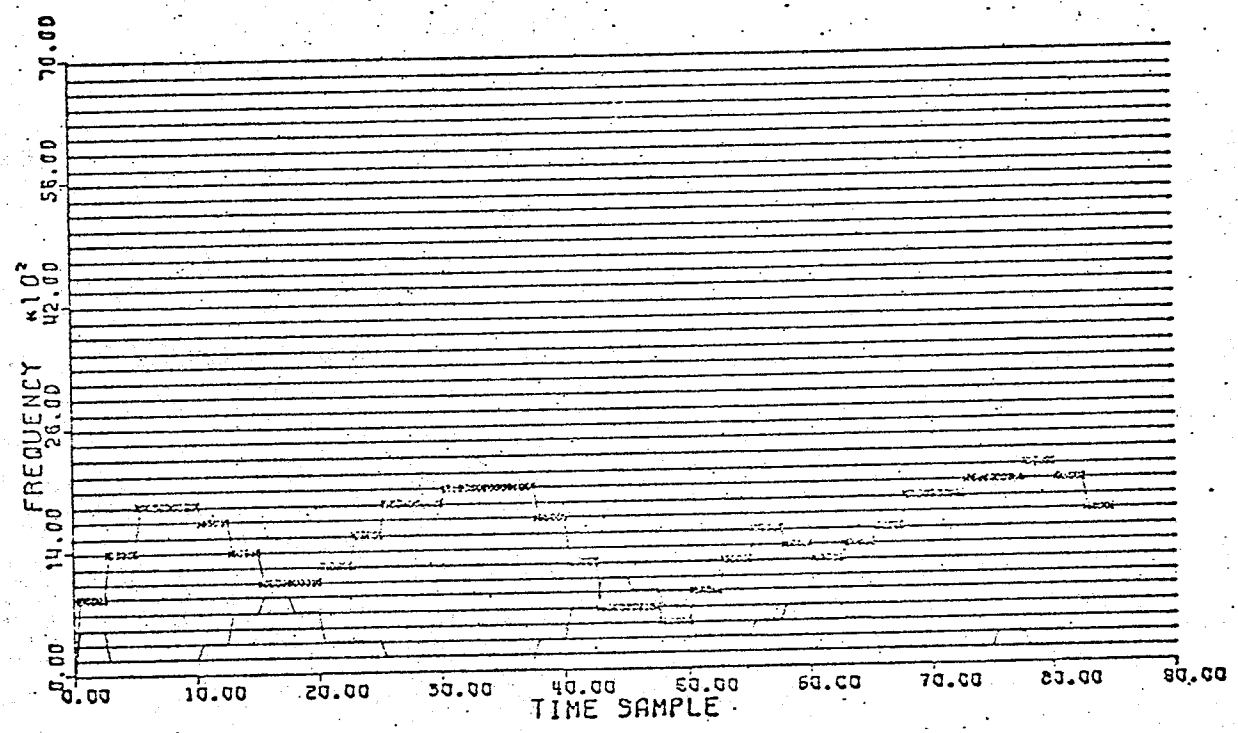
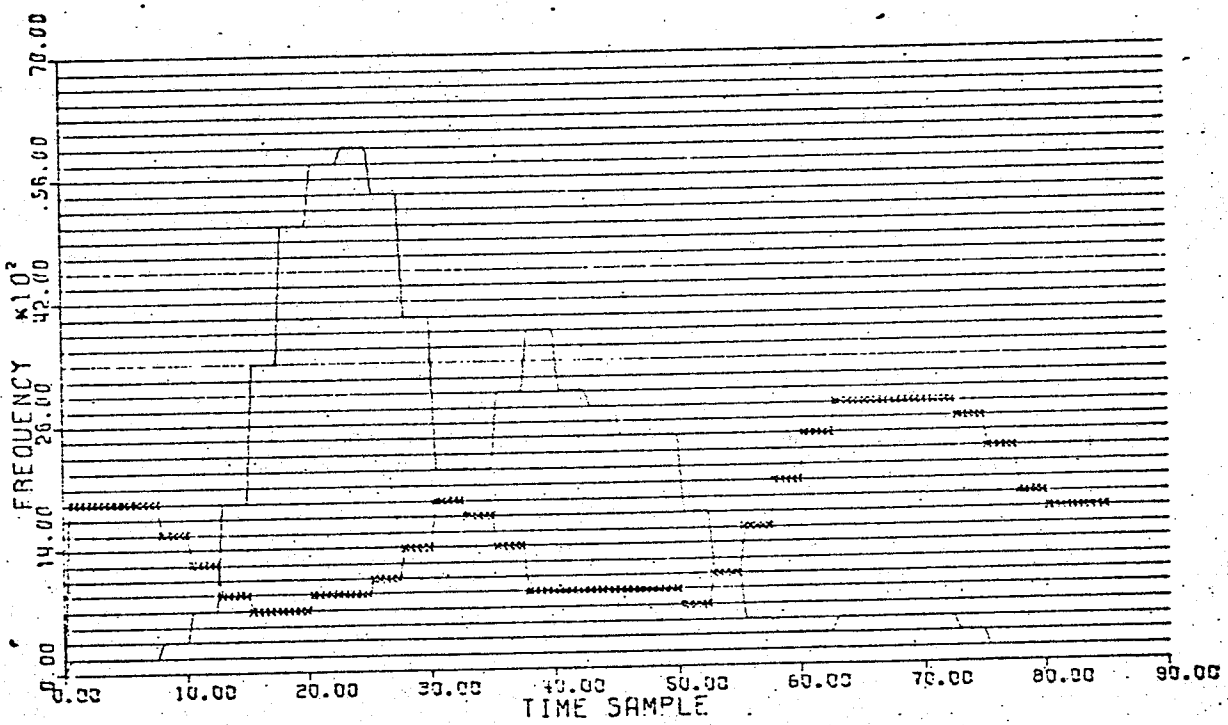


TABLE 4.3 (continued)

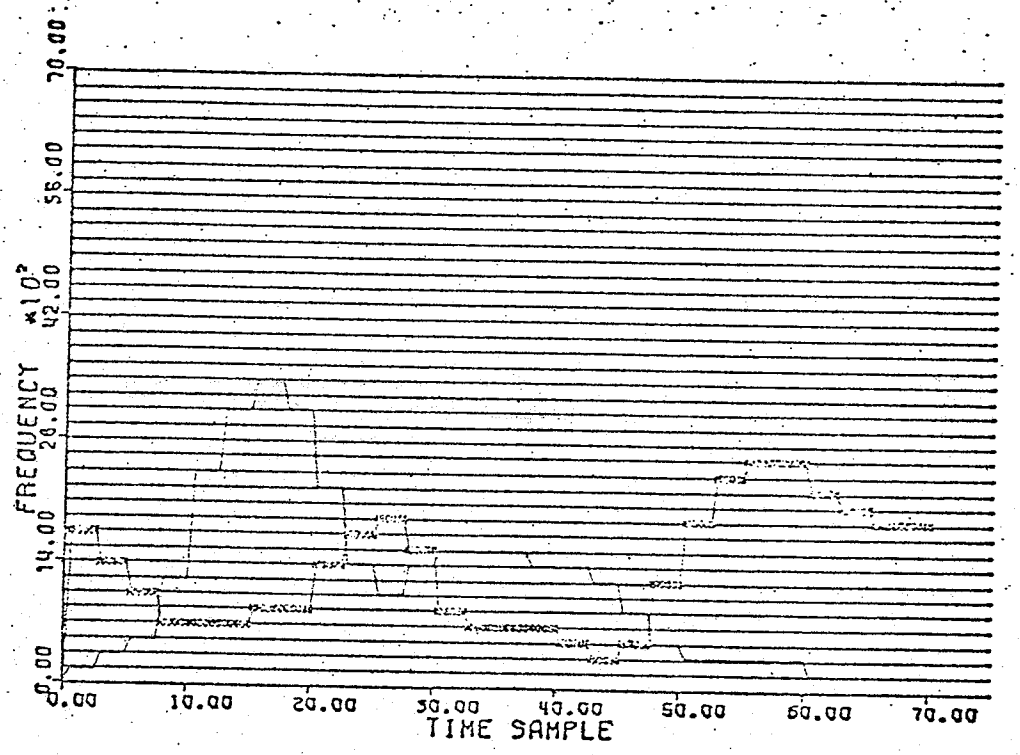
EXHIBITS BY I.R.



RESONANCE BY A.Y.



RESONANCE BY D.C.



RESONANCE BY I.R..

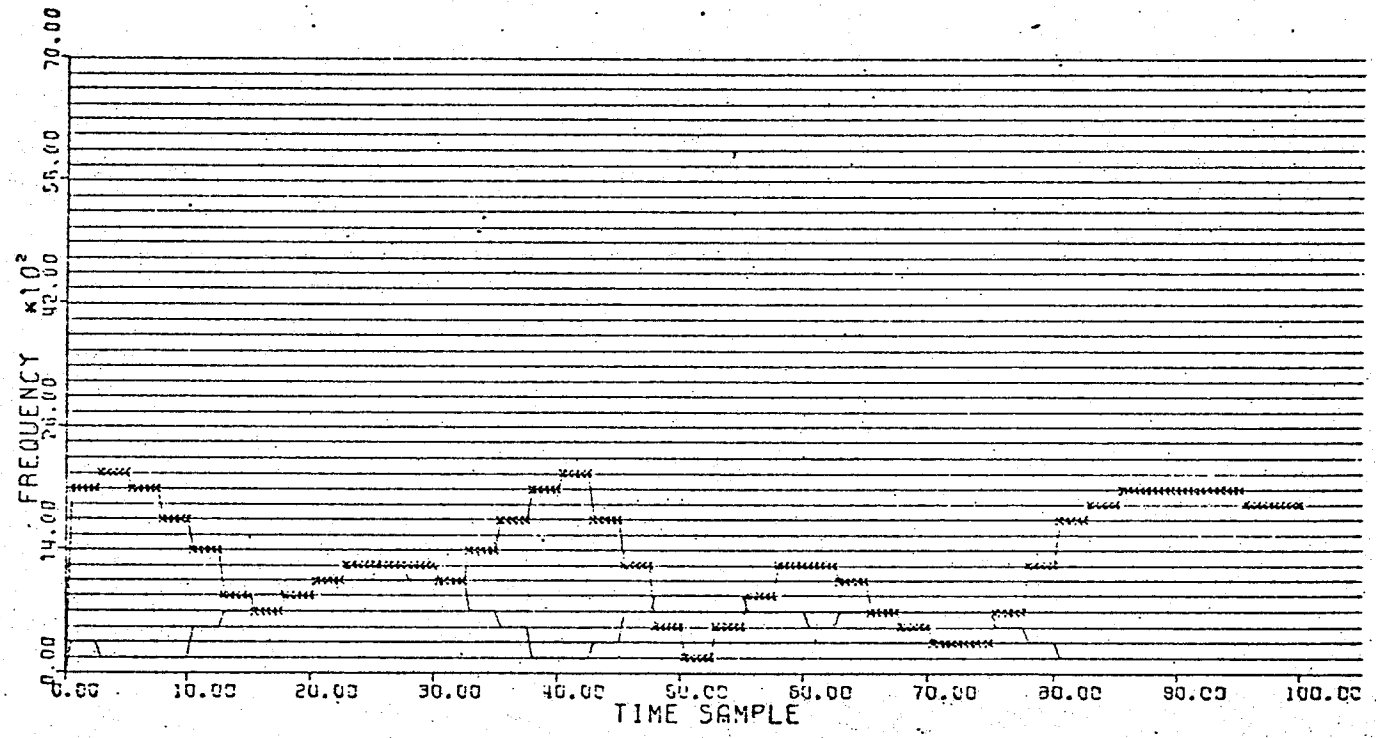
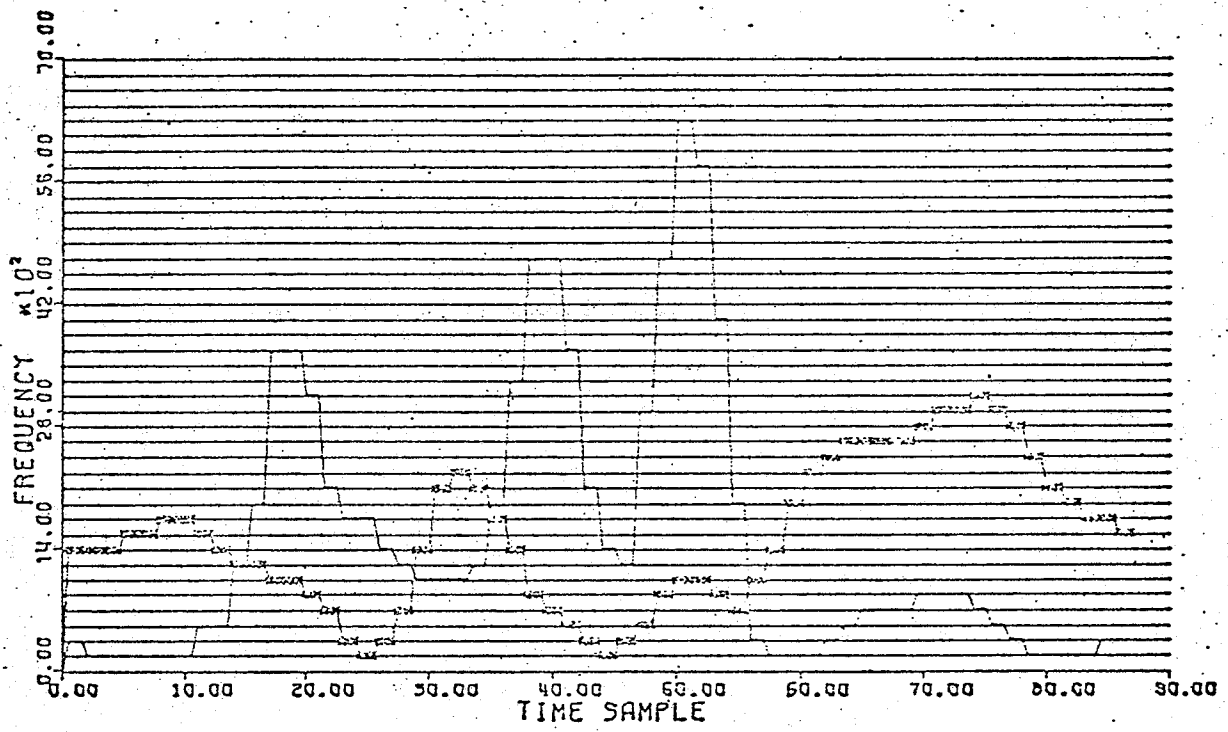


TABLE 4.4

EXHIBITS BY A.Y.



EXHIBITS BY D.C.

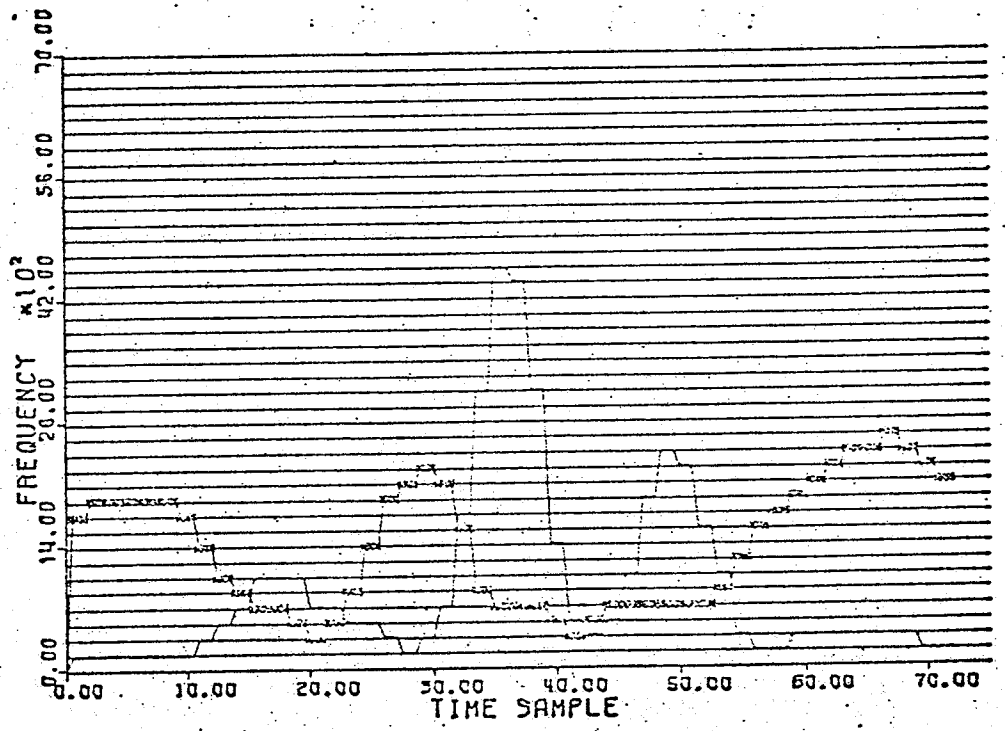


TABLE 4.4 (continued)

EXHIBITS BY I.R.

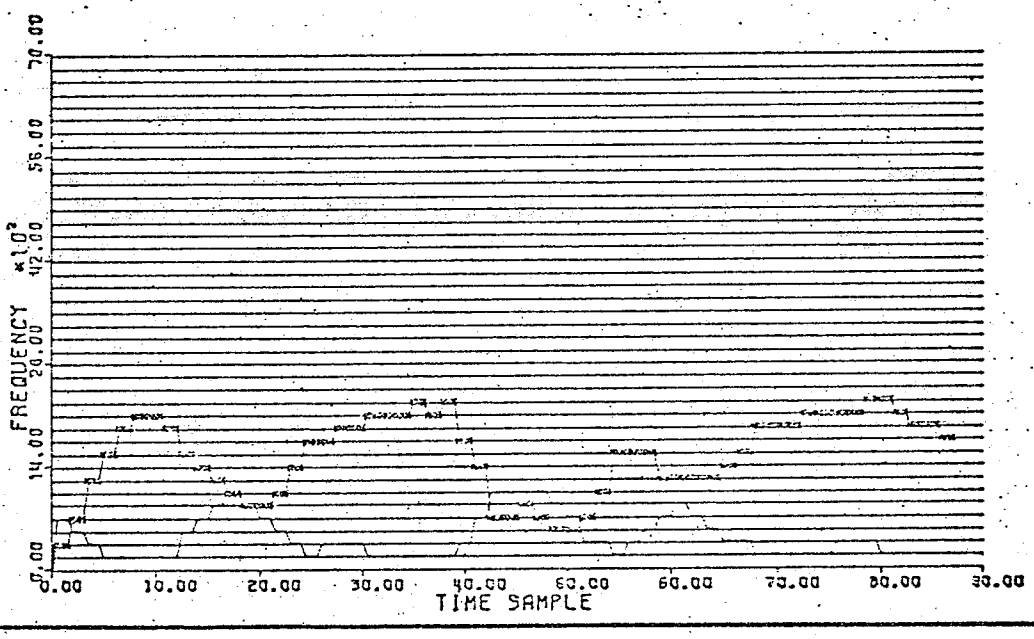


Fig. 4.3

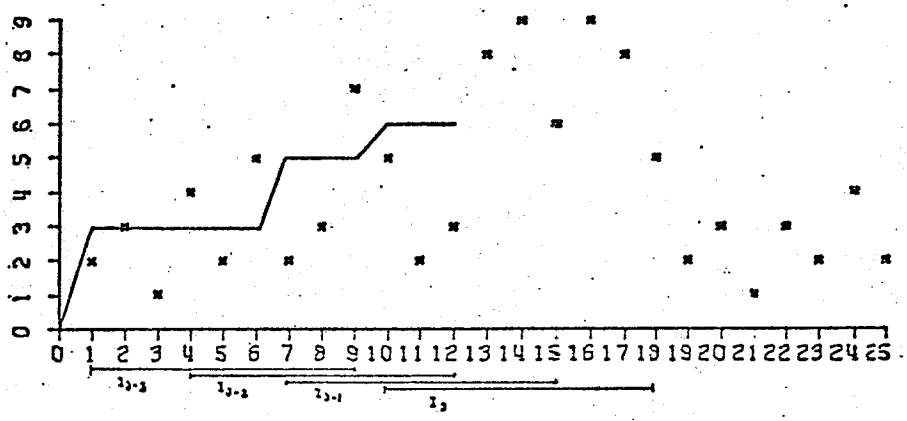


Fig. 4.4

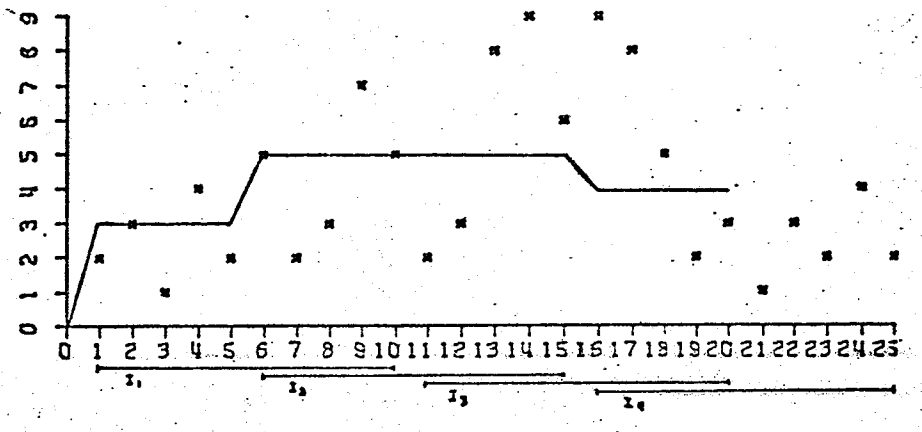
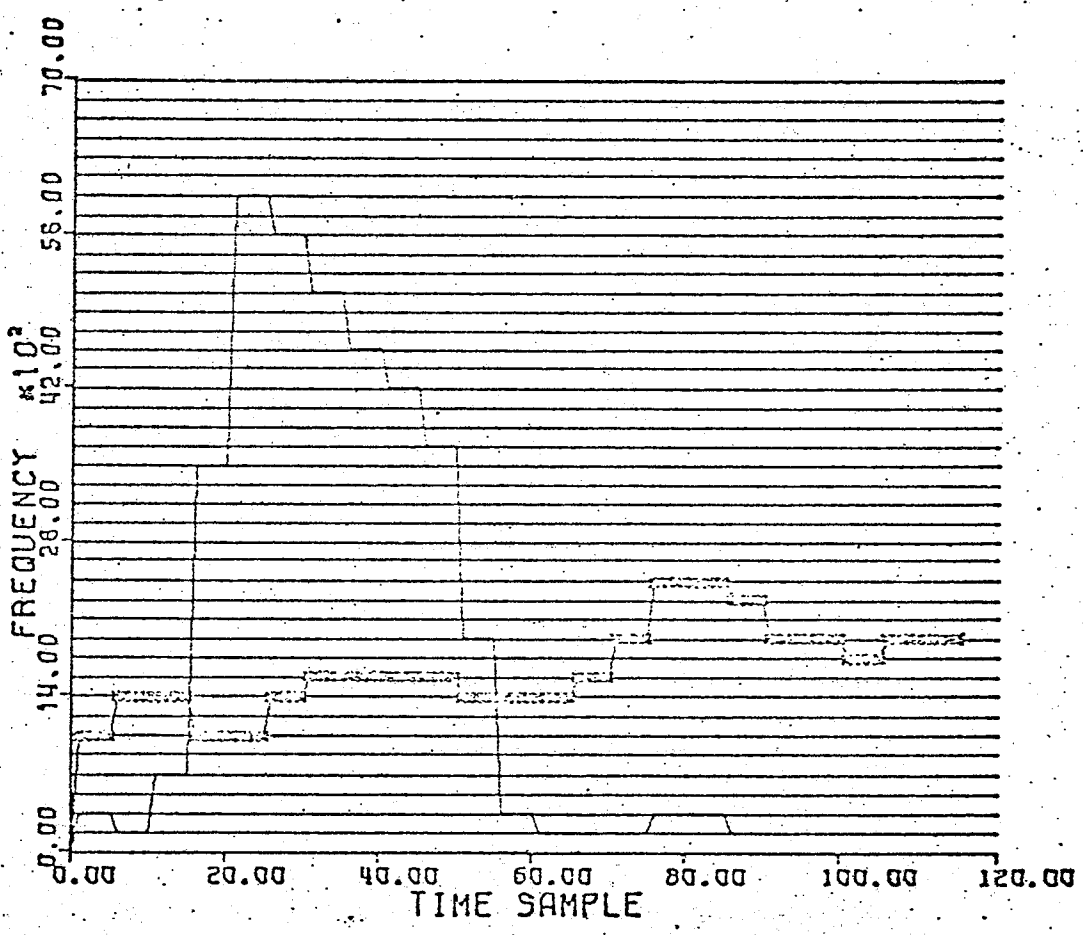
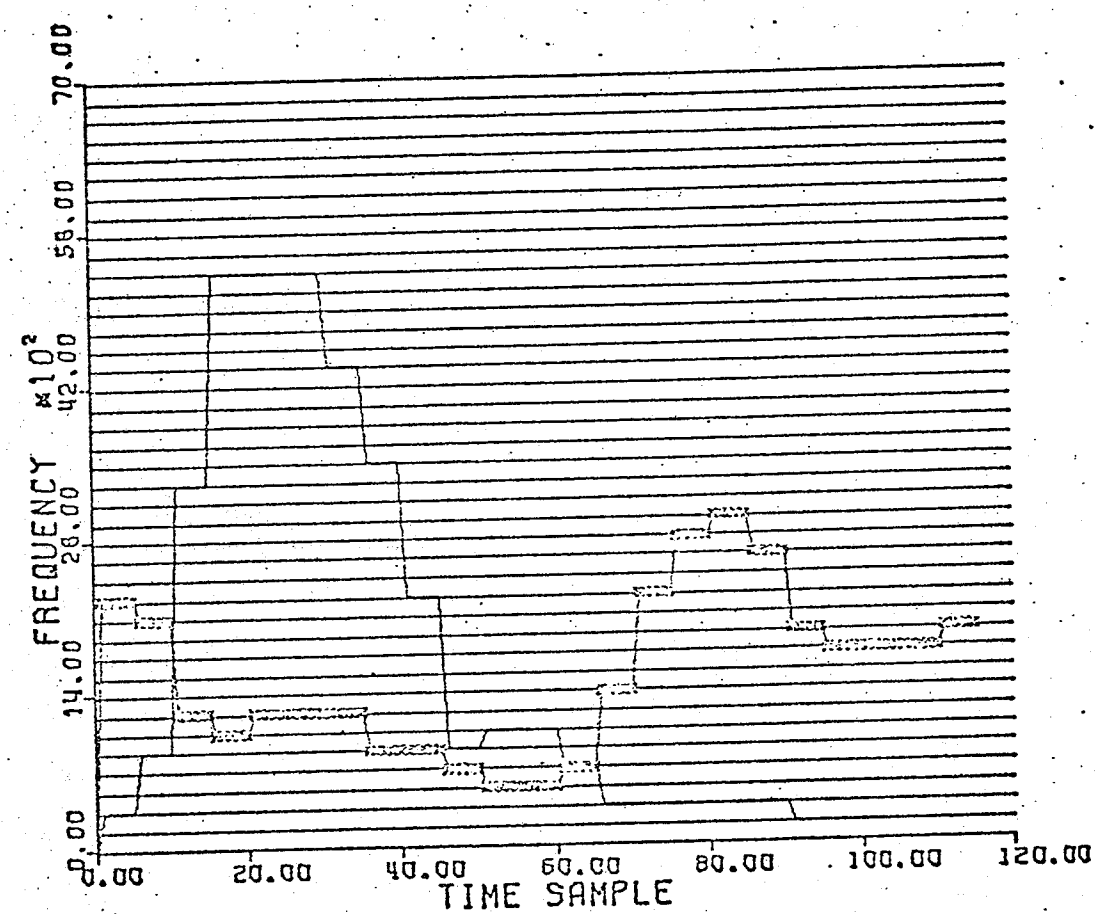


TABLE 4.5

THAT BY D.C.



PLANES BY A.Y.



PLANES BY D.C.

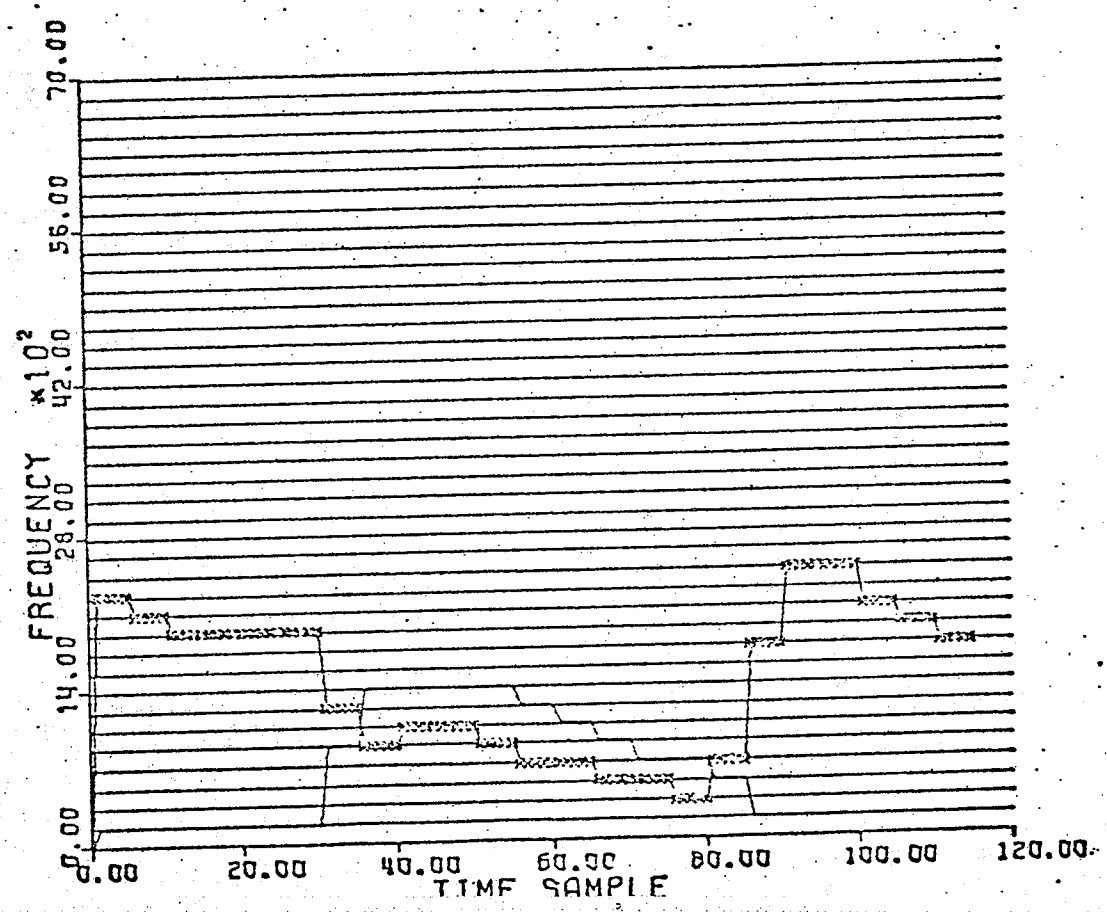
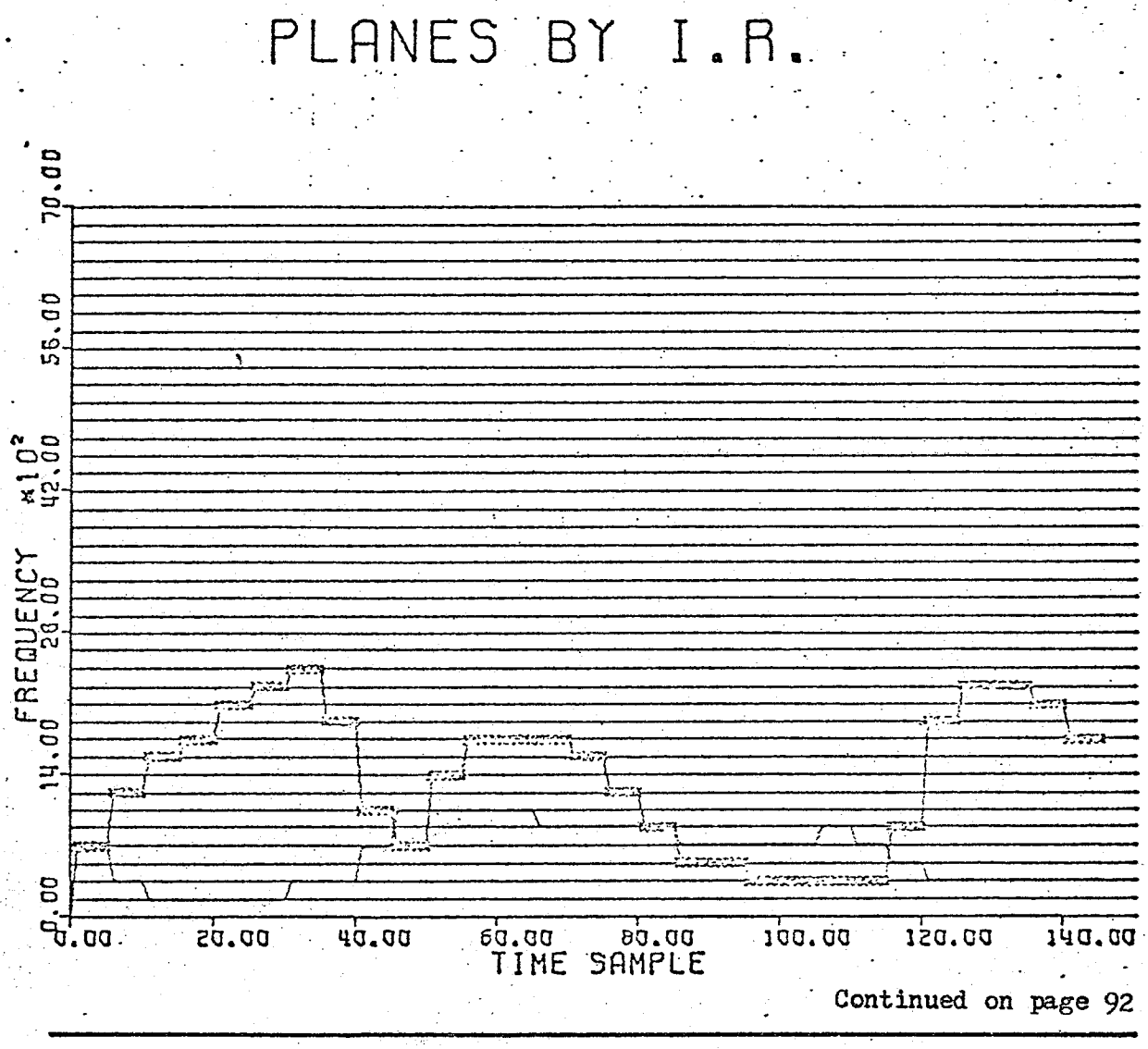


TABLE 4.5 (continued)



Continued on page 92

TABLE 4.6

RESULTS OF THE FINAL SMOOTHING ROUTINE

Speaker	Percentage Achieved
A.Y. & D.C.	83%
A.Y. & I.R.	83%
D.C. & I.R.	66%

uniqueness. A wave for a particular word is said to be unique for all three speakers, if the number of peak to peak upward and downward changes in each wave is the same. In addition to that, the wave for each word must be distinct. The last smoothing routine achieved 66% uniqueness for all three speakers. For results between individual speakers see Table 4.6. Furthermore, one should realize that 66% uniqueness does not imply that only 66% recognition can be achieved. For a full description of the recognition methods see Chapter V.

4.3 Segmentation of Input Signal Into Necessary and Redundant Data

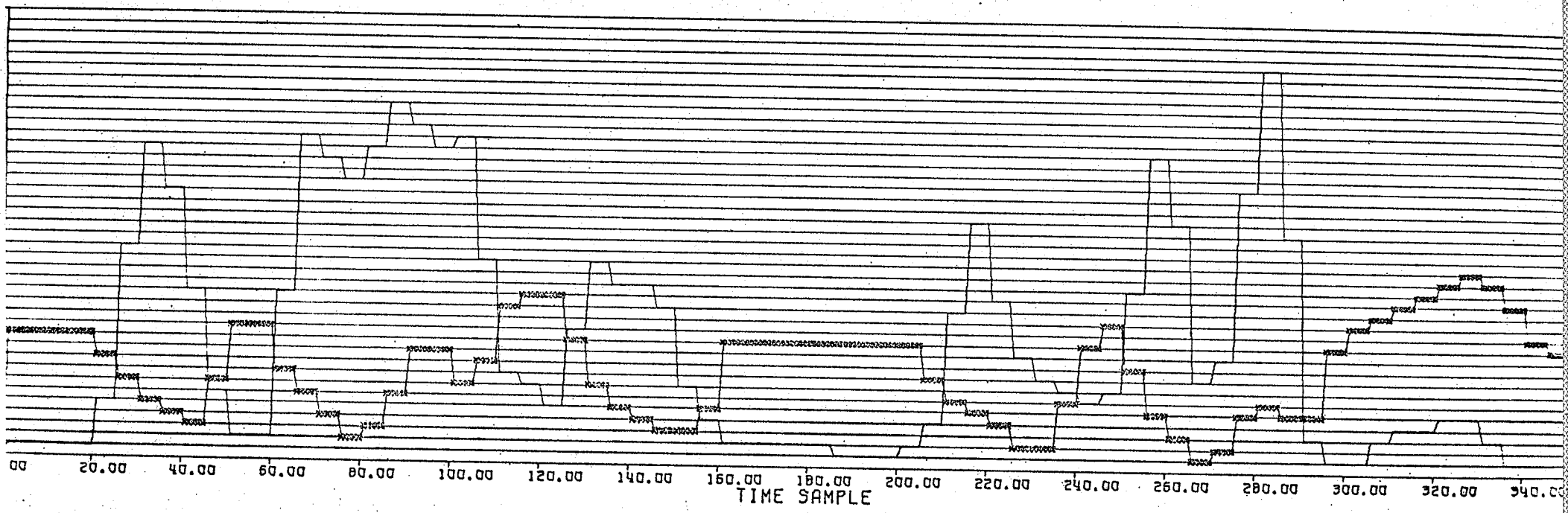
One of the more serious problems in speech recognition is detecting the word onset as well as its end. In the case of words which begin with a vowel, the author used the criterion that the onset of the word is at the point at which the amplitude crosses the third frequency scale line or 525 frequency mark. This assumption was based on observation of the graphs and proved to be correct in all cases in which the word begins with a vowel. However, in cases in which a word begins with a fricative as "F" in "flying" or plosive like "P" in "planes" the previous criteria cannot be used. These phonemes have low amplitude and high frequency characteristics. In these special cases, increase in frequency curve from the zero or noise level, which is 1925 cps in this case, would have to be detected and marked as the onset point for that specific word. At this time, it should be noted that the value of 1925 cps which is used as the noise or inbetween word value has been obtained with the aid of the smoothing technique used in previous section. Furthermore, this value is not constant. In many cases the curve deviated up or down by as much as 525 cps. In all such cases the author assumed the change to

be stray noise. The base for this assumption was the fact that all low frequency sounds are produced with the aid of the vocal cords which are vibrating. At the same time all sounds which are produced with vibrating vocal cords are high or relatively high in amplitude. Hence, since these deviations are low in amplitude and low in frequency, then they cannot be produced by the speaker and therefore must be noise. For upward deviations, the problem was not as easy to solve as the previous case. The biggest problem arose when an upward deviation occurred at the beginning of a word. In such cases, the change was considered to be the onset of that word, and it therefore became a part of it. One remedy would be to ground, as much as possible, all the components or recognition equipment to prevent the pickup of stray noise. However, another way, depending on how sophisticated the equipment must be, is to ignore all fricative or plosive beginnings, setting the onset point at the occurrence of the first vowel. The disadvantage of this method is the fact that words like "flying" and "lying" could not be used simultaneously in the same system due to the fact that they possess the same recognition pattern.

The end point of a word was simply that point after which followed 25* 40/7000 sec. of silence. This of course places a restriction on each speaker to leave 7^{-1} seconds between each word. The length of this interval was chosen on the basis that the silent spot between "A" and "T" in "that" was 20* 40/7000 sec. for D.C. Setting the interval to a smaller value would cause the recognition mechanism to interpret the silent spot as the end of that word causing failure in the recognition.

The accuracy of our assumptions and criteria are clearly demonstrated in Figure 4.5 in which one can easily detect the word onset, end, and inbetween word noise. The routine which picks out the noise is called "PICK" and is given in Appendix A.

FIG.4.5 RESULTS AFTER THE ELIMINATION OF NOISE.



CHAPTER V

5.1 Introduction

In Chapter V a close look will be taken at the three most important parts of any recognition system. The three parts are: the speaker, the list of commands and the recognition algorithm.

In Section 2 the three main speakers in this thesis will be described and their vocal peculiarities discussed.

In Section 3 the problems which can arise from a poor choice of words will be listed. Phonetic description as well as the reasons behind our choice of the six words used in this thesis will be given. The actual recognition algorithm and its use will be described in Section 4.

5.2 The Speaker

The major question when designing speech recognition equipment is, what class of people will be the main users of the instrument. The proper way to approach this problem would be to initially limit its use to a carefully controlled group of speakers. By a carefully controlled group we mean a group of speakers who all possess similar vocal characteristics, perferably all average speakers. After thorough and precise examination of different irregularities, the system could be extended to take into account wider divergency of pitch and accent, with the final system allowing for some of the more common speech defects such as stuttering and lisping. The term average in this case

applies to persons who possess a medium pitch and a medium intensity voice. It should be noted that the speaker in our investigation who possesses these characteristics is D.C. The other two speakers exemplified the two extremes, with A.Y. possessing deep loud voice and I.R. soft high-pitch voice. In addition, I.R. possesses a slight British accent which, of course, added to the problems of recognition. At present, it is hard to determine whether the differences in the graphs of I.R. are caused by the high pitch voice or the accent. It is the author's belief that the latter is the prime reason for the differences.

Another great problem brought about by the speaker is his lack of faith in the machine, as well as his nervousness. Most speakers overarticulate their words, trying to make sure that the machine hears each individual phoneme comprising the given word. Similarly, when under nervous strain, the voice of the subject tends to vibrate or change pitch. The problems which are caused by these irregularities lie in the fact that the stresses or the pitch changes may not occur in the same place each time the word is pronounced. The reason for this is that they are not produced habitually and the speaker may forget which syllable he or she stressed previously.

At this point, it should be noted that the question of the speaker depends greatly on the type of system that we are designing. If the sole purpose of the machine was to recognize "Open" and "Close", then the speaker does not pose any great problems. In such a system, with only two words to be recognized, the recognition patterns could be made so general that only one or two main characteristics of these words would trigger off the necessary equipment which would perform the necessary command. However, if the system was to recognize a long list of commands,

then the recognition patterns would have to be made precise to ensure uniqueness. In this case the speaker plays a crucial role, since the smallest irregularity might cause the failure of recognition of that specific word, unless proper steps have been taken to compensate for such difficulties. Since the list of possible irregularities is so great, it is very unlikely that such a system can be developed in the near future. At the present moment both the speaker and the list of words will have to be chosen in such a way that no ambiguities arise.

5.3 List of Words

In any speech recognition system, the list of words to be recognized plays as important a role as the speaker. The two main problems which come about from the choice of words are sound ambiguity and the phrase ambiguity. By sound ambiguity we mean words like "to", "two", and "too". Since words are recognized on the basis of their sound, the above words could therefore not be included on the same list. Similarly phrase ambiguity concerns our inability to distinguish between such words as "icecream" and "I scream" since these phrases sound the same.

In addition to the above cases we also have such problems as words which possess the same vowel and consonant representation. An example of this are words "O.K." and "obey". These words would produce very similar output waves which would fail the recognition algorithm. The list of words which was used in the research work for this thesis was not chosen in such a way as to avoid any ambiguities, but rather to exemplify as many as possible of the difficult phonemes.

The word "that" (|ð| |a| |t|), was chosen for its soft beginning |ð|, loud and prolonged |a|, silent spot between the |a| and the |t|,

and of course, the soft ending |t|.

The word "information" (|i| |n| |f| |ə| |(r)| |m| |e| |ʃ| |ə| |n|), possesses vowel beginning, change from vowel to fricative and back to vowel in |nfə|, and, fricative to vowel ending in |ʃən|. In addition, it is comprised of many phonemes, demonstrating clearly different changes which may occur in vowel to consonant relations.

The word "exhibits" (|ɛ| |g| |z| |i| |b| |i| |t| |s|), was chosen for the phonemes |g| |z|, changes in |i| |b| |i|, and the ending |t| |s|.

The word "resonance" (|r| |ɛ| |z| |ə| |n| |ə| |n| |s|). in addition to its representation of many sounds possesses hard vowel-like beginning (|r|), fast change in sounds in |nən|, imbeded fricative |z|, and turbulent ending.

The last two words "flying" (|f| |l| |əi|*|i| |ŋ|) and "planes" (|p| |l| |e| |n| |z|) demonstrate a fricative soft beginning in |f|, and plosive beginning in |p|. Furthermore, the sound |ŋ| was of much interest, as well as the sound |en|.

It should be noted that in addition to the above reasons, the words "that", "information", "exhibits" and "resonance" when joined, form a sentence. The advantage of pronouncing a sentence instead of individual words is that the speaker does not place so much emphasis on stressing each syllable in each word, but rather concentrates on the overall fluency of the sentence. The problems which arose in this list, as will be seen later in this chapter, were connected mainly with the word "exhibits".

---

* |əi| as in "bind"

5.4 Recognition Algorithm

With the onset and ending point detection methods discussed in Chapter IV the only problem remaining is to determine a method for the recognition of the words. It should be noted that the author's objective in this thesis is not to present a 100% fool-proof recognition algorithm, but rather to discuss ways in which such an algorithm could be obtained. First of all if such a system were to be designed, the speakers would not be chosen with such extreme voice characteristics as I.R. and A.Y. However, for the purpose of our study the extremes are of great importance, and it is interesting that we have obtained recognition of five out of the six words for such diverse speakers.

As it was stated before (Chapter IV), what the human ear detects is not set frequency levels corresponding to certain phonemes, but rather frequency level changes. The very fact that the average frequency of "n" in "resonance" is 175 cps for I.R. and 700 cps for A.Y. illustrates this point. Hence our main concentration will be on frequency level changes. The changes will be divided into three main groups, i.e. small (S), medium (M) and large (L). However, to account for differences in speakers, two subgroups will be introduced. The two subgroups are medium-small (MS) and medium-large (ML). These subgroups will overlap the main groups and any change falling into one of these subgroups will be in the final stages represented by either small or medium change as in the case of (SM). It should be noted that a subgroup such as medium-large can be written as either ML or LM, and we use the former to indicate that the change is more likely to belong to the medium group while the latter indicates that the change is more likely to be of the larger variety. For numerical values of the level changes for each of these groups and subgroups see Table 5.1.

TABLE 5.1

Group Name	Peak to Peak Change
S	1 - 3
SM	4 - 6
M	7 - 9
ML	10 - 12
L	13 or over

The integer values in the Table represent number of frequency bands (175 cps) between the upper and lower consecutive peaks in the wave.

Before going into the recognition algorithm we should state one more very important observation that was made concerning noise. The observation is that any high frequency changes such as the one in "that" by I.R. at the 1925 cps level (see Table 4.6) are produced by noise, and hence can be ignored. The basis for this assumption is the fact that to produce such a frequency dip, two fricatives or plosives such as |s| |f| or |p| |k| would have to be used side by side which is a very rare case in the English language.

The method by which the values for the recognition algorithm were obtained is as follows: The starting value was taken from the 1925 cps point to the next lowest point. By ignoring the front plosives and fricatives we know that the word will always start with a vowel and hence the graph will be decreasing from our zero (1925 cps) level. Thus, the next value will be an increase followed again by a decrease. Furthermore, the final change back to the zero level will be ignored and replaced with "END".

The reason for the replacement of the final value is the fact that in contrast to the front plosives and fricatives, the end plosives and fricatives are a part of the recognition algorithm. These phonemes characterize the upper bound of each speaker's voice frequency band, which is different for each speaker. Hence, the last value, being calculated with respect to a fixed zero level (1925 cps), would be a representation of speaker's band and different in each case. It should be noted also that the lower bound is almost the same (175 cps) for each speaker. The numerical values for each word and speaker are given to Table 5.2 with the ending part given.

Another very important thing that should be noticed is that each pattern is at first compared to all the recognition patterns. However, a match may not be found because of three main reasons:

A. The pattern bears no resemblance to any of our list of words. In this case, we are unable to continue and consequently the procedure terminates.

B. The pattern is, in fact, similar to one of our list except that due to the pronunciation of one of the phonemes the pattern is too short. At present there is no way to match this correctly since the random introduction of features would obviously lead to many errors.

C. The pattern bears some resemblance to one of our list patterns but is too long. In this case, we can reasonably attempt a matching by the concatenation of a wave.

The first small change would be eliminated by concatenating the two bounding values. This means that the word must be first changed back to its numerical values to produce the proper classification. Such a problem as will be seen does come about with the word "resonance" for

TABLE 5.2

WORD	SPEAKER	CHANGES								
That	A.Y.	6	3	1	6	3				
	D.C.	5	3	1	6	3				
	I.R.	8	5	2	7	2				
Information	A.Y.	8	9	10	8	3	8	12	8	
	D.C.	7	8	10	6	2	9	13	9	
	I.R.	8	10	10	6	1	9	14	9	
Exhibits	A.Y.	8	11	12	5	1	13	7		
	D.C.	8	9	10	12	3				
	I.R.	7	8	10	6	2	6	2		
Resonance	A.Y.	8	10	9	1	1	13	6		
	D.C.	7	8	9	1	2	13	4		
	I.R.	10	4	1	7	12	6	5	10	2
Flying	A.Y.	8	4	4	9					
	D.C.	8	6	6	8					
	I.R.	6	5	8	10					
Planes	A.Y.	5	1	4	14	6				
	D.C.	6	1	4	12	4				
	I.R.	7	6	8	11	4				

I.R. The steps taken in this concatenation will be discussed later in this chapter. At the same time it should be noticed that if the elimination of the first small change does not produce any definite results, the next small change is eliminated restoring the first one. If this still does not improve the recognition then the recognition simply fails.

For symbolic values for the words see Table 5.3. The author's symbolic representations are presented in Table 5.4. As can be seen "resonance" by I.R. is too long for any pattern in Table 5.4. Hence, the original values for "resonance" are considered, i.e. 10, 4, 1, 7, 12, 6, 5, 10, 2. The first small change occurs in the 4, 1, 7 sequence. By adding four and seven together and subtracting one from them to compensate for the dip we obtain the value ten or ML. Hence, the new representation of the word is ML, ML, LM, MS, SM, ML, END. Looking at the set symbolic representation table (Table 5.4) we see that the first symbol (ML) matches all words. The next one (ML) eliminates the words "that" and "planes". The next one (LM) matches all three remaining words, however, the next one (MS) eliminates the word "flying". The next two symbols, i.e. (SM) and (ML) again match both words with the final "END" matching the pattern for "resonance". Using this method 100% recognition for the five words can be obtained.

It is unfortunate that the word "exhibits" did not produce a definite pattern, however, it is the author's belief that the cause of the failure was the fast change in "ibi" where timing and stress are very important.

TABLE 5.3

WORD	SPEAKER	SYMBOLIC REPRESENTATION
That	A.Y.	MS, S, S, MS, END
	D.C.	SM, S, S, MS, END
	I.R.	M, SM, S, M, END
Information	A.Y.	M, M, ML, M, S, M, LM, END
	D.C.	M, M, ML, MS, S, M, L, END
	I.R.	M, ML, ML, MS, S, M, L, END
Resonance	A.Y.	M, ML, M, S, S, L, END
	D.C.	M, M, M, S, S, L, END
	I.R.	ML, SM, S, M, LM, MS, MS, ML, END
Flying	A.Y.	M, SM, SM, END
	D.C.	M, MS, MS, END
	I.R.	MS, SM, M, END
Planes	A.Y.	SM, S, SM, L, END
	D.C.	MS, S, SM, LM, END
	I.R.	M, MS, M, ML, END
Exhibits*	A.Y.	M, ML, ML, SM, S, L, END
	D.C.	M, M, ML, ML, END
	I.R.	M, M, ML, SM, S, SM, END

* Note - No discernible pattern was found for the word "exhibits" and hence the word will not be included in the recognition table.

TABLE 5.4

WORD	SYMBOLIC REPRESENTATION
That	M, S, S, M, END
Information	M, M, L, M, S, M, L, END
Resonance	M, M, M, S, S, L, END
Flying	M, M, M, END
Planes	M, S, M, L, END

CHAPTER VI

CONCLUSION

Although the procedures used in our research work enabled us to recognize five out of six words, the results should not be taken as perfect considering the number of words in the English language and the number of possible speakers. When extending our system to general cases the following points should be taken into consideration.

First of all in the collection of instructions there are two types of recognition systems. One which recognizes connected speech and the other which recognizes set commands. The system which the author has presented was more command oriented than connected speech. In such a system stresses and nervousness of the speaker alter the results considerably. To obtain the characteristics of words as they appear in connected speech, the subjects would be asked to repeat a long text into the tape recorder from which the embedded words would be edited and studied. This procedure should be repeated several times and the results obtained compared, to assure that the characteristics are frequency changes and not noise.

The choice of subjects should be made in such a way that all the speakers are average or possess the same speech alternations such as high pitch, low pitch, or a foreign or regional accent. Although each speaker has his own small speech peculiarities like the harmonics in our waves, the developer should make sure that they are of limited importance.

With regard to the recognition algorithm, the ideal system would be one which would recognize words on the basis of levels of changes without storing them in memory for alternations such as the concatenation in our system. It is our objective to achieve a system which performs commands instantaneously without "learning" such as is the case with "Lisper" [20]. Such a system however is hardly feasible and the author acknowledges that beside frequency changes many more aspects such as amplitude, time and alternations of the results should be considered. This new system would work on an elimination basis. If at the end of the most important recognition process (frequency changes) two or more words ended in a tie, or the procedure failed, the second recognition method, i.e. amplitude levels would be applied. If however, this method still fails to obtain definite results, the characteristics would be put through the remaining processes until either recognition was obtained or the failure was signified.

As can be seen, much work still remains to be done in this field. It is hoped that man will soon be able to communicate with machines, computers and appliances in his most natural communication - process speech.

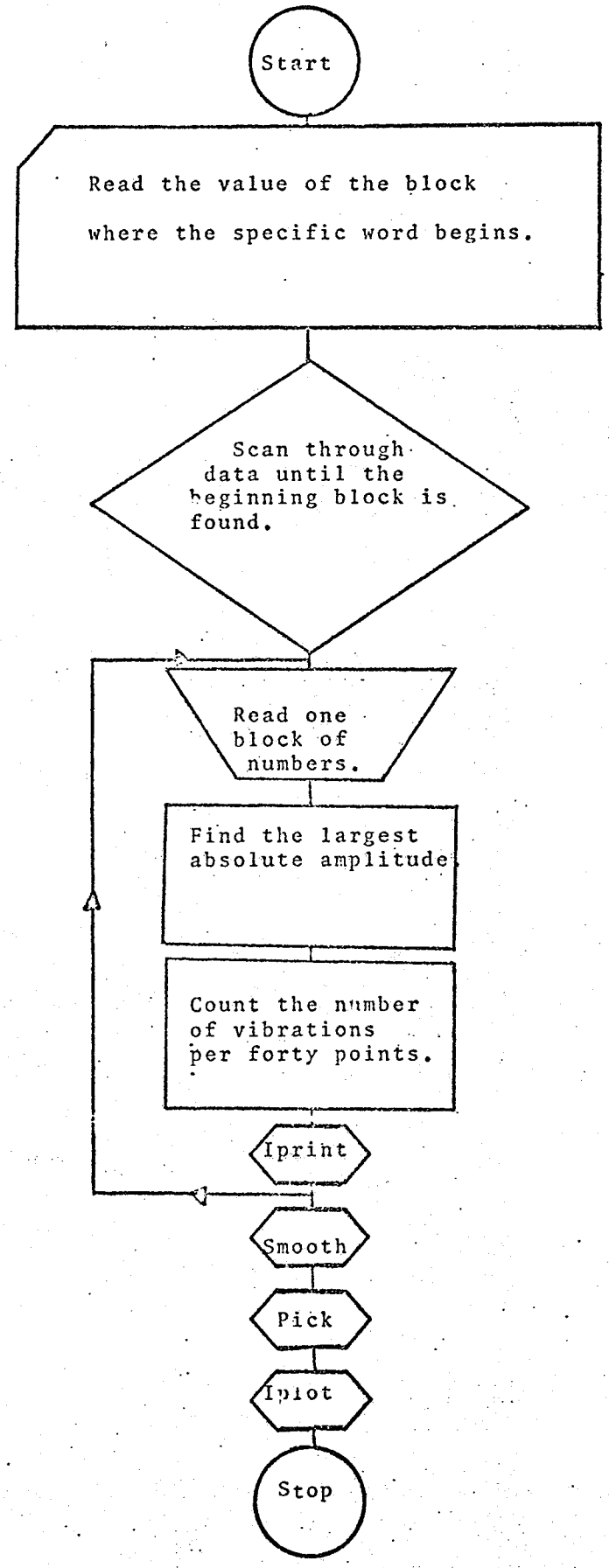
APPENDIX A

```
FORTRAN IV G LEVEL  18                MAIN                DATE = 70246          18/48/10

                C      FILTER BANK
                C
                C
                C      FORTY FILTERS IN STEPS OF 175 CYCLES PER SECOND.
                C      BLOCK IS THE STARTING BLOCK OF THE WORD.
                C      NO IS THE NO OF BLOCKS OF NUMBERS FOR THE SPECIFIC WORD.
                C      A(1200) IS THE ARRAY CONTAINING NO'S FOR THE SPECTRA OF THE WORD.
                C      TITLE IS THE SPECIFIC WORD BEING ANALYZED.
                C
                C
0001                   INTEGER FREQ,BLOCK,TIMES,TITLE(15),NOTHIN/' '/,VALU
0002                   INTEGER*2 C,B,A(1200)
0003                   INTEGER IBUF(1000)
                C      IN IS THE NUMBER OF POINTS PER TIME SAMPLE.
0004                   IN=40
                C      INC IS THE FREQUENCY BAND FOR EACH FILTER.
0005                   INC=7000/IN
                C      FREQ IS THE NUMBER OF VIBRATIONS PER TIME SAMPLE.
0006                   FREQ=0
0007                   CALL PLOTS(IBUF,1000)
0008                   CALL PLOT(0.0,2.0,-3)
0009                 1 READ(5,2,END=2000) BLOCK,NO,TITLE
0010                 2 FORMAT(2I5,15A4)
                C      II IS THE NUMBER OF THE TIME SAMPLE.
0011                   II=0
0012                 4 READ(12,3) C,B,A
0013                   IF(B.NE.BLOCK-1) GO TO 4
0014                   DO 1000 TIMES=1,NO
0015                   READ(12,3) C,B,A
0016                 3  FORMAT(2A2,200A2,200A2,200A2,200A2,200A2,200A2)
                C      I CONTROLS THE LOOP OF THE SPECTRAL WINDOW.
0017                   I=2
0018                   N=IN
                C      IAMP CONTAINS THE MAXIMUM AMPLITUDE FOR THE SPECIFIC TIME INTERVAL.
0019                 5 IAMP=0
0020                   IF(A(I-1).GT.A(I)) GO TO 102
0021                   GO TO 52
                C
                C      THIS FILTER PASSES ALL VALUES FOR WHICH THE CONDITION
                C      A(I-1)>A(I) IS SATISFIED.
                C
0022                50 FREQ=FREQ+1
0023                51 I=I+1
0024                   IF(I.EQ.N+1) GO TO 200
0025                   VALU=A(I)
0026                   VALU=IABS(VALU)
0027                   IF(VALU.GT.IAMP) IAMP=VALU
0028                52 IF(A(I-1).LT.A(I)) GO TO 51
                C
0029               100 FREQ=FREQ+1
0030               101 I=I+1
0031                   IF(I.EQ.N+1) GO TO 200
0032                   VALU=A(I)
0033                   VALU=IABS(VALU)
0034                   IF(VALU.GT.IAMP) IAMP=VALU
0035               102 IF(A(I-1).GT.A(I)) GO TO 101
0036                   GO TO 50
                C
```

Note - All programs are written by the author unless otherwise specified.

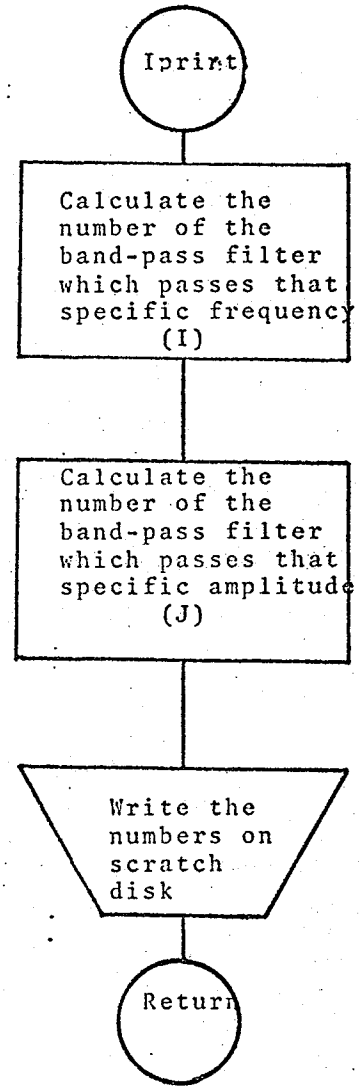
Block diagram of the main program.



```
FORTRAN IV G LEVEL  18          MAIN            DATE = 70254      23/53/52

      C     THIS FILTER PASSES ALL VALUES FOR WHICH THE CONDITION
      C     A(I-1)<A(I) IS SATISFIED.
      C
0037    200 N=N+IN
0038        II=II+1
0039        FREQ=FREQ+INC/2
0040        CALL IPRINT(IAMP,IFREQ)
0041        FREQ=0
0042        IF(N.EQ.1200+IN) GO TO 1000
0043        GO TO 5
0044   1000 CONTINUE
0045        ENDFILE 09
0046        REWIND 09
0047        CALL SMOOTH
0048        CALL PICK
0049        CALL IPLOT (TITLE,NO)
0050        GO TO 1
0051   2000 CALL PLOT(5.0,0.0,999)
0052        CALL EXIT
0053        END
```

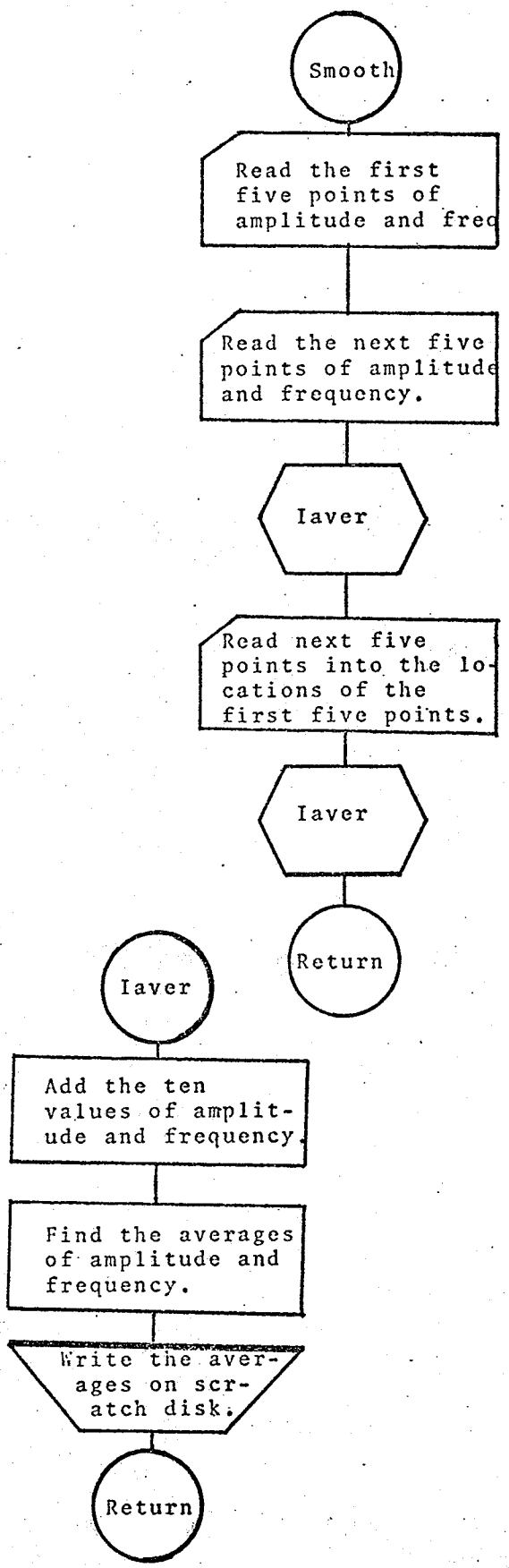
Block diagram of the subroutine "IPRINT".



```
FORTRAN IV G LEVEL 19          IPRINT          DATE = 70254          23/53/52

0001                SUBROUTINE IPRINT(IAMP,IFREQ)
                C
                C   THIS SUBROUTINE FITS THE FREQUENCY INTO A BAND-PASS FILTER FREQUENCY
                C   BAND.
                C
                C
0002                IAMP=IAMP*3
                C   CALCULATE THE AMPLITUDE BAND.
0003                I=IFREQ/175
0004                IF(IFREQ.EQ.I*175) I=I-1
0005                I=I+1
                C   J IS THE NUMBER OF THE AMPLITUDE BAND.
0006                J=IAMP/175
0007                IF(IAMP.EQ.J*175) J=J-1
0008                J=J+1
0009                WRITE(3,4) I,J
0010              4 FORMAT(2I2)
0011                RETURN
0012                END
```

Block diagram of the subroutines "SMOOTH" and "IAVER".



```
FORTRAN IV G LEVEL  18                SMOOTH              DATE = 70246          18/48/10

0001                  SUBROUTINE SMOOTH
                C
                C     THIS SUBROUTINE PERFORMS THE SMOOTHING OF THE WAVES.
                C
                C
0002                  INTEGER A(15),B(15)
                C     READ THE FIRST FIVE POINTS OF FREQUENCY AND AMPLITUDE.
0003                  READ(9,5,END=20)(A(I),B(I),I=1,5)
                C     INCREMENT THE INTERVAL WITH THE NEXT FIVE POINTS.
0004               10 READ(9,5,END=20)(A(I),B(I),I=6,10)
0005                5 FORMAT(2I2)
                C     CALL THE SUBROUTINE IAVER TO FIND THE AVERAGE OF THE TEN POINTS.
0006                  CALL IAVER(A,B)
                C     REPLACE THE FIRST FIVE POINTS WITH THE NEXT FIVE.
0007                  READ(9,5,END=20)(A(I),B(I),I=1,5)
                C     CALL THE SUBROUTINE IAVER TO FIND THE AVERAGE OF THE TEN POINTS.
0008                  CALL IAVER(A,B)
                C     GO TO BEGINNING AND PERFORM THE OPERATION WITH NEXT TEN POINTS.
0009                  GO TO 10
0010               20 ENDFILE 10
0011                  REWIND 10
0012                  REWIND 09
0013                  RETURN
0014                  END
```

```
FORTRAN IV G LEVEL 18                IAVER                DATE = 70254          23/53/52

0001                  SUBROUTINE IAVER(A,B)
                C
                C     THIS SUBROUTINE FINDS THE AVERAGE OF THE POINTS PASSED
                C     IN THE ARRAYS "A" AND "B".
                C
                C
0002                  INTEGER A(15),B(15)
                C     SET THE COUNTERS TO 0.
0003                  ITOTAL=0
0004                  ISUM=0
0005                  DO 10 I=1,10
                C     ADD THE TEN VALUES OF FREQUENCY TO ITOTAL AND THE VALUES OF
                C     AMPLITUDE TO ISUM.
0006                  ITOTAL=ITOTAL+B(I)
0007               10 ISUM=ISUM+A(I)
                C     FIND THE AVERAGE AMPLITUE AND FREQUENCY FOR THE SPECIFIC POINTS.
0008                  ITOTAL=ITOTAL/10
0009                  ISUM=ISUM/10
0010                  DO 20 I=1,5
                C     WRITE THE AVERAGES ON THE DISK.
0011               20 WRITE(1,30) ISUM,ITOTAL
0012               30 FORMAT(2I2)
0013                  RETURN
0014                  END
```

```
FORTRAN IV G LEVEL 18          PICK          DATE = 70254          23/53/52

0001                SUBROUTINE PICK
                C
                C     THIS SUBROUTINE ELIMINATES THE INBETWEEN WORD NOISE SETTING IT TO
                C     THE ZERO (1025 CPS) VALUE.
                C
                C
0002                INTEGER AMP(5),FREQ(5)
0003              1 READ(10,2,END=4)(FREQ(I),AMP(I),I=1,5)
0004              2 FORMAT(2I2)
0005                IF(AMP(3).LE.2) GO TO 4
0006              3 WRITE(9,2)(FREQ(I),AMP(I),I=1,5)
0007                GO TO 1
0008              4 IF(FREQ(3).GT.10.OR.FREQ(3).LT.6) GO TO 3
0009                DO 5 I=1,5
0010              5 FREQ(I)=11
0011                GO TO 3
0012              6 ENDFILE 09
0013                REWIND 09
0014                REWIND 10
0015                RETURN
0016                END
```

```
FORTRAN IV G LEVEL 18          MAIN          DATE = 70246          22/52/51

      C
      C     THIS PROGRAM PLOTS THE VOICE SPECTRA.
      C
      C
0001        INTEGER BLOCK,TIMES,TITLE(15),IBUF(1000)
0002        INTEGER*2 C,B,A(1200)
0003        CALL PLOTS(IBUF,1000)
0004        CALL PLOT(0.0,2.0,-3)
0005      1 READ(5,2,END=2000) BLOCK,NO,TITLE
0006      2 FORMAT(2I5,15A4)
0007      4 READ(12,3) C,B,A
0008      3  FORMAT(2A2,200A2,200A2,200A2,200A2,200A2,200A2)
0009        IF(B.NE.BLOCK-1) GO TO 4
0010        WRITE(6,2) BLOCK,NO,TITLE
0011        I=0
0012        CALL AXIS(0.0,0.0,11HTIME SAMPLE,-11,NO*1.5,0.0,0.0,20.0)
0013        CALL AXIS(0.0,0.0,9HAMPLITUDE,+9,2.0,90.0,-2047.,2047.)
0014        CALL SYMBOL(0.0,4.0,.28,TITLE,0.0,60)
0015        CALL PLOT(0.0,1.0,-3)
0016        DO 1000 TIMES=1,NO
0017        READ(12,3) C,B,A
0018        CALL IPLOT3(A,I,NO)
0019     1000 CONTINUE
0020        CALL PLOT(NO*1.5+2.,-1.,-3)
0021        GO TO 1
0022     2000 CALL PLOT(NO*1.5+5.0,0.0,999)
0023        CALL EXIT
0024        END
```

FORTRAN IV G LEVEL 18 IPLOT3 DATE = 70246

```
0001            SUBROUTINE IPLOT3(A,I,NO)
0002            INTEGER*2 A(1200)
0003            DO 100 K=1,1200
0004            I=I+1
0005            POINT=A(K)
0006            CALL PLOT(I/1200.0*1.5,POINT/2047*1.2)
0007        100 CONTINUE
0008            RETURN
0009            END
```

```
FORTRAN IV G LEVEL  18                  IPLOT                   DATE = 70246          18/48/10

0001                SUBROUTINE IPLOT(TITLE,NO)
                C
                C       THIS SUBROUTINE PLOTS THE RESULTS OF THE FILTER BANK.
                C
                C
0002                INTEGER IBUF(1000),TITLE(15)
0003                CALL PLOT(0.0,0.0,-3)
0004                CALL AXIS(0.0,0.0,11HTIME SAMPLE,-11,NO*1.5,0.0,0.0,20.0)
0005                CALL AXIS(0.0,0.0,9HFREQUENCY,+9,5.0,90.0,0.0,1400.0)
0006                CALL SYMBOL(0.0,6.0,.28,TITLE,0.0,60)
0007                I=1
0008              5 Y=I*175./7000.*5.0
0009                CALL PLOT(0.0,Y,3)
0010                CALL PLOT(NO*1.5,Y,2)
0011                I=I+1
0012                Y=I*175.0/7000.0*5.0
0013                CALL PLOT(NO*1.5,Y,3)
0014                CALL PLOT(0.0,Y,2)
0015                I=I+1
0016                IF(I.LT.40) GO TO 5
0017                CALL PLOT(0.0,0.0,3)
0018                DO 20 I=1,1000
0019                READ(9,10,END=30) K
0020             10 FORMAT(I2)
0021                CALL SYMBOL(I/20.0,K/40.0*5.0,.07,11,0.0,-2)
0022             20 CONTINUE
0023             30 REWIND 09
0024                CALL PLOT(0.0,0.0,3)
0025                DO 50 I=1,1000
0026                READ(9,40,END=60) L
0027             40 FORMAT(2X,I2)
0028                CALL PLOT(I/20.0,L/40.0*5.0,2)
0029             50 CONTINUE
0030             60 CALL PLOT(NO*1.5+2,0.0,-3)
0031                REWIND 09
0032                RETURN
0033                END
```

```
FORTRAN IV G LEVEL  18                  IPLOT2              DATE = 70246          18/48/10

0001                  SUBROUTINE IPLOT2(TITLE,NO)
                C
                C     THIS SUBROUTINE PLOTS THE SMOOTHED WAVES.
                C
                C
0002                  INTEGER IBUF(1000),TITLE(15)
0003                  CALL AXIS(0.0,0.0,9HFREQUENCY,+9,5.0,90.0,0.0,1400.0)
0004                  CALL AXIS(0.0,0.0,11HTIME SAMPLE,-11,NO*1.5,0.0,0.0,20.0)
0005                  CALL SYMBOL(0.0,6.0,.28,TITLE,0.0,60)
0006                  I=1
0007                5 Y=I*175./7000.*5.0
0008                  CALL PLOT(0.0,Y,3)
0009                  CALL PLOT(NO*1.5,Y,2)
0010                  I=I+1
0011                  Y=I*175.0/7000.0*5.0
0012                  CALL PLOT(NO*1.5,Y,3)
0013                  CALL PLOT(0.0,Y,2)
0014                  I=I+1
0015                  IF(I.LT.40) GO TO 5
0016                  CALL PLOT(0.0,0.0,3)
0017                  DO 20 I=1,1000
0018                  READ(10,10,END=30) K
0019               10 FORMAT(I2)
0020                  CALL SYMBOL(I/20.0,K/40.0*5.0,.07,11,0.0,-2)
0021               20 CONTINUE
0022               30 REWIND 10
0023                  CALL PLOT(0.0,0.0,3)
0024                  DO 50 I=1,1000
0025                  READ(10,40,END=60) L
0026               40 FORMAT(2X,I2)
0027                  CALL PLOT(I/20.0,L/40.0*5.0,2)
0028               50 CONTINUE
0029               60 CALL PLOT(NO*1.5+2,00.0,-3)
0030                  REWIND 10
0031                  RETURN
0032                  END
```

"VOICE"

LEVEL 1JAN67 COBOL F

1

```
00001      100 IDENTIFICATION DIVISION.
00002      110 PROGRAM-ID. 'BINBIN'.
00003      120 AUTHOR.  D.R.SPRAGUE.
00004      130 DATE-WRITTEN. NOV.15,1967.
00005      131 DATE-COMPILED. SEP  3,1970
00006      140 REMARKS. THIS PROGRAM CONVERTS A 12-BIT SIGNED BINARY NUMBER ON
00007      150           7-TRACK TAPE TO A 16-BIT SIGNED BINARY NUMBER ON 9-TRACK
00008      200 ENVIRONMENT DIVISION.
00009      210 CONFIGURATION SECTION.
00010      220 SOURCE-COMPUTER. IBM-360 H65.
00011      230 OBJECT-COMPUTER. IBM-360 H65.
00012      240 INPUT-OUTPUT SECTION.
00013      250 FILE-CONTROL.
00014      260     SELECT OCT-TAPE ASSIGN TO 'TAPEIN' UTILITY 2400 UNIT, RESERVE
00015      270         1 ALTERNATE AREA.
00016      280     SELECT HEX-TAPE ASSIGN TO 'TAPEOUT' UTILITY 2314 UNIT
00017      290         RESERVE 1 ALTERNATE AREA.
00018      300     SELECT PRINTER ASSIGN TO 'SYSOUT' UNIT-RECORD 1403 UNIT.
00019     1000 DATA DIVISION.
00020     1010 FILE SECTION.
00021     1020 FD  OCT-TAPE LABEL RECORD IS OMITTED, RECORDING MODE IS F,
00022     1030         BLOCK CONTAINS 2404 CHARACTERS, DATA RECORD IS REC-IN.
00023     1040 01  REC-IN.
00024     1050     02  IN-GUTS                                  PICTURE X(2404).
00025     1060 FD  HEX-TAPE LABEL RECORD IS STANDARD, RECORDING MODE IS F,
00026     1070         BLOCK CONTAINS 2404 CHARACTERS, DATA RECORD IS REC-OUT.
00027     1090 01  REC-OUT.
00028     1100     02  OUT-GUTS                                 PICTURE X(2404).
00029     1110 FD  PRINTER  DATA RECORDS ARE LINE-OUT, TITLE-1
00030     1120         LABEL RECORD IS OMITTED, RECORDING MODE IS F.
00031     1130 01  LINE-OUT.
00032     1140     02  FILLER                                   PICTURE X.
00033     1150     02  NUMBER-FLD   OCCURS 16 TIMES             PICTURE -(7)9.
00034     1160     02  FILLER                                   PICTURE X(15).
00035     1170 01  TITLE-1.
00036     1180     02  FILLER                                   PICTURE X.
00037     1190     02  NAME-1                                   PICTURE X(135).
00038     1300 WORKING-STORAGE SECTION.
00039     1310 77  X       COMPUTATIONAL   VALUE 0              PICTURE S9999.
00040     1320 77  Y       COMPUTATIONAL   VALUE 0              PICTURE S9999.
00041     1330 77  Z       COMPUTATIONAL   VALUE 0              PICTURE S9999.
00042     1332 77  LINE-CNT                                     PICTURE S999.
00043     1340 01  OPTIONS.
00044     1341     02  BRANCH-OPT                               PICTURE 9.
00045     1342     02  BLK-CNT-OPT                              PICTURE 999.
00046     1343     02  CHANNEL-OPT                              PICTURE 99.
00047     1400 01  WORK-IN.
00048     1410     02  IN-CHAR     OCCURS 2404 TIMES            PICTURE X.
00049     1420 01  WORK-OUT.
00050     1430     02  OUT-CHAR    OCCURS 1202 TIMES
00051     1440             USAGE IS COMPUTATIONAL               PICTURE S9999.
00052     1450 01  RECEIVES.
00053     1460     02  HOLD-ON.
00054     1470         03  1ST-HALF                             PICTURE X.
```

2

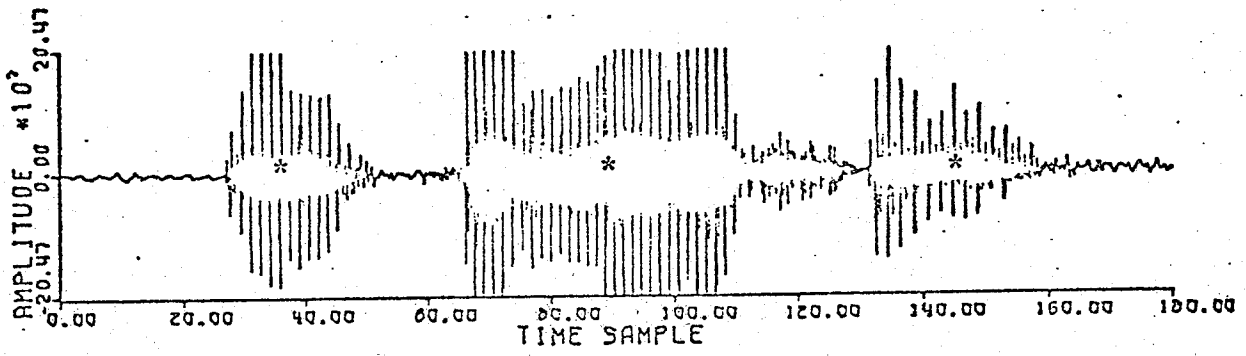
```
00055     1480           03  2ND-HALF                          PICTURE X.
00056     1490       02  BINARY-1 REDEFINES HOLD-ON
00057     1500                     USAGE IS COMPUTATIONAL      PICTURE S9999.
00058     1510       02  SHIFTER.
00059     1520           03  HI-HALF                           PICTURE X.
00060     1530           03  LO-HALF                           PICTURE X.
00061     1540       02  BINARY-2 REDEFINES SHIFTER
00062     1550                     USAGE IS COMPUTATIONAL      PICTURE S9999.
00063     1560 01    BLOCK-CNTR.
00064     1570       02  FILLER       VALUE 'RUN NO. '        PICTURE X(8).
00065     1580       02  R-CNTR                               PICTURE ZZ9.
00066     1590       02  FILLER  VALUE SPACE                  PICTURE X(8).
00067     1600       02  FILLER  VALUE 'BLOCK NO. '           PICTURE X(10).
00068     1610       02  B-CNTR                               PICTURE ZZZ9.
00069     2000 PROCEDURE DIVISION.
00070     2010 INITIALIZE.
00071     2030       ACCEPT OPTIONS.
00072     2040       GO TO READ-IN, WRITE-READ, READ-BACK DEPENDING ON BRANCH-OPT.
00073     2100 READ-IN.
00074     2101       OPEN INPUT OCT-TAPE, OUTPUT HEX-TAPE.
00075     2104 READ-IT.
00076     2110       READ OCT-TAPE INTO WORK-IN AT END GO TO EOF-7T.
00077     2120       MOVE 0 TO Z.
00078     2122       MOVE -1 TO X. MOVE 0 TO Y.
00079     2130       PERFORM CONV-RTN 1202 TIMES.
00080     2140       WRITE REC-OUT FROM WORK-OUT.
00081     2150       GO TO READ-IT.
00082     2200 CONV-RTN.
00083     2210       ADD 2 TO X.   ADD 2 TO Y.   ADD 1 TO Z.
00084     2220       MOVE 0 TO BINARY-1, BINARY-2.
00085     2230       MOVE IN-CHAR (Y) TO 2ND-HALF.
00086     2240       MOVE IN-CHAR (X) TO HI-HALF.
00087     2260       COMPUTE BINARY-1 = BINARY-1 + (BINARY-2 / 4).
00088     2270       IF BINARY-1 NOT < 2048, COMPUTE BINARY-1 = -1 * (BINARY-1
00089     2280           - 2048).
00090     2330       MOVE BINARY-1 TO OUT-CHAR (Z).
00091     2500 EOF-7T.
00092     2510       CLOSE OCT-TAPE, HEX-TAPE.
00093     2511 SWITCH-1.
00094     2512       GO TO CLOSE-OUT.
00095     2530 WRITE-READ.
00096     2540       ALTER SWITCH-1 TO PROCEED TO READ-BACK.
00097     2550       GO TO READ-IN.
00098     2560 READ-BACK.
00099     2570       OPEN INPUT HEX-TAPE, OUTPUT PRINTER.
00100     2580       MOVE 1 TO Y.
00101     2600 GO-ON.
00102     2605       MOVE 2 TO Z.
00103     2610       READ HEX-TAPE INTO WORK-OUT AT END GO TO CLOSE-2.
00104     2615       MOVE OUT-CHAR (1) TO R-CNTR.
00105     2620       MOVE OUT-CHAR (2) TO B-CNTR.
00106     2625       MOVE BLOCK-CNTR TO NAME-1.
00107     2630       WRITE TITLE-1 AFTER ADVANCING 0 LINES.
00108     2631       MOVE SPACES TO TITLE-1.
00109     2634       IF OUT-CHAR (2) > BLK-CNT-OPT GO TO CLOSE-2.
00110     2640       PERFORM MOVE-PRINT UNTIL Z > 1201.
00111     2670       GO TO GO-ON.
```

```
00112     2680 CLOSE-2.
00113     2690      CLOSE PRINTER, HEX-TAPE.
00114     2700 CLOSE-OUT.
00115     2710      STOP RUN.
00116     2800 MOVE-PRINT.
00117     2810      PERFORM LOAD-UP THRU E1 VARYING X FROM 1 BY 1
00118     2811          UNTIL X > CHANNEL-OPT.
00119     2820      WRITE LINE-OUT AFTER ADVANCING 1 LINES.
00120     2830    * MOVE SPACES TO LINE-OUT.
00121     2840 LOAD-UP.
00122     2850      ADD 1 TO Z.
00123     2860      MOVE OUT-CHAR (Z) TO NUMBER-FLD (X).
00124     2870      IF Z > 120 THEN IF X = CHANNEL-OPT MOVE 1 TO Y
00125     2871          ELSE COMPUTE Y = X + 1  COMPUTE X = CHANNEL-OPT + 1
00126     2872          ELSE NEXT SENTENCE.
00127     2880 E1.
00128     2881      EXIT.
```

APPENDIX B

TABLE 3.1 (continued) (Part B)

INFORMATION BY A.Y.



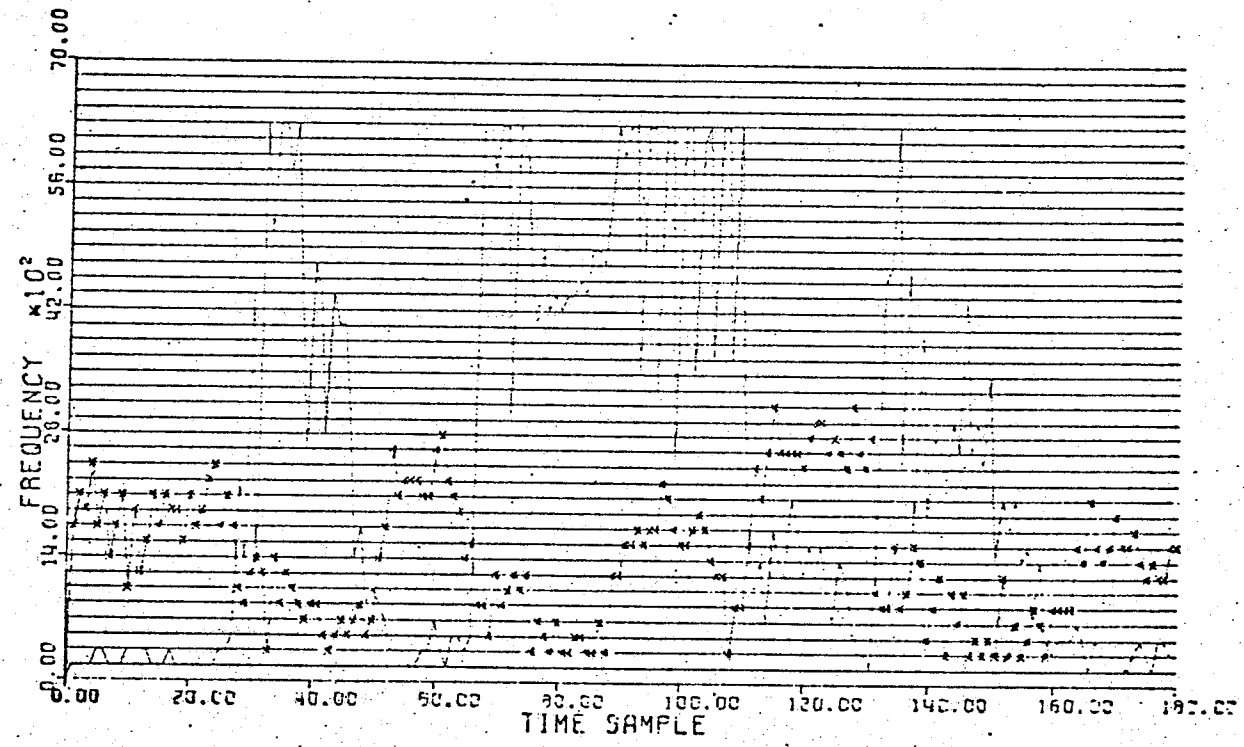
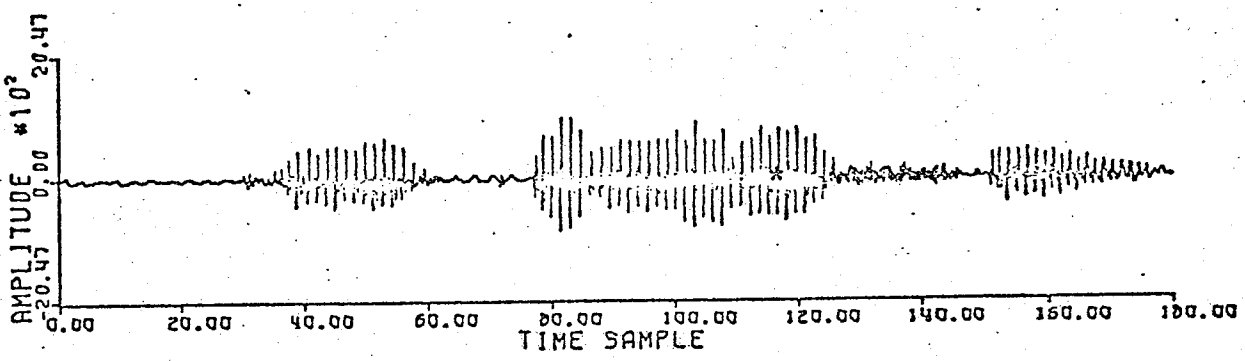


TABLE 3.1 (continued)

INFORMATION BY D.C.



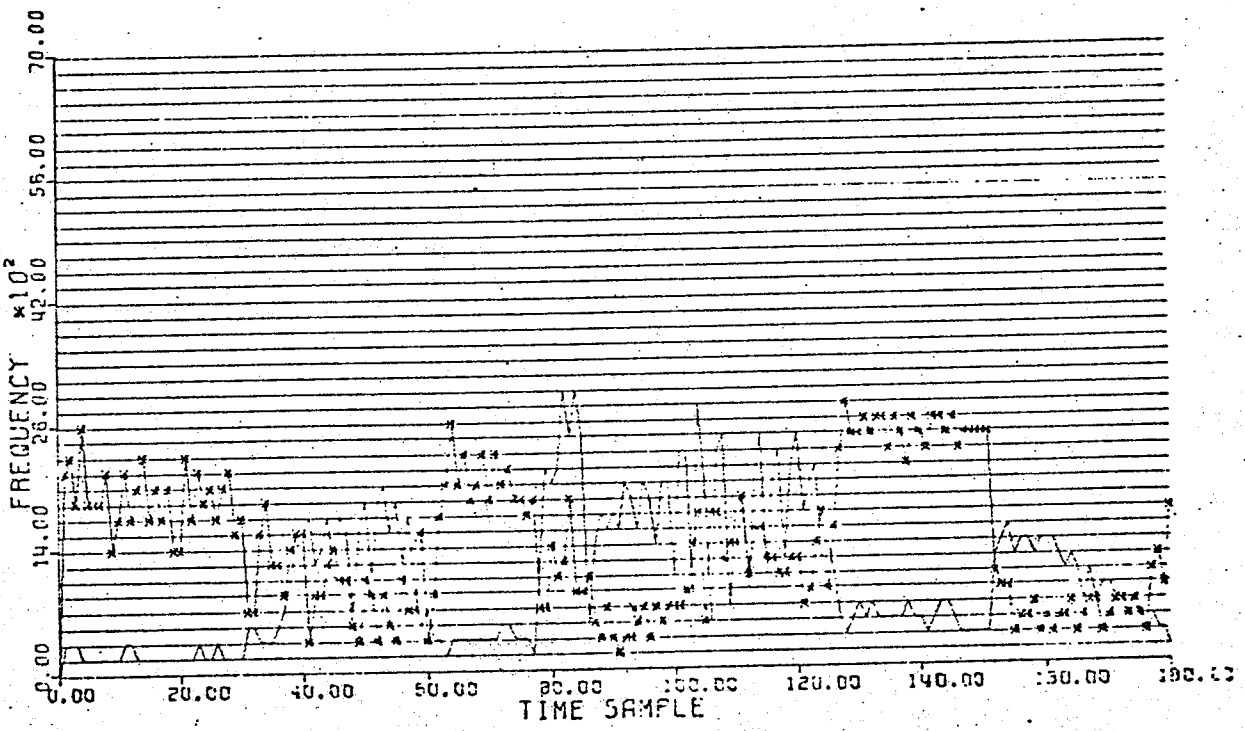
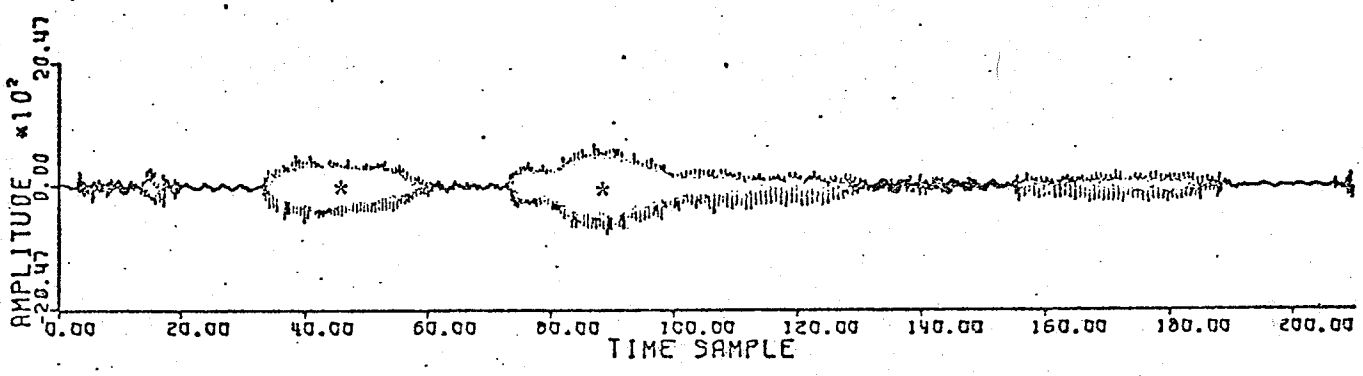


TABLE 3.1 (continued)

INFORMATION BY I.R.



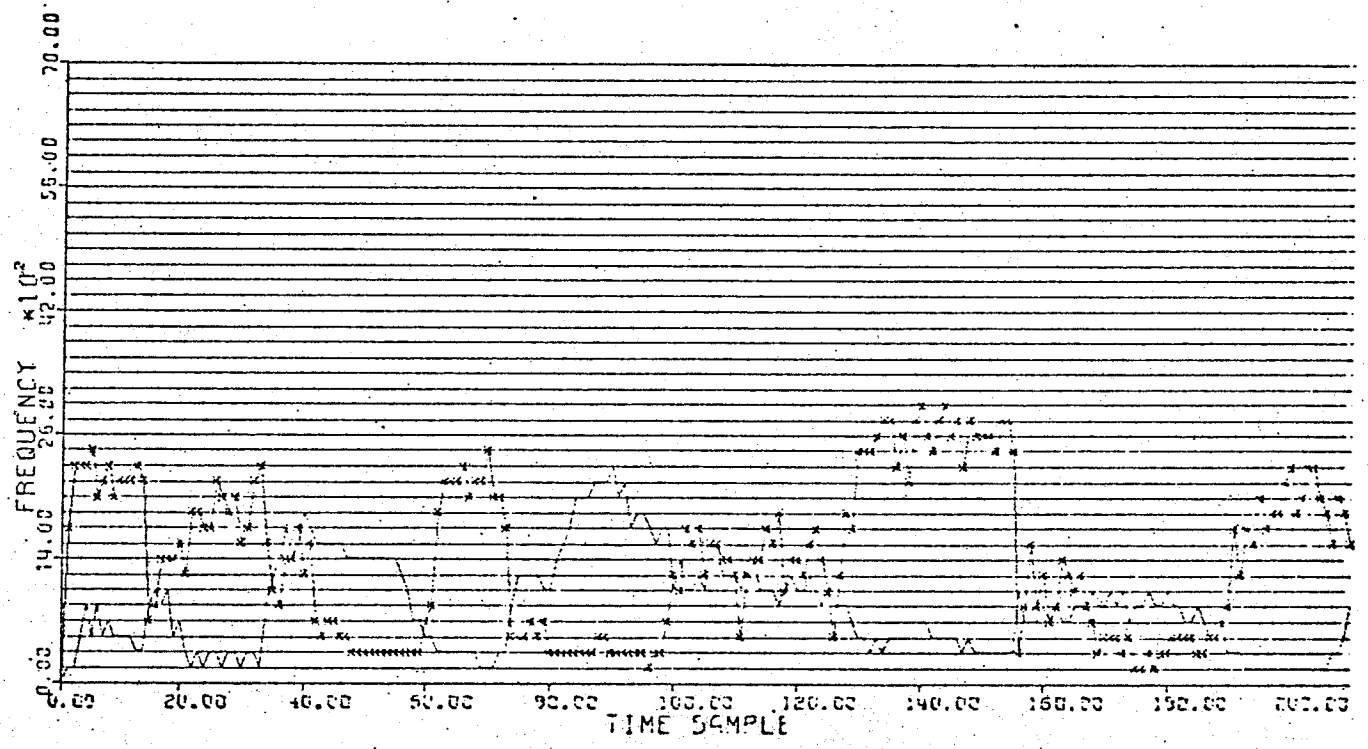
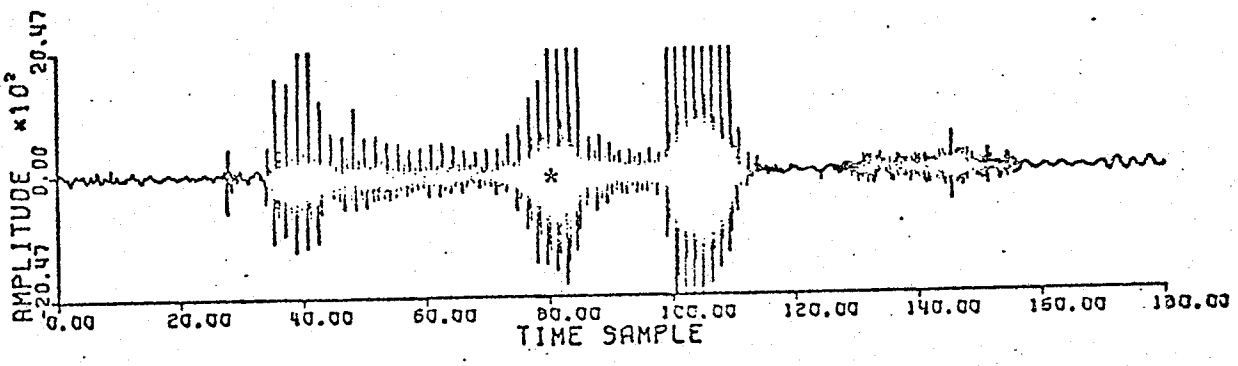


TABLE 3.1 (continued)

EXHIBITS BY A.Y.



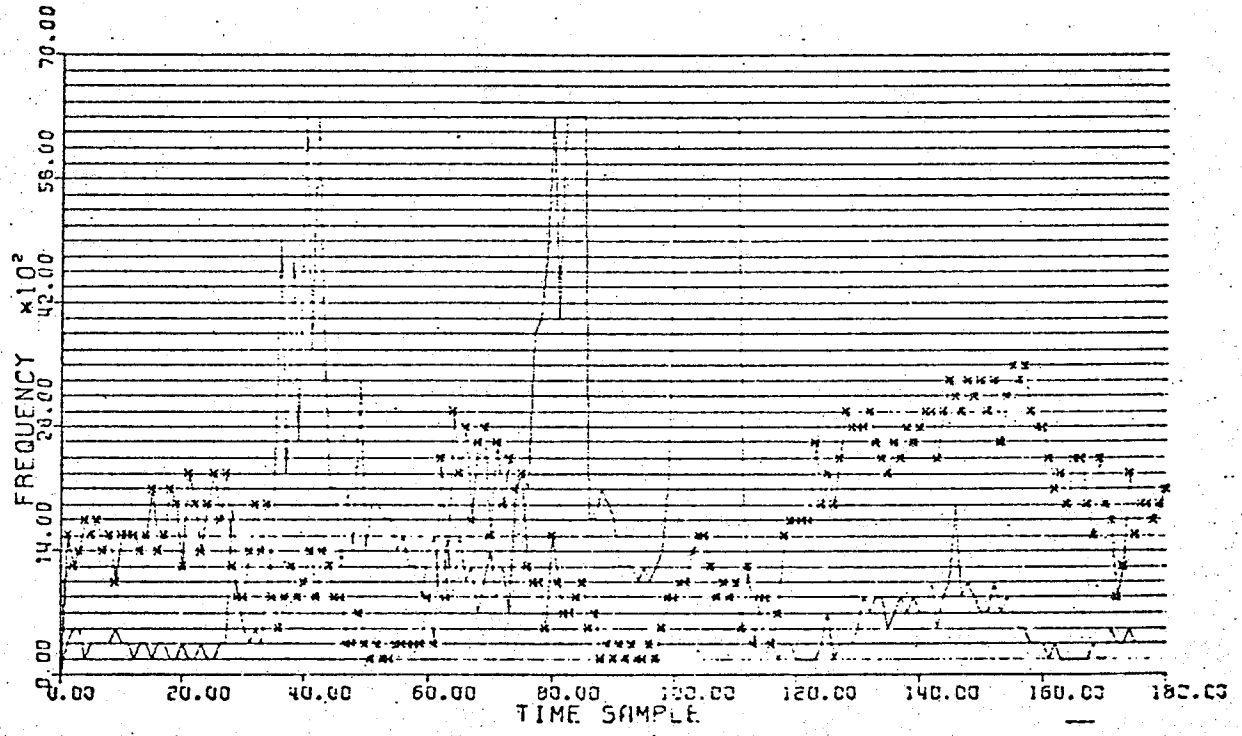
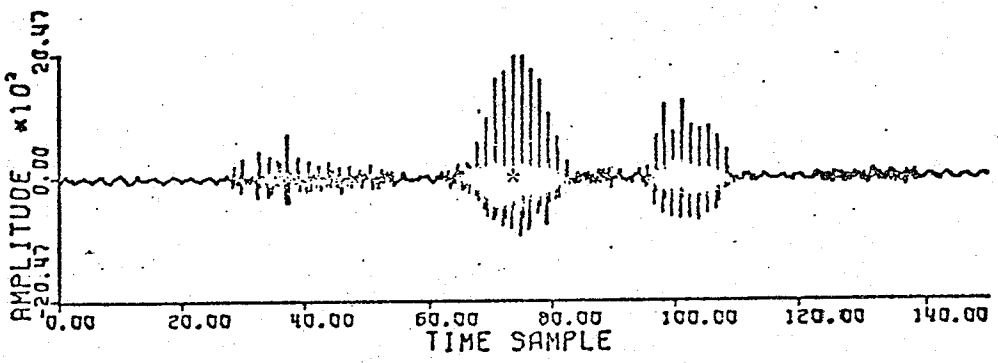


TABLE 3.1 (continued)

EXHIBITS BY D.C.



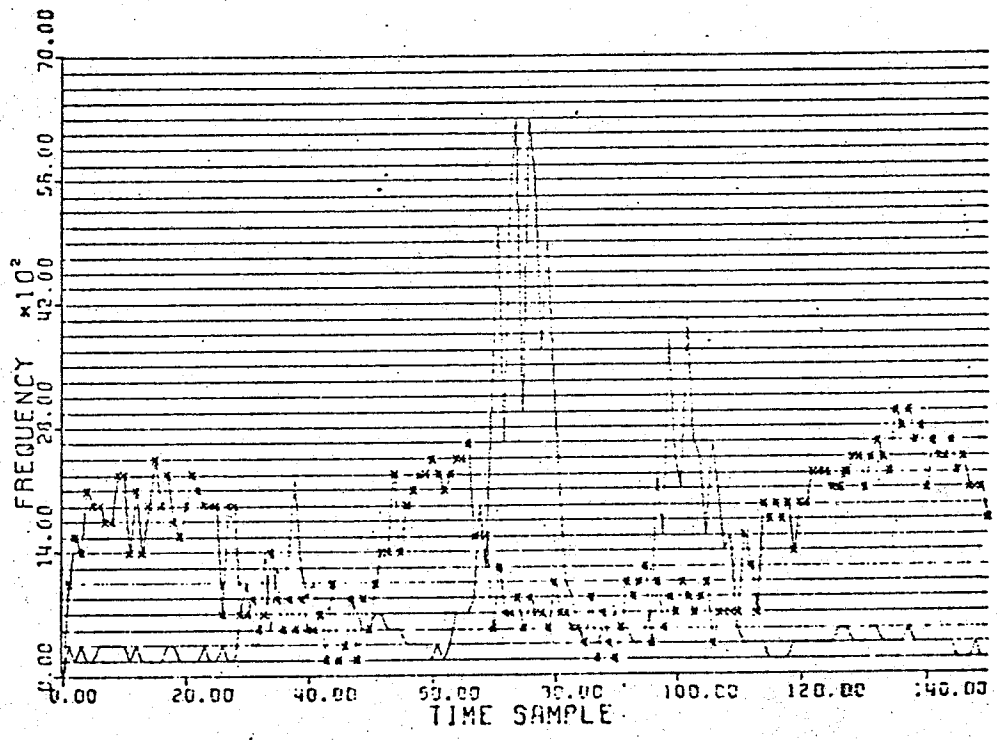
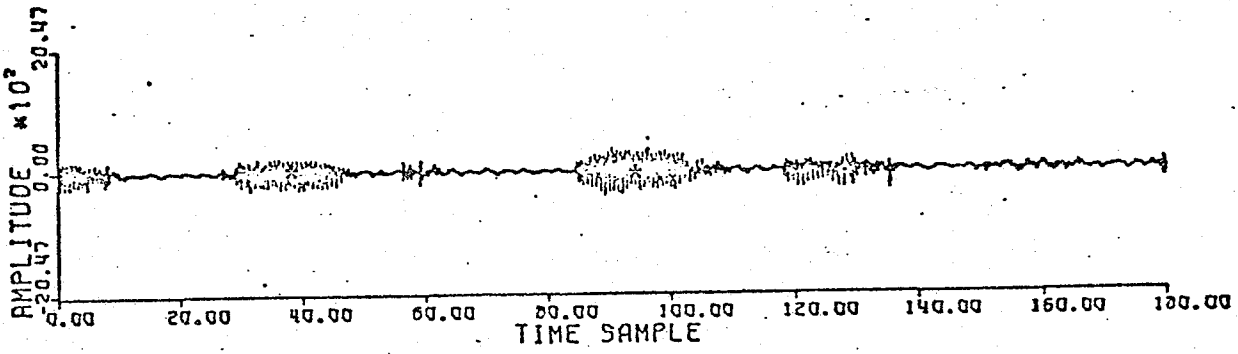


TABLE 3.1 (continued)

EXHIBITS BY I.R.



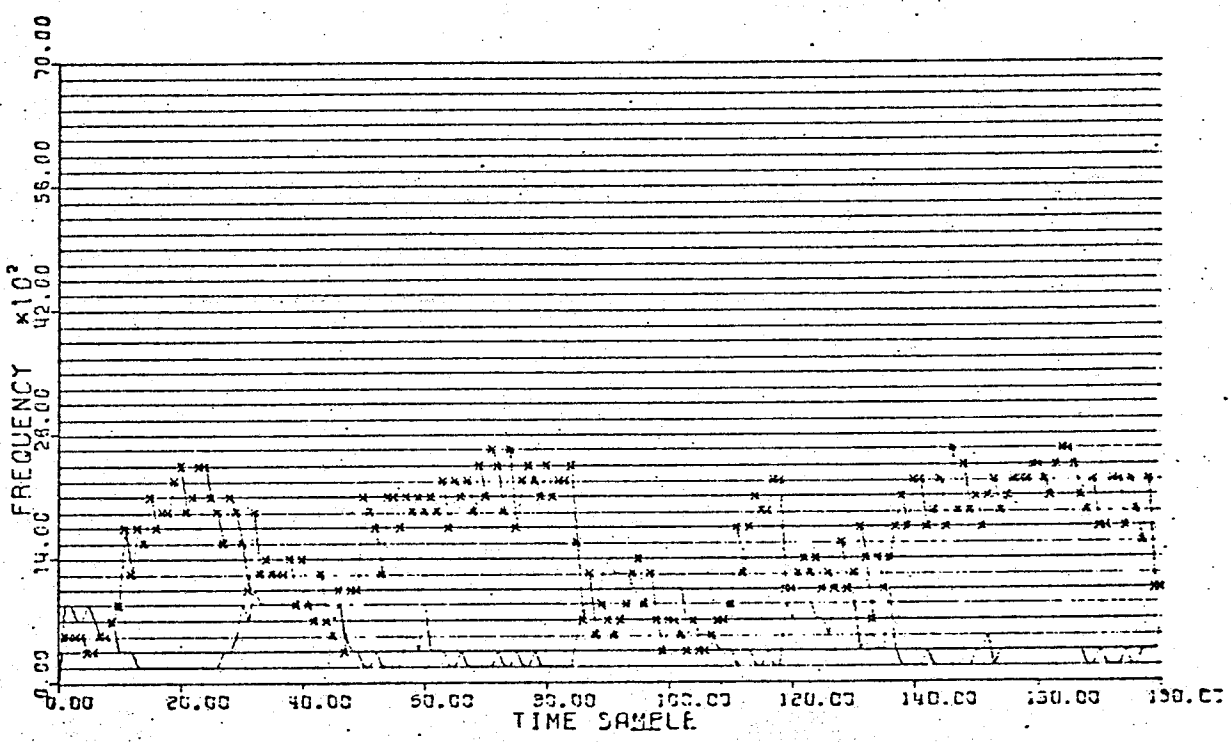
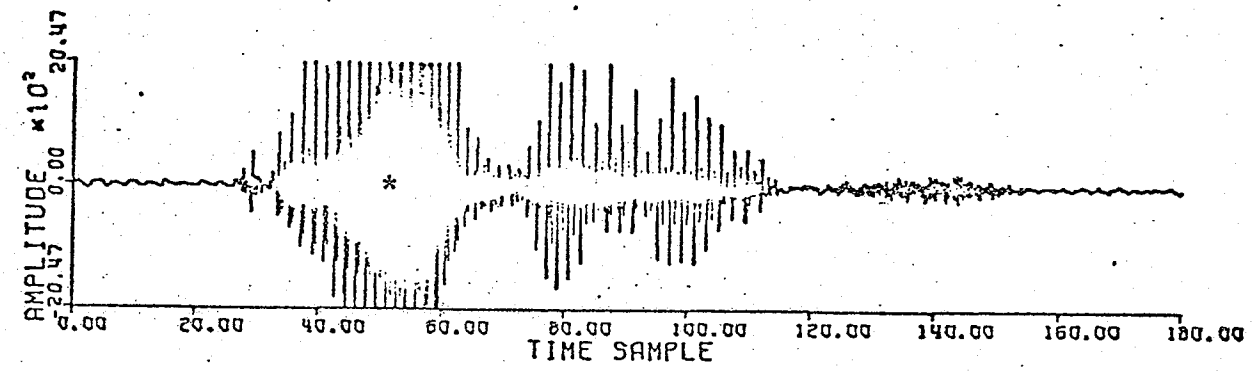


TABLE 3.1 (continued)



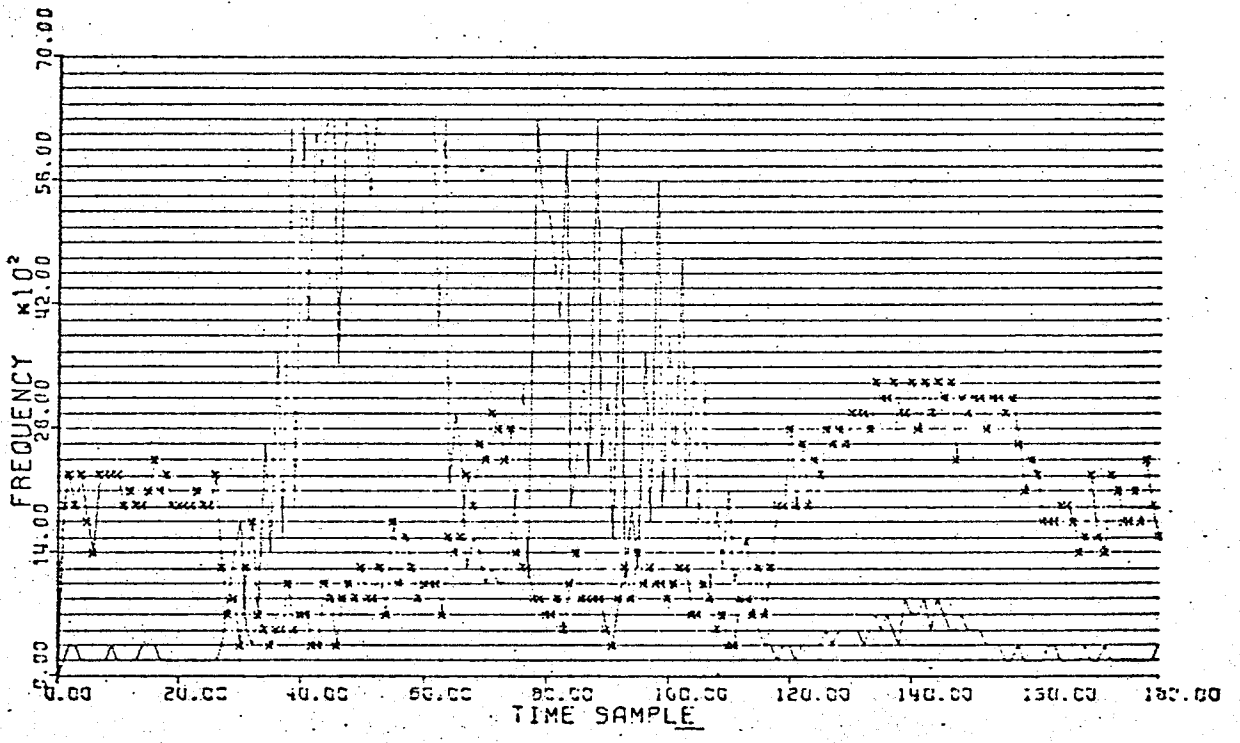
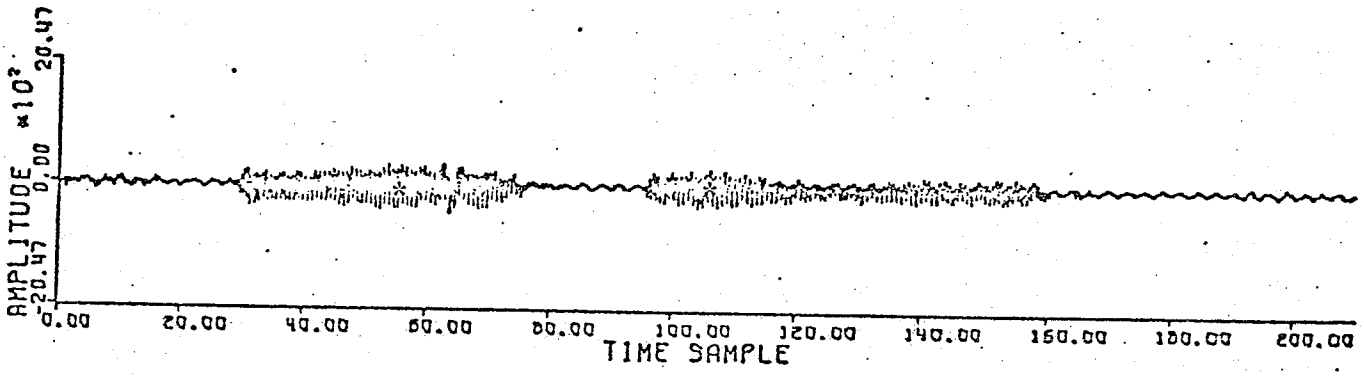


TABLE 3.1 (continued)

RESONANCE BY I.R.



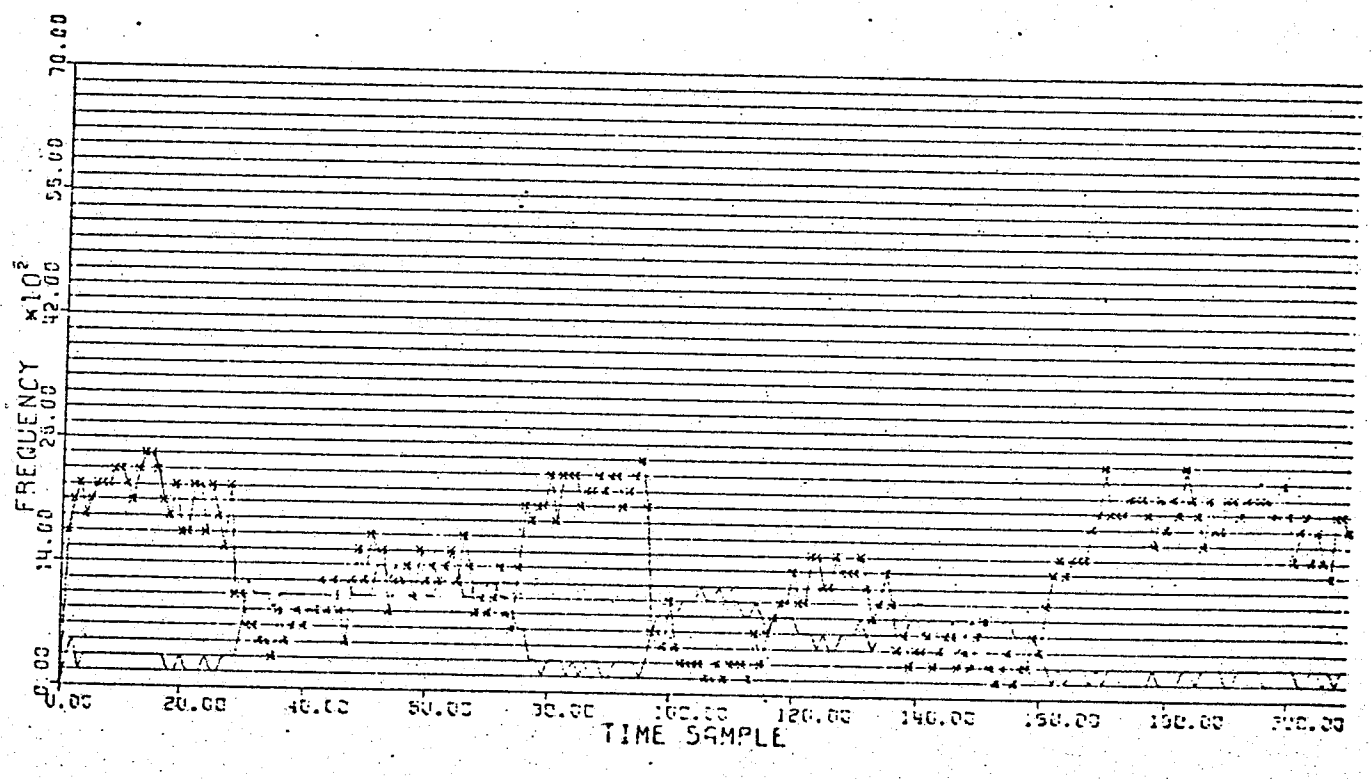


TABLE 3.1 (continued)

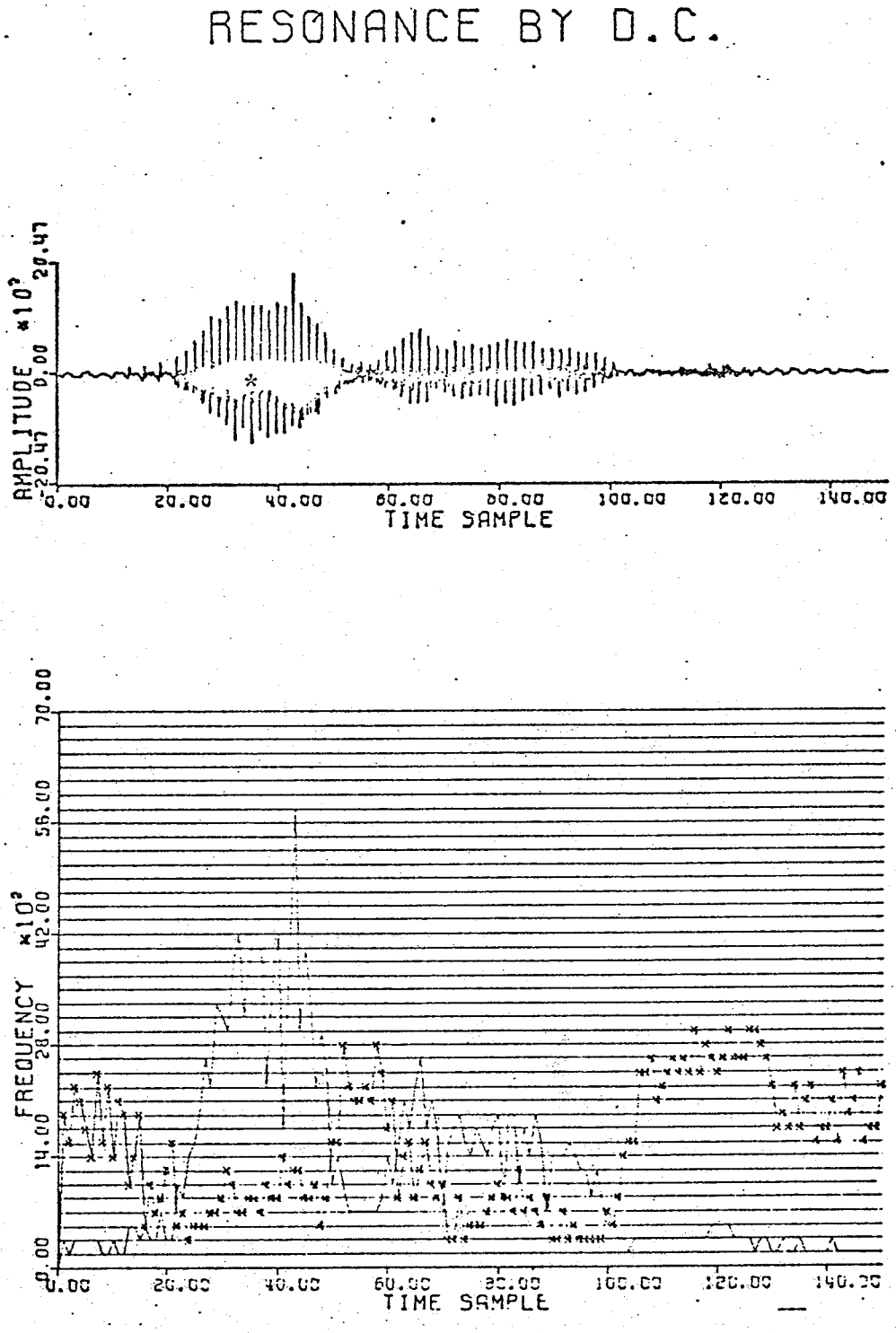
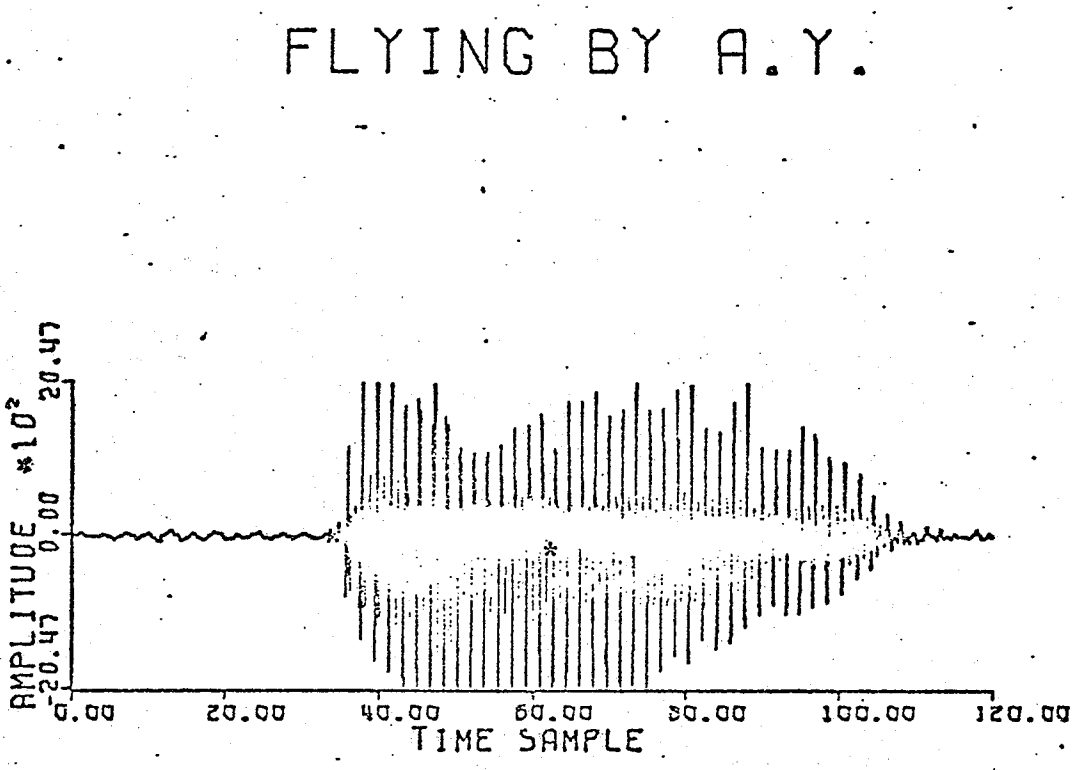


TABLE 3.1 (continued)



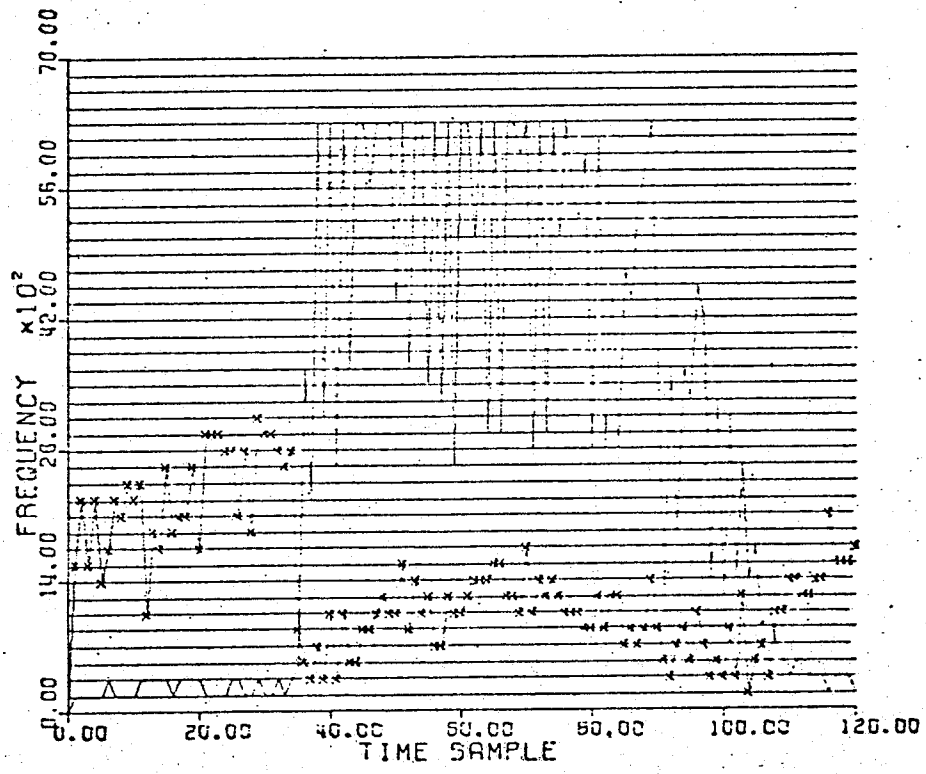
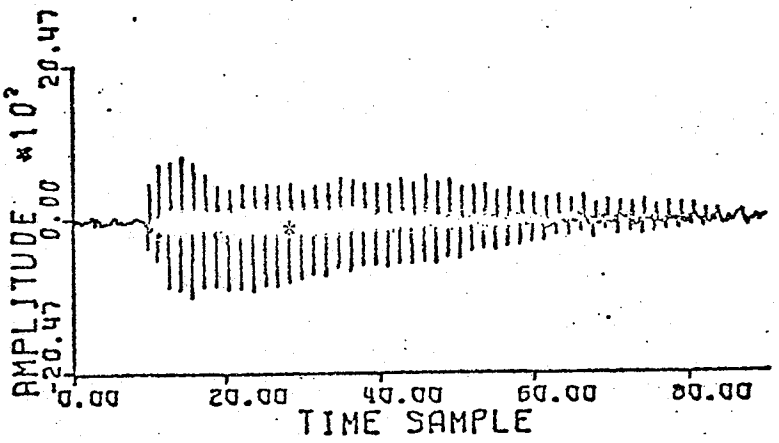


TABLE 3.1 (continued)

FLYING BY D.C.



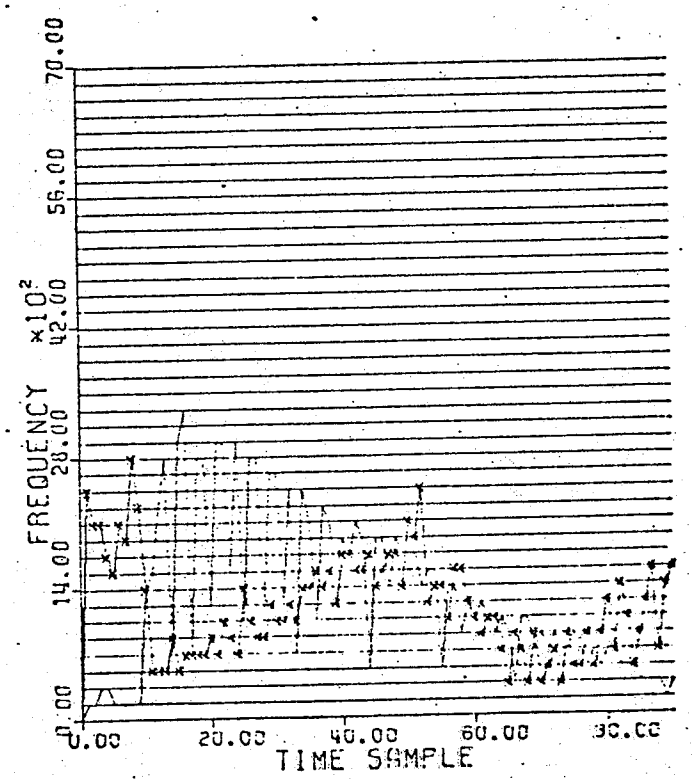


TABLE 3.1 (continued)

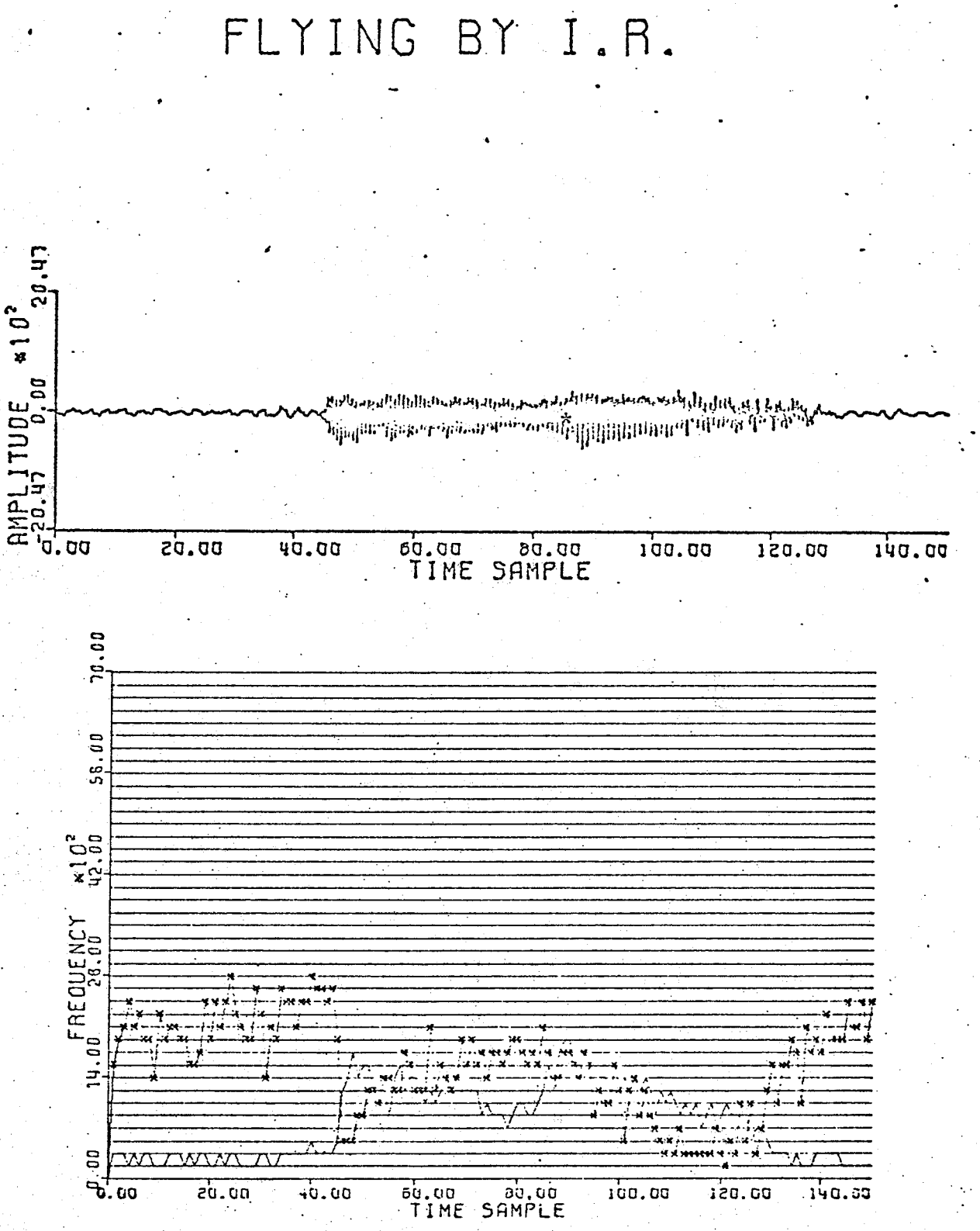
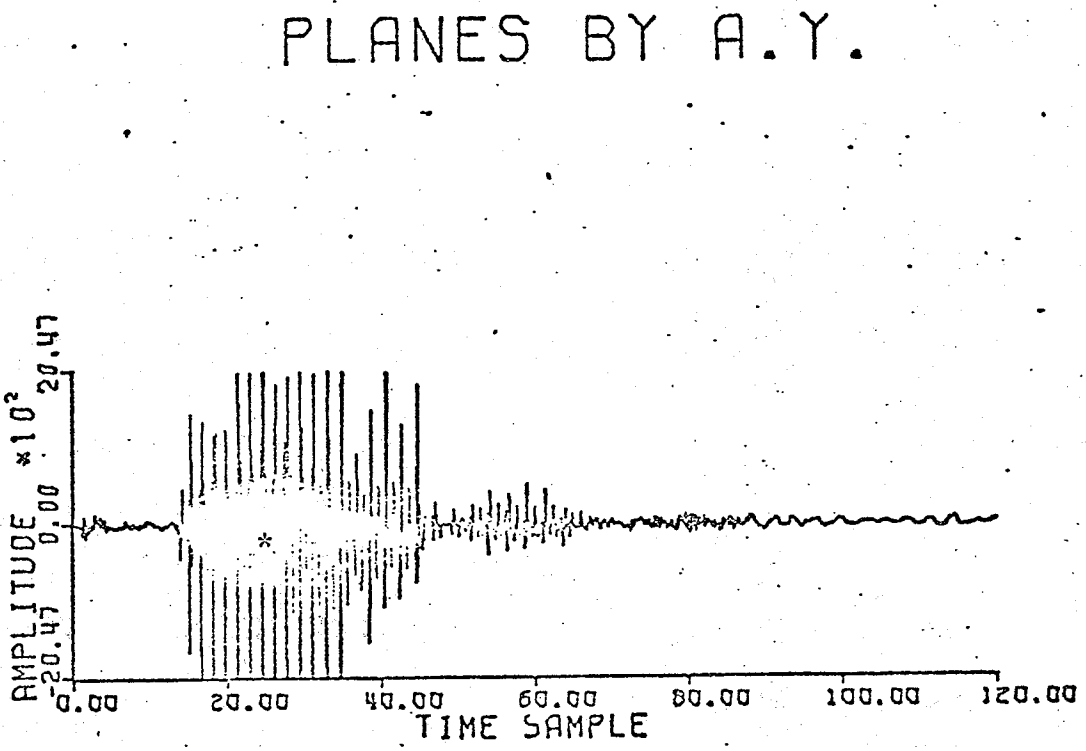


TABLE 3.1 (continued)



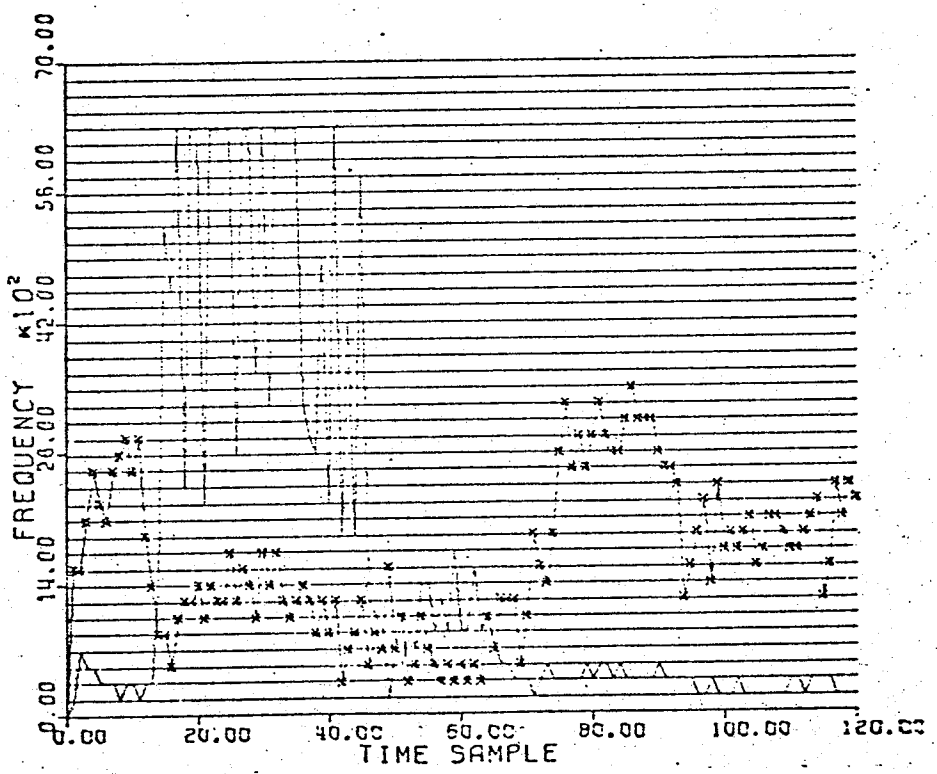
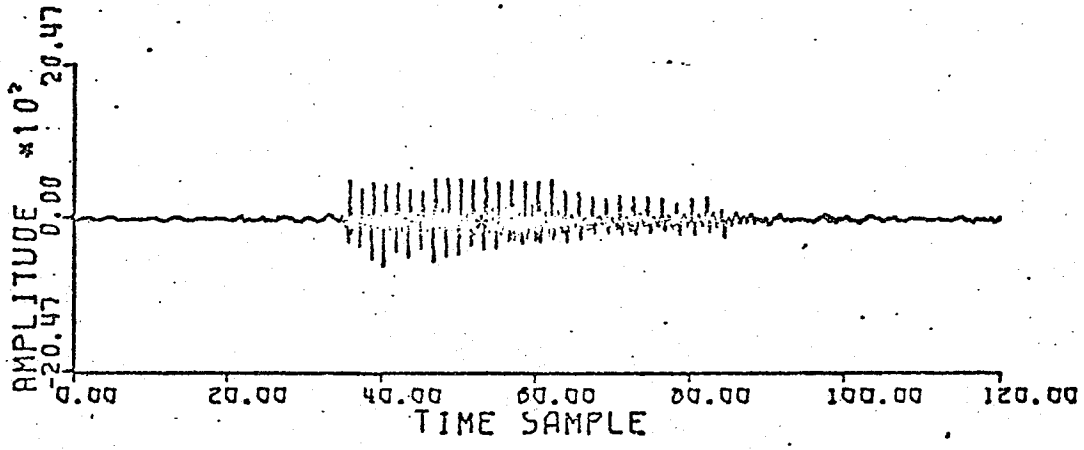


TABLE 3.1 (continued)

PLANES BY D.C.



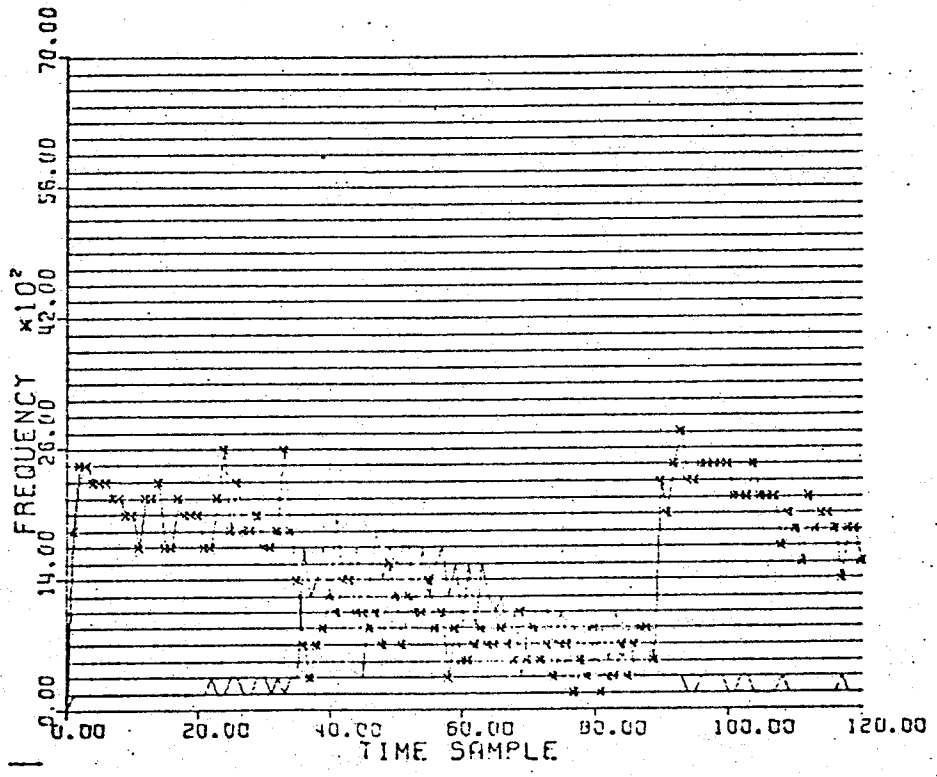
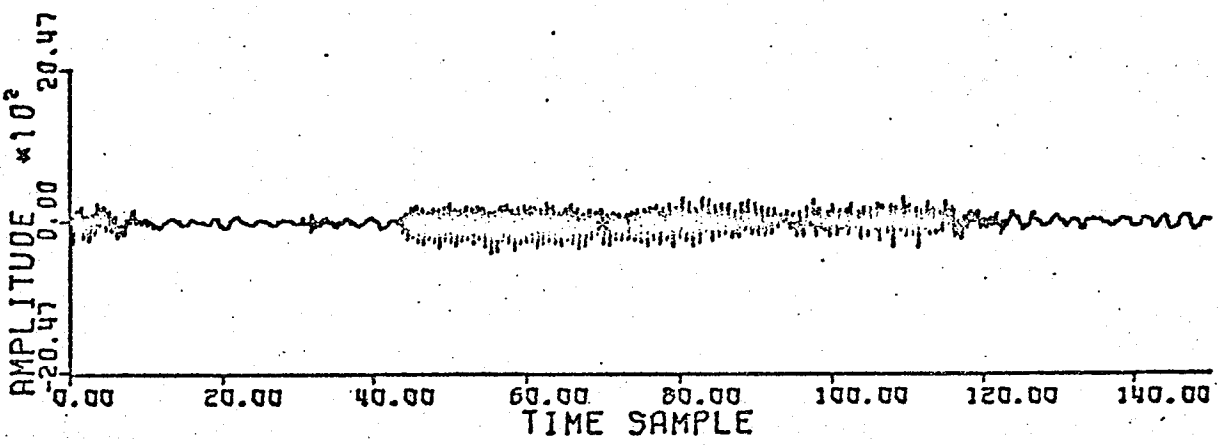
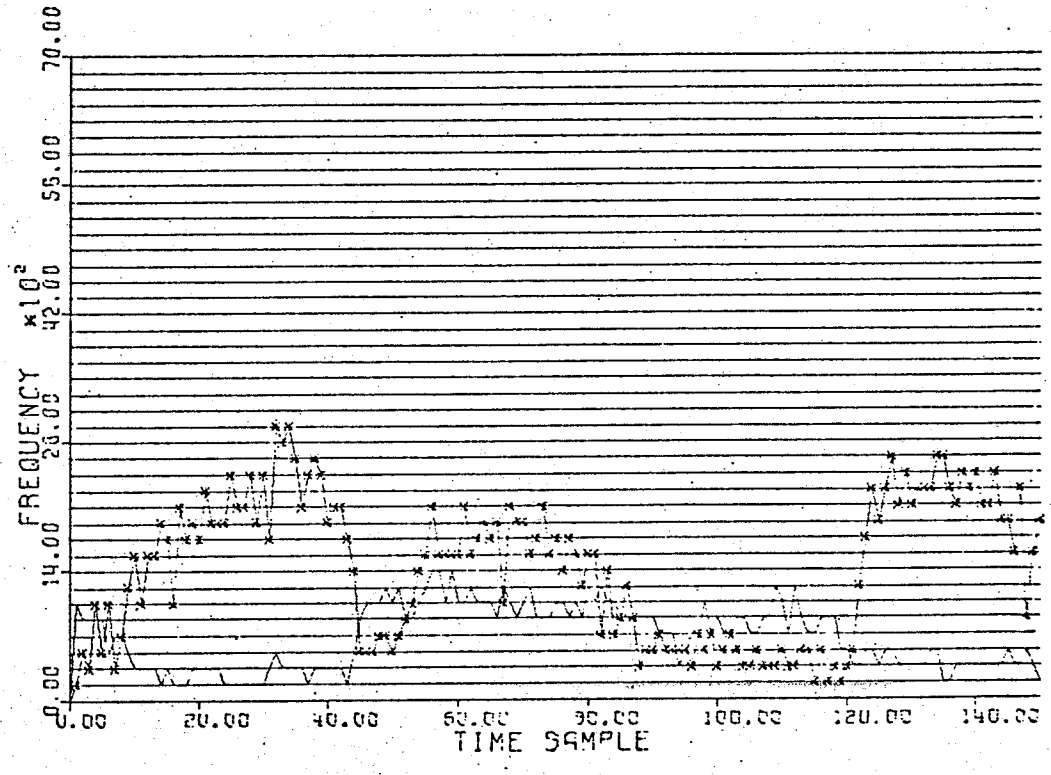


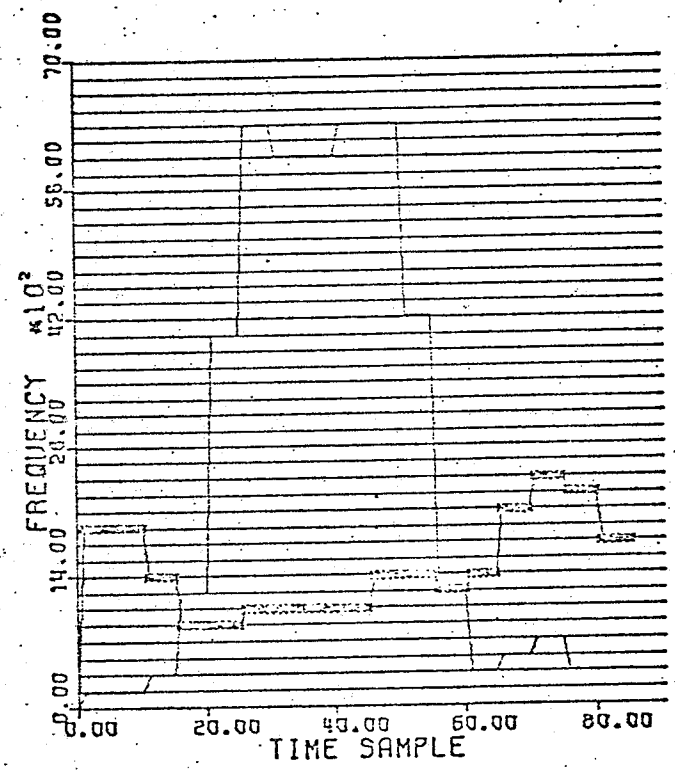
TABLE 3.1 (continued)

PLANES BY I.R.

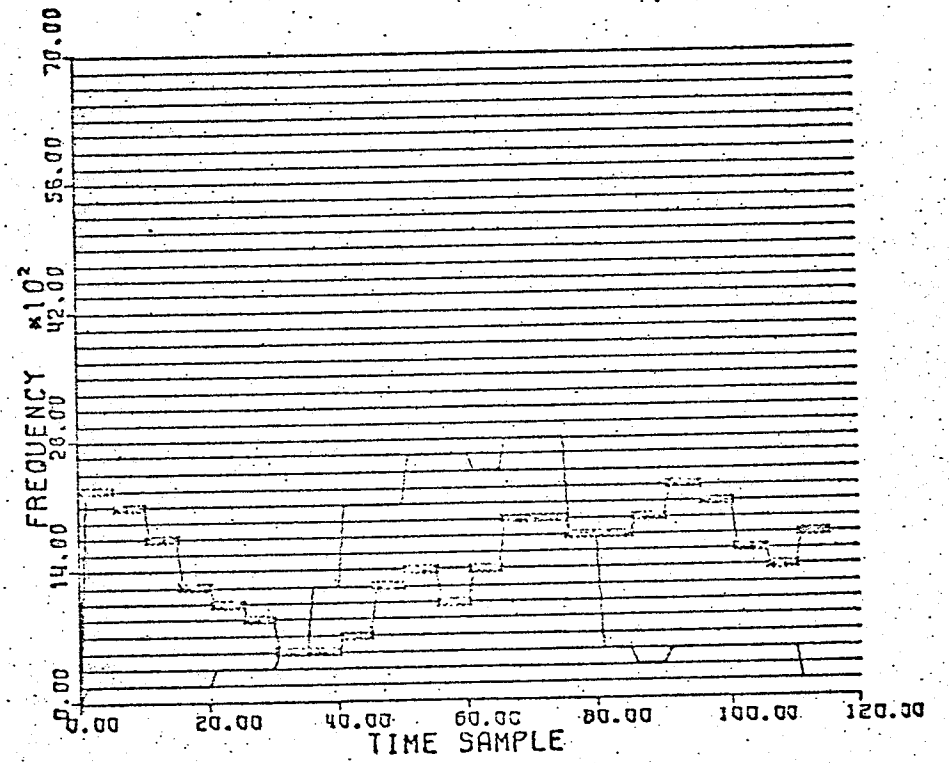




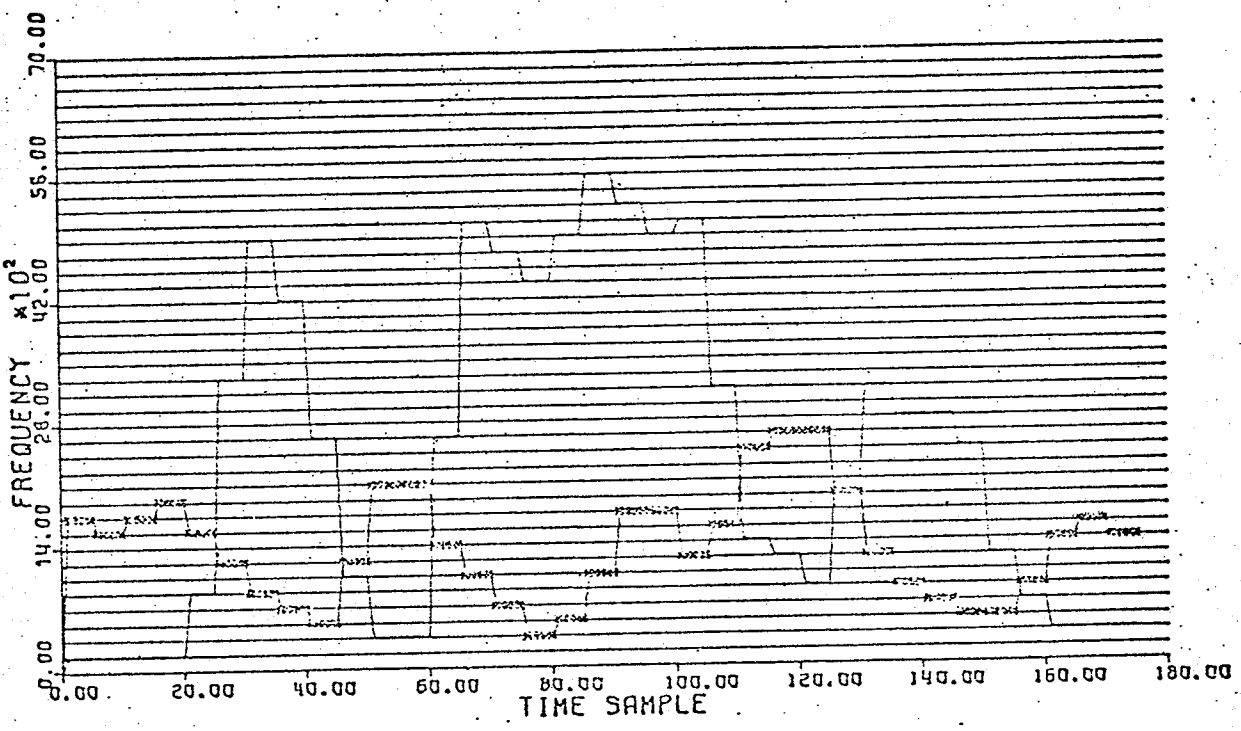
THAT BY A.Y.



THAT BY I.R.



INFORMATION BY A.Y.



INFORMATION BY D.C.

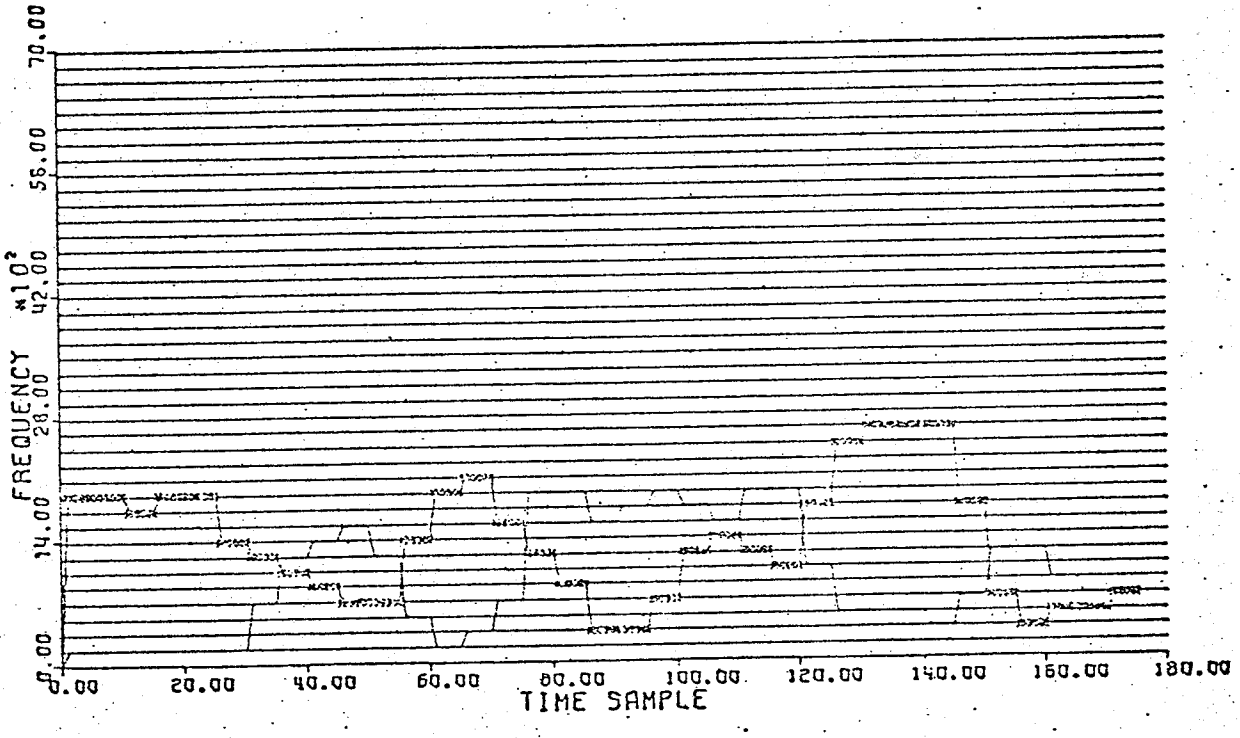
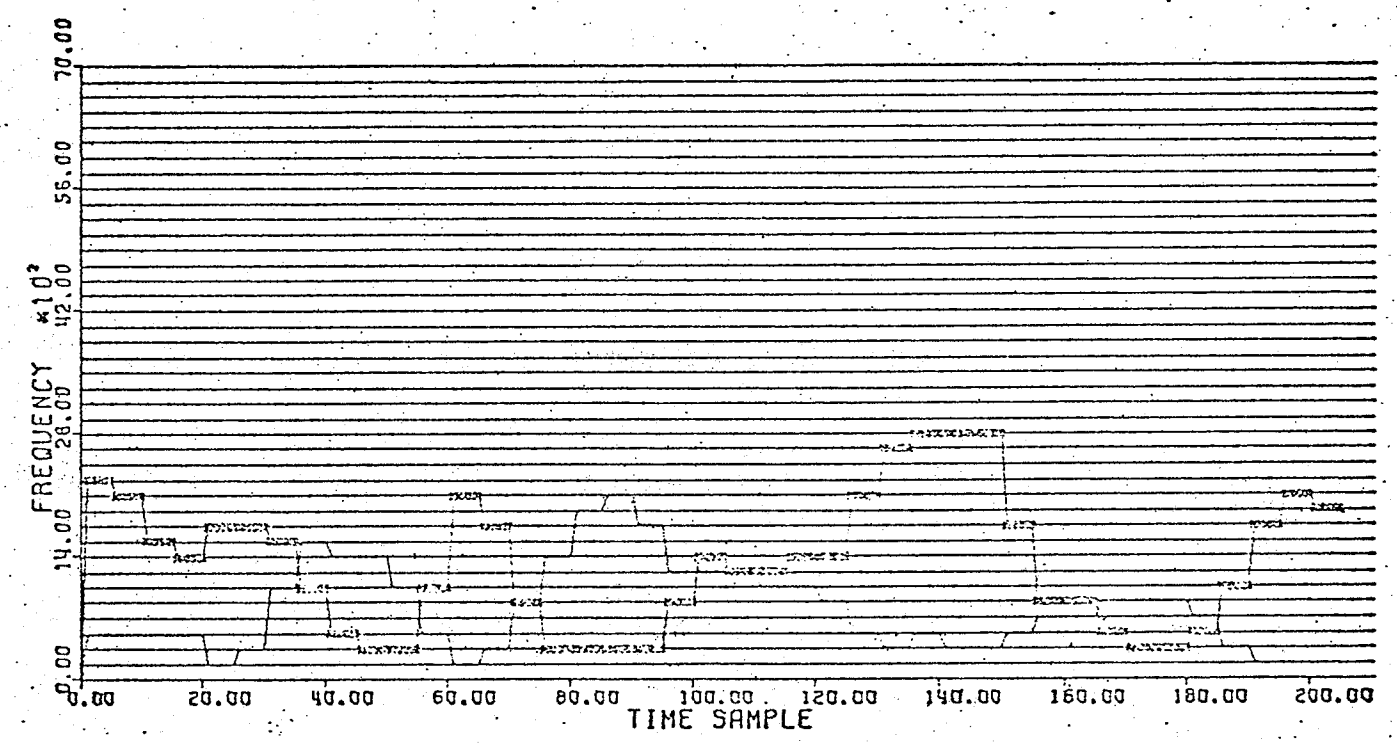
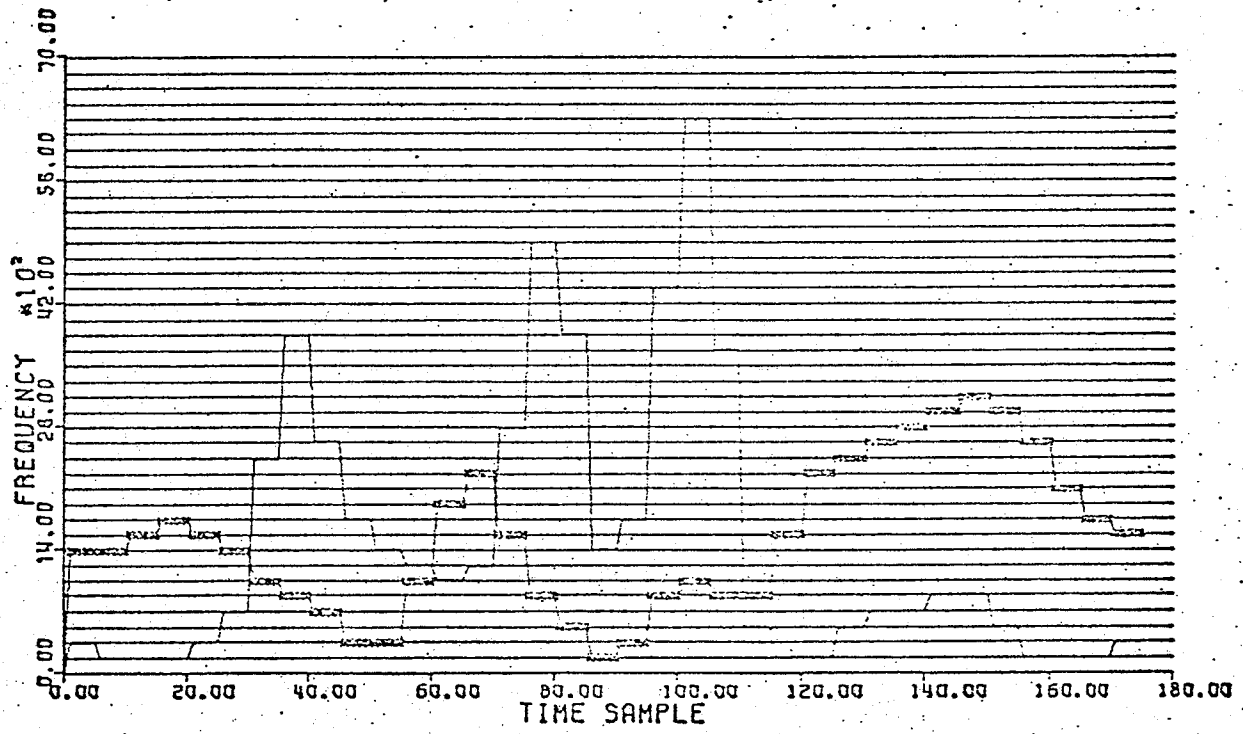


TABLE 4.5 (continued)

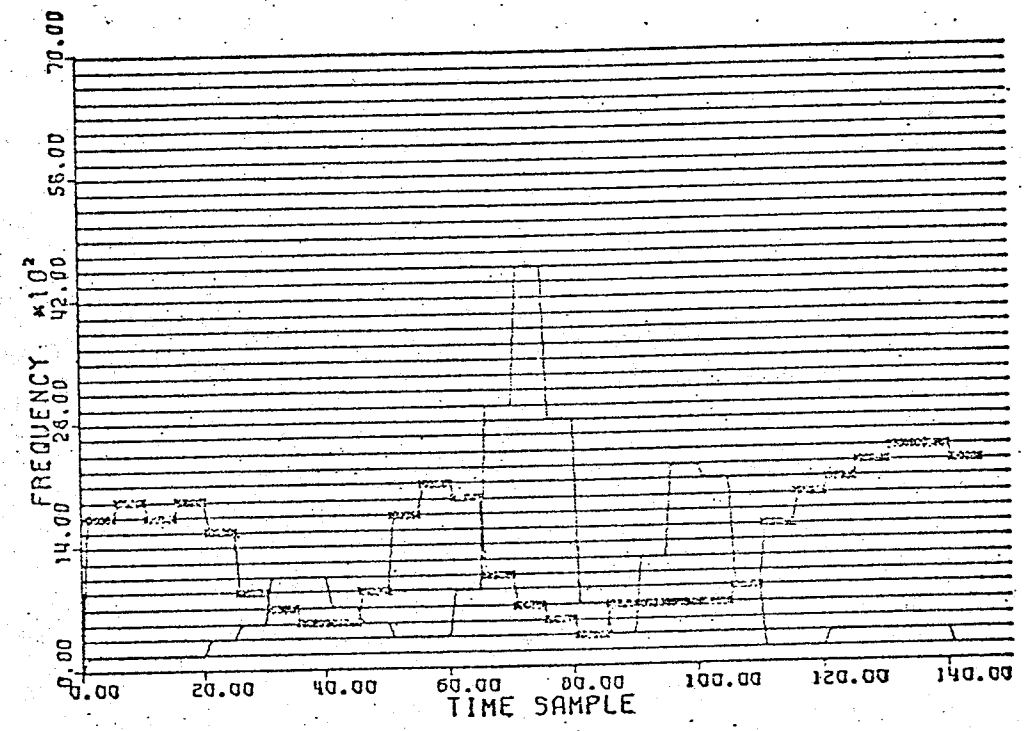
INFORMATION BY I.R.



EXHIBITS BY A.Y.



EXHIBITS BY D.C.



EXHIBITS BY I.R.

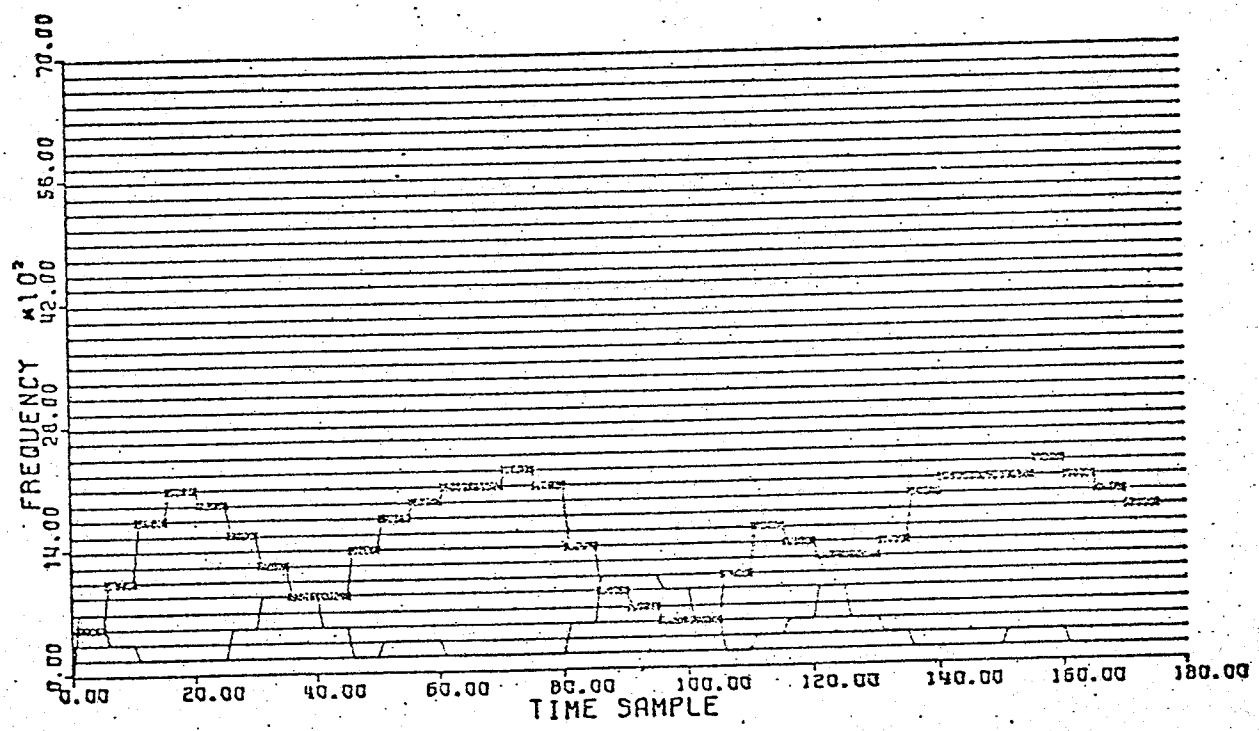
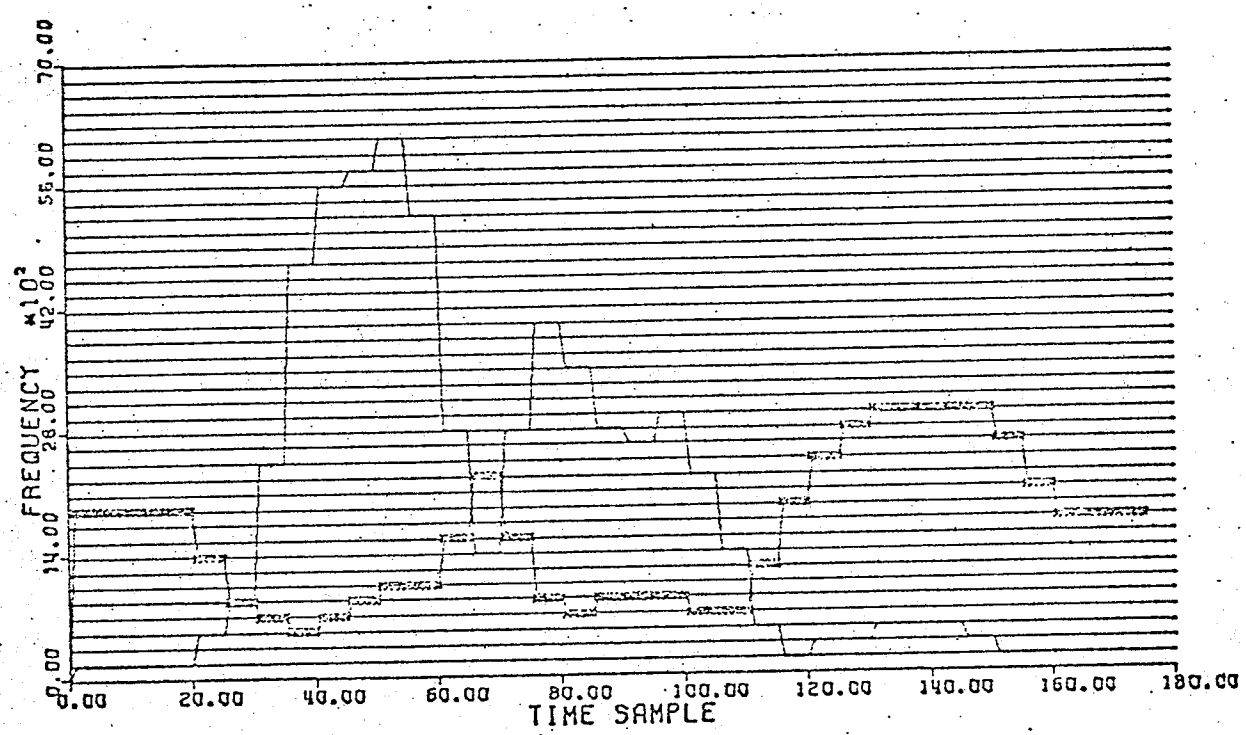
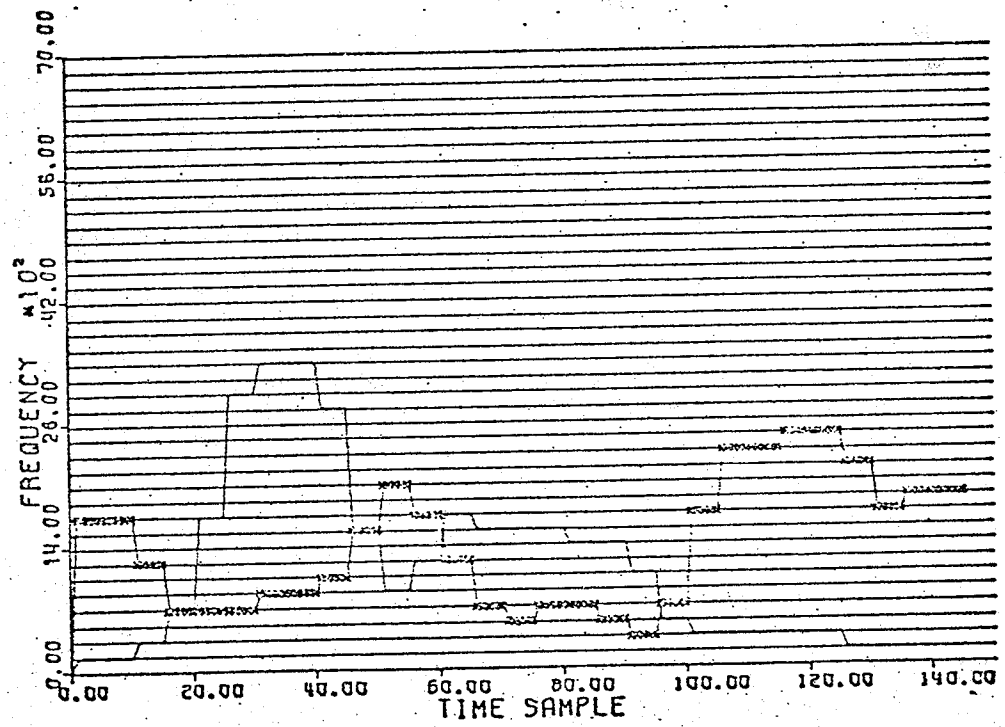


TABLE 4.5 (continued)

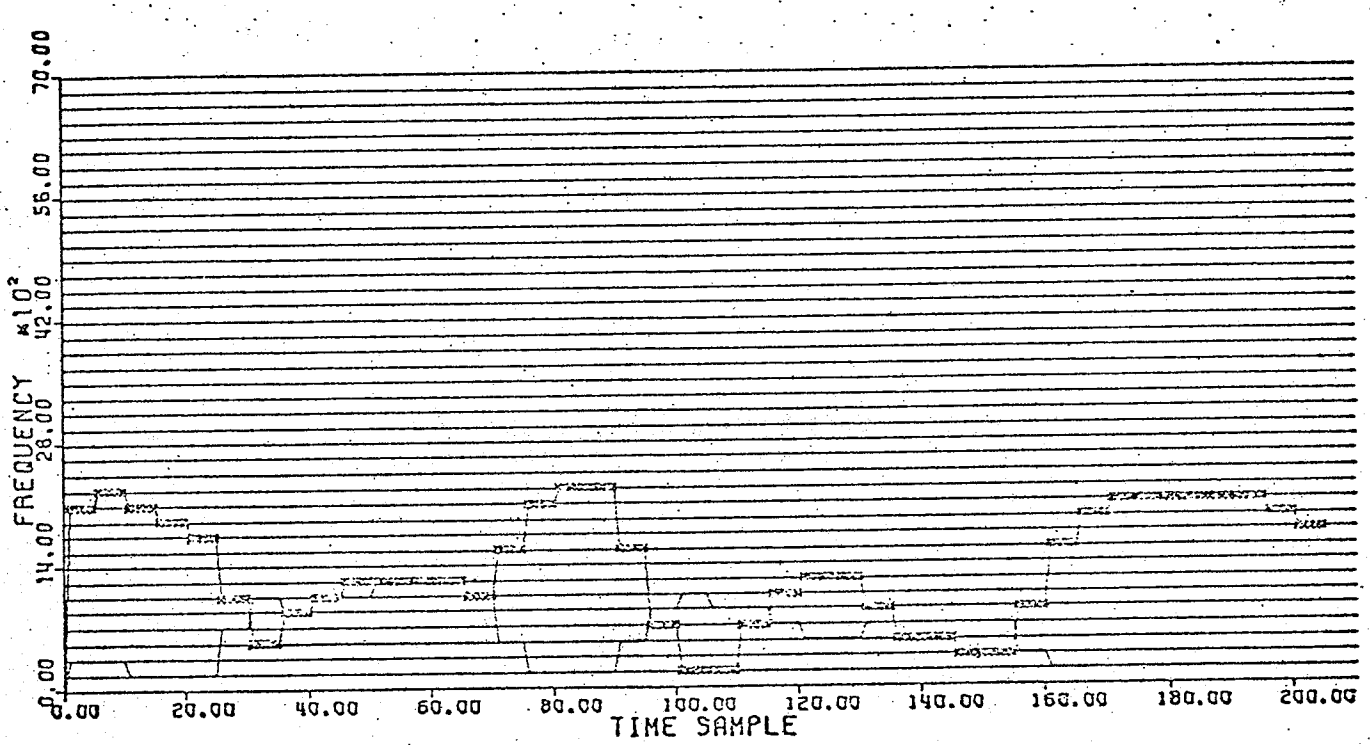
RESONANCE BY A.Y.



RESONANCE BY D.C.



RESONANCE BY I.R.



FLYING BY A.Y.

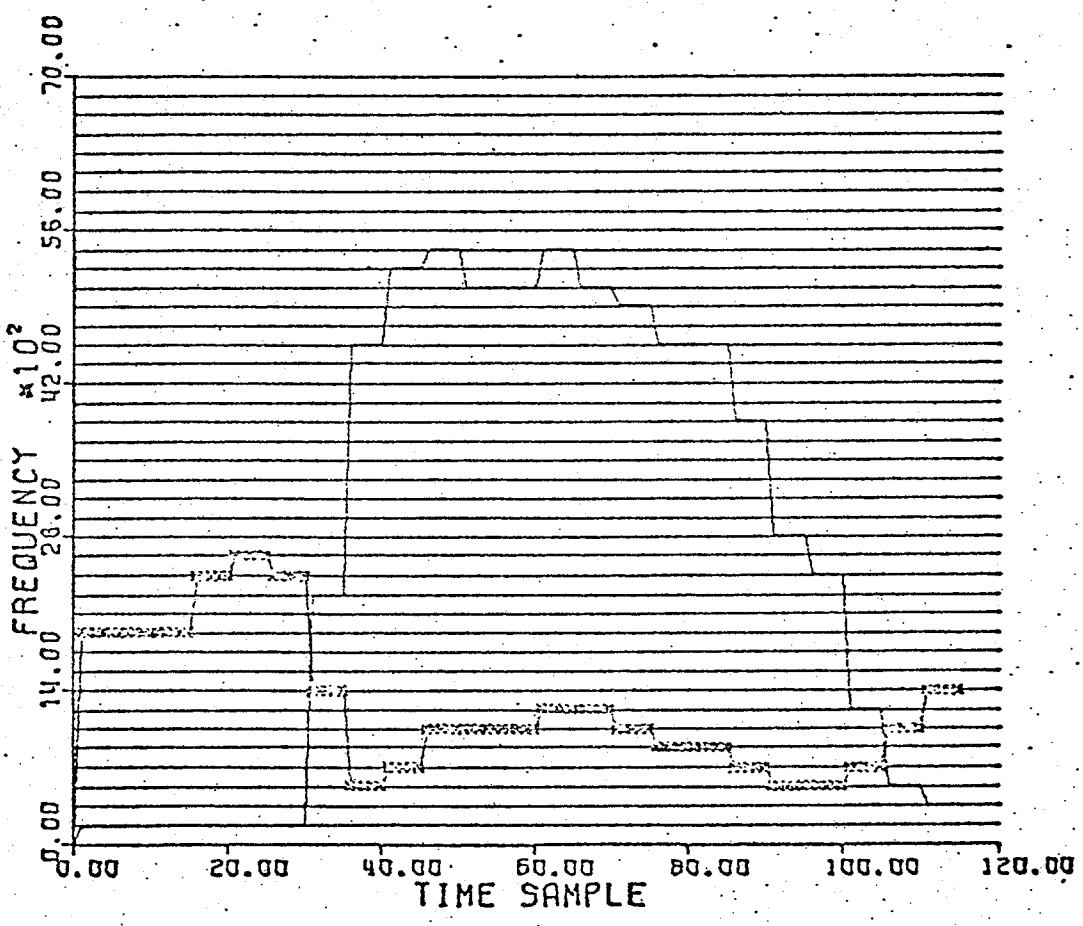
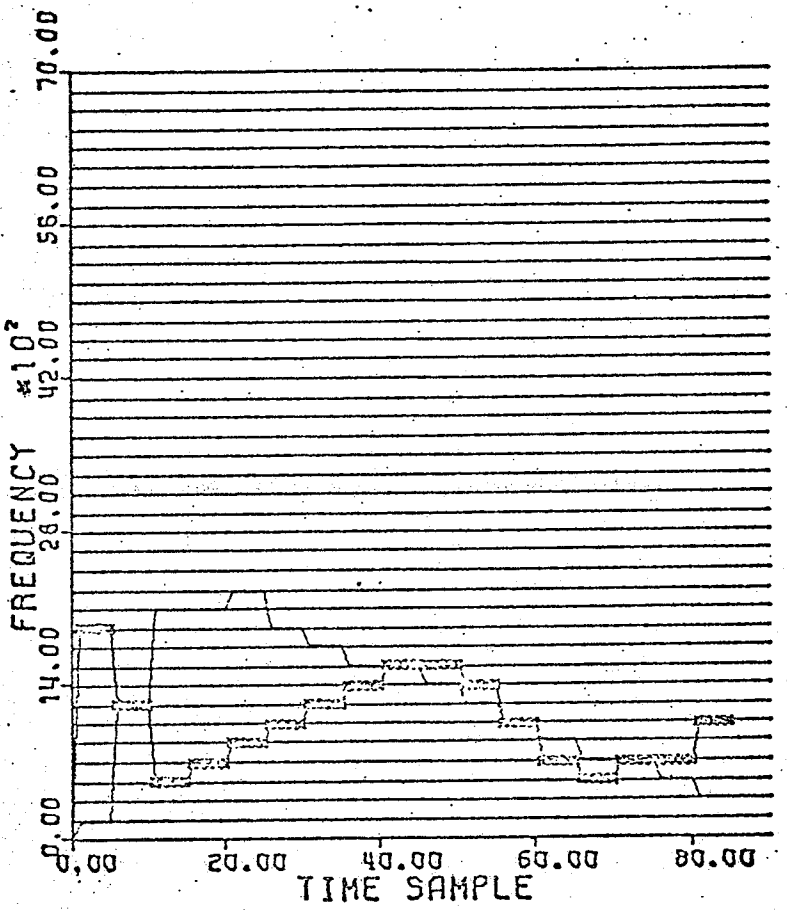


TABLE 4.5 (continued) FLYING BY D.C.



FLYING BY I.R.

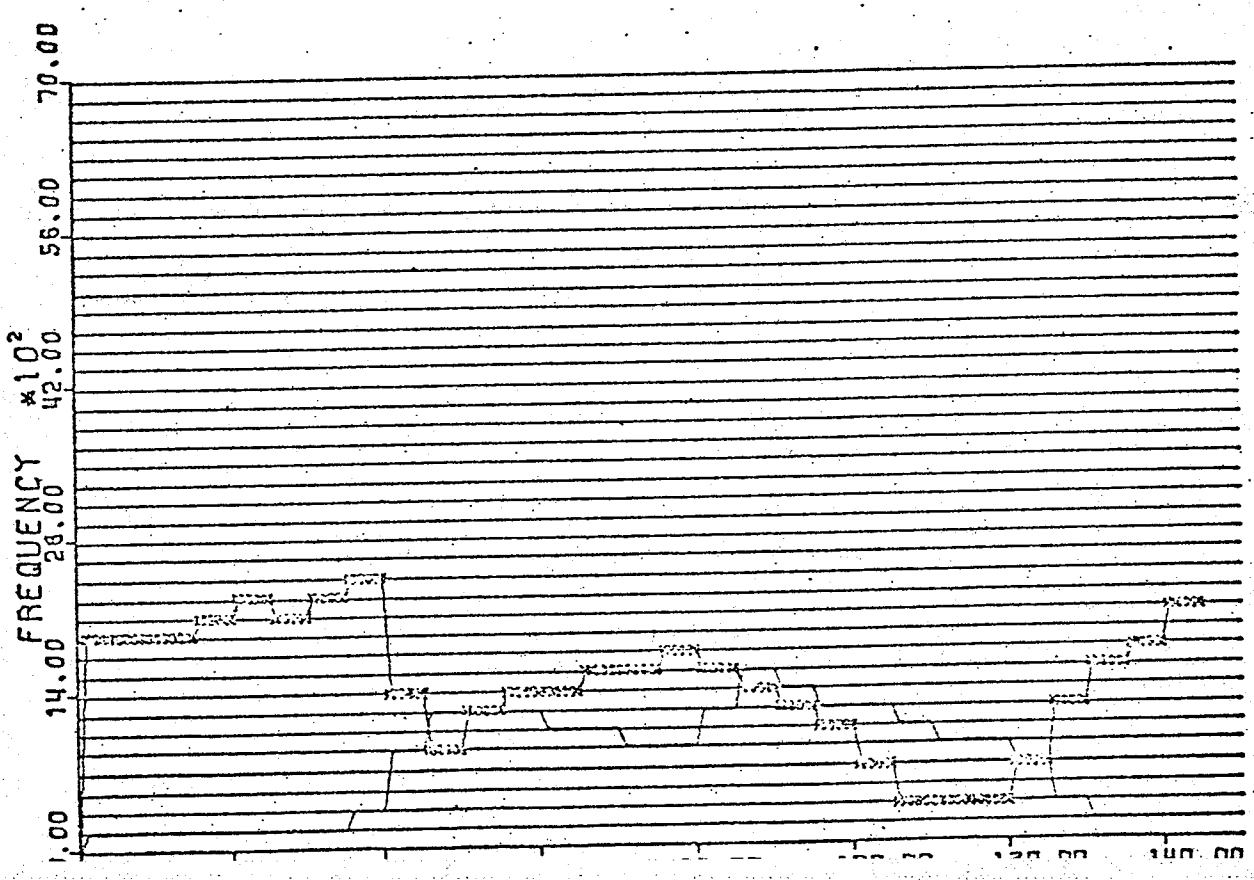
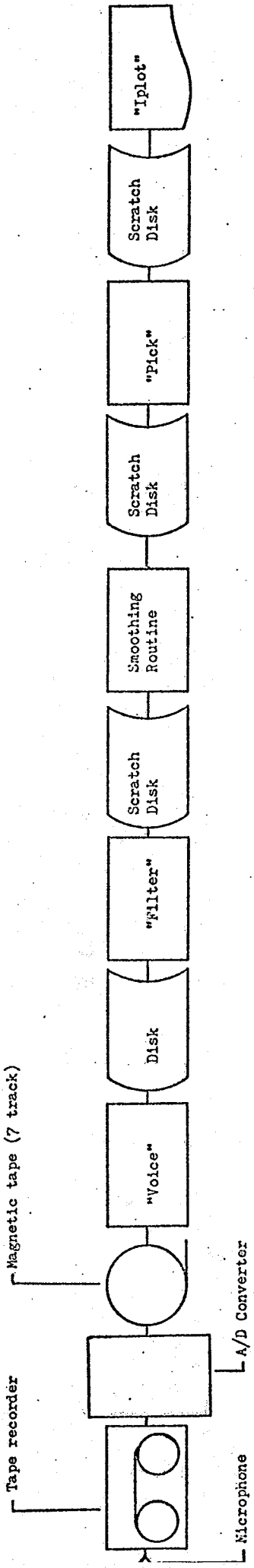


TABLE OF ELEMENTARY SOUNDS
WHICH OCCUR IN ENGLISH*

Phonetic Symbol	Key Word	Phonetic Symbol	Key Word
Vowels		Fricatives	
i	eve	h	he
I	it	f	for
e	hate	ϕ	thin
ɛ	met	s	see
æ	at	ʃ	she
a	ask	h	ahead
ɑ	father	v	vote
ɒ	not	ð	then
ɔ	all	z	zoo
o	obey	ʒ	azure
ʊ	foot	Plosives	
u	boot	p	pry
ɜ	work	t	to
ʌ	up	k	key
ə	about	b	be
Vowel-like		d	day
j	you	g	go
w	we	Affricatives	
l	let	tʃ	chip
r	read	dz	juice
m	me		
n	no		
ŋ	sing		

* The above Table (Potter, Kopp and Green, 1947) reflects American usage which is of course, more relevant to our investigation than the standard English pronunciation. It should be noted that in addition to these elementary sounds, there exist many more phonemes which have not been listed.


Tape recorder
Magnetic tape (7 track)
Microphone
A/D Converter
"Voice"
Disk
"Filter"
Scratch Disk
Smoothing Routine
Scratch Disk
"Pick"
Scratch Disk
"Iplot"

REFERENCES

1. REDDY, D.R., "Computer Recognition of Connected Speech", The Journal of the Acoustical Society of America.

2. HELMHOLZ, H. "On the Sensation of Tone" New York: Dover, 1954.

3. LINDGREN, N., "Machine Recognition of Human Language" JASA, March 1965.

4. WHITNEY, W.D., "A Sanskirt Grammar", Cambridge, Mass., 1941, pp 1-34.

5. STUMPT, C., "Die Sprachlaute", Berlin, 1926, p. 148.

6. GUNTER, R.T., "Early Science in Oxford", 6, Oxford, 1930, p. 57.

7. MILLER, D.C., "Anectodal History of the Science of Sound", New York: Mcmillan Co., 1935, p. 39.

8. RUSSELL, G.O., "The Vowel", Columbus: Ohio State University Press, 1928, pp. 6-8.

9. HALLE, M., "The Sound Pattern of Russian", Mouton & Co., 1959, p. 92.

10. DUDLEY, H. and TARNOCZY, T., "The Speaking Machine of Wolfgang von Kempelen", J. Acoust. Soc. Am., Vol. 22, 1950, p. 151.

11. PAGET, R., "Human Speech", London: Trübner & Co., Ltd., 1930, p. 19.

12. MILLER, D.C. "Anectodal History of the Science of Sound", New York: Mcmillan Co., 1935, p. 39.

13. HALLE, M., "The Sound Pattern of Russian", Mouton & Co., 1959, p. 96.

14. CRANDALL, I.B., "The Sounds of Speech", Bell System Technical Journal 4, 1925, pp. 586-626.

15. FLETCHER, H., "Speech and Hearing", New York: D. Van Nostrand Co. Inc., 1929, p. 400.

16. MOORE, C.R. and CURTIS, A.S., "An Analyzer for the Voice Frequency Range", Bell System Technical Journal, 6, 1927, pp. 217-229.

17. STEINBERG, J.C., "Application of Sound Measuring Instruments to the Study of Phonetic Problems", J. Acoust. Soc. Am., Vol. 6, 1934. pp. 16-24.

18. CHIBA, T. and KAJIYAMA, M., "The Vowel, Its Nature and Structure", Tokyo, 1941.

19. FLANAGAN, J.L., "Bandwidth and Channel Capacity Necessary to Transmit the Formant Information of Speech", J. Acoust. Soc. Am., 28, 1956, pp. 592-596.

20. BORROW, D.C., "A Limited Speech Recognition System", FJCC/68.