TEMPORAL CHARACTERISTICS OF SPOKEN CONSONANTS AS DISCRIMINANTS IN AUTOMATIC SPEECH RECOGNITION.

Thesis presented for the degree of Ph.D. in the University of Keele by P.D.Green.

ACKNOWLEDGEMENTS.

I would like to thank all those people who helped with the work described in this thesis, and with its subsequent presentation.

I am particularily grateful to my Supervisor, Dr. W.A. Ainsworth, for his constant encouragement and interest in the project, and to Drs. M.J. Underwood and J.B. Millar, who were previously members of the Speech Research Group. I am also indebted to those who gave their time to act as subjects, both in recording sounds and in listening sessions.

My thanks must also go to Stephen Hale, who was responsible for the photographic work involved in the illustrations, and to my wife Trudy and her mother, Mrs. L. Pankhurst, who typed this thesis.

The project was financed by the Science Research Council, with extra help from my mother, Mrs. J. Green.

P.D. Green,

December, 1970.

ABSTRACT.

Three time-varying functions, which can be extracted directly from the raw speech waveform, are of importance in the field of automatic speech recognition. These functions are the zero-crossing rate, the turnaround (local maximum or minimum) rate and the amplitude of the speech wave envelope. The aim of the work described here was to assess the feasibility of using these three variables to distinguish between the various consonant phonemes in English speech.

The investigation was confined to consonants spoken in isolated consonant-vowel syllables, with the consonant in the initial position. All the consonant phonems which occur in the initial position in English were spoken with each of ten vowel phonemes by four male speakers.

The three functions mentioned above were extracted from the speech wave by computer routines and displayed simultaneously using an on-line C.R.T. display. On these traces, the consonant part of the syllable could be readily distinguished by eye from that of the vowel, and the consonant was normally represented by a single peak on each trace. Further computer routines were evolved to identify these consonant peaks and extract recognition parameters describing the form of the peaks. Mistakes made by these programmes could be corrected manually from

observation of the display.

An attempt was then made to identify the consonant phoneme, using the values of the recognition parameters. The recognition algorithms took the form of modified binary threshold decision trees, and the task of designing these algorithms to fit new data was mostly automated.

Separate algorithms were constructed to recognise the utterances of each of the four speakers. For the appropriate speakers, the performances of these algorithms were very similar, about 65% of the utterances being classified correctly, with a further 25% of 'possibly' or tentatively correct identifications. The algorithms were, however, greatly speaker dependant, and performance fell off sharply when the speaker was changed.

The performance of the algorithms was independant of the vowel spoken after the consonant sound. For each speaker, satisfactory means were found to identify most of the consonant phonemes except the semi-vowel and nasal sounds.

Many similarities could be seen between the four recognition algorithms, and it was concluded that the speaker dependance might be reduced by the use of a different type of recognition algorithm coupled with normalisation of the recognition parameters.

CONTENTS.

contents continued) page number

LIST OF ABBREVIATIONS USED IN THE TEXT.

A.S.R.	Automatic Speech Recognition.
C.R.T.	Cathode Ray Tube.
C.V. Sounds	Isolated Spoken Syllables consisting of a Consonant followed by a Vowel.
db.	Decibels.
F0	Fundamental Frequency.
F1, F2 etc.	Formants (see Section I.5).
G1, G2 etc.	A 'Group' in a recognition algorithm (see Section 3.2.1).
Hz., KHz.	Hertz, KiloHertz.
I.	Intensity measure (see Section I.10).
I.L.E.	Recognition Parameters derived from the I. Trace (see Section 1.3)
I.S.D.	
I.P.D.	
I.P.S.	
I.P.W.	
n	Number of Points Constituting the Time Window used in the Smoothing Process (see Section 2.1.2).
s., ms., µs.	Seconds, Milli-seconds, Micro-seconds.
T.	Turnaround Rate (see Section I.9).
T.L.E.	Recognition Parameters derived from the T. Trace (see Section 1.3).
T.P.D.	
T.P.S.	
T.P.W.	

T.I.	Time Interval (see Section I.5).
t	Time.
t_c	'Counting Time' for T.I. rate measurements (see Section I.9).
U.V.	Ultra-Violet.
-ve.	Negative.
+ve.	Positive.
Z.	Zero Crossing Rate (see Section I.9).
Z.L.E. Z.P.D. Z.P.S. Z.P.W.	Recognition Parameters derived from the Z. Trace (see Section 1.3).
Z.T.I. Diagram.	Display of the time variations of Z., T. and I. (see Section 1.2).

INTRODUCTION.

Interest in speech research covers a wide range of disciplines and is by no means a recent development. Linguists have long been concerned with speech research as a means of discovering the underlying structure of a given language. Physiologists and Psychologists have investigated the speech communication process as a means of obtaining a better understanding of the working of the human nervous system. In the field of medicine, the aim has been to detect and correct speech defects. Likewise, Communication Engineers are concerned with improving the efficiency of speech information transfer and with bandwidth-reduction techniques.

The present study is largely concerned with automatic speech recognition (A.S.R.). Work in A.S.R. has been stimulated by the growing need for efficient man-machine communication . The development of a practical speech input (and output) device to a computer would be a great advance in this field. Speech recognition is one of the most complex pattern recognition tasks performed by a human being, and the human speech recognition apparatus is amazingly efficient and flexible, being able to function in the most adverse conditions. Since the work to be described is concerned with the question of what useful basic parameters can be extracted from speech sounds for A.S.R., it is hoped

that the results will be useful in understanding the human speech recognition process.

Before proceeding to outline the many problems found in the area of A.S.R., it is necessary to review the speech production process and the terminology of speech research.

I. 1. The Production of Speech.

The process of human speech production is described in great detail by Flanagan (21). Figure I.1. is a Schematic diagram of the important speech organs (the Vocal Tract).

The gradual outflow of breath necessary for the production of speech is created by contraction of the chest muscles. This air flow is modulated by pulsed abdominal contractions at the syllabic rate of speech.

The breath stream is modified by the action of the vocal cords (glottis), which consist of two lips of ligament and muscle. The vocal cords can be relaxed to allow the passage of air, or tightly stretched so as to close the vocal tract.

The modified stream of breath emerging from the region of the glottis passes through the vocal cavities and can be modulated in a great variety of ways to produce speech sounds. The nature of this modulation is determined by the shapes of the vocal cavities which in turn are dependent upon the

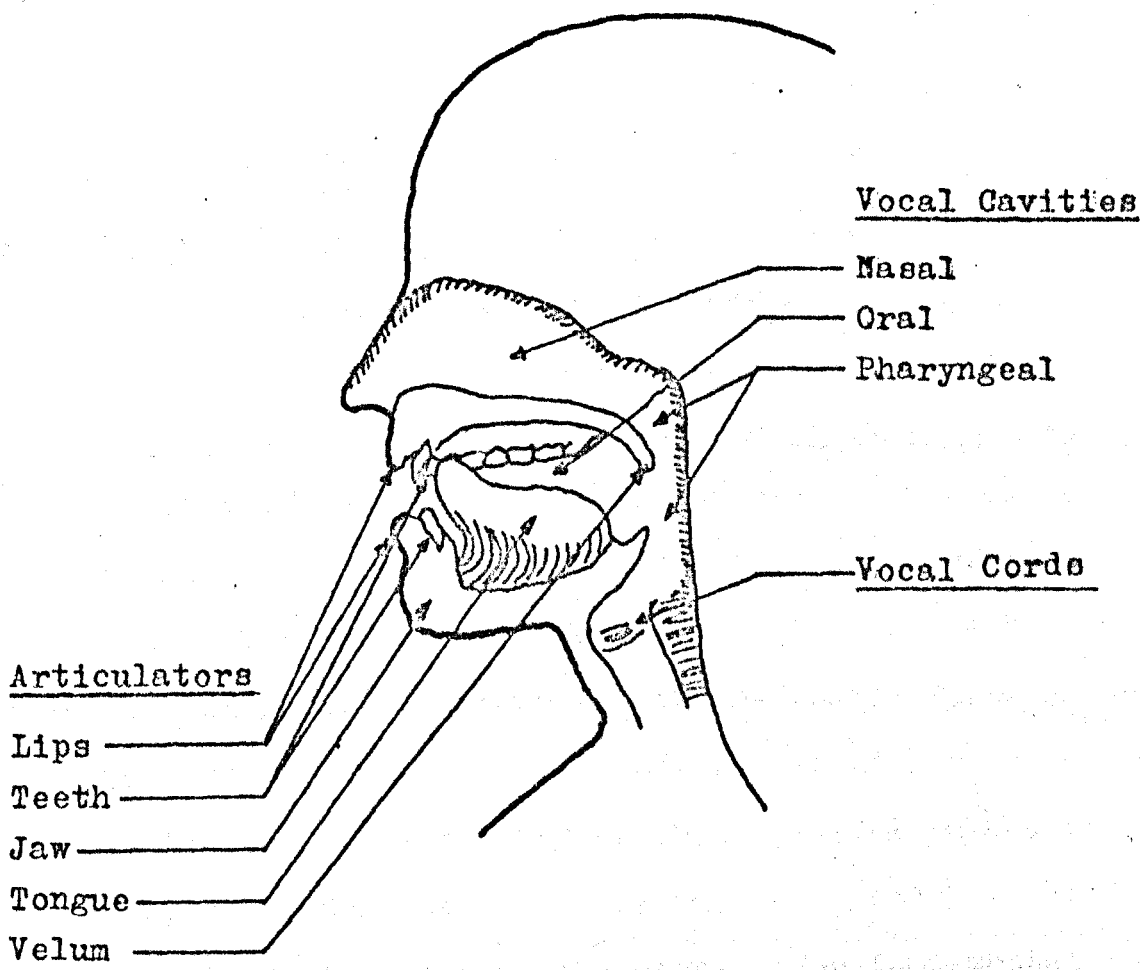


Fig. I.1 Schematic Diagram of the Vocal Tract.
(Adapted from Potter, Kopp & Green.)

positions of the speech articulators. These articulators are the lips, teeth, jaw, tongue and velum. Like the vocal cords, the lips can be used to close or partially close the oral cavity. The position of the velum similarly determines whether the nasal cavity is open or closed.

The sound pressure wave emerging from the glottis and entering the vocal cavities is termed the excitation; two distinct types of excitation are of importance in speech. Voiced sounds are excited by the vibrating action of the vocal cords. The vocal cords can be made to vibrate by a combination of glottal tension and the Bernoulli force caused by the passage of air, allowing short bursts of air to enter the vocal tract. The frequency of these glottal pulses is governed by the subglottal pressure and the glottal tension and is termed the fundamental frequency. The fundamental frequency determines the pitch of the sound.

The action of the glottal pulses sets the acoustic system above the vocal cords vibrating at its natural frequencies. The frequencies and amplitudes of these vocal tract resonances are governed by the shape of the cavities and are extremely important for the discrimination of voiced sounds.

In unvoiced (or voiceless) sounds, the vocal cords are relaxed and partially open. The sound pressure wave emerging from the glottis is noiselike and strong vocal tract resonances

do not occur, though the sound is still greatly influenced by the shape of the cavities.

I. 2. Sounds of Speech.

The articulators are capable of moving rapidly and to some extent independently to change the character of the sound produced. Only small changes in the vocal organs are necessary to produce a perceptual difference in the sound to a human listener. Of the vast number of possible articulations, only a restricted subset have meaning in any one language. These sounds are perceived by the listener as the mutually exclusive signals of the speaker's language, and are classified by linguists into categories called phonemes. The phoneme is defined as the smallest linguistic unit which can cause a change of meaning in a given language (7). The many distinguishable versions of a single phoneme are called Allophones.

Phonemes are not confined to static arrangements of the vocal tract, and a single phoneme utterance often corresponds to a sequence of articulatory events. A phoneme set can be defined for any language. A list of the English phonemes of importance in this study is given in figure 1.1. In English the phonemes can be divided into the following classes:

I. 2.1. Vowels and Vowel-like Sounds.

Vowel sounds are produced when the vocal tract is unobstructed. In normal vowel articulation, the excitation is purely voiced, the entrance to the nasal cavity is closed by the velum, and the remaining articulators are held more or less stationary. Vowels are continuant sounds (i.e. they may be sustained at will). Dipthongs are produced by moving the articulators continuously from one vowel position to another within the same syllable. Glides (/j/, /r/, /l/, /w/, /m/ and /n/) are made in a similar manner by moving the articulators smoothly from some initial position to the position for the following sound, the excitation being voiced throughout. Nasals (/m/, /n/ and /ŋ/) are produced by opening and closing the nasal tract by means of the velum and simultaneously closing the oral tract, again with voiced excitation. Since the period of time in which the nasal tract is open can be varied at will, nasals, like vowels, are continuant sounds. In English, the nasal port is normally closed except for these sounds.

I. 2.2. Fricatives (/h/, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʒ/ and /ʃ/).

In a fricative sound, a turbulent flow of air is created by means of one or more constrictions at some point in the vocal tract. The excitation may be voiced or unvoiced for

different phonemes. Fricatives are continuant sounds.

I. 2.3. Stops. (/p/,/b/,/t/,/d/,/k/ and /g/).

To produce a stop sound, a closure is formed at some point in the vocal tract, causing a short period of silence while voiced or voiceless breath pressure is built up behind this point. The vocal tract closure is then suddenly released, causing a short, sharp explosion of breath.

I. 2.4. Affricatives. (/tʃ/ and /dʒ).

Affricatives are a combination of Stops and Fricatives, in which the stop sound is exploded quickly while the articulators move into the fricative position. The excitation again may be voiced or voiceless.

The production of the individual phonemes is discussed more fully in Section 1. 4. Pitch inflections (in Chinese) and vocal clicks (South African Hottentot) are examples of speech sounds which are phonemic in languages other than English.

I. 3. The Frequency Spectrum of Speech.

The most widely used tool in speech research is the short time energy spectrum. This may be approximated by the outputs

of a filter bank, and is normally displayed visually by means of a Sonagraph (or Sound Spectrograph). The record of the spectrum of a speech sound produced by a sonagraph is called a Sonagram (Spectrogram). In this, time and frequency are plotted on linear scales horizontally and vertically respectively, and the darkness of the trace at any point is proportional to the amount of energy present at the corresponding time and frequency.

A classical study of the sonagrams of speech sounds was performed by Potter, Kopp and Green(44).The sonagrams of vowels and vowel-like sounds are characterised by several dark bars of energy corresponding to the natural resonances of the vocal tract. These energy bands are called formants. The formants are given numerical subscripts from low to high frequency, so that the fundamental frequency is termed FC ,(F.nought) the first formant F1 and so on.

Formants can also be observed in the sonagrams of the other voiced sounds, but the formant structure is generally less clearly defined than for vowels. The sonagrams of unvoiced sounds are characterised by the presence of a broad, dark region of energy (called a "fill" by Potter, Kopp and Green) due to the noiselike quality of the sound. Sonagrams of various consonant and vowel phonemes in isolated consonant-vowel syllables are shown in figures 1.4, 1.5, 1.6 and 1.7.

I.4. Problems in Automatic Speech Recognition.

No general revue of the literature on Automatic Speech Recognition will be given here; the reader is referred to the recent survey by Hyde (28). Flanagan's book(21) covers the whole range of speech research.

Despite considerable effort in the last twenty years, progress in the field of A.S.R. has been extremely slow. The difficulty of the speech recognition task can be seen in the extremely large number of **permissible variations** for each single phoneme. The acoustic representation of a phoneme varies considerably with the talker, speed of utterance, the adjacent phonemes and so on. Quite different sounds, for example, due to dialect differences, may occur for the same phoneme. The speech waveform is extremely redundant; it has been estimated that a saving of 1,000 : 1 in channel capacity would be gained by transmitting the phoneme string instead of the original spoken message. Human beings are able to exploit the redundancy in the speech wave by the use of a hierarchically organised recognition system which functions on many levels. This information about physical and linguistic constraints, knowledge of the vocal characteristics of the speaker,

the content of the message and many more perceptual clues are utilised to overcome the inherent ambiguities of speech sounds. The construction of a system with this sort of capability is at present beyond the limits of scientific knowledge, and it is likely that the development of a general purpose A.S.R. device will have to wait until the human perceptual processes are better understood.

For these reasons, only partial solutions to the speech recognition problem have been sought, and generally only limited success has been achieved. Four factors which are involved in the development of all A.S.R. systems are of interest here. These factors are by no means independent of each other.

I. 4.1. Choice of the Recognitions Task.

In many A.S.R. applications, it is possible to restrict the number of allowed articulations to a small subset of the whole range of speech sounds. The use of isolated words greatly simplifies the difficult problem of the temporal segmentation of the speech signal, and such word recognisers have many practical applications. Further reduction to the level of syllables or phonemes spoken in isolation is very useful for the evaluation of recognition techniques, though it is rarely possible to extend the work directly to connected

speech, since phoneme or syllable representations often differ greatly between the isolated and connected situations. The recognition problem can also be simplified by restriction of the vocabulary, for instance to the ten digits or to a small set of command words. In the very simple case of isolated digit recognition, a marketable device has been designed(38) and is being evaluated for automatic parcel sorting. In some instances, the vocabulary can be reduced by eliminating certain phonemes which are difficult to recognise. Restriction to a single speaker enables substantial improvement in the recognition scores, but it is largely invalid to generalise from such results since speaker differences are so great. One application where the single-speaker approach is valid is the astronant manoevering unit reported by Herscher et al (26).

I. 4.2. Choice of the Recognition unit.

Recognition at the phoneme level is the most obvious approach, and some degree of phoneme recognition is essential in a general A.S.R. device. However, since the phoneme is a unit of perceived quality, it may not always be possible to segment the speech signal directly into phoneme units. The human listener undoubtedly makes use of recognition at the level of syllables, words and higher linguistic units, and all these units must be utilised by a general purpose speech

recogniser. Recognition schemes to achieve this have recently been proposed by Fant (19) and Zagoruikol (57). In some restricted applications, such as digit recognition, syllable or word patterns may suffice without recourse to phoneme recognition.

I. 4.3. Choice of the Recognition Parameters.

By far the most popular transform of the speech wave used in A.S.R. is the frequency spectrum. There is abundant evidence that frequency information is used in the human recognition process, but it is unlikely that anything closely corresponding to the sonagram is extracted (4).
The recognition of isolated vowel sounds by means of formant frequencies and amplitudes is relatively simple (22, 25), but the many ambiguities seen in the sonagrams of connected speech limit the success of the spectral approach. Thus vowel phonemes are characterised by areas in an F1-F2 space, but these areas overlap, and their boundaries are dependent on speaker, context, etc. (3 , 8 ,20).
Martin et al (38) have shown that formant slopes, and the regions of increasing and decreasing energy, are in some respects more useful. than the formant values themselves.
The present work is largely concerned with the development of other parameters to supplement spectral information.

I. 4.4. Design of the Recognition Algorithms.

Improvement in the design of speech recognition algorithms is closely linked to progress in the fields of pattern recognition and artificial intelligence. Template matching techniques, in which the incoming pattern is compared with a set of stored patterns to find the best match, have been widely used. This method has the advantage of being relatively simple to implement, and the ease with which the stored patterns can be changed means that the system can be adapted easily to cope with changes in language, speaker, recognition unit, etc.. However, it is often necessary to distinguish between patterns which are largely similar but differ in some important detail. The presence of prominent but irrelevant data can easily mislead the recogniser, which will respond to the stored pattern with the greatest overall correlation.

Another method involves the extraction of Distinctive Features which are known to be characteristic of the speech sound elements. This approach was first suggested by Fant(18), who segmented the various phonemes by a set of binary oppositions such as voiced/unvoiced, nasal/oral, tense/lax,etc.. With a scheme of this type, it is possible to exclude irrelevant data, but the feature extractors are far from easy to construct, and the system is much less flexible than a template matching approach. It is likely that the best approach would involve

a compromise between pattern matching and feature extraction.

Self-adaptive (learning) techniques have been used in A.S.R. systems with some success. "Adaline" networks (adaptive threshold elements) for which training theorems have been evolved (41) have been used in self-adaptive A.S.R. devices by Talbert et al (52) and Damman (12). At present, however, learning systems cannot be exploited fully since the basic principles of recognition are not fully understood.

The work reported here is concerned with evaluating the usefulness of some parameters which can be extracted directly from the speech waveform in the time domain. The study of the speech waveform was urged by Dudley (15) and Chang (11), who pointed out that much useful information can be extracted from the speech waveform without recourse to extensive filtering systems. Reddy (46) has devised a phoneme recogniser for use with connected speech, in which spectral information is used only in the later stages. His results (81% correct phoneme recognition for a single speaker) compare favourably with those of other workers. The temporal characteristics of interest here are time interval and envelope measurements.

I.5 Time Intervals.

Two important time interval (T.I.) measurements can be made directly from the raw speech waveform. The central trace of figure I.2 is a drawing of a portion of the waveform for a typical vowel sound (/ɜ/). The variations in air pressure are plotted against time, and the segment comprises about 1½ glottal periods, a duration of 10- 15ms. The vowel waveform is of a quasi-stationary nature, and 'repeats' after each glottal period. Within a single period, the waveform is dominated by F1, and F2 and the higher formants appear as a fairly irregular ripple. This type of waveform occurs only when the formants are widely spaced in amplitude and frequency.

T.I. measurements are based on the identification of the zero crossing and turnaround points of the speech waveform. Zero crossings may be defined as the points in time where the speech wave crosses the zero axis of air pressure, while turnarounds are the local maxima and minima of the speech wave. Since turnarounds occur at points of zero slope, the turnaround points correspond to the zero crossings of the differentiated speech wave.

Infinately clipped speech (usually shortened to clipped speech) is made by constructing a square wave which changes polarity at the zero crossing points of the original speech wave. Differentiated and clipped speech is constructed in a

similar manner from the turnaround points (see figure I.2). Both forms of clipped speech thus discard all the amplitude information contained in the original speech waveform. Interest in T.I.'s originally stemmed from experiments on the perception of clipped speech. This work is reviewed in the following section.

I.6 Perception of Clipped Speech.

Clipped speech was the most severe type of amplitude distortion investigated by Licklider (34). Following this work, the classical study on the perception of clipped speech was conducted by Licklider and Pollack (35). Three basic circuits- an integrator, a differentiator and an infinate clipper- were linked to process speech in various ways.

The major finding from these experiments was that the intelligibility of clipped speech was only slightly less than that of normal speech, although the clipped speech was harsh and unnatural in quality. Differentiation of the speech wave prior to clipping gave slightly better intelligibility scores than clipping alone, while integration prior to clipping reduced the intelligibility drastically. The intelligibility of all forms of clipped speech was not greatly effected by further integration or differentiation after the clipping process.

The work of Licklider and Pollack was extended by

Ainsworth (1) to take account of the polarity of the zero crossings (+ve. to -ve or -ve to +ve.). Ainsworth found that the intelligibility of clipped speech was improved if the distinction between +ve. and -ve. going zero crossings was maintained, though the shape of the pulses signaling the occurence of zero crossings was relatively unimportant. The data for this experiment consisted of phonetically balanced (P.B.) words (16), and Ainsorth was able to show that while the intelligibility scores for individual phonemes did not differ greatly, the vowel-like sounds were less often confused than the fricatives.

Further work by Licklider (36) dealt with the effect of quantising the T.I.'s between zero crossings. A reduction from 20,000 to 10,000 quanta per second produced a drop from 96% to 91% in the articulation scores, which decreased rapidly on further reduction of the number of quanta.

Tanaka and Okamoto (53) also investigated some forms of T.I. distortion. In an attempt to combat noise, these workers used a 'slice level' about which to clip, rather than zero. Syllable articulation fell from 90% for a slice level at -50db. below the speech level to 75% at -30db and 45% at -20db. When the negative-going edge of a differentiated and clipped speech waveform was allowed to vary randomly in time subject to various maximum delays, syllable articulation with a slice level at -40db fell from 98% for a 1μs delay

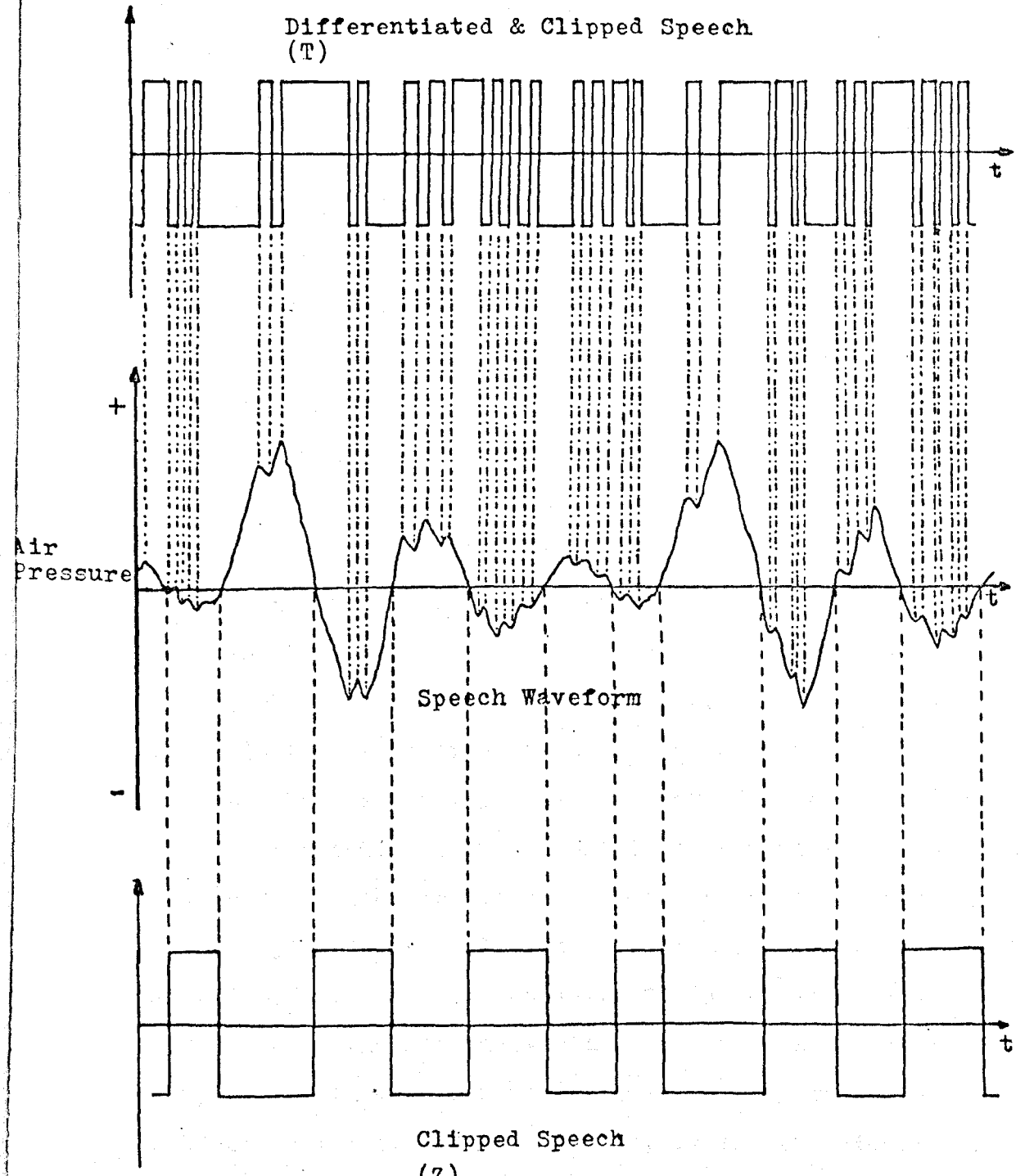


Fig I.2 Time Intervals.

to 84% for 10μs. and 55% for 100μs.

This perception work showed that the amplitude information in a speech wave may be discarded completely without seriously impairing its intelligibility, provided that the zero crossing or turnaround sequence is conveyed fairly accurately. This fact is a remarkable tribute to the flexibility of the human speech analysis system, and illustrates the extreme redundancy of the speech wave. In connected speech, the clipping process preserves the basic temporal sequence of events in the original wave, and hence any perceptual clues present in this ordering. The identification of isolated clipped vowels is much more difficult than that of connected clipped speech. (1,55).

These experiments on the perception of clipped speech have led to much work on the analysis of T.I. information, and to attempts to use T.I.'s in A.S.R. This literature is reviewed in the following sections.

I. 7. The Relationship between T.I.'s and The Frequency Spectrum.

The surprisingly high intelligibility of clipped speech is partially due to the retention of much of the spectral information in the original waveform by the clipping process. In figure I.2., it appears that the zero crossings are largely governed by F1, while most of the turnarounds are due to F2. This relationship was first postulated by Chang (10). However, this direct inverse relation between T.I.'s and Formants only holds in simple cases. Generally it has been shown that the dominant frequency of the speech wave remains the dominant frequency of the clipped (zero crossing) version (23, 56). When two formants are close enough in intensity (within 5 db.) so that neither dominates the speech wave, both formant frequencies are apparent in the clipped speech spectrum (23).

Sonagrams of clipped speech confirm that the clipping process preserves most of the important frequency componants, though some formants, especially at high frequency, may be lost, and much narrowing or broadening of the formants occurs.

The effect of the rectangular edges of the clipped speech wave is to produce inter-modulation and spurious high frequency componants due to the harmonics of the square wave. As a result, sonagrams of clipped speech have a blurred appearance; examples of such sonagrams are given by Underwood (55).

The spectrum of clipped speech was utilised by Licklider et al (35), who attributed the discrepancy between the intelligibilities of clipped, and differentiated and clipped speech to the closer approximation which sonagrams of the latter showed to those of the original speech wave. Similarly Ainsworth (1) explained his results for different transforms of clipped speech by showing that the transforms which were more difficult to recognise had destroyed the most spectral information. Thomas (54) was concerned with the importance of F2 in speech perception. He pointed out that the higher intelligibility of differentiated and clipped speech might be due to an emphasis of F2 information.

Chang et al (10) devised a T.I. display on an oscilloscope equivalent to the sonagram, in which the frequency dimension was replaced by the inverse of the T.I. length, and the brightness of any point measured the frequency of occurrence of a T.I. of the appropriate length in the appropriate time segment. The display was found to emphasise F1 when zero crossings were used and F2 for turnarounds. The two displays

were superimposed to form the "Intervalgram". While the F1 and F2 variations could generally be picked out by eye on this display many spurious tracks not related to formant movements were also observed.

The simplicity of Chang's device illustrates the great advantage of T.I. measurements. The filter banks needed to extract frequency spectra are complex and costly, and great care is needed to counter the bandwidth-dependant delays in the responses of individual filters. T.I. measurements, on the other hand, can be extracted by simple circuits or computer routines, and there is no delay problem.

There have been many attempts to extract spectral information by means of T.I. measurements. One way of improving the relationship between T.I.'s and Formants is to perform a coarse pre-filtering of the speechwave prior to the clipping process. Sakai and Doshita (49) used zero crossing information from the outputs of two filters chosen to match the range of the first two formants. A similar approach has been used by many workers (9 ,42,43,45, 5). Peterson (42) showed that the correlation between the mean zero crossing rate and formant frequencies could be much improved by this technique, but the measurement was still inaccurate under certain conditions. Though efficient formant trackers can be constructed in this way (9), the compromise between spectral

and T.I. methods negates some of the merit of direct T.I. measurement by destroying the simple temporal relationship to the original wave.

Another method of refining the measurement of T.I.'s in vowel sounds is to make use of the fine structure of the speechwave within individual glottal periods. In figure I.2., it is apparent that the influence of F1 on the T.I.'s is greatest for the first few intervals after the glottal pulse, and becomes less marked in the later portions of the glottal cycle. The fine structure of T.I's has been analysed mathematically for two simple cases by Scarr (50). For a fixed larynx frequency and a variable single resonance, and vice versa, Scarr found that first two T.I.'s after the glottal maximum gave the best formant measurement; while for two resonances of equal amplitude at the third and fourth harmonics of the larynx frequency, these T.I.'s gave a good measure of the mean of the two resonant frequencies. Scarr was able to confirm these findings by experiment.

The fine structure of vowel sounds was also utilised in the parallel studies by Underwood and Millar (55,40). Both these authors improvedtheir results by using pitch synchronous gating to ignore the latter portions of the glottal cycle. Stover(51) indicated that only the first 3 ms. of the glottal period are necessary for perception. Later work by Underwood et al(33)

involved the measurement of both F1 and F2 by similar techniques. Using synthetic speech Lavington (32) has shown that for constant larynx frequency and energy, the number of turnarounds is in roughly inverse proportion to the average of F2 and F3.

I. 8. T.I. Analysis of Vowel Sounds.

The first attempt to identify phonemes by means of T.I. information was due to Sakai and Inoue (48). The basis of this work was the probability distribution of T.I.'s, the T.I. histogram. Histograms for both zero crossings and turnarounds were produced for five Japanese vowels. Generally one or two peaks were found in the histograms and the positions of these peaks afforded some distinction between the vowels without being reliable indicators of the formant frequencies. The following work by Sakai and Doshita (49,14), in which two frequency filters were used to give T.I. histograms for both the FI and F2 regions, achieved 94% correct recognition of the five Japanese vowels for a single male speaker and 90% for a single female speaker.

Bezdel (5, 6) obtained slightly better results than Sakai and Inoue in a similar study of English vowel sounds. Again, the T.I. histogram was used, and the results were comparable to those obtained by spectral analysis (22).

The most thorough study of T.I. measurements in English vowel sounds was that of Millar (40) and Underwood (55). The aim of this work was to assess the merit of the T.I. approach, and therefore T.I. information was used alone, without recourse to pre-filtering. Millar was largely concerned with vowel identification, while Underwood attempted to synthesise clipped vowels and voiced phrases from T.I. statistics. The 1st. order T.I. statistics (T.I. histograms) were found to be of little use for identifying or synthesising the 12 English vowels. Since these measurements pay no attention to the order in which T.I.'s occur, the 2nd. order T.I. statistics were examined by means of diagram extraction and display apparatus, and both vowel recognition scores and the perception of synthesised clipped vowels were found to be improved by this means. Millar found that the 12 vowel phonemes could be separated into four groups fairly reliably, but it remained difficult to distinguish between individual vowel phonemes. Underwood further extended his synthesis work to trigram(3rd order) statistics, and found that clipped vowels synthesised from this data were nearly as intelligible as the original clipped speech. The T.I. statistics were found to be influenced markedly by pitch changes, and this was counteracted to some extent by the use of pitch synchronous gating. Millar and Underwood concluded that the ordering of T.I.'s was of great importance for speech

analysis and perception.

I.9. T.I. Analysis of consonant Sounds.

The T.I. histograms of some consonant phonemes were examined qualitatively by Millar (40). Only the continuants, fricative and nasal sounds, were considered. A fair amount of visual discrimination was possible between the histograms. The noiselike quality of voiceless fricatives led to a single, wide peak composed of very short T.I.'s, while for the voiced fricatives additional peaks due to T.I's of a much longer duration were observed. The positions and shapes of these peaks afforded some distinction between the individual phonemes. The T.I. histograms of nasals were dominated by the effect of the fundamental and nasal frequencies, with very long T.I's predominating. The T.I. histograms of the voiceless fricatives /s/ and /ʃ/, the nasals /m/ and /n/ and the voiceless stop /k/ were also examined by Sakai and Inoue (48). The histogram for /k/ was shown to be dependent on the vowel following the stop sound. T.I. histograms in conjunction with pre filtering and duration measurements were used fairly successfully for consonant recognition by Doshita (14), who reports 70% correct recognition of unvoiced consonants.

Since consonants do not generally have the quasi-stationary properties of vowel sounds, it is difficult to extend the sort

of T.I. statistics used by Millar and Underwood to cover all the consonant phonemes. The fine structure between pitch periods is non-existent in unvoiced sounds, and is less clearly defined in voiced consonants than in vowels (13). The most useful T.I. measurements which are easily applicable to the whole range of speech sounds are the T.I. rates. These rates may be expressed as the number of T.I's occurring within a short "counting time", tc, roughly equivalent to the shortest time (one glottal cycle), in which the speech waveform is capable of changing significantly. These functions will be termed Z. (zero crossing rate) and T. (turnaround rate), after Lavington (32). Time interval rates were first used by Chang (10,11).

Zero crossing measurements played an important part in the general A.S.R. system devised by Reddy (46), who calculated the number of zero crossings and the standard deviation of this number occurring within a speech segment which was to be associated with a single phoneme. In this process, the speech segment was first associated with one of four phoneme groups, "vowel-like", "fricative-like", "nasal-liquid-like", and "stop-like". Fricative-like segments were identified partially by the presence of a large number of zero crossings, and nasal-liquid-like segments had a very small number. Further classification to individual phonemes was largely accomplished

by spectral measurements, but zero crossings were used in the separation of fricative-like sounds. The standard deviation of zero crossings was mainly used in connection with the preliminary identification of "sustained" and "transitional" parts of the speech wave.

Lavington (30,31,32) has obtained promising results using a Z/T measurement space, with tc = 10 ms. It is possible to associate some vowel and consonant phonemes with separate and distinct areas in the Z/T space. 95% accuracy was obtained with a small (16 word) vocabulary of isolated syllables over a large number of speakers, though the range of phonemes was very restricted. Lavington aimed to develop this work towards a spoken digit recogniser, and was able to distinguish between all the initial consonants occurring in the 10 digits, provided no distinction was made between /f/ and /θ/.

I. 10. The Amplitude Envelope.

Although experiments on clipped speech have shown that the removal of all the amplitude information in the speech wave does not greatly effect the intelligibility to the human ear, amplitude information can be used with advantage for A.S.R. purposes. A brief study by the author of amplitude modulated clipped speech, in which the original speech envelope was retained, indicated that the intelligibility of isolated

vowels was improved slightly, and that the quality of these vowel sounds was more "natural".

Speech envelope variations, like T.I's, can easily be extracted from the original speech waveform. The envelope function used in this thesis is given by the highest value amplitude modulus found in a single counting period of length tc., and will be called I.(intensity). A similar measure was used in later stages of the work by Lavington (30,31).

Amplitude measurements played an important part in the work of Reddy (46). Together with zero crossing information, intensity measurements enabled the speech wave to be divided into segments which could be identified with a single phoneme. Both mean value of the intensity of a segment and the standard deviation of this parameter were used. Vowel-like segments were found to have the highest intensity values and fricative - like the lowest, with the nasal-liquid-like sounds forming an intermediate class.

Other A.S.R. systems which make use of amplitude information were reported by Gold (25), Talbert et al (52) and Damman (12).

I. 11. Outline of the Present Study.

The aim of the work presented in this thesis was to investigate systematically the properties of the three temporal parameters Z. T. and I. in consonant sounds, in order to

evaluate their potential for A.S.R. purposes.

An efficient A.S.R. device would undoubtedly make use of some compromise between the frequency and time domain approaches, in order to obtain the best of both worlds. In order to effect the best compromise, it is necessary to know the relative virtues of both methods when used alone. While the properties of T.I's in vowel sounds have been thoroughly investigated (40,55), and compromise solutions have been used to good effect (9 ,14,33), the position is by no means clear for consonant sounds, though the work reviewed in the preceeding section shows much promise. It is clear that the traditional "sonagram" approach is less well suited to consonant sounds than to vowels, since formant patterns are less clear or non-existent, and context dependence is increased.

This work therefore follows that of Millar (40) and Underwood (55) in relying on a high signal to noise ratio (49 db), rather than the use of filters to counteract noise, in order to obtain direct temporal transforms; and also in being based exclusively on individual phoneme recognition. Any flexible, large vocabulary device must be able to distinguish between phonemes to some extent: this is, of course, true of human listeners (21,39). It was hoped that the addition of the amplitude parameter (I.) would enable distinction between the whole range of English phonemes.

In Reddy's work (46), the envelope measure was used mainly to make broad distinctions between phoneme classes.

The availability of a versatile, high speed computing system based on a Digital P.D.P.8 and 338 programmed display (see appendix 2) made it possible to examine the variations in time of the three functions Z., T. and I. simultaneously for a single utterance. Once the consonant representations on the Z., T. and I. traces had been identified, various recognition parameters were extracted from each trace. This parameter set was then used in an attempt to recognise the phoneme. All the processing was done on the computer, though many of the operations, especially in the earlier stages, could just as easily have been accomplished using hardware. Since it was not intended to construct a practical recognition device, no attempt was made to make the parameter extraction process fully automatic. A compromise was reached between the processing time saved by better programmes and the time taken to develop these programmes. Manual correction of programme errors was greatly simplified by the flexibility of the computing system.

The general approach to the problem is discussed in Chapter 1 of this thesis. This chapter also includes a review of the properties of each of the consonant phonemes, and an indication of how these factors should influence the behaviour

of Z., T. and I. Chapter 2. describes the method and software used in the investigation. The results are presented in Chapter 3., and the research is discussed in Chapter 4.

CHAPTER I.

1.1. Data.

The investigation was confined to consonants spoken in the initial position in isolated consonant - vowel syllables (C.V. sounds). The restriction to isolated syllables greatly simplified the problem of segmenting the consonant representation from its surroundings. The initial position was preferred since some of the consonants (especially /h/, /j/, /r/ and /w/)are difficult to pronounce in the final position.

Consonant sounds spoken in the final position, or in continuous speech, can differ in many respects from those uttered in C.V. syllables. The method would require much revision to cover the more general case. A discussion of the possibilities for extension of the work will be found in section 4.1.

The C.V. sounds were recorded on magnetic tape using high-quality equipment. Details of the recording method are given in appendix (1). The C.V. syllables were always spoken and fed to the computer in pairs, since this suited the storage capacity of the P.D.P.-8. An added advantage of this procedure was that differences in stress and pitch were often observed between the first and second utterances of a pair of C.V. sounds. It would obviously be useful if the recognition

algorithm could be made agnostic to these changes.

Subjects were given no special instructions with regard to uttering the C.V. sounds, except to give approximately equal stress to the consonant and the vowel. Subjects were allowed to speak at their own speed and were asked to repeat a sound if they considered that they had made an error in pronunciation.

All the consonants were dealt with except the nasal /ŋ/ (as in"sing"), which is difficult to pronounce in the initial position. The 23 remaining consonants were spoken in pairs in an ad hoc order with each of the 10 vowels in turn. The 10 vowels used covered the whole of the F1-F2 range.

Figure 1.1.(a) lists the consonants in their order of utterance. To simplify the computer printout and display, the consonants were numbered from 1 to 23 in this order. The computer printout for each of the consonants is shown in the fourth column of figure 1.1.(a). Figure 1.1.(b) lists the 10 vowels used and their equivalent computer output. The computer output symbols are used in many of the figures in the following sections.

For a single subject, 2 sets of C.V. sounds were recorded for each vowel. Since the consonants were spoken in pairs, this gave 4 examples of each possible C.V. utterance and 40 repetitions of each consonant.

CONSONANTS				VOWELS		
No.	Phoneme	'as in'	Computer Output	Phoneme	'as in'	Computer Output
1	p	pay	P	i	heed	EE
2	t	to	T	I	hid	I
3	k	key	K	ɛ	head	E
4	f	for	F	æ	had	A
5	θ	thin	TH	ɜ	herd	ER
6	tʃ	chin	CH	ʊ	foot	U
7	h	hid	H	u	hoot	OO
8	b	bed	B	ɑ	hod	O
9	d	do	D	ɔ	hoard	OR
10	g	go	G	ɑ	hard	AR
11	v	vote	V			
12	ð	then	DH			
13	dʒ	job	J			
14	s	see	S			
15	ʃ	she	SH			
16	z	zoo	Z			
17	ʒ	azure	ZH			
18	j	you	Y			
19	r	read	R			
20	l	let	L			
21	w	win	W			
22	m	me	M			
23	n	no	N			

Fig 1.1 Table of the Phonemes studied with their Equivalent Computer Output.

Utterances by 4 male speakers were investigated. Of these, the subjects C.W.T., W.A.A. and M.A. spoke standard English, while subject P.D.G. had a strong Yorkshire accent.

1.2. The Z.T.I. Diagram.

Figure 1.2. shows the type of display used throughout this study to examine the behaviour of Z., T. and I. The diagram has been photographed directly from the 338 screen(see Appendix (2)).

In figure 1.2., time runs along the abscissa, and the distance on the axis between successive points represents a single counting interval of length t_c (the value of t_c was normally 6.4 ms., see section 2.1.2). The Z.,T. and I. traces are displaced vertically, with Z. at the bottom, T. in the centre and I. at the top. The ordinates on the Z. and T. traces are the numbers of Zero Crossings or Turnarounds found in a single counting interval, while the I. ordinate represents the highest value of the modulus of the envelope found within this period. This form of display will be referred to as the Z.T.I. diagram.

Figure 1.2. shows the Z.T.I. diagram for 2 utterances of the sound /t3/ by subject C.W.T. The traces have been smoothed and scaled in the manner described in section 2.1.

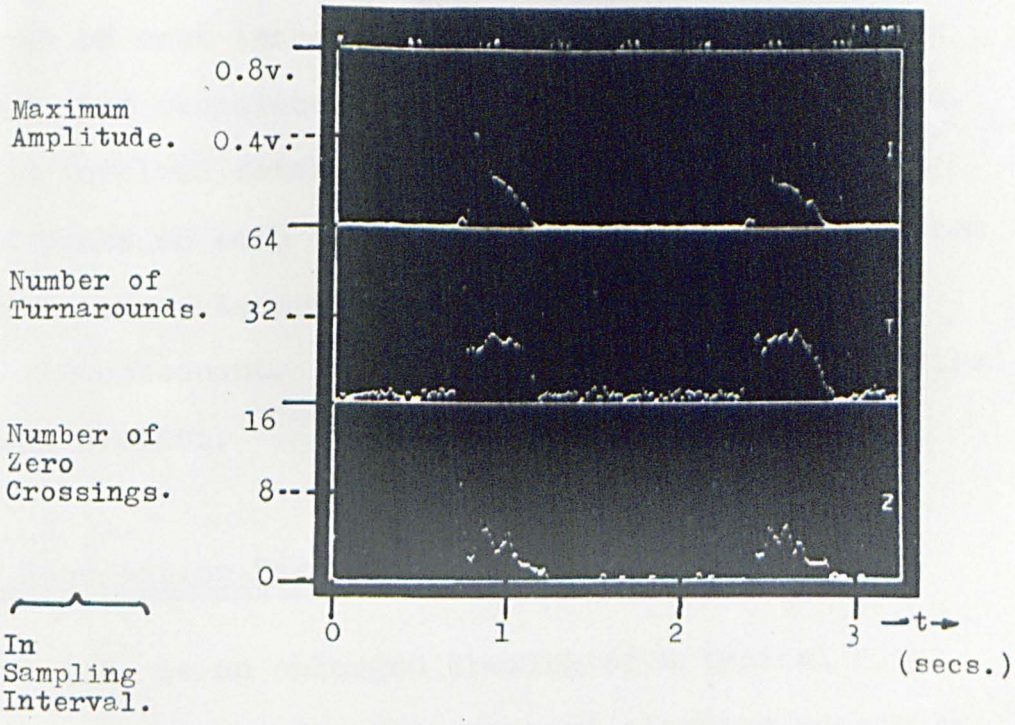
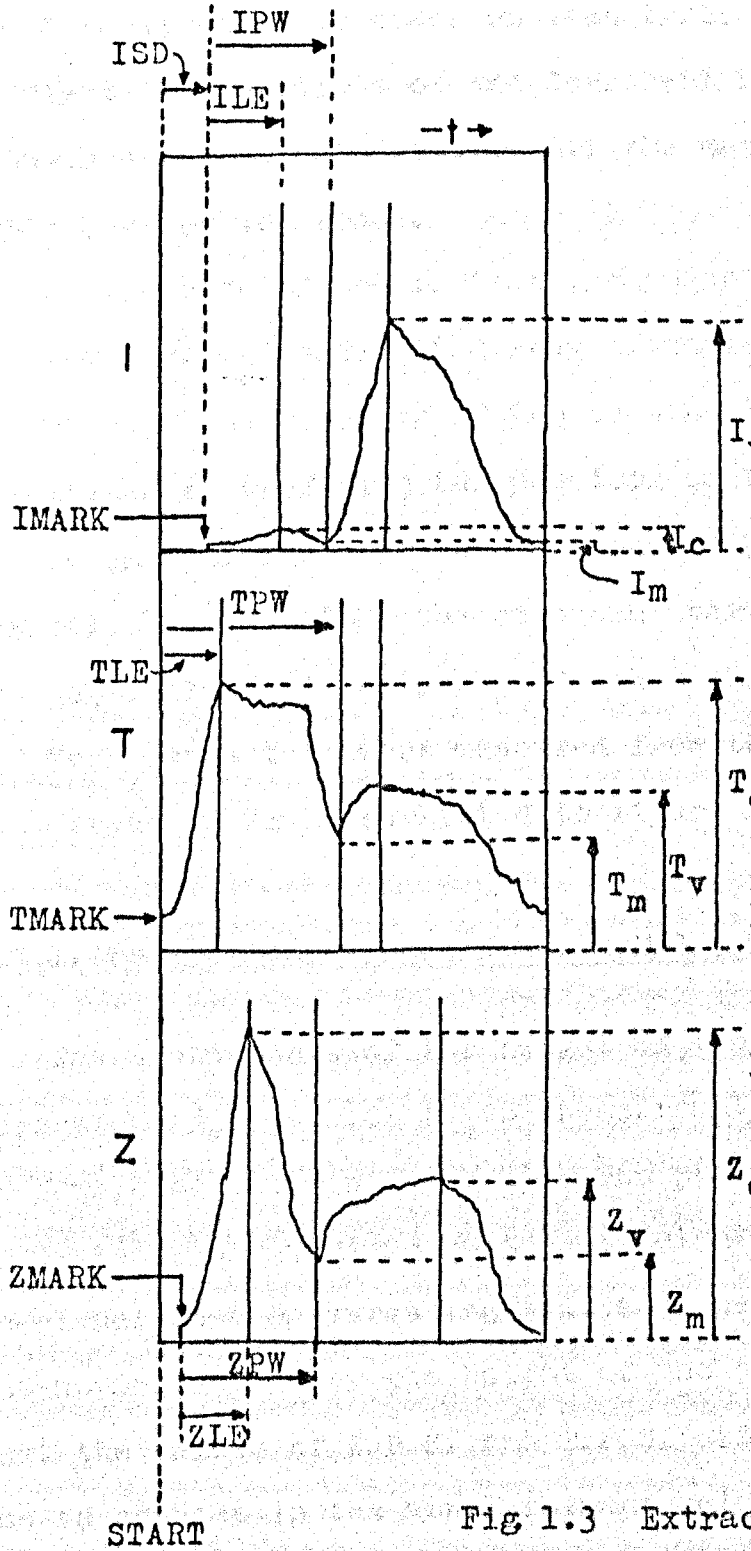


Fig. 1.2 Z.T.I. Diagram for two Utterances of /t3/ by C.W.T.

On each trace it is a simple matter to pick out by eye the points on the Z.T.I. diagram where the utterance begins, and to distinguish between the consonant and vowel parts of the sound. The consonant appears as a single peak on each trace, and there is a minimum position corresponding to the boundary between the consonant and the vowel. It was found that consonants in C.V. sounds could normally be treated as a single peak on each trace, though in some cases this peak was absent, and occasionally multiple peaks were observed. The method involved determining the positions of these consonant peaks on each trace and measuring parameters from them. These parameters were then used in an attempt to identify the consonant. The parameter set used is described in the next section.

1.3. The Recognition Parameters.

Figure 1.3. is an enlarged drawing of a typical Z.T.I. diagram for a C.V. sound. The apparent starting points of the sound on the 3 traces often differed slightly. In Figure 1.3 these positions are called Z.MARK, T.MARK and I.MARK. The diagram has been drawn with t=o corresponding to the earliest of the 3 start markers to occur - T.MARK in this case. The value of the Z. and I. traces has been set



Parameters

1/ Z Peak Size: $ZPS=(Z_c/Z_v).100$

2/ Z Peak Drop: $ZPD=(Z_m/Z_v).100$

3/ Z Peak Width: ZPW

4/ Z Leading Edge: ZLE

5/ T Peak Size: $TPS=(T_c/T_v).100$

6/ T Peak Drop: $TPD=(T_m/T_v).100$

7/ T Peak Width: TPW

8/ T Leading Edge: TLE

9/ I Peak Size: $IPS=(I_c/I_v).100$

10/I Peak Drop: $IPD=(I_m/I_v).100$

11/ I Peak Width: IPW

12/ I Leading Edge: ILE

13/ I Start Delay: ISD

Fig 1.3 Extraction of the Recognition Parameters.

to zero in the intervening period, (See Section 2.1.4). The 3 vertical lines drawn on each trace indicate from left to right the positions of the Consonant Maximum, the Minimum between consonant and vowel, and the maximum value on the vowel part of the trace.

In the case of the I. trace, changes in the envelope of the vowel as the sound died away often led to a false estimate of the position of the consonant peak. This was eliminated by confining the position of the consonant peak to the part of the utterance before the position of the overall maximum, since the consonant part of the I. trace was generally of lower amplitude than that of the vowel.

Four parameters were measured from the consonant peak on each trace. These are listed in figure 1.3.

1.3.1. The Consonant Peak Sizes Z.P.S., T.P.S. and I.P.S.

The measurement of the height of the consonant peak corresponded most closely to the parameters used by other workers,(See Section I.9). The peak size was obtained by dividing the maximum consonant height by the maximum vowel height and expressing this as a percentage.
The normalisation by the height of the vowel maximum meant that the values of Z.P.S. and T.P.S. differed between vowels due to changes in the vowel formant values.

The normalisation was intended to stabilise the measurement for those consonants with a variable articulator position, such as /h/, /g/ and /f/, when the formant pattern tended to be similar to that of the following vowel. These consonants were expected to be the most difficult to identify.
The differences in the vowel height were at any rate much smaller than those occurring between the consonants.

1.3.2. The Consonant Peak Drops, Z.P.D.,T.P.D and I.P.D.

A crude measure of the magnitude of the change between the consonant and the vowel was obtained from the height of the Minimum dividing the 2 parts of the C.V. sound. This was again normalised by the maximum vowel height and expressed as a percentage.

1.3.3. The Consonant peak widths Z.P.W., T.P.W. and I.P.W.

The duration of the consonant part of the C.V. sound on each trace was measured as the number of points between the positions of the starting point of the trace and the Minimum between the consonant and vowel. Thus Z.P.W. = 20 was equivalent to a Z. duration of 20 x tc ms.

1.3.4. The Consonant Onset times Z.L.E., T.L.E and I.L.E.

The time taken for the Consonant Peak to rise to its maximum position was measured as the number of points between the appropriate start marker and the position of the Consonant Maximum. This gave a measure of the abruptness of the onset of the sound.

1.3.5.

For some consonants, mostly /f/, /θ/ and /h/, the apparent starting point on the I. trace occurred considerably later than on the Z. and T. traces (see section 3.1.9.1). This was utilised for recognition of these sounds by measuring a 13th. parameter, the I. Start delay, I.S.D. I.S.D. was taken as the number of points between the overall start of the sound and I. MARK.

In cases where no distinct peak occurred on a particular trace, all the relevant parameters were set to zero except the onset time (L.E.), which was made equal to the number of points between the start marker and the overall maximum of the trace.

These parameters, particularly the peak widths were, of

course, far from independent of each other. Limitations and possible modificationsof the parameter set are discussed in Section 4.2. The method used for extracting these parameters is described in Section 2.1.

The next section presents a more detailed revue of the production and spectral properties of the various consonant sounds and considers how these factors should influence the Z.T.I. diagram.

1.4. Behaviour of Z.T. and I. in Consonant Sounds.

1.4.1. General.

Unlike the vowels (and vowel-like consonants), consonant sounds generally possess at least some degree of fricative (noiselike) modulation. The presence of this voiceless modulation means that the behaviour of a consonant sound on the Z.T.I. diagram can be widely different from that of a vowel.

When fricative excitation predominates, the formant structure of the sound disappears, and the sound can be considered as band limited white noise. The sonagram is then characterised by the broad, dark region of energy known as a "fill". When this occurs, the Zero Crossing and Turnaround

rates will be similar, and will presumably respond to some frequency in the "centre" of the energy band. Since the energy band often extends to frequencies above the vowel formant range, the Z. and T. rates can be much higher than those of the following vowel in a C.V. sound. The value of the parameter Z.P.S. will become very large, since the zero crossing rate of the following vowel will be restricted by its F1 value. (See Section I.7)

All the consonants except the Glides and Nasals can be divided into cognate phoneme pairs. Each pair of phonemes is made using roughly the same set of articulator positions, corresponding to a single formant pattern. Cognate phonemes differ in the type of modulation present: voiceless consonants are produced primarily by fricative excitation, while voiced consonants combine fricative and vocal chord modulation.

The presence of vocal chord excitation in voiced consonants means that these sounds retain a formant structure to some extent. Some formant bars can usually be observed in the sonagrams of these sounds, though the pattern is generally less well defined than for a vowel, and some of the formants (especially the higher ones) may be lost. Often broad regions of energy due to the fricative component of the sound can also be seen.

In a strongly voiced sound, with a clear formant pattern, the Time Interval measurements will respond with vowel-like behaviour, Z. being influenced most by the value of F1 and T. by the higher formant values. Since F1 generally appears more clearly than the higher formants in the sonagrams of voiced consonants, the Zero Crossing rate will be more influenced by the voicing than the Turnaround rate, and the values of Z.P.S. will be lower for voiced than for unvoiced sounds.

The extent to which the formant pattern influences the sonagram and Z.T.I. diagram of a consonant sound will depend on the relative proportions of voiced and voiceless modulation present. This ratio can vary between speakers, between phonemes, between different utterances of the same phoneme, and sometimes within a single utterance. In general the Z.T.I. diagrams of voiced consonants will show more vowel-like features than those of the voiceless consonants, though the latter may also be influenced slightly by the formant pattern, since weak formants are sometimes observed in their sonagrams.

Voiceless sounds generally have a lower amplitude value than voiced sounds, and this should be reflected in the values of the parameter I.P.S. The consonant peaks (especially the I. peak) on the Z.T.I. diagram will be more distinct for voiceless than for voiced sounds, since there must be a sudden

change at the onset of the vowel for voiceless sounds at the moment of switching to voiced excitation. The prominance of the I. peak and the sharpness of the change on the I. trace to the following vowel will again depend on the relative amounts of voiced and voiceless modulation used.

The duration of the consonant sound will be reflected in the widths of the consonant peaks on the Z.T.I. diagram, and the sharpness of the onset of the consonant sound will govern the L.E. values. These consonants with a longer duration are more likely to reach a "steady state" position, producing flat topped peaks.

Variations in stress can also have a large effect on the sonogram of a consonant sound. In addition to changing the overall amount of energy present in the sound, in many ways variation in stress can effect the formant pattern produced. Increase in stress will generally lead to a larger I. peak. Sometimes increased stress also leads to the appearance of additional formant bars at higher frequencies in the sonagrams. This will generally increase the Turnaround rate. Quite different Z.T.I. diagrams may be produced for stressed and unstressed versions of the same utterance.

The following sub-sections describe the production and spectral properties of each of the consonant phonemes, and indicate how these may effect the Z.T.I. diagram.

Most of the data on consonant spectra is taken from Potter, Kopp and Green (44).

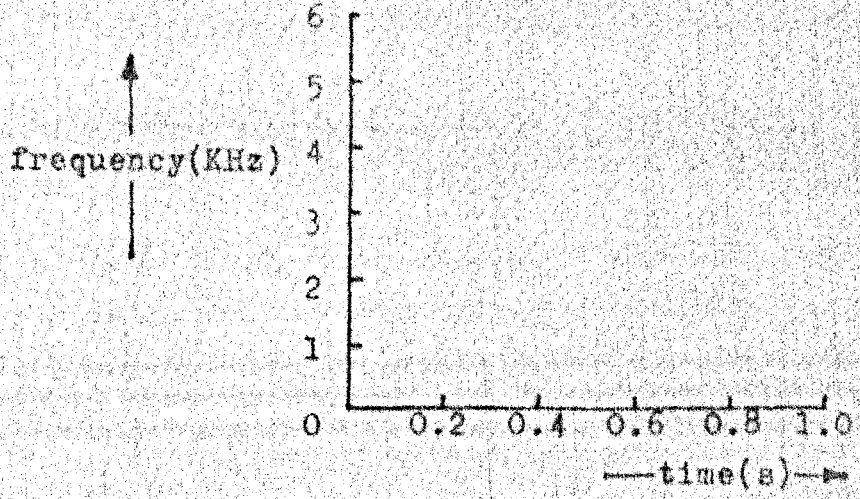
1.4.2. The Stop Sounds.

The phonemes /p/, /t/, /k/, /b/, /d/ and /g/ belong to the class of speech sounds known as stops. Stops are produced by :

(a) Stopping the breath stream by forming a closure at some point in the vocal tract.
(b) Building up breath pressure behind the point of closure.
(c) Exploding the breath by quickly reopening the vocal tract.

Stop sounds cannot be sustained and are necessarily of short duration. For this reason the consonant peaks on the Z.T.I. diagram for stop sounds should be narrower than those for other types of consonants. The sudden release of breath in the explosion of the stop should give rise to low values of the onset times (L.E.) on the Z.T.I. diagram.

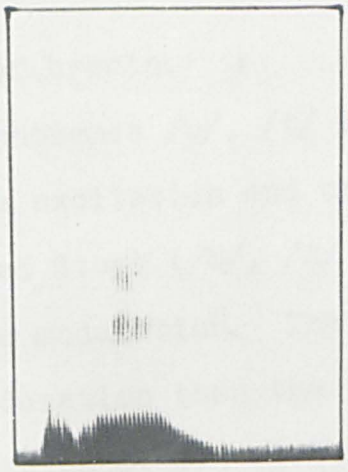
Typical sonagrams of the stop sounds in isolated C.V. syllables are shown in figure 1.4. Since the consonants were always spoken in the initial position, the stop gap (the period of silence while breath is built up) cannot be used as a recognition clue, and the sonagrams are characterised by a short "spike fill" due to the sudden



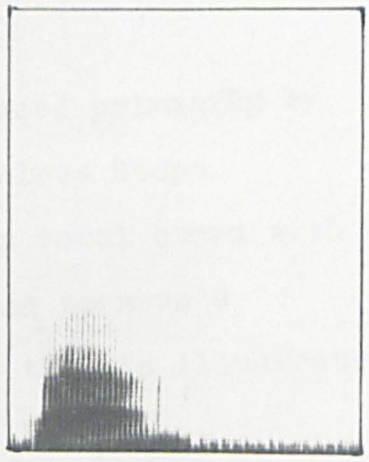
Sonagram Scales.

(same throughout)

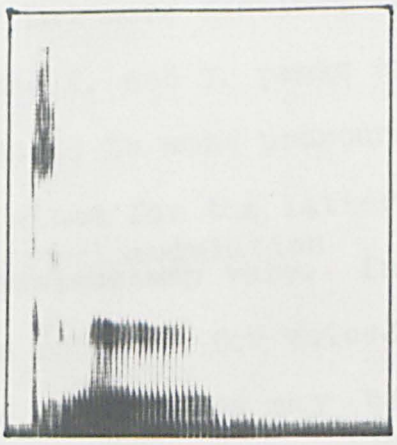
Fig. 1.4 Sonagrams of C.V. Syllables spoken by C.W.T.
-Stops



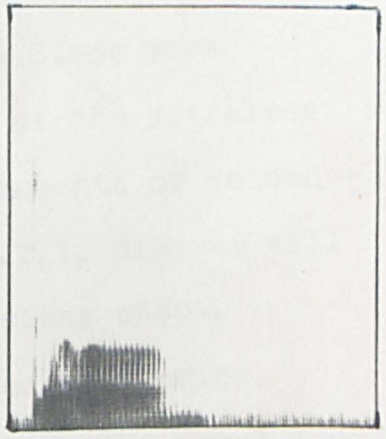
(a) /pI/



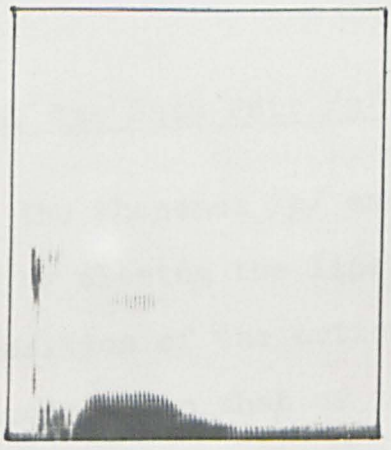
(d) /bɑ/



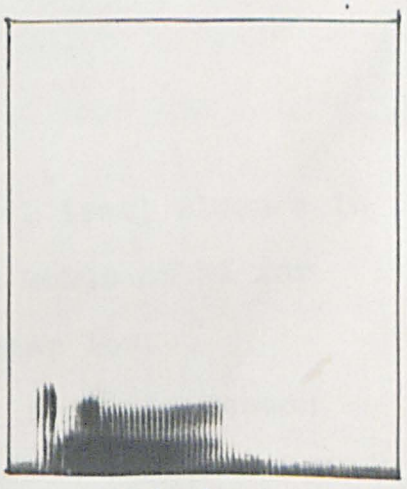
(b) /tʒ/



(e) /dʒ/



(c) /kI/



(f) /gɔ/

release of breath.

The phonemes /p/, /t/ and /k/ are produced primarily by fricative excitation and are known as Voiceless Stops. The Voiced Stops (/b/, /d/ and /g/) combine vocal chord with fricative modulation. The Voiced Stops tend to have a shorter duration than the unvoiced stops - this is illustrated by the widths of the spike fills in figure 1.4. The I. peaks for the Voiceless Stops were expected to be smaller but more distinct than those of the Voiced Stops, while the Z. and T. peaks of the Voiceless Stops were expected to be more pronounced than those of the Voiceless Stops, since for the latter the relative amounts of voiced - and voiceless modulation may vary. In general the Z.T.I. diagram will be more variable for voiced than for voiceless stops.

The stop sounds may be divided into cognate pairs, each pair having the same point of closure in the vocal tract. One member of each pair is voiced and the other voiceless.

1.4.2.1. The Stop Pair /p/ and /b/.

In the phonemes /p/ and /b/, the vocal tract closure is formed by closing the lips. The natural value of F1 for this position of the articulators is rather low, corresponding to that of the vowel /u/. For this reason

the Z. peaks should be small, especially for the voiced stop /b/, where the influence of F1 is greater. If the modulation was totally voiced, the value of Z.P.S. should not exceed 100 except for following vowels with very low values of F1. The Z. peaks for /p/ should also be smaller than those of the other unvoiced stops owing to the influence of F1.

Sonagrams of /p/ have spike fills which are normally darker towards the bottom of the pattern, indicating a concentration of energy at low frequencies. Figure 1.4(a) is an extreme example of this. This implys that /p/ will have smaller T. peaks. To a lesser extent this will also be true of /b/.

1.4.2.2. The Stop Pair /t/ and /d/.

In this phoneme pair, the vocal tract closure is made by holding the tongue against the back of the teeth. This corresponds to a medium value of F1, similar to that of the vowel /æ/. The value of F1 should hold down the height of the Z. peak for the voiced stop /d/. In a totally voiced sound, the values of Z.P.S. would be less than 100 for back vowels (high F1), about 100 for mid vowels and greater than 100 for front vowels. The F1 influence will be less strong for the unvoiced stop /t/.

Sonagrams of /t/ (e.g. figure 1.4(b)) show a large amount of energy towards the top of the pattern in the higher frequency range. This means that the T. peaks for /t/ should be fairly high ; the height of these peaks for /d/ will depend on the relative amounts of voiced and unvoiced modulation, but should generally be greater than for the other voiced stops.

1.4.2.3. The Stop Pair /k/ and /g/.

The stops /k/ and /g/ are produced by placing the back of the tongue against the roof of the mouth or the velum prior to exploding the stop. The exact position of the closure varies with the following vowel. This means that the F1 value associated with /k/ and /g/ tends to be similar to that of the vowel with which the stop is pronounced. This variability in F1 should be reflected in the Z. peaks for /g/ (voiced), and to a lesser extent for /k/. For a totally voiced /g/, Z.P.S. should be about 100 for all vowels.

The frequency band for /k/ at which there is most energy also varies with the following vowel, but normally lies between the positions for /p,b/ and /t,d/. This can be seen in figure 1.4(c). In general the values of T.P.S. for the /k,g/ pair should lie between those obtained for the /p,b/ and /t,d/ pairs.

1.4.3. The Fricative Sounds.

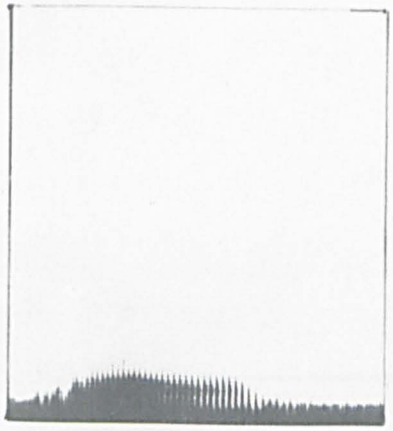
The phonemes /h, f, v, θ, ð, s, z, ʃ and ʒ / are usually classified as Fricative sounds. Fricatives are produced by:-

(a) Forming a small opening at some point in the vocal tract.

(b) Emitting a continuous stream of breath through the restricted opening.

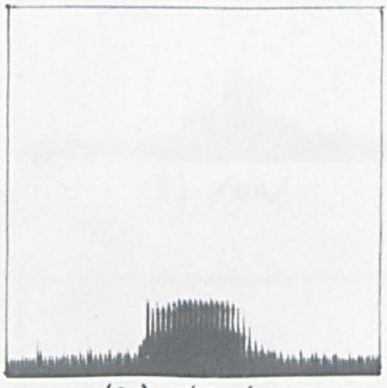
Unlike stop sounds, the Fricatives can be sustained indefinitely and sometimes attain a "steady state" position. Their duration is normally longer than that of the stops. This implies that the Z.T.I. diagram for a fricative should show consonant peaks which are both wider and flatter than those for a stop sound. The onset time (L.E.) for these peaks may be greater than that for the Stop sounds, since there is no sudden explosion of breath for a Fricative.

Typical sonagrams of Fricative sounds in C.V. syllables are shown in figure 1.5. The sonagram patterns of the Fricatives are referred to as "fills". The examples of figure 1.5 show that the duration of the fill varies a good deal, while generally remaining greater than that of the spike fill of a stop sound (figure 1.4.). The total amount of energy in the fill may be quite large (as in figure 1.5. (h),/ʃ/,) or very small (figure 1.5(a),/h/). In the latter case, the consonant peaks on the Z.T.I.diagram may become indistinct.

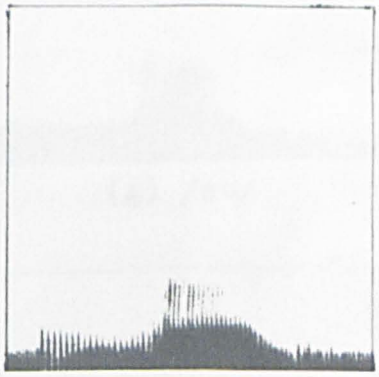


(a) /hɔ/

Fig. 1.5 Sonagrams of C.V. Syllables Spoken by C.W.T. -Fricatives.

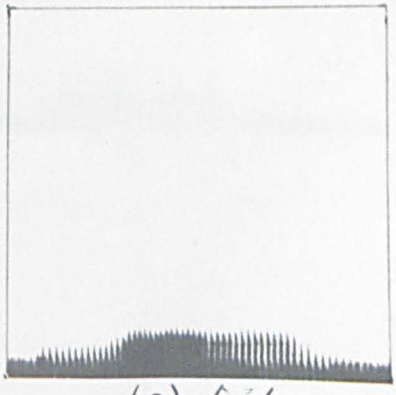


(b) /fu/

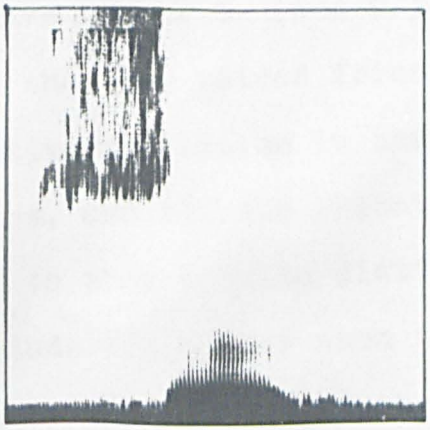


(c) /vɯ/

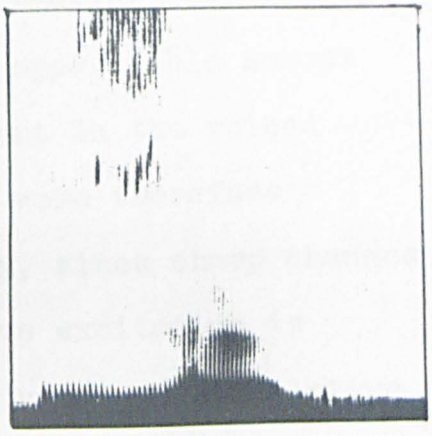
(d) /θɜ/



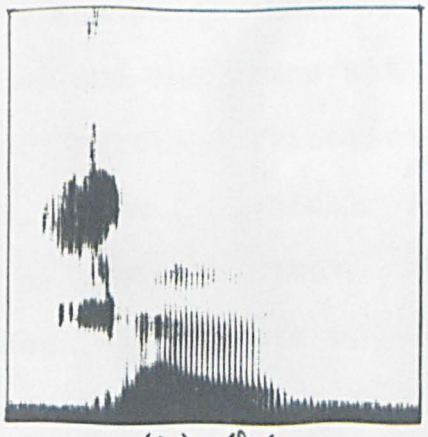
(e) /ðɜ/



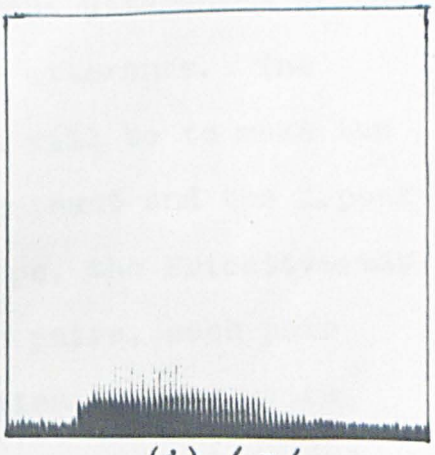
(f) /sɥ/



(g) /zɥ/



(h) /ʃʒ/



(i) /ʒa/

The phonemes /f, θ, s, and ʃ/ are produced by frictional excitation and are classed as Voiceless Fricatives, while the Voiced Fricatives /v, d, z and ʒ/ are made using some measure of vocal chord modulation. The phoneme /h/ is sometimes voiced and sometimes voiceless. The voiceless fricatives should have higher Z. (and T.) peaks and smaller but clearer I. peaks than the voiced fricatives. An appreciable amount of fricative modulation is normally present in the voiced fricatives, and all the Fricative sounds were therefore expected to show a quite distinct I. peak, since sharp changes in amplitude will occur when the fricative excitation is terminated at the onset of the following vowel. The relative amounts of vocal chord and frictional modulation in Voiced Fricative sounds can vary between different utterances of the same sound and sometimes within a single utterance. The effect of increased fricative modulation will be to make the Z. and T. peaks (especially T.) more prominent and the I. peak smaller but more distinct. Like the stops, the Fricatives may be divided into cognate voiced-voiceless pairs, each pair being made from the same articular position. These pairs are /f, v/, /θ, ð/, /s, z/ and /ʃ, ʒ/. Sometimes the voiced and voiceless versions of /h/ are considered to be a fifth cognate pair.

1.4.3.1. The Fricative /h/.

/h/ is made by forming a small opening at some point in the region of the glottis. /h/ is a variable sound ; the articulators tending to assume the position for the sound which precedes or follows it. This means that the natural F1 position of the initial /h/ in a C.V. sound will approximate to that of the following vowel. For an /h/ where voiced excitation predominates, the value of Z.P.S. should be about 100. When fricative modulation predominates, the portion of the /h/ spectrum with the greatest energy concentration will also be dependant on the vowel, making T.P.S. about 100.

According to Potter, Kopp and Green, (44), the initial /h/ is usually voiceless. The total amount of energy in the /h/ sound varies widely with differences in stress, but in most of the examples considered there was very little energy to be seen in the /h/ fill (see figure 1.5.(a)). The extreme variability of the /h/ will be reflected in the Z.T.I. diagram.

1.4.3.2. The Fricative Pair /f, v/.

In the case of the /f,v,/ pair, the opening is formed between the lower lip and the upper teeth.

The natural F1 value for this articulator position is very low, approximating to that of the vowel /ʌ/. This can be seen in the sonogram of figure 1.5(c). The Z. peaks for /v/ will therefore be very small, Z.P.S. being smaller than 100 for a totally voiced sound. In the case of /f/, the energy is widely scattered over the spectrum, but the darkest areas of the sonagram are often towards the baseline. This should restrict the size of the T. peaks for /f/. As figure 1.5(b) shows, there is sometimes very little energy present in the /f/ sound, though occasionally the amount of energy is much greater.

1.4.3.3. The Fricative Pair /θ, ð/.

/θ/ and /ð/ are produced by forming a small opening between the tip of the tongue and the upper teeth. The exact position of the opening varies from speaker to speaker. This will be reflected in greater speaker to speaker variation in the Z.T.I. diagram.

The natural F1 position for the /θ, ð/ pair is again rather low, but is slightly higher than that of /f/ and /v/. It corresponds most closely to that of the vowel /a/. Small Z. peaks were therefore expected for /ð/, which was generally strongly voiced, though Z.P.S. may be a little larger than for /v/.

In /θ/ the energy concentration is again in the lower portion of the range, as in the case of /f/. Like /f/, /θ/ often has a very low overall energy level, as shown in figure 1.5 (d).

1.4.3.4. The Fricative Pair /s,z,/.

In the /s,z,/ pair the opening is formed between the tongue and the alveolar ridge. The opening may be made either with the tip or the blade of the tongue. There is also a narrow opening between the upper and lower front teeth which contributes to the high frequency domination of the spectrum for these sounds. The duration of /s/ and /z/ is often very long.

The F1 position for the /s,z/ pair lies in the middle of the range, at the position of the vowel /æ /.
This means that strongly voiced examples of /z/ should have relatively small Z. peaks. The amount of voicing present in /z/, however, varies widely and can sometimes change during a single utterance.

As figure 1.5(f) shows, the energy in /s/ is concentrated in the higher regions of the spectrum. This implies very high T. peaks for /s/. The Z. peaks for /s/ will also be large, since there is very little energy in the F1 region.

/z/ also has a large amount of energy at HF besides the F1 bar. This can be seen in figure 1.5(g). For this reason, /z/ was expected to have large T. peaks, similar to those of /s/.

1.4.3.5. The Fricative Pair /ʃ,ʒ/.

In this cognate phoneme pair, the opening is made between the tongue and the anterior portion of the palate. In some cases the tip of the tongue is used, while in others the opening is formed with the tongue blade. As in the case of the /s, z/ pair, there is also a narrow opening between the upper and lower teeth, causing additional fricative modulation of the sound.

The position of F1 for /ʃ/ and /ʒ/ is in the top of the range, approximately coincident with that of the vowel /i/. This means that /ʒ/ should have the highest Z. peaks for voiced fricatives, though like /z/, the amount of voicing in /ʒ/ may vary widely.

The spectrum of /ʃ/ is similar to that of /s/, but for all the subjects considered, the region of greatest energy was slightly lower than that of /s/ * (compare figures 1.5 (f)

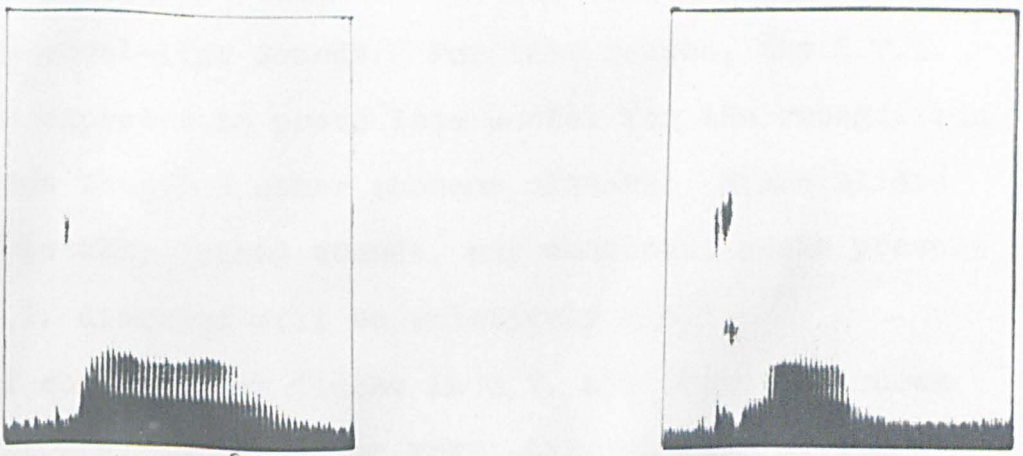
* According to Potter, Kopp and Green, (44) the reverse should be true (/ʃ/ higher than /s/).

and (h)). The T.P.S. values for /ʃ/ should therefore be slightly smaller than those of /s/, though remaining quite large. Similarly T.P.S. for /ʒ/ should be slightly smaller than for /z/. The Z. peaks of /ʃ/ will again be very high, since there is little energy in the FI region.

1.4.4. The Affricative Sounds.

The phonemes /tʃ/ and /dʒ/ are generally known as Affricative sounds. They may be considered as a combination of a stop and a Fricative sound. The /tʃ/ is formed by closing the vocal tract in the /t/ position, then exploding the stop and moving to the /ʃ/ position. Since /t/ and /ʃ/ are voiceless sounds, the combination /tʃ/ is also voiceless. /dʒ/ is made in a similar way from the voiced stop /d/ and the voiced fricative /ʒ/ and is classed as a Voiced Affricative.

Figure 1.6. shows typical C.V. sonagrams of the two Affricatives. The sonagram patterns of the Affricatives resemble those of the Fricatives(/ʃ/ and /ʒ/) involved, but tend to be of somewhat shorter duration. The initial Stop spike can sometimes be seen. There is generally less energy present in the Affricatives than in their Fricatives /ʃ/ and /ʒ/.



(a) /tʃɑ/ (b) /dʒa/

Fig. 1.6 Sonagrams of C.V. Syllables spoken by C.W.T.- Affricatives.

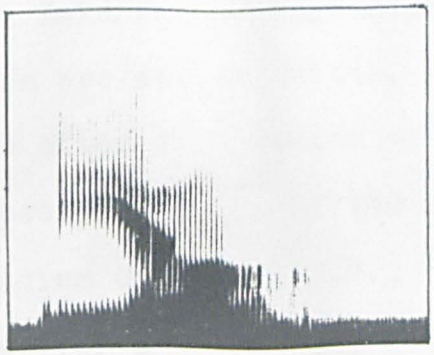
1.4.5. The Glides.

The phonemes /w/, /j/, /r/ and /l/ belong to the class of speech sounds known as Glides. In the initial position, Glides are produced by moving the articulators from some starting position towards the position for the following sound. When the Glide is said in the final position, these events occur in the reverse order. The excitation is voiced throughout the movement for all Glides. Glides are essentially vowel-like sounds. For this reason, the Z.T.I. diagram was expected to prove less useful for the recognition of the Glides than for other phoneme classes. Since Glides are almost totally voiced sounds, any consonant peaks present on the Z.T.I. diagrams will be relatively small.

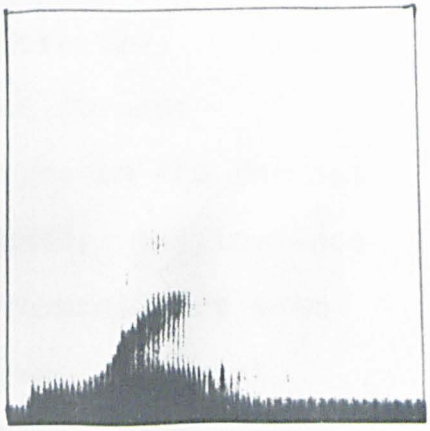
Typical sonagrams of Glides in C.V. syllables are shown in figures 1.7(a) to (d). The vowel-like quality of the Glides is shown by the presence of strong formant bars. The curves of the formants towards the following vowel as the articulators move to the vowel position can be clearly seen. The length and slope of the formant curves vary a great deal, but the curves are quite smooth and merge gradually with the vowel formants. These formant movements should not produce consonant peaks on the Z.T.I. diagram, though Z. and T. will change with the formants.

Fig. 1.7 Sonagrams of C.V. syllables spoken by C.W.T.

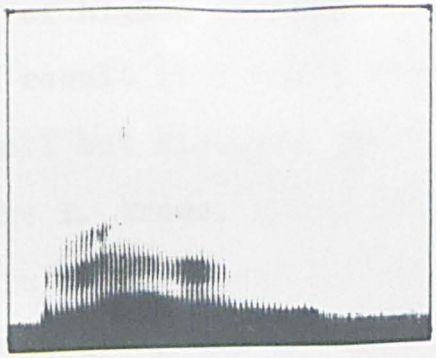
- Glides and Nasals.



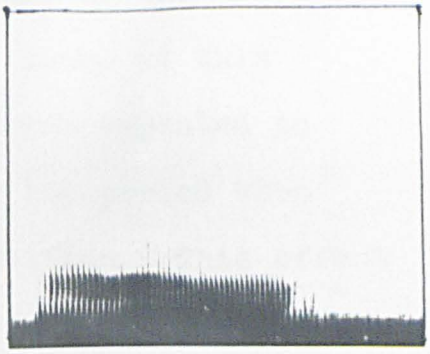
(a) /jɑ/



(d) /wæ/

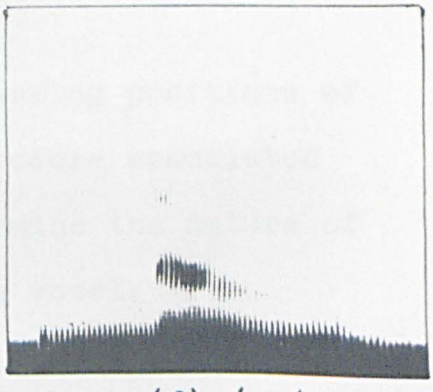


(b) /ru/



(e) /mɔ/

(c) /li/



(f) /nu/

From the sonagrams of figure 1.7, it is apparent that in isolated C.V. syllables the starting position of the Glide is normally held for an appreciable time before the articulators are set in motion. The onset of formant movement is often accompanied by abrupt changes in the formant values, especially those of the higher formants. For instance in the sonagram of figure 1.7., the higher formant bars seem to disappear for a short time before the formant movement begins. This would appear in the Z.T.I. diagram as a sharp minimum on the T. trace. In figure 1.7(d), there is a sudden appearance of higher formant bars at the onset of the movement. This would result in a sharp rise in the T. trace at this point. Small but distinct peaks were therefore expected to occur on the T. trace, these peaks covering the period when the articulators are held in the initial position. This effect might also occur on the Z. trace, to a lesser degree.

Distinct peaks were not expected to occur on the I. trace since the Glides are voiced throughout and the envelope should change only gradually.

The four Glide sounds differ in the starting positions of the articulators. These positions are therefore associated with different formant patterns which determine the nature of the formant movements towards the following vowel.

/w/ has an initial articulator position similar to that used in making the vowel /u/, and therefore has a rather low F1 value.

/j/ is made by placing the lips in approximately the same position as for the vowel /I/, while the other articulators assume the positions for /i/. The F1 value for /j/ is rather high, similar to that for /i/.

In /r/, the articulators take a similar position to that of the Stops /t,d,/, but the tongue is moved back towards the centre of the hard palate. The natural F1 value varies between that of /a/ and /æ/.

In /l/, the lips assume the position for the vowel /ʌ/, but while the tongue is placed against the alveolar ridge, or adjacent hard palate. The F1 value generally resembles that of the following vowel, but is somewhat lower.

From the natural F1 values for the Glides, /w/ would be expected to have the smallest Z. values, /j/ the highest, with /r/ lying between the two. The Z. values of /l/ will depend on the following vowel, but will generally also lie between those of /w/ and /j/.

1.4.6. The Nasal Sounds.

The 3 phonemes in English which are characterised by nasal resonance are /m/, /n/ and /ŋ/. /ŋ/ was not dealt with in the present study because of the difficulty of pronouncing it in the initial position.

The nasals are produced by opening the nasal port and closing the mouth cavities. A continuous stream of voiced breath is then emitted through the open nasal cavity. At the onset of the next sound the nasal cavity is re-closed.

Sonograms of typical nasal sounds in C.V. syllables are shown in figure 1.7(e) and (f). Since the excitation is voiced throughout, the nasals have vowel-like properties characterised by strong formant bars similar to those of the Glides. The nasal resonance produces an additional voice bar along the baseline of the sonograms (the Nasal Formant, F.N.). The levels of the Z. and T. traces will therefore be low for a nasal sound.

In the nasal sounds there is no gradual formant movement like that of the Glides. The sudden closure of the nasal port produces an abrupt change in the formant bars similar to the change which often occurs in a Glide at the onset of formant change. This will produce distinct peaks on the T. trace, and to a lesser extent on the Z. trace. The Z.T.I.

diagrams will in general be very similar to those of the Glides.

In /m/, the lips are closed as in the stop phase of the /p, b,/ pair, while for /n/, the mouth cavity closure is in the oral cavity, like that of the Stop pair /t,d./.
The natural value of F1 is therefore low for /m/ and in the centre of the range for /n/. The Z. peaks, if any, for /m/ should therefore be smaller than those of /n/.

CHAPTER 2.

2. Introduction.

The experiments described in this study were performed using a Digital P.D.P. - 8 computer. Details of the computing system will be found in Appendix (2).

The first section of chapter 2 describes the method used to extract the parameters introduced in Section 1.3. from a C.V. sound. In section 2.1., several algorithms are described for picking out the positions on the Z.T.I. diagram of the various points needed to calculate these parameters. No attempt was made to make these algorithms foolproof: it was decided to keep the programming fairly simple and to tolerate mistakes, which could easily be detected and corrected manually with the facilities available. Lack of time and computer storage prohibited the development of algorithms which would require no manual checking. The system was not intended to comprise a practical consonant recogniser, and the algorithms would in any case be of little use in a real recognition situation. The algorithms were designed to be about 90% efficient (i.e. to make one mistake in 10 trials).

2.1. Processing of a Single Pair of C.V.Sounds.

Figure 2.1. is a block diagram of the software used to process a pair of C.V. sounds.

The 2 utterances were played directly into an A.to D. converter, and the first part of the programme performed an on-line extraction of the Zero Crossing and Turnaround Rates and the Envelope Variations. These functions were held in core. A short time smoothing average was then executed and the 3 functions were scaled and adjusted to allow simultaneous display on the 338 screen.

The computer then attemped to locate the point on each of the Z.T. and I. traces corresponding to the start of the first C.V. sound. The Z.T.I. traces for the whole of the period of extraction were displayed on the 338 screen (giving a diagram similar to that explained in Section 1.2.) and the estimated starting points were indicated by vertical lines superimposed on the display. Figure 2.2. shows the Display at this stage in the processing of 2 utterances of /tʒ/ by subject C.W.T.

The programme now entered a Push-Button control subroutine. This enabled the correction of any errors made in estimating the beginning of the sound. As figure 2.2. shows, the correct starting point could easily be found by eye.

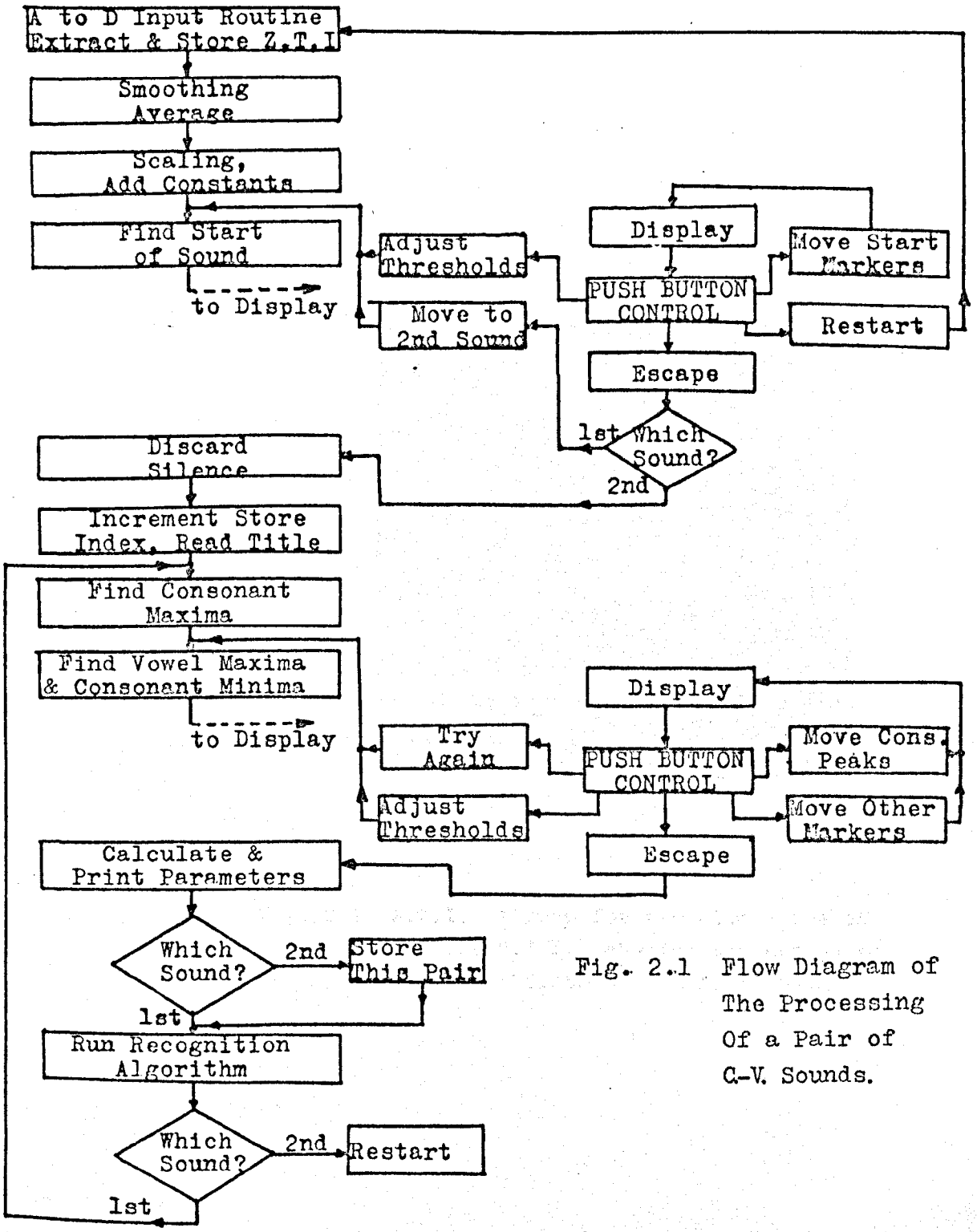


Fig. 2.1 Flow Diagram of The Processing Of a Pair of C-V. Sounds.

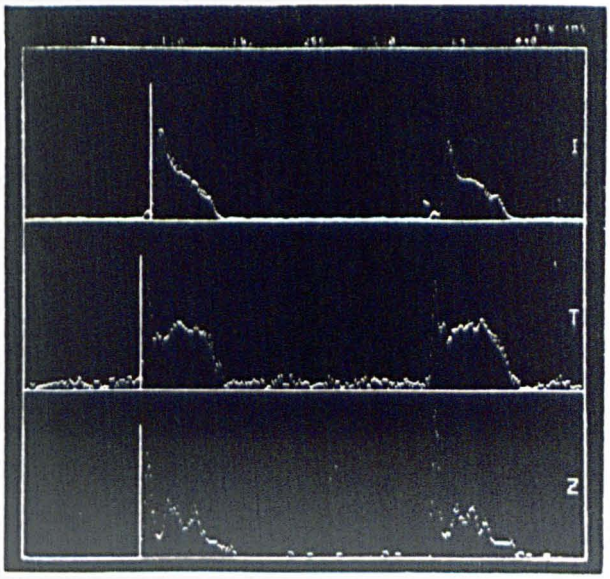


Fig. 2.2 Z.T.I. Diagram for two Utterances of /t3/ by C.W.T., showing the estimated Starting Points of the first sound.

Errors could be corrected by :

(a) Re-running the start-finding algorithm with adjusted Thresholds. For instance in figure 2.2. the I. start marker has been wrongly placed due to an inappropriate value of the I. Threshold.

(b) Moving the Start Markers manually by means of the Push Buttons.

(C) Re-running the A. to D. extraction programme. This was necessary when the wrong portion of Tape had been played.

When the start markers were satisfactorily positioned, the procedure was repeated for the 2nd sound of the pair. The 3 functions were then truncated at a fixed distance from the start for each sound, and the buffer stores were rearranged to remove as much of the unwanted silence as possible. Figure 2.3. illustrates how this was done. The Title of this pair of sounds was then read in via the teletype and the index governing the permanent storage location of the Z.T.I. diagram on Dectape was incremented.

The computer next attempted to find the position on each trace of the Consonant and Vowel Maxima, and the Minimum between the consonant and vowel parts of the first C.V. sound. These points were indicated by vertical lines superimposed on the Z.T.I. display and the computer again entered a Push Button control sequence. Errors in the positioning of the

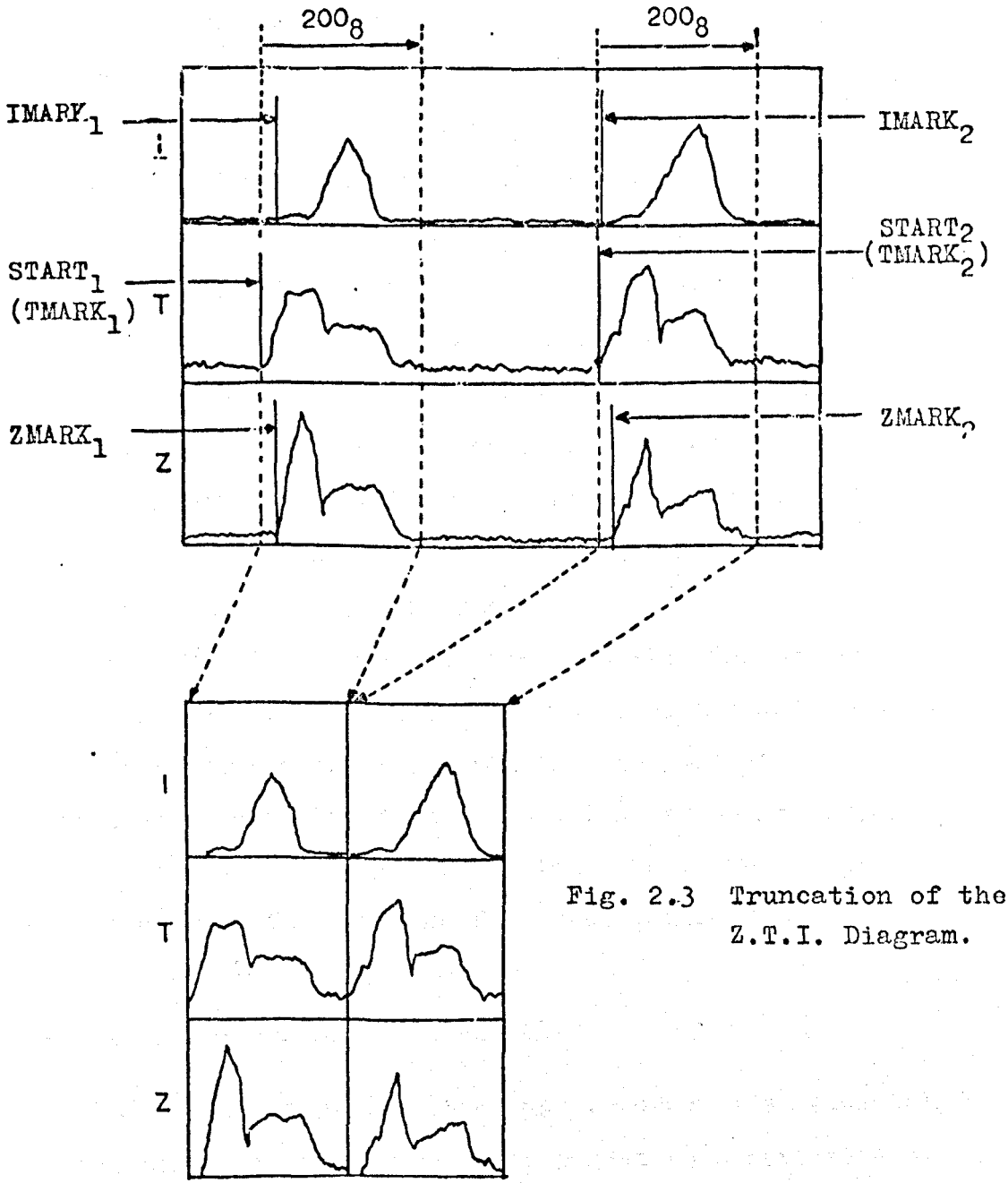


Fig. 2.3 Truncation of the Z.T.I. Diagram.

peaks and minima could be corrected by :

(1) Moving the consonant peak markers manually by means of the Push Buttons and recalculating the positions of vowel maxima and minima. The consonant peak-picking algorithm was not Threshold dependent and was performed prior to estimating the position of the other markers.

(2) Adjusting the Thresholds used in the algorithms for finding the Minima and Vowel Maxima.

(3) Moving the Minima and Vowel Maxima markers manually.

When the markers were satisfactorily positioned, the 13 parameters were calculated and printed out on the Teletype. The parameter set was then run through a recognition algorithm, and the result of the recognition attempt was also printed. This procedure was then repeated for the second sound of the pair. When the parameters of the 2nd sound had been printed, the Z.T.I. diagram and the parameter sets of both sounds were placed in permanent store on Dec-Tape.

The software used in the various processing stages is discussed more fully in the following sections.

2.1.1. The A. to D. Input Routine.

The speech input from the tape recorder was accurately backed off so that silence corresponded to a zero reading on the A. to D. converter. Checks for drift were made at

intervals of a few minutes. Samples of the speech wave were taken every 50μs (corresponding to a band width of 10 K.Hz.) to an accuracy of 9 bits. Zero Crossings were detected by a change in polarity of the speech wave and Turnarounds by a change in the polarity of the slope. After a fixed counting interval (tc.), determined by the No. of samples taken, the numbers of Zero Crossing and Turnarounds found were stored together with a measure of the envelope determined by the largest amplitude sample occurring within this period. The Zero Crossing and Turnaround counters were then cleared and the routine was repeated. The 3 functions extracted were stored sequentially in separate lists, and the input routine terminated when these lists were filled.

A block diagram of the programme is shown in figure 2.4. The last digit in the sample reading was ignored when looking for Zero Crossings and Turnarounds in order to eliminate switching point error in the A. to D. converter. The check for maximum amplitude was made only when a Turnaround had been found, since the maximum must occur at a peak.

The lists each contained 1000_8 locations. The duration of the input routine was therefore about 3.3s. with tc=6.4 ms.

The basic time between samples was 50μs when no Turnaround or Zero Crossing was found, the computer running as fast as possible. When a Turnaround or Zero Crossing occurred,

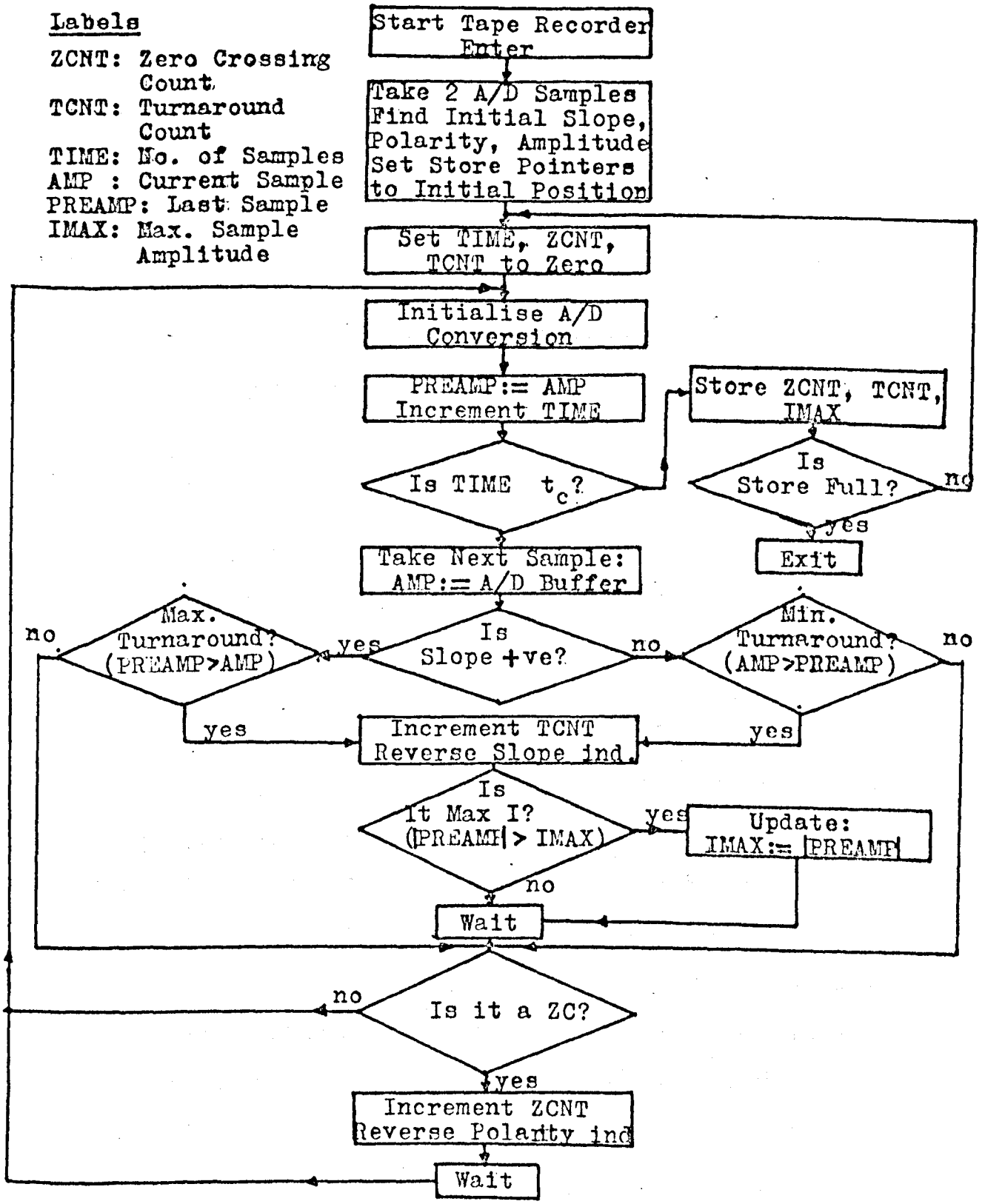


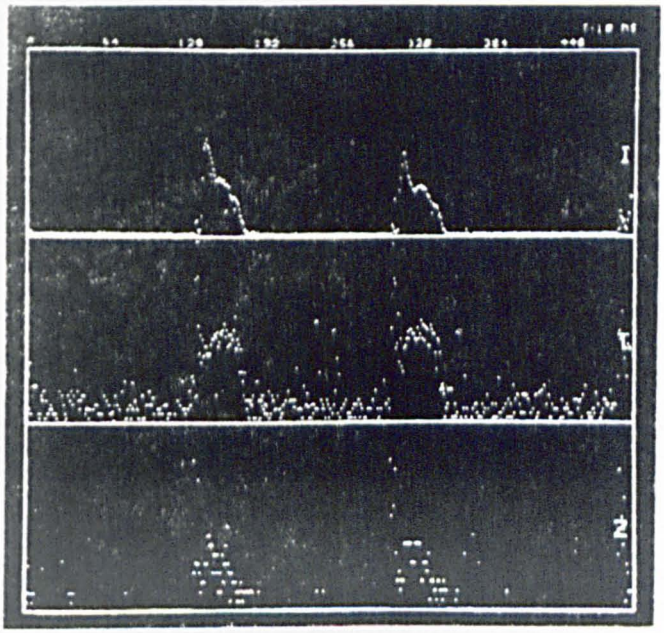
Fig. 2.4 Flow Diagram of the A-D Input Routine.

it was necessary to take time off to increment the corresponding counter, and to make the amplitude check in the case of a Turnaround. The computer therefore skipped one sample cycle by incrementing the time counter and waiting for the remainder of the sample time before continuing. The smallest period between Zero Crossings and Turnarounds which could be measured was therefore 100/s., though the time intervals could be measured to an accuracy of 50/s.

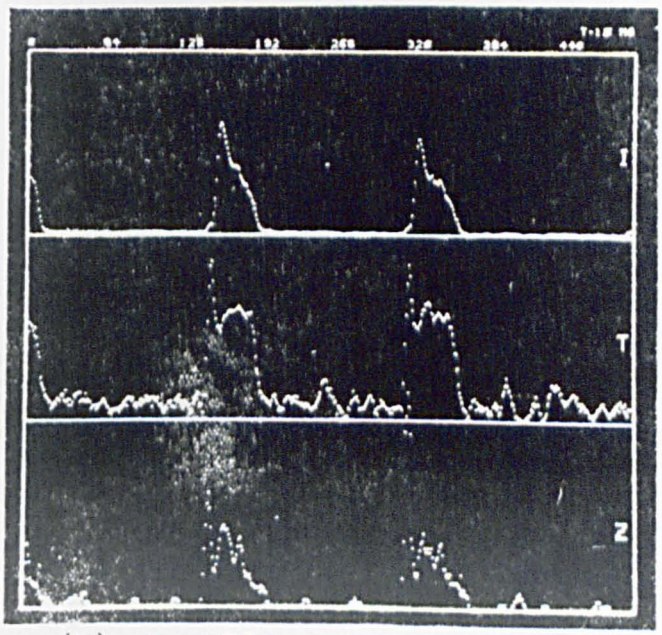
2.1.2. Smoothing Average.

The most popular interval (tc) between measurements of the Zero Crossing and Turnaround Rates in the Literature is 10 ms, roughly equivalent to a single pitch period. It was found that with this interval the raw Z.T. and I. traces were too irregular to be of much use - in particular it was hard to distinguish between the consonant and vowel part of the sound. Figure 2.5.(a) shows the unaveraged Z.T.I. diagram for the pair of /t3/ sounds of figure 2.2. with tc = 10 ms.

A short-time smoothing process was necessary to clear up the traces. The smoothing took the form of the replacement of each point by the average of n points round it. This had the disadvantage of throwing away some of the information contained in the rapid fluctuations of the traces, as seen in



(a) t_c = 10ms., n = 1



(b) t_c = 10ms., n = 5

Fig. 2.5 Effect of Variation of t_c & n.

figure 2.5.(b), (tc = 10 ms, n = 5) but this could be remedied to some extent by making tc smaller. The best compromise was found by experiment to be tc = 6.4 ms. and n = 5. The sampling time was then still less than 1% of tc. This combination was used in all the subsequent Z.T.I. diagrams.

2.1.3. Scaling.

The amplitude of the Z.,T. and I. functions was approximately equalised and constants were added to T. and I. to enable them to be plotted simultaneously on the 388 screen. Thus :

$$Z. := 2Z \text{ or } 4Z \ *$$

$$T. := T + 500_8$$

$$I. := I + 1200_8$$

---

* The utterances by subjects W.A.A. and M.A. generally had lower Z. values than those of subjects C.W.T. and P.D.G. The Z. scaling factor was therefore set to 4 for subjects W.A.A. and M.A. and .2 for subjects C.W.T. and P.D.G.

2.1.4. Routine for Finding the Start of a Sound.

Figure 2.6 is a block diagram of the algorithm used to find the starting points of a pair of sounds. The apparent starting point often differed slightly between the 3 functions Z.T. and I., and therefore each was treated separately. In particular, onset of the consonant on the I. trace was sometimes later than on the Z. and T. traces. The start pointer was therefore moved along from the beginning of the store (t = o) until both the Z. and T. values were above their respective Thresholds. The threshold values could be adjusted by using the Push Buttons.

The Pointer was next extrapolated back along each trace in turn until the trace fell below its Threshold, and then continued backwards until a significant minimum was found. The start marker for this trace was then put at this point. A significant minimum was defined as a minimum for which the slope did not reverse direction again at the next point, (see figure 2.7) thus excluding very small fluctuations in the trace. This procedure yielded start markers, Z.MARK, T.MARK and I.MARK. on the 3 traces. When these had been confirmed by eye, the start of the sound was set at the earliest of the 3 markers to occur. The other 2 traces were set to zero between this point and their respective start

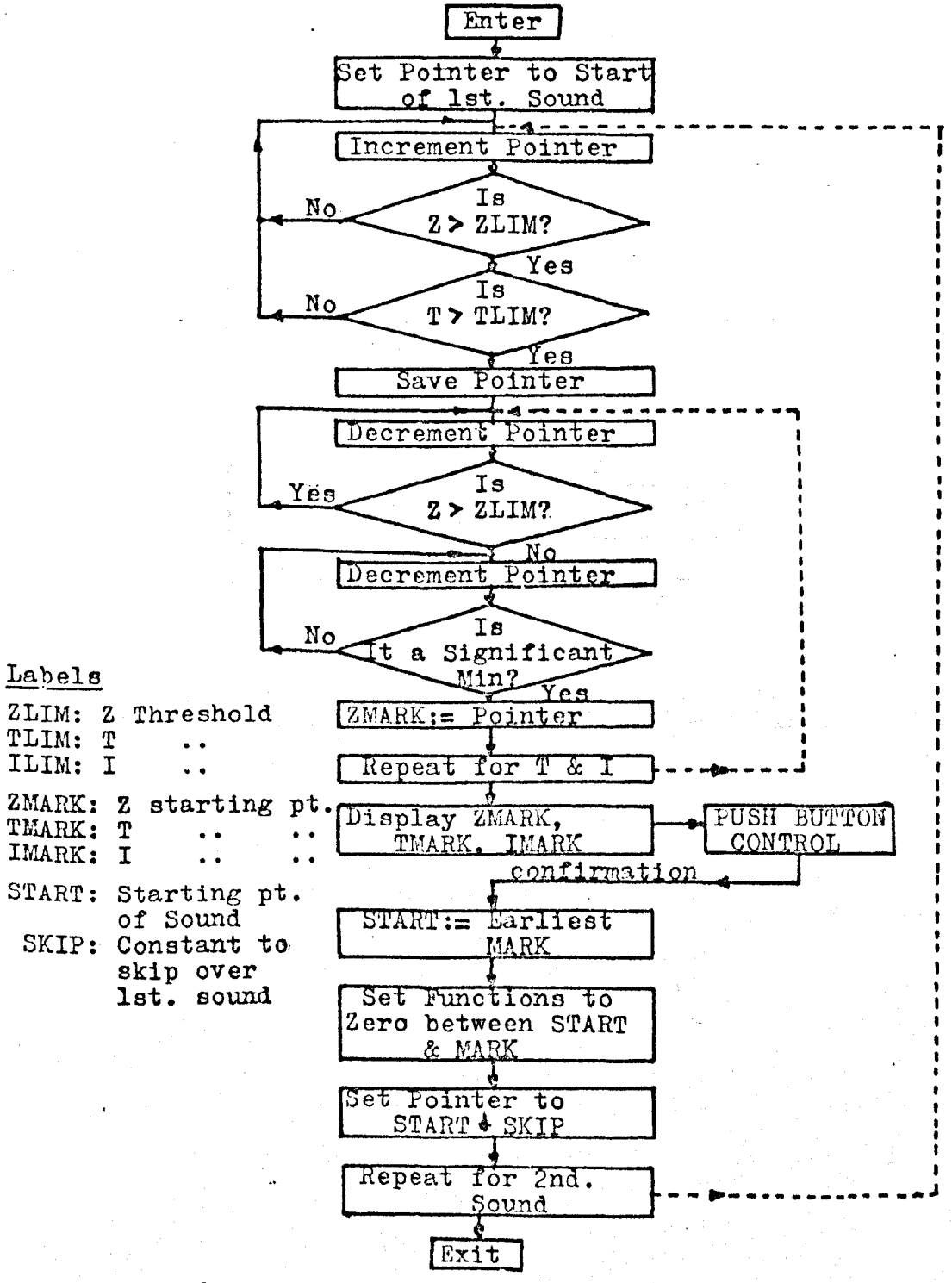


Fig. 2.6 Flow Diagram of the Starting Point Algorithm.

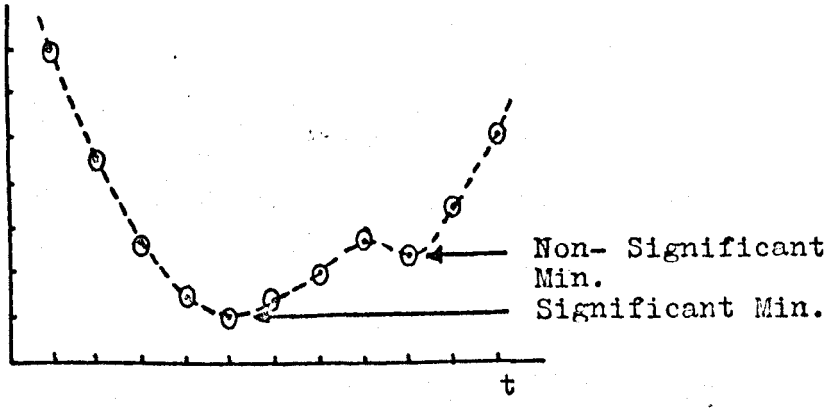


Fig. 2.7 Significant & Non- Significant Minima.

marks, (see figure 2.3.). On the I. trace, this interval was the parameter I.S.D. The widths and onset times of the sounds were measured from the start marks of the appropriate traces.

The procedure was repeated for the 2nd sound starting at a fixed distance from the start of the first sound.

2.1.5. Routine for Finding the Position of a Consonant Peak Maximum.

A block diagram of the consonant peak picking algorithm is shown in figure 2.8. An example of the working of the algorithm is shown in figure 2.9.

The procedure was to isolate the position of the consonant peak between upper and lower limits. The first step was to find the position of the overall maximum on each of the Z,T. and I. traces. Almost always one or more of these 3 maxima occurred on the vowel part of the sound. The upper limit for the consonant peak position on the Z. and T. traces was therefore set at the "latest" of these 3 maxima, (figure 2.9 (a)). In the case of the I. trace, it was necessary to restrict the consonant peak to the part of the sound before the maximum (see section 1.3.).
The upper limit on the I. trace was therefore placed at the position of the I. maximum. The lower limit on each trace

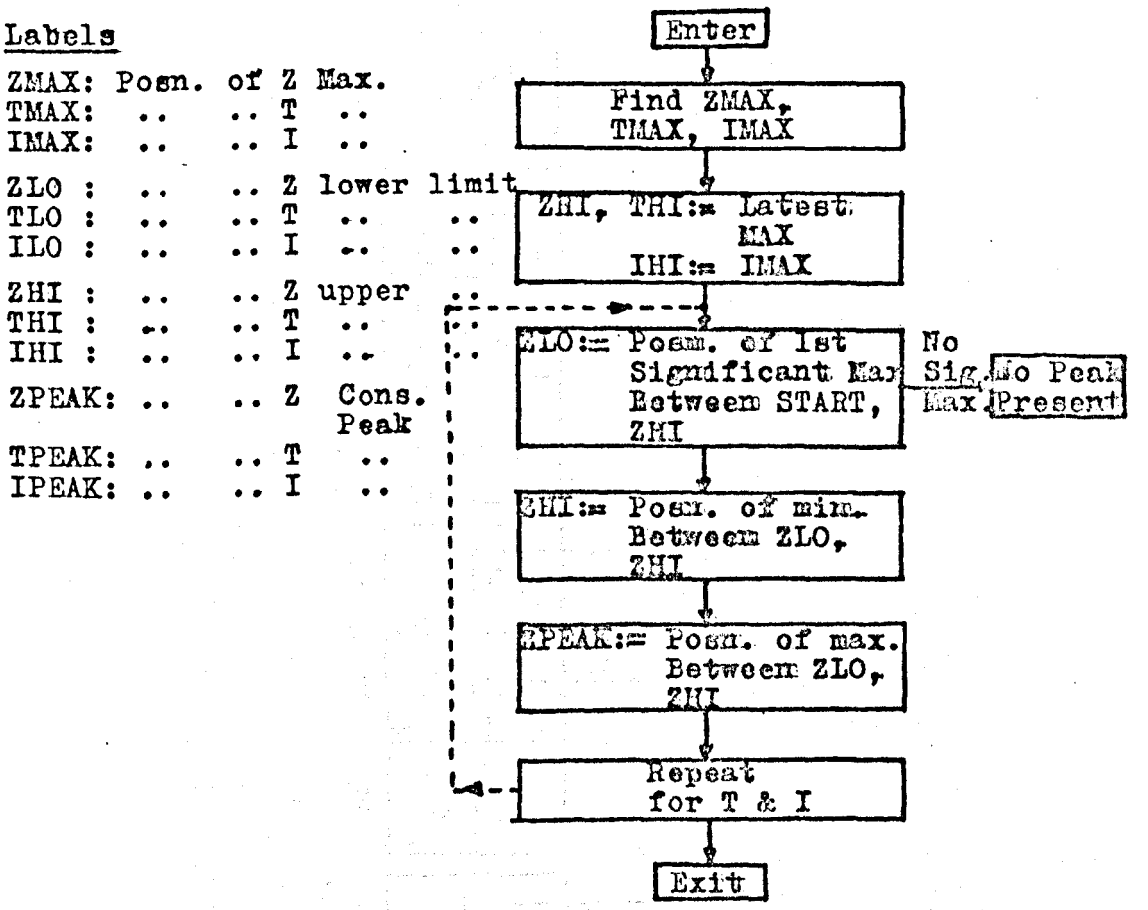


Fig. 2.8 Flow Diagram of the Consonant Peak- Picking Algorithm.

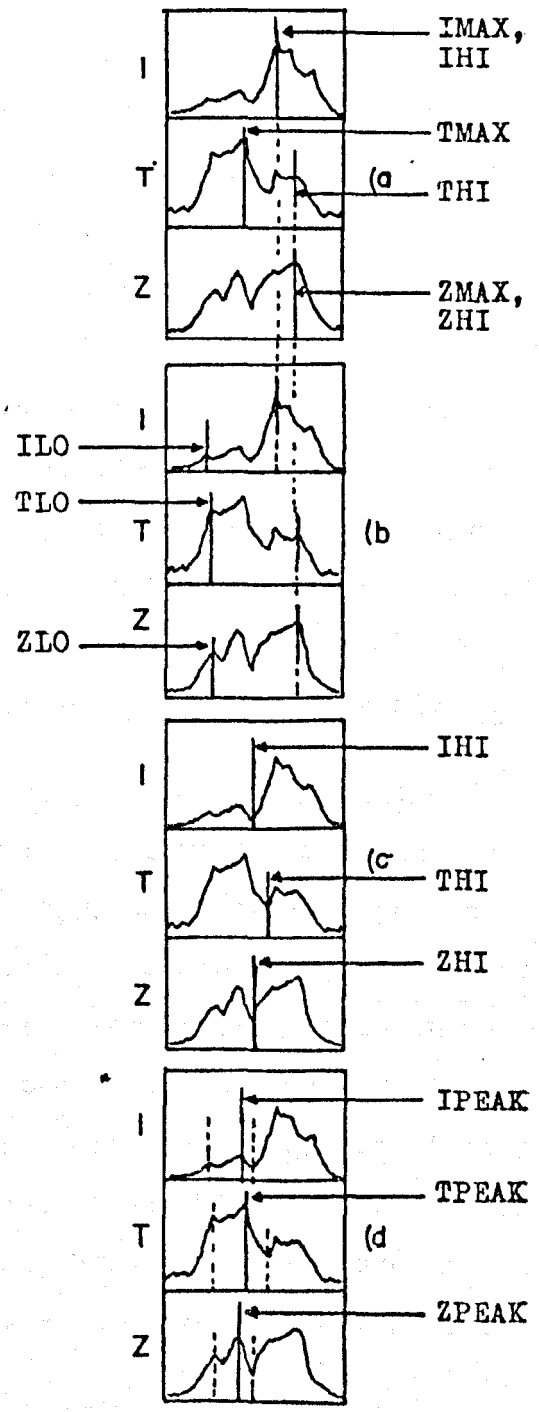


Fig. 2.9 Example of Consonant Peak Picking.

was then set at the position of the first significant maximum to occur. (A significant maximum was defined in a similar way to a significant minimum). This is shown in figure 2.9.(b). If no significant maximum was found before the upper limit, this was taken as an indication that no consonant peak was present.

The upper limit was then moved back to the minimum amplitude position between its previous position and the lower limit (figure 2.9c). The position of the maximum between the upper and lower limits was then taken as that of the consonant peak Maximum (figure 2.9(d)).

2.1.6. Procedure for Finding the Minimum and Vowel Peaks.

A block diagram of the algorithm used for picking the position of the C.V. Minima between the consonant and vowel parts of the Z,T. and I traces, and the vowel Maxima is shown in figure 2.10. The Minimum position was found first : the lower bound for the minimum position was placed at the position of the Consonant Maximum. The higher bound was advanced a set distance (60_8) from this point, large enough to ensure that it was placed on the vowel part of the trace.

The higher bound was then moved backwards until the value of the trace went above Threshold. This was meant to ensure

Labels

ZMIN : Posn. of Min Between Cons. & Vowel On Z Trace

TMIN : Ditto On T Trace

IMIN : .. .. I ..

ZVPEAK: Posn. of Vowel Max on Z Trace

TVPEAK: Ditto. T ..

IVPEAK: .. I ..

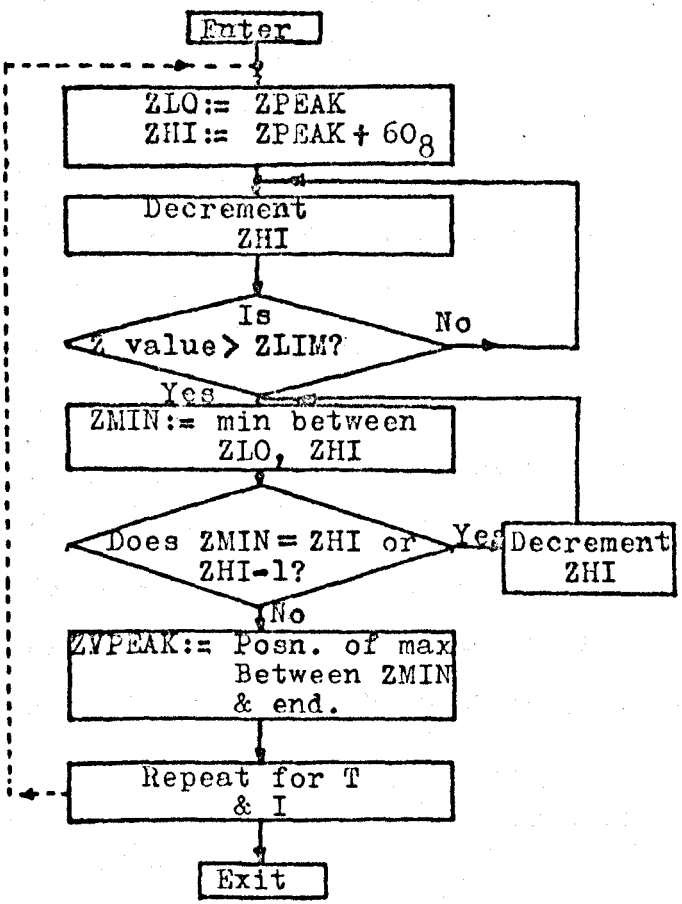


Fig. 2.10 Flow Diagram of the Algorithm for Finding the Consonant- Vowel Minimum & the Vowel Maximum.

that the C.V. Minimum was not placed in a false position on the trailing edge of the vowel. The C.V. Minimum was then put at the position of the minimum value between the upper and lower limits. This point was rejected if it occurred at the upper limit or 1 place in front of it, since this usually happened when the upper limit was still on the trailing edge of the vowel. In this case the upper limit was again moved backwards until a satisfactory C.V. Minimum was found.

This procedure worked well on the Z. and I. traces, but often failed on the T. trace. Since the vowel often had a longer duration on the T. trace, and fell away more gradually with many fluctuations, the minimum still tended to be placed on the trailing edge. For this reason the position of the upper limit for T. was taken to be the final value for I.

Once the C.V. Minimum had been found, the position of the Vowel Maximum was obtained by finding the maximum value between the C.V. Minimum and the end of the trace for this sound.

2.2. Recognition Algorithms.

All the Recognition Algorithms used were Binary Threshold Decision Trees. Figure 3.101 shows part of the algorithm used for the subject C.W.T. The path taken through the tree

was determined by sequential threshold decisions. If the parameter value was below or equal to the threshold, the left hand path was taken ; if it was greater than the threshold, the programme chose the right hand path.

Thus in figure 3.101, the first node(top of the diagram) asked whether I.P.S. was equal to 0. If this was true, the next query was "is T.P.S. greater or less than 101", and so on until an end point was reached. Each end point was associated with a particular phoneme or group of phonemes. In figure 3.101, the phonemes printed before the bracket at an end point were more likely to occur than those following the bracket. Thus "K(T" indicates a high probability for /k/ and a low probability for /t/. These 2 recognition classes will be termed "probable" and "possible" identifications.

The binary decision tree method is a very simple form of pattern recognition scheme. Each decision corresponds to a boundary plane in a parameter hyperspace. If each pattern category is associated with a single end point, the tree is equivalent to a parameter list method, since each category is then associated with a particular area in parameter space. In the present study , the method was generalised by allowing the consonants to be associated with more than one end point, and by distinguishing between probable and possible identifications.

The decision tree algorithms had the advantage of being relatively easy to programme, and could readily be adjusted to fit new data. A discussion of the possibility of using more advanced schemes is given in Section 4.3.

2.2.1. Group Numbers.

The decision tree nodes were identified by 3 digit Group Numbers, prefixed by the letter G. The first digit of the group number determined to which group the node belonged (see Section 3.2.1.). Figure 3.101 is a diagram of Group 1 for subject C.W.T. Thus the Group numbers for the nodes shown in figure 3.101 all have 1 as the first digit. The next digit was determined by the decision level within the group, and the 3rd. by the number of the node in this level, reading from left to right. Thus in figure 3.101 the decision "T.P.S. above or below 101" on the top left of the diagram was referred to as G121, and so on. In figure 3.101, the last 2 digits of the Group No. are printed on the appropriate branches of the tree.

2.2.2. Design of the Algorithms.

The Decision Algorithm had to separate the sounds into 23 categories using 13 parameters. Since this involved a

great number of possible decisions, as much of the design process as possible was automated.

The Decision Algorithm programme to be adjusted was initially held in core. This could be either a "blank" algorithm (with the possible branches connected but no decisions programmed), or one which had been used previously. The data consisted of the parameter sets for the new body of sounds, and was stored on Dectape (see Section 2.1.).

The method was to look at one node at a time starting from the first decision in the tree and working down through the branches to the end points. When the optimum decision for a particular node had been found, this was inserted in the algorithm and the computer proceeded to deal with the next. New nodes could be inserted in the algorithms if necessary by adjusting the programmed connections.

Figure 2.11 is a block diagram of the procedure for optimising a single decision node. The adjustment of the node G133 of figure 3.101 will be used as an example.

The computer first read the group No. of the node. A "trap" was then programmed into the algorithm at this node; this trap was activated whenever the node was used, and caused the parameter set of the sound being run to be held in a buffer store. The parameter sets of all the sounds to be considered were read down in turn from Dectape, and each

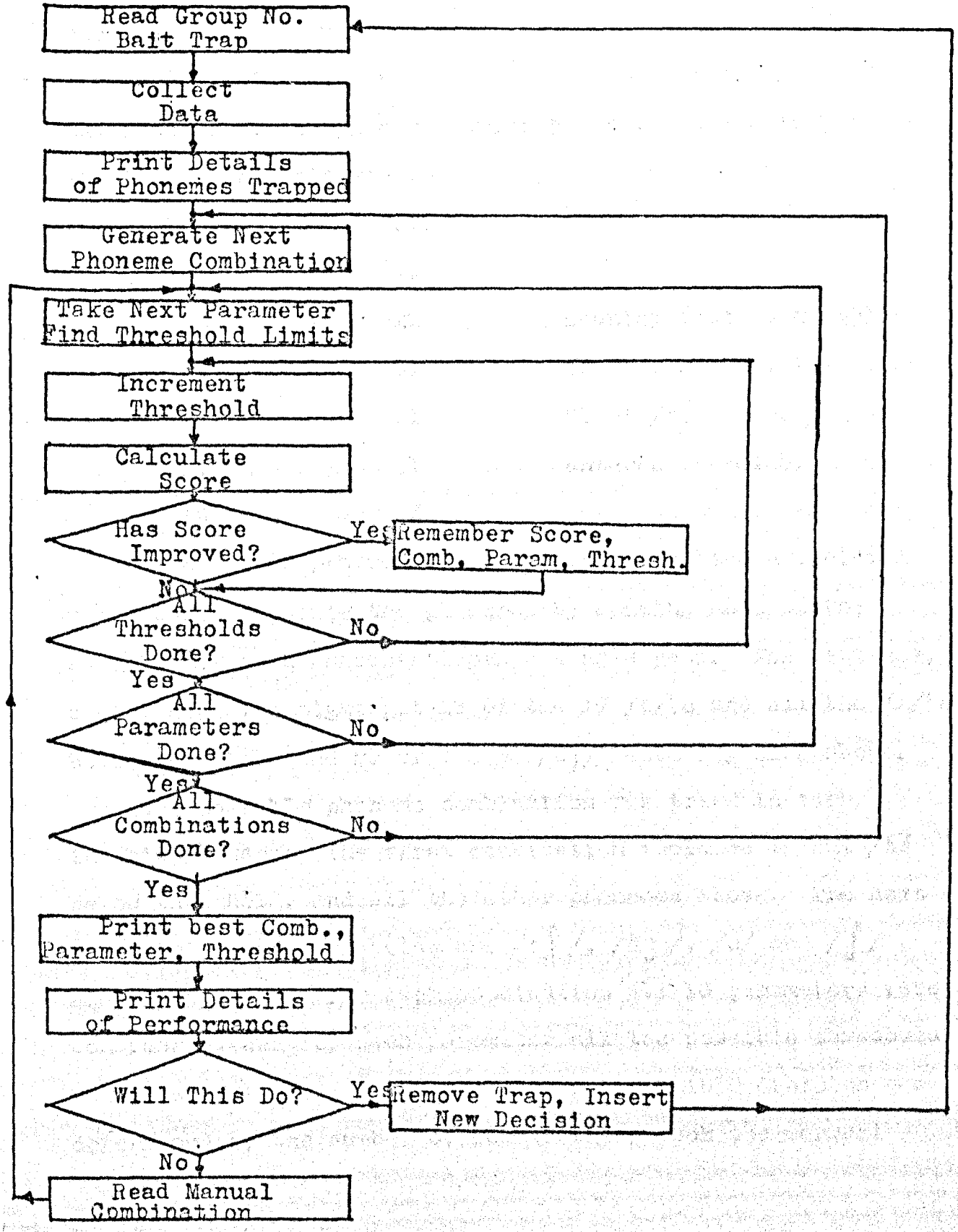


Fig 2.11 Flow Diagram for the Adjustment of a Decision Node.

sound was run through the algorithm.

When all the data had been passed through the algorithm, the details of the phonemes which had entered the trap were printed on the Teletype thus :

PHON.	C.T.
2	19
3	22
6	31
13	1
14	1
15	3

Meaning that 19 examples of /t/, 22 examples of /k/, 31 of /tʃ/ etc. had entered G133.

The computer programme now attempted to find a decision which would separate the phonemes by sending most of the examples of each consonant down a single path. For instance, a good decision might put 18 of the 19 /t/'s and all the /k/'s below threshold and 29 of the 31 /tʃ/'s entering G133 above.

Each possible phoneme combination was tried in turn. In this instance, the first combination would be to put /t/ below threshold, and all the other phonemes above. The next would be /t/ and /k/ below and /tʃ/, /dʒ/, /s/ and /ʃ/ above and so on. For each combination all 13 parameters were considered, and for each parameter all the possible threshold values were tried. The score for each possible decision was determined by the number of sounds which took the correct

path according to the phoneme combination. The computer chose the decision giving the highest score as a tentative suggestion.

The threshold limits for a single parameter were determined by the lowest and highest values of the parameter found in the appropriate data. The best threshold value was set at the centre of the highest scoring range ; thus if values of T.P.S. of 120 to 132 gave the best score, the threshold was set at T.P.S. = 126.

The suggested decision was then printed on the Teletype thus :

BELOW.... 2, 3, 13 ABOVE 6, 14, 15.

BEST PARAM..... 4 THRESH..... 6.

PHON.	C.T.
2	16
3	18
6	30
13	1
14	1
15	3

PUT IN TROG?...

i.e. putting /t/, /k/ and /ʤ/ below, /tʃ/, /s/ and /ʃ/ above, the Best decision is Z.L.E. = 6. This gives 16 instances of /t/ correct, 18 of /k/ and so on.
'TROG' is the programme name for the recognition algorithm.

Will this suffice.? An answer of Y (es) caused this decision to be entered in the algorithm. If the answer was N(o), the computer would read a phoneme combination manually, and print out the best parameter and threshold for this combination. This process was repeated until a satisfactory decision was found.

The number of possible combinations became prohibitively large when more than 7 phonemes had to be considered for a single node. For this reason, the programme ignored phonemes for which a single example was trapped. Thus in this instance /dʒ/ and /s/ would not have been considered. Instances where there were still 7 or more phonemes to be considered were dealt with manually from outset.

Most of the unsatisfactory decisions suggested by the computer involved almost all of the sounds going to one side of the node. For instance, the following decision :

BELOW....2,3,6,13,15. ABOVE...14.

BEST PARAM..... 5. THRESH..202.

PHON.	C.T.
2	19
3	22
6	31
13	1
14	1
15	2

would have a higher score than the correct decision given above, but only I /s/ and 1/ʃ/ take the right hand path. This difficulty could be remedied by giving a higher score to those decisions which divided the data more evenly, but this facility was not programmed in due to lack of time and space.

It was sometimes necessary to ignore one or more of the phonemes to find the best decision. In these cases the automatic method failed;for example

PHON	C.T.
2	3
3	4
6	27
16	2
17	4

BELOW.... 2, 3, 6, 16, 17. ABOVE.... 6.

BEST PARAM.... 5. THRESH....190.

PHON.	C.T.
2	3
3	4
6	27
16	1
17	3

PUT IN TROG..... N.

Here it proved impossible to separate /tʃ/ from the

remaining phonemes, but the decision :

BELOW..... 2, 3. ABOVE..... 16, 17.

BEST PARAM.... 3. THRESH..... 20.

PHON.	C.T.
2	3
3	3
16	2
17	4

PUT IN TROG ? .... Y

separated /t/ and /k/ from /z/ and /ʒ/, /tʃ/ falling equally between the 2 paths.

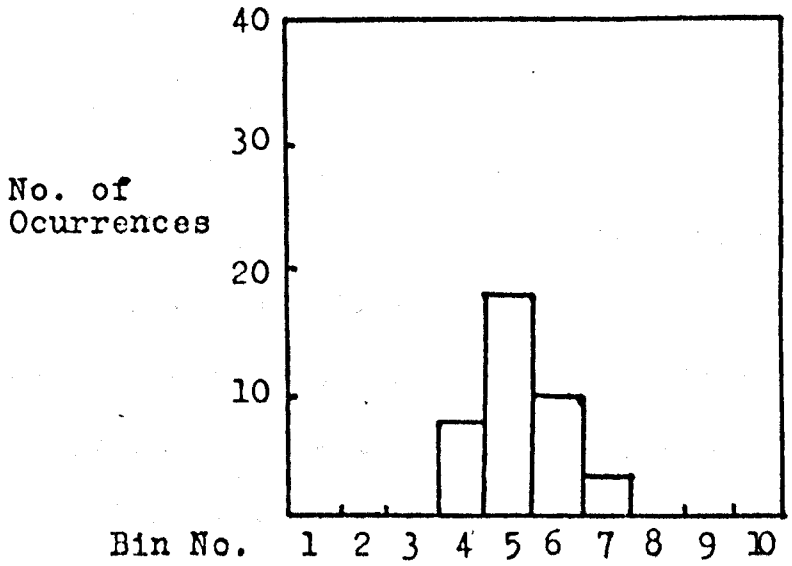
CHAPTER 3.

3.1. The Z.T.I. Diagrams.

The following section examines the Z.T.I. diagrams obtained for each consonant and each subject. The phonemes are considered in turn; the utterances of C.W.T. are dealt with in detail and subject differences are explained subsequently where appropriate.

The Z.T.I. diagrams used to illustrate this section have been photographed directly from the 338 screen. The 3 vertical lines superimposed on each trace indicate the estimated positions of the consonant and vowel Maxima and the Minimum between consonant and vowel. These lines are not plotted in cases where the consonant peak is absent. The Z. traces for subjects W.A.A. and M.A. are plotted at double the size of those for subjects C.W.T. and P.D.G.

Extensive use is also made of Histograms depicting the distribution of a single parameter for a particular phoneme and subject. These diagrams have also been photographed from the 338 screen, and their scales are somewhat blurred. The scales of the bin divisions used for each parameter are therefore set out in figure 3.1.



Scale for Each Parameter		Bin Divisions For Each Scale			
Param.	Scale	Bin No.	Scale a	b	c
1/ ZPS	a	1	0- 24	0-9	0- 4
2/ ZPD	b	2	25- 49	10- 19	5- 9
3/ ZPW	c	3	50- 74	20- 29	10- 14
4/ ZLE	c	4	75- 99	30- 39	15- 19
5/ TPS	a	5	100- 124	40- 49	20- 24
6/ TPD	b	6	125- 149	50- 59	25- 29
7/ TPW	c	7	150- 174	60- 69	30- 34
8/ TLE	c	8	175- 199	70- 79	35- 39
9/ IPS	b	9	200- 224	80- 89	40- 44
10/ IPD	b	10	$\geqslant$225	$\geqslant$90	$\geqslant$45
11/ IPW	c				
12/ ILE	c				

Fig. 3.1 Bin Divisions used in the Parameter Histograms.

3.1.1. /p/.

3.1.1.1. Subject C.W.T.

Figure 3.2 shows the Z.T.I. diagrams for two utterances of /pI/ by subject C.W.T. The consonant peaks on the Z. and T. traces were usually small and fairly narrow (the Z. peak is indistinct in figure 3.2(b)). This was due to the concentration of the energy in the lower regions of the spectrum (see section 1.4.2.1).

The striking feature of the Z.T.I. diagrams of /p/ for this speaker was the extremely large and sharp consonant peaks seen on the I. trace. In figure 3.2(b), the I. peak is much higher than that of the following vowel: that of figure 3.2(a) is less pronounced, but still has a high I.P.S. value.

Figure 1.4(a) is the sonagram for the /pI/ sound of figure 3.2(b). This sonagram shows that while a large amount of energy was present in the explosion of /p/, this energy was concentrated in a quite narrow region at the base of the pattern. In unvoiced sounds, the energy is normally spread over a much wider spectral range. The effect of this massive concentration of energy at low frequency on the envelope of the stop sound can be seen in figure 3.5, which is a U.V. recording of the raw speech waveform for the /pI/ sound of figure 3.2(b).

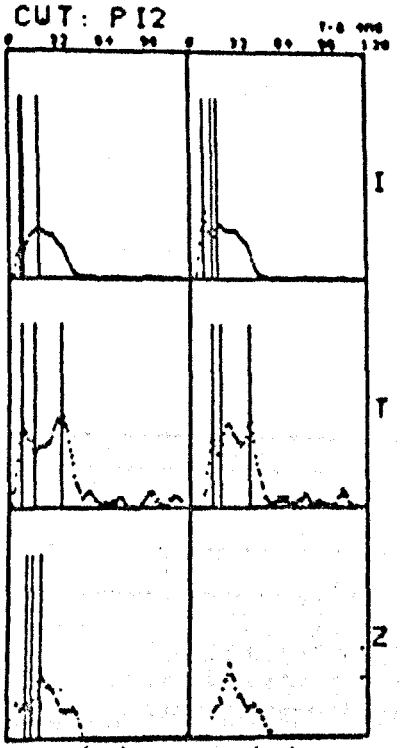


Fig. 3.2

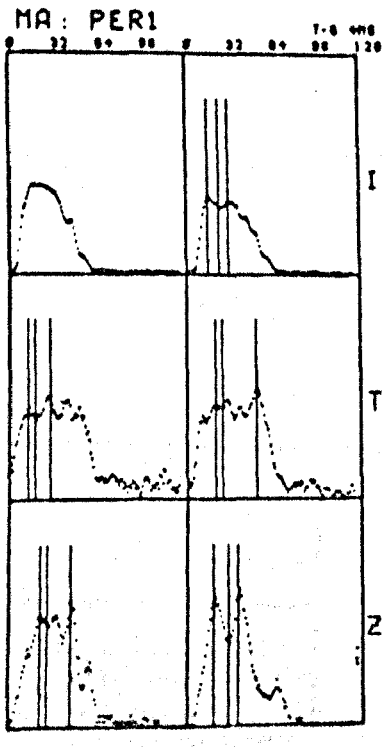


Fig. 3.4

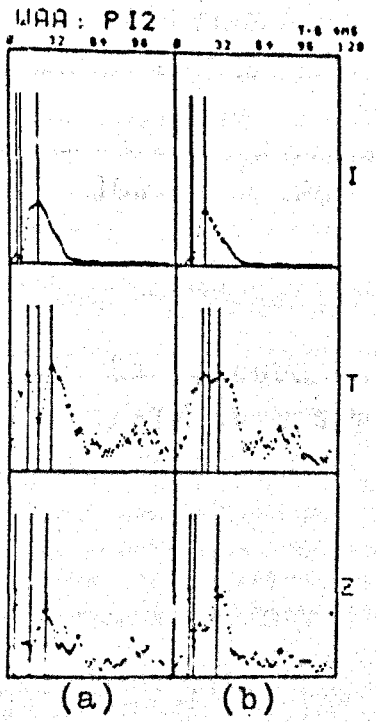


Fig. 3.3

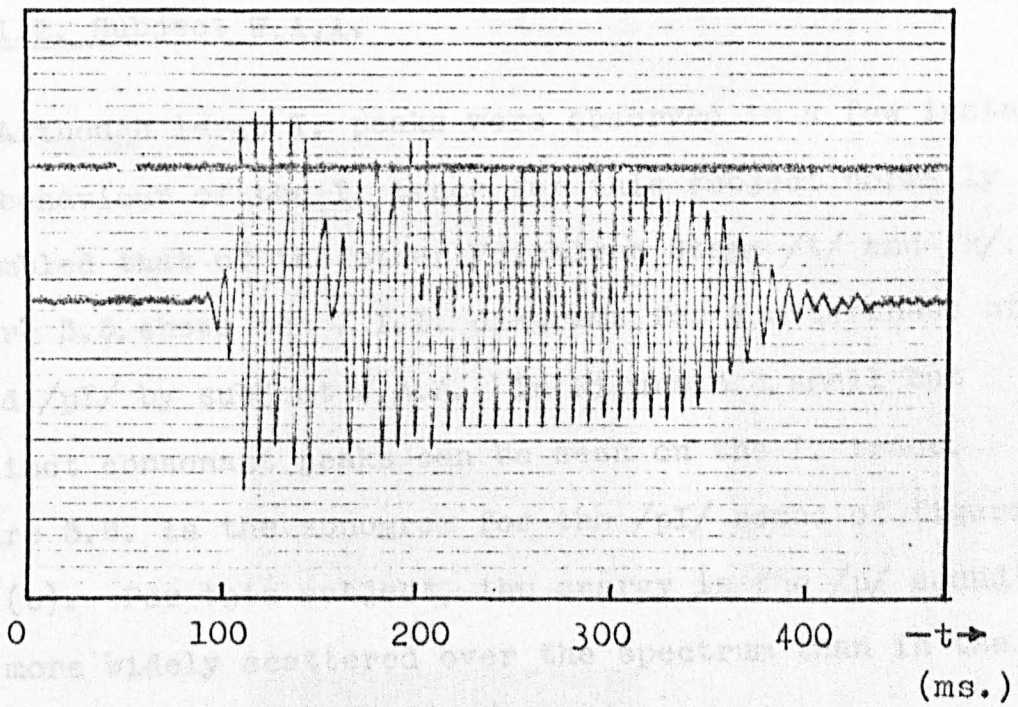


Fig. 3.5 Raw Speech Waveform for an Utterance of /pI/ by C.W.T.

Figure 3.6. shows the distribution of I.P.S. for /p/ spoken by C.W.T. Roughly 73% of the utterances of /p/ considered had values of I.P.S. greater than 70. This provided a ready means of identifying /p/, since very few of the other utterances had such high values of I.P.S.

3.1.1.2. Subject W.A.A.

Although large I. peaks were observed in a few instances, the behaviour of the I. trace for this subject normally resembled that of the other Voiceless Stops /t/ and /k/. Figure 3.3, shows the Z.I.I. diagrams for 2 utterances of the sound /pI/ by subject W.A.A. In Figure 3.3, small but distinct consonant peaks can be seen on the I. trace. Figure 3.8. is the sonogram for the /pI/ sound of figure 3.3.(b). For this subject, the energy in the /p/ sound was far more widely scattered over the spectrum than in the case of subject C.W.T.

The Z.and T. traces also showed peaks similar in form to those of /t/ and /k/; the Z. peaks were often quite large (Z.P.S. greater than 100), as seen in figure 3.3 . but tended to be narrower than those of /k/ (see figure 3.7). The T. peaks remained much smaller than those of /t/.

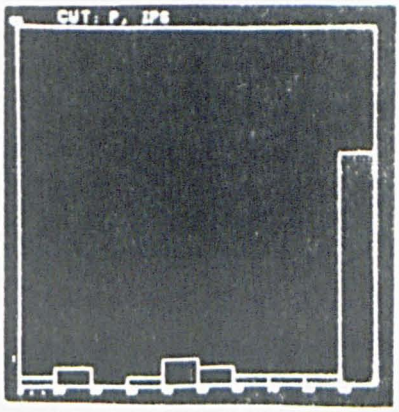
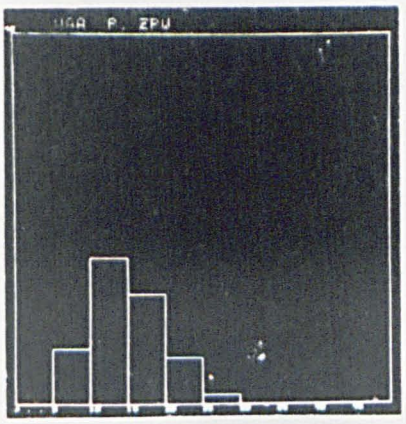
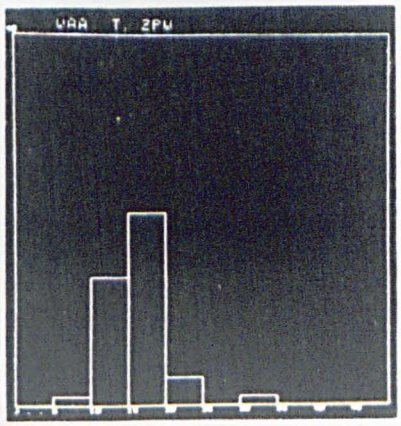


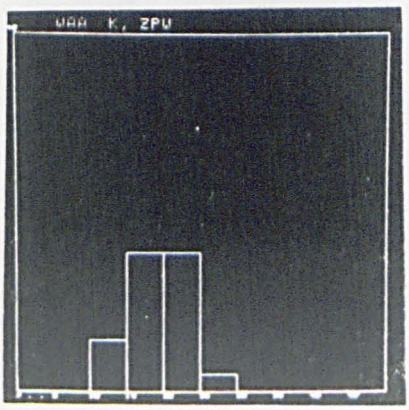
Fig. 3.6



(a)



(b)



(c)

Fig. 3.7

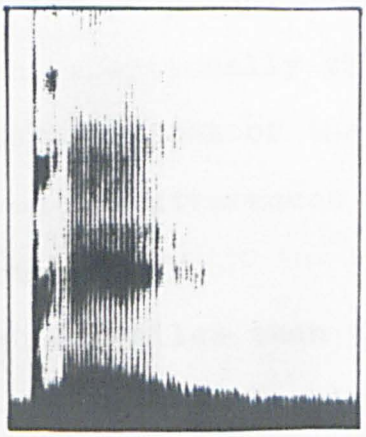


Fig. 3.8 Sonagram for an utterance of /pI/ by W.A.A.

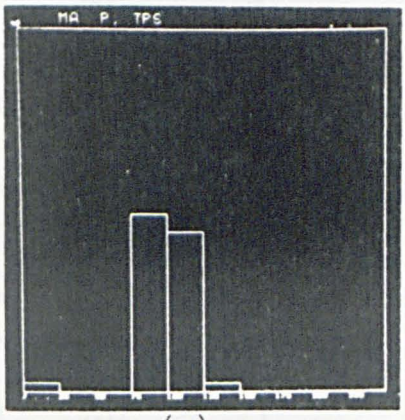
3.1.1.3. Subject M.A.

The Z.T.I. diagrams for the utterances of /p/ by subject M.A. showed variable behaviour on the I. trace. About 65% had large I. peaks similar to those of subject C.W.T., but less distinct, while for most of the remainder no distinct I. peak could be identified. Figure 3.4 is the Z.T.I. diagram for a pair of utterances of /pɜ/ in which both these types of behaviour can be seen. The overall energy level in the /p/ sound was generally low compared to that of subject C.W.T. This coupled with an exceptionally short duration led to the absence of I. peaks in some of the utterances. A similar effect was observed for utterances of other phonemes by subject M.A., notably /t/.

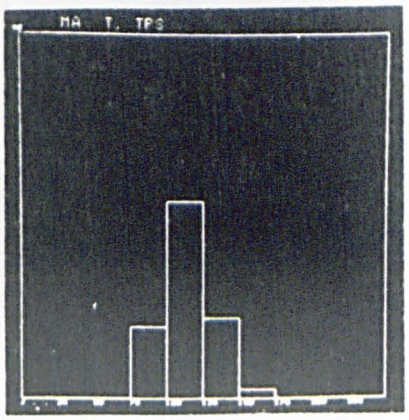
The Z. peaks remained much smaller than those of /t/ and /k/, but the T. peaks were much closer in size to these phonemes than normal (see figure 3.9.).

3.1.1.4. Subject P.D.G.

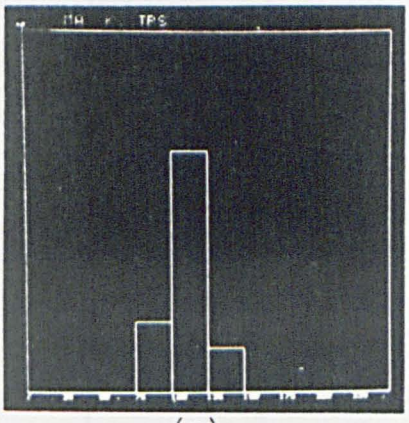
The Z.T.I. diagrams of /p/ for this subject were characterised by massive I. peaks even larger than those obtained for subject C.W.T. About 80% of the utterances had values of I.P.S. greater than 80, and the lowest I.P.S. value found was 65. The consonant-to-vowel



(a)

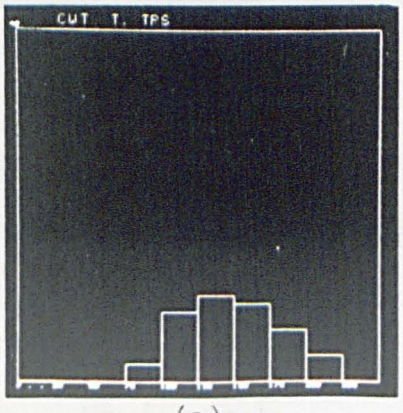


(b)

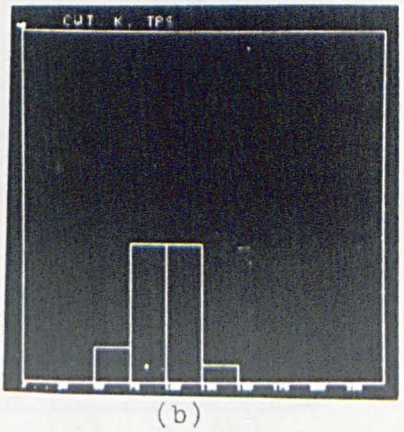


(c)

Fig. 3.9



(a)



(b)

Fig. 3.10

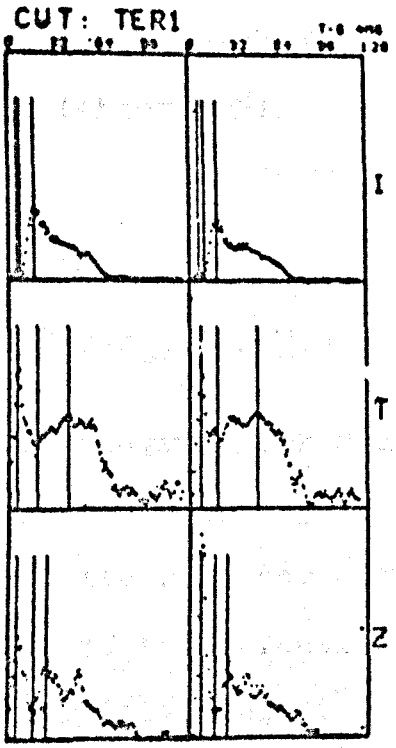


Fig. 3.11

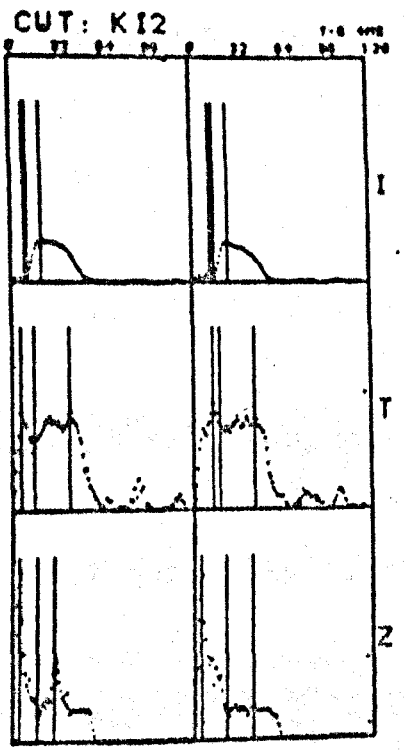


Fig. 3.13

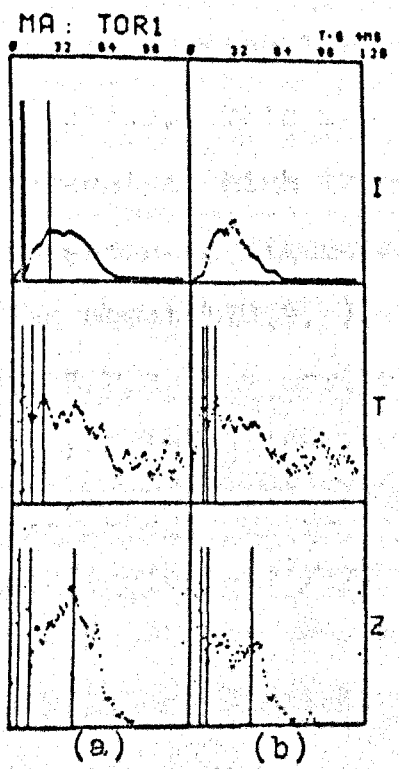


Fig. 3.12

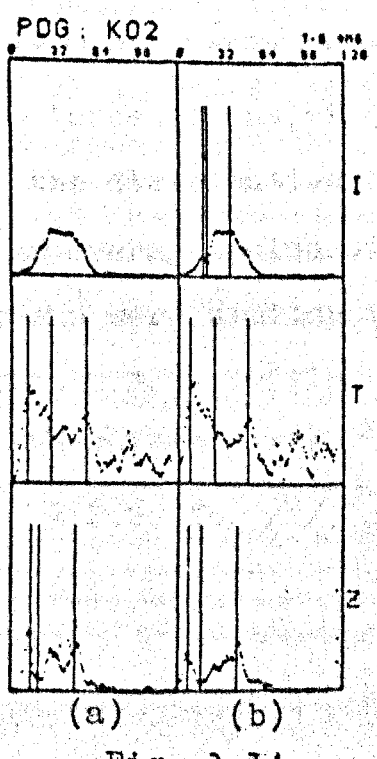


Fig. 3.14

change on the I. peak was very sharp, and I.P.D. sometimes quite low (about 30%).

3.1.2./t/.

3.1.2.1. Subject C.W.T.

Z.T.I. diagrams for 2 utterances of /t3/ by subject C.W.T. are shown in figure 3.11. The consonant peaks for /t/ were quite wide for a stop sound, but generally narrower than those of the Fricatives.

/t/ for this subject normally had the small but distinct I. peaks characteristic of voiceless sounds, though in a few cases these were absent. The T. peaks were uniformly high and sharp, /t/ having larger values of T.P.S. than those of the other Stops. This was due to the large amount of energy present at high frequencies; the distribution of T.P.S. is shown in figure 3.10(a) (the average value of T.P.S. was about 140.?,). The Z. peaks were similar in form, but Z.P.S. was more variable than T.P.S., while generally remaining larger than for the Voiced Stops. Figure 3.15 shows the distribution of Z.P.S. for /t/, and its voiced equivalent /d/.

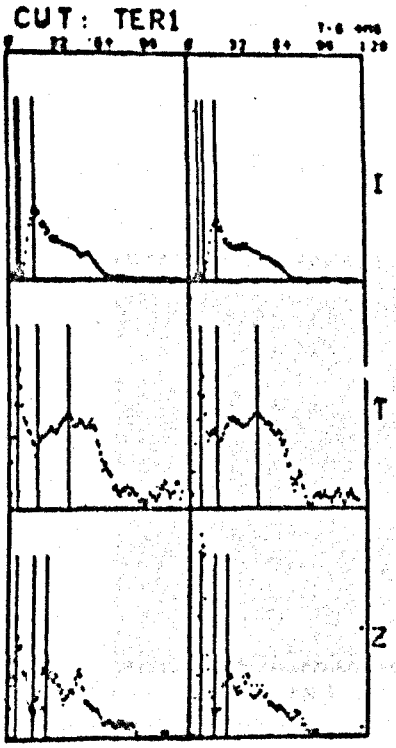


Fig. 3.11

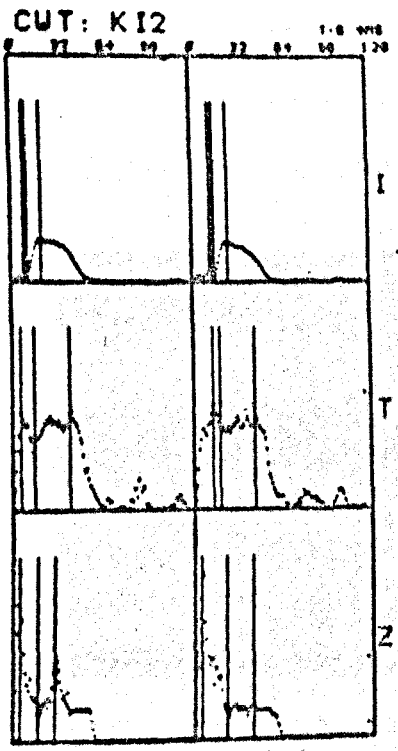


Fig. 3.13

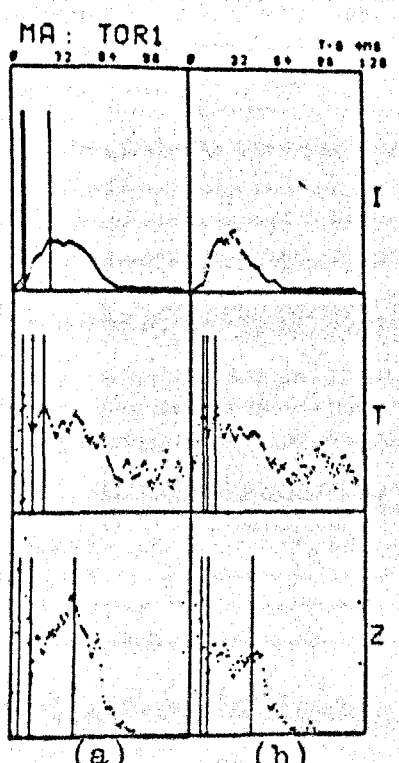


Fig. 3.12

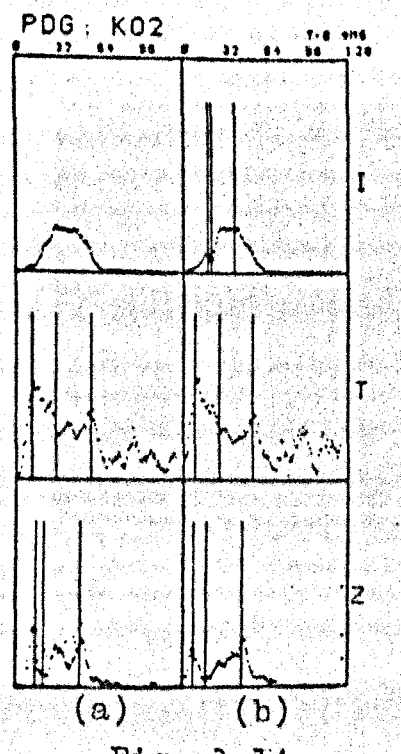
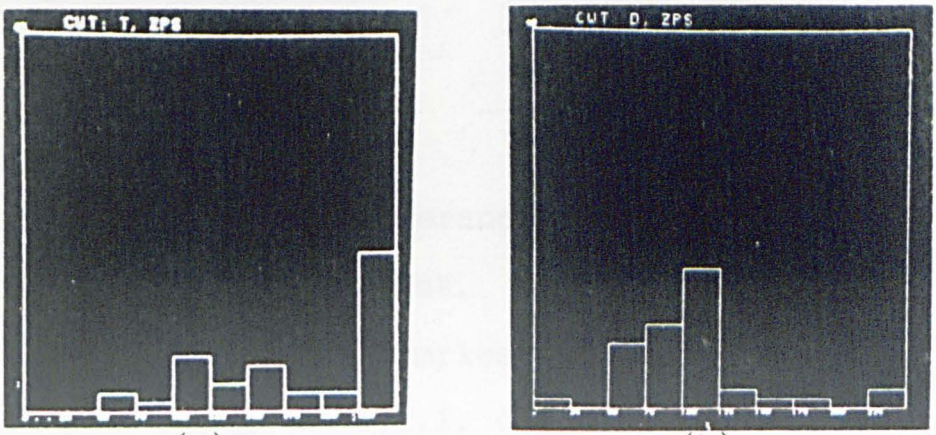
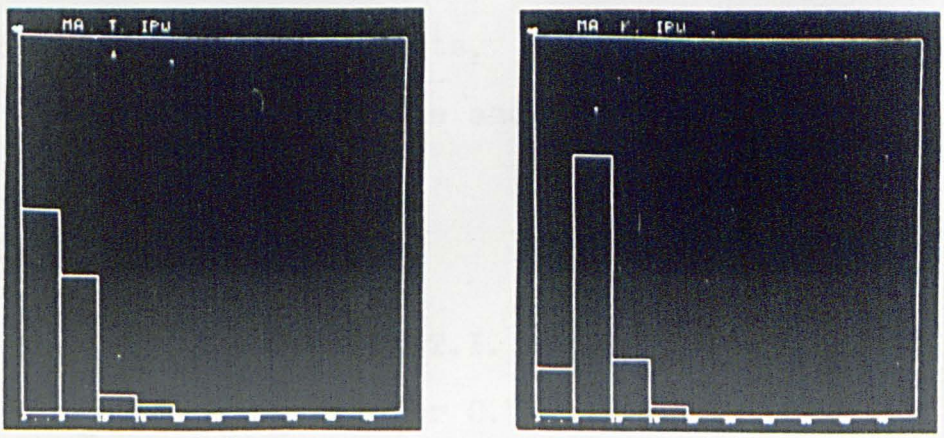


Fig. 3.14



(a) (b)

Fig. 3.15



(a) (b)

Fig. 3.16

3.1.2.2. Subject W.A.A.

The Z.T.I. diagrams for this subject were similar in form to those spoken by C.W.T.

3.1.2.3. Subject M.A.

Like /p/, about 50% of the utterances of /t/ for subject M.A. did not show a separate I. peak. The size of the Z. peak was again variable, those examples with no I. peak and a small Z. peak resembling the Z.T.I. diagram for the voiced cognate of /t/, /d/, in an exaggerated form. The T. peaks were also smaller than normal for /t/, the average value of T.P.S. being about 110. (See figure 3.10(b)).

The size of the Z. and T. peaks for this subject was ganerally reduced for all consonants. Two examples of Z.T.I. diagrams for the utterance /tɔ/ are shown in figure 3.12.

3.1.3.4. Subject P.D.G.

The behaviour of /t/ on the Z.T.I. diagram for this subject resembled that observed for C.W.T. The Z. peaks were universally large (Z.P.S. greater than 200), and the T. peaks a little bigger than those of C.W.T. The average value of T.P.S. was about 160.

3.1.3./k/.

3.1.3.1. Subject C.W.T.

The Z. and I. peaks obtained for this subject were similar to those obtained for /t/. However, /k/ generally had less prominent T. peaks than /t/, showing that the energy in /k/ was concentrated at lower frequencies, (see section 1.4.2.3).Z.T.I. diagrams for 2 utterances of /kI/ are shown in figure 3.13. Figure 3.10 shows the distributions of T.P.S. for /t/ and /k/.

3.1.3.2. Subject W.A.A.

The behaviour of /k/ for this subject was similar to that for subject C.W.T. The Z. peaks for /k/ tended to be wider than those of /p/ and /t/. The distributions of Z.P.W. for these 3 phonemes are shown in figure 3.7.

3.1.3.3. Subject M.A.

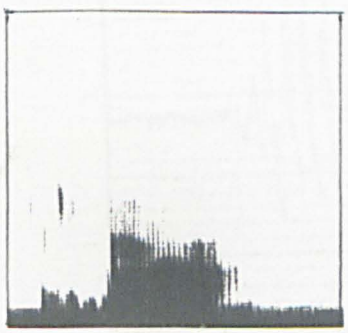
The form of the Z.T.I. diagram for /k/ was again similar to that obtained for C.W.T., but the T. peaks were generally no smaller than those of /t/, (see figure 3.9), /k/,however, showed more prominent I. peaks than /t/. Figure 3.16 shows the distributions of I.P.W. for /t/ and /k/.

3.1.3.4. Subject P.D.G.

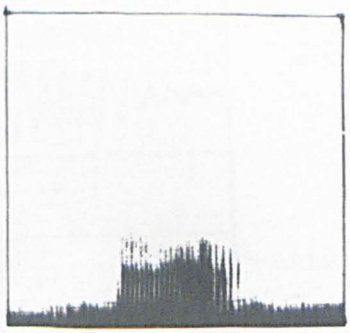
The Z.T.I. diagrams for this subject again resembled those of /k/ for subject C.W.T. Figure 3.14 shows the Z.T.I. diagrams for 2 utterances of /kɑ/. About 15% of the utterances, (for example figure 3.14(a)) did not show a separate I. peak. The soragrams for the 2 sounds of figure 3.14, shown in figure 3.17, show that much more energy was present in the second sound than the first, corresponding to a more pronounced explosion of the Stop.

3.1.4. The Voiced Stops.

The distinctive feature of the Z.T.I. diagrams for the Voiced Stops /b/, /d/ and /g/ was that for all subjects these phonemes did not normally show a distinct I. peak. The sonograms of figure 1.4 show that the duration of the Voiced Stops was very short, and the amount of energy present in the explosion of the Stop was relatively small. The envelope of the waveform generally rose quickly but smoothly into the following vowel; this can be seen in the U.V. recording of figure 3.18. I. peaks occasionally did occur for the Voiced Stops, due to increased duration and to a greater proportion of fricative excitation.



(b)



(a)

Fig. 3.17 Sonagrams for two utterances of /kɑ/ by P.D.G.

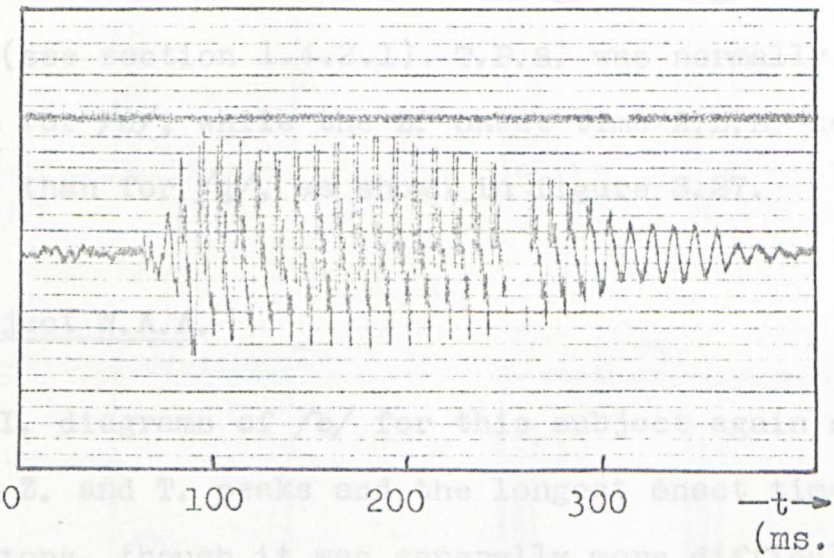


Fig. 3.18 Raw Speech Waveform for an Utterance of /dɒ/ by C.W.T.

3.1.5. /b/.

3.1.5.1. Subject C.W.T.

Figure 3.19 shows Z.T.I. diagrams of 2 utterances of the sound /bɑ/ spoken by C.W.T. As figure 3.19 illustrates, both the Z. and T. peaks for /b/ were relatively small, the Z. peaks sometimes being absent. This corresponds to the low FI value and the absence of energy at high frequencies (see section 1.4.2.1). T.P.S. was normally much smaller than for /d/, while the Z. onset time Z.L.E. tended to be larger than for /g/, as shown in figure 3.27.

3.1.5.2. Subject W.A.A.

The Z.T.I. diagrams of /b/ for this subject again showed the smallest Z. and T. peaks and the longest onset times of the Voiced Stops, though it was generally more difficult to distinguish between /b/, /d/ and /g/ than in the case of C.W.T. Figure 3.28(b) shows the distribution of T.L.E. for /b/.

3.1.5.3. Subject M.A.

The behaviour of /b/ on the Z.T.I. diagram once more resembled that observed for subject C.W.T.

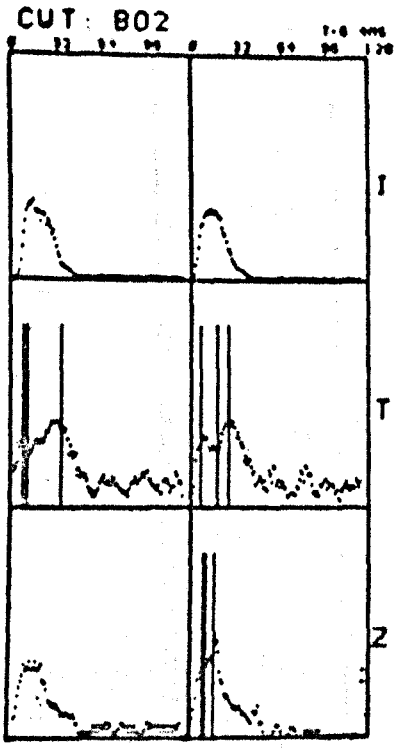


Fig. 3.19

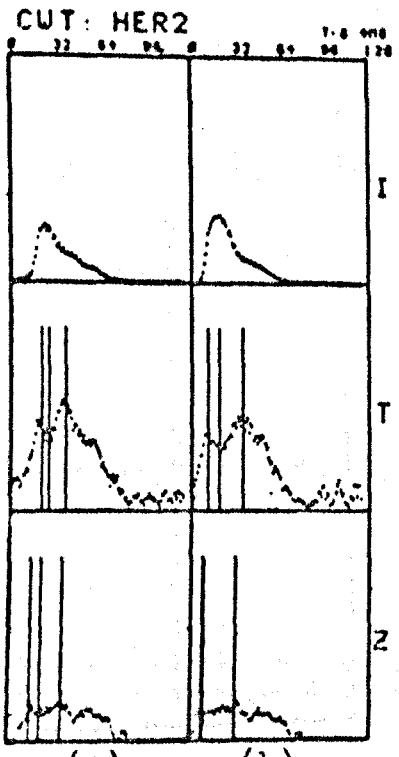


Fig. 3.21

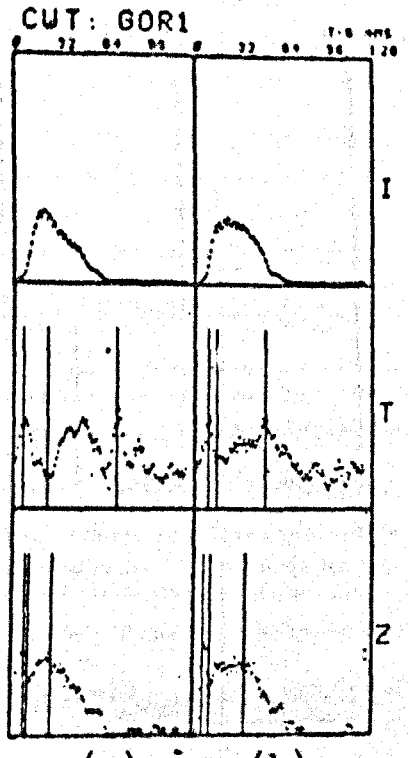


Fig. 3.20

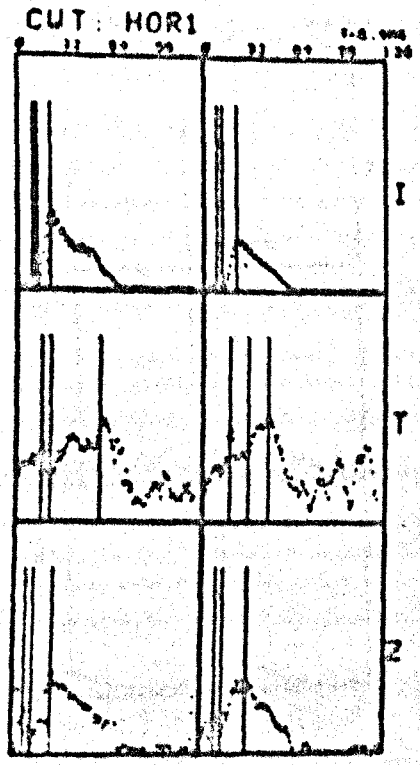


Fig. 3.22

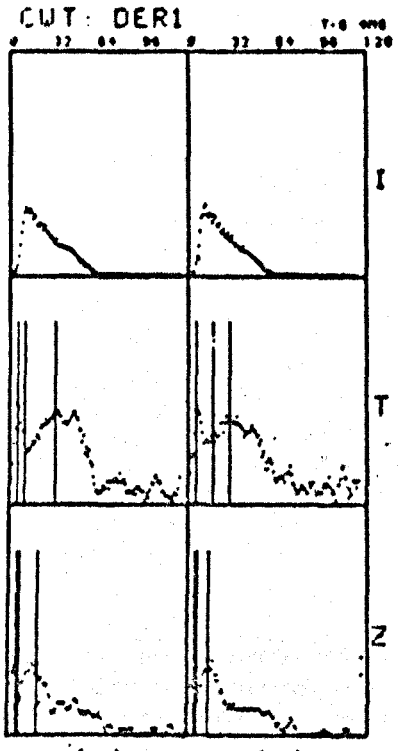


Fig. 3.23

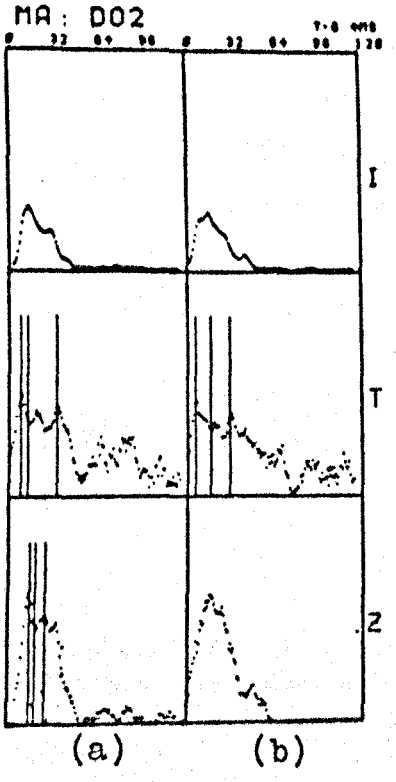


Fig. 3.25

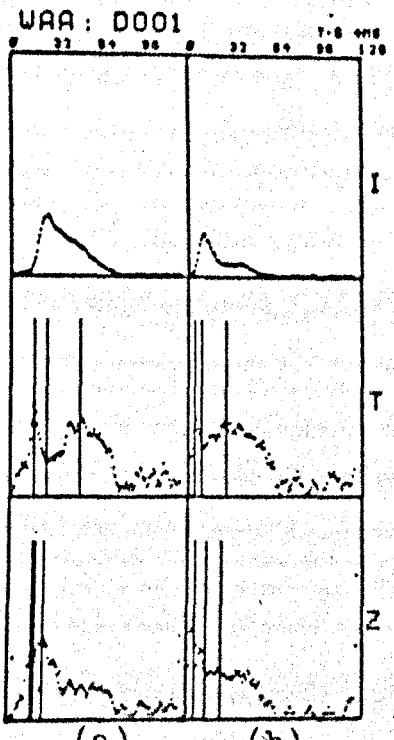


Fig. 3.24

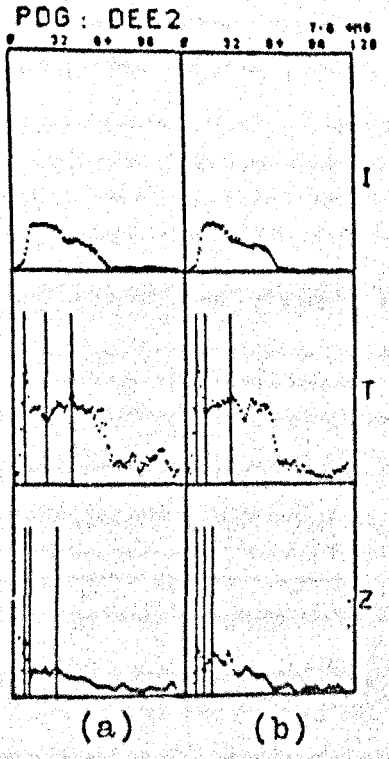
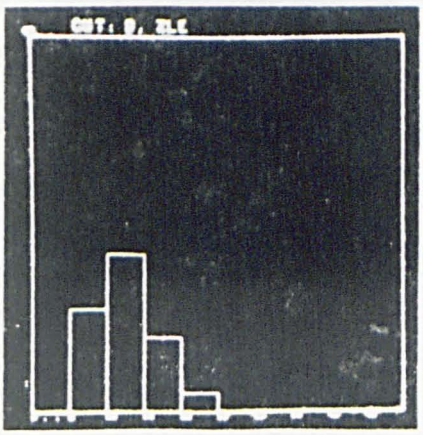
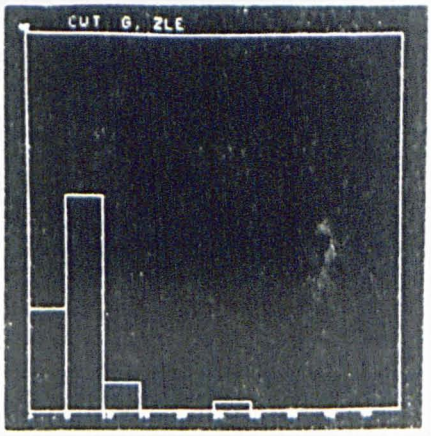


Fig. 3.26

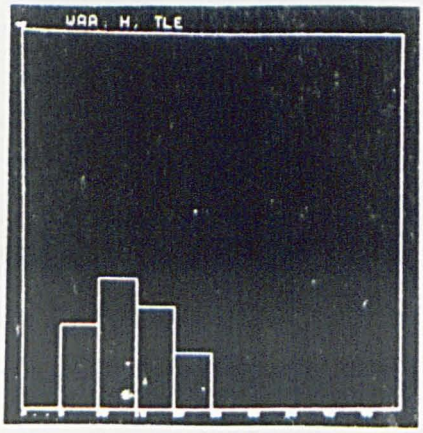


(a)

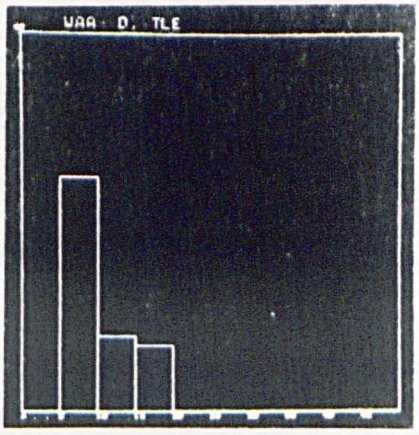


(b)

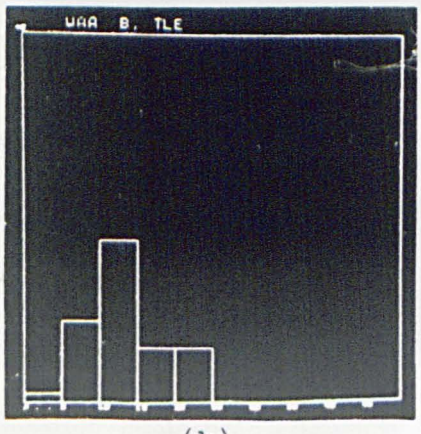
Fig. 3.27



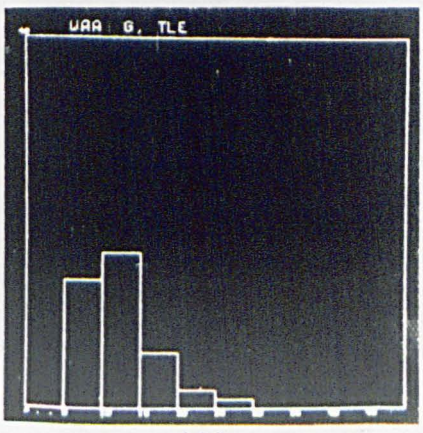
(a)



(c)



(b)



(d)

Fig. 3.28

The T. peaks were again the smallest of the Voiced Stops (see figure 3.29(a)). About 30% of the utterances of /b/ did not show a separate Z. peak. The Z. and T. peaks for /b/ were generally no wider than those of /d/ and /g/, but the Z. onset time Z.L.E. was again slightly larger than that of /g/.

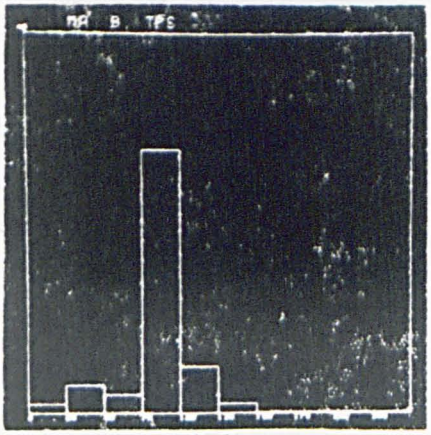
3.1.5.4. Subject P.D.G.

The Z.T.I. diagrams for /b/ could again be distinguished from those of the other Voiced Stops, as described above, and resembled most closely those of /h/. As Figure 3.30 shows, the Width of the Z. peak was often greater than that of /h/, though a few examples of both /b/ and /h/ did not show a separate Z. peak.

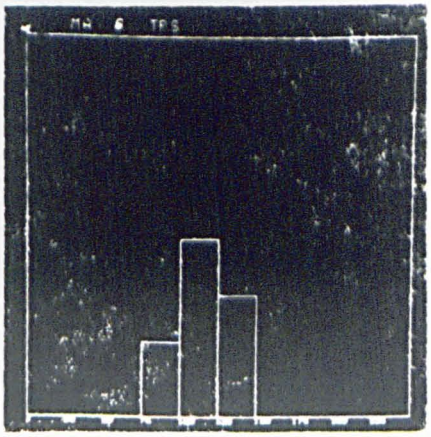
3.1.6. /d/.

3.1.6.1. Subject C.W.T.

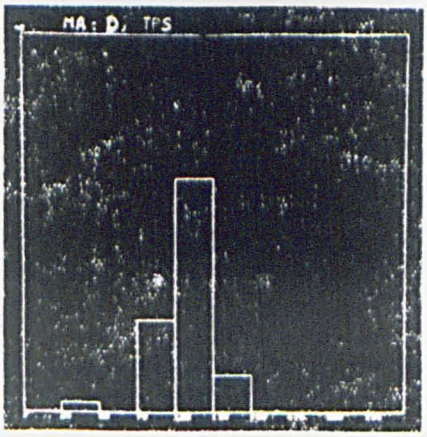
/d/ usually had the highest T. peaks of the 3 Voiced Stops showing the presence of energy at higher frequencies. The distributions of T.P.S. for /d/ and /g/ are compared in figure 3.31. The Z. peaks were similar to those of /g/, but often higher than those of /b/.



(a)

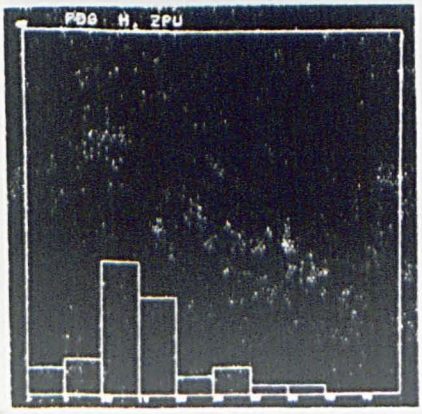


(c)

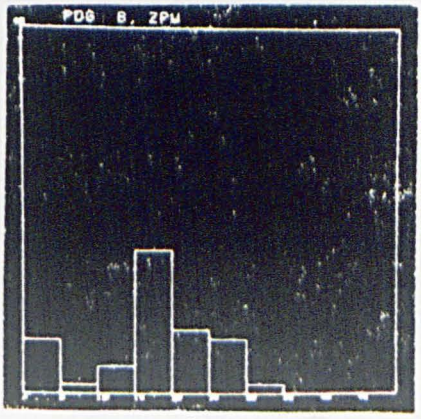


(b)

Fig. 3.29

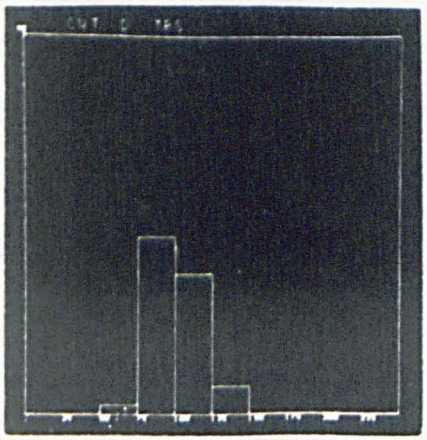


(a)

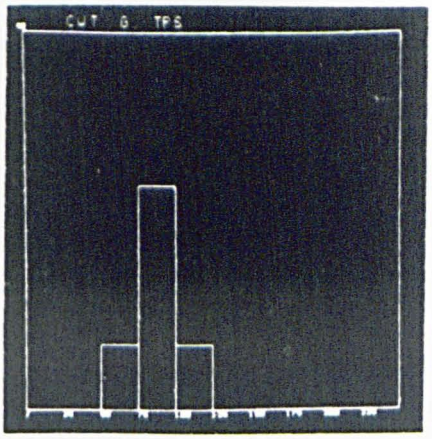


(b)

Fig. 3.30

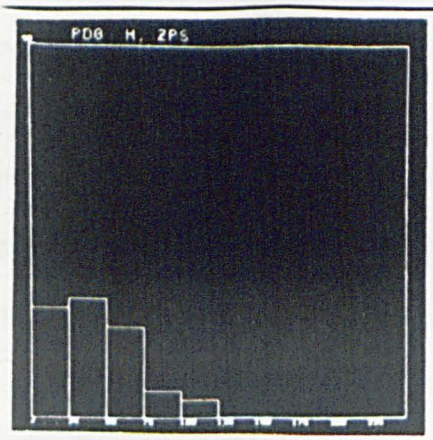


(a)

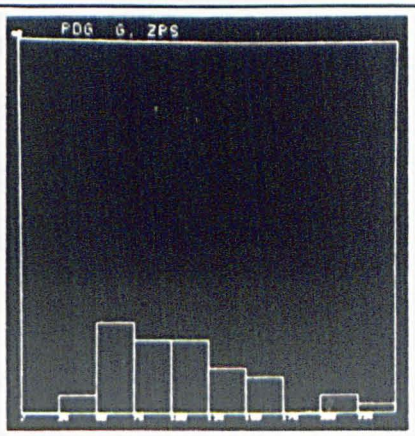


(b)

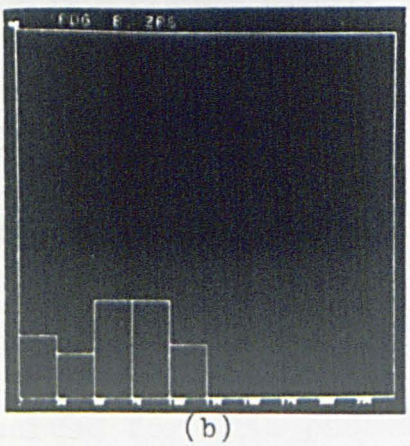
Fig. 3.31



(a)



(c)



(b)

Fig. 3.32

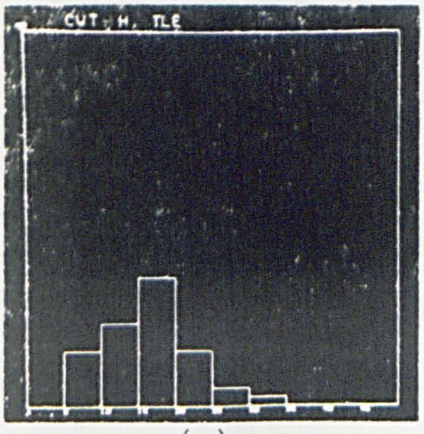
In a few instances very large Z. peaks (Z.P.S. about 200) were observed for /d/ and /g/, owing to an increase in the amount of fricative modulation. Z.T.I. diagrams for two utterances of /dʒ/ are shown in figure 3.23.

3.1.6.2. Subject W.A.A.

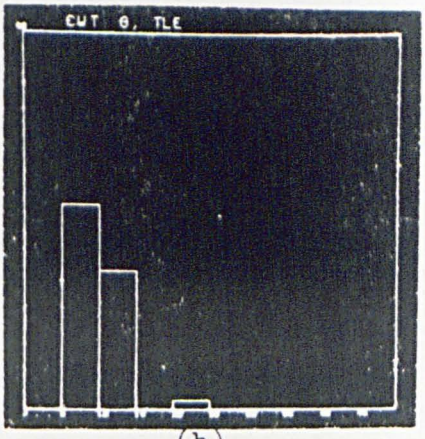
The Z.T.I. diagrams of /d/ for this subject again showed the largest T. peaks of the three Voiced Stops. The Z. and T. peaks were generally the sharpest of these 3 phonemes, (see figure 3.28). Two examples of the sound /dɯ/ spoken by W.A.A. are shown in figure 3.24. The duration of the Voiced Stops (and /h/) was somewhat greater and the Z. and T. peaks were less sharp than for subject C.W.T. This can be seen in figure 3.24 and by comparing the T.L.E. distributions of /h/ and /g/ (figure 3.28(a) and (d)) with those for C.W.T. (figure 3.33).

3.1.6.3. Subject M.A.

Z.T.I. diagrams for utterances of the sound /dɑ/ by subject M.A. are shown in figure 3.25. Like its voiceless equivalent /t/, /d/ did not show the large T. peaks normally observed for this phoneme pair. The values of T.P.S. were in fact slightly smaller than those of /g/,

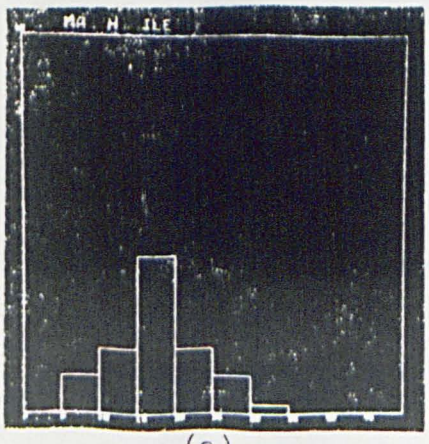


(a)

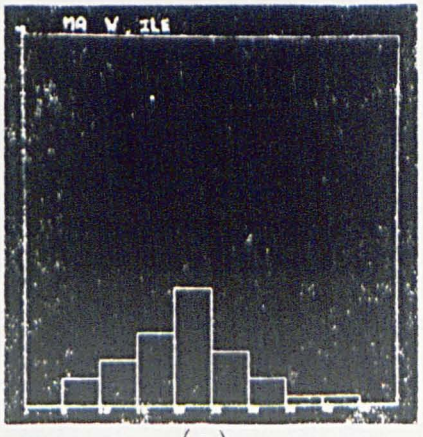


(b)

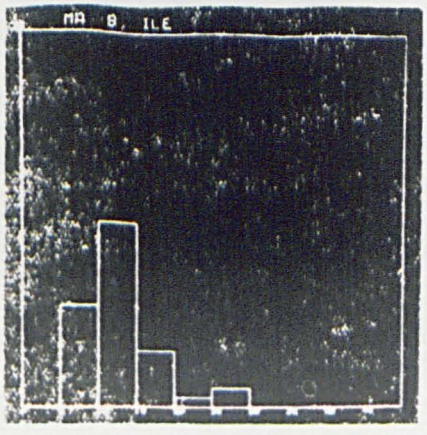
Fig. 3.33



(a)



(c)



(b)

Fig. 3.34

(figure 3.29 (b) and (c)). This was presumably due to abnormal pronunciation for this phoneme pair.

3.1.6.4. Subject P.D.G.

Figure 3.26 shows the Z.T.I. diagrams for a pair of /dI/ sounds by subject P.D.G. The /d/ was characterised by extremely large and sharp T. peaks, almost as large as those of its voiceless equivalent /t/. The average value of T.P.S. was about 150.

3.1.7. /g/.

3.1.7.1. Subject C.W.T.

Z.T.I. diagrams of 2 utterances of /gɔ/ are shown in figure 3.20. The values of T.P.S. for /g/ tended to be slightly smaller than for /d/ (see figure 3.31). The Z. peaks were generally very sharp and had the smallest values of Z.L.E. for the Voiced Stops (see figure 3.27).

3.1.7.2 Subject W.A.A. The Z.T.I. diagrams for /g/ for this speaker were similar to those obtained for subject C.W.T.

3.1.7.3. Subject M.A.

The behaviour of the Z.T.I. diagram of /g/ for this subject was again similar to that observed for subject C.W.T.

3.1.7.4. Subject P.D.G.

/g/ for this subject showed T. peaks on the Z.T.I. diagrams which were much smaller than those of /d/, the average value of T.P.S. being about 105. The size of the Z. peaks was quite variable, but these peaks were normally larger than those of /b/ (see figure 3.32).

3.1.8. /h/.

3.1.8.1. Subject C.W.T.

Although /h/ is classed as a Fricative sound, most of the utterances of this phoneme had Z.T.I. diagrams resembling those of the Voiced Stops /b/, /d/ and /g/. In particular, /h/ usually did not have a separate I. peak. Figure 3.21 shows Z.T.I. diagrams for 2 utterances of the sound /hɜ/ with this property. The absence of I. peaks for /h/ was attributed to:

(a) A relatively large amount of voicing being present in the sound.

(b) The exceptionally low energy level - as seen in the sonogram of figure 1.5(a). Occasionally this led to large I.S.D. values (see section 3.1.9).

(c) The short duration of /h/ compared to that of the other

fricatives, /h/ rarely reaching a steady state.

I. peaks did occur for a few examples of /h/, e.g. the 2 utterances of /hɔ/ of figure 3.22. This illustrates the variable nature of /h/.

The values of Z.P.S. for /h/ varied widely, but the average was about 100. Like /d/ and /g/, some utterances of /h/ had large values of Z.P.S. (about 200) due to increased frication; T.P.S. lay mostly in the range 75-100, less than for /d/.

The fricative nature of /h/ meant that the Z. and T. peaks were generally wider than those of the Voiced Stops. The onset times were also greater, and the clearest distinction was that of T.L.E. (see figure 3.33).

3.1.8.2. Subject W.A.A.

The Z.T.I. diagrams of /h/ for this subject again resembled those of the Voiced Stops /b/, /d/ and /g/. The Z. and T. peaks were almost as narrow as those of /b/, /d/ and /g/, but T.P.S. was somewhat smaller than for /d/ and /g/, while the onset time T.L.E. was generally larger than for /b/ and /d/ (see figure 3.28.) A few examples of /h/ had large values of I.S.D. due to an exceptionally low amount of energy being present in the sound.

3.1.8.3 Subject M.A.

The Z.T.I. diagrams for /h/ again resembled those of /b/ most closely. The fricative nature of /h/ again meant that the peaks were a little wider than those of /b/, and the onset times longer. The time taken to rise to the maximum on the I. trace, I.L.E., (no separate I. peak) was also greater than for /b/. The distributions of I.L.E. for /h/ and /b/ are compared in figure 3.34.

3.1.8.4 Subject P.D.G.

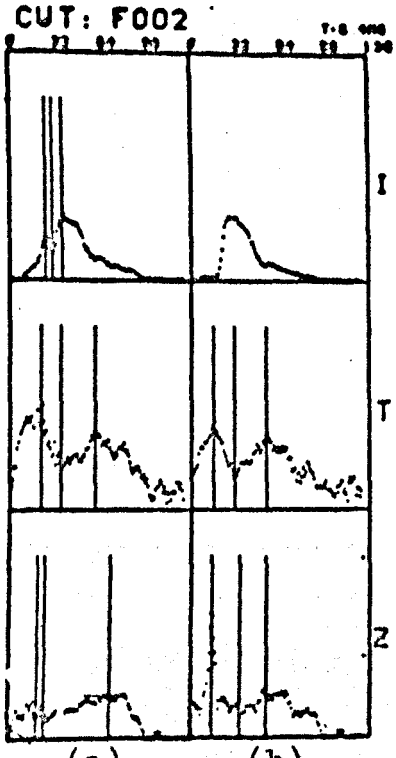
For this subject, about 70% of the utterances of /h/ gave Z.T.I. diagrams resembling those of the voiced stops, while the remainder showed distinct I. peaks which were sometimes quite large, due to an increase in the amount of energy present in the /h/ sound. A few utterances showed a large I. start delay. The Z. peaks for /h/ were less variable than for other speakers, the Z.P.S. value being mostly below 100 (See figure 3.32(a)).

3.1.9. /f/.

3.1.9.1. Subject C.W.T.

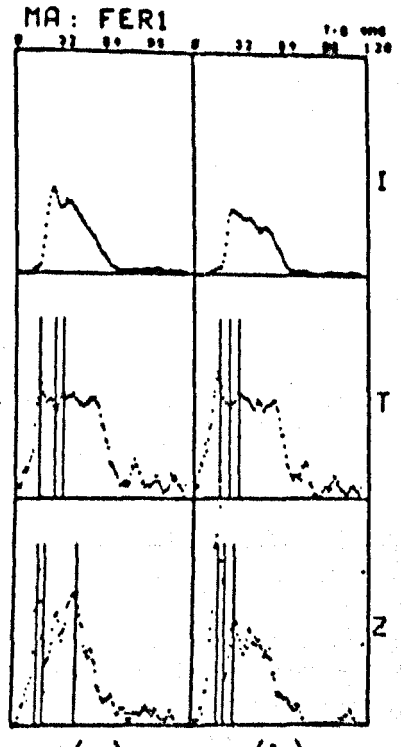
For this subject, the behaviour of the I. peak for /f/ varied widely. This was due to changes in the amount of energy present in the /f/, partly caused by changes in stress. Z.T.I. diagrams of 2 utterances of /fu/ and /fɑ/ are shown in figures 3.35 and 3.36 respectively. Many utterances had exceptionally low overall energy levels (as seen in the sonagram of figure 1.5(f)corresponding to the Z.T.I. diagram of figure 3.36(a).) Figure 3.39 is a U.V. recording of the raw speech waveform for this sound. It can be seen from figure 3.39 that the envelope of the /f/ is so small as to be indistinguishable from the silence preceeding it. The starting point of the I. trace was therefore a good deal later than that of the Z. and T. traces. This start delay was measured by the parameter I.S.D., and provided a means of separating these samples of /f/ from the other consonants. This effect was also observed in some utterances of /θ/ and in a few examples of /h/.

Other utterances of /f/, such as those of figure 3.35(a) and figure 3.36(b) had distinct I. peaks due to a greater amount of energy being present in the /f/. In a few cases, like the sound /fu/ of figure 3.35(a), these peaks



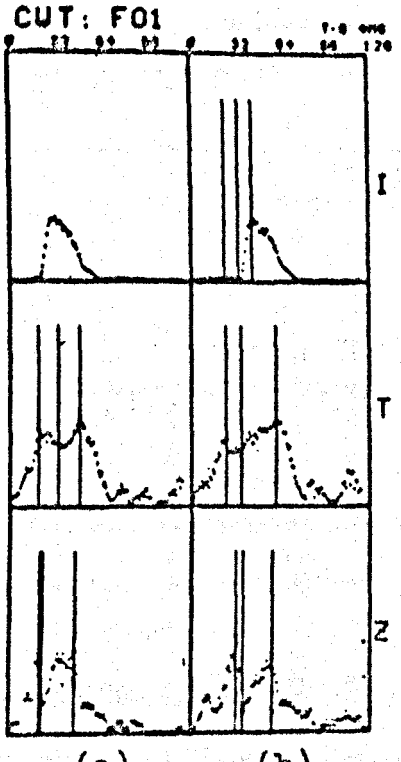
(a) (b)

Fig. 3.35



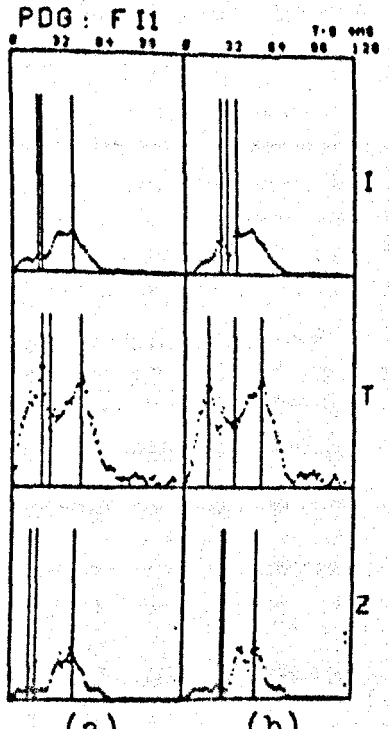
(a) (b)

Fig. 3.37



(a) (b)

Fig. 3.36



(a) (b)

Fig. 3.38

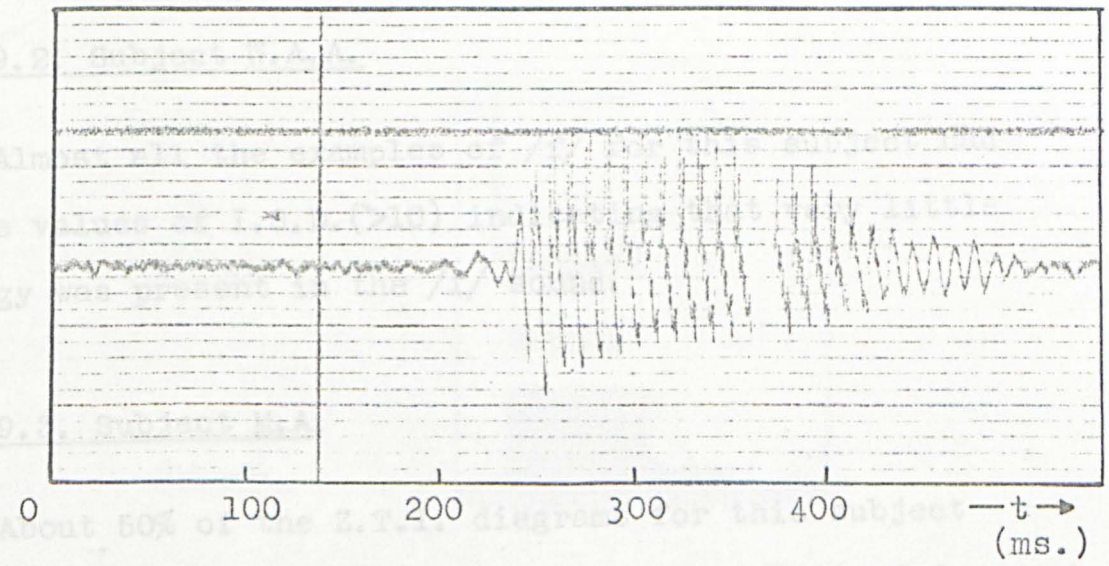


Fig. 3.39 Raw Speech Waveform for an Utterance of /fɒ/ by C.W.T.

resembled those of /p/, due to the presence of a very large amount of energy at low frequencies. The values of Z.P.S. varied over a wide range, but T.P.S. was generally quite small for a Voiced Fricative, the average value being about 100. The Z. and T. peaks for /f/ were generally quite wide, the average value of Z.P.W. being about 25 and that of T.P.W. being about 30 - 35.

3.1.9.2. Subject W.A.A.

Almost all the examples of /f/ for this subject had large values of I.S.D.(>10) indicating that very little energy was present in the /f/ sound.

3.1.9.3. Subject M.A.

About 50% of the Z.T.I. diagrams for this subject showed values of I.S.D. larger than those obtained for most of the other phonemes, but these start delays were generally shorter than those found for subjects C.W.T. and W.A.A. (The best threshold value for the I.S.D. decision fell from 10 to 5 for this subject.) Figure 3.37 shows Z.T.I. diagrams for 2 utterances of /fʒ/ by M.A.: the I. start delays can be seen for both sounds. Figure 3.40 is the sonagram for the /fʒ/ sound of figure 3.37(a). From

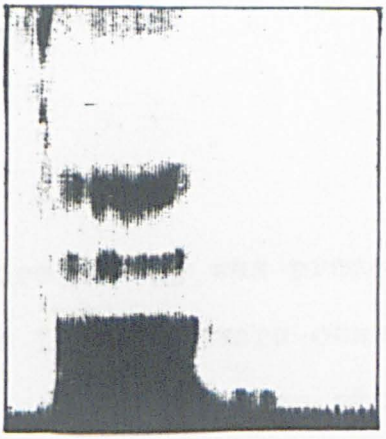


Fig. 3.40 Sonagram for an utterance of /f3/ by M.A.

figure 3.40 it is apparent that the amount of energy in the /f/ sound increases towards the end of the sound, thus decreasing the value of I.S.D. As in the case of subject C.W.T., large I. peaks were observed for a few examples of /f/ in which more energy was present in the consonant sound.

The size of the Z. peaks for this subject varied, but they were sometimes quite large, as seen in figure 3.37. This indicates the presence of more energy at a higher frequency than in the case of the previous subjects, as seen in the sonagram of figure 3.40.

3.1.9.4. Subject P.D.G.

For this subject, far more energy was present in the /f/ sound, and hence separate I. peaks were observed for most of the utterances, though 5% of the examples showed large I.S.D. values. The I. peaks were often quite large, though I.P.S. varied widely (see figure 3.41). Z.T.I. diagrams for 2 utterances of /fI/ by subject P.D.G. are shown in figure 3.38. The increased amount of energy in the /f/ sound can be seen in the sonagram for the /fI/ sound of figure 3.38(b), shown in figure 3.43. The Z. and T. peaks for /f/ remained much smaller than those of the voiceless fricatives /s/ and /f/, showing the scarcity of energy at high frequencies.

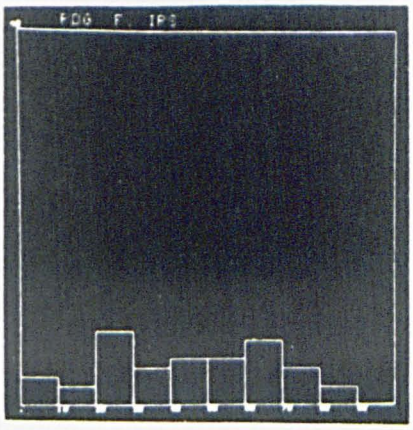
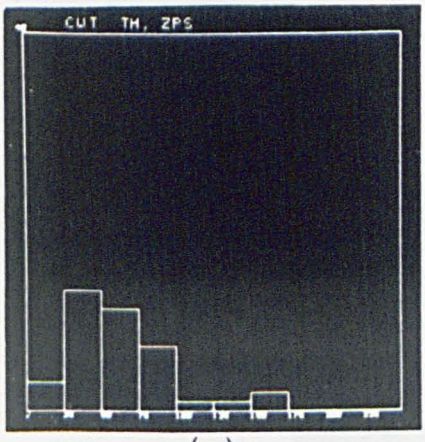
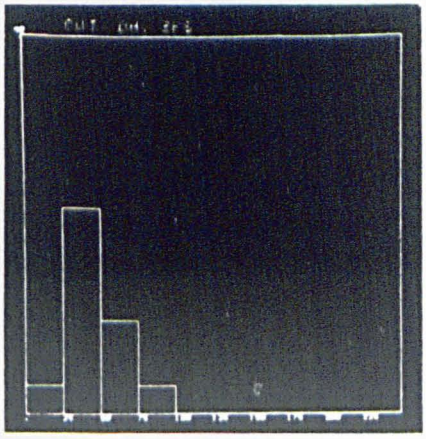


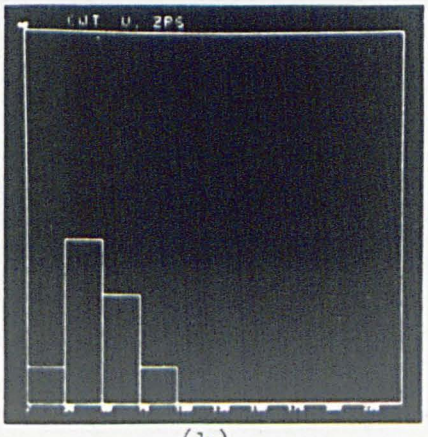
Fig. 3.41



(a)



(c)



(b)

Fig. 3.42

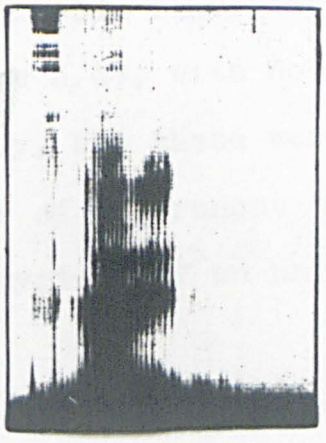


Fig. 3.43 Sonagram for an utterance of /fI/ by P.D.G..

The duration of /f/ was normally very long, and the onset times were extremely large, reflecting the fairly gradual energy built up, seen in figure 3.43.

3.1.10. /θ/.

3.1.10.1. Subject C.W.T.

/θ/, like /f/, showed varying behaviour on the I. trace due to differences in the amount of energy present in the sound. Again there were some cases, like that of the first utterance of /θɔ/ in figure 3.44, with no separate I. peak and a large I. start delay, but these were less numerous than for /f/. A larger number of utterances had a small but fairly distinct I. peak more typical of an unvoiced sound, (see figure 3.44(b)).

The Z. and T. peaks for /θ/ more closely resembled those of the voiced Fricatives/v/ and /ð/ than /f/. Figure 3.42(a) shows that the value of Z.P.S. was hardly any larger than for /v/ and /d/. This was somewhat surprising since /θ/ is classed as a voiceless sound. The very low energy level of /θ/ may be a partial cause of this, and in some cases the discrepancy may be explained by mispronunciation of the /θ/, which can easily be confused with /ð/.

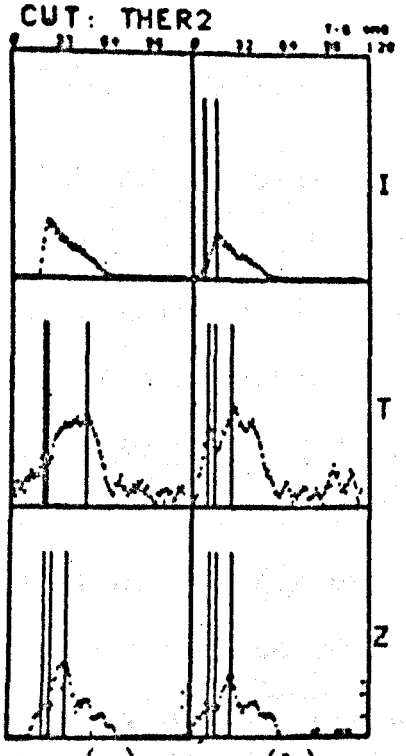


Fig. 3.44

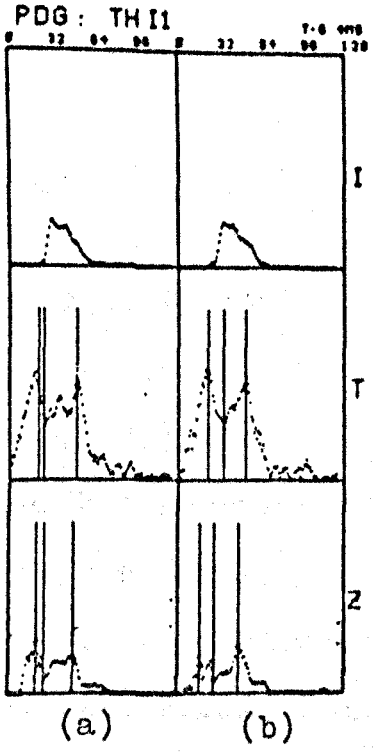


Fig. 3.46

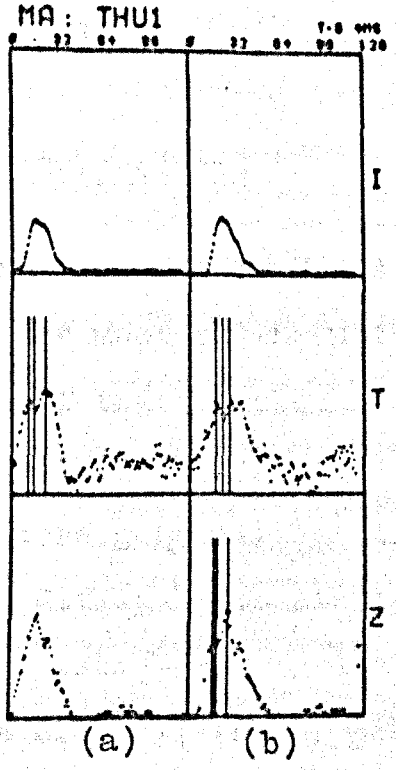


Fig. 3.45

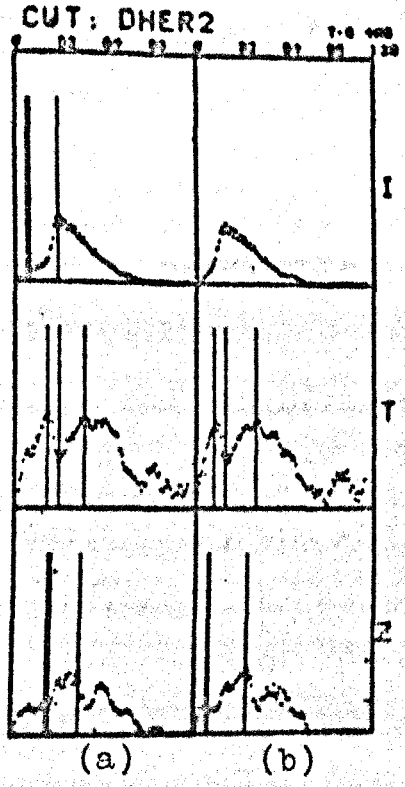


Fig. 3.47

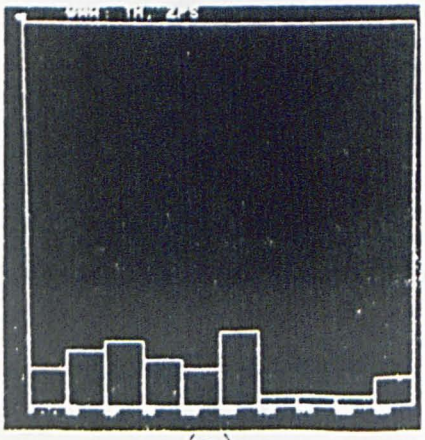
3.1.10.2. Subject W.A.A.

About 30% of the utterances of /θ/ for this subject were similar to /f/, having a large I.S.D. value (greater than 10). Most of the remainder did not show a separate I. peak. The absence of a separate I. peak was attributed to similar factors to those given for /h/ (section 3.1.8.)

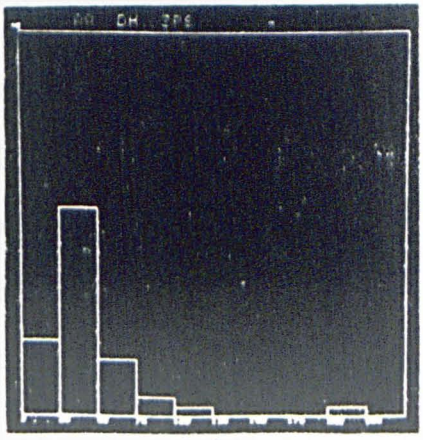
The Z. peaks were extremely variable, but normally larger than those of /v/ and /ð/, (see figure 3.48). The T. peaks were quite small, showing an energy concentration at the lower end of the frequency scale. The Z.T.I. diagrams for /θ/ generally resembled those of the voiced stops, and its duration was abnormally short for a fricative. /θ/ was very difficult to distinguish from /h/, /b/, /d/ and /g/.

3.1.10.3. Subject M.A.

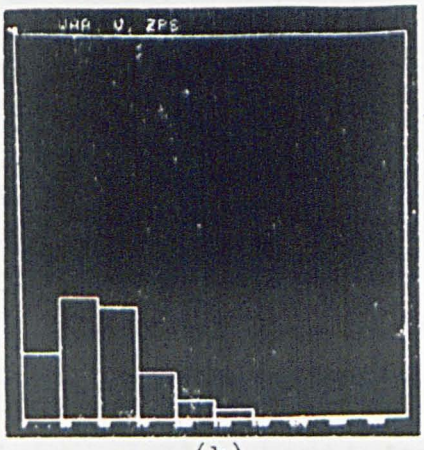
As in the case of W.A.A., no separate I. peaks were observed for most of the utterances by this subject. About 35% of the examples of /θ/ considered had I.S.D. values similar to those of /f/ for this subject. The Z. peaks, however, were generally smaller than those of /f/, and /θ/ was again difficult to distinguish from /h/, /b/, /d/ and /g/. Figure 3.45 shows Z.T.I. diagrams for 2 utterances of /θɑ/. In figure 3.45(a), the Z. peak is indistinct.



(a)

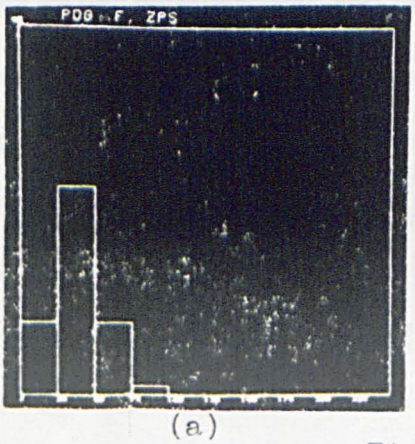


(c)

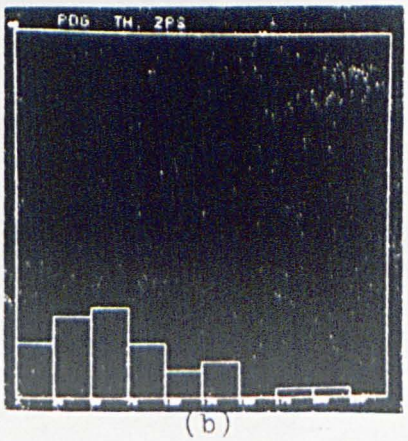


(b)

Fig. 3.48



(a)



(b)

Fig. 3.49

3.1.10. 4. Subject P.D.G.

Z.T.I. diagrams for 2 utterances of /θ/ by subject P.D.G. are shown in figure 3.46. Roughly 80% of the utterances of /θ/ showed a very long start delay on the I. trace (I.S.D. greater than 15), due to an exceptionally low overall energy level. The Z. and T. peaks for /θ/ were generally larger than those of /f/, showing a higher frequency energy concentration. Figure 3.49 shows the distribution of Z.P.S. for /f/ and /θ/.

3.1.11./v/.

3.1.11.1. Subject C.W.T.

Figure 3.50 shows Z.T.I. diagrams for 2 utterances of the sound /vu/ spoken by C.W.T. Like most of the voiced Fricatives, /v/ normally showed distinct I. peaks, indicating that an appreciable amount of fricative modulation was present. These peaks were generally more prominent than those obtained for unvoiced sounds, though the onset of the vowel was less sharp.

As Figure 3.50 illustrates, the Z. peaks for /v/ were considerably smaller than those of the vowel, the Z. trace being governed by the low F1 value for /v/. Figure 3.42(b)

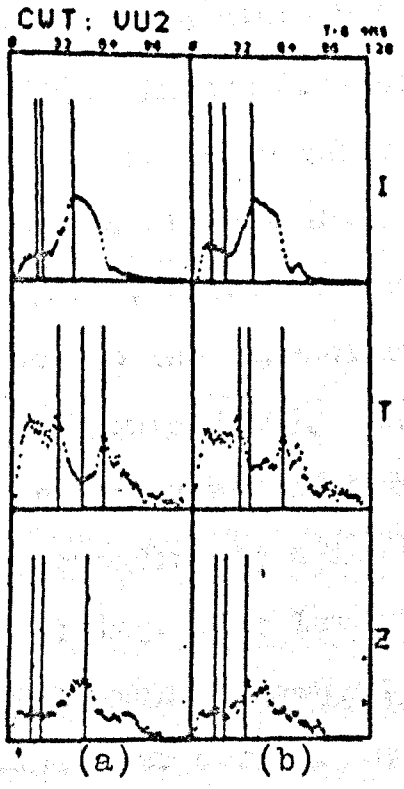


Fig. 3.50

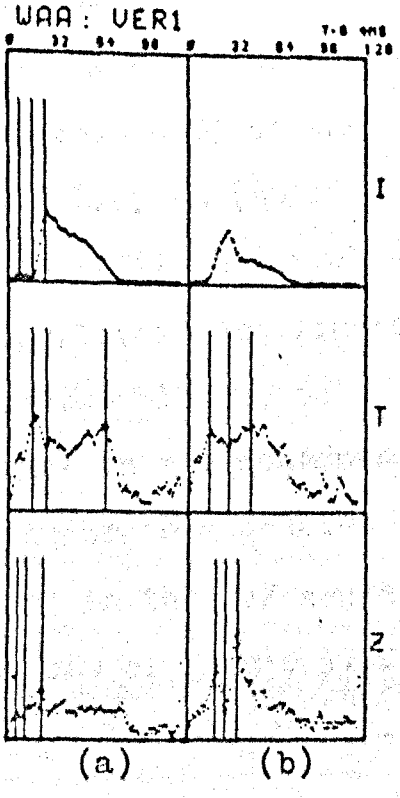


Fig. 3.51

shows the distribution of Z.P.S. for /v/. The T. peaks for /v/ were often quite large, showing the influence of fricative excitation.

3.1.11.2. Subject W.A.A.

/v/ for this subject was again characterised by very small Z. peaks. The distribution of Z.P.S. is shown in figure 3.48(b). Figure 3.51 shows 2 examples of the sound /vɜ/ by W.A.A. An appreciable number (about 25%) of the examples studied did not show a separate I. peak (as in figure 3.51(b)), while the remainder generally had smaller I. peaks than in the case of subject C.W.T.(compare figure 3.51(a) with figure 3.50). This was presumably due to a reduction in the amount of energy present in the consonant sound, corresponding to a lowered stress or to a greater amount of voicing. The low energy level in the /v/ sound of figure 3.51(b) can be seen in the sonagram of figure 3.52.

In general far more of the utterances by subject W.A.A. did not show a separate I. peak.

3.1.11.3 Subject M.A.

The Z.T.I. diagrams of /v/ for this subject resembled those of subject C.W.T. Almost all of the utterances showed a separate I. peak..

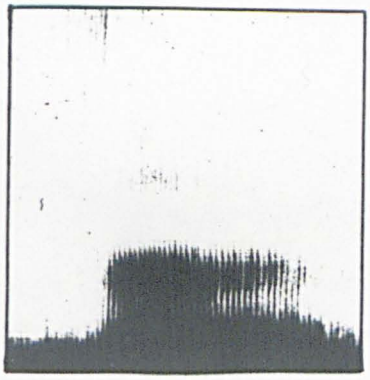


Fig. 3.52 Sonagram for an utterance of /v3/ by W.A.A.

3.1.11.4 Subject P.D.G.

The Z.T.I. diagrams of /v/ again showed very small Z. peaks. As in the case of speaker W.A.A., about 30% of the utterances did not show a separate I. peak.

3.1.12. /ð/

Z.T.I. diagrams for two utterances of the sound /ðɜ/ by C.W.T. are shown in figure 3.47. As expected, /ð/ behaved

(over

in a similar way to /v/ and it was difficult to discriminate between the two. Most of the utterances of /ð/ had distinct I. peaks as in figure 3.47(a) but sometimes (as in figure 3.47(b)) the I. peak was absent. This can be attributed mainly to differences in the amount of voicing present in the /ð/ and to a lesser overall energy level.

The distinguishing feature of /ð/, like /v/, was the extremely small (and sometimes absent) Z. peak. The Z.peaks were expected to be a little higher than those of /v/, but this was not generally the case,(see figure 3.42).

3.1.12.2. Subject W.A.A.

The Z.T.I. diagrams for /ð/ were again very similar to those of /v/. The Z. peaks were slightly smaller than for /v/, (see figure 3.48). About 20% of the utterances did not show a separate I. peak.

3.1.12.3. Subject M.A.

It was again difficult to distinguish the Z.T.I. diagrams of /ð/ from those of /v/, though these 2 phonemes could easily be distinguished from the remainder by their exceptionally low Z.P.S. values, and the presence of a separate I. peak.

As figure 3.53 shows, the Z. peaks for /ð/ tended to be wider than those of /v/.

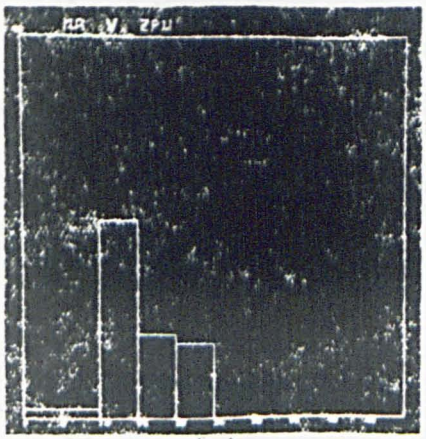
3.1.12.4. Subject P.D.G.

Just over 50% of the utterances of /ð/ for this subject did not show a separate I. peak on the Z.T.I. diagram. The presence of very small Z. peaks, and the greater duration of the fricative sound, enabled the utterances to be separated from the voiced stops relatively easily.

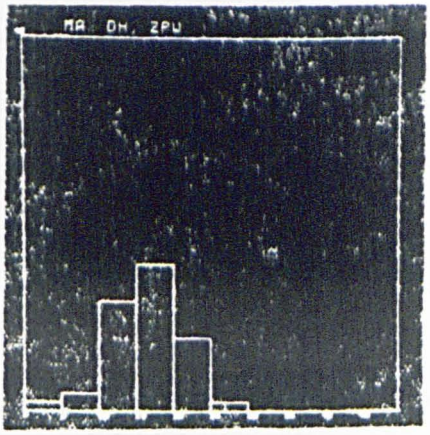
3.1.13. /s/.

3.1.13.1. Subject C.W.T.

The distinguishing features of /s/ were the massive Z. and T. peaks obtained for this sound. Figure 3.55 shows Z.T.I. diagrams for 2 utterances of /sɯ/ by C.W.T. The height of the Z. and T. peaks shows that the energy in /s/ was concentrated at high frequencies, with little energy in the vowel formant range. The duration of /s/ was often extremely large, and the peaks therefore flat topped. Most of the examples of /s/ had small values of I.P.S., the I. peaks being similar to that of figure 3.55(a).

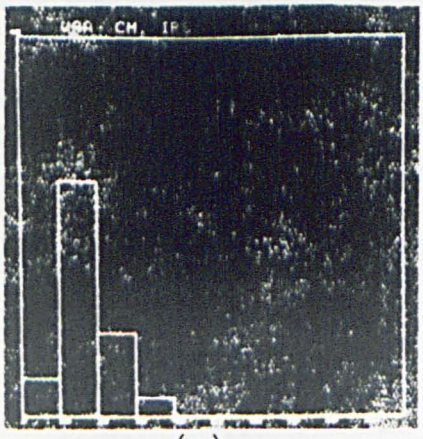


(a)

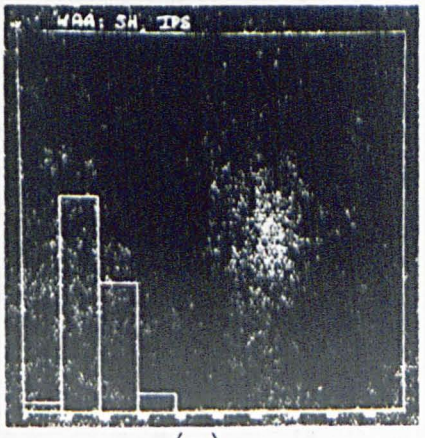


(b)

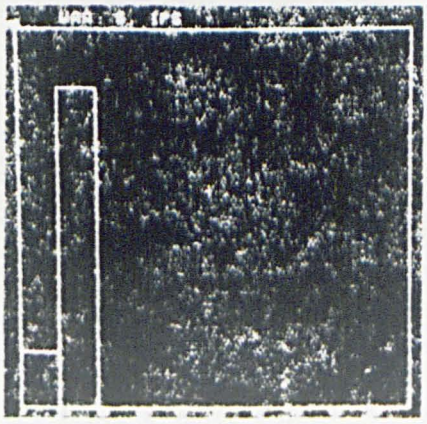
Fig. 3.53



(a)



(c)



(b)

Fig. 3.54

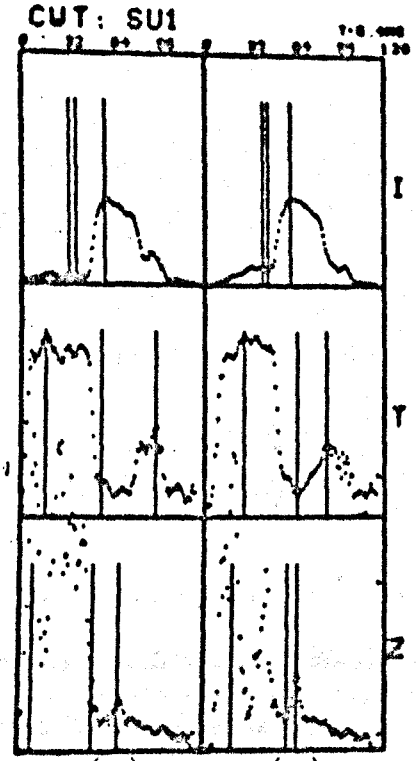


Fig. 3.55

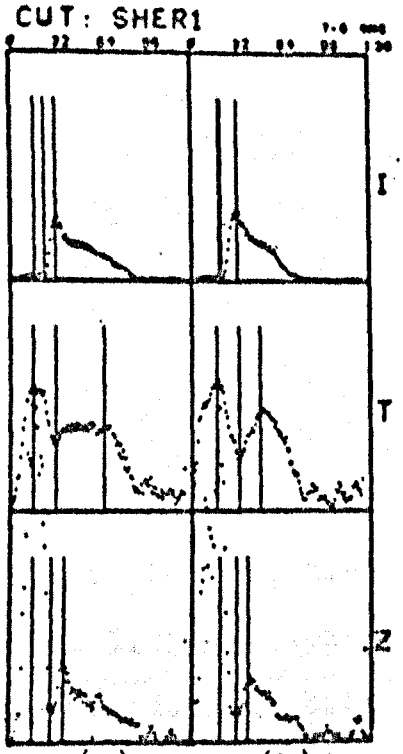


Fig. 3.57

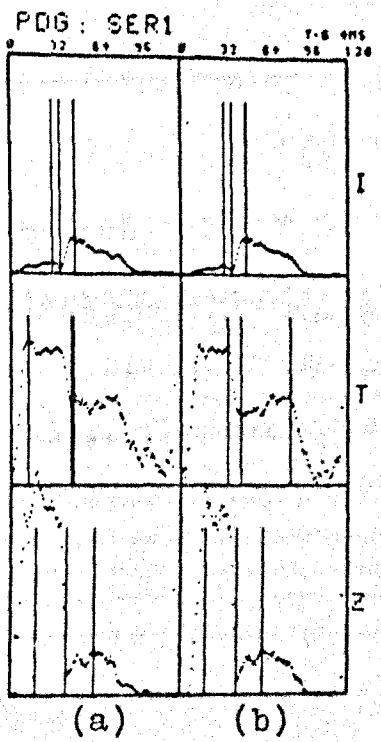


Fig. 3.56

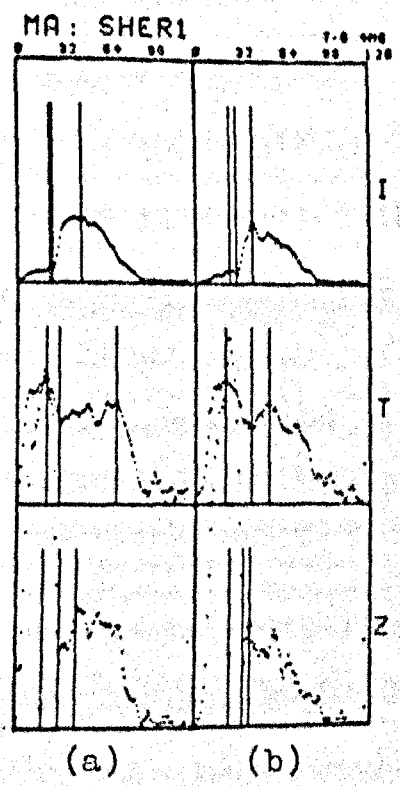


Fig. 3.58

3.1.13.2. Subject W.A.A.

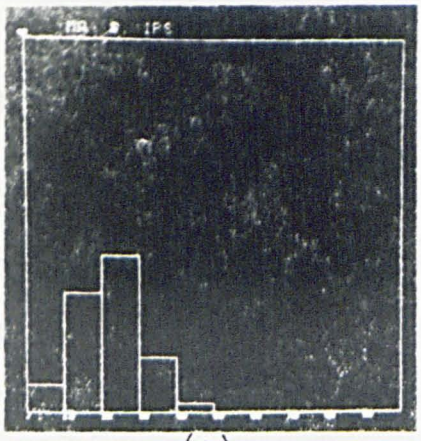
The Z.T.I. diagram of /s/ for subject W.A.A. resembled those obtained for C.W.T. The I. peaks for /s/ were generally smaller than for the other phonemes which had massive Z. and T. peaks. (see figure 3.54).

3.1.13.3. Subject M.A.

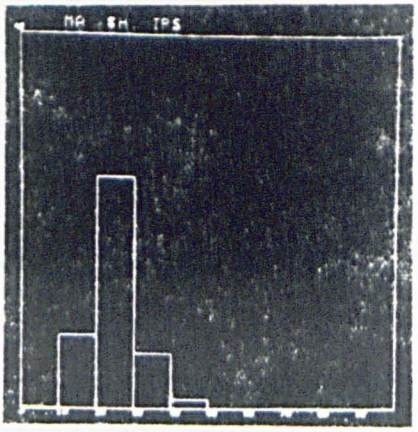
The behaviour of /s/ on the Z.T.I. diagram was again similar to that observed for C.W.T. The I. peaks were more prominent than for subject W.A.A. (Compare figures 3.54 and 3.59).

3.1.13.4. Subject P.D.G.

Figure 3.56 shows the Z.T.I. diagrams for 2 utterances of /s3/ by subject P.D.G. As figure 3.56 illustrates, the Z.T.I. diagrams for /s/ were similar to those of subject C.W.T. in a slightly exaggerated form, larger values of Z.P.S. and T.P.S. being obtained. The average value of T.P.S. was about 190.

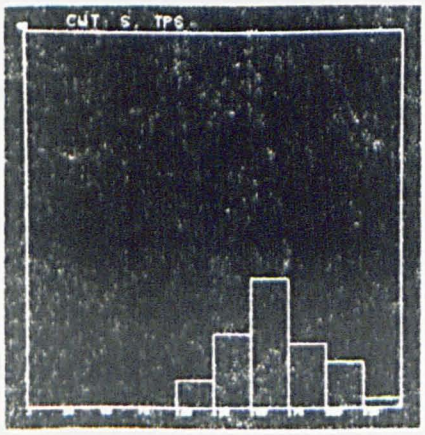


(a)

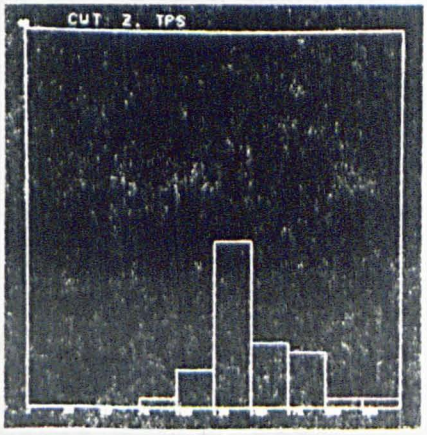


(b)

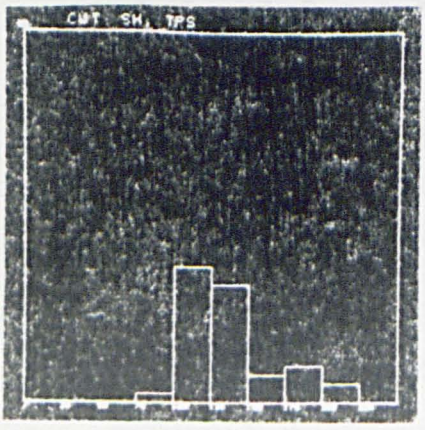
Fig. 3.59



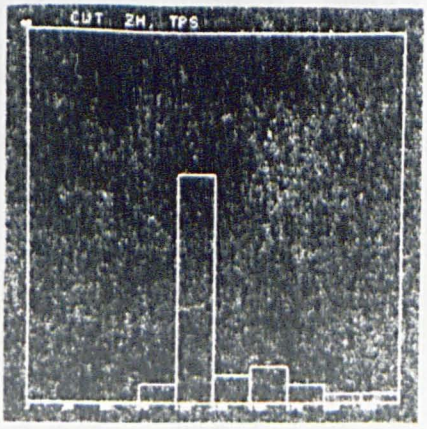
(a)



(c)



(b)



(d)

Fig. 3.60

3.1.14. /ʃ/.

3.1.14.1. Subject C.W.T.

The utterances of /ʃ/ for C.W.T. had Z.T.I. diagrams very similar to those obtained for /s/. 2 utterances of /ʃ/ are shown in figure 3.57. /ʃ/ could be distinguished from /s/ using the parameter T.P.S. which tended to be lower in the case of /ʃ/. The sonagrams of figures 1.5(f)&(h) show that the energy in /s/ was concentrated in higher regions of the spectrum than for /ʃ/. This was also true of the voiced cognates of /s/ and /ʃ/, /z/ and /ʒ/. Figure 3.60 shows the distribution of T.P.S. for these 4 phonemes.

3.1.14.2. Subject W.A.A.

The Z.T.I. diagrams of /ʃ/ for subject W.A.A. were similar to those observed for subject C.W.T. The T. peaks were again somewhat smaller than those of /s/, though this distinction was not as clear as in the case of subject C.W.T.

3.1.14.3. Subject M.A.

The behaviour of /ʃ/ on the Z.T.I. diagram was again similar to that observed for subject C.W.T. The I. peaks

were more prominent than for W.A.A. (compare figures 3.59 and 3.54). Figure 3.58 shows Z.T.I. diagrams for 2 utterances of the sound /ʃ/.

3.1.14.4. Subject P.D.G.

The Z.T.I. diagrams for the utterances of /ʃ/ by subject P.D.G. closely resembled those obtained for subject C.W.T.

3.1.15. /z/.

3.1.15.1. Subject C.W.T.

Z.T.I. diagrams for 2 utterances of the sound /z ɑ/ by C.W.T. are shown in figure 3.61. The I. and Z. peaks for this phoneme varied a good deal due to differences in the amount of voicing present. The I. peaks for /z/ were generally larger but less distinct than for its voiceless cognate /s/. When a large amount of voicing was present, as in figure 3.61(a) and the corresponding sonagram of figure 1.5(g),the Z. trace was governed by F1 and was quite small. Z.peaks more closely resembling those obtained for /s/ occurred when frictional excitation predominated. Often the Z. trace showed more than a single peak for the /z/ sound (e.g.figure 3.61(b)); this was presumably due to

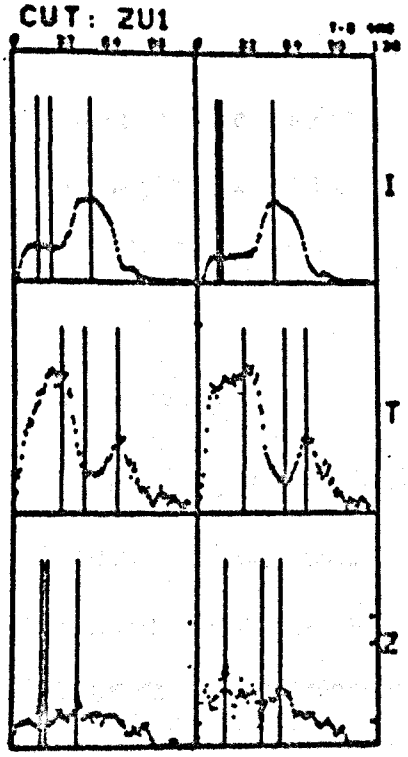


Fig. 3.61

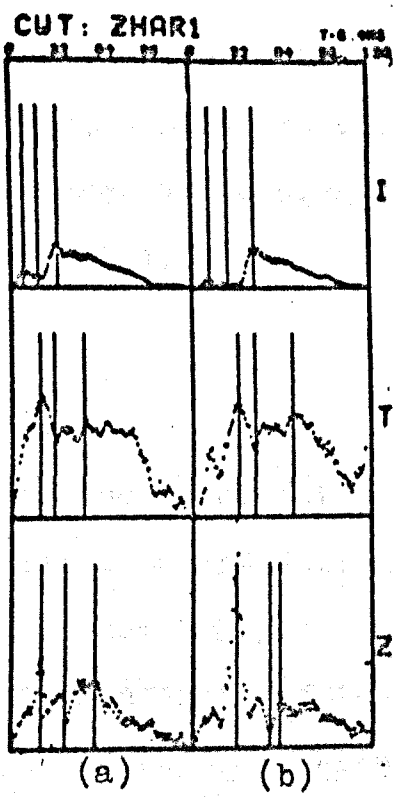


Fig. 3.63

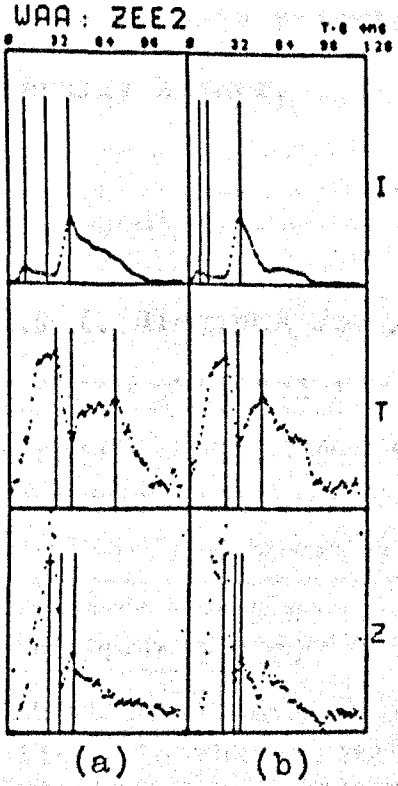


Fig. 3.62

changes in the amount of voicing within the sound itself.

The T. peaks for /z/ were very large, similar to those of /s/, showing that a large amount of energy was always present at high frequences,(see figure 3.60).

3.1.15.2. Subject W.A.A.

Z.T.I. diagrams for 2 utterances of the sound /zi/ are shown in figure 3.62. The level of voiceless modulation was generally higher than in the case of subject C.W.T. This can be seen from comparison of the sonograms of the /zi/ sound of figure 3.62(b), shown in figure 3.64, with that of figure 1.5(g).For this reason, the I. peaks were smaller, more closely resembling those of /s/, and the Z. peaks generally larger.

3.1.15.3. Subject M.A.

The Z.T.I. diagrams for /z/ for this subject closely resembled those obtained for C.W.T. The Z. peaks were generally quite small, due to the dominance of voiced excitation.

3.1.15.4. Subject P.D.G.

Variations in the relative amounts of voiced and

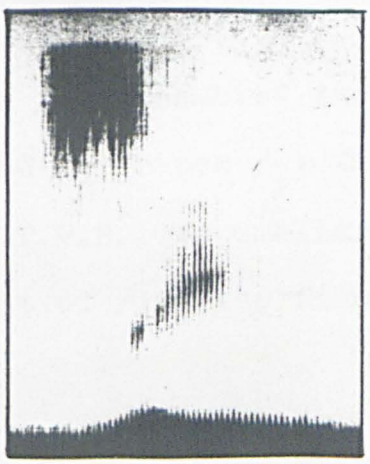


Fig. 3.64 Sonagram for an utterance of /zi/ by W.A.A.

voiceless modulation in the /z/ sound again resulted in variable behaviour on the Z. and I. traces. This is illustrated by the distributions of Z.P.S. shown in figure 3.65. /z/ could, however, be distinguished by its exceptionally large T. peaks, the average value of T.P.S. being about 180.

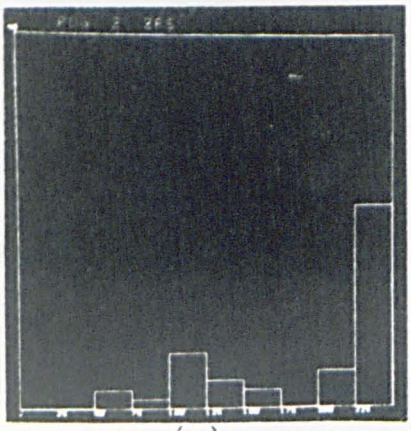
3.1.16. /ʒ/.

3.1.16.1. Subject C.W.T.

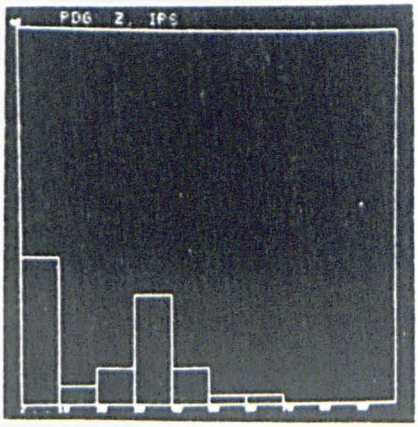
The Z.T.I. diagrams for /ʒ/ resembled those of /z/ as expected. 2 examples of the utterance /ʒ a /are shown in figure 3.63. The value of T.P.S. was smaller than for /z/, being about the same as that of /ʃ/ (see figure 3.60).

3.1.16.2. Subject W.A.A.

The Z.T.I. diagrams for /ʒ/ again resembled those of /z/. T.P.S. was slightly smaller than for /z/, but the Turnaround Peak drop (T.P.D.) was generally lower for /z/ than for /ʒ/, (see figure 3.66). This was presumably caused by a sharper change in the higher formant region at the onset of the vowel following /z/, possibly due to a greater proportion of voiceless modulation being present in /z/.

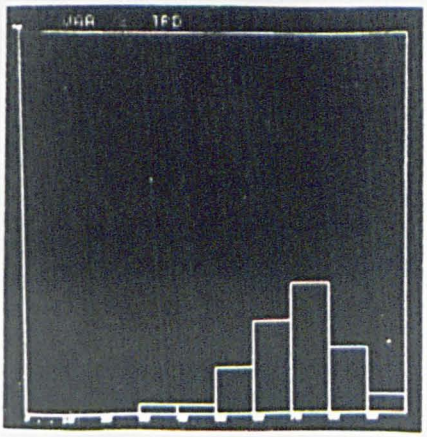


(a)



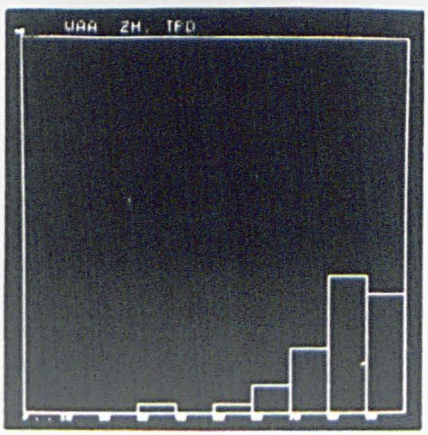
(b)

Fig. 3.65



(a)

Fig. 3.66



(b)

3.1.16.3. Subject M.A.

/ʒ/ exhibited similar behaviour on the Z.T.I. diagram to that observed for subject C.W.T. The Z. peaks were generally quite large, showing the presence of a large amount of unvoiced modulation.

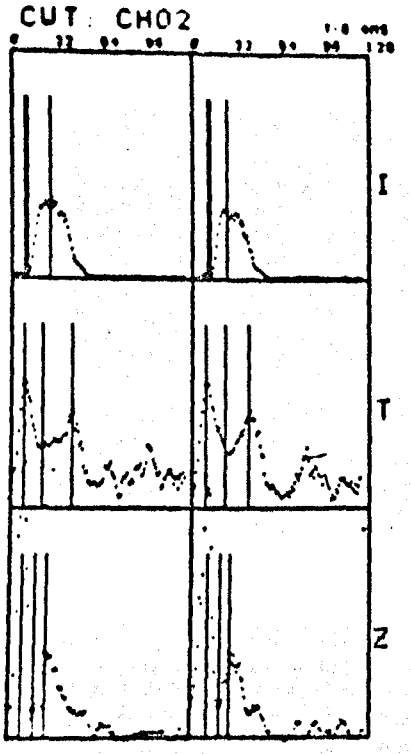
3.1.16.4. Subject P.D.G.

The Z.T.I. diagrams of /ʒ/ for this subject generally behaved in a similar way to those of C.W.T. The Z. peaks for /ʒ/ were normally less pronounced than those of /z/, indicating a greater proportion of voiced excitation.

3.1.17. /tʃ/.

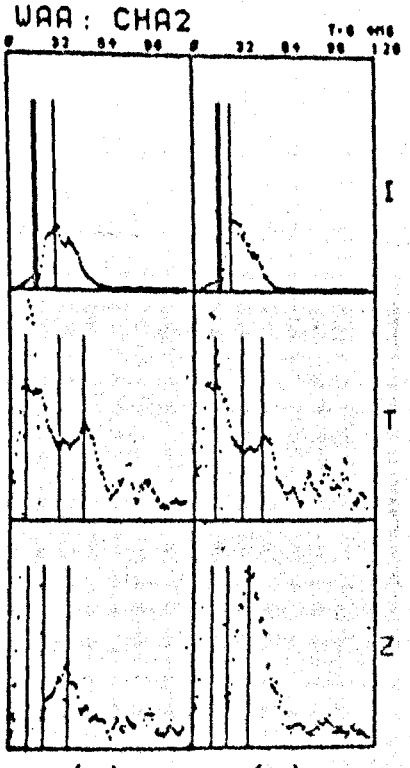
3.1.17.1. Subject C.W.T.

Z.T.I. diagrams for 2 utterances of the sound /tʃ/ spoken by C.W.T. are shown in figure 3.67. As expected, the form of the peaks was midway between those of /t/ and /ʃ/. The duration of the sound was generally longer than for /t/ but shorter than for /ʃ/. The onset times were generally smaller than those of the unvoiced Fricatives, and the Z. and T. peaks were sharper, due to the initial stop phase of the combined sound. Figure 3.69 shows the distribution of



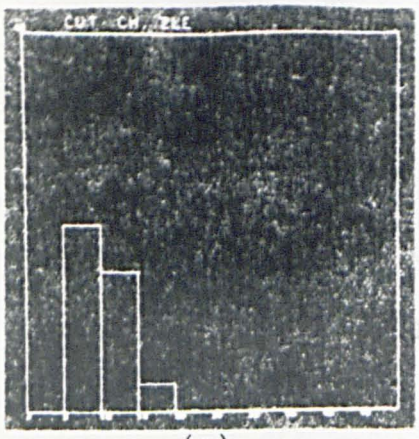
(a) (b)

Fig. 3.67

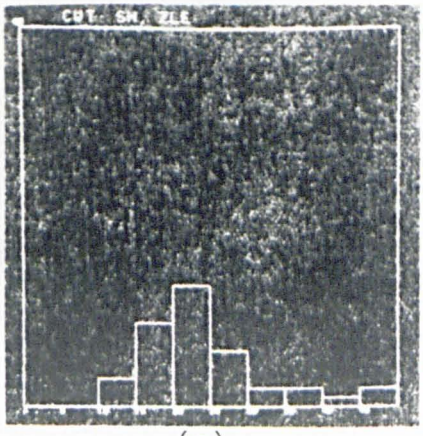


(a) (b)

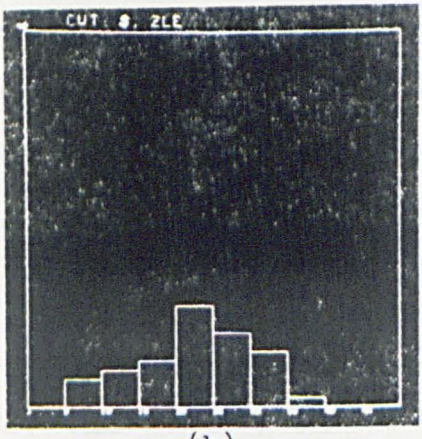
Fig. 3.68



(a)

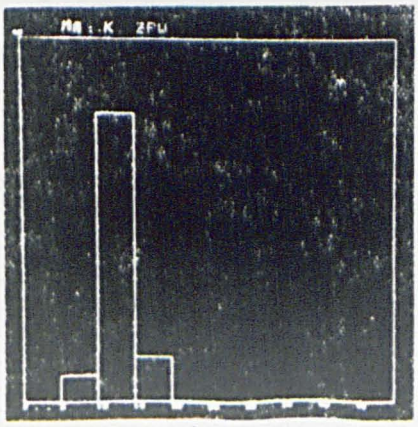


(c)

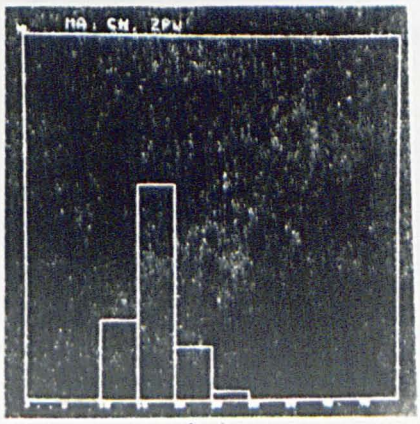


(b)

Fig. 3.69



(a)



(b)

Fig. 3.70

Z.L.E. for /t /, /s/ and /tʃ/.

3.1.17.2. Subject W.A.A.

The Z.T.I. diagrams for the utterances of /tʃ/ by W.A.A. resembled those of the fricative /ʃ/ more closely than those of the stop /t/, though the duration of /tʃ/ was generally less than that of /ʃ/.

The Z.T.I. diagrams for 2 examples of the sound /tʃæ/ spoken by W.A.A. are shown in figure 3.68. In figure 3.68, the vowel peak on the Z. trace is much larger for the second member of the pair of sounds than for the first. This was a common feature of the utterances by W.A.A., and was due to differences in pitch and stress between the first and second sounds of a pair. Figure 3.71 is a U.V. recording of 2 waveforms for the vowel /ɜ/, taken from the first and second members of a pair of /tɜ/ sounds. The increased number of zero crossings for the second /ɜ/ sound can be clearly seen.

3.1.17.3. Subject M.A.

The Z.T.I. diagrams for /tʃ/ for this subject were similar to those obtained for C.W.T. The onset times were less than those of the fricatives, while the duration of /tʃ/ was greater than that of the Voiceless Stops. Figure 3.70

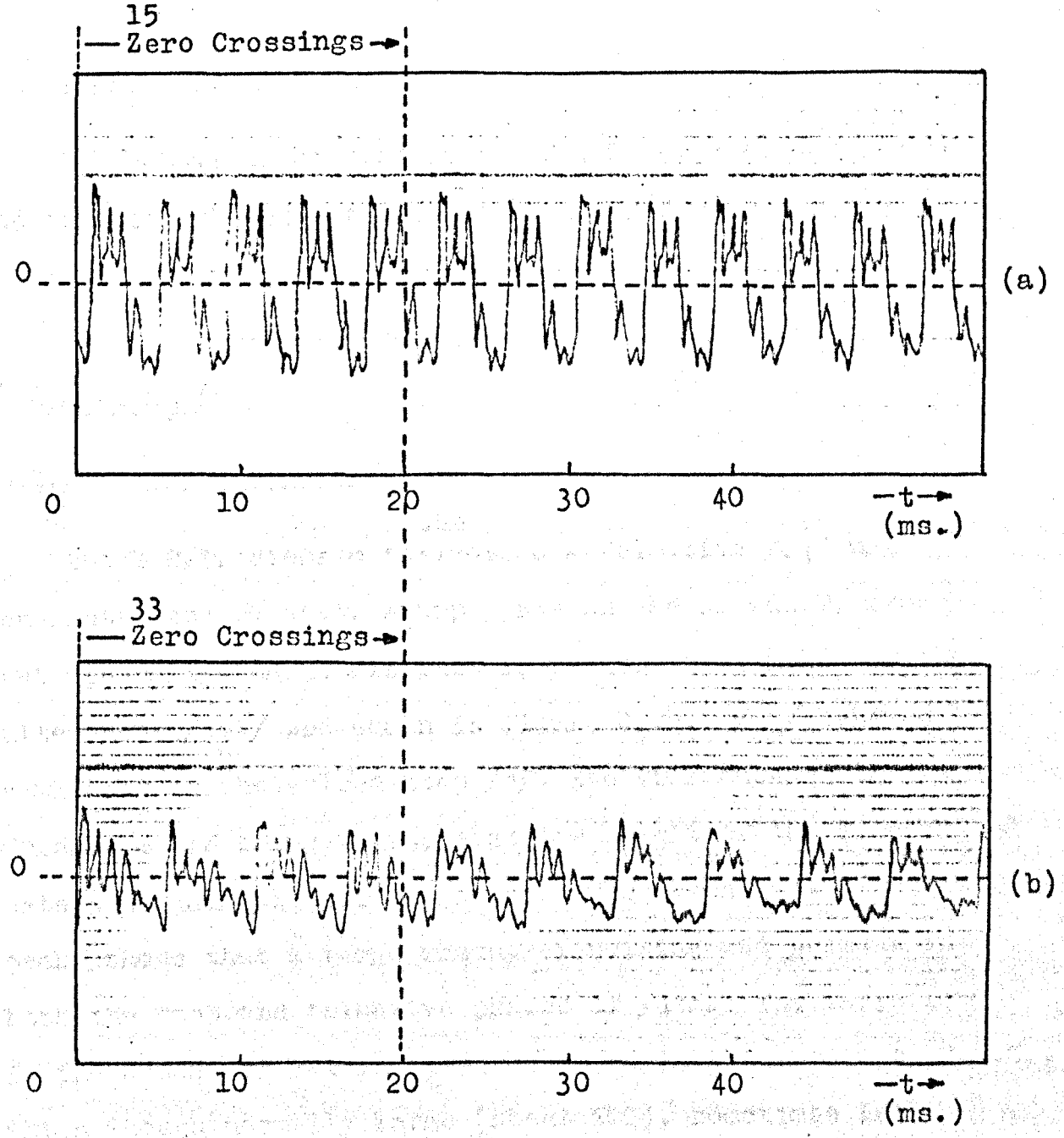


Fig. 3.71 Segments of the Raw Speech Waveforms for Two Utterances of /ʒ/ by W.A.A.
(a) In the first member of a pair of /tʒ/ sounds
(b) In the second member.

shows the distributions of Z.P.W. for /tʃ/ and /k/.

3.1.17.4. Subject P.D.G.

The behaviour of /tʃ/ on the Z.T.I. diagram was similar to that of subject W.A.A., resembling /ʃ/ more closely than /t/.

3.1.18. /dʒ/.

3.1.18.1. Subject C.W.T.

The Z.T.I. diagram for the voiced affricative /dʒ/ was characterised by high, sharp peaks on the Z. and T. traces, but generally had no distinct I. peaks. 2 examples of the utterances /dʒa/ are shown in figure 3.72. /dʒ/ behaved similarly to the voiced stop /d/, the influence of the fricative /ʒ/ being to increase the height of the Z. peak and extend the duration of the sound. The absence of distinct I. peaks shows that a large amount of voicing was present in both the stop and fricative phases of /dʒ/. The variation in the amount of voicing was reflected in the values of Z.P.S., which though normally large (about 200), sometimes fell towards the typical /d/ value (about 100).

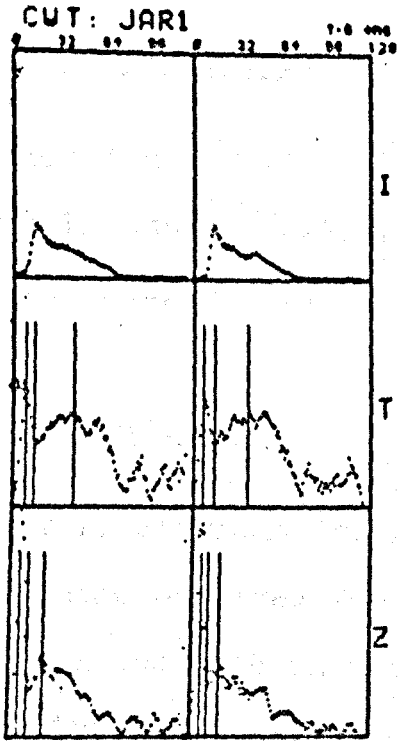


Fig. 3.72

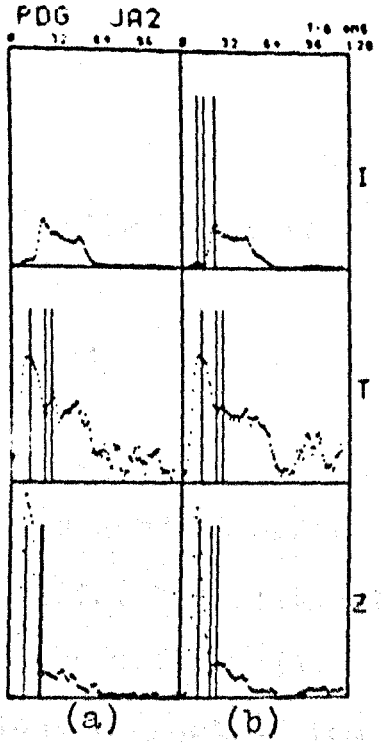


Fig. 3.74

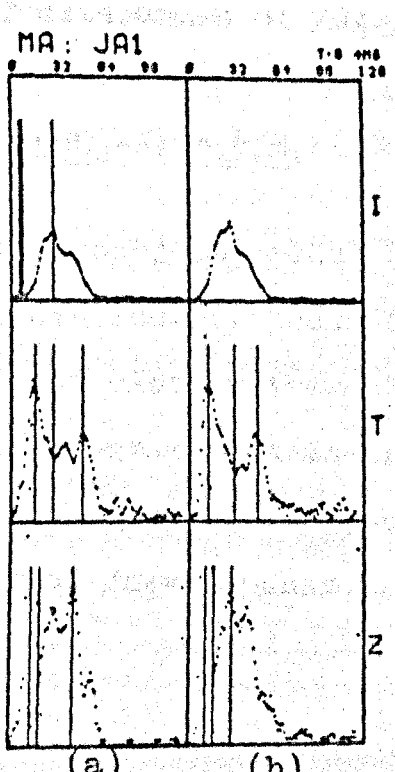


Fig. 3.73

3.1.18.2. Subject W.A.A.

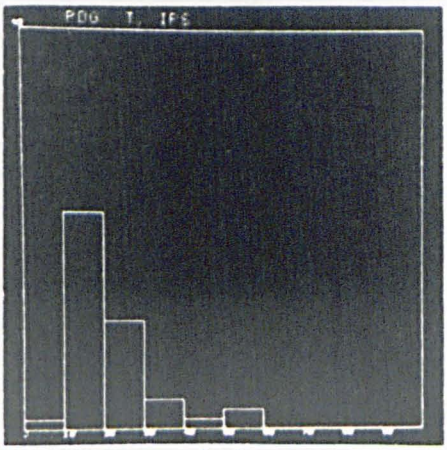
The behaviour of the Z.T.I. diagram for /dʒ/for the utterances by subject W.A.A. was very similar to that observed for subject C.W.T.

3.1.18.3. Subject M.A.

The Z.T.I. diagrams for /dʒ/ for this subject again resembled those observed for subject C.W.T., but distinct I. peaks were observed in about 15% of the utterances, showing an increased effect of the fricative part of the combined sound. Figure 3.73 shows the Z.T.I. diagrams for a pair of utterances of /dʒɑ/ by subject M.A.

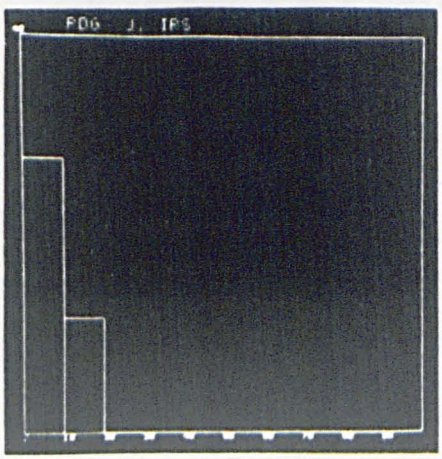
3.1.18.4. Subject P.D.G.

For this subject, about 50% of the utterances showed small but distinct I. peaks on the Z.T.I. diagram. These examples most closely resembled the voiceless stop /t/ on the Z.T.I. diagram, but their I.P.S. values were generally smaller than those of /t/ (see figure 3.75). Z.T.I. diagrams for a pair of /dʒɑ/ sounds are shown in figure 3.74.



(a)

Fig. 3.75



(b)

3.1.19. The Vowel Like Sounds. Glides and Nasals.

Since it was generally difficult to separate the 6 Glides and Nasal Phonemes from each other, this group of sounds will be discussed as a whole, rather than by individual phonemes.

The Vowel-like sounds could be distinguished from the other consonants by the behaviour of the I. trace. While the I. trace followed the gradual changes in amplitude level occurring during the production of the consonant, a number of small bumps or 'crests' frequently appeared. This effect can be seen in the Z.T.I. diagrams of figure 3.76(sound /jɑ/, speaker C.W.T.). These small variations in the envelope can also be seen in the vowel part of the Z.T.I. diagrams and occasionally in the voiced Fricatives (e.g. Figure 3.61). Figure 3.87 is a U.V. recording of the raw speech waveform for the sound /jɑ/ of figure 3.76(a). The short term variations in the envelope causing the appearance of crests on the I. trace can be clearly seen. These crests seem to be a common feature of vowels and vowel-like sounds, but did not generally occur for the Fricatives and Stops. The number and prominence of the crests varied from sound to sound and from speaker to speaker.

Although the crests do not constitute I. peaks in the usual sense, as they do not separate the consonant from the vowel

part of the C.V. syllable, their presence provided a good means of separating the vowel-like sounds from other consonants. In figure 3.76, the part of the I. trace corresponding to the initial articulator position can be easily distinguished from the following formant glide. The I. peak was therefore positioned as normal on the initial part of the trace. Sometimes, however, this distinction could not be made, (e.g. figure 3.77, sound /ru/, speaker C.W.T.). In these cases the I. peak was positioned at about the middle of the crested region. This generally gave a higher value of I.P.S. than when the peak was positioned as normal. For these sounds, the peak-picking algorithm often failed, and the I. peak marker was positioned manually.

This procedure generally gave the Glides and Nasals higher values of I.P.S. than those of the other sounds, but the most reliable difference was the value of I.P.D. When the I. peak was due to a crest, I.P.D. was only slightly smaller than I.P.S., whereas other sounds with large I.P.S. values could have much lower values of I.P.D. (e.g. figure 3.35(a)).

3.1.19.1. Speaker C.W.T.

For this speaker the crests on the I. trace were generally quite marked, often yielding high values of I.P.S.

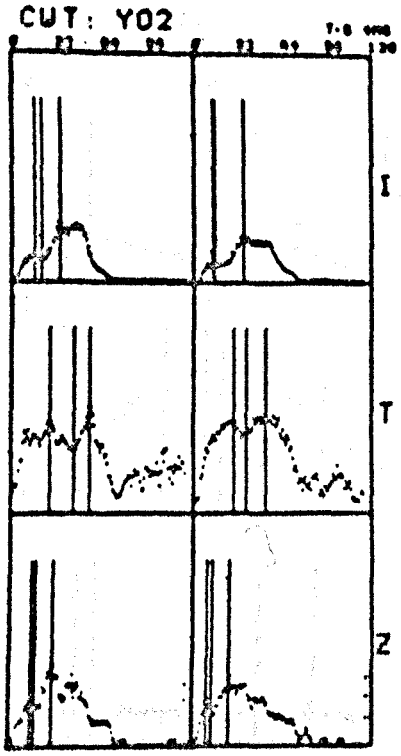


Fig. 3.76

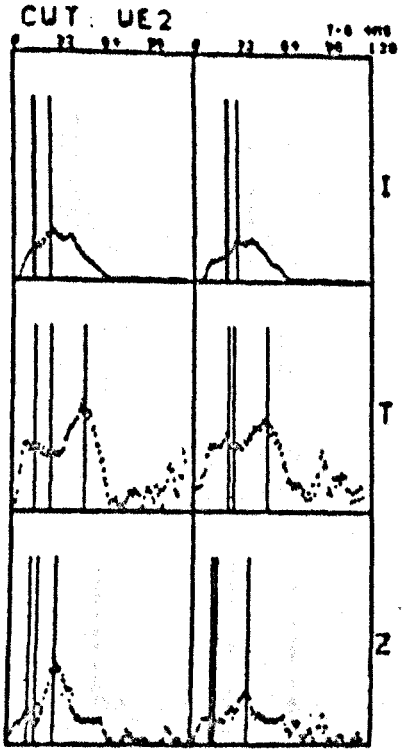


Fig. 3.78

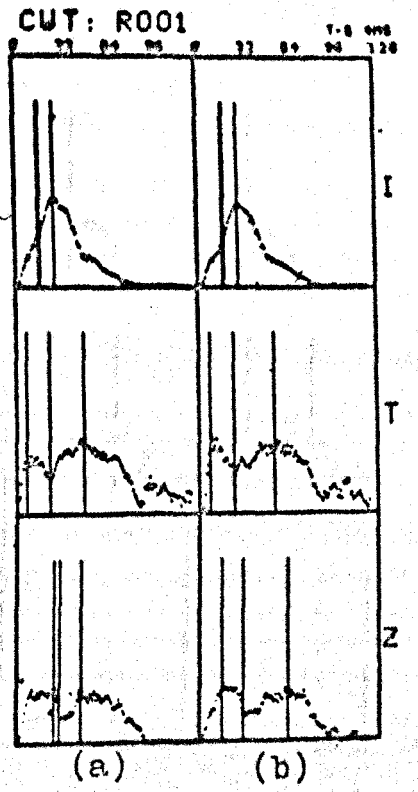


Fig. 3.77

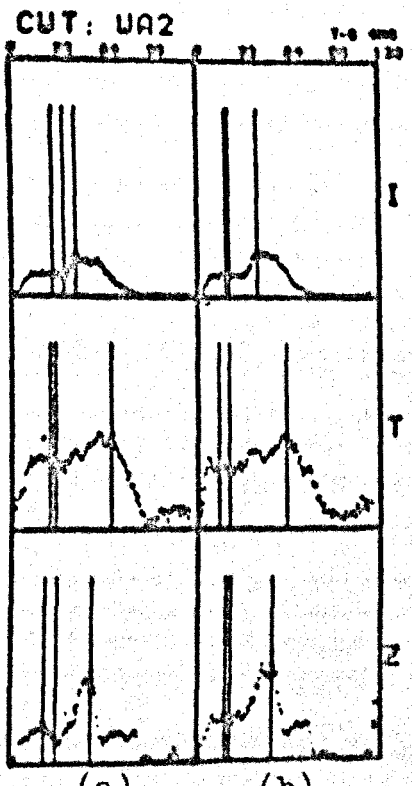


Fig. 3.79

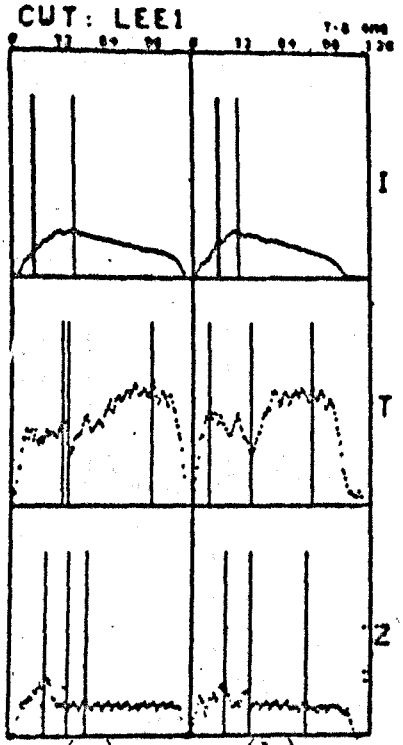


Fig. 3.80

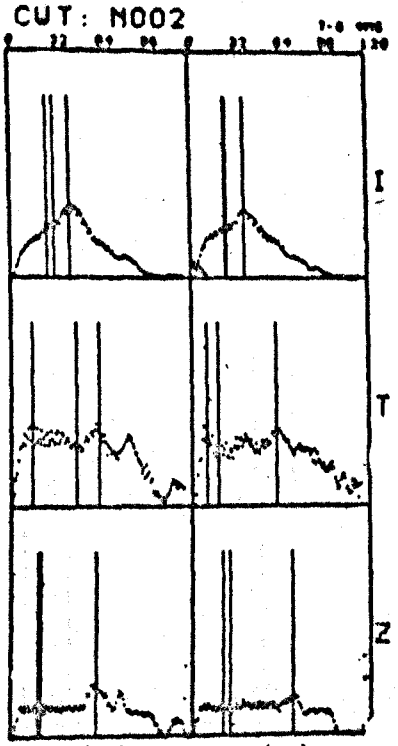


Fig. 3.82

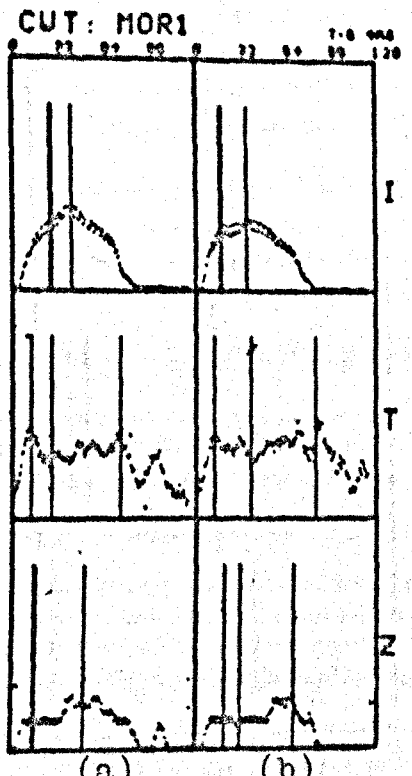
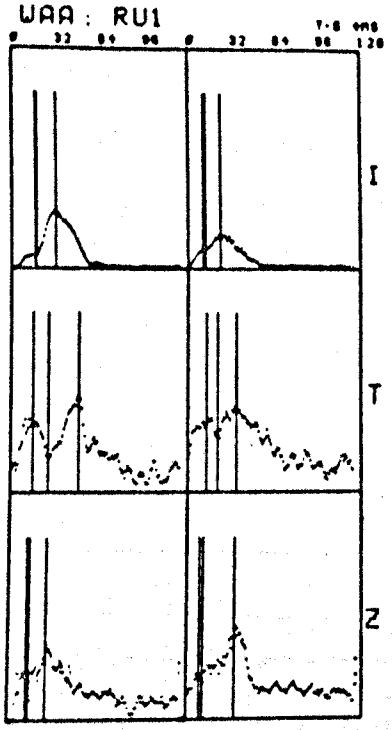
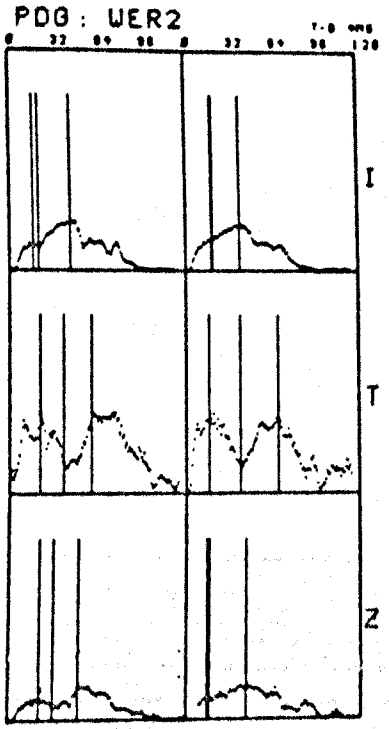


Fig. 3.81



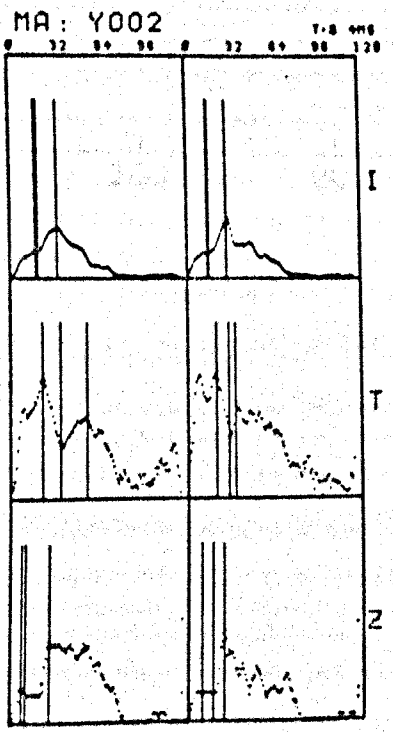
(a) (b)

Fig. 3.83



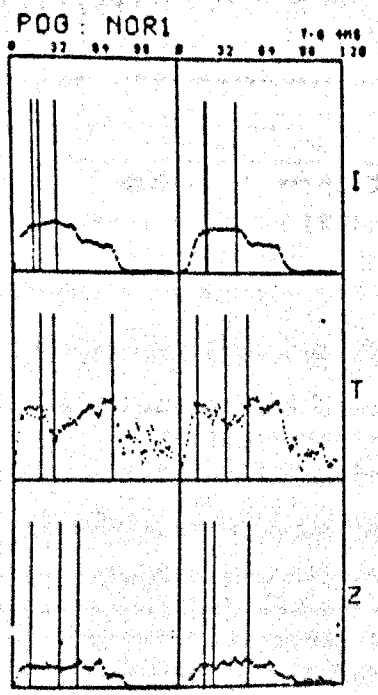
(a) (b)

Fig. 3.85



(a) (b)

Fig. 3.84



(a) (b)

Fig. 3.86

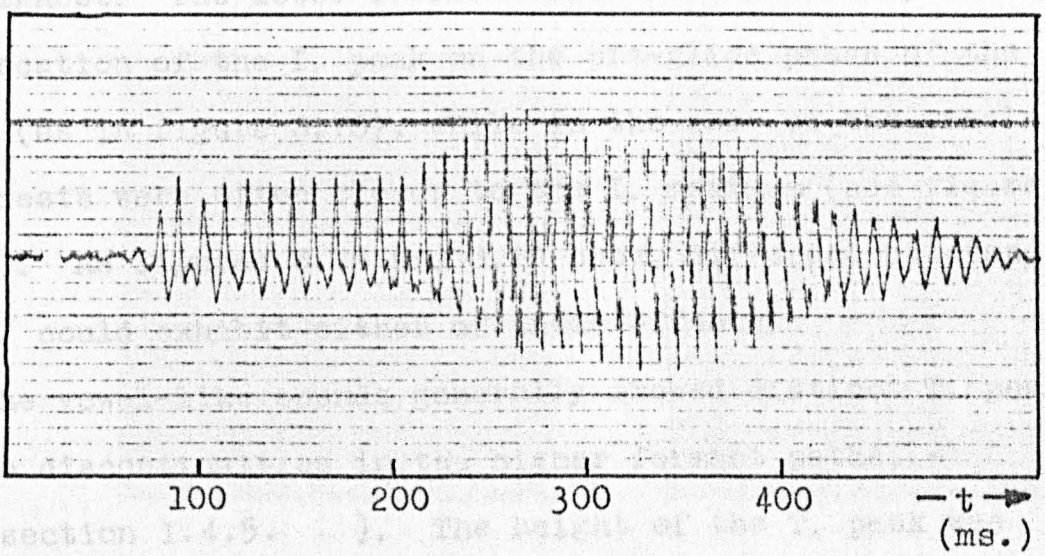
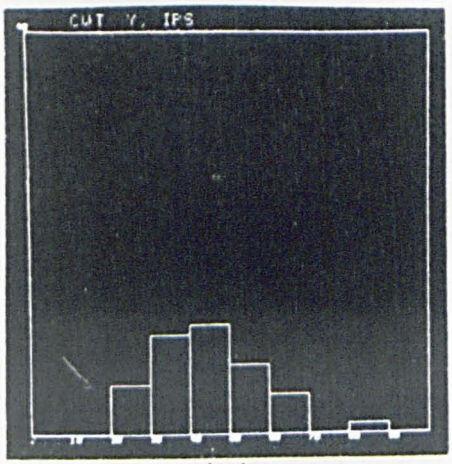


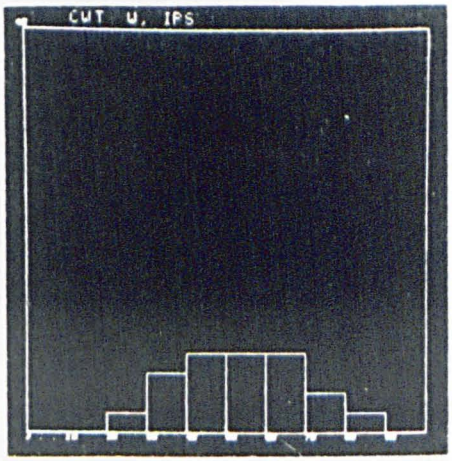
Fig. 3.87 Raw Speech Waveform for an Utterance of /jɒ/ by C.W.T.

Z.T.I. diagrams of 2 utterances of the sounds /jɑ/, /ru/, /wɛ/, /wɑ/, /li/, /mɔ/ and /nu/ are shown in figures 3.76 through 3.82. The vowel-like sounds could be distinguished f from each other to some extent using the parameter I.P.S. The distributions of I.P.S. for the 6 Glide and Nasal phonemes are shown in figure 3.88. The glides /j/ and /r/ tended to have the lowest values of I.P.S. and the nasals /m/ and /n/ the highest. The lower I.P.S. values were generally due to the location of the I. peak on the pre-glide phase of the sound (as in figure 3.76), while in the case of the nasals, the crests were often closer to the I. maximum (see figure 3.81.). As figures 3.78 and 3.79 show, different utterances of /w/ could exhibit either of these effects.

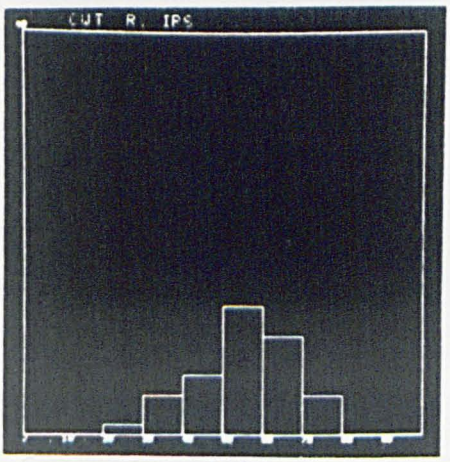
The vowel-like sounds generally showed distinct T. peaks due to discontinuities in the higher formant paths, - (see section 1.4.5.). The height of the T. peak was generally low. The values of T.P.D. were generally lower for /r/, /l/ and /w/ than for /y/, /m/ and /n/, due to a sharper higher formant change. The distributions of T.P.D. for the 6 vowel-like phonemes are shown in figure 3.89. The sonagrams of figures 1.7(b),(c) and (d) indicate that the higher formant change is more abrupt in the cases of /r/, /l/ and /w/.



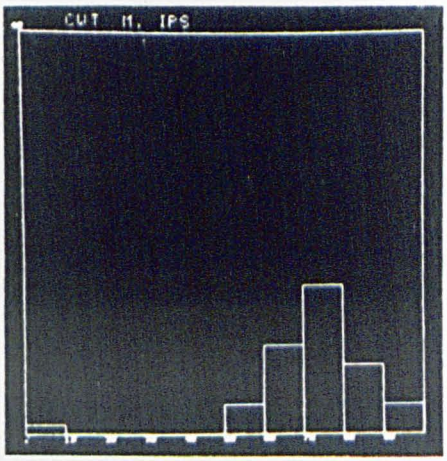
(a)



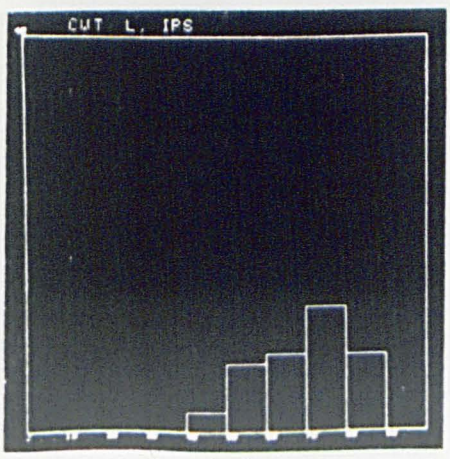
(d)



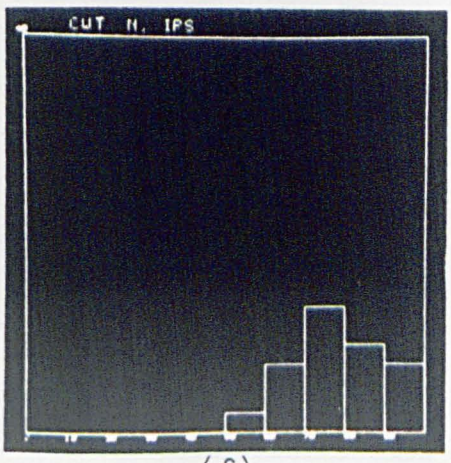
(b)



(e)

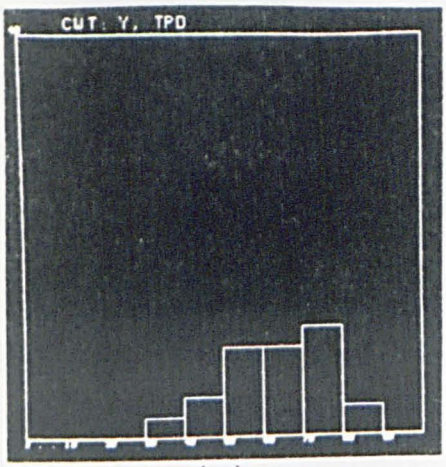


(c)

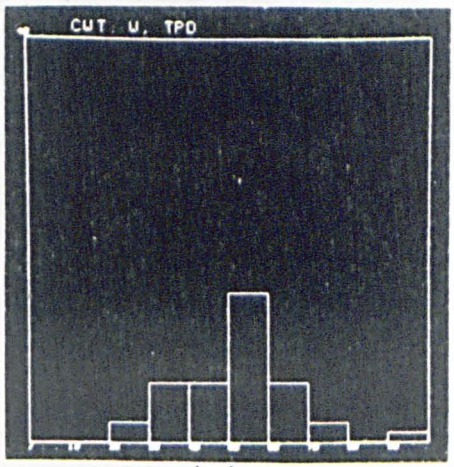


(f)

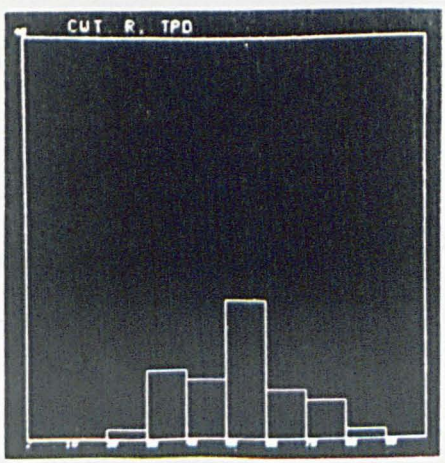
Fig. 3.88



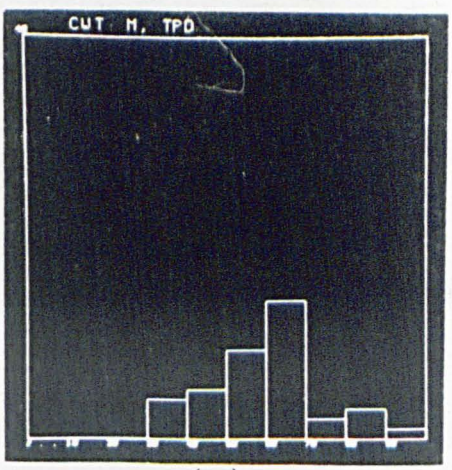
(a)



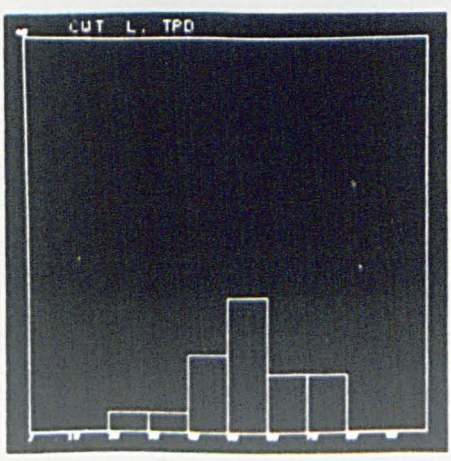
(d)



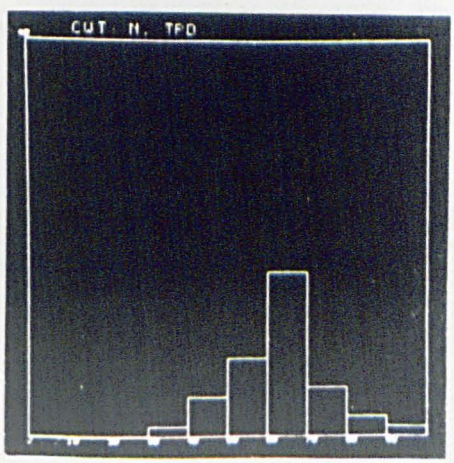
(b)



(e)



(c)



(f)

Fig. 3.89

It can be seen from figures 3.76 to 3.82 that the Z.peaks for the vowel-like sounds were often dubious, and little useful information could be gained from the Z. parameter. The level of the Z. trace was uniformly low.

3.1.19.2. Speaker W.A.A.

For this subject the 'crests' on the I. trace were less prominent and fewer in number than in the case of subject C.W.T. This is illustrated by the Z.T.I. diagrams for a pair of /rɑ/ sounds spoken by W.A.A. shown in figure 3.83. About 10% of the Z.T.I. diagrams did not show a distinct I. peak. This corresponds to the generally less clear I. behaviour for this subject. The I.P.S. values were generally lower than in the case of C.W.T. This is illustrated in figure 3.90 which shows the distribution of I.P.S. for /r/. As figure 3.83(a) shows, low values of T.P.D. still occurred for this subject, but this parameter was less useful than in the case of C.W.T.

Those vowel-like sounds which did not show a distinct I. peak were difficult to separate from /h/, /b/, /d/ and /g/, but their Z. and T. peaks were often somewhat smaller.

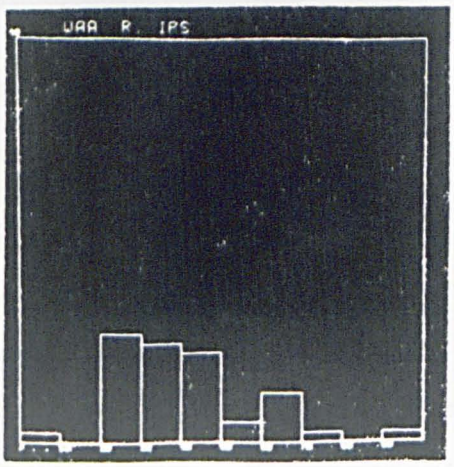
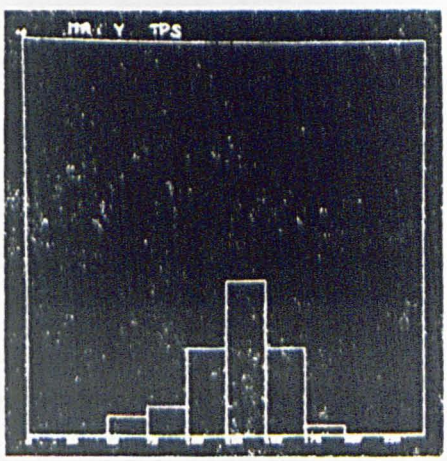
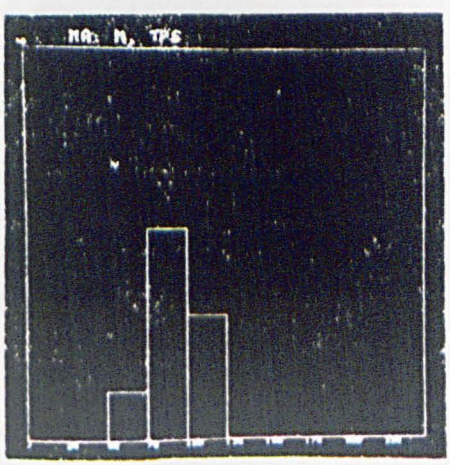


Fig. 3.90



(a)

Fig. 3.91



(b)

3.1.19.3. Speaker M.A.

The crests on the I. peak were again fewer in number and less pronounced for this subject, but the T.P.S. values were very similar to those obtained for subject C.W.T. As in the case of subject W.A.A., an appreciable number of sounds (about 10%) did not show a separate I. peak. These sounds could be separated from other phonemes to some extent by the smaller size and width of their Z. and T. peaks, and their larger I.L.E. value. Figure 3.34 shows the distributions of I.L.E. for /h/, /b/ and /w/.

The Z.T.I. diagrams for /j/ were distinguished by exceptionally large and wide T. peaks for a vowel-like sound. 2 examples of Z.T.I. diagrams for the sound /ju/ are shown in figure 3.84. The distributions of T.P.S. for /j/ and /m/ are shown in figure 3.91. The presence of large amounts of energy at higher formant frequencies and above can be seen in the sonogram (figure 3.92) for the /ju/ sound of figure 3.84(b).

3.1.19.4. Speaker P.D.G.

For this subject the Z.T.I. diagrams for the vowel-like sounds resembled those of subject C.W.T., generally in a slightly exaggerated form. The I. crests were often more

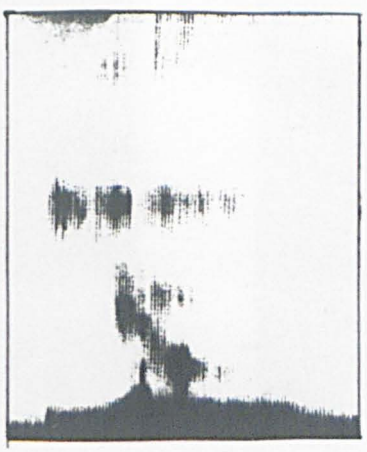
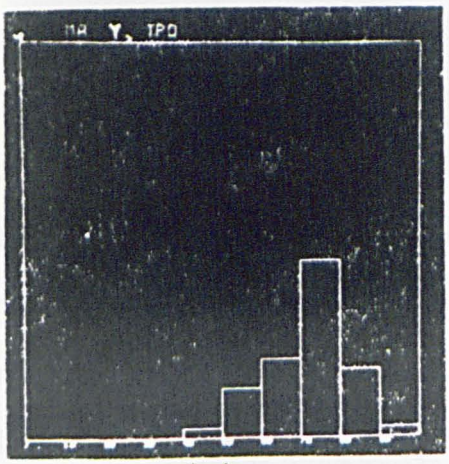
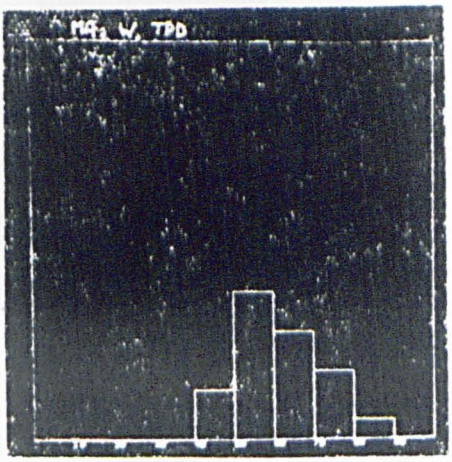


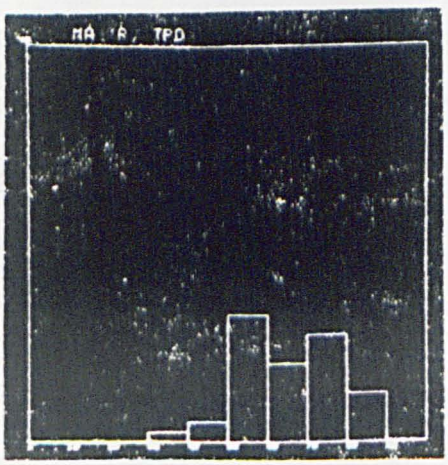
Fig. 3.92 Sonagram for an utterance of /ju/ by M.A.



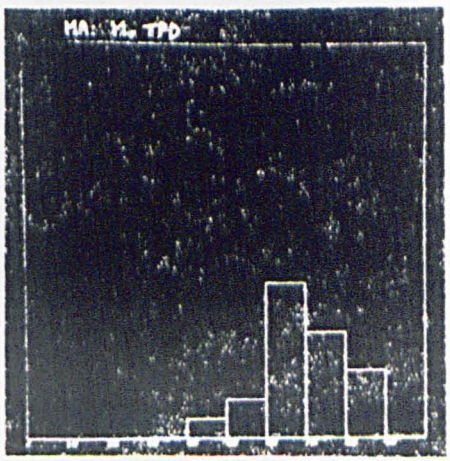
(a)



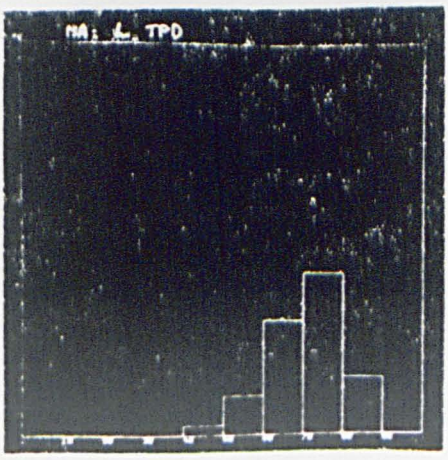
(d)



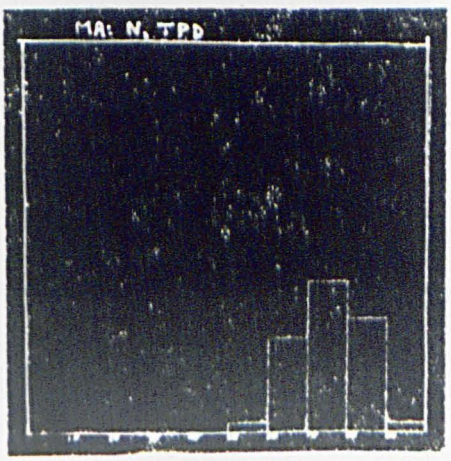
(b)



(e)



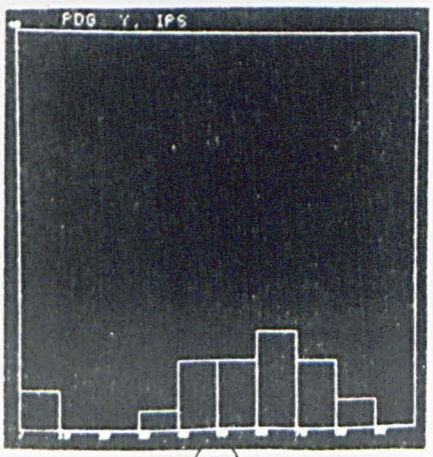
(c)



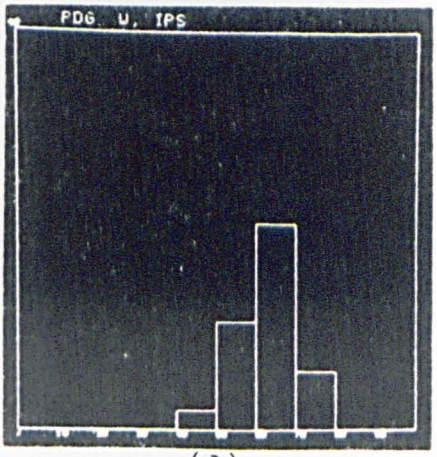
(f)

Fig. 3.93

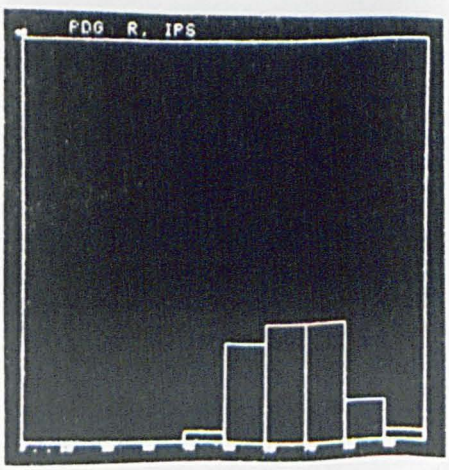
numerous than for C.W.T. and the I.P.S. and I.P.D. values were higher. The T.P.D. values were slightly lower, indicating a sharper formant discontinuity at the onset of the vowel. The distributions of I.P.S. and T.P.D. for the vowel-like sounds are shown in figures 3.94 and 3.95. The nasals /m/ and /n/ generally had the largest I.P.S. values, while /r/ and /w/ had lower values of T.P.D. than /j/. Z.T.I. diagrams for 2 examples of /wɜ/ are shown in figure 3.85. Figure 3.85 shows the sharp drop on the T. trace at the onset of the vowel. Figure 3.86 shows the Z.T.I. diagrams for 2 examples of /nɔ/, with very high I.P.S. values.



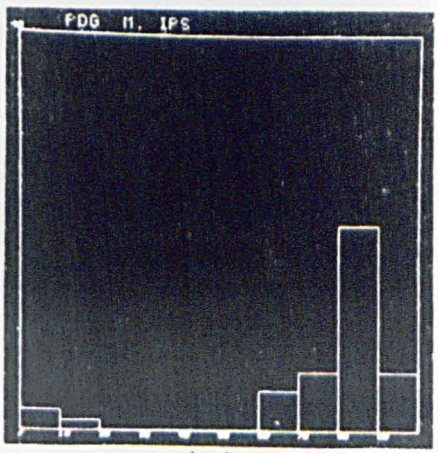
(a)



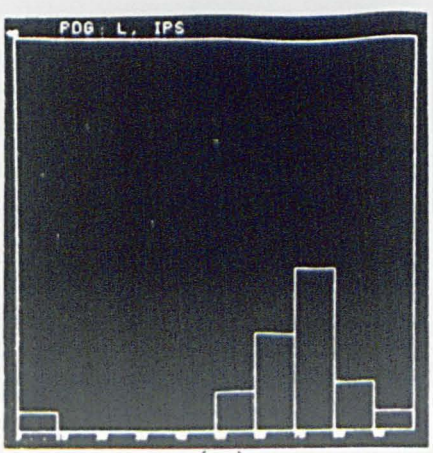
(d)



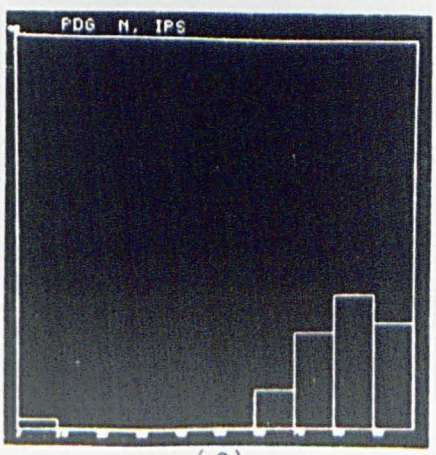
(b)



(e)

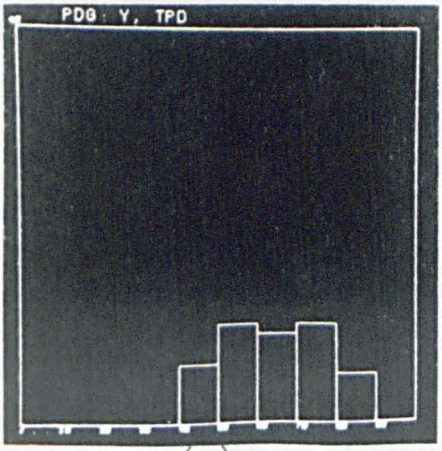


(c)

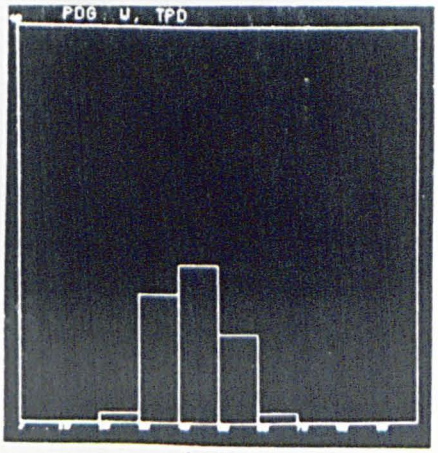


(f)

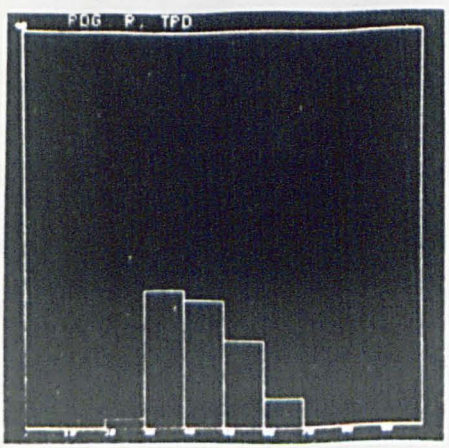
Fig. 3.94



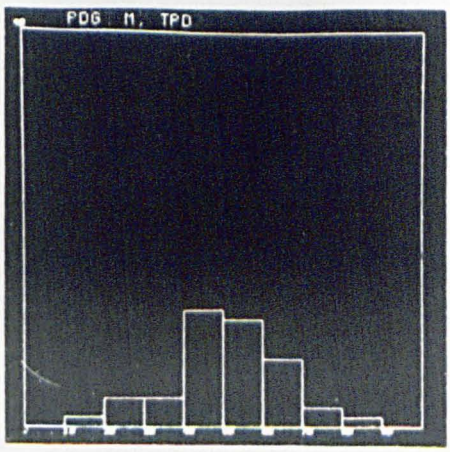
(a)



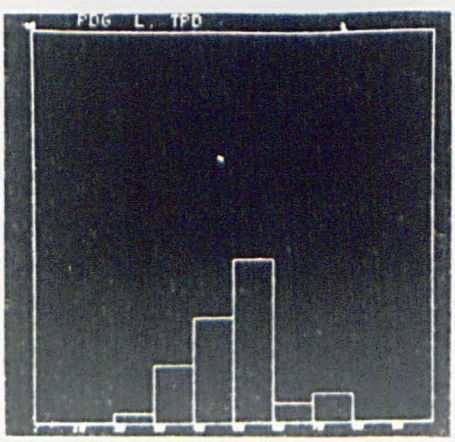
(d)



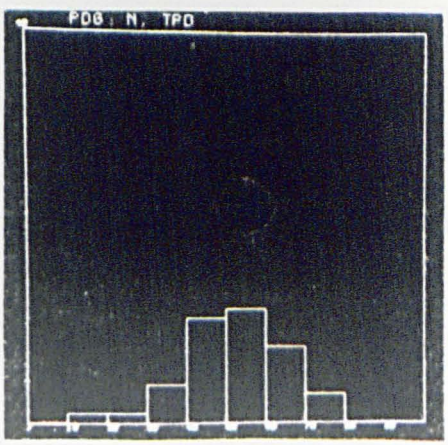
(b)



(e)



(c)



(f)

Fig. 3.95

3.2. Recognition Algorithms.

For each subject, the available data comprised the parameter sets for 4 examples of each consonant phoneme, spoken with each of 10 vowels. There were thus 40 repetitions of each consonant for a single speaker (Section 1.1.).

A recognition algorithm of the type described in section 2.2. was evolved for each subject, using the design method outlined in Section 2.2.2.

It was originally intended to design these algorithms using $\frac{1}{2}$ of the data (2 utterances of each consonant with each vowel) as a training set, and then to evaluate the algorithm performance on the remaining test set. However, the number of examples falling into each Group (see Section 2.2.1.) declined with each succeeding decision level, and beyond level 4 these were frequently insufficient numbers on which to base a meaningful decision. In these cases it was necessary to invoke the test set data in order to enable further separation of the phonemes. This procedure generally enabled distinction between individual consonant phonemes (except the vowel-like sounds), most of the end points being associated with a single "probable" phoneme. There was thus no real distinction between training and test sets, and the results presented in Section 3.3. cover the whole of the data.

Lack of time prevented the processing of further C.V. sounds to acquire more data. This could have been done by dealing with fewer subjects, but it was felt that 4 speakers were the minimum number needed to enable some assessment of subject differences. In any case, the limited storage capacity of the P.D.P.-8 would mean re-writing the algorithm design programme if larger bodies of data were to be used. The standard errors of the recognition rates are small enough to expect no major decline in the performance of the algorithms on new data.

The following sections describe the algorithms evolved for each subject, and the performance of these algorithms.

3.2.1. Separation into Groups.

The first operation in the recognition process was to assign the consonant to one of the 4 or 5 Groups. This was done by a sequential series of Group Decisions. The object of the Group separation was to reduce the number of phoneme categories to be considered by later stages of the algorithm to manageable proportions.

Each Group was designed to trap certain phonemes, though there was necessarily a large amount of overlapping between the Groups. In some cases errors in the Group decision were corrected at a later stage in the algorithm.

The sequences of Group decisions for the 4 speakers are shown in figures 3.96 to 3.99.

3.2.1.1. The Group 1 decision.

G1 included those sounds which had a large Consonant Peak on the Z. and T. traces, much higher than that of the vowel. Due to high frequency domination of the consonant spectrum, there was generally a sharp Z.P.S. distinction between these sounds and the rest.

For all subjects, G1 included the bulk of the utterances of /tʃ/, /dʒ/, /s/ and /ʃ/. Other Phonemes entering G1 varied from subject to subject as shown in figure 3.100.

The values of the Z.P.S. threshold for subjects C.W.T and P.D.G. (211 and 210 respectively) were remarkably similar, while that for subject M.A.(148) was a good deal less, corresponding to the generally lower magnitude of the Consonant Peaks for this subject. In the case of subject W.A.A., the best Z.P.S. threshold (286) gave a slightly less reliable separation than the T.P.S. decision. This was due to the variation in the height of the Z. peak for the vowel between the first and second members of a pair of C.V. sounds. (See Section 3.1.17.2.).

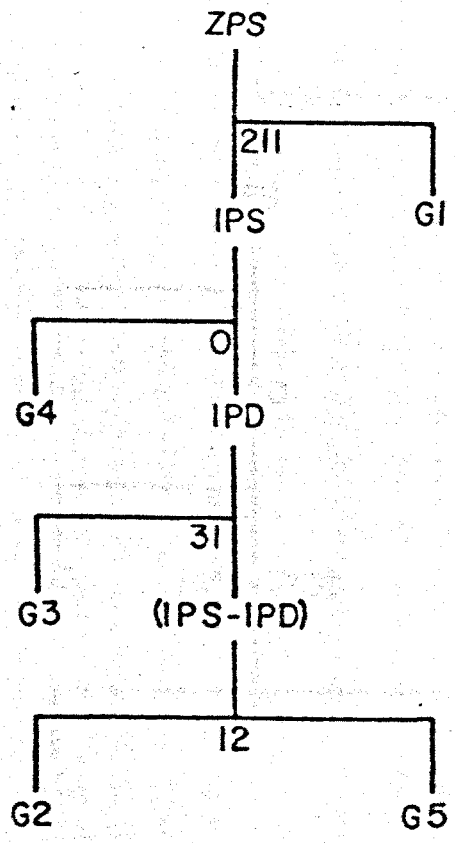


Fig. 3.96 Recognition Algorithm for Subject C.W.T. : Group Control

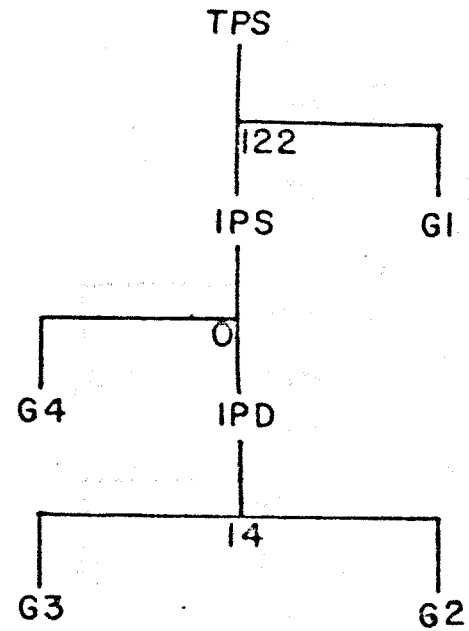


Fig. 3.97. Recognition Algorithm for Subject W.A.A. : Group Control.

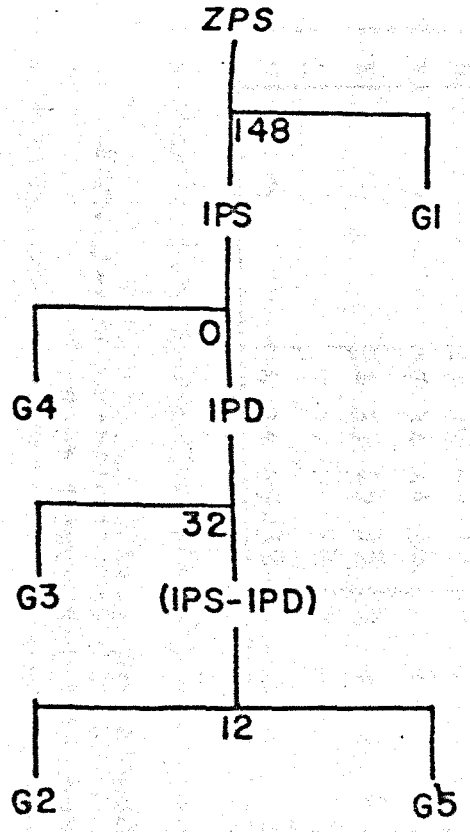


Fig. 3.98 Recognition Algorithm for Subject M.A. : Group Control.

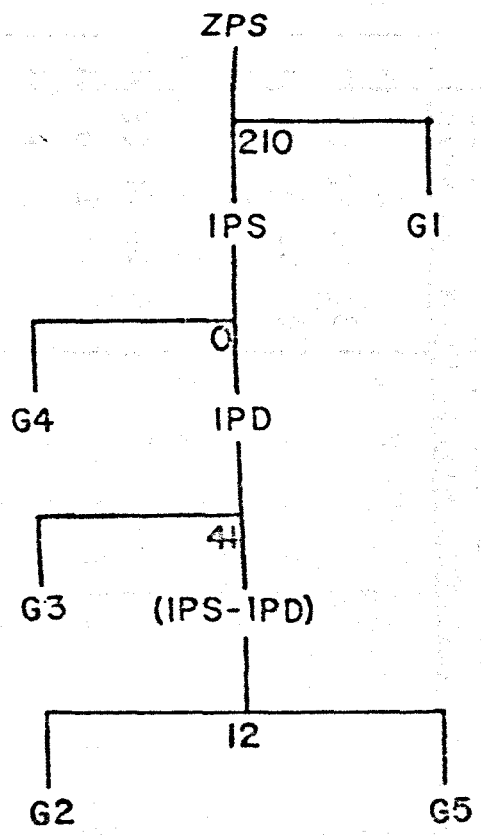


Fig. 3.99 Recognition Algorithm for Subject P.D.G. : Group Control.

	G1				G2				G3				G4				G5		
	CWT	WAA	MA	PDG	CWT	WAA	MA	PDG	CWT	WAA	MA	PDG	CWT	WAA	MA	PDG	CWT	MA	PDG
P		1			9	9	10	1	2	22	1	13	1	8	14		28	15	26
T	19	32	20	36					20	8	5	4	1		15				
K	20	6	22	5		1	1	3	20	31	12	27		2	5	5			
F	2	1	9					2	10	3	1	30	25	36	27	3	3	3	5
TH		2	1	1	1	1	1		22	7		9	11	30	34	30	6	4	
CH	39	38	40	40		2			1										
H	6	1	1				1	3	4		3	6	28	39	35	30	2		1
B	4					2	2		1	4	2	4	35	34	31	35		5	1
D	2	5	2	3		1	2		2	2	1		36	32	34	37		1	
G	6	1	1	1						3		3	34	36	39	36			
V					6		1	3	28	28	37	22	5	12	2	14	1		1
DH		1			2	1			35	29	37	18	3	9	2	22		1	
J	33	36	32	33						1	1		7	3	7	7			
S	39	40	40	40									1						
SH	39	33	40	40		2			1	5									
Z	5	39	3	25	7		2		25	1	35	12	3			2			1
ZH	5	37	30	15	1	2	2	0	29	1	6	15	3		2	4	2		
Y		2	1		31	30	27	31	8	3	5	5		5	7	4	1		
R					39	38	23	39	1	1				1	15			2	1
L		1			37	34	34	38		2				3	2	2	3	4	
W					34	33	15	35	5	1	2	1		6	21		1	2	4
M					37	37	24	38	1					3	14	2	2	2	
N		1			36	34	29	39						5	9	1	4	2	

Fig. 3.100 Distribution of Phonemes Falling Within Each Group for the Four Algorithms.

3.2.1.2. The Group 4 decision.

All the sounds not entering G1 which did not show a separate consonant peak on the I. trace were allocated to G4. G4 was mainly intended to cope with the phonemes /h/, /b/, /d/ and /g/. Apart from these phonemes, the sounds entering G4 varied widely from subject to subject (See figure 3.100). This Group was generally the most complicated, especially for subjects W.A.A. and M.A. when a greater proportion of sounds did not possess a distinct I. peak.

3.2.1.3. The Group 3 decision.

This decision performed the main separation between the vowel-like sounds (/j/, /r/, /l/, /w/, /m/ and /n/) and the remainder. The vowel-like sounds were characterised by larger values of I.P.D., due to the appearance of "crests" on the I. trace (Section 3.1.19). The values of the I.P.D. threshold varied from subject to subject with the prominence of the crests.

3.2.1.4. The Group 2 - Group 5 decision.

The remaining sounds were comprised largely of the vowel-like phonemes on the one hand, and examples of other phonemes (chiefly /p/) with an exceptionally large I. peak on the other.

The most reliable way of separating these two groups was to use the difference between I.P.S. and I.P.D. The value of (I.P.S. - I.P.D.) was generally less than 12 for the vowel-like phonemes, due to their I. crests, and greater than 12 for the remainder.

Group 5 was generally much simpler than the other groups; in the case of subject W.A.A. the Group 2 - Group 5 decision was not necessary, since the utterances of /p/ normally entered G3 or G4.

The next sub-section describes each of the 5 groups in the algorithm evolved for each subject.

3.2.2. Group 1.

3.2.2.1. Subject C.W.T. (figure 3.101.)

The first decision in Group 1 separated /dʒ/, which normally did not show an I. peak, from the remainder of the phonemes. A few examples of /t/ and /tʃ/ which had no I. peak also entered G121, together with those few voiced stops with abnormally large Z. peaks which had entered G1. The latter were separated from /dʒ/, /t/ and /tʃ/ by a T.P.S. decision and re-directed to G4. At G122, /s/ and /ʃ/ were separated out by means of their greater duration. The G133 decision separated /t/ and /k/ from /tʃ/, which had larger Z. onset

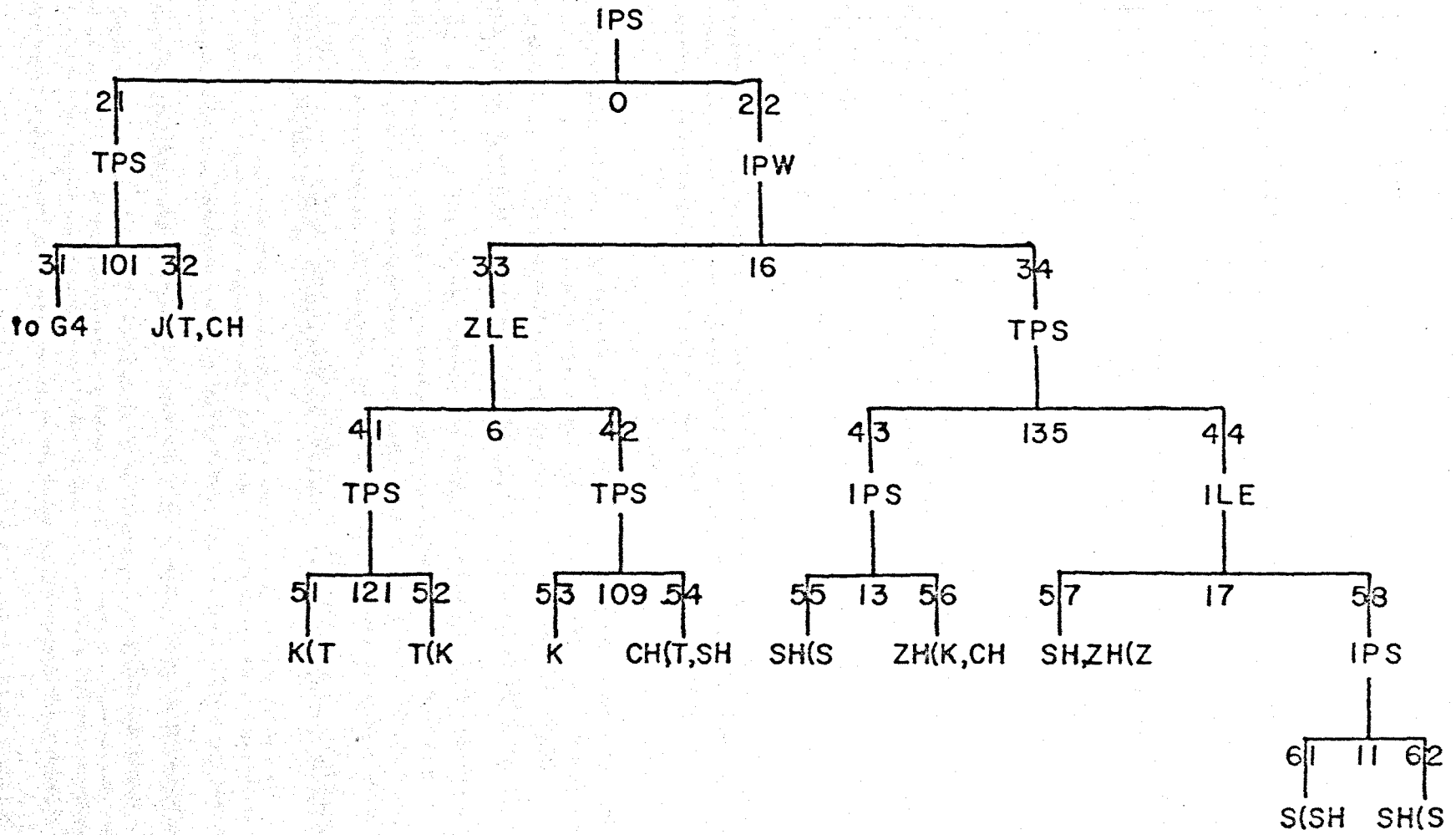


Fig. 3.101 Recognition Algorithm for Subject C.W.T. Group 1.

times : /t/ was then distinguished from /k/ by means of its higher T. peak (G141). The main distinction between /s/ and /ʃ/ was made at G134, /s/ having higher T. peaks than /ʃ/.

3.2.2.2. Subject W.A.A. (figure 3.102).

For this subject, G1 was complicated by the presence of large numbers of the voiced fricatives /z/ and /ʒ/, due to the use of T.P.S. instead of Z.P.S. in the G1 decision.

The voiced stops were this time separated from /dʒ/ using a T.P.W. decision (G121), the duration of /dʒ/ being larger than that of the voiced stops. The phonemes entering G132 were comprised mainly of /dʒ/, with a few examples of /z/ and /ʒ/. No reliable decision could be found for distinguishing /dʒ/ from /z/ and /ʒ/, but the T.P.S. decision separated /z/ from /ʒ/, /dʒ/ falling almost equally between G141 and G142.

An I.P.W. decision was again used to separate /s/ and /ʃ/ (G122), and at G133 a Z.P.W. decision separated /t/ from /tʃ/, /z/ and /ʒ/. /z/ and /ʒ/ generally had smaller Z. peaks than /tʃ/, and this fact was utilised at G144. A T.P.D. decision at G153 then separated /z/ from /ʒ/ (see figure 3.66). Most of the sounds entering G154 were examples of /tʃ/, and the Z.P.W. decision distinguished the additional utterances of /t/ and /dʒ/ from those of /z/ and /ʒ/.

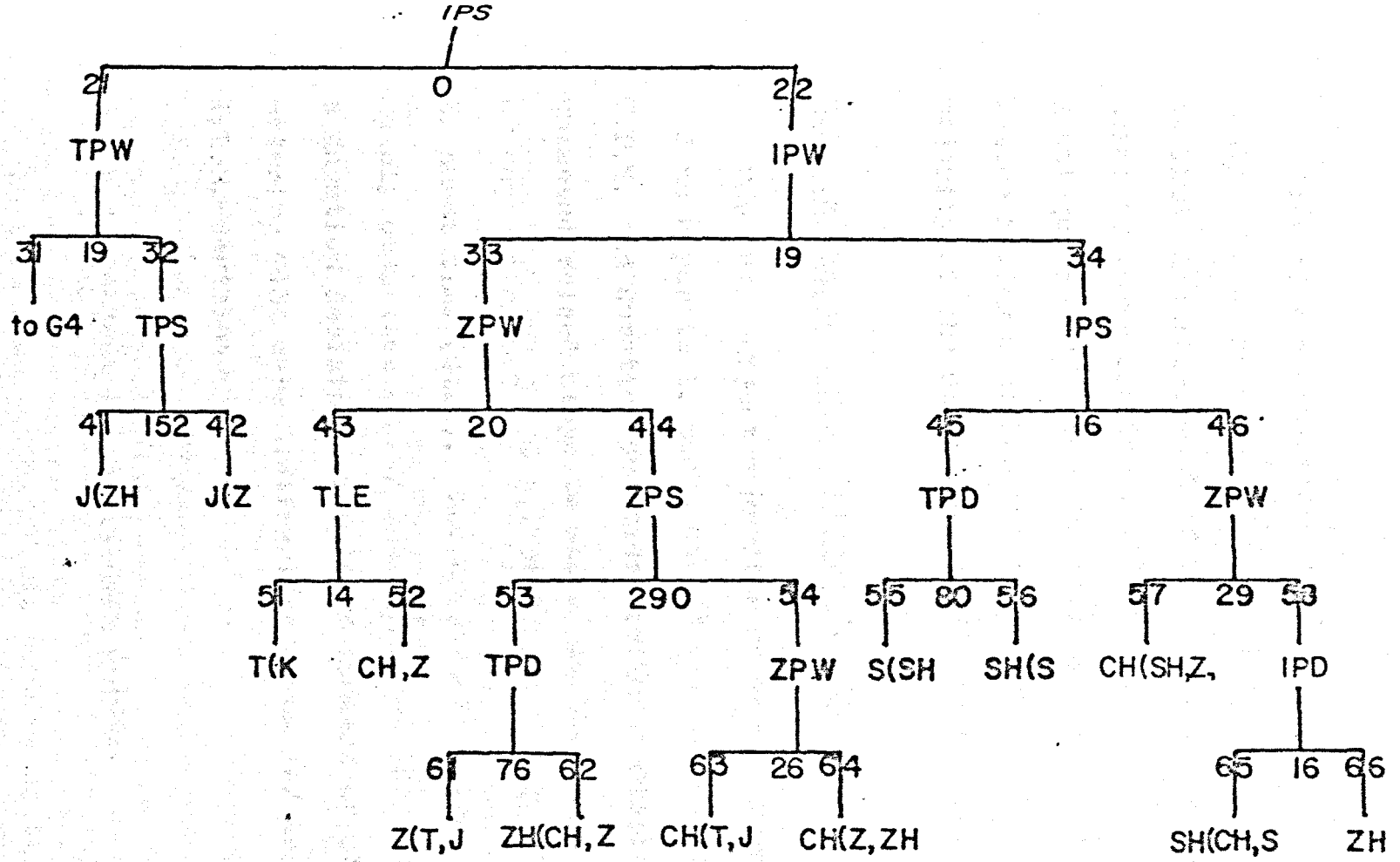


Fig. 3.102 Recognition Algorithm for Subject W.A.A.
Group 1.

Most of the utterances of /s/ terminated at G155, having smaller I. peaks than /ʃ/ (see figure 3.54). A Z.P.W. decision at G146 identified most of the remaining examples of /tʃ/ by means of their shorter duration. The I.P.D. decision at G158 distinguished /ʃ/ from the remaining examples of /ʒ/, the latter having a higher but less distinct I. peak, and the majority of the utterances of /ʃ/ arrived at G165.

3.2.2.3. Subject M.A. (figure 3.103).

For this subject a greater number of sounds entering G1 had no separate I. peak, and it was necessary to use both T.P.S. and duration decisions (G132, G131 and G131) to direct the unwanted voiced stops to G141. In this case the best means of separating out /s/ and /ʃ/ was to make use of their greater Z. onset time (G122). The utterances entering G133 were mainly comprised of /k/ and /tʃ/, and these were separated by a duration decision. The large numbers of utterances of /ʒ/ entering G134 were distinguished from /s/ and /ʃ/ using an I.L.E. decision.

3.2.2.4. Subject P.D.G. (figure 3.104).

In this case, considerable numbers of the utterances of /z/ entering G1 did not have a separate I. peak, and these were

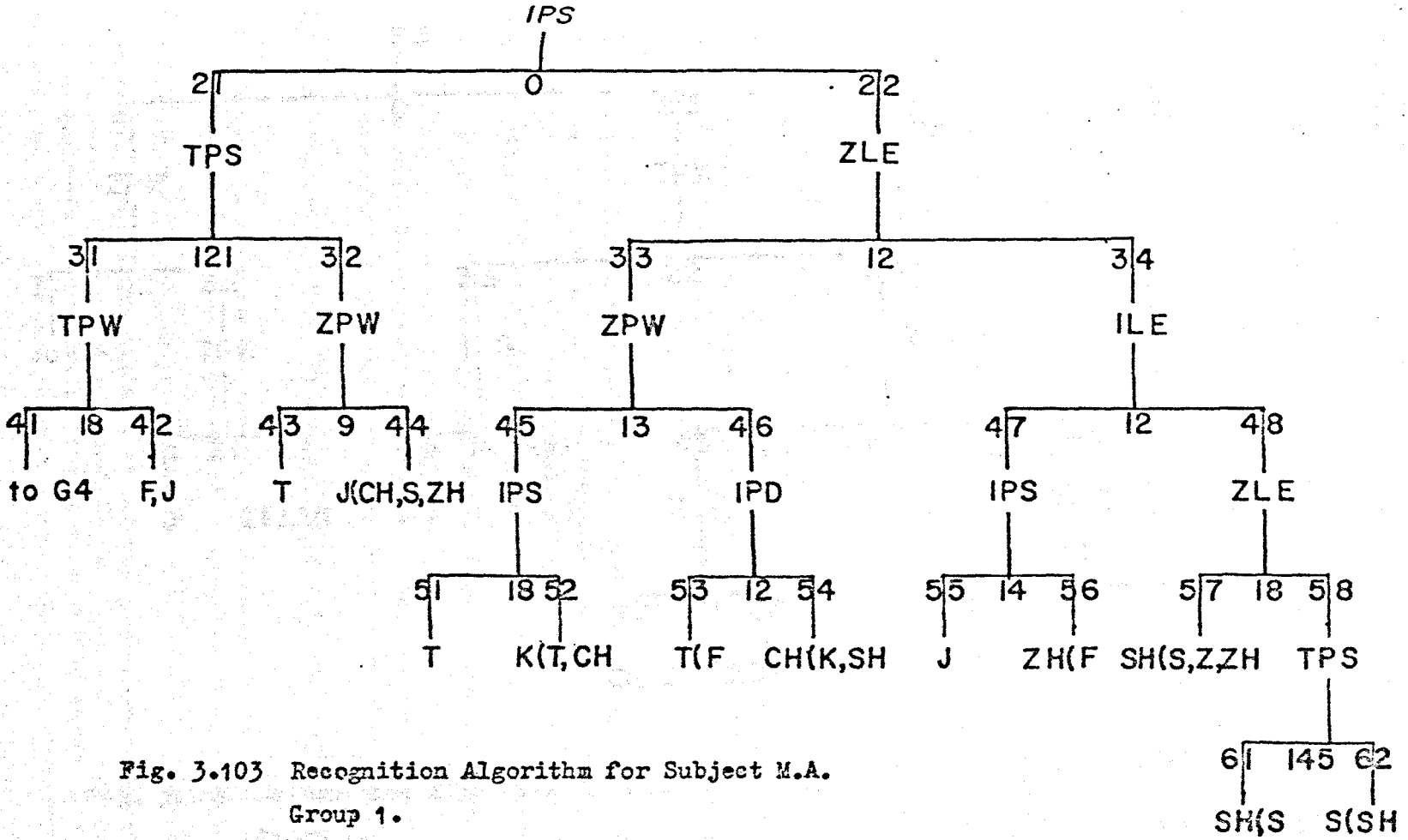


Fig. 3.103 Recognition Algorithm for Subject M.A. Group 1.

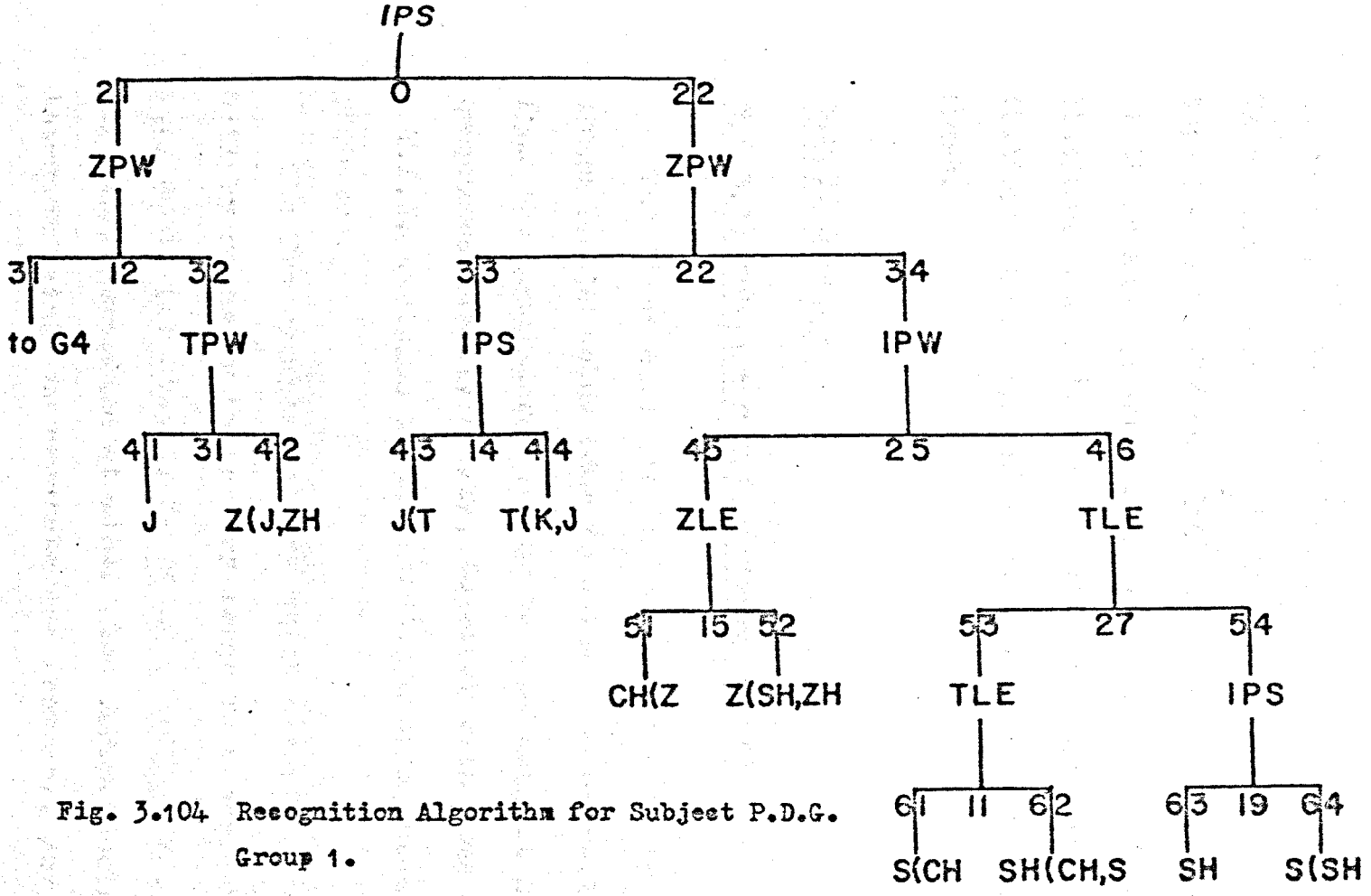


Fig. 3.104 Recognition Algorithm for Subject P.D.G. Group 1.

distinguished from /dʒ/ using a duration decision at G132. Those utterances of /dʒ/ which showed a separate I. peak, together with /t/ and /k/, were separated at G122 by means of their shorter duration. A further I.P.W. decision at G134 distinguished /tʃ/ and /z/ from /s/ and /ʃ/. /tʃ/ was then distinguished from /z/ at G145 by means of its sharper onset times.

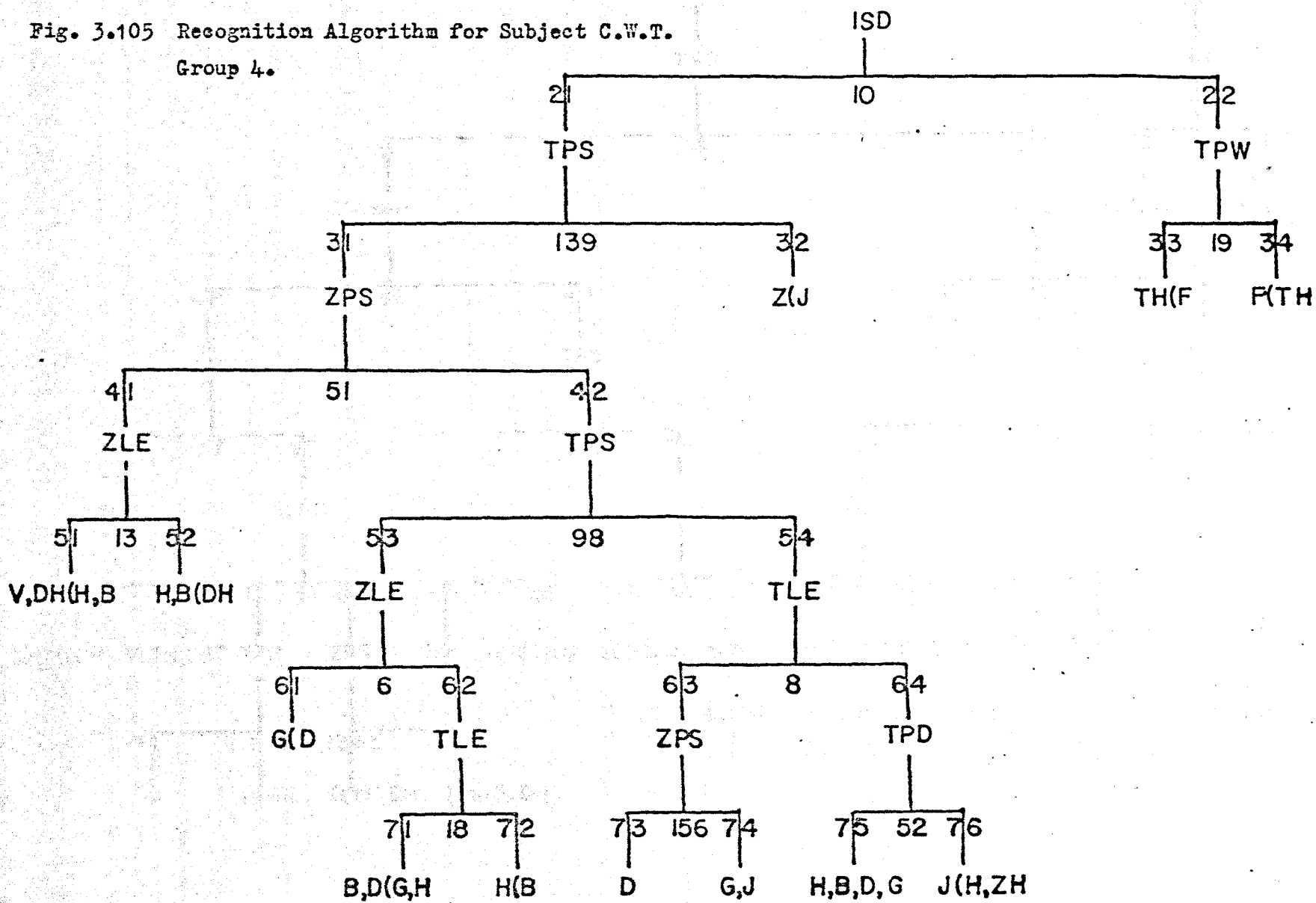
3.2.3. Group 4.

3.2.3.1. Subject C.W.T. (figure 3.105).

The first decision in G4 separated those sounds with a large I.S.D. value (due to a very low overall energy level) from the remainder. In the case of subject C.W.T., these were comprised largely of /f/, together with some examples of /θ/. A T.P.W. decision at G122 distinguished between these 2 phonemes.

The few examples of /z/ which entered G4 were removed at G121 by means of their very large T. peaks. Those sounds with very small (or absent) Z. peaks (chiefly /v/, /ð/, /h/ and /b/) were separated at G131. The T.P.S. decision at G142 was intended mainly to separate /d/ from /h/, /b/ and /g/. Most of the utterances of /g/ were then directed to G161, since these generally had very sharp Z. peaks.

Fig. 3.105 Recognition Algorithm for Subject C.W.T. Group 4.



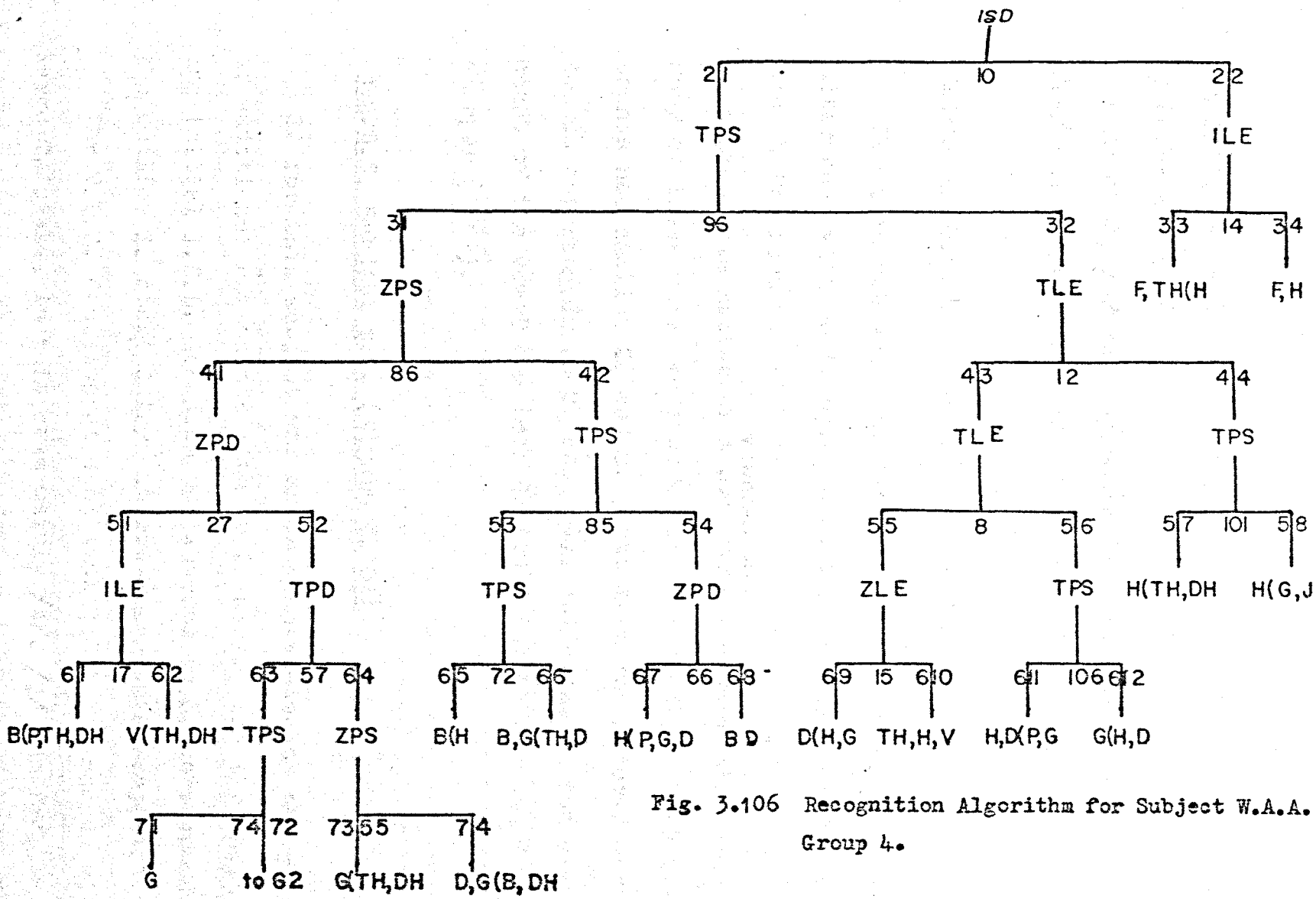


Fig. 3.106 Recognition Algorithm for Subject W.A.A. Group 4.

At G162, /h/ was separated from /b,d,and g/ by means of its greater onset times.

At G154, /d/ was separated from the remaining examples of /dʒ/ and /ʒ/ by means of its sharper T. peaks.

3.2.3.2. Subject W.A.A. (figure 3.106)

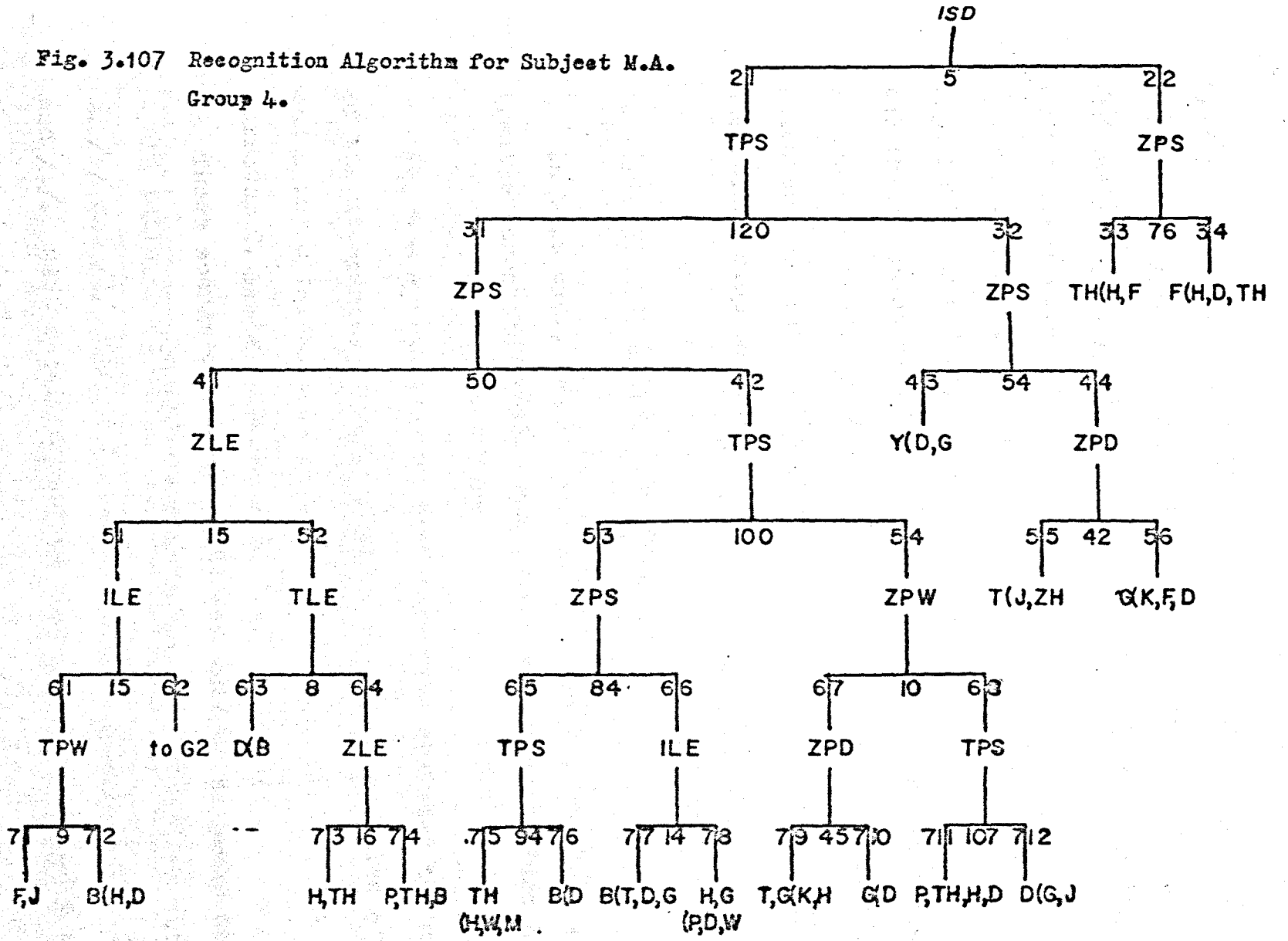
For this subject, the far greater number and variety of utterances with no I. peak meant that G4 became very complicated, though its form was similar to that for subject C.W.T.
An appreciable number of the vowel-like phonemes entered this Group, but many of these were re-directed to G2 (at G472). The vowel-like sounds had smaller Z. and T. peaks than the majority of the sounds entering G2 (G421 and 431), while the Z. cut off at the onset of the vowel was less distinct (G441) and the T. cut off sharper (G452).

3.2.3.3. Subject M.A. (figure 3.107)

G4 for this subject was again complicated by the presence of additional consonants with no I. peak. The threshold for the initial I.S.D. decision was lowered to 5, corresponding to the smaller I. start delays observed for this subject.

The T.P.S. decision at G421 isolated those sounds with large T. peaks, in this case mainly /t/, /j/ and /g/,

Fig. 3.107 Recognition Algorithm for Subject M.A. Group 4.



(/g/ for this subject often had larger T. peaks than /d/, see figure 3.29). /j/ generally had smaller Z. peaks than /t/ and /g/ (G432), while the Z. cut off for /t/ was sharper than for /g/ (G444).

The vowel-like sounds entering G4 again had small Z. and T. peaks (except /j/), while the Z. onset time was quite small and the I. onset time large (figure 3.34). Most of these sounds were re-directed to G2 via G462. The utterances of /p/ which entered this group were very difficult to separate from the voiced stop sounds.

3.2.3.4. Subject P.D.G. (figure 3.108).

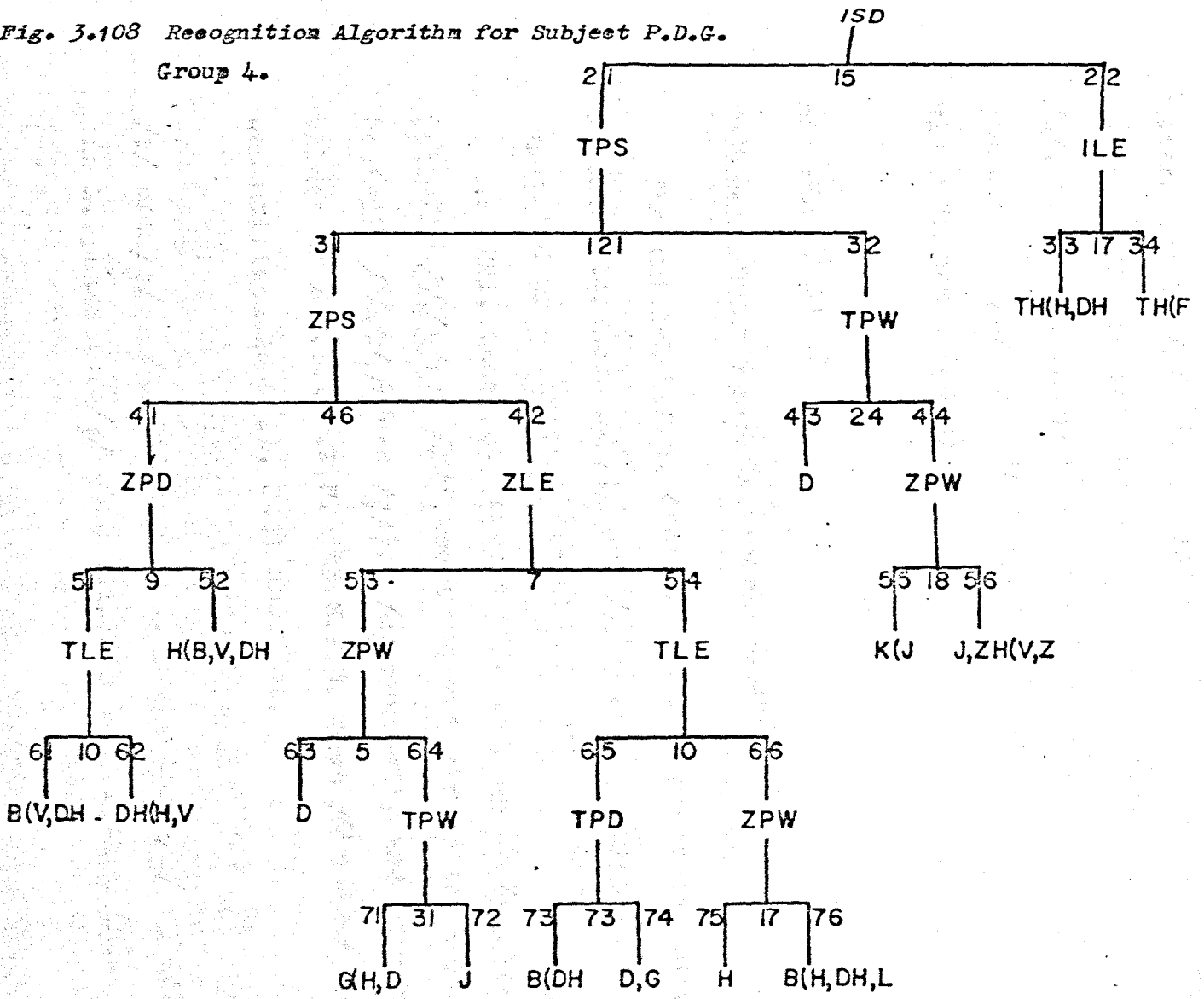
In this case, the initial I.S.D. decision was made with a greater threshold of 15. The sounds with large values of I.S.D. were mainly comprised of /θ/.

The T.P.S. decision at G421 this time isolated the majority of the utterances of /d/, together with some examples of /k/ and a few utterances of /v/, /dʒ/ and /z/. /d/ had far narrower T. peaks than the latter group (G432).

Those sounds with very small (or absent) Z. peaks were again isolated at G431. These comprised most of the utterances of /v/ and /ð/ with no I. peak, together with numbers of /h/ and /b/. G452 received most of this group which had a separate Z. peak (chiefly /h/), while the stop /b/ generally had smaller

Fig. 3.108 Recognition Algorithm for Subject P.D.G.

Group 4.



T. onset times than the fricative /ð/ (G462).

The Z.L.E. decision at G442 separated most of the utterances of /g/ from /h and b/, /g/ having the sharper Z. peaks. G454 removed about ½ of the remaining utterances of /b/, which had sharper T. peaks than /h/, and the remaining examples of /b/ were distinguished from those of /h/ by their wider Z. peaks (G466).

3.2.4. Group 3.

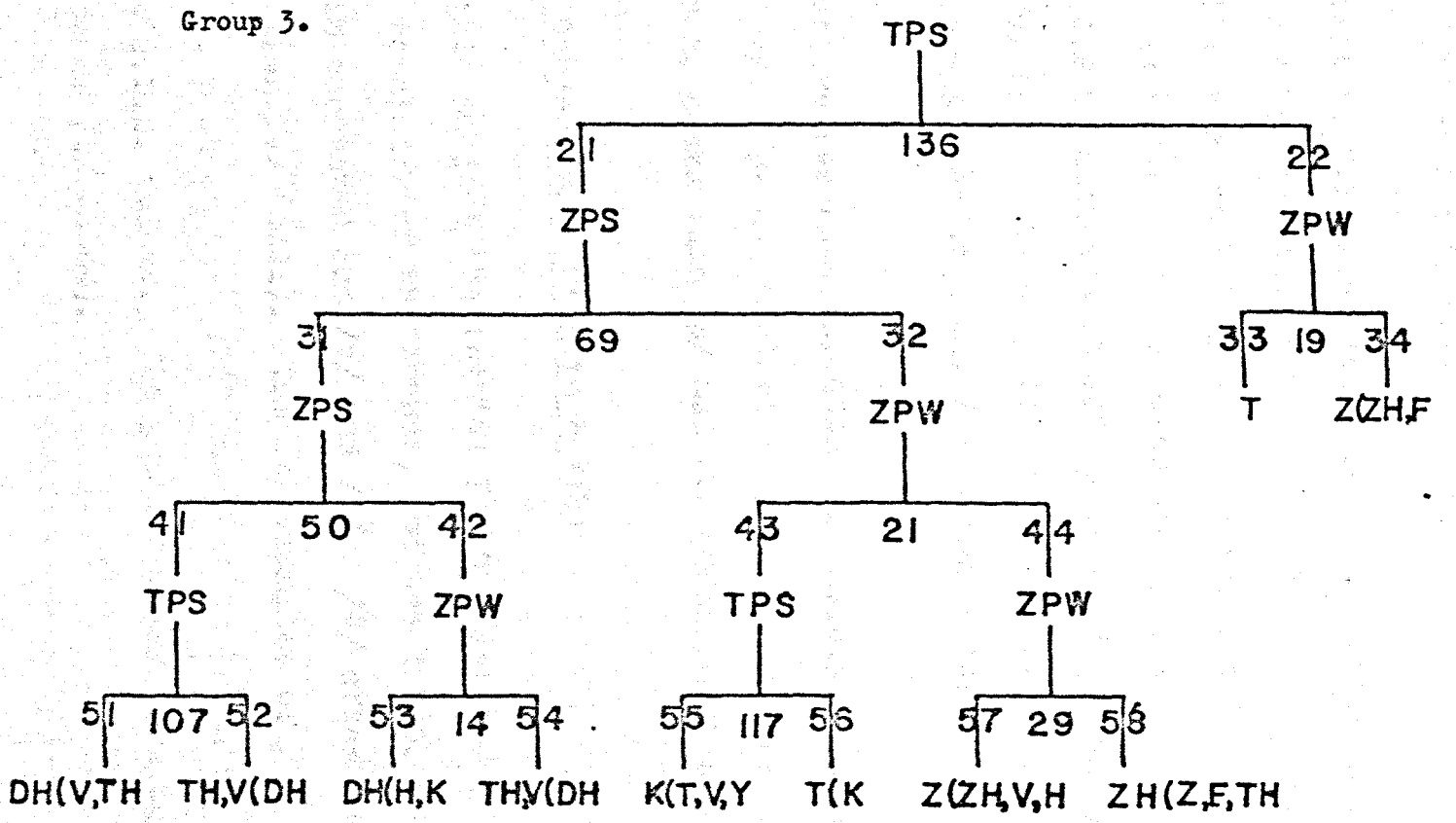
3.2.4.1. Subject C.W.T. (figure 3.109).

The initial T.P.S. decision in this group isolated most of the utterances of /t/ and /z/ present, which were then separated by a duration decision (G322).

At G221, those utterances with very small Z. peaks (chiefly /θ/,/v/and /ð/), were separated from the remainder (chiefly /t/,/k/,/z/and /ʒ/). /ð/ generally had slightly smaller values of Z.P.S., T.P.S. and Z.P.W. than /θ/ and/v/, and this was utilised in G331, G341 and G342.

A duration decision at G332 separated /t/ and /k/ from /z/ and /ʒ/. /t/ had larger T. peaks than /k/ (G343), while the Z. peaks for /ʒ/ were slightly longer than those of /z/ (G358).

Fig. 3.109 Recognition Algorithm for Subject C.W.T.
Group 3.



3.2.4.2. Subject W.A.A. (figure 3.110).

The initial Z.P.W. decision in G3 separated out those few sounds (chiefly /ʃ/) with a very long duration. At G321, those sounds with very small Z. peaks were separated from the remainder as before. These comprised chiefly /v/ and /ð/. A second Z.P.S. decision at G331 separated /v/ and /ð/ to some extent, /v/ having slightly larger Z. peaks. A further separation was made at G341, /ð/ having slightly sharper Z. peaks.

The utterances entering G332 were mainly comprised of /p/,/t/ and/k/. /p/ had smaller T. peaks than /t/, while /k/ had wider Z. peaks than /p/ and /t/ (G332, G343 and G344).

3.2.4.3. Subject M.A. (figure 3.111).

For this subject the initial T.P.S. decision isolated /z/ from the remainder of the sounds, and the Z.P.S. decision at G221 again separated out /v/ and /ð/. /ð/ often had wider Z. peaks than /v/ (G331). The few examples of /b/ with a separate I. peak entered G332 and were isolated by means of their low T.P.S. values. The peaks for /k/ were again wider than those of /t/ (G344).

3.2.4.4. Subject P.D.G. (figure 3.112).

In this case, the first decision isolated those sounds

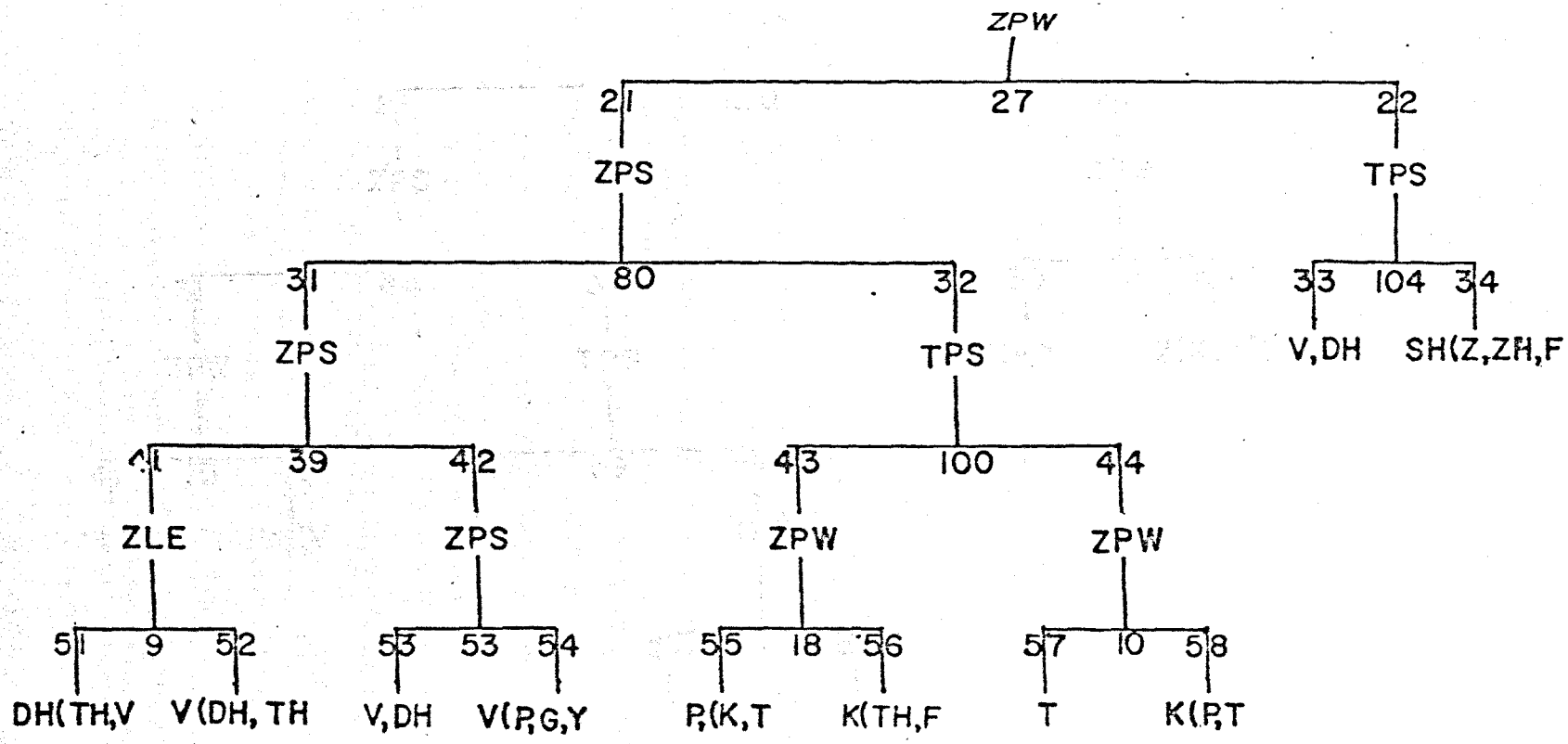


Fig. 3.110 Recognition Algorithm for Subject W.A.A. Group 3.

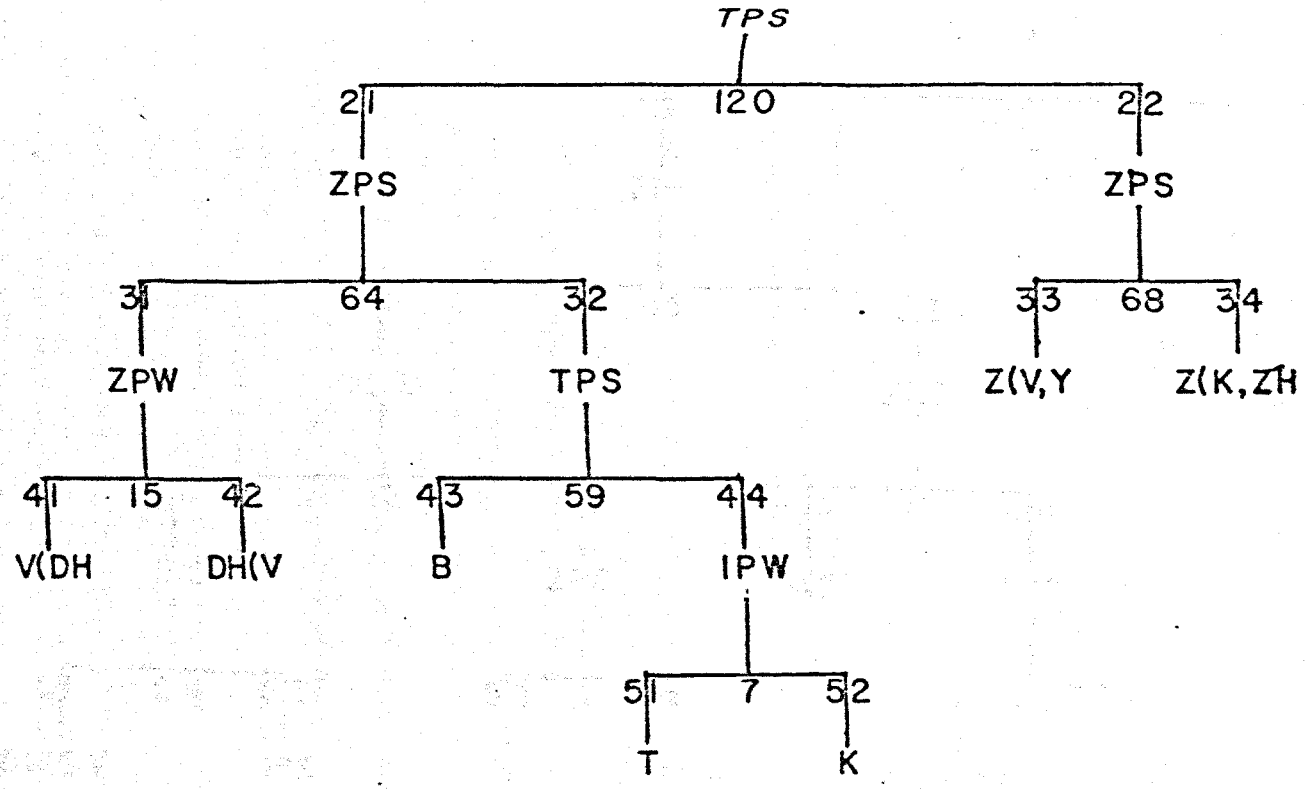


Fig. 3.111 Recognition Algorithm for Subject M.A. Group 3.

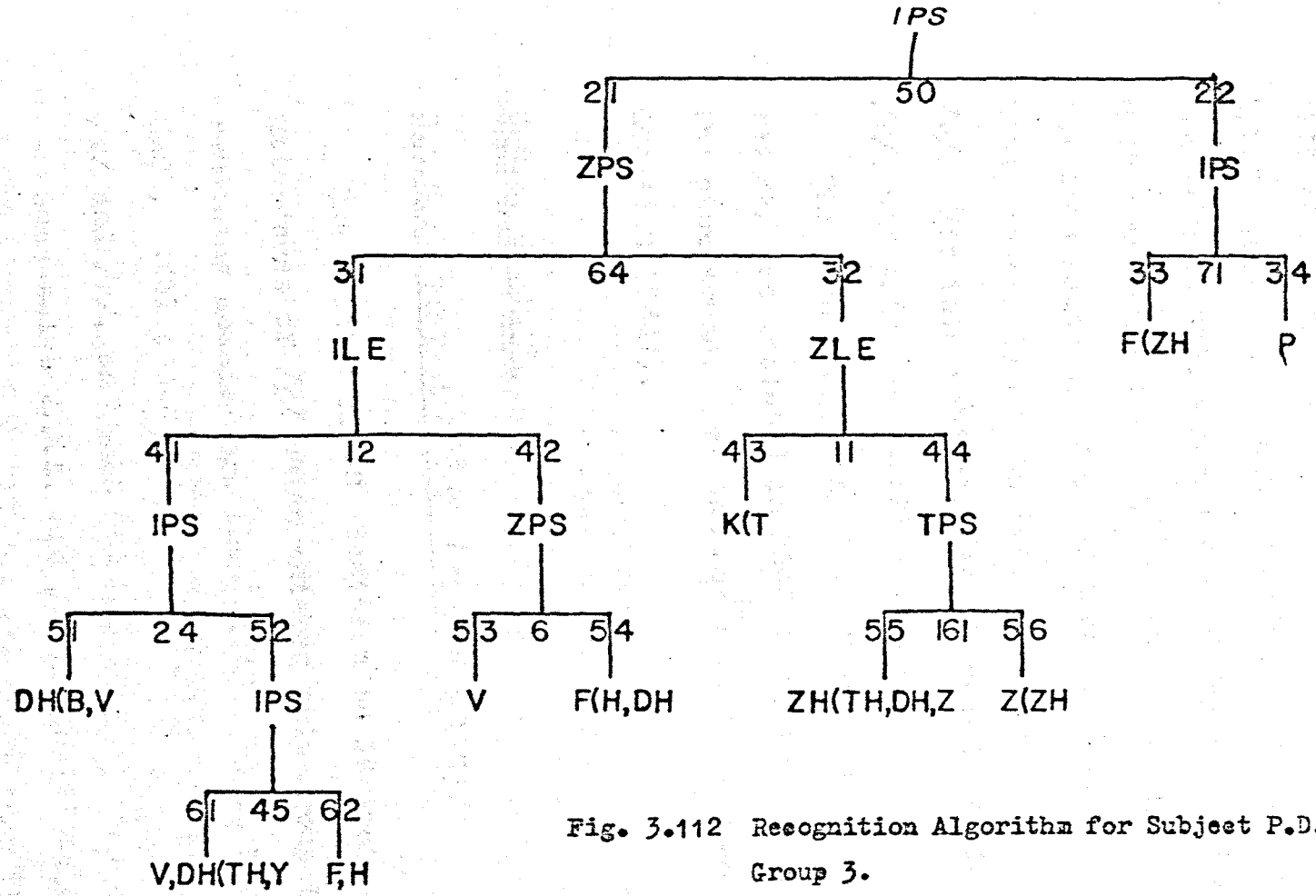


Fig. 3.112 Recognition Algorithm for Subject P.D.G. Group 3.

with very large I. peaks (chiefly /p/ and /f/). A further I.P.S. decision at G322 separated /p/ from /f/. The Z.P.S. decision at G321 again isolated /v/ and /ð/ in G331, together with some numbers of /f/, which generally had small Z. peaks for this subject. Most of these /f/ sounds were directed to G354, generally having larger values of I.L.E. than /v/ and /ð/. For this subject, /ð/ often had smaller I. peaks than /v/, and this was utilised in G341.

The sounds entering G332 were chiefly /k/, /z/ and /ʒ/, and /k/ was isolated in G343 using the sharper rise time of the stop sound. /z/ and /ʒ/ were then separated by a T.P.S. decision (G344).

3.2.5. Group 2.

3.2.5.1. Subject C.W.T. (figure 3.113).

The initial I.P.S. decision in G2 isolated most of the utterances of /p/ which entered this group in G222. The remaining sounds were then divided by an I.P.S. decision at G221, G231 containing the majority of the utterances of /j/, /r/ and /w/, and G232 the majority of /l/, /m/ and /n/.

Successive T.P.S. and Z.P.S. decisions at G231 and G241 removed those remaining sounds other than the vowel-like phonemes. Successive T.P.D. and I.P.S. decisions were then

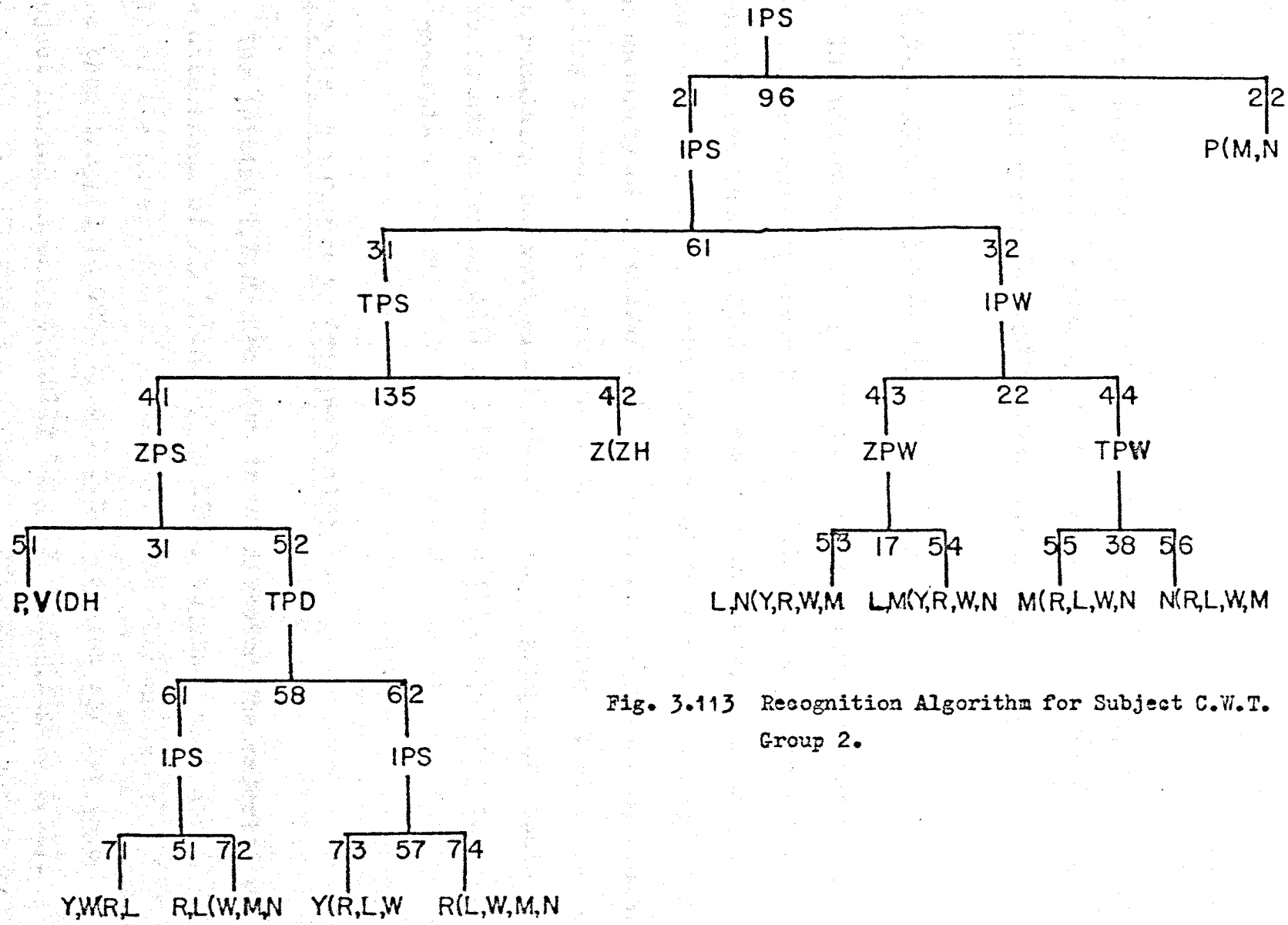


Fig. 3.113 Recognition Algorithm for Subject C.W.T. Group 2.

used to separate /j/,/r/ and/w/ to some extent (G252,G261 and G262). The sounds entering G232 could be separated to a small degree by duration decisions (G232, G243 and G244).

3.2.5.2. Subject W.A.A. (figure 3.114).

The utterances of /p/ entering G2 were again directed to G222, since most of these had very high I. peaks. An I.P.D. decision was used at G221 to separate the vowel-like sounds. Those with larger I. peaks entering G232 were largely/r/,/m/and/n/. The majority of /l/ and/w/ entered G231, with /j/ falling roughly equally between these groups. A Z.P.S. decision at G231 removed the remaining non vowel-like sounds. The 2 groups of vowel-like sounds entering G241 and G232 were then separated as much as possible.

3.2.5.3. Subject M.A. (figure 3.115).

The initial I.L.E. decision in G2 directed the majority of the utterances of /j//r/and/w/ to G222, together with about half of the utterances of /m/and/n/. The utterances of /j/ were then trapped in G234 by a T.P.S. decision /j/ for this subject having an exceptionally large T. peak (figure 3.91). A T.P.D. decision at G233 was then used to separate /r/ and/w/ from /m/and/n/ to some extent, the former having a more

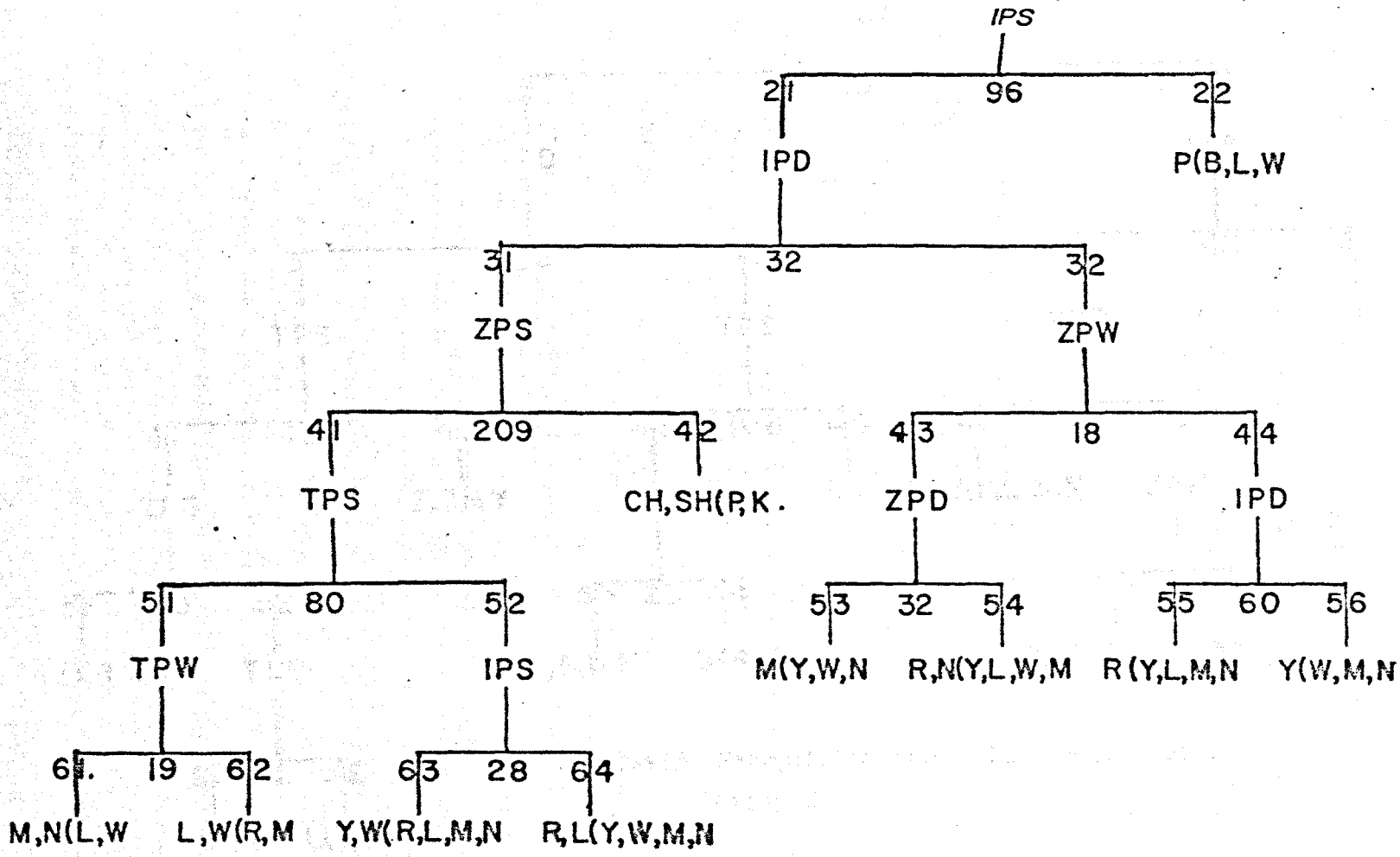


Fig. 3.114 Recognition Algorithm for Subject W.A.A. Group2.

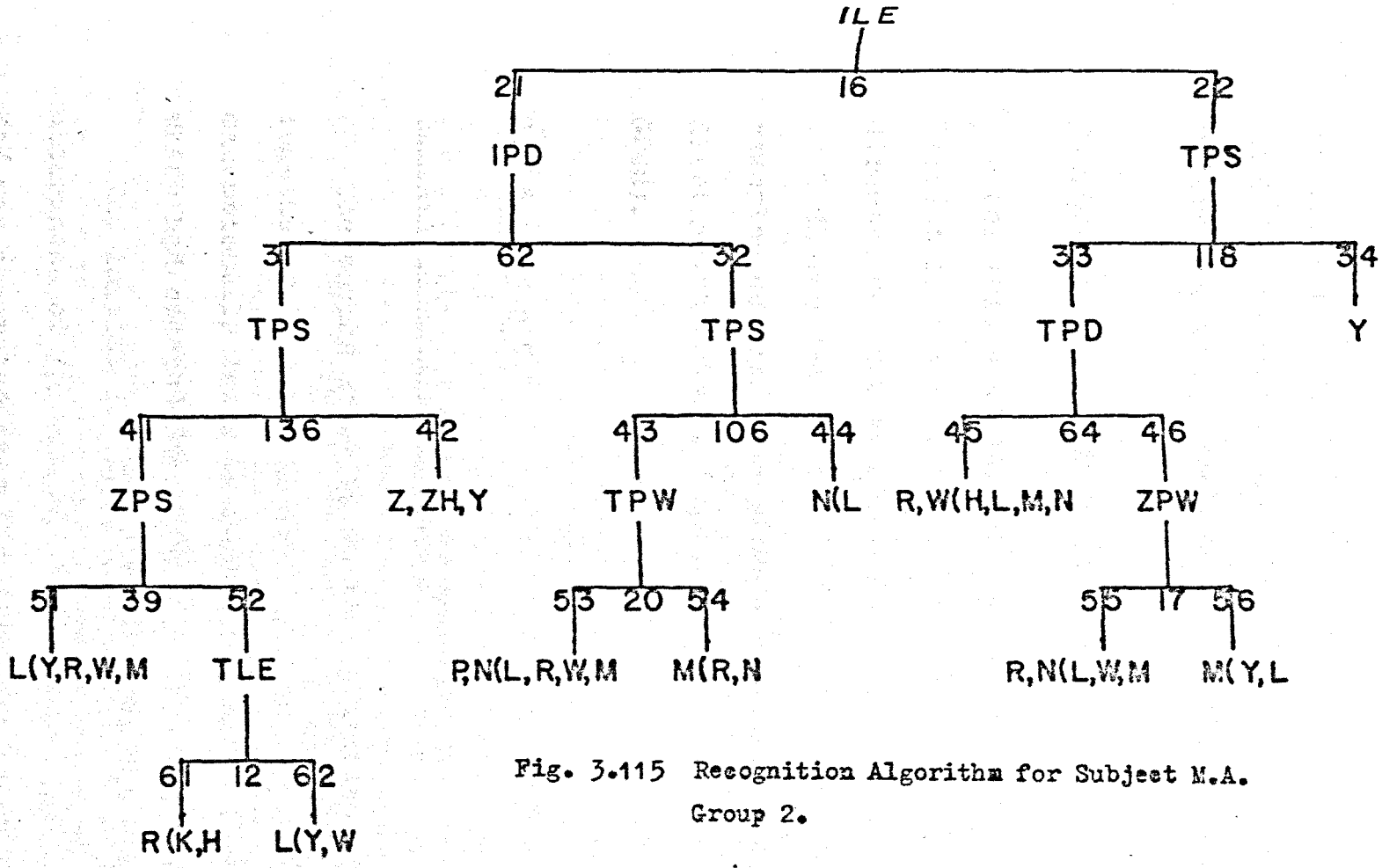


Fig. 3.115 Recognition Algorithm for Subject M.A. Group 2.

pronounced T. cut off (see figure 3.93). The I.P.D. decision at G221 directed those sounds (chiefly /m/and/n/, and the utterances of /p/ which entered G2) with higher I. peaks to G332.

3.2.5.4. Subject P.D.G. (figure 3.116) .

For this subject, /m/ and /n/ again normally had the largest I. peaks of the sounds entering G2, and these were directed to G222 by an I.P.S. decision. A T.P.S. decision at G221 removed the non vowel-like sounds together with a few examples of /j/. The parameter T.P.D. was again used to distinguish between /j/,/ɥ/,/l/ and/w/ to some extent (G231 and G242).

3.2.6. Group 5.

3.2.6.1. Subject C.W.T. (figure 3.117).

The sounds entering G5 were comprised largely of /p/, together with a few examples of /f/θ/, /v/and /ʒ/ with exceptionally large I. peaks, and some vowel-like sounds which had escaped G2.

The initial I.P.D. decision directed most of the utterances of /p/ and the vowel-like sounds to G522, where /p/ was separated from the remainder by an I.P.S. decision. At G521, /f/and/θ/ were distinguished from the remaining sounds by their large I L.E. values.

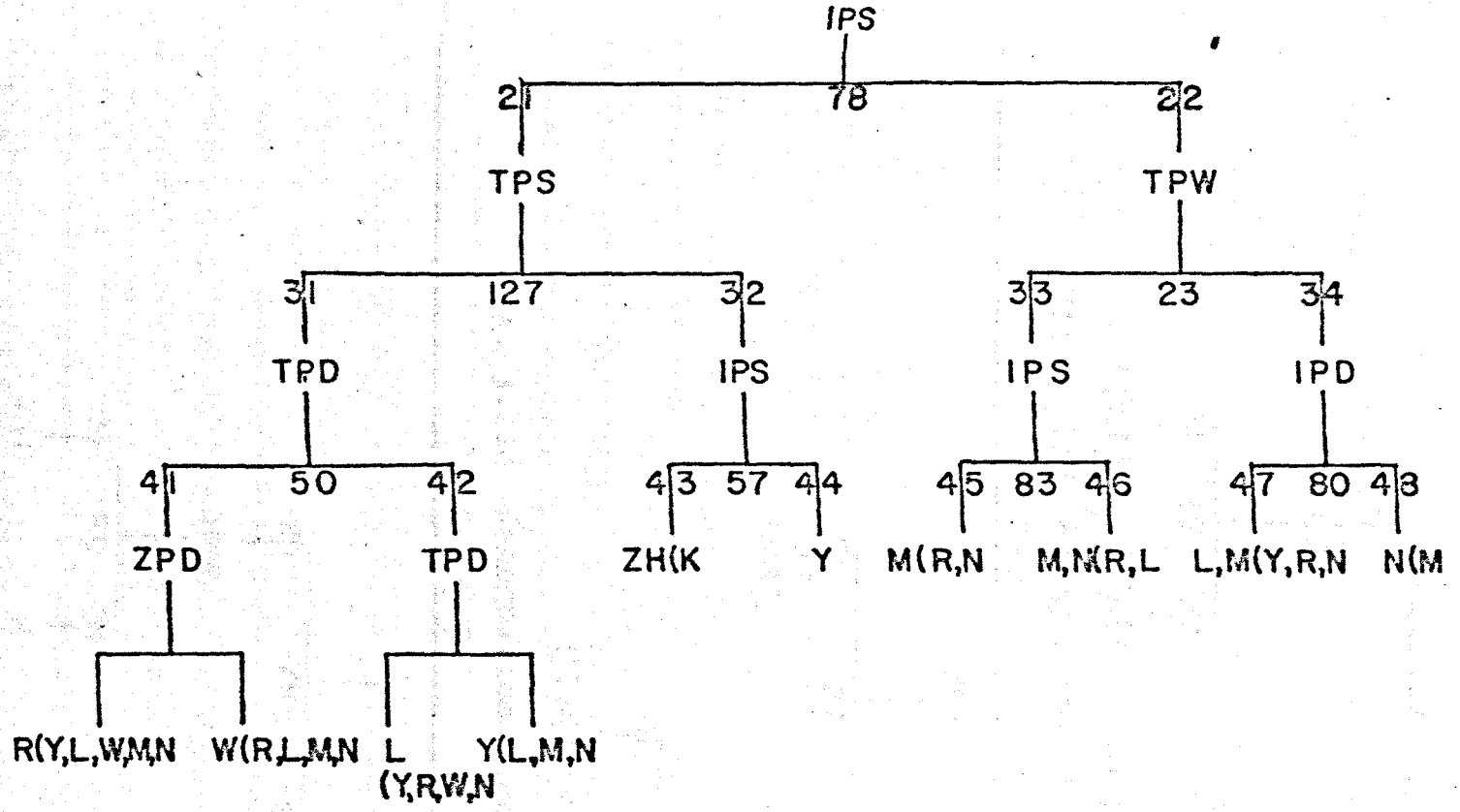


Fig. 3.116 Recognition Algorithm for Subject P.D.G. Group 2.

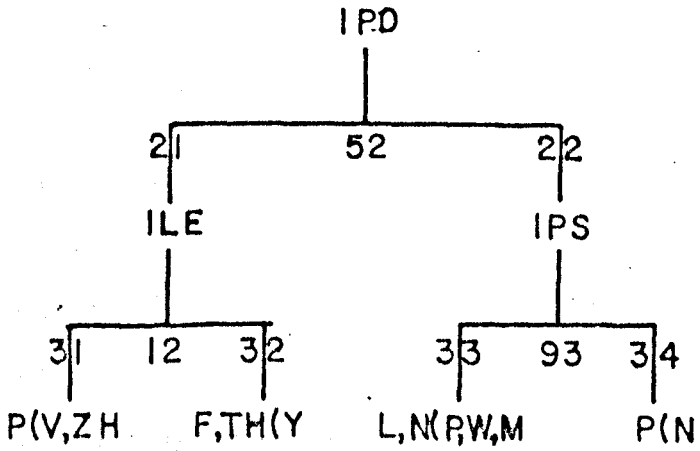


Fig. 3.117 Recognition Algorithm for Subject C.W.T. Group 5.

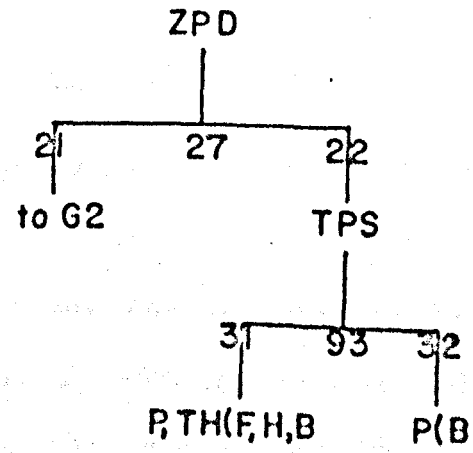


Fig. 3.118 Recognition Algorithm for Subject M.A. Group 5

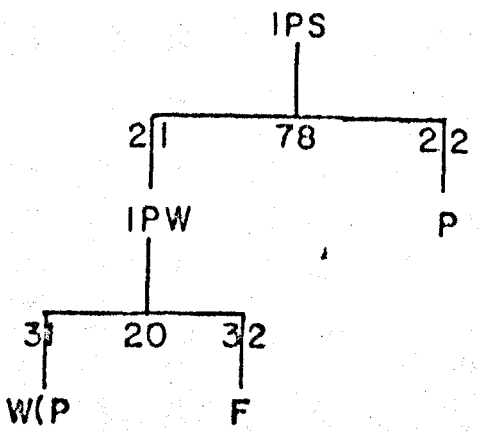


Fig. 3.119 Recognition Algorithm for Subject P.D.G. Group 5.

3.2.6.2. Subject M.A. (figure 3.118).

The vowel-like sounds entering G5 were redirected to G2 by the initial Z.P.D. decision. These sounds generally had very small or absent Z. peaks, and consequently low values of Z.P.D. Some distinction between /p/ and the remaining sounds could be made using the parameter T.P.S. (G522), most of the utterances of /p/ terminating at G532.

3.2.6.3. Subject P.D.G. (figure 3.119).

Most of the utterances of /p/ arrived in G5, together with a few examples of /f/ and the glide /w/. /p/ was isolated in G522 by an I.P.S. decision, the I. peaks for /p/ being very large for this subject. /f/and/w/ were then separated by an I.P.W. decision (G521), the duration of /f/ being much greater than that of /w/.

3.3 Performance of the Recognition Algorithms.

3.31 Overall Recognition Rates.

Figure 3.120 shows the percentage of correct identifications achieved for each subject using the appropriate algorithms. These figures cover all the available data for a single subject. Since this was a 'forced choice' recognition situation, the recognition rates can be given as a percentage of the total number of examples presented.

The first two columns of figure 3.120 show the proportions of 'probably' and 'possibly' correct identifications.(ie. an instance of /t/ arriving at an end point labelled T(K would count as a probable identification, while one of /k/ would be a possible identification). The third column gives the total of probable and possible successes. The figure for the recognition rate of human listeners given in the fourth column is derived from listening tests using the C.V. sounds recorded by C.W.T. Details of these listening tests will be found in Appendix 3. In the listening tests, no distinction between probable and possible identifications was made.

As figure 3.120 shows, the performances of the four recognition algorithms for their respective subjects were very similar. The total recognition rate of about 89% was only a little smaller than that for human listeners,

	Prob.	Poss.	Total	Human Subjects
C.W.T.	64	25.6	89.6	93.6
W.A.A.	62.5	25	87.5	
M.A.	61.4	27.2	88.6	
P.D.G.	66.8	23.7	90.5	
Average	63.7	25.1	89.05	

Fig. 3.120 Recognition Rates for the Four Algorithms With the Appropriate Speakers. (%)

though the proportion of probable successes (about 64%), which corresponds to a forced single phoneme decision, was considerably below this figure.

Both the total recognition rates and the ratio of probable to possible identifications were slightly higher for subjects C.W.T. and P.D.G. than for W.A.A. and M.A. This corresponds to the observation that the visual distinctions between phonemes on the Z.T.I. diagram were clearer for the former speakers, and to the greater simplicity of their recognition algorithms.

3.32 Effect of the Different Vowels.

The variations in the consonant recognition rates for the 10 vowel phonemes used are illustrated for the four subjects in figures 3.121 to 3.124. The rates are plotted together with their Standard Deviations; the solid lines refer to the total recognition rates (probable + possible) and the dotted lines to the percentage of probable identifications. As these diagrams show, the consonant recognition rates were to all intents and purposes independant of the following vowel. The only systematic difference observed was that the scores for /u/ (OO) were generally below the average. This was attributed to the very low values of F1 and F2 for /u/ and its exceptionally long duration.

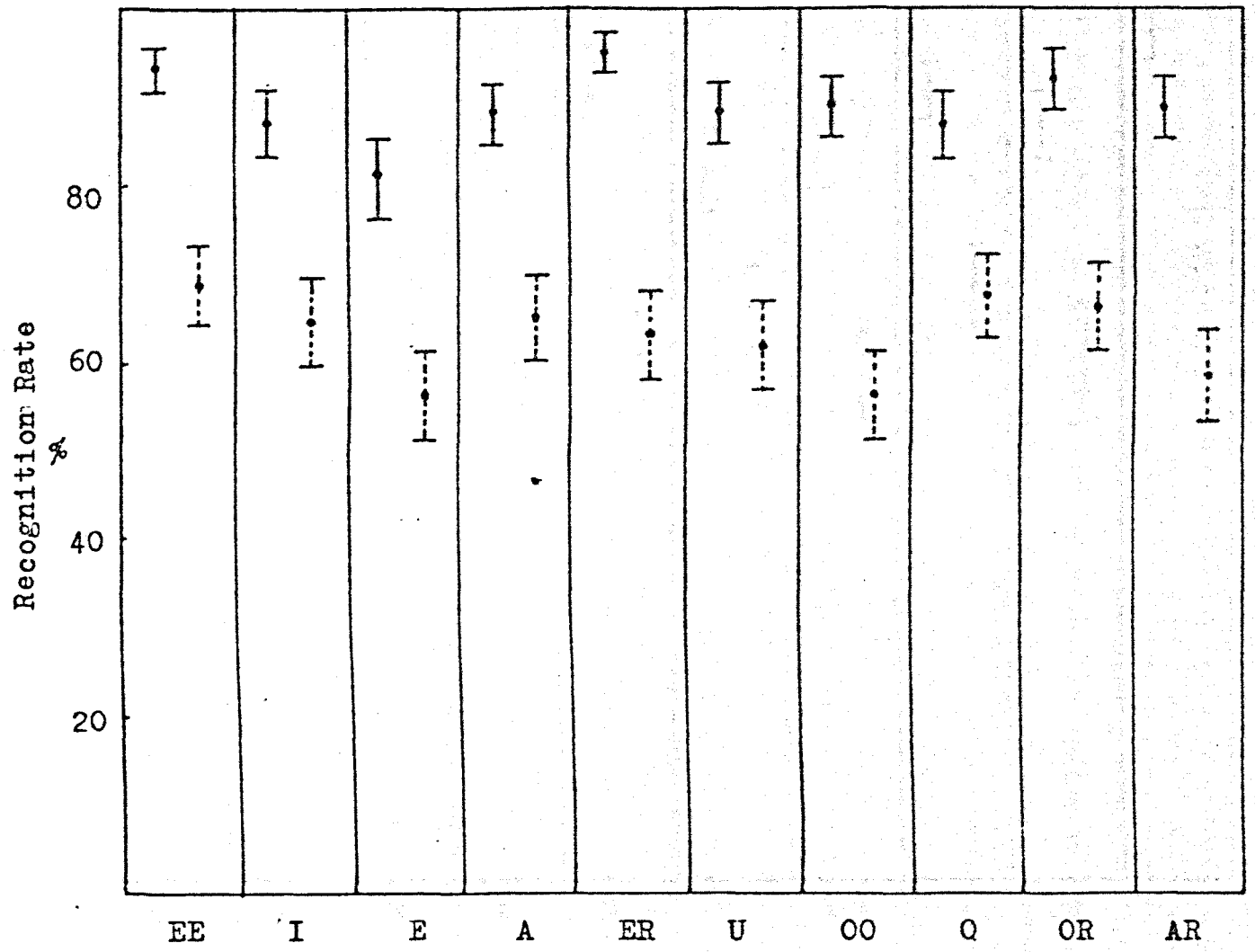


Fig. 3.121 Consonant Recognition Rates for Each Vowel: Subject C.W.T.

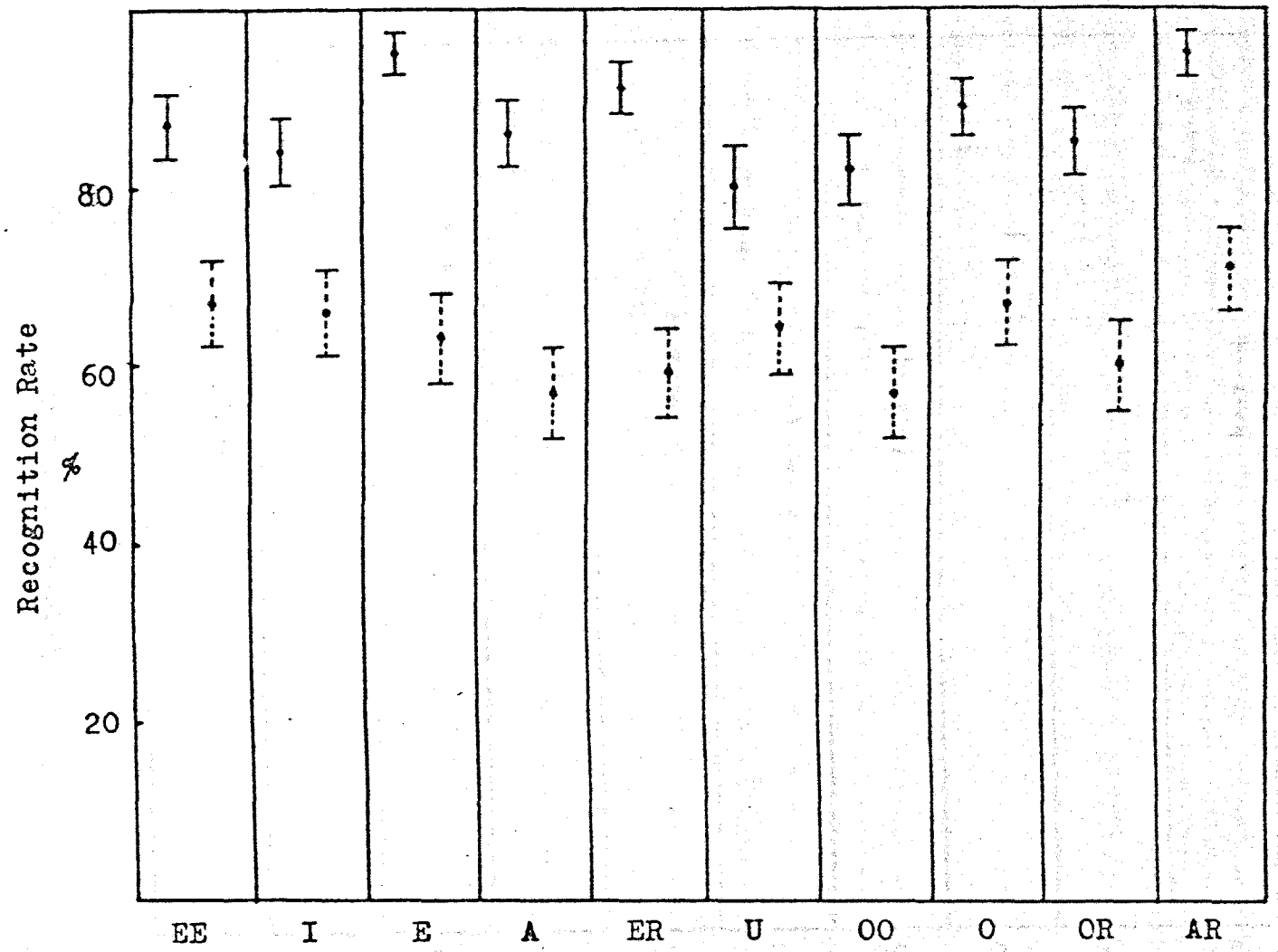


Fig. 3.122 Consonant Recognition Rates for Each Vowel: Subject W.A.A.

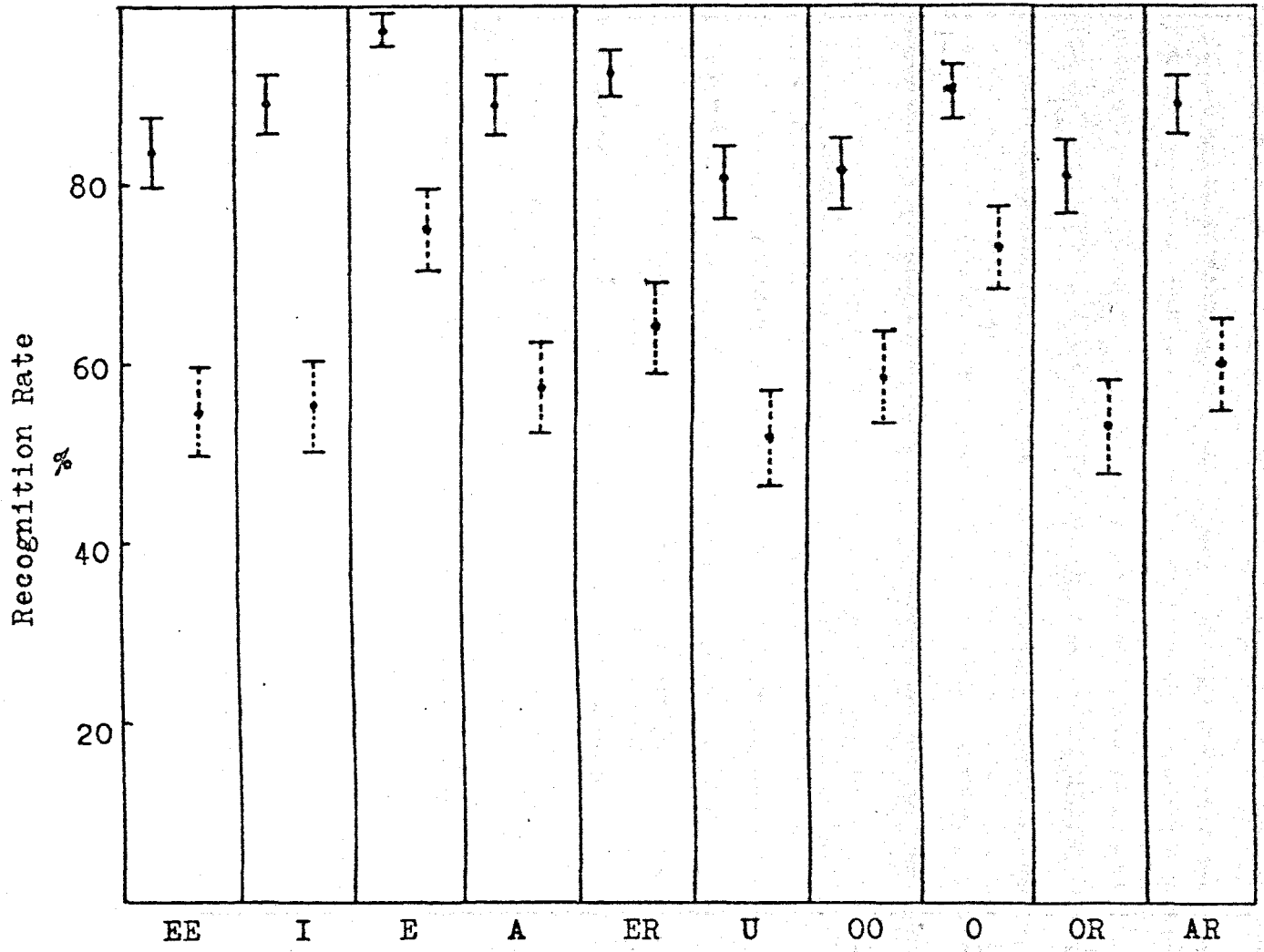


Fig. 3.123 Consonant Recognition Rates for Each Vowel: Subject M.A.

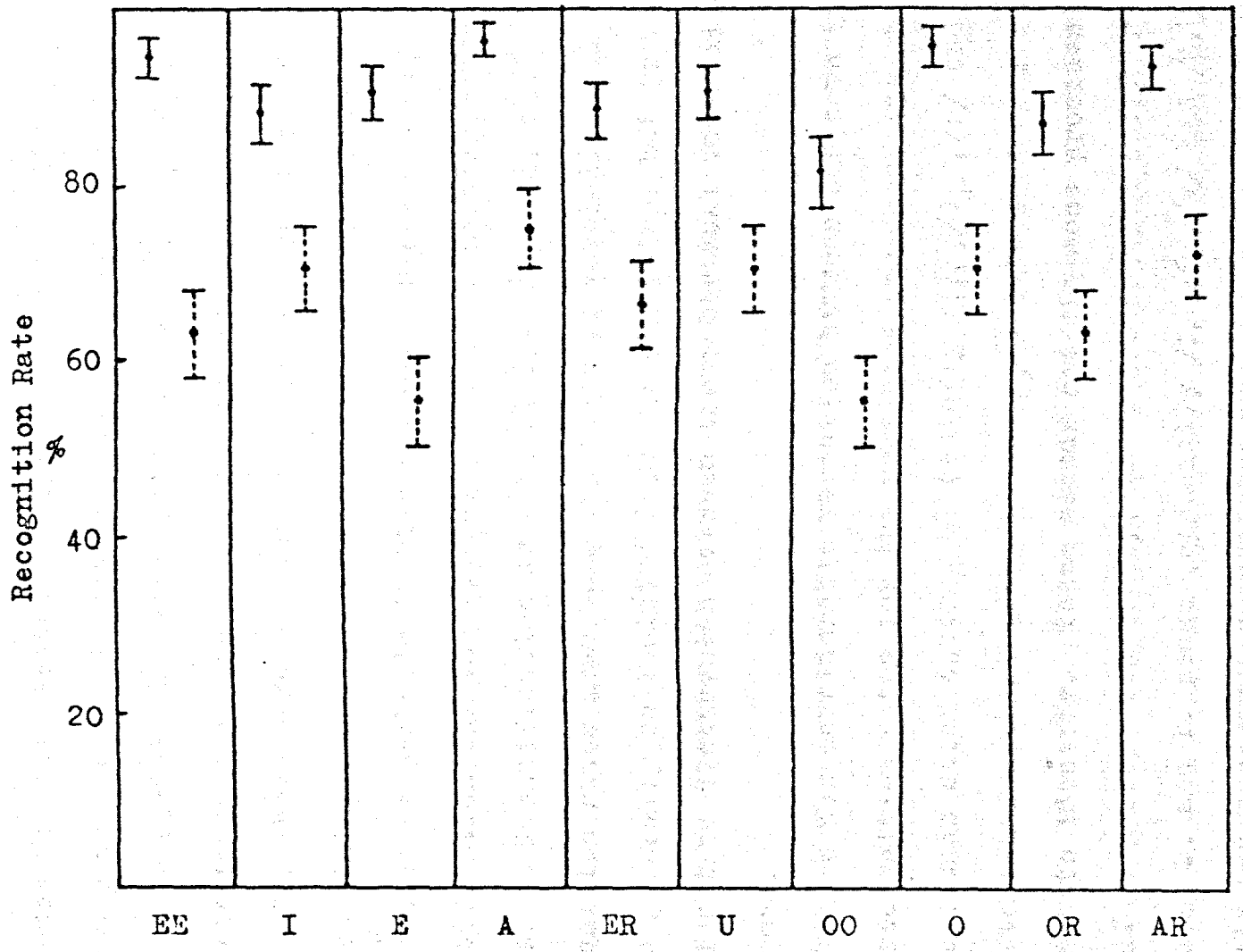


Fig. 3.124 Consonant Recognition Rates for Each Vowel: Subject P.D.G.

3.33 Recognition Rates for the Individual Consonant Phonemes.

The recognition rates for the individual consonant phonemes are illustrated in figures 3.125 to 3.128 in the same way as the vowel rates of figures 3.121 to 3.124.

As expected, the vowel-like sounds /j/,/r/,/l/,/w/,/m/ and /n/ fared the worst. Though the total recognition rates for these phonemes were no smaller than those of the remaining sounds, they generally had a much lower proportion of probable identifications. Since a possible identification for the vowel-like sounds generally corresponded to identification as a vowel-like phoneme, the recognition rates mean that the class of vowel-like sounds could generally be isolated from the remainder, but it was difficult to distinguish between the individual vowel-like phonemes.

There was considerable variation between subjects in the recognition rates for the remaining phonemes, though those sounds which entered G1 (chiefly /tʃ/,/dʒ/,/s/ and /ʃ/, together with /t/,/k/,/z/ and /ʒ/) were generally the easiest to identify. These sounds had the most prominent Z. and T. peaks. Conversely, those sounds with small or variable Z. and T. peaks (especially /h/,/θ/,/v/ and /ð/) usually had the worst recognition rates. In general, unvoiced sounds were easier to identify than voiced sounds.

For subjects W.A.A. and M.A., the increased difficulty

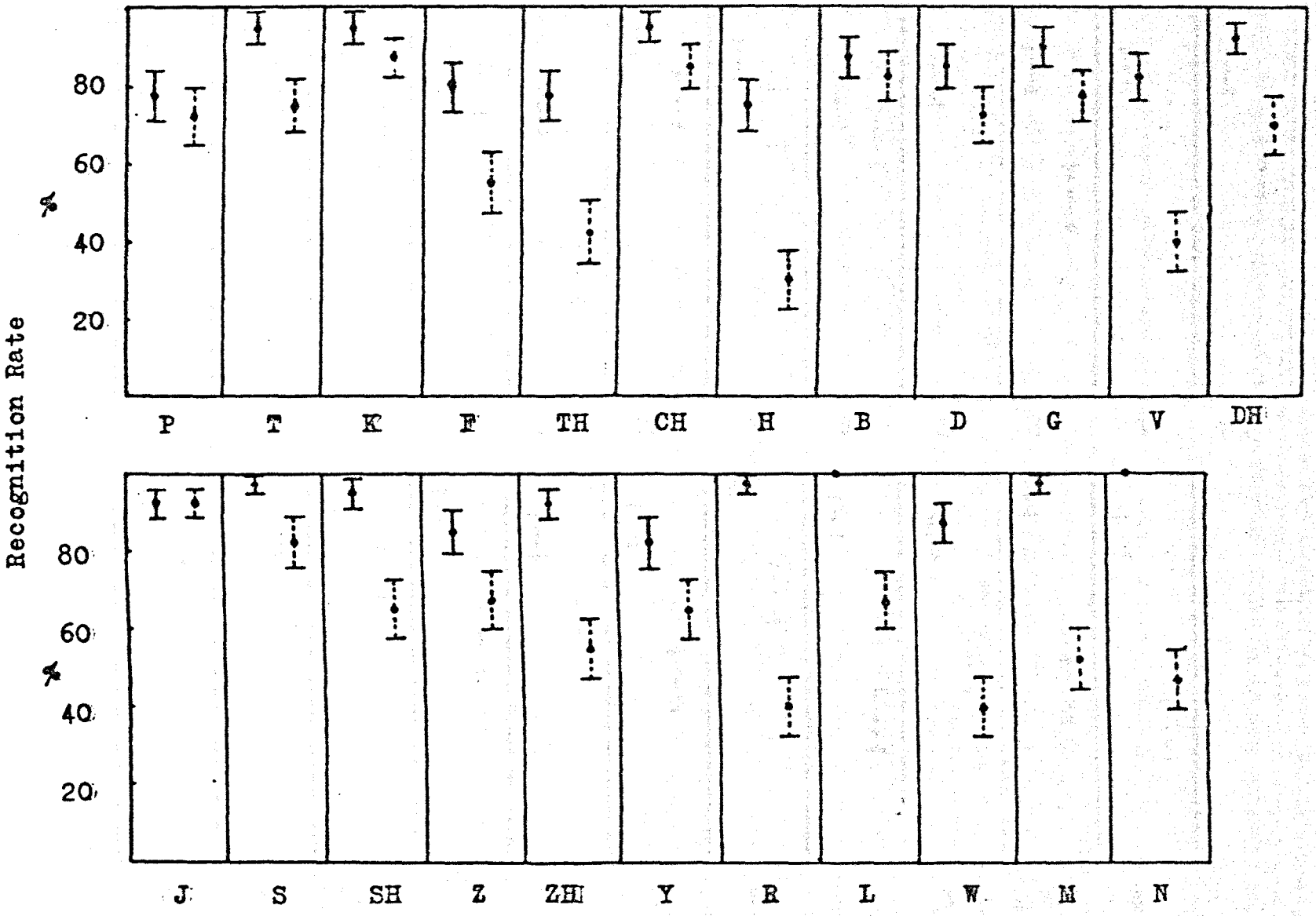


Fig. 3.125 Recognition Rates for Each Consonant Phoneme: Subject C.W.T.

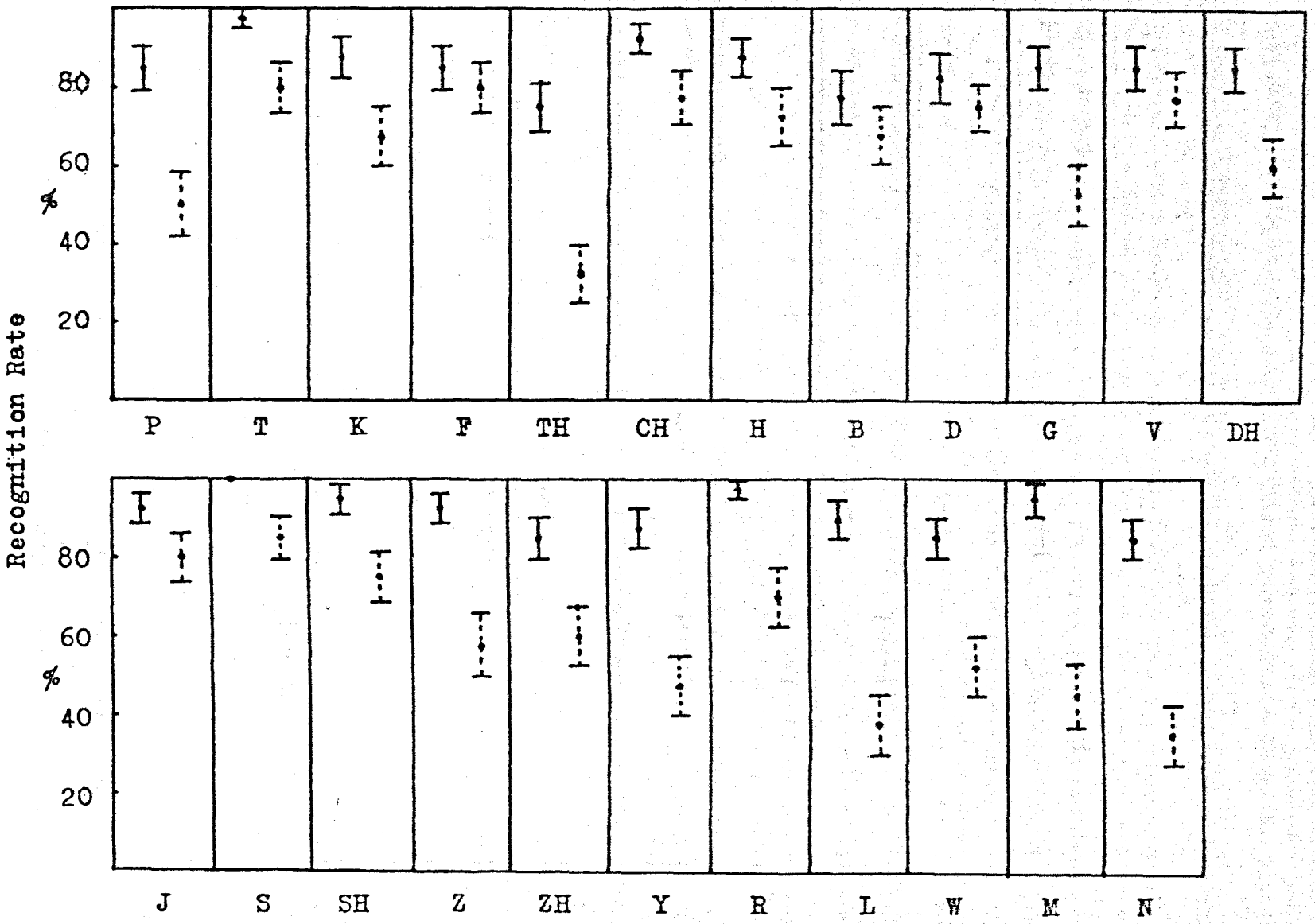


Fig. 3.126 Recognition Rates for Each Consonant Phoneme: Subject W.A.A.

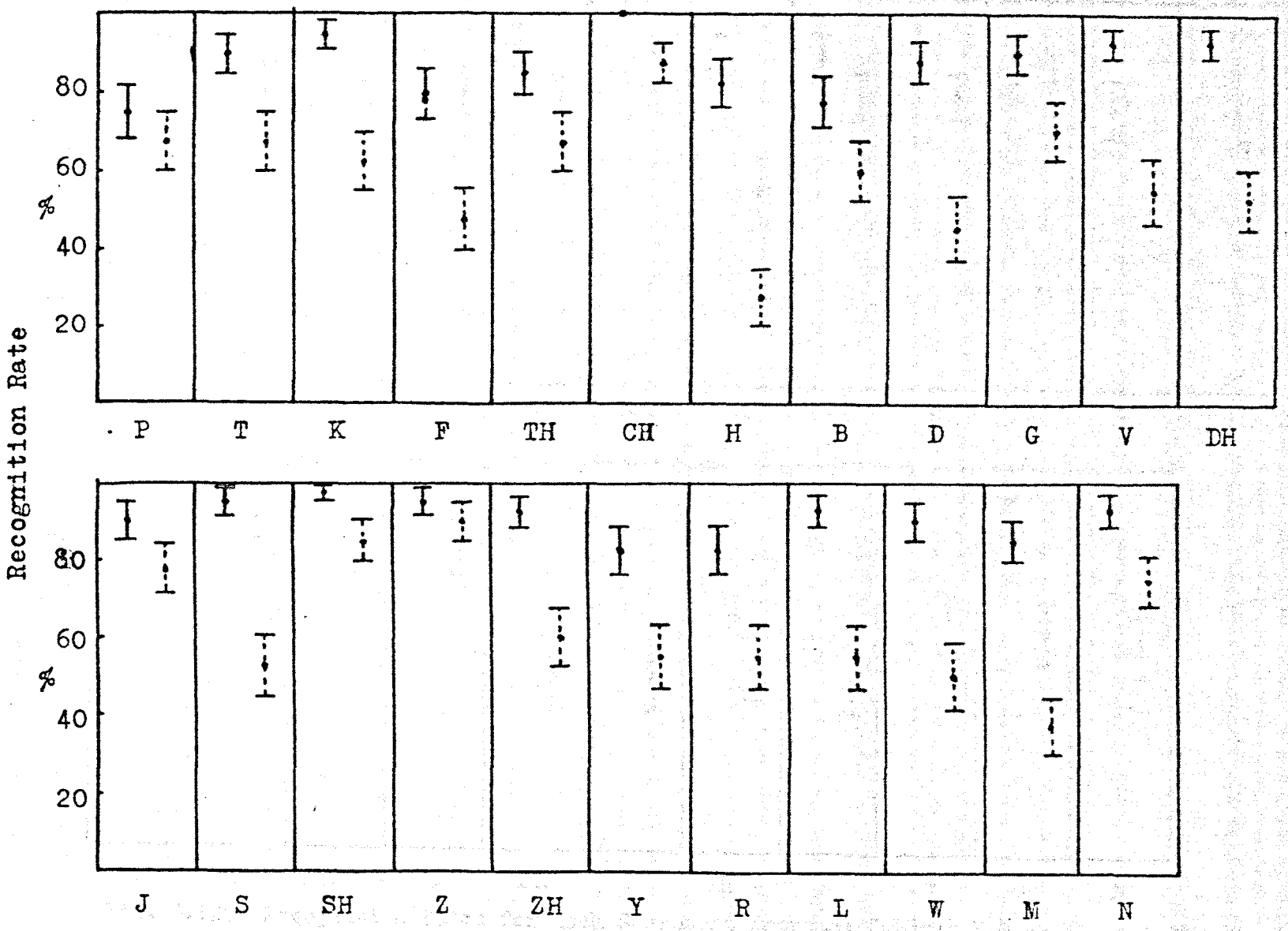


Fig. 3.127 Recognition Rates for Each Consonant Phoneme: Subject M.A.

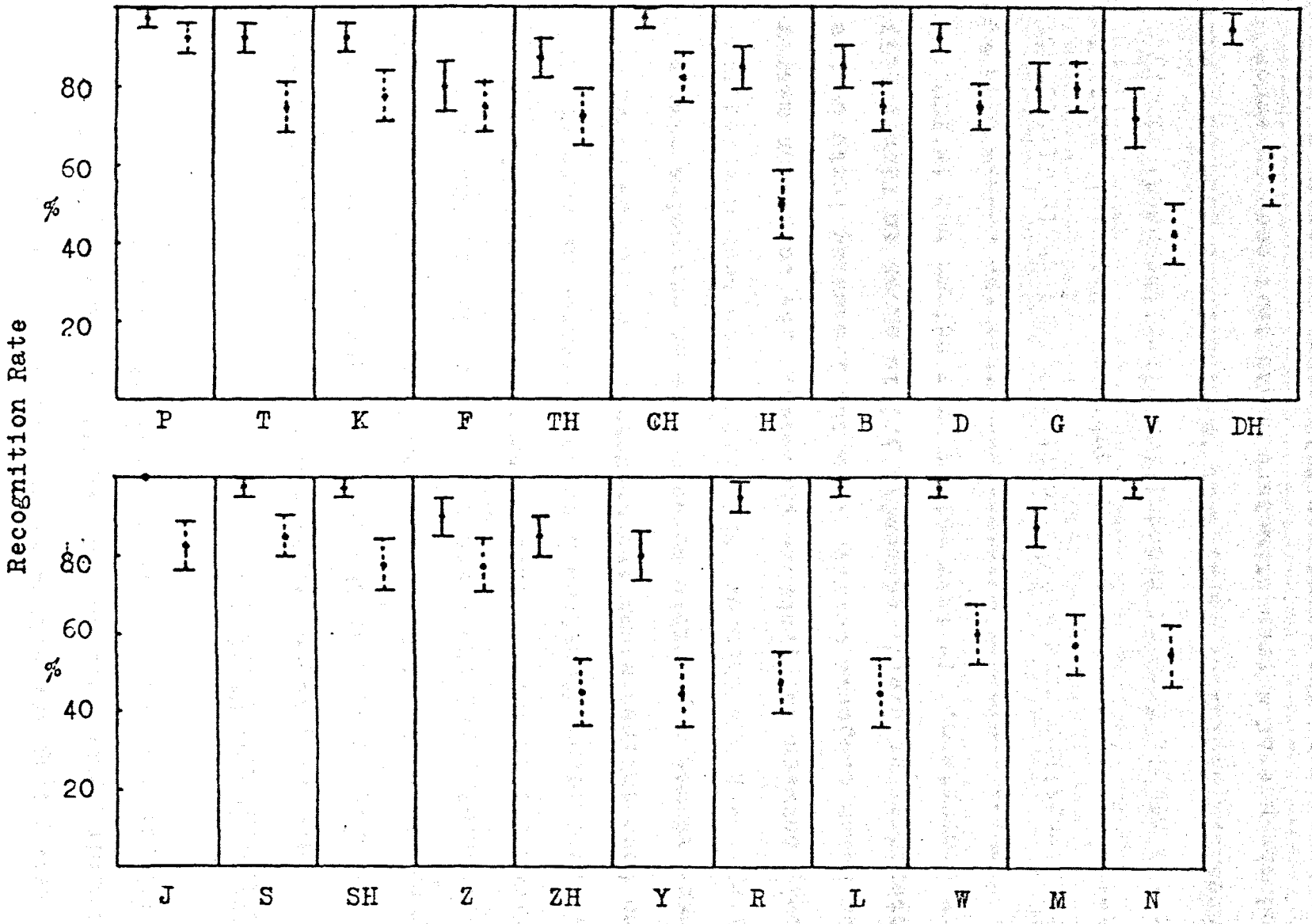


Fig. 3.128 Recognition Rates for Each Consonant Phoneme: Subject P.D.G.

in separating the voiced stops /b/,/d/ and /g/ from the remaining phonemes, and in distinguishing between these sounds, is reflected in lower recognition rates.

3.34 Confusion Matrices.

Figures 3.129 to 3.132 are confusion matrices for the utterances of each of the four subjects processed by the appropriate algorithms. The sounds presented are arranged horizontally and the responses vertically, so that each column sum is 40. In these matrices, both probable and possible identifications have been treated as correct, and are thus entered along the diagonal. In the few cases where it was impossible to associate an end point with a single probable phoneme, the errors have been divided randomly between the probable phonemes. The confusion matrix for human subjects derived from the listening tests on the utterances of C.W.T. (Appendix 3.) is shown in figure 3.133 for comparison. In this matrix, the column sum is 100.

These confusion matrices show that the errors made by the algorithms were widely scattered, and their distribution varied widely from subject to subject. This was partly due to the simple form of recognition algorithm used: binary threshold decisions imply that 'a miss is as good as a mile', and the use of a tree structure means that once an error has been made, the sound may fall anywhere within a large

	Presented																					
	P	*T*	*K*	*F*	*TH*	*CH*	*H*	*B*	*D*	*G*	*V*	*DH*	*J*	*S*	*SH*	*Z*	*ZH*	*Y*	*R*	*L*	*W*	*M*	*N*
P	31						1																
T		38			1	1					1	1				1		1					
K			38	1	2	1											1						
F				32			2											1					
TH		1		3	31		1		1							2	2		1				
CH			1			38							1			1							
H							30		3		1		1										
B	1							35			1												
D									34	2			1										
G		1		1			3	3		36													
V											33												
DH	2							1	1			37						2			2		
J				1			2	1	1	2			37	1		1							
S														39									
SH				1											38								
Z	1		1	1	2						1	1				34		2				1	
ZH							1								2		37	1			1		
Y	4				1						3	1				1		33					
R																			39				
L	1				3															40			
W																					35		
M																						39	
N																							40

Fig. 3.129 Confusion Matrix for Subject C.W.T.

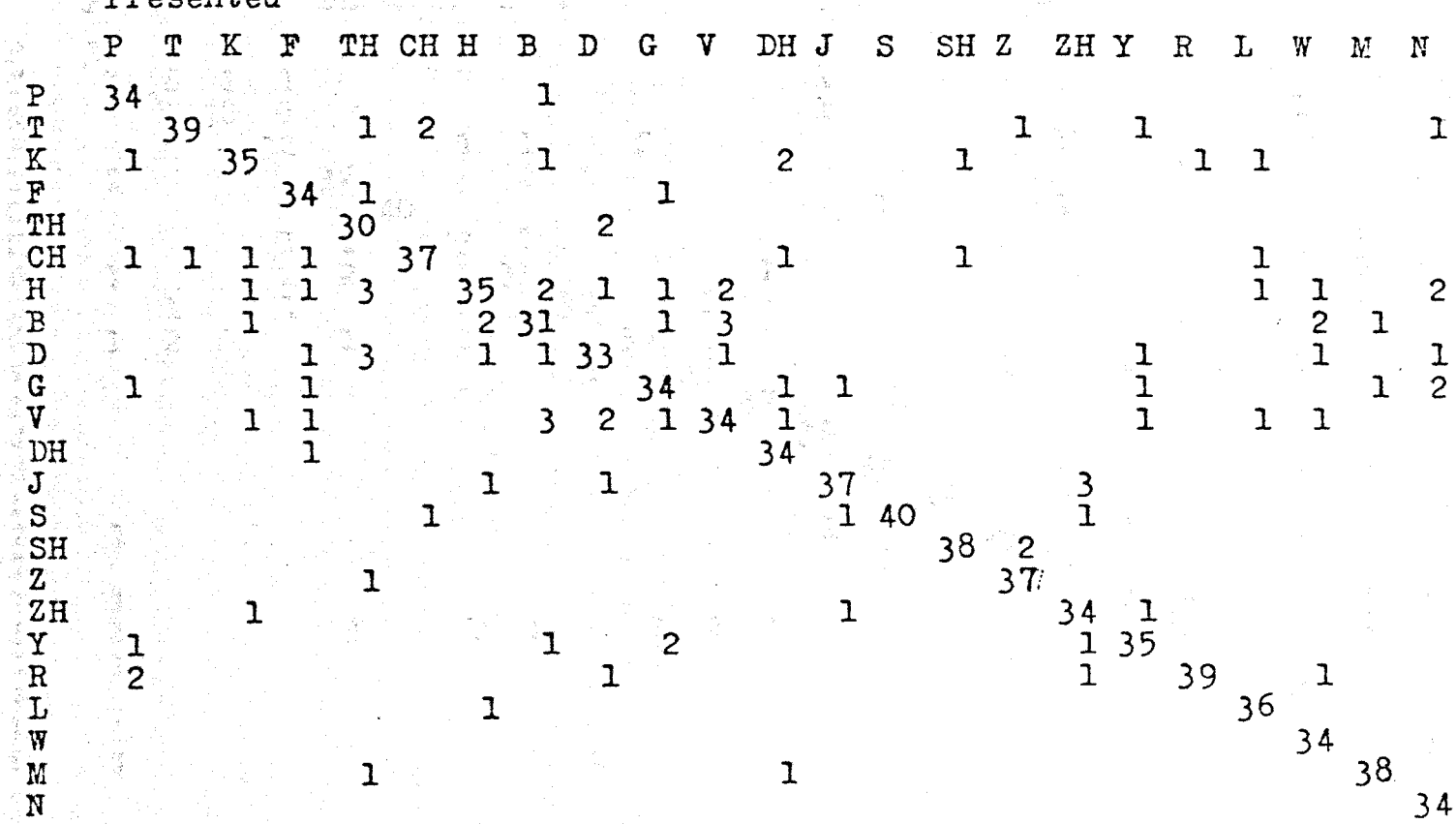


Fig. 3.130 Confusion Matrix for Subject W.A.A.

	Presented																					
	P	*T*	*K*	*F*	*TH*	*CH*	*H*	*B*	*D*	*G*	*V*	*DH*	*J*	*S*	*SH*	*Z*	*ZH*	*Y*	*R*	*L*	*W*	*M*	*N*
P	30			2	1		1	2	2			1	1						1	1	1		
T	1	36	1	1									1										
K			38	1									1								1		
F				32			1	1		2							1						
TH	2				34			2	1										1				1
CH						40								1			1						
H				1	1		33	1														1	
B	1			2	2		2	31				1										1	
D		1		1			1	1	35		1							1				1	
G	1	2	1		1		1			36									1				
V									1		37					1		1			1		
DH												37											
J													36										
S														38									
SH															39		1						
Z	1												1			38		1					
ZH														1	1	1	37	1					
Y	1																	33	1			1	1
R	1				1		1	1		2	2	1						2	33	1	1	1	
L	1																		1	37		1	1
W																					36		
M	1																		1	1		34	
N	1							1	1									1	1				37

Fig. 3.131 Confusion Matrix For Subject M.A.

	Presented																						
	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	39			2															1				
T		37														1	1						
K			37		1						1						1	1					
F				32	1		1	2			3					2		1					
TH					35						1	1											
CH			1			39									1								
H							34	1		1		1					1						
B					1			34	1	3	2							2				1	1
D								1	37	1													
G			1					1		32	1							1				1	
V				1			1				29												
DH				2			1		1	1		38											
J		1							1				40			1	1	1					
S														39			1						
SH														1	39								
Z		1		1	2	1										36					1		
ZH				1				1		2							34	2				1	
Y							1				1							30	1				
R	1			1			1				1						1		38				
L			1				1				1									39		2	
W																		1			39		
M																						35	
N																		1		1			39

Fig. 3.132 Confusion Matrix for Subject P.D.G.

	Presented																						
	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	95		1				2	1	1														
T		100																1					
K			97							3													
F				85	8			1			1												
TH				10	84						4	5											
CH			1			99									4								
H	1						90	1															
B	2				1		2	91	1	1													
D									98	1		1											
G			1				1			95			5										
V	1			4				4			86	14											
DH				1	6			1			5	78											
J						1							92				11						
S					1									98									
SH															96								
Z											1	1		1		97	1						
ZH													3	1		3	88						
Y							1											99					
R							2				3								100				
L																				98		1	1
W							1	1													100		
M							1													1		98	9
N	1											1								1		1	90

Fig. 3.133 Confusion Matrix Derived from Listening Tests on the Utterances of Subject C.W.T.

number of end points. For instance, an utterance of /r/ by subject C.W.T. should enter G2, but if it has an I.P.D. value less than or equal to 31 (say 30), it will be erroneously placed in G3 (Fig. 3.96). Once in G3 (Fig 3.109) it may be directed to one of 10 end points, and may be classified as /t/,/k/,/θ/,/v/,/ð/,/z/ or /ʒ/.

Figure 3.133 shows that a fair amount of scatter also exists in the errors made by human listeners, but most of these errors belong to a small number of important confusions, such as /f/,/θ/ and /ð/,/v/. These confusions often do not appear in the confusion matrices of figures 3.129 to 3.132, but are absorbed in the possible identifications. Figures 3.134 to 3.137 show confusion matrices for the four subjects in which the possible identifications have been treated as errors, and are classified according to the probable phoneme associated with the end point. For clarity, the 'absolute' errors shown in figures 3.129 to 3.132 are not entered, and the column sums are therefore equal to the corresponding diagonal entries of figures 3.129 to 3.132.

Most of the important confusions of figure 3.133 (human listeners) are duplicated in figures 3.134 to 3.137. The possible identifications also included considerable numbers of other confusions, which vary widely between subjects, but correspond to the similarities in the Z.T.I.

	Presented																						
	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	29											1					2					1	2
T		30	1																				
K		4	35								4							2					
F				22	8																		
TH				5	17							7											
CH		2				34									3								
H							12	1															
B							8	33		5		1											
D									29														
G									5	31													
V							3	1			16												
DH					2		2				11	28											
J		2				3	3						37				3						
S														33	9								
SH														6	26	4							
Z				3			2				2					27	10						
ZH				2	4	1										3	22						
Y																		26	12	4	2		
R																			16	2	6	4	2
L																		5	7	27	7	10	10
W																					16		
M																			4	3	1	21	7
N																				4	3	4	19

Fig. 3.134 Confusion Matrix Showing the Distribution of Probable & Possible Identifications for Subject C.W.T.

Presented	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	20	1	3					1												1	1		
T		32	4																				
K	2	3	27	1	2																		
F				32			2																
TH					13																		
CH	1	2	1			31							2		1	7	4						
H	4				2		29		2	8		2	2										
B	2				7		2	27	1			2											
D							1	3	30	3		2											
G					1		1			21		1											
V	5				4					2	31	3						2					
DH					1						3	24											
J													32			3	5						
S														34	7								
SH				1		2								6	30	1	1						
Z		1											1			23							
ZH						4										3	24						
Y																		19	4	8	2	5	5
R																		11	28	8	7	12	11
L																			7	15		3	
W																					21		
M																		3		4	3	18	4
N																							14

Fig. 3.135 Confusion Matrix Showing the Distribution of Probable & Possible Identifications for Subject W.A.A.

	Presented																					
	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	27			2			2	5											2	3	2	7	
T		27	2	1			2						3				2						
K		4	25			2																	
F				19	6		5		2														
TH				6	27		7														2	3	
CH			6			35									3								
H	3	3			1		11														2		
B		2					2	24	7	3													
D								2	18	3			2										
G			2	2					6	28													
V											22	16											
DH											13	21											
J						3							31	3			2						
S														21	2								
SH														14	34	2	3						
Z			2								2					36	6	3					
ZH				2													24						
Y									2	2								22					
R			1				4												22	5	5	9	3
L																		6	3	22	5		2
W																					20		
M																		2	6	3		15	2
N																				4			30

Fig. 3.136 Confusion Matrix Showing the Distribution of Probable & Possible Identifications for Subject M.A.

	Presented																					
	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	37																					
T		30	4										2									
K		3	31										2									
F				30			2					2					2					
TH				2	29		2					1										
CH						33										2						
H							20	2			3	3										
B							6	30			2	7								2		
D									30													
G							2		7	32												
V					2						17							2				
DH							2	2			5	23										
J		4									2		33			2						
S						3								34	6							
SH						3								5	31							
Z													3		2	29	14					
ZH			2		4							2				3	18					
Y																		18		4		3	2
R																		2	19	4	9	3	2
L																		8	10	18	6		8
W	2																		5	9	24	2	2
M																			4	2		23	3
N																						4	22

Fig. 3.137 Confusion Matrix Showing the Distribution of Probable & Possible Identifications for Subject P.D.G.

diagrams discussed in section 3.1. For instance the voiced stops (/b/,/d/ and /g/) are appreciably confused within themselves and with /h/, more so in the case of Subjects W.A.A. and M.A.

The possible identifications for the vowel-like sounds are greatly confused within themselves for all subjects. These confusions hardly occur at all for human listeners, except between the nasals /m/ and /n/.

3.35 Effect of Grouping the Phonemes.

Figure 3.138 shows the revised recognition rates of the four algorithms when the consonant phonemes were grouped into nine phoneme classes. These classes were voiced and voiceless Stops, Fricatives and Affricatives, Glides, Nasals , and the variable Fricative /h/.

The total recognition rate was improved very little by this grouping, but the number of possible identifications was approximately halved for each subject. Confusions in the absolute errors remained widely scattered, while there was a general improvement in the number of probable identifications for all classes. Figures 3.139 and 3.140 are grouped confusion matrices for Subject C.W.T. In figure 3.139, both probable and possible identifications are considered to be correct, while figure 3.140 shows the distribution of errors in the possible identifications.

	Prob.	Poss.	Total	Human Subjects
C.W.T..	78.7	12.1	90.8	95.6
W.A.A.	74.2	14.9	89.1	
M.A.	74.8	15.3	90.1	
P.D.G.	81.2	10.4	91.6	
Average	77.2	13.2	90.4	

Fig. 3.138 Recognition Rates with Grouping into Phoneme Classes. (%)

	Presented								
	U. Stop	V. Stop	U. Fric.	V. Fric.	U. Aff.	V. Aff.	Glide	Nasal	/h/
U. Stop	107		4	4	2		1		1
V. Stop	2	110	1	1		1			3
U. Fric.	1	1	144	4			2		3
V. Fric.	4	2	5	143			8	1	1
U. Aff.	1			1	38	1			
V. Aff.		4	2	1		37			2
Glide	5		4	5			147		
Nasal								79	
/h/		3		1		1			30

Fig. 3.139 Confusion Matrix with Phoneme Grouping: Subject C.W.T.

	Presented								
	U. Stop	V. Stop	U. Fric.	V. Fric.	U. Aff.	V. Aff.	Glide	Nasal	/h/
U. Stop	101			7			2	2	
V. Stop		108		1					8
U. Fric.			130	11					
V. Fric.		1	11	121	1				7
U. Aff.	2		3		34				
V. Aff.	2			3	3	37			3
Glide	2						130	26	
Nasal							15	51	
/h/									12

Fig. 3.140 Confusion Matrix of Probable & Possible Identifications with Phoneme Grouping: Subject C.W.T.

3.36 Subject Dependance.

The recognition rates achieved when the data for each subject was processed by each algorithm are tabulated in figure 3.141. All the algorithms proved to be extremely sensitive to subject changes. For the three subjects other than the one for which each algorithm was designed, the total percentage of successes fell to 50-60, and the number of possible identifications became only slightly less than that of the probable identifications. The algorithm for subject C.W.T. was slightly less sensitive to subject changes than the others.

The recognition rates for individual phonemes varied widely under these conditions, but again the vowel-like sounds showed the worst results, and the G1 phonemes the best. The performance of the algorithms remained relatively insensitive to vowel changes.

3.37 The Recognition Parameters.

It was difficult to form a quantitative estimate of the relative merits of the individual recognition parameters. Some information can, however, be obtained by considering the number of decisions employing each parameter in the four recognition algorithms. These numbers are tabulated in figure 3.142.

The total numbers of decisions comprising the algorithms for subjects C.W.T. and P.D.G. were slightly less than those for subjects W.A.A. and M.A., corresponding to the clearer phoneme-distinctions observed for the former. The I. parameters were used less frequently for subjects W.A.A. and M.A., due to the greater number of sounds with no separate I. peak, and the general reduction in size and clarity of the I. peak for these speakers.

There was no real difference in the number of occasions on which each of the three functions Z., T., and I. was used. As expected, the most useful parameters were the peak heights Z.P.S., T.P.S., and I.P.S. T.P.S. was the most frequently used parameter of all. The peak drops Z.P.D., T.P.D., and I.P.D. were used least frequently, though I.P.D. proved very useful in making the G3 decision, and T.P.D. could be used to separate the vowel-like sounds.

Of the three duration measurements, Z.P.W. was by far the most frequently used, and I.P.W. the least, though the latter measurement could not be made when there was no separate I. peak. The three onset times were used quite frequently, and appeared to be equally useful.

The additional measurements of I.S.D. and the difference(I.P.S-I.P.D.) were used once only in each algorithm (except that for W.A.A., when there was no G5), but enabled important seperations to be made.

Data	Algorithm C.W.T. Prob.	Poss.	Total	W.A.A. Prob.	Poss.	Total	M.A. Prob.	Poss.	Total	P.D.G. Prob.	Poss.	Total
C.W.T.	64	25.6	89.6	33.7	23.3	57	32.2	23.3	55.5	32.4	27.5	59.9
W.A.A.	31.1	23.9	55	62.5	25	87.5	29	19.5	48.5	23.7	23.4	47.1
M.A.	32.5	25	57.5	23.9	21.8	45.7	61.4	27.2	88.6	25	23	48
P.D.G.	34.6	32.7	67.3	29.5	28.2	57.7	27.2	26.2	53.4	66.8	23.7	90.5

Fig. 3.141 Recognition Rates Achieved by all Possible Combinations of Speaker & Algorithm. (%)

	ZPS	ZPD	ZPW	ZLE	All Z	TPS	TPD	TPW	TLE	All T	IPS	IPD	IPW	ILE	All I	ISD	(IPS IPD)	Total
C.W.T.	6		5	3	14	10	2	2	2	16	9	2	2	2	15	1	1	47
W.A.A.	7	3	7	2	19	11	3	2	3	19	5	4	1	2	12	1		51
M.A.	8	3	5	4	20	12	1	3	2	18	4	3	1	4	12	1	1	52
P.D.G.	4	2	5	3	14	3	3	4	4	14	12	2	2	2	18	1	1	48
Total	25	8	22	12	67	36	9	11	11	67	30	11	6	10	57	4	3	

Fig. 3.142 Table Showing the Number of Decisions Taken Using Each Recognition Parameter in Each Algorithm.

CHAPTER 4.

The main finding of the present study is that in the idealised recognition situation considered, a high degree of discrimination between individual consonant phonemes can be obtained from the Z.T.I. diagram. The level of performance achieved (Section 3.3), if it could be repeated in a more realistic situation, might suffice for the recognition of these phonemes in an A.S.R. device which could utilise processing on higher linguistic levels to overcome the remaining uncertainties. In this chapter, the results obtained are discussed and an attempt is made to assess the difficulties involved in extending the work to cover practical recognition tasks.

4.1. Consonant Peaks on the Z.T.I. Diagram.

As the illustrations of section 3.1 show, the Z.T.I. diagram gave a much simpler form of visual representation than the sonagram for consonant sounds. The initial consonants in C.V. sounds were for the most part represented by single, fairly well defined peaks on the Z., T., and I. traces. Occasionally, multiple consonant peaks were observed, most freqently on the Z. trace (eg. Figure 3.61). These occured chiefly in voiced fricative sounds, and were

attributed to changes in the relative proportions of voiced and voiceless modulation within the individual phoneme (Section 3.1.15).

The appearance of the consonant peaks generally tallied quite well with the predictions made in Section 1.4 . The unusual behaviour observed in some cases (eg. The abnormally large I. peaks for /p/ (Section 3.1.1) and the high I.S.D. values for /f/ (Section 3.1.9)) could be explained with reference to the sonagrams and waveforms of these sounds.

The use of isolated C.V. syllables reduced the segmentation problem to that of finding the boundary between consonant and vowel. These boundaries could readily be picked out by eye in most cases, and it was generally possible to associate the boundary with a local minimum position on each trace. An algorithm was developed (Section 2.1.6) to perform this segmentation on each trace independantly, since the apparent boundary positions could differ between the three traces. Sometimes, however, no satisfactory minimum could be found (eg. Figure 3.20), and on other occasions the existence of more than one possible boundary position (eg. Fig 3.56) necessitated manual correction with reference to the other traces. The latter effect occured most frequently on the T. trace. The lack of a more efficient segmentation algorithm was the main obstacle preventing

full automation of the parameter extraction process.

Though adequate for the purposes of the present study, the association of the consonant- vowel boundary with a local minimum was not always viable, and this approach to segmentation would probably be of little use in the treatment of connected speech. The segmentation algorithms appeared to function better for subjects C.W.T. and P.D.G. , whose pronunciation of the C.V. sounds was slower and somewhat more deliberate than that of subjects W.A.A. and M.A. The work of Reddy(46) has shown that successful speech segmentation is possible using intensity and zero-crossing measurements.

For the majority of the C.V. sounds processed, distinct consonant peaks could be identified on all three traces. Instances where the consonant peak was indistinct or absent were most common on the I. trace (about 20% of all the sounds considered), and also occured on the Z. trace (less than 5%). A separate consonant peak was almost invariably present on the T. trace. The absence of consonant peaks was attributed to a smooth transition between consonant and vowel (as in the case of the I. trace for the voiced stops , section 3.1.4), or to an exceptionally low value for the consonant sound on the relevant trace (eg. the Z. trace for /v/ and /ð/, section 3.1.11), or to a combination of these factors.

The sharpness of the transitions between consonant and

vowel on the Z.T.I. diagram was also dependant on the choice of the sampling time, t_c, and the number of points, n , constituting the time window used in the smoothing process (Section 2.12). This dependance was greatest on the I. trace. The values of t_c and n were fixed by visual inspection of the Z.T.I. diagrams for utterances by C.W.T. The increased proportion of consonant sounds showing no I. peaks for subjects W.A.A. and M.A. may thus be partially due to inappropriate values of t_c and n, though the size of the I. peaks was generally reduced for these subjects (Section 3.1.11.2)

The 'crests' seen on the I. trace for vowel- like sounds proved extremely useful for the recognition of this group of phonemes, though they did not generally identify the boundary between consonant and vowel (Section 3.1.19). Variations in the peak amplitude between glottal periods , corresponding to the crests, were observed in the raw speech waveforms for these sounds (see Figure 3.87). The low value of t_c (6.4ms., often shorter than the glottal period) could have little effect on the appearance of the crests, due to the short- time smoothing process used subsequently.

The many factors which contribute to the extreme variability seen in the sonagrams of consonant phonemes

also have a marked influence on the Z.T.I. diagram. Thus changes in stress, pitch, formant structure, context, duration and the like all effected the appearance of the Z.T.I. diagrams, and distinctly different types of consonant peaks were sometimes obtained for different utterances of the same phoneme by the same speaker. It was therefore necessary to introduce multiple end points in the recognition algorithms. These factors, however, seemed to have a less radical effect on the Z.T.I. diagram than on the sonagram, at least for a single speaker. This is exemplified by the independance of the recognition algorithms to the vowel following the consonant.

The effect of stress, duration, formant changes etc. was far more pronounced between speakers, and since no attempt was made to compensate for these changes in the measurement of the recognition parameters it was hardly suprising that the speaker dependance of the algorithms was so great. These difficulties would, of course, be compounded in a general recognition situation, where consonants uttered in the central and final positions would also have to be considered. It would then be advisable to incorporate some form of normalisation to minimise these perturbations, if a satisfactory means of doing this could be found. Duration measurements could, for instance, be adjusted according to a running phonemic or syllabic rate measurement.

Spectral features such as formants and energy fills, which can be distinguished independantly in the sonagram, merge in a complex way in the Z.T.I. diagram, though the Z. trace remains most influenced by F1 and the T trace by the higher formants. The success of the present study shows that this need not necessarily be a disadvantage. The way in which spectrally independant features interact to produce consonant peaks on the Z.T.I. diagram can give a useful representation for recognition purposes (eg. the Z.T.I. diagrams for /z/, Section 3.1.15).

4.2 Recognition Parameters.

The three functions Z,T and I. seemed to be of equal importance for consonant recognition. Parameters derived from the I. trace were used more frequently in the earlier stages of the recognition algorithms, particularily in the Group Allocation routines. Similar measurements were used by Reddy(46) to obtain a broad phoneme classification.

The consonant peak size parameters (Z.P.S, T.P.S., and I.P.S.) were determined from the maximum height attained on each trace during the consonant portion of the sound. Other workers, notably Lavington(31) have used measurement of Zero Crossing Rate etc. averaged over the duration of the phoneme. The peak picking approach introduces more

variability in the measurement, but seems to enable finer discrimination between the phonemes. One reason for this is that in many cases the consonant peaks were quite sharp, and no 'steady- state' position was reached. When this occurs, the maximum height of the trace is probably more directly related to the characteristic articulator configuration for the consonant phoneme than is the average height of the trace.

The peak size parameters were expressed as a percentage of the maximum height of the following vowel.(Section 1.3). On the Z. and T. traces, this procedure decreased the spread in the peak size values for those phonemes which are strongly context dependant, but increased the spread for the remaining phonemes. In connected speech, this simple form of normalisation would be impractical, but some adjustment of Z.P.S. and T.P.S. values according to running F1 and F2 measurements would be valuable. The present method of normalisation is very sensitive to pitch changes, especially on the Z. trace, and this greatly reduced the value of the Z.P.S. measurement for subject W.A.A.(Section 3.1.17).

The 'peak drop' parameters proved to be of relatively little use, though high I.P.D. values served to distinguish the 'crests' on the I. trace, and T.P.D. could be used to distinguish between the vowel- like phonemes to some extent. It is thought, however, that some measurement of the

abruptness of the change from consonant to vowel would be of advantage. The peak drop measurements could be improved by normalising by the consonant height rather than the vowel height, though it may not always be possible to associate the consonant- vowel boundary with a minimum position on the trace. 'Trailing- edge' duration or slope measures may be a more satisfactory approach.

The duration parameters (Z.P.W., T.P.W., I.P.W.) played a most important role in the recognition algorithms. These measurements proved to be very speaker dependant (compare figures 3.55 and 3.56), and some normalisation according to the speed of the speech would be essential in any extension of the method. For a single speaker, Z.P.W. seemed to be the most stable measurement, and this parameter was used most frequently in the recognition algorithms (Section 3.3.7). On the I. trace, the duration estimate was handicapped by indistinct consonant peaks, while T.P.W. suffered from ambiguities in the positioning of the consonant- vowel boundary on the T. trace. Sonagrams of C.V. syllables (eg. Figures 1.4 and 1.7) show more abrupt changes in the F1 region between consonant and vowel, leading to improved clarity of the phoneme transition on the Z. trace.

The time-to-rise parameters (Z.L.E., T.L.E., and I.L.E.) were also of great value for consonant recognition. The

L.E. values could probably be stabilised between speakers by a normalisation procedure similar to that suggested for the duration measurements. The three L.E. parameters appeared to possess an appreciable degree of independance, though it might be possible to discard one of these parameters without ill effect. The sharpness of the onset of the consonant sound could also be estimated from the slope of the trace prior to the consonant maximum.

The measurement of I.S.D. served to identify those consonants with a very low overall energy level for at lest some portion of the utterance. This could be done more elegantly using the average value of the I. trace prior to the consonant peak.

Apart from the peak drop and I.S.D. measurements, all the recognition parameters could be extracted from connected speech, provided the consonant segment could first be isolated. With the present parameter set, however, little information is available in cases where no consonant peak can be found. The incidence of these cases must increase in connected speech, when pronunciation is less clear, and it would probably be necessary to resort to measurements of the average value of each trace , and possibly the standard deviation of this average, after the manner of Reddy(46). For Stop sounds, the partial silence of the Stop Gap is an additional cue which was not used in this

study, but could be identified on the I. trace in connected speech.

4.3 Recognition Algorithms (Section 3.2)

Binary Threshold Decision Trees were used solely because of the ease with which they can be constructed and altered to fit new data. With multiple end points for a single pattern category, all the possible rectangular divisions of the parameter space could be implemented.

The great disadvantage of this type of recognition algorithm is the 'hit or miss' nature of the sequential threshold decisions. The introduction of 'probable' and 'possible' identification categories partially compensated for this, in giving a better picture of the true capabilities of the parameter set. Nevertheless, the sensitivity of the algorithms to numerically small variations in the parameter values was doubtless partially responsible for the extreme speaker dependance. This is evidenced by the greatly increased proportion of possible identifications when the speaker was changed(See Figure 3.141)

Ideally, the output of the phoneme recogniser should consist of an estimate of the relative probabilities that each phoneme has been spoken, which could then be utilised on the next recognition level. The phoneme recognition algorithm should also be capable of being adjusted by

feedback from other levels. There are many possible approaches to the construction of such an algorithm, which need not be described here. While decisions taken in sequence may not be the best method, the tree structure could be retained by a scheme of the following kind. On entering the tree, a number of 'marks' could be allocated, say 100, which would then be distributed at each succeeding decision node in accordance with rules dependant on distance from the threshold. Thus at the first node, a 'strength' of 70 marks might be alloted to the left hand path, and 30 to the right hand path. On the left hand path, the next decision might divide the 70 marks into 50 and 20, and so on. The number of marks collected at each end point would represent the phoneme probabilities. Such a scheme would to some degree retain the ease of adjustment of the simple recognition tree.

The Group decision sequence (Section 3.2.1) was similar for all four recognition algorithms, except that for subject W.A.A., when T.P.S. was used for the G1 decision [illegible] rather than Z.P.S., and there was no G5. Otherwise, the same parameters were used in the group allocation, though threshold values varied a good deal. The group separation corresponded only partially to separation of the phoneme classes(eg. Both the unvoiced stops /t/ and /k/, and the voiced fricatives /v/ and /ð/ commonly entered G3).

Many parallels between the four algorithms could also be seen within the groups (Section 3.2.2). While the phoneme separation was generally performed according to the same principles, there were wide differences in parameters, threshold values, and the order in which the decisions were taken. The most important cues used to identify each phoneme are summarised below.

Stops.

Short duration, Sharp peaks with low L.E. values.

Voiceless Stops. Distinct I. peaks.

/p/ Exceptionally large I. peaks.
Smaller Z. and T. peaks than /t/ or /k/.

/t/ Very high T. peaks, often large Z. peaks.

/k/ Z. peaks again prominent, T. peaks smaller than for /t/.
Often longer duration than other voiceless stops.

Voiced Stops. No I. peak.

/b/ Smallest T. peaks for voiced stops.
Often longer duration and rise time.

/d/ High T. peaks.

/g/ T. peaks of intermediate height.
Very short rise time.

Fricatives.

Longer duration and rise time.

/h/ No I. peak.

Small Z. and T. peaks.

Duration and rise time greater than for voiced stops.

Sometimes large I.S.D. value.

Voiceless Fricatives. Distinct I. peaks, I.P.S. and I.P.D. smaller than for voiced fricatives.

/f/ Often high I.S.D. value. Sometimes no I. peak.

Size of peaks variable, but very long duration and gradual rise to maximum.

/θ/ Often high I.S.D., sometimes no I. peak

Small Z. peaks, long duration.

/s/ Massive Z. and T. peaks, small I. peaks, large I.L.E. values.

T.P.S. larger than for /ʃ/.

/ʃ/ Similar to /s/, but smaller T. peaks.

Voiced Fricatives. I.P.S. larger than for voiceless fricatives.

/v/ and /ð/. Very small Z. peaks.

/z/ Z. peaks variable, but massive T. peaks, T.P.S. very high.

/ʒ/ Similar to /z/, with lower T.P.S. values.

Affricatives.

Duration longer than for Stops, shorter than for Fricatives. L.E. shorter than for Fricatives.

/tʃ/ Distinct I. peaks, very high Z. and T. peaks.

/dʒ/ No I. peaks, high Z. and T. peaks.

Vowel- Like Sounds.

'Crests' give higher I.P.S. and I.P.D. values than for other sounds(except /p/).

Smallevalue of (I.P.S.- I.P.D.).

Nasals (/m/, /n/) often have larger I.P.S. values tham Glides.

/j/ sometimes has large T. peaks.

T.P.D. values give some distinction between vowel-like sounds.

4.4 Performance of the Algorithms.(See Section 3.3)

The overall recognition scores for the four algorithms with the appropriate speakers were very similar indeed. Including possible identifications, about 90% of the consonants were identified correctly, 65% being probable identifications (Section 3.3.1). These figures include those for the vowel-like sounds, for which the method was generally unsuitable. The formant transitions which are important for the perception of vowel-like sounds (21) are not well represented on the Z.T.I. diagram, and what little distinction could be made between the glides was mainly based on the initial articulator position prior to

the start of formant movement. This position is rarely held for any length of time in continuous speech. Excluding the performance figures for the vowel-like sounds, probable identifications rose to 70-75%, and would undoubtedly rise still further if the vowel-like sounds were not considered in the formulation of the algorithm.

Satisfactory means were found for identifying most of the Stop, Fricative and Affricative phonems. Of these, the most difficult to recognise were those which showed small or indistinct peaks on the Z.T.I. diagram, and those for which the form of the peaks was most variable (eg. /h/, /f/, /θ/, /v/ and /ð/). The problem of isolating these sounds is expected to become more pronounced in connected speech, and more recognition parameters may be needed.

The performance of the algorithms was not dependant on the vowel phoneme in the C.V. sound (Section 3.3.2). This is a great advantage, since in order to 'read' sonagrams, even of isolated syllables, it is often necessary to resort to using the effect of the consonant sound on the vowel.

The strong speaker dependance of the recognition algorithms was exaggerated by the lack of normalisation of the recognition parameters, and by the type of algorithm used. While it should be possible to reduce the speaker dependance in later applications, a 'tune-in'

routine to adjust the recogniser to a new speaker will probably remain a necessity.

4.5 Conclusions.

The consonant recognition scheme described has the great advantage of speed. As no involved computations were performed, the time taken in actual processing by the computer was very small, about 100ms. for each C.V. syllable. The overall time taken was increased by the necessity of transfering data in and out of the small computer store, and by the use of manual corrections and associated displays. The need for manual intervention could probably be eliminated by more careful design of the segmentation and parameter extraction routines, and given a larger store, real time operation seems a possibility.

From the work on isolated C.V. sounds, it appears that the recognition of Stop, Fricative and Affricative phonemes can be performed adequately without reference to spectral data. The difficulties of extending this approach to deal with more realistic recognition tasks are manifold, but appear to be no more severe than those encountered in spectral recognition devices.

Appendix 1.

Magnetic Recording and Other Equipment.

The C.V. sounds were recorded on a Truvox R42 model tape recorder. The record/replay frequency response of this machine had previously been measured by Underwood(55). The response at a tape speed of 7½" per second is shown in figure A1.1 and is substantially flat from 150Hz. to 6KHz.

The microphone used was a Reslo Studio Ribbon type S.R.1.L, with a flat frequency response from 30Hz. to 20KHz. The measured signal-to-noise ration of the microphone and tape recorder was 49db.

A Revox G36 stereo tape recorder was used in the listening tests. Sonagrams were recorded with a Kay model R Sonagraph, and speech waveforms with an S.E. Laboratories type 3006 U-V recorder.

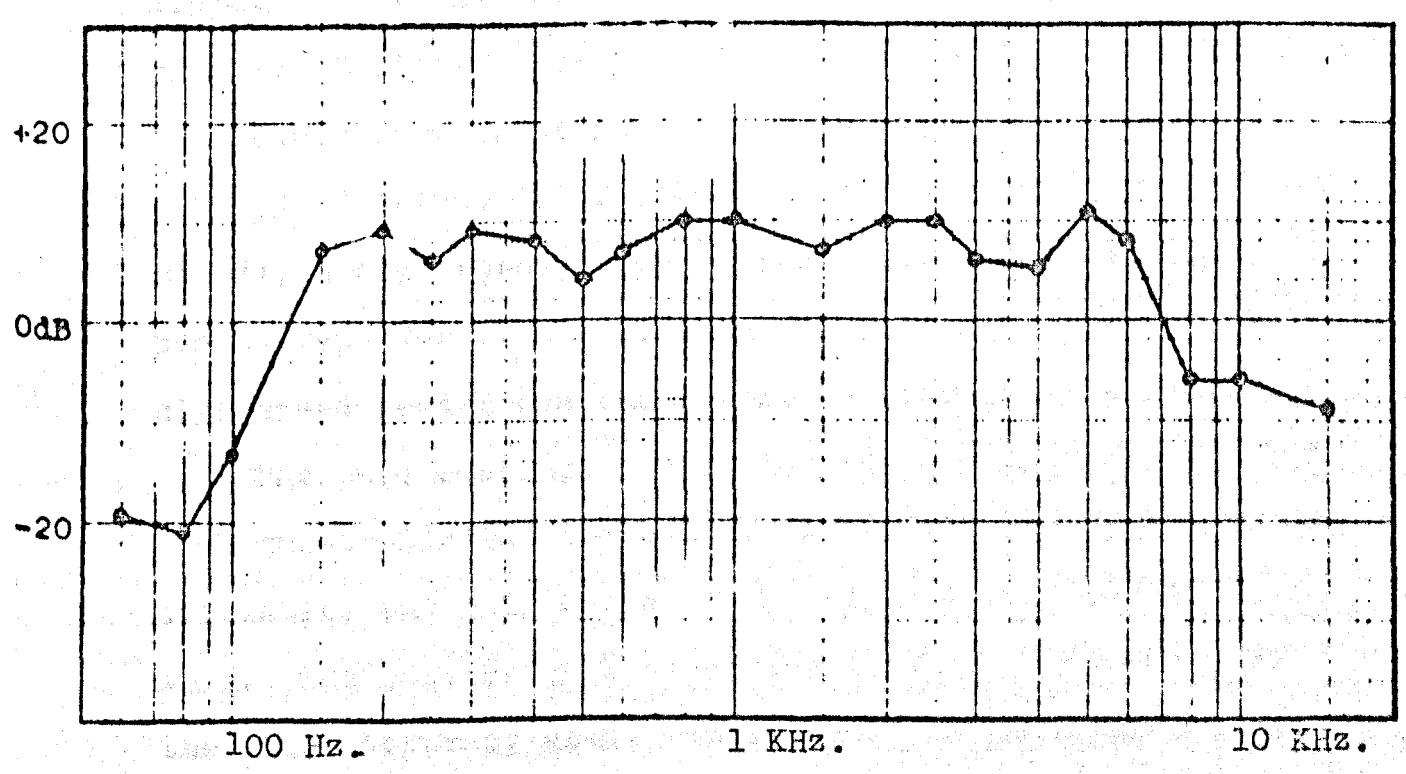


Fig. A1.1 Frequency Response of Tape Recorder.

Appendix 2.

The Computer Installation.

The computer used was a Digital Equipment Corporation P.D.P. 8. This small, high speed machine had a 12 bit word length and 8K of store. The cycle time of the P.D.P.8 was 1.5μs. The additional Extended Arithmetic Element (Type 182) greatly simplified the computation of recognition parameters.

Paper tape input and output devices consisted of an A.S.R.33 Teletype (10 characters/second reading and writing speed), a high speed reader (Type 750c) and a high speed punch (Type 75E). The reading and writing speeds of the high speed reader and punch were 69 characters/second.

The tape recorder was connected to a Digital Type 138E Analogue-to-Digital converter. A Multiplexar Type 139E was available, but only one A. to D. channel was used in this work. The conversion time of the A. to D. converter, at the 9bits accuracy used, was 13.5μs. An additional 12bit I.O. register was also available.

Magnetic Tape data storage facilities comprised two Dec Tape (Type 55) units. Both units could be used by the same program, though not simultaneously.

The cathode ray tube display used was a Digital Programmed Buffer Display type 338. The display logic functioned asynchronously with that of the computer, and the 338 obtained information from the P.D.P.8 by means of

a data break cycle. The display could be modified by means of a light pen, or by a bank of 12 push buttons. The light pen facility was not used in this study.

A Digital-to-Analogue converter designed by A.W. Wright was used in the listening tests.

Appendix 3.

Consonant Listening Tests.

The aim of these listening tests was to get some idea of human performance in a consonant recognition task as close as possible to that of the machine recognition situation.

Experiments on the intelligibility of consonant phonemes in C.V. syllables were performed by Miller and Nicely (39). In this study, the experimental conditions differed in several aspects from those of the present work:

a) Only 16 consonant phonemes were spoken. Of the 23 consonants considered in the present study, /tʃ/, /h/, /dʒ/, /j/, /r/, /l/ and /w/ were omitted.

b) Only one vowel, /a/, was used.

c) All speakers and listeners were American Females.

d) One of the aims of Miller and Nicely's work was to investigate the effect of noise on consonant recognition. The signal-to-noise ratio was therefore varied, the highest value being +12db. The ratio used in the present work was +49db.

In view of these differences, it was decided to perform a short series of listening tests. Facilities for conducting such tests under the control of the P.D.P.8 computer were available within the department. These facilities are described by Ainsworth and Millar(2 ,3).

Twelve switch boxes, each having twelve binary switches, corresponding to the 12bit word of the P.D.P.8, could be scanned automatically and their contents read into the computer by means of the I.O. register (See Appendix 2.). To cover the 23 consonant phonemes, each listener was provided with two switch boxes.

The data comprised two utterances by Subject C.W.T. of each consonant phoneme with each of 10 vowel phonemes, or one half of the utterances recorded for this subject(See Section 1.1). These sounds were re-recorded in random order on one track of a stereo Revox tape recorder. The second track carried a series of timing pulses indicating the termination of each C.V. syllable. This track was monitored by the P.D.P.8 through the A.to D. converter. The P.D.P.8 was able to stop and start the tape recorder by means of signals from the D.to A. converter connected to the tape recorder remote control socket.

The data was divided into five parts, each dealing with two vowel phonemes. These were run as five separate listening tests over a period of a few weeks. The duration of each test was about 20 minutes.

A panel of five listeners (three male and two female) was used throughout. All the listeners were familiar with the speech of C.W.T., and one had previous experience in listening to isolated C.V. sounds. A 'forced choice'

situation was implemented, since the recognition algorithms operated in this manner. The following instructions were recorded by C.W.T. and played before the start of each listening test in order to familiarise the listeners with the speaker:

"The object of this experiment is to investigate the perception of consonant sounds in consonant-vowel syllables. The sounds which you are about to hear have been produced by my voice and are recorded on tape. The experiment is controlled by a digital computer which will stop and start the tape and record your responses to each sound.

Each subject has two switch boxes in front of him, and the 23 consonants used each correspond to a single switch. The most left hand switch on your left hand switch box is not used.

Each sound consists of a consonant followed by a vowel. For instance with the vowel /ɛ/, as in HEAD, the sounds are as follows, reading from right to left on your boxes:-

/pɛ/

/tɛ/

(etc.)

When you have identified the consonant, please press the appropriate switch. This is a forced choice experiment ie. there is no 'don't know' switch. If you are in doubt about any sound, pick the 'closest'. You may assume that

all the sounds are equally likely to occur at any stage in the experiment.

During the period of waiting for the responses to be made, the light on the central control box will flash slowly on and off, as it is doing now. When everyone has pressed a switch, the light will begin to flash quickly. When this occurs, please return your switch to its original position. When everyone has done this the next sound will be played. Thank you."

The task of choosing between 23 categories was quite demanding, and learning effects were observed in the first listening test, when the response rate of the listeners was initially very slow. For this reason, the results of the first experiment were discarded, and the experiment was repeated at a later date. Because of the complexity of the task, no attempt was made to incorporate 'probable' and 'possible' identifications by allowing alternative switches to be pressed. It is thought that such a scheme would eliminate many of the errors made.

The results of these listening tests given in Figures 3.120, 3.133 and 3.138 are pooled over all listeners and all vowels. For comparison with Figure 3.133, the confusion matrix given by Miller and Nicely for a signal-to-noise ratio of 12db. is given in figure A3.1. The slightly lower overall recognition rate for the latter

experiment was probably due to the lower signal-to-noise ratio. The distribution of confusions differed widely between the two experiments, no doubt partly due to the different experimental conditions. Insufficient data was available from the present listening tests to draw any conclusions regarding these confusions.

Heard	Presented																					
	P	T	K	F	TH	CH	H	B	D	G	V	DH	J	S	SH	Z	ZH	Y	R	L	W	M	N
P	90		16	1																		
T		99																				
K	7	1	84																			
F				76	9			2														
TH	3			23	91			1														
CH																						
H																						
B								94			11	5										
D									86	9												
G									13	87		1										
V								3			78											
DH										2	11	90				2						
J																						
S														100								
SH															100							
Z									1	2		4				98						
ZH																	100					
Y																						
R																						
L																						
W																						
M																					100	
N																						100

Fig. A3.1 Consonant Confusion Matrix after Miller & Nicely
(Figures Normalised to Percentages)

REFERENCES.

Abbreviations: J.A.S.A. - Journal of the Acoustical Society of America.
S.T.L. - Speech Transmission Lab., Stockholm.

1. Ainsworth, W.A. "Relative intelligibility of different transforms of clipped speech."
   J.A.S.A. 41 p.1272 (1967).

2. Ainsworth, W.A. Millar, J.B. "A Simple time sharing system for speech perception experiments."
   Proc. 5th Decus European Seminar, Stockholm, p1 (1969)

3. Ainsworth, W.A. Millar, J.B. "Methodology of experiments on the perception of synthesised vowels."
   To be published in Language & Speech, Dec. 1971.

4. Bekesy, G.von "Experiments in hearing."
   McGraw-Hill, New York, 1960

5. Bezdel, W. "Discrimination of sound classes for speech recognition purposes."
   1967 Conference on speech communication & processing, M.I.T.

6. Bezdel, W. Chandler, H.J. "Results of an analysis & recognition of vowels by computer using zero-crossing data."
   Proc. Inst. Elec. Eng., 112, p.2060 (1965).

7. Bloch, B.
Trager, G.L.
"Outline of Linguistic Analysis".
Linguistic Soc. of America, Baltimore, Waverley Press, 1942.

8. Broadbent, D.E.
Ladefoged, P.
"Vowel Judgements & Adaptation Level".
Proc. Royal Soc.,B, 151, p.384 (1960).

9. Campanella,S.J.
Coulter, D.C.
Irons, R.
"Influence of Transmission error on Formant coded compressed speech signals".
Proc. Speech Com. Sem., STL, Stockholm, Vol 2 (1963).

10. Chang, S.H.
Pihl, G.E.
Wiren, J.
"The intervalgram as a visual representation of speech sounds".
J.A.S.A. 23 p.675 (1951).

11. Chang, S.H.
"Two schemes of speech compression".
J.A.S.A. 28 p.565 (1956).

12. Dammann, J.A.
"Application of adaptive threshold elements to the recognition of acoustic-phonetic states".
J.A.S.A. 38, No. 2, p213 (1965).

13. Dolansky, L.
Tjernlund, P.
"On certain irregularities of voiced speech waveforms".
I.E.E.E. Trans. on Audio & Electro-Acoustics Vol. AU-16, No.1,p. 51 (1968)

14. Doshita, S.
"Studies on the analysis & recognition of Japanese speech sounds".
Thesis, Kyoto University, 1965.

15. Dudley, H. "Oscillograms".
Fifth International Congress on Acoustics, Liege, paper A48 (1965).

16. Egan, J.P. "Articulation testing methods".
The Laryngoscope 58 p.955 (1948)

17. Fant, G. "Acoustic Theory of Speech Production".
Mouton & Co., 1960.

18. Fant, G. "The Nature of Distinctive Features".
S.T.L. Quarterly Progress & Status Report, 4/1966 p.1 (1967).

19. Fant, G. "Automatic Recognition & Speech Research".
S.T.L. Quarterly Progress & Status Report, April, 1970 p.16.

20. Flanagan, J.L. "Difference Limens for Vowel Formant Frequency".
J.A.S.A. 27 p.613 (1955)

21. Flanagan, J.L. "Speech Analysis, Synthesis & Perception".
Springer-Verlag, 1965.

22. Forgie, J.W.
Forgie, C.D. "Results Obtained from a Vowel Recognition Computer Programme."
J.A.S.A. 31 p.1480 (1959)

23. Fourcin, A.J. "An investigation into the possibility of Bandwidth Reduction in Speech".
Ph.D. Thesis, University of London, 1960.

24. Fry, D.B. "Automatic Recognition of Speech"
Proc. International Congress of Phonetics, Helsinki (1961).

25. Gold, B. "Computer Programme for Pitch Extraction".
J.A.S.A. 34 p.916 (1962).

26. Herscher, M.B.
Kelley, T.P.
Clodfelter, R.G.
Petroski, D.R. "Voice Controller for Astronaut Manoeuvering Unit".
Proc. 2nd National Conference on Space Maintainence & Extra-Vehicular Activities, p. V4-1, Las Vegas (1968)

27. Holmes, J.N.
Mattingley, I.C.
Shearme, J.N. "Speech Synthesis by Rule".
Language & Speech 7 p.127 (1964).

28. Hyde, S.R. "Automatic Speech Recognition Literature Survey & Discussion".
G.P.O. Telecommunications Research Dept. Report No. 45 (1968)

29. Jakobson, R
Fant, C.G.M.
Halle, M. "Preliminaries to Speech Analysis".
The M.I.T. Press, Cambridge, Massachusetts, 1961.

30. Lavington, S.H.
Rosenthal, L.E. "Some Facilities for Speech Processing by Computer".
Computer Journal 9 p.330 (1967).

31. Lavington, S.H. "Measurement Systems for Automatic Speech Recognition".
Ph.D. Thesis, University of Manchester, 1968.

32. Lavington, S.H. "Problems in Automatic Speech Recognition". Proc. I.E.E. Colloquium on Some Aspects of Speech Recognition for Man-Machine Communications. Colloquium Digest No. 1968/13.

33. Lenaerts, E.H. et al. "Talking to the Computer". New Scientist, 4th Dec. 1969, p.498.

34. Licklider,J.C. "Effects of Amplitude Distortion upon the Intelligibility of Speech". J.A.S.A. 18 p.429 (1946).

35. Licklider, J.C. Pollack, I. "Effects of Differentiation, Integration, and Infinite Clipping upon the Intelligibility of Speech". J.A.S.A. 20 p.42 (1948).

36. Licklider,J.C. "Intelligibility of Amplitude-dichotomised, time-quantised speech waves." J.A.S.A. 22 p.820 (1950).

37. Mackay, D.M. Millar, J.B. Underwood, M.J. "Discriminative Value of the Digram Structure of Speech Waveforms". Proc. 18th. International Congress on Psychology, Moscow, 1966.

38. Martin, T.B. Zadell, H.J. Granza, E.F. Hercher, M.B. "Numeric Speech Translating System". Proc. I.E.E.E./P.O. Symposium on Pattern Recognition, Washington, May 1969.

39. Miller, G.A. Nicely, P.E. "An Analysis of Perceptual Confusions among some English Consonants". J.A.S.A. 27 p.338 (1955).

40. Millar, J.B. "The Evaluation of Three Related Techniques for the Statistical Analysis of Clipped Speech".
Ph.D. Thesis, University of Keele, 1968.

41. Nillson, N.J. "Learning Machines".
McGraw-Hill, New York, 1965.

42. Peterson, G.E.
Barney, H.L. "Control Methods used in a Study of the Vowels".
J.A.S.A. 24, No.2, p.175 (1952)

43. Phillips, V.J.
Cherry, E.C. "Some Possible uses of Single Sideband Signals in Formant Tracking Systems".
J.A.S.A. 33 p.1067 (1961)

44. Potter, R.K.
Kopp, G.A.
Green, H.C. "Visible Speech".
Van Nostrand, 1947.

45. Purton, R.F. "Speech Recognition using Auto-Correlation Analysis".
I.E.E.E. Trans on Audio & Electro-acoustics AU-16, No.2,p.235 (1968).

46. Reddy, D.R. "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave".
Technical Report No. CS49, Computer Science Dept., Stanford University (1966).

47. Reddy, D.R. "Computer Recognition of Connected Speech".
J.A.S.A. 42 NO 2, p.329 (1967).

48. Sakai, T.
Inoue, S.
"New Instruments & Methods for Speech Analysis".
J.A.S.A. 32 p.441 (1960).

49. Sakai, T.
Doshita, S.
"The Automatic Speech Recognition System for Conversational Sound".
I.E.E.E. Trans. on Electronic Computers. E.C.-12 p.835 (1963).

50. Scarr, R.W.A.
"Zero Crossings as a means of obtaining Spectral Information in Speech Analysis".
1967 Conference on Speech Communication & Processing, M.I.T.

51. Stover, W.R.
"Time-domain Bandwidth Compression Systems".
J.A.S.A. 42 p.348 (1967).

52. Talbert, L.R.
Groner, G.F.
Koford, J.S.
Brown, R.J.
Low, P.R.
Mays, C.H.
"A real-time Adaptive Speech Recognition System".
Stanford Electronics Lab. Technical Report No. 6760-1.

53. Tanaka, Y.
Okamoto, J.
"Syllable Articulation at the time when the Trailing Edges of Zero-Crossing waves make Random Fluctuation."
Osaka City University, Memoirs of the Faculty of Engineering, p.75 (1964).

54. Thomas, I.B. "The Significance of the Second Formant in Speech Intelligibility".
Biological Computer Lab., University of Illinois, Tech. Report No. 10 (1966).

55. Underwood,M.J. "Time Interval Statistics in Speech Synthesis: a Critical Evaluation".
Ph.D. Thesis, University of Keele, 1968.

56. Vilbig, F
Haase, K.H. "Theoretical Investigations to Reduce Harmonic Distortion in a Clipping Process".
J.A.S.A. 29 p.776 (A) (1957).

57. Zagoruiko,N.G. "Automatic Recognition of Speech".
S.T.L. Quarterly Progress & Status Report ,April, 1970, p.16.