THE EVALUATION OF THREE RELATED TECHNIQUES
FOR THE
STATISTICAL ANALYSIS OF CLIPPED SPEECH.

Thesis presented for the degree of Ph.D.
in the University of Keele by J.B.Millar.

MAY. 1968.

PREFACE.

The work began at the same time as the setting up of a speech research group in the Department of Communication. Practical experience and apparatus were acquired gradually, and it was not until the later part of this study that both were of a standard to attack the more crucial problems of speech analysis. The work has therefore been set at the level of basic research into one method of speech analysis rather than that of the solution of more urgent problems that are holding up progress towards artificial speech recognition.

The work described in this thesis comprises one half of a two pronged attack on the problem of the statistical analysis of clipped speech. The techniques which have been developed have in many cases been applied both to speech analysis, evaluated by attempts at artificial recognition, and to speech synthesis, evaluated by subjective testing. The interaction between these two approaches has proved invaluable in the development of ideas and in the interpretation of results.

ACKNOWLEDGEMENTS.

I would like to express my appreciation to all those people who have assisted in many ways, small and great, with the work described in this thesis.

I am especially grateful to Professor D.M. Mackay for his suggestion of the topic for research and his continued supervision and encouragement throughout the work. I am also grateful for advice on many aspects of the work given by other members of the Department of Communication. In particular I would like to thank my fellow student M.J.Underwood for his constant companionship throughout the work and his willing assistance in electronic design; A.W.Wright for his expert help in all computing matters, and Dr. W.A.Ainsworth for his helpful encouragement and for reading this manuscript. I am most grateful to my mother for typing it.

I would also like to thank my wife, Sheila, for her continued patience and encouragement throughout the past two years.

This work would not have been possible without financial support provided by the Science Research Council.

May, 1968.

ABSTRACT.

Techniques are described for the statistical analysis of clipped speech in terms of the time intervals between the zero crossings of the speech waveform or its time derivative. The only statistic of this kind which had been used prior to this study was the time interval histogram. This suffers from great variability and does not have any perceptual evidence to support its use. The present studies have highlighted some causes of its limitations, especially the perturbations due to the pitch of a voiced sound. Its usefulness as a discriminator of the continuant phonemes of the English language has been shown to be fairly restricted.

An analysis of the second order, or digram statistics of the time intervals is described. The use of this for discrimination of speech sounds has some perceptual support and has been found to yield some interesting differences in patterns between various speech sounds. A real time visible speech display based on this statistical measure has been developed employing simple analogue circuitry. A wide range of samples of speech sounds have been examined using this novel method.

A technique of pitch-synchronous analysis of the time-intervals has been developed to facilitate selective rejection of noise during voiced speech. This method has been found to reduce the effects of conventional noise and of pitch perturbations in the time interval statistics. A similar technique has been developed for use with the real time visible speech display.

A PDP - 8 computer was programmed to make quantitative measurements on the time interval statistics of vowel sounds, in order that the relative discriminability of the sounds analysed by the three techniques of histogram, digram, and pitch-synchronous analysis of the time intervals could be assessed.

The results of this analysis have shown that for a given speaker and a limited set of sample utterances greater discrimination between vowels can be achieved using digram rather than histogram statistics. No significant difference in discriminative power was found between histogram and digram analysis when the set of utterances was not restricted. Pitch-synchronous analysis has been found to reduce the dependance of the statistics on pitch but to give no corresponding overall increase in vowel discrimination.

It is concluded that these time domain techniques can form a useful component of the analysis needed for automatic speech recognition, but that other types of signal processing will be required in parallel if identification is to be reliable.

CONTENTS

page number

Introduction 1

I.1 The production of speech 1

I.2 The dynamics of speech production 3

I.3 Speech analysis 4

I.4 Perception experiments using synthetic speech 5

I.5 Perception experiments using distorted speech 6

I.6 Discussion on speech perception 9

I.7 Analysis based on the intelligibility of clipped speech 10

I.8 The development of analysis using waveform information 13

I.9 The need for a statistical approach 14

I.10 The basis of the present thesis 15

CHAPTER 1 Qualitative analysis of the first order statistics of the time intervals between zero-crossings of the speech waveform. 19

1.1 Description of electronics 19

1.2 The variation of T.I. statistics in time 27

1.3 The choice of phonemes and speakers 28

1.4 Time interval histograms of control stimuli 30

1.4.1 T.I. histograms of simple auditory signals 31

1.4.2 T.I. histograms of syntheticspeech 32

1.4.3 The asymmetry of glottal excitation 33

1.4.4 The constraints of glottal repetition 33

1.4.5 The effect of gross pitch changes on the T.I. histogram 34

contents continued) page number

1.4.6 Application to T.I. measurements on real speech 36

1.4.7 The effect of monotonic pitch variations 37

1.4.8 The mixing of synthetic formants 38

1.5 First order T.I. distributions of isolated speech sounds 40

1.5.1 The display of distributions 40

1.5.2 T.I. distributions of vowel sounds 41

1.5.3 The T.I. histograms of unvoiced fricatives 44

1.5.4 The T.I. histograms of voiced fricatives 45

1.5.5 The T.I. histograms of nasals 46

1.6 The effect of variation of the duration of compilation of T.I. histograms 47

1.7 The significance of histograms of time intervals delimited by unidirectional Z.C.s 49

1.8 Conclusion 53

CHAPTER 2 The qualitative study of second order statistics of the time intervals between zero-crossings of the speech waveform 56

2.1 Methods of measurement of second order T.I. statistics 57

2.2 Digram displays of control stimuli 63

2.3 Digram displays of continuant sounds 65

2.3.1 Digrams of vowel sounds 66

2.3.2 Subjective classification of steady state vowel digrams 67

2.3.3 Digrams of fricative and nasal sounds 68

2.3.4 The effect of duration of compilation of vowel digrams 70

contents continued) page number

2.3.5 The variation of the digram of utterances of similar sounds by a different speaker 71

2.3.6 The effect of differentiation of the speech waveform on vowel digrams 73

2.4 Some changes in instrumentation 76

2.5 The digram analysis of utterances of controlled pitch by two speakers 77

2.6 The removal of unwanted information from the digram display 81

2.6.1 Noise rejection methods used by other workers 82

2.6.2 Noise rejection in the present study 84

2.6.2.1 The amplitude modulated digram display 85

2.6.2.2 Pitch synchronous gating 88

2.6.2.3 Pitch detection 90

2.6.2.4 The extent of gating in each glottal period 91

2.7 T.I. analysis after p.s.g. of vowel waveforms 95

2.7.1 Experimental method 95

2.7.2 The variation of the gating period 97

2.7.3 The effect of p.s.g. on differentiated vowel digrams 100

2.7.4 The stability of p.s.gated vowel digrams under pitch variation 103

2.7.5 The stability of p.s.gated vowel digrams under speaker variation 103

contents continued) page number

2.8 Conclusions on the qualitative study of p.s.g. 104

CHAPTER 3 The introduction of quantitative measurements 107

3.1 The reduction of scope to the study of vowel sounds 107

3.2 The general form of quantification 108

3.3 Measurements of separation or difference between statistics 111

3.3.1 Correlation coefficient 112

3.3.2 Euclidean distance 113

3.3.3 The measurement of euclidean distance in the computer 116

3.4 Experiments using euclidean distance separation measurements on the T.I. statistics of vowel sounds 116

3.4.1 The effect of duration of compilation on euclidean distance 116

3.4.2 The measurement of separation on a relative scale 119

3.4.2.1 The estimation of likely confusions between vowels on comparison of their representative statistics 122

3.4.3 Experimental results - The sounds used in these experiments 122

3.4.3.1 Vowel discrimination using unidirectional Z.C. histograms 123

3.4.3.2 Vowel discrimination using differentiated speech 124

3.5 Limitations of this analysis owing to the unweighted distance measurement 125

3.6 Improvements in the definition of the analysis space and the method of difference measurement 126

3.6.1 Empirical measurements to derive the optimum shape for the analysis space 127

3.6.2 The measurement of 'equiprobable bin divisions' 128

contents continued) page number

3.6.3 The results of equiprobable bin measurements 129

3.6.4 Methods to overcome the problem of hypersphericity 131

CHAPTER 4 Quantitative analysis using a vowel recognition system 135

4.1 Introduction 135

4.1.1 Comparison of statistics using a weighted euclidean distance 135

4.1.2 Outline of the recognition system 136

4.1.3 Experimental aims 136

4.2 Description of programs 137

4.2.1 Computation and storage of reference statistics 137

4.2.2 The computation of mixed reference sets 138

4.2.3 Comparison and recognition programs 138

4.3 T.I. analysis of vowels by three speakers 141

4.3.1 Experiments using same speaker as reference 141

4.3.2 Experiments using mixed speaker reference 143

4.3.3 The analysis of confusions in these experiments 144

4.4 Variation in recognition scores between two program runs 148

4.5 A comparison of recognition results using histogram and digram analysis 149

466 Further analysis by recognition experiments 153

4.6.1 The variation of T.I. resolution 153

4.7 The quantitative investigation of pitch synchrous gating 157

4.7.1 The results of p.s.g. on recognition scores 158

4.7.2 Pitch dependence 160

4.7.3 The effect of p.s.g. on outside-reference sounds 160

contents continued) page number

4.7.4 Review of the effect of p.s.g. on the recognition of the utterances of J.B.M. 161

4.7.5 The effect of ps.g. on the recognition of utterances by M.J.U. 163

4.7.6 Conclusions on the use of p.s.g. 164

4.7.7 Recognition scores obtained 165

4.8 Recognition based on the analysis of the T.I.s of the differentiated waveform 167

4.8.1 Within-reference utterances 167

4.8.2 Outside-reference utterances 167

4.8.3 Pitch synchronous gating 167

4.8.4 Conclusions on the T.I. analysis of differentiated speech 168

4.8.4.1 Check on the variance of each bin's contents over the twelve vowels 169

4.8.4.2 Comparison of results with subjective recognition scores 170

4.9 The effect on the recognition scores of grouping the vowels 170

CONCLUSIONS AND FUTURE PROSPECTS 174

Conclusions from the present study 174

T.I.s of speech waveforms and their histograms 174

Comparison of the quantitative analysis of T.I. histograms with that of Bezdel and Chandler 176

The uses of T.I. histogram analysis 177

The digram measure of speech waveform T.I.s 178

Evaluation of the use of pitch synchronous gating 182

Comparison of artificial and human recognition of isolated vowels 184

Parallel work 185

The way forward 188

Appendix 1 Clipping amplifier circuit 191

Appendix 2 A - Parallel time-base system 192

B - Analogue shift register stage 193

Appendix 3 Magnetic recording machines 194

Appendix 4 The computational facilities available during the present research 196

Appendix 4B The compilation of T.I. statistics in the PDP-8 199

Appendix 5 The results of subjective testing of utterances used in in analysis and auto matic recognition experiments 201

Appendix 6 Application of pitch synchronous gating to the intervalgram display 204

Appendix 7 The mathematics of mixing the reference sets 205

Appendix 8 The use of the digram display as 'visible speech' 209

References 211

LIST OF ABBREVIATIONS USED IN THIS THESIS.

A.S.R.	Analogue Shift Register.
C.A.T.	Computer of Average Transients.
C.R.T.	Cathode Ray Tube.
c/s.	cycles per second.
db.	decibel.
E.D.	Euclidean Distance.
msec. ms.	millisecond.
µsec. µs.	microsecond.
P.B.	Phonetically Balanced.
p.s.g.	pitch synchronous gating.
p.s.gated.	pitch synchronously gated.
PSG + 4.	the first four milliseconds of a glottal period.
PSG - 2.	all but the last two milliseconds of a glottal period.
sec.	second.
T.I.	Time Interval between the zero-crossings of a waveform.
Z.C.	Zero-crossing.
+ ZC.	positive going Z.C.
- ZC.	negative going Z.C.
± ZC.	positive and negative going Z.C.s.

INTRODUCTION.

The recognition of speech is but one of the many functions of which man alone is capable but which he has thus far been incapable of fully understanding or imitating artificially. It is one of the processes in the human brain which has proved to be remarkably efficient in the most adverse circumstances, when the perception of other auditory signals of similar complexity would be impossible.

It is clear that the understanding of speech is deeply involved in the physiological and psychological development of all human beings. For this reason it is considered that the human analysis system which has defied imitation is likely to do so until a far greater understanding of the functions of the elements and larger organisation of the brain has been attained, and the links between physiological and psychological observations have been established.

However, pressing humanitarian and technological problems of giving aid to the deaf and dumb and of communicating with digital computers which have 'brain-like' information handling capabilities, have caused partial answers to the basically unsolved problem of artificial speech recognition to be sought and implemented.

I.1. The production of speech.

The method of speech production is important to all research into the analysis or synthesis of speech. It is this human process which the latter field of study is attempting to imitate, and on which the former bases many of its methods. It seems likely that there is a complex link between speech perception and speech production in both directions. The disturbance of normal auditory feedback has been shown to impair speech production (36, 18). It seems likely that tactile and proprioceptive feedback also play an important part in speech production. This is evidenced by difficulty in speech production encountered while suffering the effects of anaesthesia of part of the articulatory system. These modes of feedback may also influence speech perception if imitation is induced in the listener. The basics of speech production are therefore relevant to our present study and to describe them provides a useful opportunity to review the terminology of speech description.

The acoustic power for speech production is created by muscular contraction within the chest and abdomen causing an excess internal pressure (59). The gradual outflow of air during a breath cycle is caused by contraction of the chest muscles, while pulsed abdominal contraction modulates the air flow at the syllabic rate of speech.

This modulated flow of air is further modified in a large variety of ways in the vocal tract. (See fig. I.1) These modifications are described in great detail by Flanagan (21). The glottis or vocal chords at the lower end of the tract consist of two lips of ligament and muscle which can be stretched tightly across the tract so as to close it or can be relaxed so as to allow the passage of air. The other articulators of speech are the lips, velum, tongue and jaw. The lips act in a similar way to the glottis in forming a closure or partial closure at the upper end of the tract, known as the oral cavity. They are however far more flexible as they are capable of much greater movement relative to the cross sectional area of the tract. The velum acts as a valve either opening or closing the rear entrance to the nasal cavity of which the nostrils form the front entrance. The tongue and jaw, in conjunction with the lips and cheek muscles, modify the size, shape and partitioning of the oral cavity.

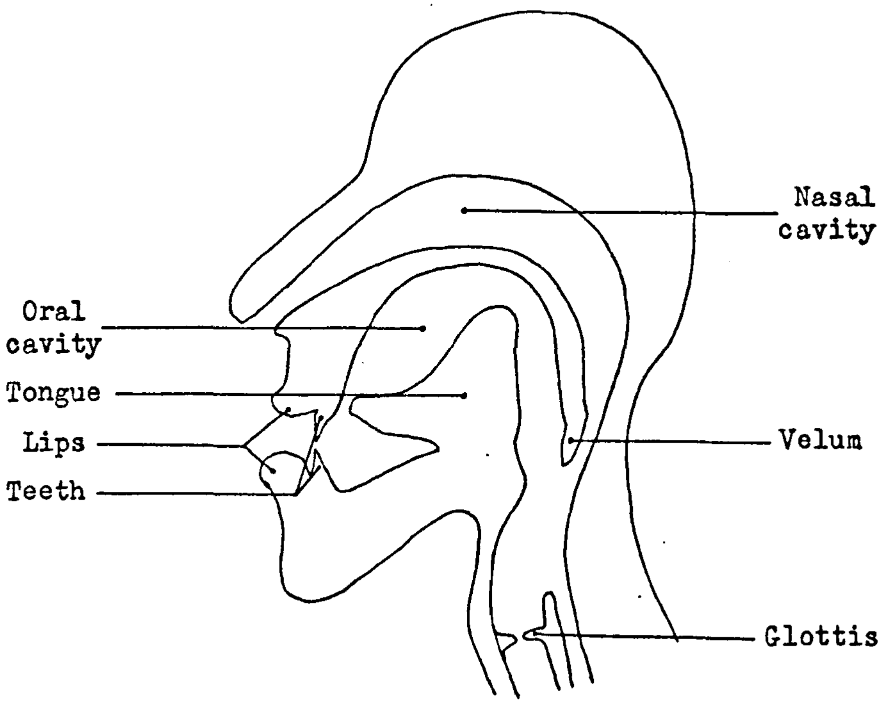


Fig. I.1 Speech articulatory organs.

These articulators are capable of rapid movement enabling the successive utterance of sounds which are perceived as the mutually exclusive signals which compose the repertoire of the speakers language. These sounds have been classified by linguists into categories called phonemes (7). The phoneme is defined as the smallest linguistic unit which can cause a change in meaning in a given language. A set of such units can be defined for any language.

I. 2. The dynamics of speech production.

The production of voiced sounds originates with the tension in the glottis. When a certain tension is achieved the combination of the Bernoulli force caused by the passage of air, and the state of tension, causes the glottis to vibrate allowing short bursts of air into the vocal tract at a frequency controlled by the subglottal pressure and the glottal tension. This frequency determines the pitch of a sound which is seen as a periodic variation of amplitude in the speech wave and can be objectively measured as the glottal period or the fundamental frequency.

The arrival of the air pulses in the tract sets up resonances in the various cavities between the articulators. The position of the articulators determines the colour of the voiced sound. In pure voiced sounds, the vowels, semivowels, and dipthongs, this colour is of utmost importance for their discrimination. This colour can be indicated by the time period or frequency of each resonant cavity. When the acoustic wave is analysed on a frequency scale, the colour is revealed as several bands of energy which are termed formants. These formants are given numerical subscripts according to their rank ordering from low to high frequency.

When the glottis is more relaxed air passes freely and various consonant sounds are produced. The simplest consonants from the production point of view are those which are caused by a single constriction of the vocal tract causing a continuous noise-like sound. These fricative sounds are produced by constrictions at the glottis itself, and combinations of tongue-tip, teeth and lips. A complex sound known as a voiced fricative can be produced by a combination of these latter constrictions and a vibrating glottis.

Stop consonants are a class of sounds which are caused by a build up of pressure behind a complete closure of the tract at some place or other, causing a silence, followed by a release of this pressure causing a short sharp burst of noise.

Nasal sounds are a subset of the voiced sounds in which the velum opens the nasal cavity and the tongue or lips close the oral cavity. The sound is emitted from the nostrils via the nasal cavity which typically has a low intensity high frequency resonance.

I.3. Speech Analysis.

The only general purpose speech recogniser which can operate with great accuracy is the human ear-brain system. For this reason most methods of speech analysis have been modelled on insights into how this system works. These insights can be gained from experiments on two levels: objective experiments to measure mechanical and electrical activity in the ear and associated nervous system, and subjective experiments to measure the perception of auditory stimuli. Some experiments of the objective type are not usually possible on the human auditory system for ethical reasons. Experiments on the electrical activity associated with auditory stimuli have been done mainly on cats and guinea pigs, with the result that the workings of auditory systems in many ways similar to the human system are partially understood. A recent review of this work has been given by Whitfield (67). The mechanical activity in the middle and inner ear of animals and humans has been studied by Bekesy (4). His work has provided the basis for an understanding of the relationship between the electrical activity observed in the lower regions of the auditory nervous system and the characteristics of the acoustic stimulation. In the study of speech these experiments form a background reference inasmuch as speech is typical of all auditory signals.

Experiments of the subjective type have been designed in several ways. Two methods which will be reviewed in more detail in subsequent sections involve the measurement of speech perception by the classification of speech sounds by subjects. One way is to use speech which has been synthesised from acoustic parameters which are thought to be important for intelligibility, and the other way is to use natural speech which has been distorted, to remove acoustic information which is not thought to be important. The subjects' classification of speech stimuli of these kinds can give important information concerning the acoustic features which are necessary for intelligibility. A further method of measuring a subject's perception of speech sounds, which uses the procedure of stimulus imitation, has been reported by Kozhevnikov and Chistovich (32). An objective analysis of the similarities and differences of the sound produced by the subject in imitating the stimulus sound was used to study the way in which speech sounds are perceived by the human ear-brain system.

I.4. Perception Experiments, using synthetic speech.

All experiments reported in the literature on the synthesis of speech are based on frequency analysis. Work on time domain synthesis has been conducted by Underwood (66) in parallel with this present study.

Synthesisers using the vocoder principle described by Dudley (15) have been built and produced speech of high intelligibility. Such experiments reveal that a quantised description of the speech spectrum is sufficient for intelligibility. An attempt to further reduce the bandwidth for speech description was made by Lawrence (35) who constructed a synthesiser to operate on the clear visual features of the spectrographic display of speech made famous by Potter, Kopp & Kopp (52). Detailed work on the synthesis of speech using these parameters has been done at the Haskins Laboratory (11, 14, 37) and by others in various laboratories. These experiments have revealed that highly intelligible speech can be produced when described by such parameters as the frequency and amplitude of the lowest three or four formants, high frequency noise and larynx frequency.

I.5. Perception Experiments using distorted speech.

Most of the experiments using distorted speech have arisen from the problems of the communications engineer who wishes to economise on the bandwidth of a speech transmission channel. To do this he requires to decrease the frequency range which he is to transmit and increase the power of the signal to be transmitted. These two processes have led to two types of distortion experiment. Firstly there is distortion of the frequency dimension by low pass, high pass or band pass filtering or more complicated processes, and secondly distortion of the amplitude of the waveform by non-linear peak clipping.

Frequency distortion.

An experiment of this form was done by French and Steinberg (24). They measured the intelligibility of speech as successively more severe high pass, then low pass filtering was applied. Kryter (33) used bandpass filters which were moved through the speech band. A band pass filter whose band width was successively widened to include more of the speech band was used by Egan and Wiener (17). Similar work to the three studies mentioned has been conducted by several workers and a large degree of agreement on areas of the frequency dimension that are important for speech intelligibility has been achieved.

Amplitude distortion.

The classical experiment on amplitude distortion of speech was done by Licklider and Pollack (40) taking the earlier work of Licklider (39) to its logical conclusion. They showed that the intelligibility of speech is not greatly reduced when its waveform is subjected to 'infinite clipping'. This process is the reduction of the speech waveform to "a succession of rectangular waves in which the discontinuities correspond to the crossings of the time axis in the original speech signal". Licklider (41) further investigated the effect of distortion on the time axis of already infinitely clipped speech. These papers will be considered in more detail.

Licklider and Pollack used three basic circuits which were linked in different combinations for the processing of speech. The circuits were a differentiator, an integrator and an infinite clipper. Their results indicated that the preclipping processing of speech had a great effect on the intelligibility of the clipped waveform but that this intelligibility was nearly independent of further processing after clipping. Differentiation or integration without any clipping gave similar articulation scores to the undistorted waveform, differentiation prior to clipping gave a very slightly smaller articulation score, simple clipping gave significantly less and integration prior to clipping gave a very small articulation score indeed.

The effect of post clipping integration or differentiation was shown to be very small. An extension of this work to cover all possible combinations of post-clipping differentiation with pulse shaping is reported by Ainsworth (2) giving the same result.

In a later paper Licklider (41) investigated the effect of distortion in the time dimension by quantising the time intervals between zero crossings. He found that a reduction from 20,000 quanta per second to 10,000 quanta per second corresponded to a drop in articulation score from 96% to 91% for pre-differentiated clipped speech. Further reduction caused considerable deterioration of the articulation score. It is interesting to note that the zero-crossings of the clipper output in Licklider's experiments were not accurately described as the zero crossings of the original speech wave. There will have been a shift in the zero level of the signal due to a.c. coupling in the clipping electronics and the magnetic recording machine used. It is not clear what difference d.c. coupling throughout would have had on the articulation score. It would of course vary slightly from one waveform to another.

Tanaka and Okamoto (63) investigated some similar forms of time interval distortion. They allowed the negative-going edge of a differentiated and clipped speech waveform to vary randomly in time subject to various maximum delays. Syllable articulation dropped from 98% for a 1 µsec. maximum, to 84% for 10 µsec, and 55% for 100 µsec. This decrement is much larger than that found by Licklider (41). A difference is that Tanaka and Okamoto used a 'slice level' about which to clip rather than zero, in order to reject noise, 40 db. below the speech level, in silent periods of the speech. In relation to this, they showed the fall of syllable articulation as the slice level was moved away from the zero level. Syllable articulation decreased from ~90% to ~75% when the level was lifted from -50 db. to -30 db, and to 45% when the level was raised to -20 db.

It was concluded that the experiments described above showed that the amplitude information in a speech waveform is of little value in conveying intelligibility, but a certain accuracy of definition of the time intervals between true zero crossings is important. The acoustic patterns of the speech which remained after clipping were the temporal sequence of zero-crossings of the waveform and the distorted spectral pattern. It is possible that cues for the high intelligibility retained by clipped speech are contained in either or both of these patterns.

I. 6. Discussion on speech perception.

The question, "How does the human analyse and subsequently recognise sounds of speech?", is still an open question after over half a century of research into this topic. It has been seen in this brief review of various approaches to the problem that much ground has been covered in discovering what is, and what is not important for intelligibility to be retained. Experiments have suggested that such different measures as a description of the short term frequency spectrum, or of the time pattern of zero crossings could be major explanations.

It seems likely that the human listener perceives speech as a result of several separate auditory cues. Not all are necessary but no one alone is sufficient. Some of these may best be described in the frequency domain, others in the time domain, and others as relative features in either domain. It should, however, be noted that human speech perception involves far more than acoustic analysis (25). The human listener can supplement auditory cues by visual ones from the speaker if they are available to him. If they are not, he still has a vast experience of speech constraints to cope with a change of speaker or narrow his choice when a single speaker is speaking. He also has experience of linguistic constraints which enable him to accurately guess unrecognised sounds.

I.7. Analysis based on the intelligibility of clipped speech.

The reason for the high intelligibility of clipped speech has been pondered by several workers. Fourcin (23) has investigated the extent to which the clipped speech retains the spectral quality of the original unclipped speech wave. He found that formant frequencies are still apparent if their intensities are within 5 db. of each other. If one formant is more intense than another by more than 5 db. it 'captures' the other by dominating the zero-crossing pattern of the waveform. Thomas (64) has recently shown that the experiments of Licklider and Pollack and several other distortion experiments in the frequency dimension indicate that preservation of the second formant is crucial for the intelligibility of speech. The superior score obtained for speech which has been differentiated before clipping is the major point in his argument from these clipped speech experiments. The +6 db./octave frequency emphasis will cause the higher formants to have more effect on the zero-crossing pattern and therefore be retained in the spectrum of the clipped signal. He reports high scores, only slightly less than those obtained by Licklider, for clipped speech which has been bandpass filtered prior to clipping, to admit only the range of the second formant.

Other workers have based their analysis on a direct measurement of the time intervals between zero-crossings. Sakai and Inoue (55) assumed that "it is reasonable to expect" that a histogram of these time intervals would enable one to "extract the indispensable factors which contribute to articulation." They produced histograms for five Japanese vowels using both normal and differentiated speech. In each case they found one or two peaks in the distribution and found that they could recognise the vowels by the positions of the peaks. They also produced histograms for consonants and showed how they varied when followed by the vowels already mentioned. The histograms of the unvoiced fricatives /s/ and /ʃ/ were shown to be independent of the following vowel but the stop consonant /k/ was seen to produce histograms very similar to those of the vowels which followed it. Histograms of /m/ and /n/ were found to be dominated by the effects of the fundamental frequency and a high pass filter was used to remove these. Their general philosophy was to avoid the use of filters in order to retain the time resolution that zero-crossing measurements afford.

Sakai and Doshita (56) have developed a more comprehensive speech recognition system. The basis of operation has however shifted,from an analysis of the histograms of the time intervals between zero-crossings,to the use of zero-crossing information from the outputs of two frequency filters. These filters were chosen to match the range of the first two formants. Their analysis is therefore on the basis of the intelligibility of two formant synthetic speech although they have not used narrow band filters to determine the formant frequencies.

Bezdel and Chandler (5) were more cautious in their approach to the analysis of zero-crossings. They said "while it is clear that there is sufficient information left for the human to recognise clipped speech, the extent to which zero-crossing information will allow machine recognition of speech has still to be established."

They analysed four English vowels /ʌ/, /ɛ/, /i/, and /u/, and the dipthong /ɒu/. They performed a qualitative analysis using a histogram with 16 bins, but reduced this to 6 bins when obtaining quantitative results using a recognition test. Histograms were compared for similarity using several decision functions, one of the aims of the study being to find the best decision function. Using a weighted Euclidean distance decision function scores of 97%, 95% and 94% were obtained for women, men and mixed groups of speakers. They further analysed clipped speech by prefiltering into two bands 300 - 1000 c/s and 1000 - 3400 c/s prior to clipping. Noticeable improvement was found in the recognition of some sounds.

Chang et al.(8) also used a measure of the time intervals between zero-crossings in order to produce a form of visible speech similar to that of the spectrogram. They produced voltages related to the time intervals between zero-crossings and displayed the variation of these voltages in real time on an oscilloscope with a long persistence phosphor. This work was essentially an attempt to reproduce the visible speech patterns of the spectrogram using zero-crossing information as this could be obtained with very simple circuitry. They showed that the intervals between the zero-crossings of the original speech wave emphasised the first formant of vowel sound whereas those of a differentiated form emphasised the second formant. A superposition of the patterns achieved by both methods gave an 'intervalgram' depicting the variations seen in the two formants. They found that the many extra tracks across the screen not related to formant movements could be largely removed by combining two or four adjacent intervals and plotting their mean value.

This combined interval measure is tending towards a measure of zero-crossing rate which many workers including Chang (9) have used as a measure of frequency.

I.8. The development of analysis using waveform information.

Without proposing that the information contained in the patterns of time intervals between zero-crossings of the speech wave provide the whole answer to the speech recognition problem, it is proposed that analysis of this information has some singularly attractive features which merit further investigation.

The study of the speech waveform has recently been urged by Dudley (16). He points out that many of the 'parameters' extracted by extensive filtering systems can be simply extracted visually in the time domain. This was also the philosophy of Chang (8), and more recently of Reddy (54), who both maintain that these 'parameters' can be extracted by simple circuitry or simple computer routines, without the constraints imposed by time-frequency uncertainty.

A large number of the features of the waveform can be extracted from a measurement of the zero-crossings of various simple transforms of the original waveform. When a differentiating circuit is fed with a speech wave, the zero-crossings of its output will give the timing of the local maxima and minima of that waveform. This enables measurement of the relative positions of the peaks in time and detection of high frequency ripple even if it is at low amplitude relative to the major components of the waveform. Similarly if an analogue integrator is used, the time course of asymmetries in the waveform can be measured. The zero-crossing pattern of the integrated waveform was found by Licklider's perceptual experiments to carry very little intelligibility, but it does give information as to the shape of the waveform and as such it could be useful in resolving ambiguities that may arise, or indicate where noise is most likely to disturb other zero-crossing measurements. The pattern of zero-crossings of the original speech wave itself of course provides information as to the dominant acoustic component of the sound. It is clear that any combination of these preprocessing conditions can be achieved by mixing in variable ratios the outputs of two or more preprocessing circuits before zero-crossing patterns are measured. A parallel combination of features extracted by the various preprocessing circuits could be achieved by analysis of the zero-crossing patterns after clipping the outputs of two or more such circuits.

The possibilities are very great and it was not possible to treat all of them in the studies to be described.

I.9. The need for a statistical approach.

The need for a statistical approach to the qualitative or quantitative measurement of speech sounds has become accepted over recent years. The advent of quantitative measures has certainly extended the use of such measurements (51) although the usefulness of statistical models has become apparent in the more qualitative areas of linguistics and phonetics. Herdan (28) reviews the use of mathematical statistics in the relationship of phonemic and phonetic categories. He quotes, in translation, a simplified picture presented by Zwirner and Zwirner (69) "Since such norms (phonemes) cannot be realised by the speech organs in exactly the same way twice, the transition from phonology as a science of the norms to phonetics as their realisation in speech must be of a statistical nature, such that the variations of a sound are distributed around their average according to the Gaussian law (normal curve), and these averages are what correspond to the norms". Herdan points out that Trubetzkoy cannot reconcile the phoneme with a 'phonetic average' of the realisations of that phoneme, as it is possible for some phonemes to have several discrete phonetic averages depending on their contextual circumstances.

It is clear that in proposing a statistical analysis of speech on a phonemic level either multimodal distributions must be accepted or a less wide ranging analysis must be attempted. The following study on isolated sounds is attempting to work within the constraints of unimodal phonetic distributions. In the investigation of zero-crossing measures as partial descriptions of such distributions it is clear that statistical measures must be made in both the time interval dimension and the ongoing time dimension. This is necessary as the time intervals between zero-crossings are measurements on a time scale which is microscopic when compared with the time scale of phonemic variations.

I.10. The basis of the present thesis.

In this thesis the development of studies conducted on statistical measurements of the time intervals between the zero-crossings of speech waveforms will be described. A limited analysis has been done on a few

English vowels by Bezdel and Chandler (5) and a wider ranging analysis on Japanese phonemes has been done by Sakai and Inoue (55). Both these studies have been based on histograms of the time intervals and some success has been reported (see section I.7.) However, the assumptions of Sakai and Inoue that the intelligibility of clipped speech suggests that the information contained in such a histogram is sufficient to extract "the indispensable factors which contribute to articulation" cannot be justified. There is no evidence that the perception of a time pattern of pulses is uniquely related to the characteristics of their first order interval distribution. Contrariwise it seemed eminently reasonable that a change in the ordering of such a time pattern, so that the first order distribution remains constant, would cause a change in the perceived auditory sensation. This was shown to be true by some preliminary experiments conducted by M.J.Underwood (66). It was an aim of this present study to investigate higher order statistics of the time interval patterns of speech to see if they provide more information relevant to the separation of speech sounds that are perceived as different by human listeners.

The time domain of the speech wave is dominated by information concerning the pitch of voiced sounds. This feature is very dependent on the speaker and his emotional state rather than what he is trying to communicate. It is appreciated that on a semantic level pitch has a considerable effect in the creation of distinctions of meaning ; but on a phonemic level in the English language, distinctions are not pitch dependent. A practical aim of this study was to overcome some of the inherent problems of time domain measurements in speech by taking advantage of the redundant nature of its waveform.

At the suggestion of Professor D.M.Mackay it was decided to evaluate the following techniques of measurement of the time intervals between zero-crossings,with a view to the discrimination of speech sounds, on a wide range of English phonemes.

1. Simple histogram analysis.
2. Second order, 'digram' analysis.
3. The further use of speech wave redundancy.

The structure of this thesis is as follows.

In chapter one the basic apparatus to examine time interval histograms is described and results of histogram measurements are presented. The way in which these zero-crossing measurements reflect distinctive waveform features are emphasised as well as histogram features which allow the original sounds to be distinguished.

In chapter two the development of apparatus to measure second order statistics is described and results presented. These results cover a wider range of speech variations due to different speakers and differently pitched utterances being used. Methods of rejecting unwanted information in these statistics are reviewed and a new method incorporating speech wave redundancy is employed.

In chapter three the growing need for a quantitative measurement on these statistics is dealt with and a suitable one employed. Limitations of this measure are found and modifications to overcome them and take further advantage of the redundancy of the signal being analysed are discussed.

In chapter four quantitative measurements on all three proposed techniques are made and more general estimates of their usefulness assessed.

In the final chapter conclusions on the usefulness of the three techniques investigated are presented. The way forward in the application of these techniques to artificial speech recognition is suggested.

CHAPTER ONE. - Qualitative analysis of the first order statistics of the time intervals between zero-crossings of the speech waveform.

Introduction.

The aim of the studies which form the basis of this thesis was to evaluate the usefulness of the statistics of the time intervals between zero-crossings (Z.C.s) of the speech waveform (to be referred to as T.I. statistics) in discriminating between speech sounds. The philosophy dominating early thought on this work was that information extracted from such statistics could be used as a partial tool for this discrimination. However the T.I. statistics alone have been used in the present studies whose central enquiry has become, "What questions about speech sounds can be simply answered by reference to these statistics?".

It is the purpose of this chapter to explore the first order T.I. measurements of speech; to illustrate any relationships to the more conventional frequency domain of measurement and to evaluate the possibilities of statistical measurements being used to discriminate speech sounds.

It was not clear from the literature what the stability of the T.I. statistics of various speech sounds is like, or how the statistics are effected by pitch in voiced sounds. The initial aim was to obtain a visual display of the T.I.s that would provide some statistical information on which measurements of stability and pitch dependence could be made.

1.1 Description of Electronics.

The electronics required to perform this analysis can be considered in four parts. Firstly there is the circuitry in which the speech waveform is processed prior to clipping; secondly the clipping amplifier itself; thirdly the post-clipping processing of the clipped waveform which is needed to define the T.I.s by the intervals between pulses and fourthly circuitry to present the measured T.I.s in a statistical form.

1.1.1 The preprocessing of the speech waveform.

The forms which this could take can be divided into two categories. There is preprocessing which will eliminate unwanted information and noise, and that which will extract specific features from the waveform to be analysed in parallel with or instead of the original waveform.

The former could take the form of a limitation of the bandwidth of the speech signal to that which is known to contain important speech information. Extensive studies on the spectral analysis of speech have been carried out by Potter, Kopp and Kopp. (52). No such bandwidth limitation was used in the early experiments although simple R - C integration was used to limit the high frequency gain when differentiation of the waveform was used. The use of sharper low pass filtering for this purpose will be discussed later.

As the system to be evaluated is essentially a temporal analysis of the speech waveform, selective filtering which involved temporal smoothing of the waveform was avoided.

Another form of noise which could be eliminated at this stage is that of spurious Z.C.s which are not truly descriptive of the major features of the waveform. The elimination of these is discussed in the next chapter. The initial approach to this problem was to measure every T.I. accurately and rely on a high signal to noise ratio to render the effect of noise negligible. The signal to noise ratio for the microphone and pre-clipping circuitry was 52 db.

Two forms of preprocessing to extract features from the waveform were used. A method of locating the start of each glottal period, defined as the position of the maximum amplitude lobe of the waveform within a certain period, was used in a later stage of these studies and will be described in a later chapter. The second process of feature extraction was differentiation of the waveform with respect to time. As seen in the previous chapter the clipping of this waveform will define the intervals between the maxima and minima of the original waveform. Differentiation can also be considered as a frequency shaping of + 6 db. per octave together with a phase shift of $-\pi/2$. This preprocessing was included,as other workers, notably Licklider and Pollack (40) and Ainsworth (2), have shown that the intelligibility of clipped speech is enhanced significantly if the waveform is differentiated prior to clipping. A further variant of preprocessing was made possible by feeding the normal speech signal and the output of the differentiator to either end of a potentiometer. The signal at the variable slider was a mixture of the two signals in a ratio dependent on its position.

The circuit used to provide the differentiation was a long-tailed pair amplifier with capacitive input and resistive feedback.

The d.c. operating position of this amplifier was adjustable to give equal positive and negative peak limiting. This was found to be necessary to avoid distortion of the Z.C.s when high frequencies of appreciable amplitude were used during testing with a sine wave. However the probability of these combinations of frequency and amplitude are quite low as shown by the long term speech spectrum given by Miller (47).

1.1.2 <u>The clipping amplifier</u>.

The most accurate and stable circuit that was built to act as a 'clipper' was a long-tailed-pair symmetrical amplifier. A diagram of the basic stage is given in appendix 1. As this circuit is basically a linear amplifier and not a trigger circuit and each stage is a.c. coupled to the next one, the clipping level was maintained by ensuring that peak clipping of a sine wave was equal for positive and negative excursions. The a.c. coupling used had a time constant of over 100 msecs. This was far greater than the time constant inherent in the tape recorders used. Four stages were required to provide sufficient gain to achieve fast positive and negative going edges to drive an Eccles Jordan bistable circuit from the signal of lowest frequency and amplitude likely to be encountered. The bistable was included as a waveform shaper to ensure that the slope of the positive and negative going edgesof the clipped waveform were independent of the amplitude/frequency characteristics of the original waveform. The four stage amplifier was extremely stable despite the fairly severe temperature changes that it experienced in the temporary buildings in which the early part of this research was done.

1.1.2.1 The setting up of the clipping amplifier for use with speech.

The initial setting up of the clipping amplifier was performed using a sine wave input. It was realised however that such a signal would not give a true impression of the performance of the amplifier when used on speech waveforms. On the basis of the long term spectrum of speech given by Miller (47), 500 c/s was chosen as the frequency most likely to dominate the average speech waveform, and therefore its Z.C.s. A 1 volt peak to peak 500 c/s sine wave was used to typify the most probable speech wave input. An output of unity mark to space ratio was achieved for this input by adjustment of each stage to give equal positive and negative peak clipping. The mark to space ratio of the output remained unity for all frequencies within the speech bandwidth at this amplitude. At an amplitude of 0.1 volt peak to peak the mark to space ratio equalled 0.8.

The setting up of the clipping amplifier using a sine wave was considered as the initial check that accurate T.I.s of the input wave were being measured. It was realised that any Z.C. perturbation due to the a.c. coupling of the amplifier stages would not be seen using such a regular signal, and also that the effects of rapidly changing amplitude of individual waveform lobes as found in speech could not be observed.

A further estimate of the clipping accuracy was made by simple comparison of the clipped waveform and unclipped waveform of several vowels. The comparison of these waveforms was facilitated by the application of the clipped waveform to the Z - modulation of the oscilloscope. In this way a regular or even continuous check on the accuracy of the clipping amplifier could be made. The error in Z.C. measurement observed when using low amplitude sine waves was not evident when similar amplitude lobes of the speech waveform were examined. It was therefore assumed that the irregular mark to space ratio experienced with a low amplitude sine wave was due to drift caused by a change in the long term energy of the signal rather than its instantaneous amplitude. The input to the clipping amplifier was thereafter controlled to give a constant average voltage input.

1.1.3 Post clipping processing.

The rectangular output waveform of the clipping amplifier was used to drive pulse circuitry which provided pulses at the Z.C.s of the pre-clipping waveform. Two simple R - C differentiators and rectifiers operated on the out of phase outputs of the Eccles Jordan circuit producing positive pulses at the positive and negative Z.C.s separately. A switch enabled either or both these sets of pulses to be transmitted to a stage of pulse shaping. The resultant waveform was a train of pulses of defined amplitude, duration and rise time at either the positive going, negative going or every Z.C. of the pre-clipping waveform.

1.1.4 Apparatus to present the T.I.s in statistical form.

There are three features of the measured T.I.s that were considered important in the study of their first order statistics. These are the lengths of the intervals, their probability of occurrence and the variability in time of intervals of very similar length which have a high probability of occurrence in a particular section of speech waveform. Two methods were used to derive the first order statistics; one of them displayed all three features whilst the other sacrificed the third feature for greater accuracy in the first and second. The former method resulted in a display very similar to that used by Chang et al.(8). A real time display of each interval was made on an oscilloscope as the vertical displacement of a spot from a horizontal line. The horizontal time axis was selected by the time base controls of the oscilloscope used. The frequency of occurrence of any one interval length or group of interval lengths was given by the brightness of the spot at a point in time or by visual integration of the brightness of a succession of spots in a horizontal plane. This depended on the speed of the time base selected.

The display was achieved in the following way. Z.C. pulses were caused to arrest the charging of a capacitor, discharge it and then allow it to recharge until the next Z.C. pulse occurred. As a constant voltage source was used to charge the capacitor, the charge on the capacitor gave an exponential transform of the T.I. that was occurring. By causing the discharge of the capacitor with the trailing edge of the Z.C. pulse, the pulse itself could sample the voltage on the capacitor by modulating the brightness of the C.R.T. to whose Y amplifier the voltage on the capacitor was presented.

1.1.4.1 A more sophisticated way of compiling and displaying the first order statistics of the T.I.s was achieved by the use of one of three preprogrammed modes of operation of the Mnemotron Computer of Average Transients (hereafter called the C.A.T.). The major raison d'etre of this computer in the department was the averaging of electroencephalographic responses, but it was capable of providing a visual display of the T.I. distributions and was used fairly extensively for this.

The C.A.T. has a core store of 400 20 bit words. An address pointer can scan the 400 word addresses at a variable linear rate using the internal analysis sweep,or at any other rate less than 800,000 words persecond using an external sweep controlled by pulses at the address advance input. The arrival of a pulse at the address reset input causes the contents of the address which is currently being scanned to be incremented; the address pointer returns to the first address,and after a dead time of 50 µs. restarts its address scan. Clearly the length of the interval between the pulses applied to the address reset input is proportional to the number of addresses that are scanned before a pulse arrives and the process is restarted.

The C.A.T. is provided with a C.R.T. display on which is displayed the contents of the core store. This is in the form of 400 spots which are lifted from a base line according to the number of counts in each word. This display is clearly in the form of a histogram. If the rate of scanning the addresses is increased to its maximum value,by use of an external pulse generator,the histogram approaches a continuous distribution.

1.1.5 Magnetic Recording of sounds.

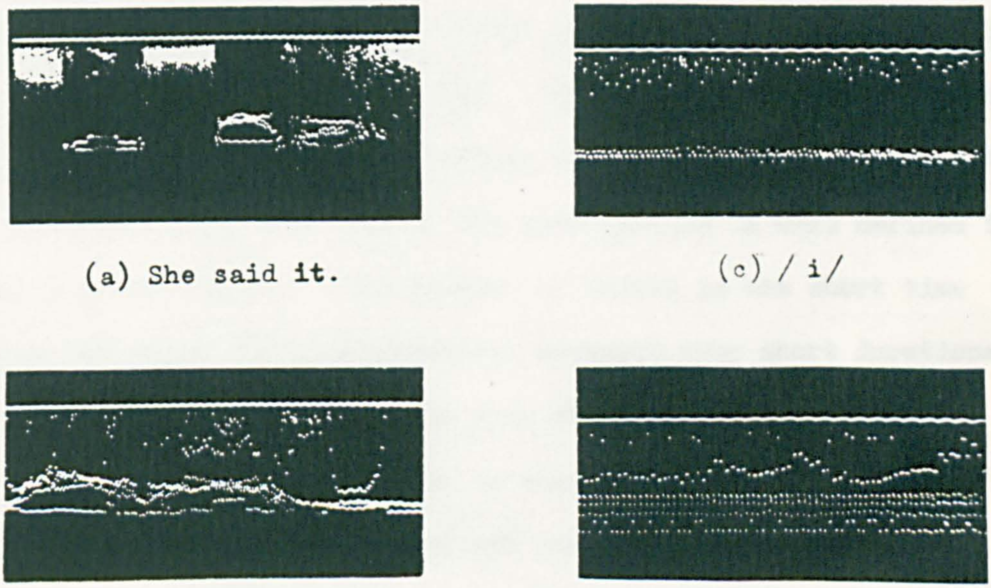
Two recording machines were used at different times during this work. They were a Siemens 12 stereo tape recorder and a Bell and Howell Language Master. Descriptions of these machines and their use will be found in appendix 3.

1.2 The variation of T.I. statistics in time.

The first method of displaying T.I. information is basically the same as that used by Chang et al to provide similar visual information to that available from the spectrogram. It was used in the present study to give the experimenter experience of the temporal characteristics of T.I. distributions as parameters of speech sounds. Several utterances were recorded on magnetic tape and played back, displaying the T.I.s on the screen of an oscilloscope. The utterances were of various durations, from single phonemes to complete sentences. These displays (fig. 1-1) are presented to give a general impression of the intervals that are present in certain speech sounds and an estimation of their stability in time. A simplification of this display to enhance its usefulness as 'visible speech', which is not the main concern of this study, is presented briefly in appendix 6.

Several important points derived from this display are illustrated in figure 1.1. It should be noted that in this and subsequent photographs, T.I.s are measured downwards from the solid horizontal line and that the T.I. scale is exponential with respect to time. The display or 'interval-gram', to use Chang's terminology, of the sentence "She said it." illustrates the stark difference between vowel and fricative T.I. distributions and minor differences between the three wowels and between the three fricatives present. The extent to which the interval distributions are constant within each phoneme segment is encouraging for the analysis of steady state continuants.

The totally vocalic utterance of "Where are you?" presents less



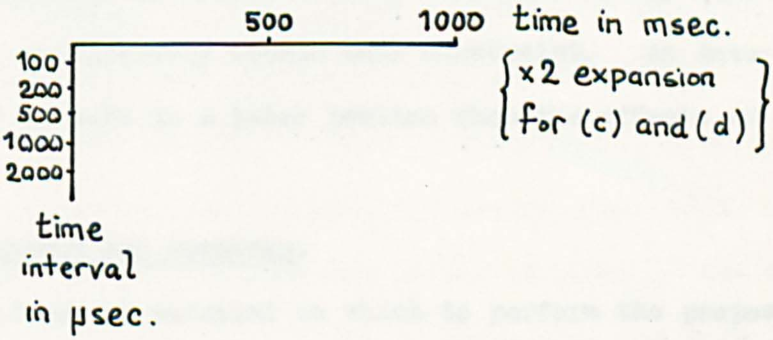
(a) She said it.

(c) / i/

(b) Where are you?

(d) / a/

Fig. 1.1 Time interval patterns of speech sounds.



Scale used for running histograms.

clear cut distinctions although smooth trends and steady state portions can be detected.

The intervalgrams of the two vowels /i/ and /a/ have a time scale which has been expanded by a factor of two. They illustrate two forms of variation of the T.I. distributions of steady vowel utterances of constant pitch. In the case of /i/ one peak of the distribution is well defined in the long time interval region. The scatter of points in the short time interval region indicates that distributions measured over short durations may differ quite considerably within the same steady utterance. This intervalgram illustrates the need either to measure distributions over periods longer than some empirically defined minimum or to describe the distribution as a histogram using bins of sufficient width to smooth the fluctuation observed.

The variation with time in the intervalgram of the vowel /a/ is slower than that of/i/. It is interesting to note the strong negative correlation between the variation of the shortest and longest intervals. The sum of the characteristic intervals is seen to remain fairly constant but three of them vary considerably within this constraint. An interpretation of this will be made in a later section when the effects of pitch are considered.

1.3 The choice of phonemes and speakers.

The choice of speech material on which to perform the proposed analysis presented several problems. It was desired to use speech sounds defined on a scale which was as widely accepted as possible. The cardinal vowels defined by Jones (31) were considered as a partial fulfilment of the above aim. Such a choice would however restrict the choice of speakers to those sufficiently trained to produce these sounds accurately. It was therefore decided to use speech sounds which could be produced by any speaker with a minimum of instruction and which would also conform as nearly as possible to those used by other experimenters in the analysis and synthesis of speech. The large volume of work done on American English was obviously unsuitable for comparison with the present work on a phonemic level due to the large differences in vowel quality.

The phoneme set finally chosen was that defined by Abercrombie (1) as "based on the accent of Standard English which is best called 'R.P.' or 'received pronunciation'.", and used by Holmes et al (30) in their work on 'Speech Synthesis by rule'. This set was found to be the best approximation to the phonemes produced by the available speakers. These were the three members of the group working on speech in the Laboratory. They were naive phonetically, but had experience in the analysis and synthesis of speech. A selection of contexts from which the phonemes could be extracted was compiled, based on agreement between the three members of the group .

These contexts consisted of several isolated words which were agreed to define a particular phoneme. When an utterance of a certain phoneme was required, the selection of words were spoken after which the speaker would articulate the required phoneme common to each of the words previously uttered. In this way the problem of memorising the positions of the articulators for the unnatural isolated utterances was overcome by

TABLE 1

Defining contexts of the phonemes uttered by speakers.

/i/ :seek,teach,peel,beam,leave,hear.

/ɪ/ :sick,pill,wit,king,fizz,dig.

/ɛ/ :shed,then,rest,sell,met,ten.

/æ/ :bad,mass,sap,tang,vat,rack.

/3/ :bird,shirk,learn,verse,were,hurt.

/ʌ/ :cup,dung,pus,fun,gust,sud.

/ʊ/ :took,bush,could,put,foot,good.

/u/ :moon,boot,tool,zoo,spoon,lune.

/ɒ/ :top,pot,shone,gong,doll,what.

/ɔ/ :law,sawn,port,fork,all,thought.

/ɑ/ :mark,car,art,barn,psalm,large.

/ə/ :the.(As distinct from /3/ .)

/v/ :eva,vote.

/z/ :easy,zoo.

/ʒ/ :azure.

/ð/ :then,this.

/f/ :if,for.

/θ/ :thin,thank.

/s/ :ass,sock.

/ʃ/ :she,shirt.

/h/ :he,hum.

/m/ :mat,mop,mood,omo,come,loom.

/n/ :nock,night,noon,ana,clan,bun.

/ŋ/ :sing,king,ring,wrong,rang,pang.

the use of the short term memory of the phoneme defining contexts. A list of the phonemes used and their defining contexts are shown in Table 1.

1.4 Time interval histograms of control stimuli.

The results illustrated in fig.1.1 confirm the findings of Chang et al.,that the T.I. distribution of speech sounds varies in a fairly smooth way in the manner of the well-known spectrogram of speech. The work described in the rest of this chapter is an extension of the work published by Bezdel and Chandler (5) and Sakai and Inoue (55), and is based on the above observed fact.

Before looking in detail at the time interval histograms, which are essentially quantised sections of the intervalgram pattern averaged over a short duration, T.I. histograms of control stimuli will be presented. The purpose of this is to reveal the accurany of operation of the apparatus described earlier in the chapter,and to familiarise the reader with the use of this analysis on simple signals,prior to investigating their use with complex speech waveforms.

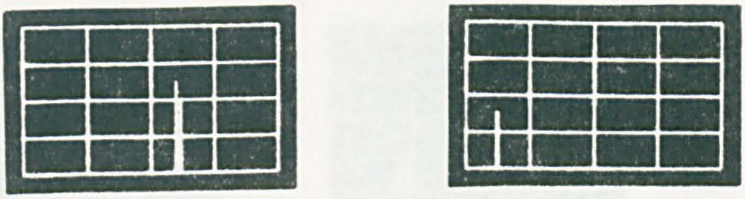
All the analysis of time interval histograms was originally done using the C.A.T. computer and will be described as such. However the actual displays illustrated in this thesis were produced using the PDP-8 computer and its associated 338 display using the C.A.T. simulating program. This line of action was taken when difficulty was experienced in printing the photographs taken from the screen of the C.A.T.. In the simulated display solid lines of variable length rather than the displacement of a dot are used to indicate the contents of a histogram bin.

The time interval scale was derived by driving the address advance of the C.A.T. externally with a variable frequency multivibrator. This could be adjusted to give any linear time interval scale within its range. In practice this was usually set to oscillate at 50 Kc/s and expansions of the scale by factors of two and four were made by use of a 'horizontal size' control. The bin widths of the histograms produced by this system were always 20 µs. but the bins were displayed more or less densely along the horizontal axis.

1.4.1 T.I. histograms of simple auditory signals.

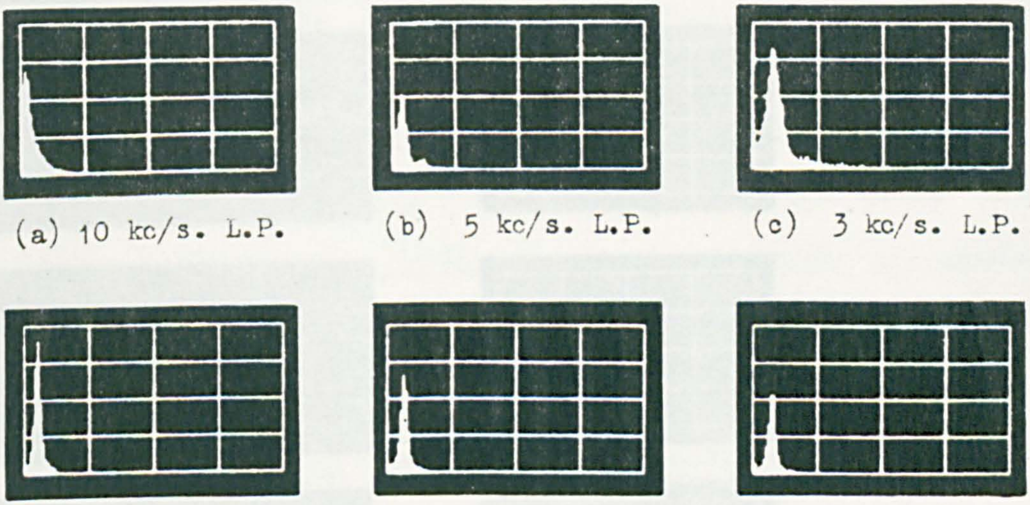
The form of the control stimuli and their time interval histograms are shown in figures 1.2, 1.3 and 1.4. Figure 1.2 illustrates the T.I. histograms of two sine waves. On the left hand side is the histogram of a sine wave of frequency 200 c/s. and amplitude 1 v. peak to peak, and on the right, that of a sine wave of frequency 1000 c/s. and the same amplitude. Only a very slight departure from the ideal unity mark to space ratio of the clipped signal is experienced at very low frequencies at this amplitude. As was mentioned in section 1.1.2.1 the signal strength for subsequent speech input to the system was maintained at a mean level of 1 v peak to peak.

Figure 1.3 illustrates the T.I. histograms of white noise generated by a Dawe white noise generator. Low pass and band pass filtered versions are presented for comparison with the histograms of noise-like speech sounds.



(a) 200c/s. (b) 1000c/s.

Fig. 1.2 T.I. Histograms of sine waves.
(4 msec. scale)



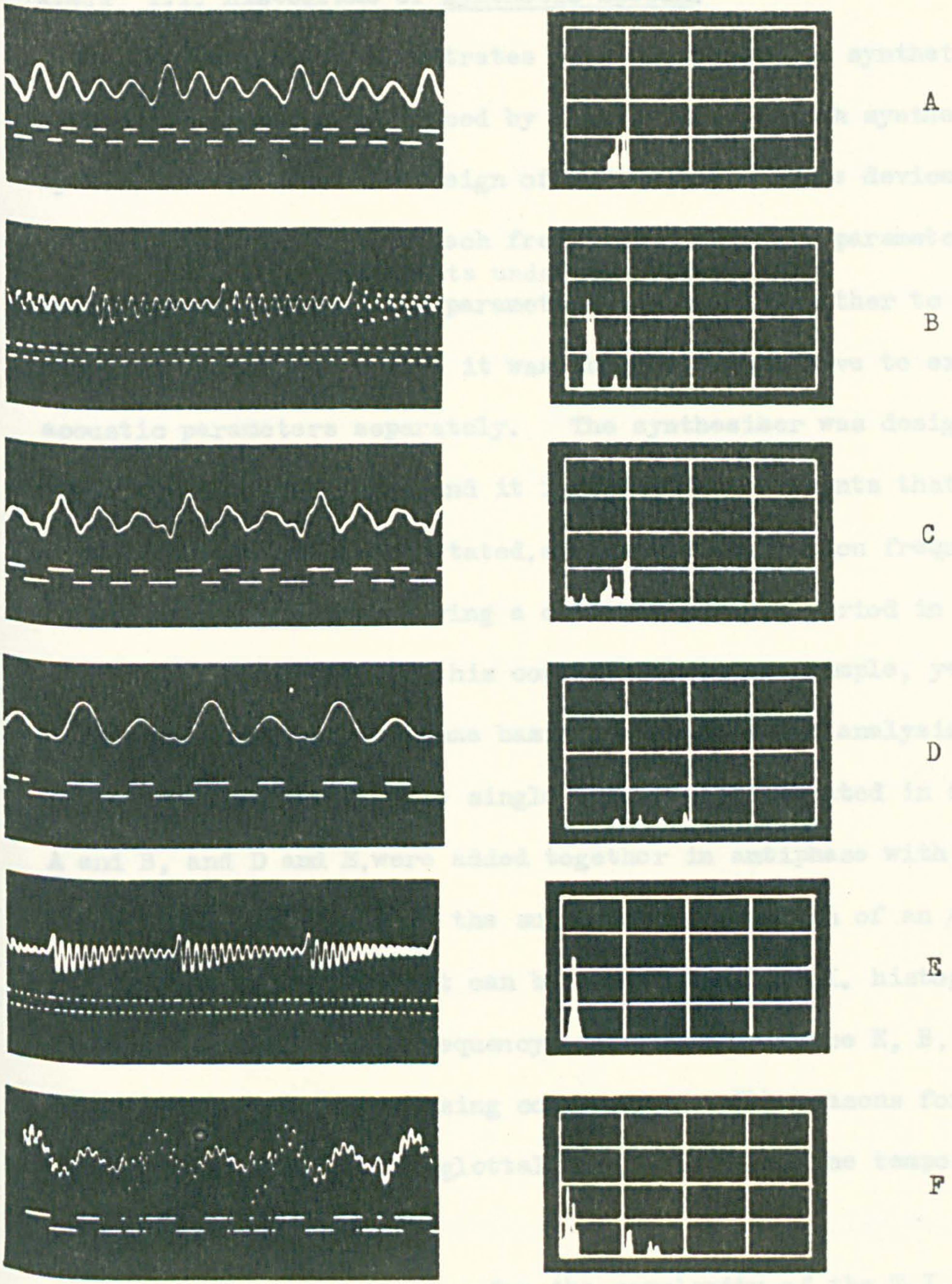
(a) 10 kc/s. L.P. (b) 5 kc/s. L.P. (c) 3 kc/s. L.P.

(d) 2-5 kc/s. (e) 2-4 kc/s. (f) 2-3 kc/s.

Fig. 1.3 T.I. Histograms of bandlimited white noise.
(4 msec. scale)

Time interval scale of the histograms.

0 1 2 msec.
0 1 2 3 4msec.
o 2 4 6 8msec.



(4 msec. time scale)

Fig. 1.4 Waveforms, clipped waveforms and histograms of synthetically produced formants.

1.4.2 T.I. Histograms of synthetic speech.

Figure 1.4 illustrates T.I. histograms of synthetic acoustic parameters of speech produced by a parametric speech synthesiser built by W.A.Ainsworth to the design of J.N.Holmes. This device is capable of producing synthetic speech from twelve parallel parameter controls. In normal operation components under the control of these parameters are mixed together to produce a synthetic waveform. Here it was thought instructive to examine some acoustic parameters separately. The synthesiser was designed on the formant model of speech, and it is individual formants that are examined here. Unless otherwise stated, a constant excitation frequency was maintained throughout, thus giving a constant glottal period in the time domain.

The purpose of this control was to use simple, yet speech-like, waveforms to illustrate some basic features of the analysis of the T.I.s of speech waveforms. The single formants illustrated in figure 1.4 parts A and B, and D and E, were added together in antiphase with relative amplitudes chosen to give the subjective impression of an /ɜ/ (part C) and an /i/ (part F). It can be seen that the T.I. histograms of single formants of decreasing frequency, (i.e. the sequence E, B, A, and D in figure 1.4) reveal increasing complexity. The reasons for this are the asymmetrical form of the glottal excitation, and the temporal constraints of glottal repetition.

These two reasons for the complexity of the T.I. histogram of a signal which has a comparatively simple specification in the frequency domain will be discussed in more detail.

1.4.3 The asymmetry of glottal excitation.

The repetitive damped sinusoid which constitutes the waveform of a single speech formant is not symmetrically placed about its zero level. This is due to the asymmetrical nature of the excitation function which is a succession of unidirectional pulses of air. In the case of the synthesised speech this asymmetry was controlled by the a.c. coupling of the signal, but to some extent this compensated for the idealised 'glottal pulse' used to excite the synthesiser. The asymmetry, so produced, tended towards the expected effect that a more natural excitation function would have on the waveform.

This asymmetry causes the interval which is half of the inverse of the formant frequency to be at the centre of a spread of intervals and not to occur alone as in the case of a sine wave. This perturbation of the intervals can be loosely associated with the bandwidth of the formant which has not been found to be an important feature in the frequency domain measurements of speech.

1.4.4 The constraints of glottal repetition.

The effects of these constraints, as seen in figure 1.4, are correlated with the relative values of glottal period, or fundamental frequency, and the formant frequency. However the effect of a finite glottal period will be considered first of all in isolation.

The facts of the case are very simple. The glottal period contains a set of time intervals which ideally are thought to be repeated within every such period of a constant sound. When the glottal period changes

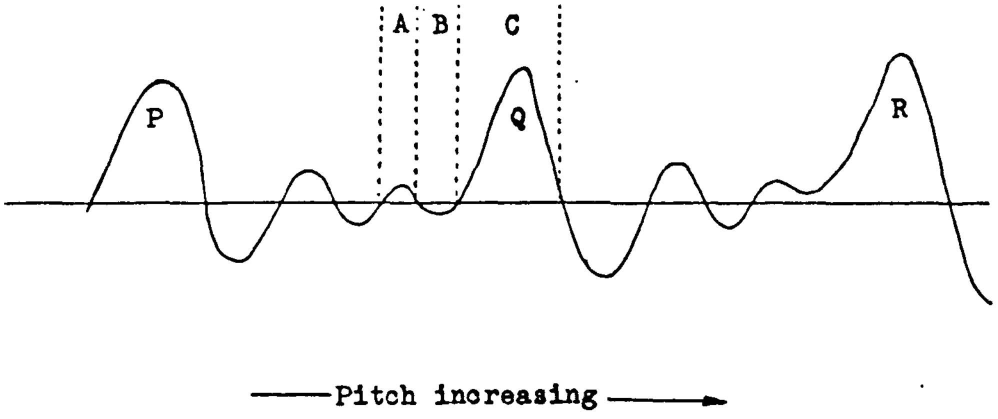


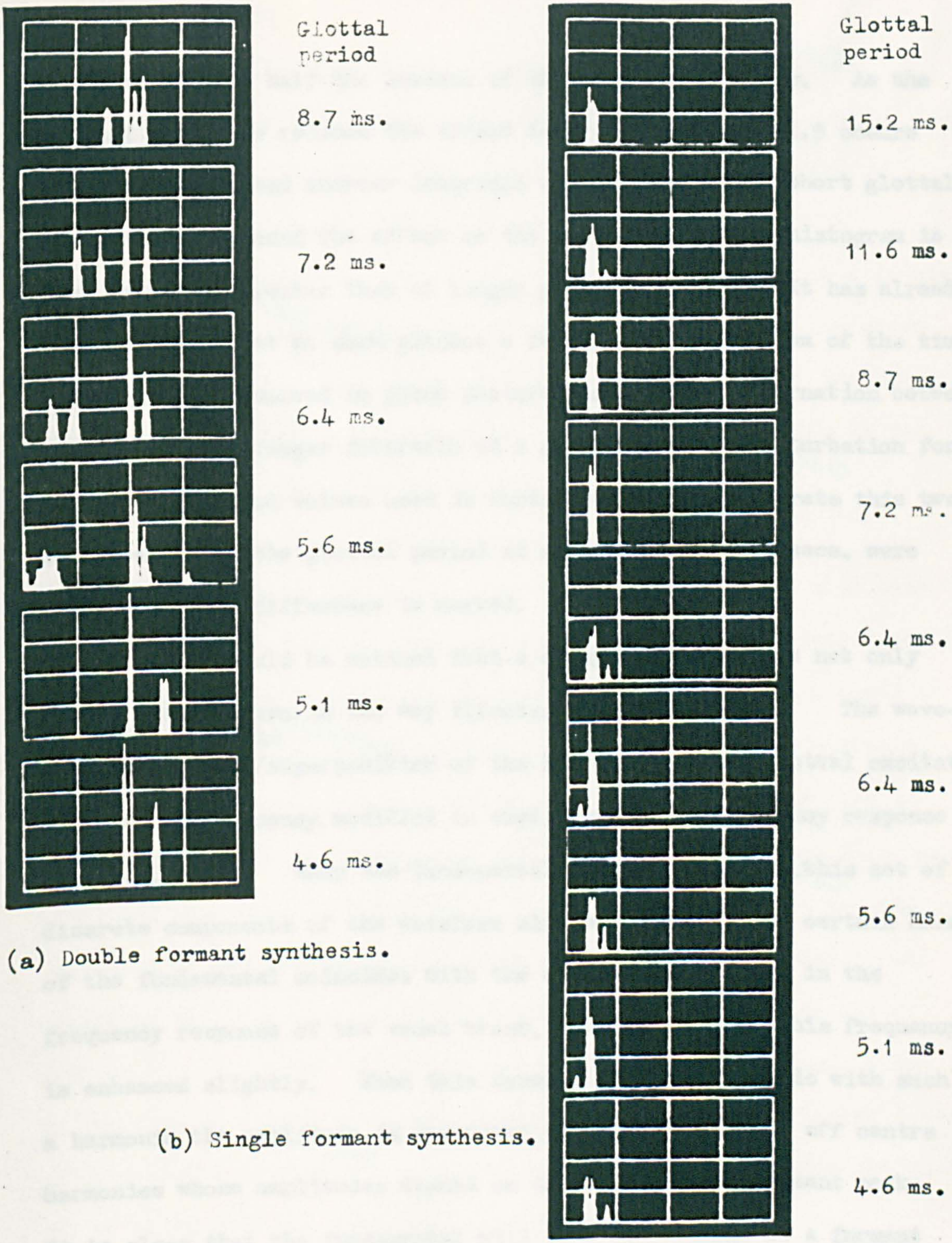
Fig. 1.5 Pitch perturbation of a voiced speech waveform.

and the sound remains the same, albeit giving a different pitch sensation, the set of time intervals must change, either to occupy a longer interval between glottal excitations or a shorter one. Figure 1.5 illustrates a typical case of a vowel waveform in which the perceived pitch is increasing and the observed glottal period is decreasing. It has been observed by viewing the waveforms of both natural and synthetic vowels that intervals corresponding to A, B and C coalesce when a shortening of the period is required,and reappear as separate intervals when a lengthening of the period occurs. If the glottal period is shortened further the interval under peak R is reduced by the same amount until a further three intervals coalesce.

The relation of this to the effects seen in figure 1.4 is as follows. When a high frequency formant has a glottal period of a certain value, the perturbation to the T.I. statistics of the change in one interval or the coalescing of three intervals is small when compared to the number of stable intervals contributing to the statistics. It is clear that when a low frequency formant has a similar glottal period the perturbation due to these 'end effects' is quite substantial. Thus these perturbations are seen to be proportional to the relative values of the glottal period and formant frequency, or more directly to the number of T.I.s there are in each glottal period.

1.4.5 The effect of gross pitch changes on the T.I. histogram.

The changes that are caused in a T.I. histogram by gross changes in pitch, and therefore in the glottal period, can be seen in figure 1.6. The histograms of a single synthetic formant are seen to contain a pre-



(a) Double formant synthesis.

(b) Single formant synthesis.

Fig. 1.6 Histogram variation with the pitch of a synthetic voiced speech wave. (4 msec. scale)

dominant peak at half the inverse of the formant frequency. As the glottal period is reduced the effect described in figure 1.5 occurs and both longer and shorter intervals occur. When very short glottal periods are produced the effect on the major peak of the histogram is seen to be far greater than at longer glottal periods. It has already been observed that at such pitches a far greater proportion of the time intervals are involved in pitch perturbations. The alternation between the shorter and longer intervals as a result of pitch perturbation for the glottal period values used is fortuitous. To illustrate this two histograms with the glottal period at approximately 6.4 msecs. were taken and their difference is marked.

It should be noticed that a change in pitch does not only affect the waveform in the way illustrated in figure 1.5. The waveform consists of the superposition of the harmonics of the glottal excitation fundamental frequency, modified in amplitude by the frequency response of the vocal tract. When the fundamental frequency changes, this set of discrete components of the waveform also change. When a certain harmonic of the fundamental coincides with the centre of a formant in the frequency response of the vocal tract, the amplitude of this frequency is enhanced slightly. When this formant does not coincide with such a harmonic the amplitude is dependent on the summation of off centre harmonics whose amplitudes depend on the shape of the formant peak. It is clear that the fundamental will be a subharmonic of a formant frequency at different values of fundamental frequency for different formants.

When the T.I. distribution is poised nicely between being dominated by one of two formants of nearly equal amplitude, this effect can cause a different kind of pitch perturbation in the distribution. This effect plus the previously mentioned pitch dependence can be seen in figure 1.6, where the histograms of a double formant sound, giving the subjective impression of an /u/, are shown. At glottal periods of 5.6 msec. and 6.4 msec. the effect of the higher formant is reflected in the histogram. The only modification to the synthesiser was a change of excitation frequency.

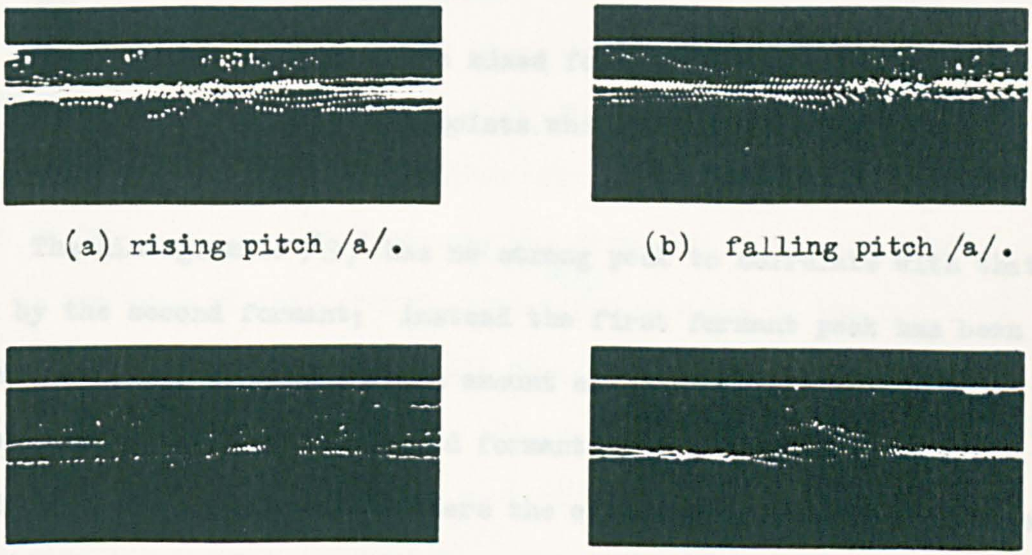
1.4.6 Application to T.I. measurements on real speech.

It was seen in figure 1.1 that there is a variation with time of the T.I. distributions. The explanation of this was reserved until now. It seems that these variations can be explained on the basis of changes in the length of the glottal period. During a subjectively steady pitch utterance, the length of the glottal period is seen to jitter back and forth around some mean value. It is assumed that the intervals which showed variation in figure 1.1 are those at the end or beginning of the glottal period which are maximally perturbed by any change in the length of the glottal period. The correlation noted between the occurrence of the longest and the shortest intervals can be interpreted as the occasional slight crossing of the clipping level which caused a pair of short intervals to be created reducing the long one of which they had previously been a part.

1.4.7 The effect of monotonic pitch variations.

To assess the effect of continually reducing or increasing the length of the glottal period on the T.I. pattern, vowels with rising and falling pitches were produced. The vowels /ɑ/ and /u/ were chosen to represent open and close vowels; the waveform of the former showing greater formant damping than the latter. Figure 1.7 depicts the variation of T.I.s in time as the pitch is falling and as it is rising. The pitch change was approximately one octave.

In both vowels the rising pitch involves the shortening of certain intervals and the falling pitch involves the lengthening of certain intervals. In the case of /u/ this is seen to be a long slow procedure involving several intervals. Two of the four characteristic intervals of the waveform are involved in major changes,and the other two in minor perturbations. In the case of the open vowel /ɑ/, the major formant being higher than in /u/, there are more intervals within the glottal period. A far smaller proportion of them are effected by the variation of pitch and fewer long slow variations in the intervals are seen. This indicates that certain vowels, especially close vowels, whose first formants do not decay appreciably in the glottal period,are effected very greatly by pitch variations such that the majority of intervals are changed. This behaviour is seen in such vowels as /u/, /ʊ/ and /i/. In most other vowels however the variation of time interval pattern with pitch is more like that of /ɑ/.



(a) rising pitch /a/. (b) falling pitch /a/.

(c) rising pitch /u/. (d) falling pitch /u/.

Fig. 1.7 Time interval patterns of vowels with rising and falling pitch. (Scales as in fig. 1.1 (c) and (d).)

1.4.8 The mixing of synthetic formants.

The T.I. histograms of the mixed formant waveforms, shown in figure 1.4 (C, F), indicate further points which are typical of the histograms of vowel sounds.

The histogram of /3/ has no strong peak to correlate with that produced by the second formant; instead the first formant peak has been split into two peaks separated by an amount approximately equal to the characteristic interval of its second formant. The reason for this can be clearly seen from the waveform where the effect of the lower amplitude second formant is to add to and subtract from the basic first formant intervals.

In the case of /i/ the same process is seen in the splitting of one of the long T.I. peaks, but also the second formant is strong enough to produce T.I.s at its own characteristic interval.

If the effects of pitch are ignored, the mixing of two formants causes similar changes in the T.I. pattern to those seen when two sinusoids are mixed.

In most unstressed vowel sounds it is usual that the second and higher formants are weaker than the lowest formant. Thus the effects of a low amplitude high frequency added to a higher amplitude low frequency are commonly seen.

When the amplitude ratio is large only the lower frequency is reflected in the T.I. distribution. As this ratio is reduced the higher frequency perturbs this distribution by lengthening and shortening the long intervals by amounts up to $1/_{fn}$ (where fn is the higher frequency). As the ratio approaches unity, a peak at $1/_{2fn}$ appears in the T.I.distribution.

During studies on the synthesis of speech on a two formant model, Delattre et al.(13) showed that for some vowels, vowel colour is critically dependent on the relative amplitudes of the formants. It is therefore unlikely in the case of these vowels that gross changes in the T.I. distribution of a particular vowel will occur, caused by changes in the relative amplitudes of the formants. The vowels whose colour is not so critically dependent on the relative amplitude of the formants are those whose two formants are well separated. The influence of these separate formants on the T.I. distribution can more easily be extracted if confusion due to variation of relative amplitudes of formants should occur. Miller (48) also states that formant amplitudes are important and sometimes critical for vowel perception.

It has proved helpful to use the simplified speech like waveforms produced by a parametric synthesiser to illustrate some basic transforms from the more traditional frequency description of speech to the time interval description. It should be noted however that it is not possible to regain the frequency description from the time interval description except in the most trivial of cases. Cobb (10) and Rainal (53) have studied the mathematics of the interval distribution for a single sine wave mixed with noise, but when additional sine waves are added, relative amplitudes and phase relationships of the frequency components greatly complicate the description in frequency terms and render any attempt to derive these variables from a T.I. distribution impossible .

Observations of the dependence of the T.I. distributions on the relative formant amplitudes revealed variations very similar to those reported by Fourcin (23) when measuring the distribution of spectral energy after infinite peak clipping of the waveform. He reported that the spectrum is dominated by the strongest formant if it exceeds the second strongest formant by 5 db. In the spectral case formant harmonics are also present.

1.5 First order T.I. distributions of isolated speech sounds.

1.5.1 The display of distributions.

The isolated speech sounds described in section 1.3 were recorded on magnetic tape and used as input to the system shown in figure 1.8.

The zero-crossing pulses were applied to the address reset input of the C.A.T. The address advance was operated using an external multivibrator. It was necessary to continue to use a rapid sweep, against which to measure the T.I.s, in order to obtain a display of reasonable horizontal dimension on the screen of the C.A.T.. The maximum internal sweep gave a histogram whose features it was difficult to resolve. At the time no suitable oscilloscope was available to be linked to the analogue output of the C.A.T.'s vertical deflection signal. This was done later, achieving much wider histogram bins and flexibility in horizontal size, controlled by the time base of the oscilloscope.

Histograms were produced for a full range of tape recorded continuant speech sounds and were recorded photographically. Spontaneous

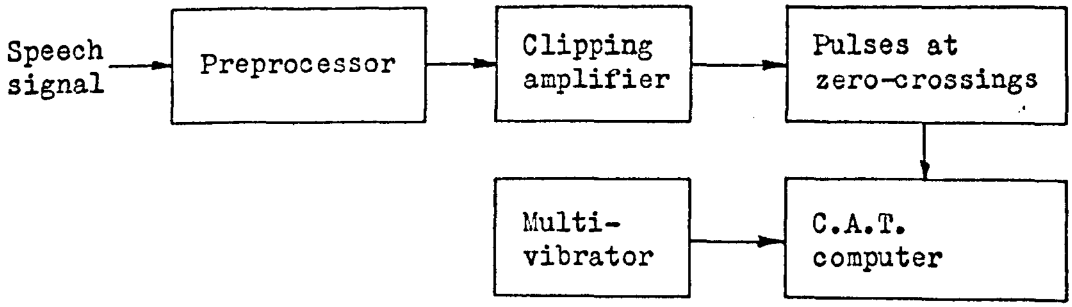


Fig. 1.8 Block diagram of the time interval histogram compilation system.

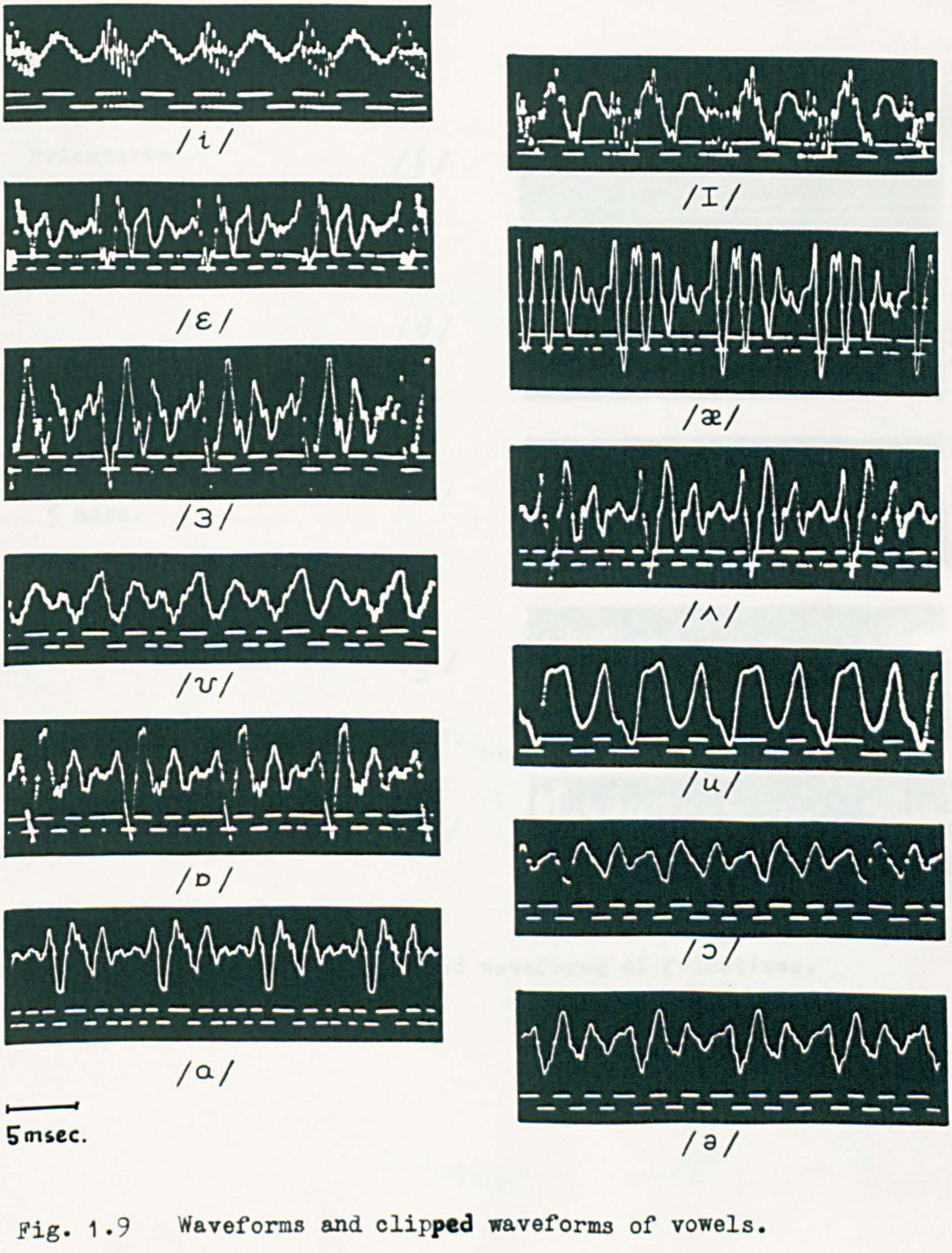


Fig. 1.9 Waveforms and clipped waveforms of vowels.

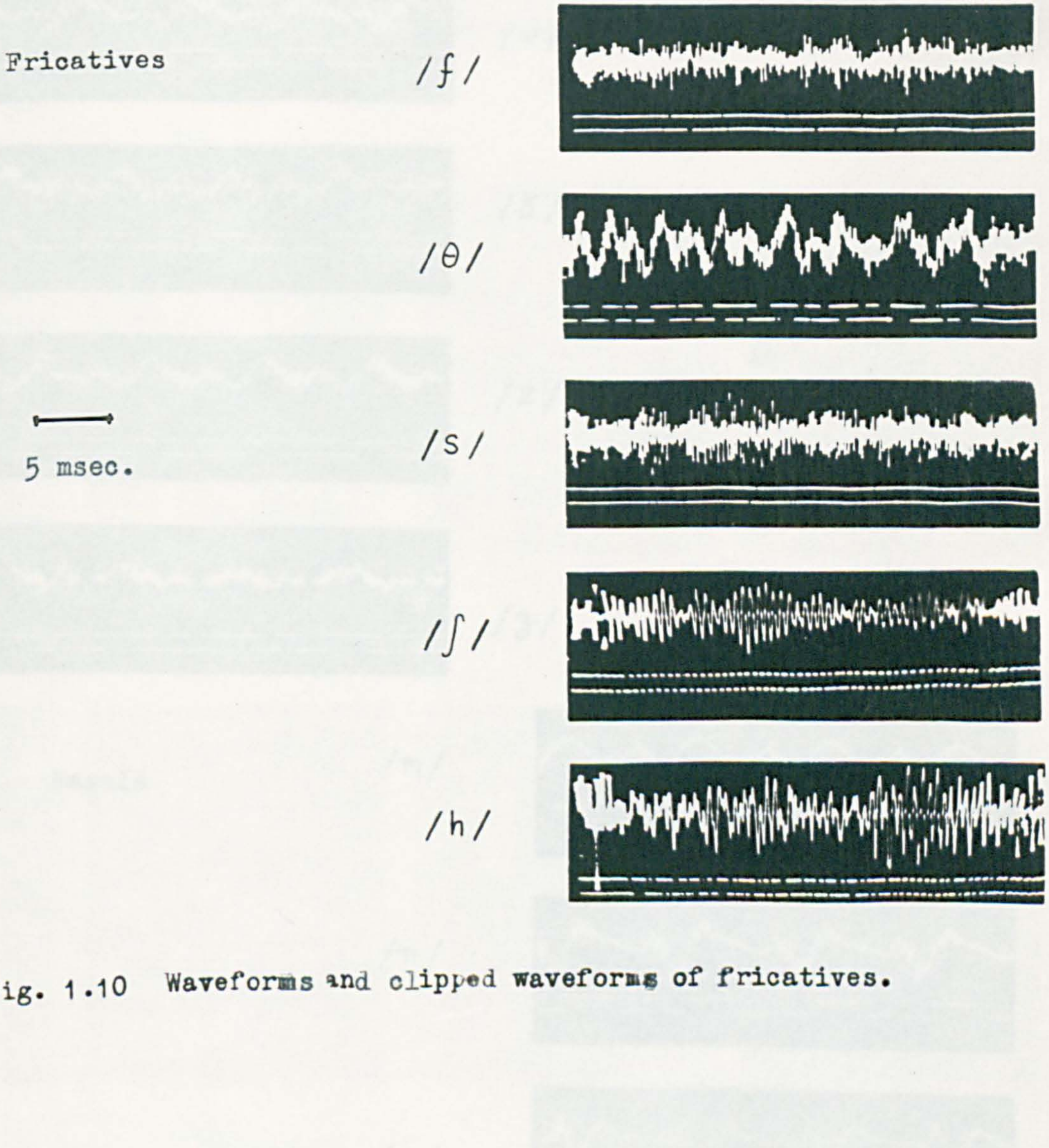


Fig. 1.10 Waveforms and clipped waveforms of fricatives.

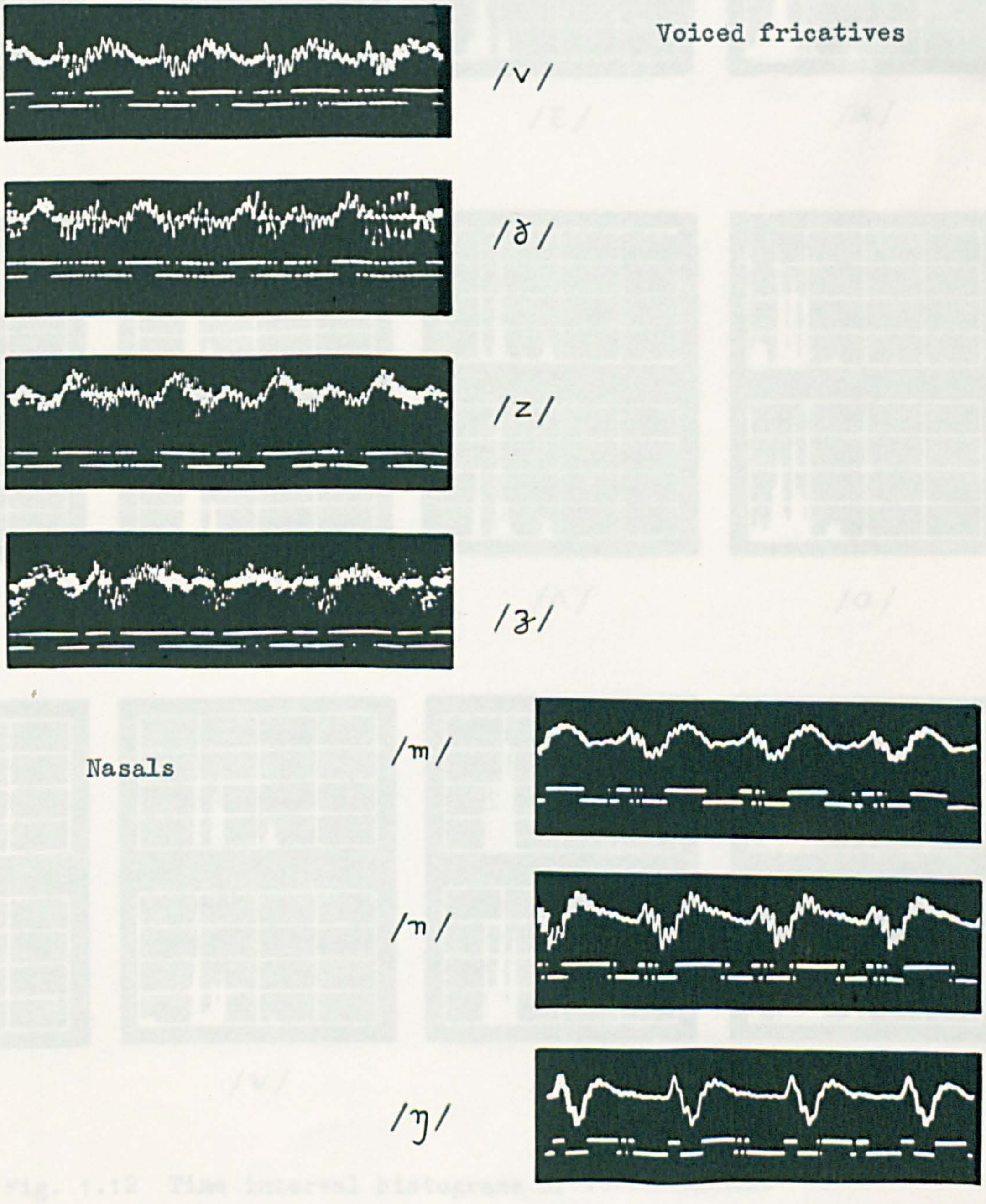


Fig. 1.11 Waveforms and clipped waveforms of voiced fricatives and nasals.

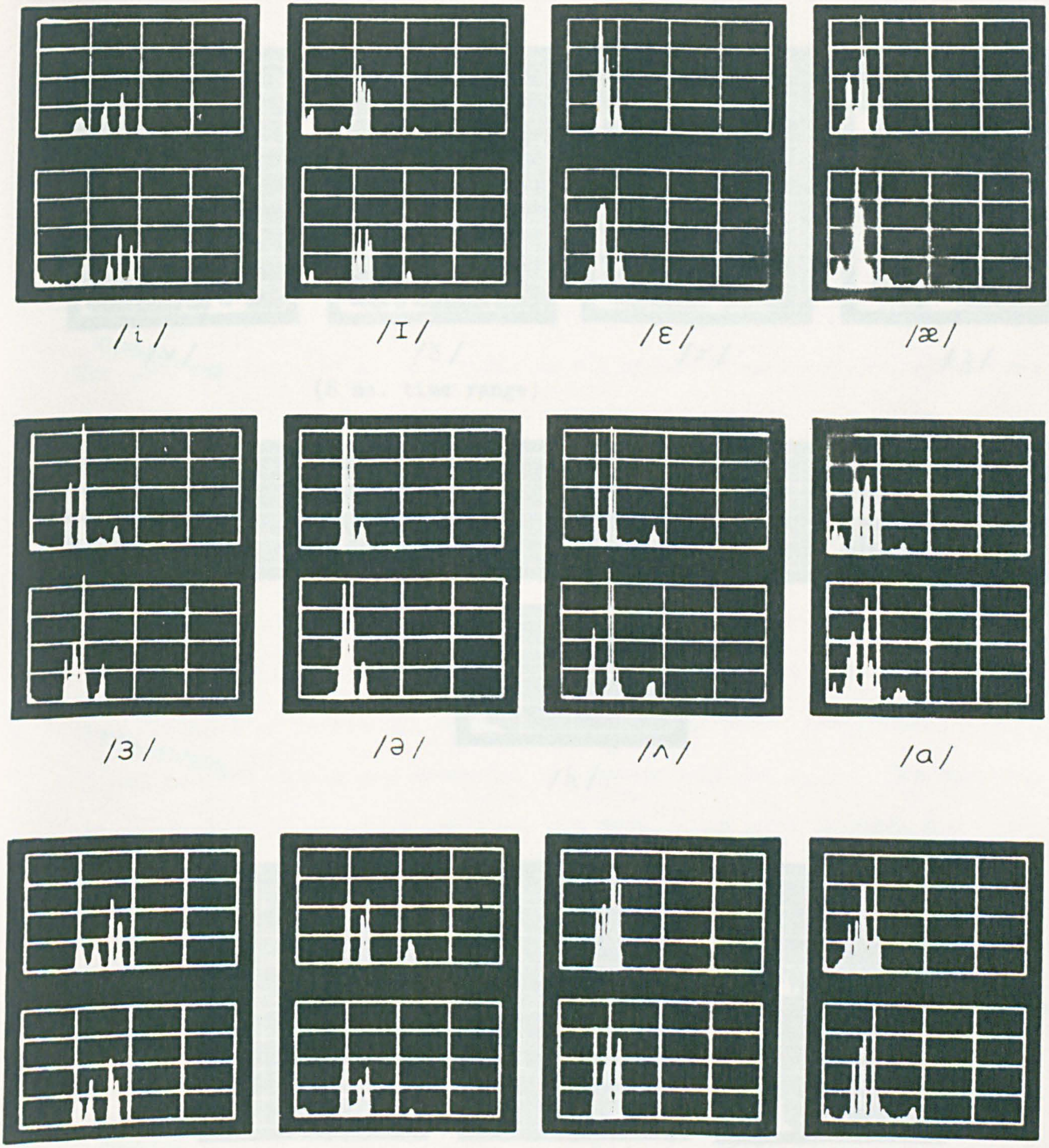


Fig. 1.12 Time interval histograms of vowel sounds.

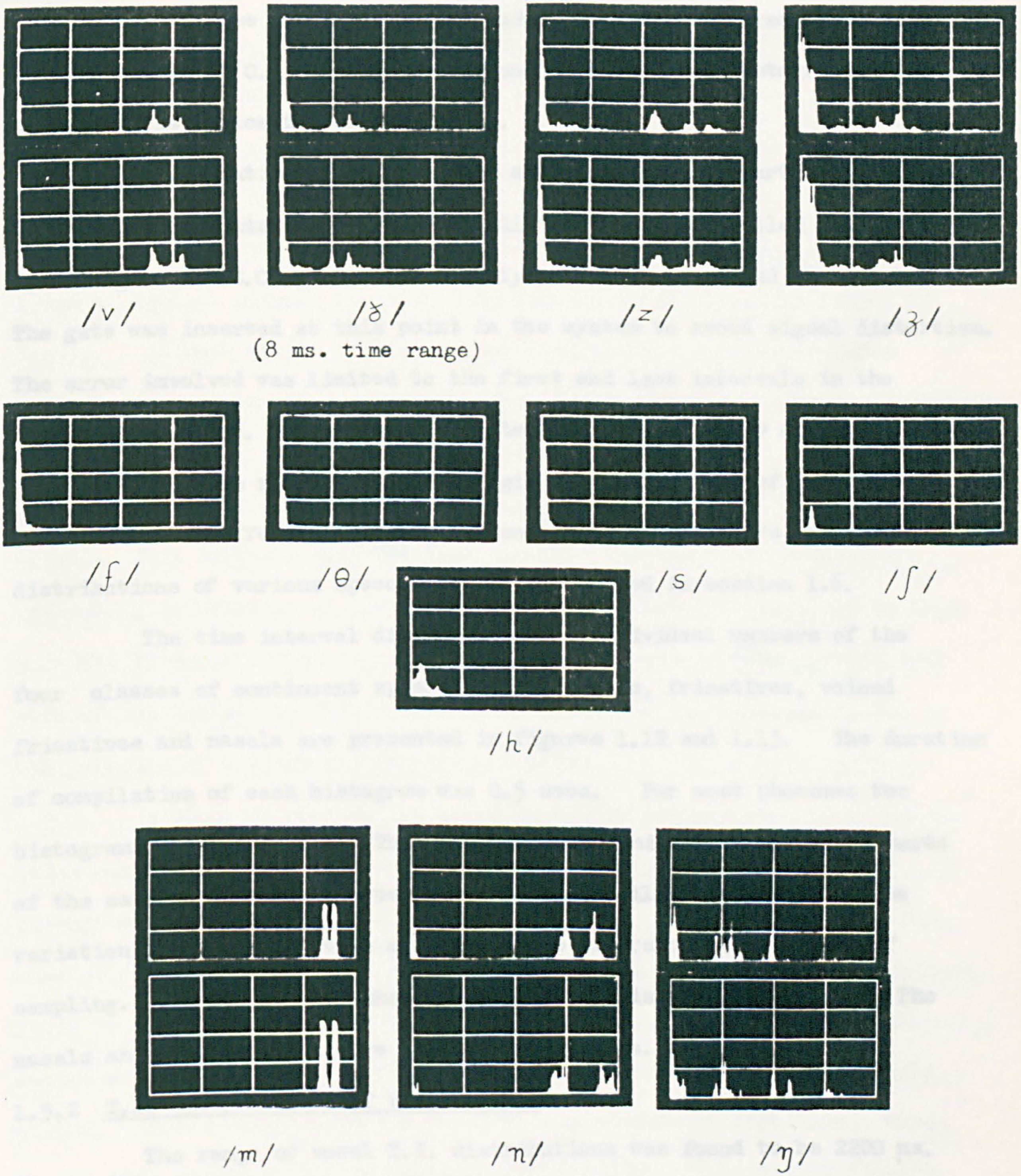


Fig. 1.13 Time interval histograms of voiced fricatives, fricatives and nasals. (4 ms. time scale, every zero-crossing counted) - except /ð/ 8 msec.

utterances of these sounds were also made and their histograms were computed on-line using the C.A.T.. These were compared with the photographic records to ensure the typicality of the latter.

The duration of a continuant contributing to a particular histogram, which was termed its 'duration of compilation', was controlled using a gate operating on the Z.C. pulses immediately before they entered the C.A.T.. The gate was inserted at this point in the system to avoid signal distortion. The error involved was limited to the first and last intervals in the compilation period. The gate was controlled by a manually operated monostable of variable refractory period, giving a wide range of possible durations. The effect of varying the on-period of this gate, on the T.I. distributions of various speech sounds is reported in section 1.6.

The time interval distributions of individual members of the four classes of continuant speech sounds, vowels, fricatives, voiced fricatives and nasals are presented in figures 1.12 and 1.13. The duration of compilation of each histogram was 0.5 secs. For most phonemes two histograms are presented. They are taken at random from different parts of the same tape recorded utterance in order to illustrate some of the variation that occurs within a steady state utterance at this rate of sampling. It can be seen that in most cases this is very slight. The nasals and certain vowels are the only exceptions.

1.5.2 T.I. distributions of vowel sounds.

The range of vowel T.I. distributions was found to be 2200 µs. The most striking feature of these distributions is their 'peakedness'.

In many cases four separate peaks can be distinguished. This is more than those observed by Sakai and Inoue (55), but they measured intervals under positive or negative lobes of the waveform only. These distributions are a combination of both. Some general features can be extracted visually and a certain amount of meaning attached to them by reference to the histograms of control stimuli (section 1.4) and the vowel waveforms (fig.1.9).

In most of the histograms the distribution centres around a maximum peak. The other peaks are often separated by a constant value; this is very clear in the first histograms of /i / and /ɑ/. In many others this pattern can be seen to emerge although largely obscured by noise. This suggests that the time interval histogram is capable of reflecting information of the vowel colour which may be equivalent to that conveyed by the two most dominant formants. It can also be assumed that some of the 'noise' in this simple model of mixing two formants is due to the presence of a third fairly strong formant.

A study of some of the vowel waveforms reveals that more information concerning the higher and weaker formants is available in the T.I. histogram than might have been expected. It can be seen clearly in the case of /i / that the second formant is not present in a constant amplitude ratio to the first formant. It has been shown by Holmes (29) that the second and higher formants are commonly excited twice within the glottal period. This variation in the relationship of formant amplitudes, which is not seen in the frequency analysis if fairly narrow bandwidth filters are used, is of considerable importance in time domain analysis. In the case of /i/, the first and second formants alternately dominate the waveform and hence the T.I.distribution.

The above is but one explanation of some of the features evident in the T.I. histogram. It is clear from the examination of the wave form of /u/ that the histogram pattern produced, not unlike that of /i/, is due rather to the interaction of the first and second formants. Intervals characteristic of their difference occur rather than intervals characteristic of one or both of them, as was seen in the histogram of the two formant synthetic /ɜ/.

The ordering of the vowel T.I. histograms in figure 1.12 is roughly according to their measured positions on a formant 1 / formant 2 (F1/F2) plane. It was seen that certain characteristics of the histogram followed grouping in this plane. As F1 increases (left to right) there is a trend towards more compact histograms on the T.I. axis. This can be seen especially in the top and bottom rows, which represent the extreme values of F2. Certain subgroups of vowels can be distinguished on the basis of histogram features. The clearest ones observed in figure 1.12 are /æ/, /ɑ/, /ɒ/; /i/, /u/; and /ə/, /ɜ/.

It will have been noted that no reference has been made to the variability of the T.I. histograms of these sounds under the typical speech variations of pitch, stress and speaker. Experiments were done to examine the effect of change of pitch and change of speaker. A change in stress could not be controlled and was not investigated seriously. The result of these experiments showed that some of the features observed in figure

1.12 remained as characteristic of a particular vowel, but in many cases the histograms looked completely different from the ones illustrated. The effect of these changes will be discussed in the next chapter where the more sensitive and unique display of second order statistics will be described. It will be seen that the first order statistics can be easily extracted from the presentation of the second order T.I. statistics used.

1.5.3 <u>The T.I. histograms of Unvoiced fricatives.</u>

The five unvoiced fricatives in English, /f /, /θ /, /s /, /ʃ / and /h /, are illustrated by one histogram each as their variation within a given isolated utterance could not be detected. In figure 1.13 they are represented by histograms displayed on the same linear time interval scale as the vowel histograms. This scale was used throughout figures 1.12 and 1.13 to illustrate the different parts of the T.I. dimension occupied by various classes of sounds. The amplitudes of the fricative histograms have been reduced by a factor of 16.

These fricatives can be considered acoustically as bandlimited white noise, and their histograms may be compared to those produced using noise as a control stimulus (fig.1.3). They differ in the band of noise present in some cases, according to the manner and place of production of this noise.

The sounds /f / and /θ / have almost identical histograms. They are both produced by a constriction at the front of the mouth, either by lip and teeth or tongue tip and teeth. The histogram consists of a peak at the very shortest intervals experienced in speech plus a slight peak at about 200 μs. This subsiduary peak was found in most of the histograms produced for these two sounds. A small number of much longer time intervals are sometimes found due to an inadvertent whistle produced with these sounds.

The sounds /s / and /ʃ / are produced by constricting a continuous flow of air by means of the tongue. Their histograms have a similar shape but are in a different position on the T.I. scale. The peak of /s / is at approximately 100 μsec. and that of /ʃ / at 200 μsec. They are both very similar to a T.I. histogram of a single high frequency formant.

The sound /h / is different from the previous four as it is produced due to a constriction at the glottis rather than at the front of the oral cavity. As such it can be likened to a whispered neutral vowel sound. The formant cavities through which the excitation noise passed are reflected in the T.I. histogram as a double peak distribution.

It was found for the utterances examined that the unvoiced fricative sounds could be discriminated by means of their T.I. histograms, with the exception of /f / and /θ /. It has been found by others (11, 32) that these sounds are only separated by their context and not by any known acoustic specification of their isolated form.

1.5.4 The T.I. histograms of voiced fricatives.

The voiced fricatives /v /, /ð /, /z /, and /ʒ/ can be likened to the four front articulated unvoiced fricatives with the addition of voicing. The fricative parts are articulated in the same way as /f /,

/θ/, /s/ and /ʃ/ respectively. If the short interval portion of each T.I. histogram is observed it is seen that this fact is reflected in them. In the case of /ʒ/ however, there seem to be extra very short intervals not found in /ʃ/.

However, the most characteristic features of the histograms taken as a whole are the very long intervals representing the voicing. These are often much longer than those found in vowels; see especially /v/, /ð/ and /z/.

Discrimination between the voiced fricatives by means of their T.I. histograms was found to be unreliable due to the variable emphasis on the friction and the voicing.

1.5.5 The T.I. histograms of nasals.

The nasals are produced in a similar way to the vowels except the nasal cavity is used instead of the oral cavity as the final resonant cavity. The spectral components of the sounds produced are highly damped revealing very clearly the periodicity of the glottis. In all the nasals there is a small amount of a high frequency component which can perturb the T.I. distributions sometimes.

The T.I. distributions produced by nasals are typified by long T.I.s but, as can be seen in the case of /ŋ/, these can be differently perturbed within the same utterance owing to the presence of high frequency. It was noted that there is an increase in the spread of time intervals recorded in the sequence /m/, /n/ then /ŋ/.

1.6 The effect of variation of the duration of compilation of T.I. histograms.

It was expected that the time taken for the T.I. distributions of different classes of speech sounds to attain statistical stability would vary. The T.I. distribution of fricatives with their very short T.I.s. distributed in a random pattern would be expected to exhibit stability fairly rapidly. Voiced sounds with their longer T.I.s. and the constraint of glottal repetition may need a longer duration.

Experiments were performed to test these expectations in case any departures from them indicated either the need for a change in the measurement parameter of 'duration of compilation', or alternative procedures for analysing voiced and unvoiced sounds in future experiments.

The effect of reduced 'duration of compilation' for histograms of six durations of the vowel /ʌ/ is illustrated in figure 1.14. The histograms of this vowel were found to be fairly typical in this respect. The four peaks evident in the 250 msec. histogram are still present in the 10 msec. histogram although their peak positions are less well defined. Departures from this behaviour were observed only in open vowels. The case of /ɑ/ has already been examined (fig.1.1) where the T.I.distribution is seen to vary slowly with respect to 10 msec. segments. Histogram peaks of such vowels,when measured in 10 msec. segments,are seen to move along the time interval dimension and sometimes disappear. The shorter term variations in the vowel /i/ (also shown in fig.1.1) average out within a duration of this length.

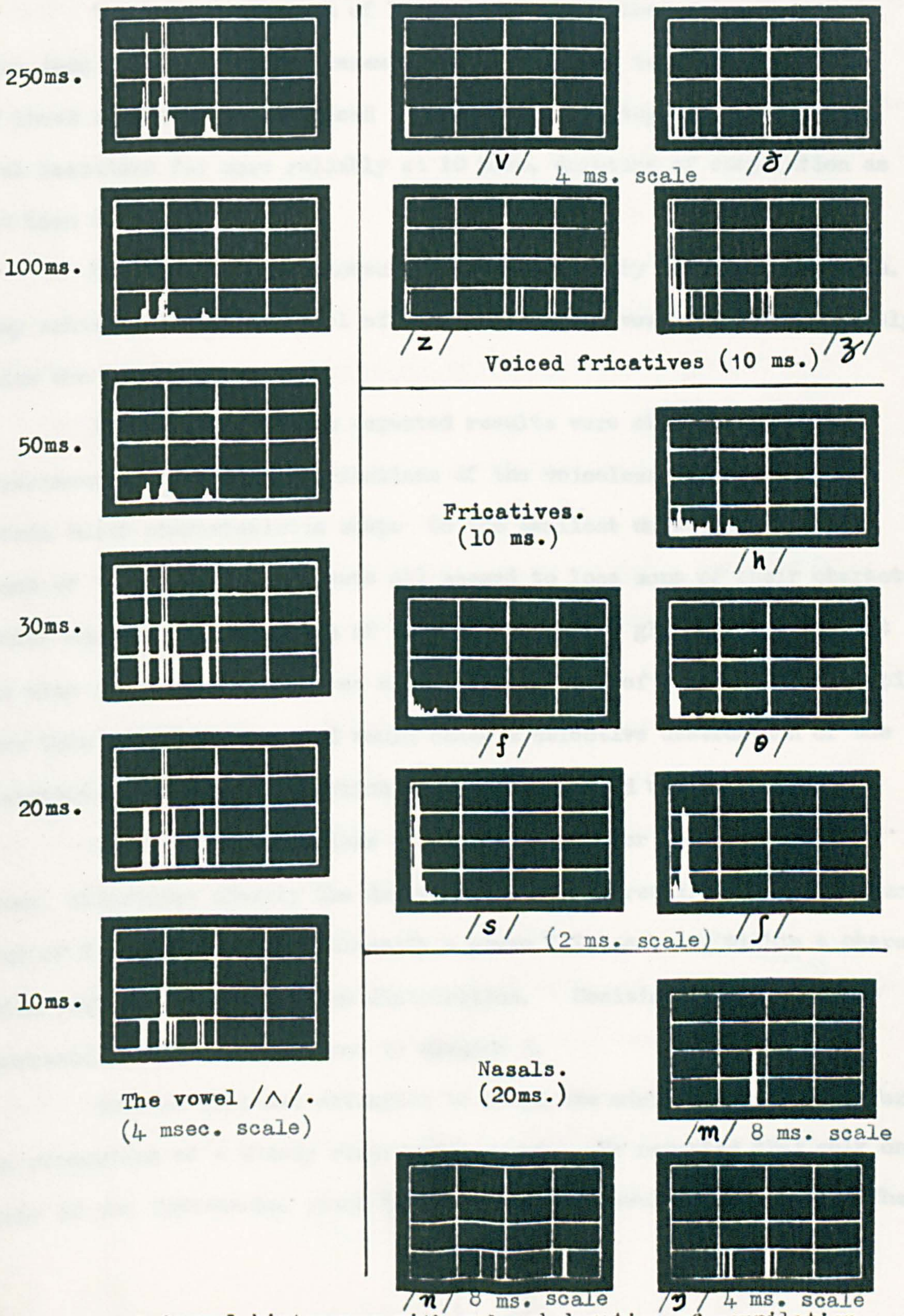


Fig. 1.14 Time interval histograms with reduced duration of compilation.

The T.I. histograms of 10 msec. of the voiced fricatives are also seen to bear a strong resemblance to the long term T.I.statistics of these sounds. The unvoiced fricative T.I. histograms retain their peak positions far more reliably at 10 msec. duration of compilation as had been expected.

The nasal sounds showed the least stability at short durations. They achieved a similar level of stability to the vowels in approximately twice the time.

It is seen that the expected results were obtained in this experiment. The T.I. distributions of the voiceless fricative sounds retain their characteristic shape to the samllest durations measured. Those of the voiced continuants all seemed to lose some of their characteristics when the duration was of the order of a few glottal periods. It was also confirmed in isolated cases that the use of a duration of compilation less than one glottal period would cause a selective destruction of the distribution, depending on which part of the period was included.

The T.I. distributions of voiced sounds for durations 10 - 30 msecs. illustrate clearly the degree of tolerance required in the measurement of T.I.s to correctly classify a given T.I. as being within a characteristic peak of the longer term distribution. Decisions based on these observations will be discussed in chapter 3.

Teacher (65) has attempted to judge the minimum time needed for the perception of a steady state vowel sound. He reported that only one cycle of the fundamental pitch was required for vowel recognition. The voiced utterances used in the duration of compilation experiments described above had a glottal period of approximately 6.5 msecs. The minimum duration over which histograms of these sounds were compiled is equivalent to 1.5 glottal periods. Only a loose comparison can be drawn between the findings of Teacher and the minimum duration of compilation for the stability of vowel T.I. statistics as no information concerning the perception of 'clipped speech' at such durations is available.

1.7 The significance of histograms of time intervals delimited by unidirectional Z.C.s.

Bezdel and Chandler (5) reported that histograms of the time intervals between successive positive going Z.C.s (+ZC), and those of time intervals between successive negative going Z.C.s (-ZC), revealed asymmetry. Experiments in this study confirmed this.

The investigation of these histograms and their comparison with those using all the Z.C.s (±ZC) is essentially a probe into the second order statistics of the ±ZC time intervals. If the ± ZC histogram indicates that N different intervals occur, then the +ZC and -ZC histograms indicate which of the possible sequential combinations of these intervals occur. The possible combinations are the N double length intervals and the N(N-1)/2 mixed pairs of the N original different intervals indicated by the ±ZC histogram. It seems therefore that if the three histograms produced by measuring intervals between ±ZC. +ZC and -ZC are taken together they will include extra information concerning the ordering of the T.I.s. As second order statistics are the subject of the next chapter, the T.I. histograms of +ZC and -ZC will simply be described here.

The T.I. histograms of all the continuants used in the previous experiments were produced for the two unidirectional ZC. conditions. The histograms were compiled over a duration of 0.5 sec. The purpose of this experiment was to look for any asymmetry and observe how it varied for various speech sounds.

The general form of difference that might be expected in the histograms is that the mean interval of each histogram will be doubled, thus distributing the +ZC or -ZC histogram about a 'centre of gravity' on the T.I. dimension of twice that found in the ±ZC case. The more interesting feature will be the change of the histogram shape involved in this process and how this change differs for +ZC and -ZC histograms.

1.7.1 Vowel histograms using unidirectional Z.C.s.

It was expected that the unidirectional Z.C. histograms of vowel sounds would show some asymmetry which could be attributed to the unidirectional form of the excitation function. This may well produce better definition of Z.C.s in one direction at particular points in the glottal period. It is difficult to detect such a tendency from the waveform alone, but this could well be observed as peak sharpening in a T.I. distribution of unidirectional Z.C.s.

The unidirectional Z.C. histograms of the twelve vowels are shown in figure 1.15. It can be seen that although certain histograms seem to have better defined peaks than their contra-directional Z.C.

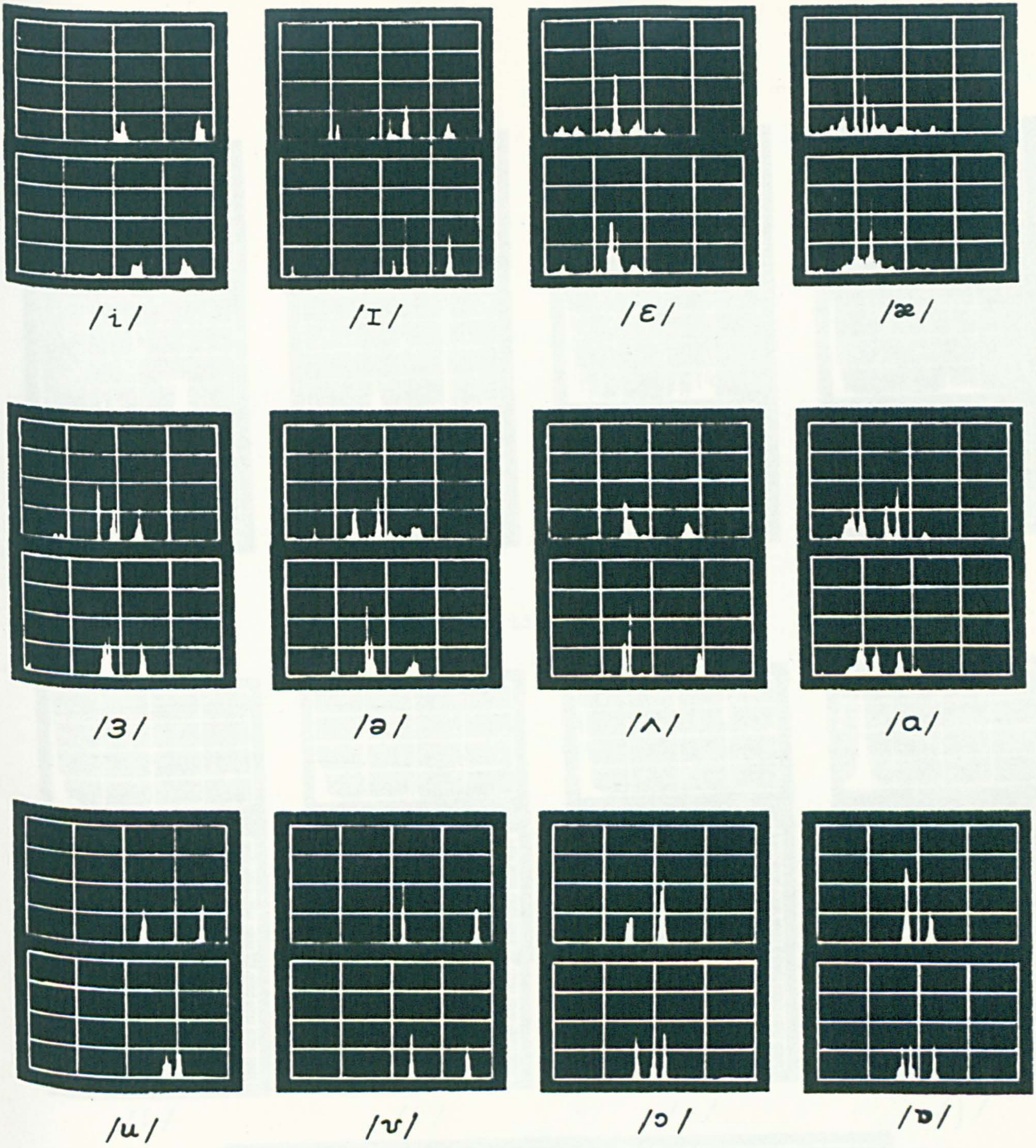
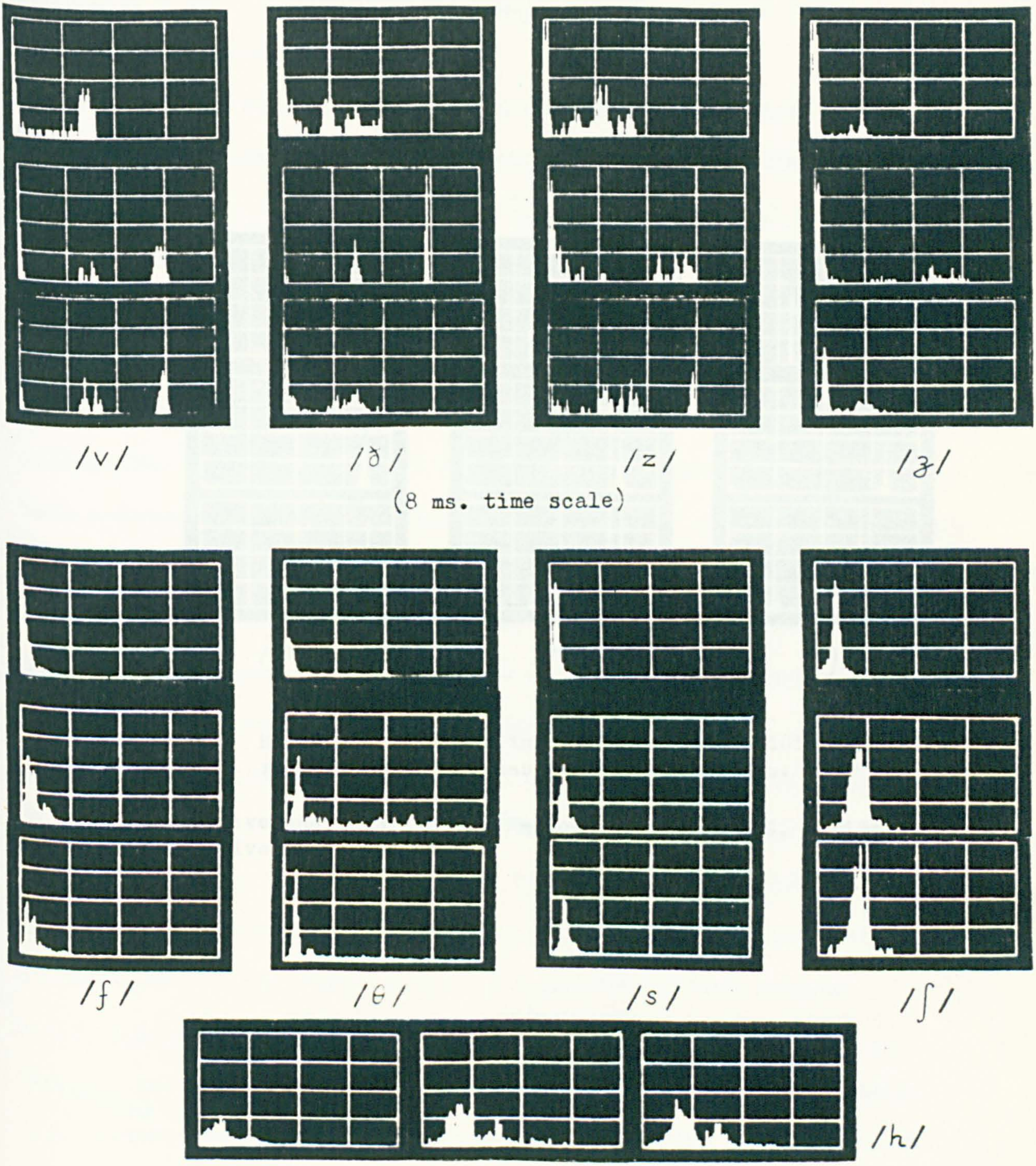


Fig. 1.15 Histograms of time intervals between unidirectional zero-crossings of vowel waveforms. (4 ms. time scale)

(positive Z.C. - above , negative Z.C. - below)



(8 ms. time scale)

(4 ms. time scale)

Fig. 1.16 Histograms of time intervals between unidirectional zero-crossings of voiced and unvoiced fricatives.

(positive and negative Z.C. - top, positive Z.C. - middle, negative Z.C. - bottom)

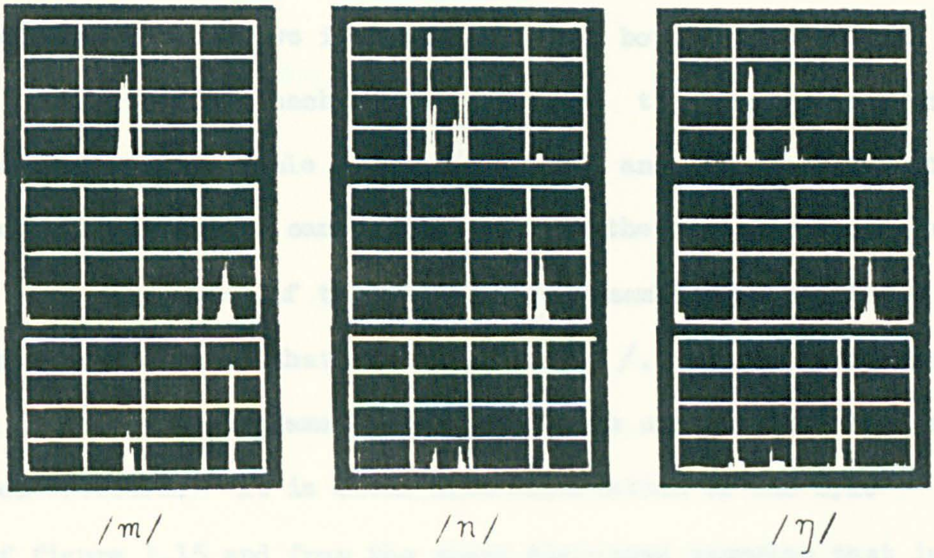


Fig. 1.17 Histograms of time intervals between unidirectional zero-crossings of nasal waveforms. (8 ms. time scale)

(positive and negative Z.C. - top , positive Z.C. - middle, negative Z.C. - bottom) versions, there is no overall pattern to suggest that the above suggested difference is observed .

The variability of order information available from the three histograms is evident from the following two examples. The four characteristic T.I.s of /*i*/, indicated by the ±ZC histogram of /*i*/, combine in sequential pairs to produce two intervals only, in both unidirectional Z.C. histograms. The area under each peak is similar; this indicates that there must be a repetitive cycle of four intervals and the exact formation of this sequence of intervals can be deduced from the three very simple histograms. The narrowness of the peaks allows unambiguous pairing. This case may be compared to that of the vowel / æ /. It is by no means obvious from the three histograms how the intervals are combined in the clipped speech waveform. It is clear from observation of the T.I. histograms of figure 1.15 and from the above discussed examples that in most cases the order information revealed is fairly small. This is due both to the simplicity of the analysis and the complexity of the subject. It should also be noted that in most vowel T.I. histograms it is not possible to derive the ± ZC histogram from the two unidirectional Z.C. histograms. It is therefore necessary to measure all three histograms to obtain the maximum order information possible by this method.

1.7.2 Fricative histograms using unidirectional Z.C.s.

The unvoiced fricatives show no change in the shape of their T.I. distributions when unidirectional Z.C.s are used. The centre of the distribution is simply shifted along the time interval scale by a factor of two. The histograms illustrated in figure 1.16 were compiled over 100 msecs. which allowed for statistical stability to be achieved without running into storage problems in the PDP-8/338 system (see appendix 4) by which the photographs were produced.

1.7.3 Voiced Fricative histograms using unidirectional Z.C.s.

The unidirectional Z.C. histograms of voiced fricatives are also illustrated in figure 1.16. In all the histograms, except -ZC /ʒ/ and -ZC /ð/, there is a peak representing two intervals of length, indicative of the presence of voicing, occurring together. This peak is at approximately 6 msecs., which is the glottal period of the speaker. This feature is strongest in the +ZC histograms. It could be explained if the low frequency component of the waveform had a more rapid +ZC than its subsequent -ZC. This is evident in some of the voiced fricative waveforms. It could also be explained by an uneven fricative amplitude throughout a glottal period.

1.7.4 Nasal histograms using unidirectional Z.C.s.

The unidirectional ZC histograms of the three nasal sounds show similar properties to each other (fig. 1.17). In each case the +ZC histogram contains a peak at an interval equal to the sum of two of the longer intervals in the ± ZC histogram. The only other intervals recorded on the histograms are very short ones. By contrast the -ZC histograms contain intermediate intervals approximately equal to the longer ones of the ± ZC histograms.

This asymmetry would indicate that when the waveform of a nasal sound crosses the zero axis in a positive direction there are several rapid Z.C.s due to a high frequency component, but when in a negative direction there is only a single Z.C.. Such a result gives information concerning the relative damping of the high and low frequency components within the glottal period. The fact that this is so clearly reflected in the unidirectional Z.C. histrograms is due to the simple form of the waveform of nasals with its widely separated components along the frequency dimension. This is an example of how the study of unidirectional Z.C. separately can give added information concerning the waveform.

1.8 Conclusion.

The study of time interval histograms on a qualitative level has resulted in many general impressions being formed concerning its usefulness as a speech sound discriminator. Most of these impressions have been gained from the utterances of a single speaker.

An attempt was made to quantify the impressions gained from viewing T.I. histograms,based on the coincidence of their peak positions. Rules were developed to decide when a peak had occurred and when two peaks were deemed to have coincided. The result of this semi-quantitative analysis revealed that it was quite common for vowels spoken by the same speaker at different times to give histograms whose peaks showed 100% coincidence if the same vowel was uttered. This coincidence never occurred when different vowels were uttered. There was however a large area of indecision where it was not clear whether similar or different vowels had produced the histograms. Deviations from 100% coincidence were also experienced when the same vowels were uttered at different pitches. The result of this study suggested that there are certain characteristics of a vowel's T.I. histogram which are characteristic of the sound itself but in addition there is that which is characteristic of the particular utterance only.

It has been found when considering the effect of pitch on both synthetic and natural vowel sounds that there are certain predictable changes in the T.I. histogram due to this constraint. It is clear that slight differences in pitch, even if formant frequencies remain constant, can cause changes in the T.I. histogram which may cause confusion between vowel sounds.

It seems that the four classes of continuant sounds studied can be distinguished from each other on the following basis. The histograms of vowel sounds typically have several fairly narrow peaks in the region of 0-2000 μs. The voiced fricatives have time intervals of both very short and very long duration, the peaks tending to be wider than those found in the vowel histograms. The unvoiced fricatives are distinctive in their lack of any appreciable number of long intervals. The nasals have very long time intervals with occasional short ones but can be distinguished from voiced fricatives by the width of their peaks. Certain of these broad distinctions could also be made from measurements of the zero-crossing rate.

The possibility of discrimination of the sounds within the four groups has been discussed above with reference to the semi-quantitative analysis performed on the vowels whose histograms are illustrated in figure 1.12,and histograms of spontaneous vowel utterances. This method was very crude and the numerical results are not much more helpful than the qualitative impressions. These results, both qualitative and semi-quantitative, cannot be compared with those of Bezdel and Chandler (5) or Sakai and Inoue (55) until some wider ranging quantitative investigations have been made. However many of the difficulties of measuring and analysing T.I. histograms have been discovered and discussed. A thoroughly quantitative analysis of the first order statistics of vowel sounds will be considered in chapters 3 and 4.

The discrimination within the three smaller groups of sounds can be summarised as follows. Some of the fricative sounds, /ʃ/ and /θ/ being the exception, could be distinguished by their T.I. histograms when taped or spontaneous utterances were presented to the analysis system. The voiced fricatives were more confused but they could sometimes be distinguished by the shape of the short interval peak of the histogram. The only cue for the discrimination of nasals by means of their T.I. histograms was the spread of the long interval peaks, but this was very dependent on speaker.

The fact that a knowledge of the histograms of time intervals between ± ZC, +ZC, and -ZC can give information concerning the shape of the waveform and the second order temporal structure, in certain simple cases, has been considered. The subject of measuring second order statistics will be developed in chapter 2.

Chapter 2. - The qualitative study of second order statistics of the time intervals between zero crossings of the speech wave.

Introduction.

The study of the second and third order statistics of the letters of the alphabet and of English words, as used in English prose, has demonstrated that more information about the probable form of the data from which they are derived can be obtained from them, than from the first order statistics (58).

These digram and trigram probabilities are defined by the equations

$$\sum_{ij} p(i,j) = 1 \qquad \sum_{ijk} p(i,j,k) = 1$$

$$\sum_{i} p(i,j) = p_j \qquad \sum_{ij} p(i,j,k) = p_k$$

$$\sum_{j} p(i,j) = p_i \qquad \sum_{ik} p(i,j,k) = p_j$$

$$\sum_{jk} p(i,j,k) = p_i$$

where i, j and k represent three sequential 'symbols' of a sequence of symbols, and p_i, p_j, and p_k are the first order probabilities of these separate symbols occurring.

When the clipped speech waveform is considered as the subject for analysis it is evident that even at the level of individual T.I.s there is temporal structure. Information concerning this structure could add to the accurate description of the waveform if incorporated in the statistical analysis.

Fourcin (23) found that the information rate required to transmit T.I. digram information was only slightly less than that needed to transmit monogram information. However it was thought likely that certain classes of sounds might contain more digram redundancy than others, and therefore to concentrate on the digram analysis of these sounds might yield some interesting results. Fourcin's results suggest that it is unlikely that there will be an advantage in using digram information for the T.I. description of all phonemes.

This chapter is concerned with the qualitative analysis of the digram structure of the various sounds of speech. A visual display was again used for this purpose. A real time display was developed in conjunction with M. J. Underwood who was responsible for most of the electronic development. This display was based on two patented ideas of Professor D. M. Mackay (42, 43). As this display is unique in its application to speech, it has been used to illustrate some of the basic problems and the variability of the T.I. description of speech, which were omitted in chapter 1.

2.1 Methods of measurement of second order T.I. statistics.

It has already been mentioned in section 1.7 that the comparison of the three types of histograms of the Z.C. intervals will yield some information concerning the sequence of ± ZC time intervals. The three histograms are those produced when every zero-crossing (± ZC), only positive (+ZC), or only negative (-ZC) zero-crossings are used to delimit each time interval. As the intervals being measured in the latter two cases are combinations of sequential pairs of ± ZC intervals, the sequence information is clearly second order.

By combining discussion of this measure of second order information with that of further measures of the same information, it is hoped to answer the question left open by Bezdel and Chandler (5).

"As the effect of asymmetry in the speech wave was not known, positive, negative and total zero-crossing distributions were analysed separately. .... it was noticed that distinct asymmetry does exist in certain channels. If this feature is to be used to advantage, further investigations will be required, as it is far from obvious how this additional information can be used to improve recognition scores".

These further measures were in fact two implementations of the same idea. It was to display the second order T.I. statistics in a two dimensional array, such that the coordinates of a point in the display defined a sequential pairing which existed in the T.I. pattern (42). A similar display has been used by Fetz and Gerstein (19) for the display of the second order statistics of neural spike intervals.

One of the methods of compilation of such a display is very similar to that used in compiling the unidirectional Z.C. histograms. The T.I. pattern of a speech wave is processed by extracting pairs of sequential intervals and using them to define the coordinates of a point in the 2-D display, or using their sum to define a bin in the unidirectional histogram. In the latter case the two histograms are compiled by two of these processes going on in parallel but out of step with each other by one ± Z.C. time interval. A similar parallel process must operate in the compilation of the 2-D display to ensure that every sequential pairing is included and not every other one.

However, whereas the 2-D display describes precisely the intervals that are involved in each sequential pairing, the three histograms give only the probability of such intervals being paired. For example, if a distribution peak at Υ μsecs. occurs in the unidirectional Z.C. histogram, this only indicates that two intervals whose sum is Υ μsecs. occur together with a certain probability. If there are more than one pair of intervals giving peaks in the ± ZC histogram whose sum is Υ μsecs, then the actual pairing can only be described in more remote probability terms which are based on the probabilities indicated by all three histograms. It is therefore clear that a display of the 2-D form will preserve more of the second order information than the three histograms. The asymmetry between the unidirectional Z.C. histograms mentioned earlier (section 1.7) will always be present except in certain very simple T.I. patterns. (e.g. where the basic repetition cycle consists of two different intervals grouped separately in groups containing an odd number of intervals - e.g. 1 long + 1 short; 3 long + 1 short; 3 long + 3 short, etc.)

In proceeding with the analysis of the 2-D representation of the second order statistics, the question left open by Bezdel and Chandler was being followed up in a more powerful way than that immediately suggested by their question.

2.1.1 The parallel time base system.

Two forms of electronic circuits were used to investigate the second order T.I. distribution by means of a 2-D display. The first was a simple extension of the time base which was reset by each Z.C. pulse, as used to produce the intervalgram patterns. The second was an analogue shift register. (43).

A functional diagram of the time base system is shown in figure 2.1 and its detailed circuit in appendix 2. A parallel arrangement of two time bases was used. They were driven by the out of phase outputs of a bistable circuit, which in turn was triggered by the Z.C. pulses. Each time base could provide a variable exponential run-down, and was so arranged that after one pulse had initiated the run-down, the next pulse terminated it and caused the output to remain at a level which indicated the magnitude of the interval that had elapsed. The outputs were applied as X and Y deflections of the spot on a C.R.T. whose brightness was suppressed except for a brief interval immediately before each time base was reset for a fresh cycle.

Typical displays of vowel sounds using this system (fig.2.2) exhibited an axis of statistical symmetry about the X=Y line. This is explained by the fact that the occasional very short T.I. is missed during the dead time of the circuit. This gave roughly even chances that a particular interval would be represented by an X or Y deflection. This observation points to a basic drawback of this system. It is never known whether a given non-axial point on the display represents a short

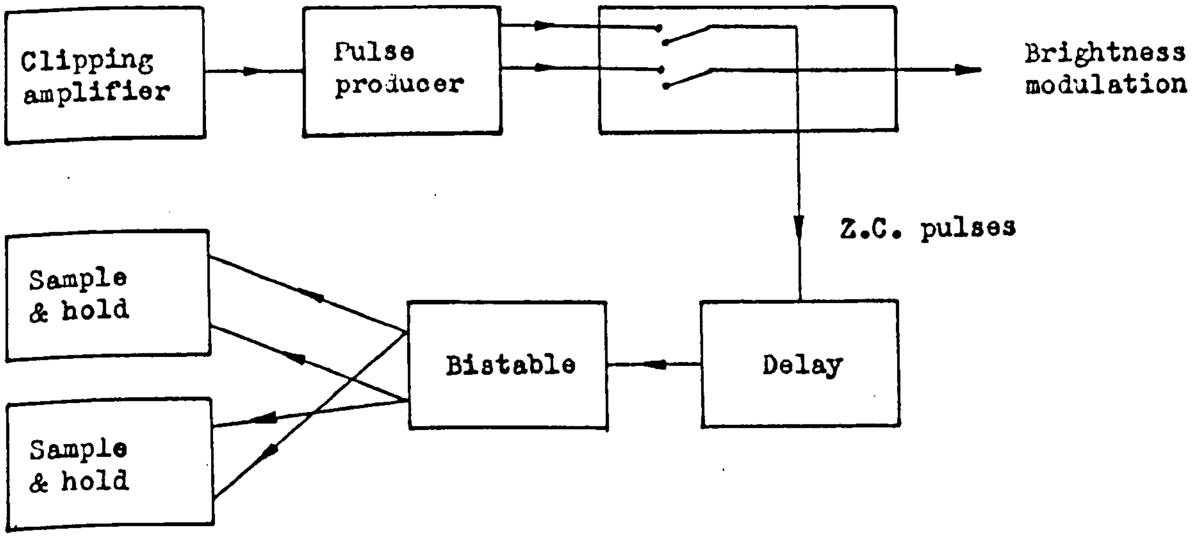
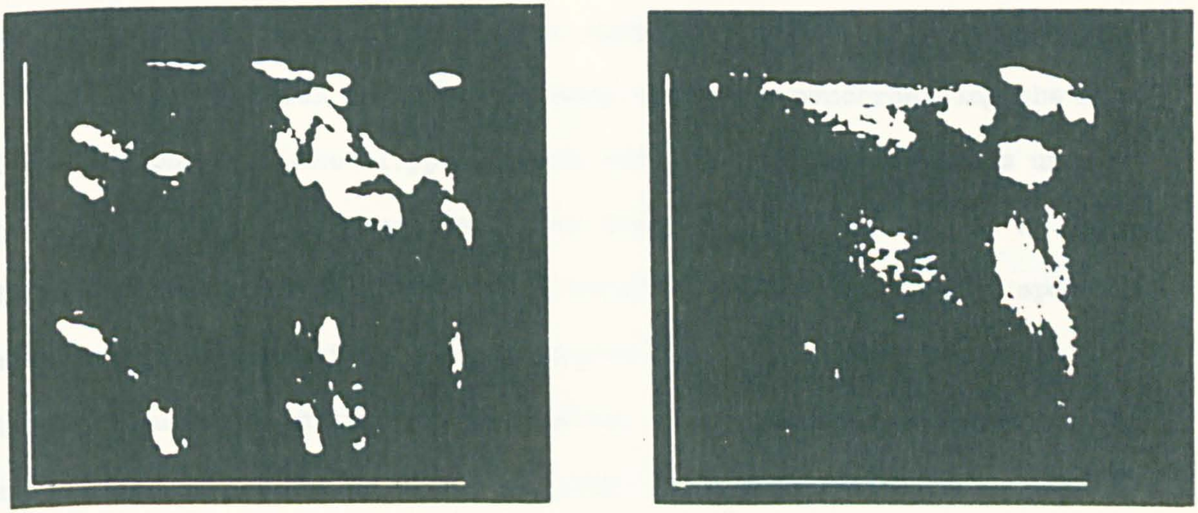


Fig. 2.1 A funotional diagram of the parallel time-base system.



(a) the vowel /æ/ (b) the vowel /ɑ/

Fig. 2.2 Typical displays using the parallel time-base system.

interval followed by a longer one or vice versa. This display simply indicates that a certain pair of intervals are adjacent . If the start of compilation of statistics was displaced by one interval the display would be inverted about the line X=Y. As the repetitive waveform of a vowel sound must contain an even number of intervals, one or other of the two possible patterns will appear. Consistent and symmetrical patterns will only appear when a fairly large number of single Z.C.s are missed.

This problem could be overcome either by synchronising the X and Y deflections with the clipped speech wave, displaying intervals under positive waveform lobes against those under negative lobes; or by introducing a time interval filter which would not allow the clipped speech to change sign for less than a time interval τ. τ may then be chosen to be larger than the dead time of the system. Both these precautions would ensure that no inverting of the display, about X=Y, occurred. Such a display still would not give any information about the direction in which ordered pairs occurred.

An experiment of playing clipped vowel sounds backwards was performed. Informal listening tests revealed that there was no change in its intelligibility. Even if the overall direction of the speech T.I.s in time is unimportant, this does not preclude the fact that long-short and short-long pairs occurring in different parts of the speech wave may be important to a specific distinction between two vowel sounds owing to their relative ordering.

2.1.2 Analogue shift register system.

It was decided that the 2-D display would contain more relevant information if the axes reliably represented the 'nth' and the 'n+1th' intervals respectively. In order to achieve this,an analogue shift register (A.S.R.) was used. (43). This was designed by M.J.Underwood. The functional diagram of this is shown in figure 2.3 and its detailed circuitry is described in appendix 2. The Z.C. pulses of the speech wave were used to reset the single time base via a chain of flip-flop delays. Each input pulse causes the sample and hold circuit S/H - 2 to sample and hold the output of a similar circuit S/H - 1. The first delayed pulse causes S/H - 1 to sample and hold the output of the time base, the second delayed pulse resets the time base. If the X reflection of a C.R.T. is driven by S/H - 2 and the Y deflection by S/H - 1,the required display will be produced. The horizontal deflection will be proportional to the nth interval and the vertical deflection proportional to the n+1th interval.

This method of display has additional advantages over the parallel time base method. The voltages representing the intervals are constantly available during the measurement of a new time interval and the timing of brightness modulation is less critical. The simple histogram may be read off as the projection of the display on to either axis rather than the projection on to the X=Y line as in the parallel time base system.

The range of the time base was controlled by a switch and a potentiometer. Although basically exponential in design it was possible to switch between a variable nearly linear run-down and a more truly exponential run-down. The latter was most useful for viewing the real

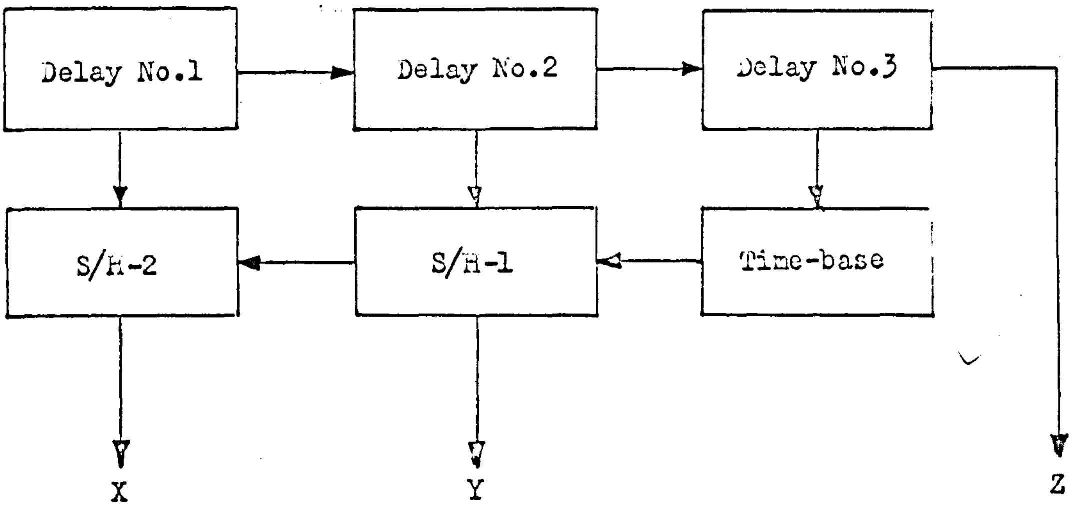


Fig. 2.3 Functional diagram of the Analogue Shift Register.

time display of continuous speech.

This form of 2-D display of second order T.I. statistics was used throughout the work to be described in this chapter and was termed the 'digram display' or simply the 'digram' of the T.I.s.

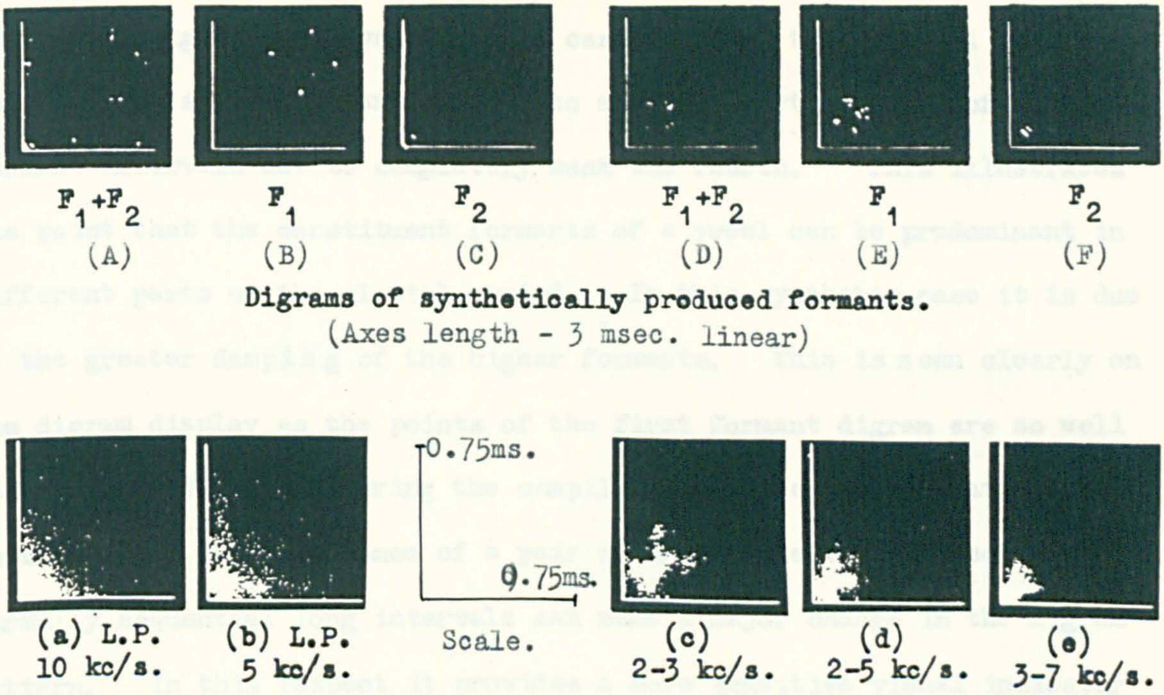
2.2 Digram displays of control stimuli.

Digram displays of familiar signals and of synthetic parameters of speech are presented to familiarise the reader with the basic capabilities of the display as distinct from the simple histogram.

A sine wave input gives a single point on the display. The distance of this point from the origin or from either axis is a measure of the frequency of the sine wave. If the frequency of the sine wave is swept from a low to high frequency the dynamic display shows the movement of the spot along the X=Y axis of the display, towards the origin.

A white noise input gives a diffuse pattern of spots whose shape can be modified by bandlimiting the noise source by low pass and bandpass filters. The digram patterns for several such band limitations are shown in figure 2.4.

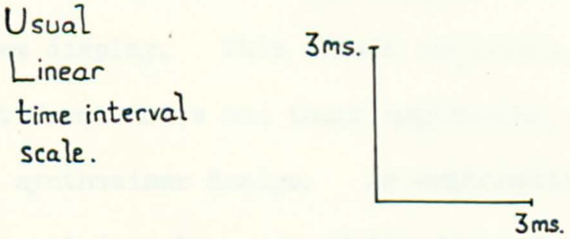
Digram displays of synthetic parameters of speech are also illustrated in figure 2.4 Figures 2.4 A and D show the digrams of a two formant synthesis of / i / and /ɜ/ respectively, while figures 2.4 B & C, and figures 2.4 E & F are their respective first and second formants. The possible changes in the T.I. distribution of a major formant when perturbed by the presence of a minor formant have been reviewed in section 1.4.8. However one or two further points are obvious



Digrams of synthetically produced formants.
(Axes length - 3 msec. linear)

Digrams of bandlimited white noise.

Fig. 2.4 Digram displays produced by control signals.



from these digram displays. In the case of /i/, the interval characteristic of the second formant is seen to slightly perturb three of the first formant intervals but to completely mask the fourth. This illustrates the point that the constituent formants of a vowel can be predominant in different parts of the glottal period. In this synthetic case it is due to the greater damping of the higher formants. This is seen clearly on the digram display as the points of the first formant digram are so well separated. When considering the compilation of the digram statistics it is clear that the occurrence of a pair of short intervals between two normally sequential long intervals can make a major change in the digram pattern. In this respect it provides a more sensitive visual indicator of the presence of a minor high frequency component than the simple histogram does.

In the case of /ɜ/ the perturbation of the first formant Z.C. pattern by the weaker second formant can be seen as a rotation of the first formant digram display. This effect indicates the relationship between the formant frequencies and their amplitudes, their phase relations being fixed by the synthesiser design. An explanation of this effect is that the second formant is a harmonic of the first formant of such an order that every other Z.C. is unperturbed when the second formant is added, but the one in between is disturbed to a degree proportional to the rotation observed. This is proportional to the relative amplitudes of the formants which are seen to be such that no intervals characteristic of the second formant alone are produced.

This point is made as the rotation effect has been observed in the dynamic display of strings of voiced phonemes. This could therefore be interpreted as the uneven onset of harmonically related formants.

2.3 Digram displays of continuant sounds.

The digram display of speech sounds was first of all viewed as a real time dynamic representation. Many interesting features concerning the movement of parts of the display were noted, similar to the one mentioned in the previous section. It was immediately obvious that to attempt to analyse the display in this dynamic form would be a large task in whichever way this was attempted. An objective approach of continually monitoring individual points or large features in the display would need a large amount of information storage,possible only with a sizeable digital computer. A subjective approach of discovering which features of the display were important to a human subject,attempting to 'read' it as 'visible speech', would need the training of a psychologist. This latter approach has since been followed up by I. K. Taylor.

It was decided to proceed with analysis of the digram display in a similar way to the previous histogram analysis. Digrams of the continuant sounds specified earlier were produced on a C.R.T. and photographed for comparison with one another. In this section the effects of pre-clipping differentiation will be examined together with the variations in the digram display due to change of pitch and speaker.

2.3.1 Digrams of vowel sounds.

The question, "Do vowel digrams discriminate between the vowels better than the vowel histograms?", cannot be answered from the present qualitative analysis. The more specific question, "Can two utterances with similar histograms be discriminated better by their digrams?" can be answered by the comparison of the two digrams and of the projection of each digram on to a single axis. An affirmative answer to this latter question would indicate that the different ordering of a similar distribution of T.I.s. could be the basis for the subjective discrimination of the sounds.

A single example of such a case is illustrated in figure 2.5. These two vowels have very similar histograms (as seen in the projection of the display on to a single axis) yet the ordering of the intervals is markedly different. The extent to which such an isolated and clear cut case is significant will be seen in the following sections dealing with the variability of the digram display.

The digrams of the vowels whose histograms were discussed in the previous chapter are illustrated in figure 2.6. The format of this figure is in accordance with the rough positions of the vowels on the cardinal vowel chart (31). These positions were determined by Miss D. Scott, a student of speech therapy. This format bears a close relation to the Fl/F2 plane already used for the illustration of vowel histograms. The front and close position corresponds to a high F2 and low Fl, the open position to high Fl,and the back and close position to low F2 and low Fl.

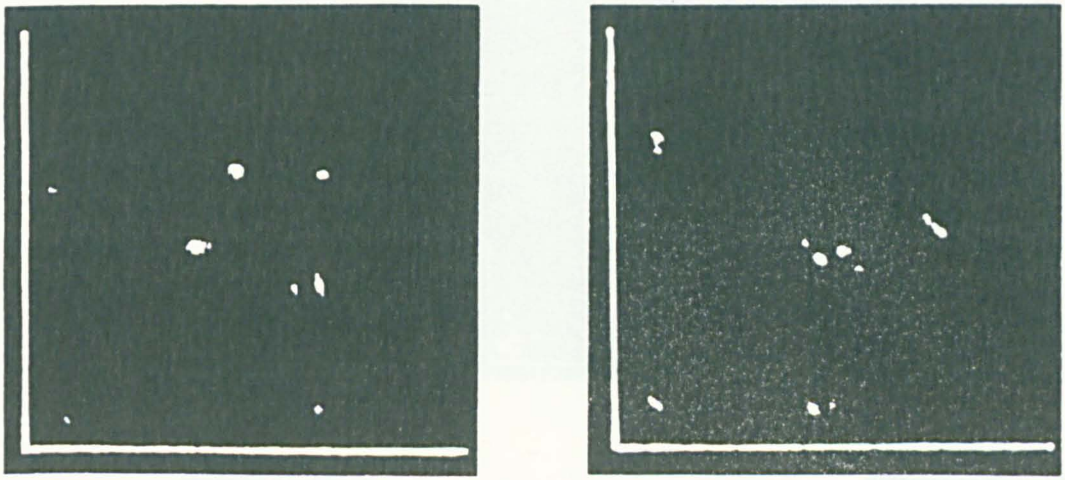
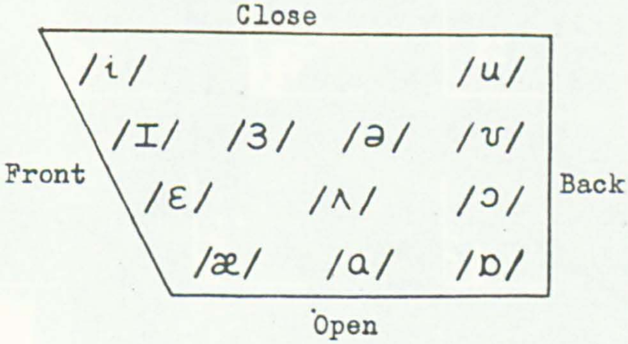


Fig. 2.5 Time interval digrams of vowels /i/ and /I/.

Twelve English vowels on the vowel quadrilateral : the format used for most of the future digram illustrations.



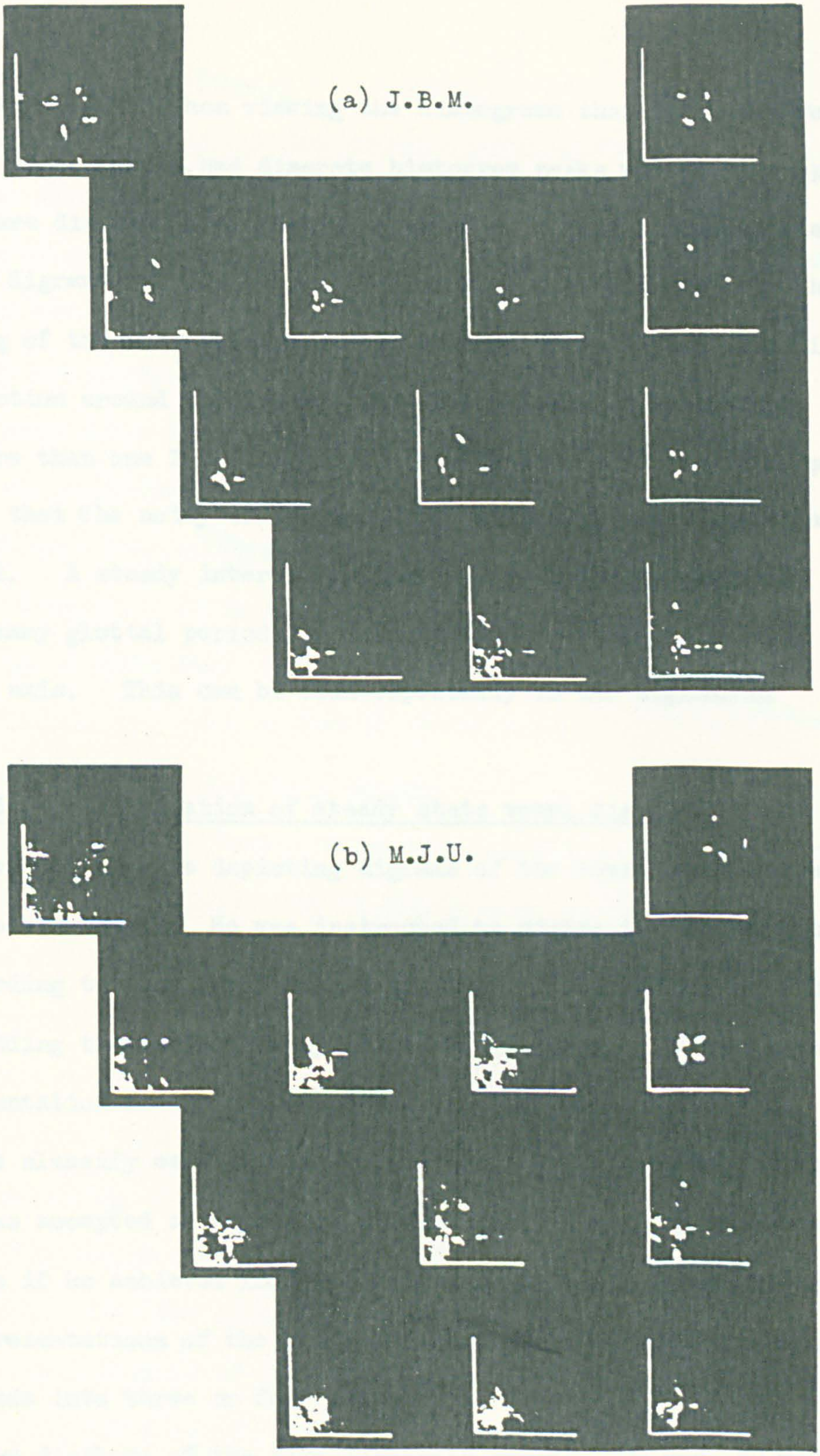


Fig. 2.6 Vowel digrams of speakers J.B.M. and M.J.U.
(axes length - 3 ms. lineær)

It was observed when viewing the histograms that certain vowels, especially the close vowels,had discrete histogram peaks whereas the open vowels had a more diffuse noisy looking histogram. This trend can also be seen in the digrams of figure 2.6a. From some of these displays the actual ordering of the intervals is easily deduced by moving in an anti-clockwise direction around the typical four dot pattern. Ambiguity arises when more than one T.I. of the same length occurs in a glottal period. It can be seen that the noisy effect is often related sequentially to a single interval. A steady interval followed by an unsteady one,when averaged over many glottal periods,produces dots in a straight line parallel to an axis. This can be seen especially in the digrams of /ɑ/ and /ɒ/.

2.3.2 Subjective classification of steady state vowel digrams.

Twelve photographs depicting digrams of the vowel sounds were presented to each subject. He was instructed to divide the twelve into subgroups according to some identifiable features of the patterns, with the aim of dividing them into as many groups as possible. He was then tested by presentation of the twelve photographs in random order. He was required to classify each one according to his prior grouping of them. His grouping was accepted as a measure of the discriminability of these static displays if he achieved 100% recognition within his groups after two complete presentations of the set. Most subjects managed to classify the twelve sounds into three or four groups.

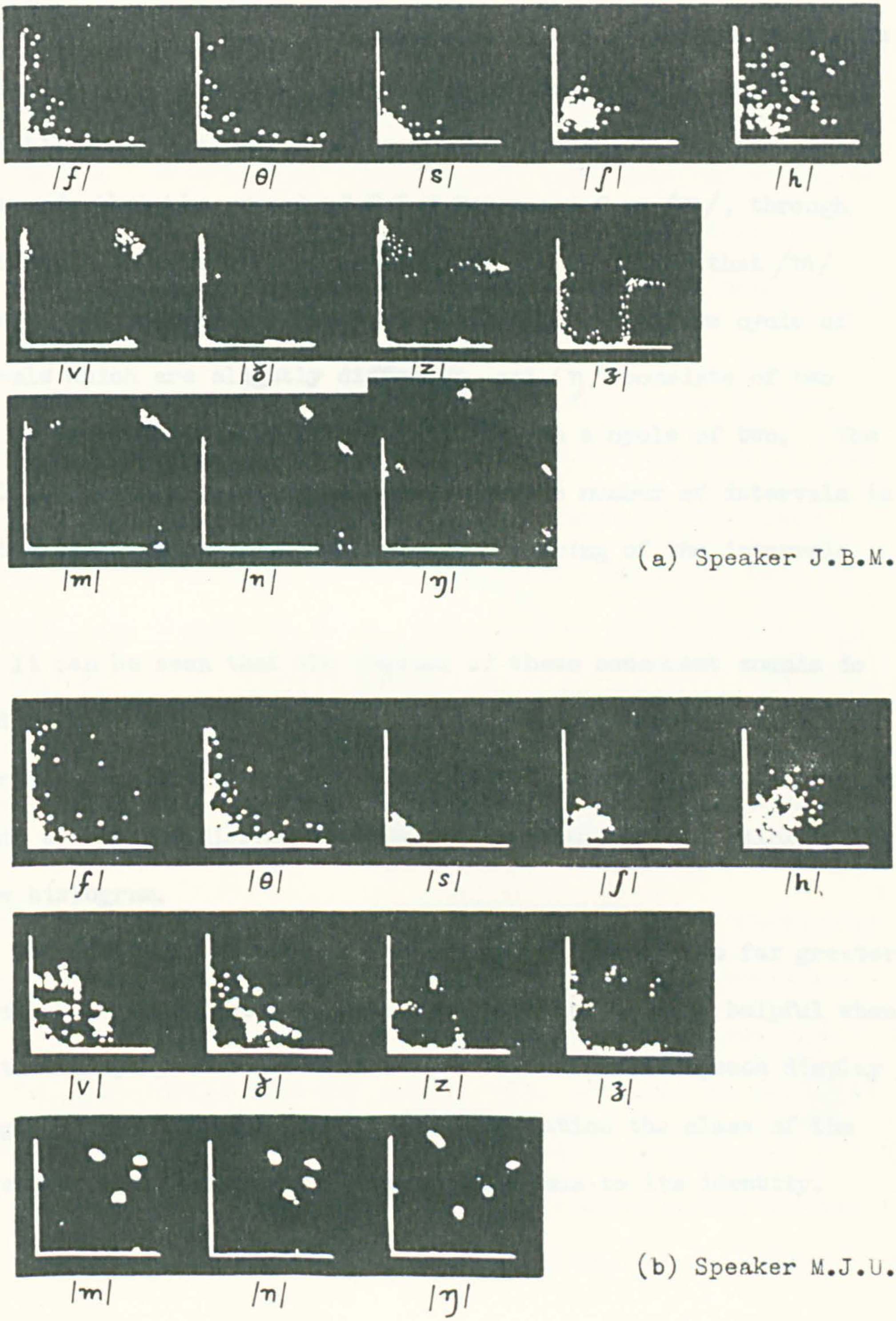
Digram displays of the time intervals between +ZC and -ZC were also presented. The results were 5 - 6 and 4 - 5 groups respectively. This apparent improvement is now discounted following studies of the effect of pitch on the vowel digram. In these combined interval cases the number of intervals per glottal period was at times reduced to two, making the sample used extremely pitch dependent. This is also a factor in the case where ± ZC are measured, but to a smaller extent.

The conclusion of this experiment is that the subjective impression of the experimenter is confirmed by the majority of the subjects: that only four categories of vowel can be discriminated by this method of display. It is admitted that certain extra cues might be present in the effects that pitch jitter has on each vowel digram if the real time display was used instead of photographs.

2.3.3 Digrams of fricative and nasal sounds.

The digrams of the twelve consonant continuants considered in the previous chapter are illustrated in figure 2.7. The voiceless fricatives have a T.I. axis expansion of X4, with the exception of /h/ which has an expansion of X2. It can be seen that there is very little extra information presented by the digram for these sounds.

The voiced fricative digrams are more interesting in that the degree of intermingling of the very short and the very long intervals can be seen. The length of the long intervals which do occur sequentially is seen to decrease with the ordering of phonemes used. This fact is not available from the histogram of the same utterances where such a trend could be detected only in the case of /ʒ/. It was therefore simply regarded as



(a) Speaker J.B.M.

(b) Speaker M.J.U.

Fig. 2.7 Fricative,voiced fricative and nasal digrams.

an isolated difference.

The digrams of the nasals show that the predominant structure is long intervals paired with long intervals. The perturbation of this pattern due to short intervals is in all cases very slight, there being no appreciable short with short pairing. It was seen in the histograms of these sounds that the spread of T.I.s increases from /m/, through /n/ to /ŋ/. It is very clear from the digram structure that /m/ consists of similar intervals, /n/ consists of a repetitive cycle of two intervals which are slightly different, and /ŋ/ consists of two intervals of greater difference also repeating in a cycle of two. The additional information given by the digram is the number of intervals in a repetitive cycle which is evident from the pairing of the intervals observed.

It can be seen that the digrams of these consonant sounds do give added information concerning the latter two classes of sounds and reveal certain trends within them. Inasmuch as these trends discriminate between the sounds, the digrams enhance the discrimination possible using the simple histogram.

The distinctions between the classes of sounds are far greater than those within the classes. This fact could prove very helpful when the real time display is considered as part of a visible speech display presenting continuous speech. In such a situation the class of the sound together with its context gives a major cue to its identity.

2.3.4 The effects of duration of compilation on vowel digrams.

The digrams of vowel sounds reveal a structure which is more pronounced than in most other speech sounds. Investigations were made to compare the temporal stability of this structure with that of the simple histogram. Several investigations were made into the effect of reducing the duration of compilation of the digram on the measurements or observations being studied at various times during the work described in this thesis.

Early in the study of the digram,small differences in the digram of the steady vowel were seen typically between 200 and 500 msec. duration of compilation. This was reported by Mackay, Millar and Underwood (44). It was later realised that these variations were due to pitch variations similar to those illustrated by the running histogram of /a/ (fig. 1.1).

The variations in the digram structure,when compiled over successively shorter periods of the same utterance,are shown in figure 2.8. Three samples using the shortest duration of compilation are given, as they would be expected to show the greatest variation. Histograms of this same utterance at the same duration of compilation were found (section 1.6) to exhibit intervals under all the peaks oflonger term histogram for every sample that was taken. The variation in the three digrams shown is similar to that experienced with the histograms. There are spots on the 10 msec. digram at all the major points evident on the 500 msec. digrams, but where the points on the 500 msec.digram

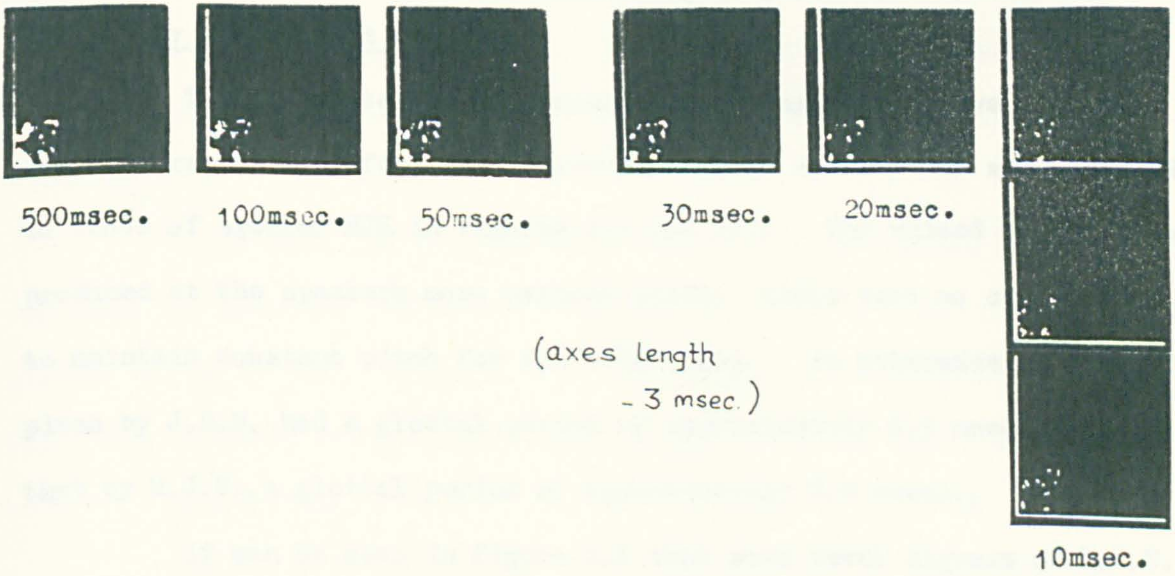


Fig. 2.8 Variation of a vowel digram with the duration of compilation.

are large, there is a correspondingly large variation from one 10 msec. digram to another.

It was concluded that there is not a gross difference between the effect of duration of compilation on the digram statistics and its effect on the histogram statistics. It can also be deduced from the results of this experiment and from observations of vowel waveforms that, in a steady vowel sound, the majority of T.I.s and their second order structure remain constant from one glottal period to the next.

2.3.5 The variation of the digrams of utterances of similar sounds by a different speaker.

To illustrate the difference in the digram displays produced from utterances by different speakers, those of speaker JBM are compared to those of speaker MJU in figures 2.6 and 2.7. The voiced sounds were produced at the speaker's most natural pitch; there were no constraints to maintain constant pitch for all utterances. An utterance of natural pitch by J.B.M. had a glottal period of approximately 6.5 msecs., and that by M.J.U., a glottal period of approximately 8.0 msecs..

It can be seen in figure 2.6 that some vowel digrams of M.J.U. resemble more nearly the digram of a different vowel spoken by J.B.M. and vice versa. This can be seen especially in the cases J.B.M. /ɒ/ and M.J.U. /ɔ/; J.B.M. /u/ and M.J.U. /ʊ/; and J.B.M. /ʊ/ and M.J.U. /u/. Note that all these pairs of vowels are adjacent on the cardinal vowel plot. In most of the other vowels the M.J.U. digrams appear more noisy and reveal that more short intervals are present.

However, some features characteristio of the vowel rather than the speaker can be seen. The three central vowels /ə/, /ɜ/ and /ʌ/ of speaker M.J.U. contain the simpler patterns of their counterparts spoken by J.B.M., in the midst of noise.

A similar comparison of the digrams of unvoiced fricatives (fig. 2.7) shows very little difference between the digrams produced by both speakers. The voiced fricatives show differences which are similar to those seen in some of the vowel comparisons. The maximum interval is shorter in the M.J.U. digrams than in the J.B.M. digrams and there are more different intervals present in a rather noisy array. Additional but more well defined intervals are seen in the M.J.U. digrams of the nasal sounds.

It is possible that most of the differences between the digrams of these two speakers can be explained on the basis of the difference in natural pitch of their voices. The similarity of the digrams of sounds which are unvoiced and therefore have no glottal pitch would be compatible with this explanation.

The noisy appearance of some digrams could be due either to pitch jitter or to noise present with the speech. In the latter case the Z.C.s of the speech wave will be perturbed maximally when the speech amplitude is low. This happens to a certain extent in each glottal period as the resonances decay. If the pitch is low and the glottal period long, the amplitude decay is most marked. Thus in this condition the proportion of Z.C.s that are likely to be perturbed by noise is greater than when the glottal period is shorter. Other differences could be caused by the shortness of the glottal period perturbing long time intervals. Such perturbation would be seen in vowels in the back-close position of the vowel chart. It is therefore noted that the simplicity of the digrams of vowels spoken by J.B.M. does not of itself add weight to their reflection of specifically vowel, rather than speaker, or pitch characteristics.

2.3.6 The effect of differentiation of the speech wave on vowel digrams.

The studies on the intelligibility of clipped speech by Lichlider and Pollack (40) and Ainsworth (2) show that if the speech wave is differentiated with respect to time before clipping,a greater proportion of the intelligibility of the original speech is retained. This suggests that better discrimination between speech sounds may be possible if the T.I. statistics of differentiated speech are used. The process of differentiation will, as seen in the introduction, mean that the T.I.s measured will be those between successive maxima (- ZC.s) and minima (+ ZC.s). The transformation to convert these extrema into Z.C.s. involves a frequency emphasis of +6 db./octave. Such preprocessing was expected to give greater weighting to the T.I.s descriptive of the second and higher formants.

The effect of differentiation on the digram form of the T.I. statistics of vowel sounds is shown in figure 2.9. The utterances used were the same as those whose digram statistics are shown in figure 2.6. The T.I. scale in both these figures is the same so that a comparison of

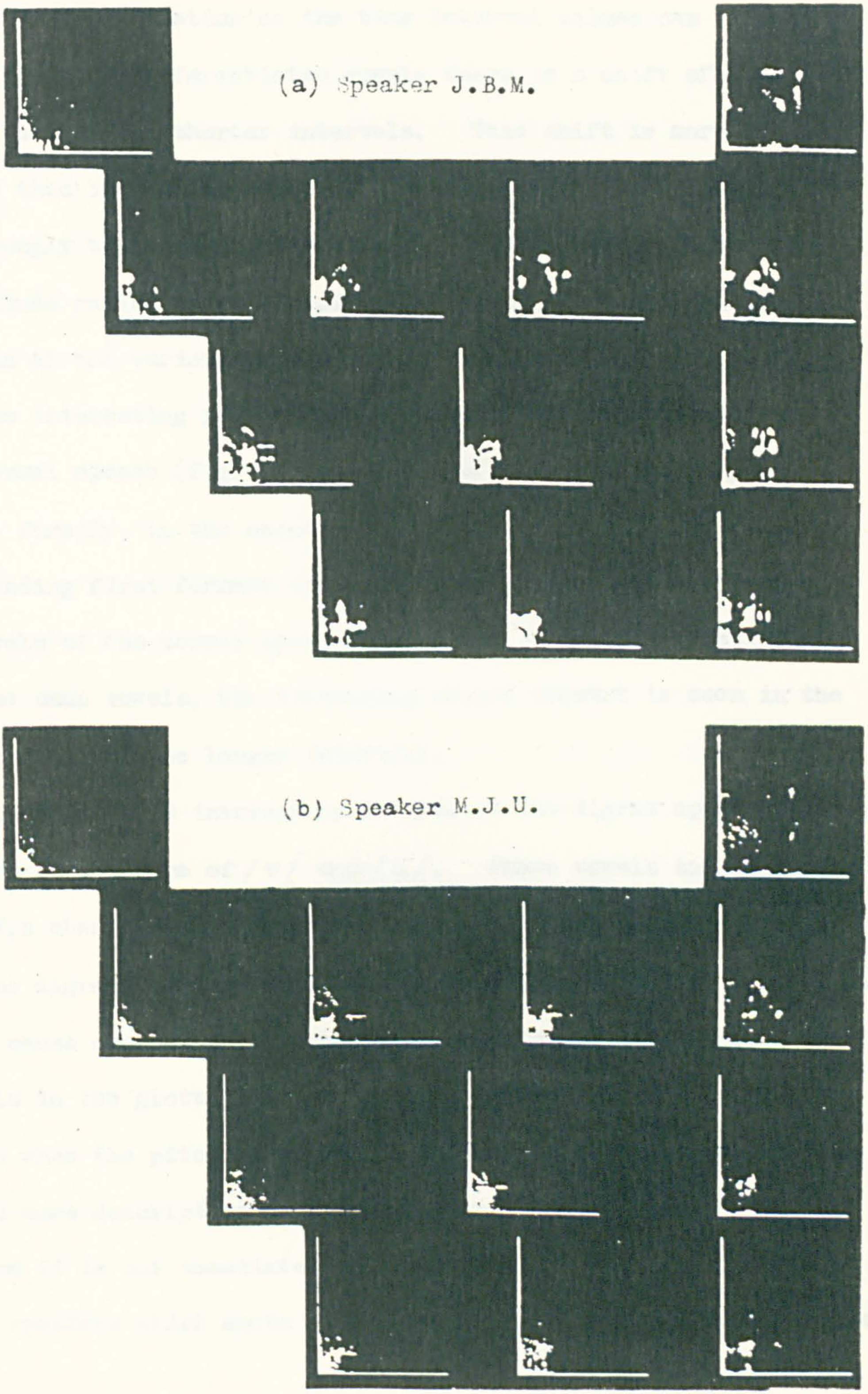


Fig. 2.9 Digram displays of differentiated vowel waveforms.

the effects of differentiation on the time interval values can be made. In all the digrams of differentiated vowels there is a shift of the distribution towards the shorter intervals. This shift is more marked in some cases than in others. The back vowels retain their long intervals more strongly than the front vowels. This is due to different formant amplitude ratios and the separation between formants which are characteristic of the various vowels.

Some interesting points can be observed by comparison of the digrams of normal speech (fig. 2.6) and those of differentiated speech (fig. 2.9). Firstly, in the sequence of vowels /i/, /I/, /ε/ and /æ/, an ascending first formant is seen in the descending magnitude of longer intervals of the normal speech digram. In the differentiated version of the same vowels, the descending second formant is seen in the ascending magnitude of the longer intervals.

Secondly, it is interesting to examine the digram appearance of the differentiated form of /ʊ/ and /u/. These vowels had very pitch dependent T.I.s when a flat spectrum was used. It was noted in section 2.3.5 that the digrams of these two vowels as uttered by M.J.U. and J.B.M. could easily cause confusion. It has also been noticed that two of the four intervals in the glottal period of these vowels are more stable than the other two when the pitch is varied. This suggests that these stable intervals are more descriptive of the vowel concerned. When normal speech is used it is not immediately obvious from the digrams of /ʊ/ and /u/ of both speakers which spots represent intervals which are stable, as all the spots are well defined. The digram of /u/ spoken by M.J.U. shows longer intervals than that of J.B.M. and only one spot occupies the same position on both digrams. It is this spot which is retained by the digram of the differentiated form of the M.J.U. utterance. The differentiated form has removed some of the speaker and pitch dependent features of the display. However, in doing so, it introduces many short intervals which complicate the display.

Such results as these suggest that an intermediate type of waveform preprocessing may be called for. If the normal speech signal was mixed with the differentiated signal in variable proportions, some optimum position might be found where high and low frequency components of the speech waveform contribute equally to the T.I. distribution. This was done using the simple mixing facility already mentioned (section 1.1.1). It was found that for individual vowels a balance between high and low frequency components could be achieved. However, this balance position was different for each vowel and it was not clear how to derive a general rule for its use. The only system which was conceived,but never implemented,was one using feedback from a measure of the Z.C.rate to control the mixing proportions. Such a system would cause a change in the mixing ratio until the Z.C. rate attained some predetermined value in the middle of the vowel T.I. range. There would however still be problems in extreme cases such as /u/.

2.4 Some changes in instrumentation.

2.4.1 The use of the Language Master.

The Bell and Howell Language Master (see appendix 3) was used for most of the recorded speech sounds in the latter part of the qualitative studies and throughout the quantitative studies. The advantages of this device were many. It replaced the awkward tape loop previously necessary when a particular sound was required many times. It enabled such sounds to be literally filed away and thus be randomly accessible rather than being in the middle of a spool of tape. It enabled random orderings of stimuli for listening tests to be reordered simply by the shuffling of the cards.

The only disadvantage was its restricted frequency response (fig.A.3.1). The low frequency deficiency did cause the ratio of first to second formant amplitudes to be modified. Certain vowels with low first formants, such as /i/ and /I/, were reproduced with waveforms whose T.I. patterns were governed more by the higher frequencies than would be expected from a device with a wider flat pass band.

The intelligibility of the reproduced waveforms, when clipped, was not impaired by this frequency response, as is shown by the results of Ainsworth (2).

2.4.2 The use of low pass filtering to remove high frequency noise.

A variable low pass filter was acquired and inserted in the preprocessing circuit to remove high frequency noise which was outside the band useful to the digram display. It was found that for vowel utterances recorded on the language master, 2.5 Kc/s was the lowest cut-off point for which no effect on the digram display of any vowels could be seen. This check was made for the more stringent condition of differentiated speech and the same cut-off value was used for normal and differentiated preprocessing. If the low pass cut-off was reduced below 2.5 Kc/s the digrams of the differentiated versions of /i / were disturbed.

2.5 The digram analysis of utterances of controlled pitch by two speakers.

Several of the differences between the digrams of the two speakers M.J.U and J.B.M. have been explained in terms of the difference between the natural pitches of their voices. It was the aim of the following experiments to discover whether two speakers could produce the same or more similar digrams by controlling the pitch of their utterances.

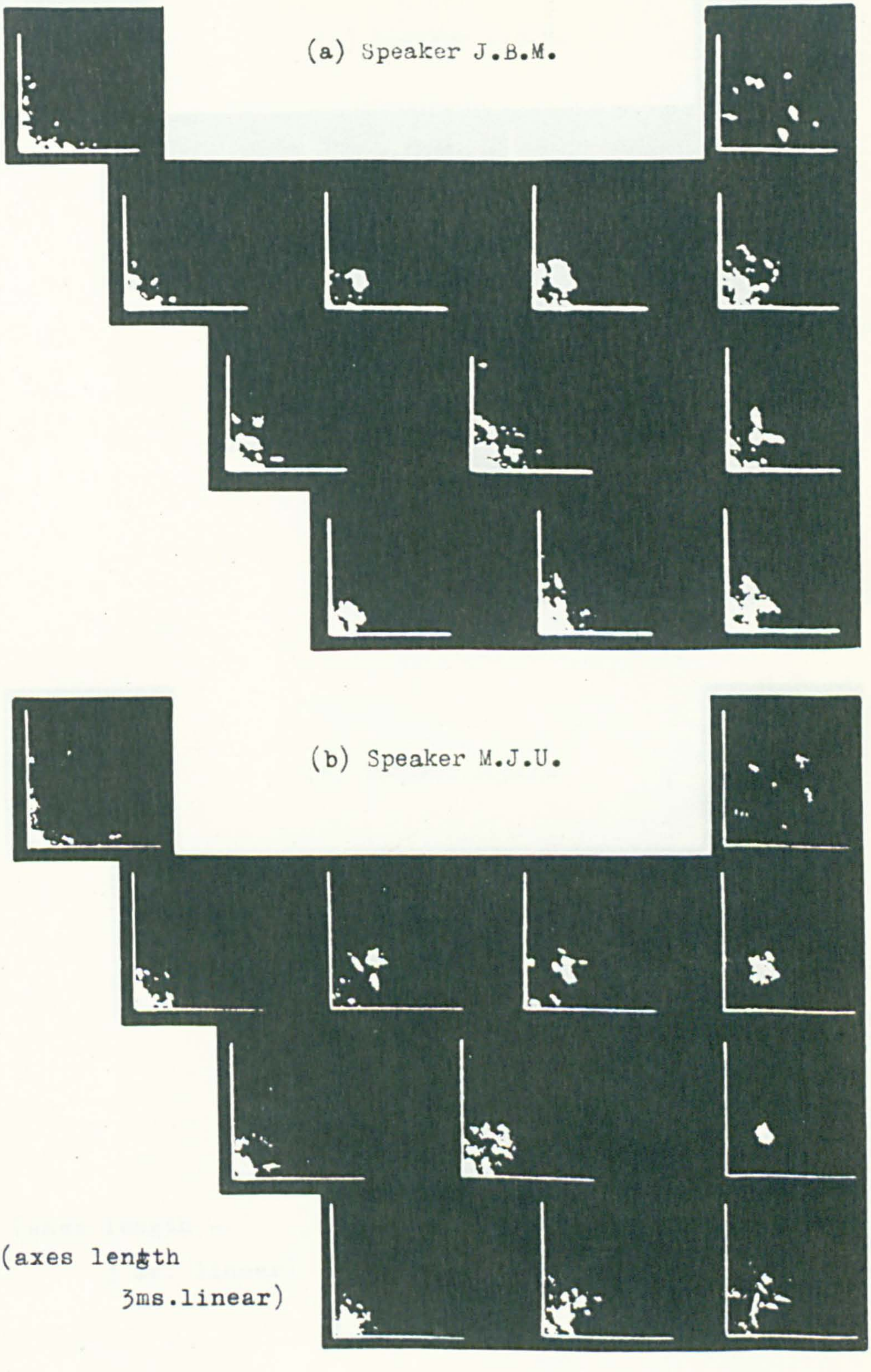
Two values of glottal period were chosen, 8.0 msecs. and 6.5 msecs. The former was a medium pitch for speaker M.J.U. and the latter was a medium pitch for speaker J.B.M. Both speakers experienced no strain in producing vowel sounds at the prescribed pitch levels.

The pitch of the utterances were controlled in the following way. A sine wave oscillator was set at either 125 c/s or 154 c/s to give a similar impression of pitch to a vowel sound whose glottal period was 8.0 msecs. or 6.5 msecs. respectively. The speaker could listen to the oscillator through headphones, prior to uttering the vowel sound, to aid his memory of the pitch required. The glottal periods of the resulting utterances were checked by viewing their waveforms on an oscilloscope.

Results.

The digrams of utterances produced by both speakers at a glottal period of 8.0 msecs. are shown in figure 2.10. Many of these digrams are far from being identical to the corresponding digram of the other speaker. The control of pitch has however resulted in some vowel digrams being more similar to those of the same vowel by another speaker. A clear example is that of /u/ which had previously shown great differences. It had been seen previously that the T.I. pattern of /u/ was very pitch dependent but what is shown here is that its dependence on the speaker is very slight, even at the level of the digram structure, if the pitch is controlled. In contrast to this case the vowel /ʊ/ has a strikingly different digram for the different speakers. The vowels /ə/, /ʌ/, /a/ and /ɔ/ also show differences. It is noticed in several of these cases that the differences are not only in the second order structure of the T.I.s but also in the first order distribution.

It is clear that the control of pitch can eliminate some of the differences in the first and second order T.I. distributions of utterances by different speakers. A greater understanding of the effect of pitch and the difference between speakers, as separate entities, was expected to be gained from an examination of the utterances of both these speakers at another controlled pitch. Accordingly, digrams of utterances having a glottal period of 6.5 msecs. are shown in figure 2.11. These results

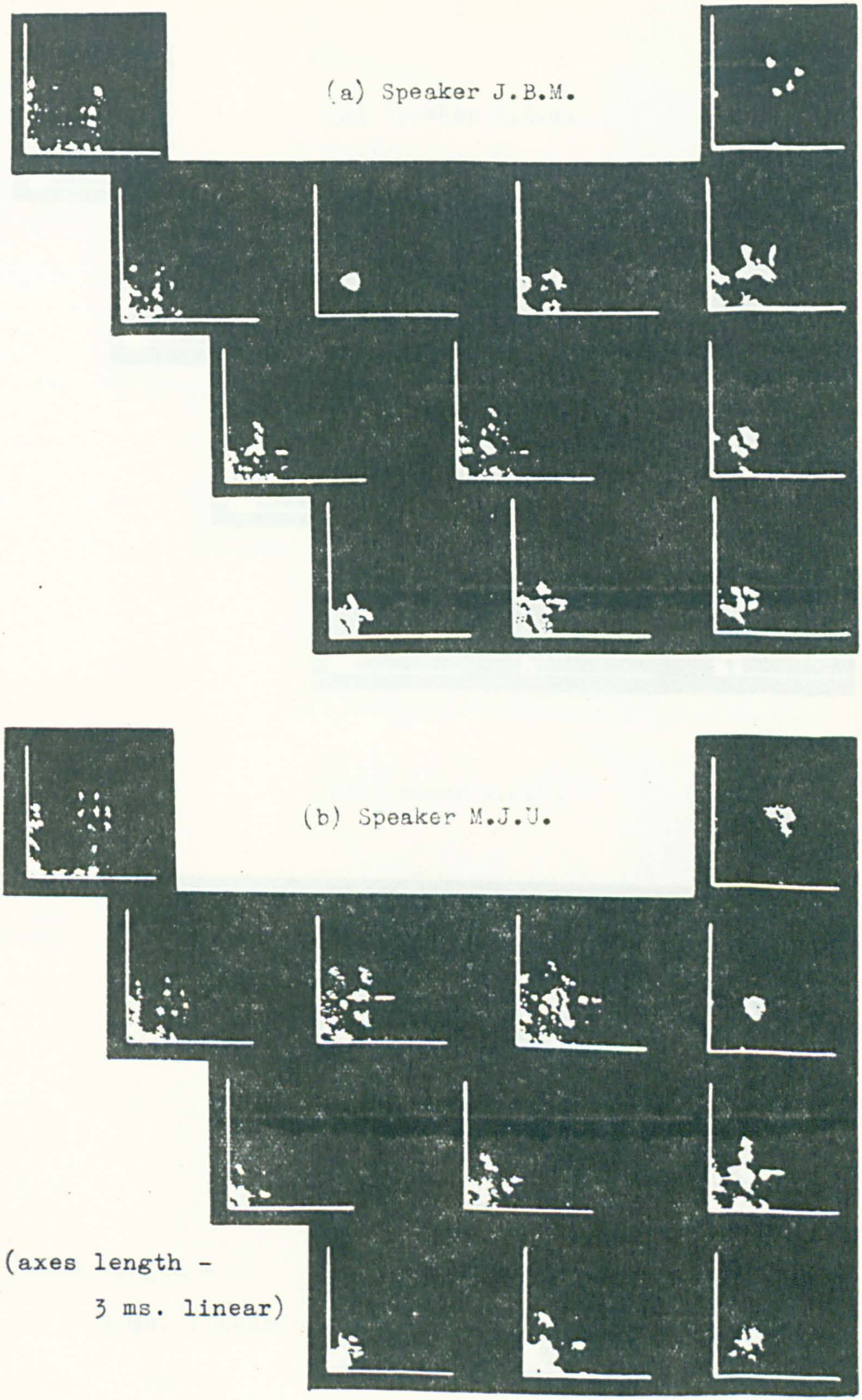


(a) Speaker J.B.M.

(b) Speaker M.J.U.

(axes length 3ms.linear)

Fig. 2.10 Digram displays of vowel waveforms with a glottal period of 8 msec.

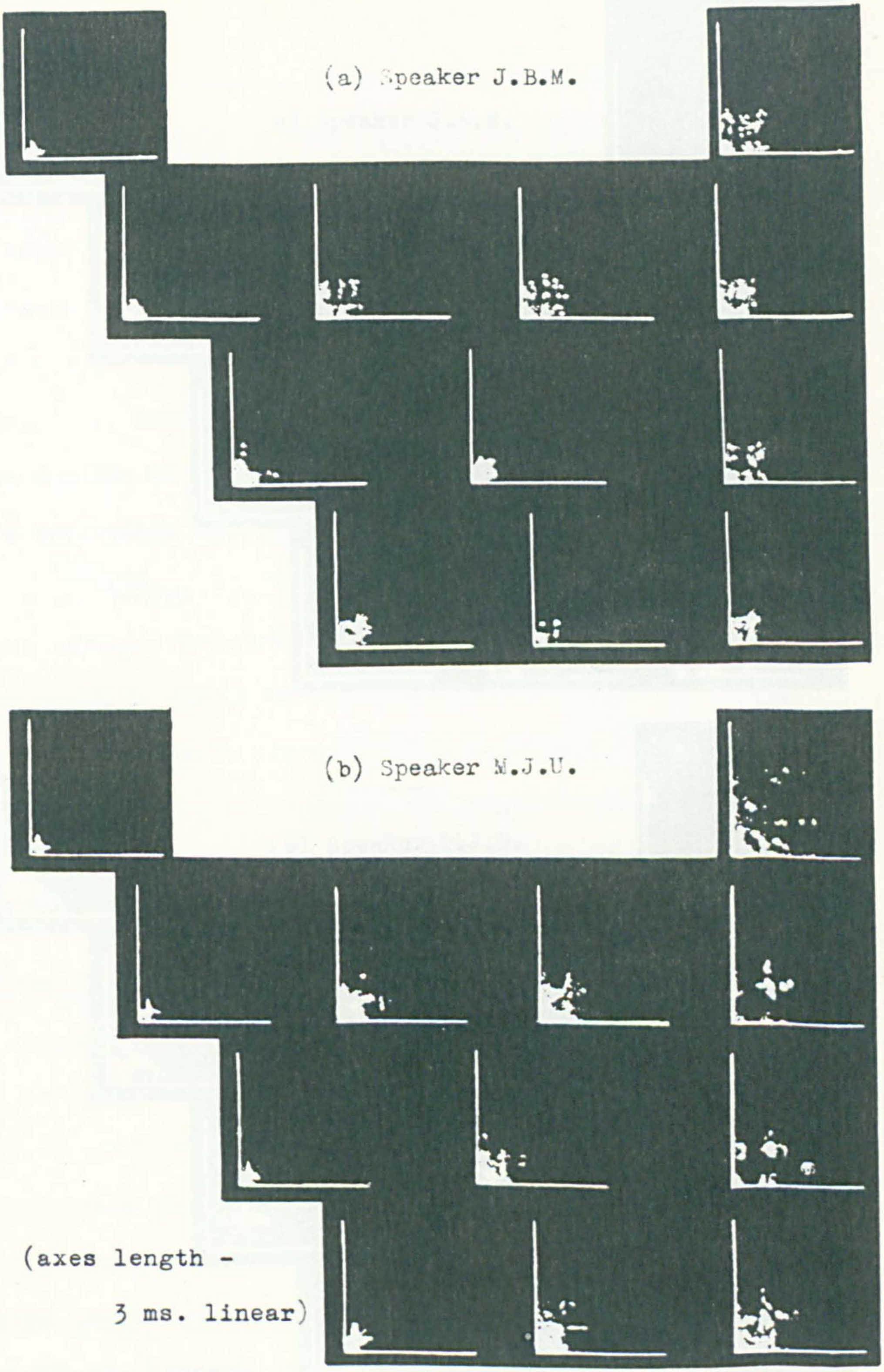


(a) Speaker J.B.M.

(b) Speaker M.J.U.

(axes length -
3 ms. linear)

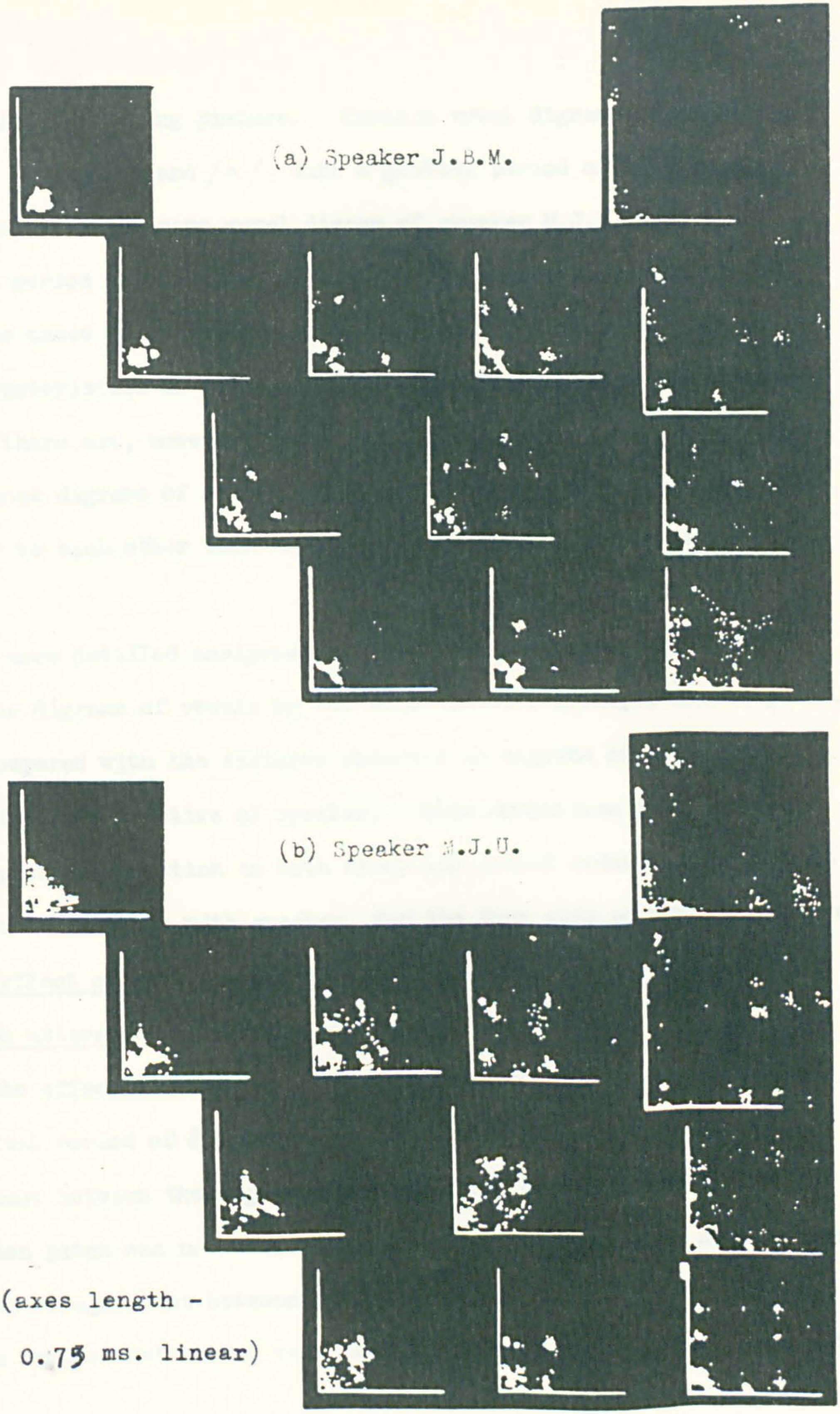
Fig. 2.11 Digram displays of vowel waveforms with a glottal period of 6.5 msec.



(axes length -

3 ms. linear)

Fig. 2.12 Digram displays of differentiated vowel waveforms with a glottal period of 8 msec.



(a) Speaker J.B.M.

(b) Speaker M.J.U.

(axes length - 0.75 ms. linear)

Fig. 2.13 Digram displays of differentiated vowel waveforms with a glottal period of 6.5 msec.

present a fairly confusing picture. Certain vowel digrams of speaker J.B.M., e.g. /i/, /I/ and /ʌ/, with a glottal period of 6.5 msecs., are more similar to the same vowel digram of speaker M.J.U. with the same glottal period than to that of J.B.M. with a glottal period of 8.0 msec. These cases point to the dominance of pitch rather than any other speaker characteristics in determining the T.I. structure of the vowel waveform. There are, however, those vowels of speaker J.B.M, e.g. /ə/ and /ɛ/, whose digrams of utterances with different glottal periods are more similar to each other than to those of M.J.U. with a similar glottal period.

A more detailed analysis was performed on the features that appear in the digrams of vowels by the same speaker, irrespective of pitch. They were compared with the features observed in digrams of vowels with the same pitch, irrespective of speaker. This comparison revealed that there is a greater variation in both first and second order T.I. statistics with pitch than there is with speaker, for the four sets of utterances used.

2.5.1 The effect of waveform differentiation on digrams of controlled pitch utterances.

The effect of waveform differentiation on digrams of utterances with a glottal period of 8.0 msecs. is shown in figure 2.12. There is more agreement between the digram patterns of the two speakers than was apparent when pitch was not controlled. It is interesting to note however that this extra agreement between speakers does not occur in the digrams of the same vowels that showed this tendency for utterances which were not differentiated. For example, the digrams of the differentiated version of /u/ are fairly different, whereas when normal speech was used, there was a large amount of agreement. Conversely there is a larger amount of agreement between the digrams of the speakers after differentiation in the case of their utterances of /ɜ/, /ʊ/ and /ɔ/.

The digrams of the differentiated version of the 6.5 msec. glottal period utterances of both speakers, illustrated in figure 2.13, are on a different T.I. scale. The patterns are magnified four times to reveal more of the structure of the T.I.s of the differentiated waveform. In certain isolated cases, for example /ɒ/ and /ə/, the fairly clear pattern for one speaker can be seen in the midst of the more complicated pattern for the other. In many vowels the first order T.I. statistics are dissimilar and the general impression is one of considerable divergence between speakers when these digrams are examined in detail.

It was found that no general elimination of the inter-speaker differences in the T.I. statistics was achieved by differentiation of the speech before clipping. It was not possible on the basis of these results to draw any more definite and positive conclusions on the effect of differentiation without describing each digram pattern and comparing them in detail. The introduction of quantitative measurements enabled more general conclusions to be drawn concerning the effect of differentiation on the perturbations of the T.I. statistics caused by different speakers and pitches. The use of quantitative measures will be described in chapters 3 and 4.

2.6 The removal of unwanted information from the digram display.

The aim of measuring T.I. statistics is to discover what relationship they have to the discriminability of speech sounds. It is known from work done on various models of speech analysis that only some of the information present in the speech wave is of importance to this discrimination. It has been pointed out (in the introduction) that the phenomenon of pitch is of no discriminatory value in the separation of voiced sounds on a phonemic level. It has been seen, however, in the preceding sections that this very factor of pitch has a large effect on the T.I. statistics. The perturbations that are caused are considered spurious when related to the overall aim of speech sound discrimination.

A further cause of spurious intervals is the effect of noise. A filter has been used to reject all noise which lies outside the frequency band of the signal of interest. No attempt has been made to reject noise within this band, which will be amplified with the speech signal and will dominate the T.I. pattern when the signal to noise ratio is low. This will occur according to the relationships between the relative amplitudes and the relative frequencies of the components of the signal and the noise, in a similar way to that discussed for the mixing of two formants (section 1.4.8).

As the study is concerned with isolated continuant sounds, the dominance of the noise during silent intervals was not a problem. In the early stages of the study, when the electronic circuits were designed, it was thought better to avoid the indescriminate T.I. distortion of differing magnitudes involved in any of the several ways of rejecting noise (see section 2.6.1). It was decided to provide a high signal to noise ratio thus keeping the electronics simple and the T.I. measurements accurate. The noise problem was left until a later stage when some experience had been gained in the accuracy of T.I. measurement required for discrimination of certain sounds, and in the sensitivity to noise of the statistics being measured.

Before describing the measures that have been taken to combat the perturbations of the T.I.s which constitute 'noise' in this analysis, the methods of noise rejection that have been used by others working on Z.C. measurements of speech will be briefly reviewed.

2.6.1 Noise rejection methods used by other workers.

The use of a trigger circuit with controlled positive and negative hysteresis.

The clipping transformation performed by this system is shown in figure 2.14 (a). A crossing of the actual zero level is not registered as a 'zero-crossing' unless it crosses by more than a defined amount '$h/2$'. Such a system effectively quietens the clipping amplifier when the input signal has a peak to peak value of less than h. This discrimination against amplitude applies to all inputs whether noise or speech, and also to all high frequency components of a complex wave, such as the speech wave, which have an amplitude of less than this level, even though the overall amplitude is greater. In addition it involves an error in the measurement of T.I.s under high amplitude lobes of the waveform, owing to the use of

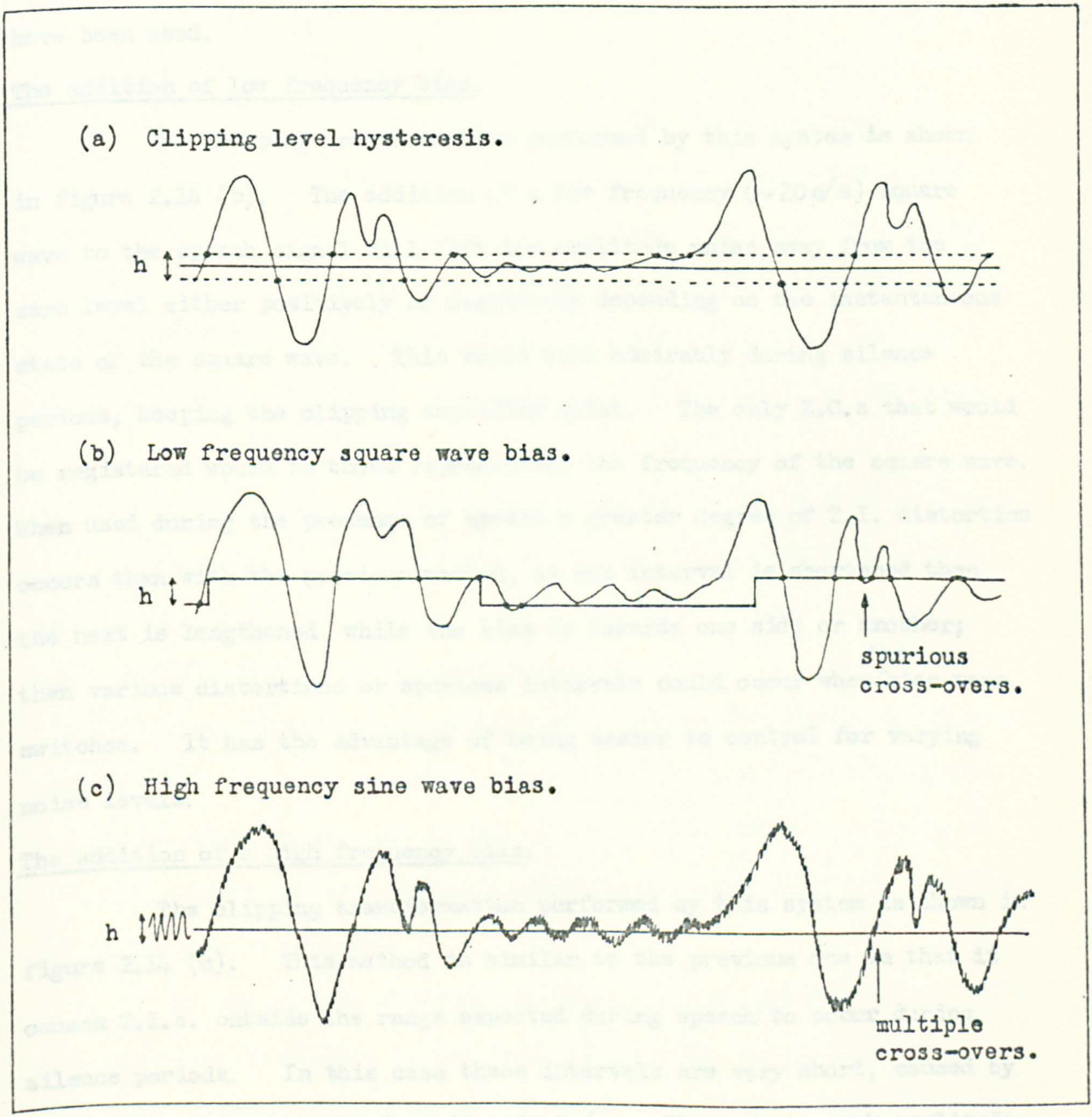

(a) Clipping level hysteresis.
h
(b) Low frequency square wave bias.
h
spurious cross-overs.
(c) High frequency sine wave bias.
h
multiple cross-overs.

Fig. 2.14 Illustration of noise rejection methods.

displaced clipping levels on the asymmetric waveform. This error is however the smallest in all the simple noise rejection techniques that have been used.

The addition of low frequency bias.

The clipping transformation performed by this system is shown in figure 2.14 (b). The addition of a low frequency (~20c/s) square wave to the speech signal will lift low amplitude noise away from the zero level either positively or negatively depending on the instantaneous state of the square wave. This would work admirably during silence periods, keeping the clipping amplifier quiet. The only Z.C.s that would be registered would be those representing the frequency of the square wave. When used during the presence of speech a greater degree of T.I. distortion occurs than with the previous method, as one interval is shortened then the next is lengthened while the bias is towards one side or another; then various distortions or spurious intervals could occur when/the bias wave switches. It has the advantage of being easier to control for varying noise levels.

The addition of a high frequency bias.

The clipping transformation performed by this system is shown in figure 2.14 (c). This method is similar to the previous one in that it causes T.I.s. outside the range expected during speech to occur during silence periods. In this case these intervals are very short, caused by the addition of a signal of ~ 20 - 50 Kc/s. The peak to peak amplitude of this signal is h, the level below which the speech signal is to be ignored. The interesting effect of the bias is that on the T.I. pattern of a speech signal with an amplitude greater than h. This system offers more scope in the way in which the spurious Z.C.s that are produced, may be processed. To ignore all the very short intervals invites more T.I. distortion than either previous method. To add these intervals into an adjacent interval,could be nearly equivalent to the first method if the phase of the clipped wave was synchronised with the original speech wave. This present method would in fact have the advantage of giving an indication of the presence of a low amplitude high frequency ripple component on a high amplitude low frequency waveform,which the hysteresis system would ignore.

If an on-line computing system was used with the high frequency bias system,a more accurate estimate of the true Z.C.s could be made by interpolation within the bursts of short intervals at each true Z.C.

2.6.2 Noise rejection in the present study.

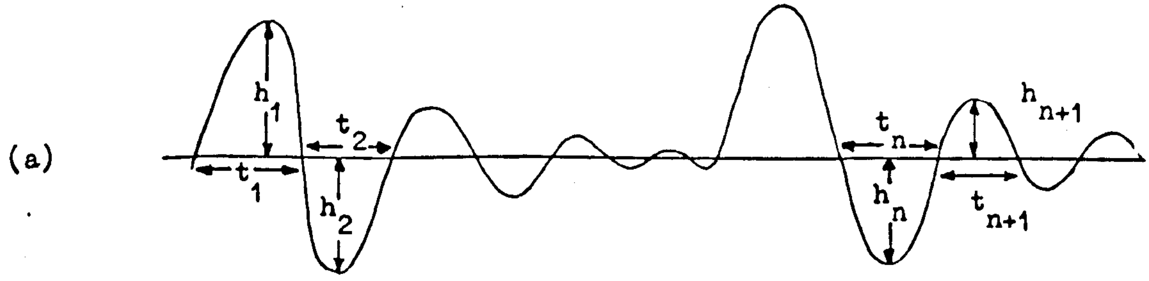
These three noise rejection methods have been reviewed, as noise rejection is considered important for a practical Z.C. measuring system working in undefined ambient noise, or for a system which is working on continuous speech where a constant signal to noise ratio is not possible due to the nature of speech.

In the present situation the need for such precautions is minimal although noise is present in the room, microphone, tape recorder and electronic circuits of preprocessing and clipping. The signal to noise ratio at the recording amplifier stage in the tape recorder was measured at +52 db. In such favourable circumstances it was decided to investigate forms of noise rejection specifically related to the speech wave; a system that will take account of the redundancy present in speech. Two forms of such a system were investigated.

2.6.2.1 The amplitude modulated digram display.

One problem was encountered with a branch study, on the digram display of continuous speech. It had become obvious in the early stages of experimentation using the digram display that it had some potential as a form of 'visible speech' (see appendix 8). Accordingly some work has been progressing in parallel with the present study on this aspect of display, used in a real-time dynamic mode. A suggestion for noise rejection, during silence or low amplitude portions of speech,was made by M.M.Taylor. An electronic system to implement the suggestion was designed and built in collaboration with M.J. Underwood. It was subsequently used by I.K.Taylor to evaluate the usefulness of the display as visible speech.

The basic idea was to display the digram on a C.R.T. as previously described, but to modulate the brightness of each spot by an amount proportional to the differential amplitude of the waveform lobes which defined the two T.I.s displayed. It was intended that in the real-time 'visible speech' display,this amplitude modulation of the brightness would separate syllables and darken the display during silent intervals. The system was evaluated separately from the 'visible speech' project as an example of a noise rejection scheme which gave a weighting to T.I.s. under various parts of the waveform depending on the instantaneous signal to



Waveform measurements involved in the Amplitude Modulated digram display.

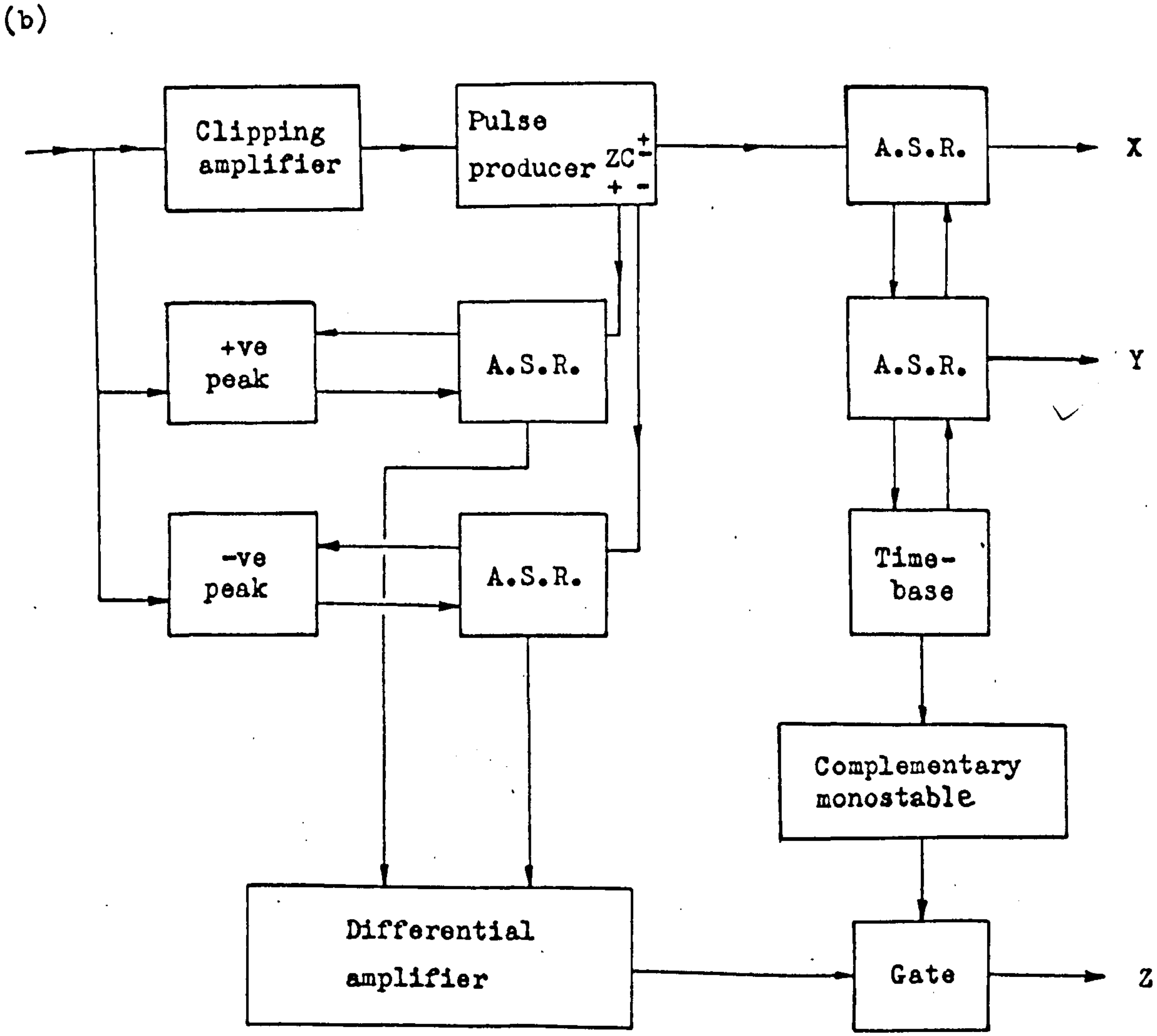


Fig. 2.15 A functional diagram of the Amplitude Modulated digram display.

noise ratio. Where the signal to noise ratio was high the intervals were measured as accurately as possible and given a high weighting, but where the ratio was low, the intervals, still measured without distortion due to the noise rejection, were given a low weighting.

Mode of operation.

The detailed aim of the system can be seen by reference to figure 2.15(a). When the interval pair, t_1, followed by t_2 are to be displayed, the brightness of the spot, with coordinates proportional to t_1 and t_2; will be proportional to $h_1 + h_2$. More generally, and illustrated here for the lower amplitude part of the waveform, the intervals t_n and t_{n+1} are displayed with weighting proportional to $h_n + h_{n+1}$. It was therefore necessary to store both h_n and h_{n+1} until t_{n+1} had elapsed.

Figure 2.15 (b) represents the function of the system designed to provide this weighting. Two analogue shift register (A.S.R.) stages were used to record the maximum positive and negative excursions, and hold voltages proportional to them until they were required. The T.I.s were measured in parallel with the amplitude measurement, and sequential pairs stored by a double stage A.S.R. as previously described. The positive and negative amplitude peaks were recorded by charging two capacitors through diodes of opposite polarity. These peak detectors were sampled alternately at every other Z.C.; the negative peak detector at the -Z.C., and the positive at the + Z.C.. Immediately after their voltages had been transferred on to an A.S.R. stage, they were reset and allowed to recharge according to the amplitude of the waveform at that time. The outputs of the two A.S.R. stages were fed to a differential amplifier. During the period of each time interval t_{n+2}, the output of this amplifier provided a voltage proportional to $h_n + h_{n+1}$. The delayed Z.C. pulse, which reset the time-base in the T.I. measurement system, controlled a gate on the output of the differential amplifier. The output of this gate provided the Z - modulation for the C.R.T.

Results.

The difference between the T.I. statistics produced using this amplitude weighting and those produced without amplitude weighting is illustrated in figure 2.16.* It can be seen that in most cases a much clearer vowel digram is achieved using amplitude weighting. Most of the noisy areas of the display have been removed indicating that they were due to low amplitude areas of the vowel waveform. Thus amplitude weighting of the T.I.s can be considered as an effective noise rejection technique. There are some disadvantages which are similar to those of the methods reviewed in section 2.6.1. The low amplitude high frequency ripple component of sounds such as /*i*/ and /I/ will be affected by this weighting. A small crossing of the axis will only be displayed with any appreciable weight if followed or preceded by a large amplitude lobe. It is seen in figure 2.16 that in the case of these two vowels, the short-short pairings have disappeared from the digram leaving only some short-long pairings in the case of /I/.

....

* These photographs were taken by M.J. Underwood.

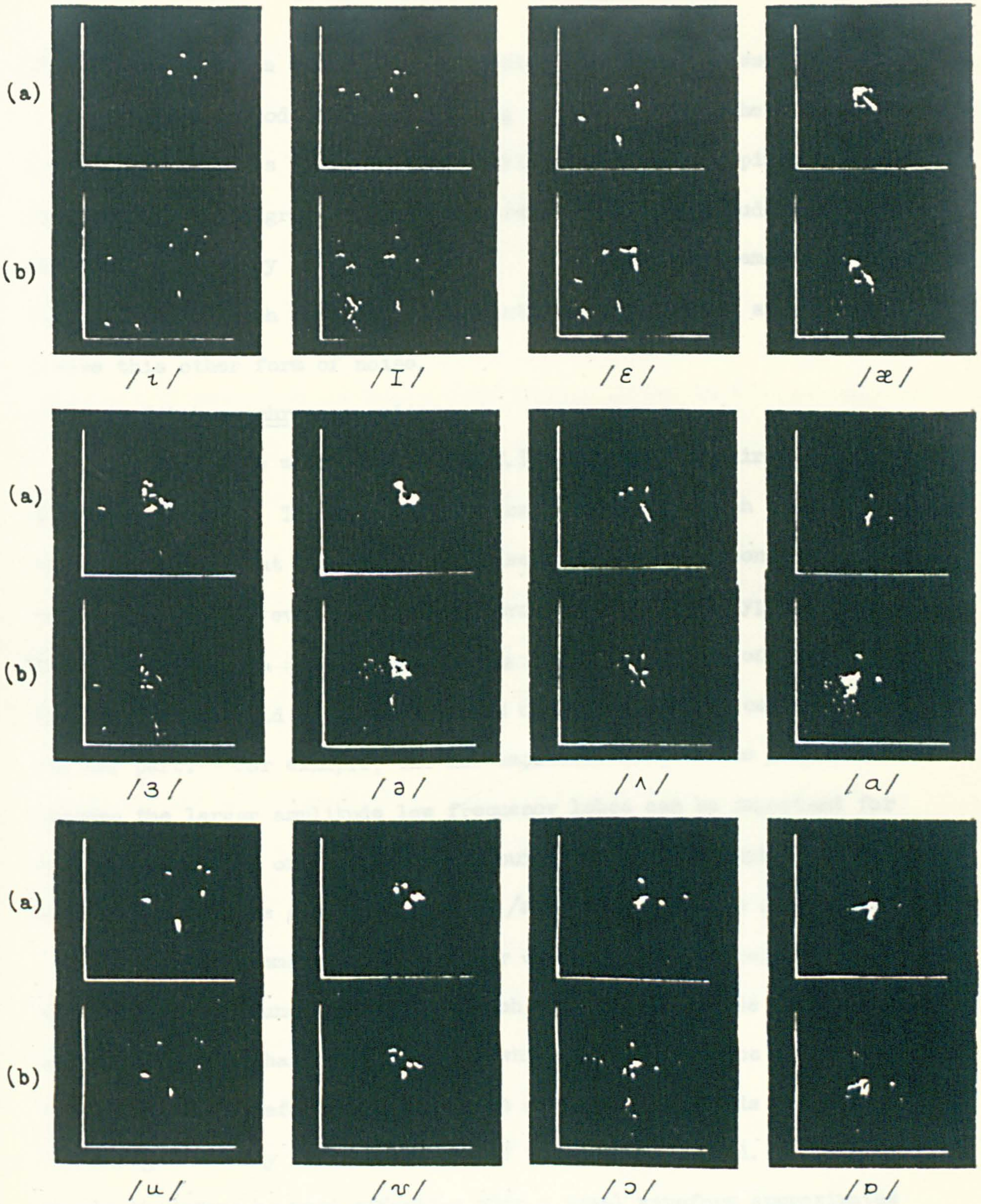


Fig. 2.16 **Vowel digrams obtained (a)with and (b)without Amplitude Modulation.**

Time interval scale 0 1 2 3 msec.

Thus stable features of the digram display have been removed.

This method also does nothing to remove that other form of noise which perturbs the T.I. distribution; that due to pitch. The six spots of the digram of /u/ which remain after amplitude weighting are known to be very pitch dependent. It therefore seemed reasonable to build more speech wave redundancy into the analysis in an attempt to remove this other form of noise.

2.6.2.2 Pitch synchronous gating.

Amplitude weighting of the T.I.s was still a fairly general purpose approach. The only information concerning speech that it used to advantage is that the signal to noise ratio is not a constant, and varies very widely even during vowel production. In a typical vowel waveform however, a low amplitude lobe [illegible] in one part of the glottal period could be important, but unimportant if it occurred in another part. For example, the low amplitude lobe of the /i/ waveform between the larger amplitude low frequency lobes can be important for the discrimination of /i/ and /u/, but a similar low amplitude lobe near the end of the glottal period of /æ/ would be quite unimportant.

There seems to be a case for devising a noise rejection technique which uses the redundancy of the speech wave caused by the constraint of pitch. A method has been developed which is based on the assumption that all vowel waveforms are a mixture of damped sinusoids whose amplitude falls significantly towards the end of the glottal period. This method should therefore be most effective when a vowel waveform approximates closely to this model.

This method can be considered as a simplification of the amplitude weighting system. The latter part of each glottal period is considered to be of low amplitude and is given zero weighting. The method is described as pitch synchronous gating (p.s.g.) as it is equivalent to gating the waveform such that only the early part of each glottal period is measured. When applied to a waveform similar to the model on which p.s.g. was based, it has a very similar effect to that of amplitude weighting, except that low amplitude crossings of the clipping level near the beginning of the glottal period are given high weighting. When applied to waveforms which do not decay so rapidly, it has the desired effect of removing some of the most pitch dependent intervals at the end of the glottal period, which constitute T.I. noise in a varying pitch signal. A similar form of analysis has been reported by David and Macdonald (12), Mathews et al. (46) and more recently by Stover (61). These workers have used pitch synchronous measurement in an attempt to reduce the bandwidth for speech transmission while retaining intelligibility, or to separate the characteristics of the glottal excitation from those of the vocal tract. The present use of the technique, for analysis solely in the time domain with the latter purpose in mind, has not been reported in the literature.

The proposal of this method posed two questions. "What is the most reliable and simple method to detect the beginning of each glottal period, and on what basis can a decision be made on the extent of the gating to be used in each period?".

2.6.2.3 Pitch detection.

The reliable detection of pitch has been the subject of many studies in recent years. Several methods have been used and they vary in the amount of information given about the excitation function. Inverse filtering of the speech spectrum (29, 49) gives the actual excitation function minus a few harmonics. The accuracy and great detail obtained using this method depends on the spectral structure being well defined. It therefore requires accurate formant tracking if it is used in continuous speech. Harmonic detection in the speech spectrum using 'cepstrum' analysis (50) can give a measure of the fundamental frequency even if no energy at this frequency exists. Two further methods operate in the time domain of the speech waveform. The length of the glottal period can be obtained from short term autocorrelation measurements on the speech waveform (27, 62.). If the additional information of the position of the start of the glottal period in time is required some form of peak detection must be applied to the waveform. (26, 54, 61).

It was clear that in order to control a gate to perform p.s.g., a pitch detection method which incorporated peak detection was necessary. A peak detector preceded by an automatic gain control was designed by M.J. Underwood. This circuit incorporated a dead-time which did not allow the circuit to record a peak within 3 msecs. of a previous one. This avoided the error of detecting two high amplitude lobes near the beginning of the period. This system was found to follow the pitch of most vowels although more difficulty was experienced in tracking the pitch of close vowels, which do not conform to the model of damped sinusoids so well as the open vowels. The performance of this circuit was improved for use with vowels of a steady pitch by introducing a variable dead-time which could be adjusted to be just less than the glottal period which was to be detected. The output of this circuit was a positive going edge when each peak was detected; this feature was termed a 'pitch marker'.

2.6.2.4 The extent of gating in each glottal period.

Having found a method to extract the start of the glottal period with sufficient reliability for the present purpose, the way in which the pitch markers should be used to control a gate on the T.I. measurement remained an open question. This question was discussed with M.J. Underwood in order that experiments on pitch synchronous synthesis of speech could be made parallel to the present study on pitch synchronous analysis of speech.

Evidence from perceptual experiments and analysis of articulatory and waveform behaviour contributed to the answer to this question.

Underwood (66) reports parallel work on the perception of pitch synchronously gated clipped speech. Slight improvement was found in the intelligibility of isolated vowels when only 6 msecs. of each 7 msec. glottal period were retained. A spectral analysis of the waveform after p.s.g., showed that formant peaks were apparent after 3 - 4 msecs. of each glottal period were included. When the whole period was included these peaks were sometimes not so well defined. Some similar perceptual experiments have recently been reported by Stover (61).

The acoustic waveform can be considered as a carrier of information concerning both the state of the articulators and the excitation function. There is nothing in the excitation function of voiced sounds which distinguishes one phoneme from another. Flanagan (20) has shown that changes in the vocal tract can have only a negligible effect on the excitation function. The state of the articulators are however most important in phoneme discrimination.

It was suggested by W. Lawrence (private communication) that the latter part of the glottal period may not be a very good reflection of the positions of the articulators for the following reason.. Studies in inverse filtering of voiced sounds to examine the glottal excitation function (49) have shown that the maximum amplitude peak of the waveform corresponds most closely in time to the rapid closure of the glottis. Holmes (29) has shown that the higher formants can be excited more than once in a glottal period; typically at the opening and closing of the glottis. The decaying resonances of each excitation due to the closure of the glottis are therefore perturbed towards the end of the glottal period by the effects of the reopening of the glottis. These effects will include a gradual build up of pressure and the re-excitation of the higher formants. It would therefore seem that there is a case, based on articulatory arguments, for looking only at the waveform which follows the maximum amplitude peak. This will be a more pure reflection of the resonances due to the positions of the articulators. Flanagan (21) states that in vowel production the glottis remains closed for between 40% and 70% of the glottal period. These figures suggest that it may be worth ignoring up to 60% of the glottal period; that is, the time when the glottis is open.

Further observations have been made of the stability of Z.C.s at various parts of the glottal period. In the majority of vowels some variation is seen in the intervals immediately preceding the maximum amplitude peak. When the stress or pitch of the utterance is changed it is most evident that such variation takes place (fig.1.5). Holmes (29) reported changes in the shape of the glottal waveform accompanying such subjective changes.

It would seem from this evidence that in some cases the removal of up to 60% of the glottal period might make the T.I. statistics less pitch dependent. The portion of the period to be gated out would be that terminating at the first Z.C. after the maximum amplitude peak. This position was chosen as it is clear from figure 1.5 that the interval under the maximum peak is very unstable in so far as its initial Z.C. is concerned. This choice differs from that of Mathews et al.(46) and Stover (61) who commenced their pitch synchronous analysis at the Z.C. before the principal peak.

Measurement of the gating period.

The control of the gating period according to the number of T.I.s which had elapsed was considered, in order to avoid the problem of fragmentary T.I.s. occurring to complete the 'on' period of the gate.

It was considered that such an approach would cause more severe p.s.g. on the vowels that were close to the assumed ideal of heavily damped sinusoids. Those which are not close to this ideal tend to have longer time intervals and thus more of the period would be included if a constant number of T.I.s per glottal period were measured. If the only aim was to remove the low amplitude portions of the waveform this approach would be helpful, but it would admit the end of the glottal period if very few Z.C.s occurred. This fact combined with the uncertainty in the number of Z.C.s per glottal period, especially in vowels such as /i/, caused this approach to be abandonned.

Other methods can be summarised as those which use a constant 'on' period, or a constant 'off' period, or a constant ratio between the two. To maintain a constant 'on' period assumes that as the glottal period increases the portion that reliably reflects the position of the articulators is constant, and the end of the period is increasingly filled with either low amplitude signal disturbed by noise or that perturbed by the glottal excitation. The perturbation of pitch, or pitch jitter, has been found by Lieberman to increase with increasing glottal period (38). Flanagan (22) has shown that as pitch and intensity increase, the width of 'glottal pulse' is often reduced, thus tending to maintain a fairly constant closed period or perhaps a fairly constant ratio of open and closed periods. The constant 'on' period method is the only one that can be implemented simply for real-time analysis; it would use a monostable controlled gate, and logic circuitry to avoid spurious fragmentary intervals being created.

The other two methods could easily be implemented when the glottal period was measured and stored. They were partly investigated later in this study when using the PDP-8 computer.

2.7 T.I. analysis after p.s.g. of vowel waveforms.

2.7.1 Experimental method.

The effect of p.s.g. on the T.I. statistics of vowels was investigated using the PDP-8 computer and three registers for the input and output of data. (see appendix 4). The T.I.s were stored in the PDP-8 store as described in appendix 4B. The program which controlled the analysis was based on one written by M.J. Underwood; modifications were added to control the duration of compilation, and later to retrieve T.I.s stored on digital magnetic tape.

The T.I. during which the start of a new glottal period was detected was denoted by adding to the value of the stored T.I. a large number, in the form of setting the most significant bit of the 12 bit computer word equal to 'one'. This was equivalent to an interval in excess of 24 msecs., therefore no confusion with long intervals was likely. In this way the T.I.s together with pitch markers were stored in the machine. A pitch synchronous analysis of the T.I.s was then carried out. The gating period was controlled in steps of 1 msec. by the switch register on the computer console. The T.I.s which were within this period were denoted by setting the most significant bit of the computer word containing their value. The state of this bit was used to control the

Z- modulation of a digram display. This display was produced by feeding the values of adjacent T.I.s into digital to analogue converters which in turn controlled the X and Y deflections of a C.R.T. This latter part of the program was a complete simulation of the analogue shift register described earlier (section 2.1.2).

The simulation of the A.S.R. and the process of p.s.g. within the PDP-8 had some great advantages. The earlier work on the analysis of digrams had involved three photographic exposures to obtain the digram pattern, plus two reference axes, on film. This had involved the use of a complex switching system. The A.S.R. simulator used a simple subroutine to write axes on the C.R.T. display. The output of the same intervals was cycled continuously enabling detailed observation of the digram patterns. The effects of p.s.g. could be seen by changing the severity of the gating on the same intervals in a non-destructive manner, thus enabling comparison and recomparison of various p.s.g. conditions. This use of the computer greatly reduced the number of experimental parameters that needed monitoring to ensure that accuracy in T.I. measurement and pitch detection were maintained during analysis of the digrams. Such monitoring was now required only while the T.I.s were being measured and stored in the computer. In all subsequent analysis the experimenter was able to devote all his attention to the effect that p.s.g. had on the digrams displayed.

The aims of this analysis.

The aim was to answer several questions that were raised by the influence of pitch on the T.I. statistics. Firstly, "What effect does varying the gating period have on the digram statistics, and does this effect vary from one vowel or class of vowel to another?". Secondly, "Does p.s.g. reduce the digram display of utterances of the same vowel at different pitches to the same pattern?". Thirdly, "Does it reveal that digrams of differently pitched utterances of two speakers can be reduced to the same pattern?".

2.7.2 The variation of the gating period.

The purpose of this experiment was to investigate the effect on the digram display of removing a successively greater proportion of the glottal period. It was expected that the removal of a small portion at the end of the period would tend to reduce the noisiness and pitch dependence of certain vowel digrams. As this portion was increased it was expected that some of the characteristic pattern of the digram would begin to disappear.

This was found largely to be true with some exceptions. Some of the digrams which followed the expected pattern are shown in figure 2.17. The vowel /I / spoken by M.J.U. with a glottal period of 6.5 msecs (which will be abbreviated to /I / - MJU - 6.5) is seen to become less noisy and finally to lose its characteristic short intervals as successively only 5 msec., 3 msec., and 1 msec. of waveform is retained in each glottal period. The p.s.g. versions of /ɒ / - JBM - 6.5 down to 2 msecs. are simply less noisy versions of the ungated form, but at 1 msec. the pattern begins to break up. The digrams of the other three utterances shown in figure 2.17 are seen to behave in a similar way.

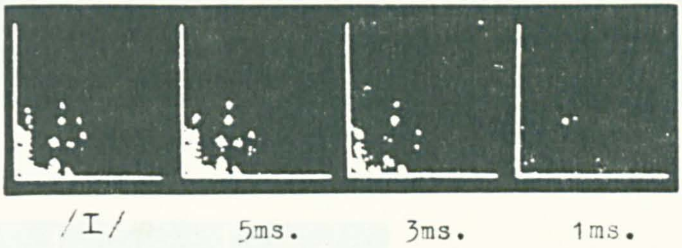
Certain vowel digrams are reduced to a single spot by severe p.s.g.. If the spot is not on the axis it must represent the first two intervals of the glottal period and illustrates the degree of stability of these intervals from period to period. An example of this is seen in figure 2.17 d. If the spot is on the axis it could represent a waveform which is dominated by a particular frequency, and the T.I. perturbations seen in the digram of the ungated form are due solely to end of period intervals. A very clear case of this has not been photographed but close approximations are seen in figure 2.18.

A further effect of reducing the gated portion of each period is for portions of the digram pattern, which seem to be equally strong in the ungated form, to disappear completely. Some examples of this are shown in figure 2.19. In these cases the latter part of the period seems to contribute a significant part of the display of a constant pitch utterance, but it may be very pitch dependent. Most of the digrams that exhibit this type of behaviour under p.s.g. are of vowels which do not conform to the model on which p.s.g. was postulated: they do not show significant damping within the glottal period.

Discussion of these results.

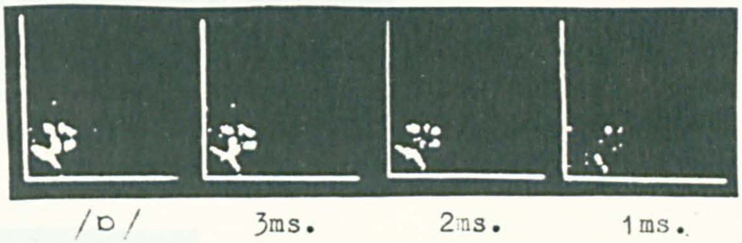
It is clear from the above results that p.s.g. does have different effects on different vowels. It is also clear that these different effects do follow certain classifications of the vowels and are not very different within these classifications.

The classes of sounds which in general are effected differently,



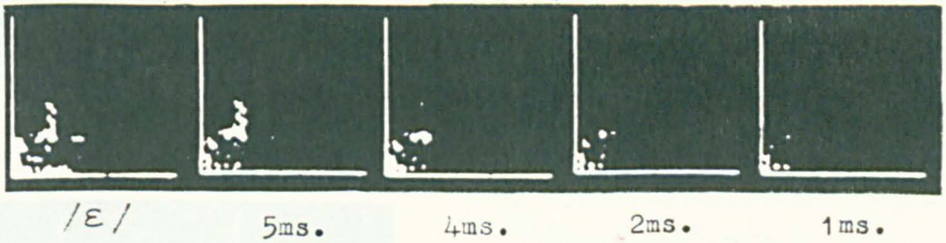
Speaker M.J.U.

6.5ms. glottal period.



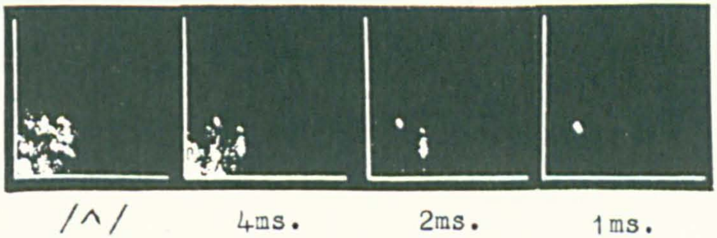
Speaker J.B.M.

6.5ms. glottal period.



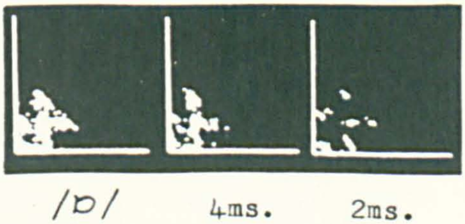
Speaker
J.B.M.

6.5ms.
glottal period.



Speaker M.J.U.

8ms. glottal period.

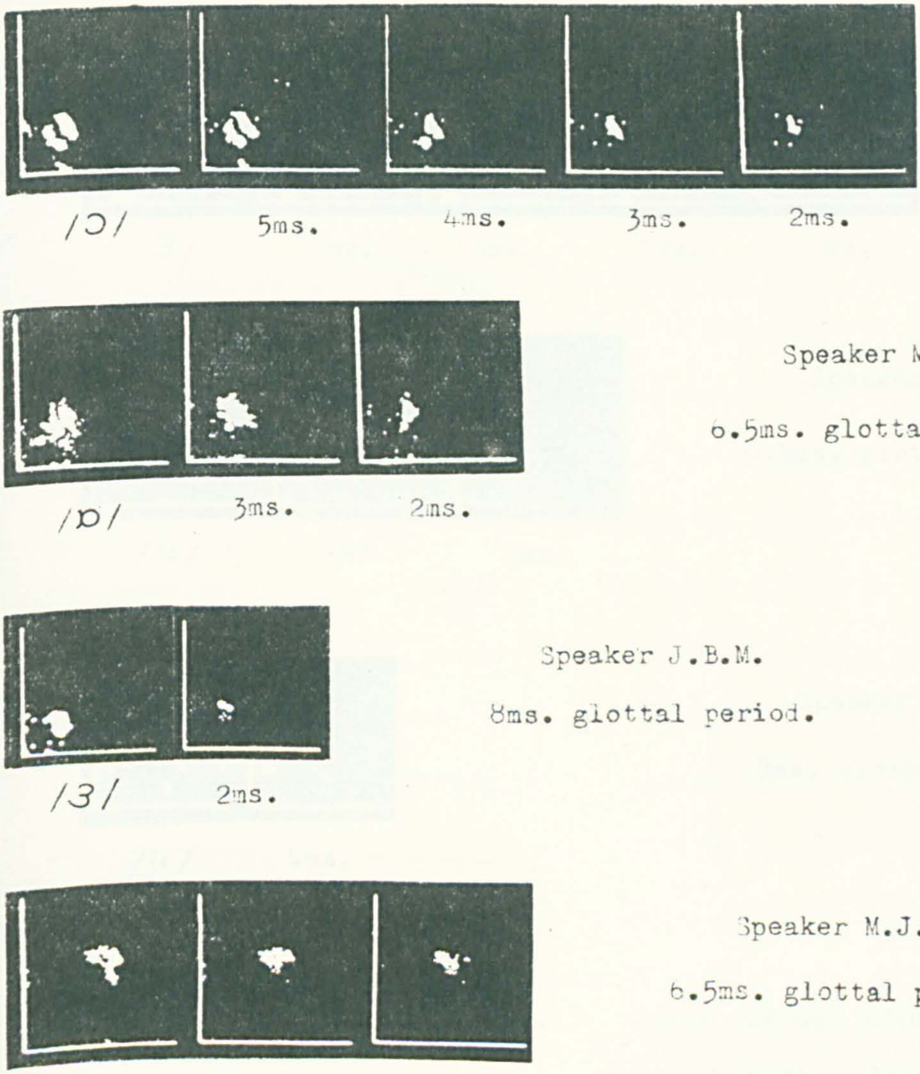


Speaker J.B.M.

8ms. glottal period.

Fig. 2.17 Vowel digrams after p.s.g..

(axes length - 3ms. linear)



Speaker J.B.M.

6.5ms. glottal period.

Speaker M.J.U.

6.5ms. glottal period.

Speaker J.B.M.

8ms. glottal period.

Speaker M.J.U.

6.5ms. glottal period.

Fig. 2.18 Vowel digrams after p.s.g..

(axes length - 3ms. linear)

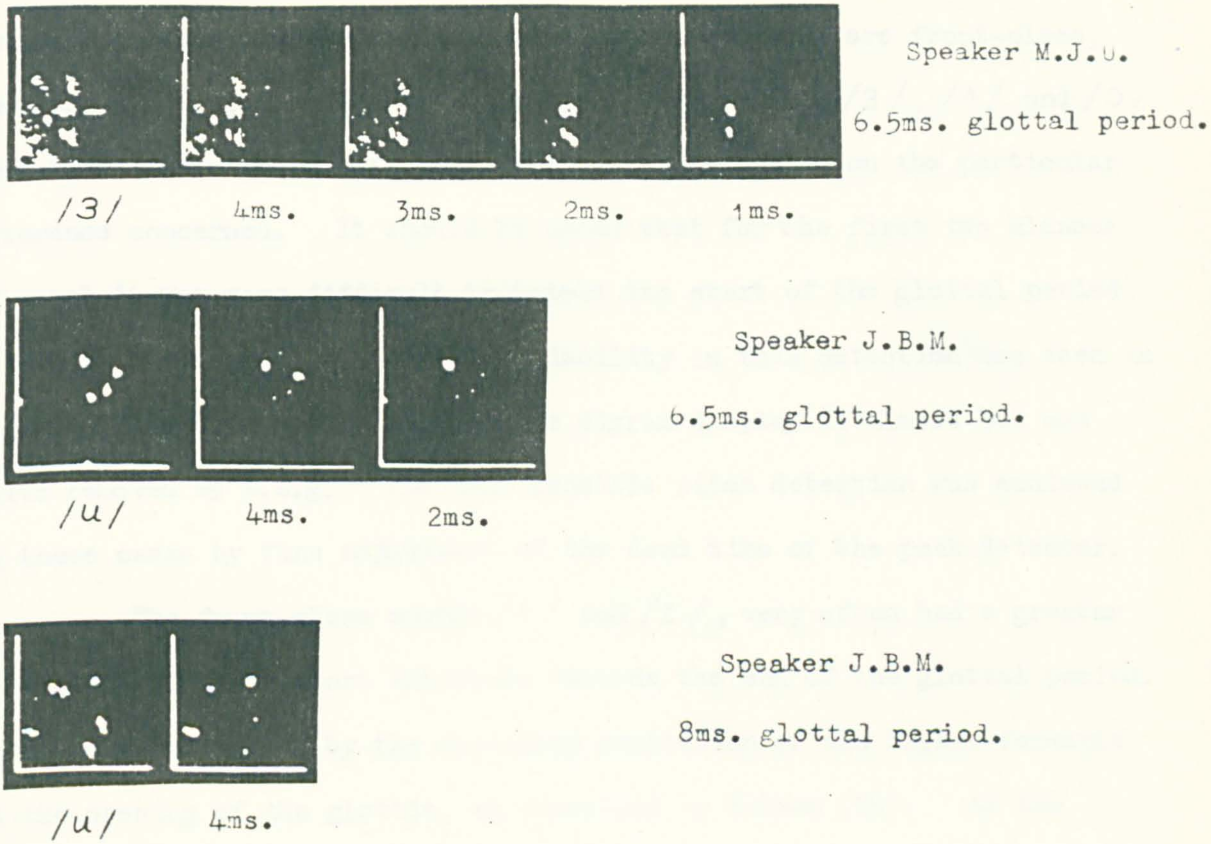


Fig. 2.19 Vowel digrams after p.s.g..

(axes length - 3ms. linear)

although certain overlapping characteristics are seen, are front-close, back-close,and open vowels. The central vowels /ə/, /ɜ/, /ʌ/ and /ɔ/ seem to be affected in less predictable ways depending on the particular utterance concerned. It should be noted that for the first two classes of vowel it was more difficult to detect the start of the glottal period reliably. Examples of slight unreliability in this detection are seen in figure 2.19 when a strong spot in the digram display is almost but not quite removed by p.s.g. The best possible pitch detection was achieved in these cases by fine adjustment of the dead time of the peak detector.

The front-close vowels,/i / and /I /, very often had a greater proportion of their short intervals towards the end of the glottal period. This can be explained by the secondary excitation of the higher formants at the opening of the glottis, as described by Holmes (29). As the relative amplitudes of first to second formant is smaller at the secondary excitation of F2,there is a greater probability of short intervals dominating the T.I. pattern at this point. The consequent effect of p.s.g. is to remove the short intervals before the longer ones. In such a case p.s.g. has little value. Pitch perturbations tend to be assimilated in the short intervals which individually contribute so little to the whole glottal period. Except in extremely long glottal periods, the amplitude remains high owing to the slow decay of the first formant resonance.

The back-close vowels, having very long T.I.s, were usually reduced by the removal of a single prominent spot or pair of spots from the display. These spots have been shown (section 2.3.5) to be very pitch dependent, therefore p.s.g. was exceedingly useful in these cases.

The open vowels showed a gradual reduction in the noisy areas of the display as the gating period was decreased. Such features as lines of dots parallel to an axis,revealing a steady change of one interval adjacent to a stable one,were removed with fairly slight p.s.g. It was for such vowels as these that p.s.g. had been suggested in place of the more elaborate amplitude weighting system. In most of these cases the use of p.s.g. was fully justified in producing simple patterns from the midst of noisy displays.

The extent to which p.s.g. can remove perturbations due to pitch, beyond those caused by noise in the case of open vowels and by long pitch dependent intervals in the case of back-close vowels, will be examined more closely in sections 2.7.4/5.

2.7.3 The effect of p.s.g. on differentiated vowel digrams.

The effect of p.s.g. on differentiated speech might be expected to be different to that on normal speech for several reasons.

The use of the differentiated waveform enables the measurement of reliable intervals earlier in the glottal period than in the normal case. This is apart from the reduction in average length of the intervals. The first interval to be measured is the one immediately after that in which the start of the period is detected. In the case of a normal waveform, measurement starts at the end of the interval under the maximum amplitude waveform lobe, whereas in the differentiated case measurement starts at the maximum of this lobe. Therefore the pitch synchronous measurement of the T.I.s of the differentiated waveform provides a slight advantage over that of the normal waveform, in that the T.I.s measured are under a part of the waveform which is deemed to be a more reliable reflection of the state of the articulators. The advantage is of course dependent on the detection of the start of the glottal period from the normal waveform, even when differentiation is included before clipping.

The T.I. statistics of the differentiated waveform might also be affected by p.s.g. differently to those of the normal waveform, as the portion of the glottal period which is gated out is that which will contain any secondary excitation of the higher formants. It could reasonably be assumed that the end of the period in the differentiated form could contain a large amount of information concerning the second formant.

Results.

In fact p.s.g. was found to have little general effect on the differentiated vowel digrams until the gating period was reduced to 2 - 3 msecs.. The use of this gating period was found to have a fairly similar effect on most of the vowel digrams (fig.2.20). The digrams of the ungated form are found in figure 2.13. In the case of /i /, /I /, /ɛ/, /ə /, /3 / and /ɔ / the digram is reduced to a much simpler pattern representing very nearly the T.I. characteristics of the second formant. An example of the usefulness of the digram over the histogram is seen in the digram of /ɔ /. A continuous first order spread of intervals between two values is seen to lie, in digram space, on a line of approximately X + Y = constant. This indicates the presence of some low

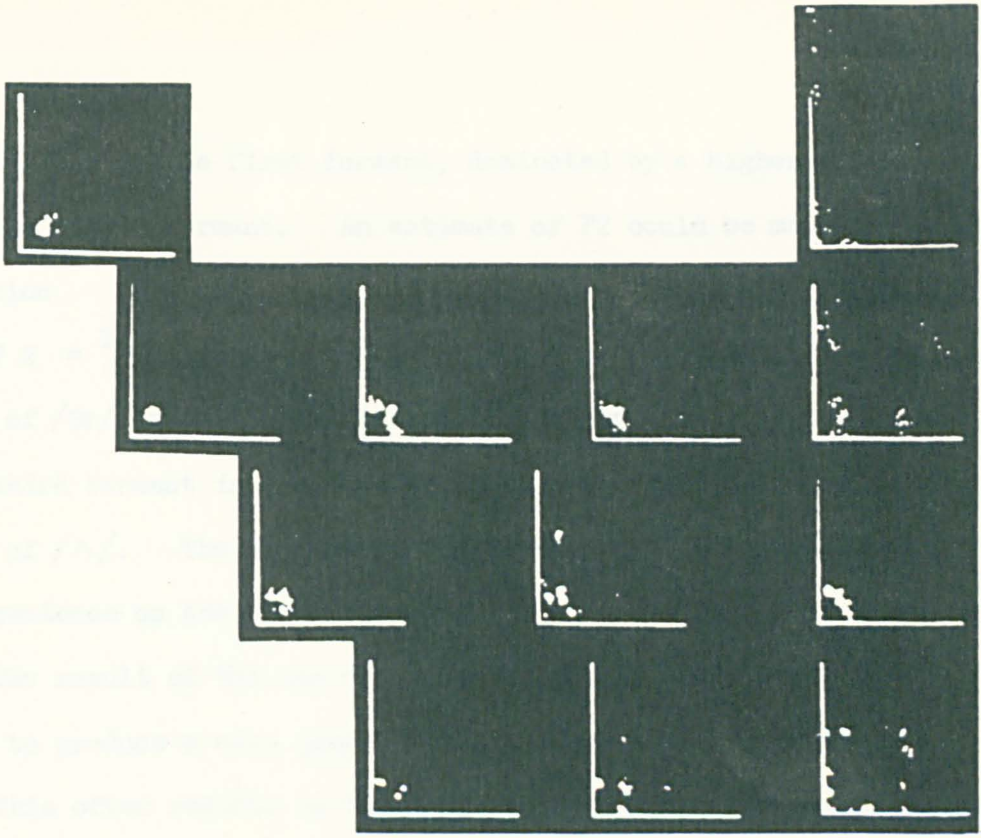


Fig. 2.20 The effect of p.s.g. on the digrams of differentiated vowel waveforms. (axes length - o.75 ms. linear)

frequency, presumably the first formant, dominated by a higher frequency, presumably the second formant. An estimate of F2 could be made by use of the equation

$$F\,2 = {}^{1}/_{C} \quad \text{where } C = X + Y. \quad *$$

The digrams of /æ/ and /ʌ/ cannot be described so simply owing to the effects of third formant in the case of /æ/, and first and third formants in the case of /ʌ/. The digrams of the vowels /ɒ/, /ʊ/ and /u/ show a larger dependence on the first formant.

The result of the use of p.s.g. on the differentiated vowel waveform is to produce a more general noise elimination in the digram display. This often results in T.I.s characteristic of the second formant being predominant. The secondary excitation of this formant, described by Holmes, seems to have little effect on the T.I. distribution.

* Strictly the value of C is given by the X and Y coordinates only at the ends of this line. That is when the low frequency component's waveform has zero slope. If the separation of the frequencies concerned is great, and the amplitude of the higher frequency is much larger than that of the lower one, the line X + Y = C is approximated by the display.

2.7.4 The stability of p.s. gated vowel digrams under pitch variation.

The effect that p.s.g. has in removing the effects of pitch on vowel digrams of two differently pitched utterances of the same speaker was investigated. Comparison was made between the digrams of utterances with glottal periods of 8.0 msec. and 6.5 msec. by the two speakers, J.B.M. and M.J.U.. Figure 2.21 illustrates the effect of p.s.g. on the digrams of some of these utterances. In all the cases illustrated, it is clear that the digrams of waveforms which are more severely gated are more similar to those of a similar utterance of different pitch, than are their respective ungated digrams. This was only found to be true in a minority of vowels. The digram pattern of utterances after gating often retained features which were particular to the utterance and perhaps to its pitch. It has been realised throughout that perturbations of the T.I. pattern due to pitch are not solely at the end of the glottal period. When a formant becomes a harmonic of the fundamental frequency, a change in the T.I. pattern can be caused in other parts of the glottal period (section 1.4.4). This perturbation is of course most pronounced in vowels where the T.I. pattern is nicely balanced between the influences of two formants. It is not clear how such a perturbation could be eliminated or accounted, for while working solely in the time domain.

2.7.5 The stability of p.s. gated vowel digrams under speaker variation.

In the light of the results of the previous section it seemed unlikely that pitch, as seen in its effect on the T.I.s of the latter part of the glottal period, could be the major cause of difference between the

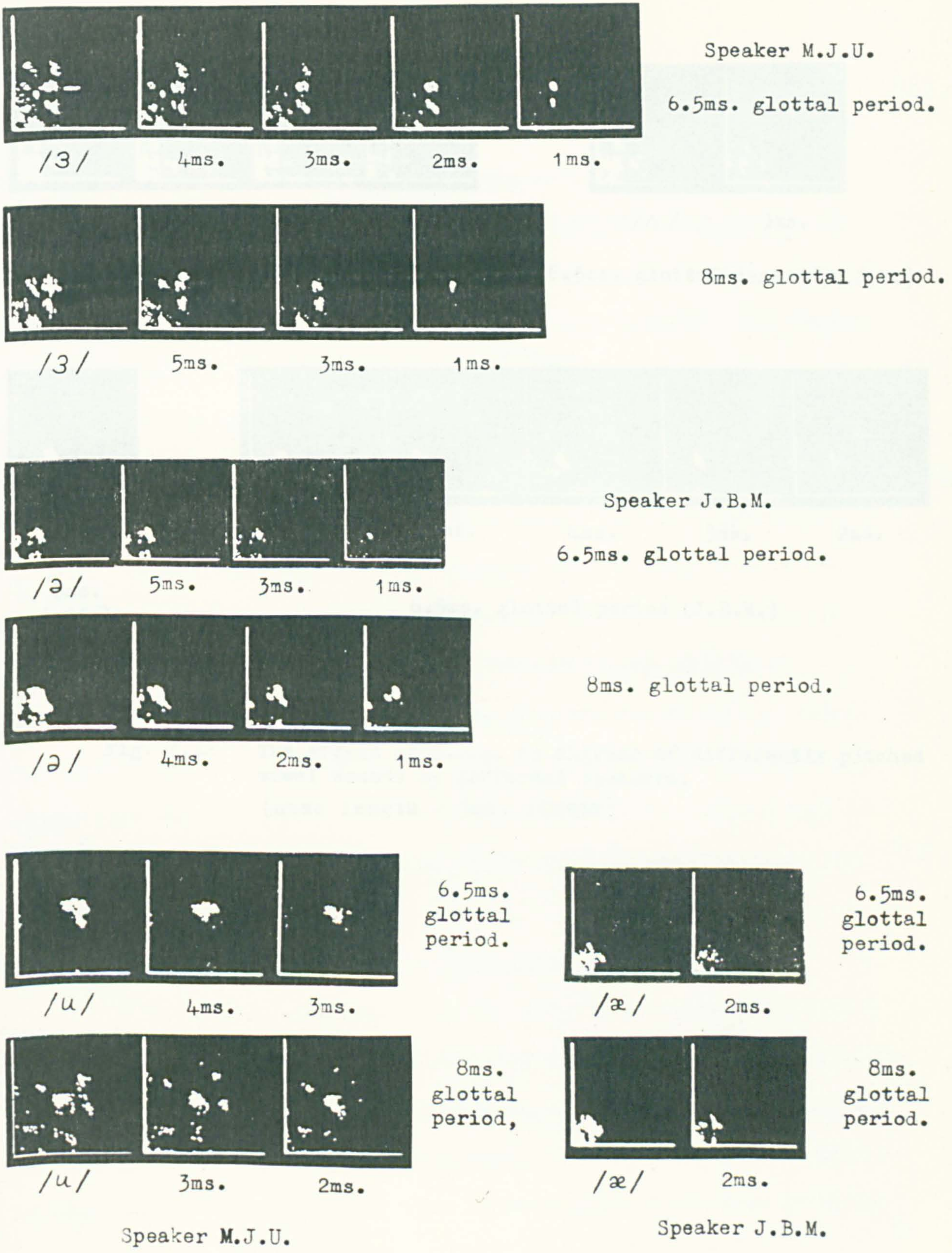


Fig. 2.21 The effect of p.s.g. on digrams of differently pitched vowel sounds by the same speaker.(axes length - 3ms. lin.)

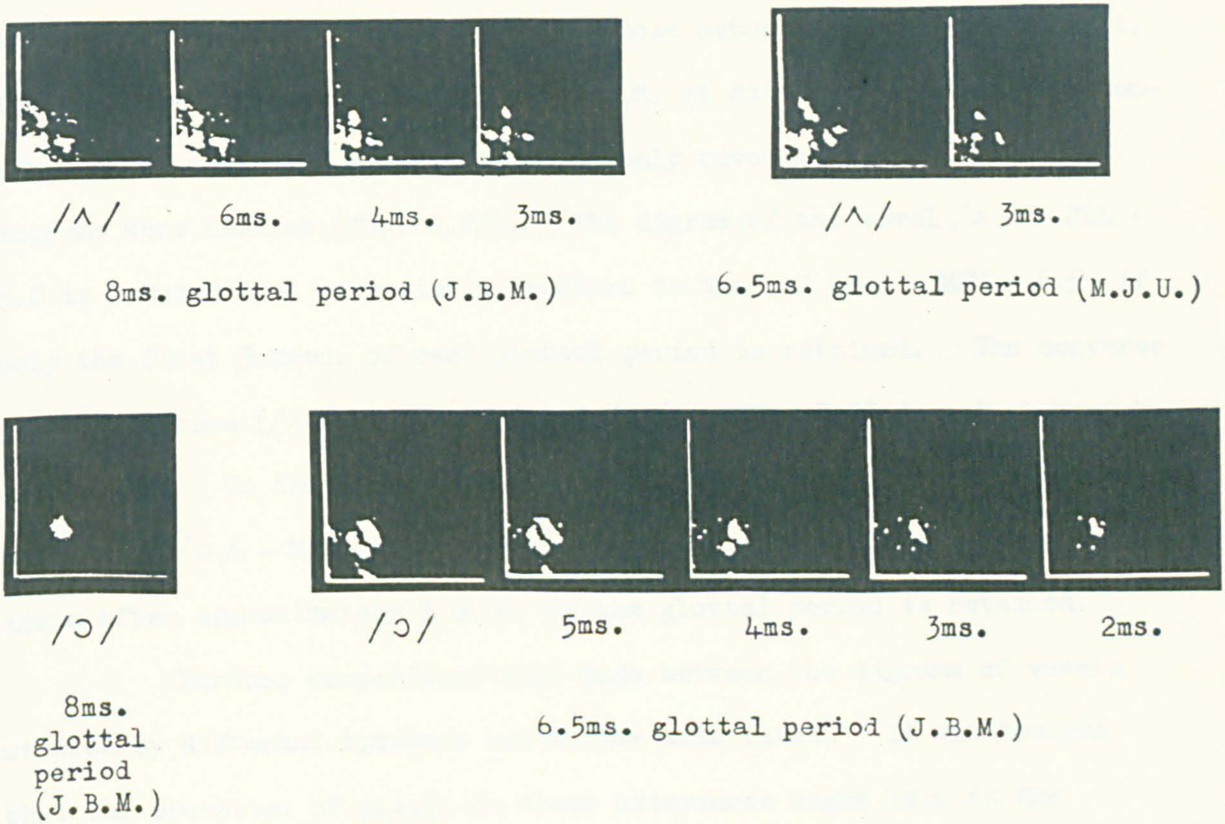


Fig. 2.22 The effect of p.s.g. on digrams of differently pitched vowel sounds by different speakers.
(axes length - 3ms. linear)

T.I. statistics of different speakers whose natural pitch was different. Indeed the utterances of M.J.U. and J.B.M. at different pitches, when compared after a certain amount of p.s.g., only revealed two vowels whose digrams were similar (fig. 2.22). The digram of the vowel /ʌ/ - JBM - 8.0 is reduced to a very similar pattern to that of /ʌ/ - MJU - 6.5 if only the first 3 msec. of each glottal period is retained. The converse pair of digrams (/ʌ/ - JBM - 6.5 and /ʌ/ - MJU - 8.0) do not show such similarity. In these converse sets of vowels however, the simple one spot pattern of /ɔ/ - MJU - 8.0 is found in the digram of /ɔ/ - JBM - 6.5, again after approximately 3 msec. of the glottal period is retained.

Further comparisons were made between the digrams of vowels uttered by different speakers but at the same pitch. It was thought that the operation of p.s.g. on these utterances might lead to the elimination of certain speaker dependent features due to the individual shape of his glottal excitation waveform. Evidence for any such elimination was very small and spread over many vowels. There were no vowel digrams that were photographed which revealed clear evidence for the above hypothesis.

2.8 Conclusions on the qualitative study of p.s.g.

When p.s.g. was applied to the open vowels which most nearly resemble the damped sinusoid model of vowel waveforms, from which the idea of p.s.g. evolved, dramatic rejection of noise on the digram display has been seen. The clear pattern which appeared has sometimes been similar to that which has emerged from noise in the digram of the same vowel but in a different utterance, of a different pitch, or from a different speaker. This is certainly not the case for all vowels. It seems that the intervals even at the start of the glottal period are somewhat pitch dependent, both in their first and second order statistics. This difference in the T.I.s at the start of the period could be explained in two ways. Firstly, there is the effect of changing relative amplitudes between formants, as the pitch changes cause alterations of the harmonic structure of the sound. Secondly, there is the effect of the a.c. coupling of the recording machine on differently pitched signals. The effective removal of one section of an asymmetric waveform will cause a slightly different a.c. zero level to be defined.

The extent to which similarity between digram patterns of similar sounds has been increased by p.s.g. of the waveform cannot be adequately measured using a qualitative photographic approach. The same problem arises as with the qualitative histogram analysis. Although there are a wider variety of features in the digram display, the human observer is not able to retain and compare full impressions of the patterns involved. In addition, the digram display uses two dimensions for the display of T.I. information,and the probability of occurrence information has to be recorded as brightness modulation. A great deal of this latter dimension is lost during the photographic process. Differences in the texture of the display, occasional drift of the d.c. brightness of the C.R.T.,and variations in the photographic process itself have contributed to this.

These problems of visual analysis led to the quantification of the method of analysis where all three dimensions were measured to any desired accuracy. The facility of an accurate memory of previous displays unperturbed by photographic limitations was a great advantage. The use of the PDP-8 computer in this analysis is described in the following chapters.

Chapter 3. The introduction of quantitative measurements.

Introduction.

The qualitative analysis of the histogram and digram statistics described in the previous chapters has indicated many of the variations that occur in the T.I. distributions of speech, even during isolated phonemes. It has also shown that a quantitative estimate of variation between statistics could be a useful aid to memory of the visual presentation, assuming that it offered a reliable measure of difference and similarity between these patterns.

This chapter is concerned with the quantification of both the statistics themselves and the relationships between them. The first problem was to decide in what units and on what scale the T.I.s should be measured, in order that the first and second order distributions would be quantised in some optimum way. The second,was to find a suitable measure of similarity or difference,which would make possible a flexible comparison of all the statistics which have been investigated.

The chapter also describes further investigations into these two forms of quantification which have the aim of making them more related to measurements of speech.

3.1 The reduction of scope to the study of vowel sounds.

On the basis of the photographic results illustrated in the previous chapter, it was clear that the digram statistics revealed more structuring of the T.I.s of vowel sounds than of the other classes of sounds examined. The nasals and voiced fricatives revealed structuring which was descriptive of their class,but within the classes the time intervals changed only in size with no obvious changes in structure. The digrams of unvoiced fricatives yielded little extra information to that obtained from the simple histogram.

It was considered that a quantitative measurement on the statistics of the twelve vowel sounds was within the storage and processing capabilities of the small digital computer available. This proved to be true,although the complexity of the analysis was restricted until magnetic tape storage became available (see appendix 4). The size of the machine was a problem because of the requirement that the statistics of all twelve vowels should be present in the machine for final analysis of the difference between them. The completion of this analysis in one program run greatly speeded up the analysis procedure, eliminated the high probability of human error in handling many paper tapes, and produced a complete matrix of the inter-relationships between the vowels.

3.2 The general form of quantification.

It has been mentioned in section 2.3 that there are two forms of quantification open to the experimenter in speech analysis; one, the subjective method involving the quantification of a subject's response to speech stimuli, the other, the objective method where specific physical measurements are made,and some form of 'artificial response' is simulated after computation has been done. The former method was employed, at the same time as this present study, by Underwood (66). The availability of on-line computing facilities being anticipated, experiments to measure the discriminability possible between vowels using the histogram and digram analyses were planned. Quantitative answers were sought to questions such as, "How does the discriminability of T.I. histograms based on a unidirectional Z.C. definition of the T.I.s compare with that possible when bidirectional Z.C.s are used?". "What effect does pre-clipping differentiation have on vowel discrimination using both histogram and digram statistics?". "Is digram analysis better for discrimination of vowels than histogram analysis?". "What effect does pitch synchronous gating have on the discrimination possible with these statistics?".

3.2.1 The scale of time interval quantisation.

Clues to the most appropriate quantising scale for the T.I.s were sought both from the earlier work of this present study, and reports of the work of others who have used the T.I. dimension in the analysis of speech.

The distributions, or narrowbin histograms, produced in the earlier part of this study were based on a linear T.I. scale. It was seen that a particular linear scale enabled a good distribution of the features of each class of sounds, but that different linear scales were required for the distribution of features of different classes of sounds.

Chang et al.(8) reported on several T.I. scales considered for use with the unqnantised 'intervalgram' display. They showed that a simple exponential curve, of time constant 0.3 ms., provided a good approximation to the subjective pitch scale of Stevens and Volkmann(60) and the articulation index scale of French and Steinberg (24). A similar exponential curve was used in the present study when the T.I.s of continuous speech were displayed in unquantised form.

Bezdel and Chandler (5) used a quantised scale affording equal accuracy of measurement for each bin, having found that,"a number of different channel distributions can produce equally successful scores".

Sakai and Inoue (55) used a quantised exponential scale on which they did not comment. They used this same scale for analysis of the original speech wave and that of the differentiated version.

It was decided to use a shallow exponential time scale in preliminary experiments on vowels. This decision was based on the facts that only vowels were to be analysed and that a fuller exponential curve would not be warranted unless high resolution of short intervals was required. The experience gained from the use of a linear scale during the experiments using the C.A.T., also suggested that a 'near linear' scale should be sufficient for the analysis of vowels.

3.2.2 The quantisation of the T.I. scale.

The linear scale that had been used in the C.A.T. was quantised in 20 μs. units. The distributions obtained had revealed that the average peak in the case of vowels had been 7 - 8 quanta wide. Some later ad hoc experiments were done using quanta of 150 μsec. and displaying the histogram stored in the core of the C.A.T. on an external oscilloscope, making possible much greater magnification of the horizontal dimension. In these experiments histogram peaks were found to be only one or two quanta wide.

The quantisation chosen, started with a bin of 120 µs. at the shortest time interval end, rising to 350 µs. at the longest time interval end. This scale of 16 bins had a range of 4.2 msecs. which was sufficient to cover all the T.I.s found in vowel sounds, whether defined by unidirectional or bidirectional Z.C.s.

3.2.3 Statistical Analysis.

The use of the term 'time interval statistics' has been rather loose thus far. In the case of the histogram analysis described in chapter 1, the bin widths used were much smaller than the significant variations in the time interval distribution, thus the histogram approximated towards this distribution. The digram analysis, described in chapter 2, used simply the digram distribution, i.e. the statistical averaging was over a period of time rather than over the time interval dimension. It was necessary to quantise the time interval dimension in order to make quantitative measurements on both first and second order distributions. From now onwards the term 'digram' will be used to denote a quantised digram distribution. The same quantisation scale and bin divisions were used for both histogram and digram statistics.

3.3 Measurements of separation or difference between statistics.

One of the major difficulties of the qualitative work was the lack of criteria on which to base the statement that one digram statistic was more similar to a second statistic than it was to a third.

The most natural form of simulation of the visual discrimination of the patterns would involve some form of feature extraction, measurement of the 'strength' of these features,and possibly a measurement of the difference between some of the features. However it seemed that a simpler approach was required to provide an initial quantitative evaluation of the discrimination possible using these T.I. statistics. Consequently two forms of difference or separation measures were investigated. They were (a) the correlation coefficient, which defines the extent and phase of any similarity between two patterns in space or time, and (b) the euclidean distance which defines a distance of separation in an N dimensional space, where N equals the number of measured parameters of the patterns. In the present case these could be the number of bins used in the statistic.

3.3.1 Correlation coefficient.

The correlation coefficient of two paterns x_i $(i = 1. \ldots N)$ and y_i $(i = 1. \ldots N)$ is given by the expression

$$C = \frac{\sum_{i}^{N} x_i y_i - N\bar{x}\bar{y}}{\sqrt{\left(\sum_{i}^{N} x_i^2 - N\bar{x}^2\right)\left(\sum_{i}^{N} y_i^2 - N\bar{y}^2\right)}}$$

where $\bar{x} = \sum_{i}^{N} x_i/N$ and $\bar{y} = \sum_{i}^{N} y_i/N$.

This measure gives the value +1 if the patterns being compared are identical, -1 if they are completely 'out of phase' with one another and 0 if their relationship is completely random.

A preliminary trial was carried out by applying this measure to T.I. histograms of vowel sounds. Histograms were compiled for eight sequential segments of each vowel utterance. The values of correlation coefficient obtained for comparisons between histograms from the same vowel utterance ranged over + 0.7 to + 0.99. Those for comparisons between histograms of different vowels ranged over - 0.2 to + 0.92. This method clearly had some potentialities as a discriminating measure. A basic disadvantage soon became apparent however. If extra bins were added to the statistics in order to accommodate further variations of the T.I. distributions, the correlation coefficient was not comparable with that computed using a smaller number of bins. This became very apparent when empty bins were added in anticipation of their use by extra long T.I.s when unidirectional Z.C. intervals were to be measured. The extra zeros in the catalogue of bin contents constituted additional similarity to another histogram with the same unused bins. This inflexibility of the measure made it impossible to compare the differences between histograms with different numbers of bins. The measurement that was required was one that would ignore the empty bins until they were used, thus rendering statistics with different numbers of bins more directly comparable.

3.3.2 Euclidean distance.

The euclidean distance measure was chosen as it overcame the problem found using the correlation coefficient. In this measure the contents of each of the N bins of the statistic are normalised then considered as excursions along each of N orthogonal dimensions of an

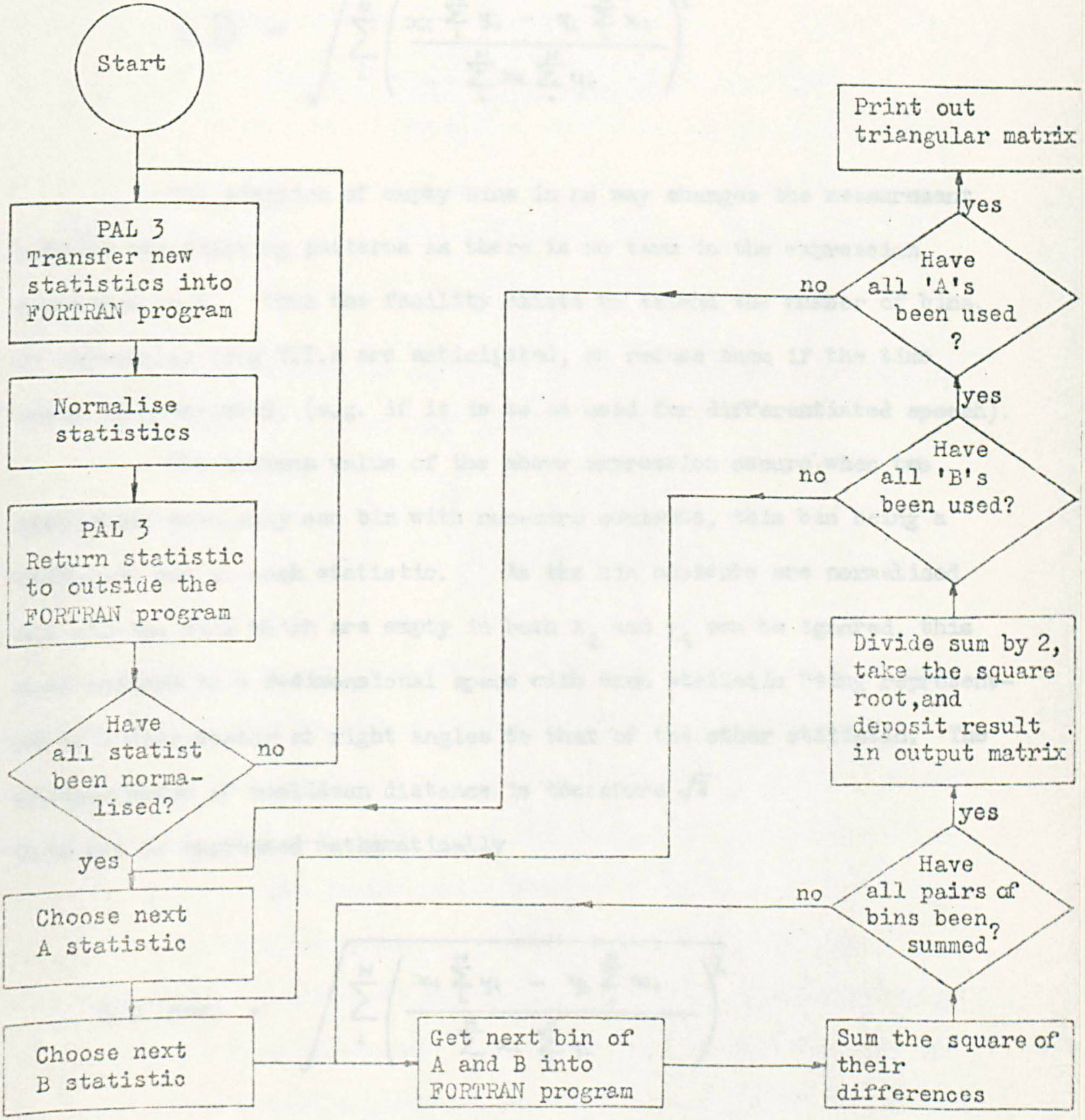

Start
PAL 3
Transfer new statistics into FORTRAN program
Normalise statistics
PAL 3
Return statistic to outside the FORTRAN program
Have all statist been normalised?
no
yes
Choose next A statistic
Choose next B statistic
Get next bin of A and B into FORTRAN program
Sum the square of their differences
Have all pairs of bins been summed?
no
yes
Divide sum by 2, take the square root, and deposit result in output matrix
Have all 'B's been used?
no
yes
Have all 'A's been used ?
no
yes
Print out triangular matrix

Figure 3.1 Flow diagram of Euclidean distance measurement program.

N-dimensional space. It is defined by the expression

$$\mathrm{E.D.} = \sqrt{\sum_i^N \left(\frac{x_i \sum_i^N y_i - y_i \sum_i^N x_i}{\sum_i^N x_i \sum_i^N y_i} \right)^2}$$

The addition of empty bins in no way changes the measurement between two existing patterns as there is no term in the expression dependent on N. Thus the facility exists to extend the number of bins, if especially long T.I.s are anticipated, or reduce them if the time scale is contracted, (e.g. if it is to be used for differentiated speech).

The maximum value of the above expression occurs when two statistics have only one bin with non-zero contents, this bin being a different one in each statistic. As the bin contents are normalised and all the bins which are empty in both x_i and y_i can be ignored, this case reduces to a 2-dimensional space with each statistic being represented by a unit vector at right angles to that of the other statistic. The maximum value of euclidean distance is therefore $\sqrt{2}$.

This can be expressed mathematically

$$\mathrm{E.D.\ max.} = \sqrt{\sum_i^N \left(\frac{x_i \sum_i^N y_i - y_i \sum_i^N x_i}{\sum_i^N x_i \sum_i^N y_i} \right)^2}$$

where $\sum_i^N x_i = x_k$ and $\sum_i^N y_i = y_j$, $k \neq j$.

This is based on the mathematical inequality $a^2 + b^2 \leq (a + b)^2$ as E.D. max. is expressed as a sum of squares; i.e. the contents of each statistic are in one bin rather than in several.

$$\therefore \text{E.D. max} = \sqrt{\frac{(x_j y_j - y_j x_k)^2 + (x_k y_j - y_k x_k)^2}{(x_k y_j)^2}}$$

$$\text{E.D. max} = \sqrt{2} \qquad \text{as} \qquad x_j = y_k = 0$$

Thus a measure of euclidean distance which has a maximum of + 1 and a minimum of 0 is defined by the expression

$$\text{E.D.} = \sqrt{\sum_i^N \frac{1}{2}\left(\frac{x_i \sum_i^N y_i - y_i \sum_i^N x_i}{\sum_i^N x_i \sum_i^N y_i}\right)^2}$$

Bezdel and Chandler (5) used both correlation coefficients and a form of euclidean distance in discrimination tests on the T.I. histograms of five vowels. After exhaustive tests they found that the euclidean distance measurements gave the better discrimination. This may well have been due to the problem mentioned above: the positive effect of bins with zero contents. They also proposed the use of a 'weighted euclidean distance' thereby giving greater importance to some of the bins of their histograms. In the present work the storage of weighting factors posed a problem until additional computer storage was obtained. The importance of some form of weighting will be discussed in section 3.6.4, and discrimination measures using weighted euclidean distance will be described in the following chapter.

3.3.3 The measurement of euclidean distance in the computer.

The computing facilities available at this time were those of the phase I installation of the PDP-8 system described in appendix 4. Programs were written to compile the statistics of the T.I. distributions within the computer as described in appendix 4 B. The measurement of euclidean distance was then achieved by a program of a hybrid type, using both the machine language, PAL 3, and the higher level, Fortran language. The flow diagram of this program is shown in figure 3.1. The output was a triangular matrix of the euclidean distances between all the vowel T.I. statistics in the machine. The maximum number of statistics that could be analysed by this program was twelve.

When the statistics of sequential segments of an utterance were analysed, only the values on the hypotenuse of the triangle were used.

3.4 Experiments using Euclidean distance separation measurements on the T.I. statistics of vowel sounds.

3.4.1 The effect of 'duration of compilation' on Euclidean distance.

This experiment was done partly as a control and partly as an extension and a quantification of observations made during the qualitative studies.

As a control it was viewed as a useful way in which to become familiar with the variation of euclidean distance (E.D.)/while varying a parameter whose effect on the T.I. statistics was fairly easily understood. This understanding was based on the experience of similar qualitative work and an elementary knowledge of statistical sampling theory.

The E.D. that was measured in the present experiment was that between digrams of sequential segments of the utterance. The duration of these segments was varied and two sets of results were measured. Firstly, as illustrated in figure 3.2 (top), the overall variation of E.D. with duration of compilation was measured. Secondly, as illustrated in figure 3.2 (bottom), the variation of E.D. for individual vowels was measured.

Figure 3.2 illustrates that the effects of a short duration of compilation,and therefore of a lack of statistical stability, are seen when this duration is reduced between 50 - 20 msecs. Below 20 msecs. the effect is most marked, as the length of a single glottal period is approached.

When the resultsof this experiment are broken down into the E.D. separation for each vowel taken separately, the majority of vowels, independently of their characteristic T.I. distribution, conform to the overall trends of figure 3.2. The obvious divergences from this, seen in the 30 msec. segments of /ʊ/ and /ɔ/, have significantly higher standard errors (~30%) than the average (~10 -15 %).

The rank ordering of the vowels at the edges of figure 3.2 is intended purely as an aid to distinguishing one vowel plot from another. No significance is placed on the order in which they occur for different durations of compilation. The differences between the level of the plots

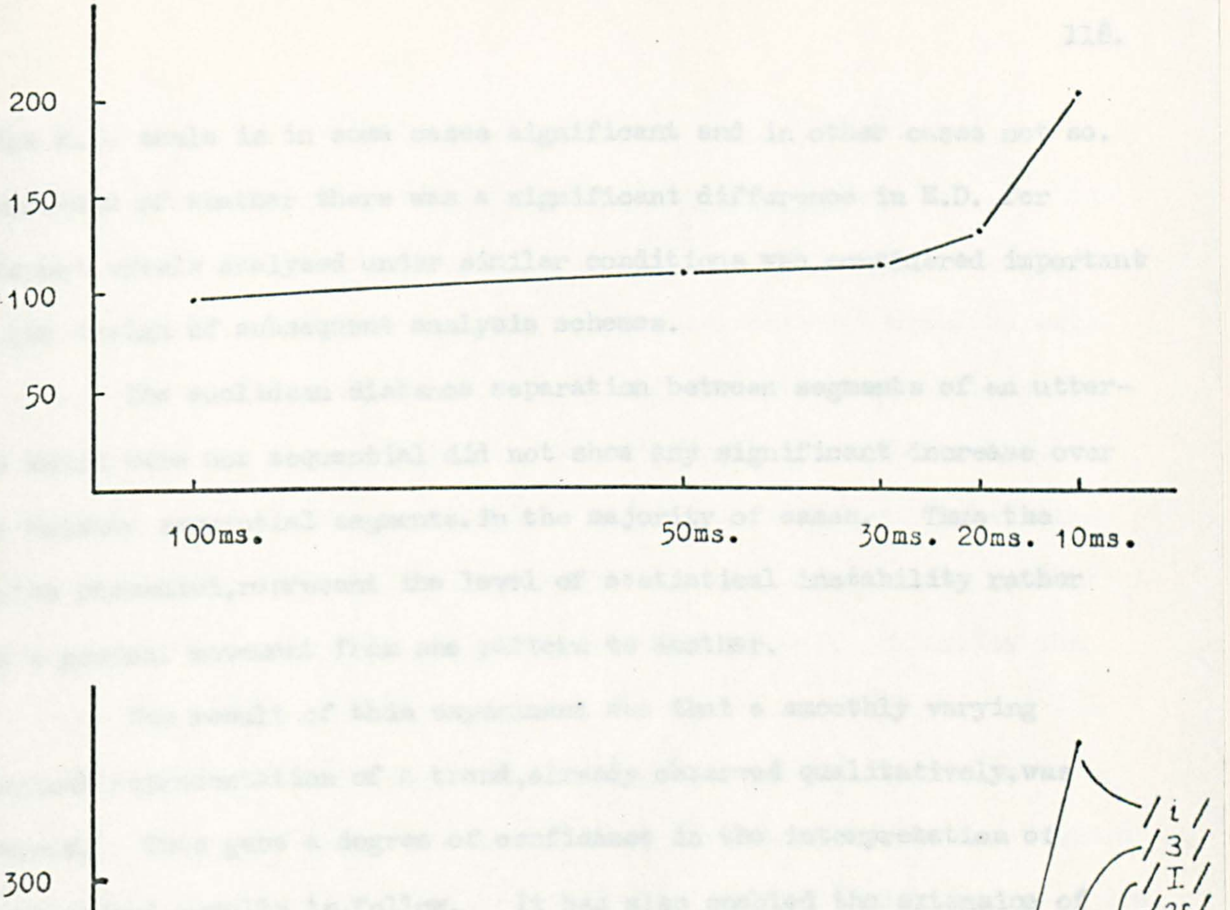


Figure 3.2 Euclidean distance between digrams of sequential segments of the same utterance (ordinate) against the duration of compilation (abscissa). Top - average for all vowels, botton - values for individual vowels.

on the E.D. scale is in some cases significant and in other cases not so. A knowledge of whether there was a significant difference in E.D. for different vowels analysed under similar conditions was considered important for the design of subsequent analysis schemes.

The euclidean distance separation between segments of an utterance which were not sequential did not show any significant increase over that between sequential segments, in the majority of cases. Thus the results presented, represent the level of statistical instability rather than a gradual movement from one pattern to another.

The result of this experiment was that a smoothly varying numerical representation of a trend, already observed qualitatively, was observed. This gave a degree of confidence in the interpretation of the numerical results to follow. It has also enabled the extension of the earlier work, to check that all the vowels behave in a similar way. However it is important to observe that different vowels analysed at the same duration of compilation have very different E.D. separation figures. It is therefore obvious that a separation of N units of E.D. does not have the same meaning for each vowel. There must therefore always be a relative scale against which vowel separations are measured.

On the basis of the results of this experiment it was possible to choose a duration of compilation of statistics, most useful for future analysis. It was desirable to use the shortest duration compatible with statistical stability, in order that rapid phonetic changes could be resolved in a full recognition system working on continuous speech.

The value of 50 msecs. was chosen.

3.4.2 The measurement of separation on a relative scale.

The separation between digrams compiled in different temporal segments of the same utterance of various vowels has been found to vary quite widely. In each measurement relating two speech categories, it is important to derive the spread in N-dimensional space within both categories that are to be compared. The speech category can be either a vowel, defined by several utterances of it, or simply one utterance of a vowel, defined by several sequential segments of it. Initially the latter case only was investigated.

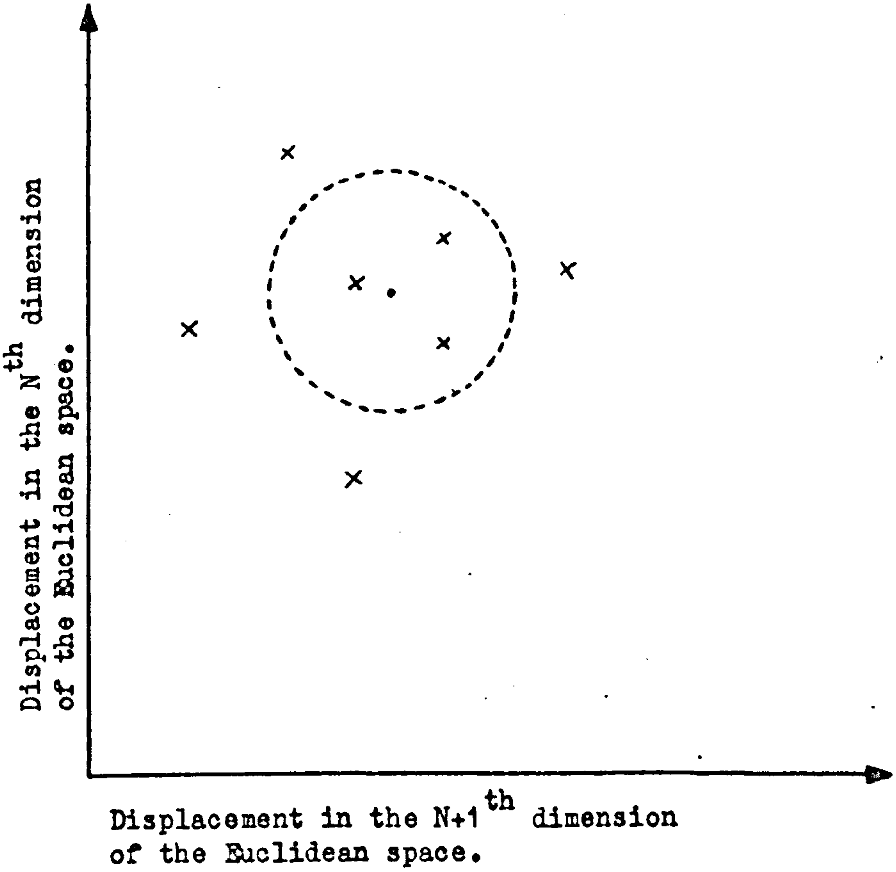
The method used in the subsequent experiments was as follows. Each utterance that was to be compared with another utterance was divided into eight sequential 50 msec. segments. A measure of the spread of the digrams of these segments in euclidean space was made, in order to define the 'territory' of each utterance. At the same time, one of these sequential digrams was chosen to represent the utterance for comparisons with the representative digrams of other utterances.

The centre of the distribution of sequential digrams was found by computing the mean digram. This was simply a digram whose bin contents were the mean of the contents of the respective bins of the sequential digrams. The E.D.s of each of these sequential digrams from the mean were computed, and their mean and standard deviation found.

The choice of the representative digram.

The digram chosen to represent the utterance in further analysis was the one closest to the mean digram in euclidean space. This digram was chosen rather than the mean digram itself for the following reason. When viewing the real-time digram display for a continuously uttered vowel sound, certain discrete movements were noticed apart from certain continuous movements. Such discrete movements could typically be caused by the loss or gain of two T.I.s in the sequence of intervals. Rather than choose a mean digram, which might be represented by a point in euclidean space which was midway between occurrent or even possible positions for the utterance concerned, the digram which was nearest to the mean digram position was chosen as typical of the utterance. This choice could be important if the sound were to be synthesised from the stored statistics in an active analysis-by-synthesis procedure. This form of representative statistic was used throughout the work described in the present chapter. In subsequent work, when the effects of mixing the statistics of several speakers or of several differently pitched utterances were investigated, the mean statistic was chosen to represent the speech category. In these later experiments there was no intention of basing synthesis on the representative statistics. They were simply a measurement used in the recognition process.

The choice of the representative digram is described in diagrammatic form in figure 3.3. The N-dimensional space has been projected on to two dimensions. In addition to the representative digram being chosen, a circle whose radius is the mean distance of the sequential digrams from the mean digram was drawn,with the mean digram


Displacement in the N th dimension of the Euclidean space.
Displacement in the N+1 th dimension of the Euclidean space.

● Position of the mean digram.

× Positions of sequential digrams taken from within the same utterance.

Fig. 3.3 A two dimensional projection of the Euclidean measurement space to illustrate the measurement of the spread of sequentially compiled digrams.

as centre. The distribution of distances of the sequential digrams from the mean digram approximated to a normal distribution. This was a sufficient condition for the statement that 50% of the digrams lay within this circle. It proved most difficult to derive a measure of the probability of confusion on the basis of overlap of such circles of different vowels. The translation of measurements of a normal distribution of distances, being a single dimensional model, to probabilities in a 256 dimensional model is not a simple matter. Such analysis was not attempted at this stage. Later on in this study this difficulty was avoided, and the probabilities found using a different method (chapter 4).

Computer program for this analysis.

Computer programs (fig.3.4) to perform the analysis described above were basically similar to that described in section 3.3.3. The major difference was in the choice of output data. The important outputs of these programs were the radius of the mean distance hypersphere in N dimensional space, the standard deviation about that mean, and the bin contents of the representative digram. It was also possible to obtain the bin contents of the mean digram and the individual E.D.s of the sequential digrams from the mean. At an earlier stage,prior to entrance into the Fortran program, the bin contents of each of the sequential digrams could be obtained if required for detailed analysis or diagnostic purposes.

The output of the above important data was via the ASR-33 teletype. The bin contents of the digrams were printed out in a two

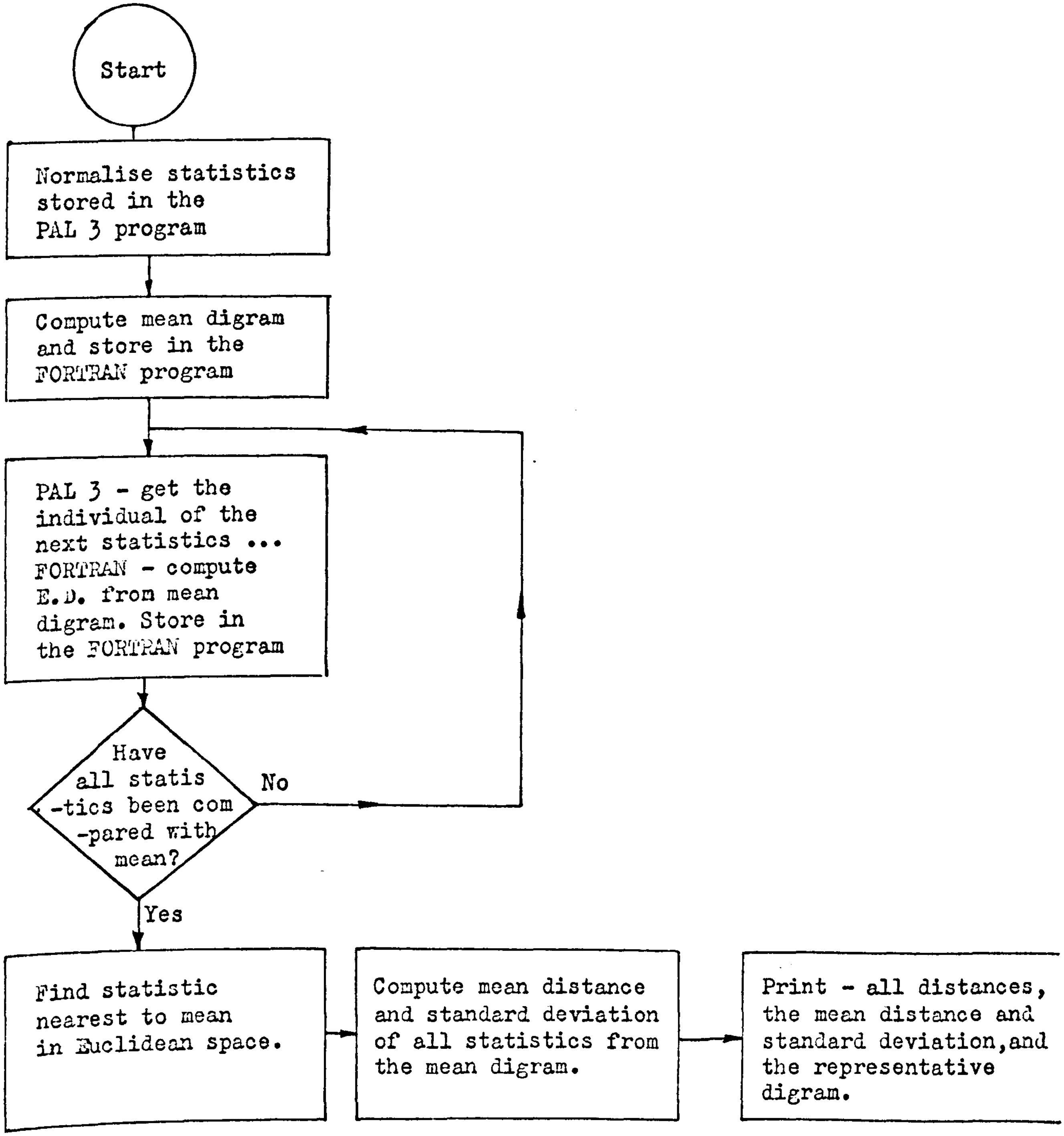

Start
Normalise statistics stored in the PAL 3 program
Compute mean digram and store in the FORTRAN program
PAL 3 - get the individual of the next statistics ... FORTRAN - compute E.D. from mean digram. Store in the FORTRAN program
Have all statis -tics been com -pared with mean?
No
Yes
Find statistic nearest to mean in Euclidean space.
Compute mean distance and standard deviation of all statistics from the mean digram.
Print - all distances, the mean distance and standard deviation,and the representative digram.

Fig. 3.4 Flow diagram for the computation of the representative digram.

dimensional array, thus facilitating comparison with earlier qualitative results. The bin contents of the representative digram were also recorded on paper tape, in preparation for re-entry into the machine for inter-utterance analysis.

3.4.2.1 The estimation of likely confusions between vowels on comparison of their representative statistics.

A further hybrid program was used to operate on the representative statistics produced on paper tape. As there was no storage space available in the Fortran program to contain these representative statistics, they were transferred into another part of the store by a PAL 3 program as soon as they were accepted from the papertape reader. The form of the program from this point onwards was the same as shown in figure 3.1.

The elements of the output matrix referred, in this program, to the inter-vowel E.D.s. These inter-vowel separations were then related to the known spread for each of the vowels. Hyperspheres of radii equal to the mean spread, and the mean spread plus one, two and three standard deviations, were constructed around each vowel digram to define several confusion thresholds. The likely confusions based on these various thresholds could then be estimated. The results were expressed as percentages of the total number of vowel pairs which were separated from one another by this analysis.

3.4.3 Experimental Results.

Sounds used in these experiments.

Three speakers (W.A.A., M.J.U., J.B.M.) recorded the twelve vowel sounds previously described, on language master cards. These utterances were used as input to the clipping circuitry, with 2.5 Kc/s low pass filtering as standard preprocessing.

3.4.3.1 Vowel discrimination using unidirectional Z.C. histograms.

The three sets of twelve utterances were processed three times. Once using bidirectional Z.C. information, and once each using the two sets of unidirectional Z.C. information. The likely confusions between vowels using these three forms of T.I. histogram were computed according to the methods outlined in section 3.4.2.

The number of confusions between vowels, when unidirectional Z.C.s of both signs are used for histogram compilation, is seen to increase more quickly as the confusion threshold is lowered, than when bidirectional Z.C. histograms are used. Confusions likely when using the two unidirectional Z.C. histograms are seen, however, to be similar. There is no evidence of an advantage in the use of either + Z.C.s or - Z.C.s. The results (fig.3.5) show that this trend is true for the utterances of all three speakers.

A supplementary experiment was done using synthetic speech (produced by the parametric synthesiser). The confusions calculated for these sounds were found to be just slightly less than those for the three sets of natural utterances in the bidirectional Z.C. case but substantially less in the unidirectional case. The reason for this lack of increase in the number of confusions for a unidirectional Z.C., T.I. histogram description of synthetic speech is not clear. Note that

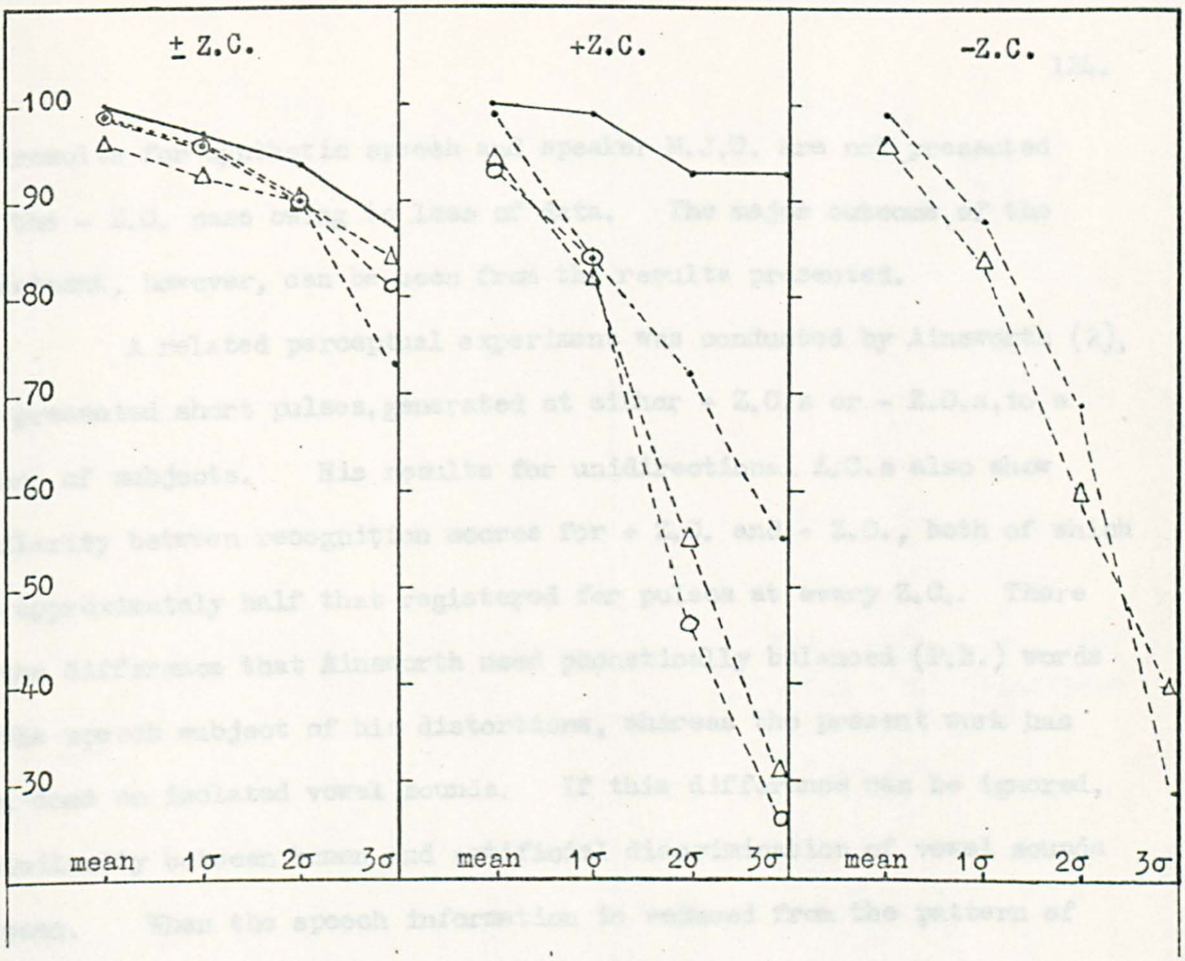


Fig. 3.5 Percentages of vowel pairs separated by T.I. histogram patterns of bidirectional and unidirectional Z.C.s, for the speech of three speakers and for synthetic speech.

Values are plotted for four confusion threshold levels.

Δ Speaker W.A.A.

o Speaker M.J.U.

• Speaker J.B.M.

Solid line joins values for synthetic speech.

the results for synthetic speech and speaker M.J.U. are not presented for the - Z.C. case owing to loss of data. The major outcome of the experiment, however, can be seen from the results presented.

A related perceptual experiment was conducted by Ainsworth (2), who presented short pulses,generated at either + Z.C.s or - Z.C.s,to a number of subjects. His results for unidirectional Z.C.s also show similarity between recognition scores for + Z.C. and - Z.C., both of which are approximately half that registered for pulses at every Z.C.. There is the difference that Ainsworth used phonetically balanced (P.B.) words as the speech subject of his distortions, whereas the present work has been done on isolated vowel sounds. If this difference can be ignored, a similarity between human and artificial discrimination of vowel sounds is seen. When the speech information is reduced from the pattern of bidirectional Z.C.s to the pattern of unidirectional Z.C.s, human discrimination scores are halved. A similar reduction is seen when the T.I. histogram information of these patterns of Z.C.s is used as the basis of artificial discrimination.

3.4.3.2 Vowel discrimination using differentiated speech.

In these experiments differentiation was used in addition to 2.5 K c/s low pass filtering in the preprocessing stage. The results for histogram and digram statistics are presented. Similar trends can be seen in the effect of differentiation on both statistics for all three speakers and for synthetic speech. (fig.3.6). There is a general increase in the number of confusions calculated at all threshold levels

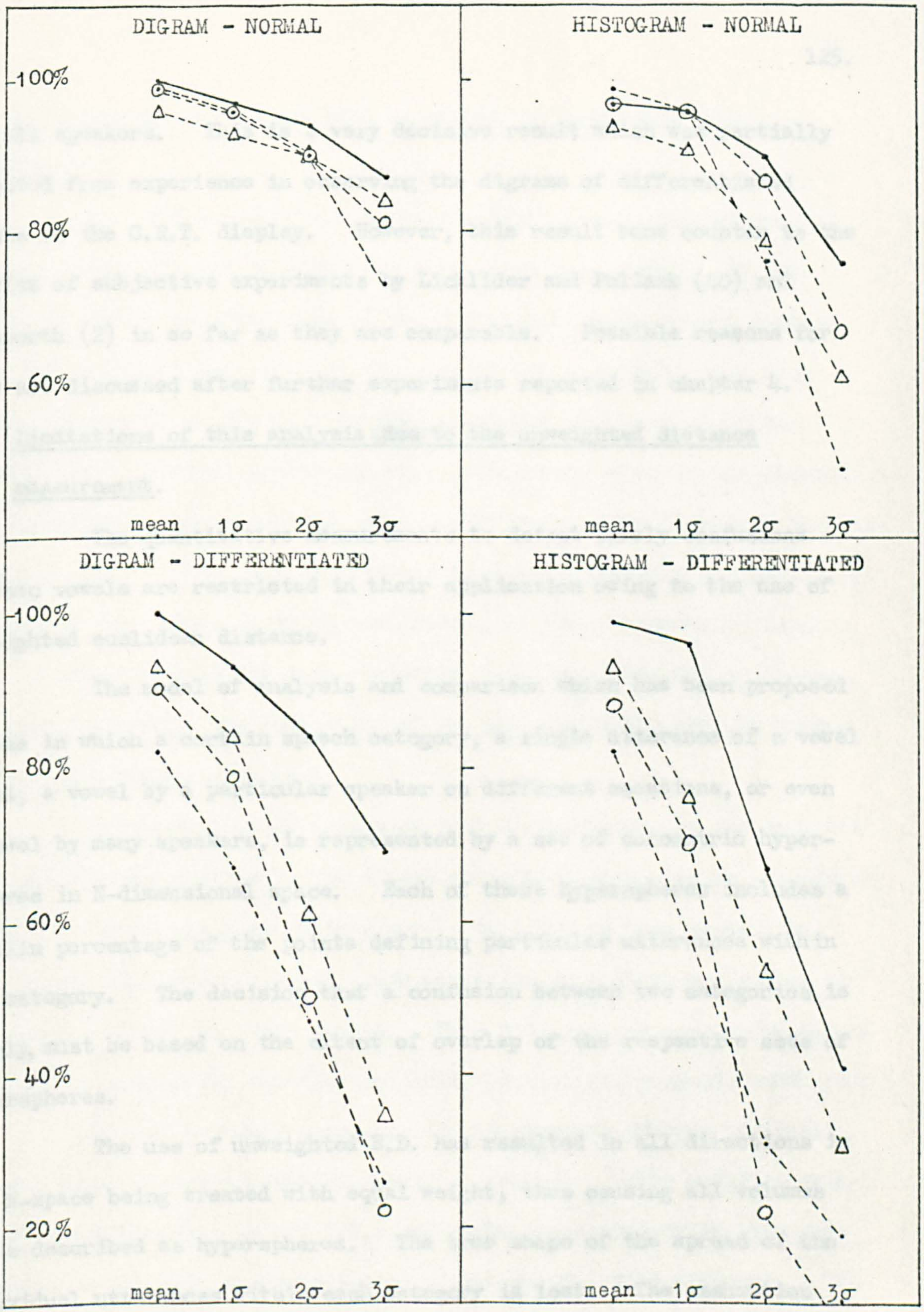


Fig. 3.6 Percentages of vowel pairs separated by T.I. histogram and digram patterns of the normal and differentiated speech of three speakers and for synthetic speech. Values are plotted for four confusion threshold levels. Code as figure 3.5.

for all speakers. This is a very decisive result which was partially expected from experience in observing the digrams of differentiated vowels on the C.R.T. display. However, this result runs counter to the results of subjective experiments by Licklider and Pollack (40) and Ainsworth (2) in so far as they are comparable. Possible reasons for this are discussed after further experiments reported in chapter 4.

3.5 Limitations of this analysis due to the unweighted distance measurement.

The quantitative measurements to detect likely confusions between vowels are restricted in their application owing to the use of unweighted euclidean distance.

The model of analysis and comparison which has been proposed is one in which a certain speech category, a single utterance of a vowel sound, a vowel by a particular speaker on different occasions, or even a vowel by many speakers, is represented by a set of concentric hyperspheres in N-dimensional space. Each of these hyperspheres includes a certain percentage of the points defining particular utterances within the category. The decision that a confusion between two categories is likely, must be based on the extent of overlap of the respective sets of hyperspheres.

The use of unweighted E.D. has resulted in all directions in the N-space being treated with equal weight, thus causing all volumes to be described as hyperspheres. The true shape of the spread of the individual utterances within each category is lost. The assumption that no weighting is required will be called the hypersphericity assumption.

The acceptance of the hypersphericity assumption had unequal effects on the analysis of histogram and digram statistics. In the latter, the number of dimensions of the hypersphere that did not contain variation was far larger than in the case of the histogram, as a much smaller proportion of digram bins than histogram bins were occupied. Therefore more spurious confusions were calculated, owing to the assumption of hyperspherical distributions, when digram rather than histogram analysis was being used. The only experiments that could usefully be done using this system were those in which comparisons were made between various types of processing on the same statistic, that is, histogram or digram.

3.6 Improvements in the definition of the analysis space and the method of difference measurement.

The assumption of hyperspherical distributions in the analysis space has clearly restricted the direct comparison of histogram and digram analysis. There are also further errors implicit in this assumption. It is likely that certain vowels have variation in a different number of dimensions compared to other vowels. This will give an uneven overestimation of the confusions actually occurring between all possible vowel pairs.

It was obviously necessary for the success of this analysis that the hypersphericity problem should be solved. This could be done by making measurements on the T.I.s and their statistics, which were more dependent on the T.I. characteristics of the signal being measured. The previous measurements had assumed hyperspherical distributions rather than the actual distributions of the vowel statistics in the analysis space.

The analysis space itself could be made more specific to the particular vowels being analysed,rather than being defined somewhat arbitrarily on the basis of general observations of the characteristics of vowel T.I. distributions.

3.6.1 Empirical measurements to derive the optimum shape for the analysis space.

The work of Fourcin (23) revealed that a long term T.I. distribution for continuous speech showed certain features which could be interpreted as the contributions of certain classes of sounds. Vowel sounds,with their strong isolated frequency components in the middle of the speech band,were suggested to be responsible for the 'plateau' in the middle of the distribution.

A computer program was written to make a similar measurement for all the twelve vowels used in these studies. The aim of this measurement was to investigate whether an empirically derived time scale could be a variable between different sets of vowel utterances, and so be useful as a 'tune-in' adjustment to a new speaker. Differences in the long term T.I. distributions for various speakers were found by Fourcin. He found them mainly in the plateau region which he attributed to vowel sounds. Therefore in studying only the vowel sounds a significant difference was expected.

3.6.2 The measurement of 'equiprobable bin divisions'.

The measurement of the shape of the T.I. distribution was approached in a different way from that of Fourcin. The aim was to specify histogram bins which would be equally filled when equal segments of each of the twelve vowels were presented for T.I. measurement and accumulative histogram compilation.

The equiprobable bin divisions are defined as follows. Let P_{ij} be the probability that a T.I. of length i occurs in vowel j, then

$$\sum_{i=i_{min}}^{i_{max}} \sum_{j=1}^{12} P_{ij} = 1$$

The equiprobable bin divisions are given by

$$\sum_{i=i_n}^{i_{n+1}} \sum_{j=1}^{12} P_{ij} = 1/N$$

where N is the total number of equiprobable bins and $i_n < i < i_{n+1}$ is the range of the nth bin. The use of this scaling of the T.I. dimension made the measurement space relative to the particular set of vowels within which discrimination was required. As the overall probability of each bin's having the same contents is the same, this choice of bin divisions provides a measurement space of maximum information capacity for the set of vowels concerned, if all bins have equal weight (68).

The computer program to perform this transformation is described by its flow diagram in figure 3.7. A T.I. histogram of the twelve vowels was compiled on any convenient T.I. scale. A large number of small bins was preferred, to reduce any errors due to linear interpolation at a later stage. This histogram was used as the input to the program. It was normalised and the probability content of each bin was found. These probabilities were compared with the desired probability limits of the set of equiprobable bins. The new equiprobable bin divisions were calculated in terms of the input bins by a process of linear interpolation.

3.6.3 The results of equiprobable bin measurements.

The variation of the overall distribution of T.I.s of the twelve vowels, when spoken by different speakers, was measured. The sets of vowels recorded on language master cards by speakers W.A.A., J.B.M., M.J.U. were used in this experiment. A L.P. filter of 2.5 Kc/s was used in the preprocessing stage as previously described.

The plots of the time intervals defining the equiprobable bin divisions for these three speakers are shown in figure 3.8 a. The plots are seen to differ for each speaker, but all approximate to a certain general shape. The two bins at either end of the distribution include all T.I.s greater than, or less than the respective extreme bin divisions.

To assess the significance of the variation observed in figure 3.8 a, three utterances of J.B.M. with glottal periods of 8.0 msec, 6.5 msec.

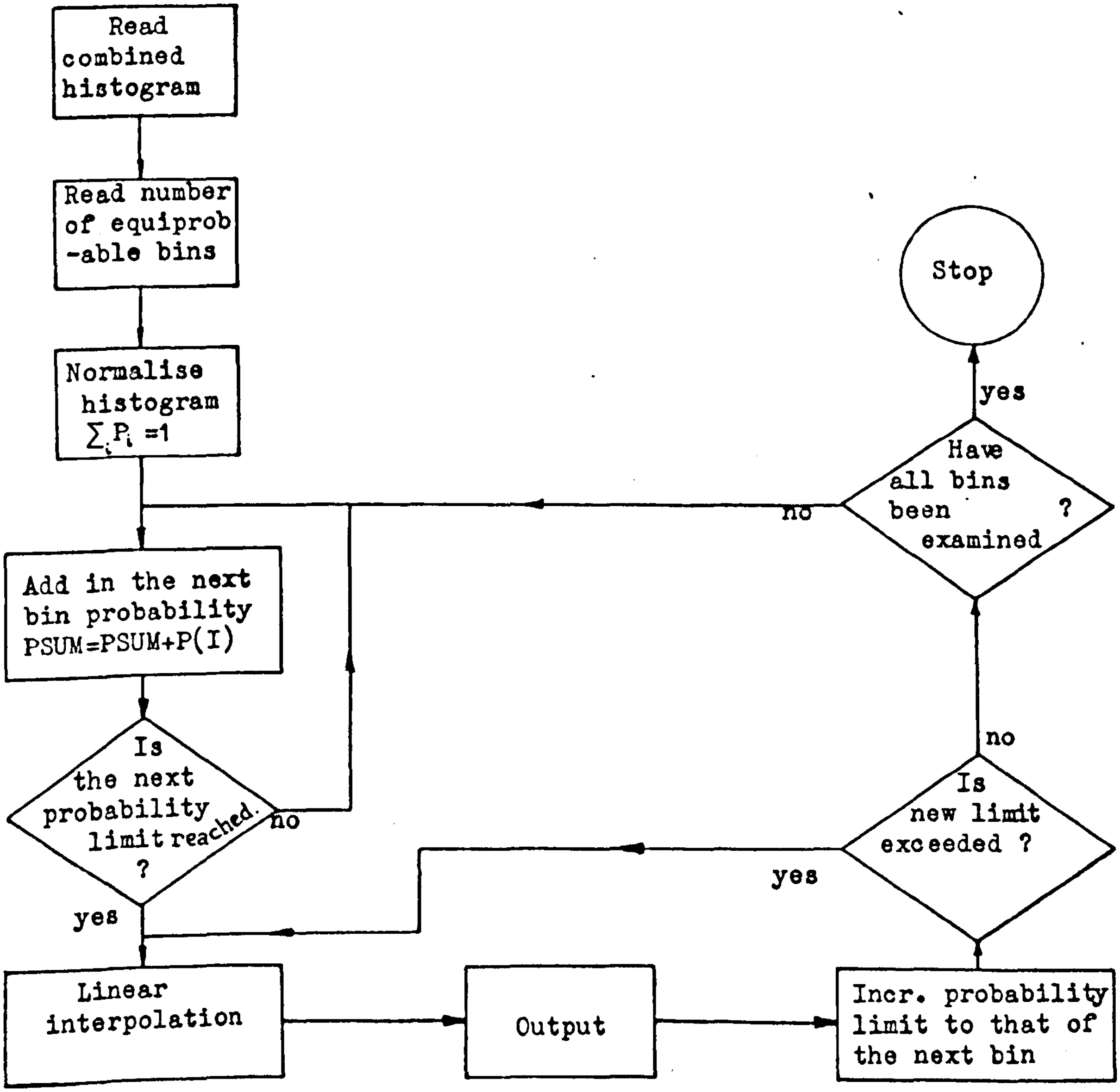

Read combined histogram
Read number of equiprob -able bins
Normalise histogram $\sum_i P_i = 1$
Add in the next bin probability PSUM=PSUM+P(I)
Is the next probability limit reached ?
no
yes
Linear interpolation
Output
Incr. probability limit to that of the next bin
Is new limit exceeded ?
yes
no
Have all bins been examined ?
no
yes
Stop

Fig. 3.7 Flow diagram of the Equiprobable bin division program.

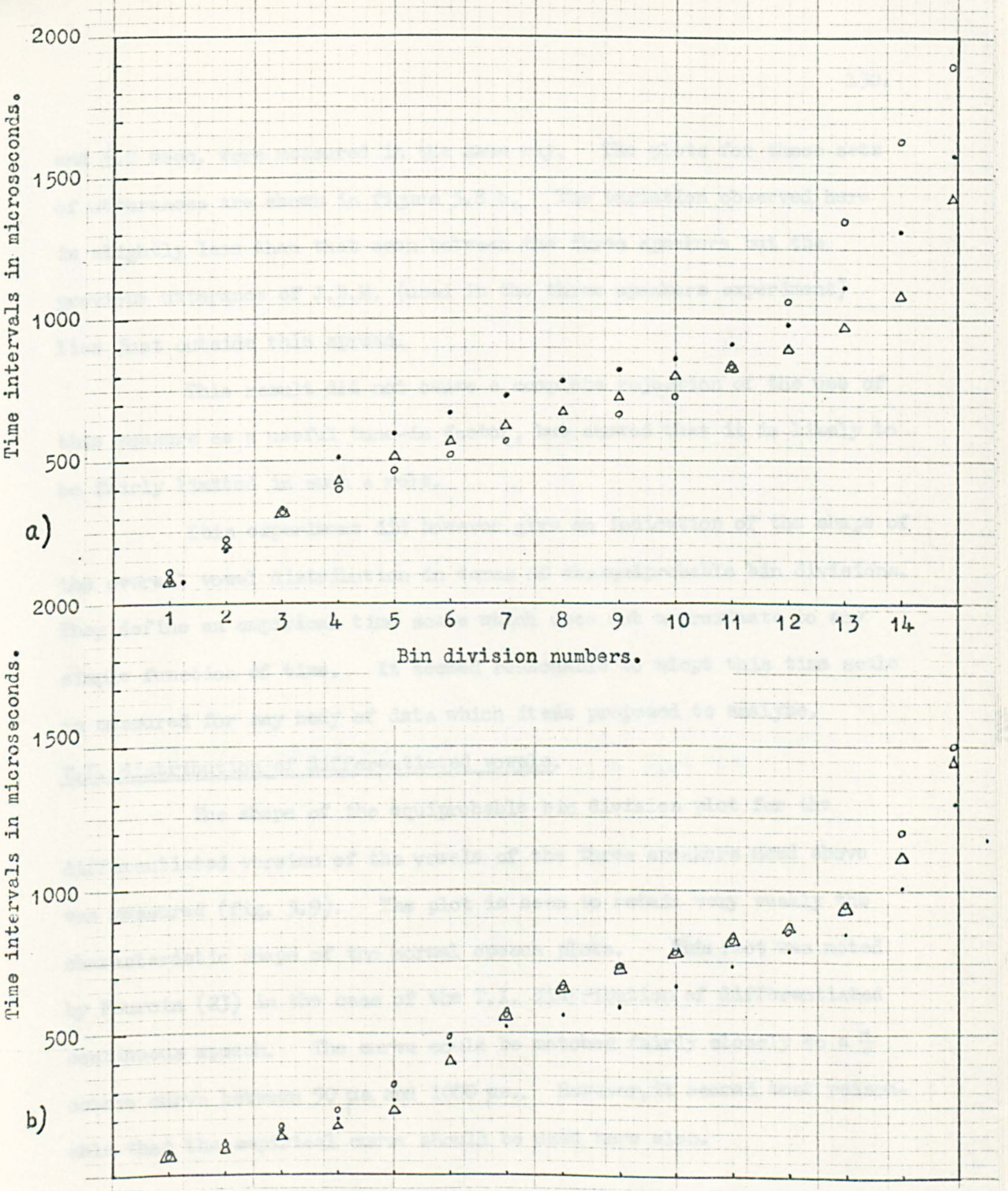


Fig. 3.8 Equiprobable bin divisions plot for the vowels of three speakers, and three sets of vowel utterances of a single speaker. Code as in figure 3.5

and 5.0 msec. were measured in the same way. The plots for these sets of utterances are shown in figure 3.8 b. The variation observed here is slightly less than that seen between the three speakers but the previous utterance of J.B.M. (used in the three speakers experiment) lies just outside this spread.

This result did not cause a complete rejection of the use of this measure as a useful tune-in factor, but showed that it is likely to be fairly limited in such a role.

This experiment did however give an indication of the shape of the overall vowel distribution in terms of the equiprobable bin divisions. They define an empirical time scale which does not approximate to any simple function of time. It seemed reasonable to adopt this time scale as measured for any body of data which it was proposed to analyse.

T.I. distribution of differentiated vowels.

The shape of the equiprobable bin division plot for the differentiated version of the vowels of the three speakers used above was measured (fig. 3.9). The plot is seen to retain very weakly the characteristic shape of the normal speech plots. This fact was noted by Fourcin (23) in the case of the T.I. distribution of differentiated continuous speech. The curve could be matched fairly closely to a $\frac{1}{3}$ octave curve between 50 µs. and 1000 µs.. However, it seemed most reasonable that the empirical curve should be used here also.

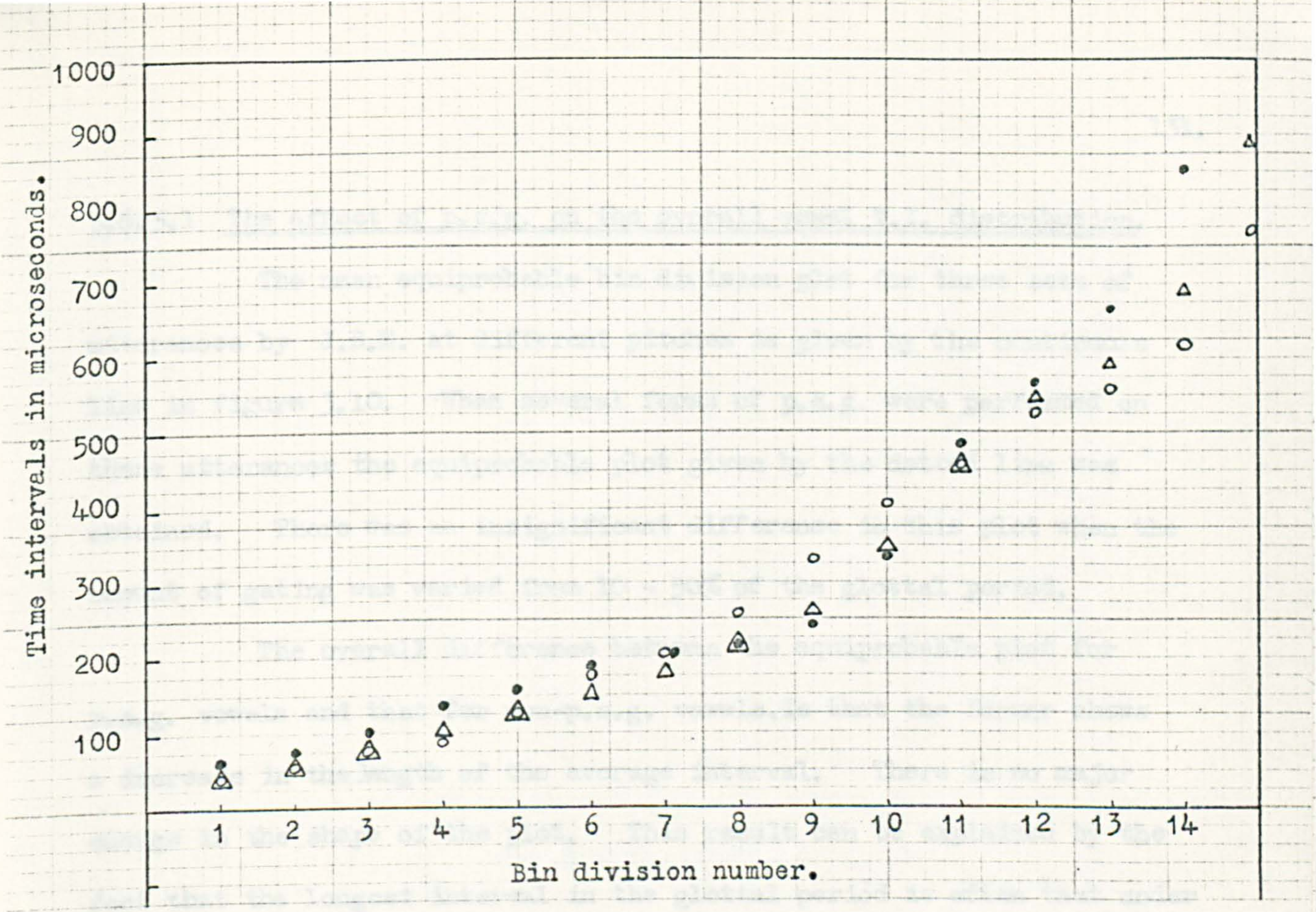


Fig. 3.9 Equiprobable bin division plot for differentiated vowels of three speakers. Code as in figure 3·5.

3.6.3.1 The effect of p.s.g. on the overall vowel T.I. distribution.

The mean equiprobable bin division plot for three sets of utterances by J.B.M. at different pitches is given by the continuous line in figure 3.10. When several forms of p.s.g. were performed on these utterances the equiprobable plot given by the dotted line was obtained. There was an insignificant difference in this plot when the amount of gating was varied from 10 - 50% of the glottal period.

The overall difference between the equiprobable plot for p.s.g. vowels and that for non-p.s.g. vowels, is that the former shows a decrease in the length of the average interval. There is no major change in the shape of the plot. This result can be explained by the fact that the longest interval in the glottal period is often that under the maximum amplitude lobe of the waveform, as it often engulfs the last two intervals of the previous period. This interval is the first to be rejected by the p.s.g. procedure. This result indicates that after this initial decrease in the length of the mean interval, any further gating effects time intervals of all values evenly.

The difference in the equiprobable plot caused by p.s.g., is within the spread experienced between different utterances of the same speaker, thus it was not considered important to vary the time scale when using p.s.g. analysis.

3.6.4 Methods to overcome the problem of hypersphericity.

The empirically derived equiprobable bin distribution was designed to make the T.I. measurement space more 'speech-shaped'.

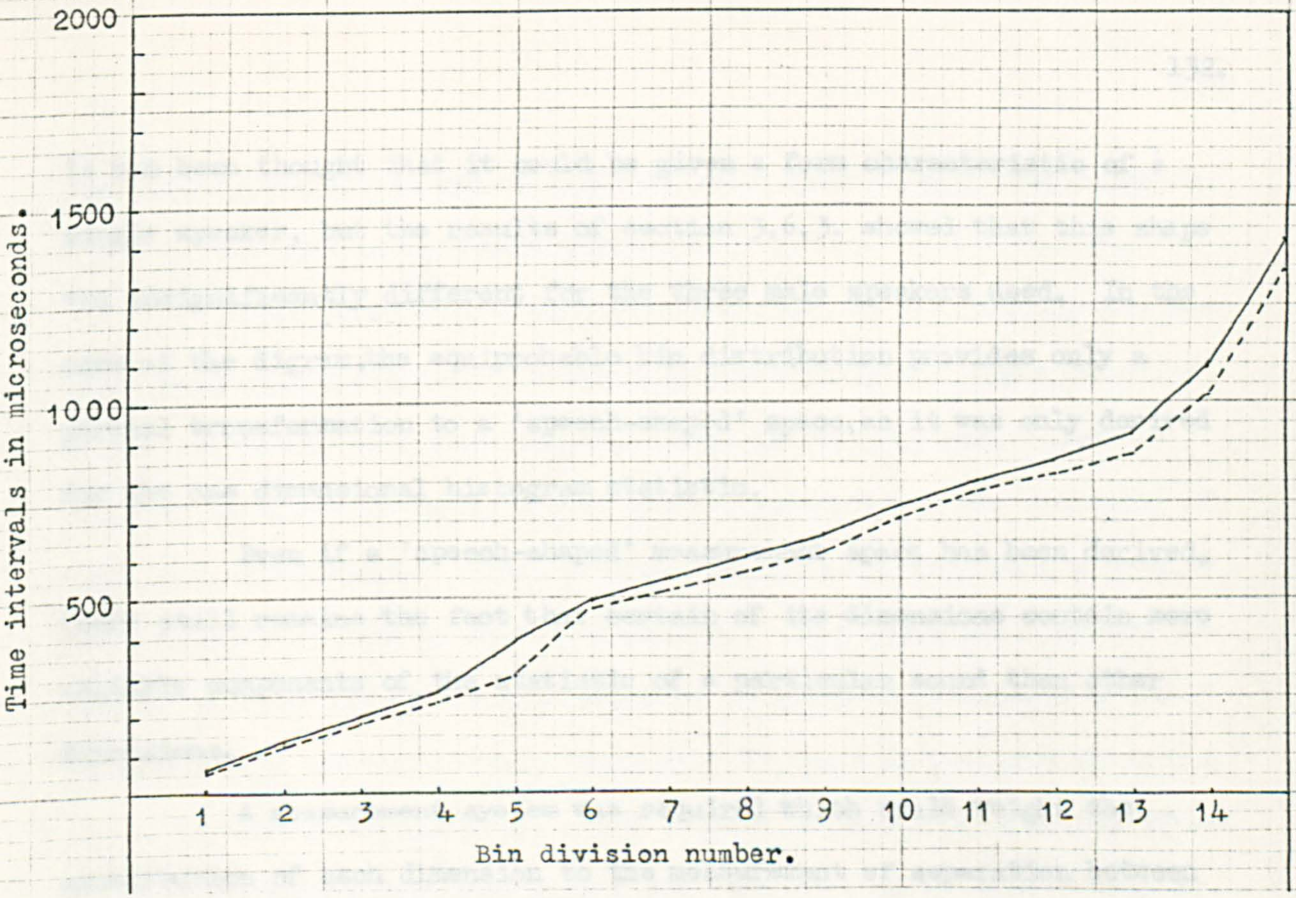


Fig. 3.10 Equiprobable bin division plot for vowels after p.s.g..

It had been thought that it could be given a form characteristic of a single speaker, but the results of section 3.6.3. showed that this shape was insignificantly different for the three male speakers used. In the case of the digram,the equiprobable bin distribution provides only a partial transformation to a 'speech-shaped' space,as it was only derived for the one dimensional histogram statistic.

Even if a 'speech-shaped' measurement space has been derived, there still remains the fact that certain of its dimensions contain more reliable components of the statistic of a particular sound than other dimensions.

A measurement system was required which would weight the contribution of each dimension to the measurement of separation between two statistics. This weighting would be proportional to the variation of the contents of this dimension in one or both of the vowels being compared. These weights for every dimension would replace the previous estimates of spread within each utterance, which were given in terms of the scalar distances between the positions,in euclidean space,of the variant forms of the statistic. In this way the shape as well as the size of the spread is measured.

Two methods of analysis of these data are possible. Firstly the spreads of two statistics could be matched dimension by dimension to check for overlapping distributions of bin contents. This method would necessitate the computation of the probability of confusion for each dimension. A final decision on the probability of confusion, that is of identifying the statistics as similar, could then be made from a combination of the results from each dimension.

The second method is a simplification of the first, by the use of a recognition system. The absolute probability that the statistics of two utterances represent the same vowel is not required. What is required is a rank ordering of the vowels, according to the proximity of their T.I. statistics. This ordering could be derived using a measure monotonically related to this probability. In a recognition system the weights of only one of the statistics to be compared (statistic A) would be used. The other statistic (statistic B) would be compared in such a way as to answer the question, "How similar is statistic B to statistic A, bearing in mind that statistic A can vary as described by its weights for each dimension?". This simplification amounts to the fact that a measure of the probability of B being confused with any member of the spread of A is computed, but not the overall probability of an A - B confusion.

This method may be implemented by using the E.D. measurement and weighting the contribution of each dimension. The resulting weighted E.D. could only become meaningful in terms of likely confusions when compared with the numbers similarly obtained when statistic B is compared with all the other statistics of the set. The comparison yielding the smallest separation in weighted euclidean space would indicate the most likely identity of statistic B.

This method is similar to that used by Bezdel and Chandler (5) in their recognition experiment on five English vowels.

The computer storage to contain the mean statistics plus their weights and each of the sequentially compiled statistics of an utterance which it was desired to recognise, was not available until phase 2 of the PDP-8 system was installed. Experiments using the second method proposed above are described in chapter 4.

Chapter 4. Quantitative analysis using a vowel recognition system.

4.1 Introduction.

The approach to the quantitative analysis of the T.I. histogram and digram statistics was modified to include empirically derived weighting of the dimensions of the measurement space, which itself was based on the equiprobable bin distribution.

A vowel recognition program based on this analysis was written for the PDP-8 (including the second phase of the installation - see appendix 4). The possibility of artificial recognition based on both histogram and digram analyses of the vowels of several speakers was investigated. More detailed studies were made on the recognition of a single speaker's utterances, when only some of these utterances were used to compile the reference statistics. The effects of pitch on vowel recognition were examined, together with the counter effects of pitch synchronous gating. Further checks were made on the effect of pre-clipping differentiation of the waveform, which confirmed the results previously reported. Finally, the effect of grouping the vowels according to an articulatory model was investigated.

4.1.1 Comparison of statistics using a weighted euclidean distance.

The weighting factor chosen to give emphasis to bins in which there was little variation was

$$\omega_i = {}^{1}/\sigma_i$$

where σ_i was the standard deviation of the contents of the ith bin, as measured in sequentially compiled statistics from a single utterance. This weighting was chosen as it represented the same variation expressed by the within-utterance spread,used in the previous analysis. This was the minimum amount of variation that could be considered as characteristic of the sound. That between different utterances and between different speakers could be added at a later stage by combining these weights derived for single utterances.

4.1.2 Outline of the recognition system.

The basic purpose of the recognition system was to compare, using the weighted E.D. measure, the T.I. statistics of a sound presented to the computer in Z.C. pulse form, with the stored statistics of all the sounds from which it was hoped to recognise the newly presented sound. These stored statistics comprised the mean statistic of an utterance, or set of utterances, of a given vowel,plus the standard deviation of each bin of that statistic. A set of twelve such statistics for all the vowels will be referred to as a set of reference statistics.

4.1.3 Experimental aims.

The aim of these experiments was to answer the following questions. "What is the relative importance of digram information as compared to histogram information,for the discrimination of vowel sounds according to their T.I. distributions?" "What effect do the perturbations of the T.I. distributions due to pitch have on vowel discrimination?". "How effective is the counter measure of p.s.g. in removing pitch effects which cause reduced vowel discrimination, from the T.I. distribution?".

These questions were not answered in chapter 3 owing to,either the restrictions of the hypersphericity assumption,or lack of time while using the earlier more laborious method. The present experiments are based on the more realistic separation measure of weighted E.D., in the context of a recognition system.

4.2 Description of programs.

The programs necessary to conduct the recognition experiments fall into two categories. Firstly, those concerned with the computation and storage of the reference statistics, and secondly, those concerned with the comparison of newly presented vowel sound statistics with the reference statistics, and the recognition of the presented sound on the basis of this comparison.

4.2.1 Computation and storage of reference statistics.

It was necessary to construct a store of reference statistics for each vowel sound that was to be recognised. These reference statistics comprise the mean contents expected in each bin and the standard deviation about this mean for each bin.

The T.I. statistics, in this case 256 bin digrams, were compiled for ten sequential 50 msec. durations of a single utterance, and were retained in the core store (see appendix 4.B). These statistics occupied over one quarter of the core store available. The mean digram and the standard deviation of the contents of each of the mean digram's bins were calculated. The original statistics were destroyed in this process owing to lack of space in the computer store. The mean digram and its standard deviations were then transferred to digital magnetic tape. When this analysis had been completed for each of the twelve vowels, a full 'single utterance reference set', that is a reference set in which each vowel is typified by a single utterance, was stored on the magnetic tape.

4.2.2 The computation of mixed reference sets.

It was required to obtain more general descriptions of the characteristics of each vowel,rather than that allowed by the measurement of several sequential segments of a single utterance. A program was written to mix the single utterance reference sets to obtain a 'mixed reference set'. The mixture could be, for example, of reference sets of single utterances at different pitches by the same speaker, or of reference sets of single utterances by several speakers. The mathematical derivation of the mixing algorithm is presented in appendix 7.

4.2.3 Comparison and recognition programs.

The function of this part of the system was to compile the statistics of a newly presented vowel and compare them with the reference set,by computing weighted E.D.s. The vowel whose reference statistic was found to be nearest to the newly presented vowel's statistic,was identified with this presented vowel.

The program measured statistics of only eight sequential segments of 50 msec. because of storage limitations. These limitations were caused by the need to recall,from magnetic tape into core store,the reference statistics of each vowel, one at a time. These reference statistics need twice the space of the basic unprocessed statistics owing to the addition of standard deviation values for each bin.

Instead of producing a mean value of these eight statistics and obtaining a single euclidean distance from each reference statistic, which would of course give the value zero if the same single utterance reference was used, all eight actual digrams were compared to each reference statistic. This preserved the within-utterance variability factor in the T.I. statistics,through to the final act of recognition. Using this method,it was not necessary to make any assumptions about the variation of the statistics within the utterance and the uncertainty in recognition that this variation would cause. The probability of the presented vowel being recognised as a particular vowel,was indicated simply by the score out of eight in the final stage.

For each presented vowel, the program computed 96 weighted E.D.s. It then searched for the minimum distance in each of the eight groups of twelve distances, which indicated the separation of the eight sequential segments of the utterance from each of the twelve vowel reference statistics. The vowels which coincided with these minima were punched out in numerical code on the high speed paper tape punch. The flow diagram of this program is shown in figure 4.1.

4.2.3.1 The use of digital magnetic tape storage.

The full PDP-8 system, described in appendix 4, was available from the outset of the use of these recognition programs. This enabled

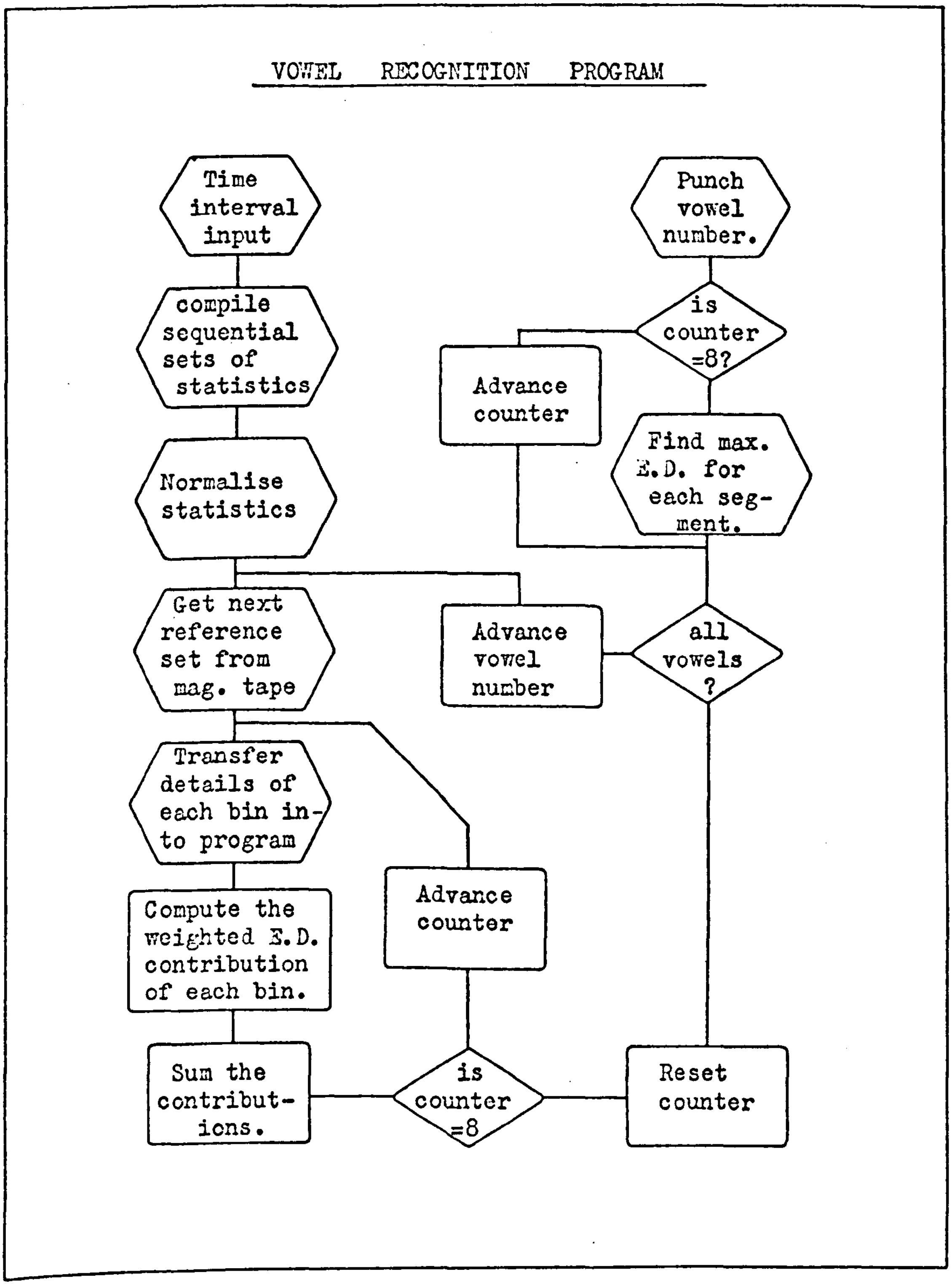

VOWEL RECOGNITION PROGRAM
Time interval input
compile sequential sets of statistics
Normalise statistics
Get next reference set from mag. tape
Transfer details of each bin into program
Compute the weighted E.D. contribution of each bin.
Sum the contributions.
is counter =8
Advance counter
Reset counter
Advance vowel number
all vowels ?
Punch vowel number.
is counter =8?
Advance counter
Find max. E.D. for each segment.

Fig. 4.1 Flow diagram of the vowel recognition program.

two spools of digital magnetic tape to be in use during the same program, although not at the same time. Initially one spool was used to store the reference statistics required by the program, and the other was used to store the values of weighted E.D., produced and used by the program. Only the results of the recognition procedure were punched out on paper tape. This magnetic record of the E.D. measurements enabled a check to be made if unexpected results were obtained. It was most useful in the early stages of this experiment. The tapes on the second unit were later used to store the T.I.s of the waveforms of all the utterances which it was desired to recognise. This was found to be exceedingly useful as it eliminated the relatively high chance of error in the continual presentation of vowel sounds to the computer from the Language Master recorder. This process became more error prone when information concerning the start of the glottal period was required for experiments on p.s.g.. Various vowel sounds required different adjustments to the dead time of the glottal period detector for this period to be detected accurately. This was in addition to the need to monitor the gain of the pre-amplifier, in order to maintain an input voltage compatible with accuracy in the clipping amplifier.

The T.I.s of each sound together with the glottal period markers were measured in the normal way. The block of core store containing them was then transferred to magnetic tape. Only when different preprocessing, or the speech of a different speaker, was required for analysis, were new intervals recorded on to magnetic tape in this way. The same intervals were used for many variants of analysis to be described in this chapter. A major advantage of this system was that the recognition programs could run untended for periods up to nine hours.

4.3 T.I. analysis of vowels by three speakers.

A series of experiments were done in which the vowel utterances of three speakers were used as inputs and references of the recognition program. These experiments can be divided into two parts; those where the reference statistics were derived from the utterances of single speakers, and those where they were derived from the utterances of all three speakers.

4.3.1 Experiments using same speaker as reference.

The purpose of these experiments was to obtain an estimate of the within-utterance variability, as measured by the recognition program. As mentioned in section 4.2.3, no measure of this variability was made explicitly during the recognition procedure, but as a sequential set of time segments from each utterance were used as the input, this variability is inherent in all the results.

The link between the variability of the statistics, and the recognition scores obtained in this experiment, must be thought of in the following terms. Variability is seen as the risk of a vowel venturing nearer to the centre of another vowel's territory, defined in the N-space of the statistics, than to the centre of its own territory. This point is made, as it is clear that it is not possible for a histogram description to be more variable than the corresponding digram description, in absolute terms. Variability of the statistics was therefore defined, as above, in terms of the phoneme-shaped hypervolumes in the N-space.

The utterances of each speaker were compared with his own reference statistics derived from the same utterances. Recognition scores were obtained for analyses using digram and histogram statistics of both normal and differentiated utterances. The time scale used in these analyses was the mean equiprobable scale for all three speakers.

SPEAKER.	DIGRAM.		HISTOGRAM.	
	Normal.	Differentiated.	Normal.	Differentiated.
J.B.M.	96.	96.	91.	87.
W.A.A.	96.	91.	94.	69.
M.J.U.	85.	78.	77.	70.

Table 2. Recognition scores (out of 96) illustrating the within-utterance variability of various analyses of the T.I.s of vowels by three speakers.

The within-utterance variability is seen to vary with speaker, type of statistic, and preprocessing of the waveform. The recognition scores (table 2) reveal that M.J.U. has more variable T.I. statistics than the other two speakers. The scores for all speakers show poorer discrimination between sounds when using histogram rather than digram statistics, and when using differentiated rather than normal speech.

This result shows that for each speaker the digram statistics of the normal speech T.I.s is the least variable of the four statistical representations of the vowel phonemes.

4.3.2 Experiments using mixed speaker reference.

The purpose of these experiments was to find what effect the two types of statistic and forms of preprocessing have on the recognition of the utterances of all three speakers, against a mixed reference. The variability of the T.I. statistics of three speakers with differently pitched voices, which had been observed qualitatively in chapter 2, was therefore being investigated. The recognition scores are given in table 3.

SPEAKER.	DIGRAM.		HISTOGRAM.	
	Normal.	Differentiated.	Normal.	Differentiated.
J.B.M.	90.	56.	85.	63.
W.A.A.	84.	73.	80.	61.
M.J.U.	62.	58.	66.	54.

Table 3. Recognition scores illustrating the effect of mixing the references of three speakers.

The most useful analysis of these figures, is in comparison with those obtained in the single speaker experiments just described. They do not show any clear distinctions between the type of statistic or preprocessing. The only pattern that can be discerned is in the averaged differences between the results of this experiment and those of the experiment using single speaker references. For example, in four cases,scores for differentiated speech show a bigger difference than those for normal speech, but in the other two cases the opposite is true. However,the averaged difference shows that the variation due to the mixing of speakers causes a greater reduction in recognition score for differentiated speech than for normal speech. Similarly the digram can be seen to be more sensitive to speaker variation than the histogram. Note that this sensitivity is defined in a similar way to 'variability' in section 4.3.1.

4.3.3 The analysis of confusions in these experiments.

It seemed reasonable to assume that the pattern of confusions between vowels, on the basis of these statistical descriptions, might differ from one group of vowels to another. The open vowels were sometimes found to have T.I. statistics that were very similar to each other. The close vowels occasionally showed such mutual similarity,depending on the relative amplitudes of the first and second formants. It has been seen that differentiation of the vowel /u/ caused the digrams of this vowel,uttered by J.B.M. and M.J.U.,to be more similar to each other (section 2.3.6). Such isolated cases may be indicative of groupings of vowels whose discriminability is dependent on speaker, type of statistic, and preprocessing, in a different way from that of other vowels. The following analysis of confusions was therefore made.

Confusions were mapped on to the cardinal vowel chart and categorised as follows. Those which occurred between adjacent positions on the chart were considered separately from those which occurred over more than one vowel boundary. In this way serious and trivial confusions were separated. The confusions were further subdivided into front-back, open-close,or diagonal shifts in the cardinal vowel space. These correspond respectively to the acoustic dimensions of first formant, second formant,and a mixture of both.

It could be predicted on the basis of the known function of differentiation on the waveform, that fewer second formant confusions should take place after differentiation. The ratio of F1 to F2 confusions should be relatively high and should increase as inter-speaker differences are included. The reverse should be true of normal speech. It was not clear prior to this experiment how histogram and digram confusions would differ.

The choice of the cardinal vowel diagram,as a space into which to map confusions,was made after a preliminary survey of the results. It was obvious that a one dimensional space was unsuitable as parallel, in addition to serial, confusion links between the vowels occurred frequently. A model based on articulatory proximity seemed a reasonable choice, as the long term aim of acoustic analysis of speech is to establish acoustic correlates of articulatory status and movement.

A relative measure such as direction was chosen in preference to specific displacements in the space. The latter was obtainable, when required, from confusion matrices which were plotted for all the recognition experiments.

Plots of the confusions between the various categories when summed for all three speakers, are shown in figure 4.2. It is a convenient shorthand when referring to front-back confusions, open-close, and diagonal confusions,to call them F1, F2 and X confusions respectively.

4.3.3.1 Analysis of trivial confusions.

The pattern of non-diagonal confusions is expressed in terms of the ratio of F1 confusions to the F2 confusions, and the change in this ratio when speaker differences are introduced into the analyses.

Confusions in normal speech histogram analyses were seen to be mainly due to F2 confusion. When inter-speaker differences were introduced, the F1 and X confusions increased, but the increase in the former was not significant for all three speakers. In the corresponding digram analyses, the emphasis on F2 confusions was increased by the speaker variations in the case of two speakers.

In the case of differentiated speech the ratio of F2 to F1 confusions was increased insignificantly when speaker variations were included in the histogram analysis. The digram analysis showed a significant swing in the opposite direction.

So it is seen that some different trends in the most likely confusions occur in the histogram and digram analyses. The F1/F2 ratio and changes in this ratio when speaker differences are included, are seen to be as predicted for all the digram analyses, but are largely against the prediction in the case of the histogram analyses. In all cases, Means and standard deviations of points plotted in figure 4.2

	F1		F2		X	
	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
Trivial confusions						
Histogram analysis						
N.S.S.	3	16	16	8	3	3
N.M.S.	15	10	16	6	22	18
D.S.S.	9	13	9	9	16	15
D.M.S.	20	9	23	17	13	11
Digram analysis						
N.S.S.	2	3	11	15	2	3
N.M.S.	4	3	19	5	13	11
D.S.S.	1	2	14	22	3	5
D.M.S.	41	14	13	2	33	30
Serious confusions						
Histogram analysis						
N.S.S.	11	17	1	2	2	2
N.M.S.	7	12	10	2	7	2
D.S.S.	10	2	17	12	19	17
D.M.S.	3	3	17	7	34	23
Digram analysis						
N.S.S.	2	3	1	2	1	2
N.M.S.	7	10	10	15	5	9
D.S.S.	0	0	12	11	3	5
D.M.S.	1	2	9	5	10	7

Table 4.2

except the histogram analysis of differentiated speech, the X confusions are also increased with speaker variations.

It has been found, by reference to the standard deviations of the original data in table 4.2, that only a small proportion of the trends shown in figure 4.2 are significant for all three speakers. The variation in the histogram confusions is seen to be greater than that in the digram confusions. The above facts tend to reduce the significance of the unexpected trends observed in the histogram recognition scores.

4.3.3.2. The analysis of serious confusions.

The higher F1/F2 confusion ratio, found for the serious confusions produced by the histogram analysis of normal speech, is seen to be insignificant when averaged over all speakers. The pattern of confusions produced by digram analysis of normal speech, does not show any significant differences between F1 and F2 confusions.

Both the digram and histogram analyses of differentiated speech produce significantly more F2 than F1 confusions. The decrease in the F2/F1 confusion ratio, when speaker differences are introduced, is not significant in either case..

4.3.3.3 Conclusions on analysis of confusions.

This analysis of the types of confusions has revealed that trivial confusions in the digram analyses always confirmed the prediction made in section 4.3.3. This result suggests that some simple transform of the two formant frequency description of vowels is present in the digram patterns, with a few exceptions which cause serious confusions which do not conform.

However, a large number of the confusions based on histogram analysis, including most of the serious ones, do not conform to the prediction, but are of low significance. In the absence of further detailed analysis based on a larger amount of data, this suggests that the histogram analysis does not produce patterns which are as simple transforms of the F1 - F2 description as the digram patterns.

4.4 Variation in recognition scores between two program runs.

The reference statistics for each utterance were compiled over ten consecutive segments of 50 msec. Subsequent recognition program runs used the same intervals plus those that followed these ten segments. The first program run was started at exactly the same interval as the compilation of the reference statistics; the second run was started 256 intervals later. This corresponds to two blocks of magnetic tape storage. This arrangement meant that the 50 msec. segments, within which the statistics were compiled, coincided during the first run with those used in the reference compilation, in fact with the first eight of the ten used. By the choice of a displacement in time equal to an arbitrary number of time intervals, the boundaries of the eight segments used in the second run were displaced from those used in the first run. In addition, use was made of the extra 100 msec. of the utterance, used in the reference compilation but not in the first recognition run.

An analysis of the difference between the recognition scores obtained on these two runs was done. The results were:-

	Mean Difference.	Standard Deviation.
Speaker. J.B.M.	3.46	3.2
Speaker. M.J.U.	4.13	3.2

If a normal distribution of these differences is assumed, 68% of the same utterance scores will be within 7 units of each other, and 95% within 10 units of each other. The unit is defined as a recognition score of 1 out of 96. Individual tests of these differences were made on the six different forms of statistics used in the following experiments, but no coherent difference was evident. There were, therefore, no grounds for expecting some statistics to give more reliable scores than others.

In the following experiments the scores are presented as the mean of the two runs, thus the above differences are equivalent to a tolerance of ± 3.5 (68% confidence) and ± 5.0 (95% confidence). These figures are presented as a guide to the interpretation of the results to follow.

4.5 A comparison of recognition results using histogram and digram analysis.

The discriminative powers of a digram analysis of the T.I.s of vowel waveforms were investigated qualitatively in chapter 2. The main concern of that study was to determine how stable the digram display was with the usual speech perturbations of changes in pitch and speaker, and also when pre-clipping differentiation was used. The digram was used simply because it was a sensitive and original method of T.I. analysis of speech waveforms. It was shown for an isolated case (section 2.3.1), that two vowel sounds could have very similar T.I. histograms but be clearly distinguished by their digram patterns. At that stage this could be shown only for isolated cases, and the overall advantage of analysis by the digram statistics, if any, could not be judged.

These experiments to be described now were designed to answer the question. "Is there any additional information in the digram statistics of the T.I.s of vowel waveforms that will improve the discrimination between vowels achieved using histogram statistics?"

Experimental procedure.

The speech of one speaker was taken at a time. Several sets of utterances of both speakers used, had been examined qualitatively: some had a controlled pitch, others were not so controlled. Those chosen for use in these quantitative experiments were three controlled pitch utterances and two uncontrolled pitch utterances for each vowel. The controlled pitch utterances were chosen to cover the pitch range of the speaker, and form the basis of a reference set of statistics. The uncontrolled pitch utterances were included to provide a measure of the difference between a single speaker's utterances at different times. The five utterances, so used, would be a test for a single speaker recognition system; to determine what range of utterances would have to be used in the compilation of reference statistics, in order to cope adequately with all vowel utterances of that speaker. The present experiments provide an estimate of the usefulness of three single pitch utterances in providing a vowel reference. The sets of utterances were labelled 1 - 5, where 1 - 3 were controlled pitch utterances with glottal periods of 8.0 msec., 6.5 msec., and 5 msec. for speaker J.B.M., and 9.5 msec., 8.0 msec., and 6.5 msec. for speaker M.J.U..

The result of a recognition run on twelve vowels was given as a score out of 96. This was the number of digrams compiled from the input set of vowels.

The form of presentation of the recognition scores, to be used throughout this chapter, is a chart with the utterance number (1 - 5) as abscissa and the score out of 96 as the ordinate. In all cases the scores for utterances 1 - 3 are linked by straight lines, and those of utterances 4 and 5 by a further straight line. These lines are not intended to convey an interpolation of recognition score between the utterances, but merely as a visual aid to distinguish the controlled pitch 'within-reference' sounds, and the uncontrolled pitch 'outside-reference' sounds.

Figure 4.3 A and B illustratesthe results for a histogram and digram analysis system respectively, for speaker J.B.M.. It is seen that sounds used as part of the reference statistics (within-reference sounds) are recognised more reliably than those which were not used as part of the reference statistics (outside-reference sounds). The scores obtained using digram statistics are seen to be near 100% for within-reference sounds, and around 30% for outside-reference sounds. This indicates that the analysis is far too sensitive to the peculiarities of particular utterances, but also that the method of combining statistics to obtain a mixed reference set has, in this case, yielded recognition results which are independent of pitch. The scores obtained using histogram statistics show some interesting differences. The within-reference

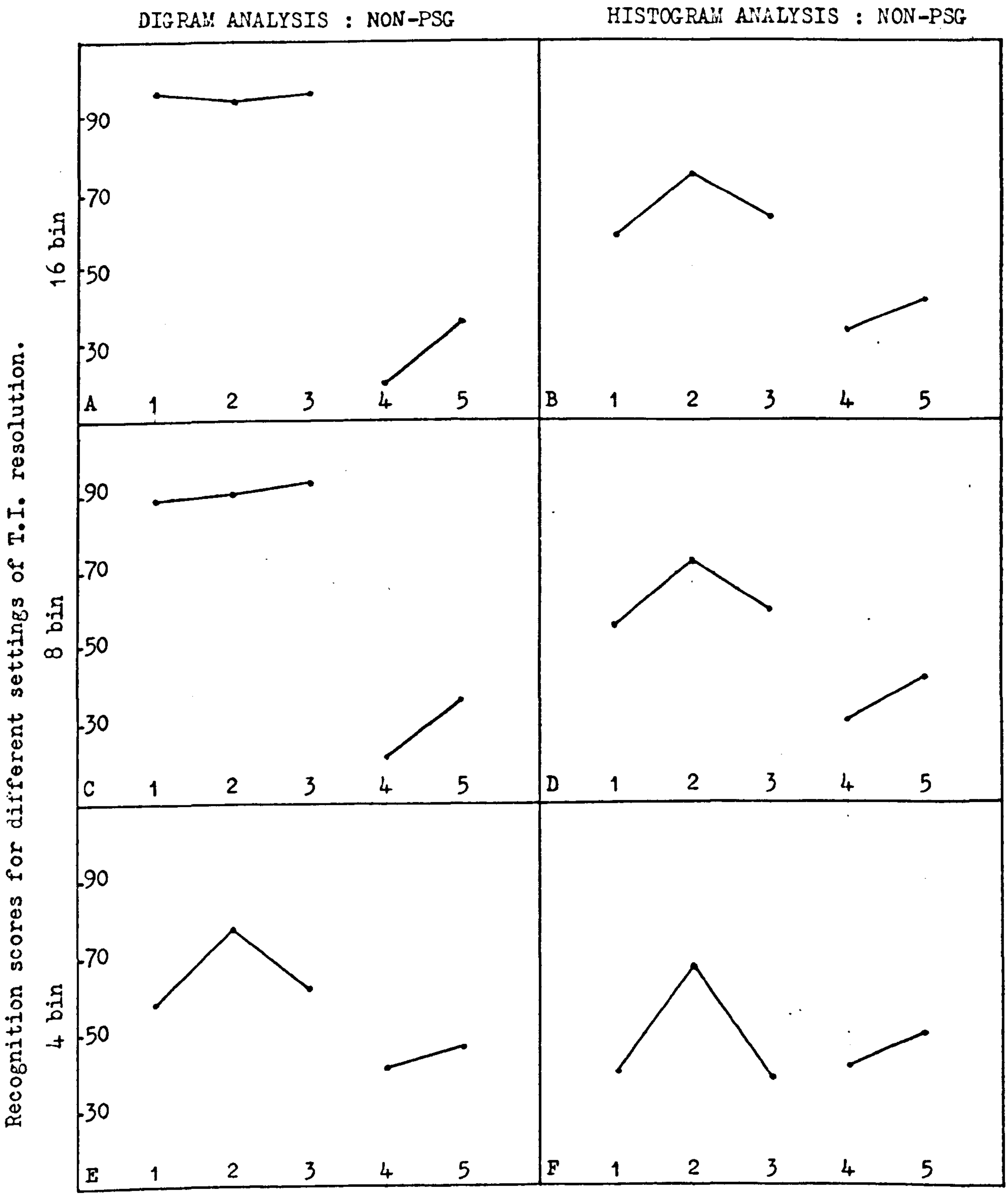


Fig. 4.3 Recognition scores obtained using digram and histogram analysis with three T.I. resolutions. (Speaker J.B.M.)

Recognition score significance limit :- I

sounds are also recognised more reliably than the outside-reference ones, but the difference between them is reduced: this reduction being mainly due to the poorer recognition of the within-reference sounds, which is seen to be very pitch dependent.

4.5.1 Comparison of results with those for another speaker.

Similar scores for histogram and digram analysis were found using a similar set of utterances spoken by M.J.U. (fig.4.4 A and B). There is a slight decrease in the scores for the digram analysis of within-reference sounds which is just significant (section 4.4.), and a similar increase for the outside-reference sounds, compared with the results of figure 4.3 A and B. The decrease in the within-reference scores could be attributed to the overall lower pitch. Within each 50 msec. segment there are fewer glottal periods, and therefore a greater proportion of this time is occupied with intervals which are under low amplitude lobes of the waveform in many of the vowels. Thus there may be a greater proportion of variability in the reference statistics.

There are greater differences between the speakers in the scores for the histogram analysis. The maximum and minimum scores for the pitch dependent within-reference sounds are similar for both speakers, but the mean level over the three utterances is 71 for M.J.U., and 65 for J.B.M. This is due to the fact, which is obvious from figure 4.4 B, that two of the M.J.U. utterances give the maximum score whereas only one did for the J.B.M. utterances. This result indicates that the pitch dependence of the within-utterance statistics is not due to a monotonic

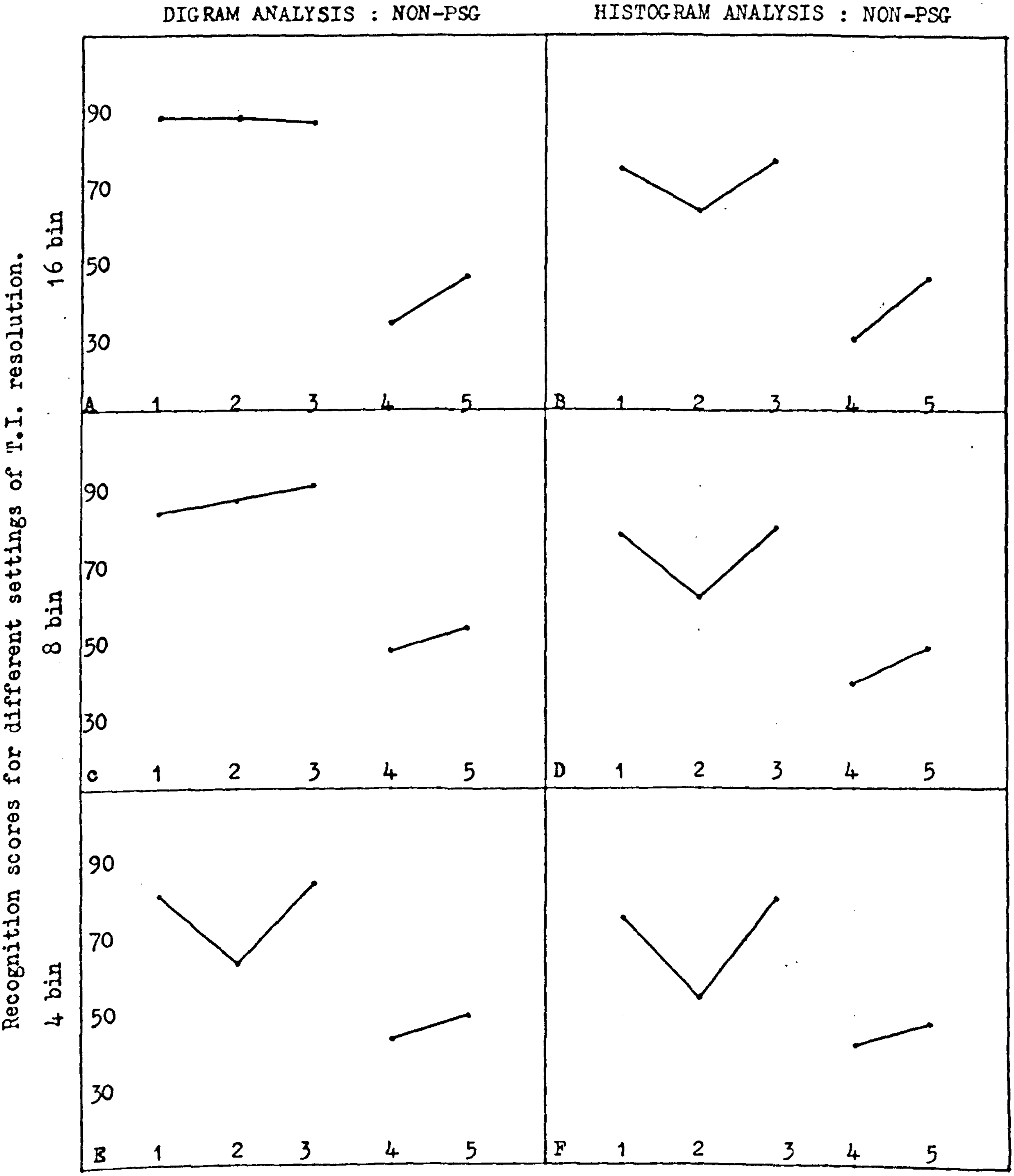


Fig. 4.4 Recognition scores obtained using digram and histogram analysis with three T.I. resolutions. (Speaker M.J.U.)

Recognition score significance limit :- I

change of the statistics with pitch, such that the mean pitch corresponds most closely to the mean statistic, but rather is due to the pecularities of particular glottal period lengths. This agrees well with the behaviour of the T.I. distributions illustrated in figure 1.6, where certain features appear, disappear and reappear as the glottal period is varied monotonically.

4.6 Further analysis by recognition experiments.

The large difference between the scores achieved using within-reference sounds and those achieved using outside-reference sounds, and the pitch dependence seen when using the histogram analysis, suggested two further variants of the analysis system. The first was to vary the T.I.resolution used in the classification of the intervals, and the second was the use of the already investigated technique of pitch synchronous gating.

4.6.1 The variation of T.I. resolution.

The reduction of the T.I. resolution was designed to eliminate the separation of utterances of the same vowel whose intervals were slightly different, and thus with a minimum of within-reference sounds, reliably recognise all utterances of that vowel by the same speaker. This action if taken too far would inevitably cause the merging of groups of vowels with similar statistics.

The method used was to divide the 16 bin equiprobable time scale into fewer bins. Reductions by factors of two and four were the simplest to implement. The results of this experiment are shown in parts C, D, E and F of figures 4.3 and 4.4. The difference between scores achieved using within-reference sounds and those achieved using outside-reference sounds is seen to have been reduced. The major component of this reduction has been the reduced score for the within-reference sounds.

For a given level of T.I. resolution, the digram analysis gives significantly higher, and less pitch dependent recognition scores than the histogram analysis, for the within-reference sounds. In the case of 4 bin resolution of M.J.U. utterances, not all the digram scores are significantly greater than the histogram scores. As the T.I. resolution is reduced there is a slight overall increase in the scores for the outside-reference sounds. In several cases this increase is only just significant over the full range of 16 bin to 4 bin T.I. resolution.

The differences in the recognition score between the analysis using digrams or histograms of the within-reference sounds, and the analysis using digrams or histograms of the outside-reference sounds, lead to the following conclusion. The greater detail of description possible in the higher information capacity digram is more specific to the particular sounds used as references, than the detail of description possible using the lower capacity histogram. The 4 bin resolution digrams, having 16 bins in all, and the 16 bin resolution histogram, have the same capacity and therefore provide a useful comparison. The results for both speakers show that both within- and outside-reference sounds have slightly higher scores when the digram is used, but for several of the utterances this difference is not significant.

The above results suggest that many of the features observed in these recognition scores can be correlated with the information capacity of the statistics irrespective of the T.I. resolution used or the order of the statistic compiled. There could be two variables of the speech sounds used which are responsible for this pattern of recognition scores. The first could be described as 'utterance peculiarities' and the second as 'pitch peculiarities'. It may well be that the former are a special form of the latter, but they are seen as distinct in the results of this experiment. The former variable is responsible for the difference in the score between the within- and the outside-reference sounds. As this difference decreases with decreasing T.I. resolution, the pitch dependence, caused by the second variable, increases in the scores of the within-reference sounds. This could not be seen in the outside-reference sounds as they are not of controlled pitch. These two variables have been separated, as they have been seen to act in opposite directions as the T.I. resolution is changed. It has been found possible to offer an explanation of only the pitch peculiarities of the utterances, and their effect on the recognition scores.

An explanation of the pattern of pitch dependence observed could be based on the same model of vowel waveform structure for which p.s.g. was postulated in chapter 2. If the effect of pitch changes is simply to perturb a rather noisy sequence of intervals at the end of the glottal period, the use of high capacity statistics could result in many bins having an odd one or two counts in them. In comparison, the low capacity statistics will on average have larger bin contents. Any shift of noisy intervals from one area of the statistic to another, will therefore cause a greater separation in the euclidean space, when a small number of bins are used to contain the contribution of these intervals to the total probability distribution of the statistic. This is based on the simple mathematical fact that $(a + b)^2 \geqslant a^2 + b^2$, as E.D. is calculated using a sum of squares.

4.6.2 Conclusion on the variation of T.I. resolution.

The desired result of reducing the difference between the scores obtained when within-reference and outside-reference sounds were used, was achieved. There were the undesirable side effects,which seem to be dependent on the capacity of the statistics,of reducing the within-reference scores with only a very slight increase in the outside-reference scores. The pitch dependence of the former was also increased.

The informational capacity of the various statistics could have been kept constant by adjusting the quantisation of the contents of the bins. An increase in this quantising unit would simply amount to the addition of noise to the bin contents and would not result in any improvement in recognition.

Quantisation of the dimensions of bin contents and of the T.I. scale, have different effects in this analysis. The former is of advantage only to the communication engineer, who requires the minimum information specification,compatible with adequate recognition. The latter defines the groups of intervals whose values are assumed to be independent of all others when E.D. measurements are made. It is seen therefore that the effects of varying the T.I. resolution could not be changed iná way advantageous to recognition by adjusting the quantisation of the bin contents.

This quantisation is sufficiently small to measure the absolute number of T.I.s involved in each statistic.

4.7 The quantitative investigation of pitch synchronous gating (p.s.g.).

The qualitative results obtained in chapter 2 suggested that there could be some advantage in the use of p.s.g. to eliminate pitch dependent perturbations of the T.I. distributions of vowels. Some digrams of the remnant of the glottal period after p.s.g., showed partial similarity to digrams of differently pitched utterances of the same vowel which had been similarly gated. The use of the recognition procedure on these digrams and the corresponding histograms, should help to decide whether p.s.g. is truly useful as an eliminator of noise and pitch effects. The mixed reference statistic will accentuate those parts of the digram which occur in other digrams of the same vowel.

Two types of p.s.g. will be described. Reasons were given in chapter 2 for the gating period to be a fixed time from the start of the glottal period. In addition to this basic form of gating, one of the other methods described was used. This method was to monitor the glottal period of the speech being analysed, and remove a fixed time segment from the end of the period. This was done by measuring each glottal period then assuming that the next period would be of similar length. The gating period was then adjusted accordingly to give an 'on' period of the glottal period minus one millisecond or the glottal period minus two milliseconds (PSG - 1 or PSG - 2). In the more orthodox method, when four or six milliseconds at the start of each glottal period were included, the gating is referred to as PSG + 4 and PSG + 6.

The PSG - 1 and PSG - 2 gatings were used on the speech of J.B.M.. It was thought that this form of gating might help to extract more relevant information from vowel waveforms which do not conform to the model on which p.s.g. was first based, i.e. a waveform without a significant amplitude decay in each glottal period. This type of waveform will occur mainly at the higher pitches when there is little time for any decay. For this reason,the method was used on the utterances of J.B.M. rather than on those of M.J.U. When very little decay occurs with a high pitch, it is typical that a large proportion of the period is pitch dependent. When very little decay occurs with a lower pitch, the T.I. information in some of the latter half of the period is quite reliable and may well be used. In this situation,if a fixed 'on' period is chosen, a smaller and smaller proportion of the duration of compilation is actually contributing to the statistics as the pitch is lowered. Therefore a fixed 'off' period could be an advantage.

The speech of M.J.U. used in these analyses was gated using only PSG + 4 and PSG + 6,as more of his vowels showed significant decay in the glottal period,owing to their lower pitch.

4.7.1 The results of p.s.g. on recognition scores.

Some preliminary experiments gave striking results. A typical set of recognition scores for the five utterances used, and the effect of PSG - 1 gating, is shown in figure 4.5. It can be seen that the effect of PSG - 1 gating has been to reduce the pitch dependence,and to increase

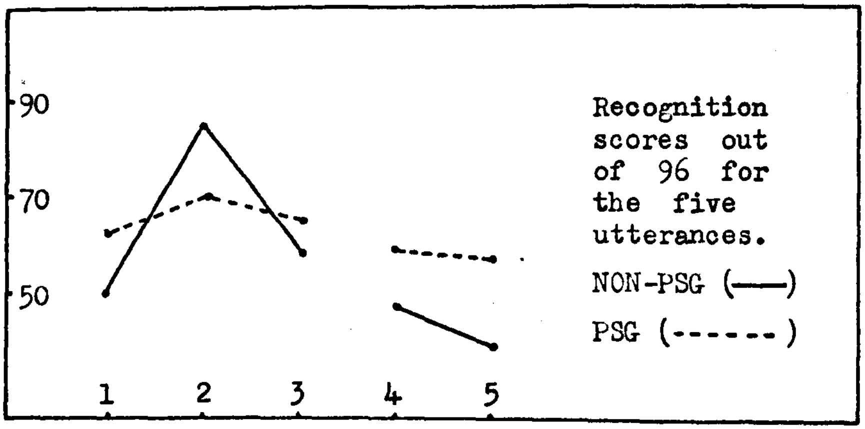


Figure 4.5.

the outside-reference recognition score. When a score fairly uniform with pitch was obtained for the gated within-reference sounds instead of a strongly pitch dependent result from the ungated sounds, there was usually little increase in the average value of the recognition score. Such results as these indicated that p.s.g. was a useful preprocessing to the compilation of T.I. statistics.

These results were obtained in preliminary experiments and are not typical of all the experiments later conducted, using different types and amounts of gating, on the speech of two speakers. Certain of their features are seen however, in many other cases. These features include the improvement of the recognition score of one but not both of the outside-reference sounds, and the smoothing of pitch dependence, for low but not high pitch and vice versa, for the within-reference sounds. The recognition scores for the p.s.gated speech of J.B.M. analysed by histogram and digram statistics are illustrated in figures 4.6 and 4.7 respectively.

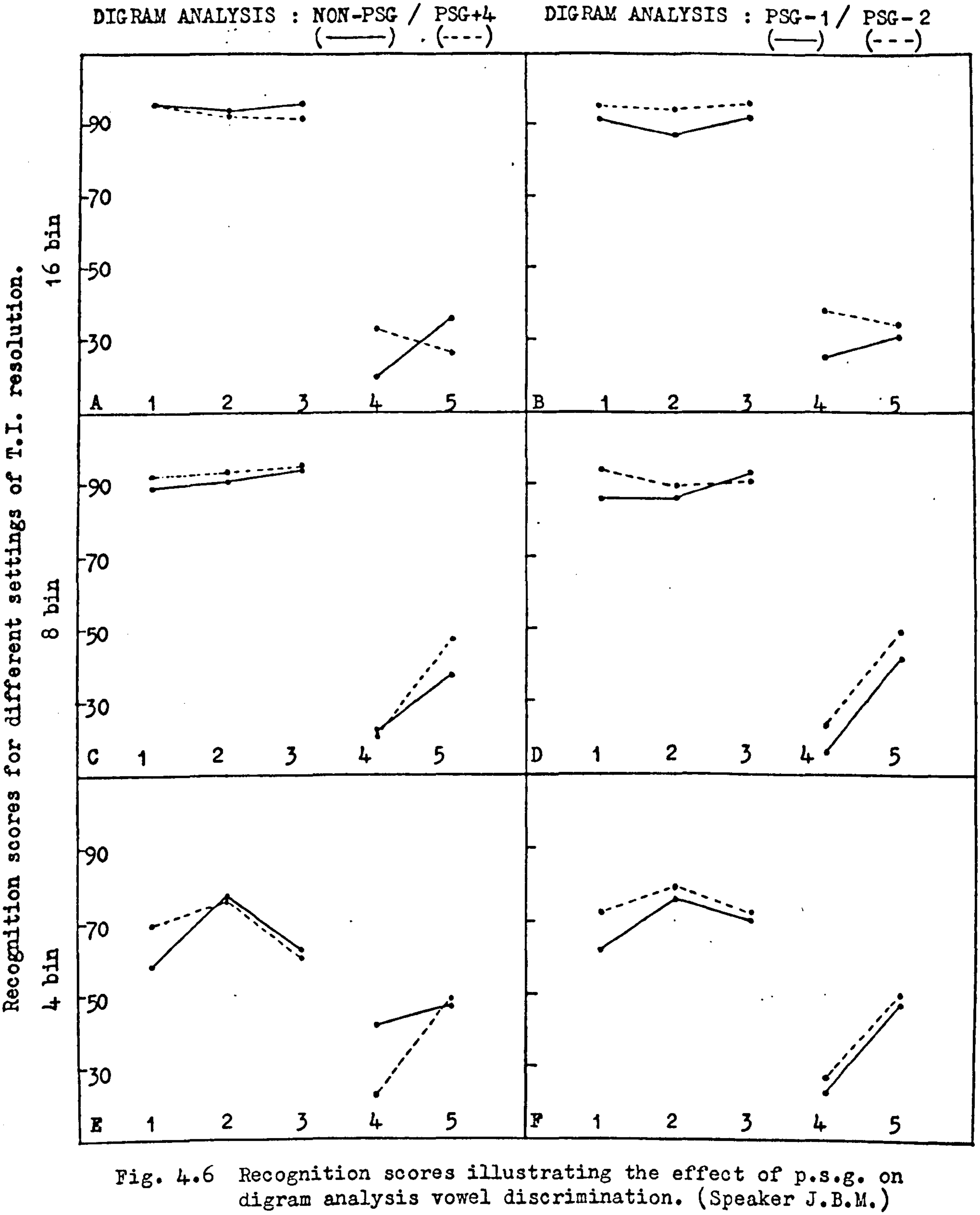


Fig. 4.6 Recognition scores illustrating the effect of p.s.g. on digram analysis vowel discrimination. (Speaker J.B.M.)

Recognition score significance limit :- I

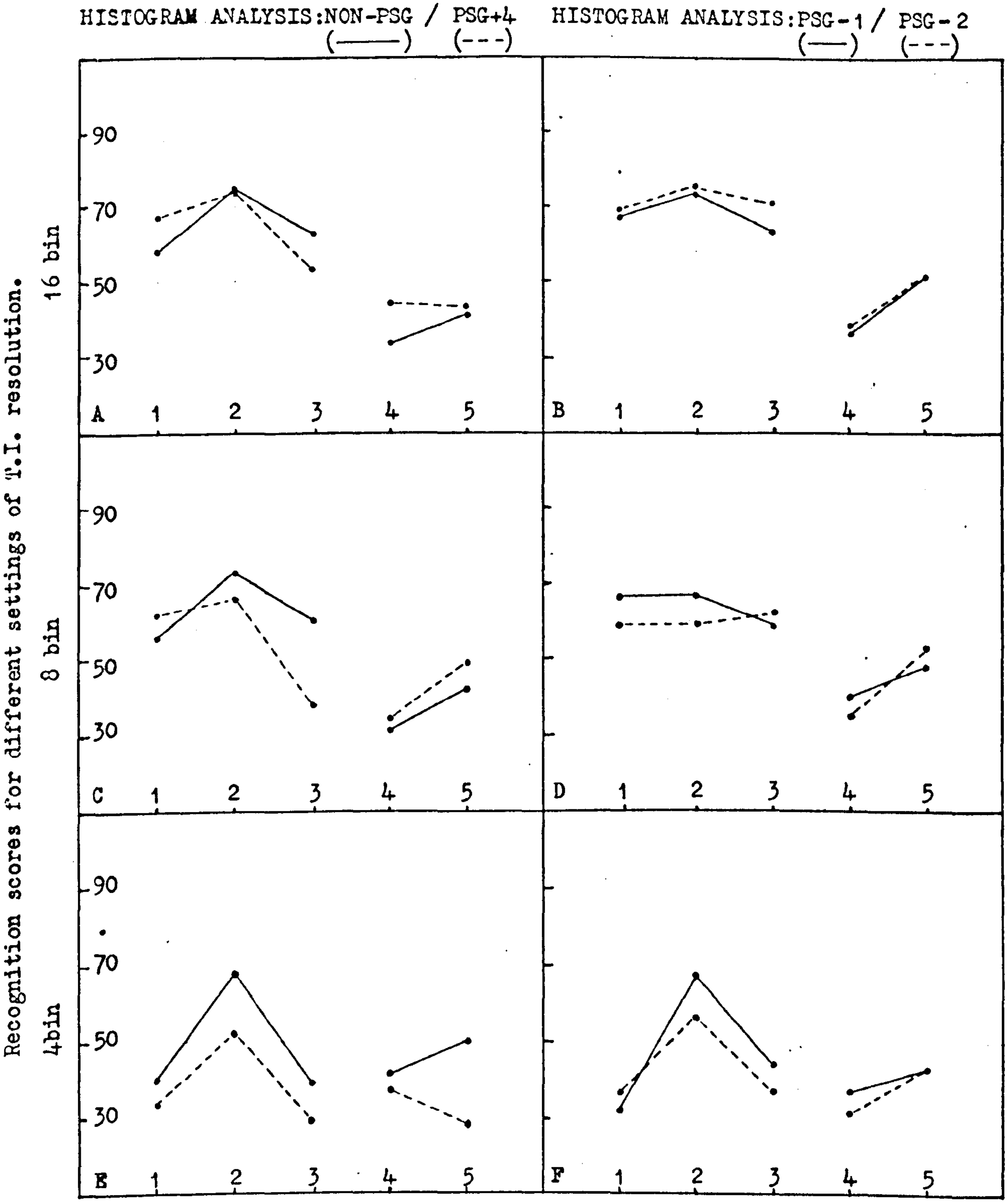


Fig. 4.7 Recognition scores illustrating the effect of p.s.g. on histogram analysis vowel discrimination. (Speaker J.B.M.)

Recognition score significance limit :- ⌶

All the scores recorded are the means of two runs of the program.

4.7.2 Pitch dependence.

Very little pitch dependence is seen in the results using the higher capacity 256 bin and 64 bin digram statistics. The steady increase in score seen in the ungated form of the 64 bin statistic, is just significant when related to the measurements of section 44. P.s.g. in this case gives largely insignificant differences in score to the ungated form. PSG + 4 and PSG -2 gating give a just significant increase in score for utterance 1.

A clear instance of reduced pitch dependence can be seen in the 16 bin digram. In all forms of p.s.g. there is no significant change in the score for the medium pitch, but for PSG -2 both high and low extremes of pitch show significant increase. PSG -1 gating also improves the score for the high pitched utterance (3), and PSG + 4 improves the score for the low pitched utterance (1). Further comment on this will be made in section 4.7.4. The pitch dependence of the recognition scores is more marked for histogram rather than for digram analyses. It becomes increasingly severe as the number of bins is reduced. Most of the forms of p.s.g. are less effective in removing the pitch dependence of these statistics.

4.7.3 The effect of p.s.g. on outside-reference sounds.

No clear picture of the differential effect of p.s.g. on the within-reference and outside-reference sounds was seen in the recognition scores. As the outside-reference sounds did not have a controlled pitch, it was not possible to predict that one form of gating should increase the score for either utterance 4 or 5. All that is clear is that, for each individual vowel utterance, PSG + 4 will retain a fixed amount in each glottal period, whereas PSG -1 and PSG -2 will retain a variable amount dependent on the pitch.

When 256 bin digram analysis was used, the recognition scores for utterance 4 were increased by all forms of p.s.g., whereas those for utterance 5 were decreased. The opposite was true for analysis by 64 bin and 16 bin digrams, with one insignificant exception. Scores for histogram analysis are increased by all forms of p.s.g. in the case of 16 bin and 8 bin histograms for both utterances 4 and 5, but are decreased by all forms of p.s.g. in the 4 bin case. In both histogram and digram analysis, the 4 bin resolution statistics gave very low recognition scores for all forms of p.s.g. on both the outside-reference sounds. Further comment will be made on this in section 4.7.6, after consideration of the utterances of another speaker.

4.7.4 Review of the effect of p.s.g. on the recognition of the utterances of J.B.M.

It was to be expected that the different forms of p.s.g. would affect differently pitched utterances in different ways. One case has already been mentioned in section 4.7.2. In this particular instance, the PSG + 4 gating gave a score insignificantly different from that obtained using PSG -2 gating of utterance 1. Both these scores are significantly better than the score in the ungated case. These gatings of the glottal period correspond to the inclusion of 50% (4 msec.) and 75% (6 msec.) of the period. A similar comparison can be made between PSG - 1 and PSG - 2 gating of utterance 3, when 80% (4 msec.) and 60% (3 msec.) of the glottal period respectively are included. The reason for PSG + 4 gating (80% of the period) giving poor results here, is not known. These two examples indicate that, for a significant increase in recognition score caused by p.s.g., the 'on' period of the gate, or the proportion of the glottal period that this represents, are not at all crucial.

There is evidence in figure 4.7 that PSG + 4 gating increases the recognition score of the low pitched utterances most, whereas the negative gating, usually PSG - 2, increases the scores of the high pitched utterances. As the capacity of the statistics is reduced, it is evident that the pattern of behaviour of the recognition scores after p.s.g. changes. PSG - 1 gives similar results to the ungated form, and the other forms of p.s.g. give scores which fall short, by a considerable margin in some cases.

A further relationship was expected between the scores obtained using PSG - 2 and PSG - 1 if the proportion of the period gated out was important. There is slight evidence of this in the 64 bin digram results. At the low pitch (utterance 1), PSG - 2 gating, which gated out 25% of the period, gave a similar score to that obtained using PSG - 1 gating on the high pitched utterance 3, which involved the gating out of 20% of the period. The gating out of 40% of this latter period, however, did not make a significant change in the recognition score.

It has been seen that all the various forms of p.s.g. have advantages in certain circumstances and that no one form can be said to be most effective in improving the recognition scores for utterances of J.B.M. There are several indications that a fixed 'on' period for the gating is not optimum, and that a fixed 'off' period of 2 msec. is more appropriate for the higher capacity statistics, being reduced to 1 msec. for the lower capacity statistics. However the differences in score were sometimes only just significant and no overall conclusion can be drawn.

4.7.5 The effect of p.s.g. on the recognition of utterances by M.J.U.

The major difference between the utterances of M.J.U. and those of J.B.M. was that the former were of lower pitch. Utterances 2 and 3 corresponded, in length of glottal period, to utterances 1 and 2 of J.B.M.. It could be predicted that p.s.g. would have a greater overall effect on these utterances than on those of J.B.M.. The following results show that this is in fact the case.

Both PSG + 4 and PSG + 6 gating increased the scores for 256 bin, 64 bin and 16 bin digram analyses (fig.4.8.) As the scores of the ungated waveform for 256 and 64 bin analyses were not as high as in the case of J.B.M., there was room for this just significant increase. The strong pitch dependence of the 16 bin digram scores is completely eliminated using PSG + 6, and partially so using PSG + 4. Note that in the case of utterance 1,with PSG + 4 gating, nearly 60% of each period is being rejected. The effect of p.s.g. on the pitch dependent scores obtained using histogram analysis is very clear and does not weaken as T.I. resolution is reduced (fig. 4.9). PSG + 6 gating was consistently better for utter-

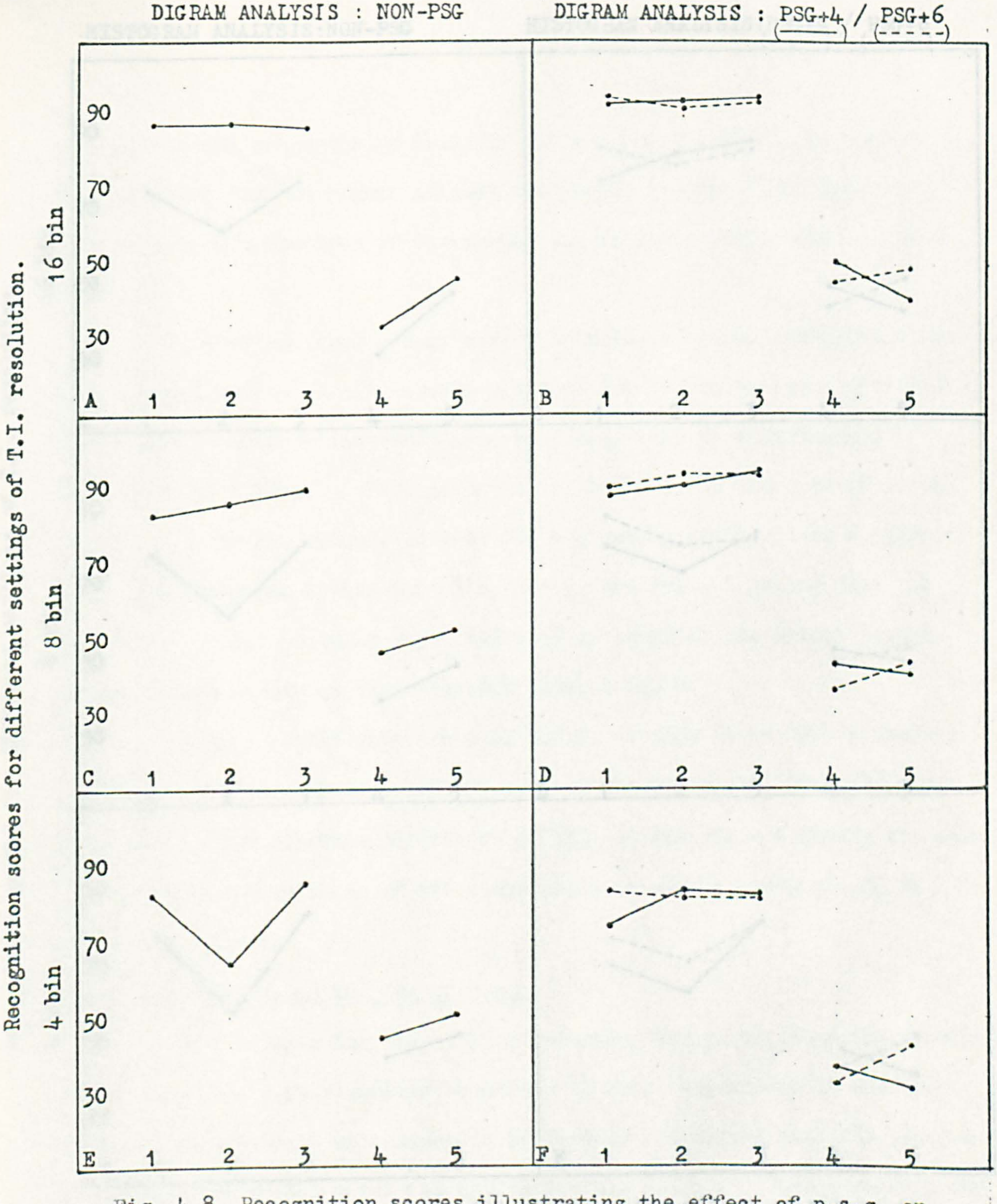


Fig. 4.8 Recognition scores illustrating the effect of p.s.g. on a digram analysis of vowel sounds. (Speaker M.J.U.)

Recognition scores significance limit :- I

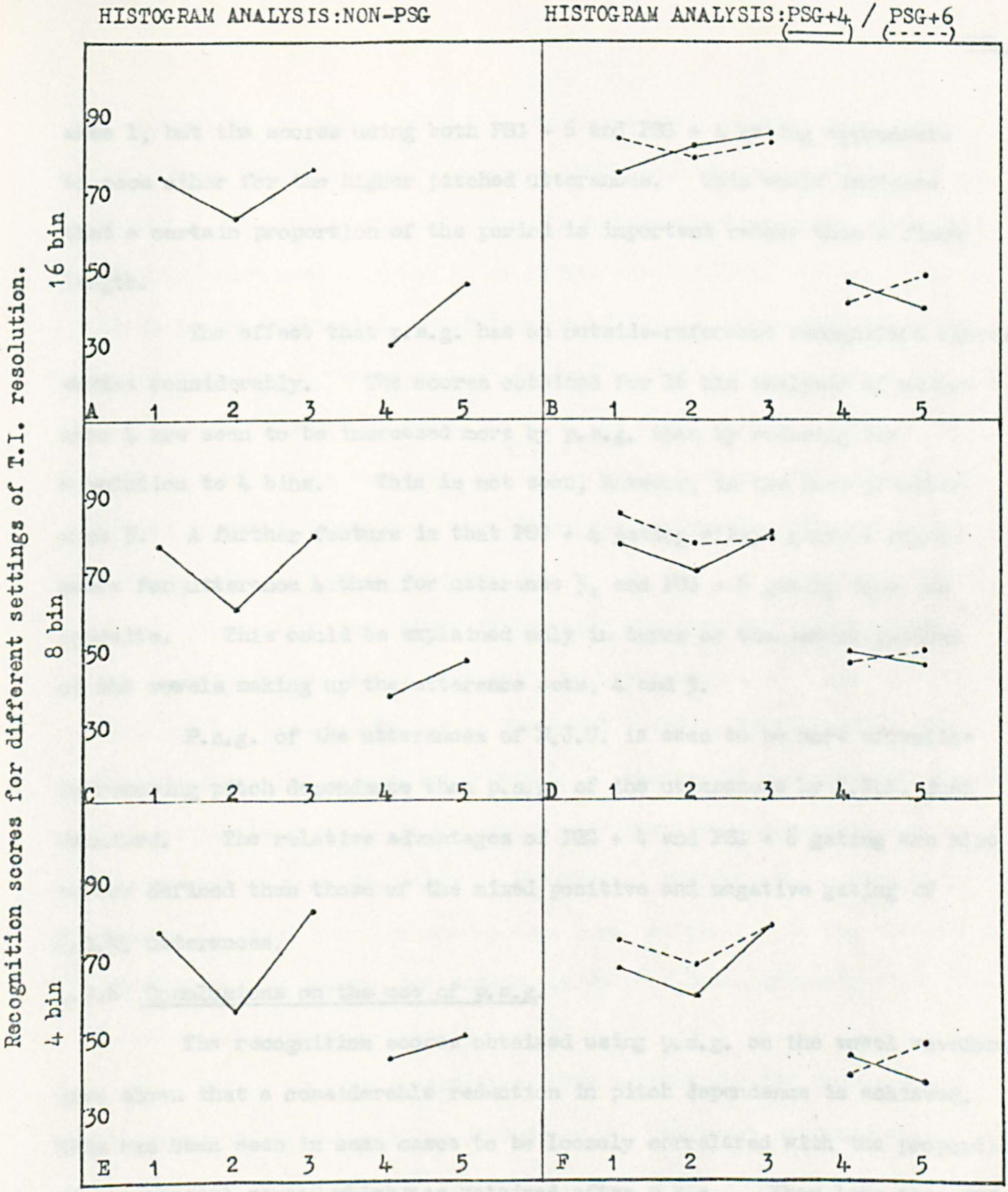


Fig. 4.9 Recognition scores illustrating the effect of p.s.g. on a histogram analysis of vowel sounds. (Speaker M.J.U.)

Recognition score significance limit :- I

ance 1, but the scores using both PSG + 6 and PSG + 4 gating approximate to each other for the higher pitched utterances. This would indicate that a certain proportion of the period is important rather than a fixed length.

The effect that p.s.g. has on outside-reference recognition scores varies considerably. The scores obtained for 16 bin analysis of utterance 4 are seen to be increased more by p.s.g. than by reducing the resolution to 4 bins. This is not seen, however, in the case of utterance 5. A further feature is that PSG + 4 gating always gives a higher score for utterance 4 than for utterance 5, and PSG + 6 gating does the opposite. This could be explained only in terms of the actual pitches of the vowels making up the utterance sets, 4 and 5.

P.s.g. of the utterances of M.J.U. is seen to be more effective in removing pitch dependence than p.s.g. of the utterances by J.B.M. just examined. The relative advantages of PSG + 4 and PSG + 6 gating are also better defined than those of the mixed positive and negative gating of J.B.M. utterances.

4.7.6 Conclusions on the use of p.s.g.

The recognition scores obtained using p.s.g. on the vowel waveforms have shown that a considerable reduction in pitch dependence is achieved. This has been seen in some cases to be loosely correlated with the proportion of the glottal period which was retained after p.s.g.. When less than 50% of the period has been retained, results no better than the score for the ungated waveform have been recorded. The results achieved when p.s.g.

was applied to the speech of M.J.U., who has a lower natural pitch than J.B.M., were found to be more coherent throughout the range of statistics examined. The application of p.s.g. to the outside-reference sounds has shown that when some increase in score has been achieved, it has been limited to the higher T.I. resolution statistics. This can be clearly seen in the case of utterance 4 in figures 4.8 and 4.9; the same tendency is evident in figures 4.6 and 4.7. On the basis of this evidence, it may be concluded that the increase in score for outside- reference sounds, achieved by reducing the T.I. resolution, is not further enhanced by p.s.g.

4.7.7 Recognition scores obtained.

Throughout this study it has been assumed that T.I. analysis alone is insufficient for a complete artificial recognition system. This assumption has been reinforced by the great variability observed in the T.I. statistics, especially that due to pitch. However, it has also been found in subjective experiments, when the clipped vowel sounds, used in the analysis experiments, have been presented to a number of subjects for recognition (appendix 5), that very low scores have been achieved. It was therefore thought useful to compare the scores achieved by human subjects with those achieved by statistical analysis.

Direct parallels cannot be drawn between all the scores, owing to differences in the recognition procedures. The human subjects were presented with a sequence of twelve different vowel sounds in random order. This set of twelve vowels by a single speaker was repeated three times during a 30 min. listening session. The average recognition score for these presentations was taken as representative of the recognition which is possible.

The artificial recognition of these sounds was the result of the comparison of the T.I. statistics of a set of utterances with a reference statistic, derived, either from the same utterances, or from similar utterances by the same speaker. It could therefore be postulated that the within-reference sounds recognition scores are roughly equivalent to a learned response of a subject who has studied the utterances concerned, and has established a relationship between the defining symbols of these vowel phonemes and the sounds themselves. This situation was not tested by the subjective experiments.

The outside-reference recognition scores are roughly equivalent to the response of a subject who has had the chance to 'tune-in' to the speaker briefly, but is then presented with new utterances which he has to recognise. This situation is much nearer to the subjective tests which were done.

It is therefore proposed to compare the outside-reference scores with the human recognition scores. The mean values of the outside-reference scores for speakers J.B.M. and M.J.U. are summarised in table 4.

	DIGRAM.	HISTOGRAM.
16 bin.	35%	38%
8 bin.	41%	42%
4 bin.	48%	48%

Table 4.

The subjective score of 47% for the same twelve isolated vowels indicates that similar scores can be achieved using both recognition systems, the human and the artificial.

4.8 Recognition based on the analysis of the T.I.s of the differentiated waveform.

The scores obtained by the analysis of the differentiated waveform are given in figure 4.10. The results for ungated waveforms are given by the points linked by solid lines and can be compared with those obtained for normal speech waveforms in figure 4.3.

4.8.1 Within-reference utterances.

When high capacity statistics are used in the analysis of within-reference utterances there is very little difference in the scores for both preprocessing conditions. When lower capacity statistics are used, the scores for the differentiated speech are lower than those for the normal speech. This difference becomes significant only in the case of 8 bin and 4 bin histograms.

4.8.2 Outside-reference utterances.

The increase in recognition score for the outside-reference utterances, observed for normal speech when 4 bin resolution was used, is not observed in the differentiated case.

4.8.3 Pitch synchronous gating.

The effect of PSG + 4 gating on differentiated waveforms was investigated. The results are given by the points joined by the dotted lines in figure 4.10. No significant increase in score is gained by this

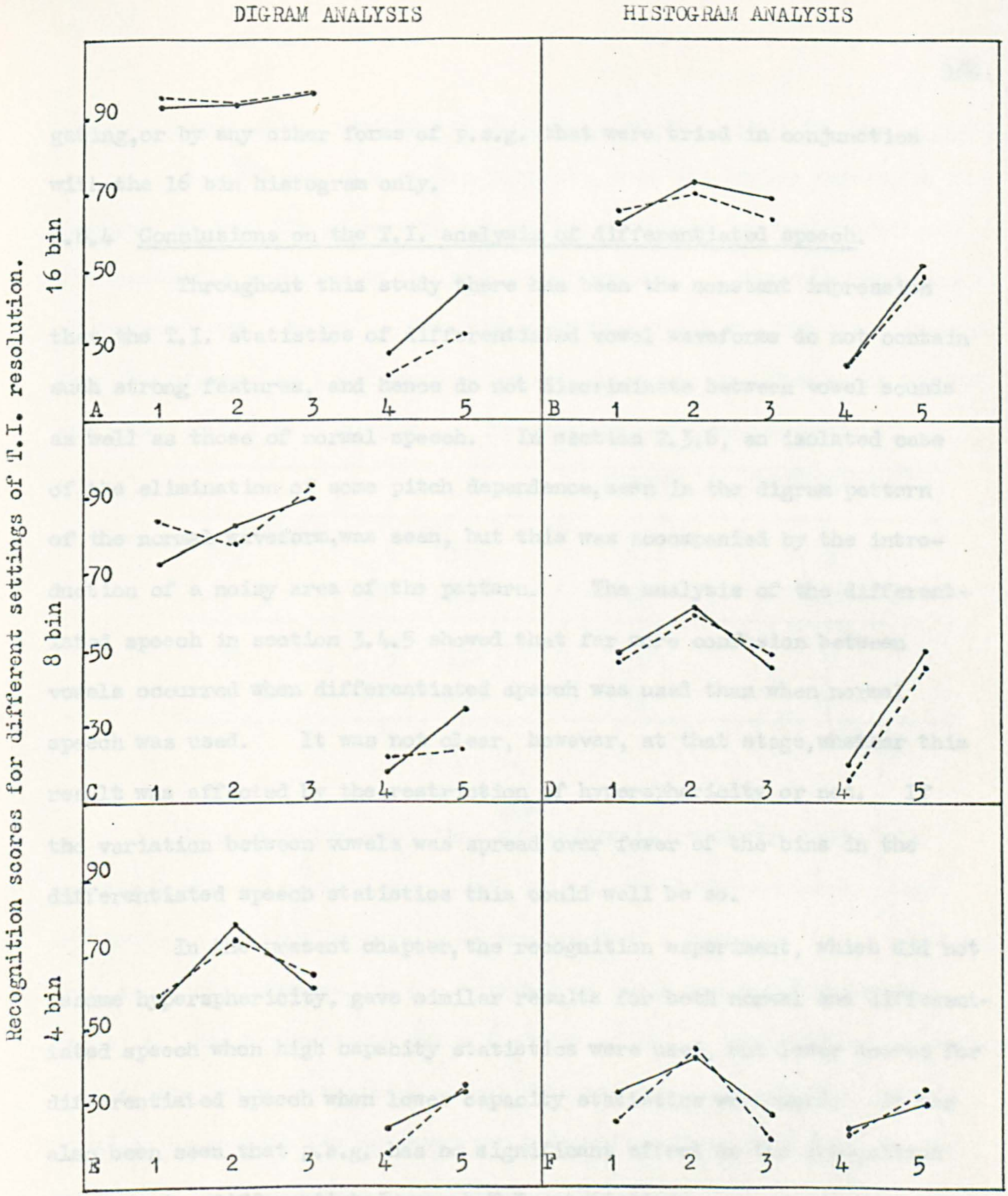


Fig. 4.10 Recognition scores obtained using digram and histogram analysis of differentiated vowel waveforms. (Speaker J.B.M.)

Recognition score significance limit :- I

gating,or by any other forms of p.s.g. that were tried in conjunction with the 16 bin histogram only.

4.8.4 Conclusions on the T.I. analysis of differentiated speech.

Throughout this study there has been the constant impression that the T.I. statistics of differentiated vowel waveforms do not contain such strong features, and hence do not discriminate between vowel sounds as well as those of normal speech. In section 2.3.6, an isolated case of the elimination of some pitch dependence,seen in the digram pattern of the normal waveform,was seen, but this was accompanied by the introduction of a noisy area of the pattern. The analysis of the differentiated speech in section 3.4.5 showed that far more confusion between vowels occurred when differentiated speech was used than when normal speech was used. It was not clear, however, at that stage,whether this result was affected by the restriction of hypersphericity or not. If the variation between vowels was spread over fewer of the bins in the differentiated speech statistics this could well be so.

In the present chapter,the recognition experiment, which did not assume hypersphericity, gave similar results for both normal and differentiated speech when high capacity statistics were used, but lower scores for differentiated speech when lower capacity statistics were used. It has also been seen that p.s.g. has no significant effect on the recognition scores using differentiated speech T.I. statistics.

4.8.4.1 Check on the variance of each bin's contents over the twelve vowels.

A check was made on the variance, over the twelve vowels, of the contents of each bin of the 256 bin digram statistic for both normal and differentiated versions of two sets of utterances. Two arbitrary levels of variance were chosen. The number of bins which had variances greater than the higher level and those which had variances in between the higher and lower levels were counted. The results are given in table 5.

		NORMAL.	DIFFERENTIATED.
UTT.	> HIGH	16	9
1	LOW - HIGH	44	20
UTT.	> HIGH	18	7
2.	LOW - HIGH	57	30

Table 5.

This showed clearly that a given level of variance was shown by a greater number of bins in the normal than in the differentiated speech statistic. The difference is of the order of a factor of two. Although both types of preprocessing are analysed on their own equiprobable bin scale, all the bin contents are not equivariable. This fact could explain some of the lower recognition scores obtained for differentiated speech. It might be possible to extract more information from the differentiated speech T.I. statistics,if weighting proportional to the observed variance in each bin was applied in the E.D. analysis. The results of listening tests on the sounds used in these recognition experiments showed,however, that the perception of clipped isolated vowels was affected very little by predifferentiation. This is contrasted with the results of Ainsworth using PB words (appendix 5).

4.8.4.2 Comparison of results with subjective recognition scores.

If the subjective recognition score is compared with the mean outside-reference scores, it is seen that the 53% recorded in the subjective experiment is consistently higher than any of the scores obtained (Table 6). It is noted that the scores shown in table 6 indicate that the differentiated waveform T.I. statistics give the optimum discrimination when a maximum number of bins are used. This is the opposite of the results for normal speech. As fewer bins contain the variations which separate the vowels, reducing the number of bins is likely to cause the cancellation of a greater proportion of this variation than in the case of normal speech.

	DIGRAM	HISTOGRAM
16 bin	37%	40%
8 bin	28%	36%
4 bin	31%	30%

Table 6.

4.9 The effect on the recognition scores of grouping the vowels.

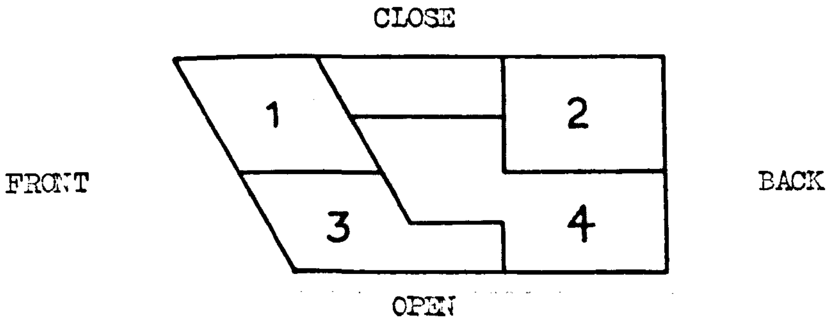
It was apparent from the qualitative studies on the T.I. histogram that the T.I. statistics of some vowels were more similar to one another than to those of the rest of the vowels. This suggests that certain groups of vowels might exist, within which, confusions based on

T.I. statistics might be quite common, but inter-group confusions might be quite rare.

A test was performed using the confusion matrix of the recognition experiment based on a 16 bin histogram analysis of ungated within-reference utterances of speaker J.B.M.. It was found that a minimum number of inter-group confusions were recorded if the following four groups were established.

1. /i/, /ɪ/.
2. /ʊ/, /u/.
3. /ɛ/, /ɑ/, /ɒ/, /æ/.
4. /ʌ/, /ɜ/, /ə/, /ɔ/.

This grouping corresponds to the following areas on a cardinal vowel diagram.



It can be seen that the proposed groups are continuous on this articulation chart.

4.9.1 The results of grouping the vowels.

A computer program was written to recalculate the recognition scores from the confusion matrices of each recognition test, ignoring confusions within these groups. The result of this further analysis is shown in figures 4.11, 4.12. The nature of this analysis meant that there could only be an improvement in the recognition scores, but any such improvement is at the cost of only discriminating four categories of vowel instead of twelve.

The improvement recorded for the within-reference sounds seems to reach a ceiling when 16 bin digrams are reached, descending the capacity scale. The improvement for the outside-reference sounds is fairly uniform. The results for the 4 bin histogram show that scores of between 70 and 90 were obtained for both within- and outside-reference sounds. The only value in increasing the capacity of the statistics was to increase the scores of the within-reference sounds. The effect of including some p.s.g. as well is negligible, as no form of p.s.g. for either speaker was found to give significant increase in the scores of the outside-reference sounds when low capacity statistics were used.

The increases in scores obtained, when grouping the vowels of J.B.M. after a recognition experiment on the T.I.s of their differentiated waveform (fig. 4.13), were in many cases less than those found in the case of normal waveform analysis (fig. 4.11). The increases of the scores for the outside-reference sounds were consistently smaller. The low scores of the 8 bin and 4 bin histogram analyses of differentiated within-reference sounds were not increased relative to the normal speech scores by the process of grouping.

4.9.2 Conclusion on the effect of grouping the vowels.

The grouping of the twelve vowels into four categories, compatible

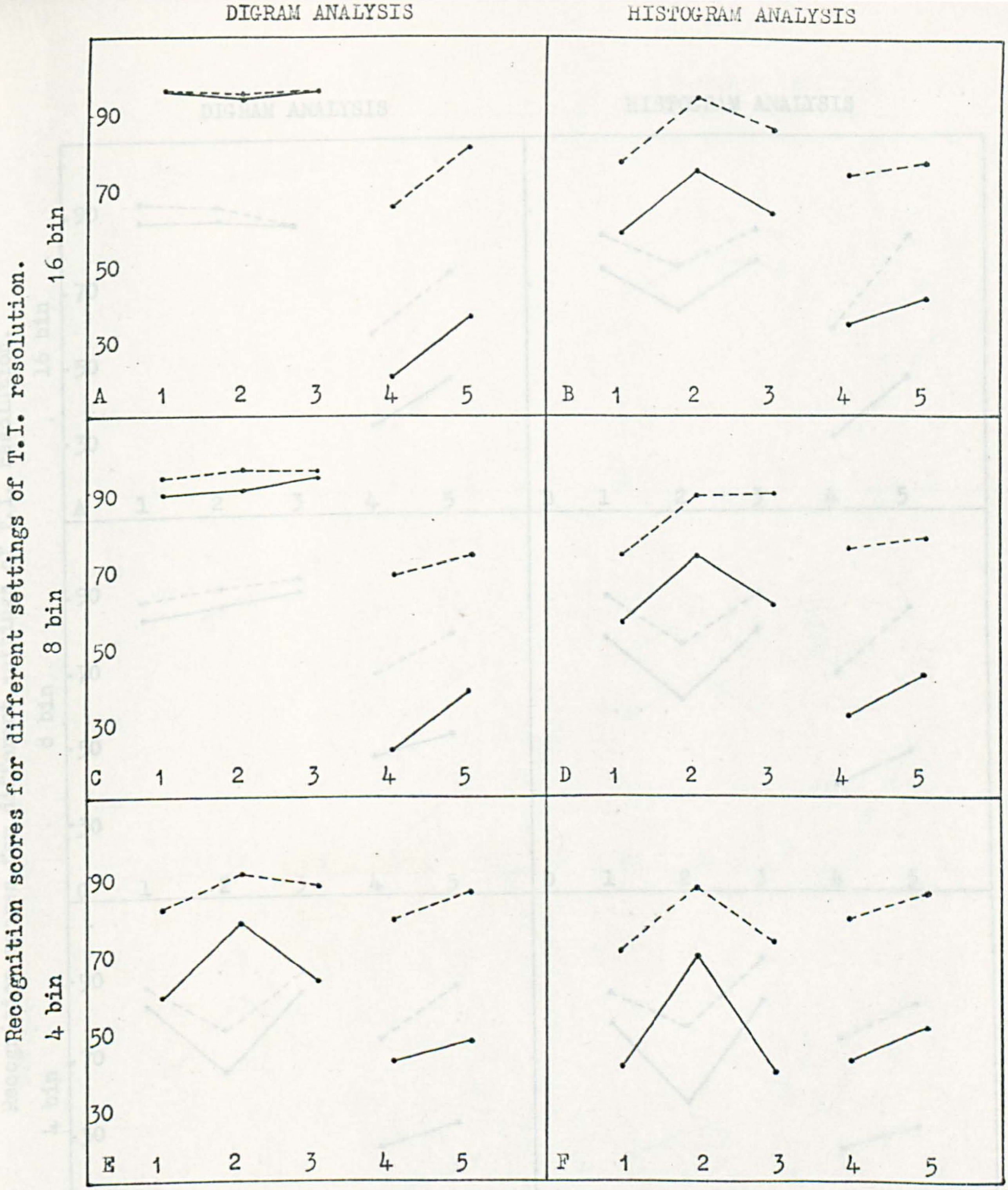


Fig. 4.11 Recognition scores illustrating the effect of grouping the twelve vowels into four groups on their discriminability using digram and histogram analysis. (Speaker J.B.M.)

Recognition score significance limit :- I

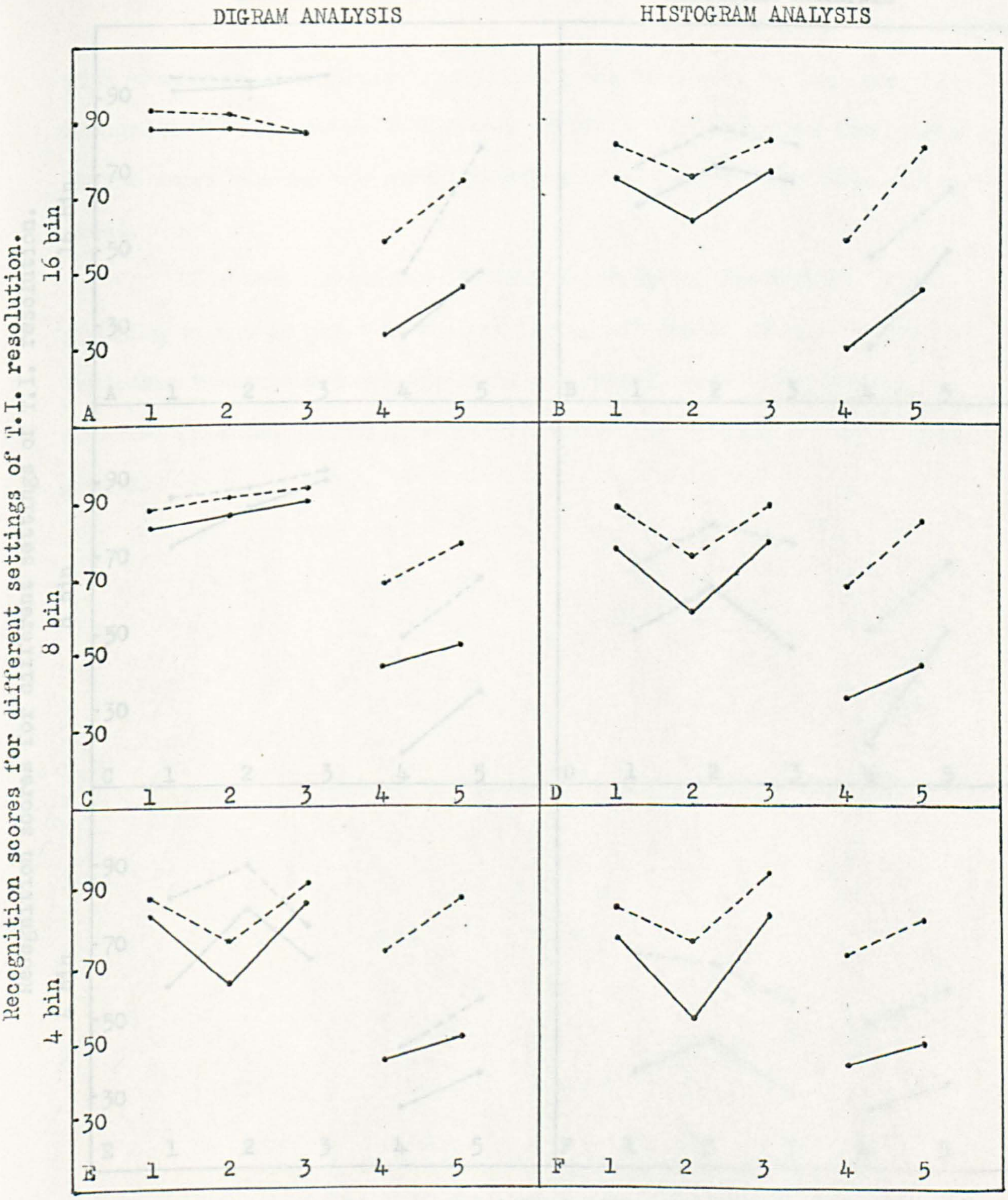


Fig. 4.12 Recognition scores illustrating the effect of grouping the twelve vowels into four groups on their discriminability using digram and histogram analysis. (Speaker M.J.U.)

Recognition score significance limit :- I

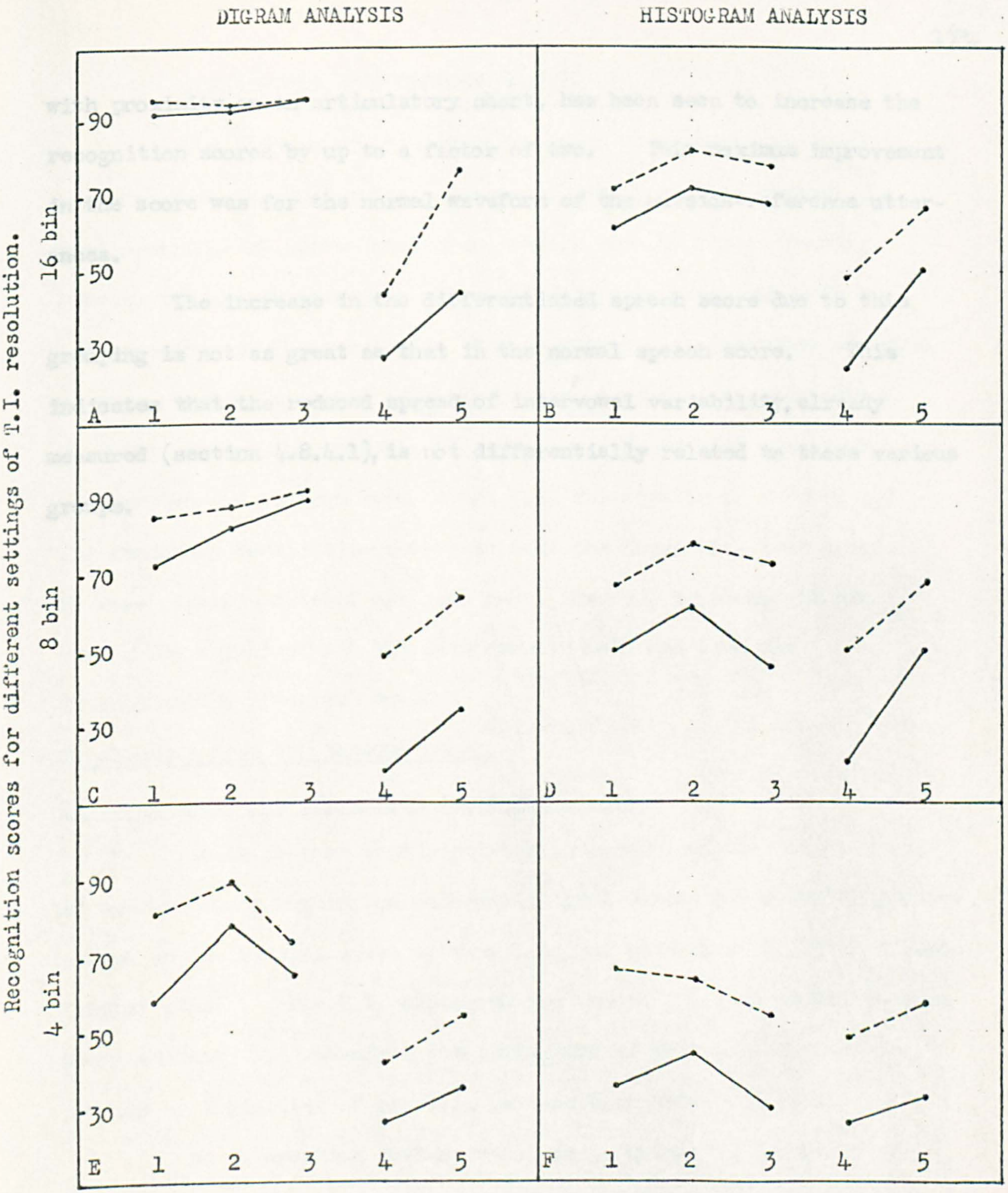


Fig. 4.13 Recognition scores illustrating the effect of grouping the twelve vowels into four groups on their discriminability using digram and histogram analysis of the differentiated waveform. (Speaker J.B.M.)

Recognition score significance limit :- ⌶

with proximity on an articulatory chart, has been seen to increase the recognition scores by up to a factor of two. This maximum improvement in the score was for the normal waveform of the outside-reference utterances.

The increase in the differentiated speech score due to this grouping is not as great as that in the normal speech score. This indicates that the reduced spread of intervowel variability, already measured (section 4.8.4.1), is not differentially related to these various groups.

CONCLUSIONS AND FUTURE PROSPECTS.

The purpose of this study has been to investigate various techniques for analysing the sequence of time intervals between the zero crossings of speech waveforms, with a view to discriminating between the speech sounds. These measurements have been made independently of their relation to the conventional frequency analysis of these sounds, by means of which a large amount of discrimination has been shown to be possible (11, 48). On several occasions similarities of pattern between the T.I. description of a sound and the frequency description have been seen, and these have been noted. No more rigorous attempt has been made, however, to associate the two. The object of all the measurements used has been the discrimination of speech sounds.

Conclusions from the present study.

T.I.s of speech waveforms and their histograms.

It is obvious that statistical measures of the T.I. patterns of speech cannot regain the information lost during the clipping process, except in the special cases of very long and very short T.I.s of a band-limited signal. The T.I. histogram represents T.I.s in isolation from their context, and therefore the histograms of only simple waveforms provide an indication of the T.I. pattern that they represent.

An interesting feature, which is important to the study of speech via its T.I. pattern, has emerged from the study of voiced waveforms (1.5.2). Not all frequency components of the complex waveform are present at the same time with the same relative amplitudes. Formant decay times vary, and their times of excitation relative to the glottal waveform are displaced with respect to one another. It is therefore possible for the T.I. pattern to be influenced by frequency components of low average amplitude,which would not affect the T.I.s if all components were present with constant relative amplitudes.

The additional frequency information available in the time domain did not result in T.I. histograms which could be used to distinguish the twelve vowel sounds examined. The result pointed rather to the fact that only broad classifications could be made on the basis of the T.I. histogram (1.5.2).

The only speech sounds,in which there was found a clearer relationship between the T.I. histogram and the frequency components, were the fricatives and voiced fricatives (1.5.3/4). In both these groups of sounds the general frequency description is simple. The only merit of the T.I. histogram analysis is its simplicity compared with frequency analysis. This latter point has already been made, concerning T.I. analysis in general, by Chang et al.(8).

These comments about the T.I. histogram are expressed from the point of view of relating it to other representations of speech which are known to characterise various sounds. The complexity of many of the histogram patterns made it necessary to use a quantitative analysis to estimate the discriminative power of this simple time domain measure.

The quantitative analysis for the twelve vowel sounds, as spoken by two male speakers gave the following generalised results. In most cases the scores obtained for within-reference sounds were highest for the maximum T.I. resolution of 16 bins, and for outside-reference sounds the 4 bin resolution often gave the maximum score. The spread of scores for the differently pitched utterances of the within-reference sounds was between 60% and 80%, this spread being reduced to 70% to 80% for the optimum p.s.g.. The scores for the outside-reference sounds were 40% to 50%. When the vowels were grouped into four categories, discrimination between these categories was increased to give within-reference scores of 78% - 94%, and outside-reference scores of 78% - 85%.

Comparison of the quantitative analysis of T.I. histograms with that of Bezdel and Chandler.

Some comparable figures from the work of Bezdel and Chandler (5) are 97% for male speakers within the reference group and 87% for male speakers outside the reference group. There are however several differences between their experiments and those of the present study. The major difference, in terms of its effect on the recognition scores obtained, is that their five vowel categories were each represented by only a single vowel, although spoken by different speakers. The nearest equivalent scores of the present study are for the four vowel categories, but in this case each category was represented by a wide phonemic range within the category. Vowels on the peripheries of these categories would be expected to be more readily confused with those in other groups than if each group were represented by a single vowel. Although the vowel categories of Bezdel and Chandler were not the same as those in this study, this fact could explain some of the difference in the scores. A further difference is that the present experiments were done for the utterances of a single speaker, and only three sets of utterances of that speaker were used to compile the reference statistics. A more comprehensive reference set would be expected to cause an increase in the scores of the outside-reference sounds, but would, at the same time, reduce those for the within-reference sounds. It has been seen that the within-reference sounds in the present study represent a rather artificial situation; that of a learned response to a particular set of utterances (4.7.7).

The use of many speakers to define the reference statistic is a further difference. If the small differences between the overall T.I. distributions of vowel utterances for three different speakers (section 3.6.2) are compared to the level of T.I. resolution which has been found optimum for the recognition of outside-reference sounds, it can be seen that the additional generality of reference, obtained by including more preselected adult male speakers, may outweigh the variability introduced.

The uses of T.I. histogram analysis.

The T.I. histogram is a simple measure but it is not likely to provide a basis for any perceptual clues. This can be deduced from the fact mentioned above, that it represents T.I.s in isolation. Sequences of given T.I.s in any order will give a similar histogram, but as Underwood (66) has shown, they may show little perceptual similarity. Although virtually nothing can be claimed for this form of analysis on the basis of perceptual experiments, the broad discriminations described above suggest that it could have some uses in speech analysis. It could be used in a limited vocabulary analysis whose constituent phonemes happen to coincide with the distinguishable T.I. patterns on any suitable T.I. scale. This has been demonstrated in a particular case by Bezdel and Chandler (5). In a more general purpose analysis, T.I. histograms could play a partial role either as a preliminary feed-forward device, as described by Mackay (45), to control the appropriate analytic action, or as a refinement of a crude frequency analysis. An example of the former use could be the separation of voiced fricative from vowel. A Z.C. rate measure could separate most of the other classes of sounds, but a crude histogram would be necessary to separate these two classes on the basis of T.I. measurement.

The digram measure of speech waveform T.I.s.

The discriminatory power of this measure was expected to be greater than that of the histogram of the T.I.s for two reasons. Firstly, it has a slightly longer term significance than the histogram, although still very much shorter than the time scale of the smallest perceptual units of speech. Secondly, the clear perceptual difference between a sequence of short intervals followed by a sequence of longer ones, and a mixture of both in one sequence, is reflected by differences in the digram pattern but not in the histogram. The digram is thus capable of containing more perceptually relevant information than the histogram. This has been seen explicitly in isolated cases. It is also more sensitive to any change in the T.I.s than is the histogram. Such a change causes different relationships with the preceding and following T.I. For example, the insertion of two short intervals between two previously adjacent long ones would cause a completely different digram pattern. The corresponding histogram patternswould be slightly different, but would overlap if the long intervals did not suffer a significant change in length.

The voiced sounds were seen to have more structured digrams than the fricative sounds (2.3.1/2/3). Digrams of voiced fricatives simply showed that in most cases the long and short intervals were not separate in the glottal period but mingled together in some way. Digrams of nasals emphasised the difference in spread of the intervals. Such a measure as the moment of the display about X = Y would be a useful discriminant of the individual nasal sounds analysed. The digram patterns of vowels gave a general impression of some significant structure, but when visual discrimination of the various digram structures was attempted, discrimination of the twelve vowels into only three or four categories was possible (Section 2.3.2).

The variability of the T.I. statistics, when either the pitch of the voice or the speaker was changed, was investigated using the digram representation. Digrams of utterances by two male speakers were seen to be very different in some cases (2.3.5), but some of the differences were explained on the basis of the pitch difference involved. The control of pitch was seen to eliminate some of these differences in both the first and second order T.I. distributions (2.5).

The effect of differentiation before clipping was also investigated. The expected greater emphasis on higher frequencies caused similarity between digrams to be seen in pairs and groups of vowels which are different to those in the normal speech case (2.5.1). No general elimination of inter-speaker differences in the T.I.statistics was achieved by differentiation before clipping.

The variation of the duration of compilation of the digram statistics showed that the major features of both first and second order distributions remain constant from one glottal period to the next, assuming that period remains constant (2.3.4).

Preliminary quantitative measures relating digram and histogram analysis showed that the least variable statistical representation of speech wave T.I.s was the digram of normal speech (4.3.1). However, when speaker differences were included in the analysis, the digram was seen to be more sensitive to these differences than the histogram (4.3.2). It was also noted that the statistics of the differentiated waveform were also more sensitive to these differences. An analysis of the confusions involved in these experiments showed, amid very variable results, that the digram analysis is a simpler transform of the two formant analysis of vowels than is the histogram analysis (4.3.3.1).

Quantitative measures of the difference between the vowel digram patterns of several utterances of single speakers, showed that they did provide more discrimination than the simple histogram for the within-reference sounds, but not for the outside-reference sounds. Scores of 90% - 100% were achieved for the former, but only 40% - 50% for the latter (4.6). Analysing the twelve vowels in the four groupings caused the latter score to be increased to 70% - 80%, the former remaining the same (4.9.1). The lack of any increase in the scores for the outside-reference sounds when using digram analysis indicates that although there is structure present in the T.I. pattern of the waveform, it is no more invariant with respect to the vowel than the T.I.s themselves.

It was noted that the scores obtained for the within-reference sounds were mainly dependent on the capacity of the statistics, rather than on the order of the statistics or the T.I. resolution used. In particular, the scores for 16 bin histogram and 4 x 4 bin digram analyses were very similar. It was seen however, that the outside-reference scores were higher for the 4 x 4 bin digram analysis. As this score was most similar to that for the 4 bin histogram analysis, it was assumed that the level of T.I. resolution is important for the determination of these sounds, rather than the capacity or order of statistics.

The effect of differentiation on the T.I. statistics and their variability, first investigated in the case of the digram (2.3.6), was further studied for both histogram and digram statistics using the quantitative measures. The intermediate unquantified result of section 3.4.5, where the T.I. statistics of differentiated vowels were seen to give very poor discrimination, was largely due to the experimental procedure. The results of section 4.8 showed that for high capacity statistics, scores nearly equivalent to those obtained for normal speech could be obtained. The scores for the low capacity statistics were, however, significantly below those for normal speech. Further, the scores for outside-reference sounds were not increased by reduced T.I. resolution.

A factor, which could explain these low scores for differentiated speech, was found by examining the variance of the contents of each bin of the statistic over the twelve vowels. It was found, in the case of the statistics examined, that the normal speech statistics have twice as many bins with a given variance as differentiated speech statistics. Further analysis incorporating an equal-variance bin distribution along the T.I. scale was not performed.

Evaluation of the use of pitch synchronous gating.

It was seen early in this study that the effects of pitch on T.I. statistics are very great. Two reasons for this have been given:

1. End effects in the glottal period, due both to the varying length of the period, and the typically low signal to noise ratio in most low pitch vowel sounds.
2. Varying relationship between the amplitudes of the formants due to changes in their harmonic structure as the pitch varies.

Perturbations classed under heading '1' have been investigated, and largely removed, by the technique of pitch synchronous gating (p.s.g.). Those classed under heading '2' have not been investigated in this study. They could be studied by performing a T.I. statistical analysis using p.s.g. on broad band filtered speech waveforms. The relative amplitude factor may then be avoided by analysing only one formant region at a time. This approaches closely to the method chosen by Sakai and Doshita (56) although this reason is not given for their choice. Such a method would also be a closer simulation of the human ear-brain system than either frequency or time domain methods.

The use of p.s.g. alone to eliminate some of the pitch dependent features of T.I. statistics has been successful. This technique has removed both noisy areas and particularly pitch dependent spots of the digram display of normal (2.7.2.) and differentiated waveforms (2.7.3.).

The stability of digrams of p.s.gated waveforms,when the pitch or speaker was changed (2.7.4/5.), was found to be little improved over the ungated case. It is assumed that perturbations under heading '2' above, could be a major cause of the remaining instability with pitch.

Quantitative measures of the effect of p.s.g. on vowel discriminability were made on the speech of two speakers (4.7.4/5). Its effect was seen more clearly in the recognition of utterances of the speaker with the lower pitched voice. No definite conclusions on a fixed rule to give the optimum severity of gating within each glottal period were possible from the results. There was slight evidence for all three proposed rules at various points in the results. It was found however, that in general,p.s.g. reduced pitch dependence, although in a large proportion of cases there was no significant increase in the overall score. The optimum scores for the ungated outside-reference scores, those for the low T.I. resolution analysis, were not improved by p.s.g..

The application of p.s.g. to the analysis of these sounds using higher T.I. resolution, occasionally increased the scores more than reducing the resolution.

A further application of p.s.g., which has not been within the scope of this work, is in the extraction of formant frequencies from bandlimited speech waveforms. The method of measuring Z.C. rate gives the nearest harmonic of the fundamental, whereas the mean value of the T.I.s of a p.s.gated waveform could give a more accurate value. An application of pitch synchronous analysis for this purpose has been investigated by Scarr (57), with successful results. (See the following section on Parallel work).

Comparison of artificial and human recognition of isolated vowels.

When the recognition scores for the outside-reference sounds were compared to the scores achieved by human subjects listening to the clipped speech waveform, they were seen to be nearly equivalent (4.7.7). It must be emphasised that the artificial recognition system, in this case, has the advantage. It is analysing a steady state signal which is unnatural to the human listener. It has been found, over several listening tests, that the spread of absolute scores for natural, clipped, and differentiated and clipped vowels is extremely large, varying between 20% - 60%, but their relative scores are fairly stable. It was noted that the utterances giving the highest scores had a more variable pitch within each utterance than those which gave the lower scores. It is possible that the human listeners were gaining cues from pitch variation. The process of elimination of pitch dependent features in the time domain,used in the artificial analysis,may be analagous to the human process of establishing what is the true vowel quality when the variability due to pitch can also be extracted. The pitch synchronous analysis carried out in this study did not, however, cause such a large difference in the recognition score.

It can therefore be concluded that the artificial system using an extremely simple T.I. statistical analysis can perform as accurately as the human system on completely steady state vowel sounds, but the latter is capable of far greater accuracy if variation of such parameters as pitch is included. Presumably the variation would cause a reduction in the artificial recognition scores.

Parallel work.

Work relevant to the topic of this thesis that has been done prior to this study has been reviewed in the introduction. Some relevant work which has been carried out in parallel with this study, both in conjunction with it and independently of it, will be reviewed here.

Work in conjunction with this study has been done by Underwood (66) on the synthesis of speech from a statistical description of the time intervals between zero crossings. One of his major findings has been that the T.I. description of speech is on too microscopic a level to be suitable as a unit for the synthesis of speech. The time scale of linguistic variations is of the order of tens and hundreds of milliseconds, whereas the T.I. values extend from tens of microseconds to a few milliseconds. Synthesis of isolated vowel sounds was attempted, using first, second and third order statistics. Each of these statistics in turn had a greater temporal span than the previous one, and showed some improvement in the intelligibility of the sounds produced.

If synthesis of recognisable speech from T.I. statistics had proved possible at the first or second order stage, then there would seem no reason why artificial recognition should not be achieved on the basis of histogram or digram statistics. As this was not the case, however, the extraction of discriminative features from the T.I. statistics must be viewed with more caution.

Parallel, but independent work has been done on the speech waveform by Stover (61) and Reddy (54), and on T.I. analysis by Bezdel (6), Scarr (57) and Lavington (34).

Stover has arrived at the same conclusions as the author concerning the redundancy in the speech wave caused by pitch, and has found evidence that only 3 msecs. at the start of the glottal period is necessary for perception. He also uses the redundancy of repetitive glottal periods which is parallel to the use of histograms of very short duration of compilation. His aim is that of economical speech transmission rather than analysis, but the preservation of intelligibility by these time domain distortions is relevant to both.

Reddy uses the speech waveform as the starting point for an analysis system which includes measurements of Z.C. rate, intensity, and pitch synchronous fourier series expansion of the waveform. The only overlap with the present work is the use of pitch synchronous analysis and pitch detection using a peak detector. He does not use a T.I. statistical analysis. All his analysis is done within a computer.

Bezdel (6) has developed a system based on his earlier T.I. histogram analysis (5). He has made modifications to reduce the variability of the T.I. histogram with various speakers, to make possible more accurate recognition. The only preprocessing employed is the addition of h.f. bias to reject noise. Adaptive processing is used at the level of compilation of the histogram. This processing includes the movement of the histogram bin divisions, and the variable weighting of the bins, which are very powerful techniques. The differences in the rate of arrival of T.I.s in these variable bins are made the basis of speech pattern detection. These difference parameters are also adaptive to provide a common reference pattern for the presence of every sound, specified by a network linking certain histogram bins.

This development of the use of T.I.s for speech recognition is complementary to that described in this study. It relies on a great deal of adaptive control whereas the present study has sought to answer some of the problems of T.I. variability in a static way. The strategy of the present study does not eliminate the use of adaptive controls, but would reduce the complexity of such controls by building general rules concerning T.I. variability into the preprocessing.

Scarr has found that a good measure of the first formant of a vowel sound can be extracted from a T.I. measurement; that of the second interval in each glottal period of the prefiltered waveform. His prefiltering eliminates all but the first formant frequency range for all vowels. This finding is compatible with the general observations of this study concerning the stability of intervals in various parts of the glottal period. It is this interval that was chosen to be the first interval measured during pitch synchronous analysis of the unfiltered waveform. Lavington has analysed a limited vocabulary of words by means of several software routines in an all computer analysis system. He has used such parameters as the Z.C. rate of both the waveform and its derivative, and the difference between these two. A very crude form of T.I. histogram using three bins, fundamental frequency detected by means of autocorrelation and amplitude peak analysis have all been included in the time domain measurements. Spectral analysis has also been done using two different routines. High recognition scores have been achieved using these parameters.

The Way Forward.

The present work has concentrated on extracting the maximum amount of information about speech sounds from their T.I. structure alone. As such it has led to the conclusion that the use of these measures is restricted, owing to their great variability and their short-term significance in the speech signal. The attempt to extract as much information as possible from the T.I. statistics has however provided some useful pointers. Some of the complicating factors limiting time-domain measurements have been found and techniques of measurement have been developed, which could usefully be incorporated in any further experiments using T.I. analysis. The great variability of the T.I. measures used, has indicated that they can only be applied in restricted tasks such as the preliminary subdivision of speech sounds, or for making fine distinctions after a crude frequency analysis. They could also be used in small vocabulary speech recognition systems where the constraints of the sequential relations between phonemes contain a large amount of additional information, or where the total number of phonemes is small.

A line of study which is likely to extend this work in the direction of speech recognition is the combination of the information obtained through individual channels or features of the T.I. statistics with other one dimensional measurements. The aim of such a study would be to observe acoustic events in simply extracted parameters of speech. These parameters could be short, or longer term amplitude measurements, the rate of change of energy in a certain frequency band, frequency change detectors, and T.I. measurements of a less detailed type than those made in this study. The ultimate aim of such a study would be to obtain correlates of these measures with articulatory movement. It is thought that the combination of simple feature extractors of this type, is far nearer to a simulation of the likely human analysis of speech waves, than a detailed measurement in any one dimension, such as the study just described.

Appendix I.

Clipping Amplifier circuit.

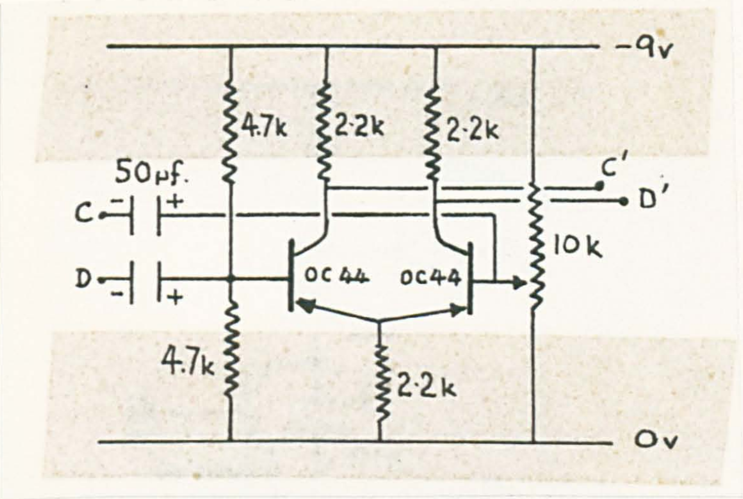


Fig. A 1.1 A single stage of the clipping amplifier.

The clipping amplifier comprises four long-tailed pair amplifier stages (fig. A 1.1). They were connected in differential form with a.c. coupling, of 50 μf. into two 4.7 K parallel resistors, giving a time constant of approximately 100 msecs.

The earlier work on histograms was done using OC 71 transistors in this circuit. The OC 44 transistor gave a better gain per stage (10 - 20). The non-linear amplification was made symmetrical about the zero signal level by adjusting the 10 K potentiometer in the base lead of the second transistor of each stage.

Appendix 2.

A. Parallel time-base system.

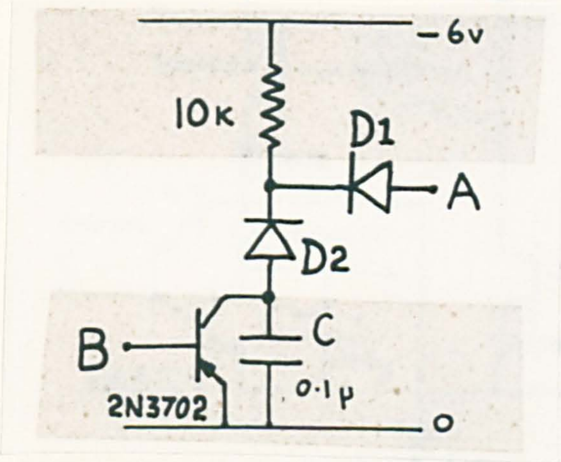


Fig. A 2.1 Single time-base and hold circuit.

Operation.

As two time bases were being used, there was no need for a separate hold circuit: the time base capacitor itself was used to store the voltage. The holding properties were achieved by maintaining a positive level on A, which caused D 2 to be reverse biased and provide high impedance to further charging of the capacitor C. A silicon transistor was used as the discharge path for C, as its low leakage current helped maintain the charge in C until it was discharged by a negative pulse at B.

B. Analogue shift register stage.

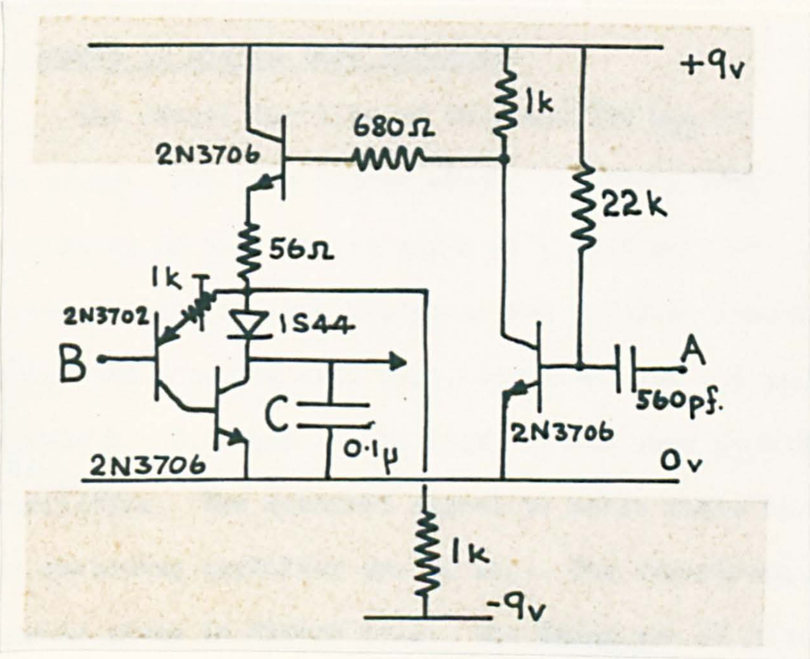


Fig. A 2.2 A single analogue shift register stage.

A positive pulse at A, causes the voltage on B to be transferred accurately on to capacitor C. The variable 1 K potentiometer was used as a fine control to compensate for the difference in base-emitter voltage drops in the two transistors following input B. When connected as a multistage shift register, the voltage on C provides the input (B) to a further stage.

Appendix 3.

Magnetic recording machines.

A3.1 Siemens 12 stereo tape recorder.

The stereo facility of this machine was not used in the present study. The four tracks simply allowed greater economy of tape. The facilities of the machine which were used were the microphone and electrical inputs, and the headphone and external speaker outputs. The microphone was a moving coil type, supplied with the machine. When used within 1 - 2 inches of the lips, it had very good signal to noise characteristics. The measured signal to noise ratio for this microphone and the recording amplifier was 52 db.. The record/replay frequency response is given in figure A3.1. The major use of this machine was, however, when the record amplifier was used as a buffer amplifier for the high output impedance Language Master. The response of the record amplifier was flat over the vowel frequency range, for which this combination was used.(fig. A3.1).

A3.2 Bell and Howell Language Master.

This machine, designed primarily for language instruction, was used in the majority of the experiments. It operates by recording and replaying on a short strip of magnetic tape which is affixed to a piece of card. This method of storing acoustic records of sounds made the retrieval of required sounds, and repetitious play-back of them, much less time consuming and arduous than using conventional spooled tape. There was also the great advantage of being able to shuffle the ordering of sounds for use in listening tests.

The frequency response of this system, for electrical recording, is shown in figure A3.1. The response for microphone recording was found to be substantially similar. This response is seen to be only just adequate for the vowel sounds. The machine was certainly not capable of storing clipped speech signals for use in listening tests. In these cases the clipping was performed after reproduction, or the clipped signal was recorded on the Siemens recorder.

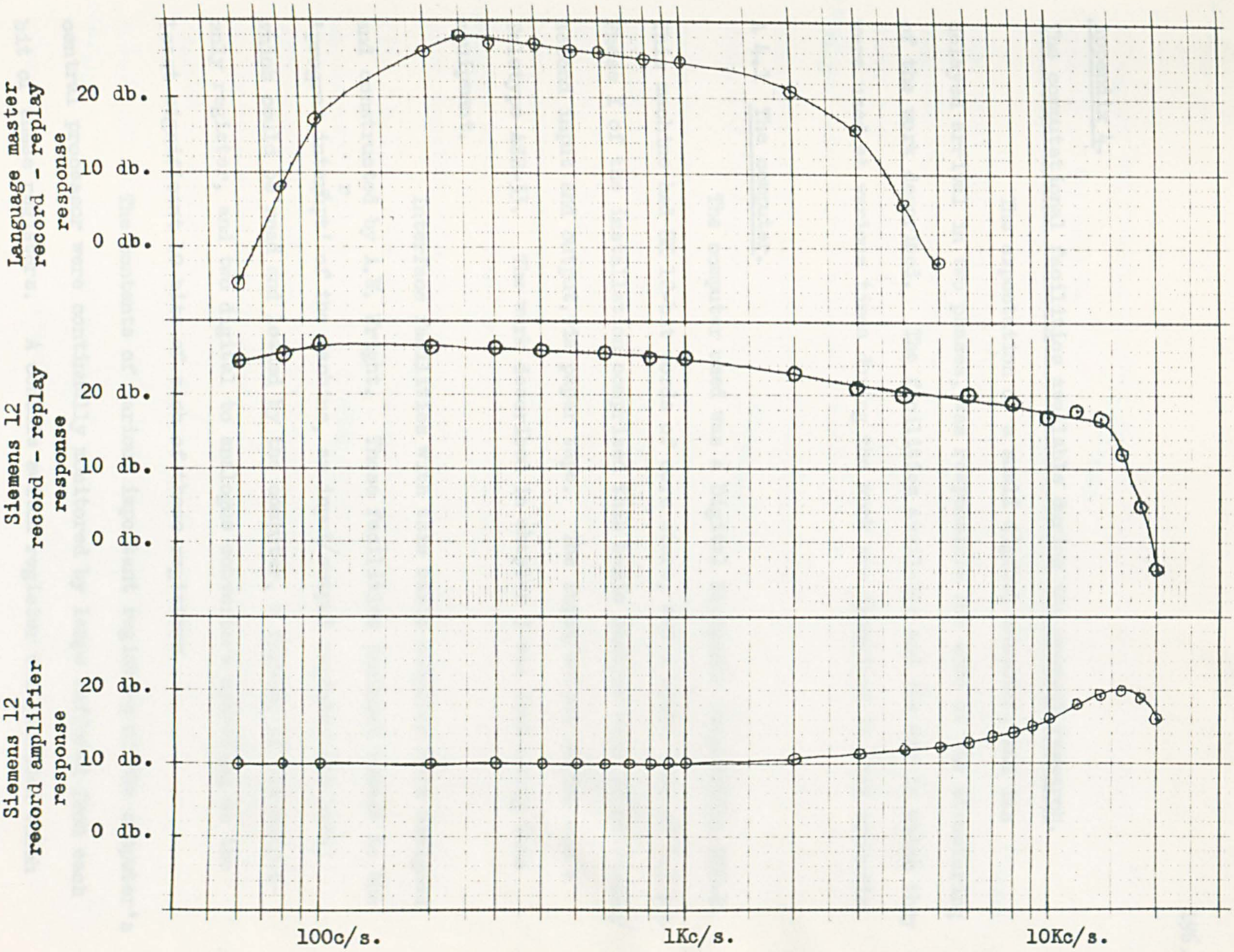


Fig. A3.1 Frequency response curves for magnetic recording machines.

Appendix 4.

The computational facilities available during the present research.

The expectation of a small digital computer, and its delayed arrival in two phases, was responsible for some of the structuring of the work described. The facilities available and the way in which they were used at various times during the work are described in this appendix.

A 4.1 The computer.

The computer used was a Digital Equipment Corporation PDP-8. This machine had 8K 12-bit words of core store, and a cycle time of 1.5 µs. Phase I of the installation comprised this basic machine with 10 character/second input and output, via paper tape. The input/output device was a Teletype ASR-33. The work described in chapter 3 was done using this equipment.

Interface facilities with this basic computer were designed and constructed by A.W. Wright. These facilities included access to the 'program interrupt' of the machine, an input/output register (12 bits) which could be read and loaded by the computer, a further 12 bit output-only register, and two digital to analogue converters operating on the least significant 10 bits of each of these registers.

The contents of various important registers of the computer's central processor were continually monitored by lamps buffered from each bit of these registers. A console switch register was provided which could be read by the computer. This could be used not only for the debugging and starting of programs, but also as a control during the running of the program.

Phase 2 of the installation comprised high speed paper tape input (300 characters/second) and output (50 characters/second), magnetic tape storage, and a 338 buffered display. Two magnetic tape decks were controlled by a single TC∅1 control unit, enabling the use of two tapes within a single program, but not at the same instant in time. This part of the second phase of equipment was most important to the work described in chapter 4, which would not have been possible using the basic machine's storage alone. The 338 buffered display was used to provide a visual check on the statistics being manipulated within the machine. A program was written to display the contents of any part of the magnetic tape storage of digram statistics on the 9 x 9 inch C.R.T. display area. A whole set of 12 vowel digrams could be displayed at once, together with axes, using the layout of the cardinal vowel diagram (fig. A 4.1). This not only provided a very quick check on the statistics on the tape, but produced a display which revealed features of adjacent vowel digrams which had not been previously noticed. This arrangement of the vowels was then replicated for the presentation of the vowel digrams in chapter 2.

A 4.2 The production of visual displays using the computer and a C.R.T.

The displays of T.I. statistics presented in chapters 1 and 2 were, in the main, produced under computer control. The photographs which had been taken from the screen of the C.A.T. and from a C.R.T.,

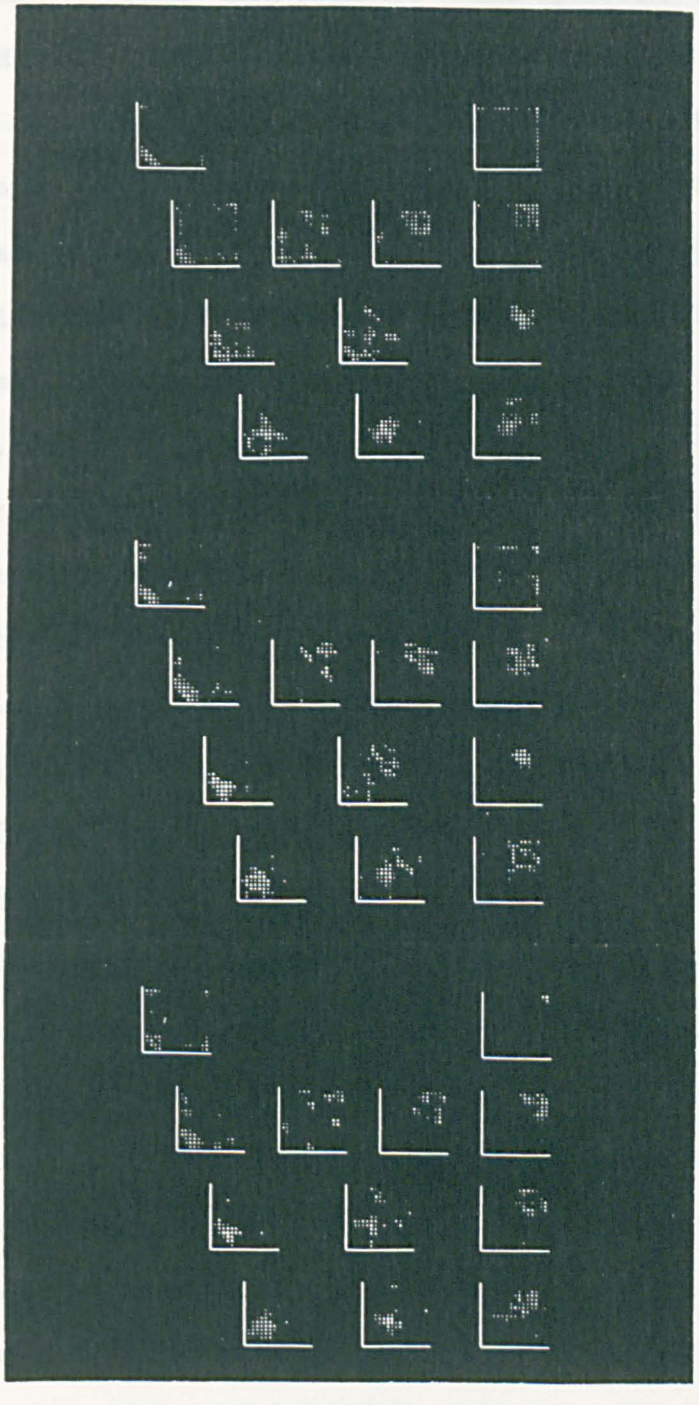


Fig. A4.1 Three examples of the display of 256 bin digrams on the 338 display.

driven via an analogue shift register, most of which had never been printed, were not easy to 'read' when printed. The operation of the C.A.T. and the A.S.R. were simulated within the PDP-8 solely to achieve better photographic records of the qualitative experiments. The only feature of the analogue digram display that was not simulated was the exponential time base. This was most useful when viewing unvoiced fricatives and vowel sounds together,but when studied in isolation an appropriate linear scale was adequate.

The T.I. histogram photographs were made more readable by representing the contents of each bin by a vertical line of length proportional to the contents, rather than the single spot produced by the C.A.T. display. The photographs were produced on the 338 buffered display.

The A.S.R. was simulated using a program written by M.J. Underwood for the synthesis of speech from statistics stored in the computer (66). T.I.s stored in the computer were loaded into the two registers (mentioned in section A 4.1) in sequential pairs. Voltages corresponding to the least significant 10 bits of these registers were derived by the D/A converters and applied to the X and Y plates of a C.R.T.. The most significant bit of one of these registers was used to control the Z modulation of the C.R.T.. This latter facility was included in order to perform pitch synchronous gating on the statistics. The pitch markers, which had been stored with the T.I.s, were used to control the setting of this most significant bit in order to provide the amount of p.s.g. specified by the setting of the computer console switch register. The details of operation of this program are given by Underwood (66).

Modifications to this program were made in order to control the 'duration of compilation' of the statistics from the stored intervals, and to obtain the intervals from magnetic tape rather than via a T.I. measurement program from a train of Z.C. pulses.

After each presentation of the digram statistics a subroutine was entered which plotted axes for the display. This facility saved considerable effort compared to the photography using the analogue system, which required complicated switching of signals for the presentation of axes.

Appendix 4 B.

A 4 B. The compilation of T.I. statistics within the PDP-8.

T.I. values were put into the computer by running a continuous time measurement program, which could be interrupted, the time sampled, reset, and restarted, on the arrival of a 'program interrupt' pulse (fig. A 4 B.1). The Z.C. pulses produced by the electronics described in chapter 1 were buffered into the 'program interrupt' line via a Schmitt trigger. The measurement program measured the T.I.s to the nearest 6 µs. (4 times the computer cycle time). As each T.I. was measured its value was stored in the core store in the form of one interval per word. A certain area of core was designated for T.I. storage, and when this was filled the program halted. When phase 2 of the equipment was in use, 4 K of the store was filled with T.I.s and immediately transferred to magnetic tape, from where they were available for any subsequent experiment

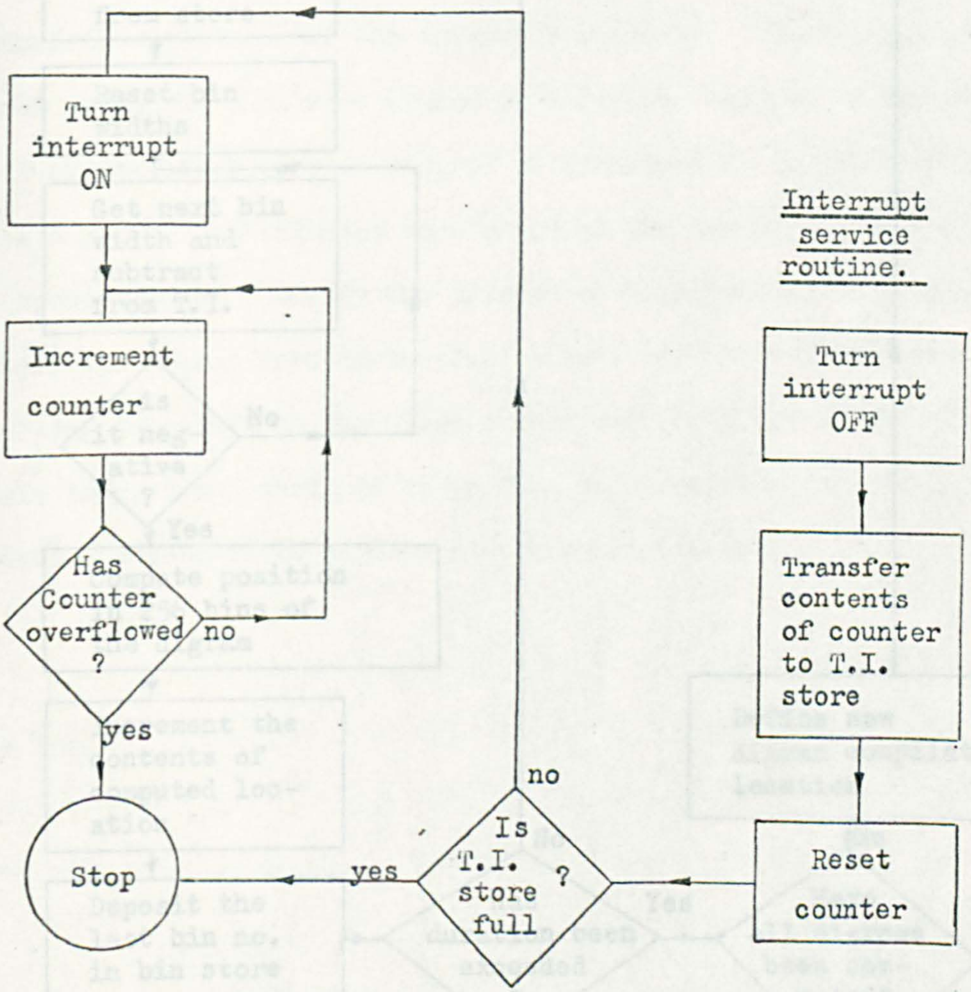


Fig. A4B.1 Flow diagram of the T.I. measurement program.

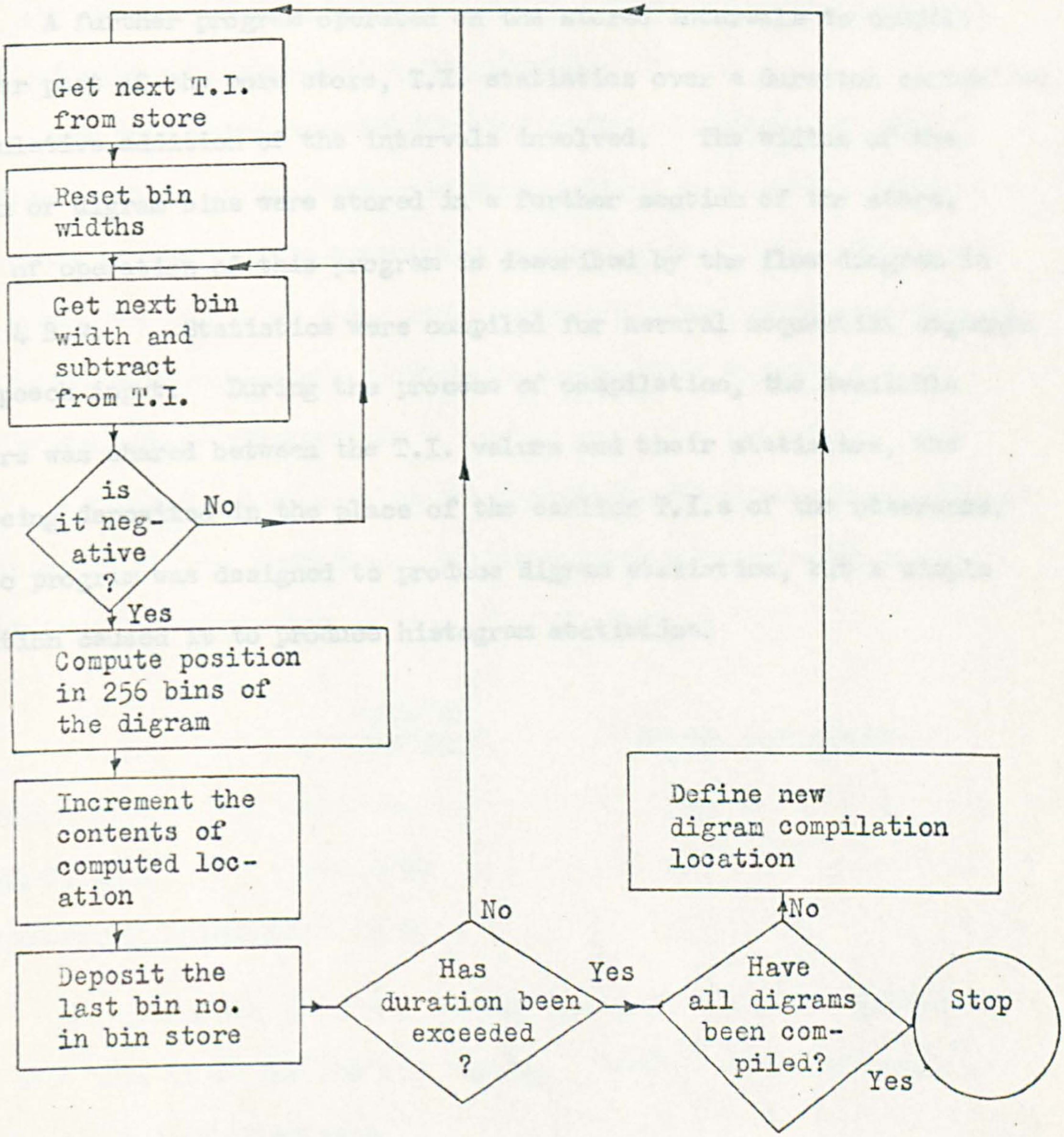

Get next T.I. from store
Reset bin widths
Get next bin width and subtract from T.I.
is it negative ?
No
Yes
Compute position in 256 bins of the digram
Increment the contents of computed location
Deposit the last bin no. in bin store
Has duration been exceeded ?
No
Yes
Have all digrams been compiled?
No
Yes
Define new digram compilation location
Stop

Fig. A4B.2 Flow diagram of the compilation of T.I. statistics from T.I.s stored in the computer.

on them.

A further program operated on the stored intervals to compile, in another part of the core store, T.I. statistics over a duration controlled by accumulative addition of the intervals involved. The widths of the histogram or digram bins were stored in a further section of the store. The mode of operation of this program is described by the flow diagram in figure A 4 B.2. Statistics were compiled for several sequential segments of the speech input. During the process of compilation, the available core store was shared between the T.I. values and their statistics, the latter being deposited in the place of the earlier T.I.s of the utterance. The basic program was designed to produce digram statistics, but a simple modification caused it to produce histogram statistics.

Appendix 5.

A.5. The results of subjective testing of utterances used in analysis and automatic recognition experiments.

The utterances recorded by speakers M.J.U., W.A.A., and J.B.M. on Language Master cards were presented in three conditions of distortion to a group of subjects in a series of listening sessions. The sounds were presented in random order via a power amplifier and a loud speaker. This output electronics was constructed and used by Underwood for similar subjective experiments.

A typical set of results are presented, with the results obtained by Ainsworth (2) for comparison, as his results were obtained using almost identical apparatus.

	ISOLATED VOWELS.	P.B. WORDS. (Ainsworth).
NATURAL.	61%	98%
NORMAL CLIPPED.	47%	85%
DIFFERENTIATED CLIPPED.	53%	99%

The articulation scores for the isolated vowels are considerably less than those for the P.B. words. This was expected owing to the unnatural lack of context.

A large range of values for the recognition of isolated vowels was obtained. Clipped vowel recognition varied between 20% and 60% for different sets of utterances. The relative scores for the three forms of processing remained fairly constant.

The most interesting difference is that whereas in the experiment on P.B. words, differentiation prior to clipping gave a score similar to that of natural speech, in the experiment on isolated vowels, this is not the case. The score is nearer to the normal clipped score. This result was also found by Underwood (66) who has done more extensive subjective testing of this.

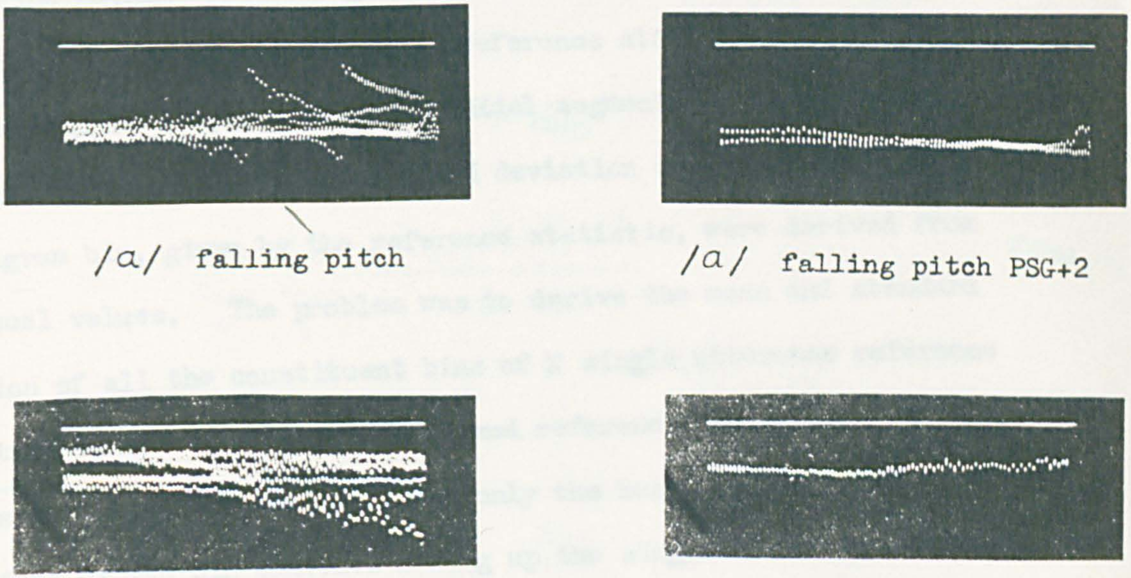
A reason for this result can be found by considering how differentiation effects characteristics of speech important for intelligibility, which are present in continuous speech, but are not present in isolated vowels. The formant transitions are certainly one of these characteristics. Work on the synthesis of speech on a formant model conducted at the Haskins Laboratories, New York, by Cooper (11), Delattre (14) and Liberman (37), and extended more recently at the Communication Department, Keele, by Ainsworth (3), has shown that the perception of different consonant-vowel pairs relies largely on differences in the second formant transition, the first formant always falling, on moving away from the vowel. The rôle of the second formant in the intelligibility of speech has also been investigated by Thomas (64). As the clipping of the waveform gives emphasis to the most dominant formant, the articulation scores obtained by Ainsworth (2) for normal and differentiated pre-clipping waveforms can be explained in terms of the importance of the second formant in defining the consonant-vowel transitions. This point is made by Thomas with reference to the similar experiments of Licklider. As no such transitions occur in isolated vowels, the emphasis on the second formant owing to differentiation may have less effect on the intelligibility of these clipped sounds. This is not to say that the second formant has no effect on the intelligibility of isolated vowels, this is highly unlikely; but it may have a greater effect on the intelligibility of the consonant-vowel transitions than on the steady state vowel segments.

Appendix 6.

A.6. Application of pitch synchronous gating to the intervalgram display.

The intervalgram display described by Chang et al.(8) was an attempt to provide similar visual information concerning continuous speech,to that available from the more complex spectrogram. They found that a better approximation to the formant patterns of vowels could be achieved by combining adjacent intervals in groups of two or four and taking their mean value, thus approaching a Z.C. rate measure. It has been seen that much of the perturbation of the T.I.s of an utterance with a varying pitch,is caused by those at the end of the glottal period. An alternative method of obtaining a more stable measure of vowel colour would be to reject the most variable intervals rather than smooth out the variations.

Some examples of the performance of this system,on the intervalgram of voiced sounds with a varying pitch,are shown in figure A 6.1. It was found that only stable intervals remained when severe p.s.g. was applied to the waveform, but that it was necessary to reject some stable intervals in order to avoid including unstable ones.



/ɑ/ falling pitch

/ɑ/ falling pitch PSG+2

/ɔ/ falling pitch

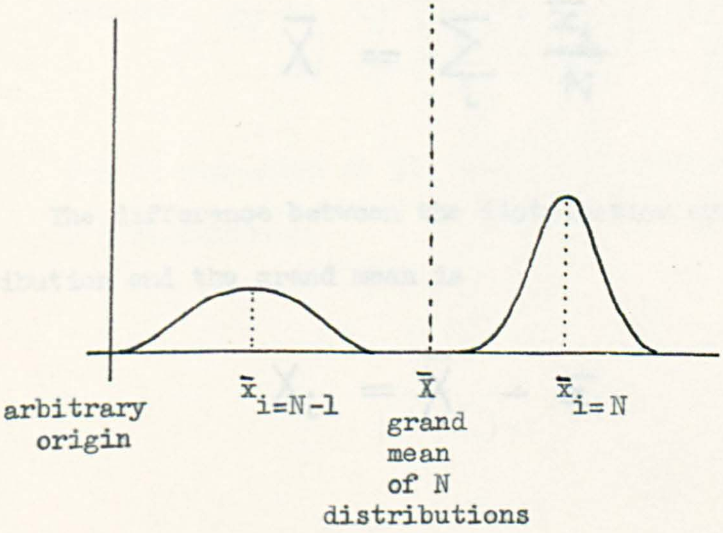
/ɔ/ falling pitch PSG+2

Fig. A6.1 Illustration of the effect of p.s.g. on the intervalgram display of vowels with falling pitch.

Appendix 7.

A.7. The mathematics of mixing the reference sets.

The mathematical background for the mixing algorithm was as follows. The single utterance reference statistics were compiled from ten digrams compiled from sequential segments of the utterance. Therefore the mean value and standard deviation of the contents of each digram bin, given by the reference statistic, were derived from ten actual values. The problem was to derive the mean and standard deviation of all the constituent bins of N single utterance reference statistics in order to derive the mixed reference statistic. It was not possible to do this directly as only the means and standard deviations of the bin contents making up the single utterance references were available. The original statistics were destroyed as described in section 4.2.1.. Thus the following analysis of the problem was required.



Consider the combination of N distributions, each of n elements. Let the distributions be typified by the subscript i, and the elements by subscript j,and the displacement of each element from some arbitrary origin be x_{ij}.

The variance of all elements of all distributions about the grand mean is

$$\sum_i \sum_j \frac{(x_{ij} - \bar{X})^2}{nN - 1}$$

The individual distribution means are defined by

$$\bar{x}_i = \sum_j \frac{x_{ij}}{n}$$

The grand mean being the mean of the individual distribution means is defined by

$$\bar{X} = \sum_i \frac{\bar{x}_i}{N}$$

The difference between the distribution mean of the ith distribution and the grand mean is

$$X_i = \bar{X} \dot{-} \bar{x}_i$$

$\bar{X}$ can be replaced in the expression for the 'grand variance' by $X_i + \bar{x}_i$.

The grand variance becomes

$$\sum_i \sum_j \frac{(x_{ij} - \bar{x}_i - X_i)^2}{nN - 1}$$

$$= \sum_i \sum_j \frac{(x_{ij} - \bar{x}_i)^2 + X_i^2}{nN - 1}$$

as $\sum_j (x_{ij} - \bar{x}_i) = 0$ by definition.

There are therefore two terms contributing to the grand variance.

$$\text{Term one} = \sum_i \sum_j \frac{(x_{ij} - \bar{x}_i)^2}{nN-1}$$

If nN is large and n $\gg$ N then

$$\frac{1}{nN - 1} \approx \frac{1}{N}\left(\frac{1}{n-1}\right)$$

$$\therefore \text{Term one} \approx \frac{1}{N} \sum_i \left[\sum_j \frac{(x_{ij} - \bar{x}_i)^2}{n-1} \right]$$

Term one can be expressed as the mean of the variances of the individual distributions.

$$\text{Term two} = \frac{\sum_i \sum_j X_i^2}{nN - 1}$$

As nN is large, nN - 1 ≈ nN, and term two can be approximated by

$$\text{Term two} \approx \sum_i \frac{X_i^2}{N}$$

This is a spread function of the N means of the individual distributions. It was found therefore that by summing these two terms, the mean of the variances of the single utterance reference statistics, and this spread function of their means, a fairly accurate estimate of the variance of the mixed reference statistics could be obtained. Note that this analysis refers to each bin of the statistics separately.

Appendix 8.

A. 8. The use of the digram display as 'visible speech'.

The potentialities of the digram display as a visible representation of continuous speech were not investigated formally during the present study. Informal experience of the visual feedback of continuous speech provided by the display, has made possible the following preliminary assessment of its usefulness as a 'visible speech' system.

The immediate impression on comparing the running histogram display, that is Chang's 'intervalgram', and the real time digram display, is that of a greater spread of information, and of patterns and changes in patterns which catch the eye in the latter. The extent to which these patterns are discriminative of the vowel sounds has been estimated in the quantitative studies. It is, however, unlikely that information transmitted visually is processed in a point by point manner as in the euclidean distance measurements.

One feature of the greater spread of information, and the emergence of clear patterns, that is considered important, is that much of the movement evident during non-steady state sounds involves more than one area of the display. The running histogram display often gives the impression of independent parameters with little coordination, even though vertical displacements of one spot must be accompanied by the opposite displacement of another if the number of T.I.s in the glottal period is constant. Typical movements in the digram display often involve positively correlated displacements at right angles to each other, in addition to the negatively correlated ones which conserve the length of the glottal period.

Some typical movements are:-

1. Along the diagonal X = Y as the dominant frequency of the sound changes.
2. Outwards parallel to an axis as pitch decreases, and in again as pitch increases.
3. Rotation of the pattern as the quality of the sound changes, but the dominant frequency remains constant.

Other more complicated patterns occur during transitions between consonants and vowels: some of these seem to be quite repeatable. In addition to the actual patterns involved, the tempo of pattern change correlates well with tongue and lip movements.

References.

Abbreviation : J.A.S.A. - Journal of the Acoustical Society of America.

1. Abercrombie,D. English phonetic texts. Faber and Faber, 1964

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

3. Ainsworth,W.A. Perception of stop consonants in synthetic CV syllables. To be published in Language and Speech.

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

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

6. Bezdel,W. Discriminators of sound classes for speech recognition purposes. 1967 Conference on Speech communication and processing, M.I.T.

7. Bloch,B. Trager,G.L. Outline of linguistic analysis. Linguistic Society of America,Baltimore. Waverley press, 1942

8. 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)

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

10. Cobb,S.M. The distribution of intervals between zero-crossings of sine wave mixed with random noise, and allied topics. I.E.E.E. Transactions on Information theory, I.T.-11 p.220 (1965)

11. Cooper,F.S. Some experiments on the perception of synthetic speech sounds. J.A.S.A. 24 p.597 (1952)

12. David,E.E. MacDonald,H,S. Note on pitch synchronous processing of speech. J.A.S.A. 28 p.1261 (1956)

13. Delattre,P.C. Liberman,A.M. Cooper,F.S. Gerstman,L.J. An experimental study of the acoustic determinants of vowel color;observations on one- and two-formant vowels synthesised from spectrographic patterns. Word 8 p.195 (1952)

14. Delattre,P.C. Acoustic loci and transitional cues for consonants. J.A.S.A. 27 p.769 (1955)

15. Dudley,H. Remaking speech. J.A.S.A. 2 p.165 (1928)

16. Dudley,H. Oscillograms. Fifth International Congress on Acoustics, Liege,1965 Paper A48

17. Egan,J.P. Wiener,F.M. On the intelligibility of bands of speech in noise. J.A.S.A. 18 p.435 (1946)

18. Fairbanks,G. Selective vocal effects of delayed auditory feedback. Journal of Speech and Hearing Disorders. 20 p.333

19. Fetz,E.E. Gerstein,G.L. An RC model for spontaneous activity of single neurons. Res.Lab.Electronics, M.I.T.,Quarterly Prog. Report No.71 (1963)

20. Flanagan,J.L. Some properties of the glottal sound source. Journal of Speech and Hearing Research. 1 p.99 (1958)

21. Flanagan,J.L. Speech analysis,synthesis, and perception. Academic press Inc., New York, 1965

22. Flanagan,J.L. Perceptual criteria in speech processing. Proc.Speech communication seminar,Stockholm, 1962 Paper D2.

23. Fourcin,A.J. An investigation into the possibility of bandwidth reduction in speech. Ph.D. thesis (1960) University of London.

24. French,N.R. Steinberg,J.C. Factors governing the intelligibility of speech sounds. J.A.S.A. 19 p.90 (1940)

25. Fry,D.B. Automatic recognition of speech. Proc.International Congr.Phonetics,Helsinki (1961)

26. Gold,B. Computer program for pitch extraction. J.A.S.A. 34 p.916 (1962)

27. Gill,J.S.	Automatic extraction of excitation function of speech with particular reference to the use of correlation methods.	Third International Congress on Acoustics, Stuttgart, 1959 Paper 703.
28. Herdan,G.	Statistics of phonemic systems.	Proc.4th.International Congress on Phonetic sciences,Helsinki,(1961)
29. Holmes,J.N.	An investigation of the volume velocity waveform at the larynx during speech, by means of an inverse filter.	Proc.Speech communication seminar,Stockholm, 1962 Paper B4.
30. Holmes,J.N. Mattingley,I.C. Shearme,J.N.	Speech synthesis by rule.	Language and Speech 7 p.127 (1964).
31. Jones,D.	An outline of English phonetics.	Heffer and sons,Ltd., Cambridge,U.K., 1956 ,p.36.
32. Kozhevnikov,V.A. Chistovich,L.A.	Speech:Articulation and Perception.	U.S.Dept.of Commerce, Clearinghouse for Federal scientific and technical information,Washington.
33. Kryter,K.D.	Speech bandwidth compression through spectrum selection.	J.A.S.A. 32 p.547 (1960)
34. Lavington,S.H. Rosenthal,L.E.	Some facilities for speech processing by computer.	Computer Journal 9 p.330 (1967).
35. Lawrence,W.	Synthesis of speech from signals which have a low information rate.	Proc.1952 Symposium on Applications of Communications theory. p.460.
36. Lee,B.S.	Effects of delayed speech feedback.	J.A.S.A. 22 p.824 (1950)
37. Liberman,A.M. Delattre,P.C. Cooper,F.S. Gerstman,L.J.	The role of consonant--vowel transitions in the perception of stop and nasal consonants.	Psychological monograms 68 No.8 p.1 (1954).

38. Lieberman,P. Perturbations in vocal pitch. J.A.S.A. 33 p.597 (1961)

39. Licklider,J.C.R. Effects of amplitude distortion upon the intelligibility of speech. J.A.S.A. 18 p.429 (1946)

40. Licklider,J.C.R. Pollack,I. Effects of differentiation, integration,and infinite peak clipping upon the intelligibility of speech. J.A.S.A. 20 p.42 (1948)

41. Licklider,J.C.R. Intelligibility of amplitude-dichotomised, time-quantised speech waves. J.A.S.A. 22 p.820 (1950)

42. MacKay,D.M. Sequential data display apparatus. Brit.Prov.Pat.Spec.43489 (1965).

43. MacKay,D.M. Improvements to or relating to electronic information--storage circuits,with special application to automatic waveform recognition. Brit.Prov.Pat.Spec.52722 (1965).

44. 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. To be published in Zeitschrift fur Phonetik.

45. MacKay,D.M. Ways of looking at perception. Models for the perception of speech and visual form Ed.W.Wathen-Dunn, M.I.T. press, 1967.

46. Mathews,M.V. Miller,J.E. David,E.E. Pitch synchronous analysis of voiced sounds. J.A.S.A. 33 p.179 (1961)

47. Miller,G.A. Language and Communication. McGraw-Hill,(N.Y.), 1951 p.40.

48. Miller,R.L. Auditory tests with synthetic vowels. J.A.S.A. 25 p.114 (1953)

49. Miller,R.L. The nature of the vocal chord wave. J.A.S.A. 31 p.667 (1959)

50. Noll,A.M. Short time spectrum and 'cepstrum' techniques for vocal pitch detection. J.A.S.A. 36 p.292 (1964)

51. Peterson,G.E. Barney,H.L. Control methods used in a study of the vowels. J.A.S.A. 24 p.175 (1952)

52. Potter,R.K. Kopp,G.A. Kopp,H.G. Visible speech. Dover publications Inc., (N.Y.),1966.

53. Rainal,A.J. Axis crossings of the phase of sine waves plus noise. Bell systems technical journal 46 p.737 (1967).

54. Reddy,D.R. Computer recognition of connected speech. J.A.S.A. 42 p.329 (1967)

55. Sakai,T. Inoue,S. New instruments and methods for speech analysis. J.A.S.A. 32 p.441 (1960)

56. Sakai,T. Doshita,S. The automatic speech recognition system for conversational sound. I.E.E.E. Transactions on Electronic computers. E.C.-12 p.835 (1963).

57. Scarr,R.W.A. Zero crossings as a means of obtaining spectral information in speech analysis. 1967 Conference on speech communication and processing, M.I.T.

58. Shannon,C.E. The mathematical theory of communication. Bell systems technical journal 27 p.379 and p.623 (1948).

59. Stetson,R.H. Motor phonetics - a study of speech movements in action. North Holland publishing Co.Ltd. (1951).

60. Stevens,K.N. Volkman,J. The relation of pitch to frequency:a revised scale. American Journal of Psychology 53 p.329 (1940)

61. Stover,W.R. Time-domain bandwidth-compression system. J.A.S.A. 42 p.348 (1967)

62. Sugimoto,T. Hashimoto,S. The voice fundamental pitch and formant tracking computer program by short-term autocorrelation function. Proc. Speech communication seminar,Stockholm,1962. Paper C11.

63.	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).
64.	Thomas,I.B.	The significance of the second formant in speech intelligibility.	Biological computer lab., University of Illinois, Tech.Rep.No.10 (1966).
65.	Teacher,C.	Human recognition of sustained phonemes.	Meeting of Acoust.Soc. of America, Nov.1962, Ann Arbor, Michigan.
66.	Underwood,M.J.	Time interval statistics in speech synthesis: a critical evaluation.	Ph.D. thesis, 1968, University of Keele.
67.	Whitfield,I.C.	The auditory pathway.	Edward Arnold Ltd.,London, 1967.
68.	Woodward,P.M.	Probability and information theory with applications to radar.	Pergammon press Ltd., London,1964.
69.	Zwirner,E. Zwirner,K.	Grundfragen der Phonometrie.	Berlin,1936.