TIME INTERVAL STATISTICS IN SPEECH SYNTHESIS:

A CRITICAL EVALUATION

by

M. J. UNDERWOOD

Thesis submitted for the degree of Doctor of Philosophy in the University of Keele.

April, 1968.

ABSTRACT

A speech wave that has been successively amplified and limited so that it is reduced to a rectangular form is intelligible to a human listener. Much information is retained in the temporal pattern of the time-intervals between the changes of state of such a wave. Techniques for the measurement and display of the first and second-order statistics of these time-intervals are described. The results of these analyses are used to produce synthetic clipped speech sounds. The ordering of the time-intervals within the sounds is an important factor in the perception of the sounds. Two different methods for eliminating unreliable time-intervals are described, only one of which is suitable for application to speech synthesis. Using a digital computer for the analysis and synthesis, three methods of using the first, second and third-order statistics to produce isolated vowel sounds are described. Although some of the synthetic vowels do not sound voiced, vowels produced from third-order statistics are nearly as recognisable as the original clipped vowels. Preliminary results from the synthesis of words and phrases indicate a very low level of intelligibility, as the methods of using the statistics do not give a precise enough indication of some of the key parameters in the speech signal. Measurements of the storage requirements needed to specify the different statistical analyses of clipped speech indicate that time-interval statistics are not a very economical way of specifying a clipped speech signal.

PREFACE

The work described in this thesis was conducted in the Department of Communication between October 1964 and December 1967. After a certain amount of preliminary work had been conducted, it was apparent that the topic could be adequately investigated only with the use of a digital computer. Delivery of the digital computer was delayed until November 1966, when the basic machine was delivered. It was not until April 1967, that the installation was complete. High speed data input and output devices were therefore not available until then, making the testing of programs a long process. Most of the quantitative work described in Chapters 4 and 5 was accomplished in the few remaining months of an S.R.C. Research Studenship.

In order that the work can be seen in the context in which it was executed, this thesis is a more or less chronological exposition of the project.

ACKNOWLEDGEMENTS

I would like to express my appreciation to all those people who helped in any way with the work described in this thesis and its subsequent presentation.

In particular, I am very grateful to Professor D.M. MacKay for his suggestion of the research topic and his continual advice and guidance as to its development. I am also grateful for the advice given by other members of the Department of Communication, notably Dr. W.A. Ainsworth,and A.W. Wright for his help in computing matters. I would like to record the particularly helpful co-operation I received from J.B. Millar in all aspects of the project. I would also like to thank my fiancée,Mary,for preparing most of the diagrams and suggesting modifications to the manuscript.

I appreciate the continued financial support provided by the Science Research Council.

M.J. Underwood

April, 1968.

CONTENTS

CHAPTER 1 INTRODUCTION

CHAPTER 2 FIRST ORDER ANALYSIS AND SYNTHESIS

Page

CHAPTER 3 SECOND ORDER ANALYSIS

CHAPTER 4 N-GRAM ANALYSIS AND SYNTHESIS

CHAPTER 5 PITCH SYNCHRONOUS ANALYSIS AND SYNTHESIS

CHAPTER 6 CONCLUSION

APPENDICES

CHAPTER I

The high speed of operation of modern digital computers has given rise to a growing problem of communication between man and machine. Considerable effort is being devoted to the study of symbiosis, the man-machine interaction, to make the man-machine interface more efficient.

At present, information has to be coded into a form suitable for acceptance by the computer. Less stringent requirements exist for the output of data. As speech is man's most important method of communication, much effort is directed towards research and the subsequent development of systems that will enable man and the machine to communicate by means of speech or speech-like sounds.

As well as providing a motivation for speech studies, the computer is an extremely useful tool for research into speech. Before reviewing what techniques exist for analysing speech it is relevant to consider how speech is produced.

1.1 The Production of Speech

The primary source of energy in the speech wave is the glottis or vocal chords. When the glottis is not being used for speaking, it allows the free passage of air to and from the lungs. When the glottis is partly closed, a noise is produced by the turbulent flow of air through the gap. Whispered speech is produced in this way. When the glottis is further constricted, an increase in air pressure in the lungs produced by muscular movements causes the glottis to act as a relaxation oscillator. The glottal waveform is rich in harmonics and causes the air in the vocal tract to vibrate in several modes simultaneously. The frequencies of vibration are determined mainly by the position of the tongue and lips. The resonant frequencies are called formants and can be identified as peaks in the frequency spectra of voiced sounds. Additional modes of vibration are produced when the nasal cavity is opened to produce such sounds as /m/ and /n/.

All the vowels and several consonants can be uttered continuously. Other sounds, however, require movement of the articulators for their production. Stop sounds or plosives, such as /p/ and /d/, come into this category. Stop sounds are produced when the vocal tract is closed at some point with the result that the air pressure builds up. The release of the closure causes a small explosion of air, hence the name plosive sounds.

When the vocal tract is constricted but not completely closed at some point, air forced through the gap produces fricative sounds. /s/ and /ʃ/ are examples of this type of sound. Typically they have energy spread over a very wide range of frequencies.

1.2 Speech Analysis

Hindu grammarians around 300 B.C. were the first to characterise speech sounds by the positions of the articulators necessary to produce them. This method is still used by phoneticians, and much work is in progress to relate the articulatory specification of speech to the major resonances of the vocal tract (44,51).

Helmholtz in 1862 (26) first showed that the vocal cavities could be considered as a resonator, the resonances of which determined vowel quality. It was not until the advent of the thermionic valve that it was possible to extend the techniques of frequency analysis. The first spectrographic studies of speech were published by Steinberg in 1934 (58).

The spectrograph became generally available after the war and has been used ever since for the study of speech. In the usual form of spectrographic display, time and frequency are the abscissae and ordinates respectively. The energy within a given frequency band is represented by the intensity of marking the paper. If a narrow-band filter is used, the individual harmonics of the glottal frequency can be resolved. If a broad-band filter is used, the spectrograph displays bands corresponding to the formants. A modified version of the spectrograph is currently being used to provide "voiceprints" (identification of speakers by spectrographs of their voices) with varying degrees of success (30, 64).

Parallel developments in speech synthesis have done much to encourage the use of frequency analysis techniques, and their subsequent incorporation into systems for automatic speech recognition (21).

1.3 Speech Synthesis

The most elaborate of early speaking machines was the mechanical analogue produced by Wolfgang von Kempelen about 1790 (12). As it allowed the control of several speech-like parameters it was capable of producing consonants as well as vowels.

The first analysis-synthesis system was demonstrated by Dudley in 1939 (11). The speech signal was passed into a filter-bank. The outputs of the filters were integrated and rectified to produce slowly changing control signals. These signals were used to control filters that were excited either by noise or a buzz source. The success of this device (called a Vocoder) supported the idea that the frequency components of speech were the information-bearing parameters.

After the war, the first electrical analogues of the vocal tract were produced (14, 59). The parameters for the control of these synthesisers were derived from X-ray studies of the vocal tract.

In 1952, Lawrence (32) produced the first formant synthesiser. His synthesiser was not a model of the vocal tract, but of its acoustic behaviour. Band-pass filters whose centre-frequencies could be adjusted were excited by a buzz or noise source, corresponding to the production of voiced and unvoiced sounds respectively. Lawrence showed that two formants were sufficient to produce recognisable vowels and that intelligible synthetic speech could be produced. The formant tracks were derived from spectrographic records.

Since then, many versions of the formant synthesiser have been built. When sufficient detail is incorporated in the control signals, it is possible to produce synthetic speech that is indistinguishable from real speech (29).

All the successful speech synthesisers have been models either of the vocal tract or of its acoustic behaviour. Consequently they enable various hypotheses about speech production and perception to be evaluated. It is important that the studies of speech synthesis and analysis do not go on in isolation from one another as there is much to be gained from their mutual interaction.

1.4 Time-Interval Analysis of Speech

Other techniques besides speech analysis and synthesis have been used to study speech. One of these techniques is to distort the speech so that certain features of the speech signal are removed. The intelligibility of the remaining speech is an indication of the relative significance of the information that has been removed, and that remaining.

Amongst the most notable of the experiments using distorted speech are those of Licklider (34, 35, 36). He investigated the effects on speech intelligibility of different types and amounts of amplitude distortion. In its severest form, the speech wave was reduced to a rectangular wave by successively amplifying and limiting the signal. This rectangular form of speech is called infinitely clipped speech, abbreviated to clipped speech for convenience. The only similarity between the clipped and the original signals is that the rectangular wave switches from one state to another at the same time as the original wave crosses the zero-pressure axis. In spite of its simple form and harsh quality, infinitely clipped speech was found to be intelligible, particularly if it was differentiated with respect to time before clipping.

These experiments show that the amplitude information in the speech wave is relatively unimportant and that a lot of information is carried in the temporal pattern of zero-crossings of the original signal. This latter fact has been used as a basis for further research into time-interval analysis of speech.

Throughout this thesis, time-intervals or intervals will refer to the time-intervals between the zero-crossings of a speech wave. If the speech is differentiated before clipping, changes of state of the rectangular wave correspond to the points of zero slope of the original wave. Where the speech has not been differentiated prior to clipping, it will be referred to as clipped speech or the word normal will be used to distinguish it from the differentiated form.

Chang (5) was the first person to examine the statistics of time-intervals. He postulated that the mean zero-crossing rate of voiced speech could be identified with the first formant, and the mean zero-crossing rate of differentiated speech with the second formant. Peterson (49) later showed that the relationship between zero-crossing rate and formant frequencies was not a good one unless the speech was filtered first to separate the formants. Even this technique is inaccurate under some conditions (50, 63).

Chang (6) also developed a visual display of speech using time-intervals, called an Intervalgram. It was very similar to a spectrograph except that the ordinate was length of time- interval rather than frequency. The brightness at any point was a measure of the number of occurrences per unit time of time-intervals of a particular length, so that the display could be regarded as giving a running histogram of the time-intervals in speech.

The simplicity of the device demonstrates one of the advantages of time-interval analysis. Filter banks are complex and costly, and there is a delay in the response of a filter which is dependent upon its bandwidth. Basically simple circuitry can be used for time-interval analysis and there is no question of more than one channel being activated at once. The only delay is that due to the length of the interval itself.

Sakai and Inoue (53) carried time-interval analysis a stage further and produced time-interval histograms of several isolated Japanese sounds. They concluded that the peaks in the histograms of voiced sounds were correlated with the formants, though there was not a unique relationship between them. Sakai and Doshita (54) later used time-interval measurements in an automatic speech recognition device for which 90% recognition of five Japanese vowels was reported.

Bezdel (3, 4) has also used time-interval histograms as the basis of a system for automatic speech recognition. He achieved results that were comparable to those obtained using spectral analysis (3, 21).

As far as the author is aware, no attempts at using time-interval statistics for speech synthesis have been made, other than using zero-crossing rate information to provide parameters for a formant synthesiser. As the time-interval statistics are derived from a clipped speech wave, the question arises as to whether time-interval statistics can be used to produce a synthetic clipped speech wave that is intelligible. This thesis will try to answer that question.

1.5 Objectives

The work described in this thesis is part of a larger study of time-interval analysis of speech that has been followed in the Department of Communication since October, 1964.* Within this general study, the work to be described here has two main objectives.

(1) To investigate the relative intelligibility of different statistical approximations to clipped speech based on time-interval statistics of the original waveform hence

(2) To discover the relative importance of different components of the statistical specification of clipped speech.

The criterion used to evaluate the synthetic speech was that the synthetic speech should be recognised as the same sequence of sounds that were used in compiling the statistics. This is distinct from the far more stringent criteria that the synthetic speech should be indistinguishable from the original or should retain sufficient information for the speaker to be identified.

As human subjects were used in assessing the intelligibility of the synthetic speech, the work also throws some light on the factors that govern the perception of clipped speech.

*The other parts of this study are to be found in :-

"An Evaluation of Three Related Techniques for the Statistical Analysis of Clipped Speech". Ph.D. Thesis to be submitted by J.B. Millar. (Reference 45).

"Relative Intelligibility of Different Transforms of Clipped Speech". by W.A. Ainsworth. (Reference 2).

1.6 Arrangement of Thesis

The work to be described falls into two major categories. Chapters 2 and 3 are concerned with the qualitative aspects of the problem. Chapters 4 and 5 describe the quantitative measurements that were obtained using a digital computer for the analysis and synthesis of clipped speech.

In Chapter 2, time-interval histograms of vowel sounds are examined to see how they can be utilised. A flexible time-interval generator is described that enables some preliminary attempts at the synthesis of clipped speech to be made. In the light of these experiments it is possible to define some of the problems of synthesising a clipped speech wave.

New apparatus for producing a real-time display of the second-order statistics in speech is described in Chapter 3. This display shows some of the limitations of time-interval analysis. A way of overcoming them is described which, although it is not directly applicable to the synthesis of clipped speech, is the basis of a method that is developed in Chapter 5.

The results of the synthesis of some steady-state sounds and short phrases from first, second and third-order statistics are described in Chapter 4, together with some measurements on the amount of information needed to specify the statistics.

In Chapter 5, pitch-synchronous techniques for the analysis and synthesis of clipped speech are described. Two methods of using the statistical information to produce more realistic voiced sounds are described, and the more successful of these is applied to the synthesis of words and phrases.

Chapter 6 summarises the main findings of this work, and draws conclusions from them.

CHAPTER 2

2.1 Introduction

A speech wave that has been successively amplified and limited so that it is reduced to a rectangular form is highly intelligible to a human observer (35). The time-intervals between the zero-crossings of such an infinitely clipped wave are in a convenient form for statistical analysis. Fourcin (22) has investigated the long-term first-order statistics with a view to bandwidth compression. Bezdel and Chandler (3) have used the histograms of time-intervals compiled over shorter periods as the basis of a system for automatic speech recognition.

This chapter is concerned with how the first-order time-interval statistics may be used to generate synthetic speech. A generator is described that utilises such information to produce a synthetic clipped speech wave and some qualitative results are discussed.

2.2 First-Order Statistics

The first-order statistics of a set of events that can be classified into n categories is defined by the set of probabilities p_i, where p_i is the probability of event i occurring. As some event must occur,

$$\sum_{i=1}^{n} p_i = 1$$

The first-order distribution of probabilities is referred to as a histogram when a graph is drawn with the n categories as abscissa and their relative probabilities of occurrence as ordinates. Where the histograms are not normalised, the ordinate is the number of occurrences (f_i) within a category, and the sum of the contents of all the bins is the total number of occurrences of all events N. Thus -

$$\sum_{i=1}^{n} f_i = N \quad \text{and} \quad p_i = f_i/N$$

If the measurements have been made over time T, the quantity $\frac{N}{T}$ is the mean rate of occurrence of the events. Chang (5) has used the mean zero-crossing rates of normal and differentiated speech as measures of the formant frequencies.

These statistical measurements are often a concise way of specifying the overall behaviour of a set of events. Where the events are not statistically independent, the first-order statistics are still a valid measure, but are not capable of expressing the interdependence of the events.

A statistical model of the way in which the events are generated can be made using the first and higher order statistics. Such a process, where the events are chosen according to a statistical model is called a stochastic process. Information theory is based on the fact that all discrete processes (and continuous processes if quantised into discrete ones) can be considered as stochastic processes. Where the statistical properties of the source remain unchanged over a long period of time, the source is termed ergodic. Information theory is capable of dealing with ergodic sources, but as yet no theory is generally available to deal with non-stationary statistics.

2.2.1. Approximations to Written English

In 1948, Shannon (56) considered English to be written according to a stochastic process, where the events are the individual letters and space of the alphabet. He made several approximations to English using compiled statistics of letters of the alphabet.

His zero-order approximation was generated by choosing the letters at random. In this way there were as many Q's as E's in his approximation. In his first-order approximation he chose the letters according to their relative frequency of occurrence, so that there were many more E's than Q's. This first-order model of the English language allowed such unlikely combinations of letters as QXZ etc. In English, Q is always followed by U. This constraint was incorporated in his second- order model, where the letters were chosen according to the preceding letter. Although none of his approximations made sense, the general likeness to English was quite good. Some similar experiments have been reported by Guilbaud (24) for Latin.

2.2.2. Application to Synthesis of Clipped Speech

Many experiments have been performed that show that speech can undergo many types of distortion and still remain intelligible to a human listener. This is because of the high informational redundancy of speech. Amongst the most notable of these experiments are those of Licklider (34, 35, 36). By infinitely clipping the speech wave, thereby reducing it to a rectangular form, he discarded most of the amplitude information. He found that although clipped speech had a harsh unpleasant quality, it was nevertheless highly intelligible, particularly if the speech was differentiated before clipping.

The clipping process can be regarded as converting the signal from an analogue to a digital form, the zero-crossings of the original wave being signalled by changes in polarity of the rectangular wave. If the rectangular wave is differentiated and full-wave rectified to produce a series of uni-directional pulses, the spacing of these pulses is equal to the time-intervals between the zero-crossings of the original wave. These pulses can be used to control circuitry that measures the intervals between them. The intervals themselves form a continuous distribution, but as Licklider (36) later showed, the intervals can be quantised without impairing the intelligibility greatly. In this way the continuous waveform of speech has been reduced to a series of pulses separated by discrete intervals. The speech signal is now in a convenient form for statistical analysis. Can intelligible clipped speech be made from time-interval statistics in the same way that Shannon generated his approximations to written English?

Although the work of Shannon was one of the starting points for this work, there are one or two major differences due to the nature of the speech signals.

Shannon's basic symbols were the letters or words of written English. The tables he used to generate his approximations were compiled from long-term statistics, so that the long-term statistics of his approximations were the same as those of written English. Fourcin (22) has measured the long-term statistics of the time-intervals between zero-crossings in speech with a view to bandwidth compression. He found that the first-order statistics of isolated sounds varied considerably from the long-term distributions, and that there were significant differences between the first-order distributions for different sounds.

The problem of speech synthesis from time-interval statistics can be divided into two categories.

The first question that has to be answered, is whether recognisable steady-state speech sounds can be produced from time-interval statistics. It is obviously simpler to design a system that will produce steady-state sounds than it is to devise one that is capable of producing continuous speech. If the statistical analysis has discarded so much information that recognisable steady-state sounds cannot be produced, there is very little likelihood of being able to produce intelligible continuous speech. For the synthesis of steady-state sounds, it will be sufficient to analyse sustained utterances of those sounds, and use the results to control a suitable generator.

Although human speech can be considered as a sequence of discrete symbols or phonemes, the process of speech production is continuous. Because of the inertia of the articulators, they cannot move instantaneously from one position to another. It has been shown that formant transitions, corresponding to the movement of the articulators from one position to another, play an important part in the perception of some sounds, notably stops and semi-vowels (33). Any device which is going to produce intelligible speech therefore, must be capable of reproducing the effects of these movements. Holmes (28) and David (10) have shown that formant movements can be represented by a series of discrete steps, if the spacing of the steps is close enough to give an impression of continuous movement. Continuous speech, therefore, can be considered as a sequence of sub-phonemic steady-state segments.

To produce continuous synthetic speech it will be necessary to compile statistics over fairly short segments of time, typically 10 to 20 msec,and produce a new signal lasting for the same length of time. It is likely that two problems will be encountered here. The first is that comparatively short analysis times will make the statistical measurements less reliable so that less realistic sounds will be produced. The second is that although adjacent segments of synthetic speech might be individually recognisable, the very bringing together of them may produce a discontinuity at the boundary. It has been shown (19,20) that there are tolerances in the specification of vowel sounds, so that two sounds which have slightly different acoustic properties are both recognised as the same vowel. If these two sounds are brought together, however, a most unnatural discontinuity will be heard. The major problem then in producing intelligible continuous speech is to give the impression of continuity of movement of the articulators.

2.3 Perception of Clipped Speech

Before considering the analysis of a clipped speech wave, and how to produce a new sound from that analysis, it is relevant to consider the perception of clipped speech.

It is a remarkable fact that a speech wave that has been stripped of all its amplitude information is still intelligible to a human listener. Its intelligibility and the comparatively low bandwidth required to transmit it have received a fair amount of attention, but a really satisfactory explanation of its high intelligibility has not been found.

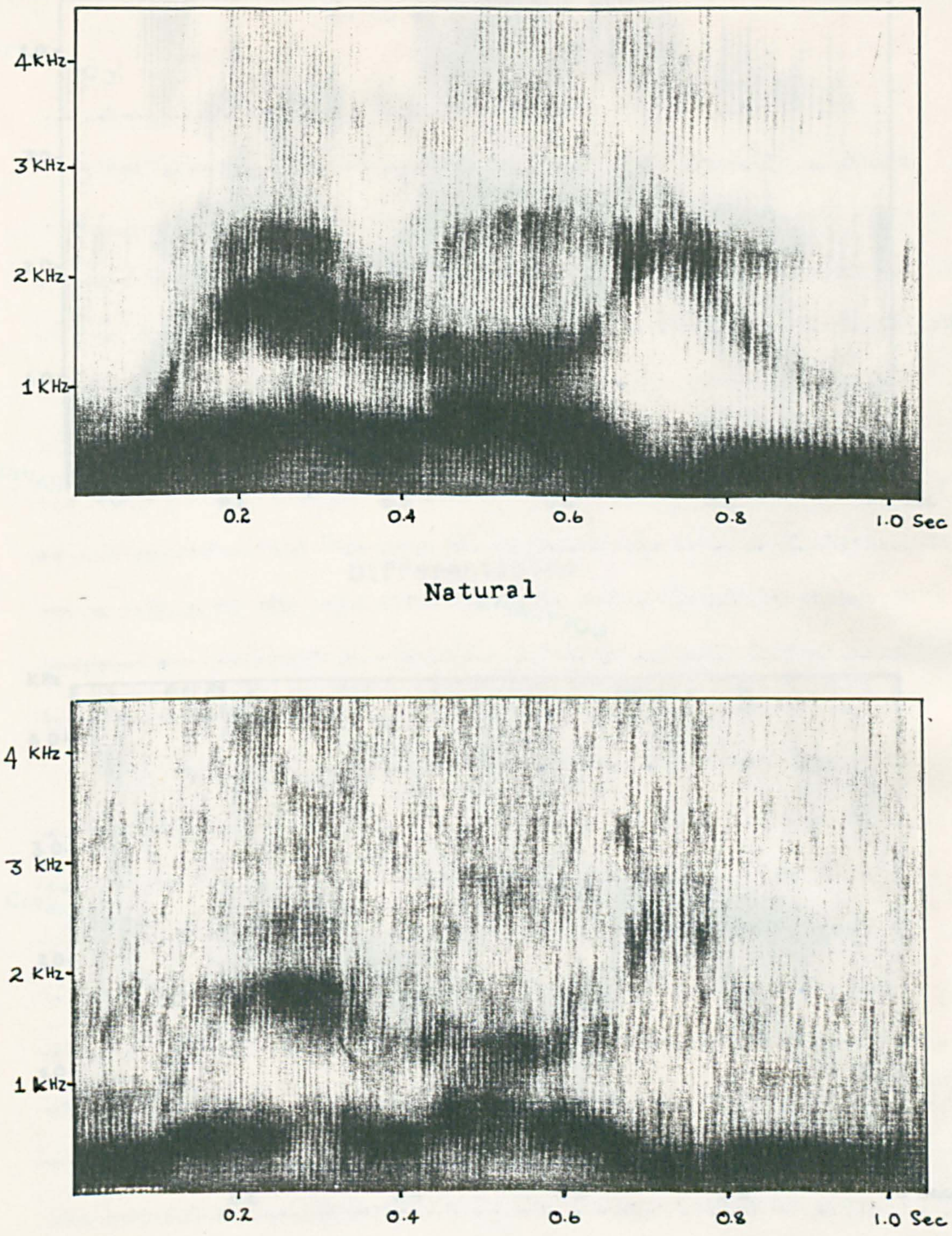
Clipping a speech wave has two major effects. The first is that much amplitude and high-frequency information is lost. The second is that the rectangular edges produce inter-modulation and spurious high frequency components. Various studies (22, 65) have shown that the frequency of the major spectral component of the clipped wave corresponds to that of the dominant frequency of the original wave. Licklider (34), in comparing centre-clipping with peak-clipping as methods of amplitude distortion, explained the higher intelligibility of peak-clipped speech by showing that the sonagrams of peak-clipped speech bore more resemblance to those of the original signal than the sonagrams of centre-clipped speech did.

Ainsworth (2) explained his results for different transforms of clipped speech by showing that the least intelligible transforms had destroyed the most spectral information.

Thomas (63) demonstrated the importance of the second formant in speech perception. He explained the higher intelligibility of differentiated and clipped speech by pointing out that differentiating the speech wave made the second formant the dominant frequency component, with the result that it survived the clipping process better.

The sonagrams shown in Figures 2.1 and 2.2 support the hypothesis that clipped speech is intelligible because some of the major frequency components come through the clipping process relatively unchanged. The sonagrams show that there is broadening of some formants and narrowing and suppression of others. The sonagram of the clipped utterance also shows harmonies of the first formant frequency as well as a general increase in the overall noise level.

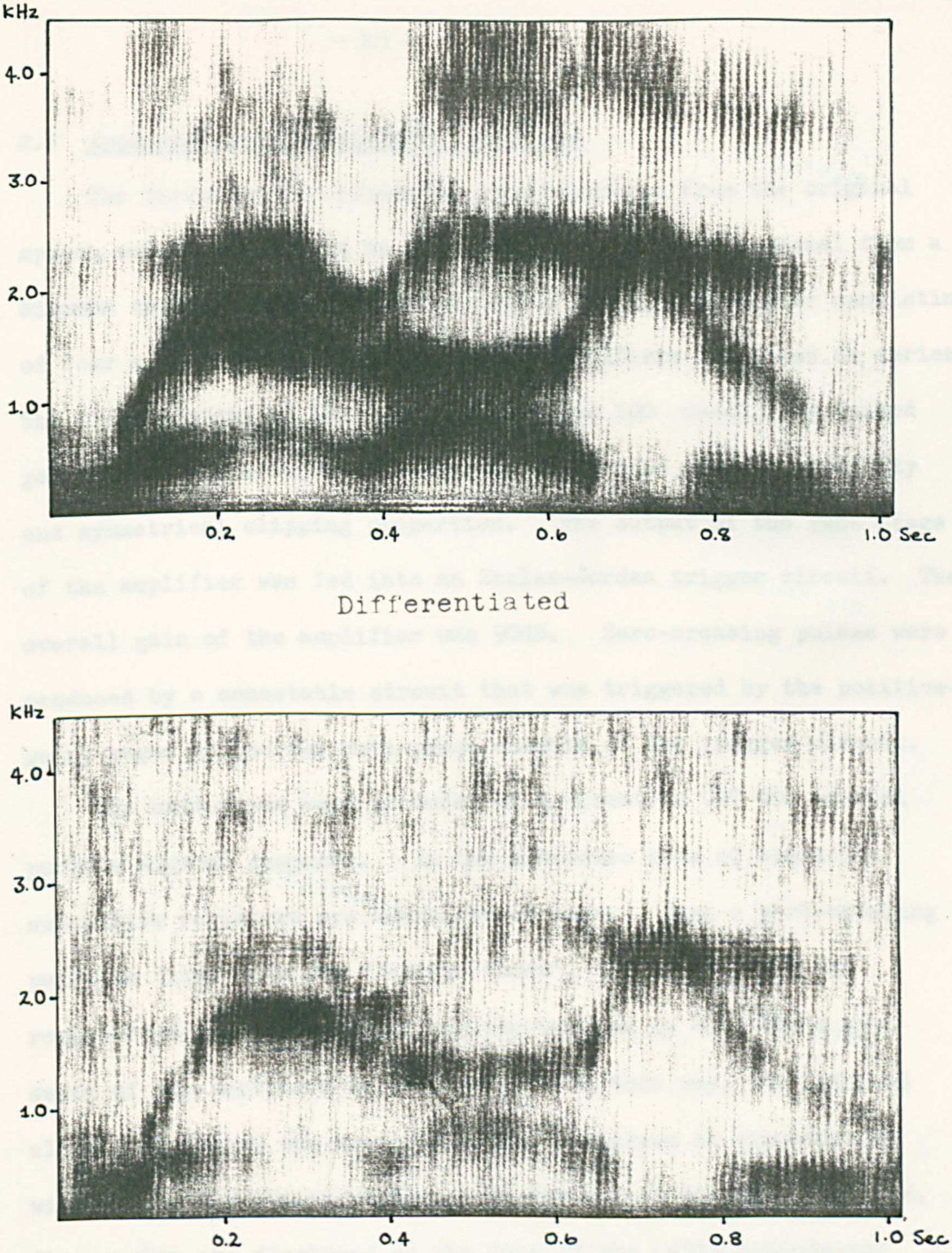
Great care had to be taken in the interpretation of these results, as the speech waveform is extremely redundant and human-beings posess remarkable powers of being able to extract a meaningful message from the slenderest of cues. A point that must not be overlooked is that the clipping process has not upset the basic temporal relationship between events in the speech wave.



Natural

Clipped

Fig. 2.1 Broad Band Sonagrams of "Where are you?"



Differentiated

Differentiated and Clipped

Fig. 2.2 Broad Band Sonagrams of "Where are you?"

2.4 Apparatus for Time-interval Analysis

The circuitry for producing zero-crossings from the original speech wave was designed by J.B. Millar. The speech signal from a Siemens tape recorder was played into a clipping amplifier consisting of four a.c. coupled long-tailed pair amplifiers connected in series. The time-constant of the a.c. coupling was 100 msec. Long-tailed pair amplifiers were used because of their good thermal stability and symmetrical clipping properties. The output of the last stage of the amplifier was fed into an Eccles-Jordan trigger circuit. The overall gain of the amplifier was 90dB. Zero-crossing pulses were produced by a monostable circuit that was triggered by the positive-going edges of the two anti-phase outputs of the trigger circuit.

The histograms were produced on a Mnemotron CAT 400 special purpose digital computer. In the histogram mode of operation successive registers are addressed in turn. When a zero-crossing pulse is fired into the "Address Reset", the contents of the register currently addressed are incremented by one, and a new sweep of the registers is initiated. In this way, the internal clock that causes the scanning of the registers is synchronised with the zero-crossing pulses. At the end of the analysis time, the results are displayed on the face of the CAT's cathode-ray tube.

2.4.1. Choice of Speech Sounds

In continuous speech, no speech sound can be considered as being steady-state. As mentioned previously, formant transitions are important cues in the perception of some sounds. Although vowel sounds can be uttered in isolation as steady sounds, in continuous speech they often undergo vowel reduction (39, 60), that is, the articulators move towards the steady-state positions, but may never get there as the articulators are moved to produce the next sound. For this reason it was considered impracticable to extract "steady-state" portions of sounds from continuous speech. Instead, steady-state phonemes were used that could be spoken in isolation. For the preliminary work, twelve vowel sounds and two fricatives were considered. The vowels were those published by Holmes (28), considered to be a representative set of English vowel phonemes, and the two fricatives /s/ and /ʃ/.

2.5 First-Order Statistics of Vowel Sounds

The twelve sustained vowel sounds were spoken and analysed by J.B. Millar. The statistics were compiled over 500 msec of speech so that the effect of compiling the statistics over an incomplete number of glottal periods was reduced. For example the mean glottal period was 6.5 msec. If 30 msec of speech had been analysed, the statistics would relate to 4.6 glottal periods. The contribution of the fractional part of the glottal period might then make a significant difference to the form of the statistics. As the results were to be used to synthesise a waveform repeating at the glottal rate, it was decided to analyse long portions of vowel sounds, so that the statistics could be considered as being compiled from an integral number of glottal periods.

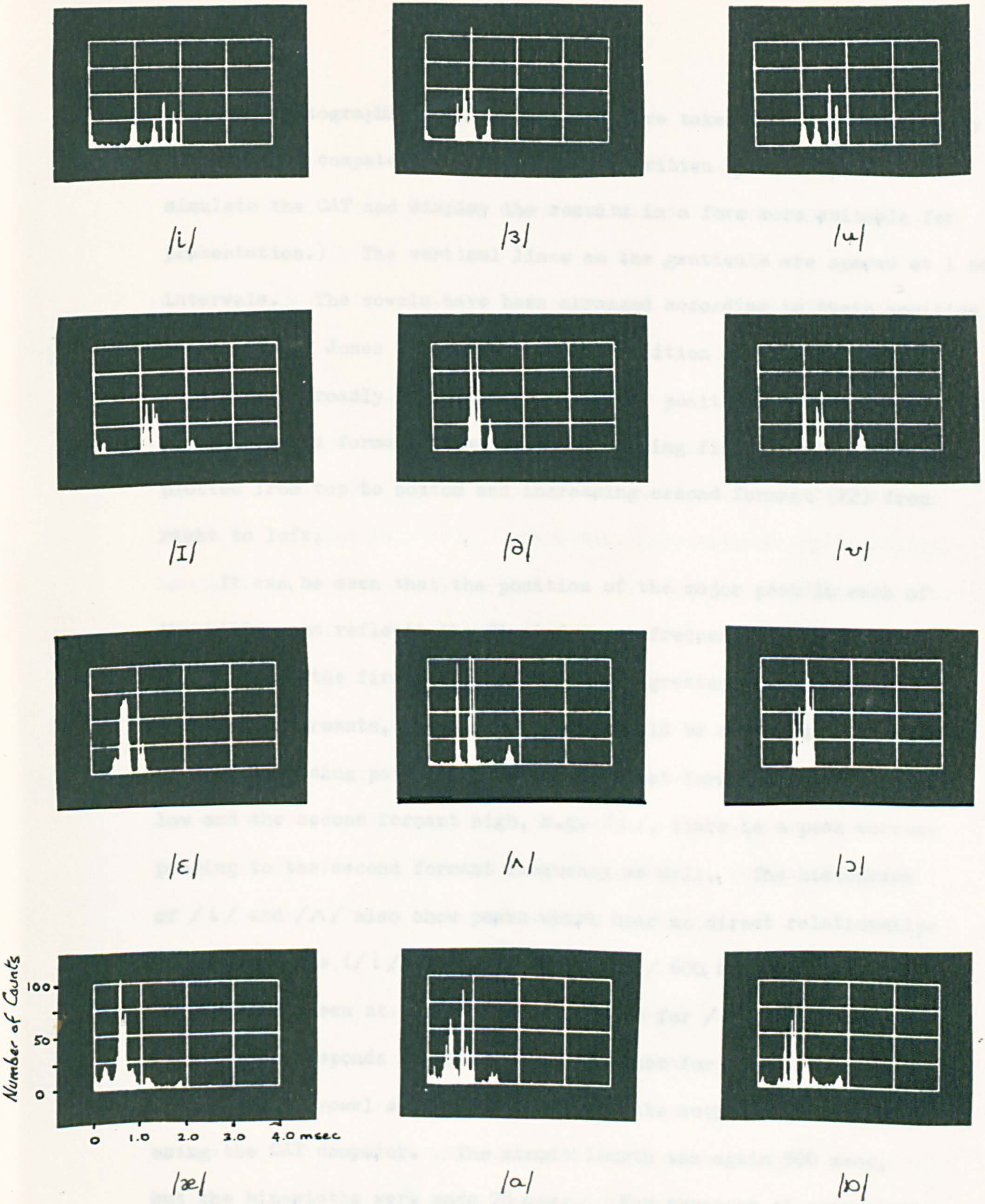


Fig. 2.3 Histograms of Vowel Sounds

The photographs shown in Fig. 2.3 were taken from the 338 display of the PDP-8 computer. (A program was written by J.B. Millar to simulate the CAT and display the results in a form more suitable for presentation.) The vertical lines on the graticule are spaced at 1 msec intervals. The vowels have been arranged according to their position on the Daniel Jones Vowel Chart. The position of the vowels on the vowel chart broadly corresponds with their positions in the first formant/second formant plane, with increasing first formant (F1) plotted from top to bottom and increasing second formant (F2) from right to left.

It can be seen that the position of the major peak in each of the histograms reflects the first formant frequency. As the amplitude of the first formant is usually greater than those of the higher formants, the first formant would be expected to dominate the zero-crossing pattern. Where the first formant frequency is low and the second formant high, e.g. /ι/, there is a peak corresponding to the second formant frequency as well. The histograms of /ι/ and /ʌ/ also show peaks which bear no direct relationships to any formants (/ι/ 300 and 2100 Hz, /ʌ/ 600 and 1250 Hz). These can be seen at 0.9, 1.4 and 1.8 msec for /ι/, (the peak at 1.6 msec corresponds to F1), and at 1.7 msec for /ʌ/.

A set of vowel sounds was spoken by the author and analysed using the CAT computer. The sample length was again 500 msec, but the bin-widths were made 78 μsec. For purposes of presentation the histograms have been normalised, so there is an equal area under each histogram. The bins that correspond to the half-periods of the first and second formants have been marked with arrows. The histograms are shown in Fig. 2.4; the first two formant frequencies of the vowels in Fig. A2.1.

Several features of the histogram are significant for speech synthesis:-

2.5.1. The Form of the Distributions

Distinct peaks are visible in all the vowel histograms, except that of /ɜ/. The number of peaks varies from one (as in /æ/ and /ʌ/) to four (as in /ι/). Where the distributions are uni-modal, the peak is spread over several bins, (typically ten) whereas in the multimodal distributions, each peak is spread over only five or six bins. This is in contrast to the peaks found in the frequency spectra of vowel sounds. There are usually two or three peaks in the frequency range up to 3KHz, corresponding to the first two or three formants. The amplitude of the formants decreases with frequency so that the first formant amplitude is the greatest. No such statement can be made about the peaks in the time-interval histograms.

2.5.2. Frequency-Time Relationship

Although the major peaks in the histograms of Figures 2.3 and 2.4 show a general trend of following the first formant, closer examination of the histograms, particularly those in Fig. 2.4, shows that the peaks in the time-interval histograms are not related to the formant frequencies in a unique way. The fact that some

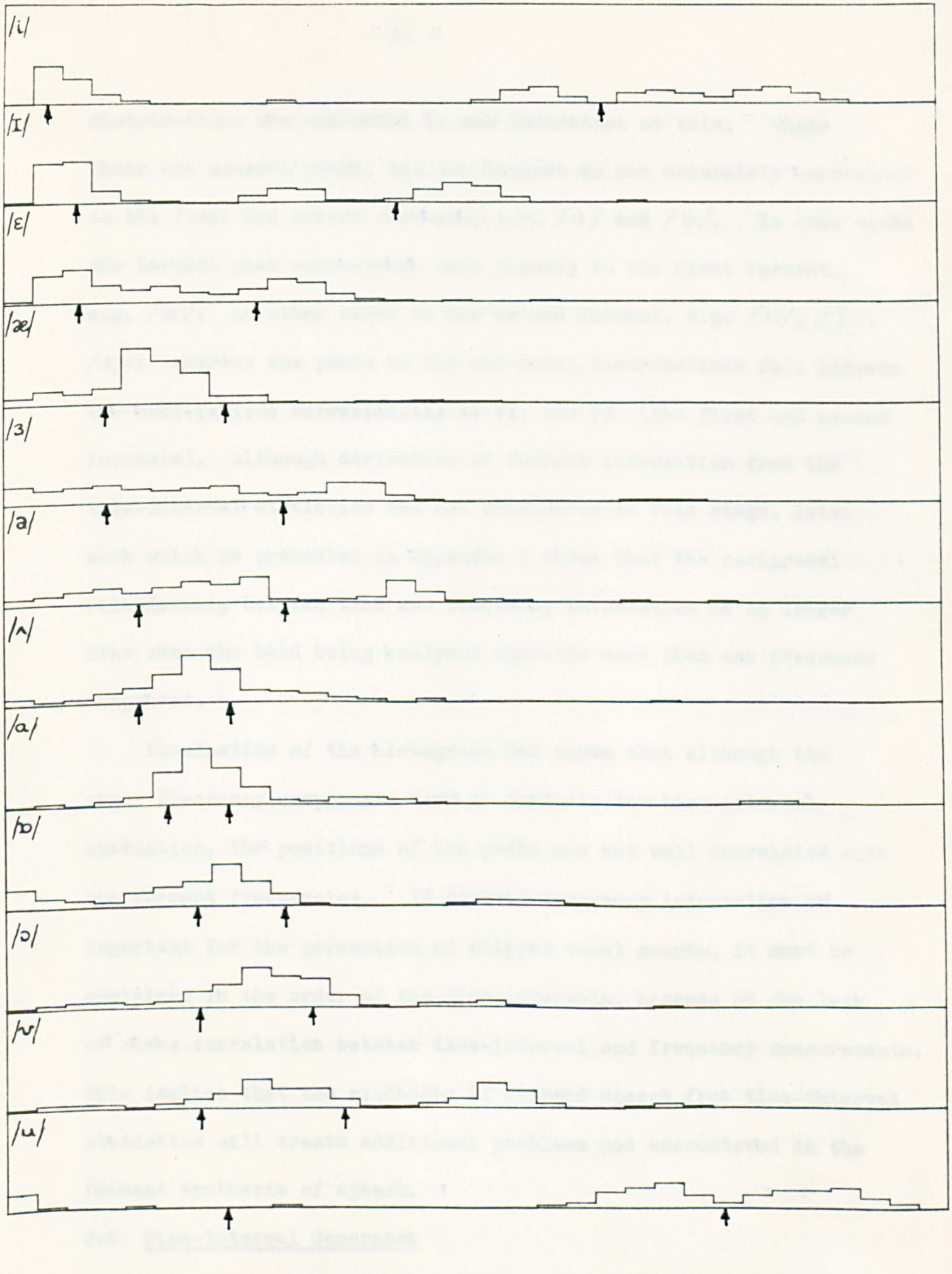


Fig. 2.4 Histograms of Vowel Sounds (Bin divisions 78μS)

distributions are uni-modal is one indication of this. Where there are several peaks, the two largest do not accurately correspond to the first and second formants, e.g. /u/ and /ə/. In some cases the largest peak corresponds more closely to the first formant, e.g. /u/; in other cases to the second formant, e.g. /i/, /I/, /ɛ/; whereas the peaks in the uni-modal distributions fall between the half-periods corresponding to F1 and F2 (the first and second formants). Although derivation of formant information from the time-interval statistics was not considered at this stage, later work which is presented in Appendix 1 shows that the reciprocal relationship between time and frequency information is no longer true when the band being analysed contains more than one frequency component.

Examination of the histograms has shown that although the major frequency components tend to dominate the time-interval statistics, the positions of the peaks are not well correlated with the formant frequencies. If formant frequency information is important for the perception of clipped vowel sounds, it must be contained in the order of the time-intervals, because of the lack of close correlation between time-interval and frequency measurements. This implies that the synthesis of clipped speech from time-interval statistics will create additional problems not encountered in the formant synthesis of speech.

2.6 Time-Interval Generator

As a digital computer was not yet available to produce complex trains of pulses, Professor D.M. MacKay suggested the design of a device which utilised histogram information directly. (The Pattern Playback (8) of the Haskins Laboratories replays directly from frequency spectrum information).

The basic principle of operation is similar to that of the function generator described by MacKay and Fisher (40), in that feedback is used between a photo-multiplier tube viewing a spot on the face of a cathode-ray tube, and the voltage controlling the movement of the spot. In the time-interval generator, however, a time-base circuit moves the spot perpendicularly towards the edge of the mask (see Fig.2.5). Disappearance of the spot behind the mask causes the rapid fly-back of the spot to the other side of the tube. Pulses from the photo-multiplier are used for spot blanking during fly-back and triggering the commutating bi-stable circuit. The time-intervals between pulses are controlled by the sweep-speed of the time-base and the distance the spot has to travel in the X-direction before passing behind the mask. Using a suitably shaped mask and driving the spot in the Y-direction as well, it is possible to produce a sequence of time-intervals of different lengths.

A transistor time-base circuit was designed to give a maximum sweep time of 3 msec when used with a modified Cossor oscilloscope with a short persistence screen. The signal from the bi-stable was passed through a four-diode gate which was controlled by a low frequency astable circuit with a variable mark to space ratio. This enabled isolated sounds of variable duration to be produced.

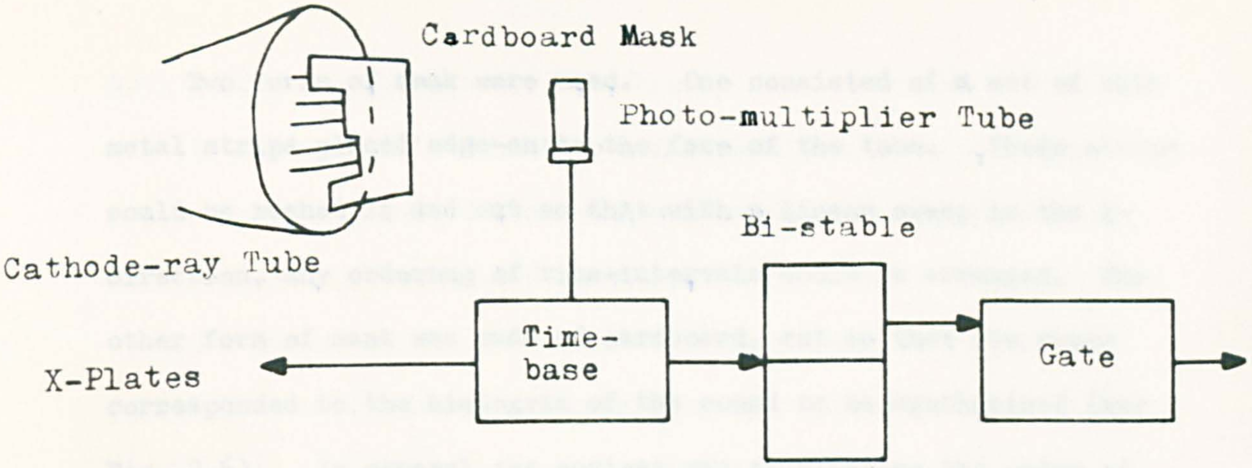


Fig. 2.5 Principle of Time-interval Generator

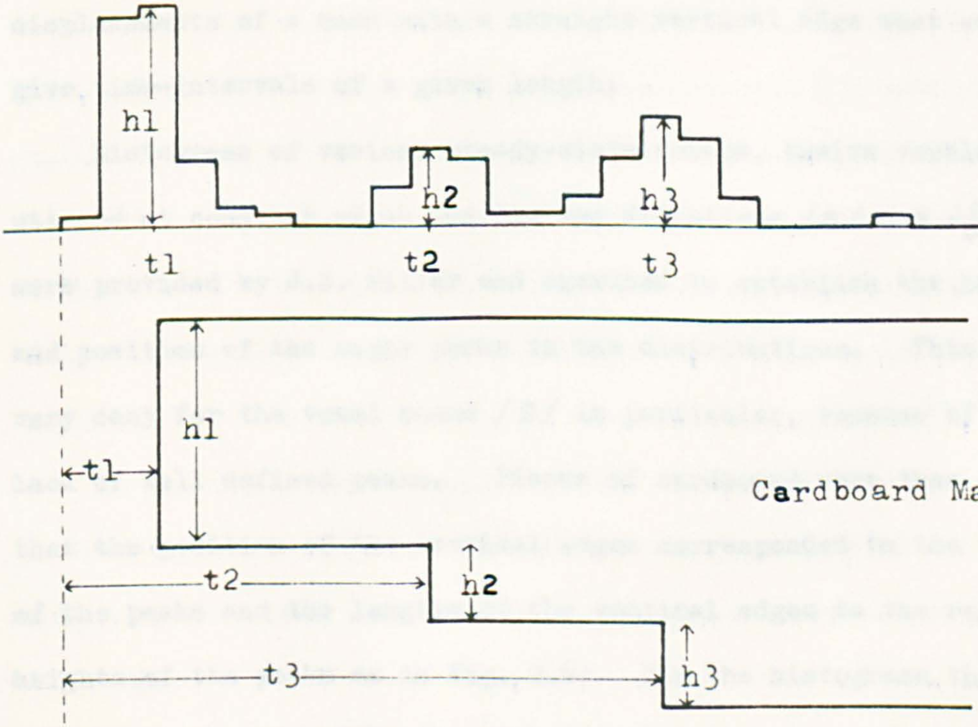


Fig. 2.6 Relationship between Histograms and Masks

Two forms of mask were used. One consisted of a set of thin metal strips placed edge-on to the face of the tube. These strips could be pushed in and out so that with a linear sweep in the Y-direction, any ordering of time-intervals could be arranged. The other form of mask was made of cardboard, cut so that its shape corresponded to the histogram of the sound to be synthesised (see Fig. 2.6). In general the easiest way of altering the order of the intervals was found to be using a fixed shape of mask and varying the form of the Y-driving voltage.

2.6.1. Preparation of Masks

The generator was first calibrated by finding the horizontal displacements of a mask with a straight vertical edge that would give time-intervals of a given length.

Histograms of various steady-state sounds, twelve vowels uttered at constant pitch and the two fricatives /s/ and /ʃ/, were provided by J.B. Millar and examined to establish the height and position of the major peaks in the distributions. This was not very easy for the vowel sound /ɜ/ in particular, because of its lack of well defined peaks. Pieces of cardboard were then cut so that the position of the vertical edges corresponded to the positions of the peaks and the lengths of the vertical edges to the relative heights of the peaks as in Fig. 2.6. For the histograms that contained only one peak, the mask was shaped to allow for the width of that peak.

2.7 Random Excitation

To see whether a first-order approximation to a clipped speech sound bore any perceptual resemblance to the wave from which its histogram had been compiled, low-pass filtered noise was connected to the Y-plates of the generator. It was found that spurious intervals were generated when the spot hit a horizontal edge of the mask. To ensure that the spot always travelled horizontally a fast-running sawtooth waveform was sampled every time the spot disappeared, and applied to the Y-plates (see Fig. 2.7).

This method produced a raucous whispered quality from the bistable, there being no steady triggering corresponding to the glottal frequency. It was very difficult to identify the vowel sounds absolutely, though there were distinct differences between some of them. /I/ and /u/ were easily distinguished because /I/ is composed mainly of short intervals while /u/ is composed mainly of long ones. The differences were not so well pronounced for vowels having less well-defined histograms, particularly the central vowels.

As expected, the fricative sounds /s/ and /ʃ/ sounded more natural than the vowel sounds, because the random method of choosing time-intervals is a better model of the way that fricative sounds are produced.

A fuller evaluation of selecting the intervals at random was left until a digital computer was available, so that comparisons could be made with sounds produced from higher-order statistical

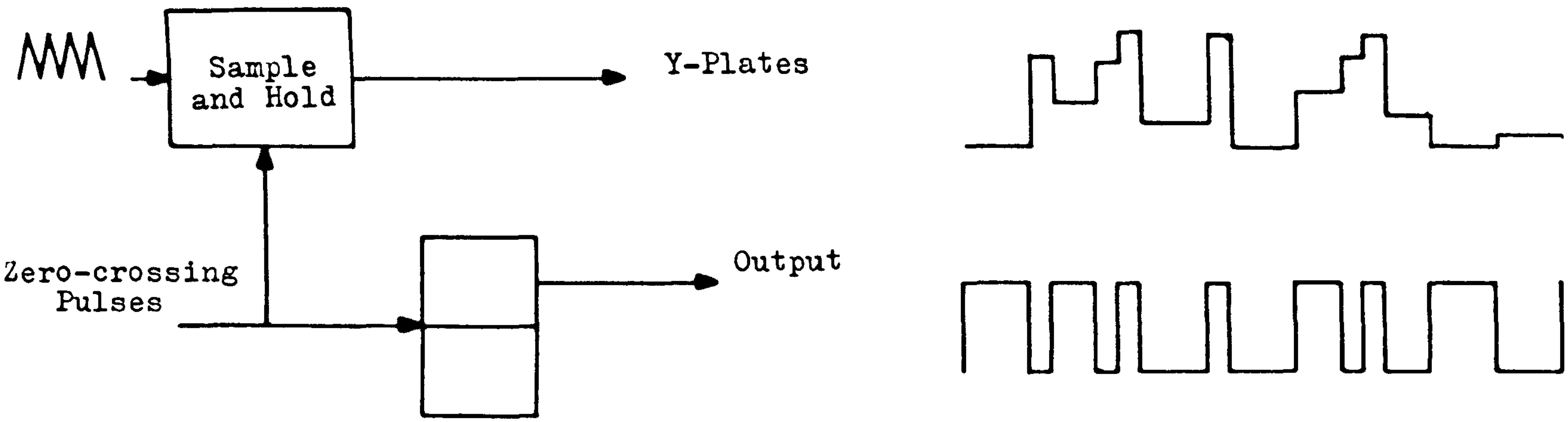


Fig. 2.7 Random Excitation

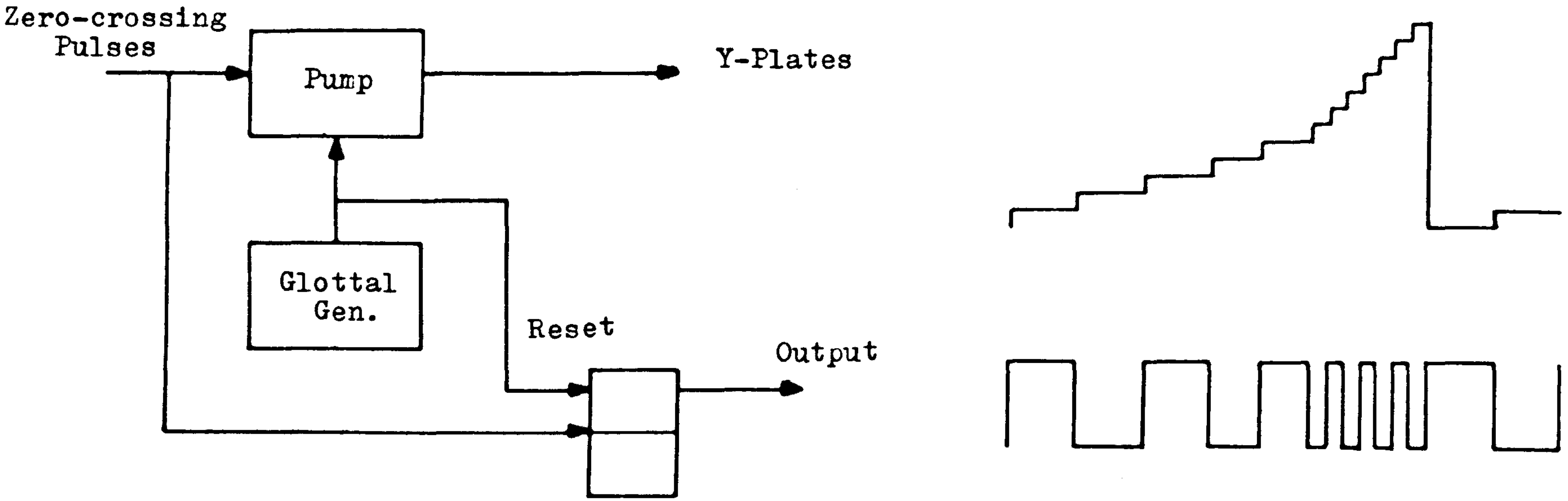


Fig. 2.8 Glottal Excitation

measurements, and non steady-state sounds could also be produced. This is described in Chapter 4.

2.8 Glottal Triggered Excitation

To obtain a better approximation to voiced sounds, it is necessary to choose the intervals in a more deterministic way, and to trigger or excite this pattern at regular intervals. This simulates more closely the behaviour of the vocal tract when it is excited by the vibration of the vocal chords.

To simulate this behaviour, the output of a linear pump circuit was connected to the Y-plates (see Fig. 2.8). Successive pulses caused the beam to be moved upwards in equi-spaced steps. Pulses from a relaxation oscillator (glottal generator) were used to reset the pump circuit at regular intervals and to set the output bi-stable to a '1' state. This ensured that there was an even number of time-intervals in every glottal period. If this was not done, there was a discontinuity in perceived pitch as the glottal generator frequency was changed. Gradually increasing the glottal frequency caused the last time-interval in every glottal period to grow shorter, until it disappeared. If there had been an even number of time-intervals in the glottal period before, there were now an odd number, so that the phase of the output signal changed every glottal period. Consequently the perceived sound underwent a marked discontinuity when the last interval in the glottal period disappeared.

Resetting the bi-stable every glottal period overcame the problem of the phase change, but the change in length of the last interval could still be perceived as the glottal generator frequency was changed. If the bi-stable was already in a '1' state at the end of a glottal period, the reset pulse would not affect it, so that a compound interval was formed from the last interval of one glottal period, and the first interval of the next. This was not a problem for the synthesis of steady-state sounds as a slight adjustment could usually be made in the glottal generator frequency to prevent the formation of compound intervals. It was envisaged that for non-steady state sounds, however, it would be difficult to ensure that compound intervals were not created as the quality of a sound changed.

2.8.1. Selection of Intervals

Adjustment of various parameters of the generator showed that vowel-like sounds could be produced, though the perceptual effects were dependent upon the order in which the intervals were selected. For example, there was a striking difference between the sequence long-short-long-short and the sequence long-long-short-short, though they had the same histogram. It was realised that the order in which the intervals were chosen was of paramount importance in the production of synthetic clipped vowel sounds and should be investigated. A rigorous investigation proved to be impossible in the time available. Consider for example, the histogram of the vowel /I/ in Fig. 2.4. The heights of the three major peaks of that histogram are in the ratio 6:2:3. Suppose that we wish to arrange eleven time-intervals to give us the same histogram, then there are

$$\frac{11!}{6!\,3!\,2!} = 4{,}620$$ different ways in which

these eleven intervals can be arranged. Undoubtedly, at least one of these arrangements would produce an /I/-like sound (the arrangement corresponding to the order of the intervals in the original utterance of /I/) and it is possible that other arrangements would also produce recognisable approximations to /I/.

Three ways of reducing the large number of permutations of time-intervals were considered:-

1. To postulate models for the selection of intervals and apply these algorithms to producing all the vowel sounds from their respective histograms. These algorithms would then have to be modified or discarded according to the intelligibility of the synthetic vowels.
2. The number of permutations could be reduced by producing fewer intervals in each glottal period.
3. The first-order statistics contain no information about the order in which the intervals occurred in the original utterance, so that first-order approximations can produce statistically unlikely combinations of intervals. Second-order statistics giving the transition probabilities of one interval following another would indicate likely pairs of intervals. The measurement of second-order statistics is discussed in Chapter 3.

2.8.1. A Simple Model

A simple model was tried initially. The masks that had been prepared from the histograms had been cut step-wise, so that long intervals came at the bottom of the mask and short ones at the top. When these masks were put into the time-interval generator, sounds were produced that had long intervals at the beginning of each glottal period and short ones at the end. Observation of the original clipped waveforms showed that this ordering of the intervals was by no means typical of natural vowels. Nevertheless, the model was tried to see if it would produce recognisable clipped vowels.

Most of the sounds that were produced could not be identified by the author as vowels;/u/ however was an exception. These findings were confirmed by two other members of the department with a similar amount of experience of listening to clipped vowel sounds.

Narrow-band spectra of these synthetic vowels were produced and three examples are shown in Fig. 2.9, together with narrow-band spectra of the original clipped vowel. In all cases, what can be considered as the first formant in the synthetic vowel corresponds closely in frequency to the first formant in the original clipped vowel. For /i/ and /æ/ however, that is where the similarity ends. The remaining peaks in the spectra of the synthetic vowels do not correspond in frequency with those in the spectra of the clipped vowels. The spectral forms for both versions of /u/ are, however, very similar.

As had been expected, the effect of changing the order of the time-intervals had produced a change in the quality of the sound, to the extent of destroying the individual vowel colour. Any

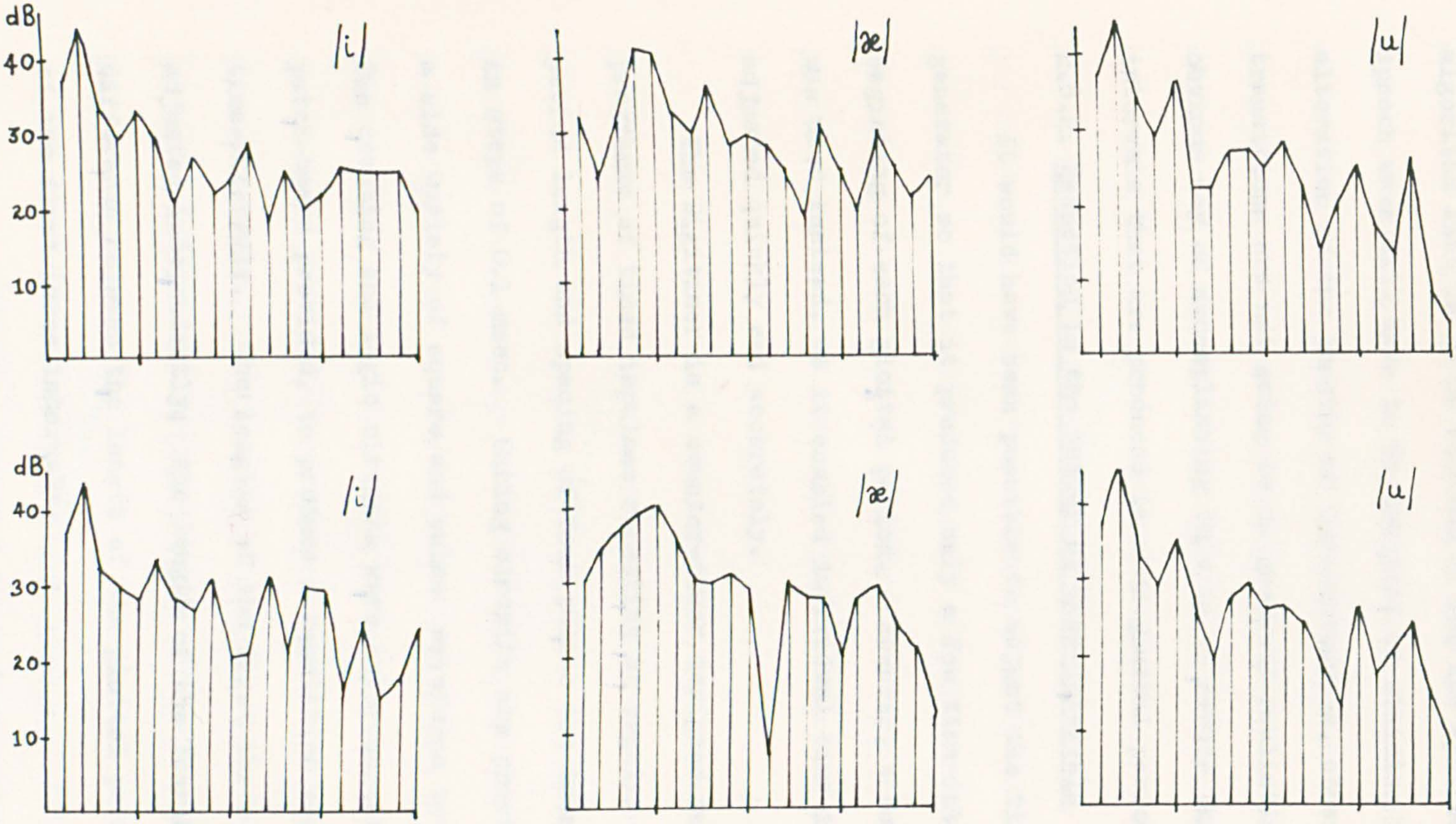


Fig.2.9 Narrow Band Spectra of Clipped and Synthetic Vowels (Fundamental Pitch 170Hz)

Upper row - Clipped Vowels

Lower row - Synthetic Vowels

algorithm that is to be successful for the synthesis of a clipped speech wave will have to be designed to minimise the effects of alteration of the lengths of time-intervals, otherwise formant-transitions are not going to be produced realistically. One obvious way of accomplishing this is to reduce the number of intervals that are produced in each glottal period.

2.8.2. Reduction in the Number of Time-Intervals

It would have been possible to adjust the time-interval generator so that it produced only a few time-intervals at the beginning of each glottal period. However, a Devices Digitimer was used instead, as it enabled individual time-intervals to be adjusted quickly and accurately.

The Digitimer is a counter-timer designed to provide a programme of timed impulses repeating at regular intervals. The period length and spacing of four pulses are independently variable in steps of 0.1 msec. Gating circuits are provided that enable a wide variety of square and pulse waveforms to be generated. The counting and logic circuits were interconnected using the patch-board provided, to produce a repetitive sequence of four time-intervals. The lengths of the first three intervals could be adjusted independently; the length of the fourth interval was the difference between the length of the glottal period and the sum of the first three intervals.

The Digitimer was adjusted to produce each vowel in turn. The intervals were chosen to produce histograms as similar to the original histograms as possible. As there were only three intervals to adjust, the similarity between the original and synthetic histograms was rather poor. The order of the time-intervals was arranged to produce sounds most like the original clipped vowels. The synthetic vowels (500 msec in duration) were recorded in random order and played on two occasions to the author and two members of the department engaged in speech research.

On average five vowels out of the twelve were recognised correctly. /æ/ and /u/ were always identified correctly; /i/, /I/, /ε/ and /ʊ/ on four out of the six presentations. The rest of the vowels, centre and back vowels, were each recognised once or twice only.

Although this was only a preliminary experiment with a small number of subjects, the results were encouraging. They indicated that recognisable clipped vowels could be synthesised from information about the peaks in the time-interval histograms. Furthermore, they indicated that only a few intervals were necessary to produce a recognisable sound. This latter point is significant as it means that the 'end of glottal period effects' described in Section 2.8 when a complete glottal period is produced can be avoided by not producing complete glottal periods of time-intervals.

2.9 Conclusion

There are two levels of speech synthesis from time-interval statistics. The first is the synthesis of steady-state sounds. Although sustained steady-state sounds are unnatural, the production of realistic steady-state sounds indicates that relevant information has been preserved in the statistics. The second level is the production of non steady-state speech sounds from time-interval statistics. Using the apparatus described in this chapter, it was not possible to produce any such sounds.

The method of selecting intervals at random was found to be suitable for the synthesis of fricative sounds and unsuitable for voiced sounds. Encouraging results were obtained when a sequence of a few intervals was repeated at a steady rate. The order of the time-intervals within a glottal period was found to be an important factor in determining the vowel quality of a synthetic vowel. Thus in order to produce a realistic vowel sound, the time-intervals have to be selected in an appropriate order. This shows the major difference between the control of a formant synthesiser and one producing a sequence of time-intervals. Adjustments to the parameters of a formant synthesiser are made in parallel. In time-interval synthesis there is only one parameter to control, (the length of the time-interval), so that the order in which the intervals are selected is all important. The major problem that has to be solved in time-interval synthesis is the development of an automatic method of selecting the intervals to produce realistic sounds.

CHAPTER 3

3.1 Introduction

In 1948 Shannon (56) showed that better approximations to the English language could be generated using higher-order statistics of the symbols (words or letters). By exploiting the informational redundancy in this way, uncommon sequences of symbols were less likely to occur than in the first-order approximation. Shannon used two techniques to generate his approximations. The first relied on published tables of word, letter, digram and trigram frequencies. In the second technique (57) he used the implicit knowledge of the statistics of the language possessed by English-speaking persons. Both of these techniques are unsuitable for generating higher-order approximations to a clipped speech waveform. The rate at which time-intervals occur is too great for humans to possess an inbuilt knowledge. If tables of digram frequencies were to exist, they would reflect the characteristics of the voice or voices used to compile them (22).

In this chapter, apparatus is described for the real-time display of the digram frequencies of time-intervals occurring in speech, and its application to the synthesis of a clipped speech waveform is discussed.

3.2 Second-Order Statistics

The simplest way of taking account of the structural information of successive intervals is to consider the set of transition probabilities $p_i(t_j)$ that a time-interval of length t_i is followed by a time-interval t_j, where i and j range over the whole set of intervals to be considered. An equivalent specification is the set of digram probabilities $p(t_i, t_j)$ i.e. the frequency of occurrence of the pair of intervals t_i t_j. They are related formally by :-

$$\sum_{j=1}^{n} p_i(t_j) = \sum_{i=1}^{n} p(t_i) = \sum_{i,j=1}^{n} p(t_i t_j) = 1$$

where $p(t_i)$ is the probability that an interval of length t_i will occur.

Without a digital computer the task is a formidable one, for if the time-interval space is divided into n quanta, the corresponding digram space has n^2 elements (22).

3.2.1. Form of the Digram Display

Professor D.M. MacKay (41) suggested an analogue solution to this problem, whereby zero-crossing pulses caused the two preceding intervals to be displayed at right angles on a cathode-ray tube.

Analogue voltages proportional to adjacent time-intervals are plotted in the X and Y directions, and accumulated spot brightness is used to represent frequency of occurrence. (cp. Kay Sonagraph where the energy within a given frequency band is represented by the intensity of marking the paper). A similar technique has been used by Fetz and Gerstein (18) to investigate the intervals between neurone pulses.

The first system to be developed employed a parallel time-base system (see Fig. 3.1). Delayed zero-crossing pulses were used to

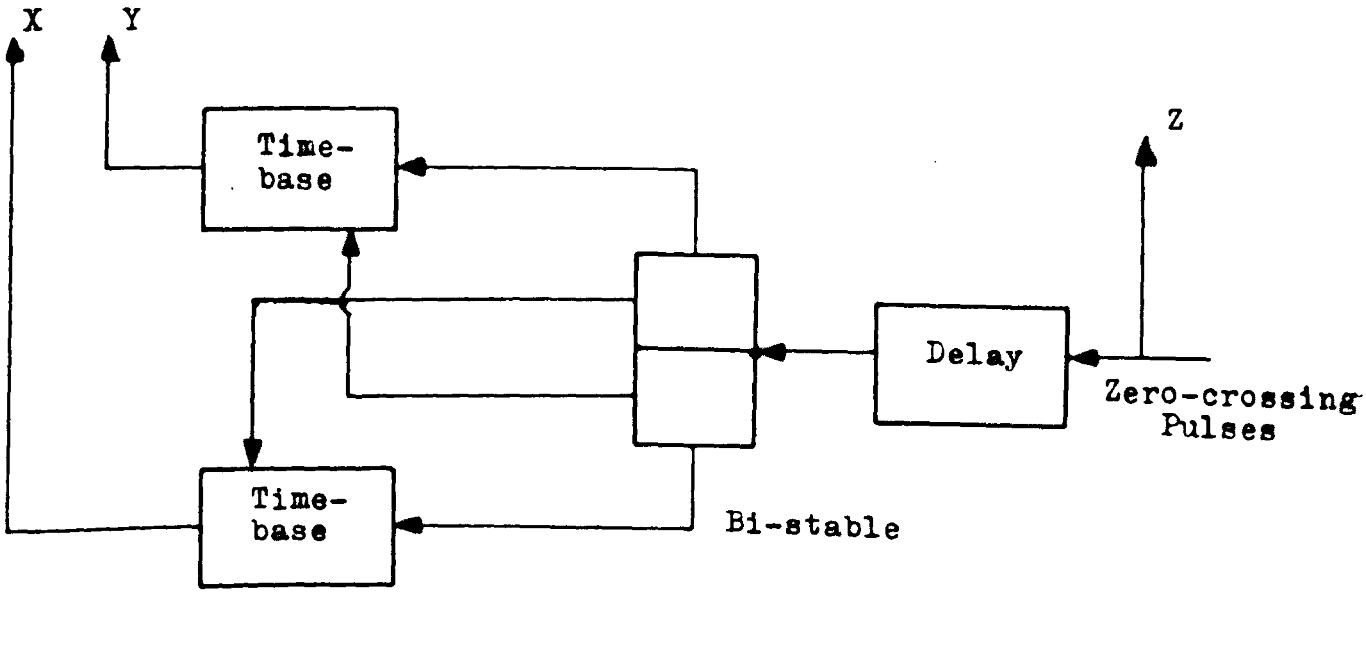


Fig. 3.1 Parallel Time-base System

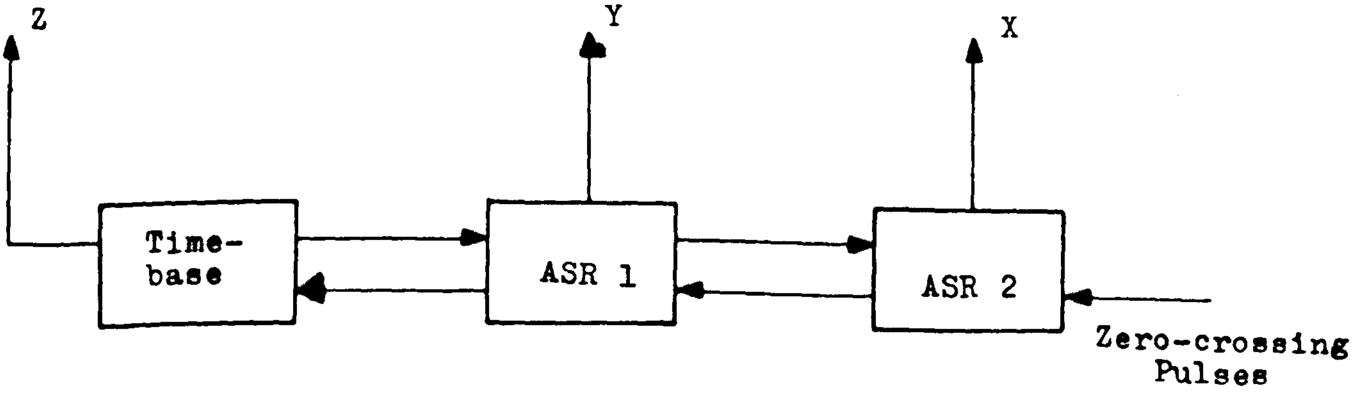
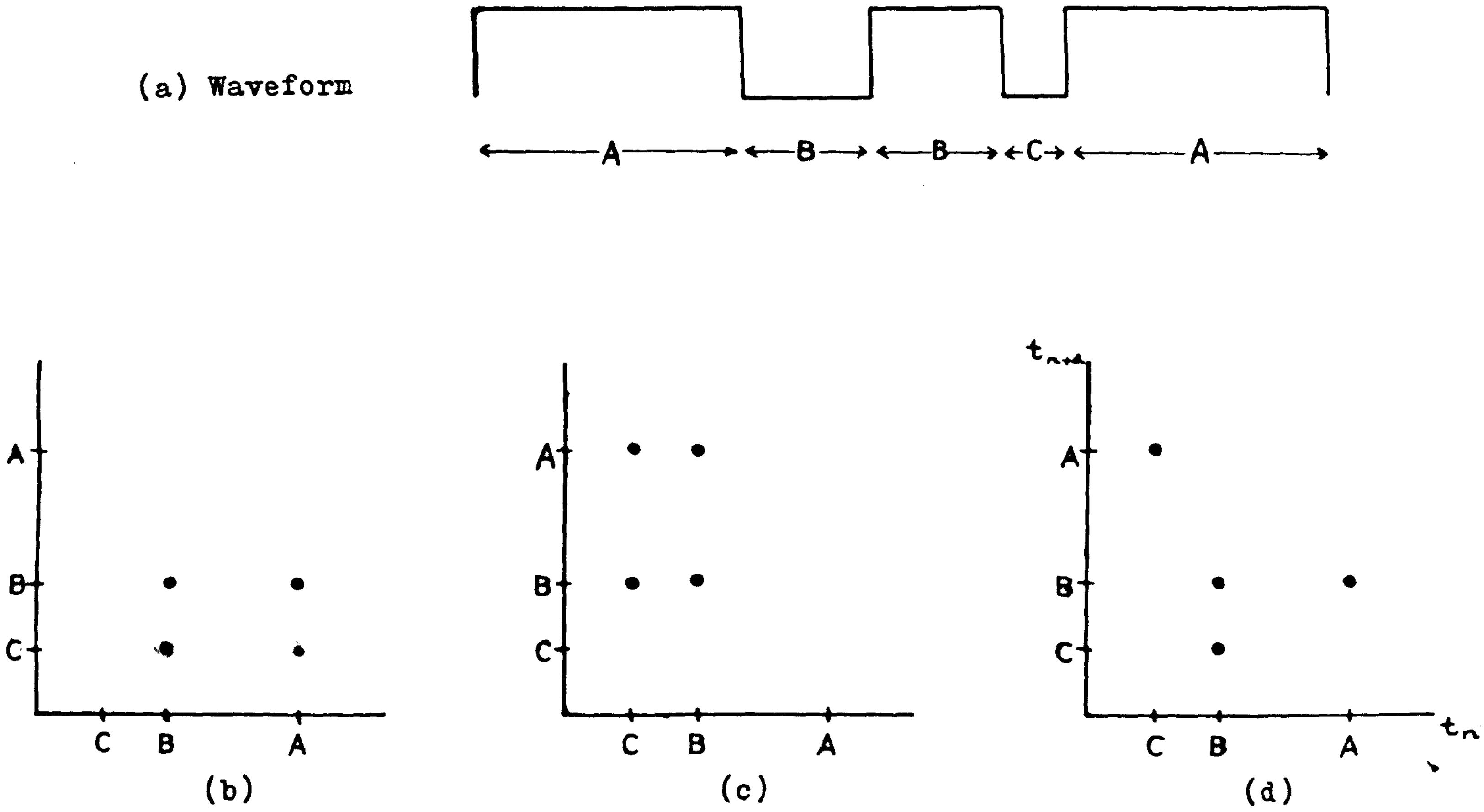


Fig. 3.2 Analogue Shift Register

reset and clamp two time-bases alternately, the two time-bases being run in anti-phase so that one zero-crossing pulse would reset one time-base and clamp the other. The zero-crossing pulse was used to brighten the spot before one of the time-bases was reset. Because of the dead-time associated with the pulse-producing circuitry, the displays had a distinctly symmetrical appearance. For voiced sounds there should be an even number of zero-crossings per glottal period (50). If two zero-crossing pulses come within the dead-time of the circuitry, the second zero-crossing pulse goes undetected, so that there are an odd number of intervals in that period. This causes subsequent points on the display to be plotted with the axes interchanged. This could have been overcome by ensuring that very short time-intervals did not occur, but there were more fundamental deficiencies in the system.

Consider the simple repetitive sequence of time-intervals shown in Fig. 3.3(a). Depending upon which of the time-intervals is chosen as the first one, the display could have either of the forms shown in Fig. 3.3 (b) or (c). Moreover, if the same pattern of time-intervals is encountered in the reverse direction, identical patterns are obtained. This arises because the display is plotting even-numbered time-intervals against odd-numbered ones (numbering the intervals t_1, $t_2 \ldots t_n$). Consequently the display is dependent upon which interval is chosen as the first one, and is independent of the direction in which the pattern is examined.



(a) Waveform

Fig. 3.3 Display of Waveform (b) & (c) using Parallel System

(d) using Analogue Shift Register

Although this form of display contains sequential information about the time-interval pattern, it was considered unsuitable for providing sequential information for use by the time-interval generator, because of the uncertainty about the absolute relationship between the intervals.

3.2.2. Analogue Shift Register

In order to overcome the disadvantages of the parallel time-base system, it is necessary to use a shift-register technique whereby the first interval of every pair is always plotted in the same direction.

Consider voltages proportional to t_n and t_{n+1} (the n^{th} and and $n+1^{th}$ time-intervals) to be stored on capacitors C_x and C_y respectively (see Fig. 3.4). In order to shift the information when the next zero-crossing pulse arrives, S_x must close to allow C_x to take up the voltage on C_y before S_y closes to allow C_y to take up a voltage proportional to t_{n+2}, the next time-interval in the sequence. In this way, the first time-interval of a pair is always represented by the voltage on C_x, and any symmetry in the resultant display is a property of the signal (42). The first-order information can be recovered by projecting the points of the display onto either axis.

Considerable effort was devoted to devising simple circuitry which would propagate the switching information from right to left, while the stored information is transferred from left to right. The final form of the circuit is shown in Fig. 3.6. Transistors T3 and T4 comprise a sample and hold circuit (Fig. 3.5) and constitute a

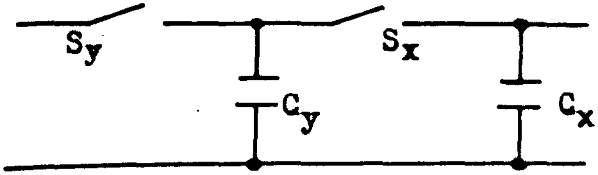


Fig. 3.4 Principle of Analogue Shift Register

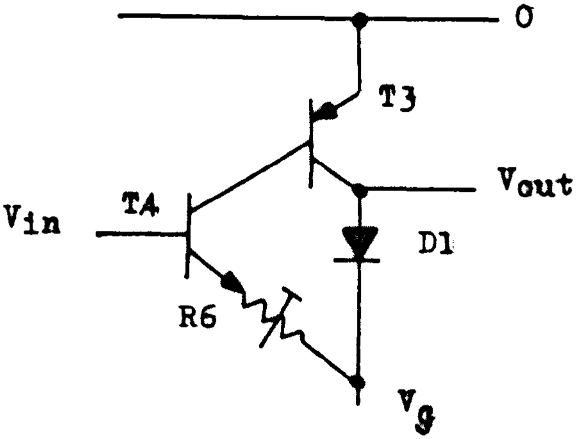


Fig. 3.5 Sample/Hold Circuit

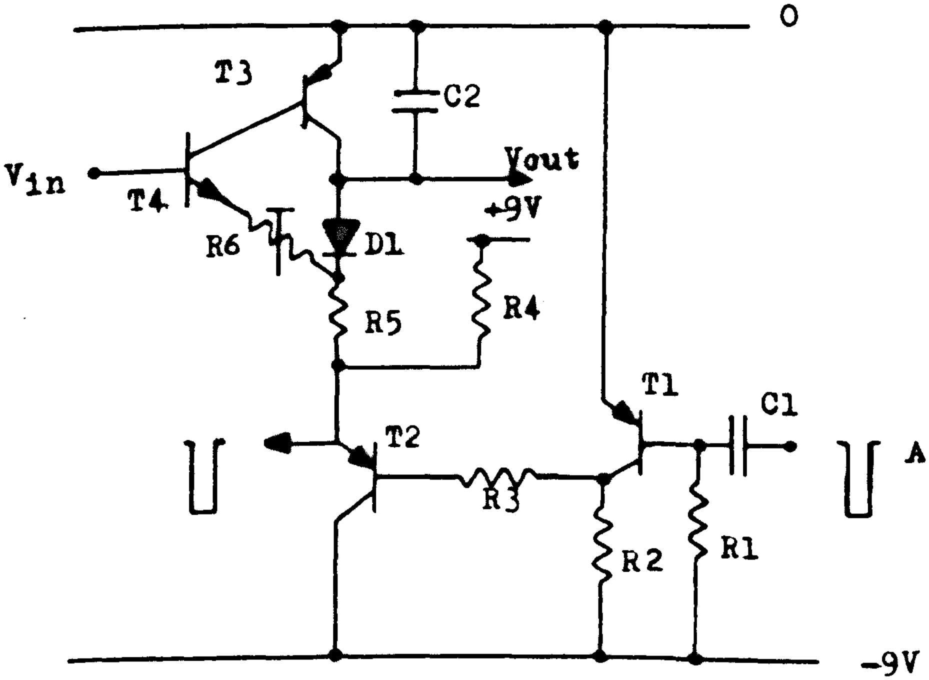


Fig. 3.6 Basic Stage of Analogue Shift Register

switched emitter-follower (37). The impedence gain from input to output for the silicon transistors used is of the order of 10,000. R6 is adjustable to compensate for the difference in voltage drop across the base-emitter junction of T3 and the diode D1 when they are both conducting. When V_g is held slightly above earth potential, T3, T4 and D1 are all reverse-biased presenting high impedence to both input and output. When V_g is made negative, T3, T4 and D1 conduct and the voltage on the input is transferred to the output.

Negative-going pulses representing zero-crossings enter at A, the trailing edge turning T1 off for a time determined by R1 and C1 (15μs). Thus a delayed version of the input pulse is produced at the collector of T1 and drives both the delay circuit of the next stage and the sample/hold circuit, via the emitter-follower T2. The voltage on the preceding stage is therefore transferred onto C2 before the preceding stage itself takes up a new voltage. Delayed pulses from A.S.R.1 (Fig. 3.2) are used to reset a simple time-base circuit and brighten the display. The usual form of time-base employed was logarithmic, so that the whole range of speech sounds could be displayed equally well.

3.2.3. Performance of the Digram Display

The complete system (Fig. 3.2) was tested using a sequence of time-intervals produced by the time-interval generator, and was found to be accurate to better than 10%. It was subsequently tested using a computer-generated pattern that produced all points in a 16 by 16 array. The accuracy was then found to be 3% of a full scale deflection corresponding to a 3msec sweep. Licklider (36) showed that the intelligibility of differentiated and clipped speech fell from 96% to 91% when the quantising rate of the clipped signal was decreased from 20 to 10 kiloquants per second. Thus an error of 100μsec in the measurement of time-intervals has a marginal effect on intelligibility. An error of 100μsec in 3msec represents a 3% error so that the performance of the display was considered adequate for the display of speech time-intervals.

3.3 Results obtained with the Display

When used "on-line" the display proved to be a sensitive indicator of tongue and lip movements, and characteristic patterns were observed on the screen. The real time nature of the display meant that the effects of variation of speech parameters (e.g. vowel quality and pitch) could be examined and evaluated rapidly.

Some examples of typical digrams are shown in Figures 3.7, 3.8, 3.9 and 3.10. The digrams in these photographs were not obtained using the circuitry just described because of the difficulty of placing axes in the correct positions on the photographs. They were obtained using a general purpose computer (see next chapter). Two output registers were loaded with numbers corresponding to adjacent time-intervals. The registers drove digital to analogue convertors, the outputs of which were used to produce X and Y deflections on an oscilloscope. The computer was programmed to draw the axes as well. The time-base used was linear, and unless specified otherwise, the length of the axes corresponds to 3 msec.

3.3.1. Vowel Sounds

Figures 3.7 and 3.8 show digrams obtained for the set of twelve vowel sounds that were used in the experiments to be described in Chapters 4 and 5. Fig. 3.7 relates to clipped vowels, Fig. 3.8 to differentiated and clipped vowels.

3.3.2. Normal Vowels

One of the most striking features is the similarity between the digrams of adjacent vowels, particularly those in the central region. The vowels have been arranged to correspond fairly well to their positions in the F1/F2 plane, so the similarity between adjacent vowels indicates the dependence of the time-intervals upon the first two formant frequencies. /i/ and /u/ both show the presence of long-long pairings of intervals, but the high second formant frequency (2100 Hz) of /i/ causes the short-short pairing not present in /u/. Passing from /i/ to /æ/ (F1 increasing, F2 decreasing) the well spaced pattern of /i/ is progressively contracted to produce the compact digram of /æ/. Similarly passing from /u/ to /ɒ/ there is a movement of the major points towards the axes, corresponding to an increasing first formant frequency.

All the vowels show definite clusters of points corresponding to repetitive triggering of the same zero-crossing pattern. Small amounts of noise or small perturbations in the glottal frequency cause changes in the zero-crossing pattern. Such perturbations are responsible for the four diffuse areas on the digram of /u/. When the waveform of /u/ was examined, it was found that there were usually

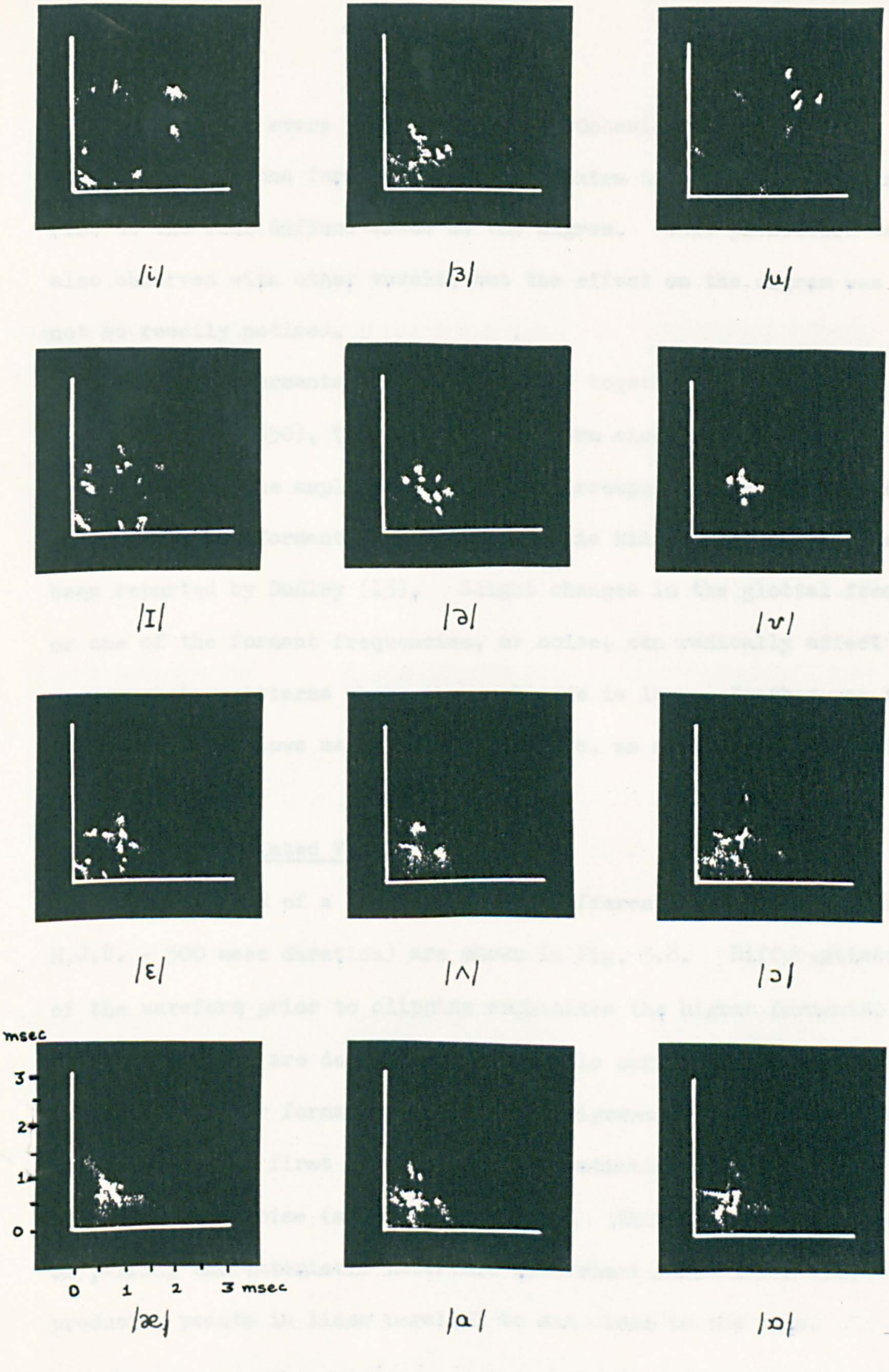


Fig. 3.7 Digrams of Vowel Sounds

six intervals in every glottal period. Occasionally, however, lobes of the second formant would cause extra zero-crossings giving rise to the four diffuse areas on the digram. This phenomenon was also observed with other vowels, but the effect on the digram was not so readily noticed.

Where the formants were fairly close together, e.g. /ɔ/ (F1 = 550, F2 = 850), the original waveform showed "beat-like" modulations of the amplitude at a rate corresponding to the difference between the two formant frequencies. This kind of phenomenon has been reported by Dudley (13). Slight changes in the glottal frequency or one of the formant frequencies, or noise, can radically affect the zero-crossing patterns where the amplitude is low. In this way the digram of /ɔ/ shows many scattered points, as well as several well-defined ones.

3.3.3 Differentiated Vowels

The diagrams of a complete set of differentiated vowels (speaker M.J.U. - 500 msec duration) are shown in Fig. 3.8. Differentiation of the waveform prior to clipping emphasises the higher formants, so that the digrams are dominated by intervals corresponding to the second and higher formants, whereas the digrams of flat speech are dominated by the first formant. Differentiation also emphasises high-frequency noise (as opposed to hum). This will have the effect of pairing characteristic intervals with short noise intervals, producing points in lines parallel to and close to the axes.

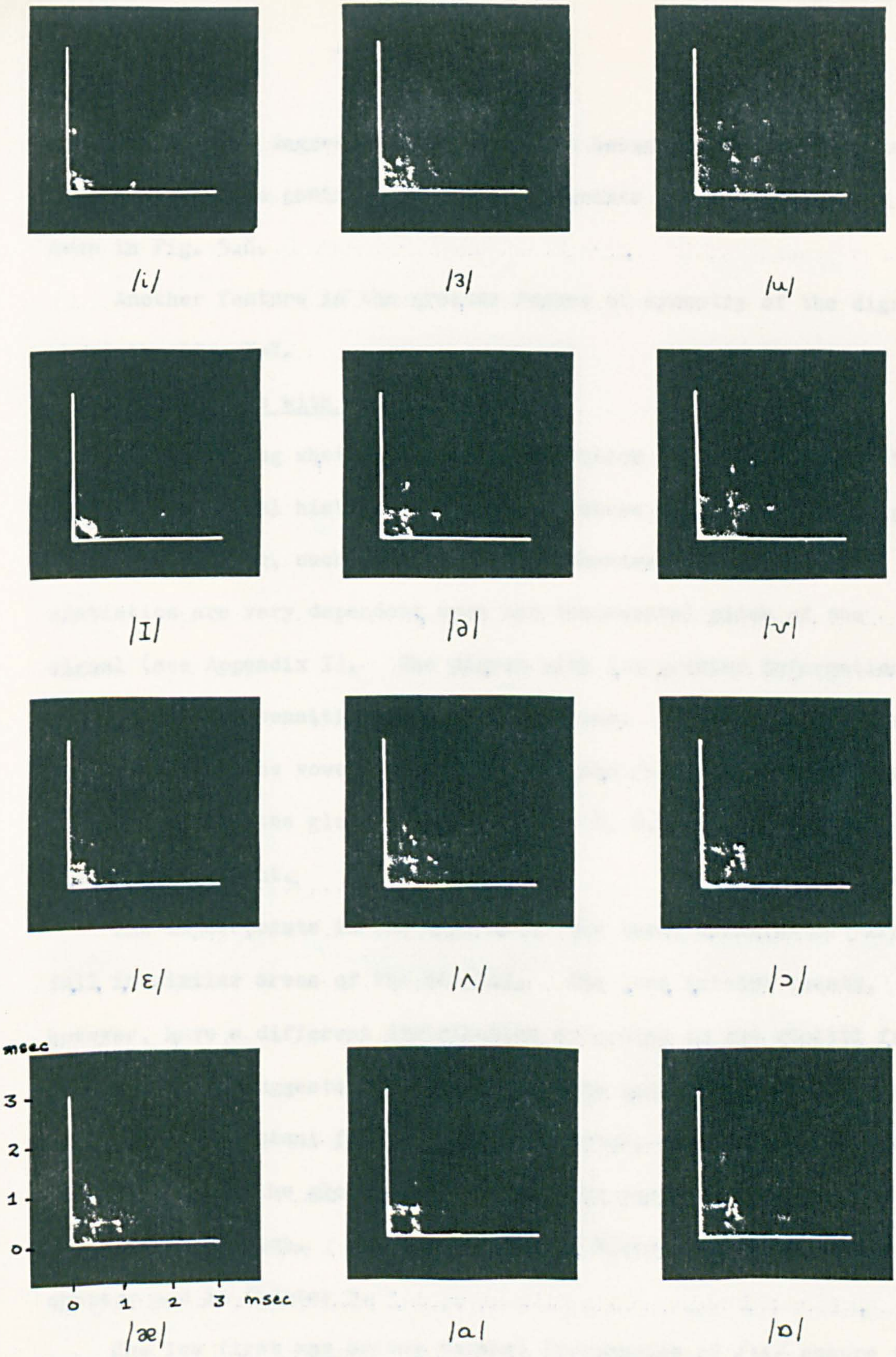


Fig. 3.8 Digrams of Differentiated Vowel Sounds

The general degree of correspondence between the second formant frequency and the position of the major points on the digrams can be seen in Fig. 3.8.

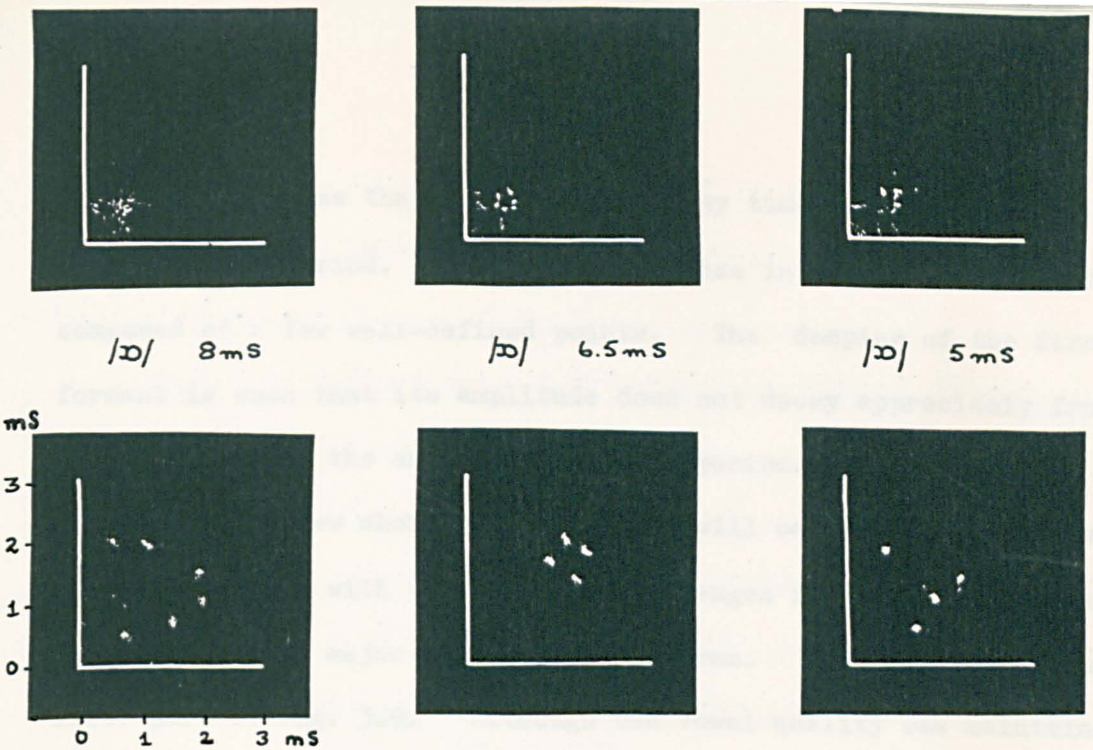
Another feature is the greater degree of symmetry of the digrams about the line X=Y.

3.3.4 Variations with Glottal Frequency

In discussing whether formant information can be reliably obtained from time-interval histograms, it can be shown that even for a closely controlled source, such as a formant synthesiser the first-order statistics are very dependent upon the fundamental pitch of the signal (see Appendix I). The digram with its greater information content is more sensitive to such variations. Two examples are shown in Fig. 3.9. The vowels shown are /ɒ/ and /u/ (duration 500 msec - speaker J.B.M.) the glottal periods being 8, 6.5 and 5 msec reading from left to right.

The major points in the digrams of the three examples of /ɒ/ fall in similar areas of the display. The less intense points, however, have a different distribution according to the glottal frequency. This suggests that there are some intervals which remain more or less constant for variations in pitch, and others, particularly near the end of the glottal period that are very sensitive to the variations in pitch. This point will be discussed later in this chapter and in Chapter 5.

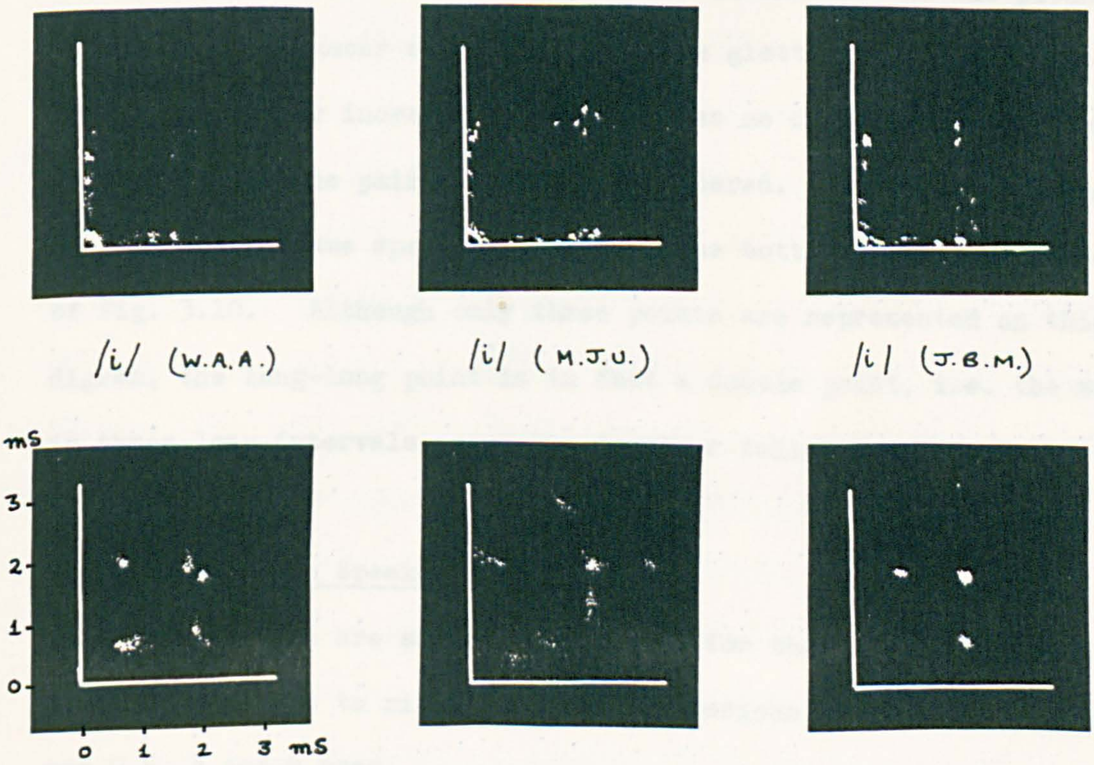
The low first and second formant frequencies of /u/ ensure that the time-intervals between zero-crossings of the wave are fairly long.



/ɒ/ 8mS /ɒ/ 6.5mS /ɒ/ 5mS

/u/ 8mS /u/ 6.5mS /u/ 5mS

Fig. 3.9 Digrams of the Same Speaker (J.B.M.)



/i/ (W.A.A.) /i/ (M.J.U.) /i/ (J.B.M.)

/u/ (W.A.A.) /u/ (M.J.U.) /u/ (J.B.M.)

Fig. 3.10 Digrams of Different Speakers

This in turn means that there are not many time-intervals in every glottal period. The digram of these intervals is therefore composed of a few well-defined points. The damping of the first formant is such that its amplitude does not decay appreciably from the beginning to the end of the glottal period. Consequently there are very few short intervals that will be affected by perturbations of the waveform, with the result that changes in pitch will cause movements of the major points of the digram. This is seen in the lower part of Fig. 3.9. Although the vowel quality was maintained as constant as possible, the effect of changing the pitch of the voice has caused appreciable changes in the digram pattern. For the lowest pitch six distinct points are observed. As the pitch is increased, the number of intervals in the glottal period falls to four. A further increase in pitch causes no change in the number of intervals, but the pairing of them is altered. A fourth utterance of /u/ by the same speaker is seen in the bottom right-hand corner of Fig. 3.10. Although only three points are represented on this digram, the long-long point is in fact a double point, i.e. the sequence is three long intervals occurring together followed by a shorter interval.

3.3.5 Speaker to Speaker Variations

Two examples are shown in Fig. 3.10 for the vowels /i/ and /u/. Reading from left to right, the glottal periods of the three speakers are 9.5, 8 and 6 msec.

There is very little difference to be observed between the three utterances of /i/. The digrams all show pairings of short with short and short with long. Only in the case of the speaker with the lowest pitch is there any absence of a steady long-long pairing.

There are marked differences between the digrams of /u/ however. This is to be expected in view of the gross differences between four utterances of the sound by the same speaker.

3.4 Limitations of the Display

As the digram display was a sensitive real-time indicator of articulatory movements, it was considered worthwhile to investigate the usefulness of the display as a visual display of speech for the deaf. It is not the object of this thesis to evaluate the usefulness of the digram display as a tool for speech analysis per se or as a visual display of speech (43,45). Nevertheless, it is worthwhile examining some features of the display inasmuch that they show the limitations of time-interval analysis of speech. It is important to consider these limitations as they may influence the way in which the statistics are used for speech synthesis.

Using the display "on-line", some vowels revealed very steady patterns, others were subject to more variation. In general,speakers with high pitched voices produced steadier displays of steady vowel sounds than speakers with lower pitched voices. Examination of the speech and clipped speech waveforms showed that the time-interval pattern was least steady in those regions where the amplitude of the speech was low. This generally occurred at the end of each glottal period, but not invariably so. The damping of the formants is such that for speakers with low pitched voices, there is an appreciable length of time at the end of a glottal period where the amplitude of the wave is low. Consequently, slight changes in pitch, or the presence of noise can cause quite severe changes in the time-interval pattern. Because of the clipping process, the digram display attaches equal weight to the intervals irrespective of how reliable they may be.

Moreover, there was a lot of information displayed so that it was not always easy to follow the movements of the major points on the display. It was felt that it would be advisable to reduce the number of points displayed and increase the reliability of those remaining.

Several methods exist for suppressing or altering the time-interval patterns in the regions where the speech wave is of low amplitude.

In his classic experiments on the intelligibility of clipped speech Licklider (35) used a high frequency bias to suppress clipped noise in the silences between words. By adjusting the bias level so that it determined the time-intervals during noise, and choosing the frequency of the bias to be above the upper limit of hearing, his subjects heard nothing during the inter-word silences. A similar method was used by Tanaka and Okamoto (62). If this method were used with the digram display many very short intervals would be generated where the speech amplitude was low, so that these points would be displayed in lines close to and parallel to the axes.

Davenport (9) used a low frequency (20 Hz) square wave bias which carried the speech wave first positive and then negative with respect to the zero-axis. This is equivalent to clipping the wave away from the zero-axis and would certainly suppress clipped noise during silence. However, extra zero-crossings would occur where the bias wave crossed the axis, and moving the clipped level away from zero does not prevent the generation of spurious intervals where the wave crosses the new clipping axis.

Bezdel (3) employed a trigger circuit with a controlled amount of hysteresis, so that a zero-crossing was signalled only when the waveform came out of the threshold region. His threshold levels were presumably chosen only to eliminate noise during the absence of a signal. It was considered that to remove the effects of small perturbations of the waveform, the threshold levels would have to be set so far from the zero axis that the zero-crossings would no longer be accurately recorded.

3.4.1 Amplitude - modulated Digram Display

A solution proposed by M.M. Taylor was to utilise the information discarded in the clipping process, by modulating the spot brightness with a voltage proportional to the peak to peak amplitude of the waveform during the intervals currently being displayed. Fig. 3.11 shows the principle of operation. This method has the advantage that the time-interval pattern is accurately preserved, time-interval pairs occurring where the speech amplitude is greater being given greater weighting than those occurring where the speech amplitude is less.

Fig. 3.12 shows a block diagram of how the idea was implemented. (This work was done in collaboration with J.B. Millar). Analogue

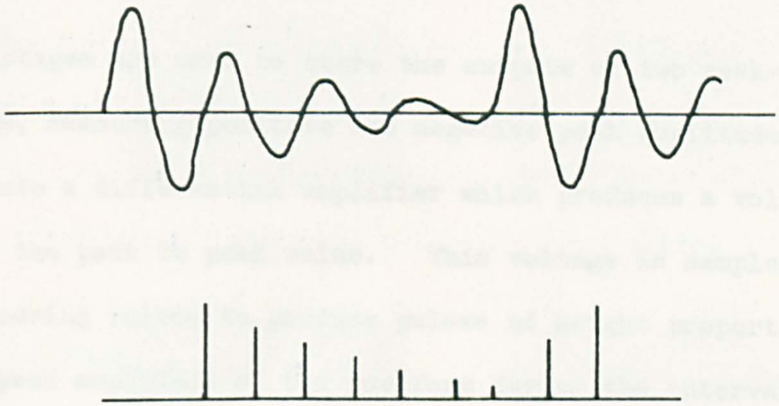


Fig. 3.11

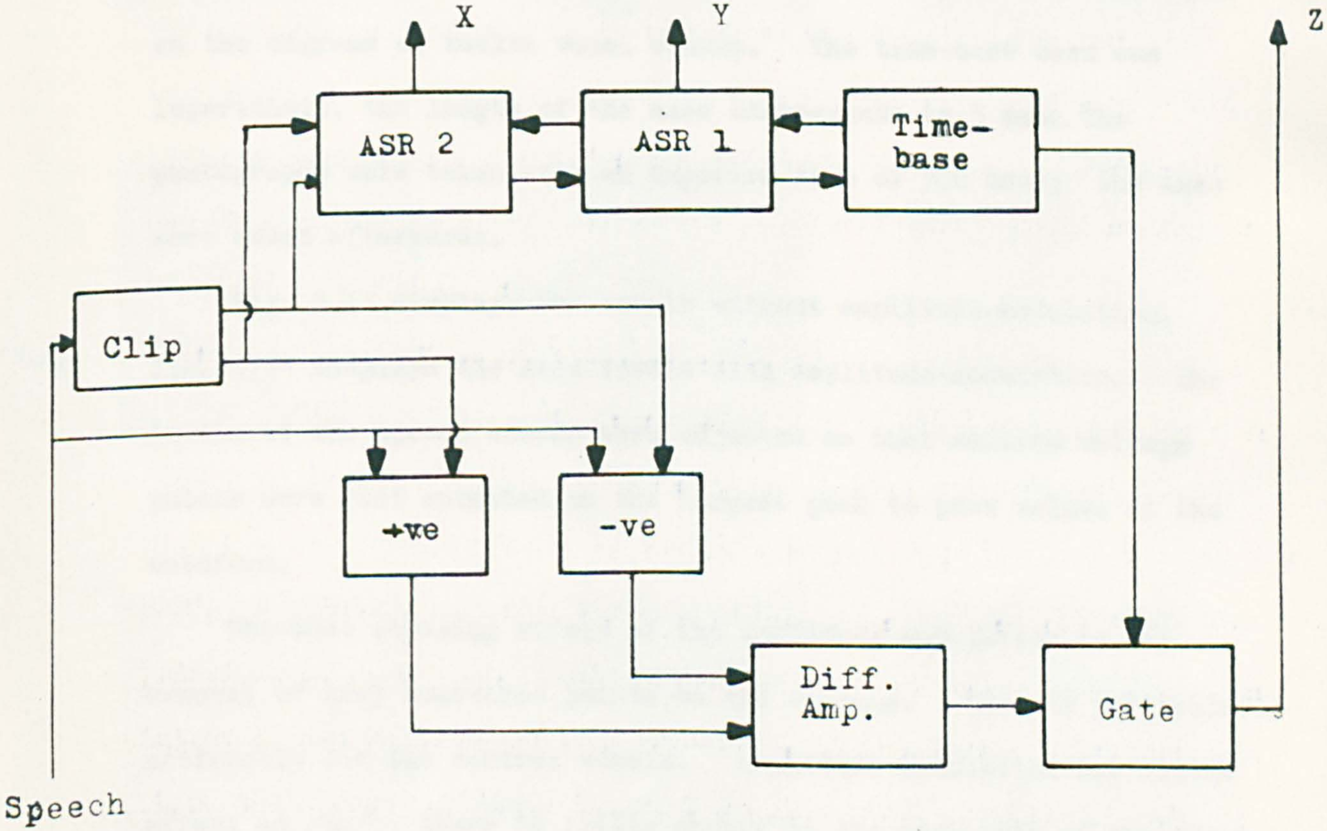


Fig. 3.12 Block Diagram of Amplitude-modulated Digram Display

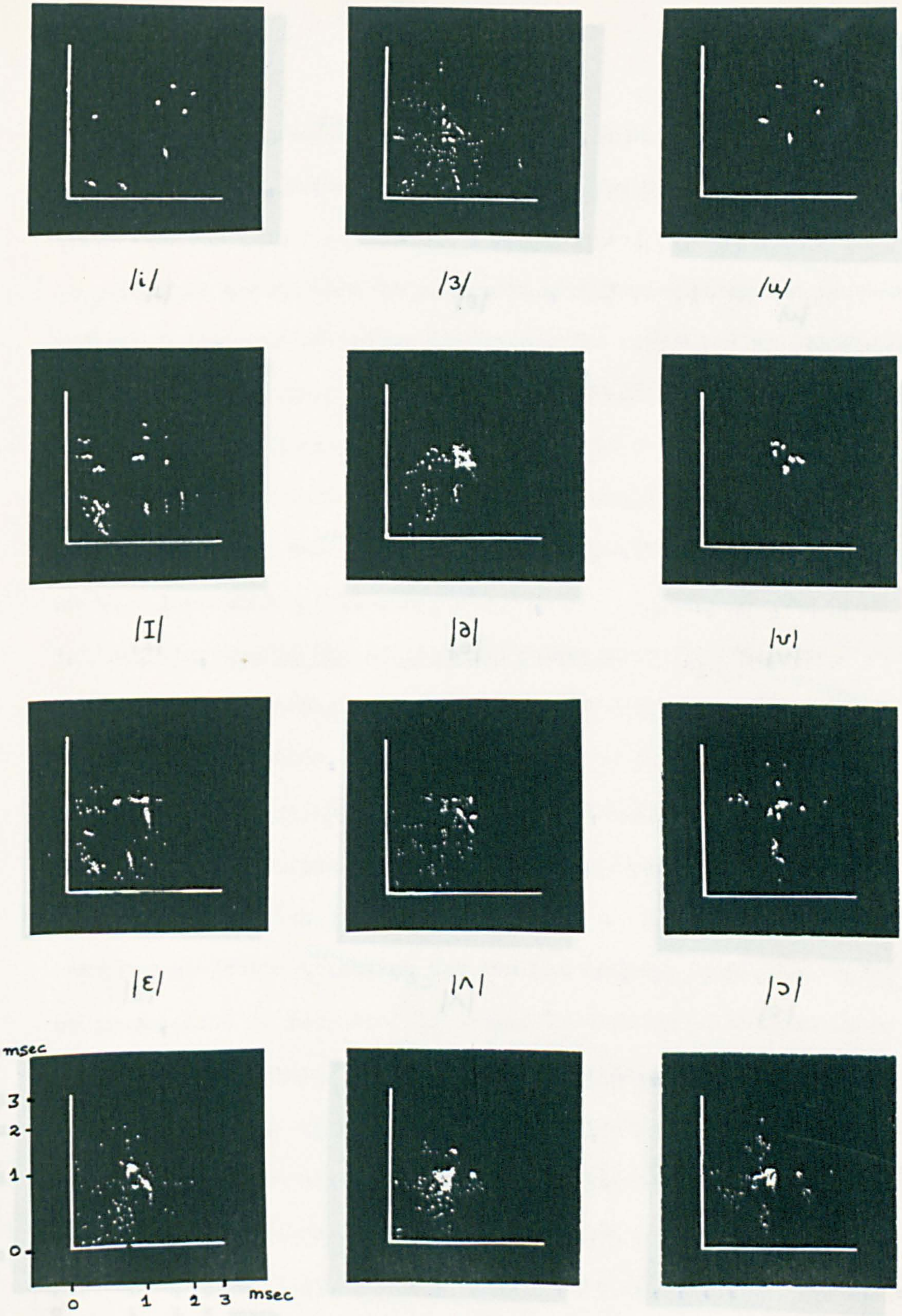
shift register stages are used to store the outputs of two peak-picking circuits, measuring positive and negative peak amplitudes. These are fed into a differential amplifier which produces a voltage proportional to the peak to peak value. This voltage is sampled by delayed zero-crossing pulses to produce pulses of height proportional to the peak to peak amplitude of the waveform during the intervals currently being displayed.

3.4.2 Results

Figures 3.13 and 3.14 show the effects of amplitude-modulation on the digrams of twelve vowel sounds. The time-base used was logarithmic, the length of the axes corresponds to 3 msec. The photographs were taken with an exposure time of 500 msec; the axes were added afterwards.

Fig. 3.13 displays the vowels without amplitude-modulation, Fig. 3.14 displays the same vowels with amplitude-modulation. The levels of the speech sounds were adjusted so that maximum voltage pulses were just obtained on the largest peak to peak values of the waveform.

The most striking effect of the amplitude-modulation is the removal of many scattered points on the digrams. This is particularly noticeable for the central vowels. Amplitude modulation has little effect on /u/; there is little change in the amplitude of the first formant during a glottal period (for speaker M.J.U.). Amplitude-modulation has, however, removed points in the digram of /i/ corresponding to the short-intervals, although these are a fairly steady feature of



/i/ /ɜ/ /u/

|I| |ə| |ʊ|

|ɛ| |ʌ| |ɔ|

|æ| |ɑ| |ɒ|

Fig. 3.13 Digrams of Vowels without Amplitude-modulation

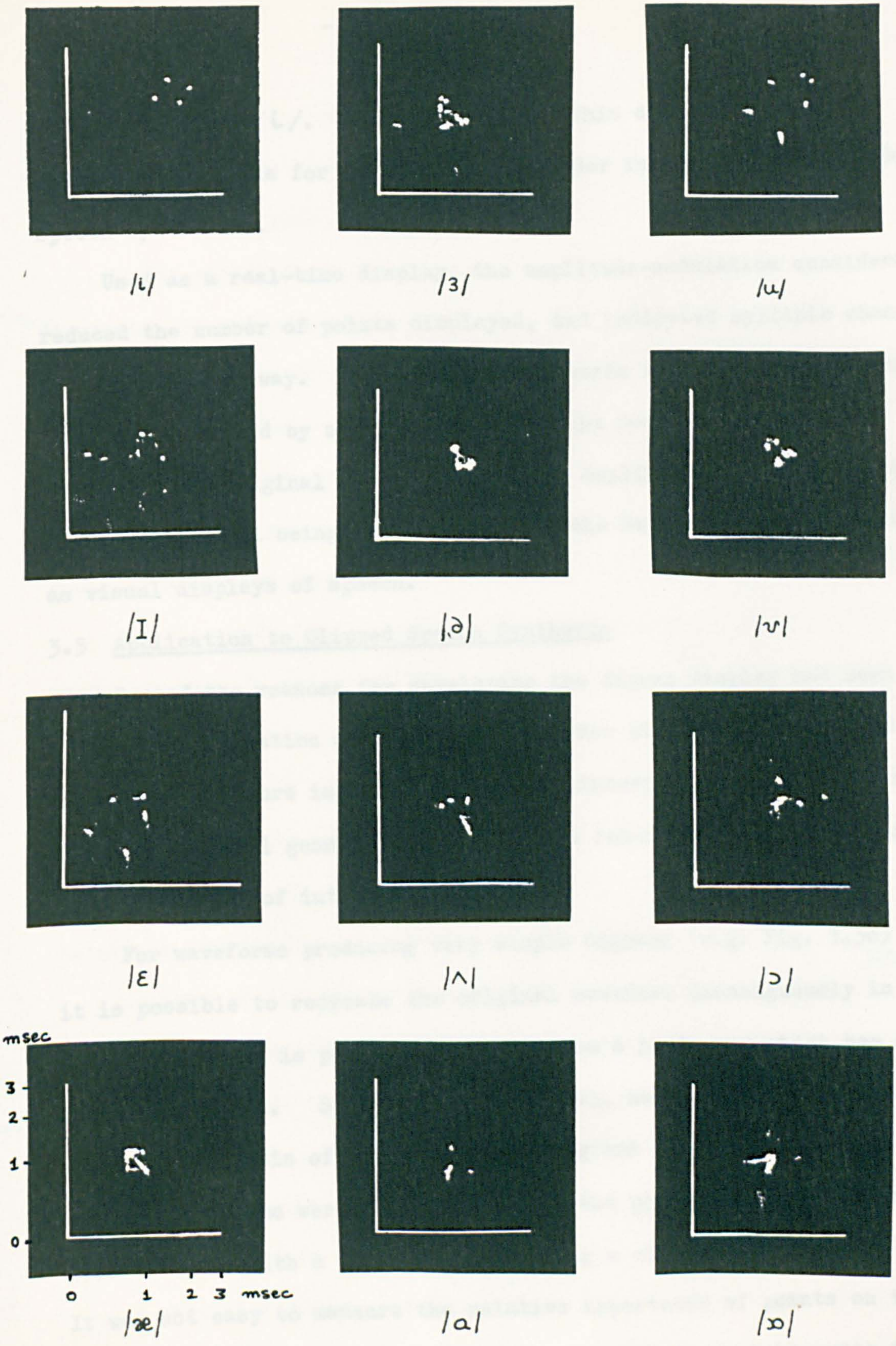


Fig.3.14 Digrams of Vowels with Amplitude-modulation

of the digram of /ι/. (For this reason, this display was not considered suitable for deriving second-order information for clipped speech synthesis).

Used as a real-time display, the amplitude-modulation considerably reduced the number of points displayed, and indicated syllabic changes in a far clearer way. For example, any words containing stop sounds were characterised by bright bursts of light near the origin.

Both the original digram display and amplitude-modulated digram display are still being evaluated within the Department of Communication as visual displays of speech.

3.5 Application to Clipped Speech Synthesis

One of the reasons for developing the digram display had been to derive information about the second-order statistics of clipped speech so that more information could be incorporated into the control of the time-interval generator. This would reduce the very large number of permutations of intervals.

For waveforms producing very simple digrams (e.g. Fig. 3.3c) it is possible to recreate the original waveform unambiguously in the same way that it is possible to do so from a histogram which has only one occupied bin. Such cases are trivial, however, when compared with the synthesis of sounds from the digrams of real speech.

Two problems were encountered when the photographs of digrams were analysed with a view to synthesising a clipped speech wave. It was not easy to measure the relative importance of points on the display because of the limited range of contrast available with the photographic paper. Secondly, for some vowels the points formed continuous bands on the photographs. This could be overcome to a certain extent by reducing the exposure time of the photographs.

Having established the important features, it was necessary to programme the time-interval generator accordingly. As the time-interval generator was essentially a first-order machine, the second-order constraints had to be incorporated either by altering the mask shape or the voltage driving the spot in the Y-direction. Without complex function generators it was not possible to choose intervals randomly according to the digram structure.

For the production of vowel sounds, the pulses from the generator were also used to drive the digram display, so that it was possible to choose the intervals to give similar digrams to the original utterance. Altering the length of one interval by moving the adjustable mask, however, caused the movement of two points on the digram display, corresponding to the pairing of the interval being changed with the intervals immediately preceding and following it.

Because of the difficulties of making accurate measurements on the digrams and synthesising signals on a microscopic level, no serious attempts were made at producing vowel sounds according to second-order statistics. Using these techniques it was difficult to see how a successful algorithm could be devised and tested that would enable all the vowel sounds to be synthesised from their digram information.

A partial solution would have been to construct a generator that worked directly from digram information. Such a generator could have been of a similar form to the time-interval generator but utilising a different kind of mask. The mask envisaged would cover the whole of the face of the cathode-ray tube, having holes made in it corresponding to the brightest points on a digram. By means of analogue shift-register stages it would be possible to produce an X-deflection proportional to the last interval and to arrange for a sweep in the Y-direction to choose the next one.

The complexity of a generator operating on these principles did not in itself prevent the construction of such a device, but rather it was felt that the production of speech-like sounds at such a microscopic level could best be undertaken with the aid of an on-line digital computer which was shortly to become available.

3.6 Conclusion

The usefulness of the display for providing information for generating a clipped speech wave was not as great as originally hoped, owing to the limitations in recording the digram results and in the control of the time-interval generator. The characteristic patterns obtained for the vowel sounds suggested that significant improvements would be made to synthetic clipped vowels when second-order information was used in their synthesis. The display has shown the degree of variability to be found in the statistics and has helped the author to obtain a better understanding of some of the problems of time-interval analysis of speech. It is hoped that the amplitude-modulated digram display will overcome some of the shortcomings of the original display. Extensive testing will be necessary to evaluate its usefulness as a visual display of speech.

CHAPTER 4

4.1 Introduction

The synthesis of a clipped speech wave involves the processing of large amounts of data, particularly when second-order statistics are considered. The computation of yet higher-order statistics and the synthesis of speech sounds from them was not directly possible without the construction of specialised apparatus. It was felt that a more powerful technique was required to process and synthesise clipped speech, particularly if non steady-state sounds were to be considered. It was shown earlier that the process of infinite clipping could be regarded as a transformation of the speech signal from an analogue to a digital form. The arrival of on-line digital computing facilities within the department meant that the qualititative work described so far could be re-examined more quantitatively and extended further to include higher-order statistics.

Modifications and improvements to the existing apparatus were necessary in order to interface it to the computer. These are described in the first part of this chapter, together with the automated system that was developed in the Department of Communication for obtaining subjects' responses to psychophysical stimuli. The synthesis of vowel sounds from first, second and third-order time-interval statistics, and the subjective results obtained are described later in the chapter. The method is extended to non steady-state speech sounds, and the information rate requirements for such a system are discussed.

4.2 Apparatus

4.2.1 The Computer

The computer was a Digital Equipment Corporation PDP 8 with two memory fields, each field containing 4,096 locations. The cycle-time of the computer memory was 1.5µsec so that it was fast enough to handle the input and output of speech data without introducing significant quantizing errors. A standard interface was provided that enabled direct connections to be made to the computer. This interface was utilised for the transfer of time-interval data to and from the computer. Additional input/output facilities were provided, the most important of these for the present work being two 12 bit registers, one of which could be loaded and read by the computer, the other one being an output register only. Direct connection was possible to these registers and they were used for measuring the fundamental period of the incoming speech, and for gating the synthetic speech on output. They were also used for producing digram displays.

4.2.2 Analysis System

A block diagram showing how the zero-crossing pulses were derived from the speech wave is shown in Fig. 4.1. The incoming speech wave passes through the processor (see Section 4.2.4) into the clipping amplifier. This is substantially the same circuit as that described previously (Section 2.4), except that a Schmitt trigger circuit was used as the final stage to produce constant rise and fall-times. The clipped speech was gated by a control signal derived from the original speech wave in the speech-operated gate (S.O.G.). Pulses were produced

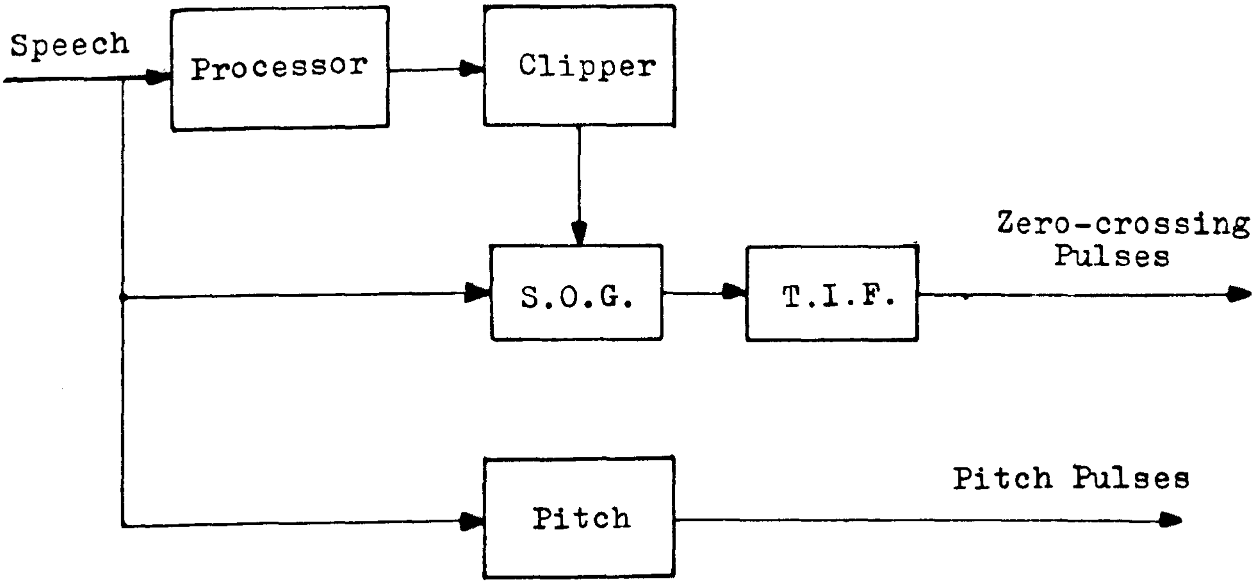


Fig. 4.1 Block Diagram of Analysis System

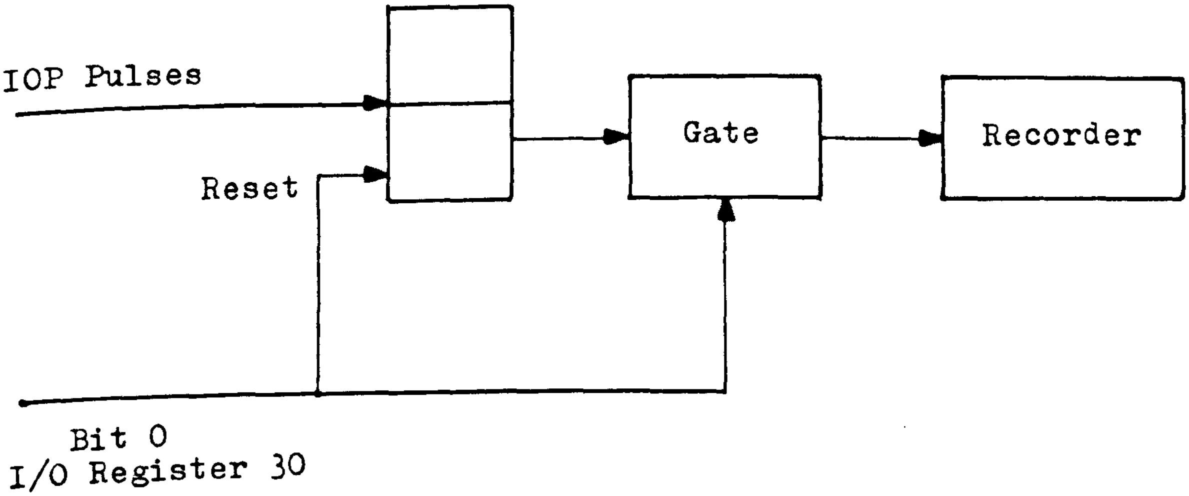


Fig. 4.2 Block Diagram of Synthesis System

at every zero-crossing by the time-interval filter (T.I.F.) which prevented time-intervals shorter than 30µsec being transmitted (see Section 4.2.7). A simple fundamental pitch extractor was incorporated in the system; its operation is explained more fully in Section 4.2.5.

4.2.3 Synthesis System

The original plan had been to use the digital computer to perform the statistical analysis of the clipped speech and then produce control signals for the time-interval generator according to different synthesis algorithms. The purpose of the time-interval generator, however, is to produce pulses defining time-intervals. As computer routines had to be written to measure the time-intervals occurring in real speech, it was considered more straightforward to modify these routines so that they could also produce correctly spaced pulses. In this way the cathode-ray tube part of the time-interval generator was no longer required and the computer was interfaced directly to the bi-stable. A block diagram of the Synthesis System used is shown in Fig. 4.2.

Computer input/output pulses were used to trigger the bi-stable, whose output was passed through a four diode gate. The four diode gate was controlled by another bi-stable that was either set or cleared by the most significant bit of a 12 bit output register on the computer. For producing glottal-triggered sounds, the bi-stable was always set to a '1' state whenever the gate was opened so that the correct phase relationships could be maintained (see Section 2.8). The signal which came out of the gate could be in one of three states, +1, -1, or 0. The alteration of the apparatus so that the signal had this form, was considered necessary for producing glottal-gated speech sounds. The reasons are discussed more fully in Chapter 5.

4.2.4 Processor

A block diagram showing the operation of the processor is shown in Fig. 4.3. The first stage consists of a long-tailed pair amplifier, either output of which could be selected. This was incorporated so that the correct phase relationship of the speech could be maintained (see Section 4.2.5 on pitch extractor). The buffer stage was designed to match a variable high-pass/low-pass filter which was used for some experiments. Switching was employed around the operational amplifier so that it could be used an an integrator or differentiator as well as a straight-through amplifier. The time-constants for the integration (10^{-3}sec) and differentiation (10^{-4}sec) were chosen to ensure that none of the emphasised speech signals were clipped before passing into the clipping amplifier. It was necessary to take this precaution because of the a.c. coupling used within the clipping amplifier.

4.2.5 Pitch Extractor

A block diagram of the pitch-extractor is shown in Fig. 4.4. It works by detecting the large voltage excursions which occur at the beginning of each glottal period and is similar to the pitch extractor described by Gold (23) and more recently by Thomas (63). A fast-acting automatic gain control (Mullard circuit modified to give a response time of around 20msec) was used to ensure that the trigger circuit was always fed with a constant amplitude signal. When the

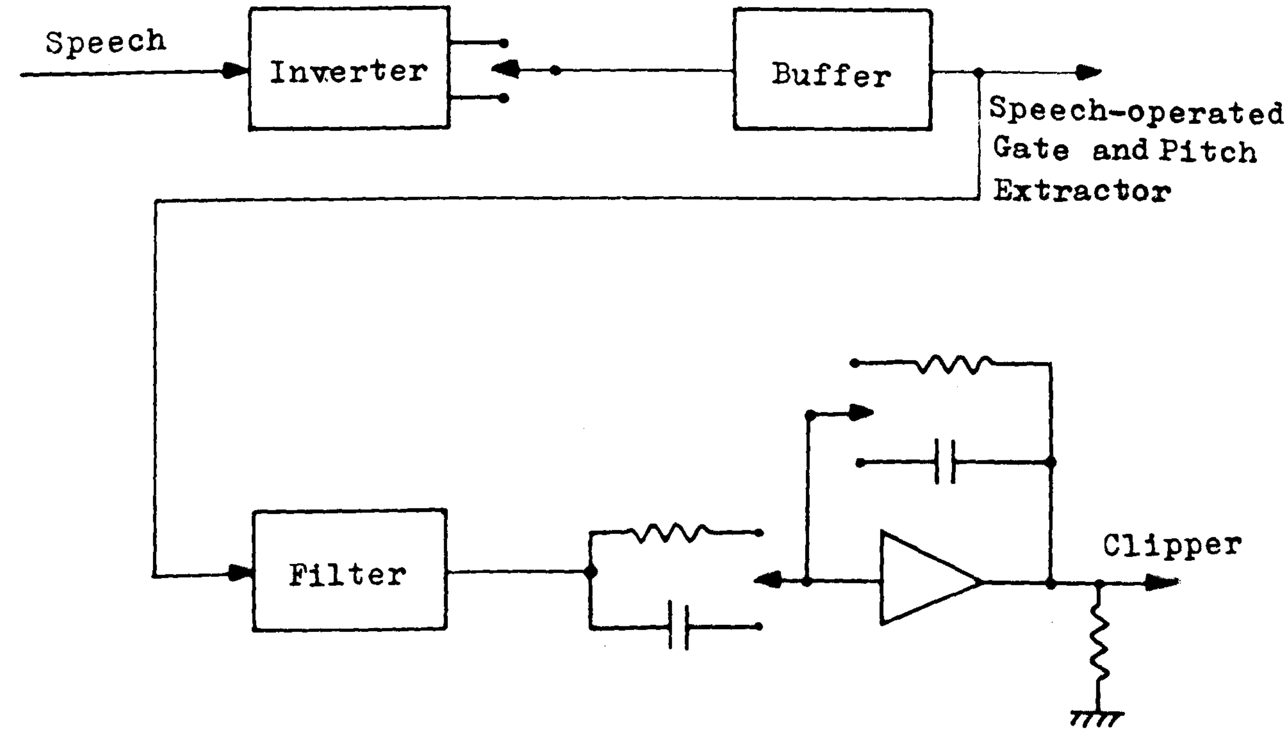


Fig. 4.3 Processor

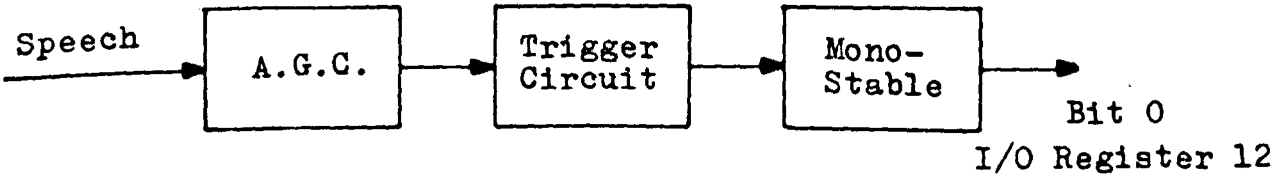


Fig. 4.4 Pitch Extractor

threshold was exceeded, the trigger circuit fired, causing the mono-stable to be triggered. This in turn caused one bit of a register within the computer to be cleared. The 'on' or refractory period of the mono-stable was 3.5msec so that the bit in the register was not cleared every time the threshold level was exceeded.

The circuit worked satisfactorily for glottal frequencies up to 250 Hz, even for waveforms of vowels like /u/ where there was little damping of the waveform within a glottal period. Additional software checks were provided to minimise faulty operation of the pitch extractor.

4.2.6 Speech-operated Gate

In the earlier experiments using the CAT computer, it was not possible to control accurately the length of the analysis time. So that non steady-state sounds could be studied, it was decided to control the analysis-time by computer program. To ensure that only speech sounds were analysed by the computer, and for experiments where the speech did not need to go into the computer, it was necessary to design some gating circuitry. There were sufficient gain and noise in the amplifier for it to produce clipped noise at its output even when the input terminals were shorted, so it was decided to utilise the digital nature of the clipped speech signal and use a digital AND gate for controlling it.

A block diagram of the speech-operated gate (S.O.G.) is shown in Fig. 4.5. The incoming speech signal was rectified, amplified and integrated, the output of the integrator being fed into a Schmitt trigger circuit. The output of the Schmitt was a '1' whenever the incoming signal exceeded the threshold (which could be adjusted). The delay

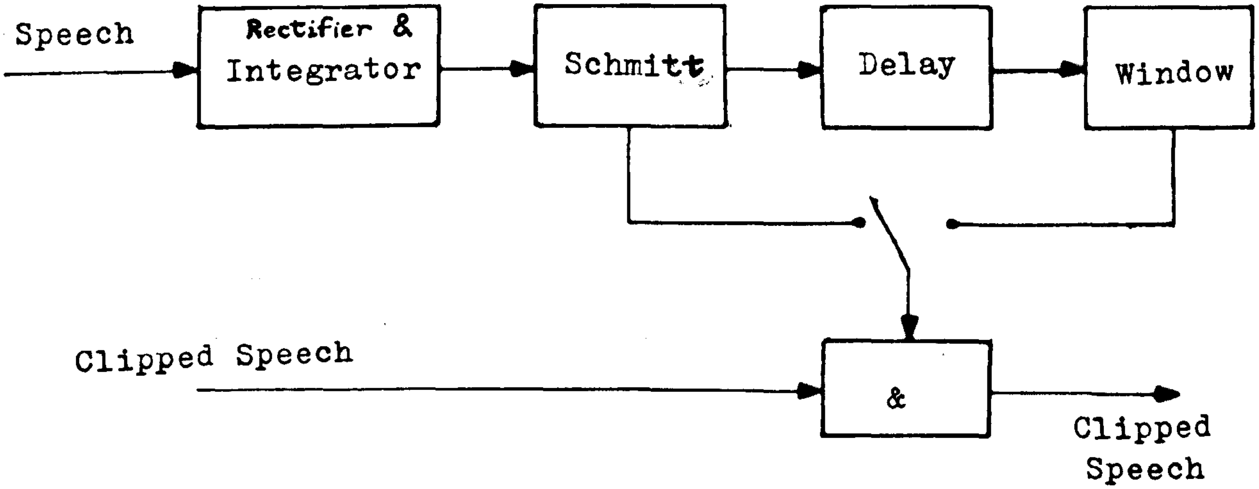


Fig. 4.5 Block Diagram of Speech-operated Gate

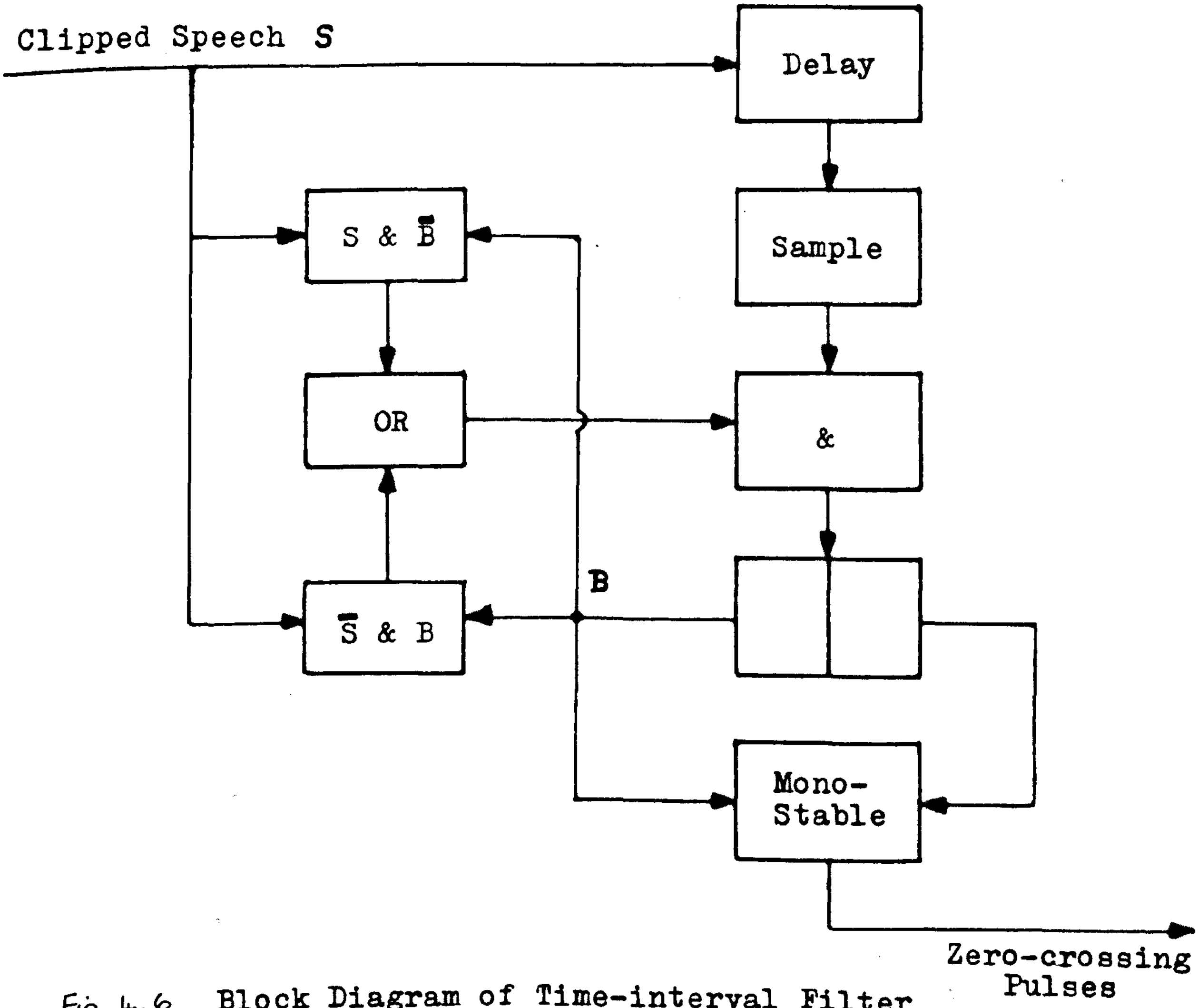


Fig. 4.6 Block Diagram of Time-interval Filter

between the application of the signal and the triggering of the Schmitt was dependent upon the setting of the threshold control and the amplitude of the signal. For a typical setting (so that the circuit was not triggered by tape noise) the delay was estimated to be around 30msec. There were two ways in which the clipped speech signal could be gated by a control signal in the AND gate.

1. The output of the Schmitt could be used directly.
2. The Schmitt output could be used to trigger a mono-stable of adjustable pulse-length (0.2 sec to 1 sec) after a delay of 0.25 sec using the output of the variable mono-stable as an input to the AND gate.

This latter arrangement was particularly useful when the Language Master was used (see Section 4.2.8).

4.2.7 Time-Interval Filter (T.I.F.)

The time-interval filter was designed to prevent two zero-crossing pulses occurring closer together than 30μsec - the time required by the computer program to store the last time-interval and check that the computer store had not overflowed with speech data. Without T.I.F., if two zero-crossing pulses occurred within 30μsec of one another the second pulse would go undetected by the computer. For a voiced sound, this meant an odd number of time-intervals in that glottal period, so that the next glottal period would be out of phase with respect to the previous one. This change of phase could be detected perceptually and occurred more often with differentiated speech because of the high frequency emphasis. If it occurred often enough it tended to destroy the colour of a clipped vowel sound, making human recognition of the vowel difficult.

The principle of operation of T.I.F. is shown in Fig. 4.6. The rising and falling edges of the clipped speech waveform trigger a mono-stable. After a delay of 30μsec (pulse-width of the mono-stable) the clipped speech wave and the bi-stable are both sampled to see if they are in the same state. If they are in different states (S =0, B = 1 or S = 1, B = 0), a pulse is applied to the bi-stable to cause it to change state. If the clipped speech had changed phase twice during the 30μsec, both the clipped speech and the bi-stable would be in the same state and no pulse would be applied to the bi-stable. In this way, the output of the bi-stable is phase-locked to the clipped speech, except that it lags 30μsec behind and cannot change state more than once in 30μsec. The rising edges of the two outputs of the bi-stable trigger a mono-stable (pulse length 8μsec) and these pulses are used as input to the computer.

The removal of all time-intervals shorter than 30μsec did not appear to have any deleterious effect on the intelligibility of clipped or differentiated and clipped speech. In as much as it prevented unwanted phase-changes (when the clipped speech was converted to a pulse form and then back to a rectangular form) it improved the quality.

4.2.8 Language Master

The sounds which were to be analysed were either recorded directly or transferred onto Language Master cards. The Bell and Howell Language Master is a machine primarily intended for use in language laboratories, but was found to be extremely useful in speech research, inspite of some technical shortcomings. Essentially, it is a tape recorder, except that the tape is cut into short lengths and mounted on small cards. Instead of the usual tape guides, there is a slot which takes the cards. Pushing the card into the slot depresses a micro-switch which turns the amplifier and motor on. The replay/ record head is directly opposite the capstan. This fact and the voltage surge in the output circuit when the card is pushed in make the signal recorded at the beginning of the card unreliable. For this reason a delay of 250 msec was built into the speech-operated gate for use with the Language Master.

The frequency response of the machine is not good either. It is shown in Appendix 2, and is deficient at both low and high frequencies. The response for microphone recording is not substantially different. The time-interval patterns are mainly determined by the dominant frequency components (see Appendix 1 and earlier analysis) and for vowel sounds, these fall in the range 200Hz to 2.5 KHz, so that use of the Language Master would not be expected to affect the time-interval patterns much. Ainsworth (2) used a similar machine to measure the intelligibility of various transforms of clipped speech. Where his work overlapped that of Licklider's, the results were in fairly good agreement. This suggests that the use of the Language Master does not have a deleterious effect on the perception of clipped speech.

It was decided that the convenience of being able to replay sounds in any order by shuffling the cards, outweighed the technical shortcomings of the machine, and the Language Master was used throughout in the analysis and preparation of all the sounds used in the listening tests.

4.2.9 Computer Input/Output Routines

The flow charts for these routines are shown in Figures 4.7 and 4.8, and are largely self explanatory. One or two details however, need explanation.

Input Routine (Fig. 4.7)

The time-intervals were measured by the computer using the Program Interrupt facility. Whenever the Program Interrupt is enabled, a change in level on the Program Interrupt (P.I.) line causes the program to transfer control to the bottom of the lower memory field. The routine starting there was made responsible for storing the value of the last time-interval, checking that the memory had not overflowed and initiating the measurement of the next time-interval. The measurement of the time-interval was a counting operation, the counter being incremented every 6μsec. When the counter overflowed, after an interval of 24 msec, this was interpreted as the end of the speech data. Earlier work had shown that the longest time-intervals that could be expected were of the order of 4msec, so that an absence of zero-crossing pulses for 24 msec was a good indication that the speech operated gate had turned off. (The instructions ION and IOF refer to the enabling and disenabling of the P.I., so that zero-crossing pulses were handled during the measurement routine only).

Before the intervals were stored in memory (one word was utilised for each time-interval) Bit 0 of I/O Register 12 was 'OR'ed with the

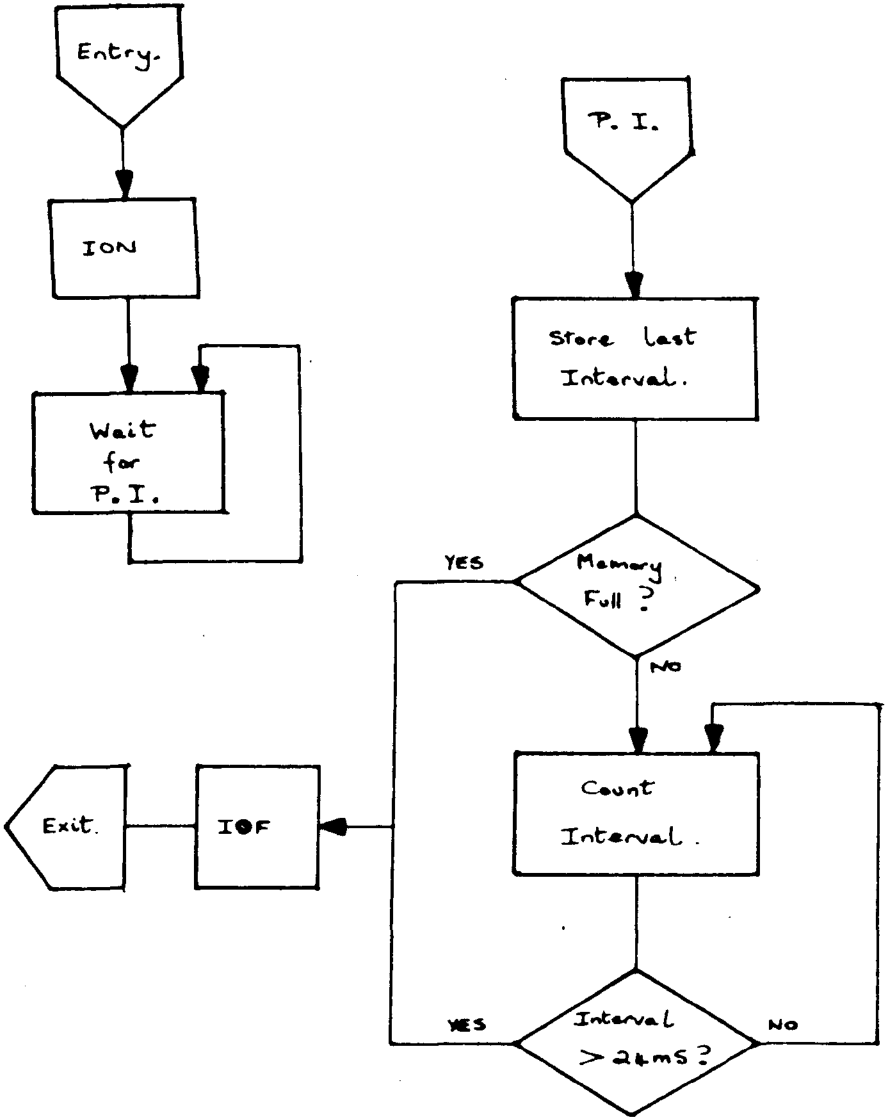


Fig. 4.7 Measurement Routine

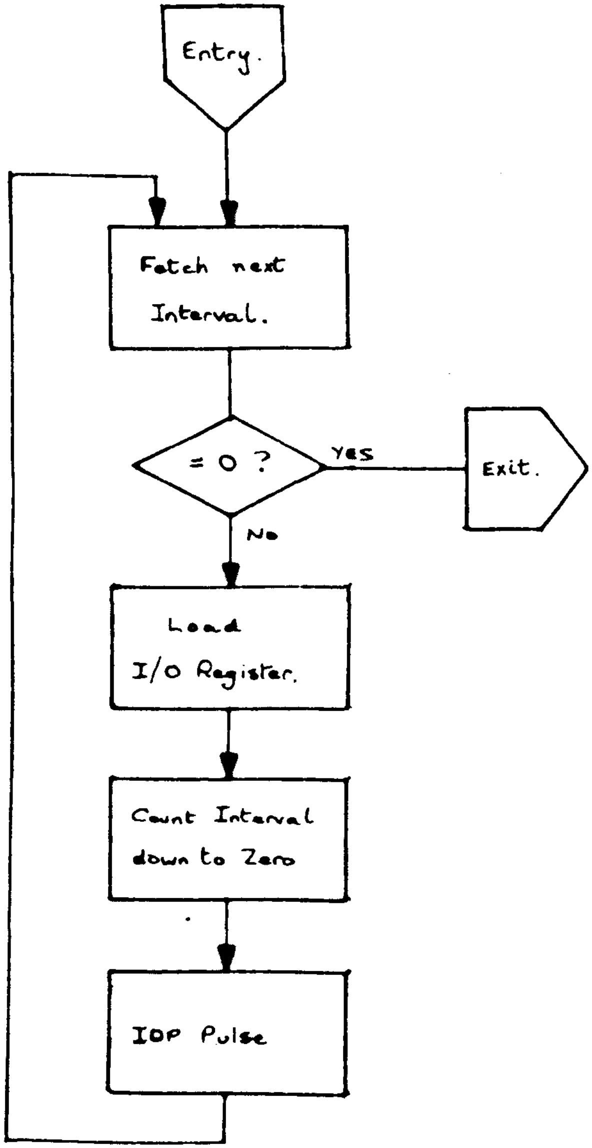


Fig. 4.8 Output Routine

time-interval. This bit was examined in subsequent operations to extract the fundamental pitch.

Output Routine (Fig. 4.8)

This is essentially the converse of the measurement routine. Successive intervals were fetched from the store and examined to see if they were zero. A value of zero signified that there was no more time-interval data to follow. The interval was loaded into I/O Register 30 so that the gate controlling the output signal could be set accordingly. Then, using the same counting method as the measurement routine, the interval was counted down to zero, when an output pulse was generated.

4.3 Choice of Speech Sounds

The earlier experiments with the time-interval generator had indicated that voiced sounds were going to be the most difficult to synthesise. It was decided to concentrate in the first instance on the synthesis of the twelve isolated vowel sounds used previously.

Observation of histograms and digrams of vowel sounds (see Section 3.3.4) showed that there were quite distinct differences between different utterances of the same sound by the same speaker. A contributory factor to this was the variation in fundamental pitch. To minimise the effects of these variations thus making it possible to make valid comparisons between different methods of synthesis, a standard set of vowel sounds was recorded and used for all the experiments. The recordings were made by the author in the following way.

A fundamental frequency of 120Hz was chosen as being typical of the author's normal speaking voice. A 120Hz sine wave was played into headphones worn by the author while he articulated the twelve vowels in isolation. They were recorded on a Truvox R42 Tape recorder with a substantially flat frequency response from 100Hz to 14KHz and subsequently transferred to Language Master Cards. By producing the sounds in this way, it was possible to ensure that all the vowels had the same fundamental frequency, and that the fundamental frequency remained fairly constant throughout the length of the recording. The sounds were produced by referring to the /h/--/d/ words listed in Appendix 3.

4.3.1 Nature of Stimuli

It was found that the Language Master was not suitable for recording clipped speech sounds, because of its limited frequency response. All stimuli for use in listening tests were recorded either directly from the Language Master or from the computer onto a Truvox R42 Tape recorder. The on-axis response of the tape recorder and its loudspeaker (electrical recording at 7½ inches per second) is shown in Fig.A2.1 (Appendix 2), and is substantially free from any pronounced bumps or dips in the range 150Hz to 5KHz.

A complete experiment consisted of up to ten groups of sounds, each group containing twelve vowels, and each vowel being presented once within each group. By shuffling the Language Master cards, it was possible to arrange the vowels in a different order in each group, so that subjects were not able to learn the order in which the sounds were presented. The usual arrangement of the groups within an experiment was of the form A,B,C,D,A,D,C,B,A, where A was a control group and B, C and D were three methods of synthesis or type of distortion.

Each vowel sound lasted 500 msec. This duration was chosen because it was found to provide a sufficiently long signal for subjects to judge without being unnaturally long. Laver (31) used a slightly shorter duration (400 msec) in his study of phoneticians' consistency in identifying vowel sounds. Ahmend and Fatechand (1) found the intelligibility of differentiated and clipped vowels decreased when the duration of the vowel was made less than 120 msec. The most likely effect of the 500 msec duration was to change short vowels into long ones with similar acoustic properties (see Section 5.6).

The four diode gate that was used for controlling the length of the stimuli introduced no clicks on switching, but was turned on and off abruptly. No overall amplitude modulation was applied to the stimuli. No attempt was made to smooth the clipped speech output either. The sharpness of the edges was limited by the high-frequency response of the tape recorder's loudspeaker.

4.3.2 Listening Tests

In view of the large number of experiments that were needed to evaluate the intelligibility of the synthetic vowels, and the large amount of data associated with each experiment, it was decided to automate as much as possible the collection and analysis of data from listening tests.

A system was designed and built within the department that would handle up to twelve YES-NO responses from twelve subjects. The system was interfaced to the computer so that 'on-line' modifications of stimuli could be performed. For the experiments described in this thesis however, the apparatus was not used in this mode. The computer was used to collect the data from the twelve switch boxes and process it to leave it in a suitable form for further analysis. Each subject had in front of him a row of twelve toggle switches. Underneath each switch was a label defining a vowel sound by means of an /h/--/d/ word. The words were in the same order as in the list in Appendix 3 with HEED at the left hand end and THE at the right hand end.

It was not possible to supply each subject with a pair of headphones, so that the experiments had to be conducted using the loudspeaker of the tape recorder. The experiments were conducted in a room which was not ideal acoustically. It measured 20 ft. by 10 ft. by 11 ft., with one of the long walls being a sliding wooden partition. As the room had not been treated with any sound-absorbing material, it was acoustically 'live', and this probably contributed to the fact that the sound pressure level at the places in the room where the subjects sat was constant within 6dB. Because of the relatively small volume of the room, it was thought that the reverberation would not affect the intelligibility appreciably.

4.3.3 Listening Test Procedure

The subjects were given instructions of the following nature.

"In front of you are twelve words, associated with each of these is a switch. The twelve words each define a vowel sound. I want you to identify the sound you hear with the vowel in one of the words and then push the appropriate switch down.

Please try and make a decision, however difficult that might be. If you find the task completely impossible push none of the switches down. When you have made your decision please remove your hands from the switches, so that I can judge when everyone is ready to continue. I will then push a button and this light will flash. This is a signal for you to return your switches to the up position. Failure to return a switch will cause the light to continue flashing, and the computer will read in no more data. The experiment cannot continue until all the switches have been returned to their up position".

No conscious attempt was made to force the subjects into making a quick decision, though some subjects reported that they felt the atmosphere was tense. A typical experiment lasted between twenty and twenty-five minutes.

4.4. Time Quantisation of Clipped Speech

Using the computer, it was hoped that the analysis and synthesis could be extended to third-order or trigram statistics. As the most convenient form of storing the statistical information was to use one computer word for each element of the statistical array, the maximum size of an array was 4,096 words. (One memory field was used to store the array, the other field contained the program and time-interval data). If one memory field were used for a trigram in this way, the time-intervals would have to be quantised into sixteen divisions. The earlier experiments using the CAT computer employed quantisation of 100 μsec or less, but examination of the histograms of vowel sounds of several speakers showed that time-intervals up to 3msec in length were likely to occur. Thus thirty 100 μsec quanta would be required, and this could not be accommodated within the computer.

Before coarser quantisation was utilised in the analysis-synthesis system, it was necessary to see if coarser quantisation could be applied to the original cipped vowels. The experiments of Licklider (36) and Tanaka and Okamoto (62) suggest that the vowel sounds could tolerate coarser quantisation than the consonant sounds and still be recognised. It was most unlikely that synthetic vowels prepared using 300 μsec quantisation would be recognisable if the original clipped vowels quantised at 300 μsec were not recognisable.

The investigation of the effects of quantisation was only necessary for the clipped vowels, and not for the differentiated and clipped vowels. Examination of the histograms of differentiated and clipped vowels of several speakers showed that very few time-intervals longer than 1.6 msec occurred, so that 100 μsec bins could be used. Licklider (36) had shown that this quantisation hardly affected the intelligibility.

4.4.1 Method of Quantisation

There are two ways of classifying the intervals into categories or bins, synchronously or asynchronously.

In the synchronous method, each interval is measured from the last zero-crossing pulse and classified accordingly. In the asynchronous method "zero-crossings" can only occur at specified points in time determined by a free-running clock. As the "zero-crossings" can only occur at the next clock pulse after their actual occurrence, an error in the classification of the interval can occur depending upon the relative timing of the clock pulses and the zero-crossing pulses. This would have the effect of broadening the peaks of a histogram of a vowel sound. For this reason, and because it was easier to implement, all the quantisation of time-intervals was done synchronously. In as much as all the time-intervals were measured using a 6μsec counting routine which was scaled down from the computer's internal 1.5μsec clock, the measurements can be said to be asynchronous. The maximum error in the measurement of a time-interval cannot exceed 6μsec however, and this is small in comparison to the quantisation used in the compilation of the statistics.

A computer program was written to perform the synchronous quantisation of time-intervals. Once a time-interval had been categorised, the program replaced that interval by one of length half-way between its bin limits. Suppose that the quantisation was 100μsec and that a particular interval was 567μsec long. It would be replaced by an interval 550μsec long.

4.4.2 Experiment 1

It was decided to find the effects of quantisation in the range 0 to 500μsec, so a tape was prepared of the twelve clipped vowel sounds quantised at 0, 100, 200, 300, 400 and 500μsec. The subjects for this experiment were five members of the department who had had some previous experience of listening to clipped vowel sounds.

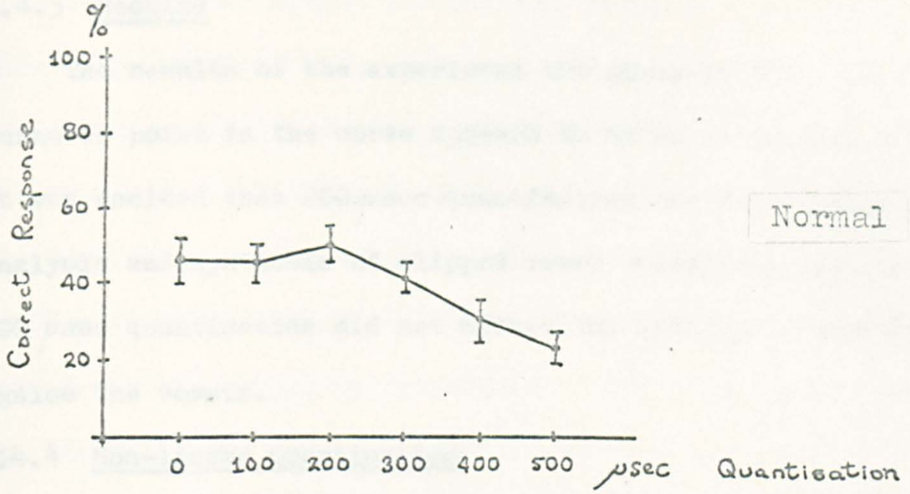


Fig. 4.9 Results of Perceptual Experiments with Quantised Clipped Vowels

4.4.3 Results

The results of the experiment are shown in Fig. 4.9. The turnover point in the curve appears to be in the region of 250 μsec. It was decided that 200 μsec quantisation could be employed for the analysis and synthesis of clipped vowel sounds, in the knowledge that 200 μsec quantisation did not affect the ability of subjects to recognise the vowels.

4.4.4 Non-linear Quantisation

As the speech was quantised synchronously after the intervals had been measured and stored, it was possible to use a non-linear scale for the classification of the time-intervals. The 200 μsec linear quantisation used for the undifferentiated speech was two coarse for fricative sounds; it was impossible to distinguish between them. A scale to accommodate both voiced and unvoiced sounds would need finer resolution for shorter intervals, and coarser resolution for the longer intervals if only sixteen categories were to be employed. Moreover, it was possible that the different classification of the intervals might lead to more economical storage of the statistics (see Section 4.7).

A commonly used scale in psycho-acoustics employs one third octave bands, so it was decided to use such a logarithmic scale in place of the linear one, and to investigate its effect upon the recognition of clipped vowels.

4.4.5 Experiments 2 and 3

Two tapes were prepared to find the effects of this logarithmic quantisation on clipped, and differentiated and clipped vowels. The first tape compared clipped vowels with no quantisation, 200 μsec linear quantisation and logarithmic quantisation. The second tape compared differentiated and clipped vowels with no quantisation, 100 μsec linear quantisation, and two forms of logarithmic quantisation. The difference between the two forms of logarithmic quantisation was the time-scale over which they operated. One was a logarithmic scale for the range 200 μsec to 3.2msec (log low), and the other for the range 100 μsec to 1.6msec (log high).

The subjects were ten unpaid undergraduates of the University, with no previous experience of listening to clipped vowels.

4.4.6 <u>Results</u>

The results of the two experiments are shown in the following tables.

	% Correct	
Natural	43	
Clipped	24	<u>Experiment 2</u>
200μsec linear	24	Clipped Vowels
Logarithmic	3	

Table 4.1

	% Correct	
Clipped	33	<u>Experiment 3</u>
100μsec linear	37	Differentiated and
Log (low)	46	Clipped Vowels
Log (high)	43	

Table 4.2

4.4.7 Discussion

Experiment 2

The results of this experiment and those of experiment 1 show that 200 μsec linear quantisation makes little difference to the recognition of clipped vowels, although the absolute level of performance of the inexperienced subjects is only half that of the more experienced ones. The results for the individual subjects were examined to see if there was any evidence of learning by the subjects during the experiment. As many subjects' scores were found to decrease as increase for the repetition of the same quantisation within the experiment. The subjects responded more quickly as the experiment continued, but this was not reflected in their scores. (see Section 5.6.1 for further discussion of learning effects).

The results for logarithmic quantisation are strikingly poor. Only 20% of the vowels were put into their correct F1 and F2 categories (see Section 4.5.4 for a fuller explanation) by the subjects compared to over 50% for the linear quantisation. This suggests that the quantisation of the intervals longer than 1msec (where the linear and logarithmic time-bases have the same slope) is too coarse for maintaining vowel quality.

Experiment 3

The results for this experiment show the reverse trend to those of Experiment 2; the logarithmic quantisation produces better results than the linear quantisation, indeed better results than the original differentiated and clipped vowels. Examination of the confusion matrices for the experiment showed that the increase could be attributed to more correct responses to /i/ and /æ/ for the logarithmic quantisation. No reason for this increase could be found. Further testing would be necessary with more subjects and other speakers' voices to see whether the effect is widespread.

On the basis of these experiments, it was decided to use 200 μsec linear quantisation for the analysis and synthesis of clipped vowels, and logarithmic for differentiated and clipped vowels. Log (high) was chosen as it provided better time-interval resolution.

4.5 The Synthesis of Vowel Sounds

The experiments with the time-interval generator, selecting intervals at random according to their first-order statistics, showed that a most unvowel-like sound was produced. This was because there was no regular triggering of a sequence of intervals, corresponding to glottal excitation. With the apparatus available, it had been impossible to extend the synthesis to incorporate higher order statistics. As the order of the statistics is increased it becomes more possible to re-create the original waveform unambiguously.

For example, a statistical array of order m would be capable of storing unambiguously a steady waveform that repeated itself after m intervals, because the specification of any one point in the m-dimensional space of the array would require the knowledge of m intervals. It is possible, however, for an m-dimensional array to store unambiguously information about a waveform that repeats after more than m intervals.

Consider for example, the repetitive sequence of intervals of different lengths a, b, c, d. The co-ordinates of occupied elements in digram space would be (a,b) (b,c) (c,d) and (d,a). Suppose that a synthesis algorithm selected intervals at random according to their digram statistics, and that the first interval the algorithm chose was one of length b. The algorithm would then search the histogram defined by x=b, where x corresponds to the length of the n^{th} time-interval. (The digram can be considered as a set of histograms of intervals following a particular interval (see Section 3.2).) The only non-zero element in that histogram would correspond to an interval of length c. The algorithm would then search the histogram defined by x=c and so on. The resultant sequence of intervals would be b, c, d, a recurring, which is the same sequence as the original.

If the time-intervals are divided into sixteen categories, and a repetitive sequence of sixteen different intervals existed, it could be represented by a digram of 256 elements, and a random synthesis algorithm of the type just described would be capable of re-creating the original waveform. If however, two of the intervals in the sequence were of the same length, the histogram (in the digram) corresponding to that length of interval would have two occupied elements. On some occasions therefore, the algorithm would choose the shorter interval, on other occasions the longer one, so that the original waveform would not be re-created exactly.

The histograms shown in Fig. 2.4 indicated that several intervals of the same length appeared in each glottal period, so that exact re-creation of the waveform would not be possible without going to an order of statistics equal to the number of intervals in the glottal period. Moreover, slight variations in the speech parameters can cause

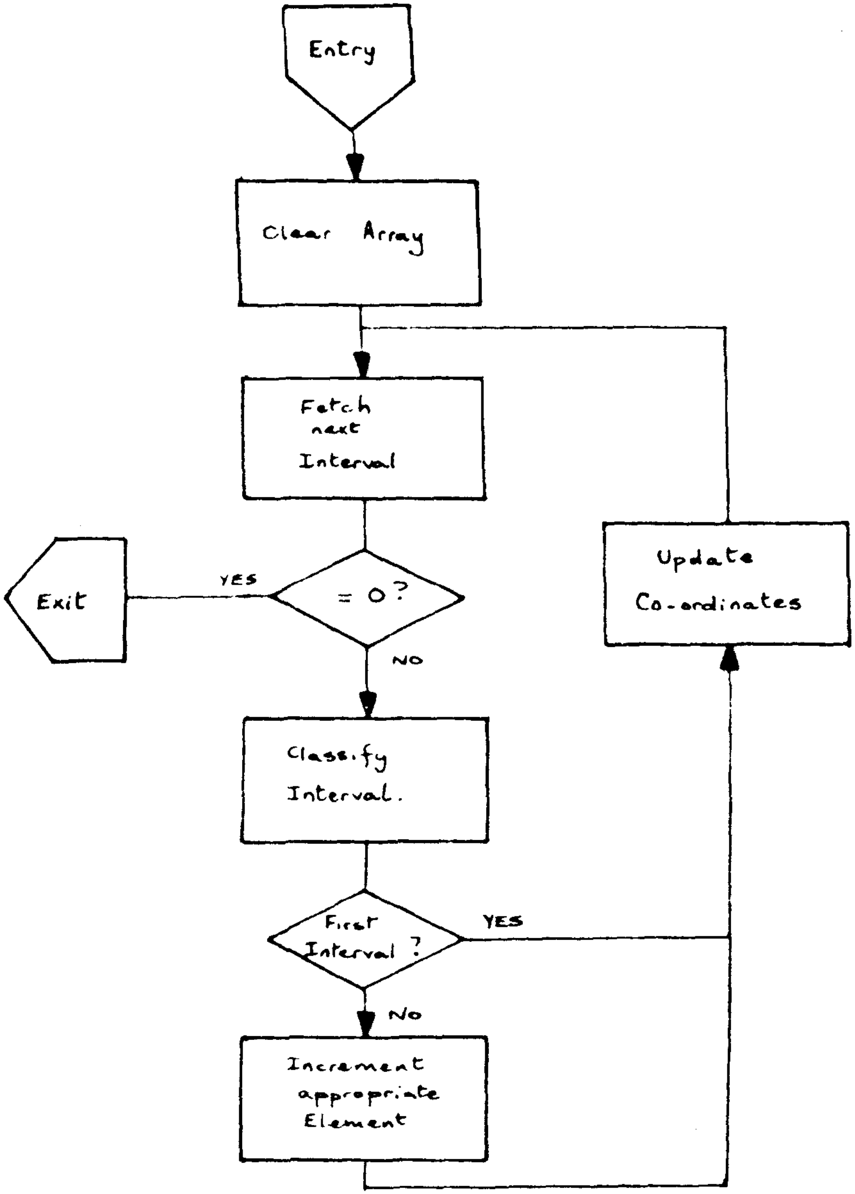


Fig. 4.10 Compilation

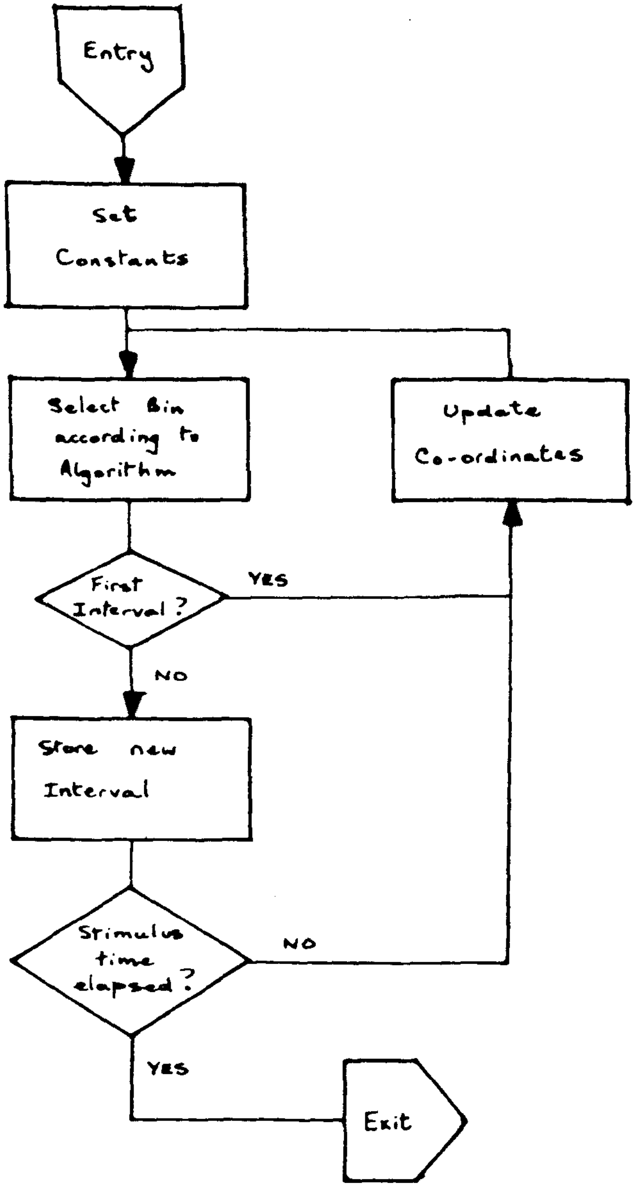


Fig. 4.11 Synthesis

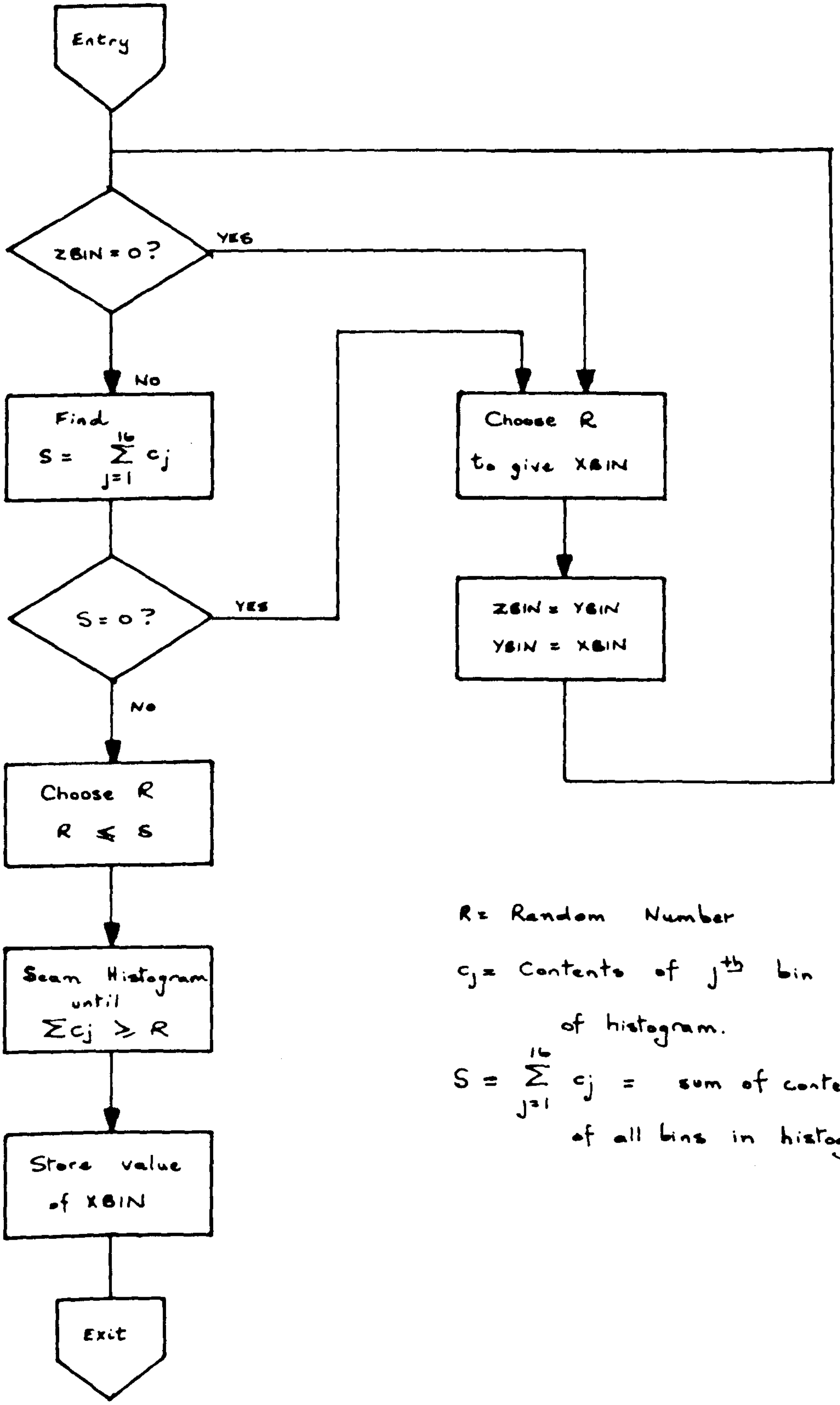
marked changes in the zero-crossing pattern from one glottal period to the next, making the re-creation of the original wave more difficult.

The purpose of this investigation, however, is not to re-create the original waveform but to see how much statistical information is required to produce a recognisable approximation to a speech signal.

4.5.1 Method of Synthesis

A computer program was written to perform the kind of synthesis described in the previous section. Flow charts for the compilation and synthesis routines are shown in Figures 4.10 and 4.11 and a more detailed flow chart for random synthesis from trigram statistics is shown in Fig. 4.12.

A pseudo-random number routine was used to select the intervals randomly from the statistics in the following way. Any element in the trigram could be specified by three co-ordinates, called ZBIN, YBIN, and XBIN, where ZBIN referred to the first of the intervals of the triplet and XBIN to the last. Having established values for ZBIN and YBIN using the random number routine, the sixteen bin histogram defined by ZBIN and YBIN was examined to see if any of its elements were non-zero. If all the elements were zero, a new value was chosen for YBIN, (after the last value of YBIN had been transferred to ZBIN) and the new histogram searched. When a non-zero histogram was found, its contents were summed and a new random number generated whose value was less than the sum of the contents of the histogram. The contents of successive elements of the histogram were summed until they exceeded the value of that random number. In this way a bin with a large count



R = Random Number

c_j = Contents of j^{th} bin of histogram.

$S = \sum_{j=1}^{16} c_j$ = sum of contents of all bins in histogram

Fig. 4.12 Random Algorithm Synthesis

was more likely to be selected than one with fewer counts in it. An interval of length corresponding to XBIN was stored, and the values of YBIN and XBIN became the values of ZBIN and YBIN respectively for the selection of the next interval. (For the digram, only two co-ordinates were required so that the updating of the co-ordinates was the equation of YBIN to XBIN. For the histogram, the intervals were chosen independently of one another so that there was no updating of the co-ordinates after each interval had been selected).

After 500 msec of intervals had been selected and stored, the synthetic sequence was outputted and recorded on the tape recorder. The statistics were also compiled over 500 msec.

Each histogram was checked for zero contents as it was quite possible that even after the algorithm had produced several intervals, it would come to the point where the compilation routine had finished so that there would be no continuity in the statistics. Such points will be referred to as 'dead-ends'. In the event of the synthesis algorithm arriving at a dead-end, it would update the co-ordinates and try again. This meant that if trigram statistics of the synthetic wave had been compiled, there would be three points (corresponding to this discontinuity) in the synthetic statistics that had not been present in the original statistics. As there would typically be over five hundred intervals in a 500msec utterance, this discontinuity was considered to be relatively unimportant.

4.5.2 Quality of the Vowels

There was a noticeable improvement in the difference between the

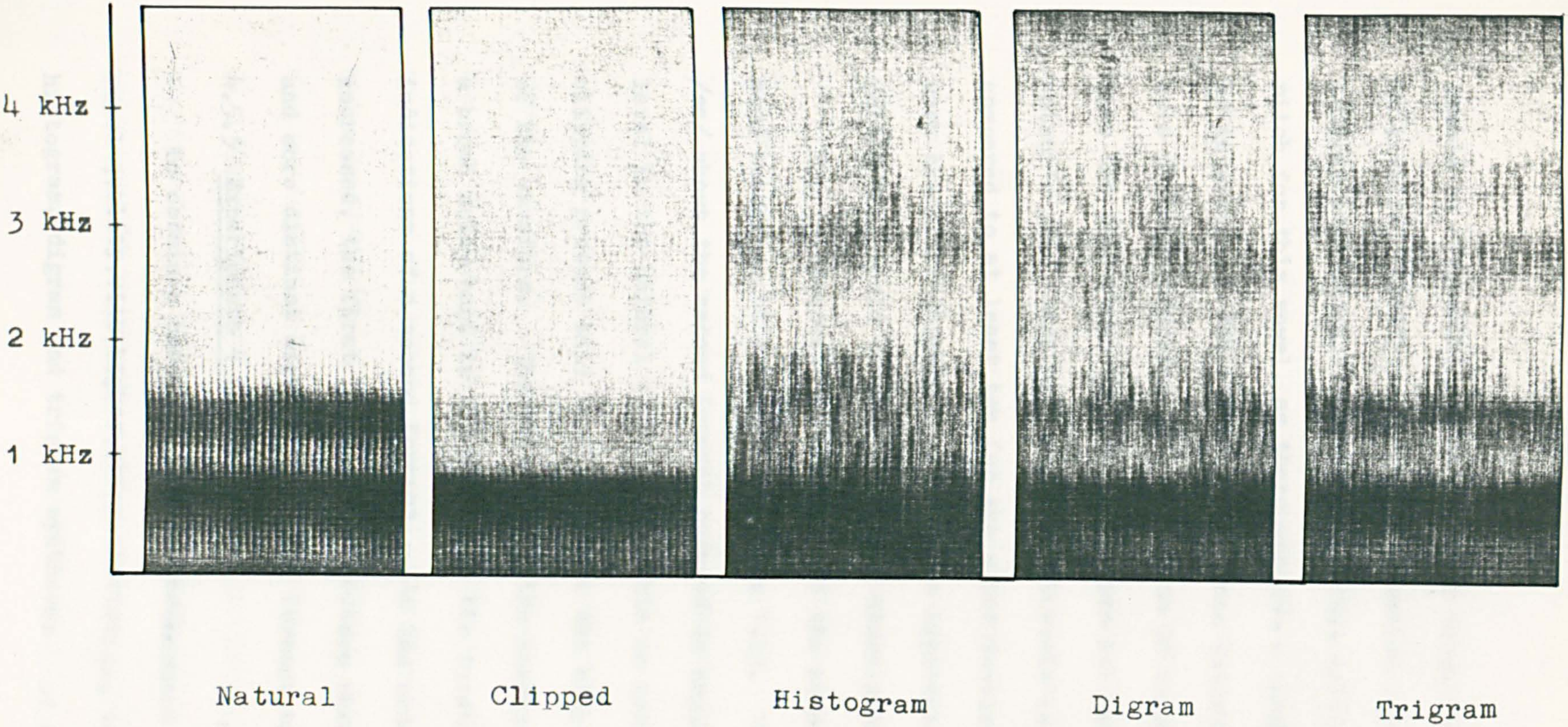


Fig. 4.13 Broad Band Sonagrams of Natural and Synthetic /æ/

vowels as high-order statistics were used, but all the sounds retained an unvoiced quality. The only exception was /u/ synthesised from trigram statistics. It was impossible to establish a fundamental pitch for this vowel, as there was not a steady repetitive sequence of intervals. Nevertheless, the long intervals combined with a more definite ordering than histogram or digram synthesis produced, gave the impression that the sound was not unvoiced. (For the author's /u/ there are only six intervals in a glottal period, compared to at least ten for the other vowels. The trigram statistics were therefore capable of storing an appreciable part of each glottal period that could be resynthesised unambiguously).

The improvement in structure of the sounds can be seen in wide-band sonagrams of /æ/ shown in Fig. 4.13. The sonagram of clipped /æ/ shows the second formant reduced in amplitude compared to its level in the natural utterance. This is not surprising as the clipping process will tend to remove the higher frequency components of the waveform. The sonagram of the histogram synthesised /æ/ shows a broad noisy band in the region of the first formant, and faint indications of a second formant. As the order of the statistics is increased, the first formant band becomes narrower and better structured and more distinct signs of a second formant emerge.

4.5.3 Experiments 4 and 5

To obtain a more quantitative assessment of the improvement of vowel quality with higher order statistics, tapes were prepared of histogram, digram and trigram synthesis of normal and differentiated vowels. These were played to two groups of subjects. Group I were twelve members of a local youth club, aged between thirteen and seventeen, with no previous experience. Group II were five members of the department with some experience of listening to and identifying clipped vowel sounds.

4.5.4. Results

The results of the experiments with the two groups of subjects are shown in Fig. 4.14. The lower curves with the error bars (± 1 standard error) represent the percentage of vowels identified correctly; the upper curves represent the percentage of vowels classified into their correct first and second formant groups. The horizontal lines at the right of each graph represent the subjects' scores for the vowels from which the statistics were compiled.

The grouping of the vowels was employed for three reasons.

1. The inexperienced subjects in particular, had great difficulty in identifying the synthetic vowels, and they were particularly instructed not to be too concerned about pressing what they thought was the correct switch. Their difficulty in performing the task is reflected in their comparatively low scores. It was hoped that grouping the vowel sounds in the analysis of the results would give a more accurate indication of the subjects' performance.

2. The comparatively small number of subjects. Two or three subjects identifying a sound as /ɜ/ instead of /ə/ for example, were sufficient to change the small number of correct responses. The major

difference between /ɜ/ and /ə/ in normal speech is that of duration, but as all the stimuli were of the same length, this clue was missing and confusion arose. It was considered more realistic to group the sounds according to their formant frequencies, so that the very obvious confusions would be ignored.

3. The subjects had different accents, depending upon the part of the country from which they came, so that there was some disagreement as to how the key words were pronounced. This would obviously affect the number of correct responses. By grouping the vowels it was hoped that individual differences due to accents would be minimised.

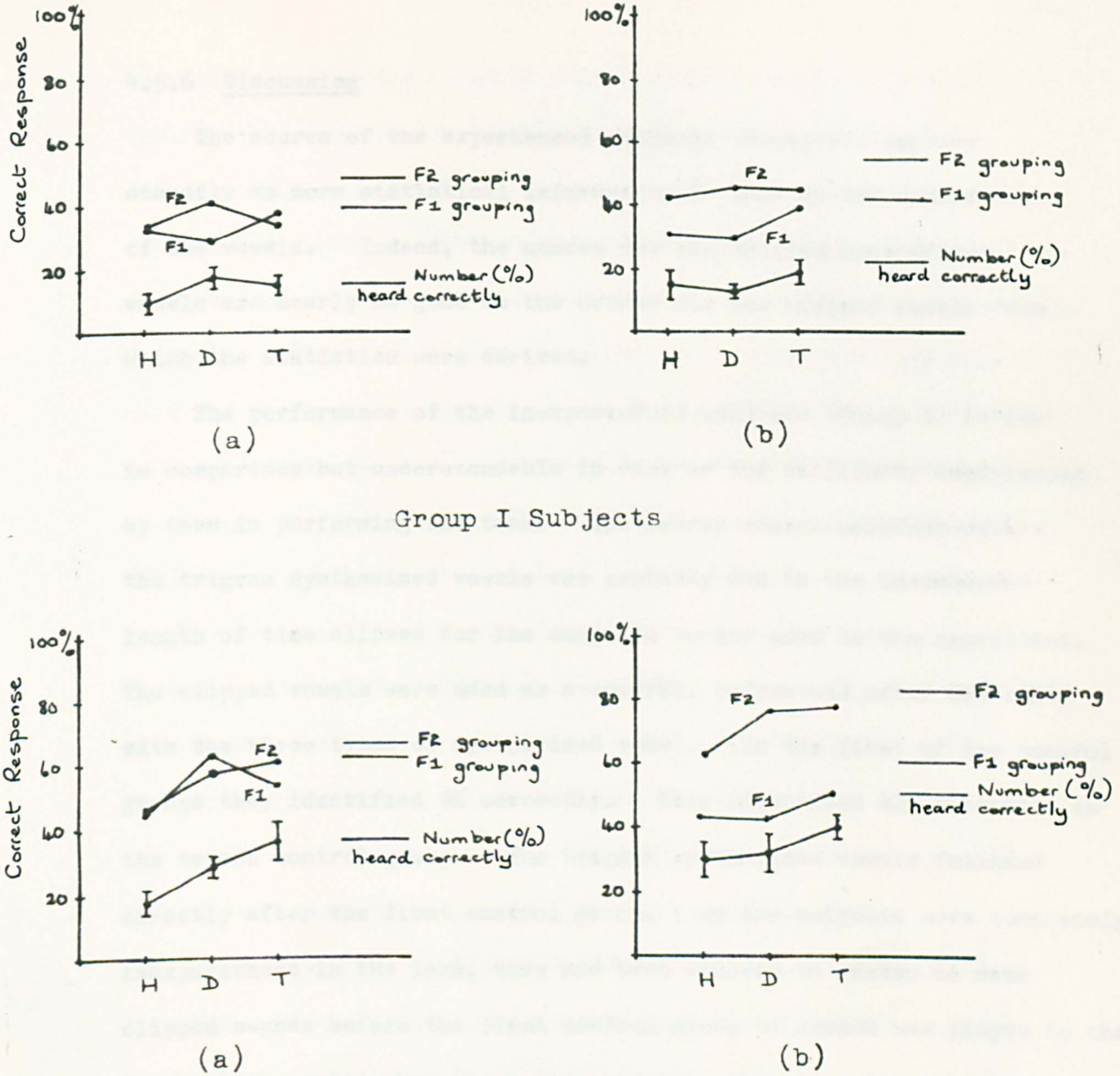
The second formant grouping was a very broad one and corresponds to the phonetic classification front, centre, back.

FRONT	/i/, /I/, /ɛ/, /æ/
CENTRE	/ɜ/, /ə/, /ʌ/, /ɑ/
BACK	/ɒ/, /ɔ/, /ʊ/, /u/

The vowels were divided into four first formant categories.

/i/, /u/
/I/, /ɔ/, /ʊ/
/ɛ/, /ɜ/, /ə/, /ɒ/
/æ/, /ʌ/, /ɑ/

The two groupings were chosen after examination of the confusion matrices for clipped, and differentiated and clipped vowels. This is discussed more fully in Section 5.6.



(a)

(b)

Group I Subjects

(a)

(b)

Group II Subjects

Fig. 4.14 Results of Perceptual Experiments for Histogram (H), Digram (D) and Trigram (T) Synthesis

(a) Normal (b) Differentiated

4.5.6 Discussion

The scores of the experienced subjects (Group II) improve steadily as more statistical information is used in the synthesis of the vowels. Indeed, the scores for the trigram synthesised vowels are nearly as good as the scores for the clipped vowels from which the statistics were derived.

The performance of the inexperienced subjects (Group I) is low in comparison but understandable in view of the difficulty experienced by them in performing the task. The poorer scores obtained with the trigram synthesised vowels are probably due to the inadequate length of time allowed for the subjects to get used to the experiment. The clipped vowels were used as a control, before and after the tests with the three types of synthesised vowel. In the first of the control groups they identified 9% correctly. They identified 21% correctly in the second control group. The trigram synthesised vowels followed directly after the first control group. As the subjects were completely inexperienced in the task, they had been allowed to listen to some clipped sounds before the first control group of sounds was played to them. It should be noted that these subjects were the only ones to show such a marked learning effect, and it could not have been predicted from the results of previous experiments using phonetically naive subjects.

With the exception of the learning effect, all the scores show a tendency to increase as the amount of statistical information utilised in the synthesis is increased. The scores for the trigram synthesis are not much below those for the clipped, and differentiated and clipped vowels from which the statistics were derived. There is probably insufficient data for the slopes of the curves to have any real significance, except that the experienced subjects showed a greater improvement as the amount of information was increased. In all cases,better scores were obtained with clipped vowels where the speech had been differentiated prior to clipping. This is in agreement with the results of Licklider (35) and Ainsworth (2).

Ainsworth used the monosyllabic words comprising the PB lists (15) to measure the intelligibility of clipped speech and found that only 1.1% of the vowels were misheard. Although the two subjects with whom he obtained these results were both present in the group of experienced subjects, they were no better than the remainder of the experienced subjects in identifying the clipped vowels correctly. This reinforces the criticism that the nature of the listening tests was rather artificial, and that the task of identifying isolated clipped and synthetic vowels sounds was by no means easy.

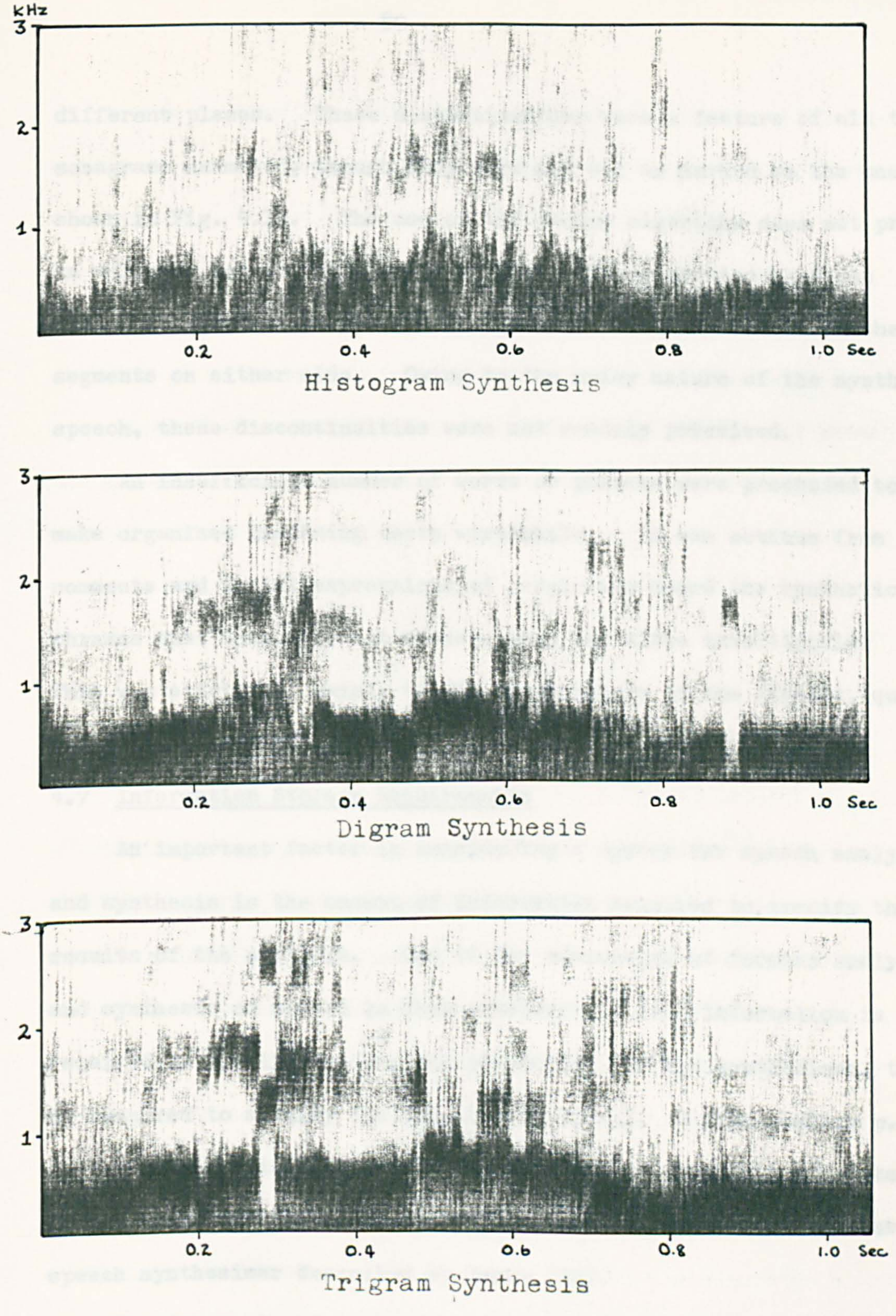
4.6 Synthesis of a Phrase

The recognition of the synthetic vowels was a difficult task, especially for the inexperienced subjects. The analysis and synthesis of different material such as the PB words would have provided data on all types of speech sounds, as well as being a less artificial type of stimulus. It was not possible however to use the existing system to process the PB words for several reasons. The greatest limitation was the size of the computer store, which although it was large enough to store some of the words, was not large enough to store all the words which contained fricative sounds. This was particularly so for speech that had been differentiated first. Moreover, it was felt that it was necessary to introduce some glottal triggering to make voiced sounds more realistic before extensive testing of a system on all classes of speech sounds.

The programs were slightly modified, however, so that non steady-state sounds could be processed. The main modification was to the timing routine, so that statistics of intervals were compiled every 20msec. This was considered to be the longest segment that could be used, consistent with retaining information about the lengths of silent periods and noise bursts in some stop-like sounds (25, 28). The phrase "Where are you?" was chosen because it was voiced throughout and all the time-intervals could be stored within the computer.

Perceptually, the quality of the phrase improved with the order of statistics used for synthesis, although people who had not heard the phrase before were not able to recognise it easily. Broad band sonagrams of the three synthetic versions are shown in Fig. 4.15.

The first formant structure becomes better defined as the order of statistics is increased, and approximates well to the first formant of the original utterance (see Fig. 2.1). Both digram and trigram versions show abrupt discontinuities in the formant pattern of some 20msec segments, near the end of the digram versions, and one third of the way through the trigram version. The phrase was processed again, and new spectrograms prepared. Abrupt discontinuities were visible in these spectrograms too, except that they occurred in



Histogram Synthesis

Digram Synthesis

Trigram Synthesis

Fig. 4.15 Broad Band Sonagrams of Synthetic "Where are you?"

different places. These discontinuities were a feature of all the sonagrams examined, though they were not all as marked as the ones shown in Fig. 4.15. The use of the random algorithm does not preclude an unlikely sequence of intervals from being generated that has different spectral properties to the sequence of intervals in the segments on either side. Owing to the noisy nature of the synthetic speech, these discontinuities were not readily perceived.

An insufficient number of words or phrases were processed to make organised listening tests worthwhile. It was obvious from the comments and facial expressions of people who heard the synthetic phrases that they were not speech-like, let alone intelligible! This was attributed mainly to the noisy nature of the signals, quite unlike human speech.

4.7 Information Storage Requirements

An important factor in considering a system for speech analysis and synthesis is the amount of information required to specify the results of the analysis. One of the advantages of formant analysis and synthesis of speech is that considerably less information is required to specify the control parameters for the synthesiser, than is required to specify the speech wave itself. A high quality p.c.m. speech link requires typically 50,000 bits/sec whereas 4,500 bits/sec is all that is required to specify the control parameters for the speech synthesiser described by Holmes (29).

To compare the amounts of information needed to specify the original clipped speech wave and the first, second and third order statistics derived from the speech wave, the following analysis-synthesis system (based on the one described in Section 4.6) was considered.

Suppose that time-interval statistics were compiled over 20 msec segments and that only one bit was required to specify the contents of an element in the statistical arrays, i.e. the elements were classified as being either empty or full. Table 4.3 shows the amount of information that would be required to specify the contents of every element fifty times a second.

Histogram	800 bits
Digram	12.8 Kbits
Trigram	204.8 Kbits

Table 4.3

Considering that 5kbits/sec and 10kbits/sec are all that are required to specify the clipped, and differentiated and clipped signals quantised at 200 μsec and 100 μsec respectively, the only saving in bandwidth would be offered by the histogram statistics. Although no experiments were conducted to estimate the effect of quantising the contents of the elements of the statistical arrays, it is thought that five bits would be sufficient, so that the histogram information would occupy no more storage space than the clipped speech signal.

The time-interval statistics could be stored more economically if only information about the occupied elements was stored.

Examination of the digram in quantised form showed that many of the 256 elements were empty. The following experiment was carried out to investigate the number of occupied cells in histograms, digrams and trigrams.

4.7.1 The Number of Occupied Elements

A computer program was written to compile histogram, digram and trigram statistics every 20msec of speech. The number of occupied cells in each array was counted, and after three minutes of speech had been analysed, the maximum number of occupied elements of each of the three arrays was printed out. (The speech was a recording made by the author reading from a book). The experiment was performed for normal and differentiated speech, the quantisation being 200 μsec and 100 μsec respectively. Although 200 μsec quantisation was too coarse to give adequate resolution of fricative sounds, it was used to give an indication of the information storage requirements.

The maximum number of occupied cells in statistical arrays compiled over 20msec are shown in Table 4.4.

	Histogram	Digram	Trigram
Normal	13	25	30
Differentiated	12	27	36

Table 4.4

These results show that transmitting information about all 4,096 elements in a trigram is clearly unnecessary when less than 40 elements are occupied. As Fourcin (22) has shown, gross time-interval statistics are dependent upon a number of factors, notably characteristics of the speaker's voice, and the type of material he is reading, so that the figures quoted above are in no way intended as absolute maximum values, but rather as an indication of the maximum numbers likely to be achieved.

The results in Table 4.4 are used to calculate the amount of information needed to specify the non-zero elements of the arrays. The results are shown in Table 4.5. The equation used was

$$H = NAB$$

where H = Information needed in bits/sec

N = Number of samples / second = 50

A = Number of non-zero elements

B = Number of bits to specify an element = 4 for histogram

= 8 for digram

= 12 for trigram

	Histogram	Digram	Trigram
Normal	2.6k	10k	20.2k
Differentiated	2.4k	10.8k	24.2k

Table 4.5

These results show a considerable saving in the amount of information required to specify the trigram statistics, a slight reduction for the digram, and an increase for the histogram statistics. Compared to the figures of 5kbits/sec and 10kbits/sec though, the histogram method is still the only method that is more economical, as long as the contents are specified to no more than 8 or 9 bits accuracy for the normal and differentiated forms respectively.

4.7.2. Distribution of Contents amongst the Occupied Elements

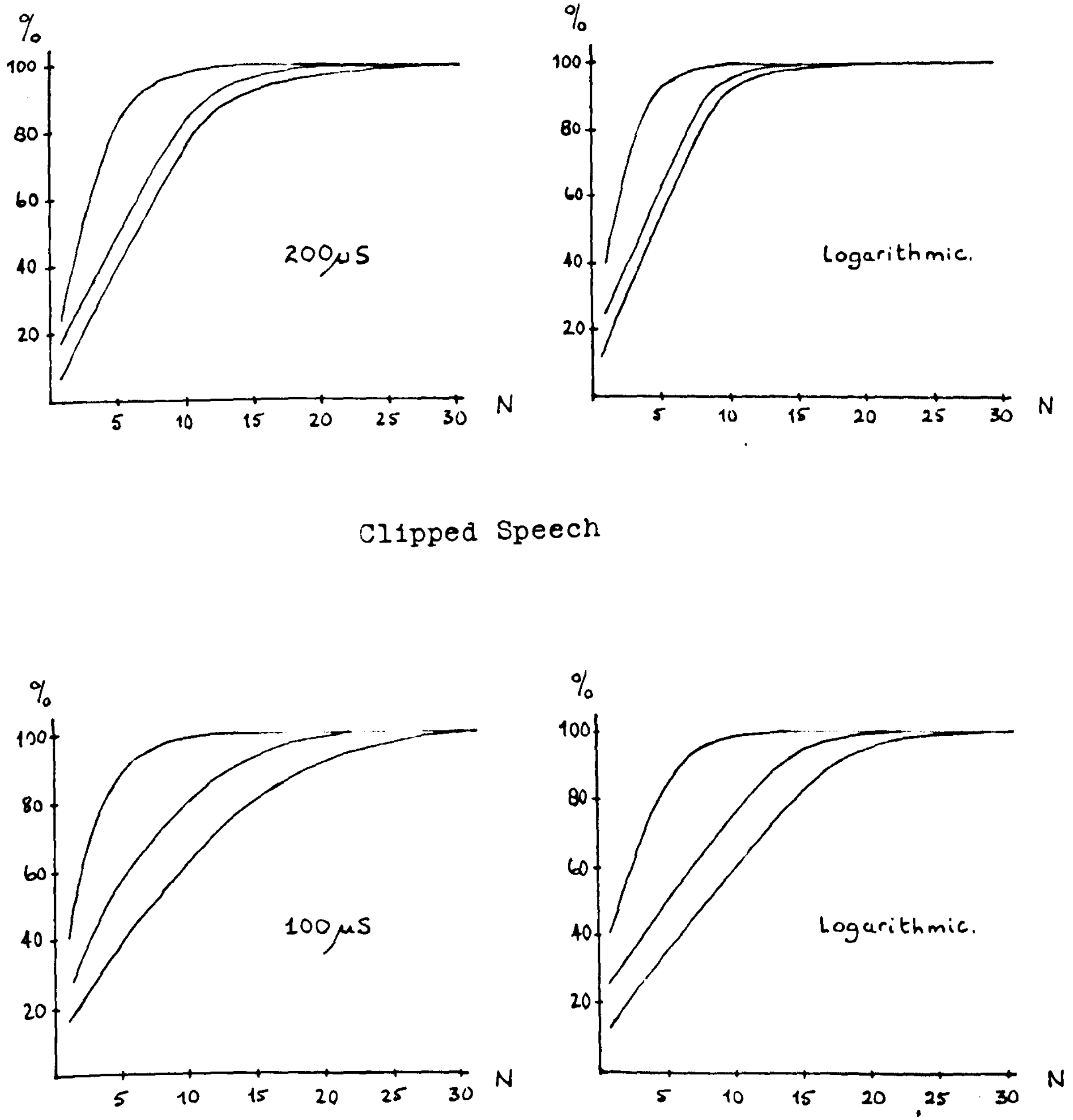
One way in which the information rate could be reduced is to specify only those elements that account for the majority of all the contents of the elements. This would amount to a kind of noise rejection, i.e. cells with only one count in them would be ignored. (The details of a program to do this were never worked out. It was considered as a hypothetical way of reducing the amount of information to specify the major features of the statistics).

The program described in Section 4.7.1 was modified to compute the distribution of contents amongst the cells. The experiment was performed with a shorter sample of speech than before. The modifications to the program meant that it took longer to perform the calculations, and as the speech was only sampled intermittently (enough speech was read in to fill one of the memory fields) it was impossible to analyse exactly the same signal as before. It was anticipated that the different results would be obtained for the maximum number of occupied cells because of this.

4.7.3. Results

The distributions obtained are shown in Fig. 4.16 for both linear and logarithmic quantisation, normal and differentiated speech.

The curves for the differentiated form are flatter for both the linear and logarithmic quantisation. This suggests that the time-intervals in differentiated and clipped speech are more evenly distributed



Clipped Speech

Differentiated and Clipped Speech

Fig. 4.16 Percentage of Contents represented by N Cells with Largest Contents plotted against N

through the range 0-1.6msec than the intervals in clipped speech are through the range 0-3.2msec. The histograms of Fig. 2.4 confirm this; the long time scale required for clipped speech is necessary to accommodate the longer intervals formed by the higher amplitude low frequency components.

As predicted, the curves yield different results for the maximum number of occupied cells and this makes comparison between linear and logarithmic quantisation difficult. A better comparison can be made by examining the figures for the number of occupied cells that constitute 80% of the statistics. The information rate required to specify 80% of the statistics, allowing one bit for the contents is given in Table 4.6.

	Histogram	Digram	Trigram
Normal lin	1,000	4,000	6,600
" log	800	3,200	4,800
Differentiated lin	800	4,000	9,000
" log	1,000	4,400	8,400

Table 4.6

80% was chosen as being sufficiently far from the top of the curve, to be independent of the precise number of occupied cells, yet represent a substantial proportion of the statistics. An incomplete set of statistics will have more 'dead-ends' (see Section 4.5.1). Consequently, the random algorithm will generate more atypical sequences of intervals.

The difference between the channel capacity requirements for the logarithmic and linear quantisation are not very great, except for trigram statistics of normal clipped speech, where a 27% reduction in the information requirements could be accomplished using logarithmic quantisation. The subjective experiments however, showed that logarithmically quantised clipped vowel sounds were not recognisable.

The distributions shown in Fig. 416. and the substantial savings in storage requirements achieved for digram and trigram statistics when 80% of the statistics are considered suggest that the second and third-order statistics are composed of several elements containing very few counts.

4.7.4. Discussion

Although the information requirements have been calculated for a limited amount of speech of one speaker only, they serve to indicate that except for first-order statistics, clipped speech (with and without pre-differentiation) can be stored more economically by suitable quantising of the original signal, than it can by storing statistics of the time-intervals. The most economical way of transmitting the histogram information is to specify the contents of all sixteen bins, rather than specifying the occupied ones only.

As stated previously, no system was designed that would calculate 80% of the time-interval statistics and then synthesise a signal from those statistics. Using a random algorithm for synthesis, several 'dead-ends' would be found, unless the statistics were reduced in such a way as to preserve the continuity. This problem occurs only for statistics of higher order than the first.

Conclusion

The results obtained with the experienced subjects in particular show that vowels synthesised from trigram statistics are recognised nearly as well as the original clipped vowels. The scores for the trigram synthesised vowels show a marked improvement over those produced from first-order statistics.

The sonagrams of the synthetic vowels show that the formant structure of the vowels improved as more sequential information was used in their synthesis. The sonagrams of the phrase that was synthesised show that abrupt discontinuities were generated in the spectral components at segment boundaries. Although these discontinuities were not heard, their existence suggests that the random algorithm is not the best method of using the time-interval statistics.

As far as the quality of the sounds is concerned, the random algorithm is most unsatisfactory as it does not give the impression of voiced sounds. As stated earlier (in Section 2.8), a more deterministic way of selecting the intervals is required to incorporate the constraint, that for a voiced sound, there is a regular repetitive pattern of time-intervals.

The information required to specify the time-interval statistics is, with the exception of first-order statistics, more than that needed to specify the original clipped signal if five bits are used to specify the contents of the cells of the array.

Judged on the grounds of quality, intelligibility and economy of specification, this method of using time-interval statistics to produce voiced sounds leaves much to be desired.

CHAPTER 5

5.1 Introduction

The method of synthesis described in Chapter 4 had three main shortcomings:-

1. The synthesis algorithm was unable to reproduce the glottal excitation of the original waveform.
2. The time-intervals were all weighted equally in the compilation of the statistics whether they were reliable or not.
3. The amount of information required to specify the statistics was in most cases greater than that needed for the original wave.

Attempts to overcome these shortcomings are described in this chapter.

The amplitude-modulated digram display (see Section 3.4.2) showed that if the time-intervals were weighted according to the amplitude of the wave, a much clearer display resulted. Although it was not possible to perform the same kind of weighting using the computer, a gating technique is used to remove intervals where the amplitude is low. Some results are presented to show the effect of this operation on the perception of real (as opposed to synthetic) clipped vowels. This technique is incorporated into the design of two algorithms that produce glottal triggering. The more successful of these two algorithms is modified to allow the synthesis of words and some preliminary results are discussed.

5.2 Pitch - Synchronous Gating

One of the disadvantages of time-interval analysis is that the zero-crossings in regions where the signal to noise ratio is low are liable to be affected by noise and other perturbations of the signal. Several methods for overcoming this have already been discussed in Chapter 3. The solution offered there, to improve the digram display, was to modulate the intensity of the spots with a signal derived from the peak to peak amplitude of the wave during the intervals being displayed. This improved the appearance of the digram display, but was unsatisfactory for deriving information for speech synthesis. An alternative way of removing unreliable intervals was required for use with the computer.

The damping of the formants for most vowel sounds means that the wave is of low amplitude at the end of each glottal period, particularly if the pitch of the speaker's voice is low. This, coupled with the effect of the next glottal excitation, means that time-intervals near the end of the glottal period are likely to change in length. The time-intervals formed by low amplitude parts of the wave are by no means confined to the end of each glottal period, but it was generally observed that the intervals near the end of a glottal period were more variable than those in the earlier part. The effects of the removal of these intervals (pitch-synchronous gating) on the digram statistics, and the perception of clipped vowel sounds are described below.

5.2.1 Method of Pitch Detection

A method of fundamental pitch detection was required that would indicate the beginning of each glottal period, so that the intervals immediately preceding it could be gated out. Although refined techniques such as auto-correlation and cepstrum analysis were known to give accurate pitch measurements (48) they would not give indications of the beginning if each glottal period, without additional information from a peak detecting circuit. As most of the waveforms to be analysed were those of sustained vowels, many of the tracking problems usually associated with pitch extraction techniques were not encountered. The simple level detector circuit described in Section 4.2.5 incorporating a fast-acting automatic gain control operated satisfactorily on all the vowel waveforms that were used in the listening tests. These included /u/, where there was very little damping of the first formant, so that it was sometimes difficult to estimate the glottal period from the oscilloscope trace. Some trouble was experienced in measuring longer glottal periods but this was largely overcome by increasing the 'on' time of the monostable.

A phase-change switch was provided in the processor, and this was always set so that the level detector fired on a positive air pressure level rather than a negative one. (The phasing of all the recording equipment used to prepare the Language Master cards was checked to ensure that thiswas so). A positive level was chosen because most of the waveforms examined started with a positive excursion from zero pressure. One of the exceptions to this was the /I/ waveform shown in Fig. 5.1.(The performance of the pitch detector was investigated with the level detector working on negative levels and was found to give less satisfactory performance due to the asymmetry of the speech wave).

Some of the differentiated vowels revealed two envelope maxima during each glottal period. This was probably due to excitation of the second and third formants at the opening as well as the closing of the vocal chords (27, 46). The pitch extractor was arranged to work on the unprocessed speech signal so that it did not trigger on both these maxima. This meant that the gate always turned on slightly earlier in each glottal period when differentiated speech was used.

5.2.2 Method of Pitch-synchronous Gating

A computer program was written to perform the pitch-synchronous gating. It was written in such a way that the gating could have been carried out in real time using simple circuitry. The computer was used, however, as it enabled different amounts of pitch-synchronous gating to be carried out on the same stored time-intervals. This was particularly useful for examining the effects of pitch-synchronous gating on the digram statistics.

The pitch extractor fired during the first positive-going excursion of the wave in each glottal period. The gate was arranged to turn on at the next zero-crossing pulse. (In this way, the first time-interval of the glottal period was discarded, and the gating introduced no shortening of any time-intervals). This zero-crossing pulse was also used as the beginning of the glottal period, because it was found to be more stable in position than the preceding pulse.

Having established at what point the gate turns on, there are three ways of determining where it should be turned off:-

1) after a certain number of zero-crossings,

2) after a certain length of time,

3) after a certain percentage of the glottal period.

Some earlier experiments with the Digitimer at producing pitch-synchronous gated sounds employed the technique of counting zero-crossings. This was not very satisfactory as small changes in the waveform increased or decreased the number of zero-crossings from one glottal period to the next. The 'on' time of the gate fluctuated accordingly, imparting an additional roughness to the speech. Moreover, it was difficult to make comparisons between the vowels as the number of intervals in a glottal period varied considerably from vowel to vowel. (Six intervals in /u/ might occupy the complete glottal period, whereas six in /æ/ might occupy only 2msec.

Methods (2) and (3) are merely different ways of specifying the length of time for which the gate is held on. After discussions with J.B. Millar, it was decided that the pitch-synchronous gating would be performed by making the gate switch at the next zero-crossing after a pre-determined length of time. This ensured that no intervals were effectively changed in length by the action of the gate. This method has the advantage over method (1) in that the gate 'on' period is more constant in length from one glottal period to the next. Using method (2) it was only the next zero-crossing after a specified length of time that determined when the gate turned off, and not all the previous zero-crossings in the glottal period.

The clipped speech output was controlled by a four diode gate operated by a bi-stable (see Fig. 4.2.). The clipped speech input to

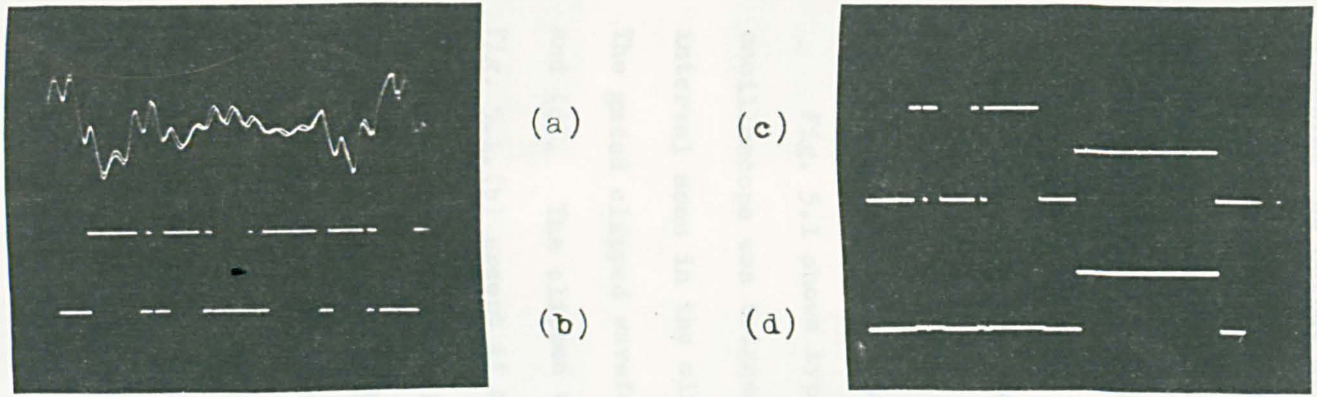


Fig. 5.1 Pitch-synchronous Gating on /I/ Waveform

(a) Original Waveform

(b) Clipped Waveform

(c) Pitch-synchronous Gated Waveform

(d) Gating Waveform

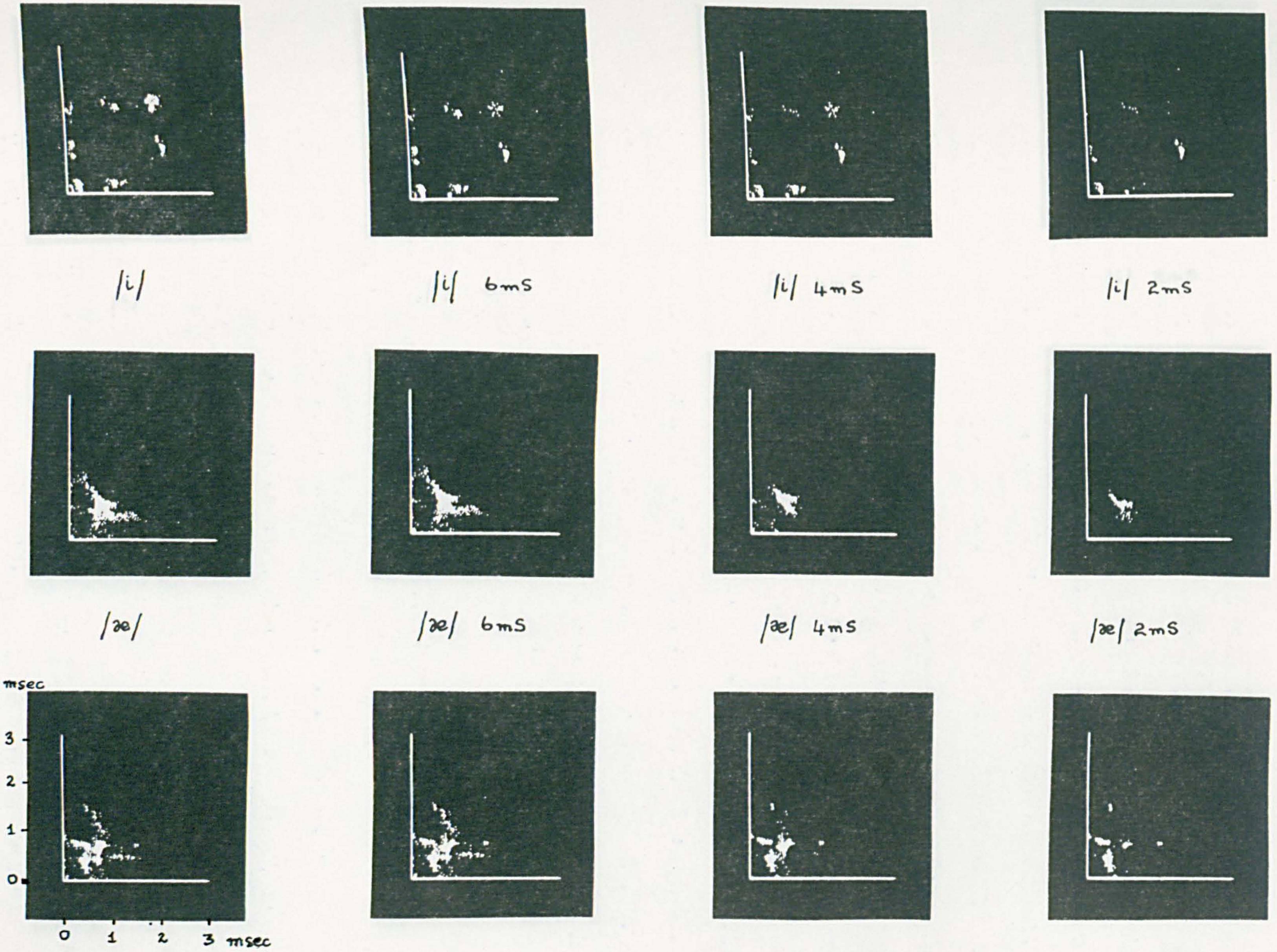
the gate ran between ± 1 volt, so that when the gate was turned off, clamping the output at 0 volts, a three level signal was produced. This was done so that the amount of speech in each glottal period could be more accurately controlled. If the input signal had used 0 volts as one of its two levels, the gate could have been turned off with the clipped speech already at 0 volts. The effective length of the time-interval would then have been from when the signal last changed from its '1' state to zero, until the instant when the gate opened again, With long intervals, this could have introduced a considerable error.

Fig. 5.1 shows typical waveforms obtained. In Fig. 5.1(a) the oscilloscope was triggered by the pitch extractor, so that the first interval seen in the clipped version (Fig. 5.1(b)) is incomplete. The gated clipped waveform and gating signal are shownin Fig. 5.1(c) and (d). The clipped waveform is the same as the clipped wave in Fig. 5.1.(b) except it has been reversed in phase, to be in the same phase as the original speech signal.

5.2.3. Effects of Pitch-synchronous Gating on Digrams

The digrams of pitch-synchronous gated vowels are shown in Fig.5.2, and those of differentiated vowels in Fig. 5.3. The digrams of the complete waveform are shown on the left-hand side. The figures under the remaining digrams refer to the amount of speech that was included in each glottal period for the compilation of the digram. The photographs relate to 500msec of speech.

Both figures show that there is very little difference between the digrams of the complete waveform and those compiled with 6msec

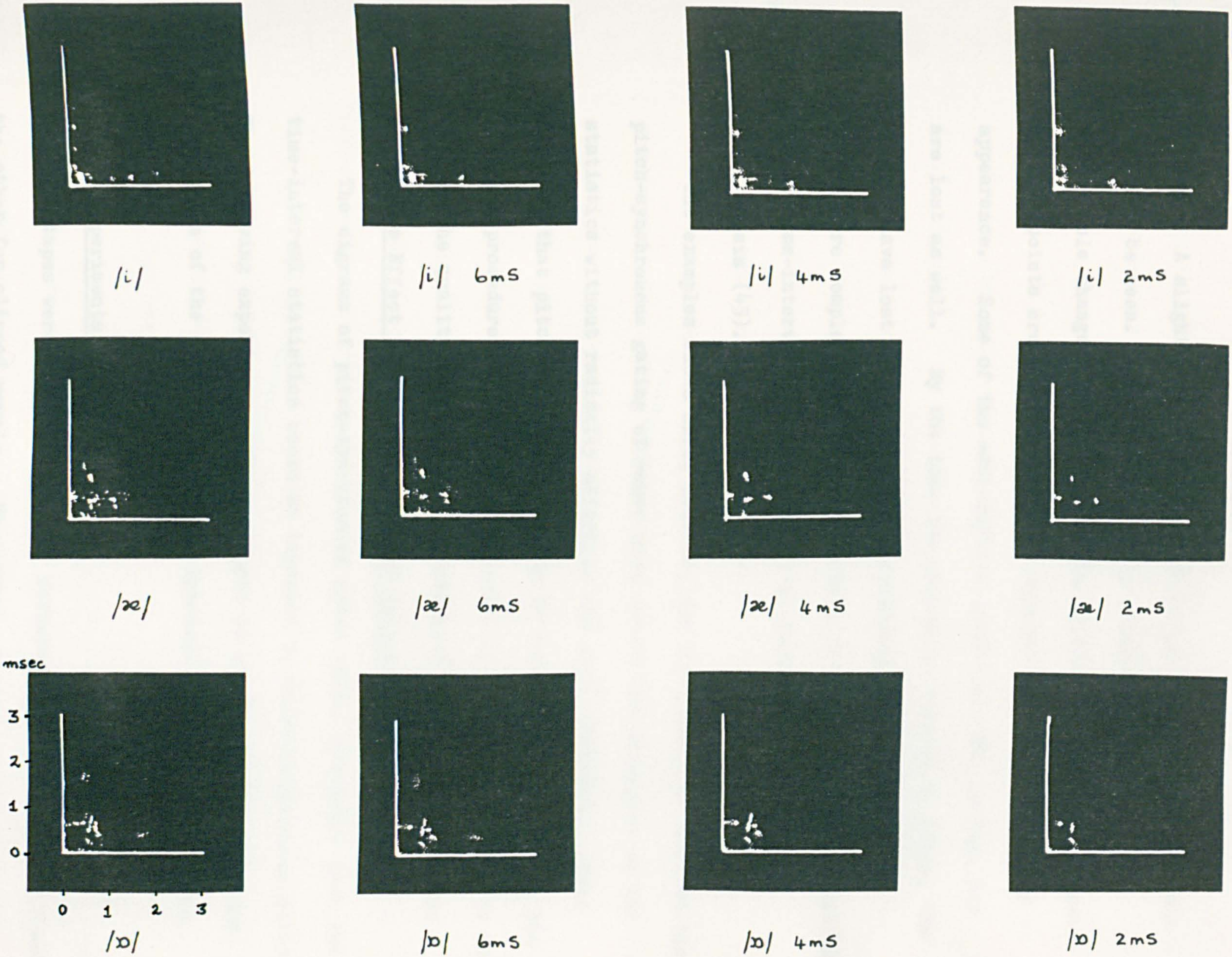


/i/ | /i/ 6mS | /i/ 4mS | /i/ 2mS

/æ/ | /æ/ 6mS | /æ/ 4mS | /æ/ 2mS

/ɒ/ | /ɒ/ 6mS | /ɒ/ 4mS | /ɒ/ 2mS

Fig. 5.2 Digrams of Pitch-synchronous Gated Vowels



/i/ /i/ 6mS /i/ 4mS /i/ 2mS

/æ/ /æ/ 6mS /æ/ 4mS /æ/ 2mS

/ɒ/ /ɒ/ 6mS /ɒ/ 4mS /ɒ/ 2mS

Fig. 5.3 Digrams of Pitch-synchronous Gated Differentiated Vowels

gating. A slight reduction in the blurring of some points is all that can be seen. When the gating is reduced to 4msec, quite appreciable changes are visible in the digrams. Some of the more diffuse points are lost, leaving the digrams with a much cleaner appearance. Some of the well-defined points of /æ/ in Fig. 5.3 are lost as well. By the time the gating is reduced to 2msec, the digrams have lost more of their characteristic points.

A more complete review of the effect of pitch-synchronous gating on the time-interval statistics is to be found in J.B. Millar's Ph.D. thesis (45).

The examples shown here, however, are sufficient to indicate that pitch-synchronous gating of 4msec does reduce the noisiness of the statistics without radically affecting the major features. This suggests that pitch-synchronous gating is worth incorporating in the analysis procedures to reduce the information storage requirements and improve the quality of the vowels synthesised from the statistics.

5.2.4 The Effect on the Perception of Clipped Vowels

The digrams of pitch-synchronous gated vowels indicated that the time-interval statistics could be improved by pitch-synchronous gating. The following experiments were performed to investigate whether the perception of the clipped vowels was improved by pitch-synchronous gating.

5.2.5 Experiments 6 and 7

Two tapes were prepared, one for differentiated and clipped vowels, the other for clipped vowels. The amount of pitch-synchronous gating was varied in 1msec steps and eight groups of vowels were ordered as follows:- complete, 1msec, 6msec, 3msec, 7msec, 4msec, 5msec, 2msec, complete. The subjects for the experiments were nine unpaid undergraduates with a small amount of experience of listening to clipped vowels (two previous experiments).

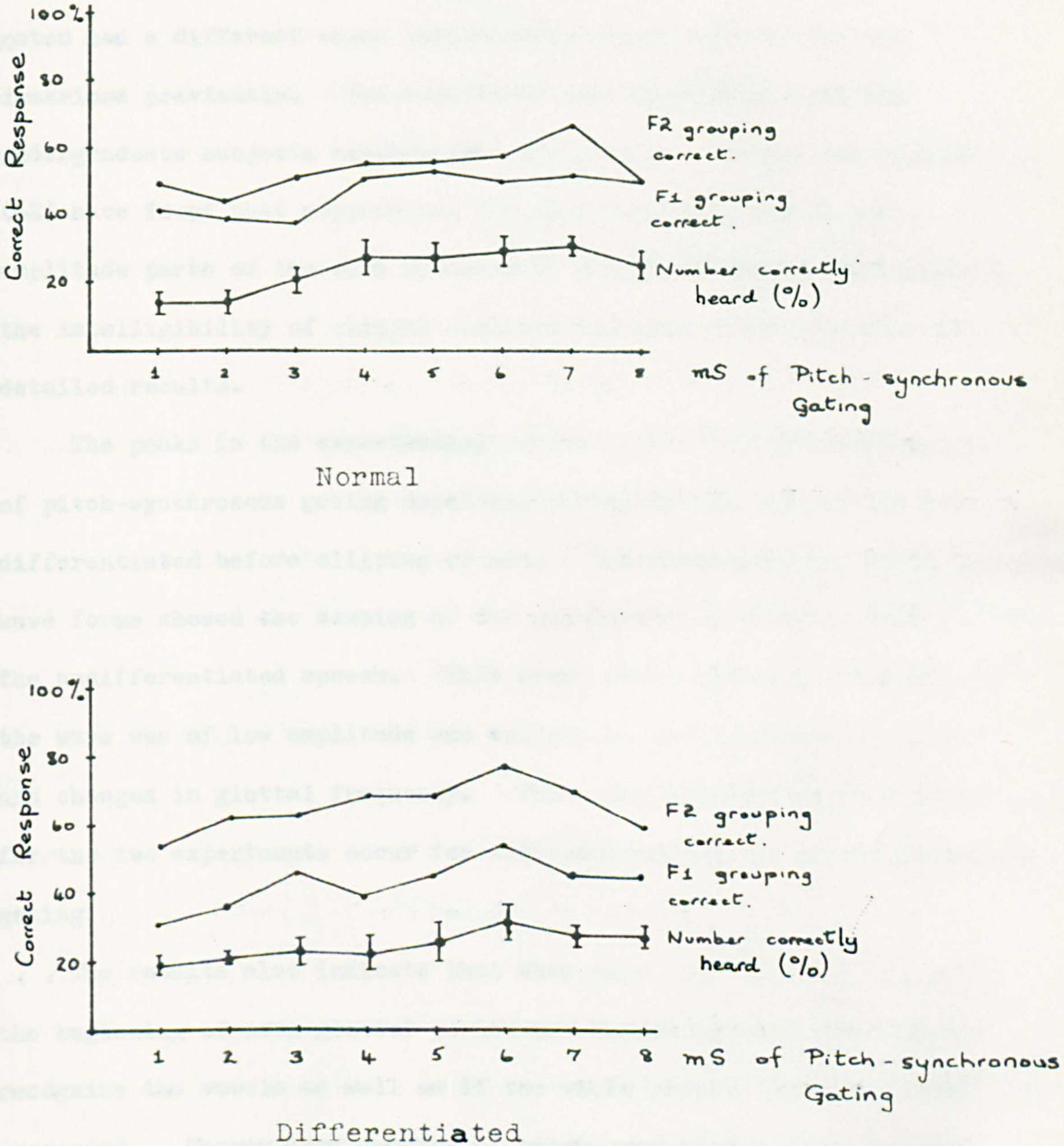
5.2.6. Results

The results of the two experiments are shown in Fig. 5.4. The scores for 8msec are the scores for the complete clipped vowels with no pitch-synchronous gating. As expected, all the scores improve as the amount of speech within each glottal period is increased. The striking feature of both sets of results is the peak in the curves, occurring at 7msec for clipped vowels and at 6msec for differentiated and clipped vowels. (The peak at 3msec in the F1 grouping curve for clipped vowels is thought to be of no significance in view of the small number of subjects used.)

The high frequency emphasis produced by differentiation has improved the F2 grouping score but has had very little effect on the number heard correctly or the F1 grouping score.

5.2.7 Discussion

The most interesting feature of the results is the lower scores that are obtained when there is no pitch-synchronous gating. This indicates that the time-intervals near the end of the glottal period are in some way atypical of the vowel sound, so that better scores are obtained when they are removed. In some preliminary experiments with pitch-synchronous gating using the Digitimer the operation of the gate



Normal

Differentiated

Fig. 5.4 Results of Perceptual Experiments with Pitch-synchronous Gated Vowels

was reversed so that only those intervals near the end of the glottal period were heard. The author found that clipped vowels sounds so gated had a different vowel quality from those gated in the way described previously. The results of the experiments with the undergraduate subjects confirm this indirectly. Tanaka and Okamoto (62) have found that suppressing the zero-crossings during low amplitude parts of the wave by means of a high frequency bias improved the intelligibility of clipped Japanese syllables, but published no detailed results.

The peaks in the experimental curves occur for different amounts of pitch-synchronous gating depending on whether the speech had been differentiated before clipping or not. Examination of the differentiated wave forms showed the damping of the waveform to be greater than for the undifferentiated speech. This meant that a greater proportion of the wave was of low amplitude and subject to perturbations by noise and changes in glottal frequency. This could explain why the peaks for the two experiments occur for different amounts of pitch-synchronous gating.

The results also indicate that when only 3 or 4msec of speech at the beginning of each glottal period are heard, subjects are able to recognise the vowels as well as if the whole clipped vowel had been presented. Narrow band spectra of pitch-synchronous gated clipped vowels were prepared; two typical examples are shown in Fig. 5.5.

In both cases no well defined peaks are seen until 3msec of pitch-synchronous gating. As more intervals are included the spectral

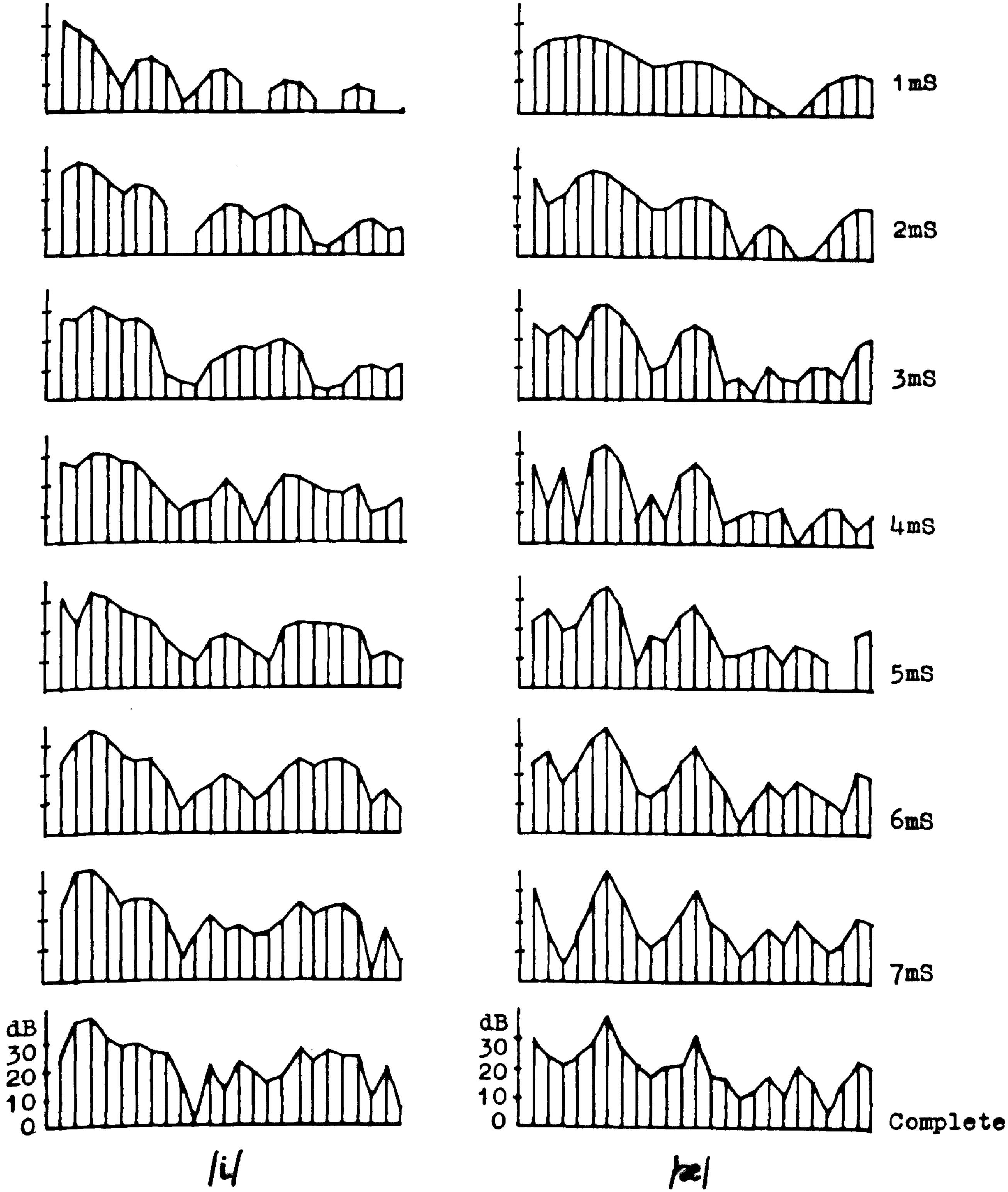


Fig. 5.5 Narrow Band Spectra of Pitch-synchronous Gated Vowels (Fundamental Pitch 120 Hz)

peaks remain in the same position, though there are signs of some inter-modulation components. It was found that the positions of the peaks corresponded well to the formant peaks of the original sounds. In some cases the shape of the spectral distributions for pitch-synchronous gated clipped vowels was closer to that of the original vowels than the spectrum of the complete clipped vowel was e.g. 7msec /æ/.

David and McDonald (10) reported that when more than 50% of a glottal period of a vocoder output was removed, the speech became unintelligible. As they gave no details of the glottal frequency and damping of the filters used, it is difficult to compare their results with those obtained with isolated clipped vowels. Morozov (47) found that the intelligibility of isolated vowels decreased as the glottal frequency was increased above 300 Hz. Stover (61) has interpreted this as meaning that only 3msec of speech at the beginning of each glottal period need be presented to ensure reasonable intelligibility of the speech signal. The results with the isolated clipped vowels would support this, though more experiments with different speakers are needed to show its generality.

5.3 Implications for Synthesis

The results of the perceptual experiments with pitch-synchronous gating indicate that there are certain time-intervals or groups of intervals that can be considered as atypical of the vowel sound. The removal of these intervals in the compilation of time-interval statistics has two advantages:-

1) The statistics refer to a more reliable set of intervals so that it should be possible to generate more realistic synthetic clipped speech using pitch-synchronous gating, particularly if some glottal triggering is incorporated in the synthesis of voiced sounds.

2) The information storage requirements are reduced, as some of the atypical intervals have been eliminated.

The experiments also indicate that there is no need to produce complete glottal periods of intervals. This overcomes some of the problems outlined in Chapter 2.

It seems desirable therefore, to incorporate pitch-synchronous gating into the routines for the compilation of time-interval statistics and the synthesis of a new wave from them.

5.3.1 Glottal-triggered Algorithm

The random algorithm described in Section 4.5.1 was a poor model for producing voiced sounds from time-interval statistics. Although the sounds from which the statistics were compiled were voiced, third-order statistics provided insufficient information for the glottal excitation to be reproduced. A different statistical model for selecting the time-intervals was therefore required, which would incorporate glottal triggering as part of the algorithm. The perceptual experiments with pitch-synchronous gated vowels had shown that recognisable clipped vowels could be produced that had only a few time-intervals specified within a glottal period. It seemed likely therefore, that incorporating:-

1) some degree of pitch-synchronous gating into both the analysis and synthesis parts of the program, and

2) some means of glottal triggering or excitation would produce more realistic clipped vowel sounds.

5.3.2 Method

The analysis and synthesis parts of the program described in Section 4.5.1 were modified to incorporate pitch-synchronous gating and glottal triggering.

The amount of pitch-synchronous gating was made variable in 1msec steps. In the compilation of digram and trigram statistics, the last interval that was 'on' in one glottal period was assumed to be followed immediately by the first interval of the next glottal period, i.e. the 'off' intervals were completely ignored. The algorithm was designed in this way to prevent 'dead ends' being reached in every glottal period even though it brought together intervals which normally did not occur next to each other.

In order to produce glottal triggering, an extra histogram was compiled of the first time-interval occurring in all the glottal periods examined. The synthesis routine chose the first time-interval of each glottal period at random from this histogram. The algorithm then chose 'on' intervals at random according to the order of statistics being used, until a pre-determined number of milli-seconds of speech had been produced. An 'off' interval was then calculated to make the time upto the measured glottal period. The first interval of the next glottal period was chosen at random from the histogram of first-intervals, and the process repeated until 500msec of speech had been produced. In this way, the algorithm was triggered regularly around a likely sequence of intervals.

5.3.3. Quality of Vowels

The voiced effects produced by the glottal-triggered synthesis algorithm when applied to the set of twelve vowel sounds were not very convincing. In general a steady voiced effect was not heard, except when only two or three time-intervals were produced in a glottal period, using digram or trigram statistics. The algorithm made little difference to the histogram synthesised vowels except when the 'on' period was occupied entirely by the first interval. As the length of the 'on' period was increased, the random nature of the signal became more apparent, and the one steady interval at the beginning of each glottal period was masked. The same was true for the vowels synthesised from digram or trigram statistics, except that the same two or three intervals were produced at the beginning of every glottal period.

Because the first interval was not always of the same length and classified into the same bin, the following intervals were not always the same from one period to the next. It sometimes happened that intervals of the same length occurred in different parts of the glottal period of the original wave, followed by intervals of very different lengths with the result that the synthetic wave sounded as if it was not voiced. Suppose the original sequence had been ABCDADBAC, etc. As the algorithm contained a random element, it could just as well produce a sequence ABCA as ADBC as ACDA. This effect would be made worse if the first interval A was sometimes of length B.

The lack of a fundamental pitch was particularly noticeable for the vowels that had been differentiated prior to clipping, as in general they had many more time-intervals.

5.3.4 Experiment 8

Only one listening test was carried out with synthetic vowels produced by the glottal-triggered algorithm, because of the failure of the algorithm to produce a convincing voiced effect.

A test tape was made to test synthetic vowels produced from histogram, digram and trigram statistics with 4msec of intervals in each glottal period. 4msec was chosen as Experiments 6 and 7 had shown that clipped vowels with 4msec of pitch-synchronous gating were as recognisable as the complete clipped vowels.

The subjects were five unpaid undergraduates who had taken part in the previous pitch-synchronous gating experiments.

5.3.5. Results

The results of the experiment are shown in Fig. 5.6. The horizontal bars at the right refer to the scores the subjects obtained with the original clipped vowels.

The number of sounds recognised correctly falls very slightly as the order of the statistics is increased, but this cannot be considered significant in view of the small number of subjects. (The results for the histogram synthesis show a greater spread, reflecting greater uncertainty). The scores for the synthetic vowels fall well below the scores for the original clipped vowels, though it should be noted that the number of clipped vowels heard correctly was higher than usual.

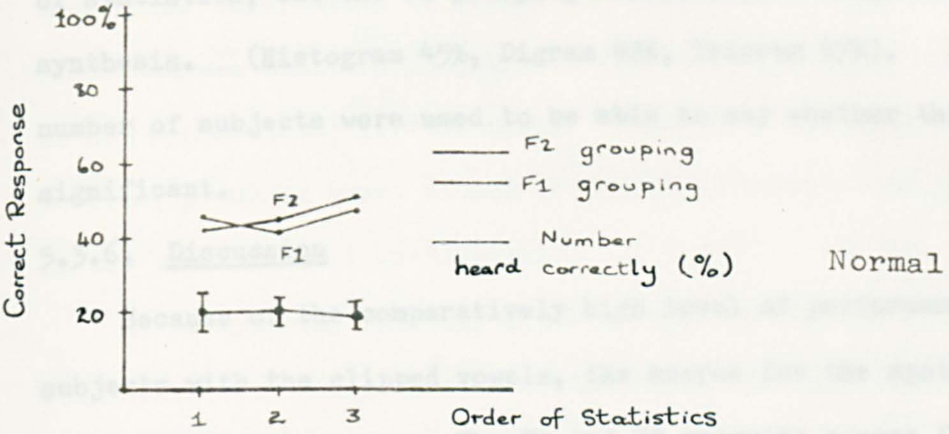


Fig. 5.6 Results of Perceptual Experiments with Glottal-triggered Algorithm

The correct F2 grouping scores improve steadily with the order of statistics, but the F1 grouping scores show a slight dip for digram synthesis. (Histogram 45%, Digram 42%, Trigram 47%). An insufficient number of subjects were used to be able to say whether this dip is significant.

5.3.6. Discussion

Because of the comparatively high level of performance of the subjects with the clipped vowels, the scores for the synthetic vowels appear relatively poor. The F1 and F2 grouping scores for synthesis are, however, as good as the F1 and F2 grouping scores obtained by some of the same subjects for clipped vowels in an earlier experiment. (Experiment 6 - pitch-synchronous gating). When this comparison is made, it indicates that subjects are able to recognised clipped vowels synthesised from third-order statistics as well as they are the clipped vowels. This finding is in agreement with the results of Experiment 4.

The better performance of the subjects with the clipped vowels in this experiment is, however, unexplained. Compared to Experiment 4, the percentage of vowels identified correctly has risen from 25% to 40%; the F1 and F2 grouping scores have risen from 51% and 52% to 55% and 65% respectively. The F1 and F2 grouping scores have not risen as much as the percentage identified correctly. This suggests that the effect could be due to the small number of subjects used, although the grouping scores are less susceptible to fluctuations caused by similar vowels being confused.

In view of the poor quality of the glottal-triggered sounds, no further experiments were carried out using this algorithm. Even though glottal triggering had been introduced, there was insufficient sequential information to remove the random nature of the signal. The algorithm was modified slightly to reproduce the first glottal period it had synthesised, for 500 msec. Successive attempts at synthesis, however, produced vowel sounds of different quality due to differences in the ordering of time-intervals.

It is unlikely that any great savings in storage requirements would have been found in view of the extra histogram needed to store statistics of the first interval in each glottal period.

5.4 Maximum Algorithm

The experiments with the random and glottal-triggered algorithms had confirmed the earlier finding with the time-interval generator of the need for selecting the time-intervals in a more deterministic way. The synthesis of clipped vowels using the Digitimer (Section 2.8.2) had been encouraging in that realistic sounding vowels could be produced with only three intervals in a glottal period, if these intervals were representative of the peaks of the histogram. The peaks of the histogram had been selected by eye, and the intervals arranged to produce the most realistic sound. This process would have been difficult to implement on the computer without designing a algorithm that performed some feature extraction on the statistics. The histograms and digrams that are illustrated in Chapters 2 and 3 show that the form of the statistics varies considerably from vowel to vowel making the construction of a general algorithm very complex.

5.4.1 Method

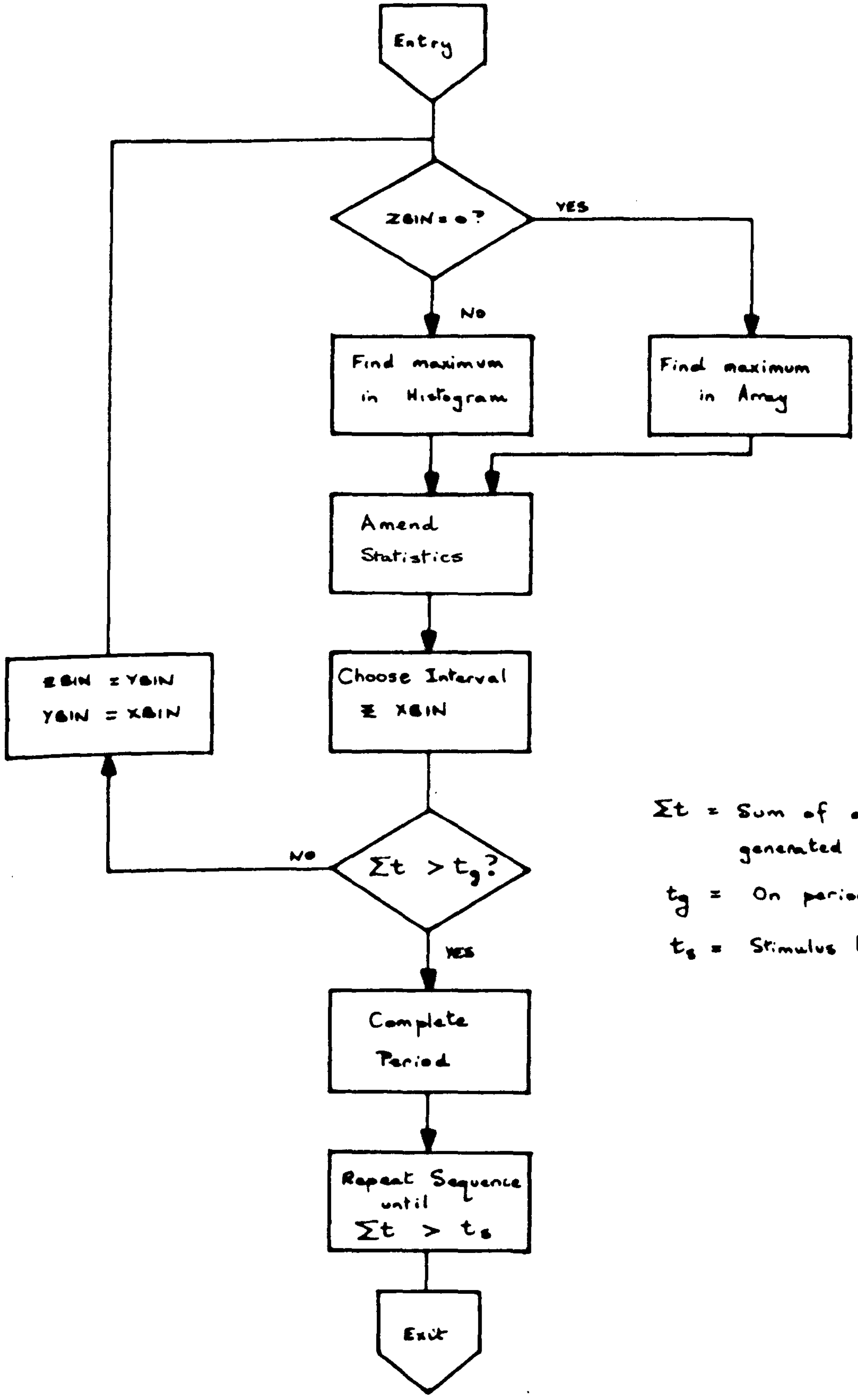
A simple algorithm was designed in the following way. The major features of both histograms and digrams were the peaks in the distributions. It was likely that the intervals that comprised these peaks were important in giving the original clipped vowel sound its colour. The maximum algorithm to be described is based on the hypothesis that the maximally occurring intervals, or groups of intervals occur near the beginning of the glottal period, and the less likely ones near the end. The waveforms of clipped vowel sounds showed that intervals near the end of a glottal period were more susceptible to noise. If the statistics were compiled over a long enough period, these noisy intervals would be spread amongst the cells of the statistics and would not produce significant peaks. By selecting intervals from the elements with the greatest contents the synthesis algorithm will perform some kind of pitch-synchronous gating by discriminating against unlikely intervals or groups of intervals.

The algorithm cannot work by continually choosing an interval corresponding to the element with the greatest contents without modifying the contents, as it would continually be choosing the same interval or groups of intervals. The algorithm therefore works in the reverse way to the compilation routine by subtracting a number from each element every time it produces an interval corresponding to that element. To achieve the correct normalisation, the number subtracted is equal to the number of glottal periods examined in compiling the statistics. Pitch-synchronous gating is applied to the synthetic clipped vowel, and the same intervals are repeated at the glottal rate to produce a steady voiced-like sound.

The flow chart of the maximum algorithm for synthesis from trigram statistics is shown in Fig. 5.7. To determine the first interval in the glottal period, the complete array is examined to find the element with the greatest contents. The co-ordinates of this element (ZBIN, YBIN, XBIN) are calculated and its contents reduced by the number of glottal periods of speech analysed. An interval is selected according to the value of XBIN and the values of ZBIN and YBIN take the previous values of YBIN and XBIN. To determine the next interval the histogram defined by ZBIN and YBIN is examined to find the element with the greatest contents. If two elements have the same contents, the element corresponding to the longer interval is selected.

The process continues until enough intervals have been produced to fill the 'on' period. An interval is chosen to complete the glottal period (measured when the original speech was analysed). That synthetic sequence of intervals is then repeated until the required stimulus length has been exceeded. The length of the synthetic utterance is always made as long as the length of time over which the statistics were compiled.

The algorithms for histogram and digram synthesis are correspondingly simpler. The three versions of the algorithm were all written so that they stopped if they came to a histogram with zero contents, as this was an indication that something was wrong. In practice, this never occurred for the vowel sounds analysed over 500 msec. When shorter analysis times were used, the trigram synthesis routine was the most likely to stop (Section 5.5.2).



Σt = Sum of all intervals generated so far.

t_g = On period

t_s = Stimulus length.

Fig 5.7 **Maximum Algorithm Synthesis**

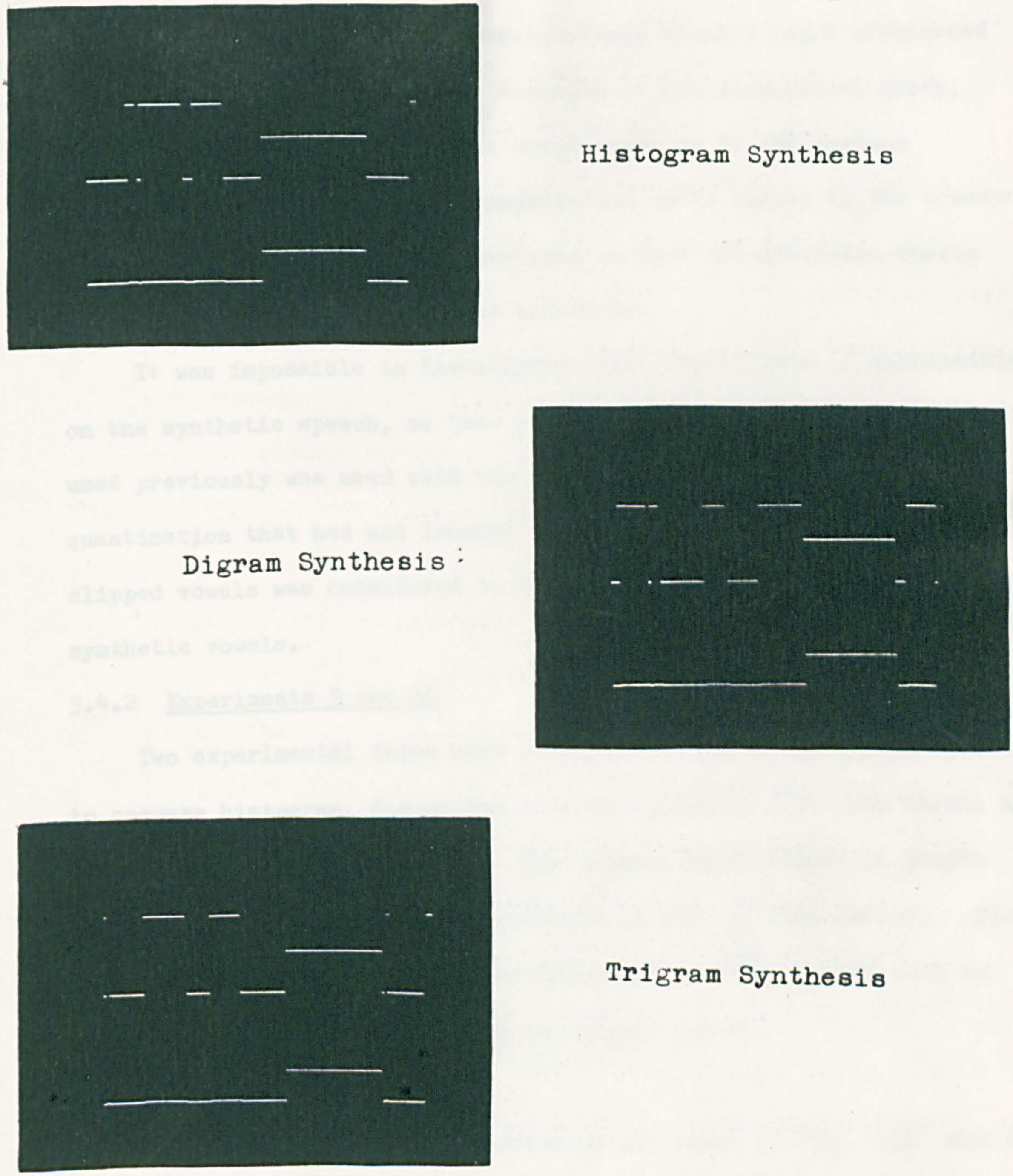
This algorithm was considered to have the following advantages:-

1) It discriminated against unlikely intervals or groups or intervals and thus performed a kind of pitch-synchronous gating.

2) Although all the elements of the statistics were examined to determine the element with the greatest contents, the contents of some elements were never decreased. (This occurred because less speech was synthesised than used to compile the statistics.) As the algorithm never selected those elements, they could just as well have been zero. This has the effect of reducing the storage requirements without creating 'dead ends'. For example, the contents of all elements could have been reduced by a fixed number or ratio, but this would have produced 'dead ends' in the digram or trigram statistics. The maximum algorithm effectively accomplishes a reduction without introducing 'dead ends'.

3) It produces a sound with a steady pitch.

The maximum algorithm can be thought of as reducing the information storage requirements by drawing a new zero-line or threshold across the statistics, so that all elements with contents less than this threshold are ignored. This threshold arises because the statistics are compiled from all the time-intervals occurring, while the synthesis algorithm selects intervals for a portion of each glottal period only. Thus the elements of the statistical arrays are incremented more than they are decremented. This means that the relative contents of elements with contents greater than the new zero are altered, e.g. the ratio 10:8 is not the same as the ratio 8:6. It was anticipated that this alteration in the relative importance of elements would not affect the validity of the algorithm greatly.



Histogram Synthesis

Digram Synthesis

Trigram Synthesis

Fig. 5.8 Synthetic /I/ Waveforms produced by Maximum Algorithm

The quantisation of the time-intervals clearly has a pronounced effect upon the contents of the elements of the statistical array, and this will in turn modify the sound produced by the maximum algorithm. Thus logarithmic quantisation would result in the broader bins having relatively larger contents so that the synthetic vowels would be dominated by the longer intervals.

It was impossible to investigate fully the effects of quantisation on the synthetic speech, so that the same quantisation that had been used previously was used with the maximum algorithm. A particular quantisation that had not lowered the intelligibility of the original clipped vowels was considered to be the most likely to produce realistic synthetic vowels.

5.4.2 Experiments 9 and 10

Two experimental tapes were prepared of synthetic clipped vowels to compare histogram, digram and trigram synthesis from both normal and differentiated clipped vowels. The stimuli were 500msec in length as before, with 4msec of time-intervals in each glottal period. The subjects were five members of the University teaching staff with no previous experience of listening to clipped speech.

5.4.3 Results

The results of the two experiments are shown in Fig. 5.9. For the undifferentiated speech, the number of vowels identified correctly falls as the amount of sequential information is increased. This was mainly due to /ɜ/ being heard as /ɔ/ for the vowels synthesised from trigram statistics. /ɜ/ and /ɔ/ have similar first formant frequencies, so

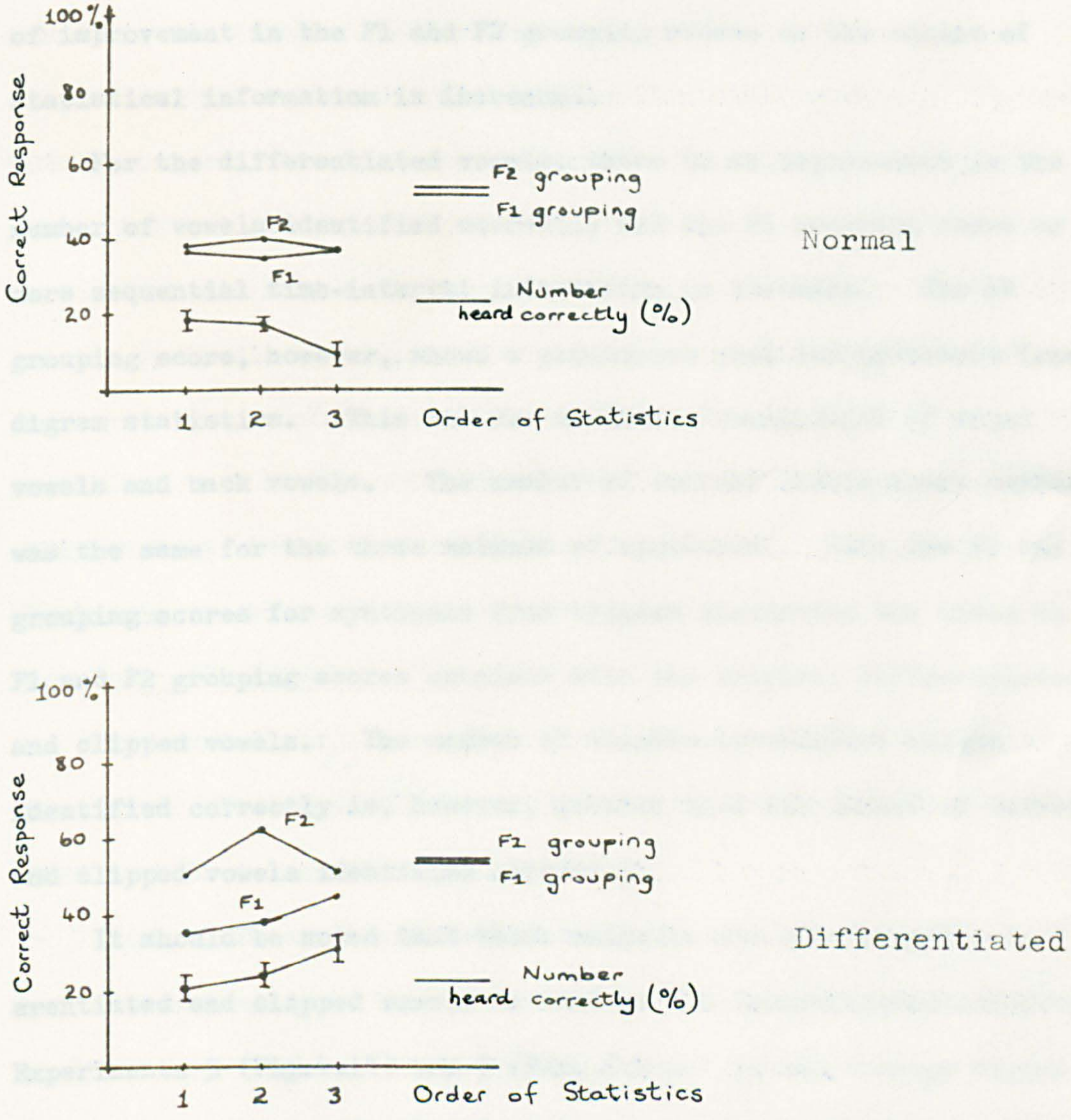


Fig.5.9 Results of Perceptual Experiments with Maximum Algorithm

that confusion between them is not unexpected. The other striking feature of the results for the undifferentiated vowels is the lack of improvement in the F1 and F2 grouping scores as the amount of statistical information is increased.

For the differentiated vowels, there is an improvement in the number of vowels identified correctly and the F1 grouping score as more sequential time-interval information is included. The F2 grouping score, however, shows a pronounced peak for synthesis from digram statistics. This was due to better recognition of front vowels and back vowels. The number of central vowels heard correctly was the same for the three methods of synthesis. Both the F1 and F2 grouping scores for synthesis from trigram statistics are close to the F1 and F2 grouping scores obtained with the original differentiated and clipped vowels. The number of trigram synthesised vowels identified correctly is, however, greater than the number of differentiated and clipped vowels identified correctly.

It should be noted that these subjects did not recognise differentiated and clipped vowels as well as the inexperienced subjects in Experiments 5 (Fig.4.14) and 7 (Fig. 5.4). If the average figure for the differentiated and clipped vowels recognised correctly is used (30%), it will be seen that the score for synthesis from trigram statistics is approximately equal to it.

5.4.4 Discussion

Although the experiments were carried out with a small number of subjects for only one speaker's voice, the results for the undifferentiated speech show that the sequential information provided by the higher-order statistics does not make the vowel sounds any more recognisable. Moreover, the level of performance for histogram synthesis is almost as good as that achieved for the original clipped vowels. The author found, however, that his performance with the synthetic vowels was far below that he achieved with the original clipped vowels (typically three, four and five right out of twelve, for histogram, digram and trigram synthesised versions compared to ten or eleven right out of twelve for the original clipped vowels.) This confirmed the subjective impression that higher order statistics made less difference to the vowel quality when the maximum algorithm was used. Using the random algorithm, the author's scores were typically three, five and seven right out of twelve.

No formal listening tests were performed using other speakers' voices, but similar scores were obtained by the author using sets of vowel sounds of two different speakers.

Examination of the subjects' responses showed that most of the synthetic and original clipped vowels (75%) were classified as /ɜ/, /ɑ/, /ɔ/ or /u/. /ɜ/ was heard for /ɜ/, /ə/, /ʌ/, and /ʊ/; /ɑ/ for /æ/, /ʌ/ and /ɑ/; /ɔ/ for /ɒ/ and /ɔ/; /u/ for /i/ and /u/. These confusions are all ones that can be explained in terms of similar formant frequencies. The confusion between /i/ and /u/ arises because the clipping process discards most of the second formant information in /i/, leaving an /u/-like sound.

The subjects' responses for the differentiated vowels showed a bias towards long vowels except that there were very few confusions between /i/ and /u/; /i/ was more often confused with /I/. No explanation of the peaks in the F2 grouping curve could be found. Further experiments would be necessary to test its significance. Indeed more testing is necessary to investigate fully the usefulness of the maximum algorithm.

5.4.5 Information Storage Requirements

No detailed experiments were carried out to investigate the storage requirements for the elements in the arrays actually used by the maximum algorithm. It was suggested in Section 5.4.1 that the maximum algorithm effectively produced a new zero-line or threshold below which all elements were considered to have zero contents. The algorithm had therefore extracted some features from the statistics, which meant that less information was required to specify them.

The statistics of the clipped vowel sounds were analysed to determine the economy in storage space provided by the maximum algorithm. Table 5.1 shows the average number of cells occupied in a 20msec segment by all the vowel sounds.

Number of occupied cells	Histogram	Digram	Trigram
before synthesis	9	21	30
after synthesis	4	8	7

Table 5.1

The amount of information required to specify this number of bins is 800, 3,200, and 4,200 bits/second for histogram, digram and trigram respectively.

These preliminary results indicate that a substantial saving in the number of elements that need to be specified for voiced sounds can be accomplished by using the maximum algorithm. The figures, however, relate to the average number of elements that need specifying and were obtained from a very small repertory of speech sounds. If continuous speech were analysed, the overall saving would not be as great as it would be inappropriate to use the maximum algorithm for unvoiced sounds.

5.5 Synthesis of a Phrase

The results of the experiments with the maximum algorithm indicated that the synthetic vowels were nearly as recognisable as the original clipped vowels if the speech was not differentiated before clipping. (The author was not able to recognise the synthetic vowels as well as the original clipped vowels). Thus one of the objectives outlined in Chapter 2 had been achieved for one speaker's voice only. Clearly there was a need to extend the method to the analysis and synthesis of a whole range of speakers' voices. However, it was felt that a system which was only capable of producing isolated vowels was a long way short of producing connected speech.

The computer analysis and synthesis programs were modified to allow a preliminary investigation of the synthesis of connected speech from time-interval statistics.

5.5.1 Method

Fig. 5.10 shows a flow chart for the modified system. The maximum algorithm was clearly unsuitable for synthesis of unvoiced sounds.

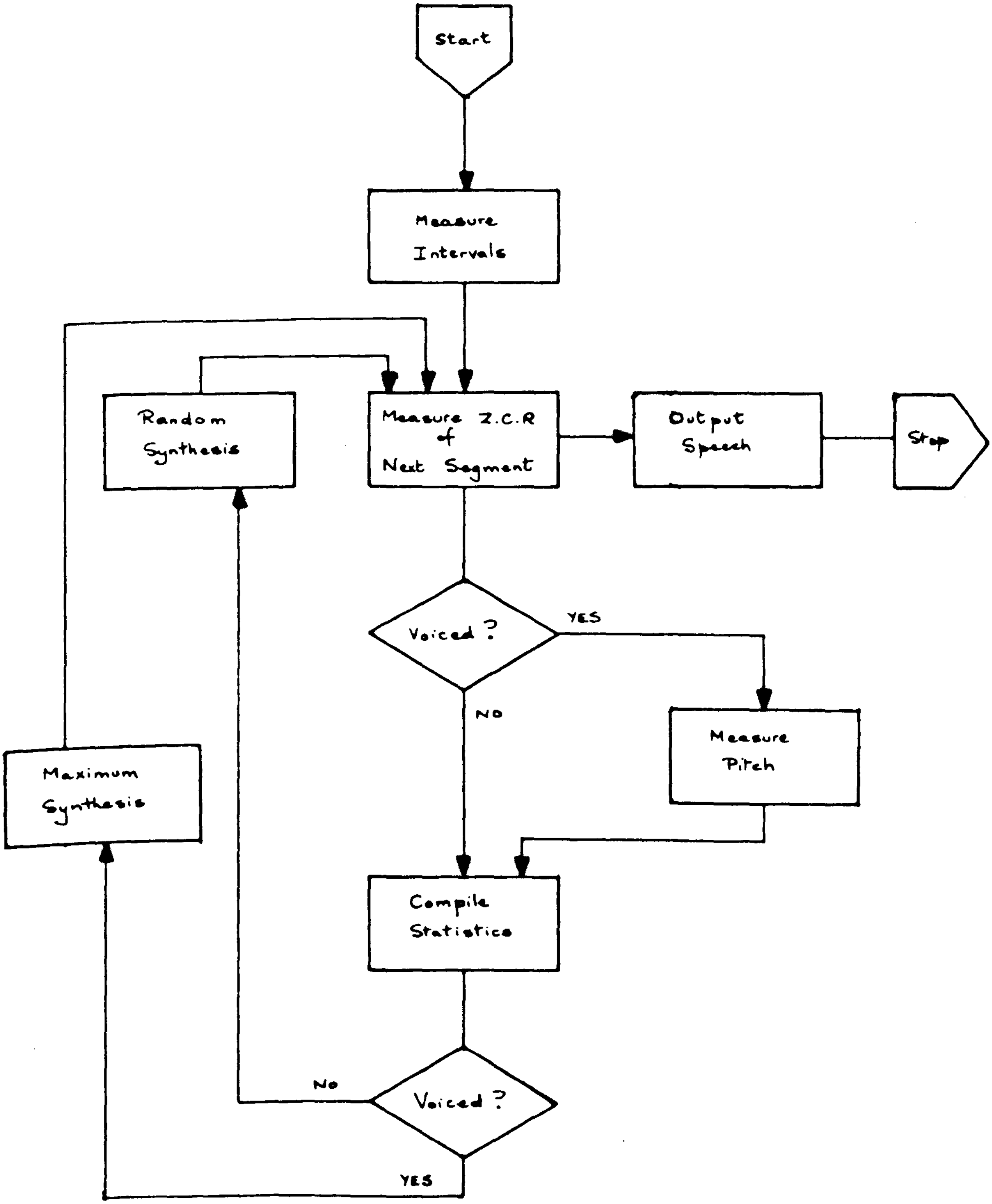


Fig.5.10 Voiced/Unvoiced Analysis-Synthesis

Earlier experiments had indicated that realistic fricative sounds could be produced using the random algorithm. It was therefore necessary to distinguish between voiced and unvoiced sounds. Liljencrants (37) has shown that a voiced/ unvoiced decision can be made fairly reliably by measuring the mean zero-crossing rate. If every zero-crossing pulse is considered, a count of more than 3,200 zero-crossings per second means that the sound is unvoiced. This criterion was employed in the computer program to determine which synthesis algorithm to use.

The number of zero-crossings in a 20msec segment was examined to determine if the sound was voiced. If it was, the pitch of the voice was measured and the length of the segment increased so that it included a whole number of glottal periods. When the time interval statistics had been compiled the appropriate algorithm was selected and a new time-segment of intervals was generated. The process of analysis and synthesis of segments of speech continued until all the original speech had been processed. The synthetic version was then played.

5.5.2 Results

The voiced/unvoiced analysis-synthesis system was first applied to the synthesis of isolated vowel sounds.

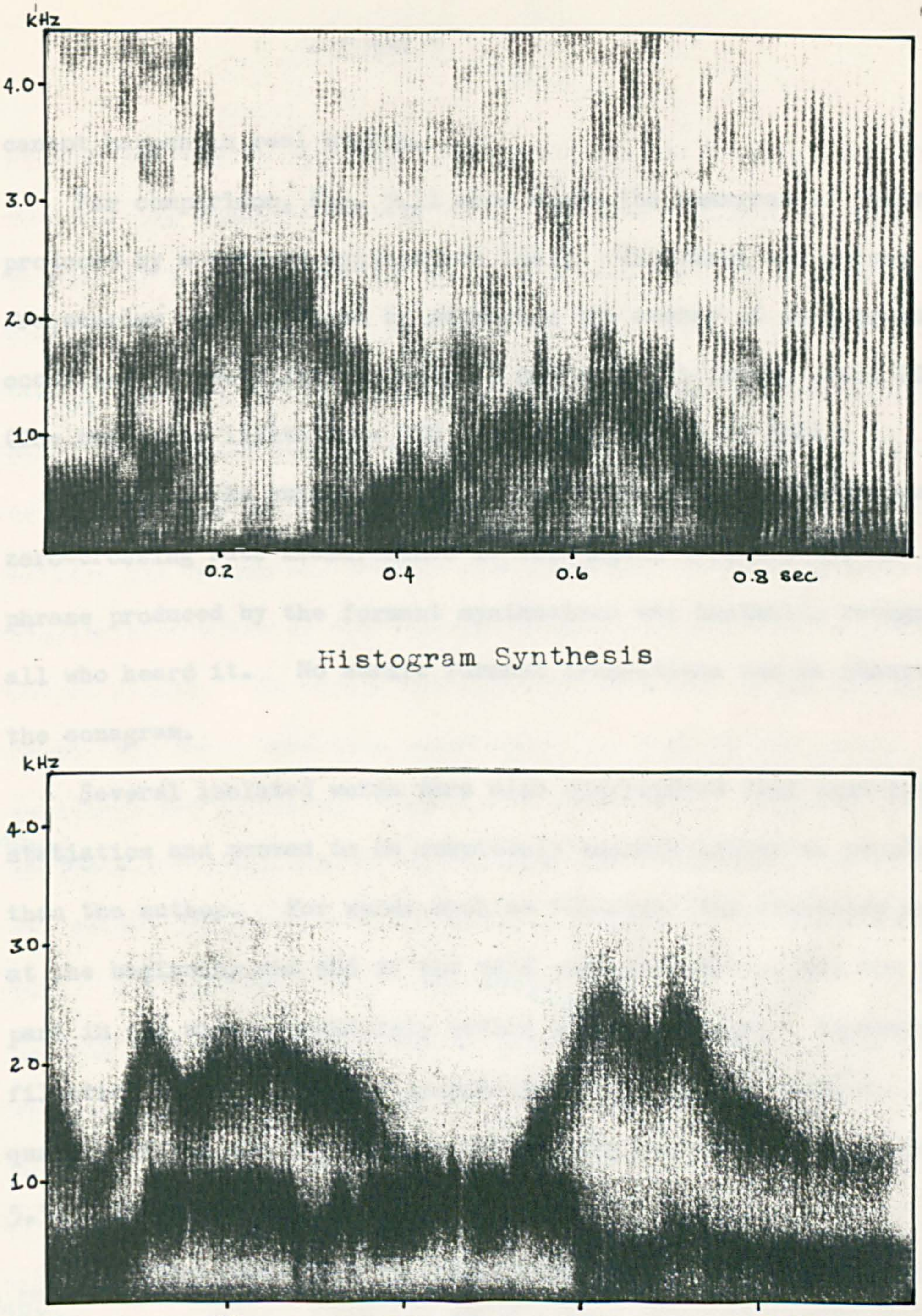
Instead of the steady sound that had been produced when the time-segment was 500msec long the synthetic vowels had a rough irregular quality. This effect or warble was present irrespective of whether histogram, digram or trigram statistics were used. The effect was more noticeable for those vowels that had a comparatively large number of intervals in each glottal period. It was found that the warble could be decreased by producing less intervals in each glottal period, but only in a few cases, e.g. /u/, could it be completely eliminated. Sonagrams of the synthetic vowels showed that the warble in the sound was probably due to adjacent time-segments having different spectral components. Although the statistics were compiled from contiguous segments there was sufficient variation from one segment to the next for the maximum algorithm to produce different sequences of intervals in adjacent segments. The warble was caused by discontinuities at the segment boundaries. If the segments were made long, e.g. 200msec, the discontinuities at the segment boundaires could be heard distinctly. Shortening the length of the time-segment merely increased the pitch of the warble. The shortest that a time segment could be made was the length of one glottal period and the warble could still be heard.

Moreover, the synthesis from digram and trigram statistics often came to a halt (Section 5.4.1). As there were only a few occupied cells in the arrays, and only a percentage of all the cells were examined in the selection of each new time-interval, it was possible for the routine to encounter a 'dead end'.

Whilst the warble was noticeable and detracted from the quality of the vowel sound, it was not known how important this would be when a longer piece of speech was synthesised. The phrase "Where are you?" was processed using the voiced/unvoiced analysis-synthesis system.

The quantisation of the speech was made 150μsec so that fricative sounds could be produced more realistically. This meant that all of the time-intervals in speech greater than 2.4msec in length were categorised in one bin. As very few were found in the words analysed, this was not considered an important limitation. 4msec of time-intervals were produced in each glottal period and the time segment had to be increased to 30msec to prevent the routine halting during synthesis from digram and trigram statistics. The phrase was played to several people, most of whom were able to recognise it eventually. There was very little difference between the histogram, digram and trigram versions. (The results of Experiment 9 showed that there was little difference between the synthetic vowels produced from histogram, digram and trigram statistics.) It was impossible to synthesise a differentiated form of the phrase to see if sequential time-interval information improved the synthetic speech.

Sonagrams of the histogram, digram and trigram synthesised versions were prepared and found to be very similar. The histogram version is illustrated in Fig. 511. The general form of the sonagram is similar to that of the original clipped utterance except that the 'formant' regions are in a different position and discontinuities in the formants can be seen at the time-segment boundaries. These discontinuities imparted a rough unnatural quality to the speech, making recognition of the phrase difficult. The abrupt transitions of the formants from one time-segment to the next are most unnatural as they correspond to instantaneous changes in position of the articulators, something that



Histogram Synthesis

Formants derived from Mean Zero-crossing Rate

Fig. 5.11 Broad Band Sonagram of Synthetic "Where are you?"

cannot happen in real speech.

For comparison, Fig. 5.11 also shows the sonagram of "Where are you?" produced by a formant synthesiser (28). The first and second formant frequencies were obtained by measuring the number of zero-crossings occurring in two filtered bands of the original phrase every 20msec. (The band-pass limits were 100 to 900 Hz, 900 Hz to 3KHz).

Although the unreliability of deriving formant frequencies from zero-crossing rate measurements is well known (49, 50, 55, 63) the phrase produced by the formant synthesiser was instantly recognised by all who heard it. No abrupt formant transitions can be observed in the sonagram.

Several isolated words were also synthesised from time-interval statistics and proved to be completely unintelligible to people other than the author. For words such as 'Charles' the fricative sounds at the beginning and end of the word were detectable, but the voiced part in the middle completely defied identification. Low-pass filtering of the synthetic speech at 5KHz, helped to improve the quality of the speech, but the intelligibility was not improved.

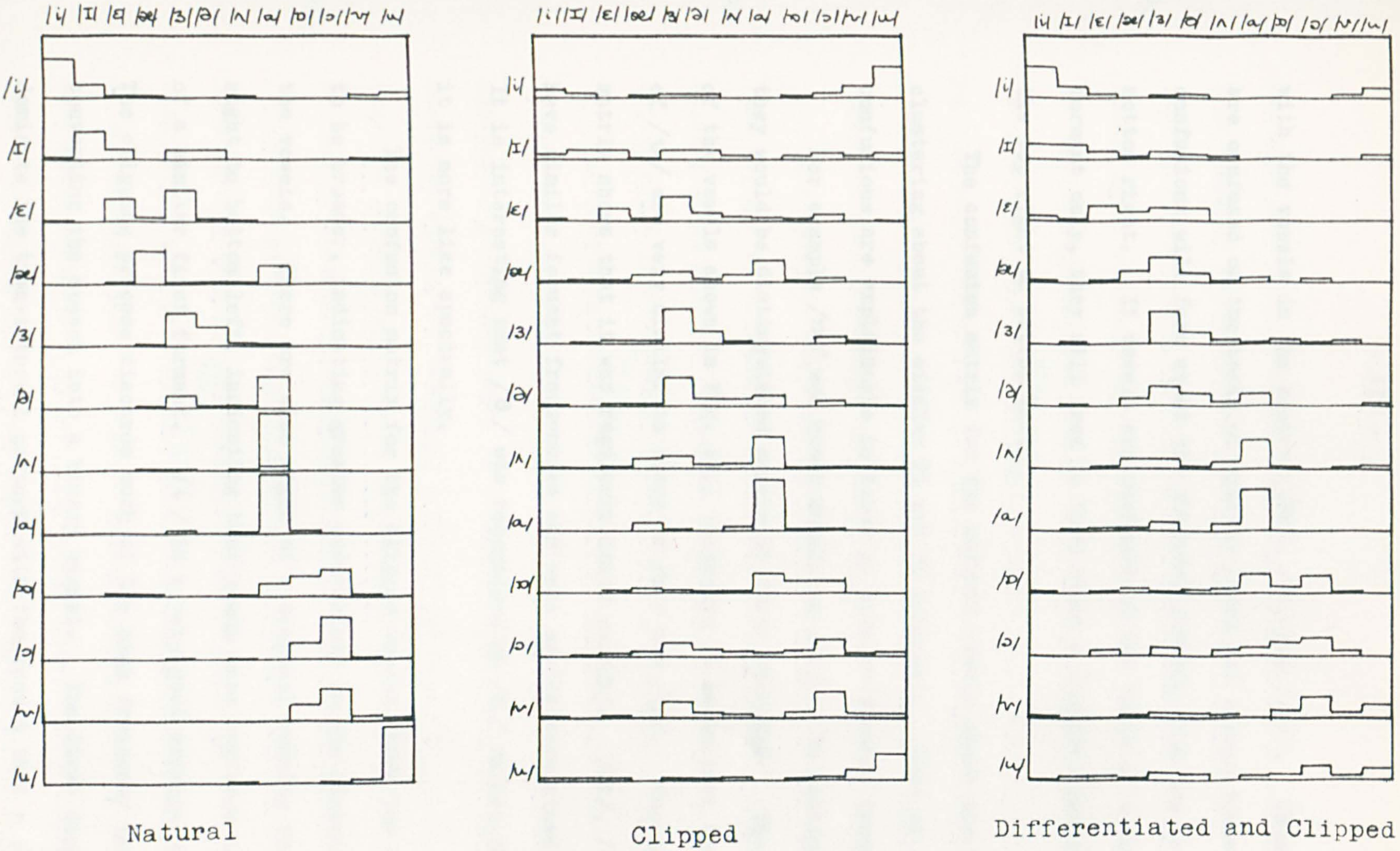
5.5.3 Discussion

The discontinuities in the formant regions observed in the sonagrams of the phrases synthesised from time-interval statistics suggest that the statistical model used for synthesis is not good enough. The greatest weakness is the failure to track continuous movements of the articulators. There are also discontinuities in what might be termed intermodulation components, e.g. harmonics of the 'formants'. The discontinuities in these too, will be perceptible to a human listener. To a certain extent they can be removed by filtering, but the basic problem of giving an indication of the positions of the articulators remains.

It is significant that the pitch-synchronous gating and the repetition of the same glottal period have been used by Stover (61) to accomplish time-domain bandwidth-compression. The major difference between his system and the one described here is that he does not compile statistics of the time-intervals from which to derive his output signal. He claims 70% intelligibility with the modified rhyme test for a bandwidth requirement of 1,200 bits/second. Although no exhaustive listening tests were performed with the voiced/unvoiced analysis-synthesis system, the preliminary experiments carried out by the author indicated that the intelligibility of PB words was around 10%. The inference to be drawn from the comparison is that the maximum and random algorithms do not produce a sufficiently good approximation to the original clipped speech wave.

5.6 Perception of Isolated Vowel Sounds

A large amount of data was obtained on the responses of subjects to isolated vowels, both natural and clipped. Confusion matrices are shown in Fig. 5.12. The results have been normalised so that the separation of the horizontal lines represents 100% correct response. The columns represent the perceived vowels, the rows the vowels presented. The vowels have been placed in order of descending F2, which means that low F1 is associated with the top and bottom of the matrix and high F1



Natural

Clipped

Differentiated and Clipped

Fig. 5.12 Confusion Matrices for Natural and Clipped Vowels

with the vowels in the centre (/æ/, /ʌ/ and /ɑ/). Thus if vowels are confused on the basis of similar first and second formants, confusions will form about the diagonal running from top left to bottom right. If vowels are confused on the basis of a similar first formant only, they will tend to form about a diagonal running from the top right to bottom left.

The confusion matrix for the natural vowels shows the results clustering about the similar F1 and F2 diagonal. Most of the confusions are explainable in terms of similar formant frequencies.

For example /ɒ/ was heard mainly as /ɔ/. In continuous speech, they would be distinguished mainly by their duration. The F1/F2 plot of the vowels shown in Fig. A3.1 (Appendix 3) shows that the formants of /ʊ/ are very similar to those of /ɒ/ and /ɔ/. The confusion matrix shows that it was regularly heard as /ɔ/. /æ/, /ʌ/ and /ɑ/ have similar formant frequencies and were mostly recognised as /ɑ/. It is interesting that /ə/ was recognised as /ɑ/ rather than /ɜ/ which it is more like spectrally.

The confusion matrix for the clipped vowels shows the distributions to be broader, indicating greater uncertainty in the classification of the vowels. There are also signs of a diagonal running from top right to bottom left, indicating that vowels were confused on the basis of a similar first formant. /ɪ/ is a very good example of this. The clipping process discards much of the high frequency information by converting the speech into a binary signal. The first formant tends to dominate the time-interval pattern with the result that a clipped /i has lost some of its timbre and sounds more like /u/. /ɛ/ is most commonly recognised as /ɜ/; they have similar first formants. /ə/ is mostly recognised as /ɜ/. This contrasts with the natural vowels where /ə/ was mostly recognised as /ɑ/. /æ/, /ʌ/ and /ɑ/ are mostly recognised as /ɑ/. It is interesting to note that the perception of /ɒ/ has moved from /ɔ/ towards /ɑ/. /ɔ/ is recognised mainly as an /ɔ/ or /ɜ/, which have similar first formants. Similarly for /ʊ/. /u/ is mainly recognised correctly with a small spread of results amongst the other vowels.

This analysis indicates that the perception of formants or formant-like regions is an important factor in the perception of isolated clipped vowel sounds. Where the clipping process has discarded information, there is a tendency for vowels with a similar F1 to be confused.

The confusion matrix for the differentiated and clipped vowels is more scattered in nature than that for the natural vowels, but is similar to it in that there is a tendency for the peaks to cluster about the similar F2 diagonal. The effect of differentiating the waveform prior to clipping is to emphasise the high frequency components, so that F2 becomes the dominant frequency component of the clipped wave.

A much greater proportion of the /i/'s are heard as front vowels, than was the case for the clipped vowels. /æ/ and /ə/ are confused with /ɜ/ - they have similar second formant frequencies, whilst /ʌ/ is still mainly heard as /ɑ/. (The second formants of /ʌ/ and /ɑ/ were measured to be 1,300 and 1,000 Hz respectively.) /ɒ/ was heard mainly as /ɑ/ or /ɔ/. The three vowels have similar second formant frequencies, though the stimulus length of 500msec probably indicated a long vowel /ɑ/ or /ɔ/ rather than /ɒ/. /ʊ/ and /u/ were mainly recognised as /ɔ/ - they have similar second formant frequencies. (The second formant amplitude of /u/ was so low that when the speech was differentiated, the sound was nearly lost in the noise and it was difficult to discern any /u/-like quality when it was clipped).

Most of the confusions found for the isolated clipped vowels with and without differentiation before clipping indicate the importance of the formant concept. Where the clipping has removed information about a high frequency component, a vowel is heard as another one with a similar first formant. When the speech is differentiated before clipping emphasising the high frequency components, confusion exists between vowels with a similar second formant. The apparent domination of the time-interval pattern by the strongest formant has already been pointed out in Chapter 2. The confusions between vowels with similar formants suggests that it is the dominant formants that are retained which determine the vowel colour.

5.6.1 Variation in Results

Considerable variation was found from one group of subjects to another, in the number of clipped vowels they were able to identify correctly. To a certain extent this was due to the small number of subjects used. (Considerable difficulty was encountered in finding subjects who were willing to come more than once or twice. As they were unpaid - except for cups of tea and coffee - there was very little incentive for them to continue coming to experiments which were both

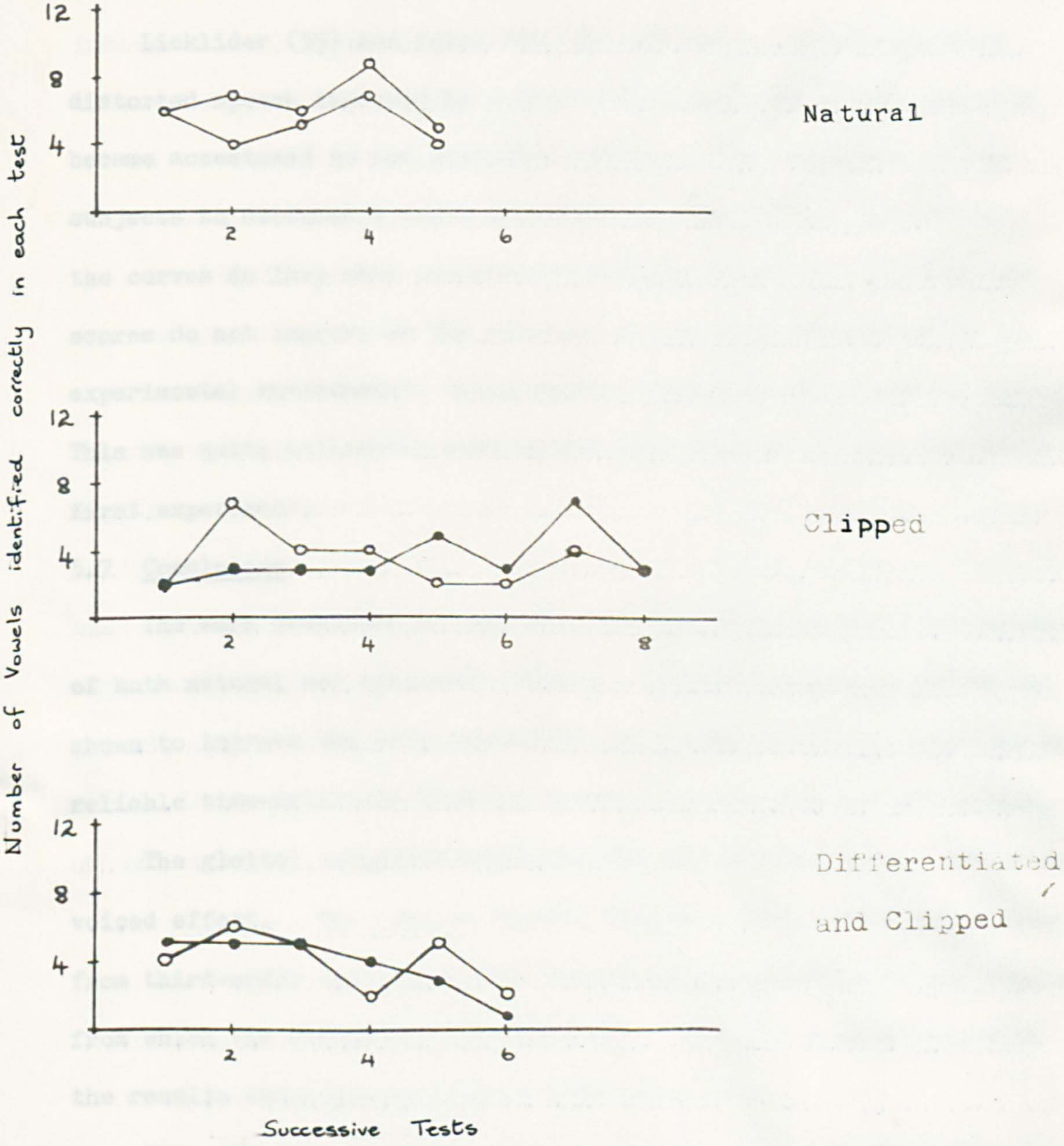


Fig. 5.13 Responses of Subjects to Successive Tests

difficult and not very interesting after the initial novelty had worn off).

Licklider (35) had found that his subjects' performance with distorted speech improved by a factor of around 50% as the subjects became accustomed to the distorted speech. The responses of two subjects to successive tests are shown in Fig. 5.13. In none of the curves do they show pronounced learning effects. Although the scores do not improve as the subjects become more accustomed to the experimental environment, their speed of response was found to increase. This was quite noticeable even within five minutes of starting their first experiment.

5.7 Conclusion

The work described in this chapter has improved the intelligibility of both natural and synthetic vowels. Pitch-synchronous gating was shown to improve the intelligibility of clipped vowels by removing unreliable time-intervals from the latter part of each glottal period.

The glottal triggered algorithm did not produce a very realistic voiced effect. The results showed, however, that vowels synthesised from third-order statistics were nearly as recognisable as the vowels from which the statistics were derived. This is in agreement with the results obtained using the random algorithm.

When the time-intervals were chosen in a more deterministic way there was little improvement in the intelligibility of the normal synthetic vowels as the amount of sequential information was increased. The vowels synthesised from histogram statistics were nearly as recognisable to naive listeners as the original clipped vowels. If the speech was differentiated before clipping,third-order statistical information was necessary to bring the scores up to the level of the original signal. The preliminary results on the amount of information needed to specify the elements used by the algorithm showed a considerable reduction for digram and trigram statistics. Thus for isolated steady-state sounds the maximum algorithm was reasonably successful.

It was found to be inadequate for the synthesis of continuous speech, because of its inability to track movements of the articulators continuously. The discontinuities which had passed unnoticed when the random algorithm was applied (because of the noisy nature of the signal) were very noticeable when synthesis was executed according to the maximum algorithm. Although the maximum algorithm was simple in nature, it is difficult to see how it could be improved, since some of the discontinuities were found to be caused by the change in length of only one interval in the 'on' period.

CHAPTER 6

6.1 Time-Interval Analysis

Although the work described in this thesis has been mainly concerned with the synthesis of speech-like sounds from time-interval statistics, one or two points emerge regarding time-interval measurements.

The reciprocal time-interval/frequency relationship that holds for a simple waveform is no longer valid for complex waveforms such as speech, where there are several frequency components in the signal being analysed. The peaks in the time-interval histograms do give an indication of the formant frequencies, but they are not an accurate measure of them. In a recent paper, Scarr (55) has shown that the first time-interval after the start of each glottal period of a speech wave filtered to remove the second and higher formants does give a reliable measure of the first formant frequency.

The poor noise immunity of time-interval measurements means that time-intervals that are not characteristic of the signal are given an equal weighting in the compilation of the statistics with those that are. The amplitude-modulated digram display showed that a much clearer digram resulted when the time-intervals were weighted according to peak to peak amplitude of the wave. A subsequent experiment by P.D. Green within the Department of Communication has shown that amplitude-modulated clipped vowels are slightly more recognisable than ordinary clipped vowels.

Pitch-synchronous gating which is a less sensitive method of suppressing unreliable time-intervals was also found to improve the intelligibility of clipped vowels when the time-intervals at the end of each glottal period were removed.

6.2 Synthesis of Steady-State Sounds

The three algorithms that were used for the synthesis of clipped speech from time-interval statistics were found to be applicable in different ways.

The random algorithm was not a satisfactory method for the production of voiced sounds because it did not incorporate the constraint that a sequence of time-intervals should be regularly repeated. The results of the listening tests showed that the vowels synthesised from trigram statistics were nearly as recognisable as the original clipped vowels. This is an interesting finding as the synthetic vowels did not have the voiced quality of the original. The sonagrams of the vowels synthesised from trigram statistics revealed dark bands corresponding to the first and second formants. Although no formal listening tests were conducted, fricative sounds retained their distinctive quality when the quantisation of the time-intervals was made sufficiently fine. The random algorithm represents a better approximation to the way fricative sounds are produced than it does to voiced sounds.

The glottal-triggered algorithm was devised to produce more realistic voiced sounds. It was found to produce a voiced effect for only those vowels with comparatively few time-intervals in each glottal period. As with the random algorithm, there were too many time-intervals in the glottal periods of most vowels for the algorithm to produce consistently an appreciable part of each glottal period.

The maximum algorithm was found to be the most successful method of producing isolated vowel sounds. The clipped vowels produced from histogram statistics were recognised almost as well as the original clipped vowels. Where the speech had been differentiated prior to clipping, the vowels synthesised from trigram statistics were nearly as recognisable as the original differentiated and clipped vowels.

The common finding obtained with all these synthesis algorithms is that third-order statistics are sufficient to produce synthetic vowels that are classified almost as well as the clipped vowels from which the statistics were derived.

6.3 Synthesis of Words and Phrases

When the random algorithm was used to synthesise words and phrases, a definite improvement was heard in the quality of the sound as more sequential time-interval information was utilised. The words and phrases were nevertheless, still unintelligible. The sonagrams of the synthetic utterances showed that the synthetic signals had a better formant structure as the amount of statistical information used in their synthesis was increased. Fairly good continuous formant transitions were observed, and the few discontinuities that were found were not heard, as they were masked by the inherently noisy nature of the signal.

Although the maximum algorithm had produced isolated vowel-like sounds that were nearly as recognisable to phonetically naive subjects as the original clipped vowels, it did not produce intelligible voiced sounds in words and phrases. This was mainly attributed to the discontinuities in the signal at the segment boundaries, although sonagrams of the synthetic utterances also showed that the formant bands were displaced frequency-wise with respect to their positions in the original utterance. In view of the ear's sensitivity to changes in an auditory signal, it is possible that the perceptual effects of the discontinuities could be reduced by the use of finer quantisation of the time-intervals. A smaller difference would then exist between intervals representing adjacent bins of a histogram. It is more likely, however, that the discontinuities are produced by slight changes in the statistics from one segment to the next causing a long interval to be chosen in one segment and a shorter one in the next segment. This could still arise if the maximum algorithm was used with a finer time-interval quantisation, as it is a limitation of the algorithm rather than the quantisation.

In evaluating the synthesis of speech from time-interval statistics, it is interesting to examine the work described in a recent paper by Stover (61). Although the aim of his work is bandwidth-compression, he employs similar techniques to those used in this work. To achieve bandwidth-compression, he utilises the redundancy of the speech waveform. He discards redundant information in three stages. The first stage involves clipping of the speech and quantising at a 100μsec rate. He then employs a pitch-synchronous gating technique to ignore all but the first 3msec of a glottal period, and achieves further economy by transmitting information about every fifth glottal period. At the receiver, the first 3msec of that glottal period is then re-circulated at the glottal frequency to produce the output signal. Unvoiced sounds are produced by circulating the information at a pseudo-random rate.

There are three major differences between his scheme and the one described here. The first is that he is able to transmit information at a much slower rate from the transmitter to the receiver than if he sent the whole signal. He achieves 70% recognition of words in the modified rhyme test with an information rate of 1,200 bits/second. This is in contrast with 800 bits/second required to transmit one bit about each element of a histogram every 20msec.

The second difference is that he employed some amplitude-modulation of the reconstituted clipped signal at the receiver. Unvoiced sounds were produced at a lower volume level than voiced sounds, and the amplitude of the wave in each glottal period decreased exponentially with time. There is very little evidence to show that amplitude-modulation of clipped speech improves its intelligibility appreciably. Cherry (7) has reported that modulating clipped speech with the envelope of the original wave improves the naturalness, but has very little effect upon the intelligibility. The experiments of Licklider (35) showed that the processing of clipped speech had a much smaller effect on the intelligibility than the processing of the original wave before clipping.

The third and most significant difference is that Stover employed direct transmission of certain time-intervals rather than statistical approximations. His system (VONAX) was judged to give acceptable results. It is tempting to deduce that the poor quality and intelligibility of the speech synthesised from time-interval statistics is because too much information is discarded in compiling n-gram statistics. It was proposed earlier (Section 2.9), however, that the major problem in synthesising clipped speech from time-interval statistics was in the development of a suitable algorithm for selecting the time-intervals.

To produce intelligible synthetic speech, the ear has to be supplied with a signal comparable in key respects with that from another human being. Both of the algorithms evidently lacked some of the key features required. The random algorithm produced a very noisy signal, and the maximum algorithm did not given an impression of continuous movement of the articulators. Further algorithms need to be tried and tested with more than one speaker before the synthesis of clipped speech from time-interval statistics can be shown to be impracticable. The indications are that a successful synthesis algorithm would have to take into account the signal in the preceeding and following segments in order to produce adequate continuity of the signal from one time-segment to the next.

6.4 Information Storage Requirements

The amount of information required to specify the results of a statistical analysis of speech is clearly dependent on the quantisation of the time-intervals, the length of time over which the statistics are compiled and the nature of the statistics (histogram, digram or trigram). It also depends on the characteristics of the speaker's voice to a certain extent. The indications from the limited set of measurements that were made are that the statistical specification of clipped speech is not informationally concise, unless the measurements are restricted to first-order statistics only.

6.5 The Perception of Clipped Speech Sounds

The ordering of the time-intervals in a synthetic clipped speech sound is a key factor in its perception by a listener. The results of synthesis using the random algorithm show that as the amount of sequential time-interval information was increased, the intelligibility and formant structure of the sounds also increased. The synthesis of words and phrases using the maximum algorithm produced abrupt discontinuities in the perceived signal at the segment boundaries. These discontinuities could be seen in the sonagrams of the synthetic utterances. Most of the confusions between the natural clipped vowel sounds, both normal and differentiated, could be explained as confusions between vowels with either similar first or second formant frequencies.

There was insufficient data from the listening tests with the synthetic vowels to make an exhaustive study of the frequency spectra of the synthetic vowels worthwhile. Nevertheless, the indications from the perceptual experiments point towards the importance of the frequency spectrum in the perception of clipped speech sounds. This is in agreement with the findings of Licklider (34) and Ainsworth (2).

6.6. Suggestions for Further Work

This study has left several aspects of the problem untouched.

The most obvious developments are in the extension of the number of voices and synthesis algorithms. Such questions as the optimal quantisation of the time-intervals and methods of improving the quality of the synthetic speech are also worthy of further investigation.

In view of the complex relationship between time-interval measurements and the major frequency components of the speech wave, it is unlikely that a simple algorithm will produce realistic formant transitions, particularly as the ordering of the time-intervals is an important factor in the perception of the sound. The results of the perceptual experiments suggest the importance of the formant concept in explaining the perception of clipped speech, so that a more direct way of producing the formants would seem appropriate. In view of this fact and the complexity of the system to produce a synthetic clipped speech wave, it is suggested that a better way of utilising time-interval measurements is in the derivation of control parameters for a formant synthesiser.

APPENDIX 1

Frequency and Time-Interval Analysis

For a sine wave clipped about its zero axis, both time-interval histograms and zero-crossing rate will give a true measure of the frequency of the sine wave, the accuracy of the histogram measurement depending upon the quantisation. When a d.c. bias is applied to the sine wave, the zero-crossing rate will still be an accurate measure of the frequency of the sine wave, whereas the single histogram peak will split into two depending upon the amount of d.c. bias and the quantisation.

Mixture of Two Sine Waves.

When two sine waves are added together, the relationship between the time-intervals and the component frequencies becomes more complex. Peterson (49) has shown how the reading of a zero-crossing frequency meter changes with the relative amplitude of the two component sine waves. It was anticipated that the behaviour of the time-interval histogram with the variation in amplitude would be more complex because of the greater amount of information in the histogram.

A computer program was written to solve the equation

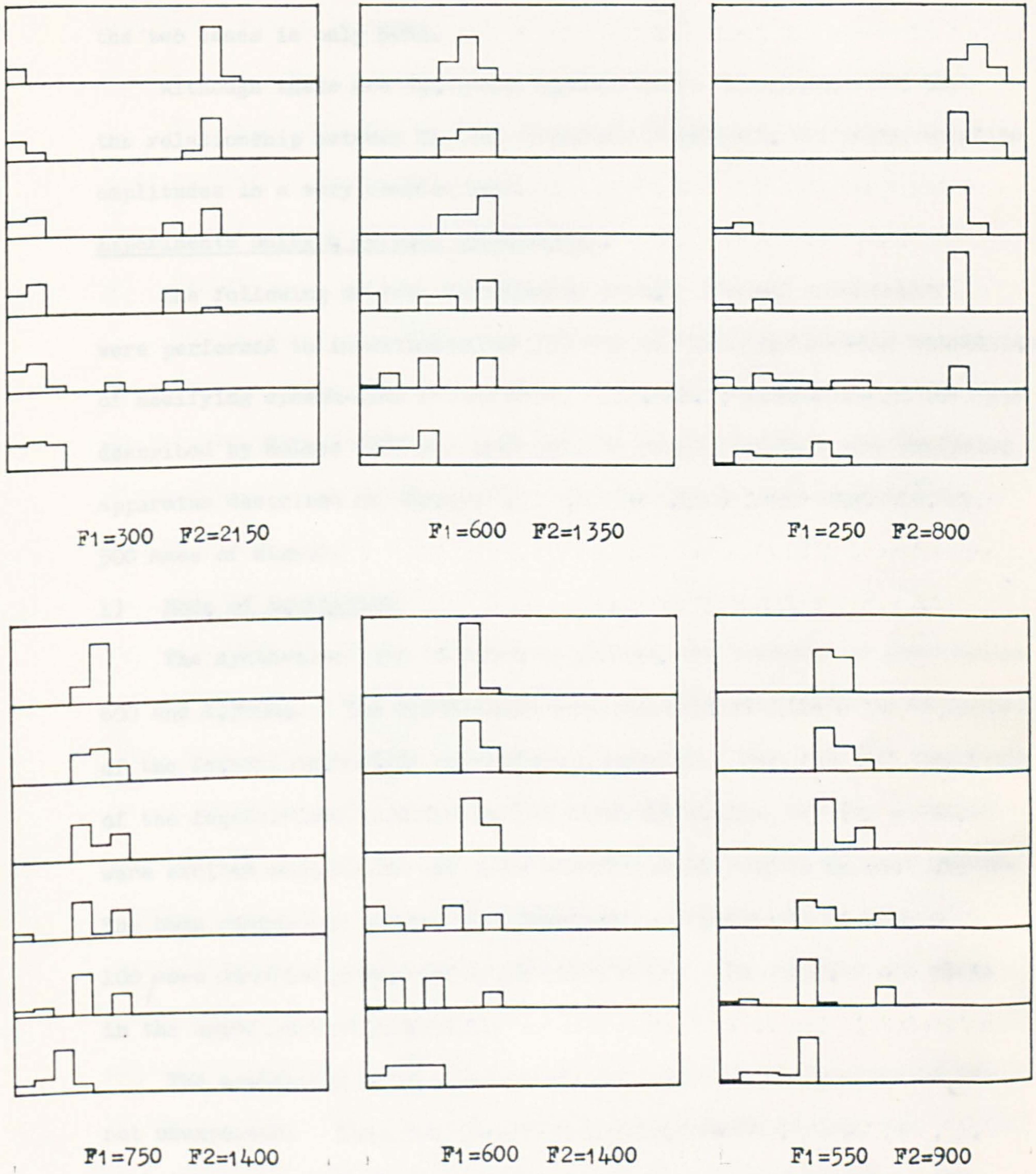
$$A_1 \sin F_1 t - A_2 \sin F_2 t = o$$

and produce a histogram of the time-intervals between the roots of the equation. A_1 and A_2 are variables that represent the amplitudes of the sine waves with frequencies F_1 and F_2. The waves were started in anti-phase to correspond to the phasing of the first two formants in a formant synthesiser (16). The histogram was compiled from the time-intervals occurring in one second of the signal. The frequencies F_1 and F_2 were chosen according to the first two formant frequencies of the vowels /i/, /æ/, /ɒ/ and /u/. Two second formant values were chosen for the vowel /ɜ/ differing by 50Hz. The relative amplitude of the two sine waves was varied in 3dB steps from A_2/A_1= -14dB to +1dB. The normalised histograms are shown in Fig. A1.1; the time-interval bins are 150μsec wide.

The behaviour of /i/ (F_1 = 300, F_2 = 2,150) is predictable, for as the second formant amplitude is increased more short intervals are produced, shortening the longer ones accordingly until the second formant time-intervals dominate the pattern.

The effect of increasing the second formant amplitude in /æ/ (F_1 = 750, F_2 = 1,400) is to split the peak corresponding to $1/2F_1$ intp two parts. This effect is even more prominent for /ɒ/ (F_1 = 550, F_2 = 900) where there is considerable lengthening of an appreciable number of intervals before the distribution concentrates into a peak corresponding to $1/2F_2$. The behaviour of /u/ (F_1= 250, F_2= 800) is no more predictable in that it produces a very scattered distribution for A_2/A_1= -5dB and -2dB.

The two examples of the central vowel (F_1 = 600, F_2 = 1,350, 1,400) are of particular interest because the distributions are quite different in the region A_2/A_1 = - 8dB to +1dB, although the difference in F_2 in



F1=300 F2=2150

F1=600 F2=1350

F1=250 F2=800

F1=750 F2=1400

F1=600 F2=1400

F1=550 F2=900

Fig. A1.1 Histograms of Sum of Two Sine Waves

the two cases is only 50Hz.

Although these are extremely simple cases, they illustrate that the relationship between the two frequency components and their relative amplitudes is a very complex one.

Experiments using a Formant Synthesiser

The following control experiments using a formant synthesiser were performed to investigate the effects on the time-interval statistics of modifying speech-like parameters. A formant synthesiser of the type described by Holmes (28) was used and its output fed into the analysing apparatus described in Chapter 4. The histograms were compiled for 500 msec of signal.

1) Mode of Excitation

The synthesiser was adjusted to provide two formants of frequencies 600 and 1,350Hz. The synthesiser controls allowed continuous adjustment of the formant amplitudes and larynx frequency. The relative amplitude of the formants was adjusted in 3dB steps as before, and the formants were excited with either the synthesiser's noise source or buzz source. The buzz source (or larynx tone generator) waveform was a pulse of 100 μsec duration repeating at 120 or 130 Hz. The results are shown in the upper part of Fig. A1.2.

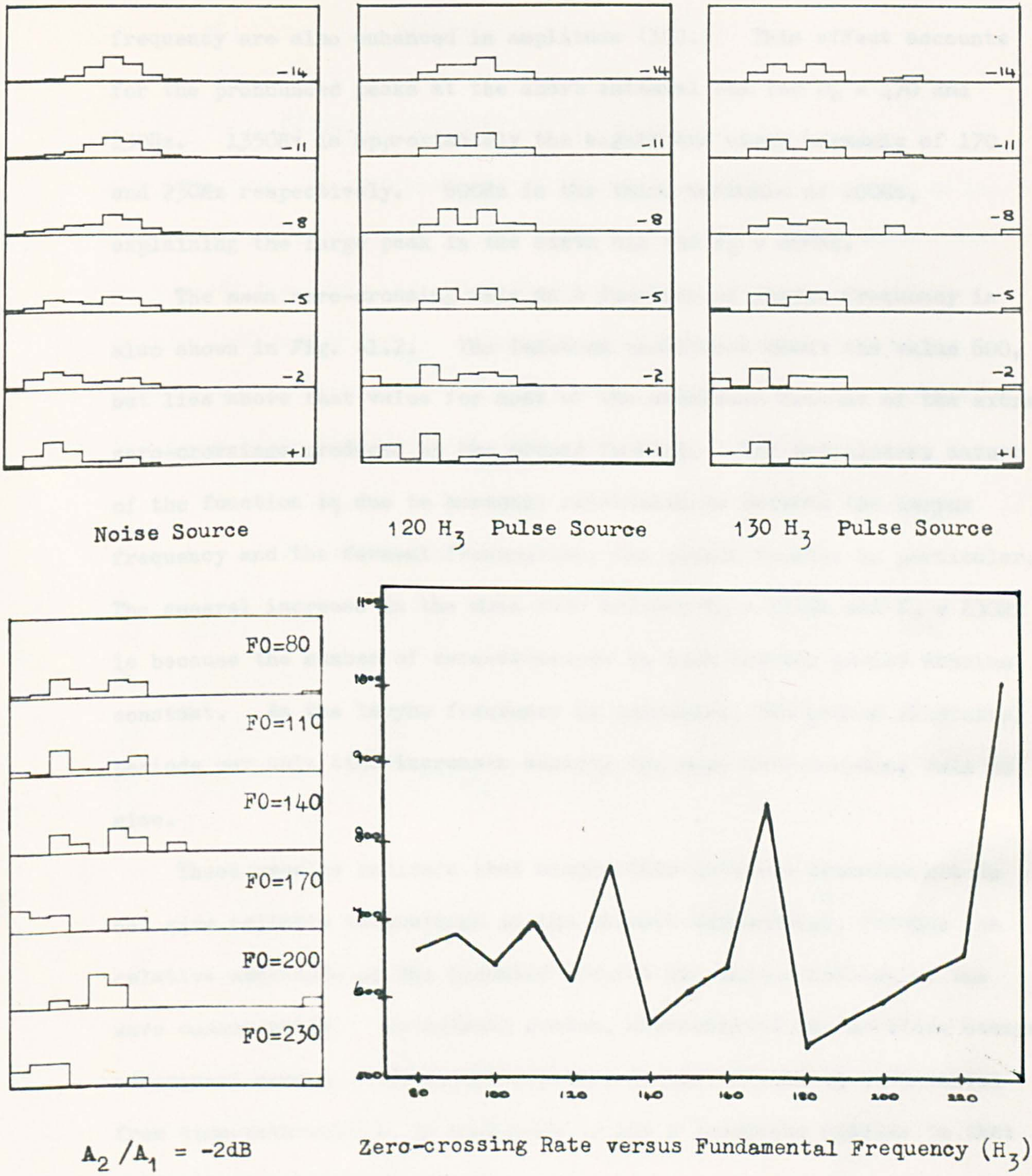
The appearance of the histograms for noise source excitation is not unexpected. When the second formant amplitude is low, the distribution is approximately gaussian about the half-period of the 600Hz component. As the second formant amplitude is increased the distribution becomes broader and more skew. Further increase causes the distribution to become bimodal, the other peak corresponding to the half-period of the 1350Hz component.

The distributions for the 120 and 130Hz larynx excitation are very similar, although the positions of the peaks do not correspond very well to the formant frequencies.

2) Variation of Larynx Tone Frequency

The amplitudes of the first and second formants were adjusted to produce a realistic /3/ (A_2/A_1 = -2dB), and the frequency of the larynx tone generator was varied in 10Hz steps from 80 to 230Hz. Histograms compiled at 30Hz steps are shown in the bottom left-hand corner of Fig. A1.2.

The most striking feature is the gross change in the histograms as the larynx frequency is altered. There is a greater similarity between the distributions for F_o (the larynx frequency) = 80, 110 and 140Hz than there is between those for F_o = 170, 200 and 230Hz. This is due to the construction of the synthesiser. The formant amplitude controls precede the formant frequency controls so that the amplitude settings are independent of any harmonic relationship between the larynx frequency and the formant frequencies. If a formant is a harmonic of the larynx frequency, successive excitations by the glottal pulse will cause oscillations in phase with the oscillations produced by the previous glottal pulse. This will increase the formant amplitude. This effect will be more pronounced at high larynx frequencies where there is a greater formant amplitude remaining at the end of each glottal period. (In human speech, formants that are harmonics of the larynx



Noise Source

120 H_3 Pulse Source

130 H_3 Pulse Source

A_2/A_1 = -2dB

Zero-crossing Rate versus Fundamental Frequency (H_3)

Fig. A1.2 Effect of Excitation on First-order Statistics

frequency are also enhanced in amplitude (17).) This effect accounts for the pronounced peaks at the short interval end for F_o = 170 and 230Hz. 1350Hz is approximately the eighth and sixth harmonic of 170 and 230Hz respectively. 600Hz is the third harmonic of 200Hz, explaining the large peak in the sixth bin for F_o = 200Hz.

The mean zero-crossing rate as a function of larynx frequency is also shown in Fig. A1.2. The function oscillates about the value 600, but lies above that value for most of the abscissae because of the extra zero-crossings produced by the second formant. The oscillatory nature of the function is due to harmonic relationships between the larynx frequency and the formant frequencies, the second formant in particular. The general increase in the mean rate between F_o = 180Hz and F_o = 230Hz is because the number of zero-crossings in each glottal period remains constant. As the larynx frequency is increased, the number of glottal periods per unit time increases causing the mean zero-crossing rate to rise.

These results indicate that simple time-interval measurements do not give reliable indications of the formant frequencies, because the relative amplitude of the formants affects the zero-crossings of the wave considerably. In natural speech, asymmetry of the waveform causes additional errors. To obtain reliable formant frequency information from time-intervals it is necessary to use a technique similar to that recently reported by Scarr (55).

APPENDIX 2

Frequency Response of Tape Recorder

The frequency response of the Truvox R42 tape-recorder shown in Fig. A2.1 was measured at a tape speed of 7½ inches per second. The sine waves that were used were recorded electrically well below the maximum modulation level. The on-axis response of the built-in loudspeaker was measured with a Dawe Sound level meter. The response was found to be substantially flat from 150Hz to 6KHz.

Frequency Response of Language Master

The frequency response of the Language Master is shown in Fig.A2.2. The test signals were recorded electrically, and the electrical output was measured at the headphone socket. The maximum signal to noise ratio (at 250Hz) was measured to be 36dB.

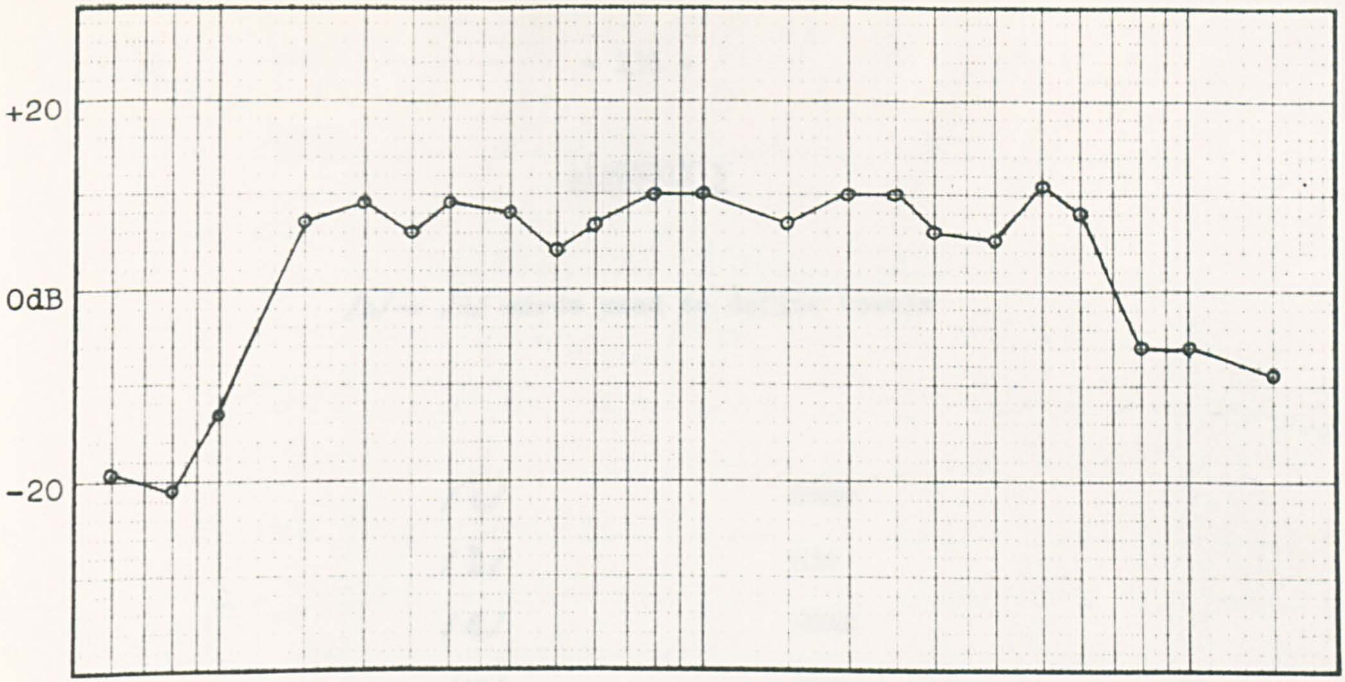


Fig A2.1 **Frequency** Response of Tape-recorder

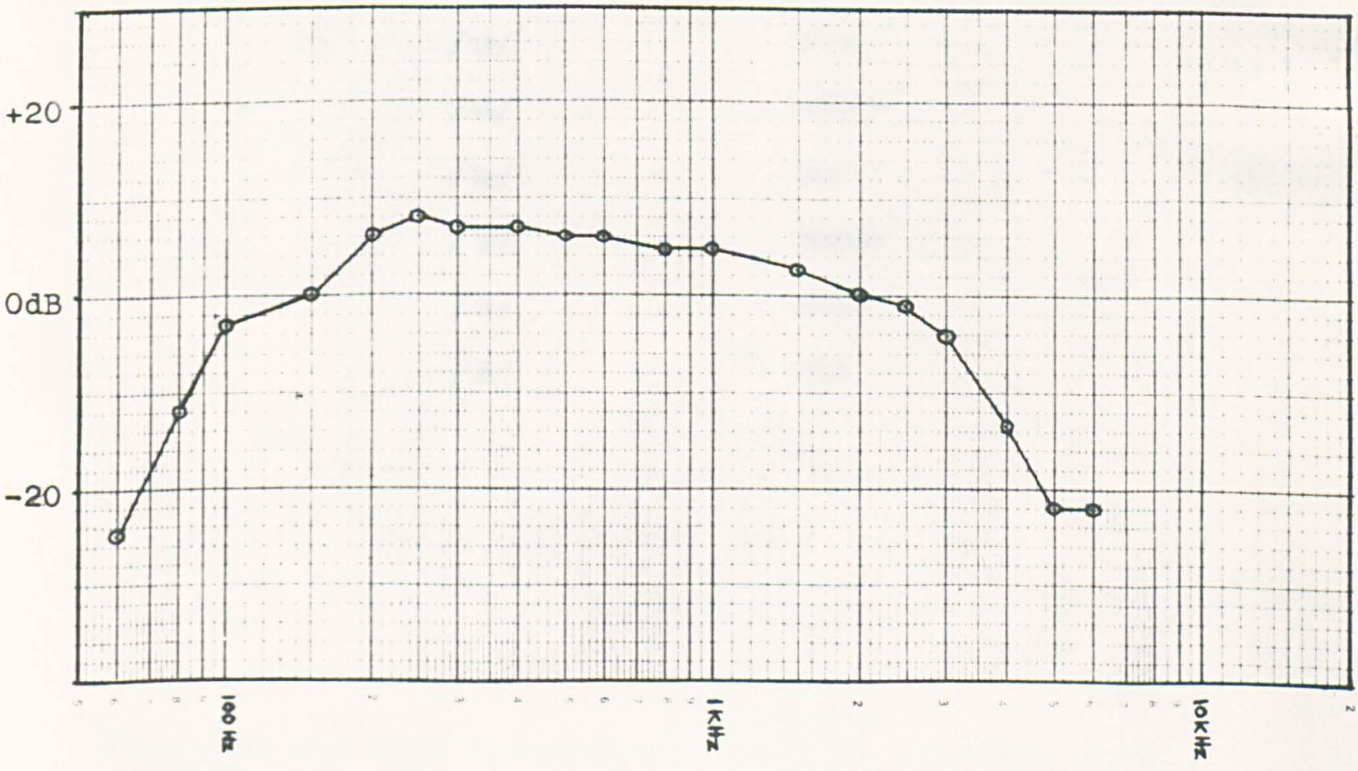


Fig A2.2 Frequency Response of Language Master

APPENDIX 3

/h/-- /d/ words used to define vowels

/i/	HEED
/I/	HID
/ɛ/	HEAD
/æ/	HAD
/ɜ/	HEARD
/ʌ/	HUT
/ʊ/	HOOD
/u/	WHO'D
/ɒ/	HOD
/ɔ/	HOARD
/ɑ/	HARD
/ə/	THE

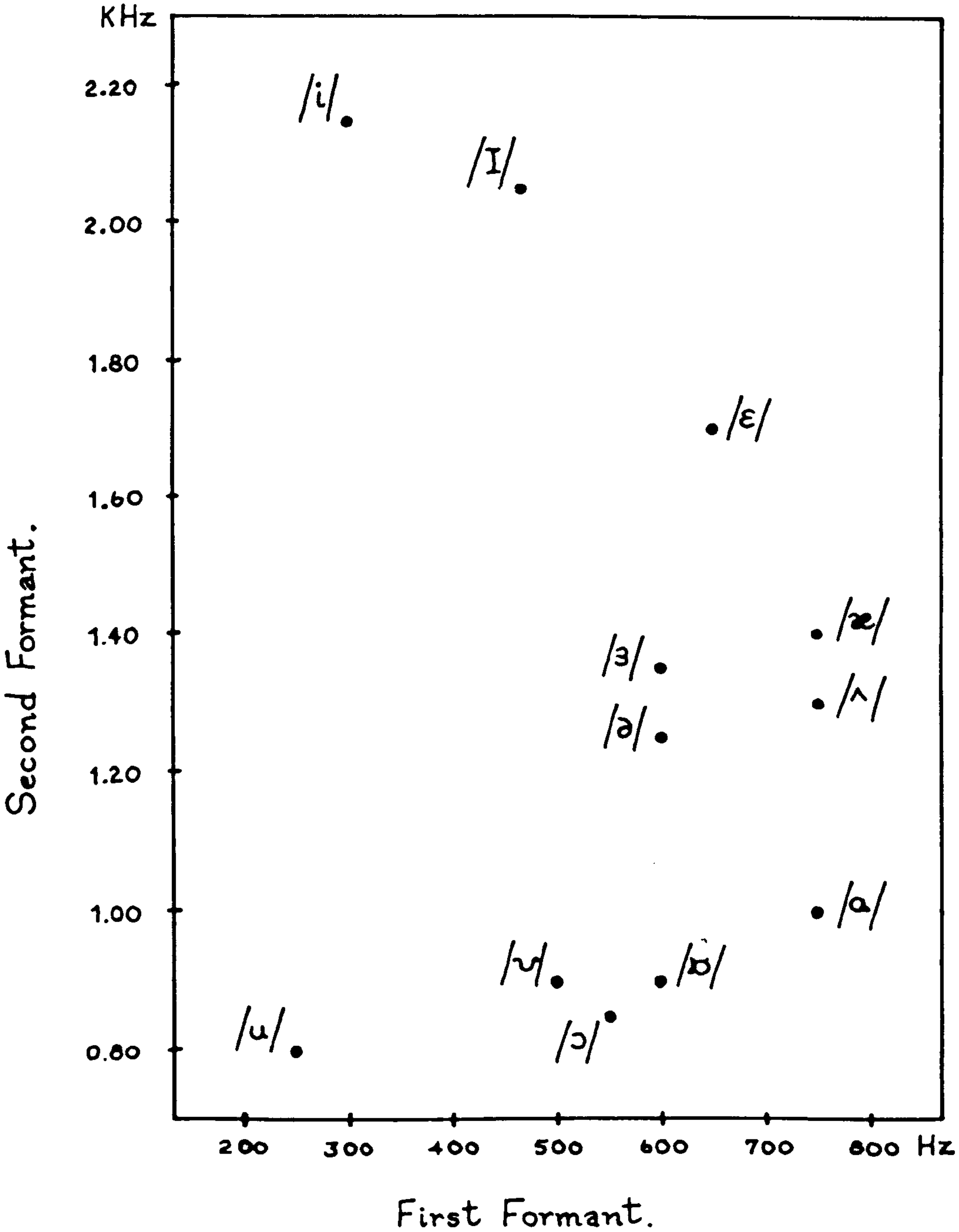


Fig A3.1 First Formant/Second Formant Plot for Vowel Sounds.

REFERENCES

1. R. Ahmend & R. Fatehchand: "Effects of Sample Duration on the Articulation of Sounds in Normal and Clipped Speech"

   J. Acoust. Soc. Am. 31, p.1022 (1959)

2. W.A. Ainsworth: "Relative Intelligibility of Different Transforms of Clipped Speech".

   J. Acoust. Soc. Am. 41, p.1272 (1967).

3. W. Bezdel & H.J. Chandler: "Results of an Analysis and Recognition of Vowels by Computer using Zerocrossing Data."

   Proc. IEE, Vol. 112, p.2060 (1965).

4. W. Bezdel: "Discriminators of Sound Classes for Speech Recognition Purposes."

   Paper B8, 1967 Conference on Speech Communication and Processing, Boston.

5. S.H. Chang: "Portrayal of Some Elementary Statistics of Speech Sounds".

   J. Acoust. Soc. Am. 22, p.768 (1950).

6. S.H. Chang, C.E. Pihl, J.Wiren: "The Intervalgram as a Visual Representation of Speech Sounds."

   J. Acoust. Soc. Am. 23, p.675 (1951).

7. E.C. Cherry: "On Human Communication."

   p.296, Science Editions Inc., New York,1961.

8. F.S. Cooper, J.M.Borst & A.M. Liberman: "The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech."

   Proc. Nat. Academy of Science. 87,p.318 (1951).

9. W. Davenport: "A Study of Speech Probability Distributions."

   M.I.T. Tech. Rept. No. 148, 1950.

10. E.E. David & H.S. McDonald: "Note on Pitch Synchronous Processing of Speech."

    J. Acoust. Soc. Am. 28, p. 1261 (1956).

11. H. Dudley: "Remaking Speech".
J. Acoust. Soc. Am. 11, p.169 (1939).

12. H. Dudley & T. Tarnoczy: "The Speaking Machine of Wolfgang von Kempelen."
J. Acoust. Soc. Am. 22, p. 151 (1950).

13. H. Dudley: "Speech Analysis by Waveform."
Paper A48. 5th Int. Congress on Acoustics, Liege, 1965.

14. H.K. Dunn: "The Calculation of Vowel Resonances and and Electrical Vocal Tract."
J. Acoust. Soc. Am. 22, p.740 (1950).

15. J.P. Egan: "Articulation Testing Methods".
Laryngoscope 58, p. 955 (1948).

16. G. Fant: "Acoustic Analysis and Synthesis of Speech with Applications to Swedish."
Ericsson Technics No. 1, 1959.

17. G. Fant, K. Fintoft, J. Liljencrants, B. Lindblom & J. Martony: "Formant-Amplitude Measurements."
J. Acoust. Soc. Am. 35, p.1753 (1963).

18. E.E. Fetz & G.L. Gerstein: "An R.C. Model for Spontaneous Activity of Single Neurons."
Research Laboratory of Electronics, M.I.T. Q.P.R. No. 71, October, 1963.

19. J.L. Flanagan: "Estimates of the Maximum Precision Necessary in Quantising Certain "Dimensions" of Vowel Sounds."
J. Acoust. Soc. Am. 29, p.533 (1957).

20. J.L. Flanagan: "A difference Limen for Vowel Formant Frequency".
J. Acoust. Soc. Am. 27, p.613 (1957).

21. T.W. Forgie & C.D. Forgie: "Results obtained from a Vowel Recognition Computer Program."
J. Acoust. Soc. Am. 31, p.1480 (1959).

22. A.J. Fourcin: "An Investigation into the Possibility of Bandwidth Reduction in Speech."
Ph.D. Thesis, University of London 1961.

23. B. Gold: "Computer Program for Pitch Extraction".
J. Acoust. Soc. Am. 34, p.916 (1962).

24. G.T. Guilbaud: "What is Cybernetics?"
Heinemann 1959 p.74.

25. M. Halle, G.W. Hughes, J.P. Radley: "Acoustic Properties of Stop Consonants."
J. Acoust. Soc. Am. 29, p.107 (1957).

26. H.V. Helmholtz: "On the Sensations of Tone."
2nd English Edition, Dover Publications Inc., New York 1954.

27. J.N. Holmes: "An Investigation of the Volume Velocity Waveform at the Larynx during Speech by Means of an Inverse Filter."
G13, 4th Int. Congress on Acoustics, Copenhagen, 1962.

28. J.N. Holmes, J.N. Shearme & I.G. Mattingley: "Speech Synthesis by Rule."
Language and Speech, 7 (3) July-September 1964.

29. J.N. Holmes: "Some Recent Research at the Joint Speech Research Unit of the British Post Office."
IEEE International Communication Conference June 1966.

30. L.G. Kersta: "Voiceprint Identification."
Nature 196; p.1253 Dec. 29th,1962.

31. J. Laver: "Variability in Vowel Perception".
Language and Speech, 8 (2) April-June 1965.

32. W. Lawrence: "The Synthesis of Speech from Signals which have a Low Information Rate." in Communication Theory edited by Willis Jackson, Butterworth, London, 1953.

33. A.M. Liberman, R.C. Delattre, L.J. Gerstman & F.S. Cooper: "Tempo of Frequency Changes as a Cue for Distinguishing Classes of Speech Sounds."
Journal of Experimental Psychology, 52, p.127 (1956).

34. J.C.R. Licklider, D. Bindra & I Pollack: "The Intelligibility of Rectangular Speech Waves."
Amer. J. Psychol. 61, p.1 (1948).

35. J.C.R. Licklider & I. Pollack: "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech."
J. Acoust. Soc. Am. 20, p.42 (1948).

36. J.C.R. Licklider: "The Intelligibility of Amplitude-Dichotomised, Time-Quantised Speech Waves."
J. Acoust. Soc. Am. 22, p.820 (1950).

37. J. Liljencrants: "A Few Experiments of Voiced/Voiceless Identification and Time Segmentation of Speech."
Paper C8, Vol. 1, Proceedings of the Speech Communication Seminar, Stockholm 1962.

38. J. Liljencrants: "The Useful 2-Stage Complementary Amplifier."
Quarterly Progress Report of the Speech Transmission Laborator, Royal Institute of Technology, Stockholm, April 1966.

39. B. Lindblom: "Spectrographic Study of Vowel Reduction."
J. Acoust. Soc. Am. 35, p.1773 (1963).

40. D.M. MacKay & M.E. Fisher: "Analogue Computing at Ultra High Speed."
Chapman and Hall, 1962.

41. D.M. MacKay: "An Improved System for the Visual Display of Speech and other Sequential Data."
Brit. Prov. Pat. Spec. 43489, 1965.

42. D.M. MacKay: "Improvements in or Relating to Electronic Information Storage Circuits, with Special Applications to Automatic Waveform Recognition"
Brit. Prov. Pat. Spec. 52722, 1965.

43. D.M. MacKay, J.B. Millar & M.J. Underwood: "The Discriminative Value of the Digram Structure of Speech Waveforms."
Proc. 18th Intern. Congress Psychol. Moscow 1966 (to be published).

44. P. Mermelstein & M.R. Schroeder: "Determination of Smoothed Cross-Sectional Area Functions of the Vocal Tract from Formant Frequencies."
Paper A2, 5th Int. Congress on Acoustics, Liege 1965.

45. J.B. Millar: "The Investigation of Three Related Techniques for the Statistical Analysis of Clipped Speech."
Ph.D. Thesis to be submitted, University of Keele, 1968.

46. R.L. Miller: "Nature of the Vocal Chord Wave."
J. Acoust. Soc. Am. 31, p.667 (1959).

47. V.P. Morozov: "Intelligibility of Singing as a Function of Fundamental Voice Pitch."
Soviet Phys.-Acoust. 10, p.279 (1965).

48. A.M. Noll: "Short-time Spectrum and 'Cepstrum' Techniques for Vocal Pitch Detection."
J. Acoust. Soc. Am. 36, p. 292 (1964).

49. E. Peterson: "Frequency Detection and Speech Formants."
J. Acoust. Soc. Am. 23, p.668 (1951).

50. G.E. Peterson & G.R. Hanne: "Examination of Two Different Formant Estimation Techniques."
J. Acoust. Soc. Am. 37, p.224 (1965).

51. E.N. Pinson: "Computing Vocal Tract Shapes to Yield Specific Tract Transfer Functions."
Paper A37, 5th Int. Congress on Acoustics, Liege 1965.

52. D.R. Reddy: "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave."
Tech. Report CS49, Sep. 1966 Dept. of Computer Science, Stanford University.

53. T. Sakai & S. Inoue: "New Instruments and Methods for Speech Analysis."
J. Acoust. Soc. Am. 32, p.441 (1960).

54. T. Sakai & S. Doshita: "The Automatic Speech Recognition System for Conversational Sound."
IEEE Trans. on Electronic Computers, December 1963, o. 835.

55. R.W.A. Scarr: "Zero Crossings as a Means of Obtaining Spectral Information in Speech Analysis."
Paper C13, 1967 Conference on Speech Communication and Processing, Boston.

56. C.E. Shannon & W. Weaver: "The Mathematical Theory of Communication."
University of Illinois Press, Urbana 1949.

57. C.E. Shannon: "Prediction and Entropy of Printed English"
Bell System Technical Journal Vol. XX, January 1951, p.50.

58. I.C. Steinberg: "Applications of Sound Measuring Instruments to the Study of Phonetic Problems."
J. Acoust. Soc. Am. 6, p.16 (1934).

59. K.N. Stevens, S. Kasowski & C.G.M. Fant: "An Electrical Analogue of the Vocal Tract".
J. Acoust. Soc. Am. 25, p. 734 (1953).

60. K.N. Stevens & A.S. House: "Perturbation of Vowel Articulations by Consonantal Context."
J. Speech and Hearing Res. 6, p.111 (1963).

61. W.R. Stover: "Time-Domain Bandwidth-Compression System."
J. Acoust. Soc. Am. 42, p.348 (1967).

62. Y. Tanaka & J. Okamoto: "Syllable Articulation at the Time when the Trailing Edges of Zero-Crossing Waves Make Random Fluctuations."
Osaka City Univ. Mem. Fac. Eng., p. 75 (1964).

63. I.B. Thomas: "The Significance of the Second Formant in Speech Intelligibility."
AF-33(615)-3890, AF Grant 7-66 Technical Report No. 10, University of Illinois. AD654 326.

64. R. Vanderslice & P. Ladefoged: "Voiceprint Mystique".
J. Acoust. Soc. Am. 42, p.1164 (A) (1967).

65. F. Vilbig & K.H. Haase: "Theoretical Investigations to Reduce Harmonic Distortion in a Clipping Process."
J. Acoust. Soc. Am. 29, p.776 (A) (1957).