

ADAPTIVE SPECTRAL ANALYSIS
FOR SPEECH-SOUND RECOGNITION

A thesis submitted by Michel Lecours
for the degree of Doctor of Philosophy
in the faculty of engineering of the
University of London.

The Electrical Engineering Department,
Imperial College,
London, S.W.7.
May 1967

Abstract

ADAPTIVE SPECTRAL ANALYSIS FOR SPEECH-SOUND RECOGNITION

A number of speech researchers have noted that it might be useful to allow the parameters of speech spectrum analysers to vary according to the characteristics of the speech sounds being processed. In particular, it seems clear that the use of narrow band filters should enhance the formant pattern of vowels, and that wide band-width filters should bring out the noise-like structure of stop-consonants.

In this thesis, the reasons for using more than one type of analyser at a time or an analyser with variable parameters for the purpose of speech recognition are explored. Arguments in favour of such a system, called here "adaptive spectral analysis", are drawn from the theories of signal analysis and from studies in speech and aural perception.

A demonstration of a real-time adaptive spectral analyser for the purpose of automatic speech recognition is discussed. The conclusion is reached that one way of doing it is to use a pattern recognition algorithm and compare the results obtained for various designs.

This method is used to compare, on the one hand, an analyser having 32 active, 100 c/s bandwidth, gaussian filters and, on the other hand, an analyser consisting of sixteen 200 c/s filters having similar filtering characteristics. Pattern recognition experi-

ments are performed with the four stop-consonants /p/, /b/, /t/, /d/ taken two by two and the pairs of vowels /a/, /ɔ/ and /e/, /i/; this data was taken from recordings of the four above-mentioned consonants followed by the sounds /a/, /ɔ/, /et/, /it/, as said by a few different speakers. Significant differences in recognition performance are obtained in the experiments. Narrow band filters are found to be preferable for recognizing vowels or for recognizing consonant-vowel frequency transitions. Wide-band filters appear to be better for classifying stop-consonants when only the data on the main body of the consonants is used.

A few results of experiments with an analyser having 64 filters of 50 c/s bandwidth are also given.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Mr. J. J. Sparkes, for his encouragement, his advice and, in particular, for his very pertinent questions and comments, and also Professor C. E. Cherry, in whose section this work was accomplished, for his interest and support.

Thanks are due to many people who were associated with this work: to Dr. R. S. Thomas for having suggested at first to realize an adaptive speech-spectrum analyser, to Dr. Turski for having discussed with me the possibility of using pattern recognition methods in this work, to D. J. Goodman for his cooperation and his advices concerning the use of the digital data taking equipment.

I am grateful to Professor D. B. Fry and Dr. Fourcin for their comments which eventually brought me, at one time, to modify completely the course of my work.

Financial assistance was granted by the Board of Trade for two years, then, by the Quebec Ministry of Education and the Quebec hydro-electric Commission (Hydro-Quebec) for one year each. Financial assistance was also offered by the National Research Council of Canada.

Finally, I would like to thank my wife for having typed the manuscript, developed the photographs and having helped me all through the writing of this thesis.

Michel Lecours

TABLE OF CONTENTS

<u>Section</u>	<u>Page</u>
Abstract	 2
Acknowledgements	4
INTRODUCTION	12
1. BACKGROUND	
1.1 Introduction	 13
1.2 Aims and motives	 14
1.3 The nature of speech	 15
1.4 Studies on speech recognition	 23
1.5 Speech perception	 27
1.6 The main difficulties in speech recognition	 29
2. A MATCHED ANALYSER	
2.1 Introduction	 33
2.2 Matched filtering	 33
2.3 A matched analyser	 34
2.4 An example	 35
2.5 Conclusion	 37
3. TIME-FREQUENCY RELATIONS IN SIGNAL ANALYSIS	
3.1 Introduction	 42
3.2 Time-frequency relationships in analysis	 42
3.3 A theory of sampling and of spectral analysis	 44

3.4	The uncertainty relation		46
3.5	Conclusion		51

4. TIME-FREQUENCY RELATIONSHIPS IN SPEECH AND AURAL PERCEPTION

4.1	Introduction		54
4.2	Time-frequency relations in aural perception		54
4.2.1	Auditory perception		54
4.2.2	The "uncertainty relation" and auditory perception		57
4.2.3	Summary and conclusion		63
4.3	Time-frequency characteristics and speech perception		64
4.3.1	Introduction		64
4.3.2	Speech characteristics		64
4.3.3	Conclusion		67
4.4	Adaptive spectral analysis		68
4.5	Definition		69

5. THE IMPERIAL COLLEGE AUDIO SPECTRUM ANALYSER

5.1	Introduction		71
5.2	Filtering		71
5.3	General description		74
5.4	Redundancy in the frequency domain		76
5.5	Conclusion		77

6. IMPLEMENTING ADAPTIVE FILTERING

6.1	Introduction		83
6.2	The function generator		83
6.2.1	Synchronisation		84
6.2.2	The speech wave		85
6.2.3	Control		86
6.3	The segmenter or pre-detector		87
6.4	Spectrograms		89
6.4.1	Apparatus used		89
6.4.2	Photographs of spectrograms and comments		90
6.4.3	Conclusion		94
6.5	General points		95
6.6	Conclusion		95

7. THE DESIGN OF THE EXPERIMENTS

7.1	The problem		111
7.2	The general lines of a possible solution		111
7.3	Subjective assessment		112
7.4	Some crude device		114
7.5	The basic requirements		115
7.6	A learning system		116
7.7	Application to our problem		119
7.8	Discussion		119
7.9	Definition		120

8. BRAVERMAN'S ALGORITHM AND MODIFICATIONS

8.1	Introduction		123
8.2	Basic assumptions		124
8.3	Braverman's algorithm		125
8.3.1	Construction of separating hyperplanes		126
8.3.2	Elimination of redundant matrix elements		127
8.3.3	Testing the decision function		127
8.4	Modifications		127
8.5	Definitions		128
8.6	Conclusion		129

9. THE NATURE OF THE EXPERIMENTS

9.1	Introduction		134
9.2	The Data		135
9.2.1	The speakers		135
9.2.2	Recording the data		135
9.2.3	Speech sounds selected for study		138
9.3	The experiments		139
9.3.1	Consonants or vowels?		139
9.3.2	Number of categories		140
9.3.3	Segmentation		141
9.3.4	Voicing		142

9.4	Controlling the amplitude		143
9.5	The positioning of speech sounds ...		146
9.5.1	Constraints introduced by speech		146
9.5.2	The positioning method chosen		149
9.5.3	The positioning technique ...		150
9.5.4	Verifications		151
9.5.5	Discussion		151

10. THE SIGNIFICANCE OF THE RESULTS

10.1	Introduction		154
10.2	Randomness and statistical significance		154
10.3	Randomness in the learning process		157
10.4	Randomness in the positioning process		158
10.5	The standard test of signifi- cance		160
10.5.1	Introduction		160
10.5.2	The absolute performance ...		160
10.5.3	The relative performance ...		162
10.5.4	Testing the results for significance		164
10.5.5	Conclusion		165
10.6	Repeating the experiments with new learning samples		166

11. THE RESULTS

11.1	Introduction	 167
11.2	The "p-b" pair	 168
11.3	The "t-d" pair	 174
11.4	The "p-t" pair	 181
11.5	The "d-b" pair	 186
11.6	The "p-d" pair	 190
11.7	The "b-t" pair	 195
11.8	The vowels	 200
11.9	The significance of the results	 202

12. CONCLUSION

12.1	Discussion of the results	 206
12.2	Adaptive spectral analysis	 208
12.3	The recognition performance	 210
12.4	Suggestions for further work	 211
12.4.1	Completing the present study	 211
12.4.2	Different pattern recog- nition algorithms	 212
12.4.3	Analysis by synthesis	 213
12.4.4	A continuously adaptive analyser	 214
12.5	Limitations in physical effort and computing time	 214
12.6	Conclusion	 217

APPENDICES

A	THE SPEECH SOUNDS USED IN THE EXPERIMENTS	 219
B	RECORDING THE DATA	 221
C	THE EXPERIMENTS WITH THE 50C/S MODE	 235
D	THE ALGORITHMS	 238
	Bibliography	 249

INTRODUCTION

This thesis is concerned with optimizing the characteristics of spectrum analysers for the purpose of automatic speech recognition.

In all problems linked with pattern recognition or with detection, one is confronted with the fundamental difficulty of finding the best possible set of measurements.

In spectral analysis, this set of measurements is constituted by samples of the energy present throughout the time-frequency plane; these samples, however, can be taken in different ways according to the filters chosen for the spectrum analyser.

In this thesis, the possible uses of a variable parameter spectral analyser are investigated for the purpose of recognizing simple speech sounds. A general pattern recognition method is used for the recognition of a few selected speech sounds. A higher reliability is obtained by using, within limits, a greater number of narrow band-width filters (a better frequency resolution) when analysing sounds or parts of sounds characterized by frequency resonances. A higher reliability is also obtained, again within limits, by having a better time resolution in the analysis and sacrificing the frequency resolution for sounds or parts of sounds characterized by a ^{plosive} ~~noise-like~~ structure.

Chapter 1

BACKGROUND

1.1 Introduction

Although speech is something we use in our everyday lives and something we consider quite simple, it is fair to say that the mechanisms of speech perception are still unknown today in spite of the intensive research that has been carried out in this field.

The audio spectral analyser has shown itself over the last twenty-five years to be an instrument of great importance for the study of speech. This instrument usually consists of a bank of adjacent filters in the audio-frequency range, which separate the frequency components present in the input signal. If the analyser is not meant to operate in real time, one can store the signal and play it back again and again while the frequency range of interest is being swept by a single filter.

The spectral analyser, as has been pointed out by a number of writers including Potter,⁵² Steinberg⁵² and Gabor,²² goes a long way towards representing the operation of the ear but it cannot be said to simulate it. In particular, as Cherry argues, these instruments are constant-parameter devices and it would be much more accurate to compare the ear to a set of these analysers, since it seems the ear possesses this property of "matching" itself to the signal it receives.

1.2 Aims and motives

Scientific interest in the nature of speech started to awake in the last two centuries. Previously there had been some attempts to produce "talking machines" but mostly for the purpose of amusing and entertaining. In the eighteenth century, "von Kempelen" designed an elaborate model of the vocal tract which is said to have been fairly successful. In the next century, theories of vowel production were proposed by students of acoustical phenomena such as Wheatstone (1837), Willis (1829) and Helmholtz, to be followed early in this century by the work of Sir Richard Paget on "Human Speech".

But it is in the last thirty years with the introduction of the "vocoders" ~~or "voders"~~ (Dudley, 1939) and after the studies on "visible speech" (Potter, Kopp, Green), when possible applications of those studies for bandwidth compression and speech recognition were understood that the research in this field has intensified.

However, the fruits that speech researchers expected to pluck easily are still far from being ripe. Instruments able to compress speech into a narrower bandwidth have not come into widespread use because of our inability to reproduce noise-free and natural sounding speech and to achieve a high ratio of bandwidth compression at the same time. On the speech

recognition side, no one has been able to find the "invariants", which are assumed to characterize speech sounds, and to produce a reliable speech-sound recognition system.

All this work, however, as well as the important contributions to this field by phoneticians, psychologists and physiologists has brought about a better understanding of speech, a deeper insight into its nature.

1.3 The nature of speech

It is probably useful at this stage, since this thesis is going to be concerned with speech to give as brief a summary as possible on the nature of speech and the development of speech research.

Speech is produced by the modulation of an air stream by the larynx and ^{its filtering by} the resonant cavities of the mouth, the nose and the throat. It can be said that the larynx pulses- or the breath pulses in the case of whispered speech-form the carrier wave which is modulated by the different constrictions of the vocal tract. On one end of the realm of speech sounds, for vowel sounds, the vocal cavities are open in a more or less fixed position and modulate the flow of air; at the other end of the realm, the stop-consonants are constituted by a transient modulation of a burst of breath. In between these two extremes, all the intermediate speech sounds, namely the diphthongs, the transitionals, the semi vowels and the fricatives are to be found.

Wheatstone on one hand, and Willis, Herman and Scripture on the other, were the first to propose theories of vowel production generally called harmonic and inharmonic theories of vowel production, which (67) Helmholtz showed to be equivalent in practice. These researchers thought vowel sounds were characterized by a fundamental frequency and a large number of harmonics. They seem to have considered single harmonics as sufficient to characterize vowel sounds and it is only in this century that Paget definitely established the two formant pattern, although the presence of multiple resonances had been suspected before.

These studies in acoustics constituted the first attempts to synthesize speech. There are two other approaches to the studies of speech: speech analysis and perception studies. This brief review will be mainly concerned with the progress made in speech analysis and synthesis.

Graham Bell produced in 1873 his phonautograph, a mechanical precursor of the electronic oscilloscope and used it to show the wave-forms of speech (speech in a visible form). This kind of display is useful to show the temporal structure of speech, the presence of the larynx pulses which excite the resonant cavities into oscillations, which gradually die down until the next larynx pulse. Pains-taking observations were made, measurements were done in order to characterize speech

sounds (Crandall and Sacia, 1925), but this type of visible display of speech is not very useful for recognition purposes. A speech wave in a given environment will be reflected many times before it reaches the ears of the listener. The phase characteristics of the speech wave will vary considerably but the ear will usually notice very little difference, which shows that the ear is not very sensitive to phase differences and that phase information in speech can be disregarded for the purpose of speech recognition.

Dudley introduced in 1939 an instrument called the vocoder whose purpose it was, as can be inferred from its name, to "code voice signals" and to obtain a certain degree of band-width compression. Ever since, vocoders have been a current topic for research. The idea is to transmit a set of low-frequency signals proportional to the energy contained in fixed frequency bands, which can then be coded and transmitted through a much smaller bandwidth than the original speech-wave. The output of a pitch detector is used to control a buzz-tone at the far end providing the carrier wave on which the set of previously mentioned low-frequency signals is modulated; a random-noise generator is used for that purpose when no voicing is detected. The intensity of the remade speech is determined by the total sound power at the near end. The original Dudley Vocoder could transmit a 3 kc/s voice signal through a 300 c/s channel, achieving thus a bandwidth compression

ratio of 10 to 1. The resulting speech was noisy but quite intelligible. This type of vocoder is usually called a channel vocoder; its main shortcoming resides in the excitation of the remade speech since it is difficult to devise a reliable system that will switch from the buzz to the hiss tone at the appropriate instant. One can avoid this problem by transmitting directly the base band of speech and by using it as the excitation source at the far end; one obtains in this way Voice-excited Vocoder, which seem to be the most promising. Other types of vocoders, proposed or developed, include the resonance vocoder which is based on formant tracking; its main advantage is to secure an impressive bandwidth compression ratio (20 to 50) but does rather poorly on consonants. There have been attempts also to store a certain number of fixed patterns at the far end and to send only the information necessary to group them in the appropriate sequence.

In recent years, as the search for a speech bandwidth compression system has intensified, more elaborate vocoders have been produced. Gold (1964) has attempted to introduce "speech-like phase" in the remade speech. Golden (1963), simulating a voice-excited vocoder on a digital computer, claims to have obtained speech of excellent quality "often not distinguishable from speech appropriately band-limited", but needed to transmit directly the baseband from 250 to 925c/s and

thus could only get a bandwidth compression ratio of 3 to 1. One feels, at the light of these results, that vocoders could at least come into commercial use in the next years if the communications requirements justify it, but one would hope to do much better than has been realized up to now, when more is known about speech.

The introduction of speech spectrographs was probably the biggest single impulse which speech studies have ever been given. Experiments performed at Bell Labs in 1945 were able to show that the visible spectrograms formed patterns which a human being could use to identify speech sounds fairly reliably. ⁽⁵¹⁾ The speech spectrograph provided a very practical tool and enabled researchers to check easily hypotheses and theories which had just been surmised before and to gain an understanding of the constitution of speech sounds.

Then, started the search for the invariants in speech. The approach which seemed the most promising and which was to carry the main stream of speech research for many years was the one known today as "formant-tracking" by which it was attempted to characterize speech sounds mainly by the value of their frequency components. Formants, however, are not as invariant as one might wish (in particular the third one); measurements have shown that if they are fairly constant for a given speaker, they vary somewhat with a different one, quite a lot for people of different sex and between

children and adults. The human ear seems to have the property to adapt itself to individual accents or to dialects. For instance, Broadbent and Ladefoged (1957) have shown that synthetic utterances of bit, bet, bat, but, preceded by a synthesized introductory sentence can be differently recognized by a proportion of listeners according to the way the formants of the introductory sentence are manipulated; although the implications of this experiment should be discussed in more detail than is done here, it can be taken for granted that the human system can tune-in on the voice of a speaker.

Spectral analysis is an important tool in speech research today and most researchers use it in one way or another. But, for all their usefulness, speech spectrograms tend to submerge us into a sea of more or less significant details for our work. For one thing, it has been shown they can be used as well for the identification of a person than for the classification of speech sounds (~~although in the first case a contour display is normally used~~); it is difficult to separate what is essential to intelligibility from what characterizes an individual's speech and from what confers speech its liveliness and naturalness. This fundamental multiplicity in the type of information speech conveys creates the need to separate the different properties of speech and it is in order to solve this problem that the "Analysis-by-Synthesis" approach was

conceived. One way to synthesize speech is by designing dynamic analogues of the vocal tract, work in which MIT seems to have specialized; another way, which the Haskins Laboratories have explored, is to synthesize speech sounds from stylised spectrograms.

Dunn demonstrated in 1950 an electrical vocal tract and in the subsequent years, considerable effort was devoted, in particular at MIT, to producing dynamic models of the vocal tract with adjustable parameters (1953, 1960, 1961). Other systems of this type are the Parametric Artificial Talking Device (PAT) developed by W. Lawrence⁽⁶⁸⁾ and the device known as OVE developed by G. M. Fant. More recently, the use of digital computers to control the different parameters of the artificial vocal tracts has permitted to extend the scope of this type of work. The research in this field has provided us mainly with a knowledge of the physical constraints under which our vocal organs operate.

The other approach, the one developed by Haskins Laboratories³⁸, is to synthesize sounds by adding together pure tones of the required frequency and amplitude. Their instrument called a "Pattern Playback" generates a pure tone of 120c/s and the first 50 harmonics when photo-electric cells are excited by a thin ray of light reflected by a painted plastic sheet; the electrical oscillations are amplified and converted into sound waves. It is then possible to synthesize speech sounds

by painting their spectrograms on a plastic sheet so that the light is reflected when there is paint and excites the appropriate oscillator. These painted spectrograms can be made as simple or as complicated as one wishes, the only limitation being the fixed frequency of the oscillators, which causes the speech produced to lack of naturalness. The Haskins Laboratories have, over the years, used this instrument to generate speech sounds from stylised spectrograms and thus to isolate and define systematically the important cues in speech. The only negative comment that can be passed on their work is that there is no certainty that the elements that they have isolated are important in practice.

38

The Haskins Laboratories have investigated the role of the first and second formants in the perception of vowel sounds, and the rules governing the importance and influence of the third formant. They have found that the second and third formant transitions play an important part in the identification of the stop and nasal consonants by the ear (1956). They found later on (1956) that the position of a burst of noise was also a clue; this led them to the concept of correlation between the second formant transition and the manner of articulation of the consonant, and to the definition of a "frequency locus" for each type of articulation.

In the following years research was mainly devoted to semi-vowels, the perception of fricatives, the effect of third formant transitions and the distinction between voiced and voiceless stop-consonants in initial position.

Jacobson, Fant and Halle proposed in 1952 an original and fruitful approach to the problem of speech perception. They work under the assumption that a listener identifies a sound by making a sequence of two-dimensional (binary) judgments on the perceptual qualities of the perceived sound; they have isolated a number of perceptual qualities according to which sounds can be uniquely classified: voiced/unvoiced, turbulent/non-turbulent, acute/grave, and so on.

1.4 Studies on speech recognition

All speech recognition systems have to rely to some extent on spectral analysis; there are however many different approaches and innumerable variations within the realm of each approach. Perhaps the first serious attempt to recognize speech sounds by mechanical means was made in 1950 by Jean Dreyfus-Graf who divided the audio-frequency band of interest between six filters and connected their output to deflecting coils operating on a pen recorder.

Further attempts were done at the Bell Telephone Laboratories by Davis, Biddulph and Balashek (1952) and by Dudley and Balashek (1958). Both experiments were

concerned with the recognition of digits from 0 to 9. The first experiment, which gave for a single speaker an accuracy of around 97 %^o, was based on the axis-crossing counts in two frequency bands, one over and one under 900 c/s. For the second experiment, the method used was to cross-correlate the input signal after analysis by ten adjacent 300 c/s wide filters to a set of 10 predetermined "average" patterns, stored in the instrument. For one speaker, the reliability was nearly perfect, deteriorated rapidly with a larger number of speakers, but improved to 90 %^o after these speakers had learned to constrain their speech.

Wiren and Stubs (1956) used the principle of distinctive features proposed by Jacobson, Fant and Halle in their speech recognition work. Their system would make a series of binary decisions on the incoming signal and classify it as:

- voiced or unvoiced
- stop or fricative
- non turbulent or voiced turbulent
- vowel or vowel-like
- acute or grave

This system was fairly successful and its proponents were able to claim for nine vowels in single syllable words a reliability in recognition of about 90 %^o.

Since this thesis has partly to do with speech recognition it is in order before going any further to

have a closer look at the two main recent attempts to recognize speech sounds.

Martin and al. have developed a speech analyser using "Analog Threshold Logic" (ATL) for feature abstraction. They have tested its capacity to recognize consonants in utterances of Consonant-Vowel-Consonant (CVC) words by a number of talkers. "The ATL element originally modeled after the biological neuron has an output which is lineary proportional to the net sum of excitory and inhibitory inputs provided that this sum is greater than some adjustable threshold". Interesting characteristics of this system are that it does not require to segmentate the speech into phonemic units and that some quantitative measurements of probability can be maintained through the logical operations. The recognizing networks operate from the logarithmetized output of 19 low-Q bandpass filters. Exceptionally good results have been obtained for six speakers pronouncing fourteen fricatives and vowel-like consonants with ten following vowels since the response accuracy was over 95 % for half the consonants studied. Martin has found that the features which are the more invariant and the more easily abstracted by machine are the spectral regions of increasing and decreasing energy. He shows a certain insistence in his report to claim that he has done away with formants; it would be more exact, however, for him to state that this method has the virtue to define form-

ants and to give useful additional information on the rate of increase and decrease of energy in a given frequency band.

Hemdal and Hughes⁶⁵ have tried to implement on a digital computer a speech recognition program based on the distinctive features approach of Jakobson, Fant and Halle. They were concerned with the recognition of ten vowels and nine diphthongs in Consonant-Vowel-Consonant (CVC) groupings. The features chosen were, as reported by Lindgren:

- 1- acute/grave : (second formant F2 closer in frequency to F3 / or closer to F1);
- 2- compact/diffuse : (F1 higher in frequency than threshold $\approx$ 500 c/s / or lower);
- 3- flat/plain : (F1 + F2 higher in frequency than threshold / or lower than threshold);
- 4- tense/lax : (longer duration and greater departure of the formant frequencies from a neutral position / shorter duration and lesser departure from a neutral position);

the thresholds required normalization as they varied for each speaker. For words spoken in isolation, this method had an accuracy comparable to the one of a human listener and, although the results were poor for continuous speech, this experiment certainly confirmed that a speech recognition system based on the distinctive features can be implemented.

For the sake of completeness, it would be necessary to mention a large number of projects on automatic speech recognition, but the purpose of this short review was to survey what seem to be the most significant attempts. The progress in the field of speech-sounds recognition has been slow but steady and there is good reason to believe that some practical applications will evolve in the near future. Most projects studied until now have tried to recognize speech at the acoustic level; everyone agrees however that this is nothing but a first step, and that unless only very limited applications are envisaged, acoustic information will have to be supplemented by some linguistic information. Fry and Denes were apparently the first ones to suggest that linguistic constraints must be built into speech recognition machines and to show that even a limited amount of linguistic information could drastically reduce the number of errors.

1.5 Speech perception

Speech seems to awake in us resonances like no other signal can: there seems to be very close ties and a great affinity between the hearing mechanisms and speech sounds. It is interesting to consider that ~~speech is such a redundant code for our hearing system that it~~ can undergo transformations which would seem to us to be destructive (such as being infinitely clipped)*

* much spectral information is preserved in clipped speech

and still remain highly intelligible. A number of experiments have also pointed out the very close relations between the hearing and the speech producing mechanisms; for instance, the experiments in shadowing have shown that a person can repeat with very little error a text which is being read in his ear at a normal speed. These experiments have led researchers to put forward two theories of speech perception: the Motor theory and the Analysis-by-Synthesis theory; both of them see the hearing and speech producing mechanisms closely linked together at some level of our subconscious.

The Motor theory³⁹ is based on experiments, which have been made at Haskins Laboratories, and which point out that perception follows the articulation rather than the acoustic value of a signal. In other words, two sounds produced by quite different articulatory movements but being acoustically similar will be perceived as two very different sounds following the articulation necessary to produce the signal rather than its acoustic value. Thus, when one synthesizes speech and gradually varies the transition at the start of the second formant, it is found that instead of the sound being perceived going gradually from the "d" to the "g", there is a sharp boundary between the two phonemic classifications - which will vary slightly however from

person to person. These experiments have given birth to the Motor Theory of speech perception according to which one recognizes sounds by reference to the articulatory movements necessary to produce them.

Stevens and Halle (1958) have proposed an Analysis-by-Synthesis model of speech perception. According to this model, the brain, following certain rules, synthesizes a number of patterns which enable one to perceive. The brain is thus seen as being continually learning and as adapting itself in the face of new experience. Stevens and Halle are of the opinion that speech is "anchored equally in the motor and in the auditory system of man" and that a similar set of rules is used to generate speech sounds and to perceive them.

1.6 The problems in speech recognition

The human hearing system can cope with a variety of signals, with music, various noises, with differences of languages and dialects. The human being is characterized by the apparent redundancy with which he is built and which confers him a very large degree of adaptability. But a practical machine able to recognize speech sounds needs to be built exclusively in function of the specialized task it has to perform and to have as little redundancy as possible.

The difficulty lies in the fact that it is quite possibly the redundancy of our auditive organs which enables us to hear so well. It would be very surprising

if the ear did carry a running Fourier analysis or a cross-correlation analysis on the incoming signals, methods which are for us analytically simple, straightforward and economical. It is much more likely that the human ear perceives through a multiplicity of processes which cannot be very readily and economically simulated. The following chapters point out, however, a few facts known or suspected about perception.

A speech recognition system does not necessarily need, of course, to simulate the workings of the human ear but it is a reasonable way to do it and it is not certain that there is any other way. For one thing, the successes and failures of the past years have pointed out that speech recognition by machines should be modeled on the operation of the human hearing system, i.e. that greater attention should be paid to linguistic factors, that perception should be understood as a process of continuous learning and that a speech recognizing machine should be in some ways adaptive to different speakers.

Recognition at the phonemic level is the basis of most recognition systems that are seriously considered. There is evidence however that the human being, by a running probability analysis, reduces the range of possibilities for the incoming signal, selects a few more important phoneme for perception, interpolates and extrapolates in accordance with grammatical rules the meaning of the sentence and the subject discussed.

Non-sense words are harder to understand than those in common use and it is our everyday experience that words out of context are less easily understood than in a normal conversation. Phoneme recognition is undeniably an essential part of the human recognition process; the enormous memory capacity of the brain, however, renders it less dependent on correct phoneme identification. But a practical machine to recognize speech must take into account the economy realized with recognition at the phonemic level; there are in english, for instance, perhaps 30,000 commonly used words, around 2,000 syllables, but only about 40 phonemes; hence, the necessity to make the most out of phoneme recognition.

One must then segmentate speech in its phonemic constituents; the difficulty lies in the fact that natural speech is not as nicely segmented as one would like: successive phonemes influence one another and the transition between them is gradual.

Linguistic constraints should then be taken into account. This would improve the performance of speech recognition machines at the cost, however, of a much greater complexity. Factors such as intonation and stress contribute very much to speech intelligibility and certainly deserve attention.

Speech recognition involves then much more than just being able to find acoustic invariants between individuals and there is little hope to see on the

market in the near future an instrument able to recognize continuous speech. The problem of introducing linguistic constraints in the analysis seems at present far too intractable, but work is going on on the acoustic characteristics of speech sounds. If one cannot hope in the light of present achievements to recognize speech with machines as easily as human beings are able to, one can hope at least to recognize non-sense syllables with an accuracy comparable to what a human listener is able to do.

Chapter 2

A MATCHED ANALYSER

2.1 Introduction

The purpose of this chapter will be to introduce empirically the notion of a matched speech spectral analyser, that is, an analyser whose filters are matched to the signal to be analysed.

In the next chapters we will see how this matched analyser fits in the fields of signal analysis, speech and aural perception, and finally, how a potentially matched spectral analyser has been implemented.

2.2 Matched filtering

If one wishes to determine the presence or the absence of a signal of known form in a background of noise, one can choose a filter which will give the maximum signal to noise ratio (S/N) at the output, when the signal is presented at the input. In the case of a signal of known form in the presence of additive white gaussian noise, this type of filter is called a matched filter and gives at its output the correlation function of the input with the signal of known form mentioned above.

More generally, in a background of additive noise of frequency spectrum $S(\omega)$ and for a known input signal $M(\omega)$ the frequency response of the filter which will

maximize the S/N ratio at the output is of the form

$$K(\omega) = C e^{-j\omega t} \frac{M^*(\omega)}{S(\omega)} \quad \dots(2.01)$$

where C is a constant. And if $S(\omega)$ is normalized to be equal to the constant C, one obtains:

$$K(\omega) = e^{-j\omega t} M^*(\omega) \quad \dots(2.02)$$

(Wainstein, Zubakov)⁶⁶

2.3 A matched analyser

Is it worthwhile to use a matched spectral analyser for speech analysis, that is, an analyser which can match its filters to the different constituent parts of speech?

In communication and radar applications, one will always use filters which have the optimal bandwidth in order to detect a signal of known form in a background of noise, which often happens to be approximately gaussian. In that type of work, the advantage of using a matched filter instead of, for example, an ideal filter with very sharp cut-off and optimal bandwidth is usually small. For these conditions, Wainstein and Zubakov have calculated in the case of a pulse train in a white noise, an improvement of approximately 1 db in the S/N ratio, improvement which becomes quickly much more significant, however, as one gets farther away from the optimal bandwidth, which would, of course, never

arise in radar. But it is very nearly what happens when analysing the spectrum of speech signals: for example, a formant is not a signal of known form but can vary between wide limits in frequency and in duration; the constituent parts, the building blocks of a speech sound are not well defined and have only been described in very general terms; their frequency and time characteristics are variable. Furthermore, it is not desired to detect the presence of signals in a background of white gaussian noise, but, as will be explained later, in a background of noise represented by all the other signals. Moreover, a spectrum analyser does not consist of only one filter, and it is a very different problem to analyse the advantage to be gained from a matched filter for detecting the presence of a given signal and the cumulative effect of a better match being offered by a whole spectrum analyser to an incoming signal.

2.4 An example

What is meant here can be illustrated by the following example: let us define a time-frequency plane and divide it in squares or rectangles of area $\Delta f \cdot \Delta t$; everyone of these rectangles represents the time and frequency characteristics- windows- of the filters in the spectrum analyser. Let us suppose for reasons of simplicity that these filters have a rectangular frequency

and a rectangular time characteristic despite the physical impossibility of such an example. Let us suppose further that we have a signal which fits exactly in one of these time-frequency cells, but whose position in time and frequency can vary around the center position t_0, f_0 . For this case, which we will call of the matched analyser, one can compute the probability distribution of the amplitude response of each cell if the signal ^{with equal probability} varies from $f_0 - \Delta f/2$ to $f_0 + \Delta f/2$, and from $T_0 - \Delta t/2$ to $T_0 + \Delta t/2$. (figure 2.01).

Let us suppose then that there is another choice for a time-frequency cell: one having the same area but half the bandwidth and, of course, a time window twice as long. We will consider two other cases: case two in which the main cell of interest of the analyser is not centered on f_0, t_0 (figure 2.02), and case three, in which the main cell of interest in the analyser is centered on f_0, t_0 (figure 2.03).

Figures 2.01, 2.02, 2.03 show for these cases the approximate probability distribution of the amplitude response of each cell being equal to or greater than the indicated normalized amplitude value.

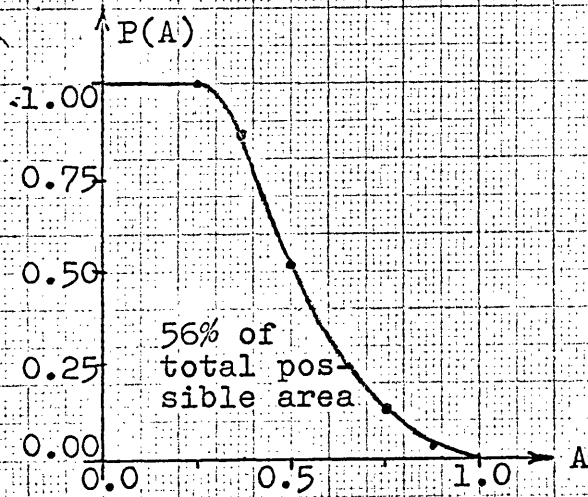
It is difficult to evaluate the advantage of matching a spectrum analyser to the incoming signal; these diagrams can help to visualize how the energy is separated between the different cells; it can be appreciated that the dispersion of the energy is very much

reduced by using matched filters. On one hand, the part of the energy which, on the average, will be left in the central box will be much larger in the matched case ($56^{\circ}/_{\circ}$ compared to $43^{\circ}/_{\circ}$ or $35^{\circ}/_{\circ}$). On the other hand, the rest of the energy is better dispersed in the adjacent boxes so that, in the matched case, no other filter gets more than $10^{\circ}/_{\circ}$ of the energy while, in the second case (figure 2.02), one other filter gets $35^{\circ}/_{\circ}$ and the others $12^{\circ}/_{\circ}$ each. In the third case (figure 2.03), two filters get $25^{\circ}/_{\circ}$ each. Matched filters minimize the probability of incorrect detection as well as of incorrect non-detection. It can be seen that as more of these signals appear on the time-frequency plane, the matched analyser will give a much clearer picture of the situation. Moreover, the resolution in frequency and in time will be as high as it can be useful: in the unmatched case illustrated, frequency resolution will seem to be better while being in reality the same but the resolution in time will be twice as bad.

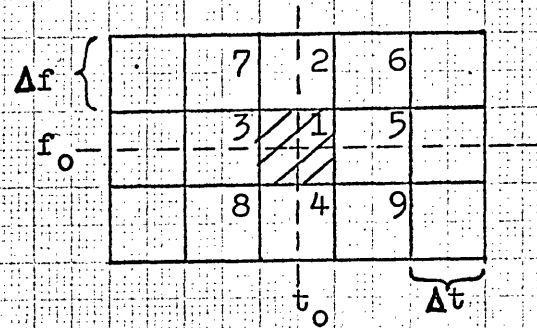
2.5 Conclusion

All this discussion rests, of course, on the assumption that a signal of the appropriate shape is presented to the analyser. Speech is not a signal, it has already been said, which lends itself to a theoretical or mathematical treatment in terms of its basic com-

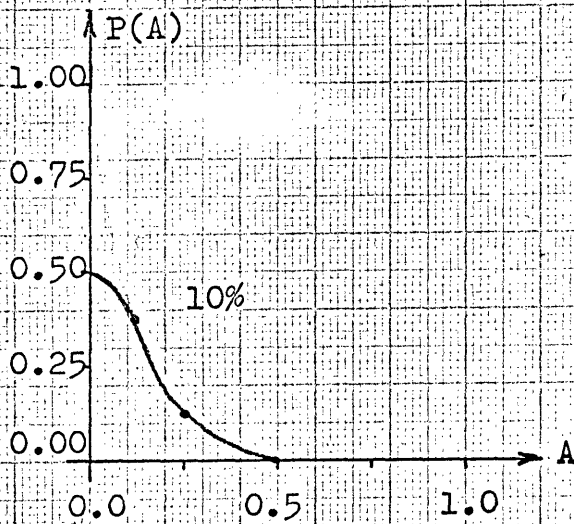
ponents and we can only define it in terms of its general properties. Whether these properties are such that a matched analyser would be of some use is a question which cannot be answered yet at this stage.



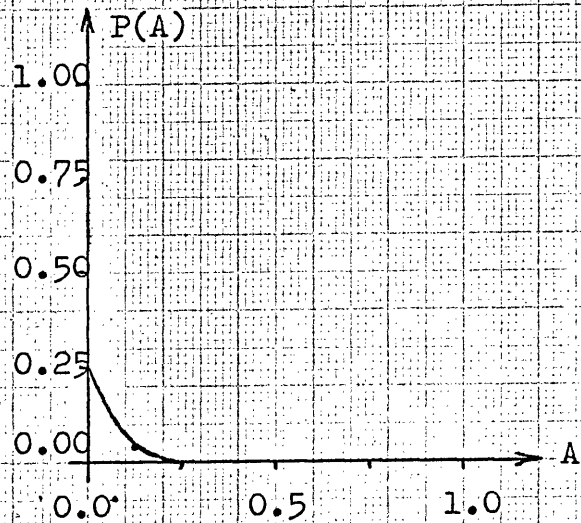
a) area no. 1.



time-frequency characteristic of analyser.



b) areas no. 2, 4, 3, 5.



c) areas no. 6, 7, 8, 9.

Figure 2.01

Probability distribution of the Amplitude response at the filters for case 1 ($P(A)$ vs A). Maximum amplitude = 1.0

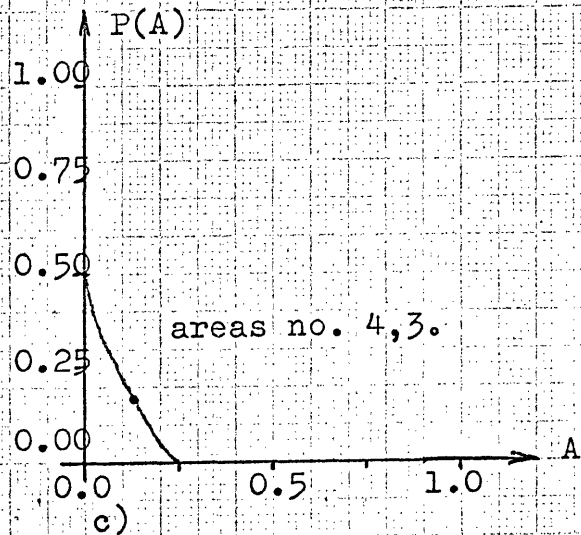
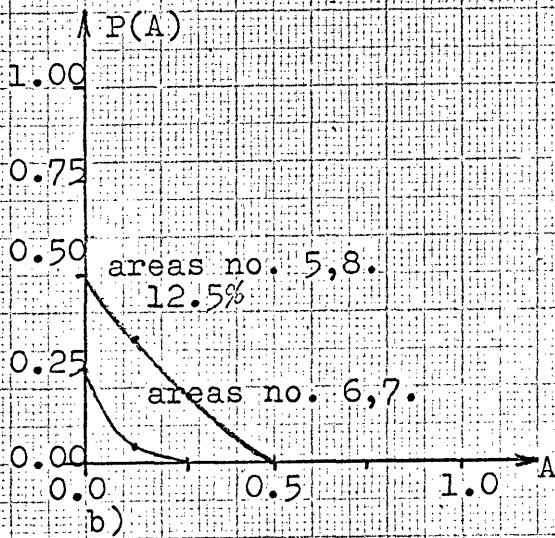
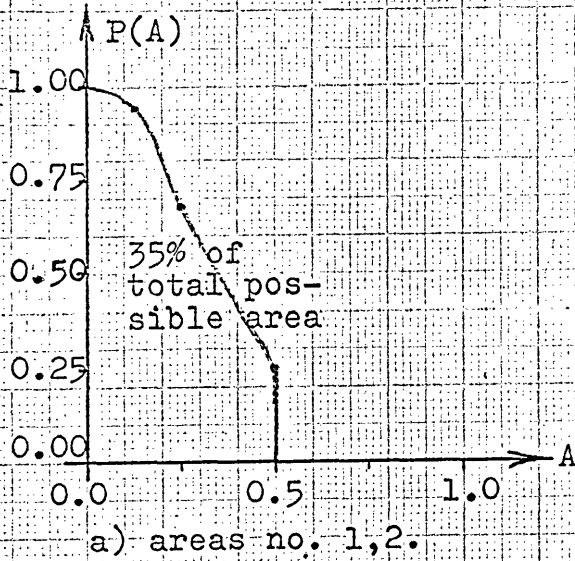
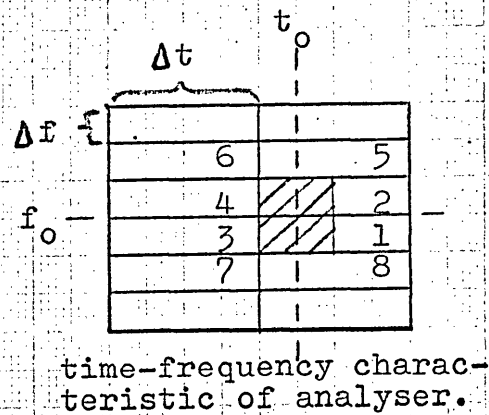
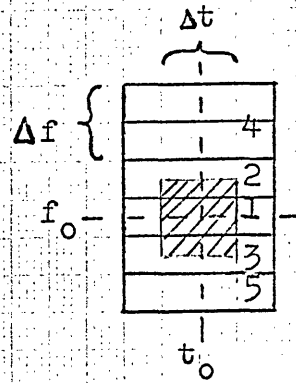
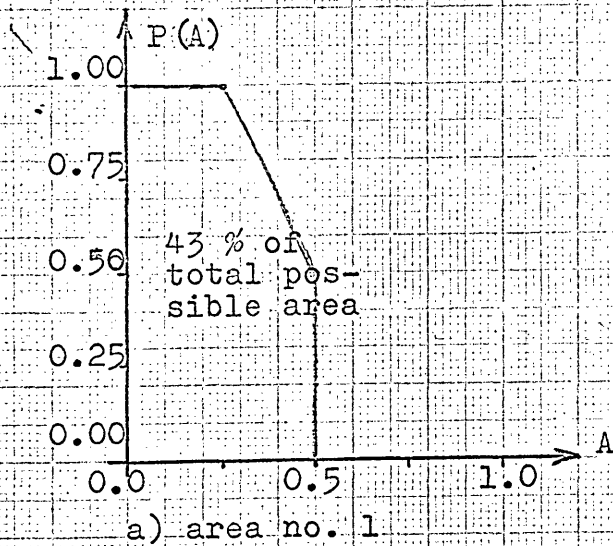


Figure 2.02

Probability distribution of the Amplitude response at the filters for case-2 ($P(A)$ vs A). Maximum amplitude = 1.0



time-frequency characteristic of analyser.

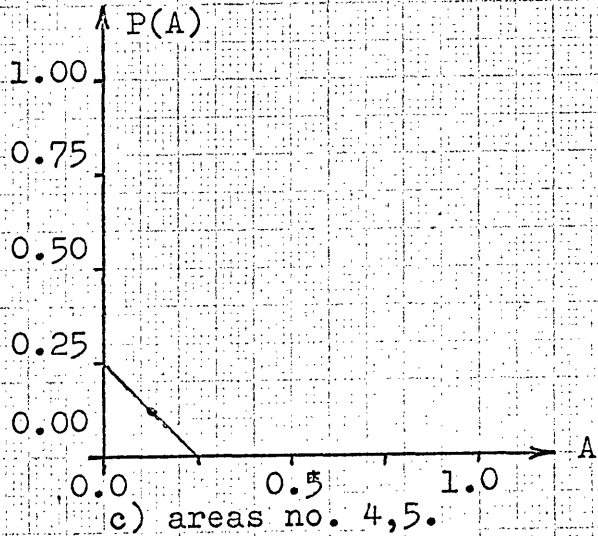
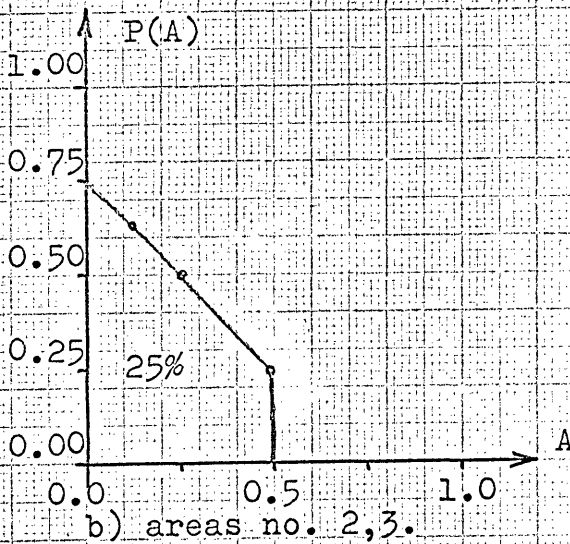


Figure 2.03

Probability distribution of the Amplitude response at the filters for case 3 - ($P(A)$ vs A). Maximum amplitude ≈ 1.0

Chapter 3

TIME-FREQUENCY RELATIONS IN SIGNAL ANALYSIS

3.1 Introduction

This chapter has three aims.

First of all, it provides the theoretical basis to the empirical ideas of the previous chapter, it develops from the concepts of frequency and of time analysis the idea of frequency-time analysis.

Secondly, it introduces the notion of sampling and shows what it means to match a spectrum analyser.

Thirdly, it introduces the concept of uncertainty in signal analysis, which will be found to be relevant in our study of speech and aural perception. Historically, the thoughts on the time-frequency relations in speech and aural perception came after similar concepts in signal analysis had been known for some time, and it is perhaps also more logical to present them in this order.

3.2 Time-frequency relationships in signal analysis

The complexity of the phenomena one has to deal with has brought about an intensive use of statistical methods-- probability, correlation methods and the like - which can be used for the study of the most vaguely defined signals, such as, for example, stochastic processes. But the engineer is limited in any exact

analytical treatment by the complexity of the mathematical tools at his disposal and the relative simplicity of the problems which these tools can satisfactorily handle. Signal analysis has traditionally been based on the two reciprocal Fourier integral relations:

$$S(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S(f) e^{j\omega t} df \quad \dots(3.01)$$

$$S(f) = \int_{-\infty}^{+\infty} S(t) e^{-j\omega t} dt \quad \dots(3.02)$$

These relations are strictly applicable to band unlimited signals of infinite duration, conditions which are of course not realised by any physical signal but which one has to assume when using Fourier analysis. Moreover, the Fourier analysis fixes the frequency spectrum of a signal for all times, which does not correspond to our physical intuition and to what our senses tell us. To illustrate this last point, the example is often given of a frequency modulated wave whose Fourier spectrum is a combination of fixed frequency components but which appear to us to move up and down in frequency as time goes on.

According to equation 3.01, it is possible to find the value of function $S(t)$ at any time "t" by integrating over an infinite frequency band. Equation 3.02 is reciprocal. It is well known that for the conditions expressed by the two previous Fourier re-

lations, the Fourier coefficients are continuous, that is, one has an infinity of data in the frequency or the time domain to characterize a function. Such signals, however, unlimited in frequency and in duration are not physically observable. Moreover, as already written, we do not intuitively consider signals as being only functions of time or of frequency but as functions of both these parameters in an unknown way. "Unknown way" does not refer here to a mathematical, but to a data processing operation - and in the case of speech signal to a psychological operation: one can from a purely mathematical or analytical point of view obtain an infinity of equivalent or interchangeable time-frequency representations for the same signal. From the viewpoint of data processing, however, or in the field of human communications from a psychological viewpoint, certain ways of analysing a signal will show themselves to be more useful, the information sought will be more directly accessible. Information refers here to the content of the message and not to the capacity of a communication link as in information theory.

3.3 A theory of sampling and of spectral analysis

The sampling theorem says that a time function band-limited to frequency F is perfectly determined by the samples taken at regular intervals at a rate of $2F$. Mathematically this can be expressed by the following

formula:

$$f(t) = \sum_{X=-\infty}^{+\infty} f(X \cdot \Delta t) \left[\frac{\sin(t - X \cdot \Delta t)}{(t - X \cdot \Delta t)} \right] \dots (3.03)$$

where $\Delta t = \frac{\pi}{\Omega}$ for a function band limited between frequencies $-\Omega$ and $+\Omega$ where all samples are orthogonal and independent. Reciprocally, it is possible to show that in a frequency domain

$$F(\omega) = \sum_{Y=-\infty}^{+\infty} F(Y \cdot \Delta \omega) \left[\frac{\sin(\omega - Y \cdot \Delta \omega)T}{(\omega - Y \cdot \Delta \omega)T} \right] \dots (3.04)$$

where $\Delta \omega = \frac{\pi}{T}$. The meaning of these two statements is that if $f(t)$ is a band limited signal, its value at any time is determined by the value formula (3.03) takes at times $t = X \cdot \Delta t$ which are Δt time units apart. Conversely, if $F(\omega)$ is a signal limited in time, its value at any frequency is determined by the value formula (3.04) takes at the frequencies $Y \cdot \Delta \omega$ which are $\Delta \omega$ apart. It should be noted that the number of samples needed in the last case is equal to the number of Fourier spectrum coefficients ($\frac{2\pi}{T}$ for positive and negative frequencies).

The sampling theorem would allow us ideally to eliminate redundancy in physical measurements although this aim is not fully attained in practice. Since it is always possible (theoretically) to define any point of the function by using these samples, it is unnecessary and useless, always from a theoretical point of view, to take more measurements; these samples are then

sufficient to measure any characteristic of the function.

Real signals are however finite in duration and in frequency. If one divided such a signal in intervals of short duration τ , and truncated the infinite frequency band required to transmit the signal to band F , one would still obtain the same rate of sampling $2F$. One would find however that each sample is a measure attached to a small rectangle of dimensions Δf and Δt in the time-frequency plane. These rectangles are the counterpart in the time-frequency plane of the Fourier coefficients or of the time-amplitude samples of the previous paragraphs: in the same way as one cannot define more than $2FT$ independent samples or Fourier coefficients, these rectangles are the fundamental units from which the signal is made and they are named for this reason "elementary areas". One can represent the time-frequency plane as divided in small rectangles $\Delta f \cdot \Delta t$; it is customary to represent the amplitude by the density of the shading and the phase by its orientation in the diagram as in figure 3.01. Certain authors prefer to speak of $2FT + 1$ sample, the extra sample referring to the DC component.

3.4 The uncertainty relation

Gabor has tried to determine the shape of the time and frequency windows of smallest elementary area. It is desired to get in this way a number of samples

which will be as independent as possible from each other; orthogonality between samples such as in the processes defined by equations 3.03 and 3.04 will not be obtained. After having transformed the signal in such a way that the negative frequencies are cancelled and the positive ones are doubled, Gabor takes as the mean value of order 1,2,...n of the epoch "t" and of the frequency "f" of the signal, the value of the moment of the same order, normalized by dividing it by the moment of order zero.

He then defines an effective duration width Δt as " 2 times the r.m.s. deviation of the signal from the mean epoch $\bar{t}$ " and the effective frequency width similarly as " 2 times the r.m.s. deviation from the mean frequency $\bar{f}$ ". This definition leads to a relation of the form:

$$\Delta f \cdot \Delta t = \frac{1}{2} \cdot \left[\begin{array}{c} \text{a form of the Schwarz} \\ \text{inequality} \end{array} \right] \geq \frac{1}{2} \dots (3.05)$$

This relation illustrates the uncertainty principle of signal analysis. In words, equation (3.05) means that it is impossible to define the time of occurrence of a signal received through a channel of bandwidth Δf , with a precision better than $\Delta t = \frac{1}{2\Delta f}$ ~~Δf~~ in the most favourable case, that is, if equation (3.05) turns into an equality. Gabor finds that the function for which the product $\Delta f \cdot \Delta t$ is the smallest takes the

form

$$\psi(t) = e^{-\alpha^2 (t - t_0)^2} e^{(2\pi f_0 t + \varphi)} \dots(3.06)$$

$$\text{and } \Phi(f) = e^{-\left(\frac{\pi}{\alpha}\right)^2 (f - f_0)^2} e^{[-2\pi t_0 (f - f_0) + \varphi]} \dots(3.07)$$

$$\text{where } \Delta t = \sqrt{\frac{\pi}{2}} \cdot \left(\frac{1}{\alpha}\right) \dots(3.08)$$

$$\text{and } \Delta f = \frac{\alpha}{\sqrt{2\pi}} \dots(3.09)$$

The time and frequency envelopes of the elementary signals are gaussian normal curves.

In order to expand a function in terms of elementary areas or "logons" as they are often called, Gabor considers first rectangles of twice the size of an area, i.e. $\Delta f \cdot \Delta t = 1$ and finds, expanding the function in terms of these:

$$\psi(t) = \sum_{n=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} C_{n,k} \exp \left[-\frac{\pi(t-n\Delta t)^2}{2(\Delta t)^2} \right] \exp(2\pi k t / \Delta t) \dots(3.10)$$

As can be seen on figures (3.01 d) and (3.01 f) the time and frequency windows overlap and the samples cannot be orthogonal. The value of the $C_{n,k}$ coefficients are obtained by successive approximations. These coefficients are complex numbers, from which a sine and cosine terms are introduced; from these one can divide each

sub-elementary area of figure (3.01 e) in its two fundamental components - figure (3.01 b) - and, again by successive approximations, find the value of the logons.

An instrument insensitive to phase, such as a spectrum analyser, can at best analyse a signal in its sub-elementary areas, as defined by $C_{n,k}$; they have the same type of time and frequency envelopes as defined by equations 3.06 and 3.07.

There are many alternative definitions of Δf and Δt . One, which Thomas used in his thesis, was proposed by Wainstein and Zubakov⁶⁶; they define $\Delta\omega$ as being half the width of the rectangle of height $S(\omega)$. They similarly define Δt as half the width of the rectangle of height $R(0)$ comprising the same area as the curve of the autocorrelation function $R(t)$. In this case, $\Delta\omega$ and Δt represent respectively the pass-band and the memory of the filter and one finds that the product $\Delta f \cdot \Delta t$ is always equal to one half.

$$\int_{-\infty}^{+\infty} S(\omega) \cdot d(\omega) = S(0) \cdot \Delta\omega \quad \dots(3.11)$$

$$\int_{-\infty}^{+\infty} R(t) \cdot dt = R(0) \cdot \Delta t \quad \dots(3.12)$$

but since ..

$$S(\omega) = \int_{-\infty}^{+\infty} e^{-j\omega t} R(t) \cdot dt \quad \dots(3.13)$$

$$\text{and} \quad R(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{j\omega t} S(\omega) \cdot d\omega \quad \dots(3.14)$$

$$S(0) = \int_{-\infty}^{+\infty} R(t) \cdot dt \quad \dots(3.15)$$

$$\text{and} \quad R(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S(\omega) \cdot d\omega \quad \dots(3.16)$$

$$\text{hence} \quad \Delta f \cdot \Delta t = \frac{1}{2} \quad \dots(3.17)$$

Slepian, Landau and Pollack have shown that prolate spheroidal wave functions as elementary functions give a better approximation for the case of band-limited signals; one needs then also $2FT + 1$ samples (including a constant) to approximate the function with a greater precision than a certain factor $12E_t^2$, where E_t^2 represents the proportion of the energy outside the time interval T . In mathematical form, if

$$\int_{-\infty}^{+\infty} f^2(t) \, dt = 1 \quad \dots(3.18)$$

$$\text{and} \quad \int_{-T/2}^{+T/2} f^2(t) \, dt = 1 - E_t^2 \quad \dots(3.19)$$

$$\text{then} \quad \int_{-\infty}^{+\infty} \left[f(t) - \sum_0^{2FT} a_n \varphi_n \right]^2 \cdot dt < 12E_t^2 \quad \dots(3.20)$$

Wolf has pointed out that the use, as Gabor does, of a linear transformation to eliminate the negative frequencies deprives the signal of much of its physical significance and he shows that the uncertainty principle should read as

$$\Delta f_+ \cdot \Delta t_+ > \frac{1}{2}$$

where the plus sign indicates the positive values of time and frequency. He proves however that $(\Delta t)^2$ as defined by Gabor can always be interpreted as the normalized second moment of the envelope of the real signal; if there is no d.c. component, Gabor's definition of Δt holds. It remains however that the uncertainty product for Gabor's elementary signals could be, in theory, ~~more than equal to~~ one half since their definition includes negative frequencies.

Kay and Silverman have confirmed the results of the work of Wolf, but have shown that it is possible, in certain conditions which are not fulfilled by physical signals, to define an uncertainty product appreciably smaller than one half.

3.5 Conclusion

The approaches described in this chapter have very much in common in their results. Except for one, which can be disregarded because of its practical inapplicability, they all point out the same kind of uncertainty between time and frequency measurements.

Gabor's treatment, as modified by Wolf, has the merit of providing a general approach to the field of signal analysis, linking the concepts of the sampling theorem to those of the Fourier analysis and to suggest a practical and realizable method of sampling. His definition, in particular, of an elementary area, is

a useful concept for spectral analysis and it will be put to practical use in the remainder of this thesis.

It has certainly been noticed that when one comes at the boundaries of the area on the information diagram which Gabor divides in elementary areas, the elementary areas at the boundary overlap somewhat with the outside regions. Prolate-spheroidal wave functions provide a way to obtain sharp frequency cut-off while having some overspill in time, which is expressed by the factor $12E^2$. It would not seem however that this would be a refinement of great importance for engineering applications and to try to take samples in terms of prolate-spheroidal wave functions would not seem to be warranted.

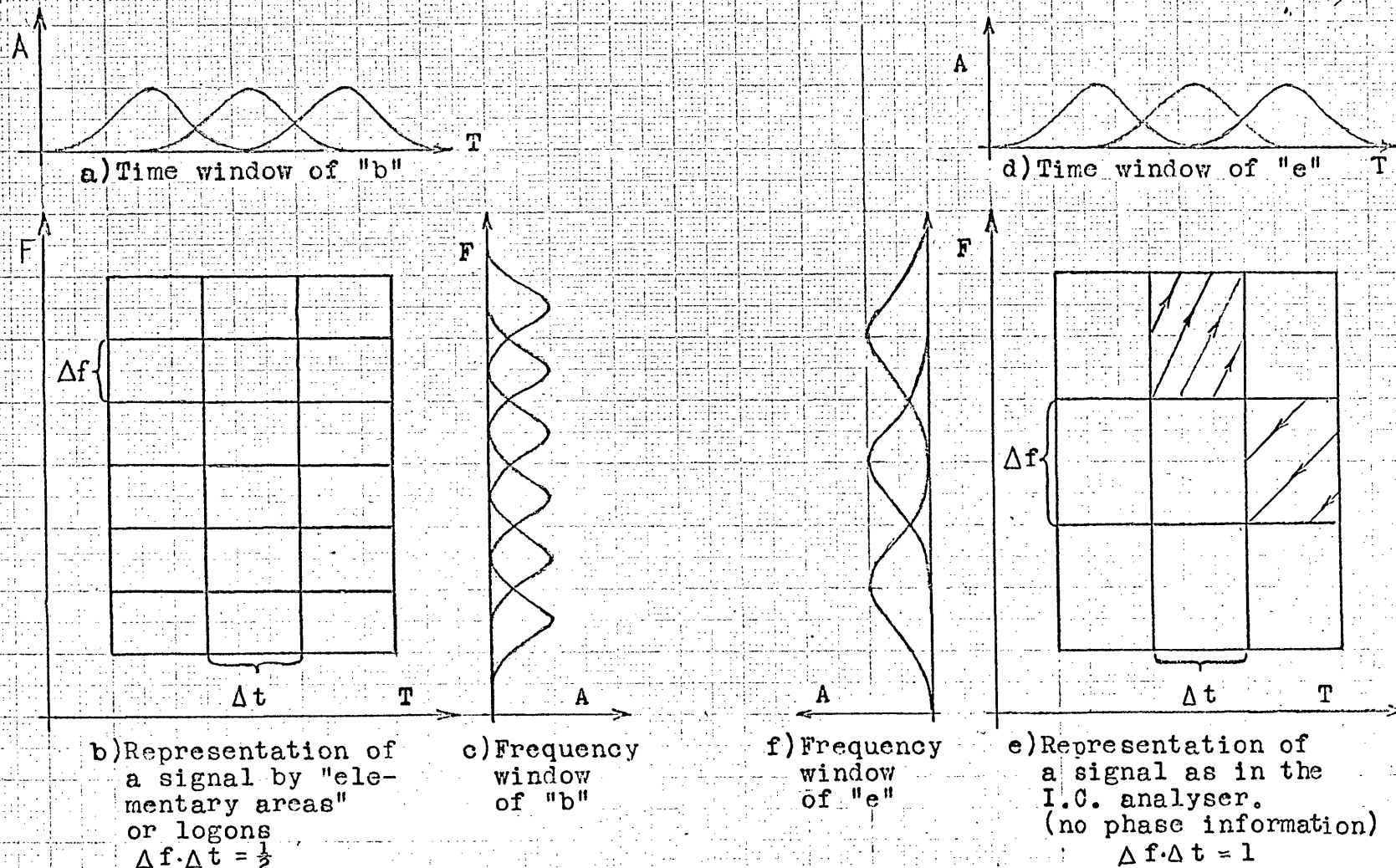


Fig. 3.01

Time-frequency representation of signals by their elementary areas with their frequency and time envelopes (windows).

A:amplitude; $\Delta f, F$:frequency; $\Delta t, T$:time

Chapter 4

TIME-FREQUENCY RELATIONSHIPS IN SPEECH AND AURAL PERCEPTION

4.1 Introduction

What has been written in the first chapter shows how complex a signal is speech, and how incomplete is our knowledge of the mechanism of speech perception. The study of speech is linked with the study of perception; speech is analysed as signal only in relation with the way it is perceived.

In particular, researchers have tried to link together the time-frequency patterns of speech and the time-frequency characteristics of the human aural perception: to define the object of our work in the light of these studies is the aim of this chapter.

4.2 Time-frequency relations in aural perception

4.2.1 Auditory perception

Helmholz was mainly responsible for originating the idea that the ear operates as a frequency analyser. Although, qualitatively, this view comes rather near to explaining the operation of the human hearing system, it is far from taking into account the complexity and the non-linearity of its operation. After the initial impressive success of the visible speech displays, this view came to be even more widely held: Gold and Pumphrey,

for example, studied the "cochlea as a frequency analyser" (1947) and tried to measure the oscillatory time constants of the resonance elements of the ear. More recent observations, however, although some of the results seem contradictory, point out that no satisfactory theoretical model of the ear has been evolved so far. One interesting conclusion for us is that the available data indicates that the ear cannot be a fixed parameter device but that a certain amount of adaptation to the incoming signal must take place. *This will be made clearer in the remainder of this chapter.*

It is difficult to present a short and concise summary of what is known of the characteristics of the ear because of the complexity of the auditory system and of the wealth of measurements which have been taken under different conditions, with different stimuli and so on. The coming paragraphs will try to present a general picture, to review some of the more important literature and in this way, to indicate the guide lines of what is known about the ear. The aspects of aural perception which have probably been the most studied (apart from speech perception) are the perception and recognition of pitch, the detectability of signals in noise, and the perception and recognition of delays.

A number of researchers such as Hirsch³², Wallack⁶² and Newman and Rosenzweig⁴⁸, Miller and Taylor have found under different conditions that a separation of 2 msec. between the time of occurrence of two brief clicks or

bursts of noise was sufficient for the ear to detect the presence of two signals. Wallack and al. have noted however that when two clicks are separated by less than 2 msec, the resulting sound appears sharper or duller according to the amount of separation, which ^{could} ~~does~~ indicate that the ear is sensitive to very small time differences indeed. These same authors note that delays of the order of a few microseconds are perceived with binaural hearing.

It is however after 10 msec. that a burst of sinusoidal tone in the medium audio frequency range (200 cps to 3,000 cps) starts to have a definite, recognizable and hence reproducible pitch, as was shown by the classical experiments of Bürck, Kotowski and Lichte (B.K.L.). In another series of experiments, they measured the delay necessary before the occurrence of the second tone after the occurrence of the first one, in order to detect the presence of two tones. Once again, they measured values of the order of 10 msec except when the frequencies of the two tones were the same or nearly so; in this latter case, the value was around 21 msec for a tone frequency of 500 c/s.

Another measure of great interest is the delay needed to ascertain which of two tones or two noises occurred first. Hirsch experimented with two signals lasting approximately 0.5 seconds and repeated at intervals of about 2 seconds, one of them being delayed at

the start with respect to the other and both of them being terminated simultaneously: he found that 17 msec were needed for a reliability of 75⁰/₀ . Broadbent and Ladefoged have commented that their own experiments and those of Christovich³⁴ show this value to be optimal because of the repetitive stimuli and the effects of learning, and suggest that a delay of 170 msec is necessary for perfect discrimination. It should be noted that such conditions arise in speech perception to recognize words like fits and fist, mist and mits, wasp and waps.

The frequency selectivity of the ear for a signal of long duration is of the order of a few cycles per second. It is however our everyday experience that the precision of our pitch discrimination improves as the duration of the signal is increased - between certain limits. Shower and Biddulph have measured the minimum perceptible frequency change in a modulated sine-wave of two seconds duration. In the frequency range from 50 c/s to 1 kc/s, they found this value to be better than 3 c/s, the performance worsening at higher frequencies; more will be said on this experiment in a subsequent paragraph.

4.2.2 The "uncertainty relation" and auditory perception

Is it possible to define in psycho-acoustics an uncertainty relation similar to the one in signal analysis and other branches of science? If one con-

siders that the uncertainty relation is a measure of the degree of doubt there is in a given perception, one cannot see how a human being could arrive at measuring the quality of his perception. A human being can compare his sensations, he can rank them in a given order, but he cannot possibly state that the indeterminacy of this measurement is so many cycles per second. And there are other difficulties brought about by the difference between the subjective and objective scales of measurements. Because the degree of uncertainty in a sensation does not easily lend itself to measurement, researchers have usually replaced it by a measure of the minimum noticeable difference between two signals. The difficulty here is that two stimuli are involved, but the advantage is that this technique enables one to get some kind of measure of an "uncertainty relation". It should be mentioned, however, that to speak of an uncertainty relation in this case is quite irrelevant since the Δf referred to here is the frequency difference between two signals separated by an interval of the order of a quarter or a half of a second.

It was Stewart who was the first to ask whether the well known fact that a pure tone gradually loses its tone quality as its duration is shortened would not be explained by an uncertainty relation of the type Heisenberg had developed in physics. He stated the following equation $\Delta f \cdot \Delta t \approx 1$ without however defining

exactly what he meant with these parameters.

It was Gabor, finally, who took the matter up and was enabled by the experiments of Bürck, Kotowski and Lichte, and those of Shower and Biddulph, to define the problem and give a more definite answer. From the result of B.K.L. that a duration of 10 msec was needed for the ear to ascertain the pitch of a 500 c/s tone, he concluded that the threshold area was characterized by the bandwidth of this signal. Noting that the ear cannot notice that the amplitude of a 500 c/s tone is doubled if this happens less than 21 msec after the start of the signal, Gabor concluded that it takes 21 msec before the ear is ready to accept another bit of data and that these 21 msec represented the other dimension of the threshold area which he computes to be 2.1 at 500 c/s and 3.0 at 1,000 c/s. Analysing then the data of Shower and Biddulph, he approximates the spectrum of the frequency modulated wave they used with elementary rectangles of area one half. These elementary areas are, however, somewhat scattered on the information diagram and Gabor divides the area of interest of the information diagram between the elementary areas and calculates the value of the threshold areas for different values of frequency. He finds an average value of 2.3 for the frequency range from 62.5 c/s to 1 kc/s and higher values for higher frequencies instead of the ideal value of $\frac{1}{2}$ if threshold and elementary areas co-

incide . From this, he concludes first of all that the size of the threshold areas on the information diagram is a characteristic of the ear, at least up to 1,000 c/s and from 20 to 250 msec. In other words, the ear can adapt the size of its threshold areas between at least 20 msec (B.K.L.) and 250 milliseconds (as indicated by the experiments of Shower and Biddulph). The size of the threshold area tells us that under 1,000 c/s, the ear is almost as perfect as any instrument insensitive to phase differences can be. In particular, it could be said to be analogous to a spectrum analyser which would have this possibility of adapting the shape of its elementary areas to suit the characteristics of the input signal.

An experimental verification of the uncertainty relation for times rather longer than 20 msec (perhaps around 50 msec) could be achieved by experiments similar to those of Shower and Biddulph. But one does not seem to be able to imagine a method of measuring the indeterminacy of the pitch for shorter time intervals. Different authors (Oettinger, ~~Christovich~~, Liang Chic-an and Sekey) have tried to replace this measurement by another one: the minimum perceptible frequency difference between two short tone bursts of duration Δt separated by a pause t_p . Although, as was pointed out earlier, Δf and Δt do not correspond to the same parameters in the uncertainty relation, it is nevertheless

apparent that they are a measure of the same phenomenon, and can be taken to be a useful indication of the shape and the limits of the uncertainty concept.

Oetinger measured the minimum perceptible frequency difference Δz in mels between two tone bursts of duration Δt separated by a pause of 0.25 sec.* He found out that the relation could be expressed by the following relation valid for durations of 1 to 100 msec.

$$\Delta z \cdot \Delta t^{0.7} = \text{constant}$$

Similar experiments were also performed by Liang Chic-an and Christovich³⁷: their results show the same trend as those of Oetinger.

Sekey has more or less repeated Oetinger's experiment with the difference that he defined a trading relationship Δf and Δt which remains constant as long as the error rate of the observers remains constant: this error rate refers to the decision of the observers as to whether the second tone burst had a higher or lower pitch than the first one. An interesting conclusion is that, considering his results, the hypothesis of the ear adaptivity is more likely to be true than one which postulates that the ear has a fixed set of resonant

*the mels scale is defined here as following the normal frequency scale up to 500 c/s and not 1,000 c/s as in the North-American norms.

elements.

Other authors came to the same conclusion through different experiments, although no suitable model of the ear has been set up yet. Green, Birdsall and Tanner, for example, have studied auditory signal detectability in noise as a function of signal intensity and duration, and compared the performance of the ear to the one of an optimal receiver knowing the shape of the signal beforehand. In one experiment, the duration was held constant and the amplitude varied, then amplitude and duration were varied for a constant energy and thirdly, the duration was varied while the amplitude was held constant. Experiment two was entirely compatible with the hypothesis of the optimal receiver for the signal of known form and experiments one and three showed discrepancies, which is not surprising since the hypothesis in any case cannot possibly be held, since there must certainly be some delay between the occurrence of the signal and the operation of the ear. The conclusion of the authors was that "the linear filter model is not an adequate model to account, on a quantitative manner, for the effect of signal duration upon detection". Creelman studied the detection of a stimulus wave form consisting of a train of damped sinusoids: "the results are interpreted as suggesting that observers are able, with high amplitude complex signals as well as with sinusoids, to match their receptive systems to the signals to be

detected. This ability seems to break down when the signals are longer than a certain range or when the signal consists of trains of highly damped sinusoids".

In another experiment concerned with the detection of a noise signal in noise, Green has found that the ear operated nearly as an optimum detector except for a constant attenuation factor: his results were consistent for signal durations of 3 to 300 msec, bandwidths of 500 to 5,000 c/s and center frequencies of 400 to 6,000 c/s.

4.2.3 Summary and conclusion

It appears then that the existence of an uncertainty relation applicable to auditory perception is virtually certain and well-known, although an exact definition - measure - of this relation has yet to be found.

This uncertainty relation has been shown by Gabor to be applicable at 20 msec and 250 msec for frequencies up to 1,000 cycles/second and it has been assumed to be applicable in the whole audio frequency range and at least in the time interval from 20 to 250 msec. Other researchers have taken measurements of a phenomenon which is obviously related; if one accepts certain assumptions, namely that the minimum detectable frequency shift is a measure of the degree of uncertainty and that the duration of the pause does not affect

substantially the nature of the results, these measurements can be taken as an approximation of the uncertainty relation or, at least, as indirect evidence of its existence. This uncertainty relation or trading relationship seems to retain its validity down to durations of one millisecond.

4.3 Time-frequency characteristics and speech perception

4.3.1 Introduction

Considering the data which has been gathered, one can wonder how does all this affect the perception of speech? How does the ear make use of its capabilities? What are the optimum time-frequency characteristics? How does speech perception require that the ear uses its adaptation faculties? If so, between which limits? No definite answers to these questions are available but it is the object of this section to provide some preliminary and partial indications.

4.3.2 Speech characteristics

Are the characteristics of the various speech sounds different enough to warrant the use of adaptive filtering?

What are these characteristics?

The realm of speech sounds goes from those phonemes, the vowels, which are characterized by very well defined

frequency resonances, to the stop-consonants with their sharp noise burst and the accompanying sudden frequency transitions which links them to the adjacent vowel. In between, one finds diphthongs, for example, where the frequency value of the resonances can already change much more rapidly than in a vowel sound; closer to the stop-consonants, one finds the fricatives, some of which are still characterized by a rather faint formant pattern while others have purely a noise structure and are distinguished by the bandwidth they occupy.

In these circumstances, the conclusion that the ear probably uses its adaptive faculties to recognize speech sounds seems difficult to escape. If one considers the two extreme cases of the vowels and the stop-consonants, one can certainly ^{advance} ~~arrive at~~ the conclusion that they are perceived in relation to different criteria: for example, it is possible and even easy to find the equivalent pitch of the vowel by matching it to a pure tone; but to match a stop-consonant and many other consonants to a pure tone and find their pitch is quite an impossible task. Halle, Hughes and Radley have noted that all attempts to find the pitch of stop-consonants are completely unconvincing and usually refer to the second formant of the adjacent vowel.*

Some additional remarks by the same three authors should be of interest here: "Physically speaking, a formant reflects organization of the acoustic energy in
* It maybe, of course, because they don't last long enough.

the frequency domain and no organization in the time domain, while a burst represents organization of acoustic energy in the time domain and no organization in the frequency domain. In the case of an ideal formant, there is no energy outside the infinitely narrow resonance frequency band, but the sound lasts for ever. An ideal burst, on the other hand, has an infinitely short duration; i.e. the energy is present only at a given instant but its bandwidth is infinite".

Between which limits would adaptation need to take place for speech perception?

On the one hand, if one considers that a signal of a millisecond duration would have a bandwidth of about 1,000 c/s, one has to agree that such short time intervals are probably not significant for speech perception; if one furthermore thinks at the distortions speech can be submitted to without being destroyed, one would tend to conclude that the perception of speech sounds does not depend on such small time differences.

On the other hand, if one thinks at the speed with which sounds vary in continuous speech, it is unlikely that the sounds can be so carefully produced by the vocal tract that they need to be analysed with a very high precision.

I would personally consider that bandwidth of 50 c/s would correspond to the maximum degree of resolution which one would require for normal speech, and I

would tend to consider that a time resolution much better than 5 msec - corresponding to a filter bandwidth of 200 c/s - would probably not be found very useful.

4.3.3 Conclusion

If the human hearing system seems to make use of an adaptive process in its time-frequency analysis in order to recognize signals, if speech seems to be the kind of signal that would require the ear to make use of adaptive filtering, would it not be logical to modify spectrum analysers so that they can adapt themselves and match their filters to the incoming signal?

As a conclusion, some remarks by Cherry would seem pertinent; commenting on spectrum analysers, he points out that this instrument is not a model of the ear although it goes a long way toward representing the process in the ear: "For one thing, writes Cherry, when measuring a spectrum, the ear is set up as a single parameter system, with a fixed set of filters; the parameters of the instrument are not controlled in any way by the signals themselves....Plosives, fricatives, vowels, et cetera may require different filtering in the spectrograph for their different characteristics to be displayed more prominently. It would be more accurate to compare the ear to a set of such visible-speech spectrographs rather than to one of fixed parameters - or to one which has the power of varying its own para-

meters according to the signals being received (e.g. the $\Delta F/\Delta T$ "aspect ratio" of its logons)".

4.4 Adaptive spectral analysis

Whilst keeping to the same rate of sampling it is possible to change the aspect ratio of the elementary areas and to trade between time and frequency resolution. The indications are that the black box which contains our model of the human hearing system also has this potentiality of adapting the aspect ratio of its receivers, and speech sounds seem to be of a different enough texture to warrant the use of this potentiality. There is no suggestion, however, that we are trying to make a copy of the human hearing system. We are merely thinking of the ear as a black box and we have decided that adaptive spectral analysis is a better way than non-adaptive spectral analysis to simulate the operation of our black box. We will then not indulge in discussions on neuron behaviour or on the critical band concept that would be necessary in order to find out how the hearing system operates in reality.

It will be realized in the remainder of this work that the adaptability that could be studied and implemented is, after all, very limited and falls considerably short of what the hearing system is presumably able to do. For one thing, the frequency discrimination of the ear diminishes as one goes up in frequency: the filters of our analyser have at any one time the same bandwidth

value.

Secondly, it could be that one would have to adopt large bandwidth filters for one part of the spectrum and narrower bandwidths for another part: it will become clear that the particular proposals contained in this thesis are much more limited in scope.

Thirdly, it must be expected that the optimum bandwidth is continuously changing. In our case, however, some fixed values had to be chosen and it was finally resolved to take bandwidths of 200 c/s and 100 c/s for the bulk of experiments. ^(64B bandwidth) This decision made implementation much simpler as can be appreciated from Chapter 6; furthermore, as will be seen in Chapter 9 and Appendix B, this permitted to have the same sampling instants in the different modes, the importance of which will become clearer later on.

Perhaps bandwidths of 250 c/s and 50 c/s would have been more appropriate or more in line, at least, with the normal practice in phonetics; the differences would have been more marked but it was thought that this constituted a first step which could be improved upon if it was shown to be of interest.

4.5 Definition

In the remainder of this work, if the analyser is said to be in the 100 c/s mode or the 200 c/s mode, or if one speaks of the 100 c/s display or the 200 c/s display, it should be understood that the bandwidth

of the filters in the analyser is 100 c/s or 200 c/s.

A definition of bandwidth is given in Chapter 5.

Chapter 5

THE IMPERIAL COLLEGE AUDIO SPECTRUM ANALYSER

5.1 Introduction

A spectrum analyser of the bank of filters type was developed by R. S. Thomas in the Signals Laboratory at Imperial College. Since this analyser was used to obtain the data for the work presented in this thesis, and since it permits practical implementation of some of the concepts presented in the last chapters, it is intended to give here a brief description of this instrument. More details will be found in Thomas' thesis: "A real-time audio spectral analyser using active filters with adjustable parameters"; unless otherwise indicated, however, the comments, graphs and equations found in this chapter are original and not taken from Thomas' thesis.

5.2 Filtering

This spectrum analyser has the advantage of using active filters whose parameters can be changed by varying the rate of sampling and by multiplying the input wave by a weighing function synchronized to the sampling wave.

The input response of the lossless bandpass active filter is of the form $K \cdot \cos \omega_0 t$ where ω_0 is the

resonance frequency of the filter. The output of this filter will be sampled at a given rate: the filter can be built in such a way as to become infinitely lossy when the sampling pulse is applied. In this case, the frequency response of the filter will be given by the following equations:

$$F(j\omega) = K \int_{-T/2}^{+T/2} e^{-j\omega t} \cos(\omega_0 t) dt \quad \dots(5.01)$$

$$= K \frac{T}{2} \left[\frac{\sin(\omega_0 - \omega) \frac{T}{2}}{(\omega_0 - \omega) \frac{T}{2}} + \frac{\sin(\omega_0 + \omega) \frac{T}{2}}{(\omega_0 + \omega) \frac{T}{2}} \right] \quad \dots(5.02)$$

which, if ω is in the neighbourhood of ω_0 and if ω and ω_0 are both large, becomes approximately:

$$K \frac{T}{2} \left[\frac{\sin(\omega_0 - \omega) \frac{T}{2}}{(\omega_0 - \omega) \frac{T}{2}} \right] \quad \dots(5.03)$$

The frequency response of such a filter has the form of a $\sin(x)/x$, and if the period T of the sampling function becomes, for example, twice as long, the bandwidth becomes twice as narrow. These concepts are illustrated in the diagrams of figure 5.01.

But if the impulse response of this filter is multiplied by a raised cosine

$$\left[0.5 \left(1 + \cos \frac{2\pi t}{T} \right) \right] \quad \dots(5.04)$$

synchronized to the sampling pulse as in figure 5.02, the frequency characteristic of the filter is described by the following equations:

$$F(j\omega) = \int_{-T/2}^{+T/2} e^{-j\omega t} K(0.5) [\cos(\omega_0 t)] [1 + \cos \frac{2\pi t}{T}] dt \quad \dots(5.05)$$

$$\begin{aligned} F(j\omega) = & K \frac{T}{4} \left[\frac{\sin(\omega_0 - \omega) \frac{T}{2}}{(\omega_0 - \omega) \frac{T}{2}} + \frac{\sin(\omega + \omega_0) \frac{T}{2}}{(\omega + \omega_0) \frac{T}{2}} \right] \\ & + K \frac{T}{8} \left[\frac{\sin(\omega - \omega_0 + \frac{\pi}{T}) \frac{T}{2}}{(\omega - \omega_0 + \frac{\pi}{T}) \frac{T}{2}} + \frac{\sin(\omega - \omega_0 - \frac{\pi}{T}) \frac{T}{2}}{(\omega - \omega_0 - \frac{\pi}{T}) \frac{T}{2}} \right. \\ & \left. + \frac{\sin(\omega + \omega_0 - \frac{\pi}{T}) \frac{T}{2}}{(\omega + \omega_0 - \frac{\pi}{T}) \frac{T}{2}} + \frac{\sin(\omega + \omega_0 + \frac{\pi}{T}) \frac{T}{2}}{(\omega + \omega_0 + \frac{\pi}{T}) \frac{T}{2}} \right] \quad \dots(5.06) \end{aligned}$$

Under the same conditions that ω and ω_0 be both large and nearly equal, equation 5.06 becomes approximately

$$\begin{aligned} F(j\omega) \approx & K \frac{T}{4} \left[\frac{\sin(\omega_0 - \omega) \frac{T}{2}}{(\omega_0 - \omega) \frac{T}{2}} \right] \\ & + K \frac{T}{8} \left[\frac{\sin(\omega - \omega_0 + \frac{\pi}{T}) \frac{T}{2}}{(\omega - \omega_0 + \frac{\pi}{T}) \frac{T}{2}} + \frac{\sin(\omega - \omega_0 - \frac{\pi}{T}) \frac{T}{2}}{(\omega - \omega_0 - \frac{\pi}{T}) \frac{T}{2}} \right] \quad \dots(5.07) \end{aligned}$$

which represents the addition of three $\sin(x)/x$ such as illustrated in figure 5.03. The result is an ~~very~~ ^{Thomas} ~~close~~ approximation to the gaussian filter (~~figure 5.04~~), a fact which is very welcomed, due to the very low side lobes of this approximately gaussian filter and to what has been written in the previous chapter about the elementary areas and the uncertainty relation.

It is shown in Thomas' thesis that to multiply the input function to the analyser by a raised cosine synchronized to the quenching pulse is equivalent to multiplying the impulse response of each filter by the same raised cosine. Figure 5.04 permits to see how near the filters of the spectrum analyser come to ~~having as time and frequency windows gaussian-type approximating the time and frequency windows of the characteristics.~~
~~gaussian filter.~~

5.3 General description

A block diagram of the analyser taken from Thomas' thesis is given in figure 5.05. It can be seen that the speech wave is multiplied in one case by the weighting function from the "external function generator" and in the other case by the 180° out of phase version of the weighting function. Each resulting signal is separately modulated in amplitude by a carrier frequency of around 15 kc/s and each modulated wave is sent to a different bank of filters.

From the spectrum analysis point of view, one has time and frequency characteristics of the type described in figure 5.04. Each of the two banks has 32 filters; their resonance frequencies are normally 100 c/s apart, and in the normal mode of analysis, when the weighting function is a 50 c/s sine wave, the frequency characteristic of the filter whose center is at ω_0 falls to zero at $\omega_0 + 100$ c/s and to about half the value of the peak amplitude at $\omega_0 + 50$ c/s, which gives a flat frequency response over the spectrum covered. All the filters in each bank are sampled simultaneously; each bank is then sampled 50 times a second, which gives a time interval of 10 msec between samples.

In the remainder of this work, the frequency band separating the points having half the amplitude of the peak of the resonance curve (6 db points) will be defined as the bandwidth of the filters of the analyser.

If one uses as weighting function a 100 c/s sine wave, the bandwidth of each filter is twice as large as in the previous case, and the sampling moments occur twice as often, at intervals of 5 msec.

The output of the analyser can be observed on a CRT screen as shown by the photographs of chapter 6. The vertical bars denote the sampling moments, the

frequency axis is in the vertical direction, the time axis in the horizontal direction and the energy in different parts of the spectrum is indicated by intensity modulation and some horizontal deflection of the electron beam on the horizontal axis.

Another set of modulators with appropriate modulating frequencies is provided so that one can choose at will the speech band up to 3.2 kc/s or the speech band going from 3.2 to 6.4 kc/s.

5.4 Redundancy in the frequency domain

The center frequencies of the filters are adjusted at 100 c/s, 200 c/s and so on up to 3.2 kc/s. If the filters have a bandwidth smaller than 100 c/s, there is not a complete coverage of the frequency spectrum. If the bandwidth of the filters is larger than 100 c/s, some of the filters should be taken out of operation and the center frequency of the others readjusted in order to maintain the appropriate rate of sampling. In the case of a bandwidth of 200 c/s, half of the filters should be taken out of operation.

When the spectrograms are intended for visual presentation, some amount of redundancy might be preferable such as in the "Kay sonograph". Since we are interested to compare the different sampling methods, it was resolved to use non-redundant representation.

The spectrum analyser had to be modified, so as to allow us to take half the filters out of operation when the 200 c/s mode was adopted: in this last case, the filters whose center frequency was tuned (effectively) to 200 c/s, 400 c/s and so on up to 3.2 kc were kept and those tuned to 100 c/s, 300 c/s were taken out of operation.

5.5 Conclusion

The Imperial College spectrum analyser is then a practical implementation of some of the sampling concepts defined by Gabor. Being, as all spectrum analysers, insensitive to phase, it loses half the information and has thus in the non-redundant mode a sampling rate of FT instead of 2FT. The Imperial College spectrum analyser is then equivalent to the last but one step in Gabor's analysis of elementary signals and corresponds to figure 3.01 b - except for the fact that the phase information is missing.

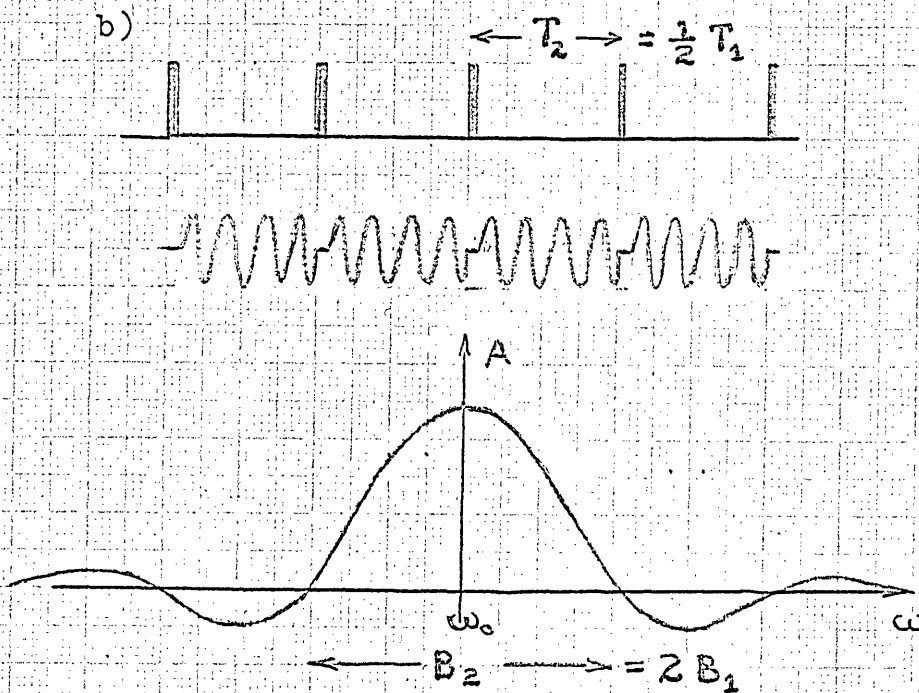
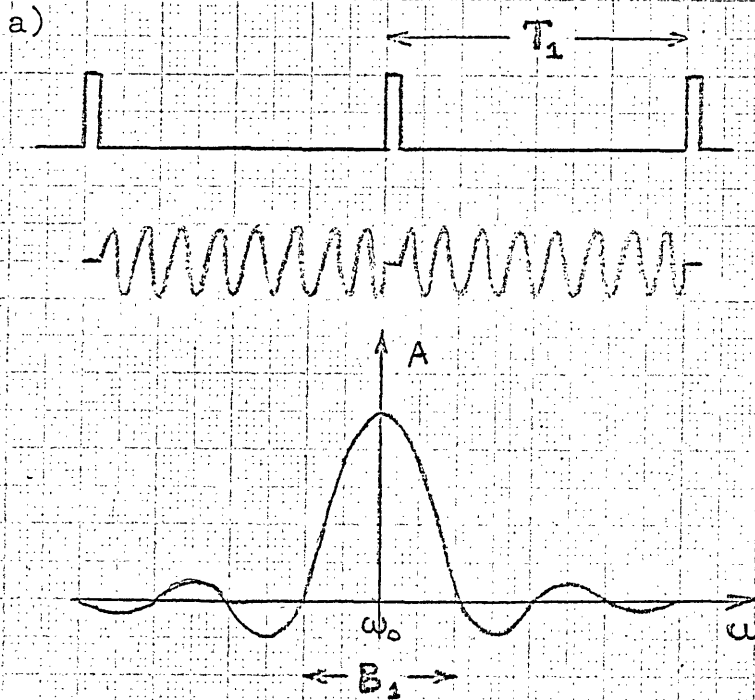
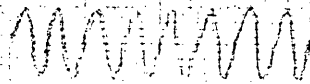
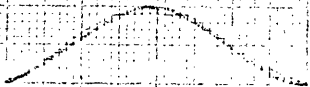


Figure 5.01

Relation between the sampling rate, the impulse response and the frequency characteristic of lossless active filters.

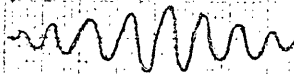


Multiplying the impulse response
of a lossless active filter



by a weighting function of the form

$$\frac{1}{2}(1 + \cos(2\pi t/T))$$



results in a modified
impulse response

which is the sum of three cosine terms:

$$K/4 \cos(\omega_0 - \pi/T)t$$

$$K/2 \cos(\omega_0)t$$

$$K/4 \cos(\omega_0 + \pi/T)t$$

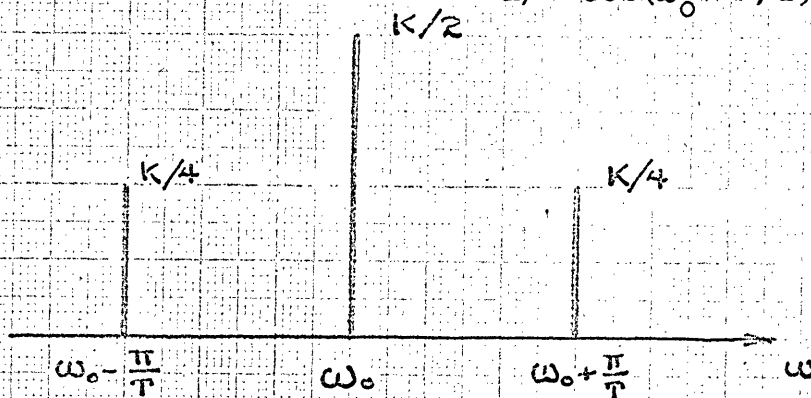


Figure 5.02

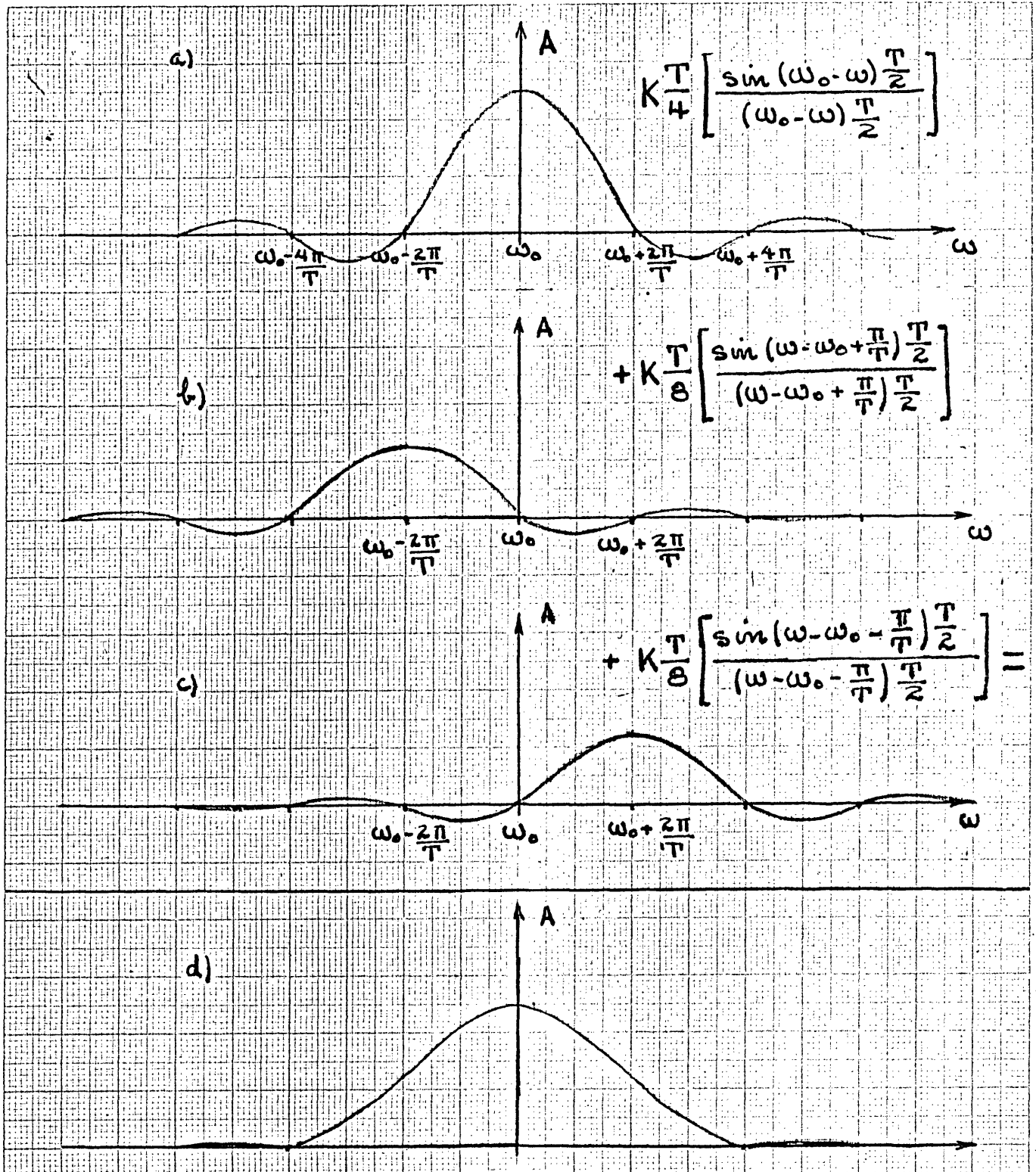


Figure 5.03

The frequency characteristic of a lossless active filter whose impulse response is weighted by a raised cosine is the sum of three orthogonal $\sin(x)/x$. ($a+b+c = d$)

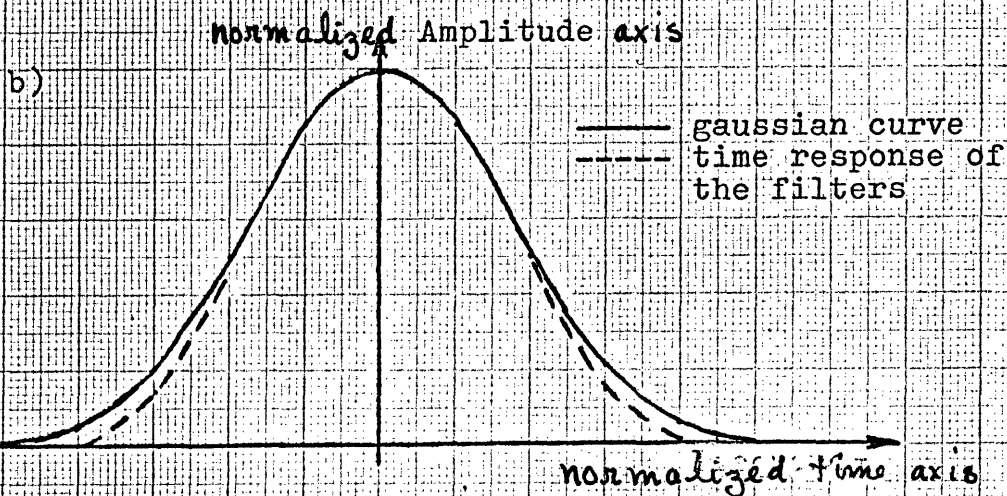
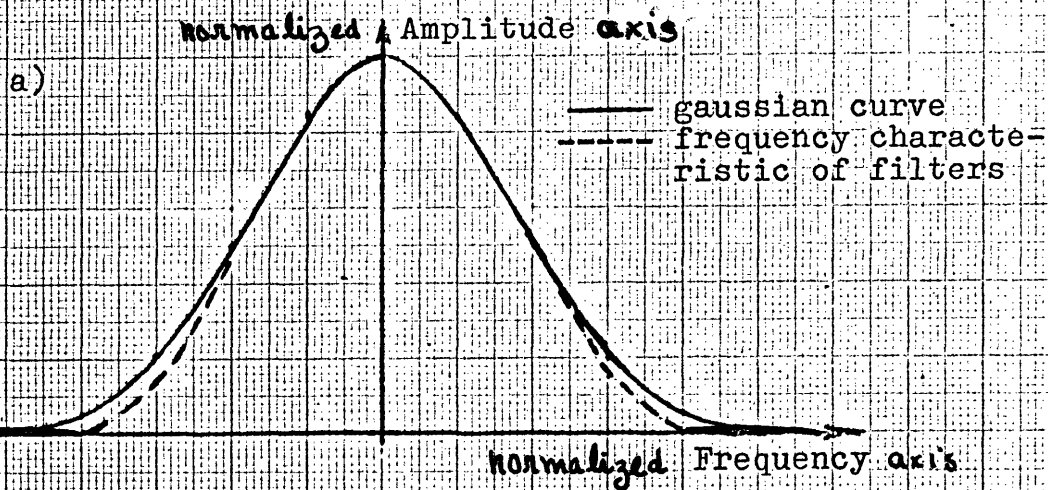
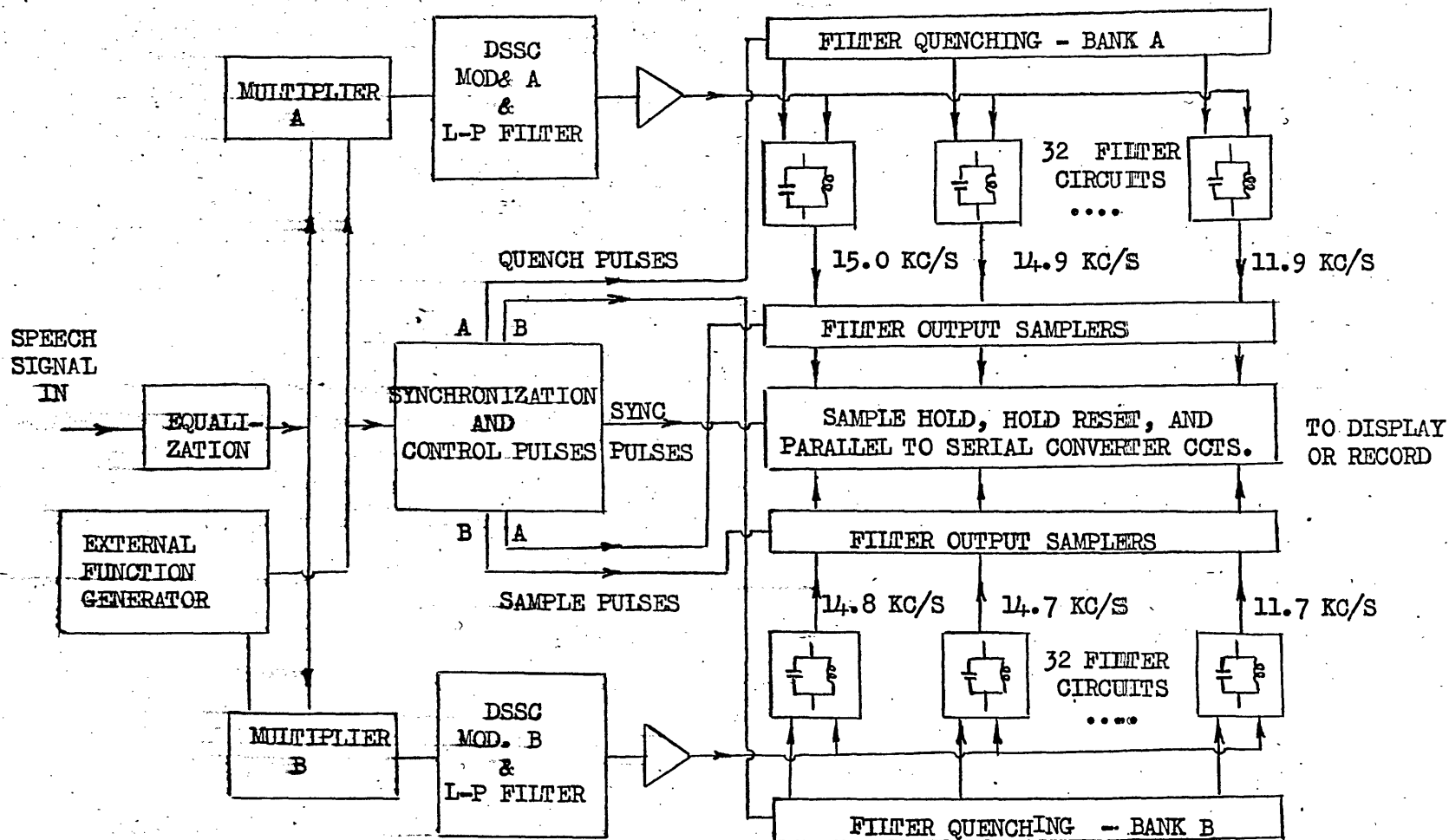


Figure 5.04

Approximation of a gaussian time and frequency response by the filters of the I.C. spectrum analyser
a) frequency characteristic of the filters
b) time response of the filters

N.B. : These curves are normalized along the amplitude axes with respect to the points of maximum amplitude and along the time and frequency axes with respect to the 6 dB points.



Block diagram of the Imperial College Spectrum analyser (from R.S. Thomas)
Fig. 5.05

Chapter 6

IMPLEMENTING ADAPTIVE FILTERING

6.1 Introduction

We intend to examine in this chapter a proposal for implementing adaptive filtering. An instrument is demonstrated which presents to the analyser the appropriate weighting function and permits it to operate in two modes and to change from one to the other according to an on-off control signal.

The bandwidths chosen for implementation were 100 c/s and 200 c/s. In practice, an adaptive speech spectrum analyser should perhaps be able to switch between 50 c/s, 100 c/s and 200 c/s or between 50 c/s and 200 c/s. However, the Imperial College analyser as presently designed, does not operate in real time with filters adjusted to a bandwidth narrower than 100 c/s.

6.2 The function generator (FG)

The role of the "Function Generator" (FG) is to present to the analyser a weighting function in the form of a raised cosine switched back and forth between frequency values of 50 c/s and 100 c/s, according to some control signal.

The FG consists of three parts:

- one which generates a 100 c/s and a 50 c/s sine wave in phase with each other and synchronized to a 50 c/s square wave;
- a second part which handles the speech wave;
- a third part which reads the output of a predetector and gives at the appropriate time the signal to switch.

6.2.1 Synchronization

In order to generate two synchronized sine waves and one square wave, the following method is used (numbers in brackets refer to figure 6.02): a 100 c/s square wave is passed through a low-pass filter (1) with cut-off frequency at 150 c/s. The same 100 c/s square wave is used to trigger a monostable (2) whose negative edge triggers in turn a binary counter (4), the output of which is a 50 c/s sine wave which is passed through a low-pass filter with a 100 c/s cut-off frequency: the resulting 50 c/s sine wave can be synchronized to the 100 c/s sine wave by adjusting the monostable. In parallel to this, the 100 c/s square wave is used to trigger another monostable (3) which provides the shifting function for a shift register stage (5) connected to the binary counter (4); once again, it is possible, by adjusting the monostable, to synchronize the 50 c/s square wave to the two sine waves as illustrated in figure 6.01 a, b, and c.

It is now possible with the proper gating system

to switch from one sine wave to the other. It will be noted, however, from figure 6.01, where the sampling points for each bank of filters are indicated, that the weighting function for bank B is always a raised cosine either of 20 msec or of 10 msec period. In the case of the bank of filters A, however, the weighting function at the switching instants is composed of a portion of a 50 c/s sine wave and a portion of a 100 c/s sine wave.

This cannot be avoided.

6.2.2 The speech wave

Filters with a 200 c/s bandwidth have an integration time twice as short as those with a 100 c/s bandwidth; hence, if one presents at their input a pure tone at their resonance frequency, the output of the 100 c/s filters when read will be twice as large as the one of the 200 c/s filters. In order to make the best possible use of the dynamic range of the analyser, it was resolved to double the amplitude of the speech wave at the analyser's input when the 200 c/s bandwidth was used. A further reason for this is that, for visible displays, the peak output amplitude of the formants would look about the same. A simple resistive circuit with proper gating was used to process the speech wave (17, 18, 14, 16). *fig. 6.02*

6.2.3 Control

A third part contains the various gates and switching functions which will read the segmenter as well as process the weighting function and the speech wave. The FG normally samples the segmenter at a very high rate through a shift resistor stage (7). When this shift register changes its state, a gate (8) causes the sampling function to change to the 50 c/s square wave so that the shift register stage remains in that state until the analyser is in a position to switch. With the present FG, this means a maximum delay of 20 msec, because of the period of the weighting function of lower frequency. By proper modifications, however, it is possible to reduce this to 10 msec and if one is ready to tolerate half the time some more abrupt changes in the weighting function, one can reduce this maximum delay still further to 5 msec; this is illustrated in figure 6.01 d, e, and f.

Two shift register stages (12, 20) are connected in series after the first shift register stage (7). Their purpose is to operate the balanced linear gates for the weighting functions and the speech wave. The amplitude of the speech wave is changed at different times for bank A and B in order for the moment of switching to coincide with the quenching and sampling pulses for the bank concerned and thus avoid the risk of

generating spurious signals. The result is illustrated in figure 6.01 h, i, j; the first two represent the speech wave with its increase of amplitude as switching occurs and the last one represents the weighting function. The function generator is presently designed to switch for minimum periods of 40 msec; it would be easy, however, to change this for a minimum period of 20 msec.

6.3 The segmenter or pre-detector

In order to demonstrate the principle of adaptive filtering, a speech segmenter was used which is fundamentally the same as the one demonstrated by Thomas in his thesis; it is similar in principle to one suggested by Liljencrants. A block diagram of the modified instrument is given in figure 6.03. A Schmitt's trigger was added to Thomas' design in order to insure that the analyser would be in the 200 c/s mode when there was no signal or when the sound level was low, the reason being that the low energy components are almost certainly sounds which one would prefer to analyse in the 200c/s mode (^{unweighted}~~implied~~ stop-consonants, for example). Secondly, an alternative path was provided with a half-wave rectifier instead of full-wave rectifiers: it was thought desirable to detect only positive going slopes (i.e. the onset of a speech sound) since it often happens that the onset of a sound is followed by a per-

iod of sharp decrease in energy and that this occasionally causes the detector to trigger for too long periods at the start of a sound; this could perhaps be avoided if the analyser could be switched to any mode for periods shorter than 40 msec.

In its present form, the detector can be adjusted to offer some discrimination between sounds which start and finish abruptly and those which do not start and terminate quite so sharply. It offers then, when properly adjusted, some discrimination between stop-consonants and other sounds.

From the data available in the literature and according to our own attempts, it does not seem unreasonable to expect that sounds which start with a stop-consonant can often be distinguished from those starting with a vowel sound on the basis rate of the rise in amplitude at the onset of a sound, and this is true, in particular, of unvoiced stop-consonants. If one considers, for example, the data collected by Crandall and reported in Fletcher⁶⁷ on different speech sounds, one can see that the average duration of the start of a vowel sound (the part characterized by a sharp increase in amplitude) is 42 msec and that only one vowel has a start period shorter than 34 msec. It can also be seen in the same reference that all unvoiced stop-consonants last less than double that time (42 msec)

so that they rise, subsist and decay in the time a vowel takes to start; voiced stop-consonants last much longer but it is quite possibly because of the ^{pre-cursive} voicing, which one can always filter out.

The fricatives would not be hard to distinguish because of their peculiar frequency characteristics, but there are still the semi-vowels and other intermediate sounds which must be taken into account.

There is little point, however, in discussing segmentation before one knows exactly how to segmentate, and whether adaptive filtering is really of any use.

6.4 Spectrograms

A few photographs of spectrographs will help to illustrate the differences between displays and to judge whether adaptive filtering could be really worthwhile.

6.4.1 Apparatus used

The following photographs were obtained by using the spectrum analyser developed by Thomas coupled to a display oscilloscope. The bandwidth of the analysing filters could be set to 100 c/s or 200 c/s or could change between these two values according to signals produced by the Function Generator and the pre-detector on reception of the speech wave.

The speech sample at the input of the spectrum

analyser had to be delayed relatively to the one presented to the pre-detector. As no electronic delay-line was available, a tape delay-line was used; its frequency characteristics were good (it transmitted tones with a precision of almost 1 c/s in 1,000 c/s) but the amplitude characteristics were rather poor; hence, the pre-detector lost some of its operating reliability and it will be noticed on some of the spectrograms that the amplitude value of the speech wave varies somewhat from picture to picture. In addition to this, because of the attenuation characteristics at low frequencies of the tape delay-line, voicing usually fails to appear.

The sample of speech to be analysed was recorded on a tape loop; the output was fed to two channels of the tape delay-line (A and B), one of which (B) was delayed more than the other; the output of channel A was used to trigger the pre-detector and the Function Generator while the output of channel B was the speech input to be analysed (see figure 6.04).

6.4.2 Photographs of spectrograms and comments

As was explained earlier, the vertical axis on the photographs is the frequency axis and goes ^{from} around 50 c/s at the bottom of the picture to 3.25 kc/s at the top. The spectrograms are to be read from the left to the right (time axis): the vertical bars denote the

sampling instants which are 5 msec apart in the 200 c/s mode and 10 msec apart in the 100 c/s mode. It should be noted that the frequency axis does not start at the very bottom of the picture but about 0.5 cm higher. The amplitude response is indicated by some deflection towards the right as well as by intensity modulation.

Figure 6.05 gives an idea of the type of spectrograms one obtains for continuous speech when the analyser operates in the adaptive mode; in the absence of any gain averaging device, a fairly constant speech amplitude must be maintained.

The differences between displays can be appreciated at a glance in the photographs of figure 6.06, which are spectrograms of the sentence: "What did you think about?" Figure 6.06a is taken with 32 filters in the 100 c/s mode, figure 6.06b with 16 filters in the 200 c/s mode. It can be observed that the small details are more apparent but that the formant pattern is considerably confused in figure 6.06b, while figure 6.06c appears to *have many advantages of its own.* ~~to share some of the advantages and disadvantages of the two other displays.~~

The photographs are illustrations of spectrograms in the non-redundant 100 c/s and 200 c/s modes and in the adaptive mode. Some of the pictures taken in the adaptive mode (figures 6.07 and 6.08) illustrate the case where the 100 c/s mode is the normal one and the 200 c/s mode is adopted only at rather abrupt amp-

litude variations of the speech wave while the others illustrate the case where the normal operating mode is the 200 c/s display.

A commentary on some of the spectrograms follows.

Figure 6.06: what did you talk about?

What: the "t" is ^{unreleased}~~imploded~~ (not pronounced); the shape of the formant pattern in the 100 c/s display would permit to infer the presence of "t" but not in the 200 c/s mode;

did: the burst of high frequency noise marking the initial "d" is more evident in the 200 c/s mode; the formant pattern in the 100 c/s would permit more easily to suspect the final "d";

d'you: the initial "d" is less susceptible to be confused with the start of the vowel in the 200 c/s display because it appears to originate at a higher frequency than the third formant of "i". The formant structure and the transition "i:" to "u:" is much clearer in the 100 c/s display;

tal(k): although there is some amplitude variation between the displays due to the delay-line, it is evident that the two lower formants are better defined in the 100 c/s display;

k'a: the downward slope of the second formant hardly appears at all in the 200 c/s mode;

bou(t): the presence of "b" before the vowel is more easily detectable in the 200 display; the

diphthong is however much easier to identify in the 100 c/s mode.

Figure 6.07: pa-ba (a)

In both cases, the formant structure is much better defined in the 100 c/s mode; the fine detail in the 200 c/s display appears to disrupt the whole picture. In the 200 c/s display, one can easily notice the presence of the noise followed by a period of relative emptiness in the case of "p" and one cannot fail to notice the presence of "b" before the start of the vowel: the frequency transitions from consonant to vowel are hardly recognizable, however, in the 200 c/s mode.

Figure 6.08: ta-da (a)

The consonant "t" is probably as good in either display, but the "d" is much easier to detect in the 200 c/s mode. The consonants "d" and "t" are both characterized by high frequency noise, but while there is an interval of relative silence between "t" and the subsequent vowel, the vowel following the "d" starts almost immediately. Hence, there is no clear frequency transition between the consonant "t" and the formants of the vowel while one can observe such transitions between the consonant "d" and the subsequent vowel.

Although the shape of the frequency transitions from "d" to "a" can be inferred from the 200 c/s display,

it must be noted that these transitions do not show up as clearly in the 200 c/s display as in the 100 c/s mode.

Figure 6.09: po-bo (ɔ)

Concerning the formant structure, the same remarks as precedently are valid. The presence of the two consonants is much easier to detect in the 200 c/s mode and they are easier to identify.

Figure 6.10: to-do (ɔ)

These spectrograms illustrate the case where the burst of noise of a consonant is almost absent and where the main clues are carried by the frequency transitions between consonant and vowel.

Although none of the consonants, in particular the "t", is very well defined in either display, there can be no doubt that they are almost undetectable in the 100 c/s mode, except for the formant transitions which do not appear clearly in the 200 c/s mode. There is, however, in the 200 c/s display some sign of a burst of noise.

6.4.3 Conclusion

There is little point to repeat more or less the same explanations for all the spectrograms. But for all the examples presented here, and these examples were taken at random and not selected for this purpose,

the same comments remain pertinent: the ^{non-redundant} 200 c/s display tends to destroy the formant pattern while the fine details and the bursts of noise which are usually present in stop-consonants do not show up as well in general in the 100 c/s display as in the 200 c/s mode (non-redundant). These remarks become even more pertinent when they are applied to continuous speech instead of non-sense syllables.

6.5 General points

At the light of what has been presented in this chapter, it becomes evident that to use adaptive filtering or different modes of analysis simultaneously acts as a multiplier of the number of cues. Can this be put to use?

The segmentation technique used here is, of course, purely empirical and almost arbitrary. It seems, however, to correspond well enough to the nature of speech and certainly permits to illustrate the principle of adaptive filtering.

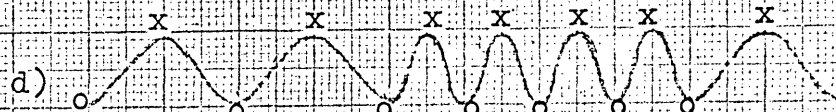
6.6 Conclusion

The preceeding discussions in this chapter, and since the beginning of this thesis, have been directed towards explaining what adaptive filtering is about and discussing its relevance. In a way, we have been going

through the same mental process the author of this work went through at first: we have been convincing ourselves that there is here a problem worth looking into and finding a justification for studying it in more detail.

This chapter was the final step in this justification process and, from now on, the thesis will be concerned with the design of the experiments and the results.

How can we assess the value of various time-frequency presentations for the study of speech sounds?



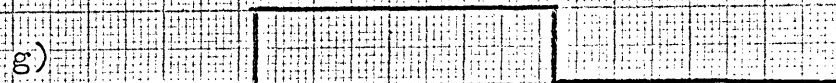
maximum delay before switching: 20 msec.



maximum delay before switching: 10 msec.



maximum delay before switching: 5 msec.



control signal.

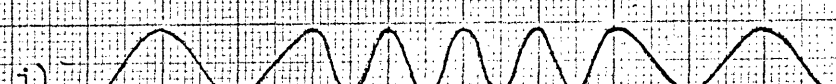


Figure 6.01

Some signals produced by the "Function Generator".

x:sampling points for bank "A".

o:sampling points for bank "B".

-----:signal inverted.

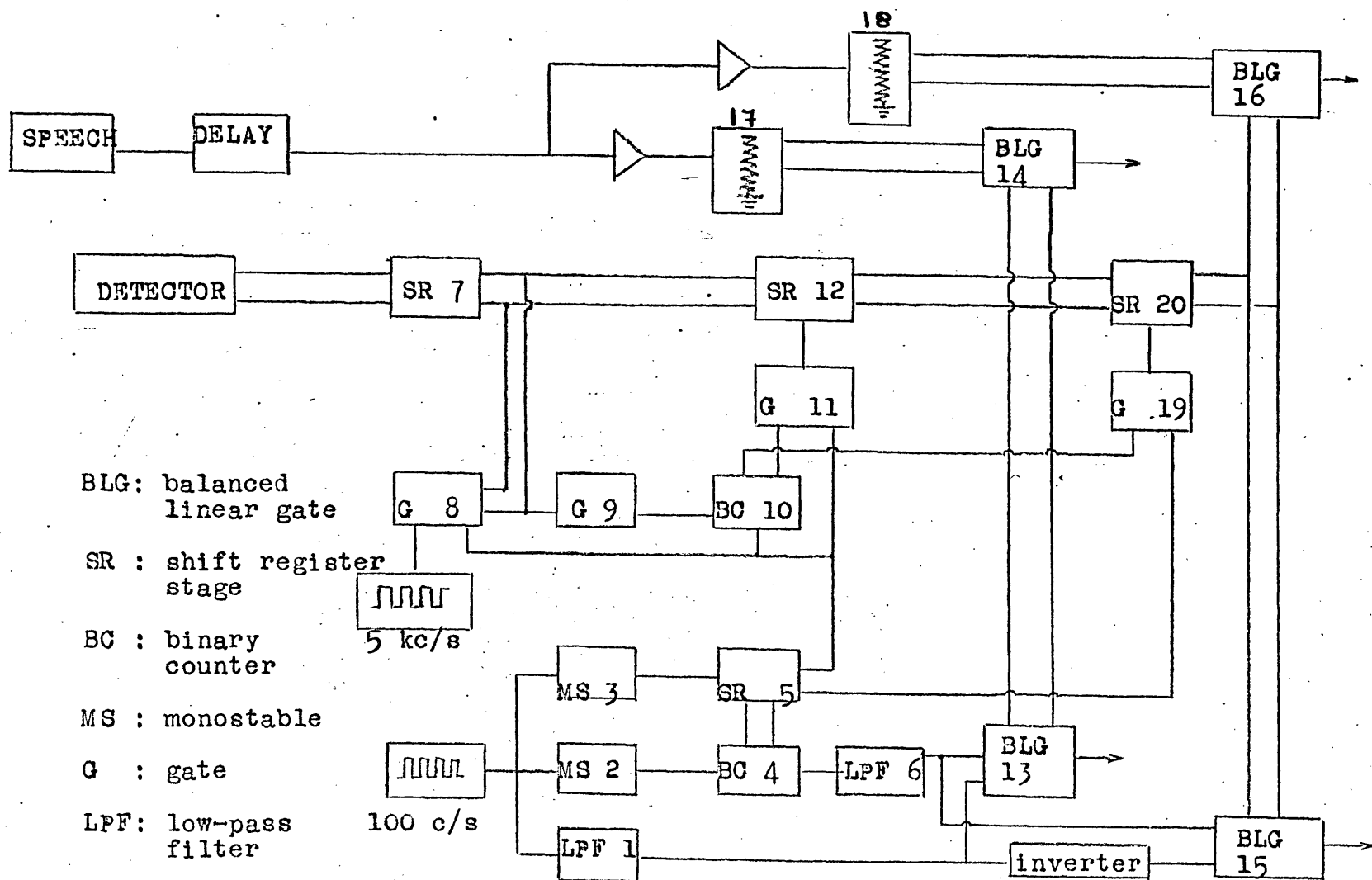


Fig. 6.02

Block diagram of the Function Generator

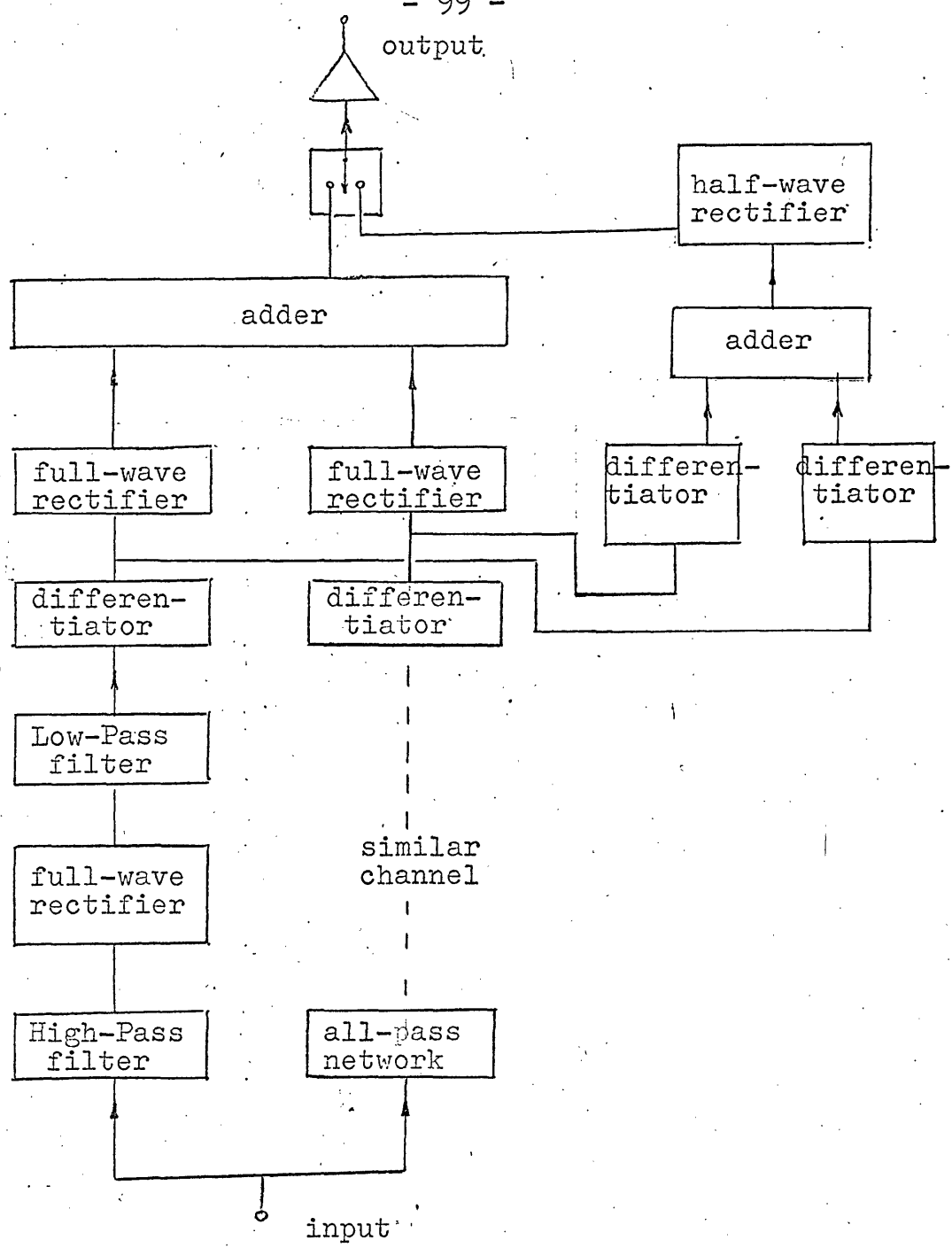
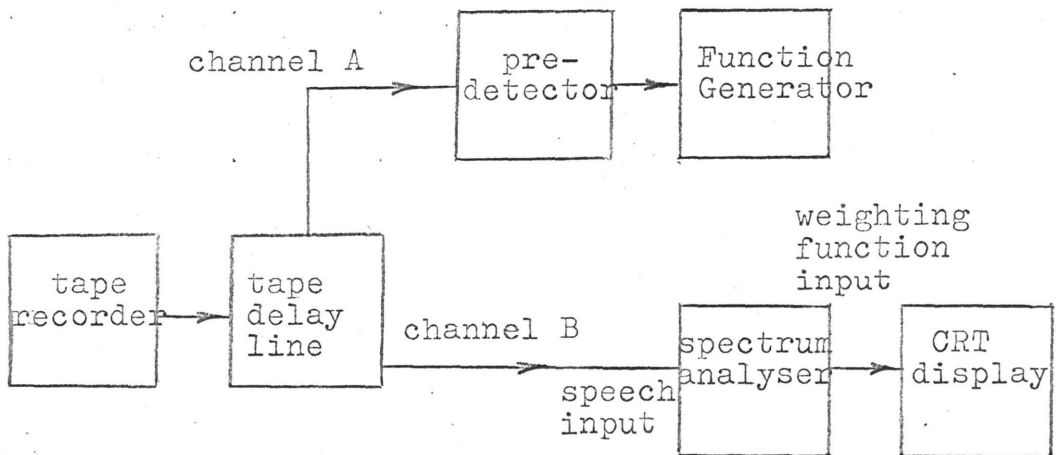


Fig. 6.03

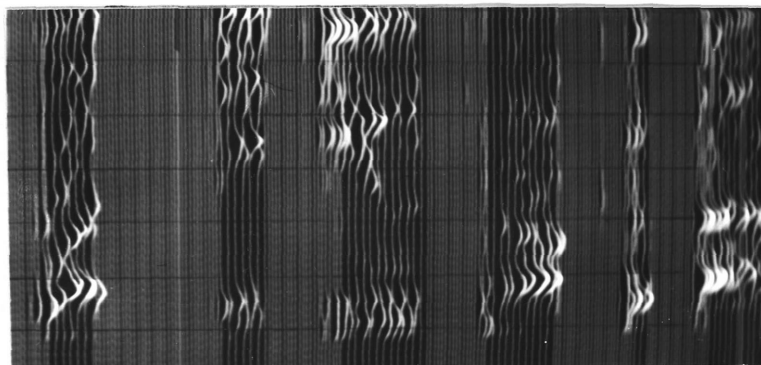
The pre-detector



Assembly for producing spectrograms
(adaptive filtering)

Fig. 6.04

what di-d' you t-a-lk a- bou(t)



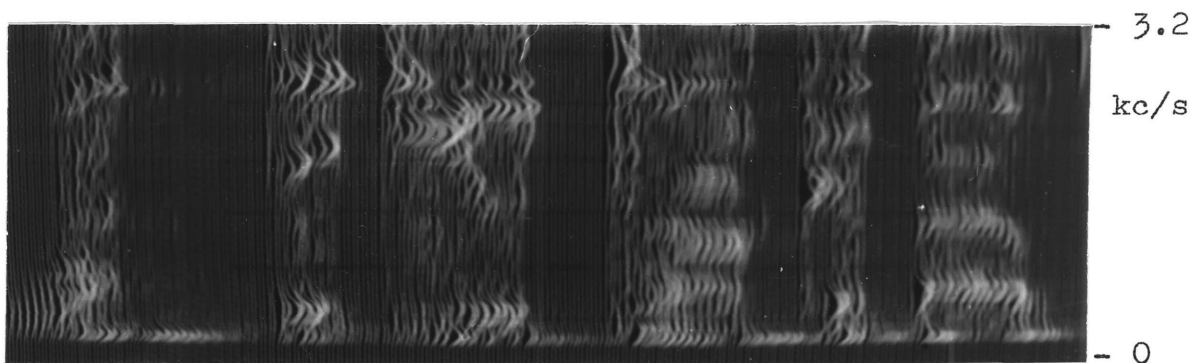
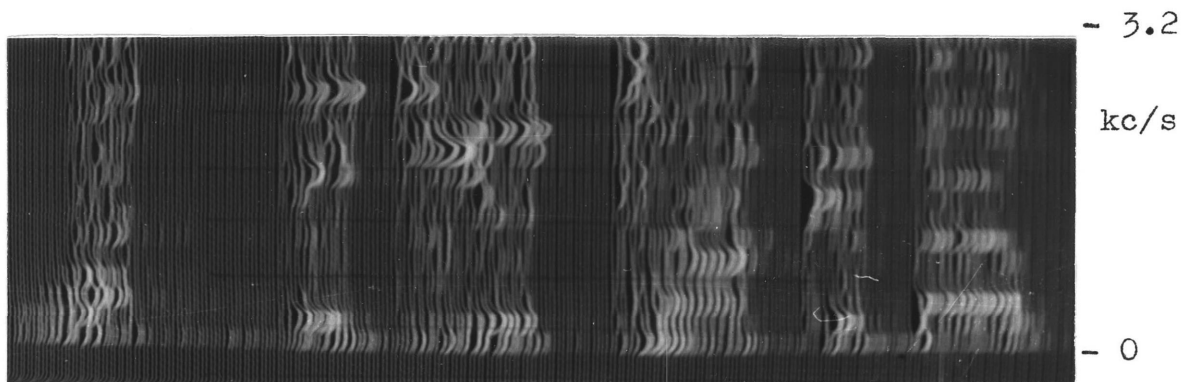
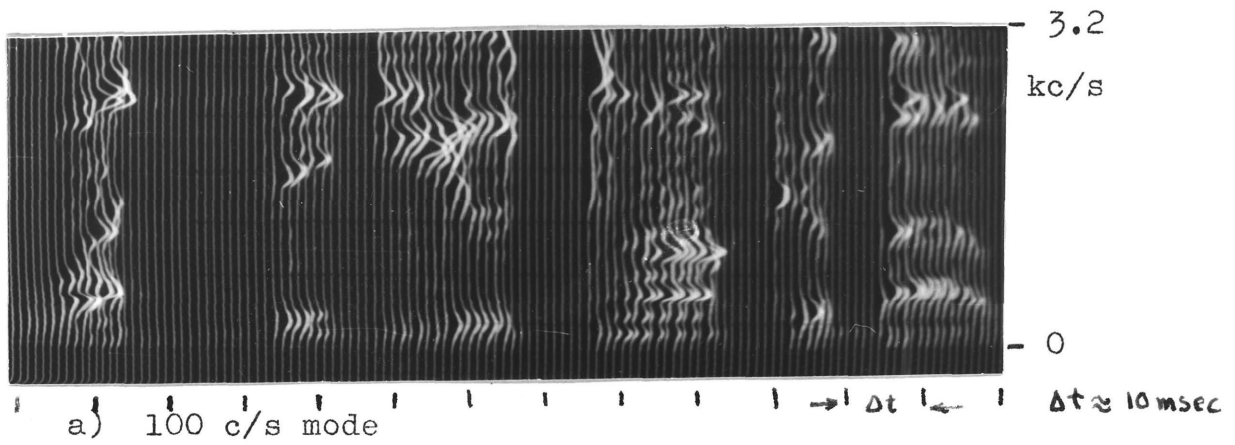
Adaptive filtering in continuous speech.

"What did you talk about?"

(100 c/s - 32 filters)/(200 c/s - 16 filters)

Fig. 6.05

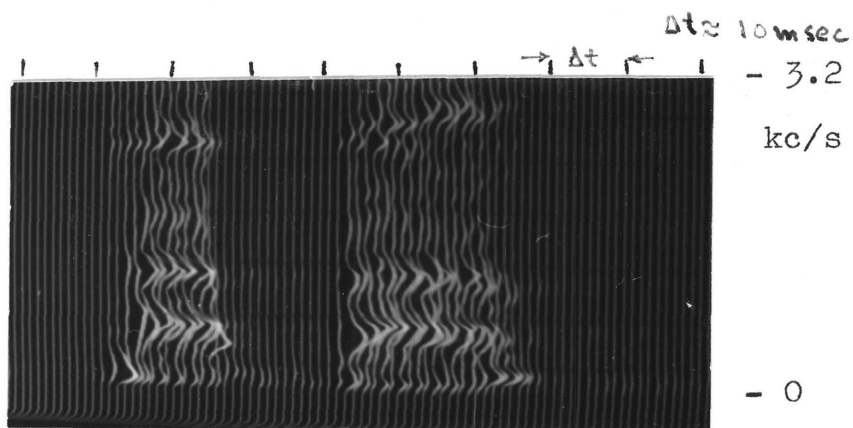
what di - d'y - ou t-a - lk a - bou (t)?



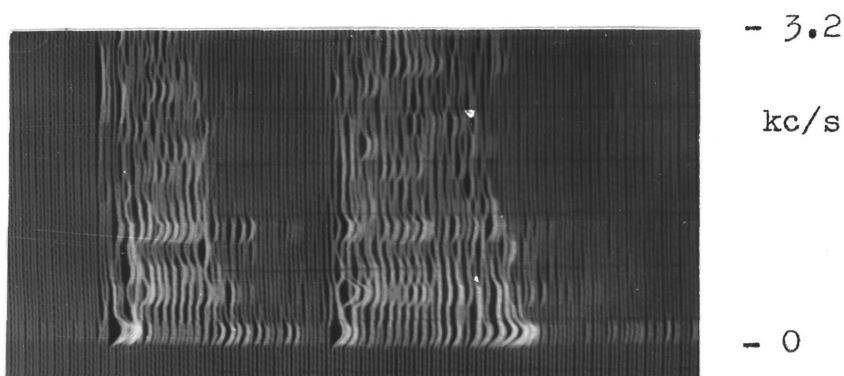
c) 200 c/s mode (32 filters)

Fig. 6.06

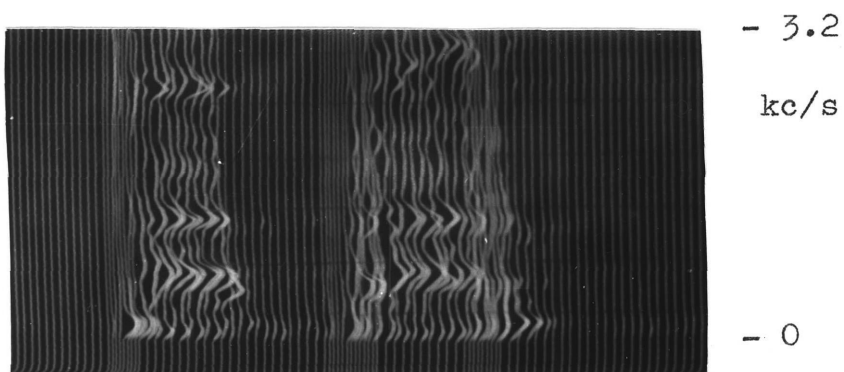
Spectrograms of "what did you talk about?"



a) 100 c/s mode



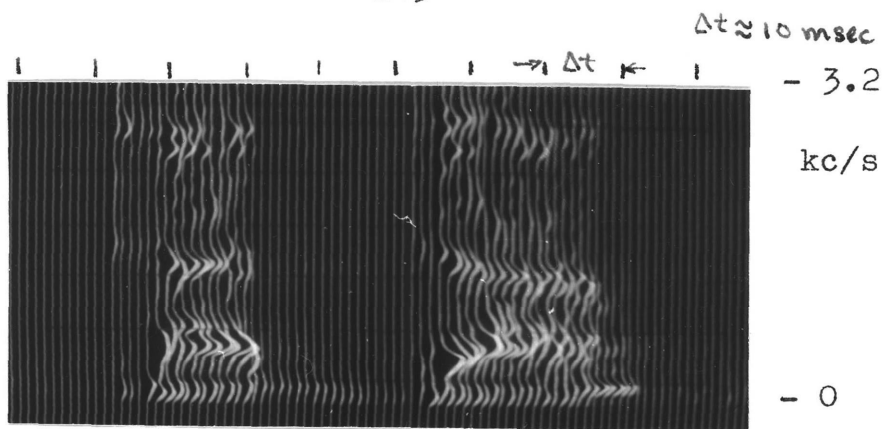
b) 200 c/s mode (16 filters)



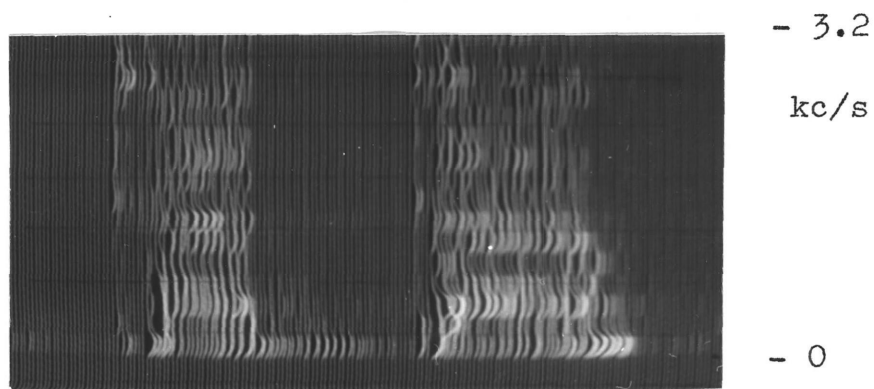
c) adaptive mode

Fig. 6.07

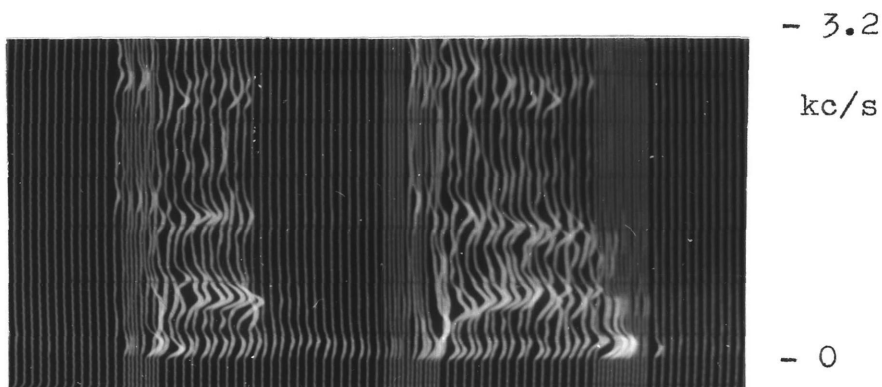
Spectrograms of "pa-ba"



a) 100 c/s mode



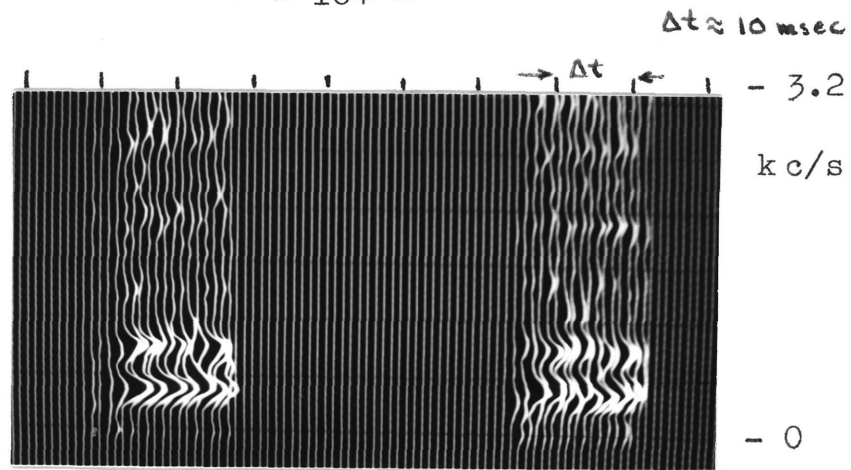
b) 200 c/s mode



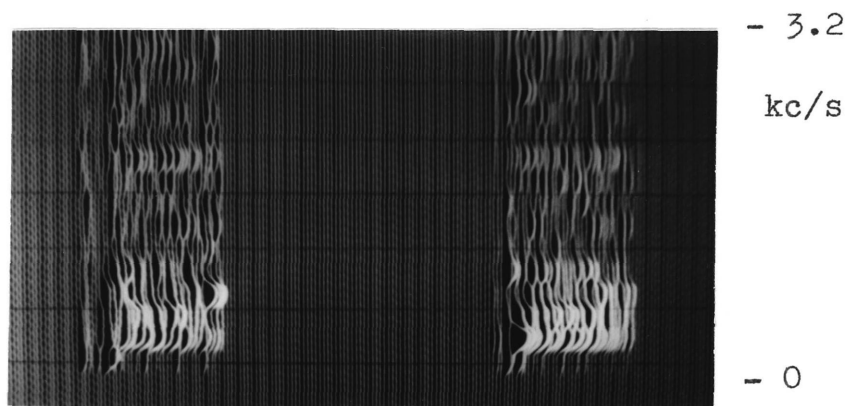
c) adaptive mode

Fig. 6.08

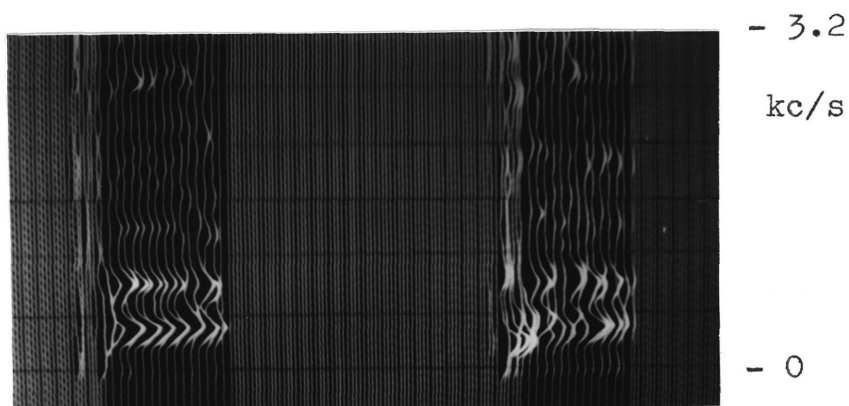
Spectrograms of "ta-da"



a) 100 c/s filter



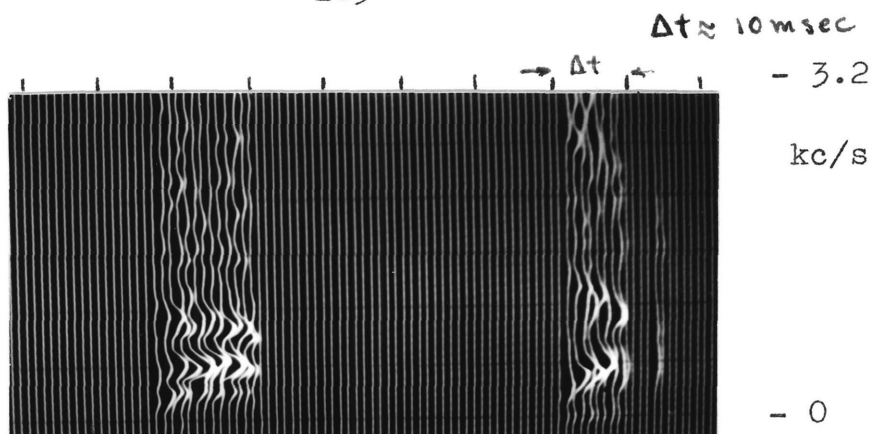
b) 200 c/s mode (16 filters)



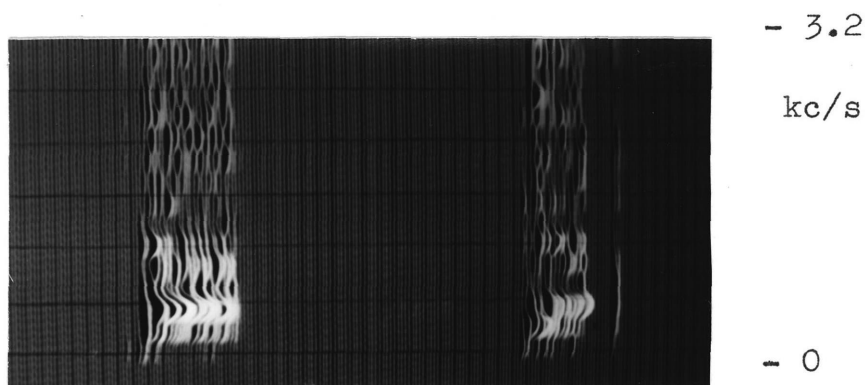
c) adaptive mode

Fig. 6.09

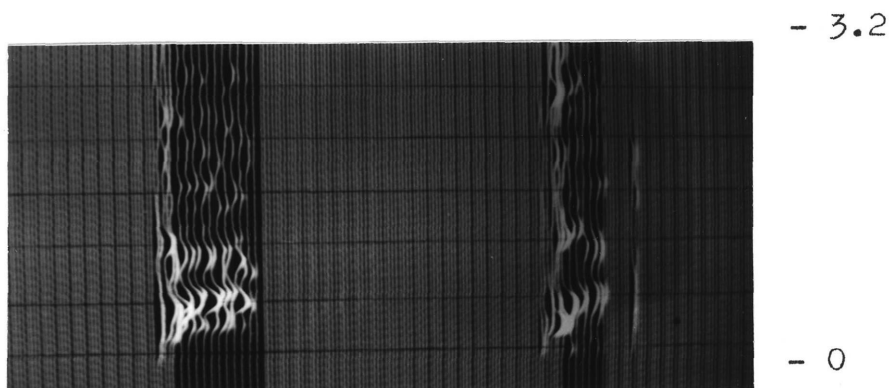
Spectrograms of "po-bo"



a) 100 c/s mode



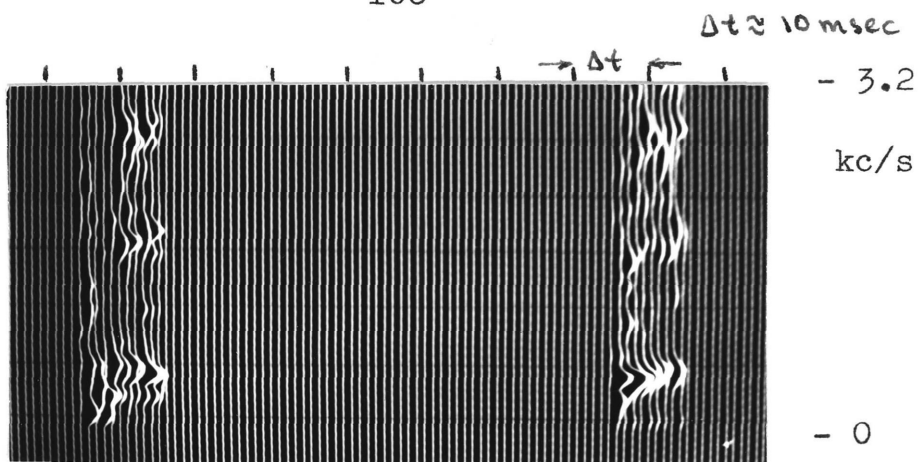
b) 200 c/s mode (16 filters)



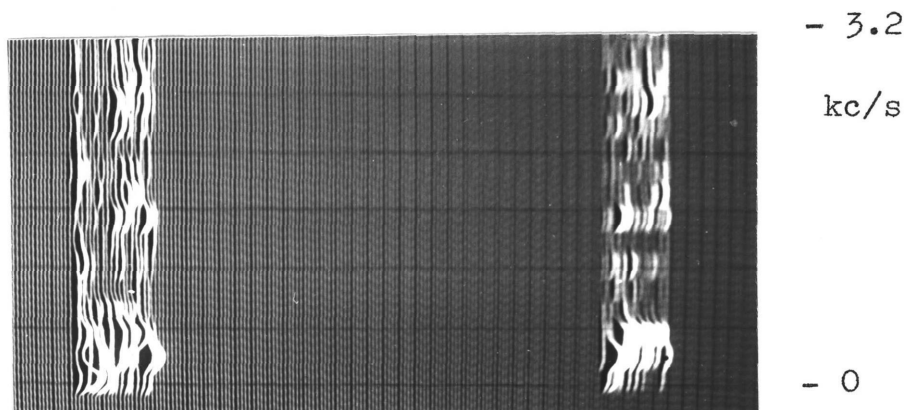
c) adaptive mode

Fig. 6.10

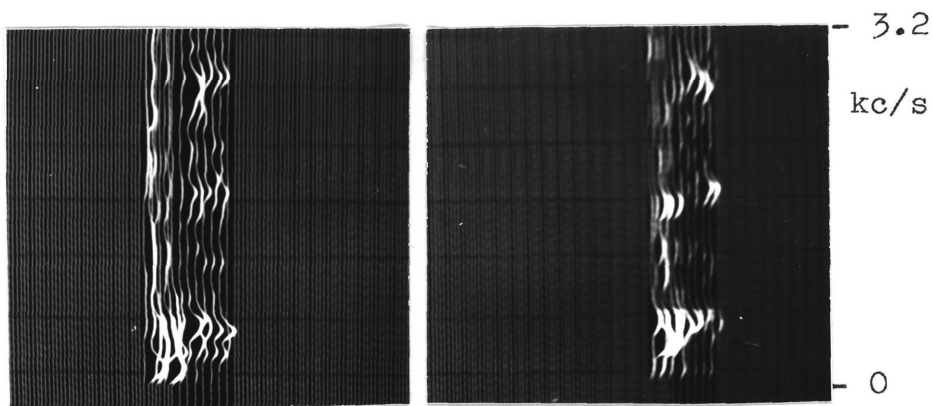
Spectrograms of "to-do"



a) 100 c/s mode



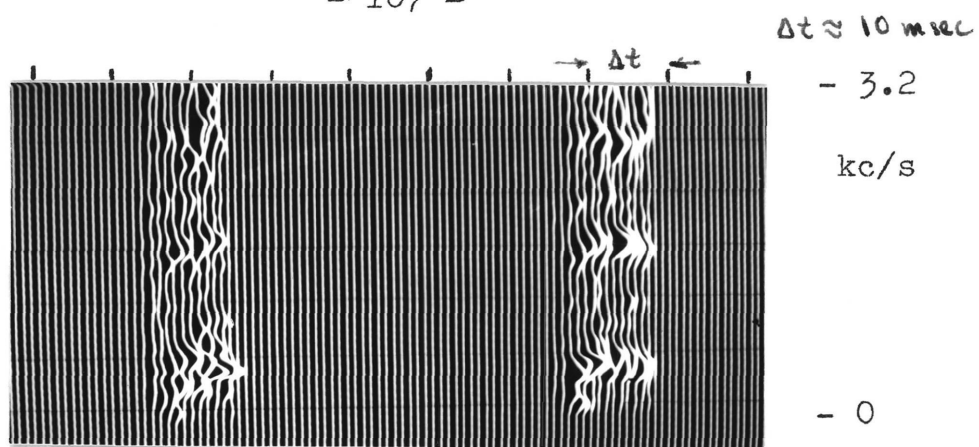
b) 200 c/s mode (16 filters)



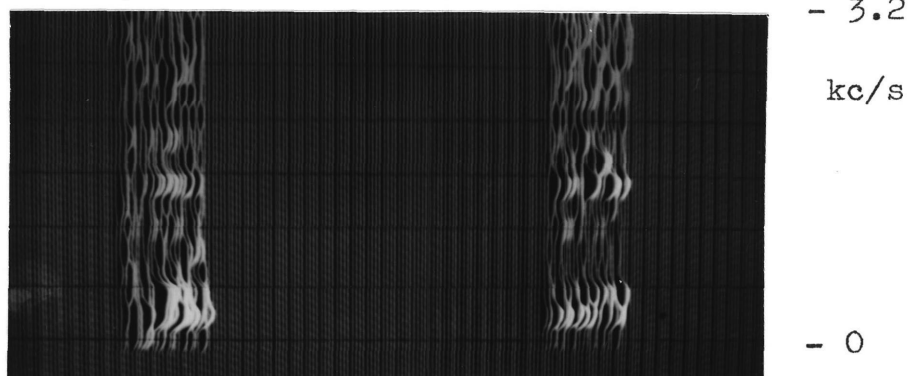
c) adaptive mode

Fig. 6.11

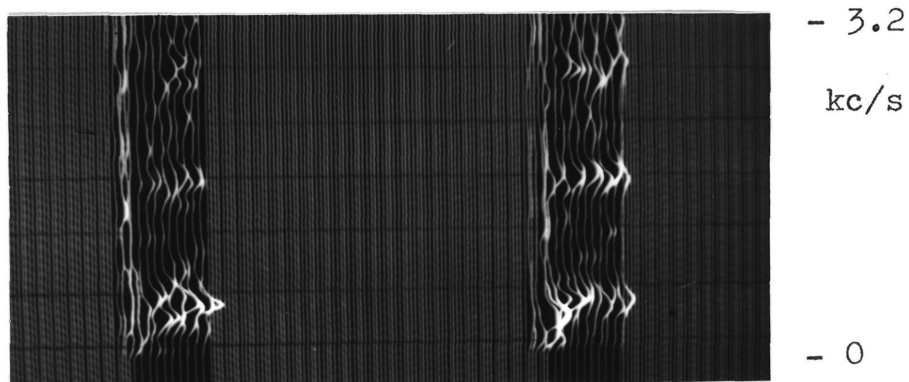
Spectrograms of "pet-bet"



a) 100 c/s mode



b) 200 c/s mode (16 filters)

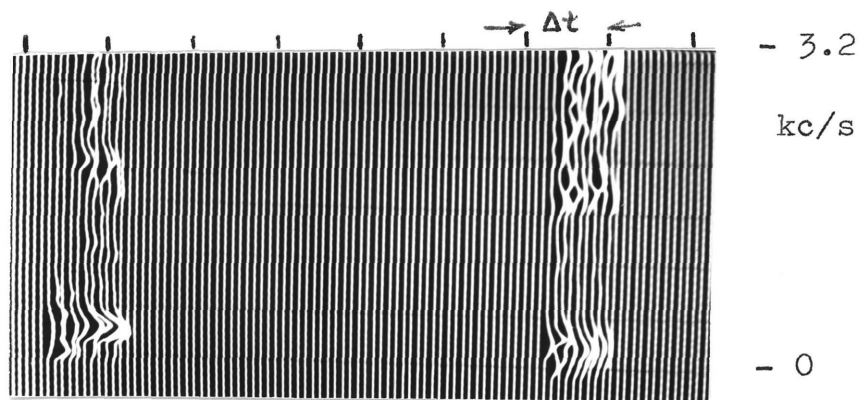


c) adaptive mode

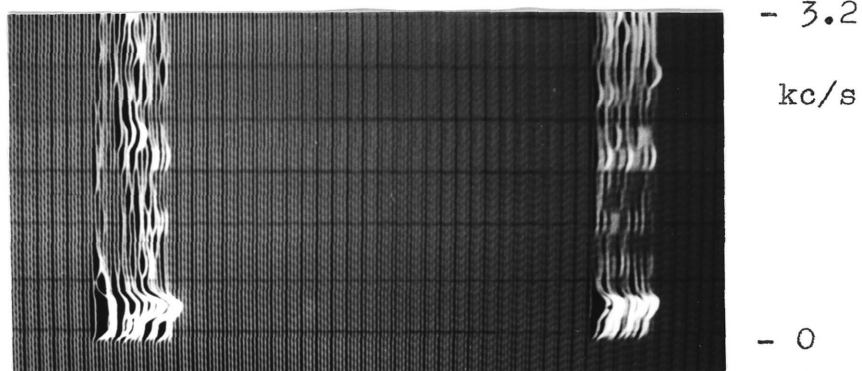
Fig. 6.12

Spectrograms of "tet-det"

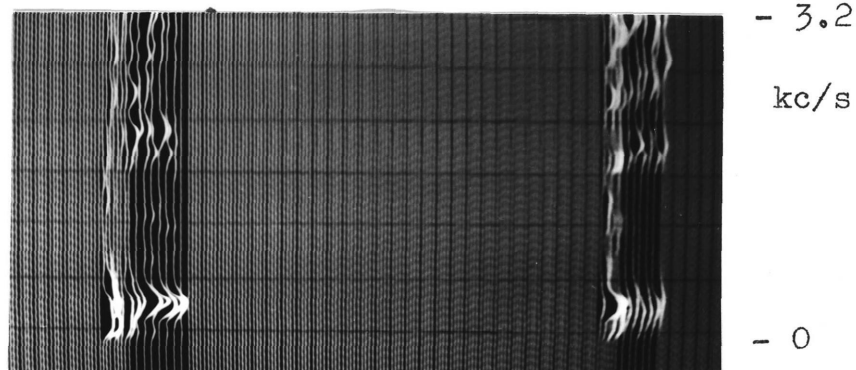
$\Delta t \approx 10 \text{ msec}$



a) 100 c/s mode



b) 200 c/s mode (16 filters)

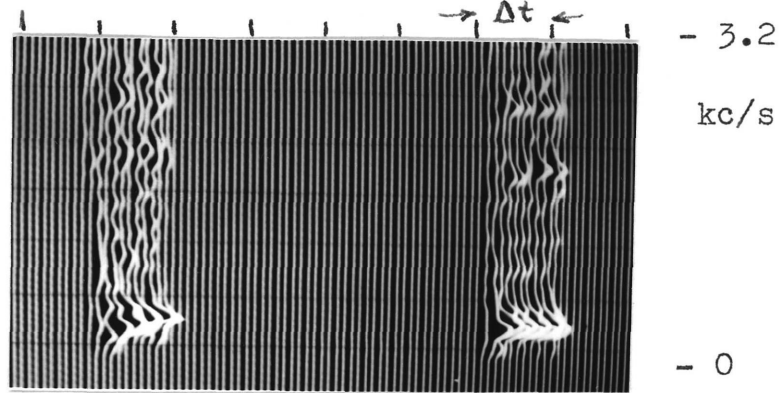


c) adaptive mode

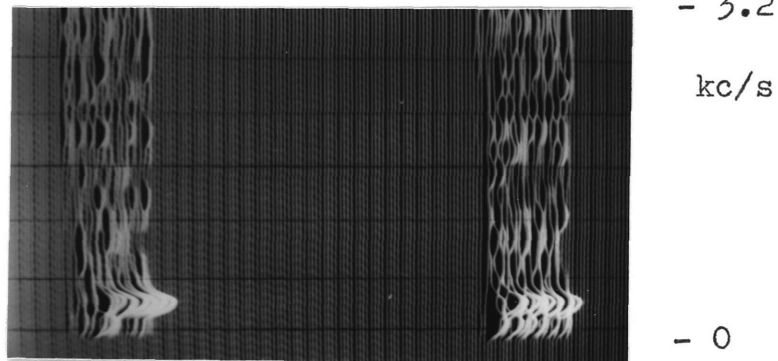
Fig. 6.13
Spectrograms of "pit-bit"

- 109 -

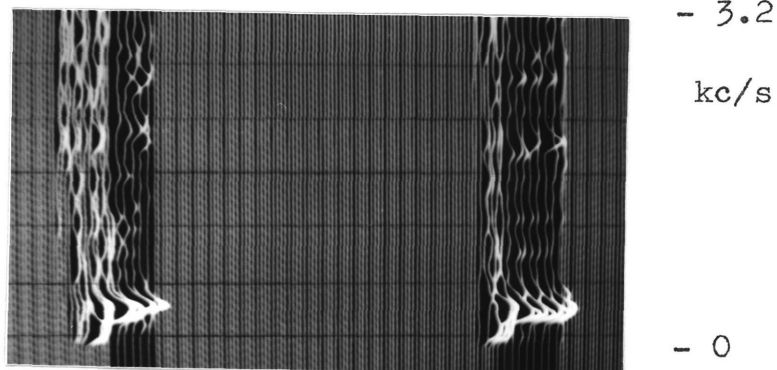
$\Delta t \approx 10 \text{ msec}$



a) 100 c/s mode



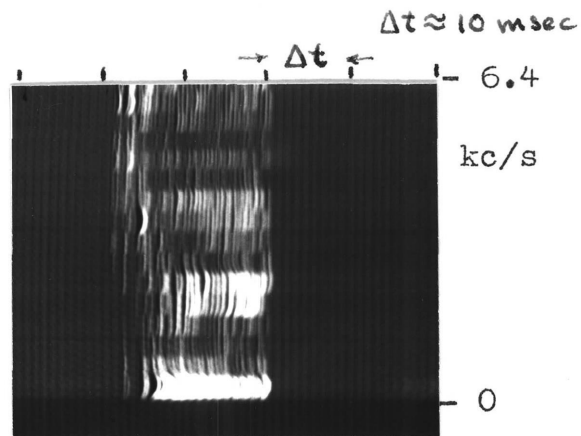
b) 200 c/s mode (16 filters)



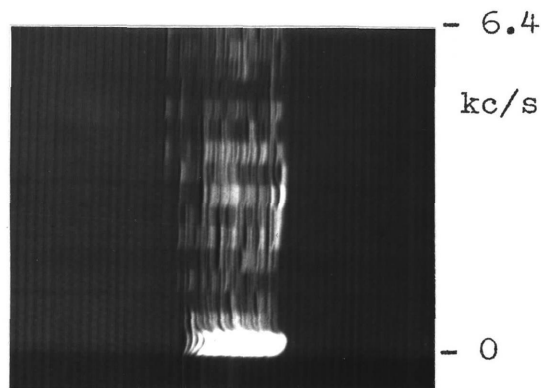
c) adaptive mode

Fig. 6.14

Spectrograms of "tit-dit"



"tit"



"dit"

Fig. 6.15

Spectrograms of "tit" and "dit"
The spectrum analyser consists of
16 filters of 400 c/s bandwidth.

Chapter 7

THE DESIGN OF THE EXPERIMENTS

7.1 The problem

How can one assess the usefulness of various time-frequency displays?

It has been shown that classes of speech sounds have different characteristics, it is suspected that the classification of speech sounds with respect to each other can be made easier if the analyser adopts the proper operating mode. It has been suggested in particular that a system where the analyser changes its mode of analysis automatically according to some control signal could be worthwhile and a demonstration of a possible system was given. The next step must be, of course, to see in which cases the analyser should be in one mode, and in which cases in the other.

Is there a way to measure reliably the differences of performance between displays?

7.2 The general lines of a possible solution

One can imagine that experiments could be done with the object of measuring how reliably the same speech sound samples could be recognized when the analyser adopts different modes of operation. If one has the

possibility of taking a great number of samples, one could take different samples of the same sounds in various modes of operation and use statistical tests of significance; but, clearly, one must also have a categorizer which will work with the same efficiency for all modes of the analyser, and whose performance will not fluctuate randomly between one experiment and the other.

Since it is the differences of performances from display to display which are mainly of interest, it is not essential that the method chosen should give a high reliability of recognition.

7.3 Subjective assessment

It would be possible to present a number of speech spectrograms to a group of human subjects, let them learn to distinguish between the spectrograms of speech sounds, confront them next with new spectrograms of the same sounds and see finally how many of these the observers are able to classify correctly for a given mode of the analyser.

One major objection to doing this is the difficulty one has of finding a method which is only marginally influenced by learning effects. In certain psychophysical experiments, it is not possible to keep a close check on the behaviour of the subjects; for example, it

could very well be that the subject is tired or lazy one day or that he is merely genial on the other day or a few minutes later and there appears to be no way of controlling these factors or of measuring them. It would be rather difficult in this case, if not impossible, to devise an experiment in which sufficient safeguards would have been introduced. There is always the possibility of using naïve subjects and thus be perhaps able to neglect the learning effect but it is clear that we cannot expect anyone unfamiliar with speech spectrograms to classify them usefully for the purpose of this work; such a study would probably tell more about the workings of the human brain than about speech.

Another objection is that any conclusion about this project would not be purely a conclusion about speech but about speech spectrograms as seen by human beings. This study might have been of great interest twenty years ago as phoneticians were starting to use speech spectrum analysers but it must be realized that it is only for technical purposes - for automatic speech recognition or for speech bandwidth compression, for example - that systems operating in real time with minimum redundancy are of crucial importance: in other words a phonetician can take his time, he can see to it that his spectrum analyser gives him a larger number of data than the theoretical minimum because redundancy is generally something very useful and the results will be used for

visual processing; for automatic speech recognition, however, or for bandwidth compression, the idea is to try to do as much as possible with as little data as possible, in order to reduce processing costs. One could, of course, always try to extend conclusions which are valid for the workings of the human brain to automatic speech recognition systems, but one would really be evading the point that the human brain and the mechanical recognition systems can probably not be equated.

There are then two serious objections to using subjective tests of the type mentioned above, one being that it appears very difficult to control factors such as learning effects and the concentration of the subject, the second one being that, even if it was possible to control these factors, the results would still not permit us to draw the conclusions we want.

7.4 Some crude device

If a human subject appears not to be acceptable because of his capacity to learn and of the fluctuations of his intellectual performance, would it be then acceptable to measure the performance of some crude device able to classify speech sounds? This idea must also be rejected and for almost the same reasons as previously.

Fundamentally, the problem is that the conclusions would be valid only for the crude device chosen; at the end of the experiments, one might modify the device and

obtain completely different results.

Moreover, how would this device be built? How would the parameters be chosen? How is one going to ensure that the results are not biased according to the conclusions one would like to be able to draw?

Nevertheless, one could, on the basis of a series of tedious measurements on spectrograms, build a categorizer which would draw the boundaries between sounds according to the same rules for all displays. This instrument would, however, probably not be simple enough to be practically realizable within the scope of this project.

7.5 The basic requirements

The classifiers of the preceeding section were rejected because they were too dependent on uncontrollable factors of one type or another; these uncontrollable factors could be learning or fatigue effects from the part of human observers and they would have influenced not only the results of subjective tests but also the design of any simple categorizer.

What is needed then is a categorizer built according to some fairly general rules, so as to minimize the differences of learning between experiments. Is there a way to control the experiments so that the only variable is the set of measurements taken on speech sounds?

In the remainder of this chapter, a method will

be exposed, which, in the opinion of the author comes rather close to this ideal: a learning system based on the principles and methods of statistical decision theory.

7.6 A learning system

The speech recognition problem can be thought of in terms not too different from the problem of detection of a signal in noise: in this last case, one would have an input function which could be either

$$f(t) = m(t) + n(t) \quad \dots(7.01)$$

or

$$f(t) = n(t) \quad \dots(7.02)$$

where $n(t)$ represents the noise and $m(t)$ the signal whose presence or absence must be detected. The speech recognition problem can be thought of as well in terms of a characteristic part for each sound and of other parts which are common to all sounds or to a number of them. For instance, if one has two possible categories and one example of a function for each category one could write down their equations as:

$$f_1(t) = m_1(t) + n(t) \quad \dots(7.03)$$

and

$$f_2(t) = m_2(t) + n(t) \quad \dots(7.04)$$

where an extra term $m_2(t)$ is included. If one had more

than two categories, one would need a more complicated notation to take into account the parts common to all categories, common to a lesser number of categories and the part which is absolutely characteristic of each category, but the principle would remain the same.

Both problems are then analogous and one can attempt to approach them according to the principles of statistical decision theory.

Let us take samples of $f(t)$: $f_2 \dots f_n$. Clearly, the probability distribution of having $f_1(t)$ and the one of having $f_2(t)$ depend on the probability distributions of all the different samples:

$$P_1(t) = P(f_1, f_2 \dots f_n) \quad \dots(7.05)$$

$$P_2(f) = P(f_1, f_2 \dots f_n) \quad \dots(7.06)$$

Let us choose the very simple case where we take only two samples f_1 and f_2 per input function, and measure the probability of having $m_1(t)$ or not for different values of f_1 and f_2 : this situation can be visualized on figure 7.01 where f_1 and f_2 are axes of coordinates forming a "sample plane" where each input function can be plotted as a point. One has then distinctive regions on the plane which are characteristic of a given type of input functions. If one has more than two types of input

functions, the plane can be divided further into more areas; it could also happen that separate regions on the plane belong to the same classification.

On the basis of the data available for learning, one must choose the decision rule, that is, one must draw lines separating the plane in two or more regions; these lines are drawn according to the probability distributions for each category. If one establishes a table of cost of errors and correct decisions, it is possible to arrive at an optimum decision rule with respect to these costs. A decision rule which gives a cost of zero to correct detection or correct non-detection and a cost of one to incorrect non-detection or to false alarm is often used in practice and is called the rule of the "ideal observer": in the case where the probabilities of occurrence and non-occurrence of a signal are equal, the decision function is drawn on the sample space through the equal probability points separating every two categories. A combination of straight lines can be chosen as in figure 7.02 for reasons of simplicity.

This constitutes the learning process.

Once the learning process is completed, new input functions are generated; for each of them, the samples f_1 and f_2 are taken, and the input function is classified according to where on the sample plane falls the point representing it. One can check whether there was an

error or not in the identification; the percentage of errors will serve as a measure of the performance of the system.

7.7 Application to our problem

Each sound will be considered as a signal with a given affiliation. Each sound will be characterized by its spectrographic samples (up to 500) which will be taken for various operating modes of the analyser. In mathematical language, we will be dealing with a multi-dimensional space (a so-called hyperspace with hyperlines and hyperpoints) but the methods remain the same as if one had only two samples. If one chooses linear decision functions as in figure 7.02, instead of having separate lines of equation $a_1 f_1 + b_1 f_2 = C_1$, one will have

$$a_1 f_1 + b_1 f_2 + c_1 f_3 + \dots + n_1 f_n = w_1.$$

7.8 Discussion

If it is possible, on the basis of learning samples, to draw something approaching this line of equal probability between two categories, or a decision function based on a similar criterion, one will have the possibility of checking the decision rule. Since the learning samples can be more or less representative of the whole universe of speech sounds of these categories, which is

presumably represented by the testing samples, the decision function can be anything from very poor to excellent.

If, however, one uses the same sound samples for learning and testing, processes them with the analyser in mode "A" and mode "B" for instance, and draws the boundaries on the basis of the sound samples reserved for learning and tests the decision function with the other speech samples, and if one finds that the performance for mode "A" is better than for mode "B", what should one conclude?

One should then conclude that, all other factors being equal, mode "A" offers a superior set of measurements and, that the information sought is more easily accessible in "A" than in "B".

One can then choose fragments of the speech wave, analyse them in different modes and find the optimum one. From this knowledge, one could possibly design a segmenter for use in an adaptive speech-spectrum analyser.

7.9 Definition

The word "sample" has been used in this chapter and will be used in the remainder of this work in three contexts:

- The word sample can be used to designate every measurement done on the speech wave, that is, in this case, the amplitude value of a "sub-elementary area" of the spectrogram.

- The expression "sample space" refers to the group of samples pertinent to a given spectrogram.

- By "speech sound sample", "sound sample" and "speech sample" are meant the specific speech utterances selected for the experiments.

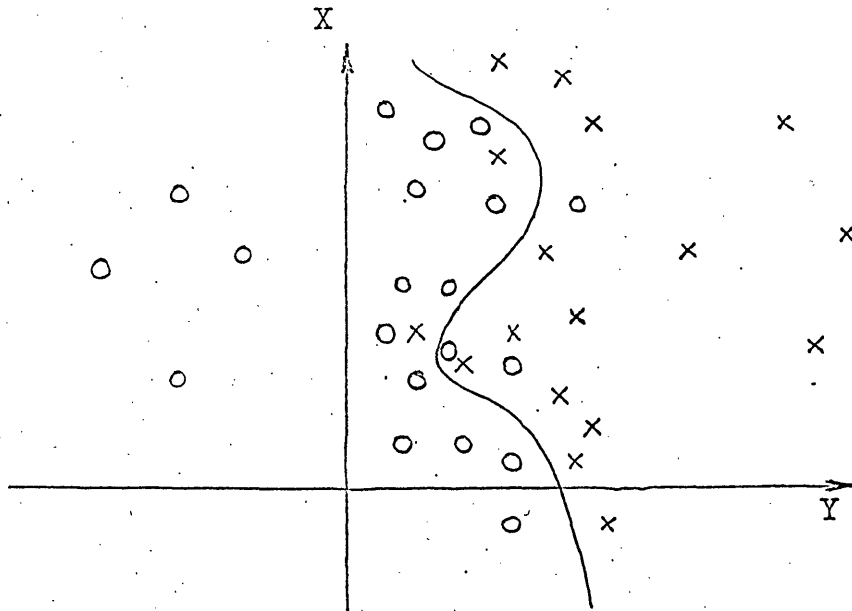


Figure 7.01

A decision function or boundary between two categories.

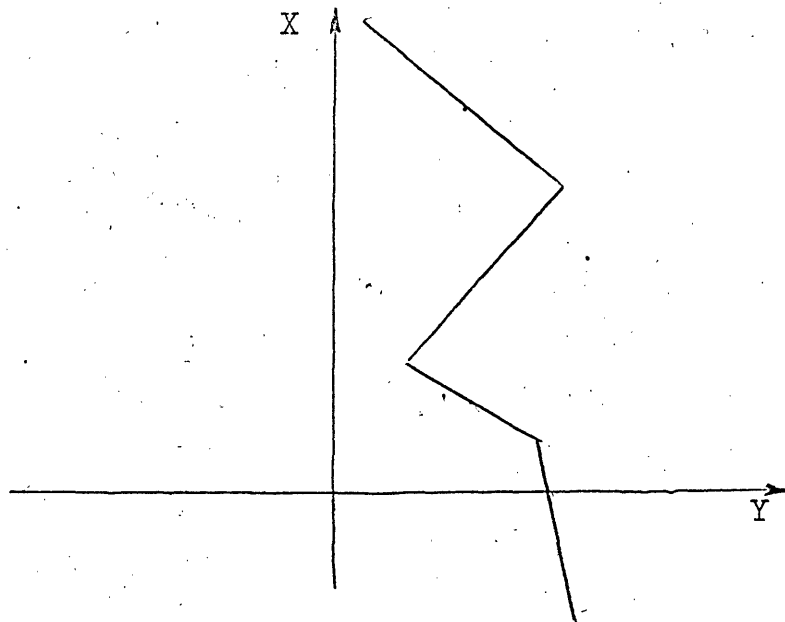


Figure 7.02

Approximation of Figure 7.01 with linear functions.

Chapter 8

BRAVERMAN'S ALGORITHM AND MODIFICATIONS

8.1 Introduction

The general pattern recognition problem can be thought of in terms of a receiver and a categorizer. The discussion of the previous chapter has been about establishing a method which would permit to equalize the effects of the categorizer and, thus, assess the value of the receiver.

A good evaluation of a probability distribution and in particular of a multivariate one can be obtained only if a large number of samples is available. For a small number of samples, unless the shape of the probability function is already known or is very simple, one cannot hope to be able to draw it precisely. It is to suspect that the probability function of many sounds is not straightforward in shape, since research on speech has failed to isolate definite clues or invariants which are necessary characteristics of each speech sound.

However, as was mentioned earlier, it is not so much the exactitude of the decision rule which is of interest, but whether speech sounds are easier to categorize in one mode than in the other.

Hence, a method was chosen which is not based on the measure of the probability distributions, but which

should give fairly similar results in a simple case. It has been proposed and used by Braverman (1962) for the recognition of printed digits. It has been adapted by Turski (1964) to stability problems in control systems. This last author has published the program he used and large parts of it were included in the programs written for the present thesis. It is the first time, as far as the author knows, that this method is applied to the recognition of speech and, in particular, with a view of assessing the value of the receiver (analyser) by equating the characteristics of the categorizer (learning process).

8.2 Basic assumptions

Braverman's algorithm is based on the hypothesis of compactness of each pattern: that is, each pattern is taken as being characterized by well-defined regions which one has to delimit in the learning process.

If one tries, then, on the basis of a few sound samples to draw boundaries between two or more categories, one will probably do as well, on average, with Braverman's algorithm than with any other method. If a larger number of samples is used, however, and if the regions characteristic of each category somewhat overlap, the boundaries drawn between categories will take weird shapes such as illustrated on figure 8.01, while a boundary

drawn through the points of equal probability between two categories would be much smoother. But, as far as the results are concerned, they are likely not to be very different in either case; the main difference might be the larger storage capacity required by Braverman's algorithm.

When only a very few samples are available for learning, it is not possible to have more than the vaguest idea of the shape of the probability distribution. Braverman's algorithm offers then as good a choice for the decision function as one can imagine.

What is, however, of primary importance for us is that this method fulfills the requirements spelt out in the previous chapter, that it provides us with an unbiased technique for comparing the usefulness of various displays.

8.3 Braverman's algorithm

Only a brief description of this algorithm will be given. Those who want to know more should refer to the original publications of Braverman. A complete program is given in appendix.

The algorithm can be divided into four parts:

1. construction of separating hyperplanes;
2. elimination of redundant rows and hyperplanes;
3. elimination of redundant pieces of hyperplanes;
4. recognition or classification of new objects.

8.3.1 Construction of separating hyperplanes

Let us try to divide an hyperspace between two categories. We are limited however to a two-dimensional representation for our diagrams but this will permit us nevertheless to illustrate the main concepts. As shown in figure 8.02, we can draw between input No 1 belonging to category "o" and input No 2 belonging to category "x" an hyperplane "a". Input No 3 (x) is read in but falls on the same side of "a" as No 1 (o) and necessitates the drawing of hyperplane "b". Input No 4 (o) is read in and, since it falls in the same region as No 2 (o), a new hyperplane is drawn "c". Input No 5 (o) is read in, falls in a region already occupied by No 1 (o) but since both belong to the same category, they do not need to be separated. Braverman rejects points of this kind but for the present work, it has been decided to store them since they can be useful to draw new hyperplanes. Input No 6 (x) is read in but falls in a previously unclassified region: this region is then classified according to this sample.

All this can be summed up in the boolean matrix of figure 8.02, where the rows denote the input sounds and the columns the hyperplanes. If the sound falls on the positive side of the hyperplane, it is denoted by a "one" and if not, by a "zero".

In order not to waste some storage capacity of the computer, redundant rows and columns of matrix A are eliminated; let us eliminate row No 1 (figure 8.02).

8.3.2 Elimination of redundant matrix elements

There are portions of the hyperplanes which are redundant. Figure 8.03 is the repetition of figure 8.02, but with two redundant pieces of hyperplanes erased. This can be stored in the computer by introducing a new matrix B which has the same number of rows and columns than matrix A; a "one" in matrix B indicates that the same element of matrix A is essential, a "zero" in matrix B indicates that the same element in matrix A is superfluous. Great care must be taken in this operation, which is very delicate: one should note, for example, that some categories such as [0,1,1] are impossible (non-existent).

There remains finally unclassified regions. By properly modifying matrix B it is possible to include them in one category or the other. Once this has been done, it is possible that some rows and columns will have become redundant: they are then eliminated. The final outcome could be as in figure 8.04.

8.3.3 Testing the decision function

The decision function can be tested by presenting to the computer new objects of the same categories.

8.4 Modifications

Braverman uses as coefficients for the hyperplanes

a series of pseudo-random numbers. Having done this, he uses an iterative method so that these planes are displaced in such a way as to be approximately half way between the learning samples of different affiliation. He uses then what he calls the method of paralleling variants, that is, he uses many computers in parallel and takes as final answer the choice of the majority; he has found that this method increases tremendously the reliability of the algorithm (from 80% to nearly 99%) in an experiment concerned with the recognition of a few printed digits.

Such refinements were out of question, however, for the present work and it was resolved to use hyperplanes drawn through the middle of and perpendicular to the hyperline joining two hyperpoints of different affiliation. This step is not unprecedented: Braverman mentions having tried it and having found an improvement in reliability; Turski also mentions it as a method which would improve the reliability of the algorithm; but both point out that this technique does not possess the same flexibility as that of the paralleling variants. To use perpendicular hyperplanes was, however, the only course open to us due to the limitations in computing time

8.5 Definitions

Confusion might arise in the following chapters

due to the use of the term hyperplane to designate either the individual hyperplanes drawn between two hyperpoints of different affiliation or the decision rule, which is after all, also an hyperplane built up from pieces of hyperplanes. The hyperplanes of the first type will then be termed individual hyperplanes or simply hyperplanes or planes while those of the second type will be referred to as decision functions, decision rules or complex hyperplanes.

8.6 Conclusion

It was mentioned earlier that certain regions in the space were left unclassified and that, in the process of simplifying the decision rule, the algorithm would give these regions the same affiliation as a convenient neighbouring region. The next chapters will deal with these problems and others and indicate where the Braverman's algorithm as modified falls short of maintaining perfect parallelism between the experiments carried out in different analysing modes. We will try to determine the best course of action in order to be the least affected by these deviations from the ideal.

A simplified flow diagram of the program will be found in figure 8.05. The program itself, written in Algol, can be found in appendix.

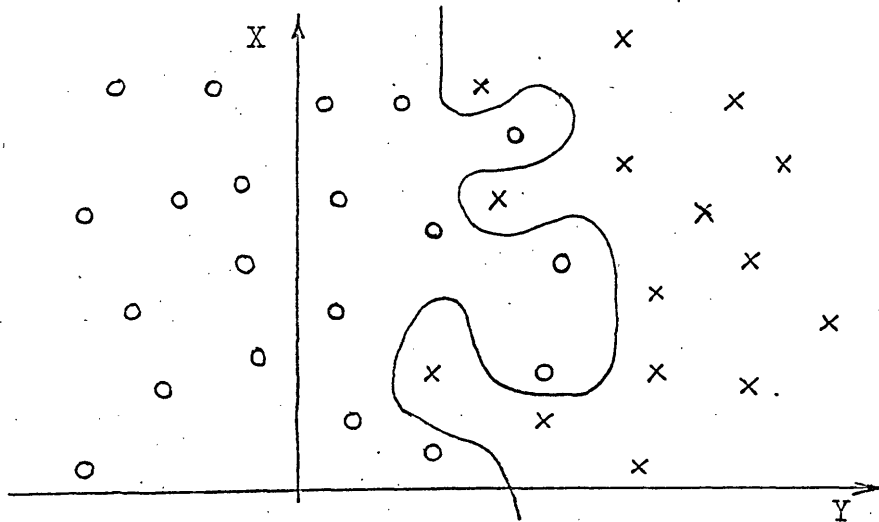


Figure 8.01

Type of decision function separating two overlapping classes, which arises from Braverman's algorithm.

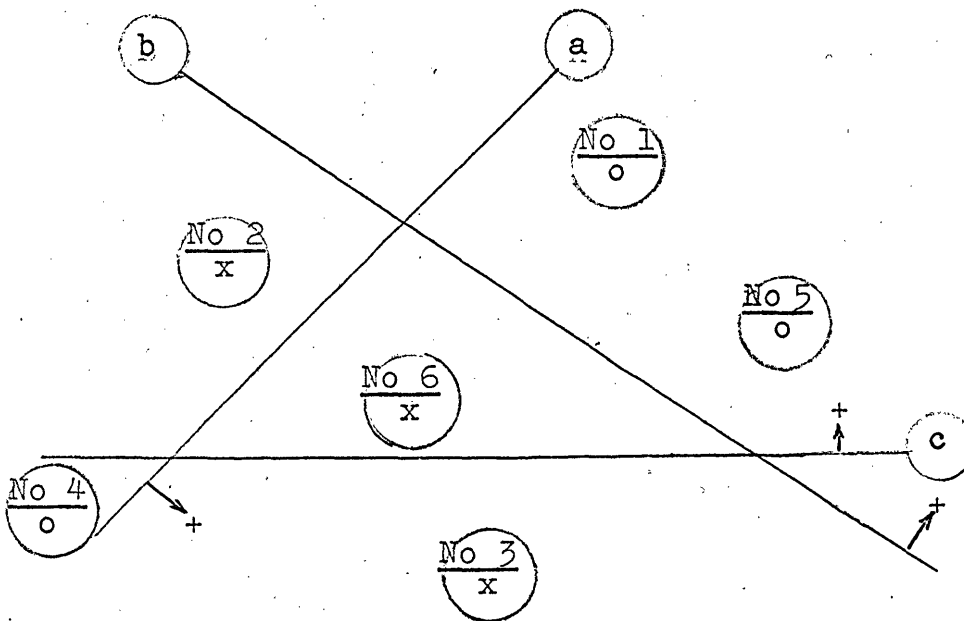


Figure 8.02

Braverman's algorithm: drawing the hyperplanes.

Boolean Matrix "A"

Inputs No.	Hyperplanes		
	a	b	c
1) o	1	1	0
2) x	0	0	0
3) x	1	0	1
4) o	0	0	1
5) o	1	1	0
6) x	1	0	0
not classified	1	1	1
	0	1	0

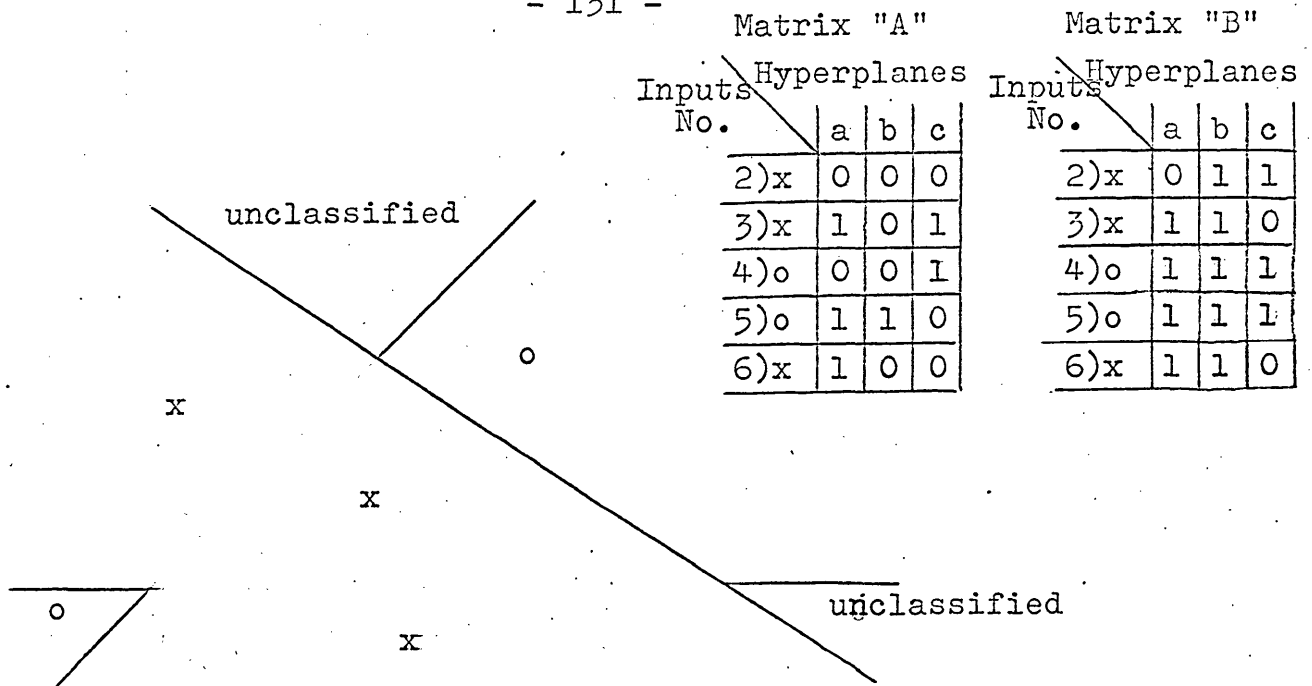


Figure 8.03

Braverman's algorithm: erasing pieces of hyperplanes and eliminating redundant input points.

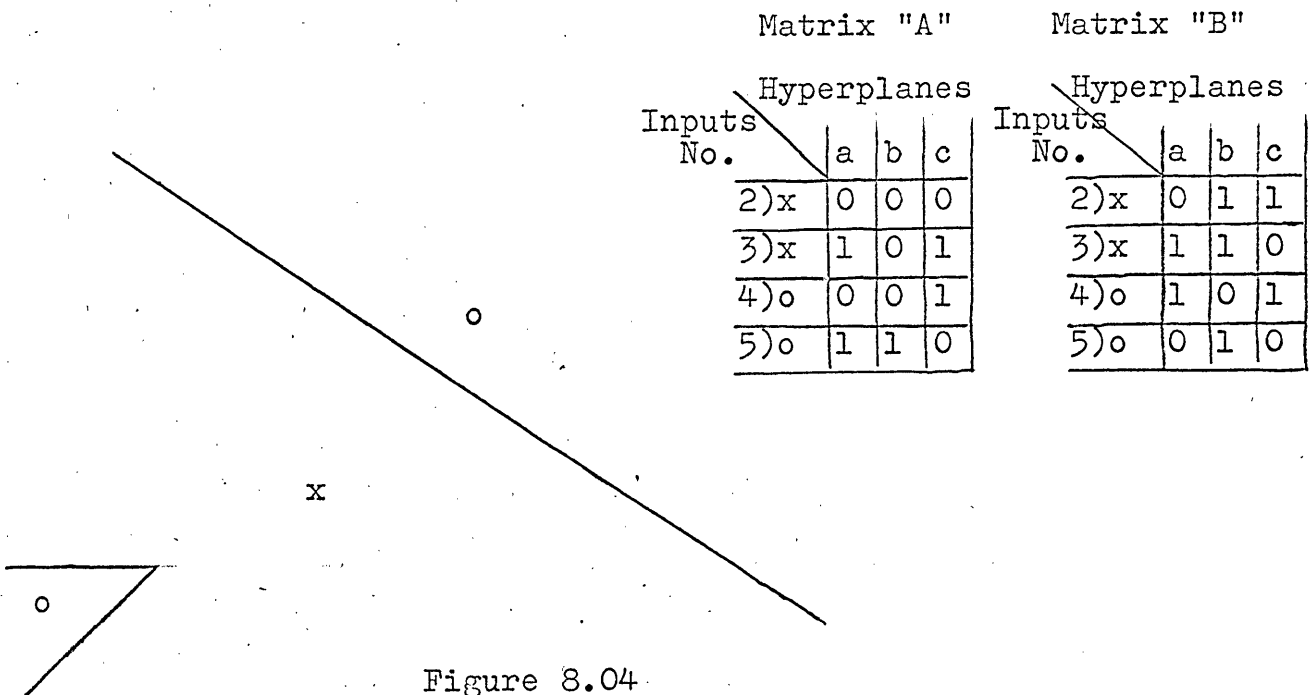


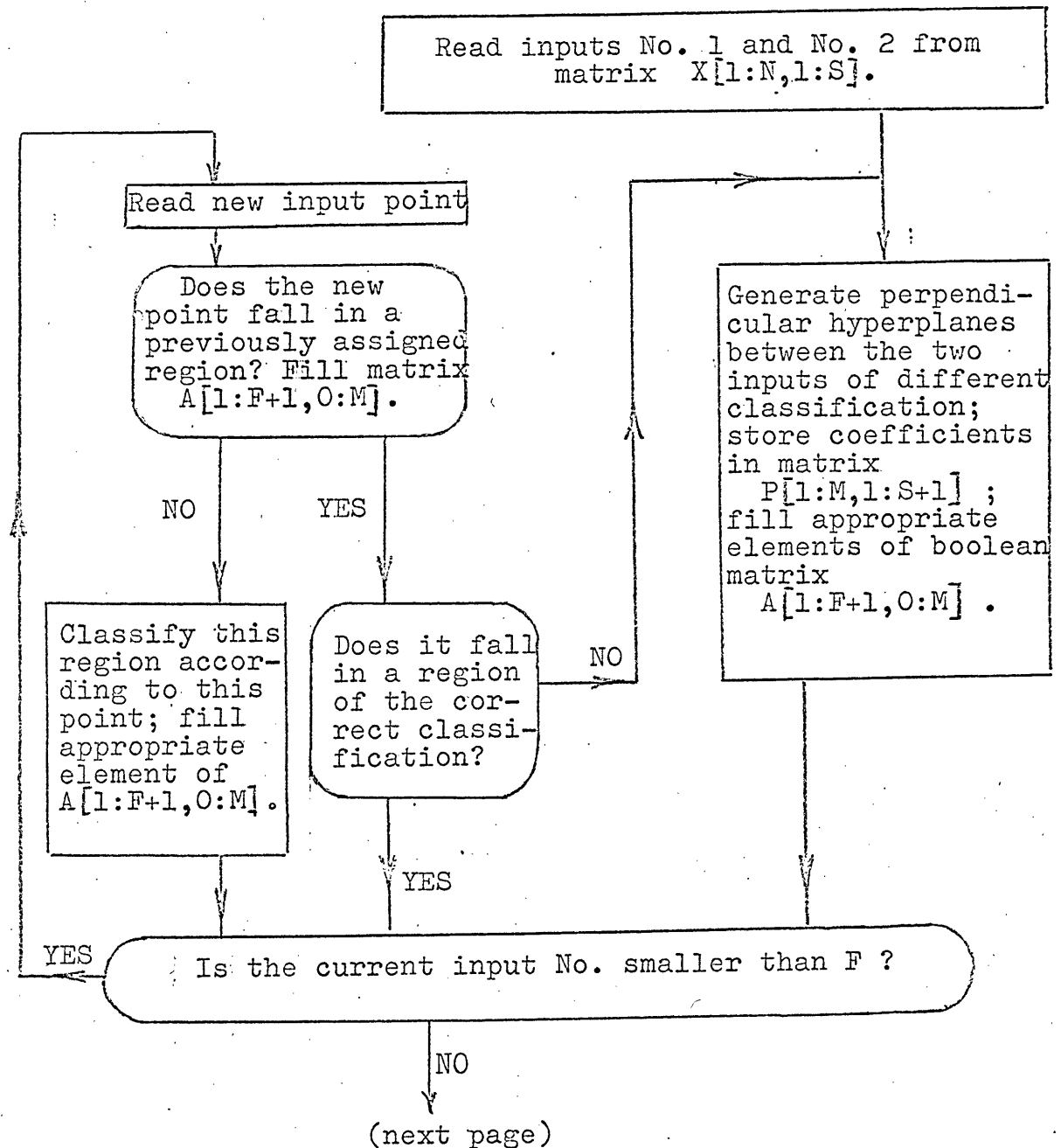
Figure 8.04

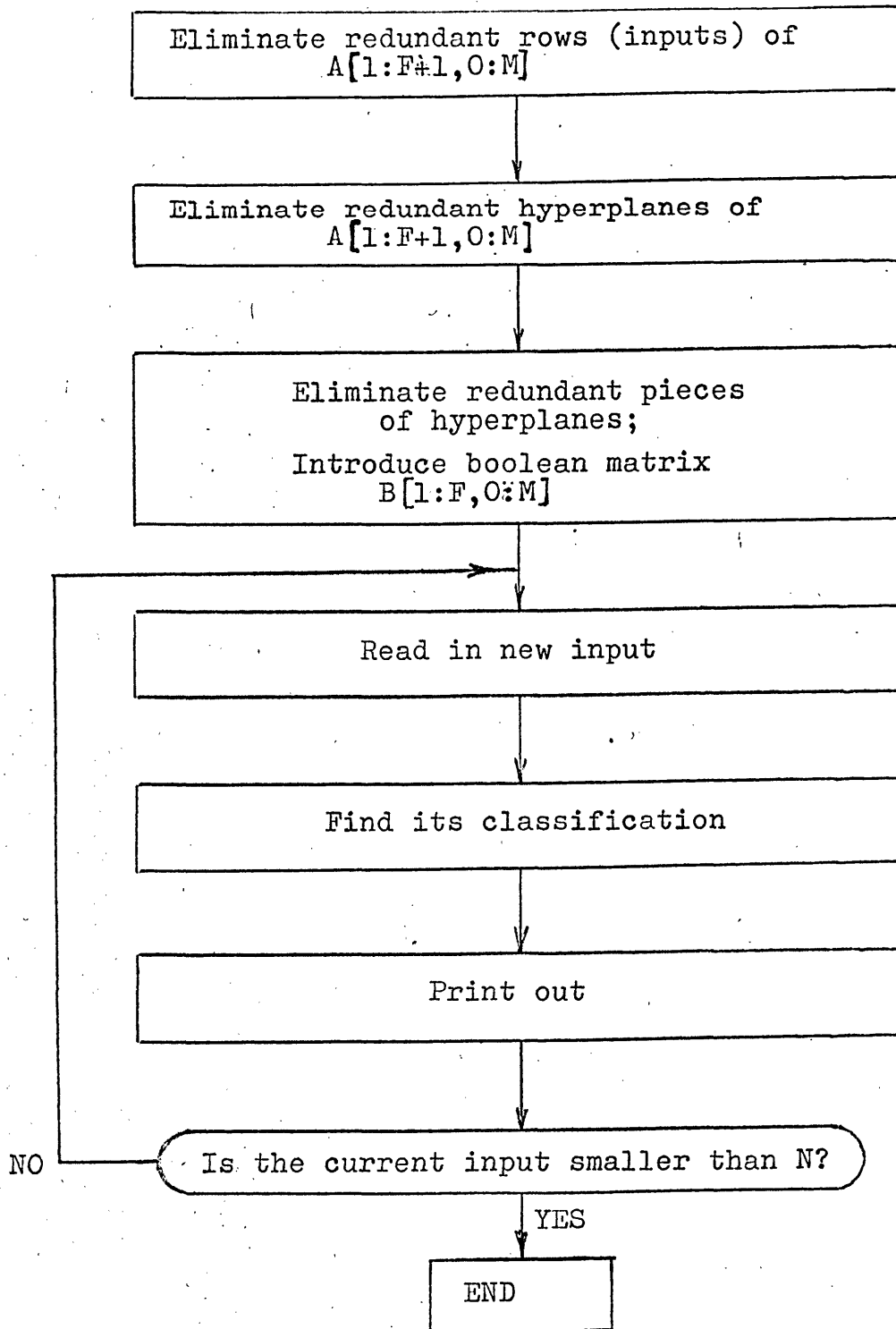
Braverman's algorithm: classifying unaffiliated regions.

Figure 8.05

Simplified flow diagram of the modified algorithm.

Comment: there are N input spectrograms, of which the first F are selected for learning ($F < N$); there is a maximum of M separating planes; there are S samples for each spectrogram; Input No. 1 and input No. 2 belong to different categories.





Chapter 9

THE NATURE OF THE EXPERIMENTS

9.1 Introduction

This chapter will be concerned with the practical points raised by the experiments. The main problems were to decide what speakers to use, which speech sounds should be recorded, which ones should be used for learning, and what part of the speech sounds should be taken for study; added to this, were the normal problems of pattern recognition, in particular the problem of positioning the samples. Finally, everything had to be done with the view in mind to economize on the physical effort involved and to make the best possible use of the computing time available.

It was not always possible to justify every step and in order to pursue the main objectives of this work, some more or less arbitrary decisions had to be taken concerning secondary points. Since it is not expected that these decisions have greatly influenced the trend of the results, they should not, in the view of the author, diminish the value of this work but rather enhance its exploratory character.

9.2 The data

9.2.1 The speakers

Four main speakers were used; they had different backgrounds although all of them had been in England for a few years; one of them was American (D.G.), one came from Ghana (P.P.), one was English (R.W.) and one was the author of this thesis, a French-Canadian (M.L.). Three other speakers, all English, were used for a few sounds only. These subjects were chosen because of their different backgrounds and their clear elocution. Since the object of this project was to study speech sounds on their generality, it was logical not to restrict the choice of speakers to those speaking a given dialect or having a particular accent although this would probably have improved the absolute recognition performance. The speakers were postgraduate students in Imperial College, they were all male and all about 25 years old.

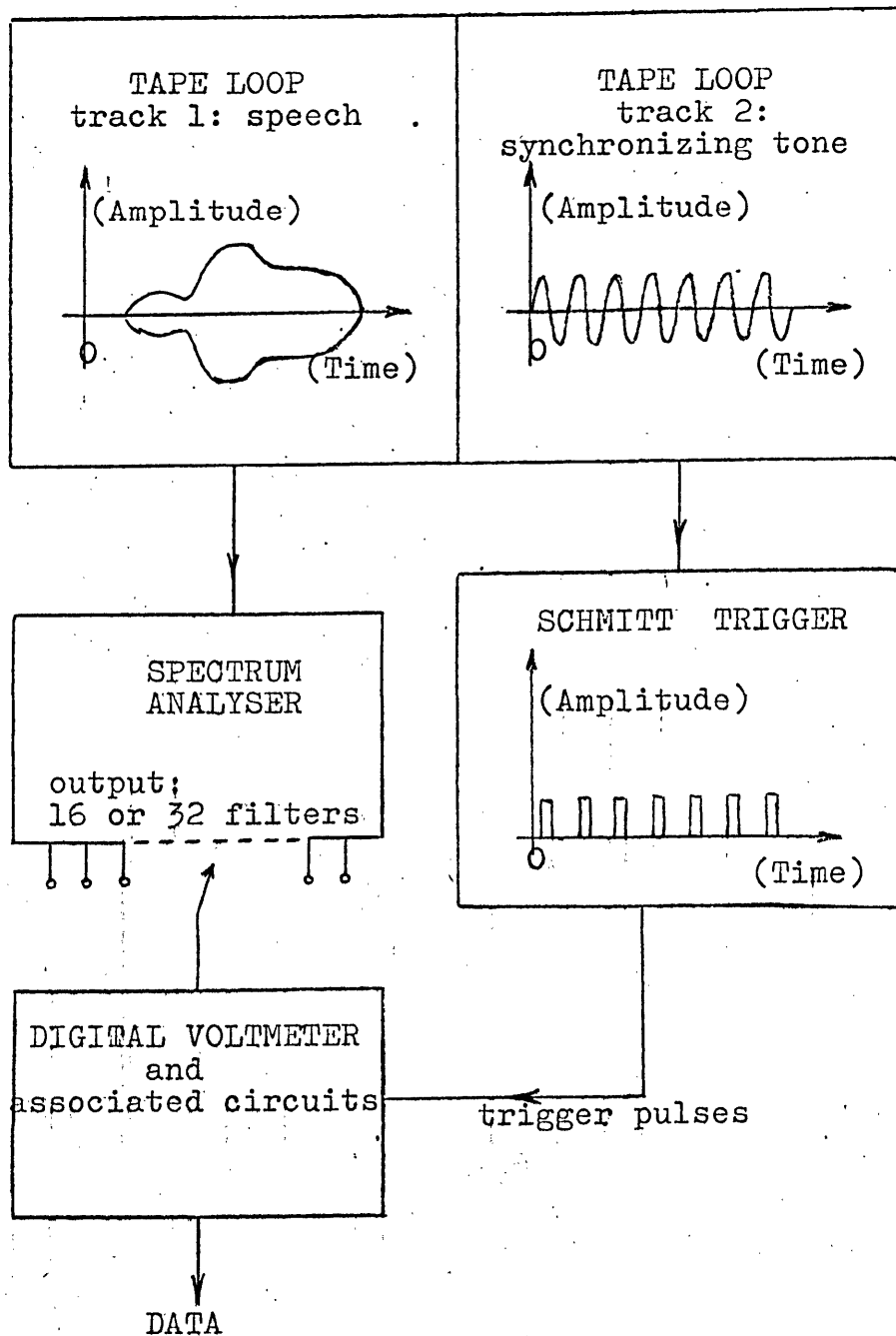
9.2.2 Recording the data

The speech sounds were recorded in normal conditions. In order to ensure, however, that they should be more or less free of noise, the recordings were made in the evenings in quiet conditions. An EMI RE 301/C tape recorder was used. The normal recording and playback speed used was 3 3/4 and for this speed the fre-

quency response of the tape recorder was almost flat from 50 c/s upwards in the range of interest (up to 3.2 kc in this case). An omnidirectional dynamic microphone (Electro-voice, model 655c) with a uniform frequency response from 55 c/s upwards was used.

The speech sound was recorded on track 1 of the tape loop; a synchronizing tone was recorded on track 2. A digital voltmeter, coupled to a paper-tape punch, was connected in turn to each filter at the output of the analyser, every time the tape loop came back to its starting point (see figure 9.01 for block diagram and appendix B for details). Since the digital voltmeter could record only at a speed of 50 times a second and that, in the 100 c/s mode, each one of the two banks of filters is sampled at this rate and, in the 200 c/s case, twice as often, a rather involved system of gates was needed to ensure that the spectrogram could be correctly reconstituted in the computer; this meant a long and tedious series of manual operations, checks and double-checks. The same synchronizing tone on track 2 could be used when recording with the analyser set in the 100 c/s or the 200 c/s mode.

To simulate the operation of the spectrum analyser with its filters adjusted to the bandwidth of 50 c/s, the tape recorder was played back at twice the speed with the analyser in the 100 c/s mode: this necessitated



Block diagram of assembly for recording the spectral data on paper tape.

Fig. 9.01

the recording of a new synchronizing tone on track 2. Proper adjustments had to be made in the program to take into account the effect which the amplitude response of the tape recorder at that speed ($7\frac{1}{2}$) and the characteristic of the frequency emphasis circuit at the input of the analyser had on the amplitude of the individual samples making up a spectrogram.

9.2.3 Speech sounds selected for study

The consonants /p/, /b/, /t/, /d/ in initial position were studied. They were followed by one of the following speech sounds /a/ as in "part", /ɔ/ as in "talk", /et/ as in "bet", /it/ as in "bit". Data was also taken with the velar stop-consonants /k/ and /g/, but was not much used, because it appeared after some time that, partly because of the physical effort involved and partly because of the limited computing time available, it was necessary to restrict the scope of this work. The two velar stop-consonants were left out because their spectrum tends to vary a lot with the nature of the sound following them and since, hence, the chances to obtain worthwhile results with the small number of samples which were available for learning did not appear too good. The remaining choice of consonants was at any rate interesting since it contained two pairs of very similar sounds /p-b/ and /d-t/, two pairs of rather different ones /p-t/

and /b-d/. The two remaining pairs /p-d/ and /b-t/ can be said to be almost at opposite ends of their category, because most of their energy is in different parts of the spectrum and one sound in each pair is voiced while the other is unvoiced. Similarly for the vowel sounds following the consonants one has two pairs of very similar sounds /a-o/ and /et-it/ representing a very high level of difficulty and four pairs /a-et/, /o-et/, /a-it/, /o-it/ representing a very low level of difficulty. A list of the speech-sound samples taken and of the speakers will be found in Appendix A. More samples from M.L. were taken because it became evident after some time that a larger number of samples had to be taken than was originally planned and one speaker was selected for that purpose. In Appendix A, a list of the vowels used for learning in the vowel experiments can also be found. About 20 samples for each consonant were taken.

In the remainder of this thesis the phonetic notation for vowels will not be adhered to, as was already done in the last paragraph.

9.3 The experiments

9.3.1 Consonants or vowels?

It can be seen from the spectrograms in Chapter 6 that the 200 c/s display destroys very effectively the

formant patterns of the vowel sounds. It seems then to be less interesting to study vowel sounds because, while it is questionable whether it would really be profitable to use larger bandwidth filters for consonants, one has to accept that the formant structure is destroyed to such an extent in the 200 c/s mode that the results of recognition experiments are bound to become poorer as the number of vowels in the dictionary of possible sounds is increased. It was then decided to do a detailed study of a few consonants and what could be called a symbolic study of a few vowels.

9.3.2 Number of categories

It was attempted at first to take the whole lot of speech sounds and to do recognition experiments involving the four consonants and a program was written with this object in mind. However, the computing time required to compile and execute this more complex program was much longer than it should be for optimal use of the Atlas computer of the London University and a great deal of data had to be stored and processed for a single program. As a result, the program was always delayed a lot at the computer and it was finally decided to take the consonants pair by pair. Looking back at the results, this seems to have been a fortunate decision: it has permitted to do a more detailed study of each

consonant in relation to every other; the number of speech sound samples available would not have been large enough to draw really significant confusion matrices or contingency tables. It became gradually clear that the behaviour of each pair of consonants could be rather different, and it is likely that to have used all of them at the same time would have confused the picture. The present work probably does not permit to have a perfect idea of the general picture, to judge what kind of performance one should expect from each display if one had to distinguish between a fair number of speech sounds, but it gives an idea of the difference existing between two sounds. It is felt that, in the present state of speech research, this is a valuable contribution and one which can be classified at the phonemic level. If the sounds had not been compared two by two, the main conclusion would have been on the behaviour of a much less fundamental group of speech sounds.

9.3.3 Segmentation

The speech sounds had to be segmented, that is, a portion of them had to be selected for study, and the segmentation had to be carried out not only between the phonemes but also inside the phoneme. For example, there are in the stop-consonant at least two parts, one which consists of a burst of noise and another which

contains the frequency transitions from the consonant to the vowel. It is normal to expect that the optimum mode of the analyser would not always be the same for all parts of the consonant. One had at first to proceed by trial and error; after some time, it became convenient to study either the first 20 msec of the consonant, or the first 40 msec, or the second part of the consonant going from 20 msec or 30 msec to 60 msec after the starting point. Anything after 60 msec could be considered as being part of the vowel. More about this will be written later on in this chapter and in the next ones.

9.3.4 Voicing

The question arose whether one should include as well in this study the time interval immediately preceding the stop-consonant which is usually empty for unvoiced stop-consonants and usually filled with ^{pre-voicing} voicing in the case of a voiced stop-consonant. It was finally decided not to for the following reasons: First of all, the purpose of this work is not to study the proportion of times a so-called unvoiced consonant is in fact unvoiced and there are simpler ways of doing it; secondly, there is no reason why ^{pre-voicing} voicing should be more easily detected with one type of filters; thirdly, it would have meant noticeably longer data tapes since the data

on a supplementary interval of fairly long duration would have had to be recorded, which would have required a lot of the computer's storage capacity.

9.4 Controlling the amplitude

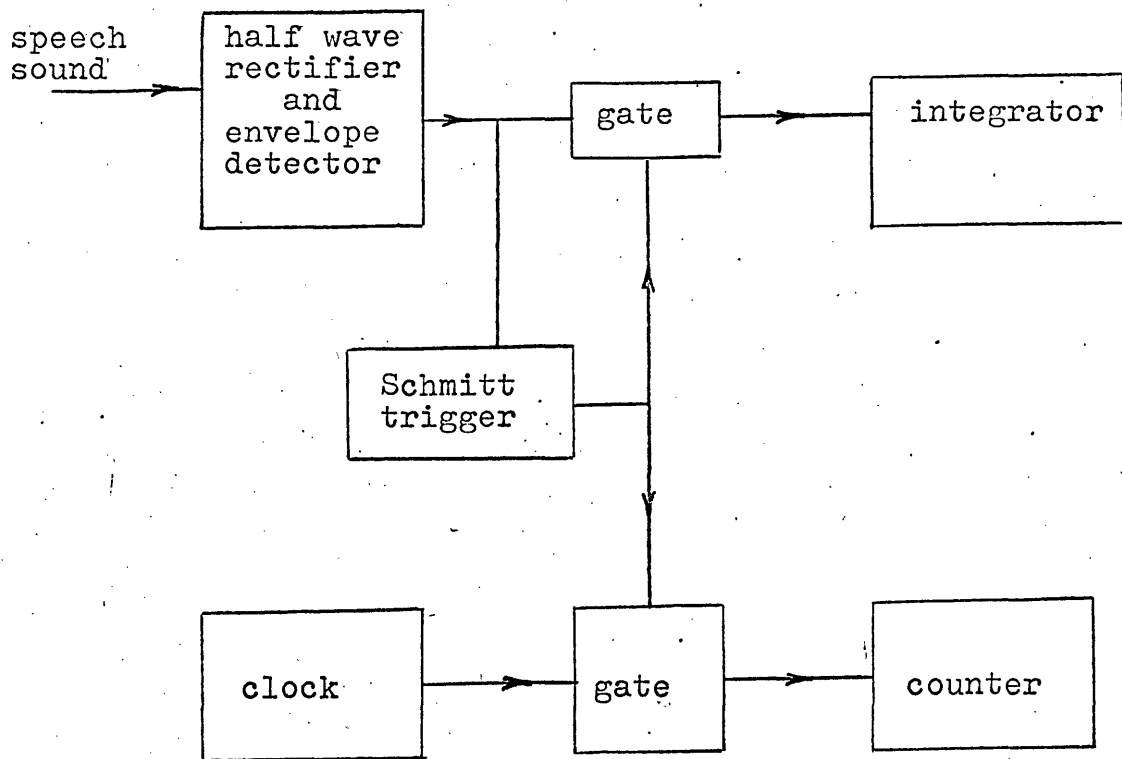
In visual pattern recognition, one has to cope with objects of various sizes and orientation. It is fortunate that in speech recognition there is no problem with orientation and there are very definite constraints applicable to the frequency, time and amplitude coordinates. From the point of view of pattern recognition, the most troublesome of these three factors are certainly the amplitude and the duration of speech sounds. The author does not know of anything that can be done to control the duration but it is certainly possible to reduce the variations of amplitude between different utterances of the same speech sounds.

The 200 c/s display would seem more likely to be adversely affected by a poor method of controlling the amplitude level of the speech wave since it would seem that the advantage of this display is to be able to follow more closely the fluctuations in amplitude, and to measure with more precision the rate of rise and decay of the speech wave.

It was attempted at first to control the level by measuring the average amplitude of the speech wave;

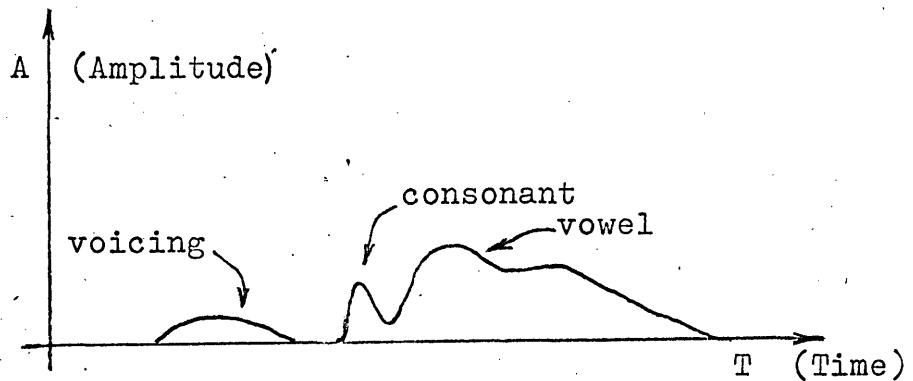
all that is needed to do is to detect the envelope of the speech wave and to integrate it while measuring its duration (figure 9.02). It was found that this method, however attractive and simple in theory, was very difficult to implement in practice and would have required a lot of development work. For one thing, what is the duration of a sound? There seemed to be no way of setting a Schmitt's trigger so that it would repeatedly give a sensible answer: for example, it would occasionally trigger on the voicing, or miss the consonant or be triggered by noise (figure 9.03). As it clearly appeared that this technique would be awkward and cumbersome, without being necessarily all that good, it was abandoned in favour of a simpler one.

The V.U. meter of the tape recorder was used to control the amplitude of the speech samples during recording as well as during the play-back. Subjects were asked to hold the microphone six to ten inches in front of their mouth and to speak at a normal level, the controls being adjusted so that the peaks of the speech wave would be about 8 dbm (reference: 0.776 volts). When playing back, the controls were set in such a way that, at maximum deflection, the needle indicated about 11 dbm. For the experiments with the 50 c/s filters, a lower level was chosen for the play-back in order not to saturate the filters.



Assembly for measuring the average amplitude of a speech sound.

Fig. 9.02



Envelope of a typical speech wave.

Fig. 9.03

In as much as could be seen, this method was not only practical, but also seemed to give fair results. The peaks of the speech wave are, of course, the vowels and it might be considered unfortunate that, in a work concerned more with the study of consonants than vowels, the vowels were used to adjust the amplitude level. There is no reason to think, however, that any other technique would give superior results and every reason to think that this technique or similar ones based on automatic gain control techniques, are so simple that they would be used in practical speech recognition systems.

9.5 The positioning of speech sounds

The superposition of the items to be classified or recognized is a basic problem in pattern recognition and one which can be solved only by empirical means. By superposition is meant here the positioning relatively to each other of the items to be compared, which is a necessary preliminary step before any measurement is taken. This step constitutes a major difficulty in recognition of visual patterns because of the different sizes and various shapes an object of the same category can have.

9.5.1 Constraints introduced by speech

The problem is perhaps not ~~much~~ simpler in speech

recognition: instead of having different sizes, one has to deal with variations in the amplitude level; sounds can be shorter or longer. Fortunately, however, the frequency coordinate is more or less fixed and the time coordinate can vary only within narrow limits.

It is particularly advantageous to study stop-consonants in initial position because it is in this case that the time constraints are the strongest: it is impossible to sustain a stop-consonant for any length of time: the data available shows that the duration of each stop-consonant is about the same for different speakers. Crandall, for example, (see Fletcher page 64)⁶⁷ has taken certain statistical measurements of consonants. Some of his results are given in table 9.01 for reference purposes. It can be seen that the duration of the consonants is very much alike for the two speakers especially for unvoiced consonants; it is likely that the larger variations for the voiced consonants are due to voicing.

It can be seen in the same reference that consonants in general vary little in duration for different speakers, that no other group of consonants, however, seems to be as good in this respect as the group of the stop-consonants in initial position. In final position, of course, stop-consonants are often ~~implied~~^{unreleased}, a difficult case which will not be dealt with in this

Sound	Speaker	Consonant duration (msec)	Transitional Characteristics	
			Low freq. c/s	High freq. c/s
ba	I	120	700	2700
	II	190	700	3100
pa	I	20	1600	3600
	II	40	900	3600
da	I	130	500	2800
	II	100	600	3200
ta	I	70		3200
	II	60	900	3000
ga	I	120	550	3000
	II	100	600	3600
ka	I	70	1200	3800
	II	80	1300	4000

Statistical data
on six stop-consonants

N.B.: this data, collected by Crandall is taken from
"Fletcher, Speech and hearing in Communication, p.64"

Table 9.01

study.

9.5.2 The positioning method chosen

It was finally decided that the particular class of speech sounds studied here could be well positioned according to their starting point. The logic of this step cannot be discussed; we would intuitively agree it is a much more pertinent positioning basis than the center of gravity or other measures used in visual pattern recognition for this purpose. The information conveyed by speech components is in no way proportional to the amount of energy they contain; any positioning method which bases itself on the energy value of speech components (center of gravity, maximization of cross-correlation functions) is bound for failure unless a proper weighting function is applied to the spectrograms, a solution which, at first glance, does not seem possible.

One further advantage of this positioning method is that it permits to maintain the parallelism between the experiments with different displays, that was talked about in the previous chapters. In addition to this, it is easy to check the results against a print-out of the data and avoid accidental mistakes.

9.5.3 The positioning technique

The technique chosen to determine the starting point of a sound was a very simple one which, checked against a print-out of the data, gave sensible results.

The starting point was taken to be the earliest point at which two adjacent samples at frequencies higher than 600 c/s would equal or exceed a threshold of 0.3 volts; this threshold was found to be high enough, in practice, not to be exceeded by occasional noise and low enough to be crossed by low energy components of speech sounds; the maximum amplitude of the samples was about 3.0 volts.

Only the frequencies over 600 c/s were considered because it was found that consonants often start with a low frequency noise pattern which seems to be somewhat random in shape and amplitude value. In order, however, to take into account this part of the sound, the program was written in such a way as to include the low-frequency data in the 40 msec period immediately preceeding what we defined as the starting point of the sound. Instead of saying, then, that we are considering the first 20 msec or the first 40 msec of a consonant, we should say that we are considering the first 20 msec or 40 msec as well as the low-frequency noise preceeding the consonant.

9.5.4 Verifications

It was possible to check the positioning technique, up to a certain extent, in comparing the 100 c/s and the 200 c/s displays because of the technique used to record the data on punched tape as explained in Appendix B. Because the same 100 c/s tone recorded on track 2 of the tape was used as synchronizing signal, and because proper care was taken in the experiment, it was possible to determine the starting point of recorded data in the 200 c/s display was usually 10 msec before the starting point for the 100 c/s display; in no case did it start more than 20 msec before the 100 c/s display or did it start later. The method chosen for positioning the sound samples was one which fully satisfied the above mentioned requirements, and agreed with the checks done against a print-out of data. If then, the positioning method itself was deficient, it can be said that both displays almost suffered as much.

Furthermore, experiments were done occasionally with slightly different positioning criteria and the results were always very similar.

9.5.5 Discussion

There is no doubt, however, that the problem of positioning has introduced in the experiment a certain amount of randomness in this sense that it is not

possible to say now that exact parallelism has been kept between the experiments with 100 c/s and 200 c/s displays. It could be argued that any error in positioning is bound to hinder the performance of the 100 c/s display more than the performance of the 200 c/s display, since this latter has already so much data in the time domain. On the other hand, it could be argued in virtue of the same reasons that the performance of the 200 c/s display might be critically dependent on the efficiency of the positioning technique. It seems then impossible to determine in which direction the results would be more likely to be biased. That the results occasionally favour one display and occasionally the other by a margin which is unlikely to be explained by chance, in addition to the precautions which have been taken and the verifications which have been carried out, seems to be a rather strong argument in favour of the validity of the positioning method adopted.

The results for the 50 c/s display cannot be as readily compared to the two others. First of all, less time was available to check and double-check. Secondly, a new 100 c/s tone had to be recorded on the second track of the magnetic tape and the sampling times did not exactly correspond to those of the two other displays.

As a conclusion, it can be said that, despite the precautions which have been taken, it is possible that some of the sounds were not positioned quite the

same respectively to the others in the 100 c/s and the 200 c/s displays. This, however, should have been taken into account in the learning process and there is every indication that the results were not biased. If one of the displays shows better resistance to mispositioning, then it is certainly a factor to take into account because the problem of positioning is inseparable from pattern recognition.

Chapter 10

THE SIGNIFICANCE OF THE RESULTS

10.1 Introduction

It has been indicated in Chapter 7 that one of the reasons why a pattern recognition method had been chosen was that such an approach made it possible to establish a direct comparison between different modes of analysis. It is now time to discuss how "direct" are in fact the comparisons one is able to make and what degree of significance one should attach to them.

These comparisons will give results which are, strictly speaking, true only for the group of samples used. Only statistical methods would permit to generalize the results and conclude with reasonable certainty that the same picture would approximately repeat itself if the same study was undertaken with another ^{similar} group of samples ~~or with the whole universe of speech sounds.~~

10.2 Randomness and statistical significance

A test of significance will only indicate the probability of being right when stating a simple hypothesis, i.e. that the averages are different. One would like however to know by how much the averages differ: a more or less exact evaluation can be done

only on the group of samples available but one can assume that the differences for the whole set of sounds are of the same order. It is unlikely that exactly the same results would be obtained if the same experiment was repeated. The present study, however, will serve as a guide as to what should be the order of magnitude of any improvement, and as to where one should or should not look.

In other words, one who wants to look at the results presented here only in terms of the standard statistical tests of significance is missing half the point. In experiments where randomness is inherent, one needs statistical methods to evaluate whether differences in average are significant or not. If, however, we can convince ourselves that randomness has been, in fact, controlled to a large extent in the experiments, then, the results should also provide an indication of the order of magnitude of the differences which one should expect.

Statistical comparisons between two different processes usually involve subjects on which an experiment is performed and a control group. One can take the example of a new medical treatment which is tried on one group while the other one is left without treatment; at the end of a given period, one sees how many persons have been cured and calculates the statisti-

cal significance of the results, usually assuming a normal distribution. At first glance, this seems to be a problem similar to ours; but a very important difference is that we do not have two groups of subjects but only one, which is handled in different ways. It is rather as if, in the previous example, it was possible after having administered the treatment to one group of persons and having seen what the results are after a given time, to bring them back to the starting point and see how many of them would get cured without the treatment, or with another treatment.

In this last case, the only doubt remaining would be whether the group chosen would be representative: perhaps the exact size of the improvement is not quite what was found, perhaps certain ill-effects had not shown up. The probability distribution of the results would, in any case, be much narrower because one would not have to account for the unknown differences inherent between **two groups**.

The results presented in this thesis should perhaps not be taken as readily as those of the hypothetical experiment previously explained. The reason is that such an experiment would permit us to conclude that the particular decision rule arrived at with the 100 c/s mode, for example, was superior to the particular decision rule that was arrived at in the 200 c/s mode.

Does this mean that the 100 c/s representation will always be better or is it just a particular case? This is a point which will be discussed in the next section.

10.3 Randomness in the learning process

It has been pointed out earlier that it was desired to submit the same speech sound samples to the same learning process for both displays, and then, test both the decision functions with new speech sound samples. How alike, however, are the learning processes? Is it possible to control them so that they are exactly the same for all displays?

It is unfortunately not possible. First of all, samples making up the spectrograms of the same sound in various displays are different - the filters have different characteristics. Secondly, the decision function arrived at is not unique: it is sufficient to change the order in which the samples come for learning to change the decision function (figure 10.01). Nothing can be done against the first factor; it is, in fact, the object of this work to isolate it and to assess it, to control all factors except this one.

It was then resolved to repeat the experiments a few times, while changing the order in which the ten speech samples were used in the learning process. If one display showed itself to be consistently better

than the other when the order of the sounds used for learning was changed a few times, then the view would be adopted that correct categorization was more easily done in one mode than in the other.

It would not have been useful to repeat the experiments more than a few times while changing the order of the learning samples. In a learning process involving ten sounds (five being of each category) there are only 25 possible hyperplanes that can be drawn at the center of and perpendicular to the line joining two sounds. If the experiment is repeated a few times with the sounds in different orders, it is certain that almost all the important hyperplanes have been tried, because there are speech sounds which will be classified without any difficulty and will not be used to draw the hyperplanes.

10.4 Randomness in the positioning method

As was pointed out earlier, there is every reason to suppose that the positioning method chosen is quite satisfactory. But slight positioning variations between the same speech samples in different modes cannot be avoided since one has different degrees of time resolution and since the filtering characteristics are different.

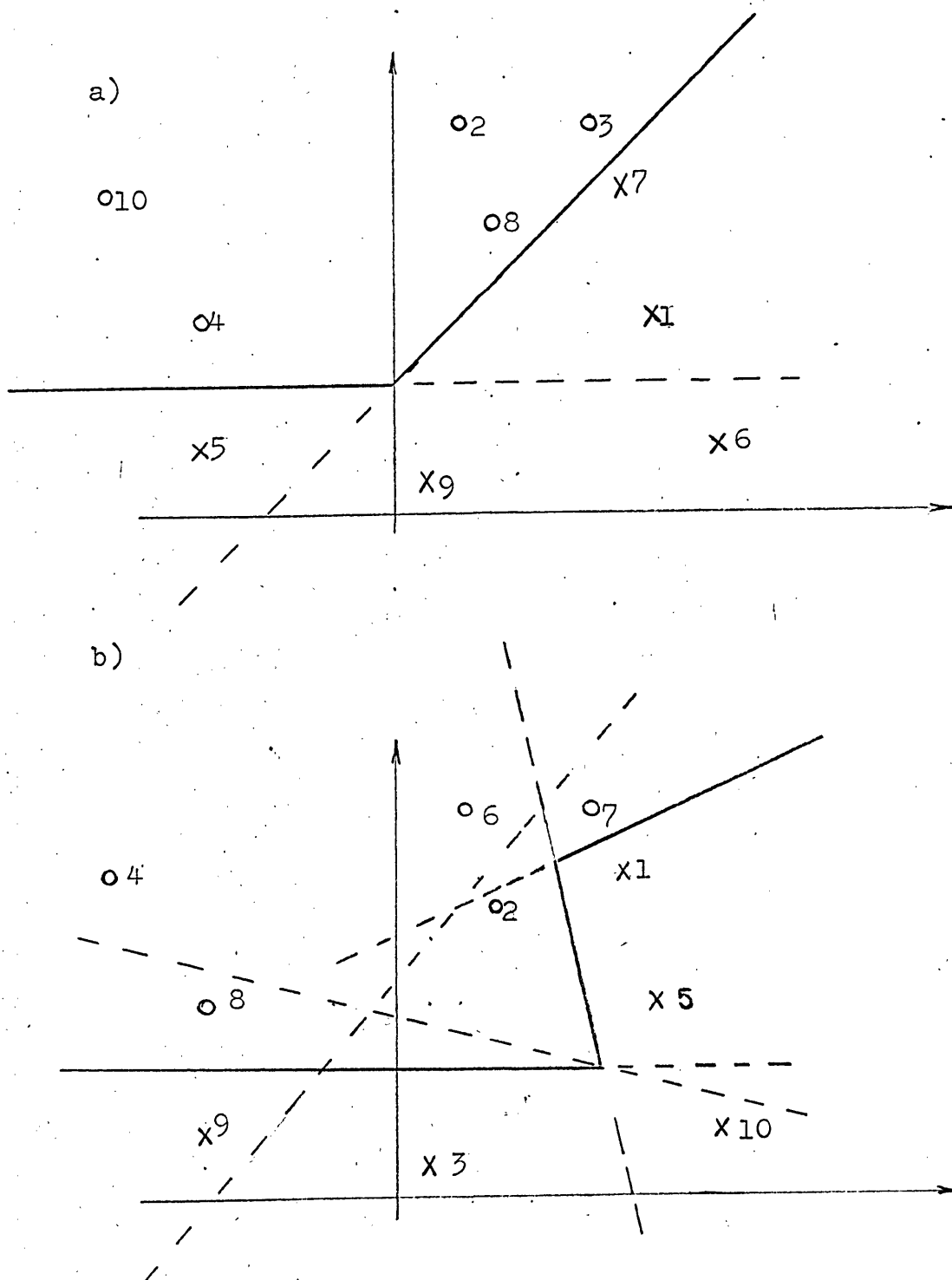


Fig. 10.01
 Changing the order of the learning samples
 changes the decision rule.
 o, x : categories
 ---- : hyperplanes
 — : decision rule

10.5 The standard test of significance

10.5.1 Introduction

In view of the randomness introduced in the experiments by various factors which have been pointed out, it becomes clear that the use of statistical significance tests is necessary. It can be hoped, however, that the results can be looked at with more confidence than these tests will indicate.

10.5.2 The absolute performance

One point, however, must be clarified before a normal test of significance based on a gaussian probability distribution is used: what would be the distribution of the results if the decision functions were random?

The experiment has been done using as coefficients of the hyperplanes pseudo-random numbers varying between + 10.0 and - 10.0, that is, in a range where the coefficients would normally vary: the hyperplanes were not then absolutely random.

It was decided to use the first twenty milliseconds of the samples of /p/ and /t/. For this case, the proportion of speech sound samples correctly classified by the decision functions had been, on average, 69% in the 200 c/s and 52% in the 100 c/s case. The distribution for the random coefficient hyper-

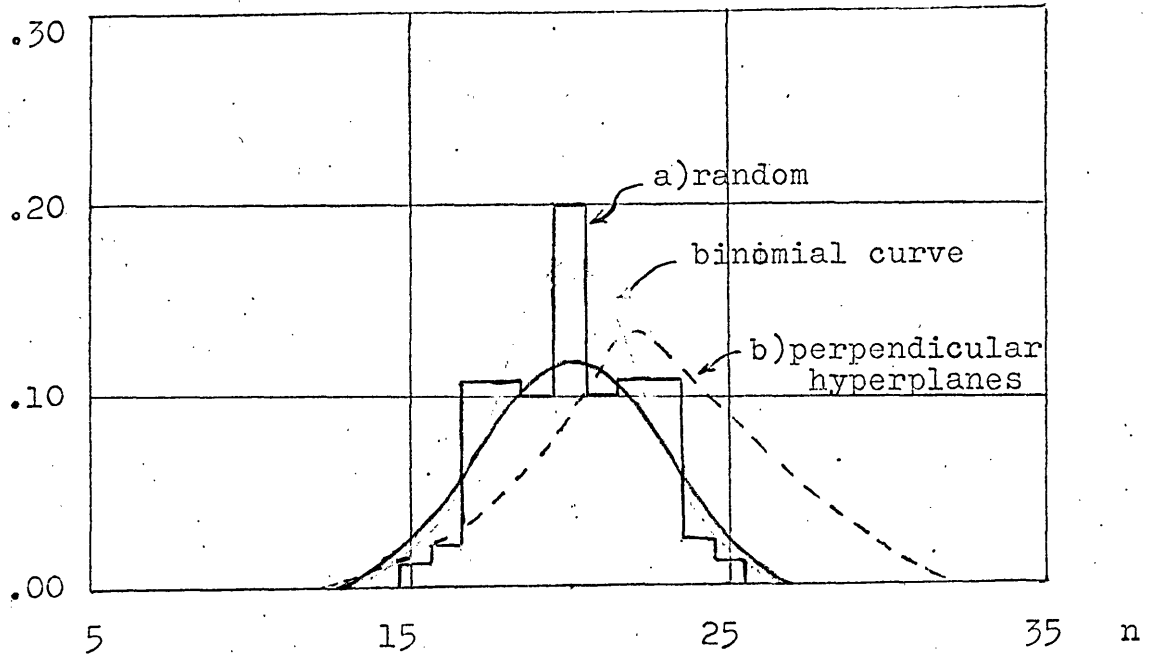


Fig. 10.02

Proportion of hyperplanes recognizing correctly
"n" sounds out of a total of forty

a) for hyperplanes with random coefficients :
(data taken from 30 hyperplanes)

b) for "perpendicular hyperplanes".

A binomial distribution is given as reference (n=40).

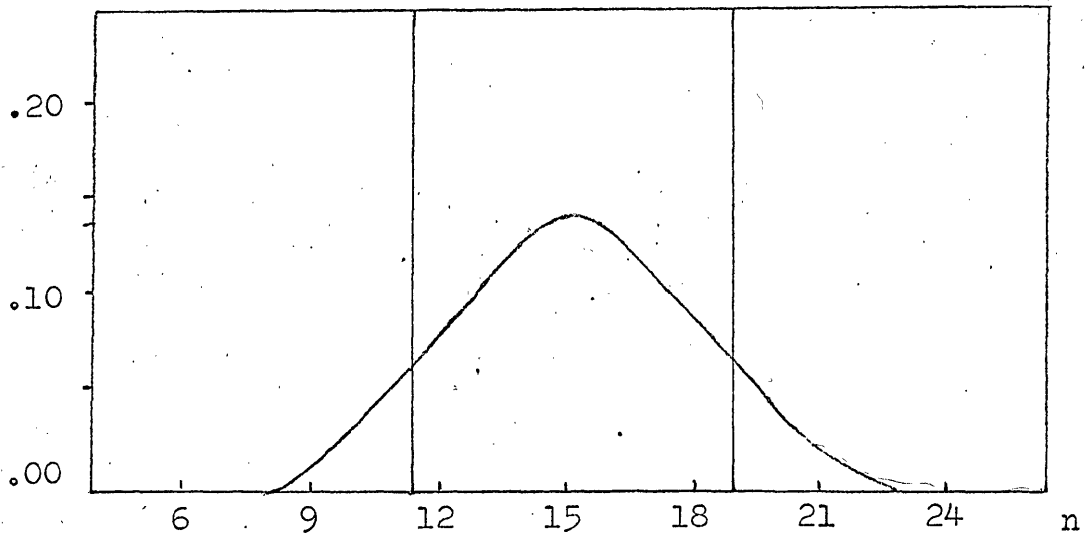


Fig. 10.03

A binomial distribution curve (n = 30)

planes was computed from 15 hyperplanes in the 100 c/s mode and 15 in the 200 c/s display. Since there was no noticeable difference between the results for the two displays, they were added in order to make up the distribution curve of figure 10.02. From this curve one can see, for example, that the probability of getting 70% correct recognition by chance out of forty samples is practically nil. This tells us moreover that the normal test of significance will permit us to evaluate the absolute recognition performance for each experiment and, to a certain extent, the differences of performance between parallel experiments.

However, instead of using around forty samples for testing as in the previous experiment, ten speech sound samples were used for learning and only around thirty were left for testing. Figure 10.03 is a binomial distribution for 30 samples. It can be seen then that, in absolute terms, there is less than 5% of the chances to have a success rate of 19/30 or more and 1% of the chances to have a success rate of more than 21/30.

10.5.3 The relative performance

When it came to decide whether there is a significant difference between two displays, it is not necessary for the performance rates to be outside the bounds of what can be got with a random experiment. In practice, however, it is expected that at least one of the

experiment	100 c/s mode	200 c/s mode	degree of significance
1	17/30	20/30	79%
2	15/30	20/30	91%
3	15/30	22/30	97%
average	15.6/30	20.6/30	91%
total	47/90	62/90	99%

Table 10.01

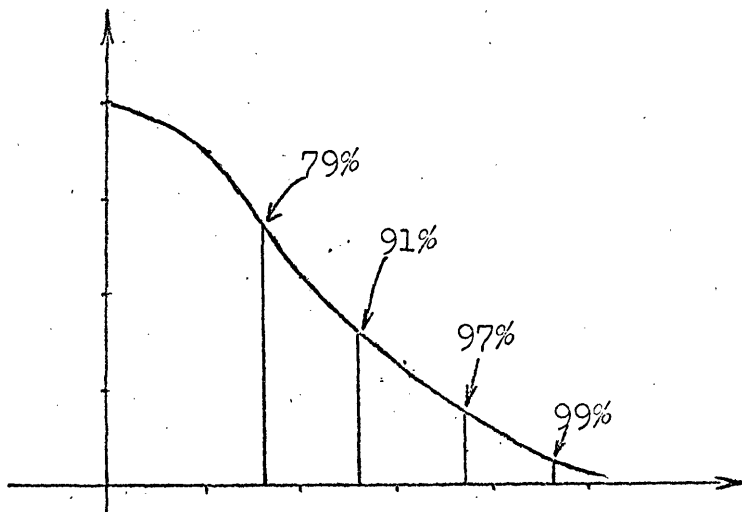


Fig. 10.04

two displays will show an unexpectedly good performance. On the basis of the ten samples available for learning, a decision function is drawn to classify the two sounds. This decision function is a prediction arrived at by following given rules. Since the rules are the same for the two modes, one can see whether the sounds let themselves be classified better in one display than in the other: if one display shows itself to be superior in a given case, it is then that the set of measurements is a better one. It could be that, on the basis of the learning samples, a decision function is arrived at which correctly classifies significantly less than 50% of the samples: this is still valid for our purpose even if the chances are that a random hyperplane would do better. As has been said earlier in this chapter, one must only be convinced that one knows the average improvement between one display and the other, that one does not have, for example, the worst possible prediction in one case and in the other case, the best possible one.

10.5.4 Testing the results for significance

Which results should be considered for the tests of significance? Suppose we are comparing two displays, that we have tested the decision functions with thirty samples and that we have twice changed the order of the samples used for learning. We have then three pairs of results, which we will assume to be those of table 10.01.

The hypothesis to be disproved is that there is no difference between the capacity of each display to arrive at a suitable decision function. We will assume, of course, that the distribution of the difference of performance for the two displays is normal and see what is the probability of obtaining the results according to this distribution. The probability of being right when rejecting the hypothesis is given in table 10.01 for each pair of experiments, for their average and their total; figure 10.04 shows the expected distribution of the differences in performance and where the results obtained fit on this curve.

Which one of these values should be chosen as being the one most likely to be true? Certainly, each case must be judged according to its own merits but, in the case of table 10.01, the fact that the repeated experiments always produced significant results pointing in the same direction adds to the significance of the results. Is it then justified in this case to take the sum of all the results? This could lead to abuses if the experiments were repeated too often since the experiments would be too similar to each other and the smallest difference could be made to become significant.

10.5.5 Conclusion

The results presented in the remainder of this thesis are strictly applicable to the speech sound samples

used in the experiments. In what measure are they applicable to the universe of speech sounds of the same category is a question which is left open. Nevertheless, the author has given in this chapter the reasons why he thinks that these results are more certain than the normal tests of significance would indicate and that they show the order of magnitude of the differences of performance which one should expect when varying the parameters of a spectrum analyser.

10.6 Repeating the experiments with new learning samples

In the experiments, five samples of each phoneme were used for learning, which meant a total of ten learning samples per experiment. A larger number of samples would have permitted to draw something approaching the optimum decision rule. One can wonder whether the small group of sounds selected for learning was really representative, and whether the same kind of results would have been obtained with a larger number of learning samples.

It was not practical to use a larger number of learning samples, but a new group of learning samples was chosen and some of the experiments, in which a difference of performance between the two displays had been found, were repeated. The results indicate, as will be seen in the next chapter, that it is reasonable to be 95% confident about the conclusions of this work.

Chapter 11

THE RESULTS

11.1 Introduction

The results for each pair of consonants with the 100 c/s and the 200 c/, will be presented first. They will be followed by an analysis of the outcome of some experiments done with vowel sounds.

The results of a few experiments performed with the 50 c/s mode will be given in appendix.

It will be noticed that the results do not conform to a nice and simple picture but that the outcome of the experiments varies as the characteristics of speech itself. All through this chapter, it will be attempted to establish a relation between the results and the characteristics of speech: the fact that such a relation can in most cases be established is one of the main arguments in favour of a high degree of significance for the results.

It will be noticed that the number of samples taken for testing sometimes varies, particularly in the first experiments. This is partly due to the fact that a smaller number of testing samples were used at first; in other cases, it was either because a part of the tape was misread by the computer or because the program discovered some fault in the data, following which either

the data tape was repaired or the program modified.

11.2 The p-b pair

The recognition performance for the experiments with the "p-b" pair can be found in table 11.01, with the appropriate degree of significance where it was computed. The figures indicate the number of sounds which were correctly classified out of a given total (for example: 21/31). This total differs somewhat between experiments with the 100 c/s and the 200 c/s displays because one or two sounds were disregarded by the computer as there appeared to be some irregularity in the data; furthermore, the initial experiments were done with a total of 24 sounds for testing but this number was increased to 31 as it became evident that the order of magnitude of the differences of performance between displays necessitated such action. In table 11.01, t_0 is the starting point of the consonant as defined in Chapter 9. Some experiments were repeated after having modified the order of the sounds selected for learning.

It can be seen from table 11.01 that, for this pair of consonants, the 200 c/s display shows itself to be vastly superior to the 100 c/s display at least for the first 40 msec. If one takes the samples characterizing the initial 40 msec, one sees that the performance of the 200 c/s was in one case 15⁰/o and in the

other case 25% superior to the performance of the 100 c/s mode; or, one could say that the results for the 200 c/s display are unlikely to be obtained by a random experiment while the 100 c/s display has not shown any capacity to classify sounds better than a random experiment. It can be noticed that the experiment with a 30 msec interval also points in the same direction.

As the results with the 200 c/s display seemed to deteriorate somewhat when the samples for the first 50 msec or the first 60 msec are considered and as there seemed to be some improvement with the 100 c/s mode, a further experiment was done considering only the samples going from 30 to 50 msec: the results were not conclusive except that they showed nothing else than randomness.

In addition to the differences in the recognition performance, Braverman's algorithm offers an opportunity to see how the classification of the sounds is arrived at by looking at the speech sounds which congregate in each region of Matrix A (refer to the chapter on Braverman's algorithm). A sound being classified in matrix A as, for example, 0010 or 1001 can be thought of as belonging to region 2 or region 9. It is interesting to see how speech sound groupings are influenced by the differences between analysing modes and between various portions of the speech sounds.

In the case of the "p-b" pair it is interesting to note that, in the experiments performed, the consonants followed by similar vowel sounds often tend to be classified in the same category when more than the initial 30 msec of the consonants are considered: the consonants followed by "a" and "o" will often be grouped together, as well as those followed by "et" and "it". In the case of the 100 c/s display, this tendency to separate consonants according to the vowel will go even further in that sense that the "it's" will often be separated from the "et's" and the "a's" from the "o's". Such a situation can be seen in table 11.02 which shows how the sounds were classified in experiment No 5 in which the data on the first 50 msec of the consonants was considered. Experiment No 5 can be compared to experiment No 1 where only the first 30 msec were considered (see table 11.03): it is apparent in this last case that the tendency of the consonants to group themselves according to the adjacent vowel is almost nonexistent.

These results can be correlated to the characteristics of the speech sounds under study. The loci of both the "b" and the "p" are in the low frequency region; the two sounds differ mainly by the establishment of a formant pattern in "b" earlier than in "p" which often consists of a burst of noise followed by a period of

Experiment No.	Time interval		Recognition Performance		Degree of significance of the differences
	from (msec)	to (msec)	100 c/s mode	200 c/s mode	
1	t_0	$t_0 + 30$	13/23	16/24	-
2	t_0	$t_0 + 40$	10/23	19/24	99.2%
3	"	"	14/29	21/31	96.6%
total	"	"	24/52	40/55	99.9%
4	t_0	$t_0 + 50$	17/23	17/24	-
5	"	"	14/29	20/31	-
6	$t_0 + 10$	$t_0 + 60$	15/23	15/24	-
7	$t_0 + 30$	$t_0 + 50$	14/25	15/25	-

Recognition performance for the "p-b" pair.
 (number of sounds correctly classified on a given total)
 t_0 : starting point of the consonants
 -: little or no statistical significance
 Table 11.01

a) 100 c/s

Region No.	pa	ba	po	bo	pet	bet	pit	bit	p	b	total
1	2	3	1			1			3	4	7
2	1	2							1	2	3
3		1								1	1
4		1	1						1	1	2
5		1	1	1	1				2	2	4
6			1	1					1	1	2
7					1				1		1
8					2	2	1		3	2	5
9						1				1	1
10	2				1		2		3		3
11							2	5	2	5	7

b) 200 c/s

Region No.	pa	ba	po	bo	pet	bet	pit	bit	p	b	total
1	6	3	2	2	4	2	4	1	16	8	24
2		4								4	4
3		1	1		1				2	1	3
4				1	1	2	1	4	2	7	9

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 5, Table 11.01, from t_0 to t_0+50 msec.)

Table 11.02

a) 100 c/s

Region No.	pa	ba	po	bo	pet	bet	pit	bit	p	b	total
1	1	2		1	2				3	3	6
2	1	1			1				2	1	3
3		2				1				3	3
4		1								1	1
5		1	2						2	1	3
6	1		1	1		1	2	2	4	4	8
7	1				1		1		3		3
8		1				1				2	2
9							1		1		1

b) 200 c/s

Region No.	pa	ba	po	bo	pet	bet	pit	bit	p	b	total
1	2	5	3	2	1	3	1	4	7	14	21
2		3								3	3
3	3				3	1	3		9	1	10

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 1, Table 11.01, from t_0 to $t_0 + 30$ msec.)

Table 11.03

relative emptiness. One would then expect that a better time resolution would permit to detect more easily the burst and the period of silence, and the results presented here demonstrate that the supplementary details which showed up in the 200 c/s display were found quite useful in the recognition process.

Considering table 11.02 it is hard not to conclude that the 100 c/s display was probably at a disadvantage because of its good frequency resolution which tended to pick up the formants and group the consonants according to the adjacent vowel. From table 11.03 it can be concluded that the formant structure is only starting to appear 30 msec after the start of the consonants while it is already predominant after 50 msec as shown in table 11.02; the results for the first 40 msec are similar to those for the first 50 msec.

11.3 The "t-d" pair

The 200 c/s mode also showed itself superior to the 100 c/s mode for distinguishing these two consonants. It can be seen in table 11.04 that the results for the 200 c/s display were always better than for the 100 c/s mode.

There are two cases where the differences in the recognition performance appear to be particularly significant. In one case, where the samples concerning the

first 40 msec of the two consonants are considered, differences of the order of 20⁰/o were noticed in favour of the 200 c/s mode. In the other case, where only the samples covering the interval going from 30 msec to 40 msec after the start of the sound was taken in consideration by the learning algorithm, differences of the order of 16⁰/o to 18⁰/o were obtained.

It is then clear enough that the distinction between the "t" and the "d" was more easily made in the experiments when a better time resolution was used. The indications are, however, that the advantages of having a better time resolution are small when only the first 30 msec of the consonants are considered but the experiment is not conclusive and would need to be repeated. There is no doubt that, if the whole consonant is taken into consideration, the 200 c/s display is much more likely to do better than the other one.

It was attempted as in the "p-b" case, to see whether the 100 c/s display causes consonants to be classified according to the adjacent vowel. It is interesting to notice that there is little evidence of such a trend not only in the results of experiment No 1 for the first 30 msec interval, but also for experiments No 2 and 3 (initial 40 msec). It is surprising to note that with the 200 c/s mode the consonants show a tendency to group themselves according to the adjacent

vowel in both experiments No 2 and 3 (initial 40 msec): the results of experiment No 3 are given in table 11.06 in support of these assertions. This difference in behaviour is probably not very significant, however, since experiment No 4 (initial 50 msec) and experiments No 5 and 6 (from $t_0 + 30$ msec to $t_0 + 50$ msec) show that the 100 c/s mode also tends to classify consonants according to the following vowel (see table 11.07).

To this must be added that, in the experiments carried out, the tendency to classify the consonants according to the adjacent vowel was much more pronounced in the case of "d" than for "t": the "t's" tended to be grouped together almost independently of the vowel while the "d's" showed a certain dependence on the following vowel. Support for this assertion can be found in tables 11.05, 11.06b, and 11.07, where it can be seen that the "t's" always tend to be classified in the same region while the "d's" mostly belong to regions characteristic of a type of vowel ("a" or "o" as opposed to "e" or "i"). Three of the four examples given here in support of this observation are experiments in the 200 c/s mode: the same tendencies can only occasionally be detected with the 100 c/s mode, since the recognition performance with that mode was so poor - or vice-versa.

In the border cases, however, the "t's" are certainly not very different from the "d's" since all the

"t's" which have been misclassified in the examples of tables 11.05, 11.06 and 11.07 have been classified with the "d's" having the same type of vowel ending.

As in the p-b case, it is possible to correlate these results with the known characteristics of the consonants. Since the loci of the "d" and the "t" are both at high frequency, the frequency values of the frequency transitions between either consonant and a vowel are not expected to be very different, and one would not think that the 100 c/s mode should perform better than the other; but since the frequency transitions between the "d" and the vowel are usually better defined than with consonant "t", it can be expected that these details will be better shown by the 200 c/s representation.

That the consonant "t" appears to be more independent of the adjacent vowel than "d" is also not surprising if one considers that the "d" falls in the category of the voiced consonants and that, while there is a period of randomness between the burst "t" and the start of the vowel, there are more pronounced frequency transitions between the burst of the "d" and the vowel; hence, the classification of the "d" is influenced by the nature of the following vowel whenever the period following the initial 30 msec of the consonant is taken in consideration.

Experiment No.	Time interval		Recognition Performance		Degree of significance of the differences
	from (msec)	to (msec)	100 c/s mode	200 c/s mode	
1	t_0	$t_0 + 30$	11/19	12/19	-
2	t_0	$t_0 + 40$	11/19	15/19	92%
3	"	"	16/27	24/30	95.5%
total	"	"	27/46	39/49	97.7%
4	t_0	$t_0 + 50$	10/19	13/19	-
5	$t_0 + 30$	$t_0 + 50$	16/26	24/30	94%
6	"	"	18/31	23/31	91%
total	"	"	34/57	47/61	98.3%

Recognition performance for the "t-d" pair.
(number of sounds correctly classified on a given total)

t_0 : starting point of the consonants

-: little or no statistical significance

Table 11.04

Region No.	da	ta	do	to	det	tet	dit	tit	d	t	total
1	2	6	1	2	1	5	2	5	6	18	24
2	1								1		1
3	3		2	1	2				7	1	8
4			1				4	1	5	1	6

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.

(Exp. No. 6, Table 11.04, 200 c/s mode, from $t_0 + 30$ msec to $t_0 + 50$ msec).

Table 11.05

a) 100 c/s

Region No.	da	ta	do	to	det	tet	dit	tit	d	t	total
1	3	4		2		3	3	3	6	12	18
2	3		3	1	4	2	2	2	12	5	17
3		2								2	2
4				1						1	1

b) 200 c/s

Region No.	da	ta	do	to	det	tet	dit	tit	d	t	total
1	1	6		3	1	5		3	2	17	19
2	3		4	1					7	1	8
3	2				3	1	3	2	8	3	11
4							3		3		3

Sounds classified according to the region in which they fall in Braverman's algorithm.
(Exp. No. 3, Table 11.04, from t_0 to t_0+40 msec)

Table 11.06

a) 100 c/s

Region No.	da	ta	do	to	det	tet	dit	tit	d	t	total
1	2	4		2		3	2	2	4	11	15
2	1	1							1	1	2
3	1		3	1					4	1	5
4					3	1	2	2	5	3	8

b) 200 c/s

Region No.	da	ta	do	to	det	tet	dit	tit	d	t	total
1	1	5	1	2		3	1	2	2	12	14
2	2		2						4		4
3	1								1		1
4			1	1					1	1	2
5					3	1	1	1	4	2	6
6							2	1	2	1	3

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 4, Table 11.04, from t_0 to $t_0+50\text{msec}$)

Table 11.07

11.4 The "p-t" pair

In the two pairs of sounds previously studied, "p-b" and "d-t", it was known that for each pair the burst of noise characterizing the consonants was in the same region of the frequency spectrum and that the locus, from which the consonant-vowel frequency transitions start, was in approximately the same position: there was then little reason to expect that the 100 c/s display would do any better than the 200 c/s display.

With the "p-t" pair, we are trying to distinguish consonants which fall in different parts of the frequency spectrum and it may be that the transition from each consonant to the vowels can serve as a basis for recognition.

Table 11.08 shows how the recognition performance varied with the portion of the consonants considered and with each analysing mode. The only difference which stands out clearly is that when only the first 20 msec of the two consonants were considered, results 10⁰/o to 25⁰/o better were obtained with the 200 c/s mode: it would then seem that there is more in recognizing "p" from "t" than picking up energy differences between low and high frequencies.

The results of experiments 4, 5, 6, 7 appear to indicate that the two modes are pretty well equivalent when the samples for the initial 40 msec of the conson-

ants are used.

Experiments 8, 9, 10 which use the data on what should be the transition interval between consonant and vowel (from $t_0 + 20$ msec to $t_0 + 60$ msec) as well as experiment 12 (from $t_0 + 40$ msec to $t_0 + 60$ msec), suggest that some advantage might be gained by using the 100 c/s mode although the results are not really significant. The recognition performance is poor however, being on average hardly better than what one would expect from a random decision rule. Either the transitions between consonants and vowels do not appear clearly enough to be of much use for recognizing consonants, which would be in accord with the characteristics of these consonants, or the number of learning samples is not high enough.

Careful study of the results shows that the nature of the vowel already influences to some extent the classification of the consonant when only the first 20 msec of the consonants are considered and that the tendency accentuates itself for the interval after 20 msec. It can be seen in table 11.09 where the results of experiment No 2 (initial 20 msec) are broken down, that some consonants followed by "a" or "o" tend to be grouped together as well as some consonants followed by "et" or "it". Table 11.10, where the results of experiment No 8 (from $t_0 + 20$ msec to $t_0 + 60$ msec) are recorded,

Experiment No.	Time interval		Recognition Performance		Degree of significance of the differences
	from (msec)	to (msec)	100 c/s mode	200 c/s mode	
1	t_o	$t_o + 20$	17/30	20/30	78.5%
2	"	"	15/30	20/30	90.5%
3	"	"	15/30	22/30	97%
total	"	"	47/90	62/90	99.95%
4	t_o	$t_o + 40$	20/29	18/29	-
5	"	"	17/30	19/30	-
6	"	"	18/28	17/28	-
7	"	"	22/29	20/30	-
total	"	"	77/119	74/117	-
8	$t_o + 20$	$t_o + 60$	21/30	18/30	79.5%
9	$t_o + 20$	"	15/29	13/30	-
10	"	"	16/30	13/30	78 %
total	"	"	52/89	44/90	90 %
11	$t_o + 20$	$t_o + 40$	15/29	16/30	-
12	$t_o + 40$	$t_o + 80$	15/30	13/30	-

Recognition performance for the "p-t" pair.
(number of sounds correctly classified on a given total)

t_o : starting point of the consonants

-: little or no statistical significance

Table 11.08

a) 100 c/s

Region No.	pa	ta	po	to	pet	tet	pit	tit	p	t	total
1	1	2	1	1	1		2	2	5	5	10
2	1	1				1	1	2	2	4	6
3	1	1	1				1		3	1	4
4	2		1						3		3
5			1	1					1	1	2
6				1						1	1
7		2		1	4	5		2	4	10	14
8							1		1		1

b) 200 c/s

Region No.	pa	ta	po	to	pet	tet	pit	tit	p	t	total
1	1	2	1	1	4	1	3	1	9	5	14
2	1	3	1		1	5		3	3	11	14
3	2								2		2
4		1		1						2	2
5			1						1		1
6				1						1	1
7	1			1	1		2	1	4	2	6

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 2, table 11.08, from t_0 to $t_0 + 20$ msec)

Table 11.09

a) 100 c/s

Region No.	pa	ta	po	to	pet	tet	pit	tit	p	t	total
1	1	5		1	2	5	1	1	4	12	16
2	1		1		1		1		4		4
3	2	1					2		4	1	5
4			1	1	1		1	2	4	3	7
5				1	2		1	1	3	2	5
6	1							2	1	2	3

b) 200 c/s

Region No.	pa	ta	po	to	pet	tet	pit	tit	p	t	total
1		2		1	1			1	1	4	5
2		1		1		1		1		5	5
3		1								1	1
4			2	1	1		1		3	2	5
5	1								1		1
6	2	1	1						3	1	4
7	2	1					1	1	3	2	5
8			1		2		1		4		4
9					1	1			1	1	2
10						4		1		5	5
11							2		2		2
12								1		1	1

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 8, Table 11.08, from t_0+20 to t_0+60 msec)

Table 11.10

shows that this tendency becomes more pronounced when the latter portion of the consonant is used.

As a conclusion it can be said that the first 20 msec of the consonants "p" and "t" appear to offer the best basis for distinguishing between them, that the formants of the following vowel certainly appear fairly well but not really the transitions from consonants to vowel.

11.5 The "d-b" pair

The number of correct classifications obtained when using the algorithm for different portions of these consonants and with the learning samples in different order can be found in table 11.11.

It is clear from these results that the most significant difference between the two displays is the superior performance of the algorithm with the 100 c/s mode when the latter part of the consonants is considered. There are good indications that, if only the first 20 msec of the consonants are considered, the 100 c/s mode will give slightly better results, but this does not seem to be of practical importance since the performance in the best case is not much higher than what one would expect from random experiments. Furthermore, if one takes the samples for the first thirty or the first **forty** milliseconds of the consonants, one obtains a

nearly equivalent performance for the two displays and generally a substantially higher reliability of recognition.

The reasons of the superiority of the 100 c/s display for the time interval going from 20 msec to 60 msec after the start of the consonant are difficult to infer from the results: it is striking however that a much greater number of hyperplanes were drawn during the learning process than in the experiments reported previously in this chapter: in experiments No 9, 10 and 11, it was found necessary to draw six hyperplanes three times, seven hyperplanes twice and five hyperplanes once. This is equivalent to forming respectively up to 64, 128 and 32 areas in the Braverman's algorithm, many of which are quite insignificant, and it is then more instructive to see how the sounds have been classified after simplification of these irrelevant boundaries. This has been done in table 11.12 for the results of experiment No 10 and shows quite a strong tendency for consonants followed by a similar ending to be grouped together. Somehow, the 100 c/s display seems to do much better. One could suggest that the 200 c/s display is less able to separate consonants followed by the same vowel than the 100 c/s display, where the frequency transitions are better defined.

It is interesting to compare the results for the

Experiment No.	Time interval		Recognition performance		Degree of significance of the differences
	from (msec)	to (msec)	100 c/s mode	200 c/s mode	
1	t_0	$t_0 + 20$	19/31	19/31	-
2	"	"	18/31	15/31	78 %
3	"	"	18/31	12/31	94 %
total	"	"	55/93	46/93	95 %
4	t_0	$t_0 + 30$	19/28	21/31	-
5	"	"	15/31	16/31	-
total	"	"	34/59	37/62	-
6	t_0	$t_0 + 40$	17/25	20/30	-
7	"	"	21/31	23/31	-
total	"	"	38/56	43/61	-
8	$t_0 + 30$	$t_0 + 50$	21/31	16/28	-
9	$t_0 + 20$	$t_0 + 60$	23/30	15/30	98.5%
10	"	"	22/30	17/30	91 %
11	"	"	23/30	20/30	80 %
total	"	"	68/90	52/90	99.5%

Recognition performance for the "d-b" pair.
(number of sounds correctly classified on a given total)
 t_0 : starting point of the consonant.

Table 11.11

-: little or no statistical significance

a) 100 c/s

Region No.	da	ba	do	bo	det	bet	dit	bit	d	b	total
1	1	4		1		3			1	8	9
2	1	1	2		2	1	1		6	2	8
3	4		2	2	2			1	8	3	11
4		4								4	4
5						1	1	1	1	2	3
6							4	1	4	1	5
7							1	3	1	3	4

b) 200 c/s

Region No.	da	ba	do	bo	det	bet	dit	bit	d	b	total
1		5	1	1	1		2		4	6	10
2		2				1				3	3
3	1	1							1	1	2
4	2		1	1					3	1	4
5	1		1						2		2
6				1	1			1	1	2	3
7	1		1	1	1	3	3		6	4	10
8							2	1	2	1	3

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 10, Table 11.11, from t_0+20 to t_0+60 msec)

Table 11.12

"p-t" pair with those for the "d-b" pair. In the former case, one had observed in the experiments quite a significant improvement in recognition for the first 20 msec of the consonants when the 200 c/s display was used instead of the 100 c/s mode, while there were only very slight signs of a possible improvement when the 100 c/s mode was used instead of the 200 c/s mode for the 20 to 60 msec interval. In the "d-b" case, however, the 100 c/s mode appears to be substantially better for the 20 to 60 msec interval and there are signs that it could also be better for the initial 20 msec of the consonants: one can then see that although one of the main differences between the two pairs of sounds is that "d" and "b" are usually voiced while "t" and "p" are usually unvoiced, the whole structure of these sounds is dissimilar, the frequency aspect being much better defined in the case of the "d-b" pair.

11.6 The "p-d" pair

As can be seen from table 11.13, the results obtained in recognition are not very exciting by the differences they show between the two analysing modes. On the whole, there are faint indications that the 200 c/s display might be better for analysing the start of the consonants - the first twenty or forty milliseconds; the results also do not exclude the possibility that some advantage would be gained by using the 100 c/s

representation for the second part of these two consonants - the period from twenty to sixty milliseconds after the start of the consonants. But these are only indications and the results themselves show little difference.

The study of the way consonants are classified indicates once again that there is very little tendency for the consonants followed by the same type of vowel to be grouped together when only the data on the first 20 msec of the consonants is used (table 11.14); such a tendency becomes, however, very pronounced in the experiments concerned with the 20 to 60 msec interval (see, for example, table 11.15). The results of experiment No 4 for the first 40 milliseconds of the consonants show that the 100 c/s display might have a more pronounced tendency to classify consonants according to the adjacent vowel than the 200 c/s mode (table 11.16).

On the whole, then, the results for this pair of sounds indicate that the two analysing modes studied are probably almost equivalent, without excluding the possibility that the 200 c/s mode might be more appropriate for the start of the consonant and the 100 c/s mode more appropriate for the remaining part of the consonants.

However, because "p" and "d" are characterized by different frequency loci and because "p" is normally

Experiment No.	Time interval		Recognition performance		Degree of significance of the differences
	from (msec)	to (msec)	100 c/s mode	200 c/s mode	
1	t_0	$t_0 + 20$	18/29	18/29	-
2	"	"	18/29	21/29	-
total	"	"	36/58	39/57	-
3	t_0	$t_0 + 40$	20/29	22/29	-
4	"	"	18/29	18/29	-
total	"	"	38/58	40/58	-
5	$t_0 + 20$	$t_0 + 60$	21/29	18/29	-
6	"	"	20/29	19/29	-
total	"	"	41/58	37/58	-

Recognition performance for the "p-d" pair.
(number of sounds correctly classified on a given total)

t_0 : starting point of the consonant

-: little or no statistical significance

Table 11.13

Region No.	pa	da	po	do	pet	det	pit	dit	p	d	total
1	2	5	3	2	2	2		3	7	12	19
2	1	1		2	1	2	2	3	4	8	12
3	2		1		2		1		6		6
4							1		1		1
5							1		1		1

Sounds classified according to the region in which they fall in the modified Braverman's algorithm (Exp. No. 1, 100 c/s mode, Table 11.13, from t_0 to t_0+20)

Table 11.14

Region No.	pa	da	po	do	pet	det	pit	dit	p	d	total
1	2	1	1		2		3		8	1	9
2		2	3	2		1			3	5	8
3	1								1		1
4	1								1		1
5	1	1			1				2	1	3
6				1						1	1
7					1				1		1
8		2		1	1	3	2	5	3	11	14

Sounds classified according to the region in which they fall in the modified Braverman's algorithm. (Exp. No. 5, 100 c/s mode, Table 11.13, from $t_0 + 20$ msec to $t_0 + 60$ msec).

Table 11.15

a) 100 c/s

Region No.	pa	da	po	do	pet	det	pit	dit	p	d	total
1	3	3	1	1	4	1	3		11	5	16
2		1								1	1
3	2	1							3		3
4		2	2	2					2	4	6
5				1						1	1
6					1	3	2	6	3	9	12

b) 200 c/s

Region No.	pa	da	po	do	pet	det	pit	dit	p	d	total
1		2		1		3		2		8	8
2	2				2		1	1	5	1	6
3	3		1		1	1	1		6	1	7
4		2	1	1			2		3	3	6
5	1										1
6	1	1		1	1				2	2	4
7			1	1					1	1	2
8					1				1		1
9								1		1	1
10							1		1		1
11								1		1	1
12	1	1		1	1				2	2	4

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 4, Table 11.13, from t_0 to $t_0 + 40$ msec)

Table 11.16

unvoiced while "d" is normally voiced, one would have expected a better reliability of recognition than was in fact obtained in the experiments.

11.7 The "b-t" pair

The main feature of the recognition results for the experiments done with this pair of consonants (table 11.07) is that they are often much better than the results obtained for the other pairs of consonants. It is not surprising that these two consonants are easy to separate since their loci are in different parts of the spectrum and since one consonant is usually voiced while the other is usually not, one can wonder why the results were not as good in the case of the "p-d" pair, where the locus of each consonant is also in a different part of the spectrum and where the "p" is normally unvoiced while the "d" is normally not.

The 100 c/s mode appears to be slightly superior for the start of the consonant - the initial 20 msec: in one of the experiments it was much better but it would seem that, on average, the advantage of using the 100 c/s display would be marginal. When the data on the initial 40 msec of the consonants is considered, high recognition performances are obtained and the 200 c/s mode would appear to be, on average, somewhat better. However, the difference which seems to be the most significant

Experiment No.	Time interval		Recognition performance		Degree of significance of the differences
	from (msec)	to (msec)	100 c/s mode	200 c/s mode	
1	t_0	$t_0 + 20$	27/32	21/32	96 %
2	"	"	27/32	26/32	-
3	"	"	25/32	24/32	-
total	"	"	79/96	71/96	92 %
4	t_0	$t_0 + 40$	27/32	29/32	-
5	"	"	27/32	28/32	-
6	"	"	(26/32)		
total (without 6)	"	"	54/64	57/64	78 %
7	$t_0 + 20$	$t_0 + 40$	23/32	20/32	80 %
8	"	"	24/32	19/32	91 %
total	"	"	47/64	39/64	93 %

Recognition performance for the "b-t" pair
(number of sounds correctly classified on a given total)

t_0 : starting point of the consonants

- : little or no statistical significance

Table 11.17

a) 100 c/s

Region No.	ba	ta	bo	to	bet	tét	bit	tit	b	t	total
1		5		1		6		3		15	15
2	6		1	1	2		1	2	10	3	13
3	2		1		3		3		9		9
4		1	1				1		2	1	3
5				1						1	1
6				1						1	1

b) 200 c/s

Region No.	ba	ta	bo	to	bet	tét	bit	tit	b	t	total
1	1	6	1	3	2	6	1	5	5	20	25
2	2		1		1		3		7		7
3	3		1	1	1				5	1	6
4	2		1	1	1		1		4		4

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 2, Table 11.17, from t_0 to $t_0 + 40$ msec)

Table 11.18

a) 100 c/s

Region No.	ba	ta	bo	to	bet	tet	bit	tit	b	t	to-tal
1		6	1	2	1	6		3	2	17	19
2	2								2		2
3	1								1		1
4	4								4		4
5	1		2	1					3	1	4
6				1						1	1
7					1		1	2	2	2	4
8					1				1		1
9					2		4		6		6

b) 200 c/s

Region No.	ba	ta	bo	to	bet	tet	bit	tit	b	t	to-tal
1		4	1	2		5		3	1	14	15
2	4		1				2	1	7	1	8
3		2		1		1		1		5	5
4	3								3		3
5			1	1	1				2	1	3
6					1				1		1
7	1				3		3		7		7

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 4, Table 11.17, from t_0 to t_0+40 msec).

Table 11.19

a) 100 c/s

Region No.	ba	ta	bo	to	bet	tet	bit	tit	b	t	total
1		6	1	2		5		3	1	16	17
2	1								1		1
3	1								1		1
4	6				1						7
5					4	1	5	2	9	3	12
6			2	2					2	2	4

b) 200 c/s

Region No.	ba	ta	bo	to	bet	tet	bit	tit	b	t	total
1	1	6	1	3	1	6		5	3	20	23
2	3		2	1	1		1		7	1	8
3	4								4		4
4					3		4		7		7

Sounds classified according to the region in which they fall in the modified Braverman's algorithm.
(Exp. No. 5, Table 11.17, from t_0 to $t_0 + 40$ msec)

Table 11.20

is that the 100 c/s mode gives better results when the second part of the consonant is taken into account, that is, in the time interval where one would look for the frequency transitions between consonants and vowels.

There is little correlation between the way in which the consonants are grouped together and the vowel accompanying the consonant as long as only the data on the initial 20 msec is taken (see, for example, table 11.18). The results for the first 40 msec, however, show very clearly a tendency which has already been pointed out with the "d-t" pair: the "t's" all tend to be classified in different groups, this grouping being often determined by the nature of the vowel accompanying the consonant. A look at the results for the "p-t" pair will show that the "t's" tend, also in that case, to be grouped together.

11.8 The vowels

It was written earlier on in this thesis, when commenting on photographs of spectrograms in the 100 c/s and the 200 c/s modes, that given a sufficiently large dictionary of possible vowel sounds, the 100 c/s representation was bound to prove itself superior. The experiments done have confirmed this expectation.

As it was not certain that significant enough differences in recognition performance would show up if the vowels were taken two by two, a program was pre-

Experiment No.	Sounds	Recognition Performance		Degree of Significance
		100 c/s	200 c/s	
I	et-it	25/31	20/31	92.5%
2	"	25/31	21/31	88.5%
total	"	50/62	41/62	96.5%
3	a-o	22/30	23/30	-
4	"	24/31	20/31	87%
5	"	29/31	26/31	94%
total	"	75/92	69/92	86%
total	et, it, a, o	125/154	110/154	98.5%

Recognition performance for the vowels.

(number of times correctly recognized on a given total)

- : little or no statistical significance
table 11.21

pared in order to do experiments with the four vowel sounds at the same time; however, the limitations in computing time did not permit to carry this program through and, finally, experiments were performed with the vowels taken two by two.

The results of these experiments can be found in table 11.21; the data covering the time interval between 40 msec and 80 msec was considered. In the experiments concerned with separating "et" and "it", a greater difference in favour of the 100 c/s display was found than in trying to separate "a" and "o"; this can probably be explained by the fact that "a" and "o" have much more stable frequency formants than "et" and "it", whose formants vary considerably in frequency between the initial and final consonants. The results can then be summarized as follows: in a simple case, the 100 c/s mode was only slightly superior to the 200 c/s mode, but in a more difficult case which was probably more similar to the type of problem one would encounter in continuous speech, quite a healthy difference of performance in favour of the 100 c/s mode was observed.

11.9 The significance of the results

In order to assess the significance of the results, a series of experiments was repeated with new learning samples. These new samples were taken from the group which has been used previously for testing and the old

learning samples were now used for testing. It can be assumed that this is almost the same as repeating the experiments with totally new samples.

The experiments which were repeated were those which had showed a difference of performance in favour of one analysing mode corresponding to a degree of significance not lower than 90^o/o.

If one should be 90^o/o confident that the general conclusions of this work are correct, one should expect a contradictory result in one case out of ten and, if one should be 95^o/o confident, in one case out of twenty.

Let us suppose that, for a given pair of consonants and a given time interval, the average recognition performance previously obtained had been 75^o/o in the 200 c/s mode and 65^o/o in the other mode. If one can be 90^o/o confident that the 200 c/s mode is superior to the other, one would intuitively expect, in one case out of ten, the performance with the 100 c/s mode to be higher than 75^o/o or that of the 200 c/s mode to be lower than 65^o/o.

Sixteen such experiments were done and, as can be seen from table 11.22, there was never any contradiction; in one case (experiment No 6 of table 11.22), the result was 0.5^o/o over the line but it can be considered equal since the experiment was carried out with 31 samples, which means that only differences in the

wrong direction in excess of 1.5°/o could be termed a contradiction.

It is then reasonable to be around 95°/o confident about the conclusions of this work.

Exp. No.	Time interval		100 or 200 c/s mode should give results higher or lower than average for other mode	Recognition performance		Average for other mode
	from (msec)	to (msec)			%	
1	t_o+30	t_o+50	d-t, 100 c/s, lower	21/31	68	77%
2	t_o+30	t_o+50	d-t, 200 c/s, higher	22/31	71	60%
3	"	"	d-t, 100 c/s, lower	18/31	58	77%
4	"	"	d-t, 200 c/s, higher	24/31	77.5	60%
5	t_o	t_o+40	d-t, 100 c/s, lower	17/31	55	80%
6	"	"	d-t, 200 c/s, higher	18/31	58	58.5%
7	"	"	p-b, 100 c/s, lower	17/30	56	73%
8	"	"	p-b, 200 c/s, higher	20/30	66	46%
9	t_o	t_o+20	d-b, 100 c/s, higher	20/31	65	49.5%
10	"	"	d-b, 200 c/s, lower	14/31	45	59.5%
11	t_o+20	t_o+60	d-b, 100 c/s, higher	21/31	68	58%
12	"	"	d-b, 200 c/s, lower	17/31	55	75%
13	t_o	t_o+20	p-t, 100 c/s, lower	17/30	56.5	69%
14	"	"	p-t, 200 c/s, higher	18/30	60	52.5%
15	t_o+20	t_o+60	b-t, 100 c/s, higher	20/32	62.5	61%
16	"	"	b-t, 200 c/s, lower	18/32	56	73.5%

Repeating some experiments
with new learning samples

Table 11.22

Chapter 12

CONCLUSION

12.1 Discussion of the results

Although the results presented in the previous chapter do not form a nice and simple pattern, it is still possible to draw some general conclusions.

One such conclusion which should be drawn is that the 100 c/s mode has shown itself superior to the 200 c/s mode for analysing vowel sounds and for classifying consonants with the help of the frequency transitions between consonants and vowels. In other words, the 100 c/s representation has shown itself more capable of classifying sounds correctly in these cases where, according to our knowledge of speech characteristics, sounds are characterized and should be differentiated by their resonance pattern.

A breakdown of the results of the experiments with consonants can be found in table 12.01, where it is indicated which analysing mode proved superior for each pair of consonants and for three different time portions of the stop-consonants. The same results are repeated in table 12.02 for the experiments concerned with the initial part of the consonant (the first twenty or first forty msec), but considering only the time interval which

	approximate time interval		
	from t_o to $t_o+20\text{msec}$	from t_o to $t_o+40\text{msec}$	from t_o+20 to $t_o+40\text{msec}$
p-b		200 c/s	equal
d-t		200 c/s	200 c/s
p-t	200 c/s	equal	100 c/s (?)
d-b	100 c/s (?)	equal	100 c/s
p-d	equal	equal	equal
b-t	100 c/s (?)	200 c/s (?)	100 c/s

Break-down of the results of chapter 11, giving the analysing mode which was found the best for different pairs of stop-consonants and different portions of these consonants. (t_o : starting point of each consonant)

Table 12.01

	approximate time interval	
	from t_o to t_o+20 msec	from t_o to t_o+40 msec
p-b		200 c/s
d-t		200 c/s
p-t	200 c/s	
d-b		equal
p-d	equal	
b-t		200 c/s (?)

Break-down of the results of chapter 11 for the initial portions of the consonants (the first 20 or 40 msec), but giving the results only for the time interval which gave the best recognition performance.

Table 12.02

gave the best recognition performance.

It can be seen from table 12.02 that the 200 c/s mode has, on the whole, shown itself superior to the 100 c/s mode when considering the start of stop-consonants and, in particular, when trying to distinguish between two stop-consonants whose loci were in the same frequency region. For the initial part of the six pairs of stop-consonants studied, the 200 c/s mode proved to be significantly superior to the 100 c/s mode for three pairs (p-b, d-t, p-t), and both displays seemed to be nearly equivalent in two other cases (p-d, b-t), and the 200 c/s display appeared somewhat better for the other pairs (d-b). In general, it can be said that the 200 c/s mode was at an advantage when considering the initial part of the stop-consonants although it might not be true in every case.

12.2 Adaptive spectral analysis

The fact that one cannot conclude that the 200 c/s display performs always better than the 100 c/s mode for the start of stop-consonants or that the 100 c/s display is always better for the latter part of the consonant makes it very difficult to implement an automatic adaptive spectral analyser. If it had been possible to conclude, for example, that one has always advantage to analyse the initial 40 msec of stop-consonants in the 200 c/s

mode, it would have been possible to think about a segmenter capable of detecting the presence of a stop-consonant. As it turns out, the situation is much more complex and the task of building a segmenter capable of isolating the correct portions of the speech sounds appears too great.

What appears possible, however, and quite useful in view of the differences observed in recognition performance in different modes, is to use in parallel a few spectrum analysers with different filtering characteristics. Let us suppose that a preliminary investigation has shown that the sound detected was a vowel: the data from the 50c/s or the 100c/s analyser would be chosen in order to take the final decision as to the exact nature of this vowel. Let us suppose that investigation had shown that it was either consonant "t" or consonant "d": the final decision would be taken using the data supplied by the 200c/s analyser. The sounds would then be classified by using a sequential decision process, choosing at every stage either the method which would have shown itself the most reliable or a weighted combination of the most reliable methods and processes.

In many of the experiments, the choice of the inappropriate analysing mode meant recognition performances which were not much better than what one would have expected from a random decision rule.

The choice of the appropriate analysing mode often meant a reliability of recognition which was significantly better than one would have expected from random experiments; an array of spectral analysers with different filtering characteristics is then a useful possibility.

12.3 The recognition performance

In many of the experiments reported in the previous chapter, recognition performances were got which were not much better than what would have been expected from random experiments.

Braverman himself had a fair amount of success in his experiments concerned with the recognition of numerals 0, 1, 2, 3 and 5. 796 numerals were prepared, 202 of which were chosen for learning; when an iterative method was used in order to bring each separating hyperplane to pass closer to the middle of the line separating the two points between which it is drawn, reliabilities of recognition varying between 79 and 89% were obtained; the longer the iterative process, the better were the results. With the method of paralleled variants, the reliability of recognition went up to 98.5%.

The relatively poor recognition performances can probably be explained by the following reasons.

First of all, the level of difficulty present in the experiments was very high: the vowels which had been

selected for study were hard to separate and the stop-consonants are easy to confuse, particularly if only a portion of the sound is taken into consideration. In some cases, however, when the level of difficulty was not so high, the recognition performance was higher as, for example, when trying to classify "b" and "t" (table 11.17). In order to check that the results would improve as the level of difficulty diminished, an experiment was done in which the algorithm tried to distinguish the sounds in "a" or "o" from those in "et" or "it": recognition performances of 22/25 were obtained in both the 100c/s and the 200c/s modes.

Secondly, the pattern recognition experiments were not carried out in optimum conditions, in that the number of dimensions was much higher than the number of learning samples. It is well known in pattern recognition that, for optimum results, the number of ~~learning samples~~ ^{learning samples} ~~dimensions~~ must be about two times larger than the number of dimensions (Allais). The number of dimensions for an interval of 20msec being around 64, and 128 for 40msec, one should have had some 200 learning samples for the experiments.

12.4 Suggestions for further work

12.4.1 Completing the present study

Obviously, much needs to be done before such a decision system as proposed above can be designed.

The work presented in this thesis has been very limited in scope: it has been limited, practically, to two analysing modes and four stop consonants in conjunction with a few vowels. New analysing modes, new sounds need to be explored, the effect of redundancy must be studied. It is also necessary to take sounds not only by pairs, but in groups of a larger size.

Eventually, it will be possible to devise an optimum method for classifying and recognizing speech sounds.

12.4.2 Different pattern recognition algorithms

We used in this work an hyperplane method; it can be argued that the hyperplane method is only an approximation of a decision process based on the measure of the probability distribution of the samples along each sampling axis. As was pointed out earlier, because of the few samples used for learning, a decision process based on the measurement of probability distributions might not have been very good. Nevertheless, the hyperplane method did not give results which were so good as to make us forget that there are other possible ways of approach.

In addition to this, Braverman's algorithm has the disadvantage of not making use of all the available information: if one sound falls in a region of correct affiliation, this sound is not used to optimize the

decision function or to modify it in any way: the decision function stops improving then as soon as all the learning samples have been correctly classified. A method based on the measurement of probability distributions would then be more appealing to the mind provided there are sufficient learning samples to render this measurement of probability distributions meaningful.

12.4.3 Analysis and synthesis

Another possible avenue for research would be to synthesize speech analysed by the Imperial College spectrum analyser, with the analyser in different modes and assess the results according to the quality of the synthesized speech. It would be possible in this way, with the help of articulation tests, to see in which condition different sounds can be best classified by listeners. In view of the fact that many of the results obtained in this project can be understood in relation with the characteristics of speech sounds, it is quite possible that the results obtained in such experiments would correspond to those which have been obtained in this research project. A new dimension besides the time resolution and the frequency resolution could be added: the number and distribution of amplitude levels. A synthesizer would not need to be built for this experiment; it could be simulated by a computer and a digital to analog converter.

12.4.4 A continuously adaptive analyser

If the ear can be compared to an adaptive spectrum analyser, it is certainly more similar to an analyser that can vary continuously the shape of its filters than to one that can switch between two or three modes. A continuously adaptive analyser would not seem however very useful at the present stage, although it might be possible to use it in psycho-visual experiments.

The Imperial College spectrum analyser can be modified in order to become continuously variable. The two main problems are to generate the weighting functions to the two banks of filters and to read the filters at the proper time. In order to do this, it would probably be expedient to use different circuits to quench and read each bank separately, the weighting function to one bank being the inverse of the weighting function of the other; the weighting function itself could be generated by a simple multivibrator whose frequency is determined by a control voltage; this multivibrator would be followed by a variable low-pass filter whose cut-off frequency would also be controlled automatically.

12.5 Limitations in physical effort and computing time

When this project was started, it was intended to study a number of consonants and vowels in different analysing modes; it was not intended to take the sounds

two by two but to have a greater number of categories; it was intended to do some more complex experiments by taking, for example, for one part of the sound the data from mode "A" and for the other part the data from mode "B" and find by how much this would add to the recognition performance; it was finally intended to take more than two analysing modes and to study the influence of redundancy.

It is possible in a few minutes to outline a program of work for quite a number of years. The main obstacles which have been encountered in this project have been the limitations in computing time and in physical effort.

The process of taking data is in itself quite a lengthy one, especially as one wants to have the most accurate measurements possible: the gain and the tuning of the filters must be checked often; their threshold varies with temperature and adjustments must be made. Experience has shown that one cannot hope to be able to collect data on much more than five speech sounds in a normal working day, which has meant something like sixty work-days to collect the data used in this thesis. If one adds to this the time lost at first before a routine was established, one can see that quite a lot of time was spent only on the process of taking the data.

Moreover, since this data comes in the shape of punched tape, it is necessary to make the necessary splices, to select the sounds one wants to use and splice them together. Every change in consonant pairs, for example, meant hours of splicing and checking. One should add to this the time spent on visual checks of the tapes.

In theory, it would have been advantageous to transfer the data on magnetic tape and effectuate the necessary transfers of data by modifying the program. When this was attempted, however, it was found that the extra routines that had to be included in the program pushed the storage capacity needed to compute the program to something well over 100 Blocks which meant, apparently, that the program was classified in another group and was consequently delayed - because it was using magnetic tape and too much storage capacity.

It has been reported earlier that it was attempted to handle more than two categories in any one program: the attempt was abandoned as the turnaround time appeared too long (ten days).

Finally the computing time available each month was limited and the amount allotted seemed to diminish as the years went by so that, at the end, for the experiments with the 50c/s mode, it was possible to send only some six programs a month, which explains why so few results are available with the 50c/s mode.

To sum up, it can be said that this research project has been deeply affected by such factors as computing time and physical effort; generally speaking, it has always been attempted to do the simplest experiments and to use the computer as little as possible. Anyone wishing to take over this work would have to tackle, first of all, the problem of the limited computing facilities available - by which is not implied that the computing facilities are not good or even not adequate for most users, but simply that they are on the borderline of being inadequate for this project.

12.6 Conclusion

It is felt that the main contributions of this work have been:

1. The discussion of adaptive filtering in its relation to signal analysis, speech and hearing;
2. the proposal to use a pattern recognition method to assess the differences in recognition performance between various analysing modes;
3. the experimental results obtained which have pointed out that a better time resolution in the spectral analysis seems to be in most cases preferable in order to classify stop-consonants, while a better frequency resolution is preferable for vowels or for identifying

consonants on the basis of the frequency transition between consonants and vowels.

4. The conclusion that the choice of the appropriate characteristics for the spectrum analyser, according to which class the speech sound to be identified is expected to belong to, will substantially improve the recognition performance, and the proposal that, for the best recognition performance, one should use a few spectrum analysers in parallel in conjunction with a complex decision rule where the influence of each analyser would be weighted according to its reliability.

This work is unfortunately left incomplete but it is felt that it still constitutes in itself a useful contribution and opens the way to more intensive and more detailed research.

APPENDIX A

THE SPEECH SOUNDS USED IN THE EXPERIMENTS

In table A.1, one can find a list of all the speech sounds used in the experiments with the 100c/s and the 200c/s modes. The speech sounds reserved for learning in these experiments are marked with an asterisk in table A.1.

For the 50c/s experiments, all the samples in "b" and "d" were used as well as the following sounds:

M.L.: pet I, pit I, pa I, pa II, tit I, to I,
po I, ta I,

P.P.: pit I, tet I, tit I, po I, ta I, to I,

D.G.: to I

In the 50c/s experiments with consonants, the same sounds as with the 100c/s and 200c/s modes were used for learning. For the vowel experiments, the following speech samples were chosen for learning:

M.L.: pet I, pit I, dit I, det I, bit I, tet I,
po I, ba II, ta I, da I

P.P.: bet I, to I, bo I, do I

D.G.: dit I, tet I

R.W.: da I

other: ta I.

	M.L.	P.P.	D.G.	R.W.	other
da (a)	I ,II ,III	I*	I	I	
do (o)	I ,II	I	I*		
det	I*,II	I		I	
dit	I*,II ,III	I	I*	I	
ta (a)	I ,II	I*	I	I	I
to (o)	I*,II	I	I*		
tet	I*,II ,III	I	I*	I	
tit	I ,II	I	I ,II	I	
ba (a)	I*,II	I*,II	I*	I	I ,II
bo (o)	I	I	I		
bet	I ,II	I*	I	I	
bit	I*,II	I	I	I	
pa (a)	I*,II	I*	I	I	I
po (o)	I ,II	I	I		
pet	I*,II	I	I	I	
pit	I ,II	I	I*	I	
ga (a)	I ,II	I	I	I	
go (o)	I	I*	I	I	
get	I*,II	I*,II		I	X,X,X,X
git	I	I	I*	I	I
ka (a)	I*	I	I*	I	I
ko (o)	I ,II	I	I		
ket	I*,II		I*	I	
kit	I	I*	I	I	

List of the speech sound samples used in the 100 c/s and the 200 c/s mode experiments.

Table A.1

APPENDIX B

RECORDING THE DATA

B.1 Introduction

A digital voltmeter coupled to a tape punch was available to record the data. The digital voltmeter was synchronized to the mains and could read at a maximum rate of 50 times per second; the tape punch could print a new row of holes on conventional computer paper tape at a maximum rate of 100 times a second.

The output of the digital voltmeter is in the form of binary coded digits: four columns of the tape are then required to record the state of every decade counter and it is possible with an eighth hole tape to record on every row the state of two decades. The modifications required by the digital voltmeter and the paper tape punch for this purpose have been carried out by D.J. Goodman. The four holes to the right represent the state of one decade and the four holes to the left, the state of the other.

It is then possible from the binary value of each row to work out the state of each decade. The output of the last storage circuit was normally adjusted for a threshold of 2.7 volts and the response of the analyser-storage circuits system became non-linear at amplitudes higher than 6.0 volts. The precision of the storage circuits being approximately

0.05 volts, it was only necessary to use two decades and since only the values up to 5.5 volts were of interest for one decade, it was decided to use 7-hole tape, which was more convenient since the programs were written on tapes of the same size.

B.2 The constraints

When the analyser operates in the 100c/s mode, each bank of filters is sampled fifty times a second. When operating in the 200c/s mode, each bank is sampled 100 times a second, which is twice the maximum speed of the digital voltmeter.

Furthermore, since the digital voltmeter is synchronized to the mains, it is either necessary to synchronize the analyser to the mains or else, to use buffer stages (storage circuits) between the analyser and the digital voltmeter. This last solution was adopted.

B.3 Synchronisation

The speech sound to be analysed was recorded on track 2 of a magnetic tape and a 100c/s tone of short duration was recorded on the other track, giving something similar to the illustration of figure B.1 a.

The synchronizing 100c/s wave then passed to a Schmitt's trigger whose output was used as input to the

Function Generator: it was then possible to get a 50c/s and a 100c/s tones of short duration and in phase with each other as well as a 50c/s square wave 90 degrees out of phase (fig. B.1 c,d,f).

The output of the Schmitt's trigger was also used to trigger a binary counter. The phase at the output of the binary counter could be reversed at will by operating a switch. The purpose of this square wave was to select, by proper gating, half the samples of one bank of filters when the 200c/s display was used. In the 100c/s mode, the action of the gating was equivalent to choosing the proper bank of filters (fig. B.1 b,e,g).

B.4 Storage

After having selected the proper samples, it was necessary to store them for 20msec - the time-interval between successive readings of the digital voltmeter. The storage circuits were a simple arrangement with a capacitor and the necessary transistor switches to quench it or to read in the value at the output of the filter. The storage circuit would hold its voltage with a precision of about 0.05 volts at the highest amplitudes for the 20msec period and performed better for lower signal amplitudes. If one intended to read filter "n", for example, one would

use for the reading the shift register stage "n" corresponding to the filter and would use for the quenching the shift register stage "n-1" occurring just before. The digital voltmeter is unfortunately synchronized to the mains and there was some risk of unprecision when the reading period coincided with the instant of quenching. In order to avoid this, an other storage circuit synchronized to the digital voltmeter was connected in series to the first storage circuit. If the reading pulse of the second storage circuit coincided with the quenching pulse of the first, the reading pulse of the second storage circuit would be slightly delayed.

The phase differences between the mains and the 100c/s tone triggering the analyser could then be disregarded: some frequency deviation due to the tape recorder could however not be avoided (about one in a thousand) and it meant that occasional samples have been lost or added (about one in a thousand). Since the samples were taken by small groups of about 12, since each group was positioned independently of the other and since the work presented here has been rather concerned with the consonants in initial position - which one finds at the start of each sample group, the effect of this error can be considered as having been minimal.

Figure B.2 shows that the relation between the output of the analyser and the second storage circuit

is almost linear.

B.5 Ordering and zeroing

It was necessary to take separate readings for each filter and for each bank of filters. In the 100c/s mode, one must read the 32 filters of each bank. In the 200c/s mode, one has to read 16 filters only in each bank, but the samples of each filter must be separated in two groups. The difficulty is to ensure that they are reconstructed in the right order.

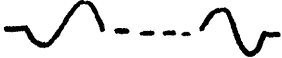
When the digital voltmeter (D.V.) operates in the manual mode, it can be triggered by an outside pulse but the reading does not take place immediately but at some later time determined by the phase of the mains voltage. Once the reading has taken place, the digital voltmeter generates a "conversion complete" pulse (c.c.) and goes to rest until it receives a new trigger pulse. In the present work, the conversion complete pulse was used to quench the second storage circuit and to trigger a monostable multivibrator whose output was sent to one terminal of an "OR" gate; to the other terminal was sent a function having the general characteristics of figure B.3 c; the output of the gate was used to trigger a monostable multivibrator which, in turn, triggered the D.V. The wave of figure B.3 c was obtained by "OR" gating B.3 a and the output of a

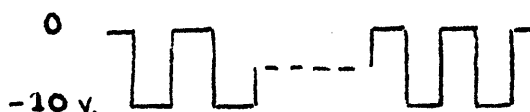
monostable multivibrator triggered by it (B.3 b). It was possible to record in this way a small group of selected samples.

A storage oscilloscope was used when recording to make sure that the samples were coming in the right order and were properly separated. After careful study, the following list of operating instructions was arrived at.

B.6 Operating instructions

(Abbreviations often refer to knobs and switches of the apparatus used).

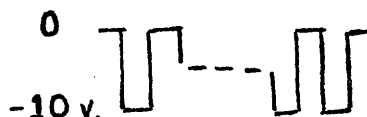
1. Record samples of speech on track two of tape loop.
2. Record  on track one.
3. Check that the input to the 50c/s low-pass filter (B.1 d) of the Function Generator is of the form



- if it is inverted, operate switch B.C. on the Function Generator box

- if it is of the form  or , return to step 2.

4. Check that the output of the POL. point (B.1 b) is also of the form



- if inverted, operate POL. switch.
 - if of another form, go back to step 2.
5. Check that both (B.1 b) and (B.1 e) keep repeatedly the same form, and monitor them all through the data recording operation.
 6. Check the triggering system.
 - 7.A) Testing and reading in the 100c/s mode:
 - 1.: switch to band "A"; see if samples come when the binary counter (B.C.) is in position 1 or 2; (for example, let us suppose position 1).
 - 2.: select filter bank "B"; the samples should appear only when the (B.C.) switch is in the other position (1 or 2); (in our example position 2)
 - 3.: Check with storage oscilloscope whether the samples No 1,3,5,7 etc. are from bank "A" or "B"; (for the purpose of our example: "A").
 - 4.: Read in appropriate order. In our example, one would proceed in the following order:
 - a. filter No.1, bank B; B.C. switch to 2
 - b. filter No.1, bank B; B.C. switch to 1
 - c. filter No.2, bank B; B.C. switch to 2
 - d. and so on.
 - 7.B) Testing and reading in the 200c/s mode:
 - 1.: See with storage oscilloscope whether samples No.1,5,9 etc. are from bank "A" or "B" and for what position of B.C. switch; (example: A-1).

- 2.: Check that the samples No.2,6,10 etc. come from the other bank but with the B.C. switch in the same position; (for example B-1)
- 3.: Check that the samples 3,7,11 etc. come from the same bank as No.1,5,9 etc. but with the B.C. switch in the other position; (example: A-2).
- 4.: Check that the samples 4,8,12 are of the same bank as 2,6,10 etc. but with the B.C. switch in the other position; (for example B-2).
- 5.: Read in a convenient order.

B.7 Recording for the 50c/s mode

The speed of the tape recorder was doubled, a new 100c/s synchronizing tone was recorded on track 1. The technique for recording the data was then virtually the same as in the 100c/s mode, with this difference that the normal spectral width of the analyser which covers the band up to 3.2kc/s when the tape is played back at the normal speed goes up effectively only to 1.6kc/s when the tape is played at twice the recording speed. It was then necessary to remodulate the speech wave to a different frequency so that the same filters could be used to take the samples of the spectral band from 1.6 to 3.2kc/s.

While doing this, one had to take into account the frequency response of the tape recorder and of the

frequency emphasis circuit at the input of the analyser. One had to multiply the samples by a weighting function so as to have the same order of amplitudes in all modes for different parts of the spectrum.

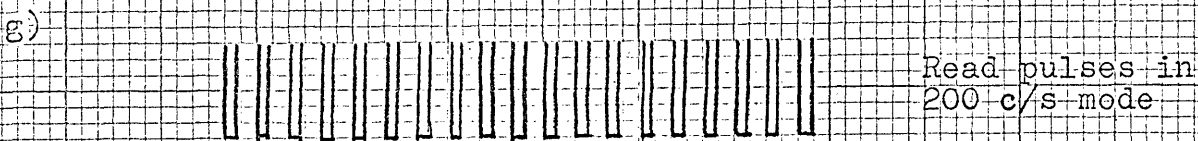
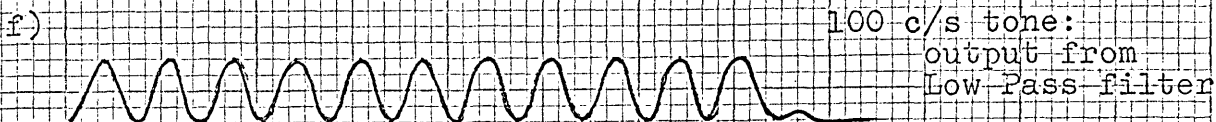
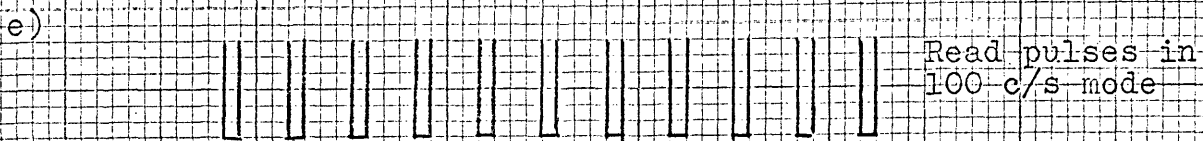
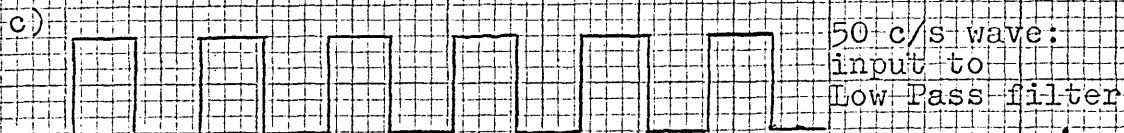
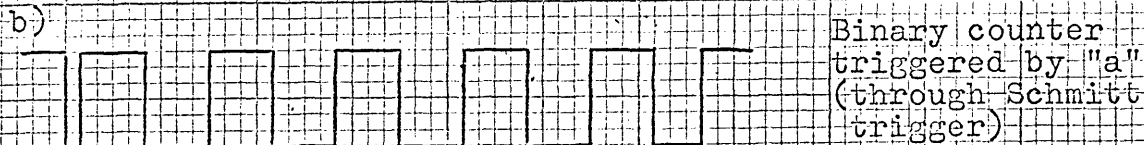
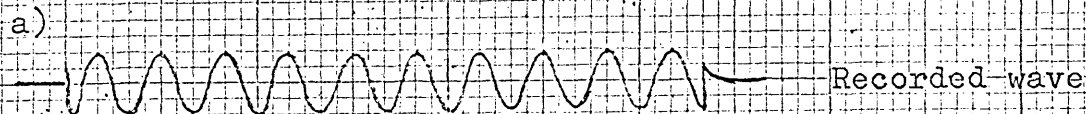
B.7 Comments

While the sampling instants in the 100c/s and the 200c/s modes were taken at the same points of the speech wave, those for the 50c/s mode were generally taken at different points because one needed to record a new synchronizing tone.

It was observed that, in many cases, the first samples taken in the 200c/s mode were referring to the time interval preceeding the first samples in the 100c/s mode by about 20 milliseconds: such a situation is depicted in figure B.1 f and h. This permitted, as was mentioned in chapter nine, a certain check to be carried out on the reliability of some positioning methods.

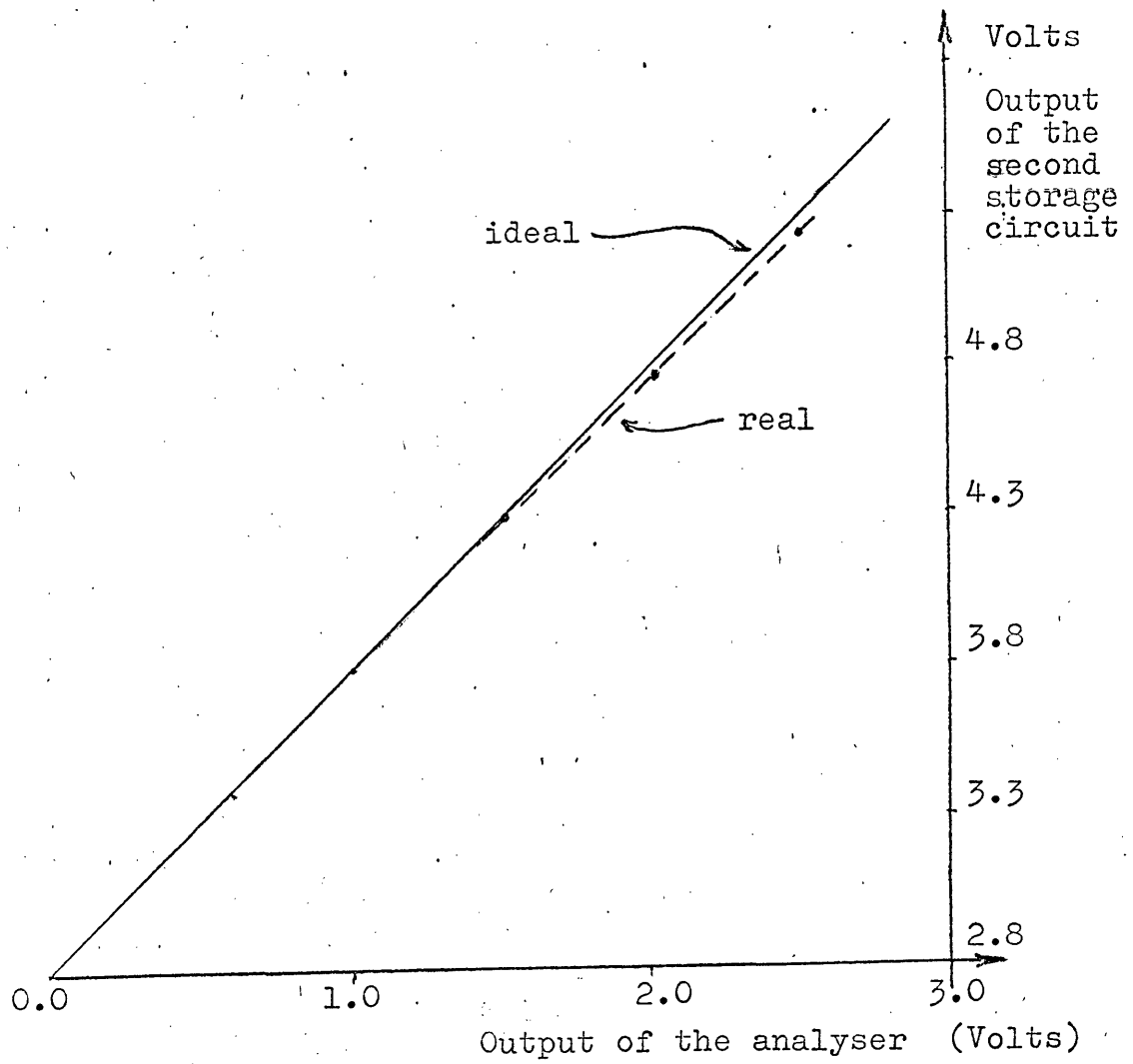
B.8 Conclusion

A block diagram of the system used to collect the data is given in figure B.4. The diagrams of the non standard circuits used in this piece of apparatus and in other instruments as well as supplementary explanations can be found in a technical manual in the Signals Laboratory.



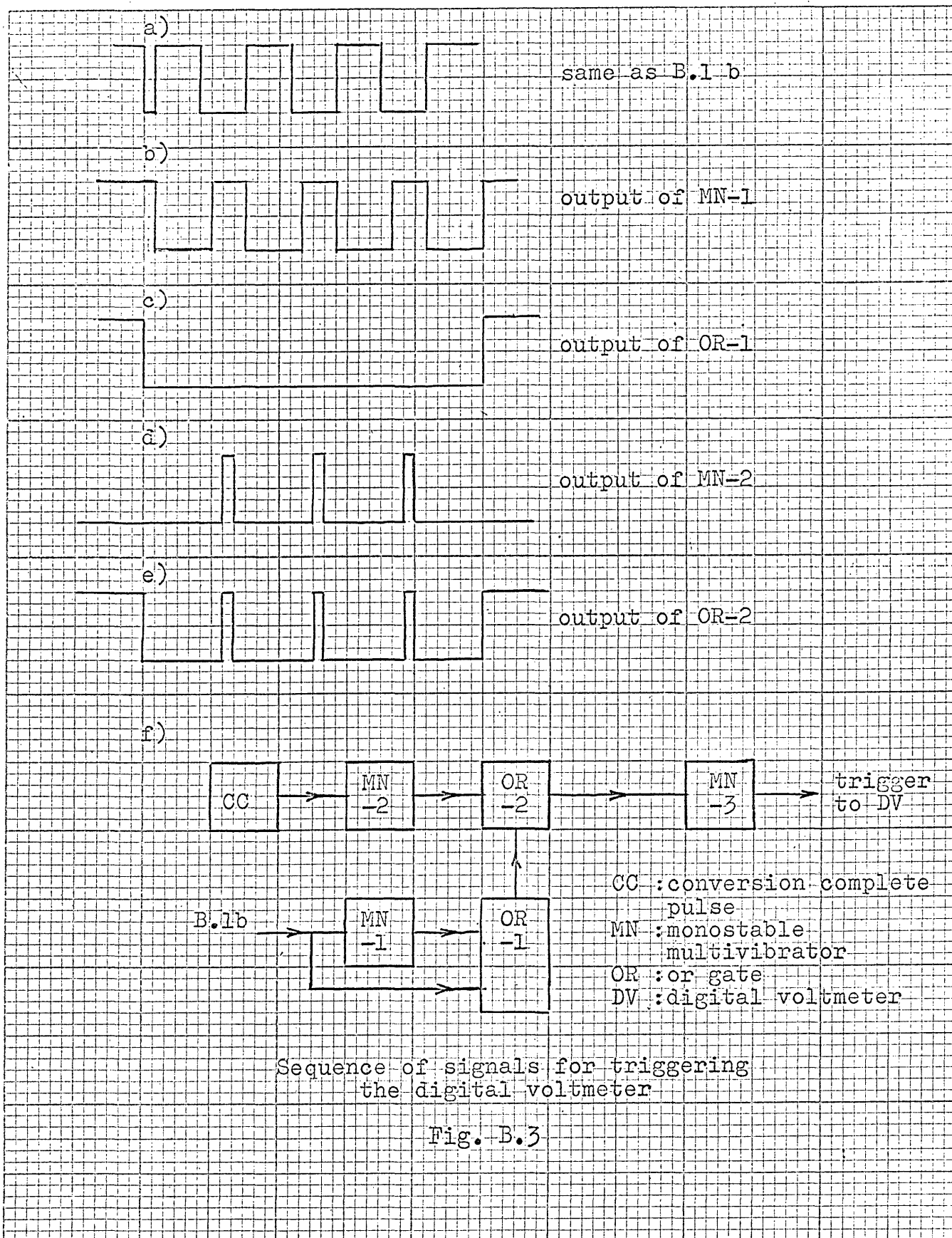
Sequence of signals for triggering
the speech analyser

Fig. B.1



Linearity of the storage circuits

Fig. B.2



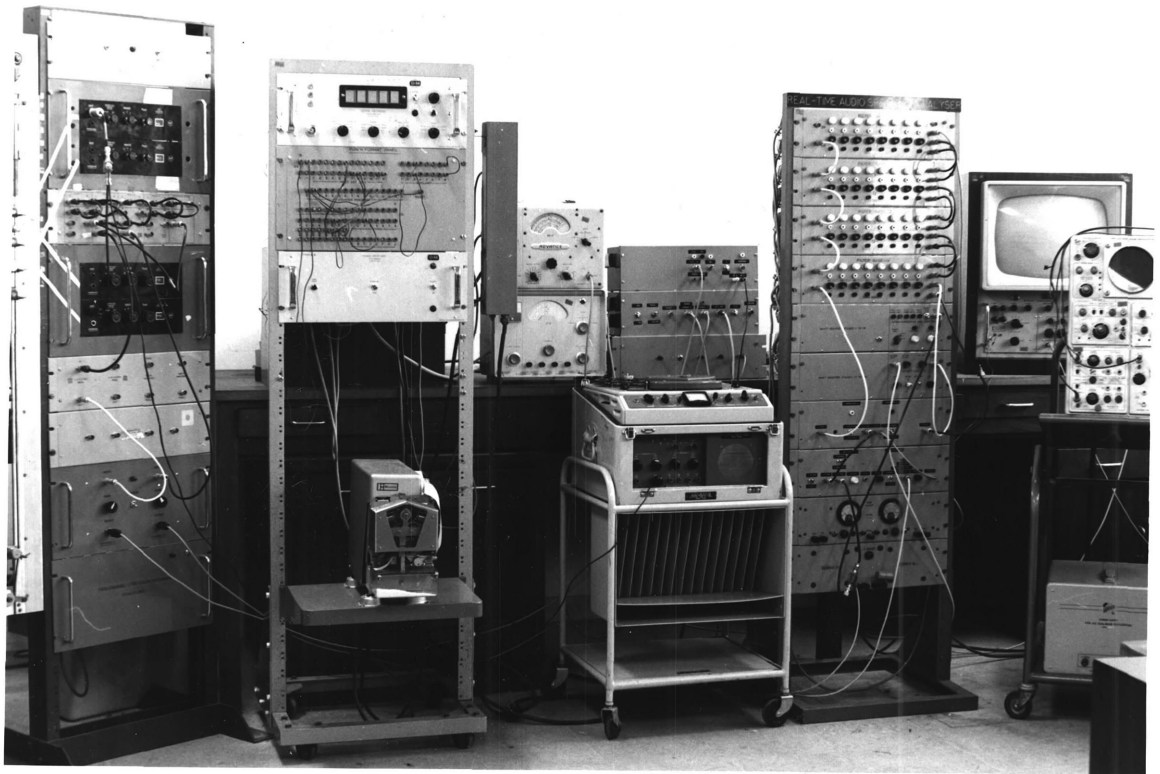


Fig. B.5

Apparatus used

From left to right:
tape delay line; digital voltmeter and tape punch unit;
signal generators; Function Generator and auxiliary cir-
cuits, pre-detector, tape recorder; spectrum analyser;
display oscilloscope; oscilloscope.

APPENDIX C

THE EXPERIMENTS WITH THE 50 c/s MODE

As has been mentioned earlier, the sampling times in the 50 c/s mode were not the same as in the 100 c/s and the 200 c/s representations since a different synchronizing tone had to be recorded on the tape loop. The extent to which this factor has influenced the results is not known.

Only a few experiments with the 50 c/s mode were carried out: the results are to be found in table C.1 as well as the results of parallel experiments in the 100 and 200 c/s modes.

On the whole, nothing permits to conclude that the 50 c/s representation is superior to the 100 c/s mode.

The 100 c/s representation has shown a remarkable capacity to separate the consonants "b" and "d" on the basis of the data covering the consonant-vowel transition period. It has been suggested that the reason for it is that these transitions show up better when the frequency resolution is finer; it would seem then that the time resolution offered by the 50 c/s mode is not high enough.

The performance of the 50 c/s mode for vowels does not appear to be better than that of the 100 c/s

mode. In the four experiments carried out, one result was exceptionally good and three were lower than the average for the 100 c/s mode.

On the whole, then, there is nothing in these results suggesting that it would be advantageous to use the 50 c/s mode. It could be, however, that the results for the vowel sounds would substantially improve if more learning samples were used since, in the 50 c/s mode, a slight change in the formant frequency can mean that a different filter is activated: more care should then be taken in drawing the decision rule.

Experiment No		Recognition performance					
		50 c/s	%	100 c/s	%	200 c/s	%
1	"b-d" from t_0+20 msec to t_0+60 msec	17/29		23/30		15/30	
2		21/31		22/30		17/30	
3		-		23/30		20/30	
total		38/60	63%	68/90	75%	52/90	58%
4	et-it	15/17		25/31		20/31	
5	"	11/17		25/31		21/31	
total	"	26/34	77%	50/62	81%	41/62	66%
6	a-o	14/23		22/30		23/30	
7	"	13/22		24/31		20/31	
8	"	-		29/31		26/31	
total	"	27/45	60%	75/92	81%	69/92	75%
total 6+7	"	27/45	60%	46/61	75%	43/61	70%

Table C.1

Recognition performance for experiments with the 50 c/s mode compared with the results for similar experiments in the 100 c/s and the 200 c/s modes.

APPENDIX D
THE ALGORITHMS

In this appendix will be found the programs which were developed for this research project. A program will first be given in complete form for experiments with the 100 c/s mode and for two categories. The numerals in the margin will correspond to notes at the end of the appendix indicating the changes that have to be made for experiments in the 200 c/s mode and for experiments with more than two categories. Comments within square brackets are not part of the program.

```
begin integer N,N1,N2,N3,N4,F,M,S,P,q1,NN,nn,t,a,  
q,y,n,s,elm,i,c,d,p,r,m,j,k,l,ab,ac,z,h;  
real Q1,Q2,Q3;Boolean lowend;  
N:= [ total number of spectrograms];  
M:= [ maximum allowable number of hyperplanes];  
d:=F:= [ number of spectrograms selected for learning];  
S:= [ number of samples per spectrogram];  
elm:= [ S * number of filters in analyser ]; 1)  
q1:= [ time interval immediately following the start of the  
       sound which will be disregarded; for 100 and 200 c/s  
       modes, 0:= 0 msec, 1:= 10 msec, 2:= 20 msec and so on];  
h:= 0 ;  
  
2)  
begin Boolean array A[1:N,0:M];array X[1:F+1,1:S],XC[1:F+1,1:M];  
for i:= 0 do for j := 1 step 1 until N do A[j,0] := true;  
[ the affiliation of each spectrogram is then inserted  
  in A[1:N,0] ];
```

[The purpose of the following procedure is to read the two decades of binary coded digits from the paper tape and retrieve their real value ;]

begin real procedure lesen;

 A: begin integer procedure lire;begin code

 2:1053,127,0,(1);

 1054,40,0,(2);

 121,127,0,(2);

 1:1054,40,0,0;

 1054,41,0,0;

 1345,40,0,9;

 1345,41,0,3;

 124,40,41,0;

 1524,0,40,0;

 356,0,-,lire;

end; a:= lire; y:= entier((a/16)+0.05); z:= a-16*y;

lesen:= y+0.1*z;end;

[The purpose of the following procedure is to select from the whole data the particular samples defining the portion of the sounds which one wants to consider in the experiment;]

begin real procedure entree(n); value n; integer n;

3)

begin array Y[1:1500];

A0:for Q3:= lesen while Q3> 0.01 do goto A0;

A1:Y[1]:= lesen; if Y[1]< 2.21 then goto A1 else

for ab := 1, ab+2 while Y[ab-2] ≥ 2.21 do Y[ab]:= lesen;m:=ab-4;

A2:Y[2]:= lesen; if Y[2] < 2.21 then goto A2 else

for ab := -2,ab+2 while Y[ab-2] ≥ 2.21 do Y[ab] := lesen;q:=ab-4;

A3:if m>q+1 then begin q:=q+2; Y[q]:=2.7; goto A3 end;q:=q-1;

for p:=1 step 1 until 31 do begin ac := p*q + 1;

for z:= 1,2 do begin

A4:Q1:= lesen;if Q1<2.21 then goto A4 else if Q1>7.0 then begin

```
write text((-lack of data, disregarded));  
print(n,2,0); newline(1);k:=n;  
for i:= h step 1 until 0 do  
for y:= k step 1 until N-1 do A[y,i] := A[y+1,i];N:=N-1;  
if n ≤ F then F:= F-1;goto A0 end else  
for ab := ac step 2 until (p+1)*q do begin  
Q1:=Y[ab] := lesen; if Y[ab] < 2.21 then begin  
print(n,2,0);print(p,2,0);print(z,2,0);  
Q3:= lesen; if Q3 < 2.21 then begin  
for ab := ab, ab+2 while ab ≤(p+1)*q do Y[ab] := 2.7;end else  
if Q3<-7.0 then begin Y[ab]:= (Y[ab-2]+Q3)/2;  
ab:=ab+2;Y[ab] := Q3;write text((-int));  
end else write text((-IRR)); end end;  
ac := p*q + 2;
```

```
A5: if Q1 > 2.21 and Q1 < 7.0 then begin  
Q1 := lesen; goto A5 end else end r; end p;  
if Q1 < 7.0 then begin print(n,2,0);  
write text((-too many data, disregarded(3c)));k := n;  
for i:=h step 1 until 0 do  
for y:= k step 1 until N-1 do A[y,i] := A[y+1,i];  
N := N-1;if n ≤ F then F:= F-1;  
QA: for Q2 := lesen while Q2 < 7.0 do goto QA;  
if n ≤ N then goto A0 end;  
for p:= 2 step 1 until q do begin  
for z:= 6 step 1 until 31 do begin  
A9:if Y[z*q+p] < 3.0 then goto A7 else  
if Y[z*q+p+1] ≥3.0 or Y[(z+1)*q+p] ≥3.0  
or Y[(z+1)*q+p+1] ≥ 3.0 then goto A8 else  
begin z:= z+1;goto A9 end; A7:end z; end p;  
A8:write text((-spectrograms n, readings p, filter z));
```

4)

```

print(n,2,0); print(p,2,0); print(z+1,2,0); newline(1);
  if p < 2 then begin write text((p/100,small,disregarded(c)));
k:=n;for i := h step 1 until 0 do
  for y := k step 1 until N-1 do A[y,i] := A[y+1,i];
N:=N-1; if n < F then F:=F-1; goto A0 end;
if q1 > 0 then NN := S else NN := S-20;
for s := 1 step 1 until NN - 20 do begin
X[n,s] := Y[entier((s-1)*elm)*(q-elm)+p+s+q1-1] - 2.7;
if X[n,s] < 0.0 then X[n,s] := 0.0 end;

```

```

if q1 > 1 then goto C1;
for j := 0,4,8,12,16 do begin ac:= p;
for s := S-19+j step 1 until S-16+j do begin ac:= ac-1;
if ac>0 then begin X[n,s] := Y[(j+4)*q-(s-S+20-j)+p+q1] - 2.7;
if X[n,s] < 0.0 then X[n,s] := 0.0 end else begin
X[n,s] := Y[(j+4)*q+1]-2.7; if X[n,s]<0.0 then X[n,s] := 0.0;
write text((V.COMP.,)); end;
end; end j; end entree;

```

```

[ The data is read in;]
for n:= 1 step 1 until F do entree(n);

```

```

[ The learning process follows;]
begin Boolean array B[1:F,1:M];array P[1:M,1:S+1];
write text((-2c) LEARNING/PROCESS(3c));

```

[The purpose of the following procedure is to draw a perpendicular hyperplane between spectrograms i and j of different affiliation; the coefficients of the hyperplane are stored in matrix P[1:M,1:S+1]; whether spectrograms i and j fall on the negative side or on the positive side of the hyperplane is stored in the boolean matrix A[1:N,0:M];]

```

begin procedure perp(i,j,k,l); value i;integer i,j,k,l;
begin integer m,n; real w; w := 0;
for m := 1 step 1 until S do begin
  P[k,m] := X[i,m] - X[j,m];
  w := -w + P[k,m]*(X[i,m] + X[j,m]) end;
P[k,S+1] := w/2 ;print(P[k,S+1],2,2);
for m := 1 step 1 until 1 do begin w:=0;
  for n := 1 step 1 until S do w:= w+P[k,n]*X[m,n];
  XC[m,k] := w - P[k,S+1];
  A[m,k]:= w- P[k,S+1] ≥ 0 end;
r:= r+1;print(r,2,0); write text((-perp(-c)));
end perp;

```

[The procedure update(i) stores in boolean matrix A[1:N,0:M] whether the spectrograms involved in the learning process fall on the negative or the positive side of hyperplane i;]

```

begin procedure update(i); integer i;
begin integer k,l;real w;
for k := 1 step 1 until r-1 do
  begin w:= 0; for l := 1 step 1 until S do
    w:= w + P[k,l]*X[i,l]; A[i,k]:= w - P[k,S+1] ≥ 0;
    XC[i,k]:=w-P[k,S+1];print(i,2,0); write text((-upd));
  end end update;

```

[The first hyperplane is then drawn;] 6)

```

B4: begin array Y[1:S]; m:=1;
if not A[1,0] = A[2,0] then begin
  r:=1; perp(1,2,1,2);p:=3; goto C4; end else
  C2: begin for m:= m+1 while A[m,0] = A[1,0] do goto C2;
  write text((-reshuffle));
  for p:= 1 step 1 until S do begin
    Y[p]:= X[2,p]; X[2,p] := X[m,p]; X[m,p] := Y[p] end;
    A[2,0]:= not A[2,0]; A[m,0] := not A[m,0]; end; goto C1;
  C4: end;

```

[The hyperplanes are drawn in order to separate the spectrograms belonging to different categories;]

```

B5: for j:= p,j while j<F and r<M do begin
  update(j);for k:= 1 step 1 until j-1 do begin
    for l := 1 step 1 until r-1 do
      if not A[k,l] = A[j,l] then goto L1;
      if A[k,0] = A[j,0] then goto L1 else if r< M then perp(k,j,r,j7)
      else write text(((c)M/TOO/SLOW(c)));
    L1: end k;
  L2: j:=j+1;print(r,2,0);print(j,2,0);write text(((r,j,B5)));
end B5;
write text(((arraysAspriorstosimplification(c)));
for z:= 1 step 1 until j-1 do begin
  for ac:= h step 1 until r-1 do write Boolean(A[z,ac]);newline(1);
  for ac :=1 step 1 until r-1 do print(XC[z,ac],2,2); newline(1);
end;newline(1);

```

[The set of hyperplanes is then simplified; this part of the program follows closely the algorithm published by Turski, to which one should refer for more details;]

```

B7: begin integer k,l,m,n;
for k := 1 step 1 until j-2 do begin
  for l := k+1 step 1 until j-1 do begin
    for m := h step 1 until r-1 do if not A[k,m] = A[l,m]
    then goto L1;
    for m := 1 step 1 until j-2 do begin
      for n := h step 1 until r-1 do A[m,n] := A[m+1,n];
    for n:=1 step 1 until r-1 do XC[m,n] := XC[m+1,n];
    for n := 1 step 1 until S do X[m,n] := X[m+1,n]
    end m; j:=j-1;print(j,2,0); write text(((j,B7)));
  L1: end l end k;
end B7;

```

```
B8: begin Boolean array a,b,c[1:r];integer i,k,l,m;  
begin Boolean procedure compare(a,b,i); Boolean array a,b;  
integer i; begin integer l;  
  for l := 1 step 1 until i do if not a[l] = b[l] then goto L1;  
  compare := true; goto L2;  
L1: compare := false;  
L2: end compare;  
begin procedure boolprod(A,B,C,i,j,k,special);  
Boolean array A,B,C; integer i,j,k; Boolean special;  
begin integer l; if special then begin  
  for l := 1 step 1 until i do C[l] := A[l] and B[k,l] end else  
  for l := 1 step 1 until i do C[l] := A[j,l] and B[k,l];  
end boolprod;
```

```
B9: begin for i := 1 step 1 until j-1 do  
for k := 1 step 1 until r-1 do B[i,k] := true;  
for i := 1 step 1 until r-1 do  
for k := 1 step 1 until j-1 do begin  
  B[k,i] := false ; boolprod(A,B,a,r-1,k,k,false);  
  for l := 1 step 1 until k-1, k+1 step 1 until j-1 do begin  
    if A[k,0] = A[l,0] then goto G1;  
    boolprod(A,B,b,r-1,l,k,false);  
    boolprod(a,B,c,r-1,0,l,true);  
    boolprod(b,B,b,r-1,0,l,true);  
    if compare(b,c,r-1) then begin  
      B[k,i] := true; goto G2 end;  
  G1: end l;  
G2: end k; for i := 1 step 1 until j-2 do begin  
  boolprod(A,B,a,r-1,i,i,false);  
  for k := 1 step 1 until i-1, i+1 step 1 until j-1 do begin  
    if not A[k,0]=A[i,0] then goto L3;  
    boolprod(A,B,b,r-1,k,i,false);  
  8) 9)
```

```

if compare(a,b,r-1) then begin
  for l := k step 1 until j-2 do begin
    for m := h step 1 until r-1 do A[l,m] := A[l+1,m];
    for m := 1 step 1 until r-1 do begin
      B[l,m] := B[l+1,m]; XC[l,m] := XC[l+1,m] end m;
    for m := 1 step 1 until S do X[l,m] := X[l+1,m] end l;
  j := j-1; end if;
L3: end k;
end i;
end B9;

```

```
[ The content of matrices A[1:N,O:M] and B[1:F,1:M] is printed
out;the decision rule is tested by classifying new spectro-
grams; ]
```

```

write text((-c)arraysA(afterlearning(2c)));
begin integer ja; ja := j;
print(r,2,0); print(j,2,0); print(p,2,0); newline(1);
for z:=1 step 1 until ja-1 do begin
for ac:=h step 1 until r-1 do write Boolean(A[z,ac]);
newline(1);end;newline(1);
write text((-c)arraysB(2c)));
for z:= 1 step 1 until ja-1 do begin
for ac:=1 step 1 until r-1 do write Boolean(B[z,ac]);
newline(1);end;newline(1);
for z:= 1 step 1 until ja-1 do begin
for ac:=1 step 1 until r-1 do print(XC[z,ac],2,1);
newline(1);end; newline(1);
for m:= F+1 step 1 until N do begin d:=m;entree(F+1);update(F+1);
for ac := h step 1 until r-1 do write Boolean(A[F+1,ac]);
newline(1);
for ac := 1 step 1 until r-1 do print(XC[F+1,ac],2,1);
newline(1)end;newline(1);
end end end end end end end end end end

```

NOTES

1) In fact, $elm := (S-20) + (\text{number of filters in analyser})$
if $q1 > 0$; the twenty extra samples are concerned with the
low frequency region immediately preceeding the consonant.

2) If more than two categories are used, this paragraph should
read as follows:

begin Boolean array A[1:N,h:M], XA[h:0];

array X[1:F+1,1:S], XC[1:F+1,1:M];

[the affiliation of each spectrogram is then inserted
in A[1:N,h:0];].

3) When experiments with the 200 c/s mode are done, the program
should be amended as follows:

begin real procedure entree(n); value n; integer n;

begin array Y[1:1500];

A0:for Q3:= lesen while Q3> 0.01 do goto A0;

A1:Y[1]:= lescn; if Y[1]< 2.21 then goto A1 else

for ab := 1, ab+4 while Y[ab-4] ≥ 2.21 do Y[ab]:= lesen;m:=ab-8;

AB3:Y[3]:= lesen; if Y[3] < 2.21 then goto AB3 else

for ab := 3, ab+4 while Y[ab-4] ≥ 2.21 do Y[ab]:= lesen;i:=ab-8;

A2:Y[2]:= lesen; if Y[2] < 2.21 then goto A2 else

for ab := 2,ab+4 while Y[ab-4] ≥ 2.21 do Y[ab] := lesen;s:=ab-8;

AB4:Y[4]:= lesen; if Y[4] < 2.21 then goto AB4 else

for ab := 4, ab+4 while Y[ab-4] ≥ 2.21 do Y[ab]:= lesen;q:=ab-8;

A3:if m>q+1 then begin q:=q+4; Y[q]:=2.7; goto A3 end;q:=q-1;

for p:=1 step 1 until 15 do begin -

for z:= 1,2,3,4 do begin ac := p*q + z;

A4:Q1:= lesen;if Q1 <2.21 then goto A4 else if Q1>7.0 then begin
write text((-lacks of data,disregarded));

print(n,2,0); newline(1);k:=n;

for i:= h step 1 until 0 do

for y:= k step 1 until N-1 do A[y,i] := A[y+1,i];N:=N-1;

```
if n  $\leq$  F then F:= F-1;goto A0 end else  
for ab := ac step 4 until (p+1)*q do begin  
Q1:=Y[ab] := lesen; if Y[ab] < 2.21 then begin  
print(n,2,0);print(p,2,0);print(z,2,0);  
Q3:= lesen; if Q3 < 2.21 then begin  
for ab := ab, ab+4 while ab  $\leq$  (p+1)*q do Y[ab] := 2.7;end else  
if Q3  $\leq$  -7.0 then begin Y[ab] := (Y[ab-4]+Q3)/2;  
ab:=ab+4;Y[ab] := Q3;write text((-int));  
end else write text((-IRR)); end end;  
A5: if Q1 > 2.21 and Q1 < 7.0 then begin ...
```

4) For experiments in the 200 c/s mode, the program should be amended as follows:

```
for p:= 4 step 1 until q do begin  
for z:= 3 step 1 until 14 do begin  
A9:if Y[z*q+p] < 3.0 then goto A7 else  
if Y[z*q+p+1]  $\geq$  3.0 or Y[(z+1)*q+p]  $\geq$  3.0  
or Y[(z+1)*q+p+1]  $\geq$  3.0 then goto A8 else ...
```

5) For experiments in the 200 c/s mode, the program should be amended as follows:

```
if q1 > 1 then goto C1;  
for j := 0,10 do begin ac:= p;  
for s := S-19+j step 1 until S-10+j do begin ac:= ac-1;  
if ac>0 then begin X[n,s] := Y[(j+10)*q-(s-S+20-j)+p+2*q1] - 2.7;  
if X[n,s] < 0.0 then X[n,s] := 0.0 end else begin  
X[n,s] := Y[(j+10)*q+1]-2.7; if X[n,s]<0.0 then X[n,s] := 0.0;  
write text((-V.COMP.,)); end; ...
```

6) For experiments with more than two categories, the program is amended as follows:

```
begin Boolean procedure vergleich(h,j,i,k); integer h,j,i,k;  
begin integer l;  
  for l := h step 1 until j do if not A[i,l] = A[k,l] then  
    goto L1; vergleich := true; goto L2;  
  L1: vergleich := false;  
  L2: end vergleich;  
B4: begin array Y[1:S]; m:=1;  
C1: if not vergleich(h,0,1,2) then begin  
  r:=1; perp(1,2,1,2); p:=3; goto C4; end else  
  C2: begin for m:= m+1 while vergleich(h,0,1,m) do goto C2;  
  write text(←reshuffle→);  
  for p:= 1 step 1 until S do begin  
    Y[p]:= X[2,p]; X[2,p] := X[m,p]; X[m,p] := Y[p] end;  
    begin for p := h step 1 until 0 do begin  
      XA[p] := A[2,p]; A[2,p] := A[m,p]; A[m,p] := XA[p] end end;  
  end; goto C1;  
C4: end;
```

7) For experiments with more than two categories, the program is amended as follows:

```
if vergleich(h,0,k,j) then goto L1 else ...
```

8) For experiments with more than two categories, the program is amended as follows:

```
if vergleich(h,0,k,l) then goto G1; ...
```

9) For experiments with more than two categories, the program is amended as follows:

```
if not vergleich(h,0,k,i) then goto L3; ...
```

Bibliography

1. Ahmend R., Fatehchand R., "Effect of sample duration on the articulation of sounds in normal and clipped speech", J.Acoust.Soc.Am., 31, 1022, 1959.
2. Aizerman, Braverman, Rozonoer, "Theoretical foundations of the Potential Function method in pattern recognition learning", Automation and Remote Control, 25, 821, 1964.
3. Allais D.C., "The selection of measurements for prediction", Technical Report No. 6103-9, Stanford Electronics Laboratory, 1964.
4. Bashkirov, Braverman, Muchnik, "Potential function algorithms for pattern recognition in learning machines", Automation and Remote Control, 25, 629, 1964.
5. Braverman E.M., "Experiments on machine learning to recognize visual patterns", Automation and Remote Control, 25, 315, 1962.
6. Broadbent D.E. and Ladefoged P., "Auditory perception of temporal order", J.Acoust.Soc.Am., 31, 1539, 1959.
7. Bürck, Kotowski, Lichte, "Der Aufbau des Tonhohenbewusstseins", Elek. Nachr.-Technik, 10, 326, 1935.
8. Caldwell W.F., "Recognition of sounds by cochlear patterns", IEEE trans. on military electronics, 7, 179, 1963.
9. Cherry C., "Pulses and transients in communication circuits", Chapman and Hall, 1949.
10. Cherry C., "On human communication", Science Edition, Inc., New York, 1961.
11. Creelman C.D., "Detection of complex signals as a function of signal bandwidth and duration", J.Acoust.Soc.Am., 33, 89, 1961.
12. Davis K.H., Biddulph R., Balashek S., "Automatic recognition of spoken digits", J.Acoust.Soc.Am., 24, 637, 1952.

13. Dreyfus-Graf J., "Sonograph and sound mechanics", J.Acoust.Soc.Am., 22, 731, 1950.
14. Dudley H., "Remaking speech", J.Acoust.Soc.Am., 11, 169, 1939.
15. Dudley H. and Balashev S., "Automatic recognition of phonetic patterns in speech", J.Acoust.Soc.Am., 30, 721, 1958.
16. Dunn H.K., "On vowel resonances and an electrical vocal tract", J.Acoust.Soc.Am., 22, 740, 1950.
17. Fant G., "Acoustic theory of speech production", 's Gravenhage, Mouton and Co., 1960.
18. Flanagan J.L., "Bandwidth and channel capacity necessary to transmit the formant information in speech", J.Acoust.Soc.Am., 28, 592, 1956.
19. Flanagan J.L. and Saslow M.G., "Pitch discrimination for synthetic vowels", J.Acoust.Soc.Am., 30, 435, 1958.
20. Fry D.B., "Theoretical aspects of mechanical speech recognition", J.Brit.Inst.Radio Engrs, 19, 211, 1959.
21. Fry D.B. and Denes P., "The role of acoustics in phonetic studies" in "Technical aspects of sounds", vol. 3, E.G. Richardson and E. Meyer, Eds. Amsterdam, Elsevier Publishing Co., 1962.
22. Gabor D., "Theory of Communication", J.Inst.Elect. Engrs, (London), 93 part III, 429, 1946.
23. Gold B., "Experiments with speech-like phase in a spectrally flattened pitch-excited channel vocoder", J.Acoust.Soc.Am., 30, 1892, 1964.
24. Gold T., "Hearing", IRE Trans. on information theory, 1, 125, 1953.
25. Gold T. and Pumpfrey R.J., "The cochlea as a frequency analyser", Proc.Roy.Soc., B 135, 462, 1948.
26. Golden R.M., "Digital computer simulation of a sampled-data voice-excited vocoder", J.Acoust. Soc.Am., 35, 1958, 1964.

27. Green D.M., "Auditory perception of a noise signal", J.Acoust.Soc.Am., 32, 121, 1960.
28. Green D.M., Birdsall T.G., Tanner W.P., "Signal detection as a function of signal intensity and duration", J.Acoust.Soc.Am., 29, 523, 1957.
29. Halle M., Hughes G.W., Radley J.-P.A., "Acoustic properties of stop-consonants", J.Acoust.Soc. Am., 29, 107, 1957.
30. Halle M. and Stevens K., "Speech recognition: a model and a program for research", IRE trans. on information theory, 8, 155, 1962.
31. Highleyman, "Linear decision functions with application to pattern recognition", Proc. IRE, 50, 1501, 1962.
32. Hirsch I.J., "Auditory perception of temporal order", J.Acoust.Soc.Am., 31, 759, 1959.
33. Jacobson R., Fant C.G.M. and Halle M., "Preliminaries to speech analysis", Technical Report No. 13, Acoustics Laboratory, MIT, 1952.
34. Kay I. and Silverman R.A., "On the uncertainty relation for real signals", Information and Control, 1, 64, 1957.
35. Konig, "Effect of time on pitch discrimination threshold", J.Acoust.Soc.Am., 29, 606, 1957.
36. Lawson J.L. and Uhlenbeck G.E., "Threshold signals", McGraw-Hill, 1950.
37. Liang chic-An and Chistovich L.A., "Frequency difference limens as a function of tonal duration", Soviet physics, Acoustics, July 1960, p. 75.
38. Liberman A.M., "Some results of research on speech perception", J.Acoust.Soc.Am., 29, 117, 1957.
39. Liberman A.M. et al., "A Motor Theory of speech production", Proc. of Speech Communication Seminar, Stockholm, 1962.
40. Liberman A.M., Delattre P.C., Cooper F.S., "The role of selected stimulus variables in the perception of the unvoiced stop-consonants", Am.J.Psychol., 65, 497, 1952.

41. Liberman A.M., Delattre P.C., Cooper F.S., Gertsman L.J., "The role of consonant-vowel transitions in the perception of stop and nasal consonants", Psych. Mono., 68, No. 1, 1954.
42. Licklider J.C.R., "Three auditory theories" in "Psychology, a study of science", Editor S.Koch, McGraw Hill, 1959.
43. Liljencrants J., "A few experiments of voiced-voiceless identification and time segmentation of speech", Report STL-QPSR-3/1962, Royal Institute of Technology, Stockholm, October 1962.
44. Liljencrants et al., "51 channel analyser for spectrum sampling", Proc. of Speech Communication Seminar, Stockholm, 1962.
45. Lindgren N., "Machine recognition of human language", parts I,II,III, IEEE Spectrum, vol.2, 1965.
46. Martin T.B. and Tavalage J.J., "Application of neural logic to speech analysis and recognition", IEEE trans. on military electronics, 7, 189, 1963.
47. Martin T.B. et al., "Speech recognition by feature-abstraction technique", Technical Document Report No. ALTRD 64-176, August 1964, Air Force Avionics Laboratory.
48. Miller G.A. and Taylor W.G., "The perception of repeated bursts of noise", J.Acoust.Soc.Am., 20, 171, 1948.
49. Oettinger R., "Die Grenzen der Hörbarkeit von Frequenz- und Tonzahländerungen", Akustica, 9, 431, 1959.
50. Paget, Sir Richard, "Human speech", Kegan Paul, Trench, Trubner and Co. Ltd., London, 1930.
51. Potter R.K., Kopp G.A., Green H.C., "Visible speech", D. van Nostrand Co. Inc., New York, 1947.
52. Potter R.K. and Steinberg J.C., "Towards the specification of speech", J.Acoust.Soc.Am., 22, 807, 1950.
53. Sacschewsky K.E. and Oestreicher H.L., "Pattern recognition as a problem in decision theory and an application to speech recognition", IEEE trans. on military electronics, 7, 186, 1963.

54. Sebestyen G.S., "Decision making processes in pattern recognition", MacMillan, New York, 1962.
55. Sekey A.R., "A study of auditory perception in the time frequency domain", Ph.D. Thesis, University of London, March 1962.
56. Shower and Biddulph, "Differential pitch sensitivity of the ear", J.Acoust.Soc.Am., 3, 275, 1931.
57. Stevens S.S. and Davis H., "Hearing, its psychology and its physiology", Wiley, 1938.
58. Stewart G.W., "Problems suggested by an uncertainty principle in acoustics", J.Acoust.Soc.Am., 2, 325, 1931.
59. Stumpf, "Die Sprachlaute", Julius Springer Verlag, Berlin, 1926.
60. Thomas R.S., "A real-time audio spectral analyser using active filters with adjustable parameters", Ph.D. Thesis, London University, 1964.
61. Turski W., "A learning automation for solving stability problems of differential equations", Computatio, 1, 57, 1964.
62. Wallack, Newman and Rosenzweig, "The precedence effect in sound localization", Am.J.Psychol. 62, 315, 1949.
63. Wiren J. and Stubs H.L., "Electronic binary selection system for phoneme classification", J.Acoust.Soc.Am., 28, 1082, 1956.
64. Wolf E., "Reciprocity, inequalities, coherence time and bandwidth in signal analysis and optics", New York University, Inst. of Math. Science, Div. electromagnetic research, Report No. EM-106.
65. Hemdel J.F. and Hughes G.W., "A feature based computer recognition program for the modeling of vowel perception", Proc. of the Symposium on models for the perception of speech and visual form, AFcRL, Nov. 1964,

66. Wamstein L.A., Zubakov V.D., "Extraction of signals from noise," Prentice Hall, 1962.
67. Fletcher H., "Speech and Hearing in communication" Van Nostrand, 1953.
68. Lawrence W., "Parametric Artificial Talking Device," the 60th meeting of the Acoust. Soc. Am., session G, October 20-22, 1960.