

VISUAL ENHANCEMENT OF TELEPHONE CONVERSATIONS.

A thesis submitted for the degree
of Doctor of Philosophy in the Faculty
of Engineering, University of London

by

Barry Stapley

Imperial College, University of London

1972

ABSTRACT.

The thesis reports the results of a study of how a videotelephone may affect the perception and use of many of the non-verbal cues that are often used in face-to-face meetings.

In most of the experiments described, two types of video image were used, namely a high definition monochrome image and a simple cartoon image. An experiment similar to that conducted by Sumbly and Pollack showed that the cartoon image was almost as good as the high definition monochrome image for lip reading. The results of an experiment on the perception of direction of gaze were used to decide the relative positions of the camera and video monitors within the laboratory videotelephone experiment. Two exploratory experiments were performed to investigate (1) the apparent proximity of a facial image on a monitor screen, (2) the relative utility of different types of video image for the recognition of a person's face.

Video-tape recordings were made of subjects conversing over the laboratory videotelephone link and these recordings were later analysed in order to investigate the subjects' looking and speech behaviour. It was found that the type of video image had negligible effect on the conversants' looking or speech behaviour, whereas the level of noise added to the audio channel and the method used to stimulate conversation had a significant effect on many of the different behavioural measures. The measured amount of eye-contact was found to be significantly greater than that predicted by a simple probabilistic model of looking behaviour. This indicated that a videotelephone user's direction of gaze was directly influenced by the direction of gaze of the other conversant.

CONTENTS.

<u>Item.</u>	<u>Page.</u>
ABSTRACT	2
CONTENTS	3
GENERAL NOTES	7
ACKNOWLEDGEMENTS	8
CHAPTER 1. INTRODUCTION	9
1.1 The possible advantages of a videotelephone system	10
1.2 Effectiveness of different modes of communication	13
1.3 The author's field of study	18
1.4 Contribution of this thesis	21
1.5 Summary and preview of later chapters	22
CHAPTER 2. VERBAL AND NON-VERBAL SIGNALLING	24
2.1 Conversation and signs	25
2.2 Verbal and non-verbal signs	27
2.3 Kinesics	29
2.4 Facial expression	30
2.5 Direction of gaze	32
2.6 Proximity, posture and orientation	37
2.7 Physical appearance	39
2.8 Speech	40
CHAPTER 3. THE VISUAL IMAGE	43
3.1 Visual communication channels	44
3.2 Video image coding	47
3.3 Videotelephone schemes	50
3.4 Control of the viewer's movement	54

3.5	Picture size and viewing distance	57
3.6	Visual images used for experiments	60
3.7	Generation of cartoon image	65
CHAPTER 4. CONTROL, RECORDING AND TERMINAL EQUIPMENT		76
4.1	Relative position of video monitor and camera	77
4.1.1.	Perception of eye-contact	78
4.1.2.	Video links	81
4.1.3.	Choice of monitor and camera positions	83
4.1.4.	Experiment to measure the perception of eye-contact	84
4.1.5.	Results	87
4.1.6.	Discussion of results	91
4.1.7	Final positioning of camera	93
4.2	Audio channel	96
4.2.1	Loudspeaking telephones	97
4.2.2	Audio system used for experiments	98
4.2.3	Noise	101
4.3	The laboratory video system	104
4.3.1	Video synchronising system	104
4.3.2	Video control unit	107
4.3.3	Video bandwidth	111
4.4	Data processing apparatus	113
CHAPTER 5. LIP READING		116
5.1	Purpose of experiment	117
5.2	Lip reading experiment	119
5.2.1	Use of video-tape recorder	119
5.2.2	Vocabulary	121
5.2.3	Visual images used	122
5.2.4	Experimental procedure	124
5.2.5	Results	126
5.2.6	Discussion of results	130

5.3	Bisensory perception	131
5.3.1	Accessory stimulation	131
5.3.2	Combined visual and audio perception	132
CHAPTER 6. IDENTIFICATION AND APPARENT PROXIMITY OF A VIDEO IMAGE		136
6.1	Visual identification of a person	137
6.1.1	Recognition of a videotelephone caller	137
6.1.2	Detection or discrimination	138
6.1.3	Visual form discrimination	139
6.1.4	Facial images used for the experiment	141
6.1.5	Types of video signal used	142
6.1.6	Test procedure	144
6.1.7	Results	146
6.1.8	Discussion of results	149
6.2	The apparent proximity of a video facial image	151
6.2.1	The perception of distance	151
6.2.2	Experimental apparatus and procedure	154
6.2.3	Results	156
6.2.4	Discussion of results	157
CHAPTER 7. THE LOOKING AND SPEECH BEHAVIOUR OF A VIDEOTELEPHONE USER		160
7.1	Introduction	161
7.1.1	Interaction with a video image	165
7.2	Stimulation of conversation	170
7.3	Preliminary experiment	175
7.3.1	Objectives of the preliminary experiment	175
7.3.2	Experimental procedure	175
7.3.3	Forms of behaviour measured	176
7.3.4	Results	178
7.3.5	Discussion of results	181

7.4	Main experiment	183
7.4.1	Objectives of the main experiment	183
7.4.2	The experiment	183
7.4.3	Statistical analysis	189
7.4.4	Results	193
7.4.5	Subsidiary experiment	199
7.4.6	Discussion of results	205
CHAPTER 8.	GENERAL SUMMARY AND SUGGESTIONS FOR FURTHER WORK	216
APPENDIX I	Number of white picture elements in a cartoon image	220
APPENDIX II	Data rate for image transmission	222
APPENDIX III	Circuit diagrams	227
REFERENCES		229

GENERAL NOTES.Abbreviations.

d.f.	-	degrees of freedom
F	-	Snedecor's variance ratio
MS _A	-	mean square of factor A
p	-	probability of null-hypothesis being confirmed
R.M.S.	-	root mean square
SS _A	-	sum of squares relating to factor A
t	-	as in Student's t test
T	-	sum of the ranks with the less frequent sign in the Wilcoxon matched-pairs signed-ranks test
X'	-	transform of X

Results presented graphically.

In chapters 4 to 7, many experimental results are shown graphically. In certain figures, lines are drawn through points representing related measurements. These lines have been introduced solely to clarify this connection and they do not represent the result of curve-fitting operations.

Where an experimental point represents the mean $\bar{x}$ of a number of independent observations, the variations of these measurements is indicated by a line between the limits $\bar{x} \pm \delta$, where δ is the standard error of the mean.

ACKNOWLEDGEMENTS.

The author wishes to express his thanks to his supervisor, Professor E.C. Cherry, for continued advice and encouragement throughout the course of this study, and also for procuring funds for the research.

Dr. J.E. Thompson, B.G. Champness and J.F. Springate read the manuscript of this thesis and made many helpful comments. The author is indebted to Mr. A.A. White for his technical assistance in maintaining the picture generation equipment and for constructing some of the author's apparatus. A debt of gratitude is also owed to all those who participated in the experiments, as either subjects or observers, the latter task being a particularly unenviable one. Sincere thanks are extended to my wife for her continual encouragement and support, and for typing the first draft of this thesis.

The project was supported financially by the Science Research Council, and the British Post Office kindly lent much of the essential apparatus required for the research. Thanks are also due to the Science Research Council and the British Post Office for providing a maintenance grant for the author.

Chapter 1

Introduction

The introduction considers some of the possible advantages of a videotelephone system. It also reviews some of the published experiments that have been performed to investigate the relative effectiveness of different modes of communication. This is followed by a brief resumé of the objectives and extent of the author's field of study. The chapter ends with a summary and preview of later chapters.

1.1 The possible advantages of a videotelephone system

For several years people have realised that in many circumstances the telephone is not as satisfactory as a face-to-face meeting, and the main reason for this has often been assumed to be the lack of a visual channel. Not long after Bell's invention of the telephone in 1875, there were many proposals for methods of 'distant vision'. Since then, many science fiction novels and films have included the idea of a system enabling persons to see as well as to talk to each other, i.e. a videotelephone. There seems to have grown a widespread belief by members of the lay public, as well as some engineers, that a videotelephone would have many advantages over a normal telephone.

The first working demonstration of a videotelephone link was in 1927, when a two way audio and visual communication link was set up between New York and Washington (IVES, 1927). This, and later demonstration models of the videotelephone, were impracticable because they required bulky terminal equipment, as well as special illumination of the users. However, present technology, particularly in the field of camera tubes (BUCK, 1968) and semiconductors, now make it possible to produce unobtrusive, robust and reliable terminal equipment, which operates without the need for special lighting. In America, Bell Laboratories have already developed a videotelephone, named Picturephone Mod II, and have given it extensive field trials. Many other national or commercial telecommunication organizations have developed, or are developing, various forms of videotelephone. Although the many different forms of videotelephone that have been developed have very different image formats and video signalling parameters, they all have one factor in common, and that is that

they all require a very large bandwidth for the transmission of the video signal, compared to the audio signal. Typically, the video signal requires 100 - 300 times the channel bandwidth used by the audio signal. This, of course greatly increases the cost of transmission and the question arises as to what is gained from using the video channel at this increased cost.

Many different advantages have been claimed for the videotelephone. It has been suggested that the videotelephone would be a good substitute for many of the meetings that now take place for reasons of business, administration, sales, etc; thus, in the future there would be 'video executive board meetings' (HEALY, 1968). If the introduction of the videotelephone did obviate much of the need for travelling to meetings, it could bring about immense social changes by allowing the de-centralization of office locations. The Electronic Industries Association, in America, have estimated that by 1980 the introduction of videotelephones will produce a saving of $\$ 6 \times 10^9$ per annum on domestic air travel, and $\$ 6 \times 10^9$ on highways, as well as other appreciable savings in police protection, postal costs and recreational costs, producing a total saving of $\$ 50 \times 10^9$ per annum (SODOLSKI, 1969). Even if these estimates were correct, they do not indicate that there would be any reduction in the amount of travelling to meetings, as with the introduction of videotelephone users may simply attempt to participate in many more meetings.

Many of the claims for the advantages of the videotelephone might be modified if there was a greater understanding of the essential differences between face-to-face meetings and telephone conversations. Naturally, the telephone (and also the videotelephone) will not permit the various forms of tactile communication behaviour that exist in face-to-face encounters (e.g. shaking hands, kissing), but this form of behaviour is not so frequent in Western cultures, as in some African and Asian societies. Also, there are important social differences between face-to-face meetings and telephone conversations. Because a face-to-face meeting may have been arranged in advance to take place at some

mutually convenient time, and may also involve one or more participants in the expenditure of time and money in visiting the other, such a meeting will be regarded as more serious and the participants more committed. In contrast, the telephone is something that intrudes at inopportune moments; it is used for brief enquiries; and does not require much effort from the user (REID, 1970). Another advantage of the face-to-face meeting over a telephone conversation is that a visitor will be able to judge the importance or status of the person he is visiting. For example, a visitor may be shown into a waiting room by a secretary, thus giving an indication of the status of the person he is visiting. A visitor may also infer certain personal attributes about his host from his surroundings (e.g. untidiness).

Although the videotelephone will not be able to provide the user with all the information that is available from a face-to-face meeting, it will provide him with more information than does the telephone. The problem, then, is to evaluate this information. One direct approach is to compare the effectiveness of interactions using the videotelephone with the effectiveness of telephone conversations. For example, it would be useful to study the relative ability of persons using a videotelephone as against a normal telephone, to tell truth from lies, to comprehend information and to produce acceptable, negotiated settlements.

Before embarking on experiments to compare videotelephones and normal telephones, it would be worth studying similar experiments that have been performed already to compare the relative effectiveness of face-to-face meetings and audio-only communication. The following section contains a brief review of such experiments.

1.2 Effectiveness of different modes of communication.

Before studying the relative effectiveness of modes of communication which do and do not allow visual communication, one might consider the function which non-verbal cues may serve in a social interaction.

Non-verbal signals may simply be used as an adjunct to the verbal signal, or they may be a necessary qualification to the verbal utterance. ARGYLE (1969) suggests that there are six main ways that non-verbal cues are used in face-to-face meetings. These are as follows:

- 1) Mutual attention and responsiveness. During a conversation, 'there must be a continuous evidence that each is attending to and corresponding to the other'. The main non-verbal cues used for this are direction of gaze, head nods, and gestural reciprocity. Although a person may be aware of these cues emitted by the other person, he may not realize what signals he is sending himself.
- 2) Channel control. Direction of gaze and head nods are used in determining who shall speak and for how long. Interactors are not usually aware of sending or receiving these cues.
- 3) Interpersonal attitudes. Attitudes, or intentions of threat and sexual desire etc., are often indicated by spatial proximity, posture, and tone of voice; such cues are often very important, as attitudes may not be apparent in polite or carefully controlled verbal messages.
- 4) Illustration. Speech may be accompanied by movements of the hands which are used to illustrate what is being said or for pointing to objects.
- 5) Feedback. So that a person may plan his next utterance, he needs to know how his utterances are received, e.g. to see if the other is surprised, agrees, is annoyed, or understands. The information is mainly obtained by studying the other's face.
- 6) Emblems. This is the name given to arbitrary gestures, which are used in place of words, e.g. head shaking. They are often used in circumstances where normal speech is difficult.

Argyle goes on to suggest that one of the main differences between verbal and non-verbal communication is that verbal signals are most useful for discussing objects, person's absent and abstract ideas, whereas non-verbal cues are better for communicating 'emotions, interpersonal attitudes and other aspects of the immediate interaction'. Thus, if the purpose of the conversation is only to obtain some form of factual information or data, it would be expected that a face-to-face meeting would have little advantage over a telephone conversation. This fact has been demonstrated with an experiment by the Joint Unit for Planning Research, (REID, 1970), who showed that subjects were able to understand the salient points of a business letter just as well when the letter was read to them over a telephone, as when the reader was clearly visible.

It would be expected, therefore, that face-to-face meetings (and by implication the videotelephone) would have an advantage over the telephone when it was important that the conversants be able to gauge each other's attitudes or reactions to certain proposals. However, even in these circumstances, the telephone is not drastically inferior, as many non-verbal cues, such as the pitch or stress of the speaker's voice, may often be sufficient to indicate his attitudes or reactions.

Because of this, it is not surprising that the few experiments performed to compare face-to-face conversations with those where the conversants could not see each other, have found little difference in the subject's communication behaviour or in the relative effectiveness of the two modes of communication. For example, COOK (1970) compared the behaviour of subjects who conversed face-to-face with that of subjects who conversed via a microphone/loudspeaker system. Apart from the utterances being shorter and there being fewer interruptions in the no-vision condition, the mode of communication produced no significant differences in the number of 'filled pauses' (e.g. "ums" and "ers"), the number of attention signals (e.g. "mm", "yes"), the speech rate or the feeling of comfort as indicated by a subjective rating scale. ARGYLE, LALLJEE and COOK (1968) performed a series of experiments to compare conversations under conditions of varying visibility. (The

visibility was altered by conversants wearing either dark glasses, face or body masks, or being separated by one way, or opaque, screens.) It was shown that subjects found no difficulty in interacting if the visibility was reduced symmetrically (e.g. both wore dark glasses or both wore masks) but subjects did find difficulty in reacting, and felt less comfortable, if one conversant could see more of the other conversant than the other could see of him.

Two exploratory experiments which are particularly interesting were carried out by the Joint Unit for Planning Research (REID, 1970). In the first experiment, subjects were asked to assess the suitability of a candidate for a travel scholarship. Eleven subjects interviewed a confederate face to face, and another eleven interviewed him by telephone. The subjects had to judge the confederate on twelve different rating scales (e.g. mature/immature, reliable/unreliable etc.). It was found that for each of the twelve rating scales there was no significant difference in either the mean or variance for judgements from the face-to-face interviews and those from the telephone interviews. However, the face-to-face interviews had a significantly higher confidence in their judgements. The second experiment was concerned with a person's ability to tell if another is lying, when he can and cannot see the other person. The experimental procedure was to instruct a person to speak for approximately five minutes about his recollections of school or college days. The five minutes of speech were broken into ten 30 second periods in some prearranged order. (When lying, the speaker had to describe himself as a more successful student than he was in reality.) Two other subjects, who were listening, had to judge which of the periods were lies and which were truth. One of the subjects listening, sat at the opposite side of a table to the speaker, while the other sat in another room and listened via a loudspeaker. There was no significant difference in the ability to judge truth or lies by the subjects who could see the speaker and by those who could not. But again, the subjects who could see the speaker were significantly more confident of their judgements, and in this particular case, it was obviously an unjustified confidence.

It is possible, therefore, that for certain types of interaction a person may have an unjustified confidence in some decisions or judgements made during a videotelephone conversation, compared with similar judgements made during a normal telephone conversation. In consequence, care must be taken when assessing subjects' (or users') judgements, as to the relative merits of videotelephones and normal telephones.

In the author's knowledge there has been only one experiment that has directly demonstrated a difference between communicating face-to-face and communicating via audio links. In this experiment by MORLEY and STEPHENSON (1969), pairs of subjects were used to simulate a wage negotiation using face-to-face communication and via a telephone system. For each pair of subjects, one was cast in the role of management representative and the other as a union representative. Both subjects were given explicit instructions as to what would be acceptable wage settlements for the sides that they each represented, as well as general background information on the simulated dispute. It was found that the management, which had the stronger case, won more negotiating victories via the telephone, than face to face. Morley and Stephenson argued that their findings are due to the difference in formality associated with the two modes of communication. Thus, using the more formal mode of communication (the telephone) there is more emphasis on the inter-party communication, whereas face-to-face allows greater emphasis on the more human, interpersonal communication.

The above experiment is similar to one by SMITH (1969) who compared the different outcomes of negotiations carried out by pairs of subjects face-to-face with those where the subjects were limited to written communication only. Each subject was cast in either role of plaintiff's attorney, or defence council, in a suit by an individual against a drug company. The results showed that there was a greater dispersion of settlements under the face-to-face negotiation, and it was argued that this was due to a 'successful attorney' being able to push the settlement further in his preferred direction when face-to-face, compared to when

restricted to written communication only. A questionnaire was given to the subjects in order to determine their satisfaction with the settlement that they obtained. No difference in satisfaction was found between the two modes of communication. This experiment was therefore, an example of where an expanded opportunity for communication did not bring greater satisfaction with the agreed terms for a negotiation, and one may conjecture that there are other forms of interaction where an expanded opportunity for communication (i.e. videotelephone versus telephone) would not necessarily bring about greater satisfaction.

None of the experiments referred to in this section dealt with face-to-face meetings where the participants needed to study jointly detailed diagrams or documents. Naturally, in these circumstances face-to-face meetings are superior to telephone conversations. Unfortunately, it is unlikely that the videotelephone will be of any advantage over the telephone in these circumstances, as the spatial resolution of the videotelephone that will be available in the near future is not enough to allow documents to be easily read.

It would be wrong to automatically conclude from the experiments mentioned in this section that there was very little difference in the effectiveness of face-to-face meetings and telephone conversations, as all that the experiments may demonstrate is that the simple measures and experimental procedures that have been used to date are inadequate for comparing two very complex forms of social interaction.

Some of the questions as to the effectiveness of videotelephones may be best answered by simple field trials. Ideally, such field trials should study as carefully as possible any change in a person's overall communication behaviour, (i.e. travelling, letter writing and telephoning), when able to use a videotelephone; and the field trials should not rely just on the user's subjective opinions of the videotelephone system. This of course does not mean that the user's subjective opinions should be ignored, as they are very important to organizations who are promoting or selling videotelephone systems.

We may conclude, therefore, that it may be difficult to demonstrate in the laboratory any improved effectiveness of communicating via a videotelephone compared to communication via the normal telephone, but that a difference may be demonstrated by using field trials. When this data on the relative effectiveness is available, planners will be able to make more realistic estimates of what influence the introduction of the videotelephone will have on travelling and location of office centres.

1.3 The author's field of study

When conversing over a videotelephone link, it must be assumed that the conversants will consciously or unconsciously try to use many of the non-verbal cues, which are normally used in face-to-face meetings. It is by allowing the conversants to see each other, and thus enabling them to use these non-verbal cues, that the videotelephone has any advantage over the standard telephone.

Non-verbal cues are often utilized during a conversation without the conversants being aware that they are using them. The use of non-verbal cues may be considered a social skill (and as such, may be improved by training, (ARGYLE, 1969)). If a videotelephone is to allow the conversants to employ their communication skills to their full potential then the videotelephone should allow the use of as wide a range of non-verbal signals as possible, and introduce as little distortion as possible into the perception of these cues.

The research work reported in this thesis was undertaken with the object of studying how a videotelephone may affect the perception and use of some of the cues that are often used in face-to-face meetings. The results of some of the experiments may prove useful to engineers who are concerned with the 'human factors' aspects of videotelephone design. The results, and also the method used to record the conversants' behaviour, may also prove useful to the many psychologists engaged in studying the use of non-verbal cues in a social interaction.

Before embarking on the series of experiments described in this thesis, it was necessary to know the non-verbal cues that are used in face-to-face meetings. Fortunately, there already exists a body of relevant work, which is specifically about the role of visual signals (i.e. non-verbal cues) that are used in social interactions. Much of the research that has been carried out was motivated by the desire to make certain social interactions more effective, and the results have been used in the training of professional social skills such as interviewing, teaching or supervision. Other research has been done by psychiatrists, who are interested in abnormal communication behaviour in social interactions.

Some of the visual signals that are used in social interactions are facial expressions, posture, gesture and direction of gaze, as well as the relative proximity and orientation that people choose. Chapter 2 contains a review of the literature on these non-verbal cues.

The author has limited himself to studying the enhancement of telephone conversations for two person interactions (dyadic interactions.) For example, during the experiments performed to investigate the perception of direction of gaze, only one facial image appeared on the video monitor screen. However, many of the conclusions drawn from the results of these experiments are applicable to video systems designed to allow more than two persons to see each other.

It would have been interesting to study how varying the many different picture parameters of a videotelephone would influence the user's perception and use of non-verbal cues. The results of such a study may have some bearing on the choice of picture parameters for commercial videotelephone systems. Such a study is beyond the scope of this thesis, and because of the high cost of variable parameter two way video systems, this type of study may only be carried out by comparatively large organizations. The author limited himself to using only two types of video image, namely a high fidelity 625 line monochrome image and a very simple 'cartoon' image. (The reason for choosing these two types of image are given in section 3.6).

Some organizations have performed subjective tests on videotelephones, which are exactly similar to the type of subjective tests used for studying broadcast television images. However, if the videotelephone is simply assumed to be a 'television' image added to a telephone channel, then erroneous conclusions may be reached as to the optimum choice of picture parameters. Thus, if the type of test is restricted to subjects observing video images, which have different picture parameters or impairments added, it may be concluded that a narrow picture format would be advantageous as it would eliminate unnecessary background, (ALLNATT, 1968). However, if a narrow picture format is chosen the videotelephone user is much more likely to move out of the field of view of the camera. Therefore, when choosing such parameters, it is essential to have subjects conversing over a link, so that the amount of body movement may be taken into account, (see sections 3.4 and 3.5.)

To stress this point further, an analogy may be made with the methods used for studying the impairments encountered on telephone circuits. When studying the effects of these impairments, they are often classified into talking effects, listening effects and conversational effects, (RICHARDS and HUTTER, 1968). In particular, the type of test

used by telephone engineers to study listening effects may be thought of as tests analogous to those used by television engineers to study the impairment or changes in picture parameters of broadcast video images. Just as telephone engineers have not limited themselves to listening tests, so should the videotelephone engineers not limit themselves to viewing tests.

1.4 Contribution of this thesis.

The purpose of this short section is to indicate precisely the original work that is reported in this thesis.

The experiments concerned with the perception of direction of gaze, (section 4.1), and with lip reading, (section 5.1), are modifications of published experiments. In the original published experiments, the stimulus was a person who could be directly observed. The originality of the author's experiments is that the stimulus source was a facial image on a video monitor screen. Furthermore, some of the video images in the lip reading experiment were of a novel form.

The exploratory experiment described in chapter 6, which dealt with the perception of facial images, used an established technique for comparing the perception of visual forms; however, this technique has not been used before in conjunction with facial images. The second experiment described in chapter 6, which is concerned with the apparent proximity of a video image, is to the author's knowledge completely original.

In the experiments in which subjects converse over a video link, (see chapter 7), many of the measures used e.g. duration of looking and eye contact etc., and the method used for stimulating conversation are the same as those used in published experiments, which have investigated subjects' non-verbal behaviour in face-to-face interactions. Again, the

originality of the author's experiments was the measurement of subjects' looking and speech behaviour while they were conversing over a video link. Mr. B.G. Champness suggested comparing the chance and measured amounts of eye contact but the author further developed the probabilistic model used for estimating the duration of chance eye contact.

In many of the experiments reported in this thesis cartoon images were used. The suggestion for using this type of image came from the author's supervisor, Professor E.C. Cherry.

1.5 Summary and preview of later chapters.

Chapter 2 reviews some of the literature that has been published on conversation and non-verbal cues. Emphasis has been given to those non-verbal cues used in face-to-face meetings, which may (or should) be available to users of a videotelephone.

Chapter 3 contains a wide-ranging discussion on the video image that may be used for a videotelephone. The topics covered include video image coding, the picture parameters of existing videotelephone schemes, picture size and viewing distance, and the control of a viewer's movement. Also included is a description of the types of video image used in the author's experiments, and the method of generating the 'cartoon' image.

Chapter 4 describes the audio, video and data processing equipment used in the author's experiments. It also contains an account of an experiment performed to investigate the perception of direction of gaze when using a videotelephone. The results of the experiment were used to determine the most suitable positioning of the camera, relative to the video monitor in a videotelephone terminal.

Chapter 5 describes an experiment performed to investigate the usefulness of different forms of video images for lip reading. The experiment was similar to the frequently cited experiment by Sumby and Pollack. The chapter also contains a brief review of the literature on bisensory perception.

Chapter 6 gives an account of two simple exploratory experiments performed to investigate: a) the identification of a facial image using different forms of video images, b) the apparent proximity of a facial image on a video monitor screen.

Chapter 7, unlike earlier chapters which deal with the perception of non-verbal cues, describes three experiments performed to investigate the use of non-verbal cues during a videotelephone conversation. The experiments were concerned with the subjects' 'looking' behaviour and some of the simple measurable aspects of their speech behaviour. The subjects' looking behaviour was analysed to find out whether it was influenced by the 'looking behaviour' of facial images on the video monitor screen of the videotelephone terminal.

Chapter 8 gives a general summary of the separate conclusions to be found in many of the different chapters.

Chapter 2.

Verbal and Non-Verbal Signalling

This chapter reviews previous work on conversation and the various signs that are used in conversation. Emphasis is given to those non-verbal signs that may be available to a video telephone user. Virtually all the work reported has been carried out by psychologists or social psychologists, but an engineer, who is involved with the human factors aspect of the design of video telephone systems, should be acquainted with the possible human signalling that may be used with video telephones. He may then consider what effect the various limitations of any particular video telephone system will have on the communicants.

2.1. Conversations and Signs

When two persons are engaged in a conversation they will communicate with a wide variety of sign activity. In fact, as CHERRY (1957) points out, there can be no communication without a system of signs. The problems of the experimenter who is observing two subjects in conversation, is to decide what sign activity is relevant. An essential prerequisite in solving this problem is to define a 'sign'. The philosophical question of determining what is a sign has been answered by the logician, Charles Sanders Peirce, who emphasised the triadic nature of the sign situation. Pierce defines a sign thus:

a sign - is something which stands for something, to somebody,
for something else, in some aspect or capacity

(See GALLIE, 1952)

This definition demonstrates the triadic nature of a sign situation, i.e. sign, standing for object, to interpretant. A logical consequence of the definition is that every sign requires another sign to interpret it, and that every sign has the potential of evoking an endless series of further interpretant signs.

Thus, in a conversation, a non-verbal signal is any variation in one communicant's performance which the other person interprets as a sign. The person emitting the sign may do so intentionally or unintentionally, and may or may not be aware of emitting a sign. For example, a participant in a conversation may drum his fingers on a desk top, and the other person may interpret this as a sign of the other's impatience. The person drumming his fingers may not do so intentionally and may not be aware that he is doing it.

NEWCOMB (1953) proposed that the simplest possible communication act may be described in terms of his A, B, X system (See Figure 2.1.), where A transmits information to another person, B, about something, X. The arrows in the figures represent attitudes (or orientations) towards the perceptual object at the arrow head. Newcomb also proposed that communicative acts usually result in increased symmetry.

An interaction between two persons may be of a very complex nature and the constraints upon the participants' behaviour may change from one moment to the next. The interaction may range from a casual exchange of pleasantries to a very intense intellectual discussion. Two persons engaged in a conversation may have different or even conflicting motives for doing so, yet they will try to establish a 'working consensus' during their interaction (GOFFMAN, 1956). Many different theories have been used to account for the pattern of behaviour during an interaction between two people. One such is game theory, which analyses interaction in terms of their outcomes (BOSTRUM, 1968). It is, of course, extremely difficult to use mathematical models for such a complex form of behaviour as a conversation. The problem of describing types of interaction has been tackled by BALES (1950), who used 12 categories of verbal utterances, although non-verbal signals were taken into account.

Throughout this thesis, only conversations where the speakers interact will be considered; i.e. where one person's utterance acts as a stimulus, eliciting a response from the other person, which in itself then becomes a stimulus.

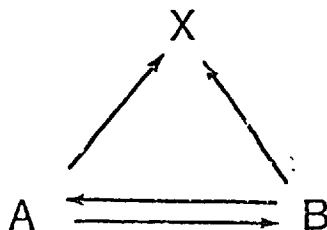


Figure 2.1. Newcomb's ABX system for minimum communication acts.

2.2. Verbal and Non-Verbal Signs

When two people are engaged in conversation, both verbal and non-verbal signs are used. By definition, verbal signs are based on phonetics and often have a written form; all other signs are classified as non-verbal. Non-verbal cues have a variety of different uses in face-to-face meetings. ARGYLE (1969) suggests that there are six main ways that they are used, and these are given in section 1.2.

RUESCH (1966) has listed many of the similarities and differences between verbal and non-verbal codification and spatiotemporal characteristics. In certain circumstances, it may be much easier, or even mandatory, for a person to indicate his intentions with non-verbal signs, as opposed to verbal signs. For instance, GOFFMAN (1959), mentions verbal taboos - thus, a person may not say "I am bored with this conversation, I wish to terminate it", but may indicate this by various gestures. ARGYLE et al (1970) showed subjects a video recording of a confederate who read three different passages in three different tones of voice and accompanying facial expressions, indicative of superior, neutral and inferior attitudes. The verbal content of these passages was also arranged to be superior, neutral and inferior. Subjects were asked to rate the various parts of the recording on superior-inferior and hostile-friendly scales. The results showed that the non-verbal cues had much more impact than the verbal ones.

Another difference between verbal and non-verbal signals is the level to which they commit a person in an interaction. Because a verbal utterance has the possibility of being written down, and thus

recorded, care must be taken over utterances in certain situations; it is not possible to 'unsay' an utterance. It is much easier to dispute the interpretation of many ambiguous non-verbal signals.

DAVITZ (1964) asked subjects to both recognise the emotional meaning of utterance and to express various emotions, while reading neutral passages. The results indicated that there are large individual variations of emotional sensitivity, and that persons who could recognise emotions well were also better able to express them; a person who is accurate in identifying the emotional meaning of vocal expressions also tends to be accurate in identifying facial expressions.

The following sections in this chapter review the literature on the non-verbal signs which are made possible by the addition of a visual channel to the normal telephone channel, and also on any possible changes in the forms of speech used. Specifically, the sections are devoted to kinesics, facial expression, direction of gaze, proximity, posture and orientation, physical appearance, and speech. Because of the nature of the visual telephone, tactile communication, such as the shaking of hands, will not be considered; neither will abnormal human sign activity be considered. An extensive review of non-verbal and verbal signals has been carried out by VINE (1970). In his review, both human and subhuman signalling is covered, as well as normal and abnormal sign behaviour in humans.

2.3 Kinesics

The term kinesics was first introduced by BIRDWHISTELL (1952) to cover the study of visible body motions; he also invented a notational system for recording minute facial and postural movements. BIRDWHISTELL (1960) introduced three different levels for the study of kinesics - prekinesics, dealing with the physiological basis of body motion; micro-kinesics, dealing with the derivation of kines (the least particles of isolatable body motion); and social kinesics, which deals with the function of these elements in communication. For example, he found that subjects could discriminate eleven positions of eye lid closure, but only four with differential meaning. He also considered that no body motion is meaningless as it will always have significance at one or more analytic levels as part of a structured pattern.

In conversations, speech is typically accompanied by kinesic markers BIRDWHISTELL (1966). These kinesic markers are synchronized with the vocal behaviour and may emphasise or modify the vocal constructions, or indicate the context of the message situation. For example, a 'pluralization marker' might be a sweep of the hand coinciding with the utterance of "you", when a person is addressing more than one person. Many such gestures are meaningless, or at best ambiguous, by themselves; they only become meaningful when the context is given or when they are accompanied by some utterance.

SCHEFLEN (1963) analysed films of a psychotherapist interviewing patients for kinesic, postural and linguistic behaviour. He found that patterns of behaviour were important and not isolated events. He later found that non-verbal behaviour had a hierarchical arrangement (SCHEFLEN 1964). The three levels of this hierarchical structure were - the point, position and presentation. The smallest unit, the point would typically be a shift of the head and this would be held for a few sentences; the next largest unit, position, is a gross postural position and this indicates the contribution that each individual is making in the group, e.g. sitting back in a chair; whereas, the presentation was the totality of one person's position, such as leaving the room.

2.4. Facial Expression

Some of the first work on facial expression was done by BELL (1806) who performed basic analysis of the facial musculature. DARWIN (1872) studied the facial expressions in both man and animal. Darwin regarded facial signals as primarily of an emotional and involuntary kind i.e. emotional expression is innate and reflects the biological structure of the autonomic system - producing baring of the teeth, weeping, etc. The evidence that he used to support this argument was that other races used similar expressions (evidence obtained by correspondence with Europeans in contact with aboriginal races) and, that animals also had similar facial expressions.

Mead criticized Darwin by pointing out that animals do not undertake to express their emotions for the benefit of other animals (MEAD 1934). Similarly, human facial gestures need not express the content of the individual's mind.

A great deal of work has been done on the recognition of facial expressions. One of the problems in this field was that many authors assumed that there were standard expressions for 'fear', 'joy' etc., and even complex emotions such as 'pity'. Thus, agreement among judges varied tremendously and varied for different expressions. A review of the early work in this field was done by WOODWORTH (1938). He also suggested that facial expressions should be ranked in order of similarity into six groups, or scale steps:

- 1) love, mirth, happy, 2) surprise, 3) fear, suffering,
- 4) anger, determination, 5) disgust, 6) contempt.

SCHLOSBERG (1952) developed Woodworth's linear scale into a circular scale. He found that the most satisfactory categorization of facial expressions was in terms of two independent dimensions, labelled pleasantness - unpleasantness, and attention - rejection. The average

error in his recognition experiments was approximately one half of a scale step. Later, (SCHLOSBERG 1954) a third independent dimension of activation was added which ranged from sleep to tension. OSGOOD (1966) also found three similar dimensions - Pleasantness, Activation, and Control; and by means of cluster analysis, suggested that there are 7 - 10 basic regions.

Much of the early work on facial expressions used photographs as a stimulus source. These photographs of frozen and sometimes exaggerated facial expressions were very unnatural. FRIJDU (1953) compared cinefilm strips of 2 - 15 seconds duration with still photographs presented for 10 seconds. It was found that the average accuracy of recognition improved by approximately 50 percent when the moving cine image was used compared with the photographs, thus showing the importance of dynamic cues in recognition.

FRIJDU (1958 & 1959) showed that the context may alter the judgements of facial expressions. He presented a series of four photographs to subjects and gave them different descriptions of the situation in which the photograph had to be judged. The situational cues sometimes led to quite opposite impressions of the expression in any particular photograph.

More recently, ethologists BRANNIGAN and HUMPHRIES (1969) have studied facial expressions in young children. They propose that many of the elements of facial expressions are probably innate, as are many of the signals of animals - an argument the same as Darwin's.

2.5. Direction of Gaze

In the following paragraphs 'eye contact' is defined as mutual gaze between two persons in the region of the eyes, and 'looking' as one of the participants in the interaction gazing at the other.

Psychologists first started experiments on the direction of gaze ten years ago; however, poets and novelists such as Shakespeare and Jean-Paul Sartre were aware of the importance of eye gaze in human relationships. (See CHAMPNESS 1970). An extensive review and bibliography of the work on gaze direction is to be found in M. Argyle's book, "Social Interaction" (ARGYLE 1969). Gaze direction is an important non-verbal signal used during conversation. ARGYLE & DEAN (1965) suggest that eye contact may have the following functions:- to obtain feedback at the end of an utterance of the other participants reactions, and also to signal whether the channel is open or closed, i.e. to help to synchronize utterances. Also some people desire to be seen and eye contact is the proof to them of being seen. Eye contact will influence the establishment and recognition of social relationships, thus, prolonged eye contact may have sexual overtones and cause embarrassment.

KENDON (1967) collected data on eye contact and looking from pairs of people engaged in conversation, by analysing films taken at two frames per second. The general results were:- the participants look at each other intermittently, usually between 28 - 70% of the time although variations of 0 - 100% were recorded; these glances vary in length, normally between 1 - 7 seconds. While participant A is looking at the other member, B, of the dyadic interaction B will only look back part of the time to produce eye contact. This eye contact is typically 10 - 30% of the total time, each eye contact being of short duration (about 1 second). Also, there is more looking by the person who is speaking and the person who is listening will look at the speaker for longer periods than he looks away. If A is about to speak, he will look away; this is probably to avoid distraction while planning or organizing his utterance. At the end of a natural phrase, A, the speaker, looks briefly at B, and at the end of the utterance there will

be a prolonged gaze. This prolonged gaze is mainly to obtain feedback on the other's response i.e. to see if A still has B's attention or if he can carry on speaking. The general pattern, therefore, is the speaker looking away and the listener looking at the speaker, while at the end of an utterance there is eye contact.

KENDON & COOK (1969) examined the gaze behaviour of 11 subjects interacting with each of 4 other subjects. Fifteen different measures were analysed e.g. frequency, mean length, and percent of the total time spent looking while listening and while talking. Subjects were found to have consistent gaze and utterance patterns, although with some of the measures, the identity of the person with whom they were interacting did have some effect.

In the two experiments mentioned above, both subjects in the dyadic interaction were observed for gaze patterns. However, in many of the early experiments in gaze directions (e.g. ARGYLE & DEAN 1965, EXLINE et al 1965, EFRAN 1968) only one subject was observed for direction of gaze, while the other member of the interaction was a confederate who was instructed to gaze continuously at the other person, see Figure 2.2.

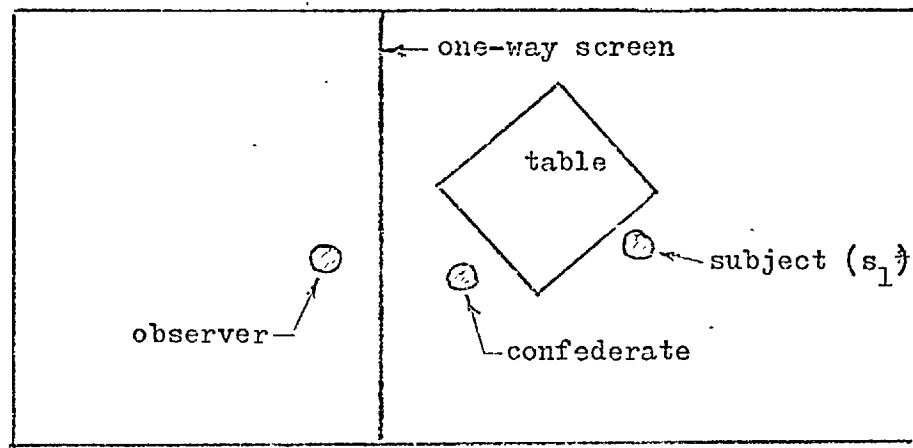


Figure 2.2 The laboratory arrangements often used for studying direction of gaze, (ARGYLE, 1969).

Thus, whenever the genuine subject looked into the confederate's face there was eye contact. This leads to a greater amount of eye contact than would usually be encountered in normal situations. * ELLSWORTH & CARLSMITH (1968) found that with positive verbal content, frequent eye contact produces more positive evaluations of the confederate by the subjects (both females), while, with negative verbal content, frequent eye contact produces a negative evaluation. Similarly, KLECK & NUESSELE (1968) showed silent films of a confederate engaged in conversation, the confederates having either high or low amounts of eye contact (80% or 15%). Subjects judged the high eye contact confederates to be friendly warmer, pleasant etc.

Many different aspects of the interaction will influence the amount of looking and eye contact, as well as various personal attributes of the participants. For instance ARGYLE & DEAN (1965) found that as the distance between subjects in a dyadic interaction increased, the amount of eye contact also increased. Subjects (S_1) sat at 2, 6, and 10 feet from a confederate (as in Figure 2.2.), who was instructed to gaze continuously at the subject. The experiment was repeated by GOLDBERG et al (1969) except that the confederate used a more natural gaze behaviour, and again they found an increase in eye contact with increased separation of subject and confederate. Argyle & Dean postulated an affiliative-conflict theory to account for this change in subject performance. However, STEPHENSON & RUTTER (1970) carried out a similar experiment, the difference being that a confederate sat in the position (S_1) normally taken by the subject. The confederate carried out a sequence of gaze behaviour, looking at the subjects eyes, ears and shoulders. They found that as the separation between the subject and the confederate increased, and hence the distance between observer and confederate increased, then observers increasingly recorded the confederate's gaze, directed at the subject's shoulder and ears, as eye contact. This shows that the Argyle and Dean results are possibly due to observer performance and not subject performance.

Results of some of the other main experiments concerned with gaze direction may be summarized as thus: Subjects look more at people they like (EXLINE and WINTERS 1965). When talking in front of two confederates, a subject will look more at the one with whom he has previously had a friendly conversation (EFRAN and BROUGHTON 1966). Subjects will look more at a confederate who is of apparently higher status when receiving approval (EFRAN 1968). Subjects look less during conversations with embarrassing content (EXLINE, GRAY and SCHUTTE 1965). Females engage in more eye contact than males (EXLINE 1963, EXLINE, GRAY and SCHUTTE 1965, KENDON and COOK 1969). Argyle and Dean found that the amount of eye contact depended also on the sex of the subject's partner; however, Kendon and Cook found interaction between the sex of the subject and the sex of the partner. Recipients of extended looks judged the looker to be more dominant than when the looker gave brief looks (THAYER 1969). Subjects look less when they have been implicated in deception and when they are lying. Unless they scored high on a Machiavellian* rating (EXLINE^{etal}/1961), people high in affiliation motivation (measured by a FIRO questionnaire) look more in non-competitive situations than those low in affiliation motivation (EXLINE 1963, EXLINE, GRAY and SCHUTTE 1965).

STRONGMAN and CHAMPNESS (1968) found that for a group of 10 subjects, the dominance hierarchies were positively related to eye contact, looking, and looking when uttering. They also made an interesting comparison between the actual amount of eye contact and the amount that would be expected if the two members of the dyadic interaction were just 'looking robots', i.e., the subjects do not interact and the eye contact just arises from chance. The chance eye contact from looking robots was calculated from the formula:-

$$\text{expected eye contact} = \frac{(\text{A's looking time}) \times (\text{B's looking time})}{\text{total time}}$$

* A Machiavellian person enjoys manipulating others for his own purposes; is artful, cunning, subtle and diplomatic.

They found that the actual eye contact measured was between 22% and 110% greater than would be expected from chance robots. CHAMPNESS (1971) has extended the above formula, by taking into account the amount of time that each member spends talking and the amount of looking and talking etc. He found that when subjects disagreed strongly, the amount of looking increased. Thus, if they were 'looking robots', the amount of eye contact would increase, but in fact, the measured amount of eye contact was less, thus demonstrating that the subjects did interact as far as a gaze direction is concerned.

An important aspect of all work on eye contact is of course the very perception of eye contact. GIBSON and PICK (1963) showed that people are remarkably sensitive to another's direction of gaze. A further discussion on this is to be found in Chapter 4.

One of the reasons that gaze direction has been so prominent in recent literature on non-verbal communication (and in this thesis) is that compared with facial expressions, posture, etc., it is much easier to measure and thus more amenable for experimentation.

2.6. Proximity, Posture and Orientation

When two persons are engaged in a conversation they will tend to choose a certain relative orientation and separation. The particular orientation and proximity will depend on the location and various personal attributes of the conversants.

SOMMER (1961) asked pairs of subjects to sit down in a room which had various seating arrangements. He found that five feet between subjects (3 feet between chairs) was a comfortable distance for conversation when they were sitting opposite each other. If the available 'opposite' chairs were more than $3\frac{1}{2}$ feet apart then subjects preferred to sit side by side. ARGYLE & DEAN (1965) noticed during an experiment on the variation of eye contact with proximity, that when the subject and confederate were sitting very close together (2 feet apart), the subject tried to lean backwards in order to increase the separation between them, whereas, when they were 10 feet apart the subjects attempted to lean forward in order to decrease the separation. The proximity that persons choose for comfortable conversation will depend on the cultural background. WATSON & GRAVES (1966) found that Arabs tend to sit much closer together than do Americans. SOMMER (1962) also pointed out that the proximity that people choose will also depend on the particular meeting place. Seating tends to be closer in larger rooms or halls than in the more intimate surroundings of smaller rooms.

SOMMER (1965) also found that the orientation that the subjects chose depended on the type of interaction. When pairs of subjects (American) were asked to sit at a rectangular table they tended to sit at opposite sides in competitive situations, side by side in cooperative situations, far apart when coacting, and together at the corners during casual conversation. COOK (1970) replicated this experiment with British subjects who were both university and non-university subjects. He found that the British subjects tended to sit further apart during casual conversation than did Sommer's American subjects. SOMMER (1959) also found that when subjects

were seated around a rectangular table, most interaction occurred between the adjacent subjects at the corner of the table. The sitting position that subjects choose at a table will depend on the position of the leader of the group, (SOMMER 1961).

MEHRABIAN (1968) asked subjects to talk to a hat rack. The subjects stood closer when they imagined the hat rack to be someone that they liked, compared with someone that they disliked. When ARGYLE and DEAN (1965) told subjects to stand at a comfortable conversation distance from a cut out, life sized photograph, they stood closer if the stimulus photograph was of someone with their eyes shut, compared to eyes open.

Although the literature referred to above demonstrates that subjects prefer to carry on their conversation at some particular separation, precisely what is signalled by proximity is not clear. PORTER, ARGYLE & SALTER (1970) carried out an experiment in which three confederates interacted with subjects at distances of 3, 4, and 8 feet; however, proximity did not account for any significant variation in any of the 21 rating scales, which the subjects had to fill in concerning their impressions of the confederates.

It is more certain what is signalled by posture. SARBIN and HARDYCH (1955) found that subjects made more consistent judgements about supposed emotional states from the postures of stick figures. When analysing films of group therapy sessions, SCHEFLAN (1965) showed that the posture that one person takes will be affected by the posture of the person to whom he is talking. MEHRABIAN (1968) found that when addressing an imaginary person whom they liked or disliked, subjects would adopt different postures. Mehrabian also found that when photographs of people in various postural positions were used as stimuli, subjects judged that relaxation and a forward lean of trunk communicated a more positive attitude.

When two persons converse by means of a video telephone, the non-verbal cues of proximity and posture may or may not be available. Thus, if the picture format only allows the face to be seen, these particular non-verbal signals will not be available. Even if the format does allow the whole body to be seen, the perception of posture and distance may be distorted. (See Chapter 6).

2.7. Physical Appearance

The physical appearance of a person will give the observer an indication of the person's age, race and social status. The clothes that people wear (particularly uniforms) often give evidence of the wearer's occupation. Many assumptions that are made about personal attributes from physical appearances may have little objective grounds for reasoning. For instance, THORNTON (1944) asked a group of judges to rate confederates who did or did not wear glasses for different personality traits. Confederates wearing glasses were rated more intelligent, more industrious, more honest and more dependable than those who did not. In Thornton's experiment, the judges only saw the confederate very briefly. However, ARGYLE and MCHENRY (1969) found similar favourable interpretations of intelligence from judges of persons wearing glasses, but this diminished the longer the judges looked at the wearer of the glasses.

It is not always easy to quickly guess the identity of a person speaking on the telephone, even if the speaker is well known to the listener. In video telephones, the addition of the visual channel will greatly facilitate the identification of the speaker. An experiment on the type of visual channel needed for the identification of the caller, is reported in Chapter 6.

2.8. Speech

The non-verbal aspects of speech that may convey information about the speaker's emotional state or attitudes are the pitch, the stress, the rate of speech, and speech disturbances; these are often referred to as the voice quality or the tone of voice.

For example, STARKWEATHER (1956), asked subjects to judge short recordings of speech for both aggressiveness and pleasantness. The speech recordings were in two forms - the first was a recording of normal speech, the second was rendered content free by frequency filtering (bandpass 100 -450 Hz). It was found that aggressive judgements were as reliable with the content free speech as with normal speech; the judgement of pleasantness was less reliable for content free speech. Subjects were also asked to judge typescripts of the recording, i.e. content only. Judgements of the typescripts for aggression and pleasantness were less reliable here, than those of normal and content free speech. By presenting recordings of both normal speech and content free speech to a linguist, trained in transcribing both pitch and stress in utterances, LIEBERMAN (1965) showed that the perception of pitch and stress was confounded by the verbal content of the utterances.

The study of speech disturbances, particularly pauses and interruptions, would seem fruitful when discussing the comparisons of modes of speech communications. Thus a communication channel that produces more interruptions and pauses might be deemed as less satisfactory, as it may imply that the synchronizing of the participant's utterance was more difficult, or that the very perception of the other's actions or intentions was impaired.

Speech disturbances take many forms, for instance: stutter, unnecessary repetition, incomplete sentences, slips of the tongue, and the frequent incursions of 'ums' and 'ers'. The inclusions of 'ums' and 'ers' are usually treated separately from other speech disturbances and are called 'filled pauses'. MACLAY and OSGOOD (1959) postulated that filled pauses may be used to keep control of the conversational 'ball'. The rationale was that a person has learnt that in a conversation, if he pauses silently and for long enough, someone else will take over as the talker, so he will fill his longer pauses with 'ers', or some other sound, in order to indicate that he wishes to continue speaking. BOOMER (1965) rejects this hypothesis, pointing out that there are normally more filled pauses when a person is asked to record a monologue. Subjects in experiments by KASH & MAHL (1965) used more filled pauses when they could not see each other; COOK (1970) found no such change, whereas subjects in JAFFEE & FELDSTERN (1970)'s experiments used a lesser number of filled pauses when they could not see each other.

ARGYLE, JAFFEE and COOK (1968) found that during a conversation subjects made more pauses and interruptions when the visual cues available were reduced. COOK (1970) compared the conversation of subjects who conversed face to face with those who conversed over a microphone/loudspeaker system. He found that there were fewer interruptions, utterances were shorter, and that there were more requests for repeats in the no-vision condition. Cook suggested that the unexpected result (more interruptions in the visible conditions) may be due to the subjects being more confident in the placings of their interruptions, compared with the no-vision condition where they had to take greater care.

Variation in the rates of speaking are mainly caused by variation in pausing (GOLDMAN - EISLER, 1961 A). In an experiment by COOK, ARGYLE and LALLJEE (1967), subjects engaged in a conversation said more when concealed than when unconcealed, and males spoke more than females. There has been some controversy as to the cause of the variation in pauses. GOLDMAN - EISLER (1961 B) showed that pauses occur before words of low transitional probability; BOOMER (1965) rejects these findings and demonstrates that the positions of pauses in an utterance are related to the grammatical structure. COOK (1969) has reported on experiments relating pauses to anxiety and speech disturbances.

In an experiment by MOSCOVICI (1968), it was shown that, compared with subjects who sat face-to-face, (whether visible or screened) subjects who sat back-to-back or side-by-side spoke in a style that was more like the written form. This indicated that the change in style of speech was due to the relative unfamiliarity of the latter two positions and not merely the withdrawal of vision. However, this experiment was carried out with French subjects and it is, therefore, not necessarily applicable to conversations in English.

When a visual channel is added to a telephone system, both verbal and non-verbal signs, used in a normal telephone conversation, may alter. COOK (1970), in his experiments on conversation under different conditions of visibility, hypothesized that there are verbal equivalents to visual signals, which would be used if the visual channel were closed, but he found no evidence of such substitutes.

Chapter 3

The Visual Image.

This chapter deals with the type of visual image that is likely to be encountered in videotelephone communication links. It reviews, briefly, the coding of video images and describes the video image aspects of the main published proposals for videotelephone systems. There is a discussion on the choice of picture size, viewing distance, and the important problem of keeping the users 'on camera'. Finally, it deals with the types of visual images used for the author's experiments, and how these images were produced.

3.1 Visual Communication Channels.

The purpose of an electronic visual communication channel is to provide a viewer with some form of visual representation of a distant scene. In the case of human communication by video telephone, the visual representation should be sufficient to allow each of the participants to recognise the various forms of non verbal behaviour in the other person.

The complete visual representation of a distant scene would require not only the three spatial dimensions, but also, further dimensions to represent the visible energy spectra at all points, thus, the binocular view of space would require a six-dimensional vector representation, (SCHREIBER, 1967). Standards for colour transmission of videotelephones have already been proposed (VALENSI, 1969) and Bell Laboratories have begun research into methods of producing colour for their Picturephone, (MOLNAR, 1969). However, in the immediate future, it seems that flat monochrome images will be used for videotelephones; such an image normally only requires a scalar representation.

FANO (1961) proposed a model of communication systems used for transmitting information about an analogue source. This model may be applied to visual communication systems; a schematic diagram of this model is shown in Figure 3.1. The function of the source encoder (in this model) is to find a representation of the visual world which satisfies a certain fidelity criterion. The fidelity criterion used for broadcast television is that a system should produce as an exact reproduction of the original scene as possible. In other circumstances, a different, less stringent fidelity criterion may be appropriate, leading to different source encoders, (PEARSON, 1965). An appropriate fidelity criterion for a videotelephone visual transmission system might be that

only the useful, non verbal behaviour of the interactors need be recognised.

The choice of fidelity criteria for a visual communication system will be influenced by several factors. One will be the cost of various alternatives. Another factor will depend on the estimation of the customer's requirements. Because broadcast television has been in widespread use for many years, many potential videotelephone customers may expect an image of very high quality for the videotelephone.

It has often been suggested that efficient use would be made of a visual communication system, if it were 'matched' to the human viewer. Attempts to put a numerical value on the information capacity of the human eye when considered as an information 'sink', have been made by JACOBSON (1951), KELLY (1962) and BUDRIKIS (1964).

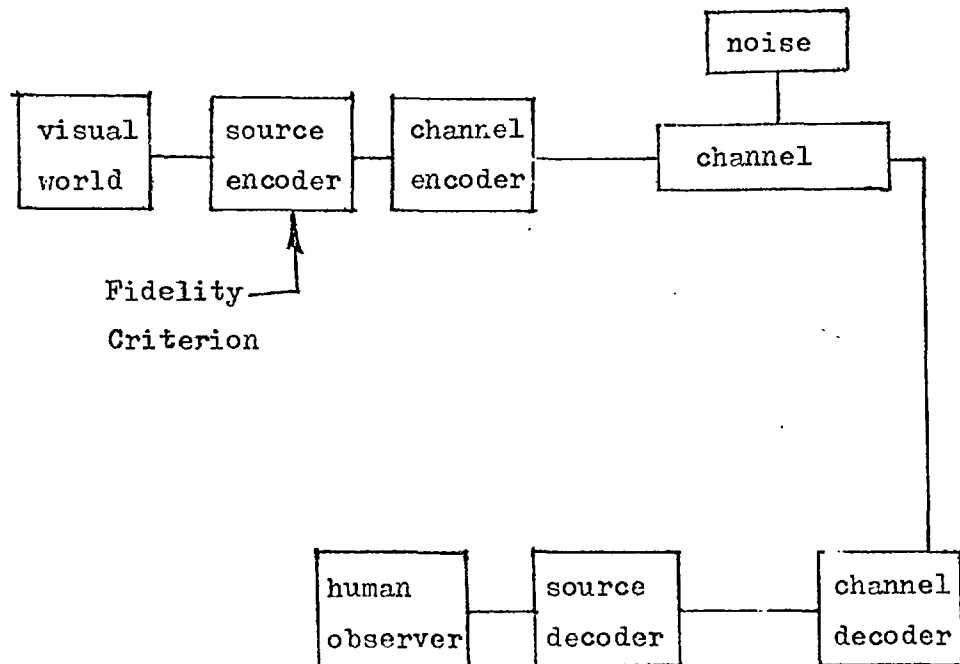


Figure 3.1 Schematic diagram of Fano's model for an image transmission system.

Jacobson arrived at a figure of 4.3×10^6 bits/sec., by multiplying the total number of acuity squares in the eye by the maximum rate at which changes in mosaic patterns could be detected (18 per second). Kelly refined Jacobson's approach by taking into account the estimated signal to noise ratio of each single pathway, arriving at a maximum figure of 800×10^6 bits/sec. A similar approach was used by Budrikis, who estimated that the capacity of the foveal region for monochrome images was 2.5×10^5 bits/sec. However, when the human viewer is regarded as a channel, measurements of recognition rates (SZIKLAI, 1956) and reading rates (PIERCE and KARLIN, 1957) suggest a channel capacity of only 40 bits/sec. All the video coding systems described in the following section require data transmission rates far in excess of this figure of 40 bits/sec. The problem of matching the visual system to the viewer is that the viewer is not just an information sink or channel, but each individual viewer is a complex being with stored memories and past experience giving rise to different expectations.

3.2 Video Image Coding.

The type of image that will be used for videotelephones, in the foreseeable future, will be very similar to that used for broadcast television i.e. raster scanned and monochrome images. Although many of the methods of image coding mentioned below were originally planned for broadcast television, they are applicable to videotelephones; however, some of the bandwidth compression systems are less efficient when used for videotelephones because of the reduced redundancy of the images used.

The first proposals for the transmission of an image by electrical signalling to a distant receiver were made in 1878 by M. Senlacq and the first demonstration of a working system was in 1881 by S. Bidwell. This was only six years after Bell first demonstrated the telephone. In Britain, the standards for rectilinear scanning used for broadcast television were established in 1935, and these have hardly changed to this day. (See GARRATT and MUMFORD(1952) for a history of television.) Nipkow, in 1884, patented a method of raster scanning for the transmission of images to a distant receiver. All broadcast television today still uses raster scanning and virtually all the proposals for more complex coding schemes for future development, still employ raster scanning.

Normal telephone channels use a bandwidth of 3kHz, and the speech signal transmitted may be coded into a digital form with a data rate of approximately 48,000 bits/per sec. The image used for the Picturephone Mod. II generates a data rate of 6.3×10^6 bits/sec., whereas broadcast monochrome television uses a bandwidth of about 5 MHz and a data rate of up to 100×10^6 bits/sec. Because the cost of transmission is related to the bandwidth or data rate used, there has been a great deal of interest in methods of reducing the very large data rates used for the transmission of visual images. An indication of this widespread interest is that bibliographies

on bandwidth compression by PRATT (1967) and ROSENFELD (1968), contain a total of 112 references, and many new papers have been published since then. Although often called bandwidth reduction schemes, they may often be applied in two distinct ways: 1) To minimise the required channel capacity for a specific image quality. 2) To maximise the image quality for a specified channel capacity. Generally, in studies on bandwidth compression of television signals, 6 bit per sample PCM is used as a standard of comparison. Thus a successful bandwidth compression system will require, on average, a low number of bits per sample. However, no one as yet has been able to produce a method which would determine the lower bound of the bits per sample needed for a particular image and fidelity criteria.

The various schemes may be classified as either statistical coding, or psychophysical coding (SCHREIBER, 1967), although the distinction is often ill defined. Statistical coding refers to those schemes which make use only of the statistical relationships among nearby samples; these coding systems are usually reversible. An example of such a system is 'run-length-coding' (CHERRY et al, 1963), and this system has been shown to be able to produce high quality images with as little as 2 bits per sample, on average (ROBINSON, 1969). Psychophysical coding refers to those schemes that try to exploit the perceptual properties of the human receiver. Normal broadcast television makes use of some of the perceptual properties of the human viewer. For example, because of the persistence of vision, it is possible to produce the illusion of continuous motion by the rapid superposition of stationary pictures. DEUTSCH (1966) demonstrated that a frame rate of 7.5 per second produces scarcely noticeable motion degradation. An example of a bandwidth compression scheme which uses psychophysical coding, is the 'dual mode' system, proposed by KRETZMER (1956). This system exploits the fact that

the viewer is relatively insensitive to errors in the specification of brightness levels in areas of high detail.

Many of the schemes require complex coding equipment and large storage facilities so that a careful cost analysis must be carried out to see if the cost of the increased complexity is less than the saving due to the reduction of bandwidth. The coding system which is to be used for the American Picturephone service, is some form of differential PCM (NAST & WELBER, 1969). The differential PCM system would require on average 3 bits per sample, whereas MOUNTS (1969) has demonstrated a frame differencing system that would require an average of only 1 bit per sample, for a typical Picturephone image.

A futuristic scheme for bandwidth compression has been suggested by GABOR and HILL (1961). They proposed that in principle, a transmitter could be built which would have sufficient intelligence to pick out the relatively unimportant areas of an image, such as grass, crowds and wallpaper. Coded signals could then be sent to a receiver which would select from some library, standard patterns for grass etc., more important details of the original scene being sent by conventional methods. The idea of using 'intelligent' transmitters may be adapted for use with videotelephone systems. Persons engaged in conversation over a video link only look at the video image, on average, 44% of the total time (see chapter 7). The proposal is that the transmitter, at either end, has sufficient 'intelligence' to recognise when the viewer is not looking at the monitor screen. When the transmitter detects that the viewer is not looking, it will signal to the other transmitter to stop sending any video signals. Over trunk routes, several of such intermittent signals may then be combined to save channel space by a system similar to that used for speech over transatlantic cables.

There are far too many proposals for the bandwidth compression of video images for them all to be mentioned here, however extensive reviews may be found in the theses by THOMPSON (1968), ROBINSON (1969) and CHITNIS (1971). Chitnis, in particular, has taken account of the different, and often misleading, figures of merit used by many authors.

3.3 Videotelephone Schemes

The first videotelephone was demonstrated in 1927 (IVES, 1927). This video link was between New York and Washington; the transmission was by means of a 250 mile wire link. The image used was 2 inches wide by $2\frac{1}{2}$ inches high, and consisted of 50 lines with a frame rate of 18 per second. The image was produced by rotating Nipkov discs. Although the apparatus was rather unwieldy, it is interesting to note that it was demonstrated as an adjunct to the telephone, whereas experiments with similar apparatus in Britain by Baird, were aimed at broadcast television.

A revival of interest started in 1956, when a videotelephone was demonstrated by Bell Laboratories: it had a postage stamp sized image and a frame rate of 1 per 2 seconds. HALL (1964) reported an experimental variable parameter television at Bell Laboratories which was used for subjective tests, in order to determine acceptable standards for a videotelephone. The results

of these tests led to a system called Picturephone Mod 1, with 275 lines, 30 frames per sec., 2:1 line interlace and 500 kHz video bandwidth. Horizontal line scanning was used with a Kell factor of 0.5. The displayed image was $4\frac{3}{8}$ inches wide and $5\frac{3}{4}$ inches high (aspect ratio 3:4), and the viewing distance was 36 inches. This particular picture format was chosen so that the speaker's head would be shown with as little background as possible. The choice of the relatively narrow picture format would seem to indicate the underlying assumption that the spectator's face was the most important thing to be seen.

When the Picturephone Mod. 1 was used in a field trial at both the 1964 World Fair and between the Murray Hill and Holmdale Laboratories, many of the users found it difficult to stay 'on camera', (CARSON, 1968). (Because of the importance of keeping users 'on camera' the whole of the next section is devoted to this problem.). Several of the users also said that they would like more resolution, a larger screen and a larger field of view. In Picturephone Mod. II, which is the most recent development from Bell Laboratories, the video bandwidth has been raised to 1 MHz. The video signal is encoded into a 6.3×10^6 bits/sec. signal for any transmission further than 6 miles. The choice of this data rate, and hence video bandwidth, was determined by the imminent availability of a 6.3×10^6 bits/sec. digital carrier system, (DORROS, 1971). The picture format has also been changed to a wider one, while the use of a silicon - diode array, for the camera tube target permits the user to adjust the field of view of his own station set by altering the size and position of the raster scan of the target. A summary of the main parameters is given in Figure 3.2. An extensive description of the Picturephone system including the standards, transmission plan, station set components, and maintenance arrangements may be found in Bell Systems Technical Journal, February 1971 (Vol. 50, No. 2).

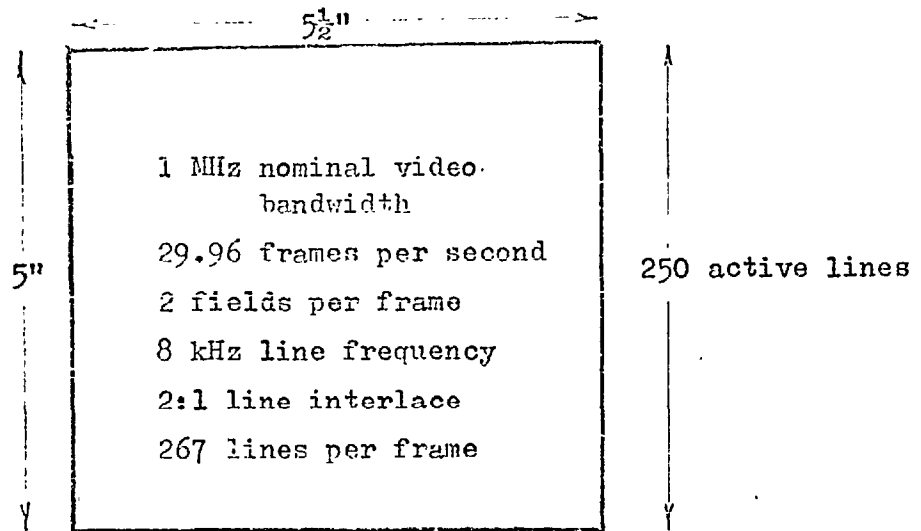


Figure 3.2 Picture parameters of the Bell Picturephone Mod.II.

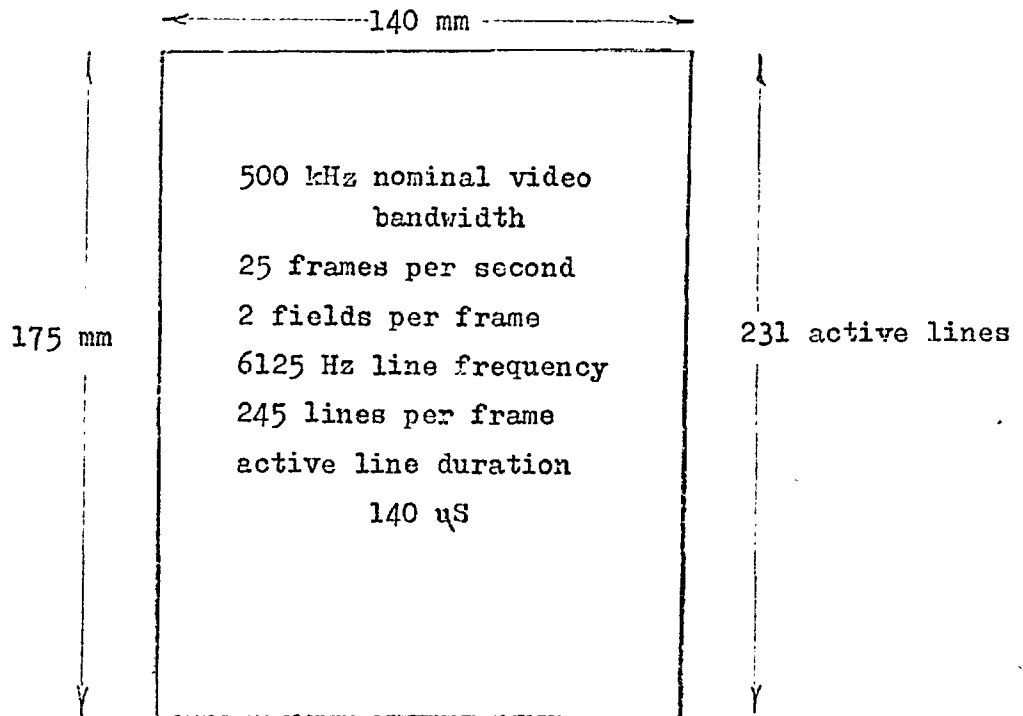


Figure 3.3 Picture parameters of the British Post Office.
experimental Viewphone system (1968).

(Both Figures 3.2 and 3.3 are drawn half full size.)

The first definite proposals for videotelephone standards in Britain were made by LUNN and MUMFORD (1965). They described a system having 228 lines, horizontal scan, and 555 kHz bandwidth. The picture format was similar to that of the Picturephone Mod.1, in that it was higher than it was wide. PURTON and LUNN (1964) adopted this system for digital transmission with a data rate of 1.6×10^6 bits/sec: to do this they used 4:1 dot and line interlace.

The British Post Office have used subjective testing techniques, very similar to those used to test broadcast television, to arrive at the bandwidth of their videotelephone system (ALLNATT, 1968). In these subjective tests, three subjects sitting side-by-side viewed a stationary picture of a head and shoulders, while the ambient illumination was kept low. In the normal videotelephone situation, one person would view a moving image of another person, while the ambient illumination would have to be reasonably high because the viewer himself is being viewed by means of a television camera. One of the reasons given for using the test arrangements is that they allow a more critical than average appraisal of the video image. A summary of the main parameters of this system, called 'Viewphone', is given in Figure 3.3. Vertical scanning was used because it was assumed that the users may attempt to send alpha-numeric material. (BURDETT (1962) found that with a unity Kell factor, intelligibility of typescript is slightly higher than when the scan direction was vertical.)

If in the future, videotelephones become very widespread, they may be used for international telephony. This would give rise to problems of compatibility of systems. As standards converters are very expensive, it would be necessary for the operators of the various systems to come to some agreement about picture parameters. One of the immediate problems would be the different frame rates used in North America and Britain.

3.4 Control of the Viewer's Movement.

In a television broadcasting situation, when a person being viewed moves, a cameraman will redirect the camera so that the image of the person will stay reasonably stationary on the viewing monitor screen i.e., the person will stay 'on camera'. In a videotelephone situation there is no cameraman, nor are there any firm proposals as yet for making the camera automatically track the user as he moves. Furthermore, if a picture format is chosen such that the user's face fills most of the screen, then any slight movement of the user will result in his facial image disappearing from the screen i.e. he will move 'off camera'.

When people use a videotelephone, they will usually sit in some form of chair. This allows the head and trunk to move as if pivoted about the waist, leading to little vertical movement of the head and shoulders but a greater amount of horizontal movement. This means that with a narrow picture format users will have difficulty staying 'on camera', (as was reported with the field trials of the Picturephone Mod.1.) It is the author's personal opinion that, if the British Post Office had used a typical moving image as a stimulus source in their subjective tests instead of the stationary image, then they would have eventually arrived at a wider picture format.

One way of reducing the amount of time that a viewer is 'off camera' would be to use a wider angle lens to increase the field of view; but this would reduce the size of the face, making the the perception of facial expressions and gaze direction more difficult.

An alternative is to reduce the user's movement by providing the user with some monitoring device enabling him to see what image is being transmitted to the other person. A mirror placed close to the camera lens would suffice as a monitoring device. However, the disadvantage of such a system is that the user would divide his attention between his own image and that of the person to whom he was talking. The author carried out some preliminary experiments in which the video monitor screen was split in two, allowing the viewer to see his own transmitted image on one half of the screen and that of the other person on the other half. This system was quickly abandoned, as the subjects spent a large proportion of their time looking at their own image, particularly if the image had been processed in some way. This behaviour was probably due to the novelty of seeing themselves on 'television' and this made the system unsuitable for the planned experiments, but no doubt this effect would decline with prolonged use.

One simple solution to the positioning of the user in front of a videotelephone is to put the monitor screen at the end of a tunnel, (see Figure 3.4). This will limit the distance that the viewer may move while still being able to see the screen. FISCHER and EBEL (1968), working in Germany, have carried out subjective tests on three different constructions of tunnel: 1) a simple air tunnel with black sides, 2) a tunnel with a convex lens on the front, 3) a solid Acrylic block. They also varied the length of the various types of tunnels; the longer the tunnel, the more it restricted the user's movement, but longer tunnels were less acceptable to the users. The test resulted in the choice of an Acrylic block of length 20 cm. To make the whole videotelephone unit appear less bulky, the tunnel was mounted vertically with a 45° mirror at the top. (See Figure 3.5). This German unit was specifically designed to stand on the floor

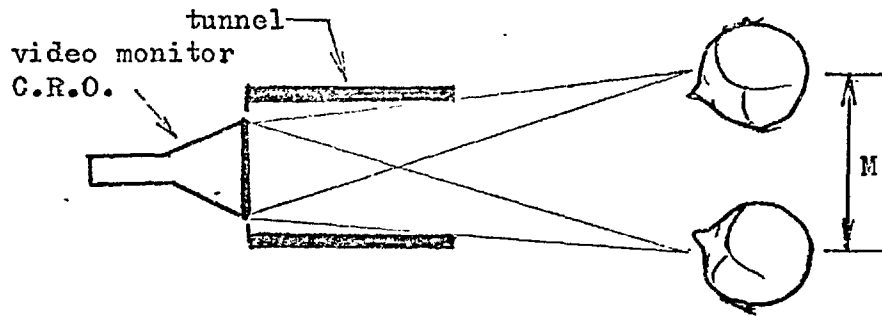


Figure 3.4 Use of 'tunnel' in front of video monitor, in order to reduce the movement of viewer.
 M is the maximum movement of the viewer which will allow an unmasked view of the monitor screen.

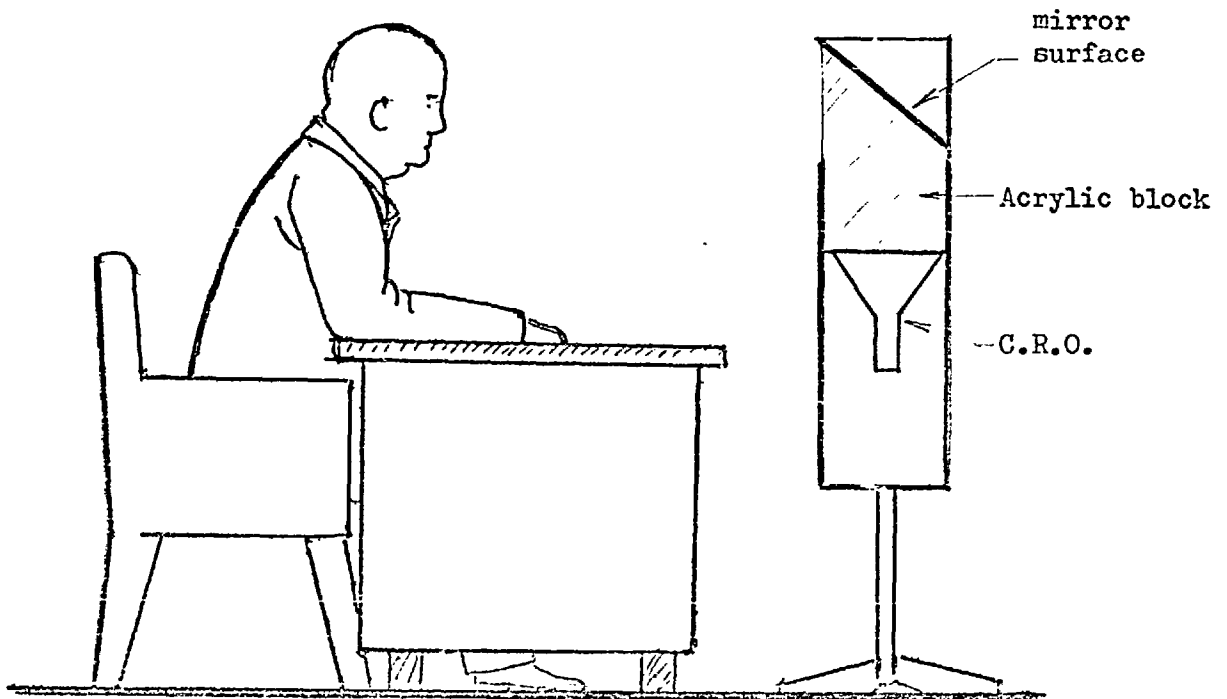


Figure 3.5 Schematic diagram to show the relative positions of the viewer and the video-telephone terminal developed by Siemens AG (Germany).

beyond the user's desk, where as the American system has been designed to stand on the user's desk.

Another advantage of using a tunnel in front of the monitor screen is that it reduces the amount of ambient light that falls on the screen, improving the contrast and enabling the cathode ray tube to be operated at a lower brightness, ensuring longer life.

3.5 Picture Size and Viewing Distance.

In the choice of a suitable screen size and viewing distance for a videotelephone, some of the factors which have to be taken into account are: cost, aesthetic compatibility with other office equipment, and subjective preferences.

JESTY (1953) found that for any particular bandwidth and size of television image, different subjects chose similar viewing distances, and that the ratio of preferred viewing distance (V) to the picture height (H) was a consistent and repeatable measure dependent on the bandwidth of the image viewed. The ratio, V/H , was found to increase as the bandwidth, and hence resolution, was reduced. For a 625 line television image with a bandwidth of 5.0 MHz, the viewing ratio was approximately 6. This value of viewing ratio has been used for subjective tests of 625 line broadcast television images. (For example, see ALLNATT and PROSSER, (1965).

BALDWIN (1951) demonstrated that visual acuity was dependent on viewing distance and the colour of the image viewed. He sat the subjects used for the acuity test at various distances from an image, which was similar in size and brightness to that of a 17 inch television screen although the image was produced by a slide projector. The average visual acuity at a viewing distance of 60 inches was approximately 1.8 times that at 20 inches (corresponding to 1.8^2 increase in the bandwidth required for a television image), and at distances above 60 inches the visual acuity was fairly constant. JESTY (1952) suggested that this was the reason why, on average, subjects are less critical of small pictures than large ones, when viewed from distances which give the same viewing ratio, V/H .

VERNON (1953) has pointed out that, as a television screen can be seen to be small compared with its surroundings, the images tend to look smaller than the original objects that they depict, where as images projected on to a larger screen at a greater distance are seen as possessing the original sizes of the objects depicted. This awareness of the diminutive size of depicted objects while viewing a small television screen normally ceases with increased experience, particularly if the material viewed is familiar. However, certain unfamiliar things such as unfamiliar three dimensional models appear more 'real' on a larger screen.

When viewing a three dimensional object, the perception of depth arises from binocular cues, perspective, shadow, overlap and past experience of the object. When viewing the image of the same object projected on to a screen, binocular cues are not available for the perception of depth. The binocular cues will instead indicate the flatness of the screen. STAFFORD and HANSEL (1953) have shown that as the viewing distance from a flat screen is increased, the binocular cues become weaker, and the perception of depth of an object depicted on a screen is improved.

In their experiment, subjects sat at distances of 3 to 24 feet from a screen and were asked to judge the relative depth of blocks of wood from photographs projected on to the screen. It was found that with binocular vision there were less errors the further the subjects sat from the screen. (Approximately 3 times the number of errors occurred when subjects sat at 3 feet, compared to when they sat at 24 feet from the screen). It was also found that judgements made with monocular regard were more accurate than those made with binocular vision, the difference being greater the closer that the subjects sat to the screen. Stafford and Hansel concluded that television screens should be much larger than those available at the time, (1953).

The various experiments reported above give contrary results as to the best choice of viewing distance. The experiments concerned with acuity would suggest a small screen close to the viewer, where as those experiments relating to the perception of depth and the impression of reality would indicate a larger screen, further away from the viewer. All the experiments were carried out with the subject just watching an image on a screen. However, in a videotelephone situation, the viewer will interact with the image on the screen, and consequently a different image size and viewing distance may be preferred. The optimum distance and size of screen may be found to be related to the 'comfortable distance for conversation' (see section 2.6), and thus related to the 'apparent distance' of the facial image on the video monitor screen (see section 6.2). And there are, of course, personal preferences which may be finally the most important factor of all.

It may be questioned as to whether a choice has to be made at all, for it should be possible for the operators of videotelephone systems to offer different designs of videotelephone sets to its users. A choice could be, for example, a small unit designed

to fit on the user's desk or a larger, pedestal mounted unit to be viewed from across the room. The first of these alternatives (i.e. small desk top unit) was used for all the author's experiments. This choice was determined by two factors; 1) The ready availability of small video monitors, and 2) that it seems likely that this type of unit will, at least initially, be more widespread than the pedestal unit. The following section describes the type of video image used in the author's experiments.

3.6 Visual Images used for Experiments.

Two criteria may be adopted when choosing the picture parameters to be used in experiments to investigate how the visual image would enhance telephone conversations. Firstly, the visual image should be similar to those used in existing vidcotelephone systems, or similar to those that are likely to be used in future systems. The results of experiments with such visual images would then be directly applicable to the design or implementation of these systems. Secondly, several visual images should be used which are different in some definable parameters. In this way the effects of the various picture parameters on the visual enhancement of telephone conversations could be determined. As there are very many different picture parameters that could be chosen, this may lead to very many experiments or unmanageable, large, factorial experiments.

Because of the various limitations of time and cost, it was decided to use two types of image: one, a high fidelity monochrome image, and the other, as simple an image as possible which would allow the recognition of the speaker's identity, facial expressions and movements etc. The choice of the high fidelity system was determined by the widespread use, and availability of, 625 line closed circuit television equipment. This would fulfill the first criterion of choice, mentioned above, as it would be very similar to the video image used in the 'Comfravision' system, (see PEARMAN and IDDLES, 1968). Although this is a higher fidelity video image than that to be used in any switched videotelephone system, future developments might result in a video image with similar quality to that used in the experiment. The main parameters of the monochrome image used in the author's experiments are summarized in Figure 3.6.

The 'simple' image used for the experiments consisted of just the form, or shape, of the speaker's face, lips, eyes, and upper parts of his body, thus removing all cues of texture and shading. VERNON (1962) stated that for the perception of form or shape, the most important feature seems to be the general outline or contour of the object. The importance of contours or edges is well known to those involved in the design of pattern recognition machines, in which it is common to incorporate a process for reducing the original image to a simpler form by edge detection. (See, for example, SARAGA and WOLLONS, 1968.)

It is possible to produce simple circuits that will reduce a typical monochrome video image so that only the edges or contours are left; the resulting image will be like a line drawing or cartoon. However, a cartoon image produced by such circuits would be different to that produced by an artist, in several different ways. Firstly, the circuitry will not be able to separate

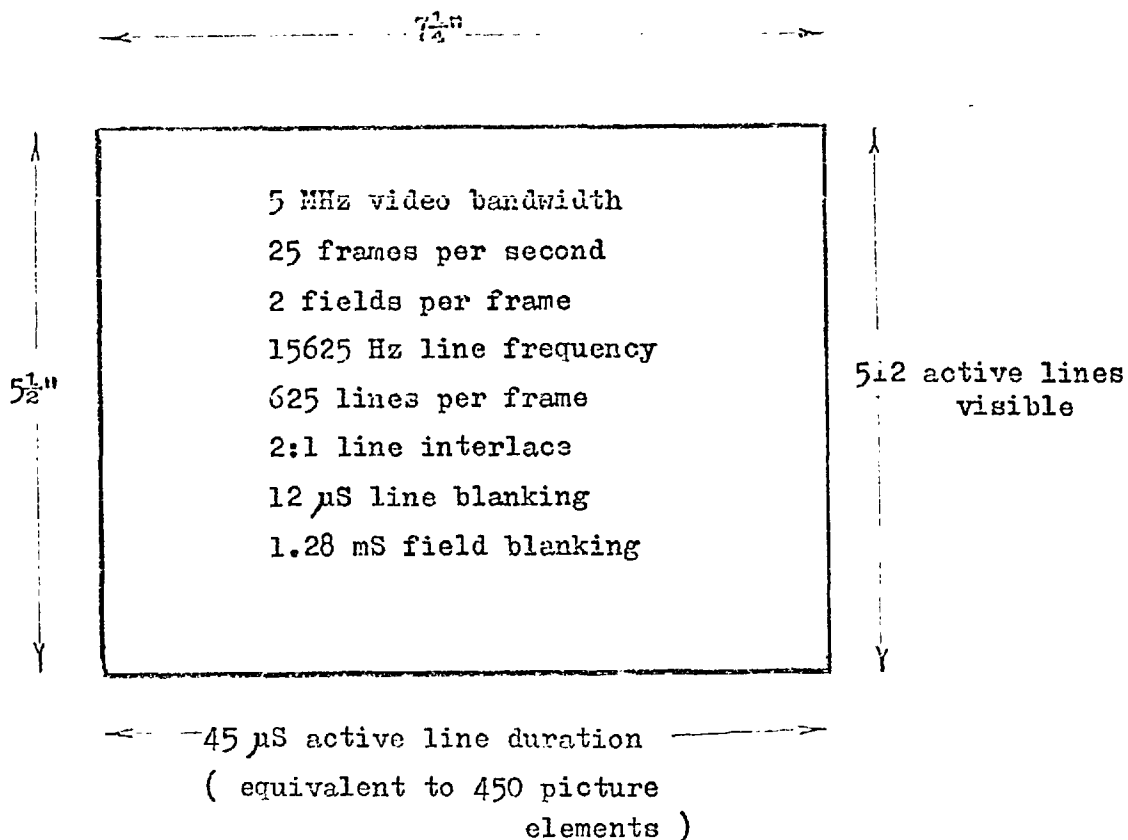


Figure 3.6 Picture parameters of the video image used in the author's experiments. (Ampex 9A video monitors were used for all perceptual and conversation experiments.) This figure is drawn half full size.

'figure' from 'ground'. Thus, if the original image had large and irrelevant variations in brightness in the 'ground', a simple circuit would detect this and include it, where as an intelligent artist would exclude irrelevant details. Secondly, many cartoon drawings by artists will emphasise or distort parts of the face or body to produce a desired effect: the cartoon may be a caricature (KRIS, 1952), and this is far beyond the ability of simple circuitry.

Examples of the two types of images used in the experiments are shown in Figure 3.7. For brevity, the two types will be referred to as the monochrome image and the cartoon image throughout this thesis.

The reason for including the cartoon image in the experiments, is that it will give some insight into precisely what features of a video image are useful for the enhancement of telephone conversations. The author does not advocate that videotelephones should use such an image. However, if such an image were used, the data rate required for transmission would be much lower than needed for a monochrome image. A comparison of the entropy and coding needed for the two types of image is given in Appendix I.



'Archie'



'Max'

Figure 3.7 Examples of the two types of images used in the experiments.

3.7 Generation of Cartoon Image.

The remainder of this chapter is concerned with the design of the cartoon generating apparatus used for the videotelephone experiments.

PEARSON (1965) has demonstrated a cartoon type of video image. His apparatus consisted of a 2-level quantizer, followed by a differentiator and full wave rectifier. The disadvantage of his method is that very different cartoons are produced if the 2 level quantizing threshold is changed slightly. Also, unwanted contours are introduced where a very slowly changing signal crosses the quantizer threshold. With this type of detector, it is virtually impossible to detect both sides of a face and the eyes with the same threshold setting. In fact, it should not really be called a contour or edge detector.

VERNON (1962) states that the perception of contour depends mainly on the sharp gradient of brightness between a surface of an object and its background. Thus, a contour or edge detector will consist of some method of differentiating a brightness signal with respect to position, followed by some form of level detector. With a raster scanned video system this differentiation is achieved by differentiating the analogue video signal with respect to time.

Circuits may be produced which achieve a very close approximation to differentiation. Their disadvantage is that any high frequency noise is greatly amplified, and thus, using these circuits gives rise to some very difficult filtering problems. This problem is reduced if a circuit that gives a first order approximation to differentiation is used. Such a circuit simply consists of delaying the analogue signal for a time τ , and then subtracting

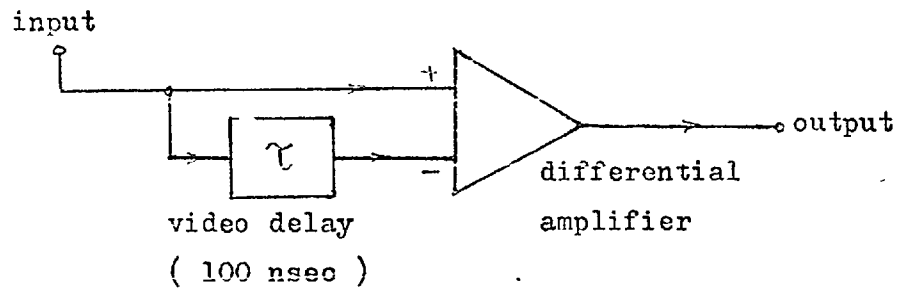


Figure 3.8 Circuit for first order approximation to differentiation .

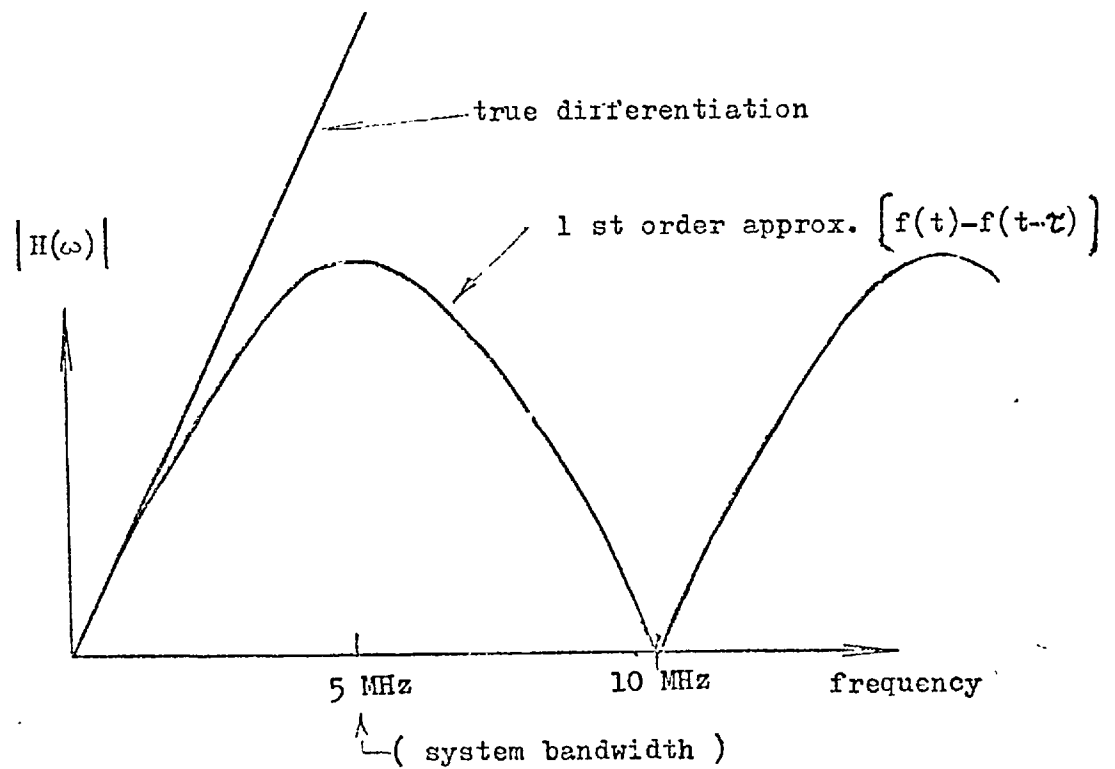


Figure 3.9 Frequency Response of Differentiators

it from the present signal $f(t)$, (see Figure 3.8). The output of the circuit $g(t)$, is given by:-

$$g(t) = f(t) - f(t-\tau)$$

This circuit has a linear phase response, and the frequency response $H(j\omega)$ is given by:-

$$|H(j\omega)| = 2 \left| \sin \left(\frac{\omega \tau}{2} \right) \right|$$

Figure 3.9 shows the difference in frequency response of a true differentiator and of a first order approximation. The higher frequency lobes of the response of the first order approximation may be relatively easily filtered out.

An edge detector was constructed using a first order differentiation with a delay time of 100 nsec. As both white to black and black to white transitions had to be detected, the differentiator was followed by a full wave rectifier (fwr). This then fed a simple threshold device (th), which detected brightness transistions above some defined threshold level. The threshold was adjustable, so that the cartoons of varying degrees of complexity could be produced. The threshold devices were followed by a monostable (M), so that the displayed edges would have a constant thickness. A schematic of the simple edge detector is shown below, in Figure 3.10, and the complete circuit is shown in Appendix III.

If the circuit mentioned in the last section were used by itself, any edges parallel to the raster scanning direction (i.e. horizontal edges in this system) would not be detected. Many of the important features are very nearly horizontal e.g. the eyes and the mouth. Horizontal edges may be detected by subtracting one line of the raster scan from the one immediately above it, followed by a full wave rectifier and threshold gate, as before.

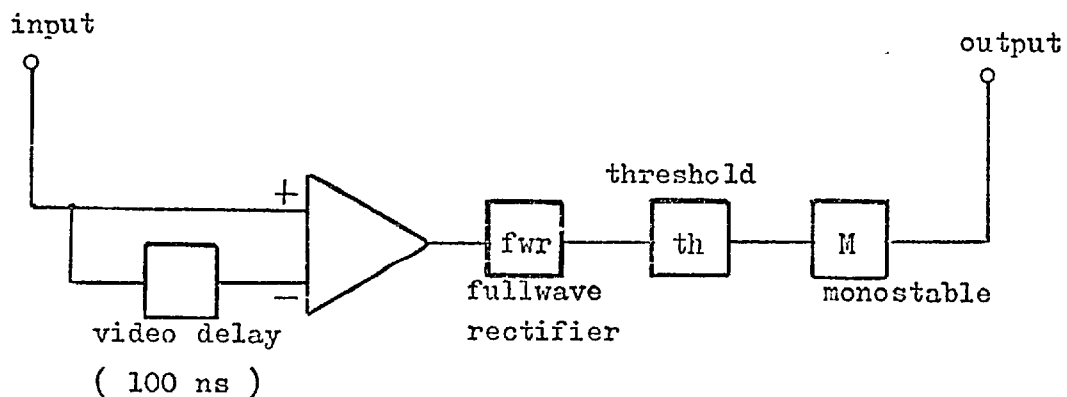


Figure 3.10 Simple Edge Detector which will detect vertical and near vertical edges.

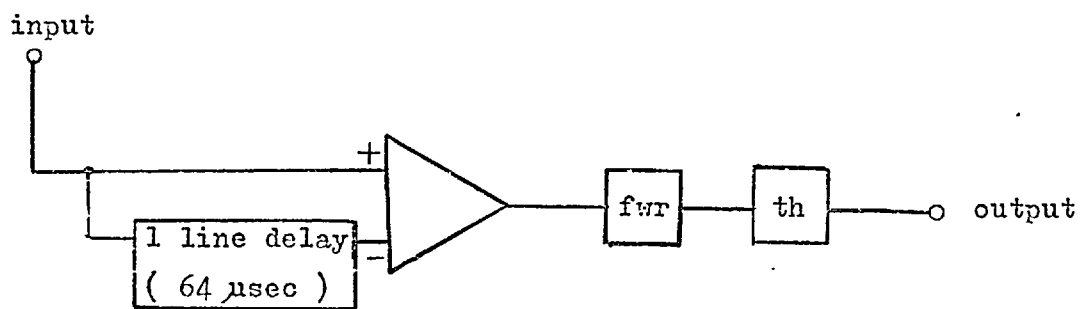


Figure 3.11 Circuit for detecting horizontal and near horizontal edges

To do this, it is necessary to delay the video signal for a time equal to one line period ($64 \mu\text{sec}$). A glass acoustic delay line was used for delaying the video signal. A schematic of a 'horizontal' edge detector is shown in Figure 3.11. Although it has been called a horizontal edge detector, it will detect any significant edge that is near the horizontal.

If all the details around the eyes etc. are to be detected, the threshold circuit must be able to detect very small transitions in brightness. There was however, a trade between the level of the threshold detector and the amount of noise introduced. When the threshold was adjusted to give adequate detail, the image was marred by noise which was snow-like in appearance. As the noise in simply organized regions of a picture is particularly disturbing (GRAHAM, 1962), a system was implemented which would remove the noise in the blank areas of the cartoon image. To do this, the noise stripping circuit made use of the fact that a cartoon line on a television screen consists of a string of adjacent picture points; thus, if all the detected picture points which do not have an adjacent picture point were removed, then much of the disturbing noise would be removed also. A schematic of a system constructed to do this is shown in Figure 3.12.

By using a one line delay circuit and a shift register, as shown in Figure 3.12, it was possible to have access to the N picture points on the line immediately above, in the same field. A simple gating arrangement would only allow a detected point if there was a detected point in the shift register. If a long shift register is used (N large), a point on a near horizontal contour line will be allowed, but the probability of allowing an unwanted, random noise point increases with the length of the shift register. If a very short shift register is used, only points on or near vertical lines will be allowed. Hence, there is a trade-off between the noise stripping abilities and the possible

nearness of a contour line to the horizontal.

The optimum length of the shift register will depend on the noise level, on the video source, and on the particular image used. Subjective tests were carried out using the flying spot scanner and a positive transparency of a face. Noise was added to the video signal to give the same subjective impression of noise level as was obtainable with the vidicon camera to be used in the experiments. No difference was detected when the shift register was between five and nine stages long. For the final construction, therefore, a seven stage shift register was used.

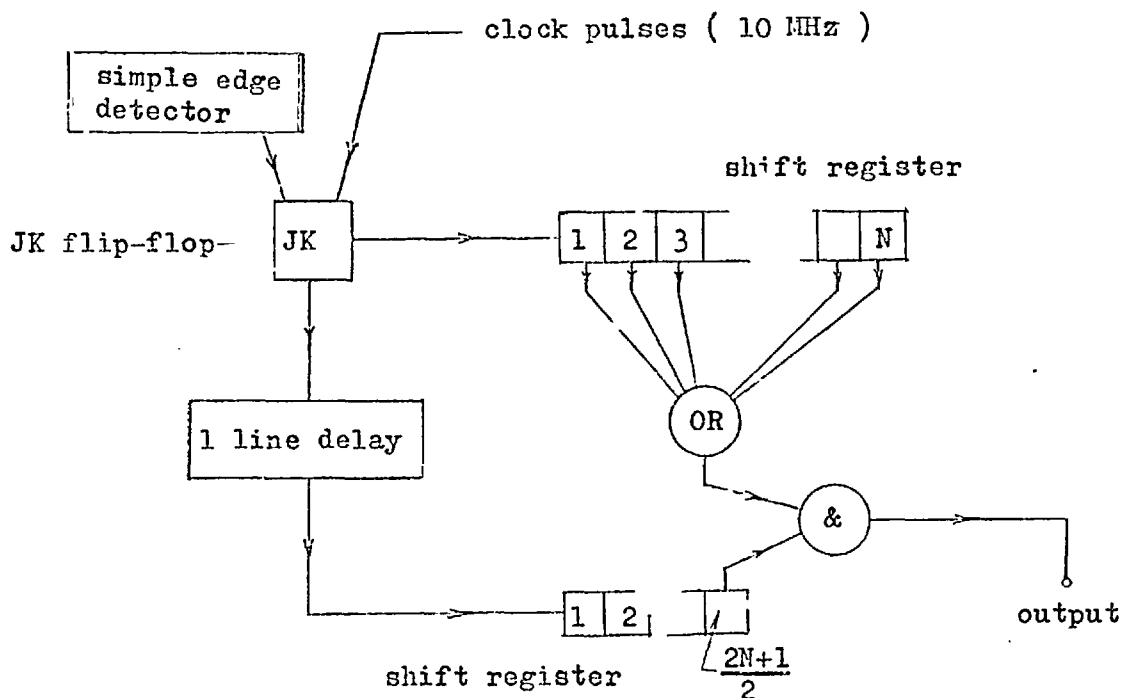


Figure 3.12 Circuit for stripping noise from a cartoon image.

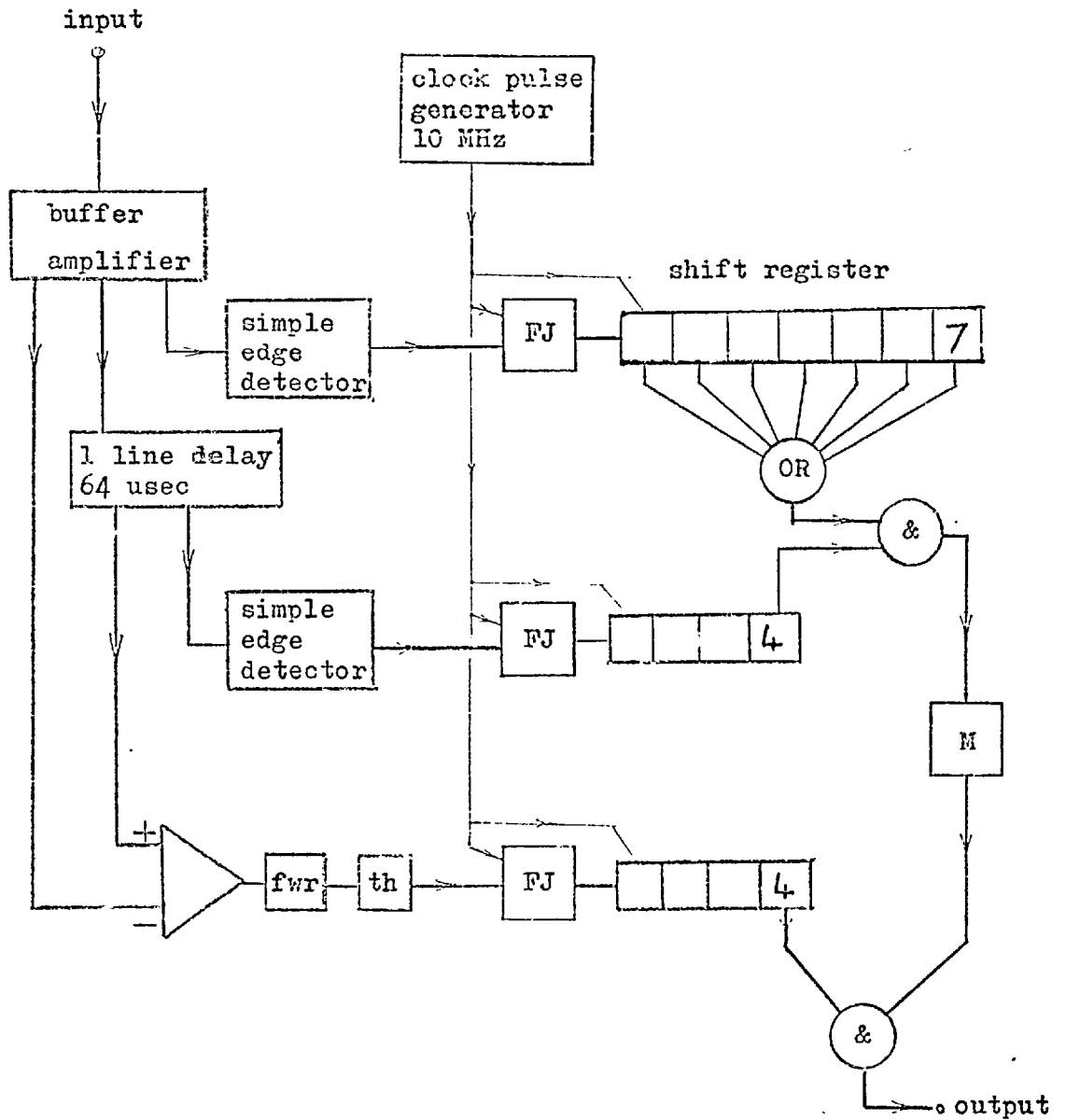


Figure 3.13 The complete cartoon generator.

With only one acoustic delay line, it is not possible to produce a system of stripping the random detected points produced by the horizontal edge detector. Therefore, the setting of the threshold level in the horizontal detector could not be so fine as that in the simple edge detector. Better noise stripping circuits may be devised if more than one delay line were available, (PAN, 1962). However, as delay lines are very expensive, the slight improvement due to the more complex system would hardly justify the increased cost.

The complete cartoon generator is shown in Figure 3.13. The configuration shown allows the output of the simple edge detector plus noise stripping to be added to the output of horizontal edge detector, while using only one acoustic delay line. As many of the detected changes in brightness tended to appear in both fields of each frame, the horizontal cartoon lines were usually two scanning lines thick. For this reason, the monostable (M), was added to make the vertical lines appear to have the same thickness as the horizontal ones.

Plumbicon and vidicon cameras have a triangular noise spectrum, with the greater amount of noise at the ^{high} frequencies, (VAN DOORN, 1966). Because of this, whenever a plumbicon or vidicon camera was used, it was found that the subjective appearance of the cartoon was improved, if the original analogue video signal was passed through a low pass filter with a cut off frequency of 3 to 4 MHz.

For an image of a person's head and shoulders (as shown in Figure 3.7) produced by a plumbicon camera with a low pass filter (3.5 MHz cut off), the noise stripping circuit reduced the total number of detected picture elements per second by about 2 to 3%. Since many of the points removed were noise in blank areas of the cartoon image, the final image was much more acceptable.

The difference between simple edge detection, horizontal edge detection and a combination of the two may be seen in Figures 3.14 - 3.17. All the photographs are single frame exposures. The video image was produced from a slide by a plumbicon camera. The picture points in all the cartoon images are at discrete intervals of position, e.g. the output of the simple edge detector is fed into a JK flip-flop which is 'clocked' at 10 MHz, (see Figure 3.12).

Data on the number of picture points produced by the different types of cartoon generator is to be found in the section on coding cartoon images in Appendix I.

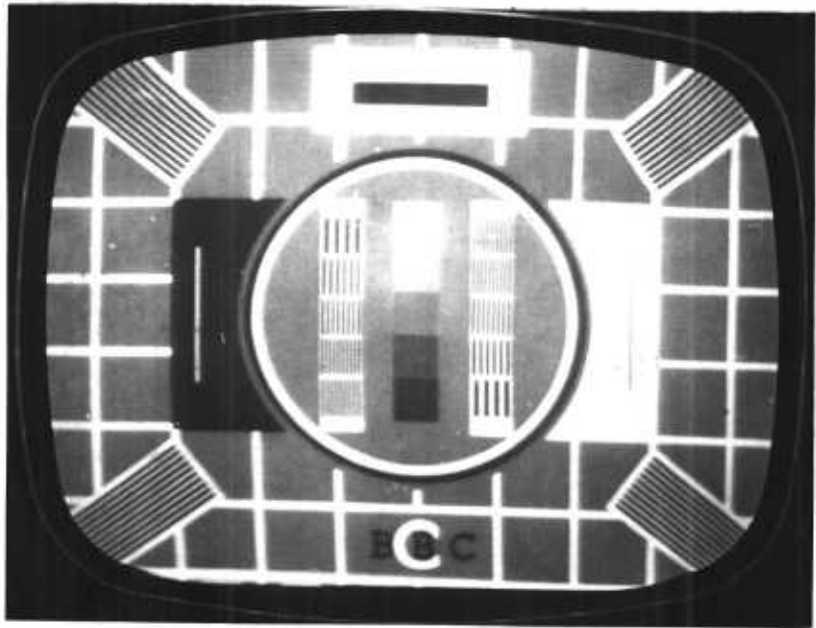


Figure 3.14 Original monochrome image.

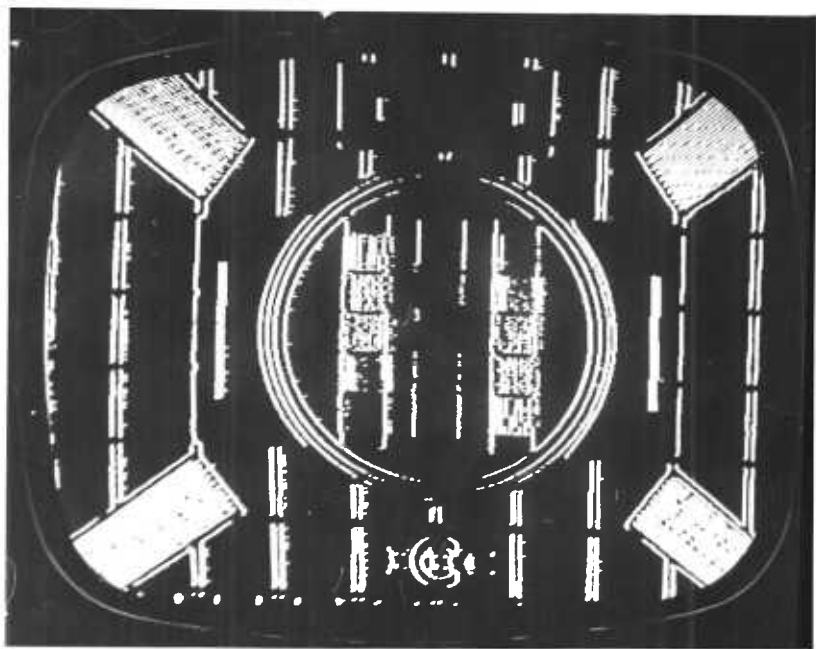


Figure 3.15 Cartoon produced by a simple edge detector.

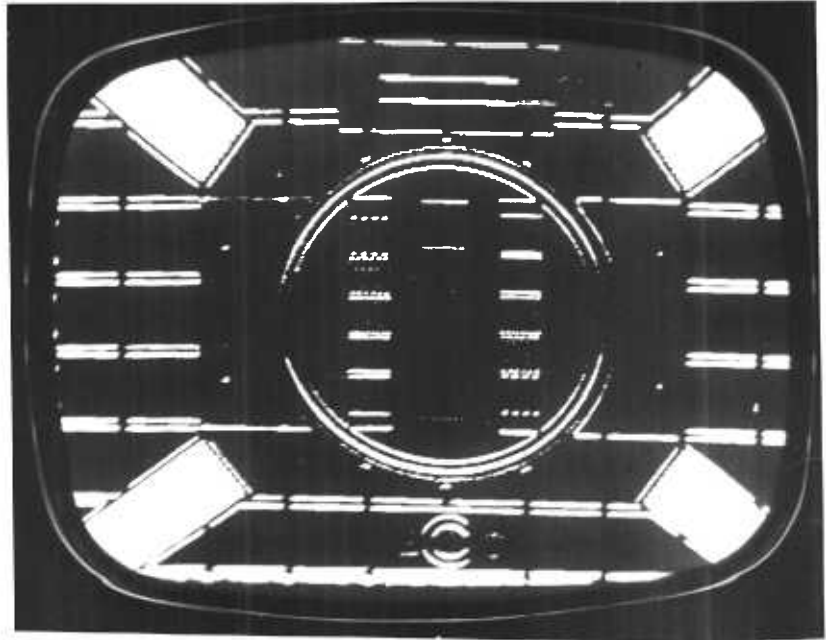


Figure 3.16 Cartoon produced by horizontal and near horizontal edge detector.

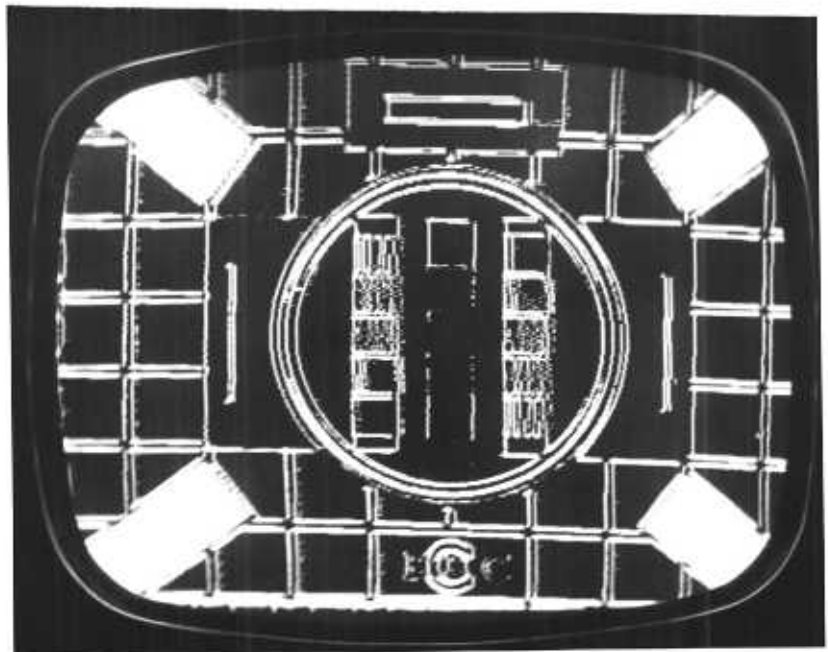


Figure 3.17 Complete cartoon image.

Chapter 4.

Control, Recording and Terminal Equipment.

This chapter describes the control, recording and terminal equipment used in the videotelephone experiments. The chapter is split into four sections. The first deals with the perception of another person's direction of gaze when conversing over a videotelephone link and describes an experiment performed to investigate this. The results of the experiment is used to decide the relative position of the video monitor and camera, in the laboratory terminal equipment. The second section deals with the audio system used, and the third with the video system constructed for the videotelephone experiments. The fourth section contains a description of the method and apparatus used to process the data obtained from the videotape recordings of subjects conversing over the videotelephone link.

4.1 Relative position of video monitor and camera.

Once the type and size of video image and the viewing distance to be used have been decided (see section 3.6), the next problem is to choose the relative position of video monitor and camera at each terminal. In commercial systems, this choice may be determined by aesthetic design, or the need for various graphic facilities. One important factor is the perception of eye contact. Hence, the relative positions of monitor and camera should be chosen so that when one viewer, (A), looks at the image on his video monitor, the person to whom he is conversing, (B), should be able to perceive that A is, in fact, looking at the monitor, and moreover for greater reality, B should perceive that A is looking at him.

The following section deals with the perception of eye contact, and this is followed by sections which describe an experiment carried out to determine the optimum choice of camera and monitor position, with the criterion that the perception of eye contact over the videotelephone link should be as accurate and realistic as possible. The results of this experiment will have important implications for other experiments described later in this thesis, namely those concerned with a videotelephone user's looking behaviour. Thus, in these later experiments, if measures of 'eye contact' are to have any significance, it must be known whether the users can accurately perceive the other user's direction of gaze.

4.1.1 Perception of eye contact.

The stimulus sources for the perception of a person's direction of gaze are the position of the pupil and the posture of the head. As GIBSON and PICK (1963) point out: the main visible features of the eyes are the pupil and sclera within a frame formed by the eyelids. The concentricity of the pupil in the sclera is one of the dependents in the perception of the direction of gaze. (see Figure 4.1). This is not the only dependent, however, as the position of the head must also be taken into account. (see Figure 4.2). Figures 4.1 and 4.2 amply demonstrate that the perception of the direction of gaze depends not only on eye form, but the eye form relative to the face form.

In the experiment reported by Gibson and Pick, a trained 'looker' gazed at various imaginary points on a horizontal line at the level of the subject's eyes. The subject was asked to say 'yes' when the 'looker' gazed directly into his eyes. The results suggested that the subjects had good discrimination of whether or not the 'looker' was gazing directly at them. There were considerably less 'yeses' from the subjects when the 'looker' rotated her eyes 2.9° away from the direct eye contact position. It is usual to call the standard deviation of the distribution of 'yeses' the 'threshold', and the mean of the distribution the constant error (CE). In the Gibson and Pick experiment, the looker varied her head position so as to look askew at the subject. With the head rotated through 30° to either side, there was a small shift in the CE, in the same direction as the head rotation. The threshold for the askew head position was very similar to that for the straight forward head position.

CLINE (1967) presented subjects with a mirror image of a trained 'looker'. The apparatus used allowed the looker to gaze

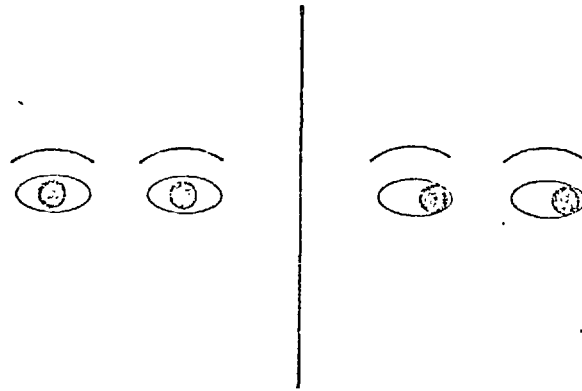


Figure 4.1 Schematic centering and off centering of the pupil in the sclera as the stimulus for the impression of being or not being looked at.

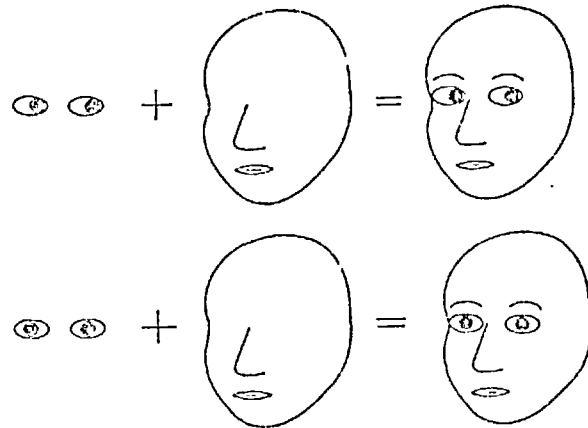


Figure 4.2 The same Eye-Forms as in Fig. 4.1 surrounded by a Face-Form which shifts the apparent direction of gaze. Eyes formerly gazing to the right now seem to gaze forward, and those formerly gazing forward now gaze to the left.

(Figs. 4.1 and 4.2 taken from GIBSON and PICK, 1963)

at one of 13 targets, which the subjects could not see. It appeared to the subjects that the looker was gazing at various points about the subject's head and face. The subject had to indicate where they thought the looker was gazing. Accuracy for judgements of gaze, directed at targets other than the subject's eyes, was considerably less for gaze directed at the eyes, thus subjects are very sensitive to eye contact. Also, it was found that the threshold for vertical rotation of the looker's eyes was significantly greater than for horizontal rotation. Unlike Gibson and Pick, Cline did find a difference in the thresholds for different looker head positions; also, the threshold for the straight forward head position was smaller than that found by Gibson and Pick. However, Cline agreed with Gibson and Pick as to a similar relationship between CE and head position.

If the eyes on a flat image, such as a photograph or portrait, appear to be looking at the viewer, then when the image is rotated slightly, the eyes still appear to look unwaveringly at the viewer. This effect is well known, as the eyes of many famous portraits, notably the 'Mona Lisa', seem to follow one around wherever one may walk. ANSTIS et al (1969) investigated this effect by performing an experiment, very similar to that of Cline's, but using a television image of a trained looker, as well as the 'live' looker. The results indicated that when the television monitor was rotated, the perceived direction of gaze seemed to move in the same direction as the monitor rotation. This would not normally be expected to occur if the rotated image was truly flat. The explanation for this result, however, was that the frontal surface of the video monitor was convex, thus making the television image less like a flat portrait and more like a three dimensional statue. For both television images and 'live' lookers, Anstis found a similar relationship between CE and head position, as did Cline, and Gibson and Pick.

McGRATH and HATCHER (1961) took a series of photographs of a looker, who fixated his eyes at a target points varying in one degree steps from the camera lens, as well as directly at the lens. The targets were on a horizontal line and a vertical line, intercepting at the centre of the lens. The photographs were taken at a distance of 5 feet. Life sized photographs of the looker were then shown to subjects, who had to judge the looker's direction of gaze. It would appear from their results that for vertical rotation of the looker's eyes, there was a considerable CE below the centre. However, McGrath and Hatcher did not carry out any statistical analysis of their results, so that it is difficult to know what emphasis to place upon them.

4.1.2 Video links.

Raster scanned images, using analogue video signals, may be expected to influence the perception of the direction of gaze. This possible influence may be due to several factors. Firstly, the image of the eye form etc. is effectively sampled in one direction and not the other. Secondly, because of the limited bandwidth of practical systems, the possible resolution of detail around the eyes will be worse than in face to face interactions. These limitations mentioned will affect the perception of the direction of gaze; also there will be noise, ghosting, filter ringing, and other aberrations which will have an effect.

One other major disturbance encountered in videophone links is that the camera and video monitor are spatially separated and, therefore, the viewer looks at the monitor display and not the camera. For example, if the camera is on top of the monitor

(see Figure 4.3) and the communicant A looks at the eyes of the facial image of B on the monitor, then the facial image that B sees of A will be a face looking down and not at him. Similarly, if the camera is at the side of the monitor and A looks at the monitor, B will see an image of A apparently looking to the side.

In such systems there can never be true eye contact, but the 'off set viewing' effect will be reduced, the nearer the camera is to the image of the eyes on the monitor.

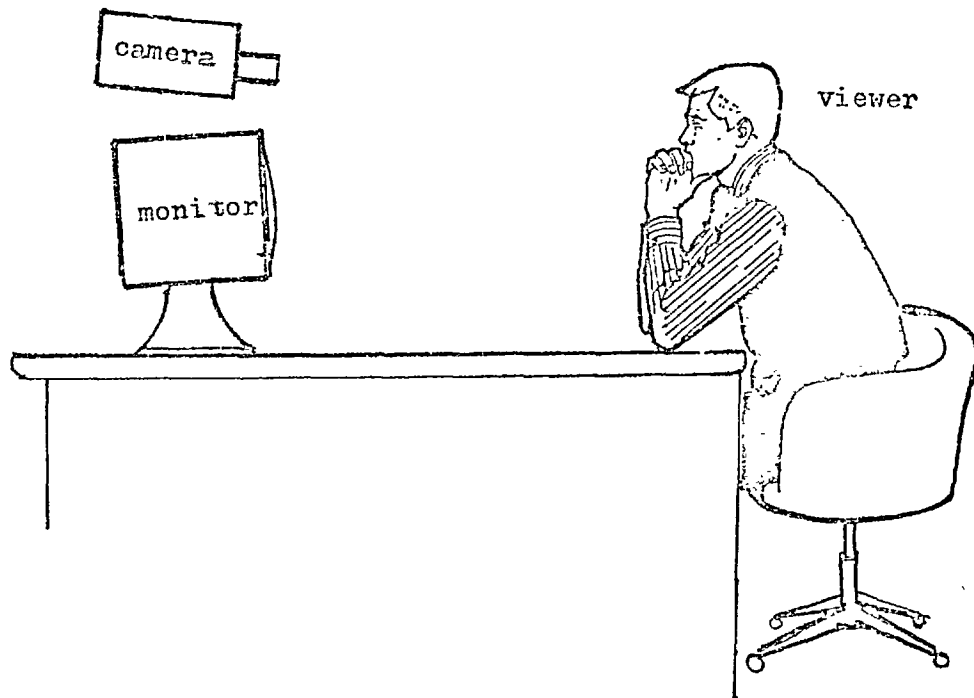


Figure 4.3 Simple videotelephone terminal.

4.1.3 Choice of Monitor and Camera Positions

As the video link under consideration is for experimental purposes, the height of the cameras and monitors may be adjusted if particularly tall or short subjects are encountered. Also as the video link is to be used only for experiments on conversation, no graphics facility need to be taken into account, when positioning the camera. Thus, the only consideration in positioning the camera and monitor is to reduce the 'offset viewing' effect.

There is likely to be a difference in the perception of horizontal and vertical rotations of the eyes. This difference, if it exists, is also likely to be affected by the direction of scanning and the limited bandwidth available, and will be different from one system to another. If data concerning the perception of eye contact for a particular video system were available, plus a knowledge of the dimensions of the camera and monitor, then it would be possible to select the relative positions of the camera and monitor such that the disturbing lack of eye contact is minimized. For example, consider a video link which has a square picture monitor and the facial image is arranged such that the eyes are in the middle of the screen. In this case, the camera may be as close to the eyes on screen whether it is placed above, below or at the sides. If experiments showed that for this particular video link horizontal rotation of the eyes was very much less perceptible than vertical rotation, the camera would be placed at the side and not above or below the monitor.

The example above was a particularly simple one for in most video links a camera immediately above, below or at the side of the monitor will not in general be equally close to the eyes of the monitor image. In such situations there is a different minimum angle of rotation between the image and the camera in the vertical and horizontal directions, thus the perceptibility of eye contact for these particular angles will have to be compared.

The following section deals with an experiment carried out to measure the perceptibility of eye contact for vertical and horizontal rotation of the eyes over the video channel to be used in future conversation experiments, and hence to decide the camera and monitor positions for these future experiments.

4.1.4 Experiment to measure the perception of eye contact

For this experiment, 3 subjects were used at each session. One of the subjects (the looker) was seated in front of a television camera. The camera was arranged to be at the same height as the subject's eyes and 1 metre away from them. (see Figure 4.4). The camera (Link type 101) used a vidicon tube and a 25 mm lens. This was the type of camera used for all the videotelephone conversation experiments. Fixed to the camera was a single line of miniature pilot light bulbs, spaced at 2.5 cm. intervals. With such a simple arrangement, it is not possible to have a pilot bulb at the centre of the lens, but a bulb (bulb X in Figure 4.4) was placed as near as possible to the lens. The subject, seated in front of the camera, was told to look directly into the centre of the lens when bulb X was lit, but whenever any other bulb was lit, to look directly at that bulb. These bulbs were lit in a random order by the operator of the experiment.

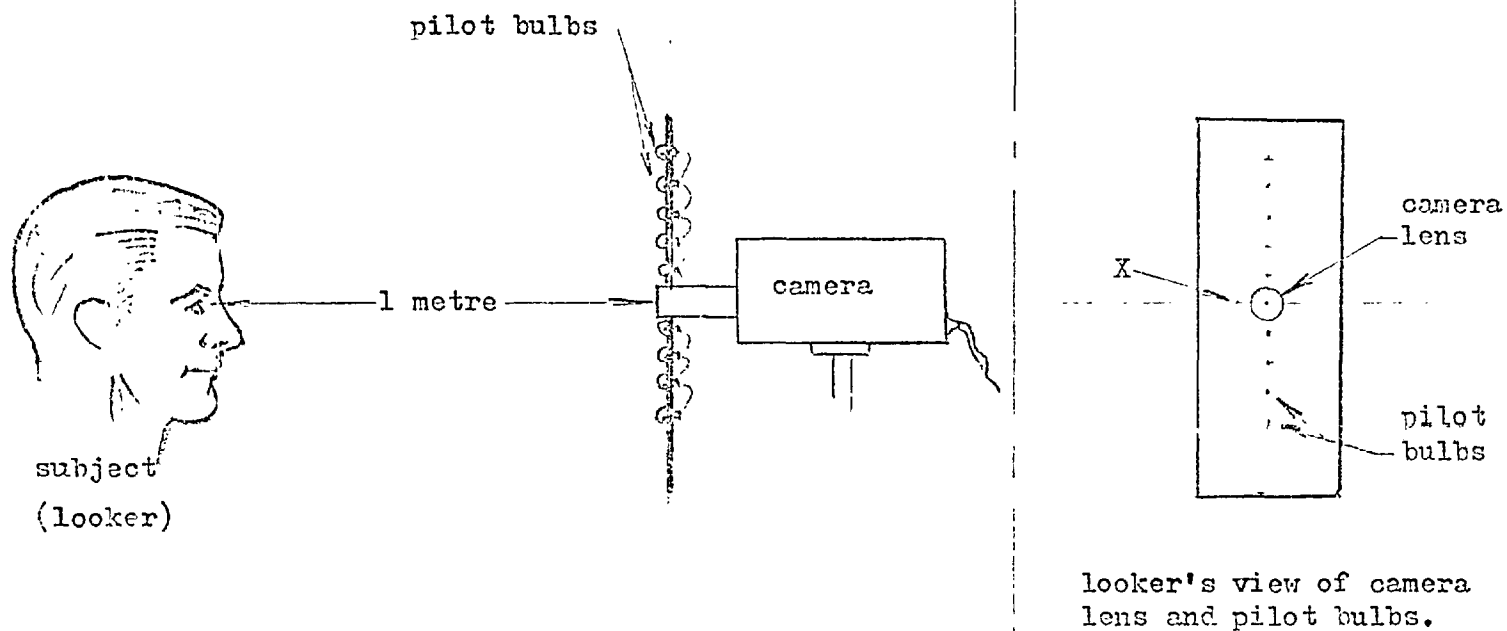


Figure 4.4 Position of the camera and looker for the experiment on the perception of a looker's gaze.

The other two subjects (viewers) were seated in front of a video monitor; the monitors used were the same as used for all the other experiments (see section 3.6). The illumination of both viewers and looker was arranged to be as similar as possible to that encountered in the videotelephone conversation experiments. The viewers were seated approximately 1 metre away from the monitors; this meant that the reduced image of the looker's face appeared the same size as the live looker's face would, at a distance of 2.5 metres. Each viewer was given a small box with two buttons on it. They were given the following instruction: "press the green button when the person is looking directly at you, press the red button when he is not looking directly at you". (This instruction was chosen because it is virtually the same as that used by Gibson and Pick.)

The screen was made blank as soon as the subjects had made their decisions. Care was taken to ensure that the subject looking at the camera had settled into a steady gaze before his image was switched to the monitors for the next trial. The scanning direction of the camera was altered after approximately 40 trials, and after approximately 80 trials the subject in front of the camera was changed with one of the subjects watching the monitor. The whole cycle was repeated three times with a short break in between each cycle. Thus, every subject was used as both looker and viewer, consequently avoiding the problem of selecting a single trained looker, who was free from ocular imbalance. Each group of three subjects made a total of at least 72 judgements on each lamp position. All subjects were university students or secretarial staff, who responded to advertisements for volunteer subjects. Only subjects who said that they had normal vision and did not wear spectacles, were used.

Separate tests were carried out with the line of pilot lights mounted vertically and horizontally. Preliminary tests indicated that the range of values used for vertical rotation of the looker's eyes needed to be greater than that for horizontal rotation. For tests with horizontal rotation, 9 pilot bulbs were used, four to the left and right of the lens, whereas, for vertical rotation, 10 were used, 5 above and 4 below the lens.

4.1.5 Results

Horizontal Movement of the Looker's Eyes

As there was no significant difference between the scores using normal scanning and the scores for reverse scanning ($\chi^2 = 2.6$), they were added together, in order to simplify the presentation of results. The results for horizontal rotation of the looker's eyes are shown in Figure 4.5. Four groups of subjects were used, therefore each point on the graph represents the outcome of 288 judgements. The standard deviation (or threshold) of the distribution shown was 2.7° and the mean of the distribution (or CE) was 0.5° .

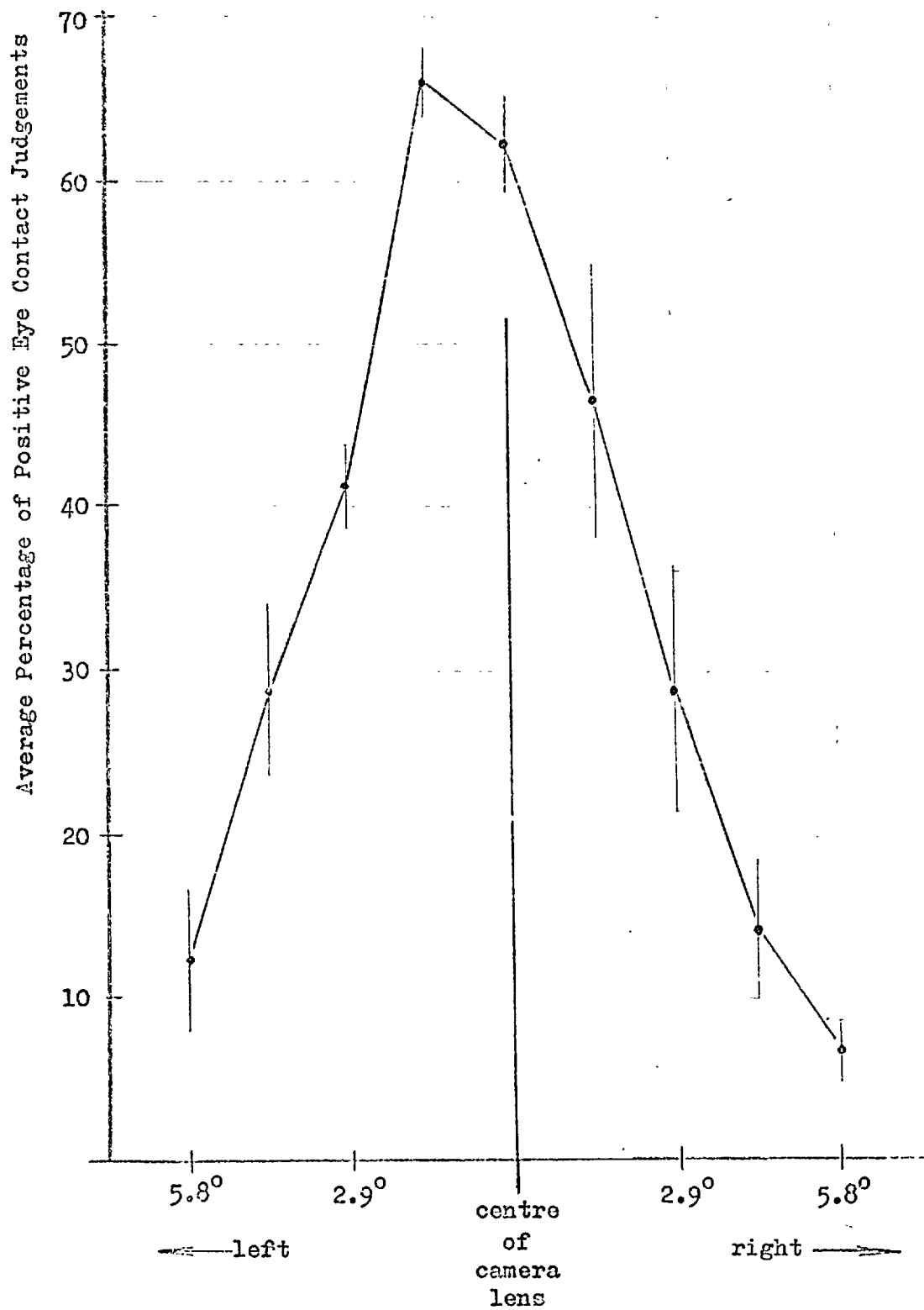


Figure 4.5

Results of the Experiment on the Perception of Eye Contact with Horizontal Rotation of the 'Lookers' Eyes.

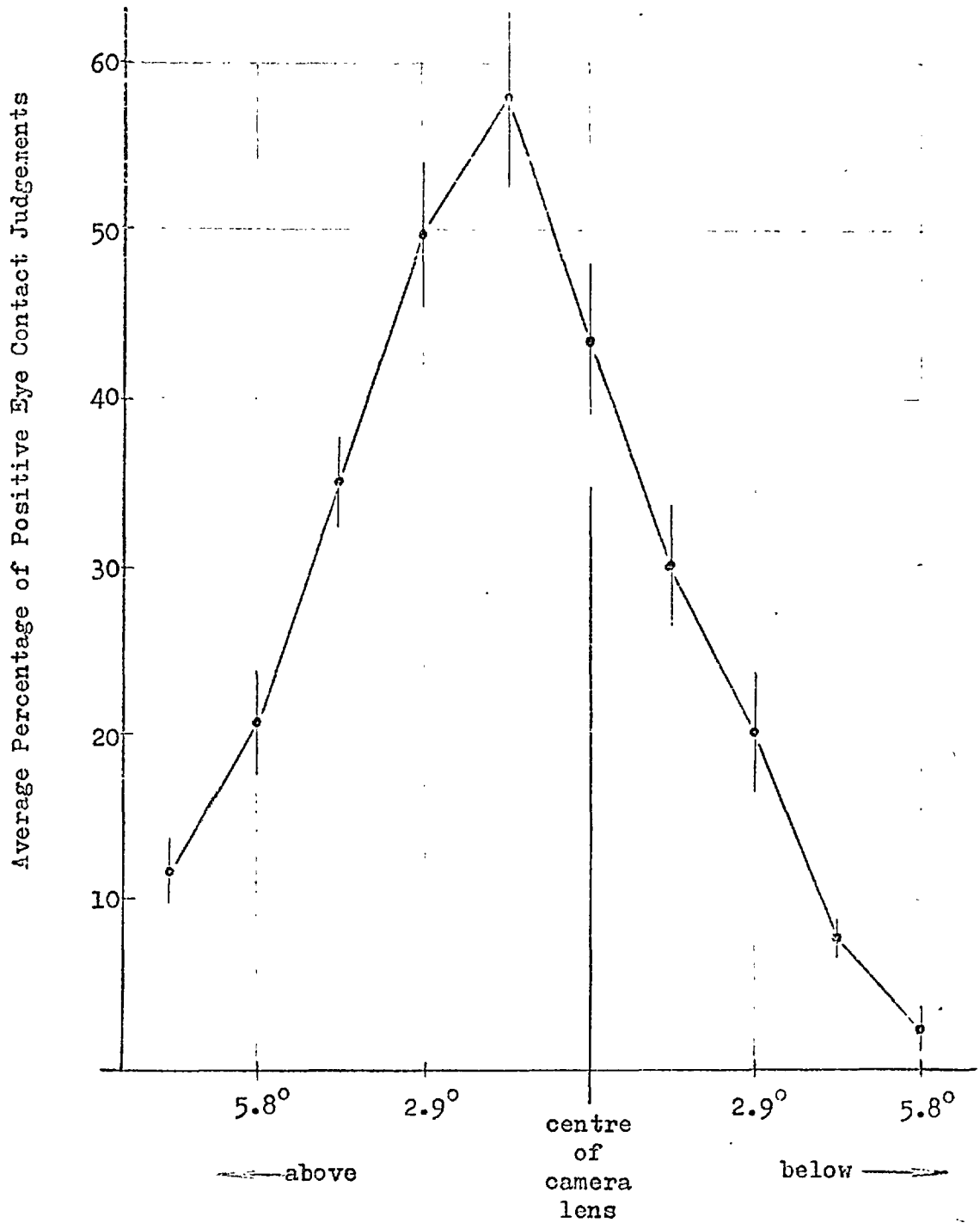


Figure 4.6

Results of the Experiment on the Perception of Eye Contact with Vertical Rotation of the 'Lookers' Eyes.

Vertical Movement of the Looker's Eyes

Again, there was no difference between normal and reverse scanning, so they were added together. The results for vertical rotation of the looker's eyes are shown in Figure 4.6. Five groups of subjects were used, so that each position on the graph represents the outcome of 360 judgements. The threshold was 2.8° and the CE was 1.5° . This CE of 1.5° corresponds to the looker gazing 2.7 cm above the centre of the camera lens. For the sake of brevity, the CE for horizontal and vertical rotation of the looker's eyes will be referred to as CEH and CEV, respectively.

To obtain a confidence level for the measured values of CE, a χ^2 test was used, on the null-hypothesis that the distribution had the same shape but that CE was equal to zero. This procedure required estimates of the numbers of positive judgements for values of rotation of the looker's eyes, between those measured. Everett's Interpolation formula was used to estimate these values (HARTREE, 1957). A similar procedure was used to calculate the confidence intervals of the means of the two distributions.

A summary of the statistical analysis is shown in table 4.1.

Table 4.1 Summary of the statistical analysis of the measured constant error values.

	CE	rejection of null-hypothesis that CE = 0			95% confidence interval of CE
		χ^2	d.f.	p	
CEH	0.5°	37.	8	<.001	$0.5^{\circ} \pm 0.4^{\circ}$
CEV	1.5°	283.	9	<.001	$1.5^{\circ} \pm 0.4^{\circ}$

4.1.6. Discussion of Results

Both the standard deviation of 2.8° for vertical rotation and 2.7° for horizontal rotation of the looker's eyes, are very close to that measured by Gibson and Pick (i.e. 2.8°) for a non-video link situation. It is interesting to compare the results of the experiment with those measured with the Bell Telephone Laboratories, 'Picturephone', see STOKES (1969). With the Bell system, it was found that the 'off set viewing effect', or 'lack of eye contact', became noticeable when the looker's horizontal rotation of the eyes exceeded 4.5° and the vertical rotation 5.5° . The large disparity in this result, and the results of the experiment reported here, are most likely due to:-

- 1) the different size and number of lines used for the video monitor image,
- 2) different method used in measuring or even defining the threshold of this effect.

As there was no significant difference in the CE's measured for reverse and normal scanning, this eliminates the possibility of the viewing conditions and any peculiarities of the video monitors, influencing CEH; however, it does not eliminate the possibility that they will influence CEV. The small CEH may be due to possible unassymetrical illumination of the looker. The much larger CEV is probably due to a combination of viewing conditions, overhead illumination of the looker, and the peculiar movement of the eyelids for vertical rotation of the eyes. This peculiar movement involves the upper eyelid tracking the upper edge of the pupil over the whole range of its vertical movement, whereas the lower lid moves downward for downward gazes but does not move upward for gazes above eye-level.

Cline found a similar, but smaller, CEV for 'live' lookers, while for photographs of the looker's face, McGrath and Hatcher found an opposite CE i.e. a CE below centre, (see section 4.1.1).

The average percentage of times that subjects thought there was eye contact over the video link was generally low. This may be partially due to the pessimistic interpretation that some subjects placed on the instructions, i.e. they pressed the green button only when they were absolutely sure that there was eye contact. Another disturbing influence is that the looker focuses his eyes on the camera lens, which is only 1 metre in front of him, yet to the viewer the looker appears to be at a distance of 2.5 metres. This meant that the looker's eyes appear to be focused at some point in front of the viewer and not at the viewer's eyes. This also accounts for some of the viewer's comments, that only one of the viewer's eyes appeared to be looking at them. This is an insurmountable problem when video telephones use small monitor screens and cameras which are relatively close to the user. There were some subjects who felt that there was never any eye contact. This helps to explain the failure of the author's earlier attempts to measure the angle of rotation of the looker's eyes for loss of eye contact by the well-known up-down technique (WETHERILL and LEVITT, 1965).

The main conclusion from the results of these experiments is, that given a square video monitor screen with the facial image approximately central, then the best place to have the camera to obtain 'eye-contact' is below the monitor screen. It may be noted that the Bell system has the camera above the monitor screen, but of course this was determined by other considerations, such as the graphics facilities. The conclusion that the camera should go below the monitor screen is an unfortunate one for future experiments with the particular equipment available, as it is the one arrangement that is impossible. This is because all the controls for the monitor are below the screen, and the position of the centre of gravity of the monitors makes inverting the monitor difficult.

4.1.7. Final Positioning of Camera

With the equipment to be used in later experiments, the shortest distance possible between the camera lens and the 'eyes' of the facial image on the monitor was 15 cm, and to the side. In this position the lack of eye contact is very noticeable. To overcome this detrimental effect in the conversation experiments, a jig for holding the monitor, camera and half silvered mirror was made. This enabled the camera lens to be at the same apparent position as the eyes of the video monitor image. A photograph of the complete terminal is shown in figure 4.7 and a schematic diagram of the relative positions of the camera, video monitor and half silvered mirror are shown in figure 4.8.

The half silvered mirror reflects the monitor image, but allows light from the viewer through to the camera. As the camera lens is surrounded by black cloth to exclude extraneous light, it is completely invisible to the viewer. The mirror allows 50% transmission of light through it, thus the aperture of the camera lens has to be opened one more stop than usual.

The video monitor was arranged to be at the end of a short tunnel. This helped to reduce the viewer's movements and to increase the contrast ratio of the video monitor screen, (see Section 3.4). The arrangement with the half silvered mirror and short tunnel has proved very successful in the videotelephone conversation experiments carried out by the author, as the users very rarely moved off-camera. The dimensions of the tunnel were 0 inches high, 8 inches wide and an equivalent depth of 11 inches.



Figure 4.7 Photograph of a person using the terminal equipment.

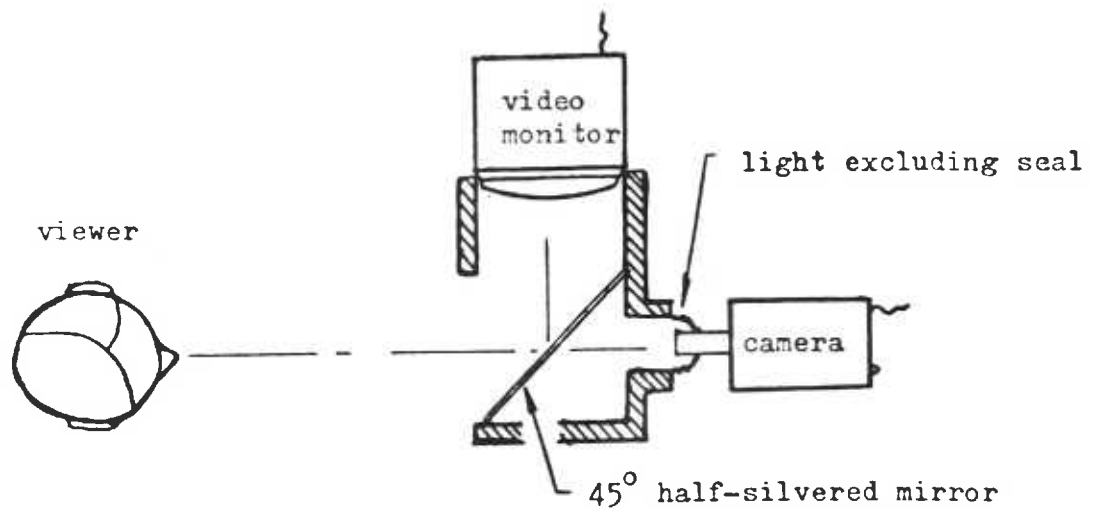


Figure 4.8 A schematic diagram of the relative positions of the camera, video monitor and half silvered mirror used in the terminal equipment.

If the conversation carried out over a video link is to be as similar as possible to a face-to-face conversation, then the eyes of the facial image on the video monitor screen should be at the same height as the eyes of a person engaged in a face-to-face conversation. When this was tried using the author's system, even illumination of the viewer's face proved difficult, so the video monitor and camera were lowered slightly. This meant that in the experiments, the centre of the monitor screen typically appeared 6° below the horizontal.

The video monitor, loudspeaker, microphone and tunnel were placed on a table, whereas the camera was mounted on a tripod behind the half-silvered mirror, and was not visible to the subjects. (See figure 4.7.) During the conversation experiments, each subject sat on an upright chair in front of the terminal equipment. The chair was placed as close as possible to the table, thus the subject's forward movement was restricted by the table and his backward movement by the chair.

The distance between the monitor screen and the subject's eyes was about 33 inches, i.e. six times the picture height. This ratio of viewing distance to picture height was used because it was the same as that used when performing subjective tests on 625 line broadcast television images. Although it was argued in section 3.5, that this might not be the optimum or preferred distance for a person to be seated during a videotelephone conversation, this distance was chosen as no work has yet been done on what is a more suitable distance.

The main source of illumination during the experiment was natural light from a window near the terminal equipment. Some control over the level of illumination was obtained by the use of venetian blinds. Typically, the range of illumination varied from 400 to 500 lux.

The author would like to point out that he does not suggest that the terminal equipment used in his experiments is the most suitable design for any future commercial videotelephone systems . The author does, however, recommend that when a decision has to be made as to the relative position of the camera and video monitor in any future systems, that a simple experiment similar to the one described be carried out.

4.2 Audio Channel

The audio system that will be in widespread use with the proposed videotelephone links, will almost certainly be some form of 'loudspeaking' or 'handsfree' telephone. The audio system chosen for the author's experiments (or any similar experiments) should resemble such a system as closely as possible. The experimental system used should also have the facility for adding some of the impairments that are likely to be encountered in commercial systems.

4.2.1 Loudspeaking Telephones

If a normal telephone handset was used with a videotelephone, the mouthpiece might obscure much of the mouth, thus nullifying some of the advantages of the videotelephone. This is one reason why loudspeaking telephones would be used; however, they also have other advantages over a normal handset which may encourage them to be more widely used in the future. A survey by FLOHRER (1970) showed that users seemed to like the following advantages:- 1) More than one person at each terminal can participate in a conversation. 2) They like to have their hands free during conversation. 3) They also like to have their hands free during dialling. This last item, in particular, has been studied by HERBERLE (1968), who showed that with normal handsets many people hold the handset too far away to hear the various tones while dialling and this gives rise to inefficient dialling. This disadvantage is completely overcome by the use of loudspeaking telephones.

Loudspeaking telephones have some disadvantages, for example local room noise is more disturbing than it is when normal handsets are used. However, many loudspeaking telephones incorporate a normal handset, which may be used in noisy surroundings, or when the conversation is of a confidential nature. The main disadvantage arises from the fact that since the acoustic-electric transducers (i.e. microphone and loudspeaker) are further from the mouth and ears than the equivalent transducers in a simple telephone handset, there must be more gain supplied in the system. As the microphone and loudspeaker are acoustically coupled, the total loop gain must be kept below some critical value. If this value is exceeded, oscillation or 'singing' will occur. The relative positioning of the microphone and loudspeaker, and the acoustical properties of the room in which they are placed, will affect the amount of acoustic coupling between microphone and speaker, and hence the amount of electrical gain permitted in the system.

The problem of 'singing' in commercial loudspeaker telephone systems is overcome by using voice-controlled gain switching, which allows full amplification only in one direction at a time (GALYAS 1969, BUSALA 1960). This introduces three main problems:- 1) although called 'voice switched systems', noises other than voices such as local room noise will cause switching, 2) clipping of the first and last syllables of a word (this is a problem of compromise in the time constants and thresholds used to detect speech), 3) difficulty of 'breaking in' during a conversation (here, once one person starts talking it is impossible for the other person to interrupt). By having some hysteresis or hang-over time in the switching, clipping of the last syllables can be avoided, but a compromise is needed with the facility to 'break-in' (see 3 above).

4.2.2 Audio system used for experiments

Because of the unknown perceptual effects that voice switching may introduce, it was decided not to use it. Great care had to be taken, therefore, in the positioning of the microphones (dynamic A.K.G. 19C), and the loudspeaker (4½ inch Goodmans). The microphones were supported on a flexible stand and were made slightly more directional by placing them in black foam tube thus reducing the acoustic coupling with the loudspeaker, which was placed on top of the television monitor, (see figure 4.7). These black foam tubes also made the microphones less noticeable on the video screen. (Note, the black foam tubes were removed while the photograph in figure 4.7 was being taken, so that the microphone would show up against the dark background.)

The rooms, in which the terminals were situated, were acoustically treated by hanging double layers of curtaining around the room. (This black curtaining also served the very useful purpose of increasing the contrast of the image for the television cameras). The gain in the audio channel was provided by a normal stereo amplifier (Leak stereo 30) (see figure 4.9).

There is a standard method of measuring the loss in a speech path between the speaker's mouth and the listener's ear for telephone connections (RICHARDS 1970). This loss can be described in terms of 'reference equivalent', which is the setting of a well defined, though arbitrary, reference system (N.O.S.F.E.R.). This system will produce the same impression of loudness, when a listener hears speech over it, as over the test speech path. This measurement system consists of a microphone, earphone and associated amplification with well defined characteristics. A similar standard method for defining the loss of loudspeaking telephone systems has yet to be established.

Because of the lack of a standard method and also the lack of standard equipment for measuring the loss of the audio system used, an arbitrary method had to be devised. Any method chosen should measure the loss between the speaker's mouth and the listener's ear with as much precision as possible. The method used consisted of taking the difference in levels measured by two half-inch capacitor microphones. One was placed in front of a simulated speaker's lips, and one near the ear of a simulated listener. The audio source was a compact 'hi-fi' loudspeaker (Goodman's Maxi) enclosed in rubber foam. A hole in the front of the foam was made to be approximately the same size as an open mouth. This source was placed in a position that a typical speaker would take, relative to the terminal equipment. At the other terminal a simulated body was placed in a position that a listener would normally adopt, i.e. about two feet from the loudspeaker

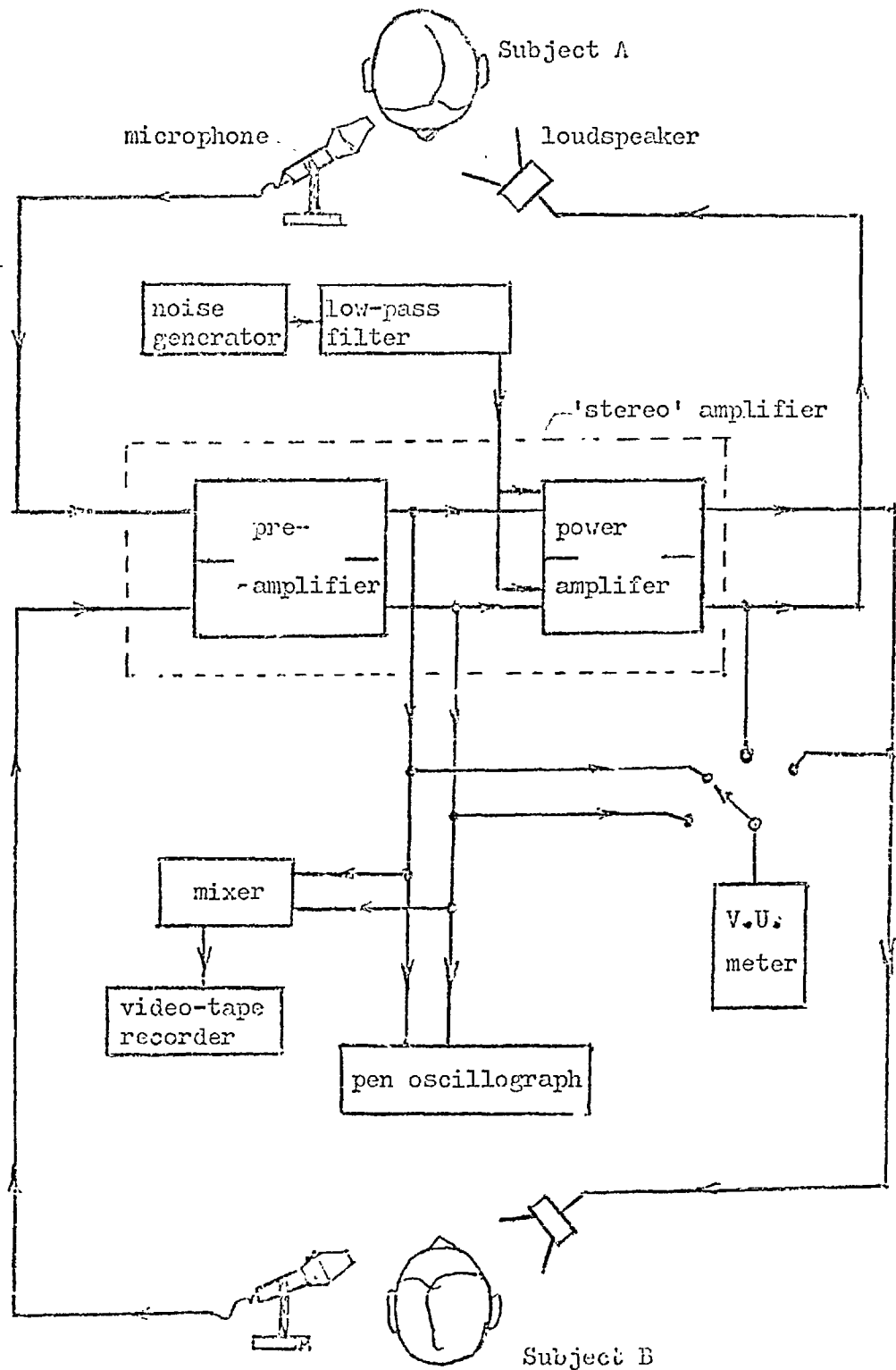


Figure 4.9 Schematic layout of the Audio Channel.

and one foot from the loudspeaking telephone microphone. The distance of the capacitor microphones from the simulated mouth and ear is shown in figure 4.10. Earlier calibration of the capacitor microphones (B & K type 4134) and their associated amplifiers (B & K type 2605) showed that the levels measured with them agreed to within $\pm \frac{1}{4}$ db over the frequency range of interest. The measured frequency attenuation characteristic is shown in figure 4.11.

The frequency pass-band of the system was approximately 200Hz to 3500Hz. This is slightly larger than that normally used for telephone channels i.e. 300Hz to 3400Hz (C.C.I.T.T., 1968). GLEISS (1968) has shown that a 100Hz band at the lower end appears to contribute as much to the naturalness of transmitted speech as does a band 5 to 10 times as wide at the upper end. This means that speech over the audio system used in the experiment should sound more natural than that encountered in normal telephone channels, but, no doubt, normal telephone channels will improve in the future.

4.2.3 Noise

There are several different ways in which the speech path between the speaker and the listener may be degraded. Typically in telephone systems, there will be loss, echo, added noise, non-linear distortion, long propagation delays and asymmetry in transmission. Although changes in any of these degradations may also change the amount that the subjects look at each other, for simplicity in this experiment only additive noise was considered.

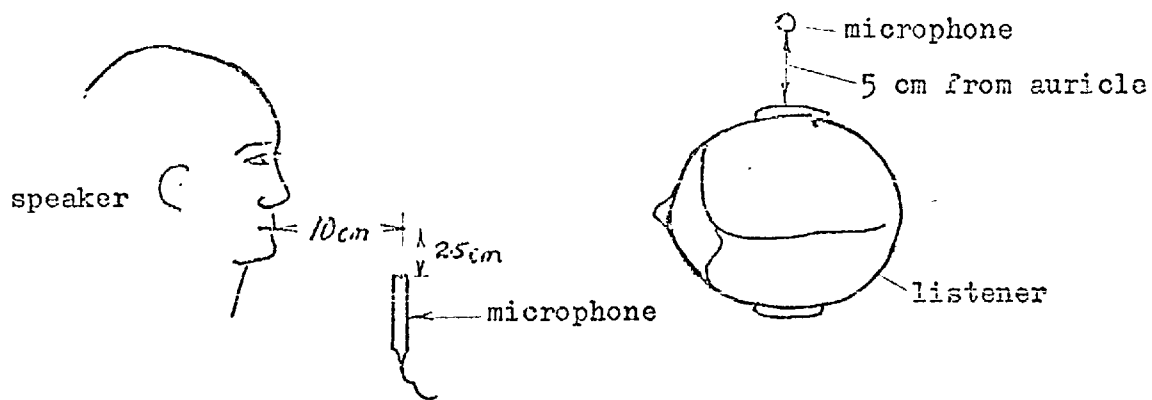


Figure 4.10 Position of the capacitor microphones when measuring the characteristics below.

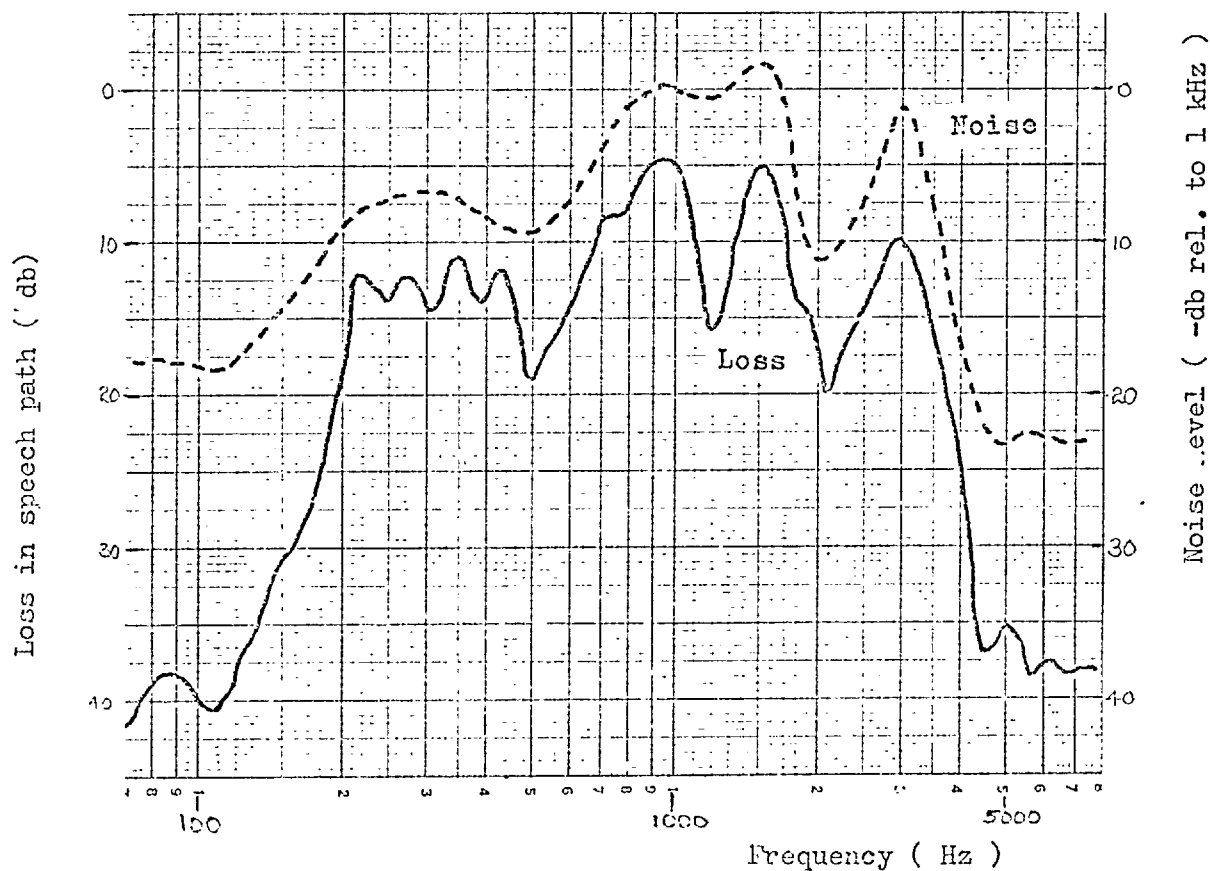


Figure 4.11 Frequency - Loss characteristic of the audio communication system, and the spectrum of the masking noise.
(noise level = $0.65 \text{ dynes} / \text{cm}^2$)

Noise from a white-noise generator (General Radio type 1390-B) was first filtered by a simple low pass R.C. filter (cut-off frequency 5000 Hz). The filtered noise was added to the audio system between the stereo pre-amplifier and the power amplifier (see figure 4.10). This allowed the speech to be recorded with little noise present. The noise levels were monitored by measuring the RMS voltage across the loudspeaker coils with the same VU meter that was used to measure the speech levels.

With the noise level at 0.1 volts RMS across the loudspeaker, the level and spectrum of the noise at the listener's ear was measured (these are shown in figure 4.11). The noise level was measured with a $\frac{1}{2}$ inch capacitor microphone, placed 5cm from the listener's ear. Using a B & K type 2603 microphone amplifier with B weighting, the noise level measured was 0.65 dynes per sq. cm., (or 70.2 dB above the standard reference level of 2×10^{-4} dynes per sq. cm.). At this level of noise the peak speech to noise level was typically - 6dB. The spectrum was measured by examining the output of the microphone amplifier with a spectrum analyser (Tektronix type 315).

4.3 The Laboratory Video System

A complete video system had to be built which would fulfil the following requirements:-

- 1) It must be able to record the video images of both the participants in the video telephone experiments.
- 2) It must be able to process the video signals in various ways for any perceptual experiments, e.g. adjustable blanking, and the addition of noise.
- 3) It must allow the rapid switching of various different video signals between different sub-units, enabling simple and efficient organisation of experiments.

The following sections deal, very briefly, with some aspects of the system.

4.3.1 Video Synchronising System

As the horizontal edge detection circuit described in section 3.7 used a fixed video delay line (64 usec delay), the line frequency had to be very stable. The noise-stripping circuit, described in the same section, required clock pulses at a rate which was at some integral multiple of the line frequency. These two requirements were met by building a quartz crystal clock pulse generator, followed by a binary divider, (see figure 4.12).

To allow the apparatus to be useful for experiments other than the author's, the clockpulse frequency was constructed to be twice the maximum video frequency i.e. at the video Nyquist rate. The video bandwidth normally used for 625 line systems is 5.5 MHz, however, for simplicity, a video bandwidth of 5.0 MHz and a clockpulse frequency of 10MHz were used.

The binary divider divided the 10 MHz clockpulse frequency by 320 to produce a 31250 Hz square wave, i.e. twice line frequency. This square wave drove a commercial synchronising pulse generator (S.P.G.). The S.P.G. (Philips EL8250) produced horizontal drive pulses and complete blanking pulses. All these pulses were as specified by the C.C.I.R. standards for synchronous scanning.

All monitors used in the experiments were modified by the author so that they would synchronise to separate synchronising pulses and not to the composite video signal. This simplified many of the video processing circuits, as it meant that the synchronising pulses did not have to be re-inserted into the video signal, if that processing circuit had removed them.

A unit was also constructed which gave adjustable blanking pulses. These pulses could be used to define any arbitrary horizontal rectangle on the video image. An example of its use is shown in Figure 5.1, where a speaker's lips have been blanked out. The unit consisted of four adjustable monostables (see Figure 4.12). These monostables MH1, MH2, MV1 and MV2, respectively, determined the horizontal position, width, vertical position and height of the required rectangular area of the screen.

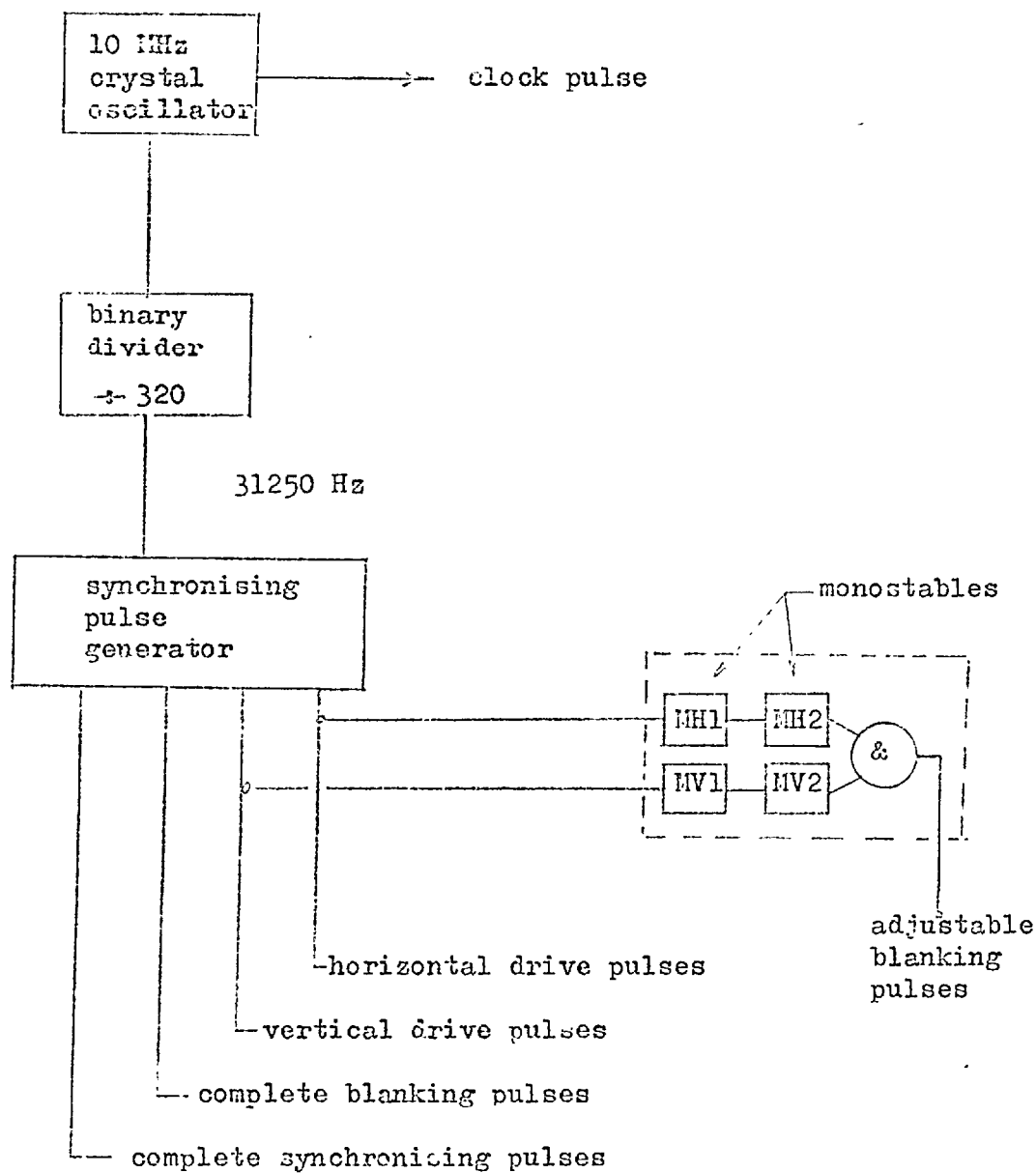


Figure 4.12 Schematic Diagram of Laboratory Video Synchronising System.

4.3.2 Video Control Unit

A video control unit was constructed, which enabled the video signals in various parts of the system to be switched and monitored. The unit contained two identical sub-units which controlled the blanking, gain and type of image that was sent to each of the two videotelephone terminals. (See Figure 4.13).

To facilitate the analysis of the non-verbal behaviour of the videotelephone users, it was necessary to record the two different images from both videotelephone terminals. However, only one videotaperecorder was available (Ampex VR 7003). Several different methods were tried, and the most satisfactory one proved to be the simple addition of the two video signals. This gave an image, which consisted of both the viewer's head and shoulders 'ghosted' on top of one another on the screen, (see Figure 4.14). The subjects were seated slightly off-centre of the camera, so that the position of their eyes did not coincide on the recorded image. This type of image proved completely satisfactory for determining the direction of gaze of both the participants in a videotelephone conversation.

For recording, the videotaperecorder required a composite video signal, so that the simple video addition circuit was followed by a system of reinserting the synchronising pulses into the video signal. (see Figure 4.13.)

A simple schematic diagram of the complete terminal and recording equipment layout is shown in Figure 4.15.

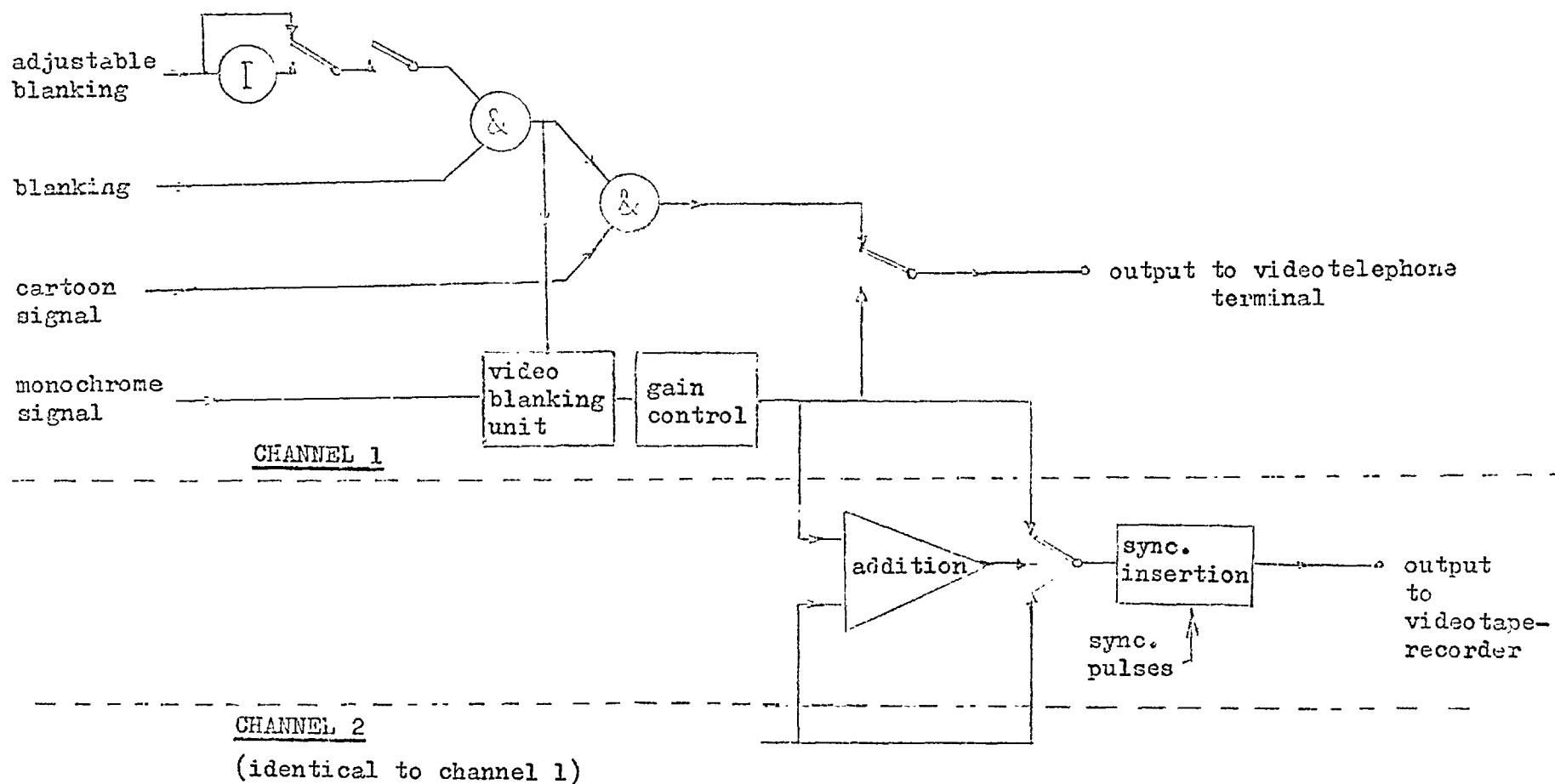


Figure 4.13 Schematic Diagram of Video Control Unit.

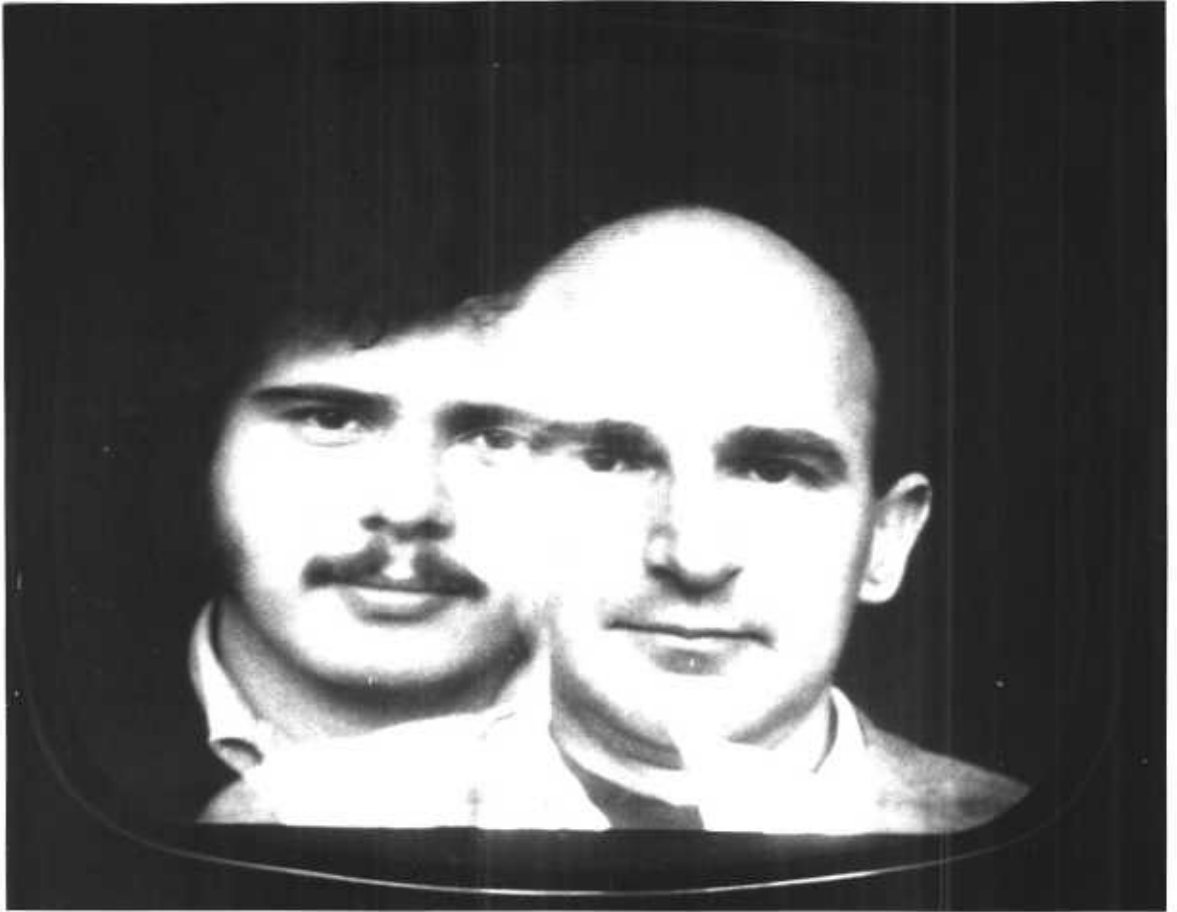


Figure 4.14 An example of the type of image that was recorded during the videotelephone experiments. This type of image enabled the non-verbal behaviour of both the participants to be studied.

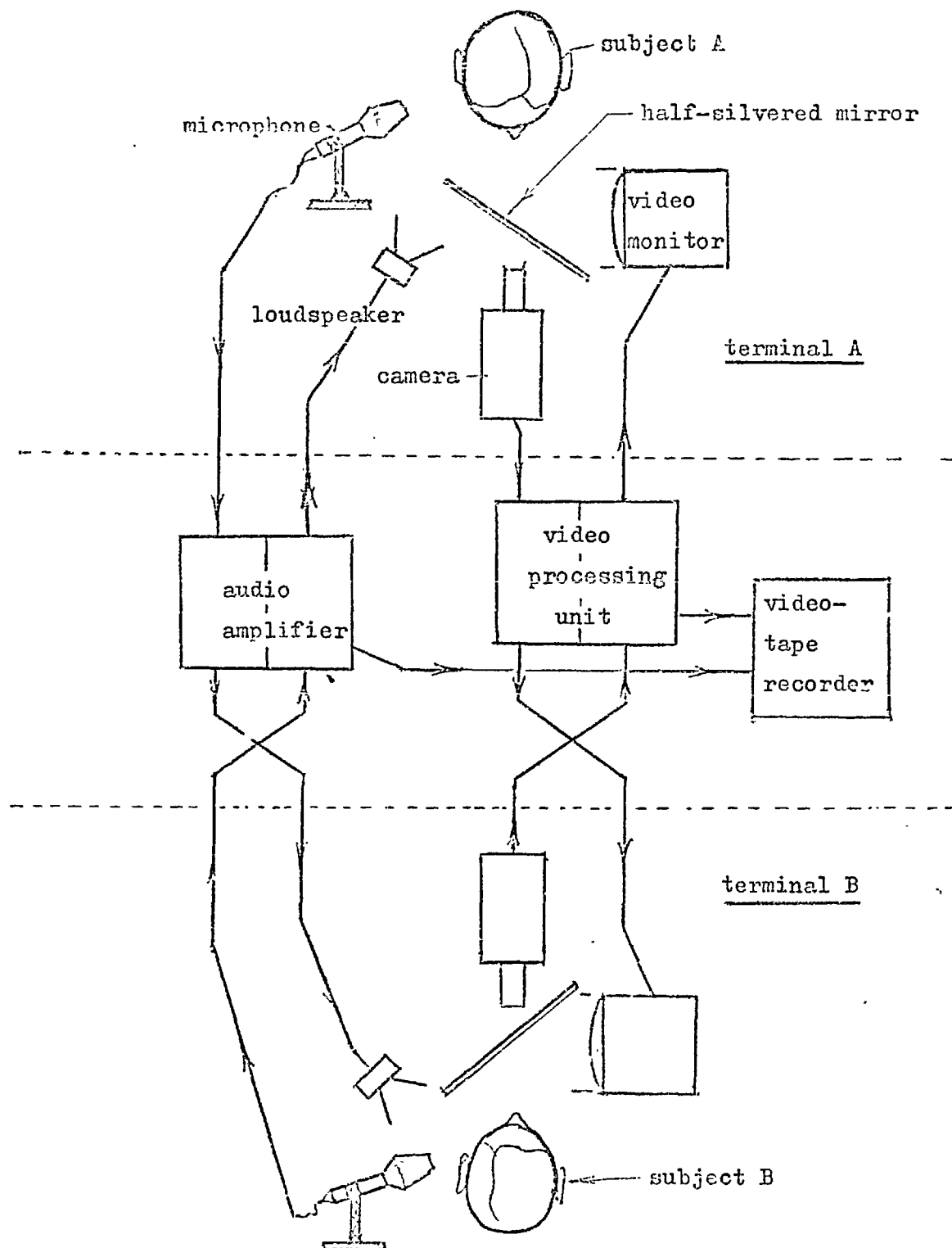


Figure 4.15 Schematic Diagram of Terminal and Recording Equipment Layout.

4.3.3 Video Bandwidth

As the system was designed for a 5.0 MHz video bandwidth, ideally a low-pass linear phase filter with a cut off frequency of 5.0 MHz should be used after each video source i.e. the cameras. Construction of two such filters (design as in WEAVER, 1958) was undertaken by the technical staff of Imperial College, but owing to an unfortunate accident, the construction was not completed before the video-telephone experiments took place.

The frequency of the system from 15 kHz to 10 MHz was measured using an Airmec type 201 sine wave oscillator and a Tektronix type 545A oscilloscope. The sine wave generator was fed into the input of the video control unit and measured at the cathode of the 'picture tube' of the video monitor. As part of the system was DC coupled, a small DC component was added to the sine wave to ensure that no clipping of the waveform took place. The results (Figure 4.16) indicated that the voltage gain was constant from 30 kHz to 300 kHz and then fell gradually until at 3 MHz it was 3dB down on the level at 15 kHz.

By using a standard resolution chart (Marconi Resolution Chart No. 1) which was fixed in front of the camera, it was found that the horizontal resolution was such that 450 'lines per picture height' were discernable; this is equivalent to a bandwidth of 5.6 MHz.

Voltage Gain
dB relative to
15 kHz level

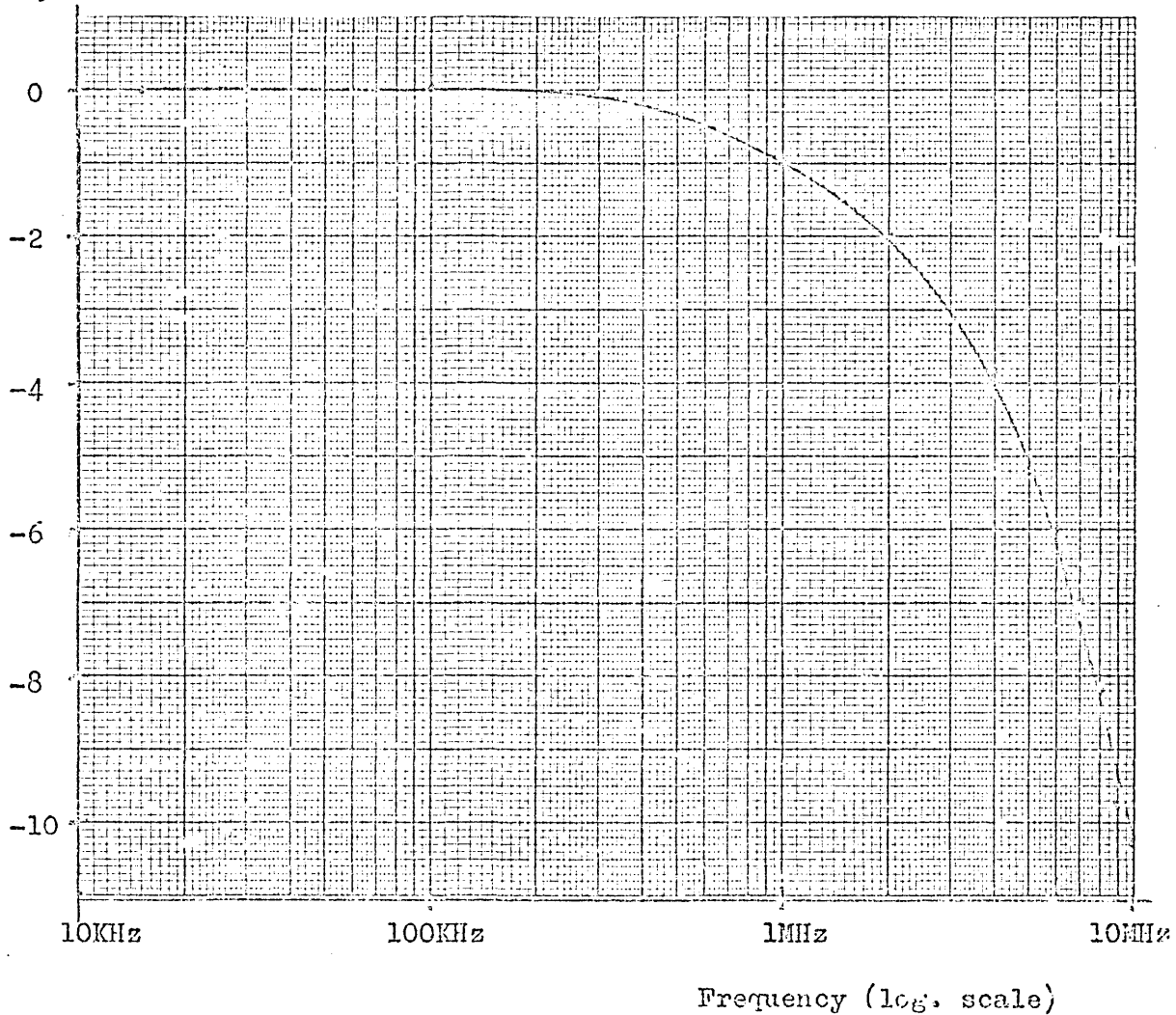


Figure 4.16 System Frequency Response
(Measured between the input to the Video Control Unit and the cathode of the C.R.T. in the video monitor.)

4.4 Data Processing Apparatus

A schematic diagram of the apparatus used for processing the experimental data is shown in Figure 4.17. In the author's experiments it was used in the following manner. Each conversation in the videotelephone experiment was recorded on to video tape. These recordings were later played to groups of observers, each observing one particular aspect of one of the subject's behaviour during the conversation. Each observer had a button which he would press when the particular behaviour that he was looking for occurred. Each observer's button was connected to one track of a paper tape punch unit, which consisted of a DM punch drive unit (type DM5021) and a Westrex, 8 track paper tape punch. The paper tape was stepped forward at a rate of 10 steps per second, and when the observer pressed his button, the unit produced holes (10 per second) along the length of the paper tape in the track to which the observer was connected. The unit could handle a maximum of 8 observers simultaneously. Therefore, the timing and duration of 4 different forms of behaviour for each of two subjects could be transcribed from video tape to punch tape with a timing accuracy of 1/10 second.

The paper tape was later read by a paper tape reader (Tally type 464). The paper tape in the reader was stepped forward at a rate of 10 steps per second. This rate was chosen because of the limited speed of the event counters which followed the reader. Each output of the tape reader gave out a single pulse, which lasted for the whole period that there were adjacent holes on the tape passing through the reader. The output, therefore, corresponded exactly to the original 'button press' of the observer connected to that track.

Various combinations of the reader outputs were then taken to a logic patch-board, enabling the coincidence of various forms of behaviour to be analysed. For example, in the videotelephone experiment described

in Chapter 7, four observers were used; two observing the two subjects' speech and the other two, the subjects' looking behaviour. Using the logic patch-board, it was then possible to analyse the amount of 'looking while listening' or 'looking while talking'.

The output of the logic patch-board was fed via buffer amplifiers to electromechanical event counters and to cumulative timers. The cumulative timers were simply digital counters (I.D.L. Scanner type 1800), which counted the output pulses from an 'and' gate, driven by the buffer amplifiers and a 10 Hz pulse train. The buffer amplifiers, event counters and cumulative counters were contained in a separate sub-unit. This sub-unit contained two buffer amplifiers, each driving an event counter and cumulative timer. The two buffer amplifiers also fed an 'and' gate, which in turn drove an event counter and cumulative timer. This sub-unit could be used by itself when only two behaviour patterns and the coincident pattern of these behaviours were needed, e.g. the looking behaviour of two subjects and the eye contact between them.

In all the author's experiments, the observer had to keep his button pressed for the whole of the time that the particular behaviour which he was observing occurred. The apparatus could be used for recording events which have no well defined duration. For example, the observer could press his button once each time a question was asked. The apparatus is therefore very useful for recording and analysing a whole range of different events and forms of behaviour that occur in conversations.

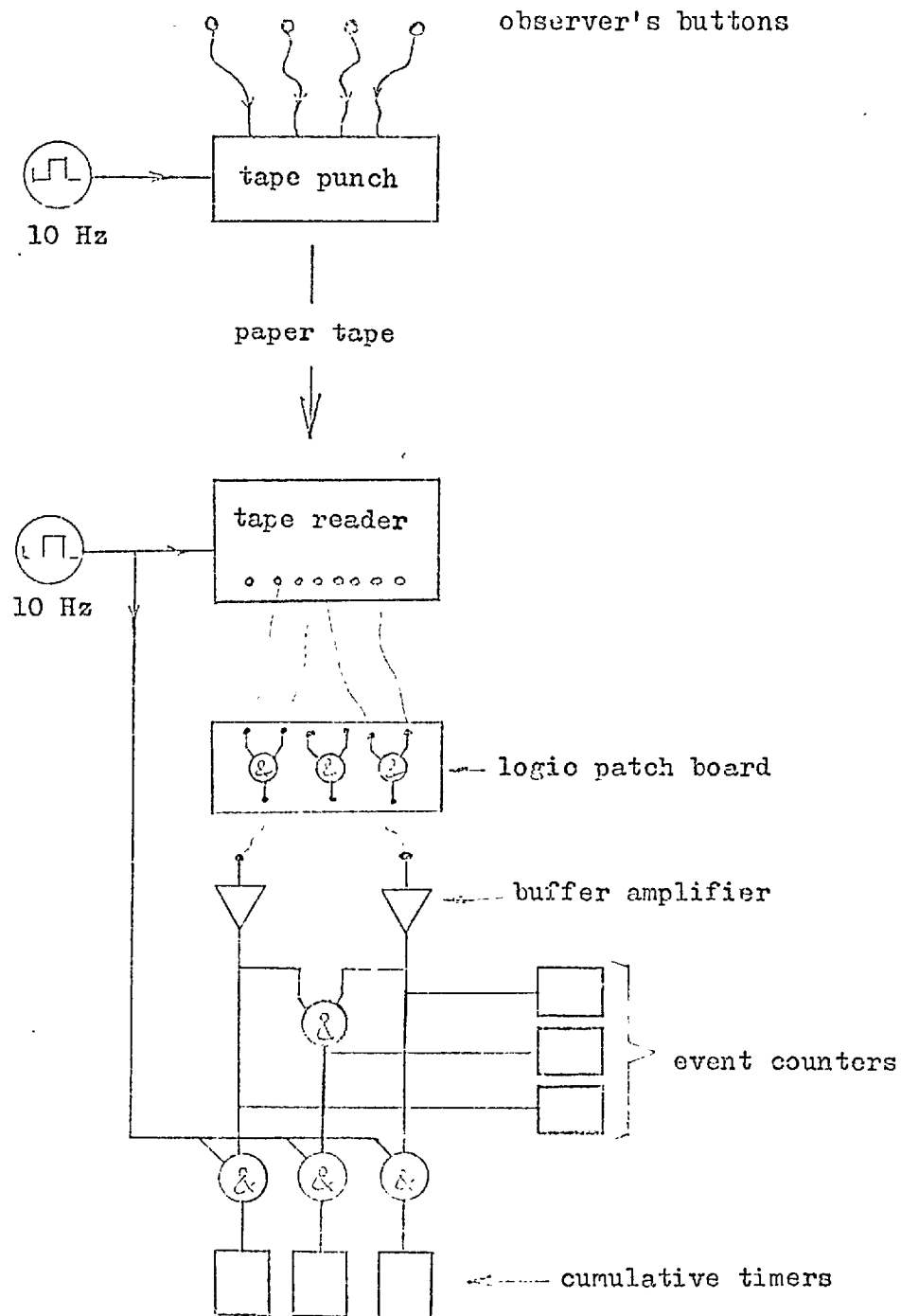


Figure 4.17 Schematic diagram of the apparatus used for processing the experimental data.

Chapter 5.Lip Reading.

This chapter describes an experiment to investigate the usefulness of different types of video images for lip reading. The results are compared with those obtained by Sumby and Pollack, who performed a similar experiment using live speakers and not recorded video images. The chapter ends with a brief review of the literature on bisensory perception.

5.1 Purpose of Experiment

There were two main purposes for carrying out the experiments described in this chapter. The first was to find out whether the video_telephone, with different types of video images, could be used for lip reading. It is often assumed that executives and managers of large corporations will be the main users of video_telephones; this will probably be true in the near future as this is the only group that could afford the expense of such a system. However, if the video_telephone proved useful for lip reading and it became much 'cheaper', the deaf and hard of hearing would be one group that could benefit from it. Because there is no compulsory registration of deaf persons, it is difficult to estimate precisely the number of persons in this rather neglected group; but GREGORY (1964) estimated that there were 2.2. million persons in Great Britain who were deaf or hard of hearing. One of the problems encountered in such a survey is that of defining 'deafness', but the definition usually used is 'evidence of disability in normal conversation'. Very little research has been done on lip reading and it is not known what proportion of deaf persons are able to lip read. The little work that has been carried out on lip reading has concentrated on a few words which are difficult to lip read, and the recognition of these words by lip reading were compared with recognition of the same words by means of electronic tactile stimulators, (see for example, PICKETT, 1963).

The second reason for performing the lip reading experiment was to clarify a classical experiment by SUMBY and POLLACK, (1954). Their experiment, reported in a paper entitled 'Visual Contribution to Speech Intelligibility in Noise', may be summarised briefly as follows. At each experimental session, six subjects were used, who were seated five feet from a speaker. Three faced the speaker, having

a full view of him, and the other three faced away. Close fitting earphones were used to carry out articulation tests for various audio signals to noise levels, and for various sizes of vocabularies. The average difference score between the two groups of subjects may then be interpreted as the visual contribution to speech intelligibility, at the particular audio signal to noise level and size of vocabulary under test.

The results showed that the visual contribution was greater at low speech to noise ratios. However, the visual contribution as defined by the difference-scores between auditory presentation alone and bisensory perception must, necessarily, be small at high speech to noise ratios. This is because there is little room for improvement with bisensory presentation compared with auditory only presentation. However, the visual contribution, relative to its possible contribution, was approximately independent of the speech to noise ratio. Also, at low speech to noise ratios the visual contribution was greater when small vocabularies were used.

Although Sumby and Pollack showed that visual cues may be used to increase speech intelligibility, they did not show which visual cues were used. Throat and facial movements may have been used as well as lip movements. Also, the visual cues may have been acting as timing cues. The aim of this experiment which was very similar to Sumby and Pollack's was to demonstrate that the main visual cue was in fact lip reading, and that cartoon images may be sufficient for lip reading.

5.2. Lip Reading Experiment

5.2.1 Use of Video-tape recorder

With the introduction of video-tape recorders, experiments similar to that of Sumby and Pollack could be repeated with greater precision and flexibility. The improved precision comes from the ability to re-record parts of the tests several times until a satisfactory speech level is achieved. It is therefore possible to produce a repeatable speech signal combined with a visual image, which remains unchanging with time or number of replays; an impossible task for a live speaker. Secondly, in live tests, as subjects have to sit near enough to see the speaker, there are dangers of direct air transmissions, due to ill-fitting earphones, giving rise to erroneous speech to noise ratios. This problem is completely overcome by using a video-tape recorder.

It has been shown by ROSENTHAL (1963) that much of the transmission of experimenter bias is carried out by unintentional non-verbal cues. To help alleviate this problem, Rosenthal recommended that experiments should be automated as much as possible. This may also be done with a video-tape recorder by recording all the instructions for the subjects on the same tape as the test stimuli.

The improved flexibility of using a video-tape recorder arises from the ability to manipulate the visual image used. Such manipulations are:- blanking off parts of the face, or 'close ups' of parts of the face. Also, as a live speaker is not needed for each test, the whole planning and performance of experiments is greatly simplified.

Another advantage is that it is possible to perform the experiments with many subjects at each session. The number of subjects is only limited by the number of video monitors and loud-speakers available.

A disadvantage of recorded stimuli is that video-tape recorders are not easily stopped and started, so that the test words have to come at fixed intervals. If the interval is too short, there will be difficulty in finding the relevant test word on the score sheet and the subject will then not be ready for the next word. If the interval is too long the subjects's attention may wander and the test would become uncomfortably long. It is necessary, therefore, to carry out several preliminary tests before deciding on a word repetition rate.

Another disadvantage is that if one recorded word is spoken in such a way as to make it harder to identify and this escapes the experimenter's notice, then this error is repeated throughout²¹¹ the tests. If live speakers are used for the tests, such errors are more likely to be randomly distributed throughout the whole test series. Also, as it is not possible to change the whole order of presentation of the test stimuli, it is necessary to record long series of test stimuli, so that any possible sequential effects may be balanced out.

The advantages of increased flexibility and precision, plus the simplification of experimental organization made possible by using a video-tape recorder far outweigh the disadvantages.

5.2.2. Vocabulary

Sumby and Pollack found that under conditions of low speech to noise ratio, the visual contribution (in terms of difference score) is decreased as the size of the test vocabulary is increased. Thus, by choosing a relatively small vocabulary, the difference between audio only and audio plus visual presentation will be emphasized, therefore making the comparison between the different types of image more precise. The other main advantage of using a relatively small vocabulary is that it simplifies the subject's search of the score sheet for any particular word.

A disadvantage of using small vocabularies is that any confusion between two or more particular words will have greater influence on the overall results than it would if a large vocabulary was used. If the same vocabulary is used to compare different audio processes etc., then confusion between particular words is not so important.

Sumby and Pollack found little difference in their results when using spondees or tri-syllable words; monosyllables gave lower articulation scores throughout. Spondees* have a greater homogenous audibility than other types of words (EGAN 1946) and they are, therefore, particularly useful for comparison tests of different types of equipment.

The vocabulary used in the test reported consists of 16 spondees chosen from a widely used list (EGAN 1946). As the list was American in origin, it was essential to remove words peculiar to their language such as, 'buckwheat' and 'sidewalk'.

The speech level during recording and playback was measured by means of a 'peak VU' meter (B & K type 2409).

* A spondee is a two syllable word with equal stress on both syllables.

5.2.3. Visual Images Used.

Four different visual images were used and these are shown in Figure 5.1. The first, (a), was a close up of the speaker's face, such that the complete image of his head just fitted the monitor screen. The speaker was facing slightly off the centre of the camera lens and this image will be referred to as 'face only'. The second type of image, (b), consisted of the speaker's head and shoulders. Both these first two images were 'normal monochrome raster scanned images', as described in section 3.6. With the laboratory video telephone link, an image similar in size to the 'face only' would occur if the user sat close to the terminal equipment and leaned forward, whereas an image similar in size to the 'head and shoulders' would occur if the user sat well back in his chair,

The third type of image, (c), was a cartoon of the speaker's face. It was produced by passing the video signal used to produce the 'face only' image through the cartoon generator, described in section 3.7.

The fourth image, (d), was the same as the first except that the lips were blanked out. The image allowed the contribution to lip reading of the area around and including the lips, to be determined. This was done by taking the difference score between the results of the tests using the first and fourth images. This method was thought preferable to just using the lips as an image as a pair of lips flapping about in isolation on a monitor screen, is too far removed from any visual experience that a subject is likely to have encountered, whereas, a face with just the lips blanked off does not look so unrealistic.



(a)



(b)



(c)



(d)

Figure 5.1 Facial images used in the lip reading experiment.

Photographs taken from the original video-tape recording used for the experiment.

5.2.4. Experimental Procedure

A schematic diagram of the interconnections between the various pieces of apparatus used in the experiment is shown in Figure 5.2. The video output of the video-tape recorder (Ampex 7003) was fed to two nine-inch monitors. These monitors were the same as those used in the video telephone terminals. Each monitor was placed on a table approximately three feet from the subject.

The audio signal was fed via a stepped attenuator to a mixer, where it was mixed with white noise from a noise generator (General Radio Type 1390 - B). The noise output was adjusted so that the stepped attenuator setting was equal to the signal to noise ratio at the output of the mixer. The output of the mixer was amplified by a commercial amplifier (Leak Stereo 30), which, in turn, drove a high quality composite loudspeaker

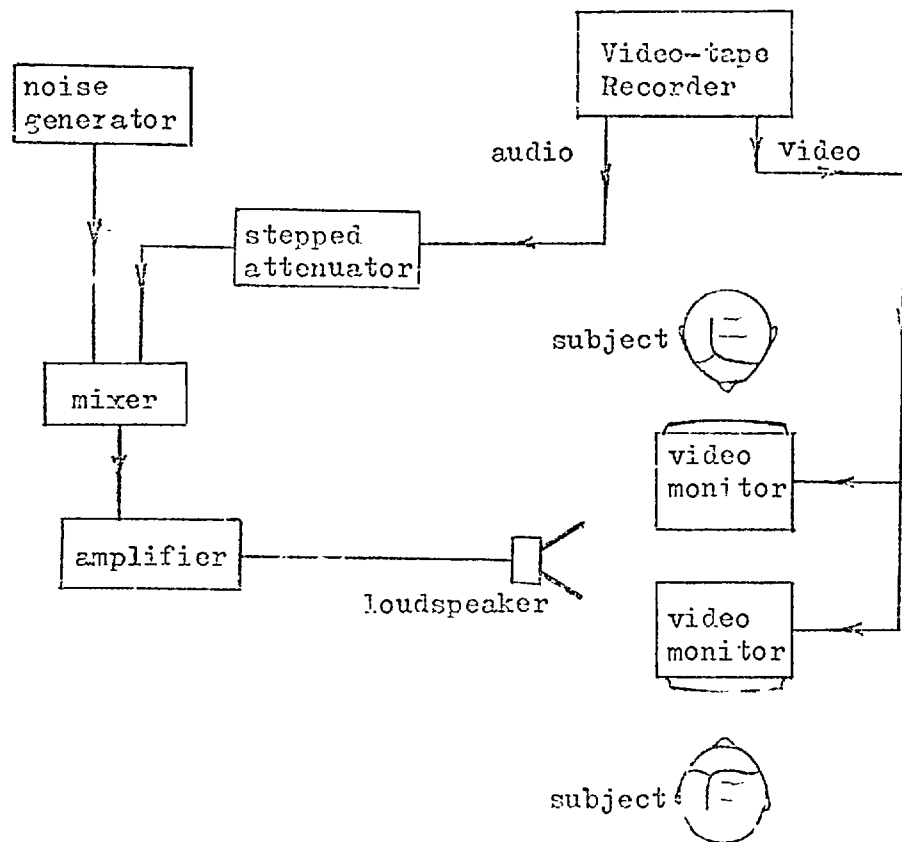


Figure 5.2 Schematic diagram of apparatus layout.

(Leak Sandrich). This arrangement meant that two subjects could be used at one time. The frequency passband of the audio system was 50Hz and 6200Hz.

The subjects were given the score sheet and allowed a few minutes to become familiar with the words. At the beginning of the recording, the speaker read the list of spondee twice, and gave all the instructions required for the test, plus a few examples.

The test consisted of sixteen blocks of sixteen spondees. In the first eight blocks, the choice was limited to the first or last eight words on the score sheet. These first eight blocks were used to ensure that the subjects were familiar with the words and their position on the score sheet and also to familiarize subjects with the different forms of visual presentation. The subjects were not informed that this was a 'practice session'. The last eight blocks used all sixteen words on the score sheet. Each block was presented with one of the four types of visual image. The order of presentation of the visual images was made non-cyclic, so that carry-over effect would be reduced. After each block was presented, taped instructions told the subjects which numbered sheet and which words were to be used next.

Each test word (a spondee) was prefixed by another spondee; this was only used as a timing cue. The speaker, normally looking down at this word list, looked up as he said the second spondee. The subjects were fully instructed about this procedure. The pairs of spondees were presented at a rate of 1 pair per 10 seconds.

Noise was switched on only during the blocks of sixteen spondees. The same speech to noise ratio was used for each type of the four visual images presented in turn, the ratios being randomly chosen from 6dB steps over the range 0 - 30dB.

A total of thirty subjects were used; drawn from university students or secretarial staff. Persons accustomed to wearing spectacles were asked to wear them.

5.2.5. Results

The results are presented graphically in Figure 5.3. Analysis of variance was performed on the data for the images used for lip reading i.e. excluding the 'blanked off lips' image. All the subjects achieved 100% correct scores for the 0dB speech to noise ratio, so this particular noise level was excluded from the analysis. The analysis of variance, shown in table 5.1., took into account the fact that each group of subjects was exposed to every type of video image, but not every different noise level. The form of analysis of variance to be used in such 'repeated measures experiment' is described in WINER (1962). As the results were in the form of 'percentages', an $X' = \arcsin \sqrt{X}$ was performed on the data, before carrying out the analysis of variance. (See EDWARDS 1968).

Although the analysis of variance showed that the type of image used was a significant factor, the difference between the scores for 'cartoon' and 'face only' visual presentation was very small. The difference between the two scores for different speech to noise ratios is shown in Figure 5.4. Also shown in Figure 5.4. is the difference between the 'face only' and 'blanked off lips' presentations. This difference score is equivalent to the contribution that lip reading and any movement of the blanked out area around the lips makes to the intelligibility of the spoken words. For the purpose of comparison, the difference score (i.e. the visual contribution) found by Sumby and Pollack for a vocabulary of sixteen spondee words, is also shown in Figure 5.4. At large noise to speech ratios, the difference score for the present experiment was approximately 80%, and this is very close to the equivalent difference score found by Sumby and Pollack (approximately 70%).

Table 5.1 Analysis of variance performed on the data from the lip reading experiment.

Factors are: N = noise level (5 different levels)
I = type of video image (3 types)

	SS	d.f.	MS	F	significance level
<u>Between</u>					
N	1.5784	4	0.3945	9.6	<0.05
subjects within groups	1.8468	45	0.0412		
<u>Within</u>					
I	0.3466	2	0.1733	10.6	<0.05
NI	0.3672	8	0.0459	2.8	n.s.
I x subjects within groups	1.4774	90	0.0164		

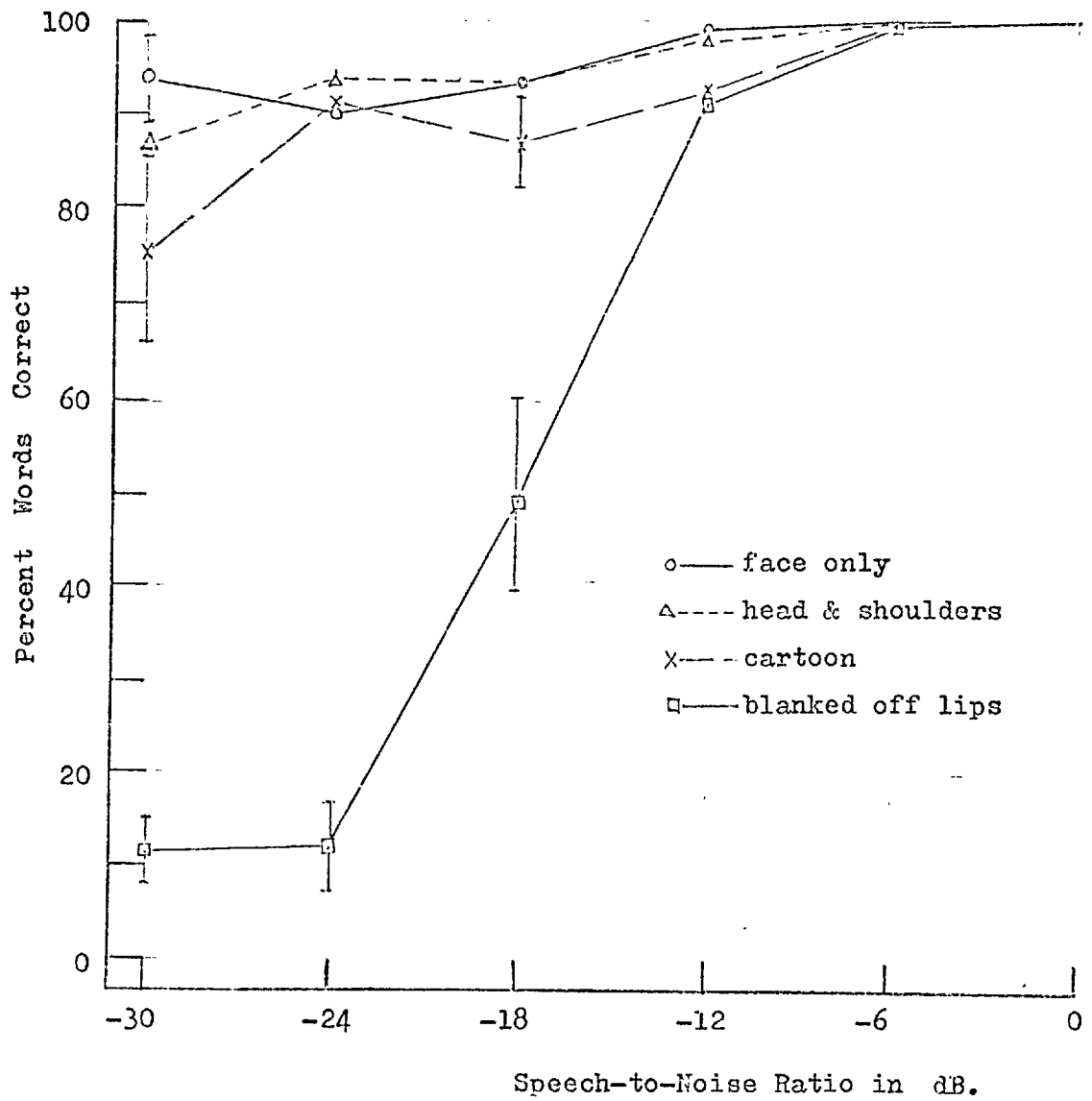


Figure 5.3 Percentage of words correctly identified at various speech to noise ratios using different facial images.

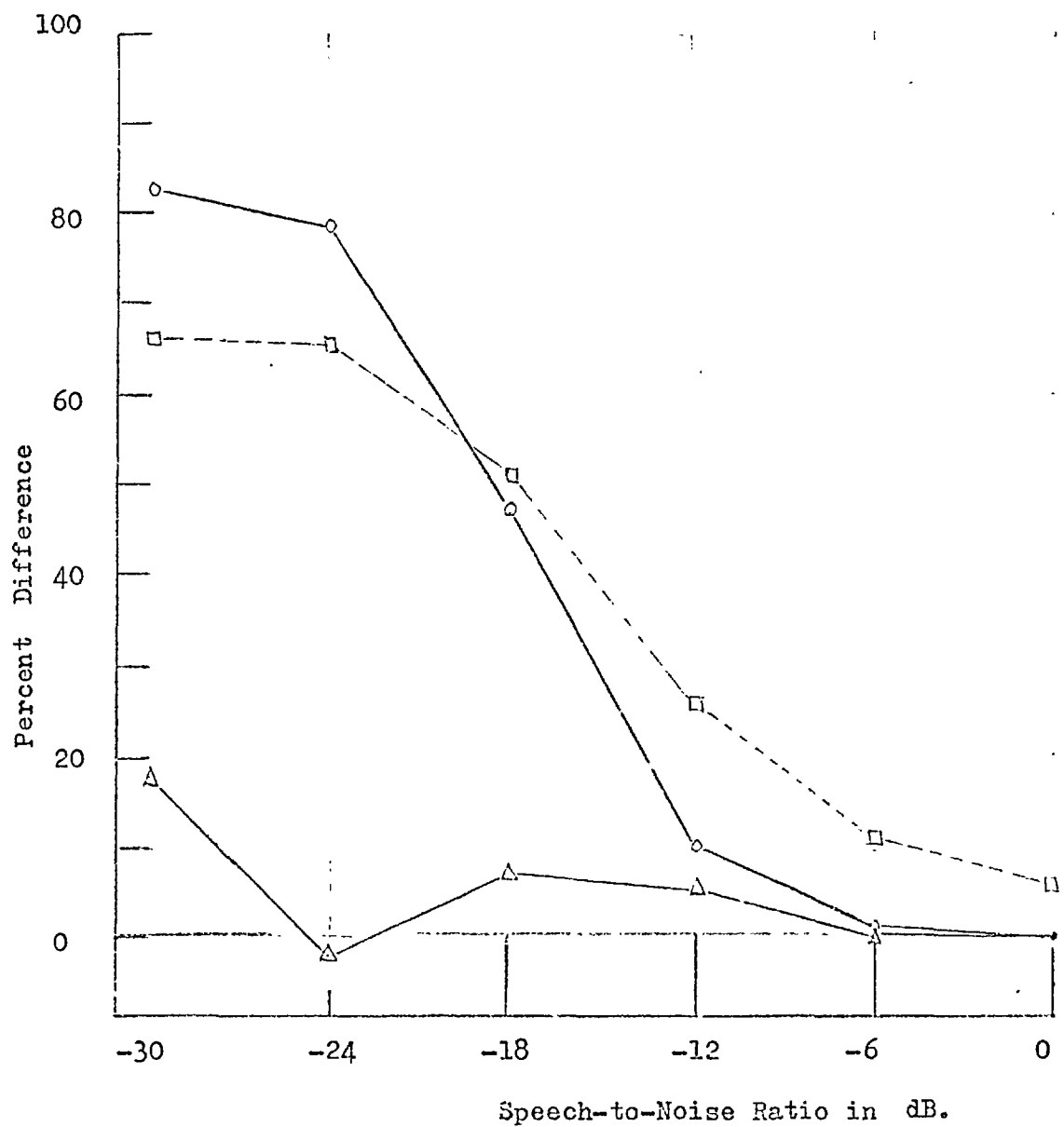


Figure 5.4 Difference scores

- difference between face only and blanked off lips images.
- △— difference between face only and cartoon images.
- difference score found by Sumby and Pollack.

If a smaller vocabulary was used, the percentage of correct responses would be normally expected to be higher than that obtained with a larger vocabulary. However, analysis of variance carried out on all the data, including the practice scores where a small vocabulary was used, showed that the size of the vocabulary was non-significant. This result of lower than expected scores in the practice sessions, was probably due to some learning taking place, and the result may be interpreted as a demonstration for the need for practice sessions before the main session.

5.2.6. Discussion of Results

The results shown in Figure 5.3 and 5.4 are very similar to those of Sumbly and Pollack, and thus it may be assumed that the main visual contribution to speech intelligibility in their experiments was lip reading.

None of the subjects, when questioned, admitted to any previous practice at lip reading and many of the subjects were surprised at their ability to lip read. It is possible that persons may use lip reading as an aid to speech intelligibility in certain noisy or quiet, solemn environments without being aware of it. An instance where a person does become aware of his perception of lip movements is when he is observing a film or television monitor and the sound and visual image are not synchronized. TAYLOR et al (1967) when studying deafness in jute weavers (jute factories are very noisy environments), found that the ability to lip read was not related to deafness.

The negligible difference between the results using a cartoon image and a normal monochrome image demonstrates how very little information need be supplied for the subject to be able to lip read.

5.3. Bisensory Perception

In the lip reading experiment reported in the earlier part of this chapter, the subjects had to perform a bisensory perception task, whereby information about which word the subjects had to identify was transmitted visually, by means of lip reading, as well as aurally.

Although the study of bisensory perception was not the motivation for that particular experiment, such an experiment is ideally suited for studying bisensory perception.

5.3.1. Accessory Stimulation

The study of bisensory perception may be split into two distinct categories. The first, normally referred to as accessory stimulation, is concerned with the perception of a stimulus via one sensory modality, whereas other, usually unrelated stimuli, are applied to another modality.

Many examples of accessory stimulation are quoted in a review of the subject by LOVELESS et al (1970). A typical example is that, when lifting different weights, subjects judged them as lighter if the lifting was accompanied by an audio tone. In another example, HERSHENSON (1962) has shown that the average reaction time to audio plus visual stimulation was less than for either singularly.

Much of the variation in the reported effects of accessory stimulation are possibly due to the wide range of primary perception tasks considered. For example, TRELSMAN & HOWARTH (1959) studied how the threshold for detecting changes in the intensity of an audio tone was affected by the timing of a visual warning signal. An experiment in which the main perceptual task was much more complex, was performed by CARMICHAEL et al. (1932), who showed how the verbal label given to an arbitrary line drawing

affected the subject's ability to accurately reproduce the line drawing.

One important field of study concerned with accessory stimulation is that which deals with the problem of noise and efficiency of working. No simple relationship has been found between noise and efficiency, as the effects of noise are very dependent on the particular task being performed. (HOCKEY 1969).

5.3.2. Combined Visual and Audio Perception

The second category of bisensory perception concerns the situation where both modalities convey relevant information. In particular, this section will deal with simultaneous stimulation of the audio and visual senses.

When a stimulus is difficult to perceive, the question arises as to whether its identification is based on a combination of evidence from both modalities, or from one only. As many experiments have shown that the probability of correctly identifying a stimulus is greater if both the ear and eye are used, compared with using either one singularly, it must be assumed that some form of combination of information from the two sensory modalities takes place. These results stimulated further experiments, which tried to elucidate which of the many models proposed for describing the combination of evidence was more accurate.

The hypothesis that has aroused most interest is that of sensory independence, namely that the eye and the ear act as independent observers of the stimulus and do not exchange information before coming to their

own decisions. If these two decisions are in conflict, the observers must choose one or the other.

POLLACK (1964) performed an experiment to test this hypothesis of independence. The audio stimulus comprised monosyllabic words recorded at five different speech to noise ratios and played back via ear-phones. The visual stimulus consisted of the same words printed on a moving belt which moved at five different speeds. Subjects had to identify the words when the stimuli were presented visually only, auditory only, or both simultaneously. The particular model tested is represented by the equation (1) below:-

$$P_{VA} = P_V + P_A - (P_V \cdot P_A) \quad \dots(1)$$

where P_{VA} = the proportion of words correctly identified with auditory plus visual presentation.

P_V = the proportion of words correctly identified with visual only presentation.

P_A = the proportion of words correctly identified with auditory only presentation.

The results for bisensory perception obtained from the experiment were slightly higher than that predicted by equation (1), but this small difference could easily be accounted for by small experiment errors.

The model used, which gave rise to equation (1), is that whenever the independent decisions taken by the eye and ear are in conflict, the observer chooses the correct one. This, of course, avoids the question of how the observer knows which of the two conflicting decisions are correct. This particular model has also been used where two different sources of information about a common stimulus are applied to one sensory modality. For example, TULLING et al (1965) used it to account for the results of an experiment where he presented words in a tachistoscope, after showing the subjects the context in which the word may be used.

GARNER and MORTON (1969) have carried out a logical analysis of the problems involved in experiments to investigate perceptual independence. They concluded that when correlated stimulus inputs are used, it is almost impossible to decide between the many different models used.

Another method of studying bisensory perception is to use shadowing* (vocalization-at-presentation). This method was used by DORNBUSH (1968), who presented subjects with visual or auditory stimuli consisting of numbers and letters. The subjects had to shadow the visual or auditory stimuli and afterwards to recall both of them. Dornbush demonstrated that equal division of attention between the auditory and visual modality was not possible.

Four different criteria that the observer may use for deciding which of the decisions is correct when the independent modalities come to conflicting decisions, have been investigated by CORCORAN and WEENING (1969). The most successful model that they proposed assumed that the two sensors accompanied their discrete decisions with a confidence measure.

CORCORAN and WEENING (1969), in their experiments, used a modulated tone masked with noise as the stimulus that had to be identified. This modulated tone was presented visually by means of an oscilloscope and aurally via earphones. As the identification of tones masked by noise using an oscilloscope is not an every day task, the subjects had to undertake long training sessions, until they reached a satisfactory level of performance.

If, however, lip reading was used as the original stimulus, this specific problem of long training sessions would be eliminated, as the results of the experiment described in the earlier part of this chapter indicated that subjects were already proficient at lip reading. Also, all the necessary equipment for an investigation into bisensory perception,

* Shadowing is the repeating out loud of items on one channel immediately as they occur.

using lip reading as the visual stimulus, already exists. For these reasons, the author suggests that further research on bisensory perception may be accomplished using the equipment described and at little cost.

If such an experiment was performed, it would not be necessary to use such extensive speech to noise ratios, or as many types of visual image, as the author did, in his experiment. By using a more careful analysis of the subject's responses, including their incorrect responses, detection theory may then be usefully applied in interpreting the results. Caution must be maintained when comparing the results of such an experiment with those of other experiments, where synthetic material was used as stimuli and where detection and not identification of stimuli was the primary task.

Chapter 6.

Identification and Apparent Proximity of a Video Image

This chapter describes two experiments, both of which were concerned with the perception of a facial image appearing on a video monitor screen. The first was an exploratory experiment carried out to investigate the relative utility of different types of video image for the recognition of a person's identity. The second experiment investigated the apparent proximity of a facial image on a video screen.

6.1. Visual Identification of a Person

6.1.1. Recognition of a videotelephone caller

When a person is called on a telephone, he will usually answer it by stating either his name, his telephone number, the name of the organization that he represents or he may simply say "hello". The caller then has to identify himself. There are a limited number of openings that are socially acceptable and the sequence of these possible openings follow a regular order (SCHEGLOFF, 1967).

If the videotelephone becomes widely used, there may evolve a set of opening rituals particular to its use. Some of the openings used in normal telephone conversation will sometimes be redundant when using the videotelephone, as the caller will recognize the person who answers his call. The caller may recognize the identity of an individual, or he may recognise the person's role from any uniform that might be worn.

The advantages of a videotelephone over a normal telephone for the recognition of a person's identity will, in most conversation, be most important in the first few seconds of the conversation. However, because of this advantage, the videotelephone may be used for purposes hitherto unconceived. For example, the police may be able to hold 'distant identity parades'.

As the recognition of a face is the recognition of a very complex form, the following section briefly reviews the literature of the perception of form. This is followed by a description of an exploratory experiment carried out to investigate the usefulness of different types of video image for the visual recognition of facial images.

6.1.2. Detection or Discrimination

Both the experiments described in this chapter investigate some aspects of the perception of facial images displayed on a video monitor. Generally, perception can be defined as follows: an organism has perceived a stimulus if the behaviour of the organism can be related to some aspect of the applied stimulus (DEMBER 1963). However, of course, perception can occur at different levels.

The purpose of the first experiment was to investigate the relative utility of various types of visual image for the recognition of a person's identity. It is, therefore, concerned with the discrimination of one facial image from another. Because the experimental method is in some ways similar to that used by MARSH (1967) for investigating the relative threshold of detection for simple shapes, it would be useful at this stage to clarify the difference between detection and discrimination. This may be done by considering a simple example. Assume a light is lit momentarily against a dark background, watched by a subject who is asked to say when he sees the light. When he does so, he is performing the process of detection. If, however, the subject is asked to say what colour the light was, or what shape it was, he would then be making a choice between alternative stimuli. Since he would presumably need to detect the presence of the image in order to make a decision about its colour or shape, the process of detection would necessarily be essential to the more complicated perceptions of discrimination (DEMBER 1963).

Further evidence of the distinction between different levels of perception arises from the studies of the neuro-physiological structures in the visual system of a cat (HUBEL and WIESEL, 1965). These studies indicated that the analysis of a retinal image occurs initially on a stage-by-stage basis, as the neural signals penetrate from one layer of cells in the cortex to another.

6.1.3. Visual Form Discrimination

The discrimination of one visual form from another is a task that most people perform many times each day. However, there has been very little work done on the perception of visual form. A reason for this was given by ARNOULT (1957), who wrote, "the greatest single barrier to the development of systematic knowledge about form perception is the difficulty involved in describing the stimulus". Only a small number of forms have names which are universally agreed upon; these are the simple geometric shapes like the triangle, circle and square etc.. With some forms, which resemble commonplace objects, there may be agreed names, but there are large variations in the possible stimuli which may be given a particular name. For example, most people may agree that a particular simple line drawing is shaped like a car, but there are very many different line drawings that could be labelled 'car'.

One of the few experiments carried out on visual form perception was performed by BOYNTON (1957). He presented subjects with shapes masked by random shapes at various contrast ratios. The subjects had to identify the particular class, from the six possible classes that were presented. The six classes were: crosses, triangles, quadrangles, pentagons, Y shapes and nothingons, which were arbitrary jagged shapes. BUTLER (1964) presented pairs of shapes tachistoscopically to subjects, who had to press one of two keys, depending on whether the shapes were similar or different. The average reaction time was taken as the measure of difficulty in discriminating the shapes. Of the shapes used, it was found that the easiest to discriminate was a circle from a triangle, and the hardest, a triangle from a diamond.

Other experimenters have investigated how much a form has to be distorted before it is not recognised as that particular form. Typical of such experiments was that carried out by VENNERS (1948), who investigated how much the width or height of a square could be altered before it was perceived as a rectangle. A similar experiment, but using more complex forms, was performed by VON DE GEER and LEVELT (1963). In their experiment, the form was generated by the movement of a spot of light.

In the experiments described in this chapter, the stimulus form is a facial image. Experiments could be performed to investigate the discrimination of facial images from other geometric forms, and this may give an insight into what features of a face are important for it to be recognised as a face. However, when receiving a call on a videotelephone, a person will usually be concerned with whose facial image appears on the screen, and not whether it is a face. A research programme has been started by GARNEAU (----) on what are the more important features of a face that are needed for identification.

There are several different ways of categorizing facial images. One approach could relate to the person's role or status, i.e. old man or child; another way could depend on the physical attributes of the face, i.e. bald headedness or beardedness. A third method of specifying a facial image is simply to give the person's identity and time of recording. If this method is used for defining the stimulus for an experiment involving identification, then the relationship between the person whose face has to be recognised and the subjects who have to perform the identification is important. Further, it must be decided how many different recordings of the facial image of the person are to be used. If only one recording of each face is used, then it should ideally be presented to each subject once only, as on further presentations

the subject may recognize that particular recording because of some aspect which is independent of the facial image, i.e. a scratch on a photograph or some detail in the background. This problem may be partially eliminated by allowing a long period between each presentation of the same image to each subject.

The following sections describe an exploratory experiment carried out to investigate the relative utility of different types of video image for the recognition of a person's identity.

6.1.4. Facial Images used for the Experiment

Photographs were taken of the entire staff and students belonging to the Communications and Electronics Section (to which the author belonged). The format of the photographs was chosen to give an image similar to that encountered on the laboratory videotelephone system. From the whole range of photographs, eight were chosen (3 staff, 5 students), and of the eight, 3 wore spectacles, 2 had beards, 1 was bald, 1 wore a moustache and 1 was a negro. By choosing more than one person who wore a beard and more than one person who wore spectacles, this eliminated the possibility that these rather prominent features alone would be used for recognizing the person.

Twenty subjects were used for the identification task. All were students belonging to the Section and had known the persons whom they had to identify for at least 9 months.

6.1.5. Types of Video Signal Used

The source of the video images used was a plumbicon camera (Philips LDH0150) with a slide holder attachment. The eight photographs used were in the form of positive transparencies.

Four different types of video image were used. Three were various forms of 'cartoon' images and the fourth, a normal monochrome video image but masked with noise. An example of each of the four different images is shown in Figure 6.1. The first type of cartoon image was produced by processing the original video signal by means of a simple edge detector (as shown in Figure 3.10). The second type of cartoon image was produced by using a circuit that could detect horizontal and near horizontal edges (as shown in Figure 3.11). The third type of cartoon image was produced by adding the first two together and also using a noise stripping circuit. This was the type of cartoon used in all the videotelephone experiments and a complete description of the method of generation of such an image may be found in Section 3.7.

The fourth type of image was a normal monochrome image (see Section 3.6). However, if this type of image had been used for this experiment (concerning the identification of facial images) at the ^{normal} signal to noise ratio then the subjects would have obtained 100% correct identifications. Because of this the signal was masked with noise at various signal to noise ratios.

The noise source was a General Radio, 1390B, noise generator. The noise from the generator was specified to be of Gaussian amplitude distribution and to have a flat frequency spectrum to an upper bandwidth of 5 MHz. A simple check was carried out on the specifications. The amplitude distribution was checked by using a Quan-Tech Level Analyser, which unfortunately only has a useful bandwidth of 1 MHz.

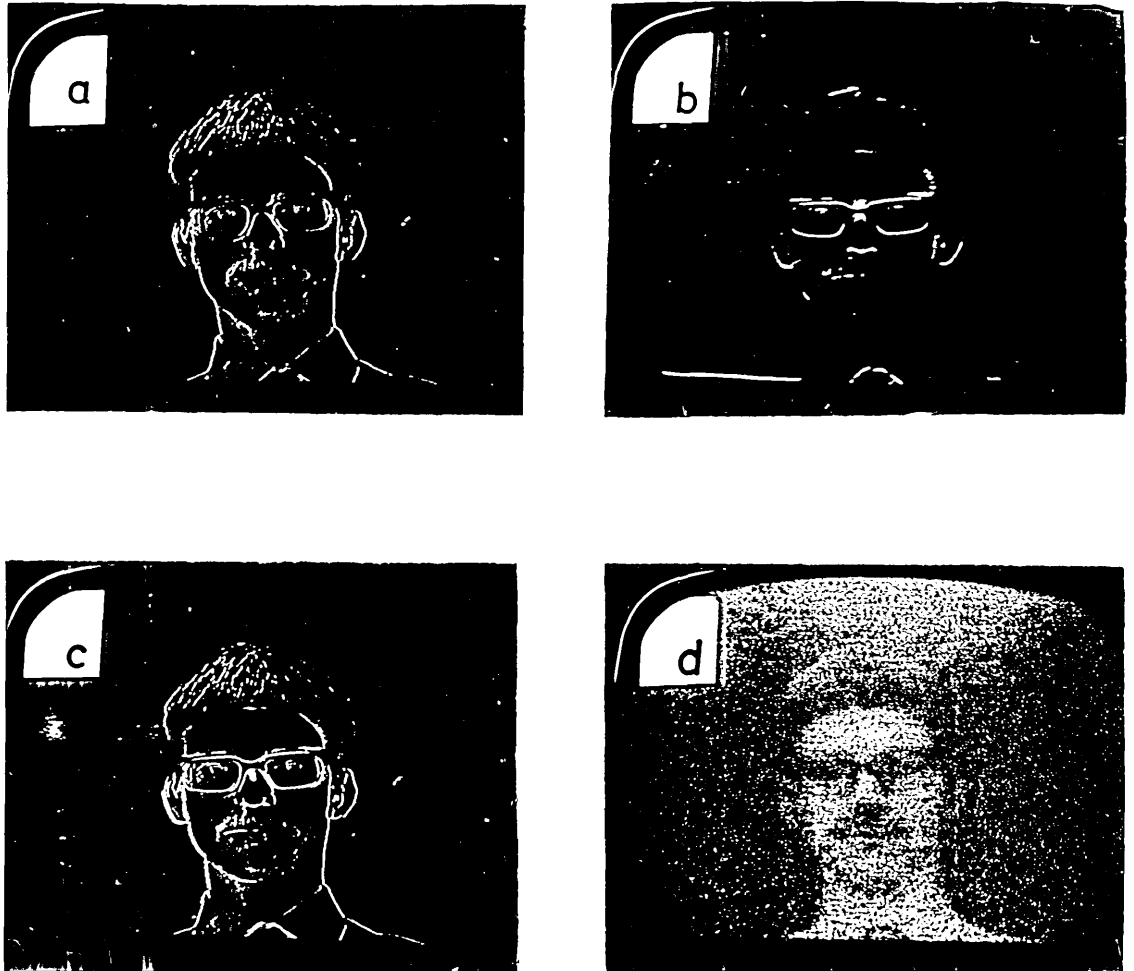


Figure 6.1 Examples of the four types of video image used in the facial recognition experiments.

The percentage of time in a fixed period that the noise signal amplitude exceeded a given level as plotted against that level is shown in Figure 6.2. The abscissa is scaled in units corresponding to a cumulative normal probability distribution. As the graph in Figure 6.2 is not a straight line, the amplitude distribution of the noise was not truly Gaussian. As no suitable commercial frequency analyser was readily available, the frequency spectrum of the noise was tested by the method employed by PEARSON (1965). This method used the property that the noise power, and hence the square of the R.M.S. voltage, is proportional to the noise bandwidth. The noise bandwidth was varied by using a number of Butterworth π section, LC lowpass filters (GEFTE, 1962) and the r.m.s. voltage was measured with a Ballantyne 320A valve volt meter. The results are plotted in Figure 6.3. The graph shows a slight preponderance of energy in the upper frequencies.

For the experiment, the noise was kept constant at 0.16 volts r.m.s. and the video signal attenuated by a wideband stepped attenuator. The signal to noise ratio was defined as the ratio of the peak signal voltage to the r.m.s. voltage of the noise.

The video monitors used were the same as were used throughout all the experiments.

6.1.6 Test Procedure

The subjects were given a list of the 8 possible faces that they would have to identify. They were asked to state the name of the person as soon as they could identify him on the monitor screen.

In the first session, only the three types of cartoon image were used. The total of 24 different images were presented to the subjects in a random order. As soon as the subject had identified the person whose

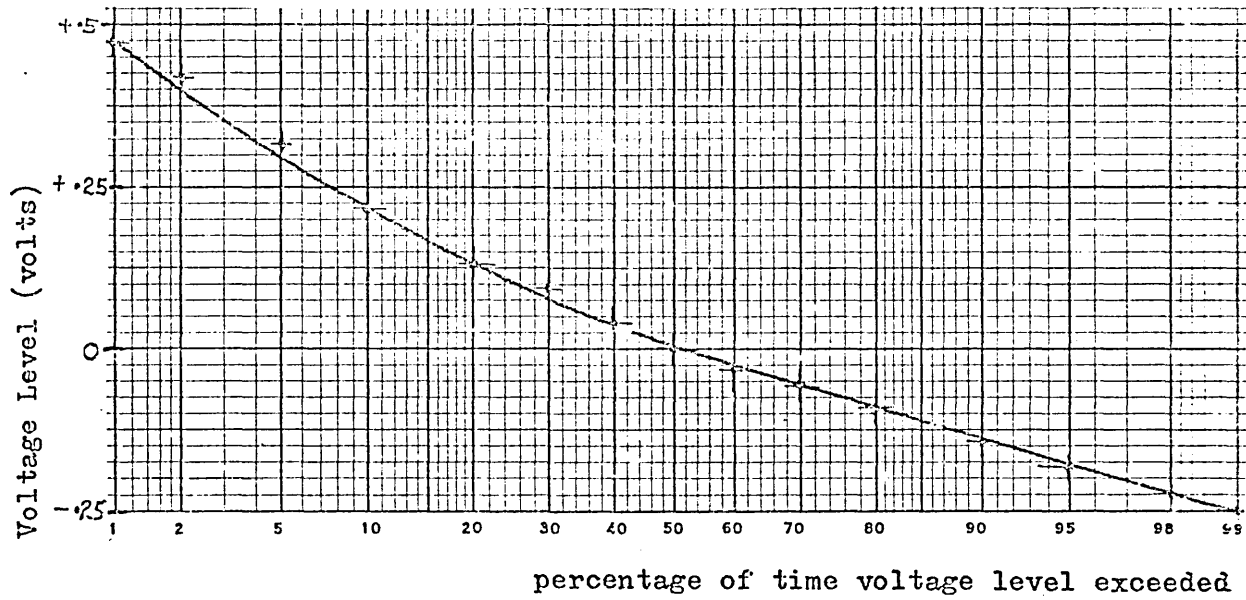


Figure 6.2 Noise Amplitude Distribution .

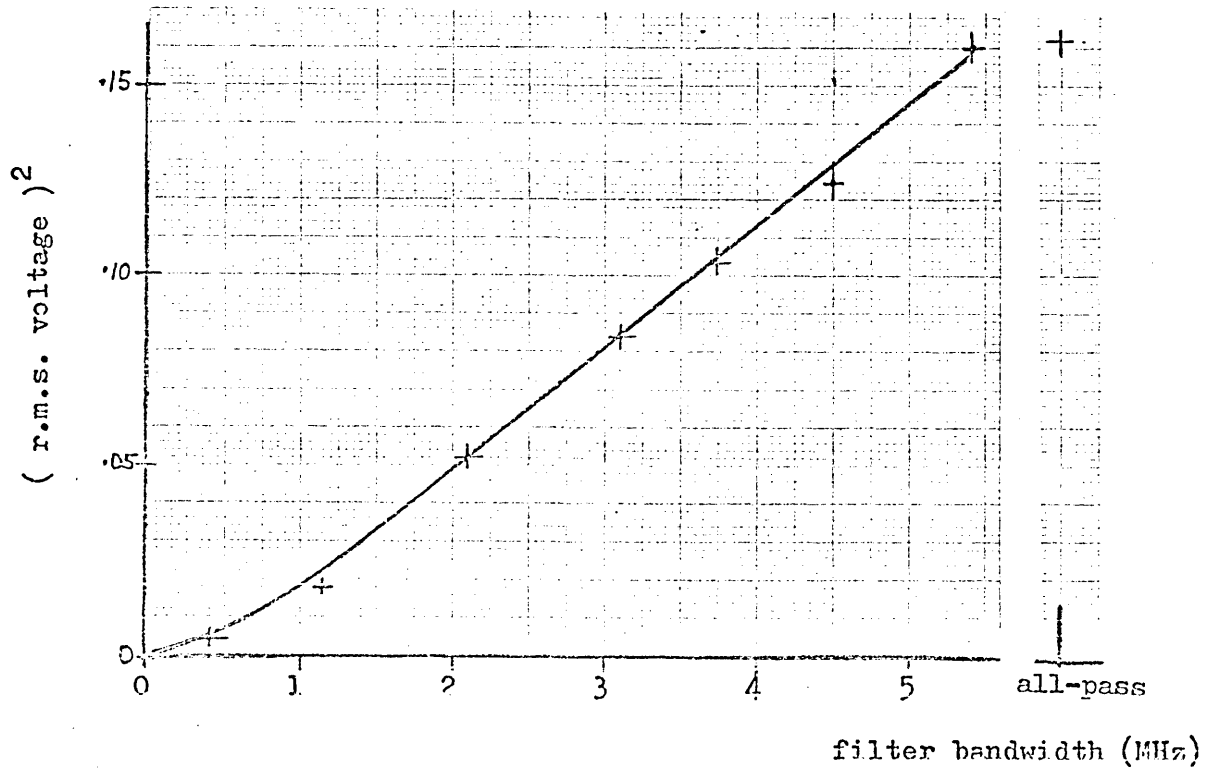


Figure 6.3 Variation of Noise Power with Bandwidth.

facial image was shown; then the next image was presented. If, however, after 3 seconds the correct identification had still not been made, then a new image was presented. The experimenter made no response to any correct or incorrect identifications made by the subjects.

The second session occurred 2 weeks later, by which time it was hoped that the subjects would have forgotten any peculiar non-facial aspects of the individual slides. In this session, the subjects were presented with a monochrome image masked with noise. The signal to noise ratio was increased in 1 or 2 db steps every 3 seconds, until the subject identified the facial image. Each of the 8 different facial images was presented 3 times. The 24 presentations were shown in a random order. This meant that some facial images were presented twice before others had been presented at all. This procedure was adopted so that the subjects could not work out, by a process of elimination, which image was likely to be shown next. Only the results of the first presentation of each facial image were recorded.

6.1.7. Results

Figure 6.4 shows the relative frequency of correctly identifying the 8 different facial images, by using the monochrome image with various signal to noise ratios. In plotting the graph shown, it is assumed that, if a subject can identify a facial image at one particular signal to noise ratio, then he will be able to identify it at all greater signal to noise ratios.

The results of using the three different types of cartoon image are shown in Table 6.1. Each entry in the bottom row of the table (i.e. equivalent monochrome image signal to noise ratio) was obtained by first finding for each individual facial image the signal to noise ratio of the monochrome image which gave the same relative frequency of correct

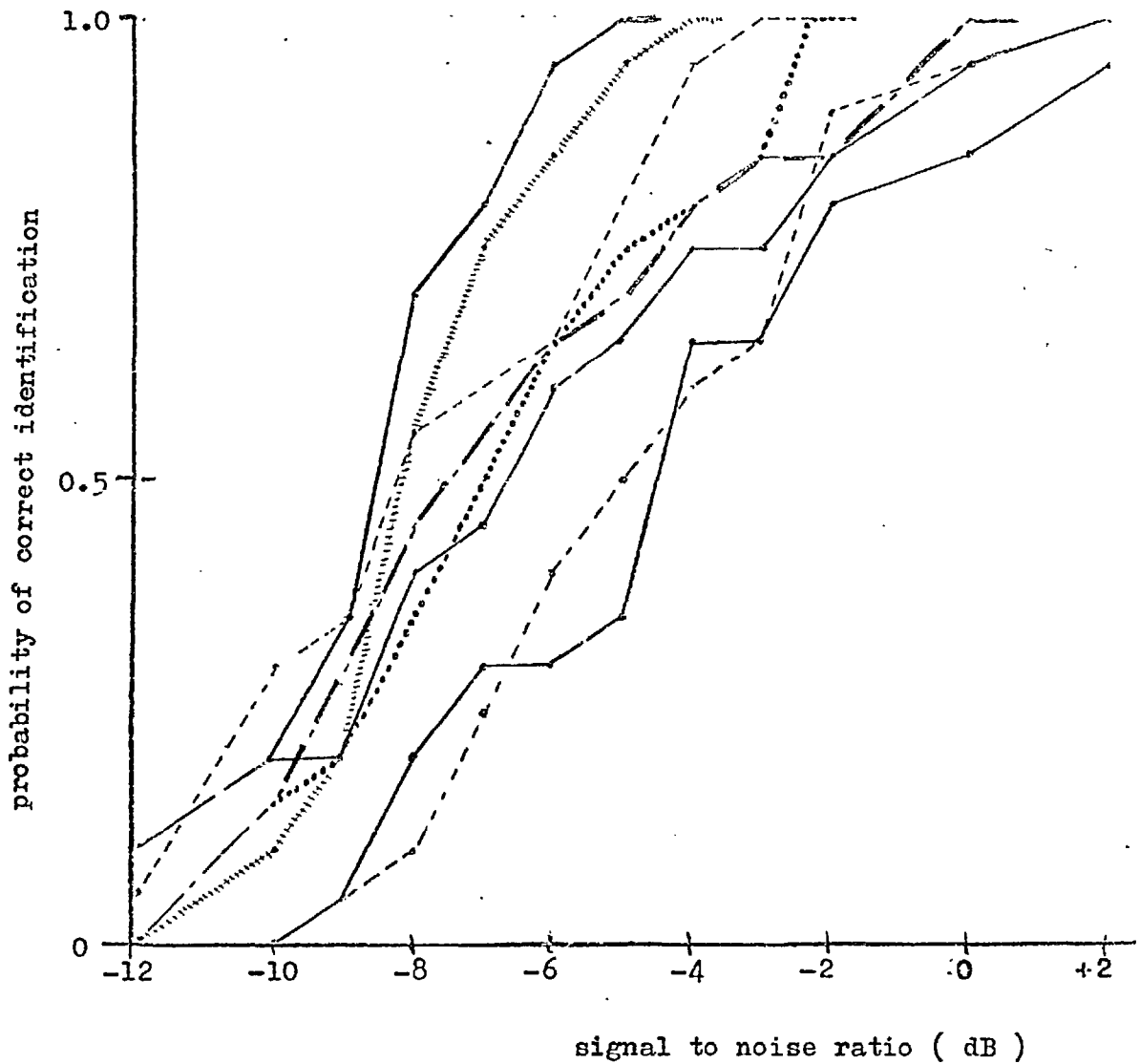


Figure 6.4 The results of the experiment to measure the probability of correctly identifying a facial image masked with noise, at different signal to noise ratios. (Monochrome video image used).

Table 6.1 The results of the experiment on the recognition of facial images using different forms of video cartoons. (Eight different facial images were used; twenty observations of each face were used for each type of cartoon.)

	type of cartoon image		
	simple edge detector	horizontal edge detector	complete cartoon
Probability of correct identification (P)	0.58	0.85	0.98
standard deviation of P	0.21	0.15	0.05
Equivalent monochrome image signal to noise ratio. (dB)	-6.6	-3.4	-1.2

identification; and then taking the average over the 8 different facial images. In some instances, the relative frequency of correct identification using one of the cartoon images was 1.0. In such cases, the lowest signal to noise ratio of the monochrome image that gave completely correct identification was used. When the average was taken in these instances, all that may be stated is that the equivalent signal to noise ratio of a monochrome image was at least equal to this average and possibly greater.

6.1.8. Discussion of Results

The high probability of correctly identifying a facial image when a cartoon type of image is used demonstrates how little information is required for such identification. In particular, when the complete cartoon generator was used, out of the total 160 visual presentations (8 facial images x 20 subjects) 157 were correctly identified. The experiment also showed that for identification purposes the complete cartoon image was as good, and possibly better, than a normal monochrome video image with a signal to noise ratio of -1.2. dB.

For 7 of the 8 different facial images, the probability of correct identification when using the horizontal and near horizontal edge detector (P_H), was higher than the probability obtained when using a simple edge detector (P_S). The average of the ratio (P_H/P_S) for all 8 facial images was 2.26 (standard deviation = 2.55). This significant difference ($p = .024$ by the Wilcoxon Signed Rank test) demonstrates that many of the important features of the face are horizontal, or near-horizontal. This result may have been expected as the mouth, eyebrows and eyes fall into this category. The result also justifies using the more complex cartoon generator for the videotelephone experiments.

As some faces might be expected to be much more recognizable than others, there may exist a correlation between the 50% threshold level for recognizing the monochrome facial images masked by noise, and the probability of recognizing the same faces when using some form of cartoon image. No such correlation existed for the faces used in the present experiment, but, as tests of significance for correlation coefficients are very insensitive with small sample sizes, it would be extremely unlikely that any significant correlation could be shown, using only eight different faces.

A criticism that may be made of the experiment is that eight different facial images is far too small a sample for them to be a typical sample and hence the results have little generality. It is because of this possibility that the experiment may only be thought of as exploratory. There are very many problems that arise when attempting a similar experiment with a larger sample of facial images, the main one being that of defining the relationship between the persons whose faces are used, and the subjects who have to perform the recognition task.

Although the experiment showed that a cartoon image is sufficient for identifying a person, what is not known is how good such an image is for identifying facial expressions. The main problem of performing an experiment to answer this question is that of accurately defining a facial expression (see Section 2.4). If such an experiment were performed using many types of image besides cartoon images, valuable data might be obtained which would help in the design of videotelephone systems.

6.2 The apparent proximity of a video facial image

6.2.1. The perception of distance

The proximity or apparent proximity between two persons is very important during conversations. An expression that is sometimes used to proclaim the merits of new types of communication equipment is that 'it brings people closer together'. The proximity and relative orientation chosen by the participants of a face-to-face conversation will depend on the meeting place and the type of conversation (see Section 2.6). However, with a videotelephone conversation, the participants have little control over their apparent proximity and relative orientation. When two persons are engaged in a face-to-face conversation, the distance between them will influence some aspects of their behaviour. In a videotelephone conversation, the apparent proximity may similarly influence certain aspects of the participants behaviour. For example, ARGYLE and DEAN (1965) showed that, in face-to-face conversations, the amount of eye contact increased as the separation between the participants increased. Therefore, we might expect that in a videotelephone conversation, using terminal equipment for which the size of the monitor screen combined with the viewing distance make the other participant seem far away, more eye contact would occur than with terminal equipment that makes the other participant appear nearer. This, of course, assumes that certain aspects of human interaction behaviour will be similar to both face-to-face and videotelephone conversations.

So that the data obtained from the videotelephone experiments (particularly on eye contact) may be correctly interpreted, it is necessary to determine how far away from each other the participants thought they were. Later in this section, an experiment will be described which was carried out to determine this apparent proximity, but first, there is a brief review of the possible cues used for the perception of distance.

There are different visual cues that may be used for the perception of proximity in face-to-face conversations. The monocular cues are as follows:

- 1) Relative Size; the perception of distance of an object will depend on the size of the retinal image produced by the object, and how this compares with past experience of that object.
- 2) Interposition; an object that overlaps another is assumed to be nearer.
- 3) Linear Perspective; parallel lines or edges are seen to converge at a distance.
- 4) Monocular Movement Parallax; when an observer moves his head, near objects move relative to far objects.
- 5) Accommodation; when the eye is focussed on a near object, objects at a greater distance are blurred.
- 6) Aerial Perspective; when the surface details of an object cannot be clearly seen, then it is assumed to be far off.
- 7) Light and Shade; the perception of distance is often affected by lighting conditions, but this effect, along with aerial perspective, has not as yet been specified in quantitative terms.

There are cues which will depend on the observer having two eyes and these binocular cues are:

- 1) Convergence; the lines of fixation from the two eyes have to converge on the object viewed. This cue is derived from the muscles which control the movement of the eyes, but it is not an important cue if the object is more than a few yards away.
- 2) Stereoscopic Vision; this cue arises from there being a different retinal image in each eye.

The list of cues above was taken from an extensive review of visual space perception by GRAHAM (1965). A great deal of research has been

done on each of these cues and also on combination of cues, particularly where they present conflicting evidence.

In a normal face-to-face conversation, a combination of some or all of the cues listed may be used by a person for estimating the proximity of the other. During a videotelephone conversation, many of these cues will not be available or may be distorted. As a small screen, and hence a small facial image, is likely to be used, the retinal image produced by such a facial image will be the same size as a retinal image produced by a 'live' face at a greater distance than the monitor screen. However, many of the other visual cues such as binocular convergence, stereoscopic vision, accommodation and interposition will indicate that the facial image is at the same distance as the monitor screen. Which of these is the more important will depend on whether the observer is just looking at a face on the screen or whether he is involved or interacting with the image.

Another factor that may influence videotelephone users' perception of proximity of the other user, is the knowledge that they may be physically separated by very many miles.

There are, of course, audio cues as well. Hence, if the output volume of the loudspeaking telephone is low, the subject will seem to be far off. This may be an important factor, but it is difficult to quantify. For this reason, the experiment described in the following section was limited to the perception of proximity using visual cues only.

6.2.2. Experimental Apparatus and Procedure

A positive transparency was produced of a person (A.W.) who was familiar to all the subjects who participated in the experiment. The format of the transparency was the same as that encountered in the videotelephone experiment (see Figure 3.7 for example of format). This facial image was displayed on a monitor (the same as that used for all other experiments) by means of a plumbicon camera plus a slide holder. The size of the facial image on the monitor screen was 2.6 times smaller than A.W.'s actual face. During the experiment, the subjects were seated at a distance of 38 inches from the monitor screen. This meant that the size of the subject's retinal image produced by A.W.'s facial image on the monitor screen, would be the same as that produced by A.W. standing 100 inches away from the observer.

The video monitor was placed at the end of a tunnel, 11 inches deep, 9 inches high and 8 inches wide. This was done because in the videotelephone experiment, the video monitor at each terminal was placed in a similar tunnel (see section 4.1.7).

A large model of the front of the monitor and tunnel was constructed, and the dimensions were arranged to be 2.6 times the actual size. This meant that when it was viewed from a distance of 100 inches away, it would form the same size retinal image as the actual monitor and tunnel, viewed from a distance of 38 inches. The whole model was arranged so that A.W. could carry it in such a fashion that, when viewed from the front, his facial image would have the same format as did his recorded image on the video monitor screen. Black curtaining hung from the bottom of the model, in order to obscure A.W.'s feet.

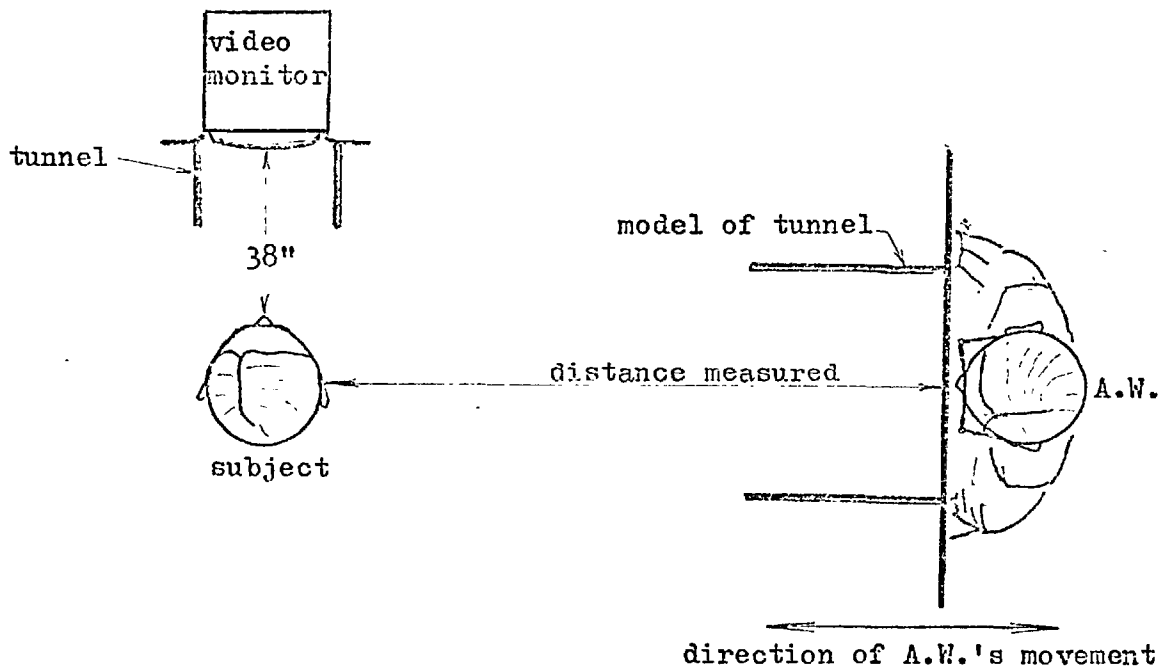


Figure 6.5 Schematic diagram showing the relative positions of the subject, video monitor and A.W., during the experiment to determine the apparent proximity of a facial image on the video monitor screen, (not to scale)

Table 6.2 The apparent proximity of a facial image on the monitor screen. The first and second column headings refer to the direction of movement of A.W. during the experiment.

	towards (inches)	away (inches)	$\frac{\text{away}}{\text{towards}}$
average distance	220	198	0.909
95% confidence limits	188-265	175-228	0.994-0.823

For the experiment, one subject at a time was seated facing the video monitor and at a distance of 38 inches away. Each subject was given the following instructions: "A.W. is going to walk towards you very slowly. When he seems to be at the same distance away as he appears on the video monitor, say NOW". A.W. then started walking slowly towards the subject, from an initial distance of 50 feet. The line of approach was 90° away from the subject's line of regard for the video monitor (see Figure 6.5); this meant that the subject had to turn his head in order to get a clear view of A.W. When the subject uttered "NOW", a note was taken of the distance between the subject and A.W., but A.W. made no change in the rate of advancement until he reached the subject. At this point, the subject was given a similar instruction as before only substituting "away from" for "towards". The procedure was then repeated with A.W. walking slowly backwards.

6.2.3 Results

Twenty-one subjects were used. One subject made use of his thumb held out at arms length, in a manner similar to a landscape artist, in order to judge the relative sizes of the monitor image and that of A.W. Because of this, his result was excluded from the analysis below; however, it is interesting to note that this subject gauged the apparent proximity of the facial image on the monitor screen to be much nearer than any other subject. (His estimate of 103 inches was very similar to one that would occur if only the relative sizes of retinal image were used to judge proximity.)

The size of the retinal image formed when A.W. was at different distances from the subject was proportional to the tangent of the distance and hence approximately proportional to the reciprocal of the distance. For this reason, when calculating the average distance etc., the reciprocal of each subject's estimation of the distance was used. By doing this, a more conservative figure was obtained for the average equivalent distance.

The results are summarized in Table 6.2. For each subject, the ratio of his two estimates was calculated, i.e. estimated distance when being approached divided by the estimated distance when being receded from. The average and confidence levels of this variation is shown in the third column of Table 6.2.

6.2.4 Discussion of Results

The results clearly show that, when a facial image is viewed on a small, monochrome video monitor screen, it must produce a larger retinal image than that produced by a 'live' face, to appear at the same proximity as the 'live' face. One possible reason for this is that as the monitor screen can be seen to be small compared to its surroundings then images produced on it tend to look smaller than the natural sizes of the objects depicted (see VERNON, 1953). Other factors, which possibly affect the perception of apparent proximity are viewing distance, screen size, as well as the very limited nature of the raster scanned monochrome images.

Consider the possible implications of these results. First, the apparent proximity of approximately 200 inches is much greater than the 'comfortable distance for conversation', which is about 60 to 70 inches (see Section 2.6). Users, therefore, may feel rather remote from each other. Second, the amount of 'eye contact' and 'looking', measured during a conversation over such a link, is likely to be higher than that measured in face-to-face conversations.

Examination of the third column of Table 6.2 shows that there was a significant difference between the results obtained when A.W. was approaching the subjects and when A.W. was backing away from them. Because, in this experiment, A.W. always moved towards the subject before backing away, it is not possible to state whether the significant difference was due to the direction of movement or the order in which these movements occurred, i.e. there may have been carry over effects. Any possible carry over effects could have been eliminated, if the order in which the direction of movement was carried out was alternated with each new subject. With such a procedure, some difficulties will arise, particularly at the beginning of each session when A.W. starts close to the subject. As the author's experiment was only to obtain an estimate of the apparent proximity, the simple procedure was adequate for this.

It would be interesting to repeat the experiment but using a range of monitor sizes and viewing distances, in order to find which combinations gave an apparent proximity approximating to the comfortable conversation distance. A criticism of such experiments, and also of the author's experiment, is that only a stationary image without sound was used, and that subjects only observed the image and did not interact with it. A more realistic estimate of proximity may be obtained by performing measurements on the forms of non-verbal behaviour that are known to be dependent on proximity, i.e. eye contact.

With a videotelephone link where the distance between the user and the camera is less than the apparent distance of the facial image on the screen, any forward or backward movement of the user will produce a distorted impression of movement to the other person. For example, with the laboratory terminal equipment, if the subjects leaned forward, their faces would typically be 31 inches away from the camera, whereas, if they leaned backwards in their chairs, the distance to the camera became 52 inches. This movement of 21 inches produced a change in the size of the facial image on the monitor screen by a factor of 1.7. As the facial image typically appeared about 200 inches away, the movement of 21 inches by the subject produced an impression of a much greater movement. The only way that this problem may be overcome is to place the camera at the same distance from the user as the apparent distance to the facial image. Unfortunately, because of the lack of space, this solution was not possible in the laboratory experiments. However, this is a factor that could be considered by designers of commercial equipment.

Chapter 7

The Looking and Speech Behaviour of a Videotelephone User.

This chapter describes three experiments performed to investigate the looking and speech behaviour of a video telephone user. The first, a preliminary experiment, measured the user's total duration of looking and the loudness of his speech when noise was added to the audio channel. The second, a main experiment, investigated the user's looking behaviour in more detail, and compared the measured duration of eye-contact with the expected duration of chance eye-contact. The third experiment measured the looking behaviour of conversants with different status.

7.1 Introduction.

This chapter describes three different experiments that were carried out to investigate some aspects of a person's behaviour when engaged in a conversation over a videotelephone link. By studying this behaviour, it was hoped that a greater understanding of the use of the video image in communication would be obtained.

In selecting what forms of behaviour to study, it was assumed that the users of a videotelephone would attempt to use many of those non-verbal cues that occur in normal face-to-face interactions. (Chapter 2 reviews the literature that has been published on these cues). The user may display some forms of non-verbal behaviour that are not visible to the other user because of the limited video image format used: e.g. the user may make some gestures with his hands which are outside the field of view of the camera. With prolonged use of the videotelephone, users may also develop non-verbal cues which are peculiar to videotelephone conversations.

Although facial expressions will be used by the participant of a videotelephone conversation, they have been ignored by the author, as no simple method has yet been devised (or is likely to be devised) to classify them. (See section 2.4.)

The terminal equipment and seating arrangements used for the author's experiments were organized so as to reduce to a minimum any movement of the subjects. This was done so that the subjects would stay in the field of view of the camera, but it also meant that cues of posture and orientation (see section 2.6) could be ignored.

The forms of non-verbal behaviour that received most attention in these experiments were those of the direction of gaze and its interaction with a person's speech behaviour. In a face-to-face conversation when a person is coming to the end of his utterance, he will usually look at the other to see how his utterance has been received. He will, therefore, obtain 'feedback' information about the other's reactions and will modify his future behaviour accordingly. Also, during the utterance the speaker will have short glances at the listener to see if he still has his attention as well as to observe his reactions. By looking at the other person at the end of his own utterance, a person will find out whether the other person has something to say or whether he, himself, may carry on talking. Conversely, to the person who is listening, the other person's looking behaviour indicates whether or not he has finished speaking. Thus, looking behaviour is used to help synchronise speech (KENDON, 1967). Looking behaviour has an indicative as well as a communicative value. Section 2.5 reviews the literature concerned with looking behaviour.

Because of its importance, the cue of gaze direction should be readily available to users of videotelephone and, wherever possible, videotelephone terminals should be designed to minimize any distortion of the perception of this cue. (See section 4.1.)

Even were a person's direction of gaze not one of the more important non-verbal cues used during a dyadic interaction (see section 2.5), how much a user looks at the video image, or whether he even bothers to look at all in some circumstances, should be of great interest to designers of such systems.

Videotelephone users may look at the video monitor screen and yet not interact directly with it as they would if it were

a person. For example, they might have prolonged gazes at the screen, independently of the video-image's gaze behaviour. Such long gazes, in a normal face-to-face interaction, would be embarrassing and may even have sexual overtones. A method that may be used to help decide whether or not a person does interact visually with the video image is described in the following section 7.1.1. In many videotelephone conversations, the video image will have some of the attributes of a person and also some of an inanimate object. For example, it is an inanimate object to engineers who are discussing its relative technical merits, or to the two users who agree to switch it off. It takes some of the attributes of a person if, during a conversation, one user is directly influenced by the direction of gaze of the video-image.

Even if a person does not directly interact with the video-image, he will be influenced by it. He will be able to infer from the image whether or not he is being observed by the other person. If a person knows that he is being observed, it will greatly influence his behaviour and state of mind. SARTRE (1943) describes the reactions of a person who discovers that he is being observed. Sartre also asserts that looking and being looked at are two different states which cannot occur at the same time, within the same person. ARGYLE AND WILLIAMS (1969) investigated what influence experimental variables had on a subject's feeling of being observed while carrying on a conversation. They found that a person being interviewed feels more observed than the interviewer. They also showed that when two subjects were conversing, the relative lighting conditions and the looking behaviour of the other subject has no influence on the feeling of being observed. This last result is not unexpected as it is the knowledge of being observed that is important and not the ability to see the observer.

ARGYLE, LALLJEE AND COOK (1968) performed a series of experiments where the subjects' visibility of each other, while engaged in a conversation, was varied. They found that the synchronising of speech was worse when vision was reduced e.g. there were more interruptions and pauses when the vision was reduced. Also, when more visual information was available, subjects felt more comfortable and better able to perceive the other subject's reactions. However, in the author's opinion, some of these effects may be attributed to the methods used for reducing the relative visibility of the subjects i.e. wearing of masks or dark glasses, and not the reduction of visibility, itself. They also found that subjects' feeling of comfort was greatly influenced by asymmetry in their relative visibility of each other. Because of this, if one person engaged in a conversation over a videotelephone link switches off the camera in his terminal equipment while the person with whom he is conversing is not able to switch off his camera, then the person being observed will feel less comfortable. (A situation similar to that of the telescreen in George Orwell's 1984.)

In all the experiments described later in this chapter, no asymmetry was introduced into the two video channels between the subjects conversing over the videotelephone.

Several authors have shown that the sex of the persons engaged in a dyadic interaction will influence the looking behaviour (see section 2.5). In all the experiments performed by the author, only male subjects were used. However, it is very unlikely that this would influence the various conclusions drawn from the experiments.

7.1.1 Interaction with a Video Image.

One of the main questions that the experiments described in this chapter hope to answer, is whether a person using a video-telephone interacts with the video image that appears on the monitor screen. It is not obvious from a cursory examination whether subjects do, or do not, interact with the image. For example, they may look intermittently at the image but their behaviour, and particularly their looking behaviour, might not be directly influenced by the image.

The question whether the subjects interact with the video image, may be answered by comparing the actual amount of eye-contact that occurs with the amount that would be expected if the two participants were just 'looking robots', i.e. if the subjects did not interact with the video image and the eye-contact just arises by chance. If the actual amount of eye-contact is significantly above or below the chance amount, then it can be assumed that subjects did interact with the video image. If, however, the actual amount was not significantly different from the chance amount, the question is still unanswered.

STRONGMAN AND CHAMPNESS (1968) estimated that the duration of 'chance eye-contact' between two simple 'looking robots', A and B, by using the following formula:

$$\text{chance eye-contact} = \frac{L_A \cdot L_B}{T} \quad \dots(1)$$

where L_A is the total amount of time that A spends looking at B
 L_B is the total amount of time that B spends looking at A
 T is the total time considered

To show that this very simple formula may lead to erroneous interpretation of results, consider the following example. Two subjects are engaged in a conversation and each subject looks at the other one as soon as he hears him start talking and stops looking as soon as he hears him stop talking. Each subject's looking behaviour is therefore determined by the other person's speech. Further, assume that the subjects never both speak at the same time, hence, during this conversation there will be no eye-contact. Then, by using the argument in the previous paragraph and the simple formula given, it would be shown that the chance eye-contact and the actual eye-contact were different. Hence, we would make the conclusion that the subjects did interact. This, of course, is a correct conclusion as the the subjects must have interacted in order for them not to have spoken at the same time, but the interaction used only audio cues and not visual cues, so it would have been incorrect using the simple formula given to conclude that there had been any visual interaction.

As users of a videotelephone link are almost certainly going to interact verbally, any useful estimate of the chance eye-contact must take this verbal interaction into account. CHAMPNESS (1971) has used a formula for estimating the chance eye-contact, taking into consideration any verbal interaction. Unfortunately, because of limitations in his measuring apparatus, he had to assume that one or other of the subjects in the conversation was always talking i.e. there were no silent periods. This assumption would have been inappropriate for the author's experiments, particularly during the 'trucking game conversations', where the silent periods were on average equal to 40% of the total time.

The following paragraphs show the derivation of the formula used by the author for estimating the duration of chance eye-contact. It uses a simple probabilistic model and, for the sake of brevity, the following symbols are used.

U_A = Total duration of A's utterances.

U_B = Total duration of B's utterances.

X_A = Total amount of time that A spends looking at E, while listening to B's utterances.

X_B = Total amount of time that B spends looking at A, while listening to A's utterances.

Y_A = Total amount of time that A spends looking at B, while A, himself, is uttering.

Y_B = Total amount of time that B spends looking at A, while B, himself, is uttering.

As before, L_A and L_B are the total amount of time that A and B spend looking at the other subject, and T is the total time of the conversation.

KENDON and COOK (1969) have shown that the amount of looking while listening (X), and the amount of looking while talking (Y), are two consistent aspects of a person's looking behaviour. Conclusions reached from using these measures are therefore more likely to be meaningful than conclusions reached using other arbitrary measures of looking behaviour.

Using the simple probabilistic approach, we may say that the total duration of chance eye-contact during all of A's utterances, will be equal to

$$\frac{Y_A \cdot X_B}{U_A} \quad \dots (2)$$

Similarly, the total duration of eye-contact during B's utterances will be equal to

$$\frac{Y_B \cdot X_A}{U_B} \quad \dots (3)$$

Now, assume that the amount of time that both subjects spend talking at the same time (joint utterances) is negligible and may therefore be ignored. This assumption was valid for the author's experiments, as the average measured amount of joint utterances was only 2 to 3% of the total time. We can now calculate the total amount of silent periods and this will be equal to $T - U_A - U_B$. The total amount of time that A spends looking at B during the silent period is $L_A - X_A - Y_A$. Similarly, for B, the total amount of time he spends looking at A is equal to $L_B - X_B - Y_B$. The total chance eye-contact during the silent periods will be equal to:

$$\frac{(L_A - X_A - Y_A) \cdot (L_B - X_B - Y_B)}{(T - U_A - U_B)} \quad \dots(4)$$

The total duration of chance eye-contact during the whole of the conversation will then be obtained by adding together the terms (2), (3) and (4) and this will give:

total chance eye-contact =

$$\frac{Y_A \cdot X_B}{U_A} + \frac{Y_B \cdot X_A}{U_B} + \frac{(L_A - X_A - Y_A) \cdot (L_B - X_B - Y_B)}{(T - U_A - U_B)} \quad \dots(5)$$

The formula, (5), given above, will give an estimate of the total duration of chance eye-contact between two subjects who interact using audio cues and not visual cues. If the actual amount of eye-contact measured was significantly different from the chance amount, it will show that the subjects used visual, as well as auditory, cues in the interaction. For example, if the actual amount was much less, then it would show that the subjects actively avoided eye-contact; this may be expected

to occur in embarrassing, or in some highly competitive, situations. If, however, the measured amount of eye-contact was greater than the chance amount, it would show that the subjects actively sought eye-contact for a variety of reasons.

When deriving the formula for chance eye-contact, the measure 'total duration of A's (or B's) utterances' was used. The word 'utterance' was used in preference to 'speech' or 'talking' because in the measurements performed in the experiments described later in this chapter, the length of an 'utterance' was taken as the time from when a person started speaking until he stopped speaking. During the period of a single utterance, the fluent speech may be interspersed with filled pauses or halting speech. Any attention signals produced by a subject when listening to the other, such as 'yes' or 'I see', were ignored. The same criteria for measuring the length of an utterance was used by COOK (1970) when investigating various aspects of a person's verbal and gaze behaviour.

7.2 Stimulation of Conversation

In any experiment where conversation is needed, there is a problem of keeping the subjects conversing naturally for a length of time. This section reviews methods used by other workers and describes the methods used in the author's experiment.

In some of the earlier work on eye-contact (EXLINE and WINTERS 1965), interviewing the subject by a confederate was used. The disadvantage of this is that the confederate's knowledge of the experiment design may influence the results. Also, during an interview, the person being interviewed may feel ill at ease and have the feeling of being observed (ARGYLE and WILLIAMS 1969).

Other methods of stimulating conversation have been developed for the testing of telephone circuits. These typically consist of the identification, by one person, of random shapes described by the other. The disadvantage of this method for the experiment described here, is that much of the time is spent looking at the random shapes and not at the other person in the interaction.

In some tests on eye-contact (for example see ARGYLE, LALLJEE and COOK 1968) the subjects were just asked to discuss a series of topics, such as books, films, home towns etc. This method is often successful for short conversations, but the conversations may become laboured or 'dry up' after some time.

MORLEY and STEPHENSON (1970) asked subjects to perform a simulated union-management negotiation. They compared the results when the negotiation was carried out face-to-face with that carried out over a telephone. Although this method produces a conversation which may be considered as very realistic, it does require that the subjects who participate in the experiment be given very explicit

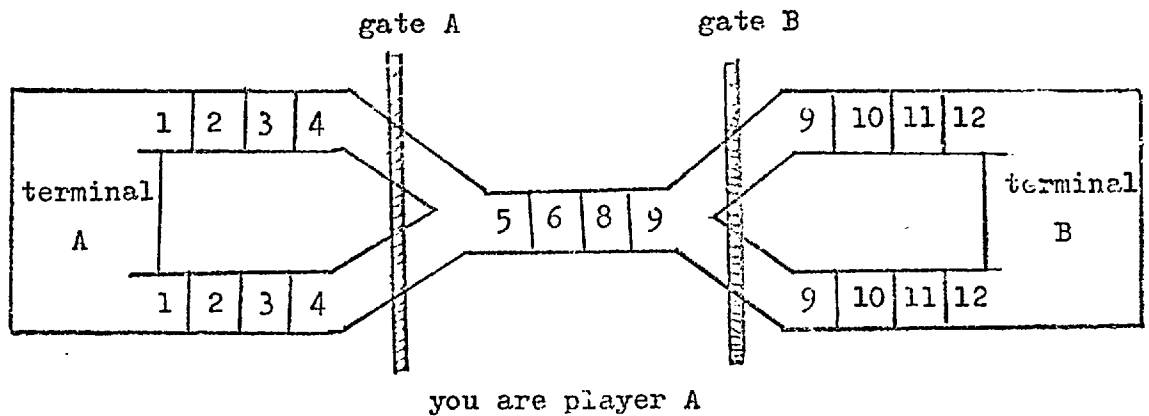
instructions as to the maximum or minimum acceptable wage settlement, as well as a great deal of background information.

The Morley and Stephenson experiment was only concerned with negotiation, but there are, of course, many different classes of conversation. When investigating the effects on conversation of various forms of distortions or limitations which occur in communication channels, it may be fruitful to use several different classes of conversation. The problem, then, is to identify the various classes of conversation possible e.g. discussion, quarrel, instruction etc., and to find methods of stimulating them in the laboratory (CHERRY, 1968).

Methods of stimulation used by the author

For the present series of experiments, a simple method was required which would make two subjects converse freely for some time and would not require many instructions. Two methods were used for the preliminary and the main experiment. The first was simply to ask the subjects to discuss their plans for the summer holidays or any films that they had seen. This first method produced a relaxed form of conversation, similar in many ways to the 'pleasantaries' often used at the beginning of a meeting between two persons.

A second method was simply to make the subjects play a simplified version of the game developed by DEUTSCH and KRAUSS (1960), which they used to study bargaining over telephone channels. Here, subjects imagine themselves to be the operators of rival transport companies. Each receives an imaginary pay-off each time his lorry reaches the opposing terminal. The layout of the roads and the control of the gates means that there must be some form of bargain reached or a deadlock occurs (see Figure 7.1 for the layout and rules). At the start of the game the rules are taken away after



Rules

1. Each player sets out from his own terminal towards the opposite one, and collects an imaginary £10 when he reaches it, then he must return to his own terminal before beginning another trip.
2. The goal is to earn more money than the other player.
3. Player A begins, then each player moves his own lorry in turn.
4. Each lorry can move only in the forward direction.
5. Each move may be 0, 1, or 2 numbered squares.
6. The gates A and B are controlled respectively by players A & B.
7. Each player can move freely through his own gate, but must ask for and receive permission of the other player before passing through his gate, in either direction.
8. The lorries cannot move toward one another on the same road, this causes collision. If collision occurs, both lorries must return to their respective terminals and begin again.

Figure 7.1 The rules and layout of the game used to stimulate conversation between subjects.

the subjects have read them for only a few minutes. This leads to further discussion between the subjects as to what the rules were, and how they should be interpreted. Subjects were also given the freedom to implement any new rules that they agreed upon. To encourage further conversation, certain aspects of the diagram showing the plan of the road and terminals were made purposely ambiguous. For example, it is not clear from the diagram whether or not squares 1 and 12 should be considered as part of the terminals.

The aim of this second method of stimulating conversation was to produce a conversation which would contain negotiation, discussion and factual enquiry. For example, if this method is successful, the subjects might be expected to discuss the rules of the game, or any new rules that they may devise themselves, and also to make factual enquiries about the position of the other person's truck, or whether or not he had permission to pass through his opponent's gate; negotiations with each other may be expected to occur when one person is required to give way to the other in order to avoid deadlock. This type of conversation may be typical of business meetings, after the initial pleasantries have been dispensed with.

In both the methods of stimulating conversation described above, the two subjects in the conversation had equal status. A third method of stimulating conversation, used in a later experiment, was designed to give one of the subjects greater status and control of the conversation than the other participant. This third method capitalized on a situation that only occurs once a year at Imperial College, namely, that it is the custom for third year undergraduates to show small groups of first year undergraduates around the postgraduate laboratories, in order to give them some idea of the kind of research carried out in the college.

On arrival at the author's laboratory, the senior student and one of the junior students were seated at each of the videotelephone terminals. The senior student only, was given the instructions for the experiment and he was given a form which he had to complete. The senior student then proceeded to describe three different shapes, which the junior student had to identify from twelve different shapes that he had before him. The shapes and labelling of the shapes were the same as those used by the British Post Office, for conversation experiments with telephone circuits (G.P.O. drawing Numbers EX28609 and EX28608). The senior student was instructed not to show the shapes that he was describing to the junior student. When they had finished the shape identification task, the senior student wrote the results on the form, and then to complete it, he had to ask the junior student for his views on the video system.

In the conversational situation, described above, the generally higher status of the third year student is augmented by the fact that he, alone, had the instructions as to the experimental procedure and by the very nature of the task. The conversation produced by this method of stimulation should be similar in many ways to a senior person interviewing a junior or subordinate. Because of his status, or role, the senior person might be expected to ask most of the questions, to give directives, and generally to have control of the conversation.

7.3 Preliminary Experiment

7.3.1 Objectives of the preliminary experiment

The preliminary experiment had the following objectives.

- 1) To see if the 'trucking game', described in section 7.2, would keep subjects conversing for up to 20 minutes.
- 2) To investigate the subjects' behaviour when different levels of noise were added to the audio channel.
- 3) To find what length each sample of conversation should be.
- 4) To see if there was any difference in the subjects' looking behaviour while playing the 'trucking game' and when engaged in relaxed conversation.
- 5) Simply to perform a trial run on the experimental procedure, using the whole apparatus.

7.3.2 Experimental Procedure

For this experiment, six different levels of noise were added to the audio channel. The noise levels were monitored by measuring the RMS voltage across the loudspeaker coils by using a B & K Electronic Voltmeter (type 2409). The six measured levels were: 0.002 (no noise added), 0.032, 0.1, 0.32, 0.56, 1.0 volts r.m.s., these being approximately equivalent to steps of 0, 25, 35, 45, 50, 55 dB. With the noise level at 0.1 volts, RMS, across the loudspeaker, the peak-speech-to-noise ratio was typically -6 dB. A spectrum of the noise is shown in Figure 4.11.

Only the monochrome image was used in the experiment. The video images transmitted between the two participants were recorded on a video-tape recorder. (See section 4.3.2 for method used.)

The two subjects in each test were placed^{at} either terminal with a copy of the rules of the trucking game. A large drawing of the 'road layout' was supplied, so that they might keep track of the position of their imaginary lorries. The rules were taken away after two minutes and the subjects were told to start playing and to continue playing for about twenty minutes, when the experimenter would tell them to stop.

The subjects were informed before the test that different audio noise levels would be used. During the game, each noise level (other than zero) was chosen. The condition of no noise occurred three times; at the beginning, at the end, and approximately the middle of each session. Each level of noise lasted for a period of three minutes, and there was a total of eight three minute periods.

When the last period of noise was finished, the subjects were asked to stop playing the game and to discuss, for about five minutes, their plans for the summer holidays. No noise was added during this period.

7.3.3 Forms of Behaviour

Speech Level

Many different methods have been established for measuring the level of speech during a conversation (RICHARDS (1969) has made a comparison of the main methods). The method used in this experiment

was to use the peak RMS reading of a VU meter (B & K type 2409). Readings were taken during the experiment of the 'peak reading' for each utterance made by the subjects. The advantage of this method over some of the more established and automated methods, is that the slamming of doors, other people's speech, and subject's non verbal utterances, e.g. laughter, tapping of the microphone etc., could be ignored. This method, therefore, measures the peaks of purposeful utterances. As the terminals were situated in laboratories which were in constant use, door slamming etc., occurred quite frequently. It was normal to take 10-15 readings per subject in each 3 minute period, and the arithmetic mean of these was used. The measurements were performed during the actual conversation.

Amount of Conversation

The audio signals from the two terminals were passed through a simple envelope detector to a pen oscillograph (A.E.I. Mk.3). It was then possible to measure the total period spent talking and the total amount of silent periods, by using a cartographer's pedometer.

Amount of Looking and Eye Contact

During a playback of the video recording, two observers each concentrated on one of the subjects engaged in the recorded conversation. When the subject, whom the observer was concentrating on, apparently looked at him, he pressed a button for the duration of that look. The observer's buttons were connected to a unit (described in section 4.4), which allowed the total amount of 'looking' and number of 'looks' to be recorded, plus the total amount of eye-contact and the number of 'eye contacts'. When a recording of one of the conversations was re-analysed for a second time, the results for looking and for eye-contact were within 5% of those obtained in the first analysis.

In the following sections, references to the amount of looking, eye-contact, and increase in speech level during the silent periods refer to the average of these measures during the three periods of silence.

7.3.4 Results

Initially, five pairs of subjects were used; one pair, however, interpreted the rules quite differently to the other four groups. This particular pair played the game in a co-operative manner, their conversation consisting only of numbers, and they hardly looked at the screen at all. Because of this, their conversation was not used in the analysis to obtain the following results. The results may be summarized as follows:

- 1) Conversation tended to become very laboured when the noise level reached the 55dB level; two of the four groups stopped talking completely.
- 2) The level of speech increased as the noise was increased. Figure 7.2 shows the average increase in speech level for different noise levels.
- 3) The time spent talking by any pair of subjects varied between 35% and 80% of the total time available. When a comparison was made of the amount of time spent talking during the different periods of noise for any particular pair of subjects, the standard deviation of the amounts was always less than 10% of the average amount. Thus, over any 3 minute period, the amount of time spent talking, by any particular pair of subjects, varied very little no matter what the noise level was, so long as it was not the 55 dB level. This did not apply when 1 minute periods were considered.

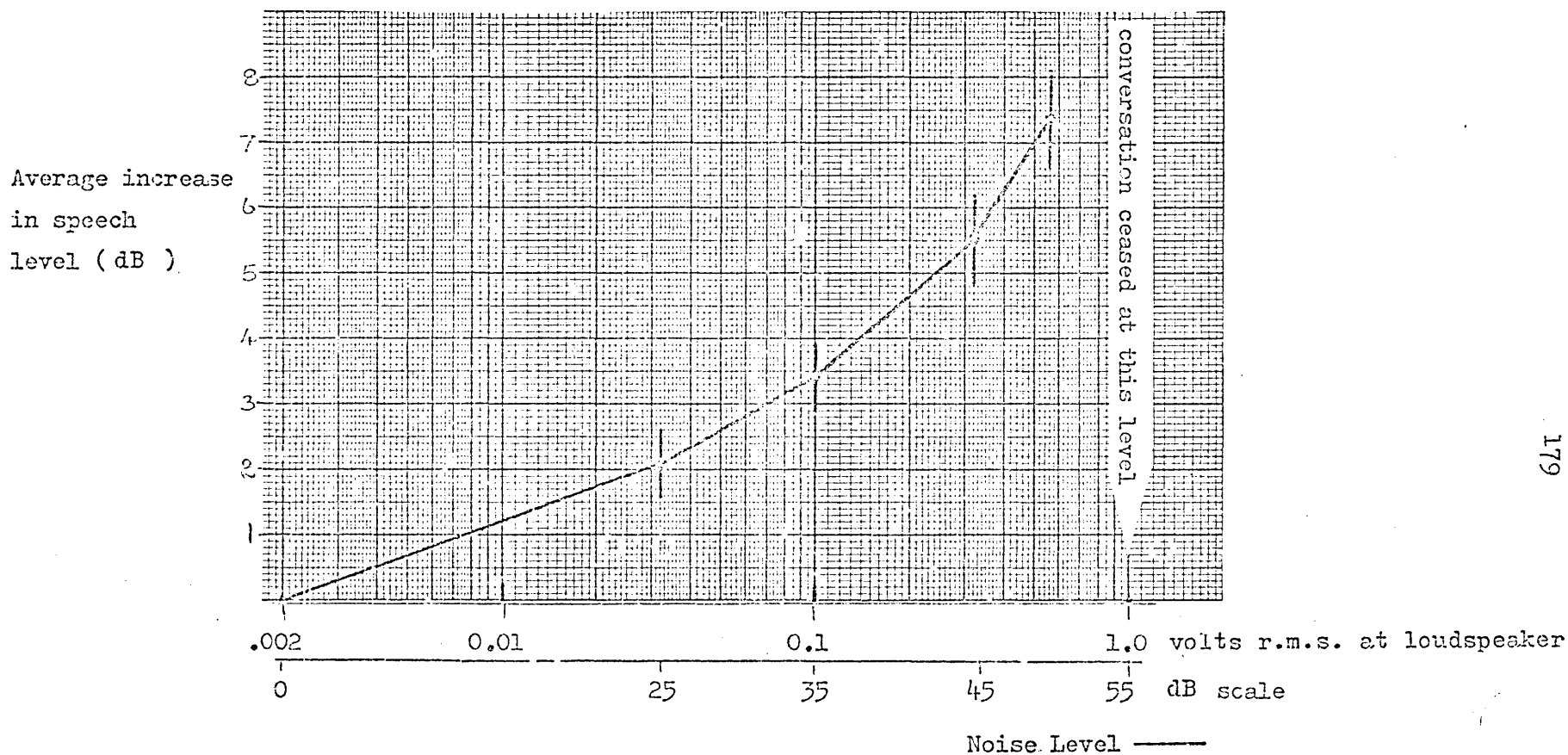


Figure 7.2 Average increase in peak speech level for different levels of noise. (10 subjects used)

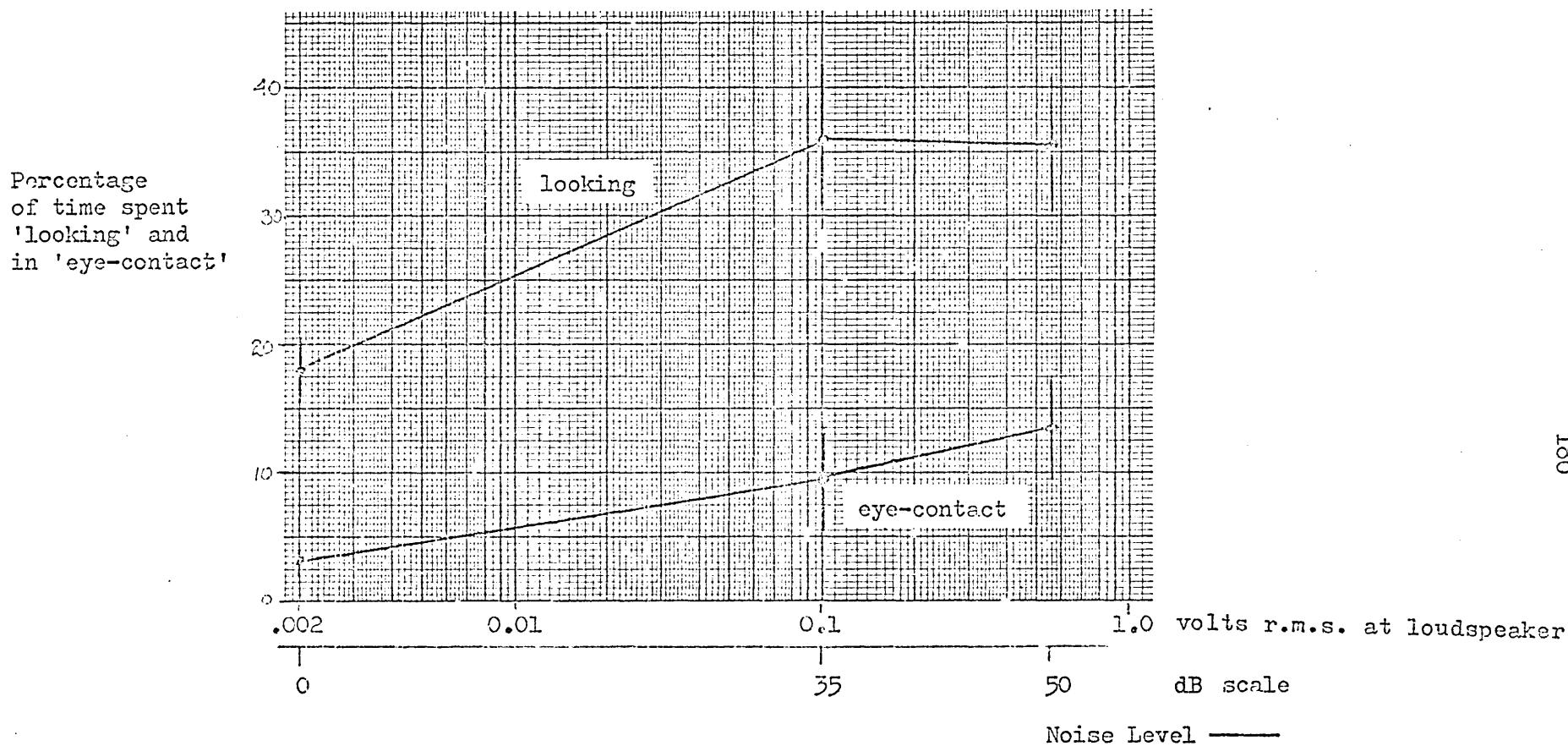


Figure 7.3 Average amount of 'looking' and 'eye-contact' for different noise levels. (8 pairs of subjects used)

Analysis of the subjects 'looking behaviour' showed a possible increase in the amount of looking when audio noise was present. Because of this, a further four pairs of subjects were used, the procedure being the same as before, but using only the 0 dB, 35 dB, and 50 dB noise levels.

The average amount of looking and eye-contact for different levels of audio noise is shown in Figure 7.3. Examination of Figure 7.3 shows that there was a negligible difference between the average amount of looking during the periods when 35 dB of noise was added, but the average amount of looking during the noisy periods was greater than during the periods when no noise was added. By simply examining Figure 7.3, it seems that there is a possible increase in the amount of eye-contact during periods when noise is added to the audio system, but a full analysis of variance would have to be carried out to be sure of this.

The average amount of time spent 'looking' during the game was 18%, while the amount during the relaxed conversation period was 46%; the difference between these two means was significant at $p < .005$ ($t=5.84$).

7.3.5 Discussion of Results

The 'trucking game' proved to be a very simple and useful method of stimulating a varied type of conversation for all, but one, of the nine pairs of subjects used. This pair played the game in a co-operative manner producing a conversation consisting mostly of numbers. A repetition of this type of conversation may be avoided in future experiments by stressing the competitive nature of the game.

That some subjects stopped talking at the 55 dB noise level is probably not a reflection of normal behaviour. If the subjects were not constrained by the experimental conditions, they may probably have stopped talking altogether, or perhaps moved to a quieter location, at much lower noise levels. The level at which subjects stop conversing will also probably depend on the nature and content of their conversation.

From Figure 7.2, it may be seen that subjects increased their speech level as the audio channel noise was increased. This increase is due partly to the subjects talking more loudly, and partly to them moving nearer to the microphones. GARDNER (1966) found that subjects, asked to repeat sentences to each other over a high quality loudspeaker link, increased the level of their speech by approximately 10 dB, for a change in noise level of 40 dB. Gardner's result though slightly larger is very similar to the result of this experiment, taking into consideration the difference in the subject's tasks, the communication link, and the methods of measuring speech level.

The little variations in the total time spent talking by any particular pair of subjects, from one 3 minute period to another, showed that 3 minutes was a suitable choice for the length of different noise periods. If a shorter period had been chosen, the amount of talking, in any particular period, would have to be taken into account when any measures such as the amount of 'looking' were needed.

The preliminary experiment reported in this section demonstrated that both audio noise level and the method of stimulating subject's conversation influenced the subject's looking behaviour. The experiment described in the following section investigates in greater detail this change in subject's looking behaviour; it also investigates how the type of video image used may influence looking behaviour. A discussion of the effect of audio noise on subject's looking behaviour appears in section 7.4.6.

7.4. Main Experiment

7.4.1 Objectives of the main experiment

This experiment had several objectives. The first was to find out whether a person using a videotelephone interacts with the visual image on the screen and to find out whether this possible interaction is affected by the type of video image, the type of conversation, or audio noise level. This is done by comparing the actual amount of eye contact and the chance amount between two users; the rationale for this is given in section 7.1.1.

The second objective was to investigate any changes in the looking behaviour which were produced by using different forms of video image, different types of conversation, or different audio noise levels.

A third objective was to see if certain aspects of the user's verbal behaviour were affected by the different experimental conditions. Only very simple (i.e. measurable) aspects of the subject's verbal behaviour were investigated i.e. the number of requests for repetitions, the number of interruptions, and the number and total length of utterances.

The objectives, stated above, are all parts of the main objective which was to obtain a better understanding of precisely what use is made of the video image by a videotelephone user.

7.4.2 The experiment

In this experiment, three different factors were varied. The first factor was the form of communication channel which the subjects had to converse over; the second factor was the method of stimulating

conversation; and the third was the level of audio noise added to the channel.

For the first factor, three different forms of communication channel were used. The first of these (M) used a monochrome video image to augment the audio channel, the second, (C), a cartoon video image to augment the audio channel and the third, (A), used only the audio channel. The main parameters of the two types of video images are given in section 3.6. In all three different forms of communication channel, 'loudspeaking-telephones' were used, and these are described in section 4.2.1.

Two different methods were used for stimulating the conversation between the subjects. These two methods were the same as those used in the preliminary experiment, i.e. to discuss their summer holidays or films that they had seen etc., and to play a simplified version of the Deutch and Krauss trucking game. The first method created a pleasant, relaxed conversation, and the second a conversation typical of a bargaining or competitive situation.

The third factor, the amount of noise added to the audio channel, was only varied while the subjects were playing the trucking game. Although it would have been interesting to vary this factor during the period of relaxed conversation, as well as during the game, to do this would have lengthened the total duration of the relaxed conversation, and thus the conversation may have become rather laboured. During the game, the audio channel noise was adjusted to be at one of two different levels; these will be referred to as the 'low' and 'high' levels. The low level of noise was the inherent thermal noise of the system plus any room noise picked up by the microphones. This noise level produced, on average, a voltage of 0.002 volt r.m.s. measured across the loudspeaker coils. The high noise level was produced by adding filtered white noise to the audio system so as to produce a voltage of 0.1 volts r.m.s. across the loudspeaker coils.

The high noise level was approximately 35 dB greater than the low noise level and during the high noise level the peak speech to noise ratio was typically -6 dB. This level of noise was chosen because the preliminary experiment indicated that it would produce a change in the subjects' looking behaviour, yet not produce any undue difficulty in maintaining the conversation. (See section 4.2.3 for the frequency spectrum of the noise and the method used for adding it to the audio channel.)

Experimental Procedure

For the experiment, the following procedure was adopted. A pair of subjects were seated at the terminals and asked to talk to each other about their summer vacation for about five minutes. They could communicate by one of the three different forms of communication channel described above. After about three minutes, they were asked to stop and each was given a copy of the rules of the trucking game. After a further one and a half minutes, the rules were taken away and they were asked to start playing and to continue until asked to stop. The subjects were informed that there would be periods when audio noise would be added and that the video image may change or even disappear. During the game, each form of communication channel (M,C, or A) was presented with each level of noise (H & L) producing six different periods in total. Each period lasted three minutes, and at the end of each period a short burst of noise was added to the audio channel.

At the end of the sixth period, the subjects were asked to stop playing the game and to talk about one of the subjects on the list in front of them. The list contained the following subjects; travel, books, films. After they had discussed one of the topics for about

three minutes, they were asked to change to another topic, and were finally stopped after a further three minutes. During these last two periods of relaxed conversation, the two forms of communication channel that had not been used at the beginning of the experiment, were used.

To finish the experiment, the subjects were asked to fill in a questionnaire. The questionnaire asked the following questions:

- 1) How easy was the conversation when audio noise was added?
- 2) How well could you communicate your ideas or feelings to the other person?
- 3) How well would you rate your own comfort?
- 4) How well could you tell how the other person was reacting to you?

These four questions had to be answered with reference to each of the three different modes of communication used (i.e. audio channel only, normal image, and cartoon image). Subjects rated their answers on a scale, 1 to 7, ranging for example, from comfortable to uncomfortable. They were then asked what was the main thing wrong with the video-phone system and what were its main advantages.

Although it is difficult for a subject to rate, for example, his comfort on such a seven point scale, it was his relative ratings for the three different modes that were of interest. Hence, a subject may have difficulty in deciding whether his comfort, for one particular mode of communication, should be rated at 4 or 5, but once this rather arbitrary decision has been made, it is easier for him to rate his relative comfort for the other modes.

The order used for the different experimental conditions is summarized in table 7.1. Six pairs of subjects were used; this number was chosen because it enabled a complete, counter balanced design to be used for the game part of the experiment, thus balancing for any first order carry-over effects (See COCHRAN & COX, 1957). The order of the three forms of communication channel used for the periods of relaxed conversation was also counter balanced. Ideally, to eliminate all carry over effects, the different periods of the game should be mixed up with the periods of relaxed conversation, but this would be very difficult to organize.

All the subjects used for the experiment had performed at least one of the simple perception or lip reading experiments. They were, therefore, familiar with the system and the different types of video image used. This should have reduced any behavior due to the novelty of using the videotelephone system for the first time. All the subjects were, however, ignorant of the particular measures used.

The whole of the six conversations were recorded onto video tape. This was later played back to four observers; two observing the two subjects' speech and the other two, the subjects' looking behaviour. Each observer pressed a button and kept it pressed while the particular behaviour that he was looking for, occurred. The observer's buttons were connected to a paper tape punch. For each three minute period of conversation, the first 15 seconds were ignored, thus giving subjects time to become aware and adapt to any changes that had occurred in the type of video image or audio noise level.

The paper tapes were then passed several times through the processing unit, described in section 4.4. For each three minute period, the total duration and number of occurrences of the following forms of behaviour were obtained: looking, eye-contact, utterance, joint utterances, looking while listening and looking while uttering. The number of 'role changes' in each three minute period was also recorded (role change refers to the change in role from speaker to listener and vice versa). All the data was punched onto computer cards. A computer

Table 7.1 Order of the different experimental conditions
used in the main experiment.

A = audio channel only

M = monochrome image plus audio channel

C = cartoon image plus audio channel

L = low audio noise

H = high audio noise

pairs	relaxed convers- ation	trucking game								relaxed convers- ation	
1	M	A.L	C.H	M.L	C.L	A.H	M.H			C	A
2	C	M.H	A.H	C.L	M.L	C.H	A.L			M	A
3	A	A.H	M.L	M.H	A.L	C.L	C.H			C	M
4	M	C.L	M.H	C.H	A.H	A.L	M.L			A	C
5	C	C.H	C.L	A.L	M.H	M.L	A.H			A	M
6	A	M.L	A.L	A.H	C.H	M.L	M.H			M	C

program was written which calculated the average and standard deviation of the percent of time, number per minute and average duration in seconds, of each form of behaviour. The program also calculated the total duration of silent periods in each three minute period, and also the amount of 'chance eye-contact' by using equation (5), given in section 7.1. Analysis of variance was also performed on the data; the form of the analysis of variance used is described in the following section.

7.4.3 Statistical Analysis

In many experiments using factorial designs, the analysis of the data is simplified if separate subjects can be used for each combination of experimental treatments. However, if the experiment contains several factors, difficulty often arises in obtaining sufficient numbers of subjects. It is often convenient to observe each subject under more than one treatment condition. These types of experiments are usually referred to as those in which there are 'repeated measures'. The different forms of the analysis of variance that should be used for experiments containing repeated measures are to be found in WINER (1962).

There are two main advantages of using a repeated measures design for an experiment, they are: 1) economy of subjects, thus enabling simpler organization of experiments i.e. instructions have to be given less often, 2) each subject serves as his own control; this is particularly useful when the responses of people to the same experimental treatment show a relatively large variability.

The model used for the analysis of variance for a simple, one factor, repeated measures experiment, where each subject receives every possible treatment of the factor, is:

$$X_{ij} = u + t_j + s_i + (s_t)_{ij} + e_{ij} \quad \dots (1)$$

where

X_{ij} = the criteria measure for treatment, j , on subject, i .

u = grand mean for the treatment population

t_j = effect of treatment, j .

s_i = constant effect of subject, i .

$(s_t)_{ij}$ = treatment subject interaction

e_{ij} = experimental error (mean equal to zero and normal distribution)

It is the inclusion of the term s_i that differentiates this model from that used when different subjects are used for each possible combination of treatments. The model is linear and additive but does not include any carry-over effects. The order of treatment on the subjects should, therefore, be counter balanced so as to reduce carry-over effects. The model, and hence the analysis, can be simply generalized for experiments with more than one factor, provided that each subject received every possible combination of treatments.

Now, consider the analysis of variance used for the 'trucking game' part of the conversation. This had two fixed factors; the first was the form of conversation channel (B) and the second, the level of audio noise (N). The different pairs of subjects may be considered as a third factor (P). Many of the measurements taken on each individual are likely to be correlated with those taken on the other subject with whom he is conversing (i.e. the total duration of the subject's utterances are likely to be negatively correlated to those of his partner). To eliminate the effects of any possible correlation and to reduce the experimental error term, a fourth two

level random factor (S) was introduced. One person from each pair of subjects was assigned to one level of the factor, and the other person was assigned to the second level of the factor. A standard library computer programme for the analysis of variance of factorial designed experiments was used to compute the Sums of Squares (SS) and Mean Squares (MS) for each factor and interaction of factors. (Note, the notation to be used in MS_A = Mean Square for factor A.)

To test the null-hypothesis that factor B had only a random effect (i.e. there was no difference produced by the differing forms of communication channel), the following statistic is calculated:

$$F = \frac{MS_B}{MS_{(BP)}} \quad \dots(2)$$

Similarly, for factor N (audio noise level), the statistic calculated for the test of the null-hypothesis is:

$$F = \frac{MS_N}{MS_{(PN)}} \quad \dots(3)$$

and for the interaction of factors N and B, (NB), the statistic is:

$$F = \frac{MS_{(NB)}}{MS_{(PNB)}} \quad \dots(4)$$

In general, for any factor Y with q treatment levels in a factorially designed experiment where there are repeated measures on n persons or n pairs of subjects etc., then the statistic used to test the null-hypothesis that factor Y has only random effect is:

$$F = \frac{MS_Y}{MS_{(PY)}} \quad \dots(5)$$

Significance levels or probabilities of Type 1 error may be calculated with the aid of tables of the cumulative F distribution, having $(n - 1)$ and $(q - 1)$ degrees of freedom. (A derivation of equation (5) may be found in WINER, 1962.)

An exactly similar analysis of variance may be carried out using the method of stimulating the conversation as a factor, instead of the audio noise level.

Where the observations of a person's behaviour are in the form of percentages or proportions (i.e. the total percent of time spent looking), these observations will approximate to a binomial distribution. Hence, the variance and mean will be related. When such measures were used, a $X' = \arcsin \sqrt{X}$ transform was carried out on the data, before performing the analysis of variance (See EDWARDS, 1968.) It is fortunate that the distribution of the F ratio in the analysis of variance is relatively little affected by inequalities in the variances which are pooled into the experimental error.

7.4.4 Results

The main results of the experiments are summarized in tables 7.2 to 7.6, and these may be found at the end of this section.

Table 7.2 gives the means of the various observations under each combination of experimental treatments. Also shown in table 7.2 (last column) is an estimate of the mean error of measuring the various forms of behaviour. This estimate was obtained by replaying a part of the video recordings of conversations by two different pairs of subjects to the observers, for a second time. For each part of the recording used, the percentage error of each measure was taken as the modulus of the difference between the observations of the first and second replays, divided by the mean of the two observations. For example, if the observed duration of looking during the first replay was L_1 , and during the second replay, L_2 , the percentage error was given by the following:

$$\text{percent error} = 200 \cdot \frac{|L_1 - L_2|}{L_1 + L_2}$$

The table gives the mean error; thus for measures performed on each individual's behaviour, it is the mean of four, separate estimates of the percentage error, whereas for joint measures (eg. eye-contact and joint utterances), it is the mean of the two, separate estimates.

Table 7.3 to 7.5 shows the analysis of variance performed on the experimental data. Only the mean square (MS) is shown as the sum of squares (SS), and the F ratio may be calculated from the given MS. The level of significance of the rejection of the null-hypothesis, that any particular factor has a random effect, is indicated by asterisk (* $\equiv p < 0.05$, ** $\equiv p < 0.01$). In these tables, the figures within brackets in the row labelled 'average' are the averages of the transformed data. When performing analysis of variance for the various forms of looking behaviour, the data for periods when there was no video image available

was excluded from the analysis.

Before performing the analysis of variance, the experimental data (or the transferred data where the original was in the form of percentages) was tested for homogeneity of variance. Only two of the many different measures were found to have heterogeneity of variance, and these were the 'mean length of each look' and the mean length of each 'looking while listening'. Inspection of the data for both these measures indicated that, for each combination of treatments, the standard deviation was approximately proportional to the mean. Because of this, a transform of the form $X' = \log(1 + X)$ was used on the data before performing the analysis of variance (BARTLETT, 1947). With this transform, the variances were stabilized as indicated by the F ratios of 3.0 for the mean length of look, and 4.2 for the mean length of looking while listening: this may be compared with F ratios of 15.1 and 44.2 respectively, before the transform was used (critical value of $F = 4.5$).

Table 7.3 shows the analysis of variance performed on the various measures of looking behaviour during the trucking game part of the conversations. Examination of this table, in conjunction with table 7.2, shows that the total duration of 'looking' and of 'eye-contact' was greater during the periods when audio noise was added to the channel. The increase in the 'duration of looking' was due to an increase in the frequency of looking and not an increase in the average length of each look. Similarly, the increase in the total duration of eye contact was due to an increase in the frequency of eye contacts and not to an increase in the average length of each eye contact. During the noisy periods, there was an increase in the total duration, frequency and mean length of the 'looking while listening', also an increase in the total duration and frequency of 'looking while talking'. The type of video image used affected only the 'number of looks per minute' and the mean length of each 'looking while talking'.

Table 7.4 gives a summary of the analysis of variance carried out on various forms of looking behaviour when the 'method of stimulating conversation' was considered as one factor, and the type of video image used as another. (Hence data for periods during the trucking game when audio noise was added to the system was excluded from the analysis.) Examination of table 7.4 shows that the method of stimulating conversation had a significant effect on every different measure of looking behaviour except for two: the mean length of each eye contact, and the mean length of each 'looking while talking'. If table 7.2 is then examined, it may be seen that every measure of looking behaviour, except the two already mentioned, was significantly greater during the periods of relaxed conversation than during the trucking game. Table 7.4 shows that the type of video image had no significant effect on any of the different measures of looking behaviour.

For periods of relaxed conversation, a Wilcoxon's signed rank test was used to test the null-hypothesis that the proportion of a subject's speaking time that he spends looking, (l_T), is equal to the proportion of his listening time that he spends looking, (l_L). For periods when the monochrome image was used, the null-hypothesis was rejected with a level of significance $p < 0.02$ ($T = 9$, $l_T < l_L$), while for periods when the cartoon image was used, the null-hypothesis was rejected with a level of significance $p < 0.05$ ($T = 13$, $l_T < l_L$). For periods when the trucking game was played, a similar null-hypothesis was not rejected.

As might be expected, the subjects did look at the monitor screen during periods when there was no video image, but the amount of looking during these periods was considerably lower than during periods when there was a video image.

Table 7.5 gives a summary of the analysis of variance performed on measures related to the subjects' speech behaviour. Examination of this table, in conjunction with table 7.2, shows that the duration of utterances was greater, but that the number of utterances was less, during the periods of relaxed conversation when compared with during the trucking game. Also, the number of role changes was less and the total duration of silent periods

(no utterance from either subject) was less during the periods of relaxed conversation. None of these measures related to the subject's speech behaviour were significantly affected by the type of video image or by whether there was a video image or not. Analysis of variance performed on these measures of the subject's speech behaviour during the trucking game, showed that neither the audio noise level nor the form of communication channel had any significant effect on them, and for this reason, the analysis of variance has not been included.

Table 7.6 shows the differences between the measured duration of eye-contact and the chance duration, predicted by using equation (5) of section 7.1.1. The table shows all the differences between measured and chance duration of eye-contact (EC - chance), as well as the rank of these differences for the three different modes of communication. The variance of the differences was heterogeneous and no simple transforms were found that would make the variance homogeneous. Therefore, it was not possible to perform analysis of variance on the results, and only a non-parametric test could be used. Wilcoxon's signed rank test was used to test the null-hypothesis that there was no difference between the measured and estimated values of duration of eye-contact when monochrome or cartoon images were used. For periods when the monochrome image was used, the null-hypothesis was rejected with a level of significance $p < 0.01$. For periods when the cartoon image was used, the null-hypothesis was not rejected, but the value of 'T' was only a little greater than the critical value for rejecting the hypothesis at the 5% significance level, (see table 7.6). The same form of test also showed that there was no significant difference between the difference (EC - chance) for periods when the monochrome image was used, and the difference (EC - chance) for periods when a cartoon image was used.

A 'sign test' showed that there was no significant difference between the measured and estimated amount of eye-contact* during periods when only the audio channel was used. (See table 7.6).

*There was, of course, no actual eye-contact during periods when there was no video image. Eye-contact is used here to refer to those instances when both subjects were looking at the blank screen (where the facial image of the other person would have been.)

When a randomization test was carried out on the difference between the measured and estimated duration of eye-contact for each pair of subjects (using periods when monochrome and cartoon images were used), only in three of the six pairs was the difference significant beyond the 5% level.

It may be concluded therefore, that on average, during periods when the monochrome image was used, the difference between the measured and estimated amounts of eye-contact was definitely significant, whereas for periods when the cartoon image was used the difference was not definitely significant, and that there were large variations from one pair of subjects to another.

The results for the number of 'requests for repetition' by the subjects during the various experimental conditions were completely unusable for comparison purposes, as only two such requests occurred during the whole experiment. These two occurred during periods when there was no video image and when audio noise was added to the channel.

An attempt was made to count the number of interruptions which occurred during the various experimental conditions, as this would be an indication as to how good was the synchronization of the subject's utterances under the various conditions. (See ARGYLE, LALLJEE & COOK, 1968). The author, however, concluded that it was impossible to distinguish objectively between interruptions due to poor synchronization between subjects and interruptions which were intentional, and that this measure was, therefore, not a reliable one.

When processing the results of the questionnaire, only the difference in the various ratings for the three different modes of communication were used. A sign test (see SIEGEL, 1956) was used to test whether the differences were significant. The results of the first four questions were as follows:-

(1) there was no significant difference between modes for the ease of conversation when audio noise was added.

(2) the 'monochrome image plus audio' was better than the 'audio only' for communicating their ideas ($p < 0.04$)

(3) the 'monochrome image plus audio' was better than the 'audio only' for their own comfort ($p < 0.05$)

(4) the 'monochrome image plus audio' was better than the 'audio only' for the ability to tell how the other person was reacting to them ($p < 0.01$)

For all four questions the ratings for the cartoon image plus audio condition lay between the ratings for the monochrome image plus audio and the audio only conditions, but was never significantly different from either. Only two subjects mentioned that there was anything wrong with this particular videotelephone system; both these subjects commented on the limited field of view. There were many different reasons given as the advantages of a videotelephone system.

Before discussing the results of this experiment, a short subsidiary experiment will be described.

7.4.5. Subsidiary Experiment.

In the two experiments described in the early part of this chapter, the methods used for stimulating the conversation between two subjects did not give either a greater status or control over the conversation. This section describes a simple experiment performed to investigate whether any asymmetry in the relative status of the subjects would influence their looking behaviour.

In this experiment, six third year undergraduates each interviewed a first year undergraduate. The method used to stimulate the conversation was of the type described in section 7.2; this method gave the senior student more control over the conversation.

Analysis of the subjects' looking behaviour showed that every junior student being interviewed looked more than the senior student. The mean total duration of the junior's and senior's 'looking' were 43% and 22% respectively, and the difference between these means was significant at the $p < 0.025$ level ($t = 2.73$). The greater amount of looking by the junior student was mainly due to the greater 'mean length of each look', (3.2 seconds as against 1.6 seconds; the difference was significant at $p < 0.025$, $t = 2.4$) and not to the frequency of 'looking'. There was no significant difference between the senior student and the junior student in either the total duration and frequency of utterances or the mean length of each utterance.

The general pattern of looking behaviour, therefore, was that the junior subject, who was being interviewed, had longer gazes than the senior interviewing subject.

	Relaxed Conversation			Trucking Game						mean error %
	A	M	C	A.L	A.H	M.L	M.H	C.L	C.H	
<u>Total duration (% of total time)</u>										
looking	7.9	43.7	42.1	2.9	4.6	13.9	23.8	10.5	23.1	4
eye contact	.2	17.4	15.0	.2	.5	3.1	8.8	3.4	7.7	10
chance eye contact		11.0	11.7			2.0	3.2	1.5	4.5	
utterance	41.8	42.3	41.5	33.3	31.2	30.9	31.3	32.1	33.4	3
joint utterance	3.4	2.2	3.4	3.3	3.2	3.3	2.4	3.4	3.5	22
looking while listening		20.9	21.1			3.9	7.2	3.2	7.7	11
looking while talking		11.8	11.8			3.2	3.4	2.8	5.9	14
silence	19.9	17.6	20.5	36.7	40.8	41.6	39.8	39.3	36.8	
<u>Number per minute</u>										
looking	5.4	9.7	9.9	2.2	3.9	5.7	8.6	3.8	6.4	8
eye contact		9.2	8.9			1.8	3.9	1.3	2.5	7
utterance	8.9	7.2	7.5	11.5	12.0	11.2	10.5	10.6	10.5	8
joint utterance	.37	3.3	4.5	4.2	5.0	4.4	3.4	4.3	4.4	11
looking while listening		6.7	7.0			3.4	5.9	2.6	4.3	12
looking while talking		7.6	7.0			3.0	4.4	2.4	3.6	11
change of roles	9.9	8.7	8.1	12.8	15.7	13.1	13.6	13.5	13.1	
<u>Mean length (secs)</u>										
looking	.7	3.2	3.0	.9	.8	1.3	1.7	1.4	2.1	14
eye contact		1.2	1.0			.4	.9	.7	1.4	15
utterance	3.0	4.5	4.0	2.2	1.8	1.9	2.1	2.2	2.1	10
joint utterance	.5	.4	.5	.4	.4	.5	.4	.5	.5	16
looking while listening		1.9	2.0			.6	.8	.7	1.0	16
looking while talking		.9	1.0			.6	.7	.8	.9	22

Table 7.2 Means of the various forms of behaviour observed, and an estimate of the accuracy of measuring the various forms of behaviour.

(A = audio channel only; M = monochrome image; C = cartoon image; L = low audio noise;
H = high audio noise)

	d.f.	looking			eye-contact EC			looking while listening			looking while talking		
		%	freq	mean	%	freq	mean	%	freq	mean	%	freq	mean
N	1	.3172**	94.31*	.0924	.0861**	16.59**	2.281	.1073**	54.24*	.0447**	.0592*	23.27*	.1581
I	1	.0211	51.50*	.0086	.0008	5.35	.927	.0028	17.40	.0060	.0001	4.77	.4909*
NxI	1	.0041	.15	.0002	.0009	1.15	.137	.0022	1.48	.0011	.0003	.00	.0611
PxN	5	.0038	6.15	.0202	.0036	1.18	.555	.0029	4.64	.0019	.0047	2.15	.5573
PxI	5	.0121	5.53	.0255	.0029	1.02	.323	.0062	3.92	.0114	.0018	.62	.0362
PxNxI	5	.0193	3.44	.0213	.0031	.53	.463	.0086	5.47	.0137	.0064	1.15	.3208
Average		17.7% (.39)	6.13	1.63 (.39)	5.7% (.17)	2.4	.84	5.5% (.21)	4.04	.77 (.24)	4.3% (.18)	3.77	.75

Table 7.3 Summary of the analysis of variance performed on the observations of the subjects' looking behaviour during the trucking game.

Factors are: N = noise level

I = type of video image (monochrome or cartoon)

P = pairs of subjects

* = $p < 0.05$

** = $p < 0.01$

(% = total duration, freq = number per minute, mean = mean length of each measure.

	d.f.	looking			eye-contact EC			looking while listening			looking while talking		
		%	freq	mean	%	freq	mean	%	freq	mean	%	freq	mean
C	1	1.911 ^{**}	307.3 [*]	.5799 [*]	.5168 ^{**}	336.5 ^{**}	1.978	1.039 ^{**}	180.7 [*]	.2648 ^{**}	.4273 [*]	254.1 [*]	.9989
I	1	.015	9.3	.0010	.0005	.8	.019	.002	.7	.0000	.0002	4.4	.3668
CxI	1	.007	13.8	.0024	.0005	.1	.204	.002	4.5	.0023	.0001	.0	.1171
PxC	5	.047	22.3	.0373	.0168	9.8	.433	.030	15.4	.0346	.0271	28.5	.5568
PxI	5	.009	1.2	.0121	.0042	5.9	.071	.010	2.8	.0073	.0009	3.6	.2057
PxCxI	5	.026	4.4	.0304	.0082	6.2	.138	.013	3.8	.0145	.0042	4.5	.0872
Average		27% (.51)	7.2	2.2 (.45)	9.7% (.26)	5.3	.82	12.3% (.31)	4.9	1.28 (.32)	7.4% (.24)	5.0	.83

Table 7.4 Summary of the analysis of variance performed on the observations of the subjects' looking behaviour during the periods of relaxed conversation and the trucking game.

Factors are: C = method used to stimulate conversation

I = type of video image (monochrome or cartoon)

P = pairs of subjects

	d.f.	utterance			no. of role changes per min.	silence %
		%	freq	mean		
C	1	.1765*	188.7*	50.91*	164.2*	.4632*
I	2	.0003	10.0	2.68	1.1	.0009
CxI	2	.0019	3.4	4.77	4.2	.0041
PxC	5	.0115	24.4	3.81	19.2	.0216
PxI	10	.0026	5.4	3.06	7.7	.0045
PxCxI	10	.0026	7.2	3.27	8.7	.0056
Average		40% (.64)	9.5	3.0	11.0	29% (.56)

Table 7.5 Summary of the analysis of variance performed on observations relating to the subjects' speech behaviour.

Factors are: C = method used to stimulate conversation

I = type of video image (monochrome, cartoon or no image)

P = pairs of subjects

Table 7.6 Difference between the measured and estimated duration of eye-contact for the three different modes of communication.

Difference (D) = measured duration - estimated duration.
(Expressed as a percentage of the total duration of the conversation.) The rank of the differences is shown so that the Wilcoxon Signed-Ranks Test may be applied.

T = Sum of 'ranks' with less frequent signs (critical value of T for 0.05 level of significance = 40).

monochrome image		cartoon image		audio only	$D_M - D_C$ rank
D_M	rank	D_C	rank		
2.65	13	2.39	13	-.03	5
23.74	17	7.47	15	.01	18
3.23	14	12.48	17	-.70	16 -
.70	7	-.83	10	.13	10
2.29	11	-.89	11	-.25	12
5.80	16	-.65	9	-.01	15
.63	6	.13	5	-.00	6
5.45	15	10.62	16	.44	14 -
-.12	2	-.06	3	-.01	$\frac{1}{3}$ -
.50	5	.44	7	-.02	$\frac{1}{3}$
-.19	$3\frac{1}{2}$	-.01	2	.25	4 -
-.06	1	-.00	1	.01	$\frac{1}{3}$ -
2.37	12	.11	4	-.01	11
27.89	18	15.61	18	1.96	17
.82	9	.21	6	.00	7
.76	8	-.63	8	.05	9
1.71	10	.97	12	-.06	8
-.19	$3\frac{1}{2}$	3.27	14	-.07	13 -
T = 10		T = 44			T = 48

7.4.6. Discussion of Results.

Looking behaviour during periods of audio noise.

The experiment showed that subjects looked at each other more frequently, and for a greater percentage of the total period of the interaction, when noise was added to the audio channel. The possible reasons for this increase (in the amount of looking) may be found if one considers Argyle's suggestions, as to the different uses that non-verbal cues have in a social interaction, (see section 1.2).

- 1) Interpersonal attitudes are often conveyed by vocal non-verbal cues; these are collectively referred to as 'tone of voice'. When audio noise is present, the tone of voice may be more difficult to perceive and the interactors will, therefore, look at each other more in an attempt to make greater use of the visual non-verbal cues, such as posture, as an indication of interpersonal attitudes.
- 2) The audio noise may make the beginning and the ending of utterances more difficult to detect, and because of this, the conversants may attempt to make greater use of the visual, non-verbal cues associated with channel control, eg. direction of gaze.
- 3) The vocal attention signals, such as 'yes' or 'I see', which are often used to signal mutual attention or responsiveness, may also be more difficult to perceive when noise is added to the audio channel. This would lead the interactors to look at each other more often, so that greater use could be made of head nods or facial expressions that indicate mutual attention or responsiveness.

In noisy environments, conversants may also look at each other more frequently to make greater use of lip reading. (The experiment described in chapter 5 demonstrated that subjects who had no previous formal practice at lip reading were able to use both the monochrome image and the cartoon image for lip reading.)

Interactors may or may not be aware that they look at each other more frequently during noisy periods, or that they make any use of lip reading, or any of the non-verbal cues mentioned above.

The level of noise used in the main experiment produced a change in the subjects' looking behaviour, but no change in the simple measures of the subjects' speech behaviour (eg. frequency and duration of utterances). If a higher level were used, the subjects' speech behaviour is likely to be different from that recorded during the main experiment, and if the much higher noise level were used, the conversation is likely to become very laboured or cease completely, (as demonstrated in the preliminary experiment). As many aspects of a persons' looking behaviour are correlated with his speech behaviour, (KENDON & COOK, 1969), the pattern of looking is likely to change as the level of audio noise is increased.

Looking behaviour during different types of conversation.

The main experiment showed that there was a greater frequency and duration of 'looking' during the periods of relaxed conversation, compared with periods when the trucking game was being played. As many aspects of looking behaviour are correlated with speech behaviour, some differences in looking behaviour would be expected, as the experiment also showed that the subjects' speech behaviour was different during these periods. While playing the trucking game, the subjects were in a competitive situation and, hence, the lower amount of looking during the game, compared with the periods of relaxed conversation, is in agreement with the results obtained by EXLINE, (1963), who showed that males look at each other less during competitive situations.

The other important difference between the trucking game and the periods of relaxed conversation is that during the game the subjects had to look down at their 'road maps' occasionally, in order to 'move their trucks' and to record their score. The subjects also seemed to gaze at the 'road maps', while planning their next move. (REID, (1971), in a survey of civil servants' communicating behaviour found that approximately

32% of two-person face-to-face meetings took place because the participants had either documents or diagrams to discuss.) Hence, the situation where the interactors have to look down at some document or diagram etc. is quite common.

As different types of conversation are likely to produce very different patterns of looking behaviour, care must be taken when comparing the results of different published experiments. KENDON, (1967), for his experiments, simply asked subjects to become acquainted with each other. The type of conversation that is produced from such interactions is likely to be similar in many respects to the type of conversation during the periods of 'relaxed conversation' in the author's experiment. In Kendon's experiments, the average total duration of 'looking' was 49% (of the total duration of the interaction), and this is surprisingly close to that measured by the author, in the periods of relaxed conversation, i.e. 44% when the monochrome image was used and 42% when the cartoon image was used.

Kendon also found in his experiments that the proportion of a subject's speaking time that he spent 'looking' (l_T), was less than the proportion of his listening time that he spent 'looking' (l_L). For the periods of relaxed conversation during the author's main experiment, the same relationship was found between l_T and l_L , i.e. $l_T < l_L$.

The author's results considered above are similar or consistent with the published results on non-verbal behaviour in face-to-face meetings. This implies (but does not prove) that the assumption made at the beginning of this chapter was correct, namely that users of a video-telephone would attempt to use many of the non-verbal cues that occur in face-to-face meetings.

Looking behaviour and type of video image.

The analysis of variances performed on 22 of the 24 measures of looking (and all of the measures of speech behaviour) showed that there was no significant difference between periods when the monochrome image was used and periods when the cartoon image was used. Because of this, the discussion above can be assumed to apply equally to when the monochrome or cartoon images are used. A further implication, therefore, is that a cartoon image is adequate for communicating many of the non-verbal cues used. This is not surprising, as a cartoon image is probably sufficient for the perception of 'head nods', some facial expressions, posture, position of the head and lip movements. However, some cues such as the direction of gaze will be more difficult to perceive accurately when a cartoon image is used. A much lower data rate would be required for transmitting a cartoon image than for transmitting a monochrome image, (see Appendix I), but such a cartoon image would almost certainly be unacceptable to a public who are familiar with high quality broadcast television.

Measured and chance eye-contact.

For periods when the monochrome image was used, the measured total duration of eye-contact was significantly greater than the chance duration, estimated using equation 5 of section 7.1.1. As the simple probabilistic model used for formulating the chance duration of eye-contact took into account the known consistent aspects of a person's looking behaviour, it may be concluded that a subject's looking behaviour was influenced by the looking behaviour of the facial image appearing on the screen. The likely explanation of this is, that when a chance eye-contact occurs, the subjects extend the period of eye-contact, for the following reasons, (from ARGLEYE & DEAN, 1965).

- 1) To try to establish a relationship between them.
- 2) To seek information of the other's reactions.
- 3) To indicate who should speak next.
- 4) They simply like to be seen, and eye-contact is the proof that they are being seen.

It is not the direction of gaze of the facial image on the video monitor screen, by itself, that directly influences the viewer's direction of gaze, as the viewer must also be aware that the facial image on the monitor screen represents someone, who is at the same time observing an image of the viewer himself. Thus, when viewing a television broadcast, the direction of gaze of the facial image on the monitor screen will not influence the viewer's direction of gaze because he knows that he is not being observed by the person represented on the monitor screen.

For periods when the cartoon image was used, there was not a significant difference between measured and chance duration of looking. However, the value of the statistic calculated for testing the null-hypothesis that there was no significance, was only a little above the critical value needed for rejecting the null-hypothesis; thus, if a larger sample were used, the difference between measured and chance duration of eye-contact may prove significant. With a cartoon image, it is not possible to perceive accurately the direction of gaze of the other conversant. Hence, if a larger sample did show a significant difference between the measured and chance eye-contact, the implication would be that viewers were able to infer the direction of gaze from other cues, such as the position of the head. This would not be unreasonable, as a conversant often turns his head away when he stops looking at the other conversant.

Subsidiary experiment.

The subsidiary experiment showed that when a junior (freshman) student was being interviewed by a senior student, the junior student had longer gazes at the senior student than vice-versa. If this is at all typical of the looking behaviour during such interviews, of juniors by seniors, then it casts doubt on the methodologies adopted by some experimenters for studying the looking behaviour in this type of interview. For example, EFRAN, (1968), in his experiments, instructed the senior interviewing person (a confederate) to gaze continuously at the junior subject.

Speech behaviour.

It was expected that the simple measures of speech behaviour would be significantly affected by the method used to stimulate conversation; thus, one expects there to be fewer but longer utterances during a relaxed conversation than while playing a game, which requires alternate moves from the participants. What is more surprising is that neither the audio noise level or the form of communication channel had any significant effect on these simple measures.

One of the difficulties of comparing the results of the author's experiments with those of other published experiments is that other research workers use a different definition of an utterance. For example, RICHARDS, (1964), defines a silence as a period of more than 350 milliseconds when neither of the conversants is speaking. He also does not exclude attention signals. Using his definition of an utterance, one would expect a given conversation to appear to have shorter utterances and have more role changes, than if the author's definition was used, (see section 7.1.2.) This would account for the differences between the author's results and those of Richards. Hence, Richards states that during a conversation the mean length of an utterance is approximately 1 second, and that there are approximately 15 role changes per minute. During the periods of relaxed conversation in the author's experiment (and using the author's definition of an utterance), the mean length of an utterance was approximately 4 seconds and the number of role changes per minute, approximately 9. During the trucking game, the mean length of utterance was 2 seconds and the number of role changes, 13.

MARTIN, (1931), first suggested that the number of repetitions asked for by a telephone user gives some clue of the difficulty encountered during a telephone conversation. RICHARDS, (1967), when analysing conversations carried out over a satellite link between New Zealand and Great Britain, found that the number of requests for repetition, in the first minute of each conversation, occurred with an approximate average frequency of

1 per 2 minutes. The extremely low number of requests for repetitions in the author's experiments implies that, even with the noise added, the 'quality' of the audio channel used in the experiments was higher than that encountered on some satellite links.

Questionnaire.

The results of the questionnaire showed that subjects rated the 'monochrome image plus audio channel' significantly higher than the 'audio channel only', for comfort, for communicating their ideas or feelings and for telling how the other conversant was reacting to them. The subjects were unaware of the forms of behaviour being measured, but because they had to carry on a conversation using different types of communication channels, it is highly probable that they realised that some form of comparison was being made between the types of communication channel. Because of this, their answers to some of the questions in the questionnaire may have been determined by what they thought the advantages should be, of adding video channel to the audio channel, instead of being determined by their experiences during the experiment.

The questionnaire only provides evidence that the subjects thought that they were better able to judge the other conversants' reactions, when they could also see him on the monitor screen, compared to when they could only hear the other conversant; it does not provide evidence that they were actually better able to judge the other's reactions. This result may be compared with the result of an experiment by REID, (1970), who gave subjects the task of telling truth from lies, when they could and could not see the speaker. The subjects were more confident that they were better able to tell truth from lies when they could see the speaker, but their actual measured performance showed that this confidence was unfounded.

The video channel will give video telephone users the opportunity to use a wider range of non-verbal cues for judging the other conversant's

reactions. DAVITZ, (1964), has shown that there are large individual differences in the ability to express or understand emotional messages conveyed by non-verbal cues, so that the wider range of cues provided by the video channel will not be necessarily utilized by all users. One advantage of the video channel is that it allows a user, who is still speaking, to immediately judge the reactions of the other conversant, whereas with an audio only communication link, a speaker often has to wait until the other person replies before he can judge his reactions.

Reliability of behaviour measures.

Of the many papers published on looking behaviour, only a relatively small number contain any discussion of the reliability of measuring a subject's looking behaviour. EXLINE (1963), obtained a measure of reliability by using two observers simultaneously recording the looking behaviour of a subject. An 'index of agreement' was derived by dividing the total duration of coincident observations by the maximum total duration of 'looking', indicated by combining the two observations. Unfortunately, because of the different methods used to measure reliability, it is not possible to compare Exline's value of 0.81 for the 'index of agreement' with the author's estimate of 4%, for the mean error of measuring the duration of looking. Other authors (for example, STEPHENSON & RUTTER, 1970) have merely given the correlation between the measures of the two observers who were both recording the looking behaviour of one subject. However, this procedure does not expose any consistent differences between the observers. STRONGMAN & CHAMPNESS, (1968), in pre-experimental trials, used two observers to measure the total duration of a subject's looking. They found a 95% agreement (5% error) between the observers; this accuracy of measurement was similar to that in the author's experiments.

MEHRABIAN & WILLIAMS, (1969), used a video-tape recorder to record many different aspects of a subject's behaviour. They gave the reliability coefficients of their measurement of 22 different forms of non-verbal behaviour, but they did not report the method of measuring or defining the reliability coefficients. The reliability coefficient of measuring eye-contact was given as 0.50, and this was one of the worst coefficients reported. As the experimenters wished to record the

subjects' leg movements as well as his direction of gaze, the definition in the region of the eyes would necessarily be poor, and this was probably one of the main reasons for the poor accuracy of measuring eye-contact. However, the author's results demonstrate that a video-tape recorder may be used for accurately recording eye-contact.

During the author's experiments, when a subject looked at a facial image on the monitor screen, it appeared to the observer who was measuring looking behaviour, that the subject was looking directly at him. This simplified the observer's task and consequently the observers did not require any special training. CLINE, (1967), showed that subjects were most accurate at judging the direction of gaze when the 'looker' was gazing directly at them; thus, the observers in the author's experiments were in the most favourable position for judging when the subjects looked at the monitor screen.

The duration of each 'eye-contact' was quite short; the average duration was approximately one second. If the same two observers were used for measuring the looking behaviour throughout the whole experiment, then any consistent differences in their reaction times would lead to errors in the measurement of eye-contact. By using as many different observers as possible (as did the author), this particular form of error in measuring the amount of eye-contact will be reduced. None of the published experiments concerned with looking behaviour seemed to have this simple procedure to increase the reliability of the measurement of eye-contact.

The reliability of many of the measures, and particularly the measures of speech behaviour, would be increased if the observers studied the videotape recordings before performing their appointed observation task. However, such a procedure would require the observers to spend much more time on an experiment and it was not possible to do this in the author's experiment.

None of the published papers concerned with communication behaviour have contained any estimates of the reliability of measuring the more detailed forms of looking and speech behaviour (e.g. looking while listening). It is not possible, therefore, to compare the author's estimates of reliability of these measures with any other.

Chapter 8

General Summary and Suggestions for Further Work

This chapter briefly summarises the results of the experiments reported in this thesis and makes some suggestions for further work.

8. General Summary and suggestions for further work.

The experiments reported in this thesis fall into two categories: the perception of various non-verbal cues via a video link, and the measurement of subjects' behaviour while they are engaged in a conversation over a video link. Experiments of the first type are an essential pre-requisite to performing experiments of the second type, as it is necessary to know whether the participants of a videotelephone conversation are able to perceive many of the non-verbal cues that are often used in a conversation.

The first experiment reported was concerned with the perception of a person's direction of gaze when observed via a video link (see section 4.1). The results showed that subjects could accurately perceive the direction of gaze via the video link; their accuracy was very similar to that reported for 'live' lookers at a distance of 2 metres. There was a slight difference between the perception of vertical and horizontal rotation of a person's eyes, but again, these differences were similar to those reported for experiments with live lookers. It was argued that the results of such experiments could influence the choice of the relative positions of the camera and picture monitor in a videotelephone station set. Experiments such as these are, of course, an essential preliminary before performing any measures of 'eye contact' between videotelephone users during a conversation over a video link: if the users cannot perceive the direction of gaze of the other participant, then any detailed measures of eye-contact are meaningless.

The second experiment investigated the subjects' ability to lip read when different types of video images were used, (see chapter 5). It was shown that a simple cartoon image of the speaker's face was virtually as good as the full normal monochrome video image for lip reading. Any results concerned with the utility of videotelephone systems for lip reading will be of some importance to the deaf and hard of hearing, as this is a section of the community that would benefit greatly from the introduction of such a system.

The third experiment reported was concerned with the recognition of facial images, (see section 6.1). The facial images to be recognized were reproduced as either half-tone monochrome video images or different forms of cartoon images. The results did show that a cartoon image is at least as good as a normal monochrome video image with a signal to noise ratio of -1.2 dB; also, that the horizontal edge detector produced a more recognizable cartoon than did the vertical edge detector. This experiment was only exploratory and further work needs to be done on the methodology of such experiments.

The fourth experiment was concerned with the perception of the relative proximity of a facial image on a monitor screen and that of a person 'in the flesh' (see section 6.2). It was shown that, when viewing a facial image on the monitor screen, a larger retinal image must be formed than that produced by a 'live' face, in order to appear at the same proximity. One suggestion for further work here is to repeat this experiment to see how factors such as video monitor screen size and viewing distance influence the perception of proximity. It seems likely that there are complex psychological and social factors which influence the perception of proximity of a video image. For example, simply knowing that the other conversant is at a distant location may influence the perception of proximity. The feeling of proximity, or presence of the other, may be an important factor when comparing face-to-face meetings with videotelephone conversations.

In chapter 7, three experiments were reported which analysed the looking and speech behaviour of subjects who were conversing over a video link. It was shown that the audio noise level, the method of stimulating conversation, and the relative status of the participants influenced the looking behaviour. It was argued that the change in looking behaviour caused by the audio noise could be accounted for by considering the uses of non-verbal behaviour, as described by ARGYLE (1969). A simple probabilistic model was developed to estimate the amount of chance eye contact. When the normal monochrome image was used, it was found that the measured amount of eye contact was significantly different from the

estimated amount; this indicated that a subject's direction of gaze was directly influenced by the direction of gaze of the person with whom he was conversing over the video link. This did not occur when the cartoon image was used. It was also shown that during the periods of relaxed conversation between participants over the video link, the total duration of looking as well as the relative amounts of looking while talking and looking while listening were similar to the published results of experiments concerned with face-to-face conversations. This possibly indicates that the users of a videotelephone link will attempt, at least initially, to use non-verbal cues in the same way as they are used in face-to-face conversations. However, with prolonged use, the videotelephone user may evolve new forms of conversational behaviour.

In this particular series of experiments the average total percent of time subjects spent looking at the monitor screen, when normal half-tone images were used, was 43% during the periods of relaxed conversation, and only 14% when the trucking game was being played. The average length of each gaze during these periods was 3.2 and 1.3 seconds, respectively. This length of gaze greatly differs from the gaze usually allowed in subjective tests of broadcast quality images. In these tests, the subjects are usually presented with a test image for approximately 10 seconds, (also, only 4 different scenes are used for the whole test series) ALLNATT & BRAGG (1968). This again cast doubt as to the utility of applying these 'broadcast' testing techniques to videotelephones, as the much shorter glances and smaller total duration of looking at the monitor screen may make the users more tolerant towards any impairments of the image.

It has been stated several times in this thesis that many of the results of the experiments may be influenced by the video image used. This is particularly so when any measure of eye contact or direction of gaze was involved. A general suggestion for further work is, therefore, to repeat any of the experiments reported in this thesis, in order to investigate what effect resolution, and particularly lower resolution,

has on any of the results. There are important practical reasons for using lower resolution video images for future experiments. Any videotelephone system that is likely to be introduced in the future will have considerably lower resolution than that of the video system used by the author. Furthermore, the wideband video 'conferencing' facilities introduced by the British Post Office (Confravision), allows up to five participants to sit in a row in front of a camera. In these circumstances, although the video signalling parameters are the same as those used by the author, the number of picture elements associated with each face is very much lower than in the author's experiments. (When viewing five persons on a Confravision video monitor, each face takes up approximately $1/50$ of the screen area of the face compared with the author's experiments.)

This thesis has only been concerned with dyadic interactions, there would be many possibilities for further ^Sresearch into videotelephone user's behaviour, if group meetings were considered.

APPENDIX I.Number of white picture elements in a cartoon image.

The number of white picture elements contained in a cartoon image will depend on two factors: (1) the original facial image, (2) the various threshold settings in the cartoon generator. By way of an example, this appendix will consider the two facial images which are shown in figure 3.7. These are two of the facial images that were used in the experiment described in chapter 6. The cartoon in the upper half of figure 3.7 'Archie', a person wearing spectacles, is typical of the more complex cartoon, which contains many white picture elements. The cartoon in the lower half of figure 3.7 'Max' was chosen as being representative of the cartoon image with relatively few white picture elements.

To measure the number of white picture elements in each cartoon the output of the cartoon generator was sampled approximately 50 nsec after each clockpulse (i.e. halfway between each clockpulse). The number of pulses per second occurring at the output of the sampler was measured using a Marconi TF 2414A/2M counter.

Table AI shows the number of white picture elements per frame that were produced when the simple edge detector, or the horizontal edge detector, or the complete cartoon generator were used. (See section 3.7 for a description of different methods of generating cartoon images). The video monitors used in the author's experiment allowed 512 scanning lines to be visible, whereas the number of horizontal elements visible was 450. (See figure 3.6). The total number of picture elements (N) visible was, therefore, equal to 230,400. The figure in brackets in table AI represents the number of white elements in any particular cartoon expressed as a percentage of N . Of the many cartoon images used in the author's experiments, none contained more than 3% of N white picture elements.

Table A1Number of white picture elements in a cartoon image.

cartoon	Archie	Max
Simple edge detector	2872 bits/frame (1.3%)	1520 bits/frame (0.7%)
Horizontal edge detector	3280 " (1.4%)	2160 " (0.9%)
Complete cartoon	5920 " (2.6%)	3520 " (1.5%)

The figure in brackets represents the number of white elements in any particular cartoon expressed as a percentage of the total number of picture elements contained in one frame.

APPENDIX II.Data rate for image transmission.

In this appendix a comparison is made between the estimated data rate required to transmit the 'monochrome' and cartoon images used in the author's experiments. Two monochrome and cartoon images will be considered; these are shown in figure 3.7. (Note the cartoon images are the complete cartoon images used in Appendix I). The comparisons of data rates are summarized in table A2.

An estimate of the number of bits per frame required for the transmission of the monochrome image, using conventional pulse code modulation, may be obtained by multiplying the number of picture elements per frame (N), by the number of bits per picture element (P), required to satisfactorily code the signal amplitude. For the purposes of comparison, it has been assumed that six bits per sample would be required for coding the signal amplitude. This figure was chosen as it is often used as a basis of comparison, (SCHREIBER, 1967). It must be stressed, however, that this simple calculation allows only rough comparisons, as the actual monochrome image used was generated by a non-quantized, non-sampled, analogue video signal.

It would be possible to transmit the monochrome image at a lower data rate than estimated above, if other methods of coding were used. A few of the coding methods that may be used to obtain data rate reductions are mentioned in section 3.2.

To transmit a cartoon image by conventional raster scanning techniques would require N bits per frame (i.e. one bit per picture element). However, because in a cartoon image only a small fraction of the total number of picture elements are white, more efficient methods of coding may be devised.

In the process of generating the cartoon images shown in figure 3.7, a system of line broadening was used to produce a more subjectively pleasing cartoon. (See section 3.7). Hence, when an edge was detected, two adjacent white picture elements were generated. Therefore, only half of the number of picture elements shown in table A1 are needed to completely describe each cartoon image. Thus, for coding purposes, the number of white elements (L) in the complete cartoon image will be taken as 2960 and 1760, for the cartoons 'Archie' and 'Max', respectively.

A saving in the data rate needed to transmit the cartoon may be obtained by simply taking into account the low probability of white picture elements. Thus, if the conventional raster scanned image is considered as a sequence of independent binary digits, and each element has a probability L/N of being white, then the first order entropy per frame of such a source is given by the following equation, (see LAEMMEL, 1953).

$$H = N \left[\frac{L}{N} \log_2 \left(\frac{N}{L} \right) + \left(1 - \frac{L}{N} \right) \log_2 \frac{1}{\left(1 - \frac{L}{N} \right)} \right]$$

Table A2 gives the results of using this formula to calculate the entropy per frame for the two cartoons under consideration. It would be possible (although not necessarily practical) to devise codes that would transmit the cartoon image with a data rate equal to the entropy. Such a system would result in a data compression of 10:1 for the cartoon 'Archie', and 15:1 for the cartoon 'Max'. Greater reductions in data rates may be obtained if the second or higher order probability distribution were exploited. (SCHREIBER, 1956). Unfortunately, the measurement of the higher order probability distributions was outside the scope of the author's project.

Many different schemes may be used for the reduction of the data rate required to transmit raster scanned line drawings. For example, LOEB (1953) proposed a 'stopping spot' process; a process which uses a variable speed sweep. A working system, which uses this principle, has been demonstrated by BELTRAMI & MOSCA (1969). It was used to transmit visual images over existing telephone networks. Some further reductions in the data rate for the transmission of raster scanned images may be obtained if multi-beam scanning is used. For example, CHITNIS (1971) compared the data compression possibilities of multibeam scanning of weather maps containing 9% black picture elements. It was found that with single beam scanning the best compression ratio was approximately 5:1, whereas with three beam scanning it was approximately 8:1.

Because the cartoon images contain so very few white picture elements some reduction in the transmission data rate may be obtained by simply sending the position co-ordinate of each white picture element. For the picture parameter used in the author's experiments, each picture element would require 18 bits to specify its position co-ordinates. Compared with conventional raster scanning, using this simple coding system would result in a data compression of 4.3:1, for the cartoon 'Archie', and 7.3:1, for the cartoon 'Max'.

Data compression schemes which use single line raster scanning do not exploit one important attribute of line drawings or cartoons, namely the continuity of the lines. FOY (1964) has derived an expression for the entropy of a line drawing which takes into account the continuity of the lines in the drawing. The calculation of entropy, using Foy's formula, requires the statistics of the arc lengths. Again, the measurement of such statistics is outside the scope of this thesis.

A coding scheme which does exploit the continuity of lines is contour following (BENJAMIN, 1953). Contour following systems may take several different forms, only one of which will be considered here.

Consider a system which has a high speed local scan to find the start of each line; once the starting point is known, only three bits are necessary to specify the next adjacent element on the line. The specification of the starting point of each line would require 18 bits, (assuming the same picture parameters as those used in the author's experiments). The total number of bits required to transmit one frame of the cartoon image would therefore be $18 \times$ (total number of individual lines), plus $3 \times$ (the total number of white picture elements). By examining a slide projection of a photograph of the cartoon 'Archie' and 'Max', it was found that they contained respectively, 67 and 70 individual lines. (If very short or single, isolated elements were ignored, these figures become 42 and 40). Using such a contour following system would result in a data compression ratio of 23:1, for the cartoon 'Archie', and 35:1, for the cartoon 'Max'.

It may be argued that any system other than raster scanning is impractical. However, the Japanese have already proposed a system of contour following, which may be used for narrow band transmission of drawings, (SEZAKI & KATAGIRI, 1965).

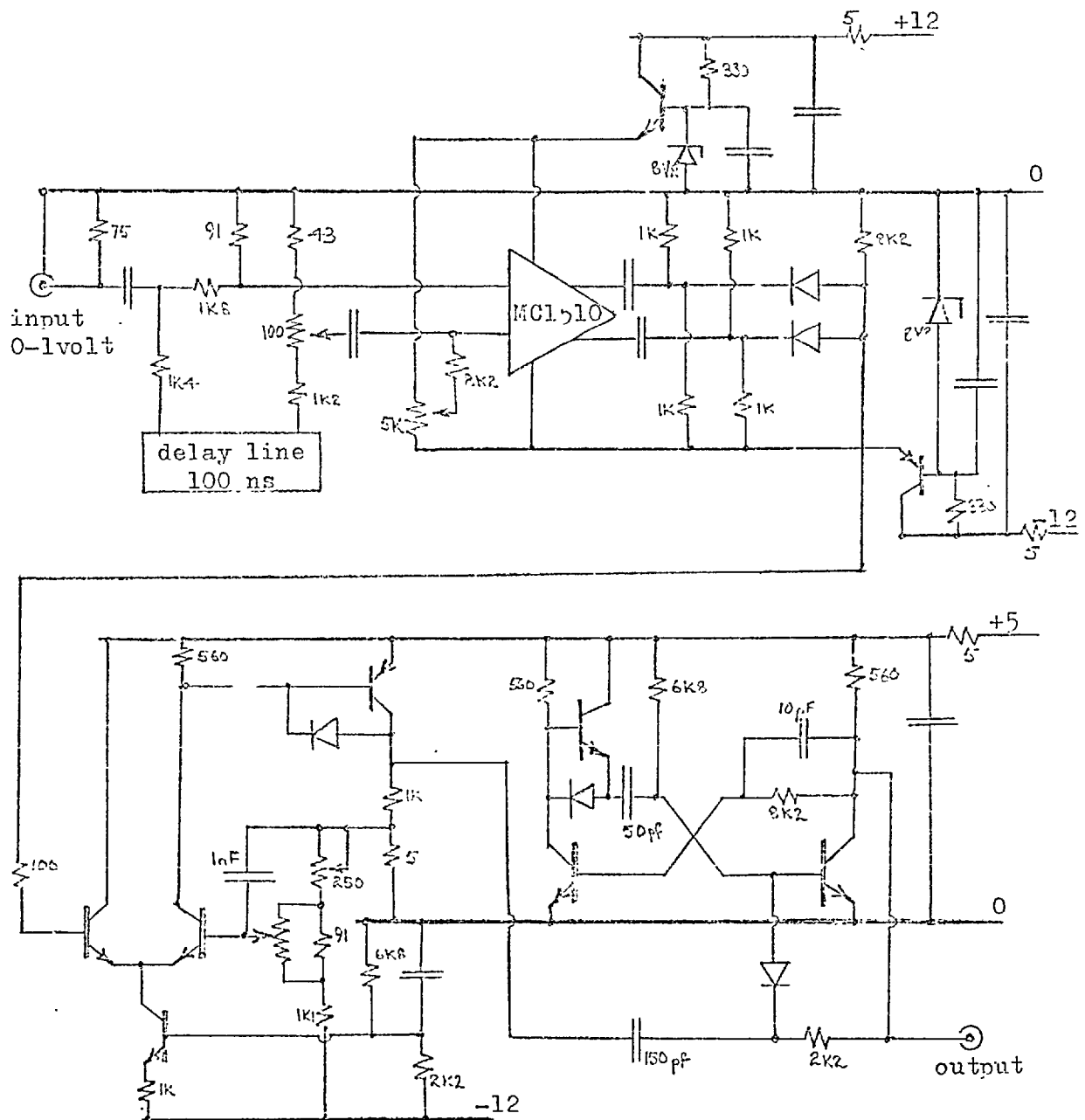
Table A2Data rates for image transmission.

	Archie	Max
<u>Monochrome Image</u>		
conventional P.C.M. (6 bits/sample)	1382.	1382.
<u>Cartoon Image</u>		
conventional raster scanning	230.4	230.4
1st order entropy of single line raster scanning	22.8	14.9
specify position co-ordinates of each white picture element	53.2	31.7
contour following plus start points	10.1	6.5

All numbers are bits/frame x 10^3

APPENDIX IIICircuit Diagram of the Simple Edge Detector.

The simple edge detector was constructed in the form of a single 'plug in' printed circuit module. The same module with a few modifications was also used as part of the horizontal edge detector.



Transistors :- NPN P346A , PNP V405A

Diodes :- AAZ13

Capacitors :- 0.1uF unless stated

Simple Edge Detector

REFERENCES.

- Allnatt, J.W.
1968 Subjective studies concerning the viewphone project.
4th International Symposium on Human Factors in
Telephony, Bad Wiesse, Germany.
- Allnatt, J.W. & Bragg, E.J.W.
1968 Subjective quality of television pictures impaired by
sinewave noise and low frequency random noise.
Proc. I.E.E., 115, No.3,
- Allnatt, J.W. & Prosser, R.D.
1965 Subjective quality of television pictures impaired
long-delayed echoes. Proc. I.E.E., 112, No. 3, p.487.
- Anstis, S.M., Mayhew, J.W. & Morley, T.
1969 The perception of where a face or television 'portrait'
is looking. Amer. J. Psychol., 82, pp. 474-489.
- Argyle, M.
1969 Social Interaction. Methuen, London.
- Argyle, M. & Dean, J.
1965 Eye-contact, distance and affiliation. Sociometry,
28, pp.289-304.
- Argyle, M., Lalljee, M.G. & Cook, M.
1968 The effects of visibility on interaction in a dyad.
Hum. Relat., 21, pp. 3-17.
- Argyle, M. & McHenry, R.
1969 The effect of spectacles on judgements of intelligence.
In 'Research project on non-verbal communication:
progress report to the SSRC for period Oct.1968-Sept.1969.
M.Argyle, Oxford University, Institute of Experimental
Psychology, mimeographed.
- Argyle, M., Slater, V., Nicholson, H., Williams, M. & Burgess, P.
1970 The communication of inferior and superior attitudes
by verbal and non-verbal signals. Bulletin de
Psychologie,
- Argyle, M. & Williams, M.
1969 Observer or observed? A reversible perspective in
person perception. Sociometry, 32.
- Arnoult, M.D.
1957 Toward a psychophysics of form. See GRAHAM (1965)
p.569.

- Baldwin, M.W.
1951 Subjective sharpness of additive colour picture.
Pro. Inst. Radio Engineers, 39, No.16, p.1173
- Bales, R.F.
1950 Interaction Process Analysis. Addison-Wesley,
Cambridge, Mass..
- Bartlett, M.S.
1947 The use of transforms. Biometrics, 3, pp.39-52.
- Bell, C.
1806 The organs of the senses familiarly described, being an
account of the conformation and functions of the eye,
ear, nose, tongue, and skin. Harvey, London.
- Beltrami, A. and Mosca, V.
1969 High speed visual communication over existing telephone
networks. I.E.E.E. Trans. Comm. Tech., 17, pp.395-403
- Benjamin, R.
1953 Discussion following the LOEB (1953) paper.
- Birdwhistell, R.L.
1952 Introduction to kinesics, an annotation system for analysis
of body motion and gesture. University of Louisville
Press, Louisville, Ky.
- Birdwhistell, R.L.
1960 Kinesics and communication. In - E. Carpenter & M. McLuhan
(eds) Exploration in Communication - An anthology,
pp.54-64, Beacon Press, Boston.
- Birdwhistell, R.L.
1966 Some relations between Kinesics and spoken American
English. In - A.G. Smith (ed.) Communication and Culture;
Readings in the Codes of Human Interaction. New York,
Holt, Rinehart & Winston, pp.182-189.
- Boomer, D.S.
1965 Hesitation and grammatical encoding. Lang & Speech, 8,
pp.148-158.
- Bostrum,
1968 Game theory in communication research. J. of Communication,
18, pp.369-388.
- Boynton, R.M.
1957 Recognition of critical targets among irrelevant forms.
see GRAHAM (1965) p.569.

- Brannigan, G. and Humphries, D.
1969 I see what you mean... New Scientist, 22nd May, pp. 106-109.
- Buck, T.M., Casey, H.C., Dalton, J.V. and Yamin, M.
1968 Influence of bulk surface properties on image sensing silicon diode arrays. B.S.T.J., 47, pp. 1827-1854.
- Budrikis, Z.L.
1964 On the channel capacity of the human sense of vision. Proc.I.R.E.E. Australia, p.228.
- Burdett, P.A.
1962 The effect of scan direction on the intelligibility of a picture. I.E.E. Report of the International TV conference, May-June, pp.27-34.
- Busala, A.
1960 Fundamental considerations in the design of a voice-switched speakerphone. B.S.T.J., 39, pp.265-294.
- Butler, J.
1964 Visual discrimination of shapes by humans. Qurt. J. of Exp. Psychol. 16, pp.272-276.
- Carmichael, L., Hogan, H.P., and Walter, A.A.
1932 An experimental study on the effect of language on the reproduction of visually perceived form. J. exp. Psychol. 15, pp.73-86.
- Carson, D.N.
1968 The evolution of the PICTUREPHONE service. Bell Lab. Record, 46, pp.282-291.
- Champness, B.G.
1970 Mutual glance and the significance of the look. Advancement of Science, March, pp.309-312.
- Champness, B.G.
1971 Patterns of Eye Gaze. Ph.D. Thesis, Exeter University, (to be submitted)
- Cherry, E.C.
1957 On Human Communication. New York, Wiley.
- Cherry, E.C.
1968 Some thoughts about our proposed research programme. Personal communication to the GPO Research Station, Dollis Hill, 3rd. DEcember.

- Cherry, E.C., Kubbah, M.H., Pearson, D.E., & Barton, M.P.
1963 An experimental study of the possible bandwidth compression of visual image signals. Proc.I.E.E.E., 51, No.11, p.1507.
- Chitnis, A.
1971 Some data compression methods for the electric transmission of monochrome pictures. Ph.D. thesis, University of London, (to be submitted in 1971)
- Cline, M.G.
1967 The perception of where a person is looking. Amer.J. Psychol. 80, No.1, pp.41-50.
- Cook, M.
1969 Anxiety, speech disturbances and speech rate. Br. J. clin. Psychol., 8, pp.13-21
- Cook, M.
1970 Non-verbal signalling in social interaction. Ph.D. thesis, Oxford University.
- Cook, M., Argyle, M., & Lalljee, M.G.
1967 The effect of varying visibility on floor apportionment Unpublished. (Institute of Experimental Psychology, Oxford University)
- Cochran, W.G., & Cox, G.M.
1957 Experimental Designs, Second Edition. John Wiley.
- Corocan, D.W.J., & Weening, D.L.
1969 On the combination of evidence from the eye and ear. Ergonomics, 12, No.3, pp.383-394.
- Davitz, J.R.
1964 The Communication of Emotional Meaning. McGraw-Hill, New York.
- Darwin, C.
1872 Expression of the Emotions in Man and Animals. London, Murray.
- Dember, W.N.
1963 The Psychology of Perception. New York; Holt, Rinehart & Winston.
- Deusch, S., Balaban, P., and Itzmon, A.
1966 Experimental results of pseudo-random dot scan. Final Report, Phase(II) No.PIBMRI-1308-66. Polytechnic Institute of Brooklyn.

- Deutsch, M., & Krauss, R.M.
1960 The effect of threat on interpersonal bargaining.
J. abnorm. soc. Psychol., 61, pp.181-189.
- Dorros, I.
1971 The Picturephone system: The network.
B.S.T.J., 50, pp.221-234.
- Edwards, A.L.
1968 Experimental Design in Psychological Research; (third
edition), New York; Holt, Rinehart & Wilson.
- Efran, J.S.
1968 Looking for approval: effects on visual behaviour of
approbation from persons differing in importance.
J. pers. soc. Psychol. 10, No.1, pp.21-25.
- Efran, J.S., & Broughton, A.
1966 Effects of expectancies for social approval on visual
behaviour. J. pers. soc. Psychol. 4, pp.103-107.
- Egan, J.P.
1946 Report for S.S. Stevens. Psycho-Acoustic Laboratories.
Harvard.
- Ellsworth, P.C., & Carlsmith, J.M.
1968 Effects of eye-contact and verbal content on affective
response to a dyadic interaction. J. pers. soc. Psychol.
10, No.1, pp.15-20.
- Exline, R.V., Gray, D., & Schuette, D.
1965 Visual interaction in a dyad as affected by interview
content and sex of respondent. J. pers. soc. Psychol.
1, pp.201-209.
- Exline, R.V., Thibaut, J., Brannon, C., & Gumpert, P.
1961 Visual interaction in relation to Machiavellianism and
an Unethical act. Center for Research on Social Behaviour
Technical Reports; No.16. University of Delaware, Newark,
DEL. Mimeo. (Abstract in Am. Psychologist, 16, p396)
- Exline, R.V., & Winters, L.C.
1965 Affective relations and mutual glances in dyads. In-
Affect, Cognition and Personality. (S.S.Tomkins and
C.E.Izard, eds.), pp.319-350. Springer, New York.
- Fano, R.
1961 Transmission of Information. M.I.T. Press & John Wiley.
- Fischer, K., & Ebel, H.
1968 Problems of how to position users in front of a video-
telephone. 4th International Symposium on Human Factors
in Telephony; Bad Wiesse, Germany. p269 .

- Flohner, N.
1970 Human aspects of loudspeaking telephones. 5th International Symposium on Human Factors in Telecommunications. London.
- Foy, W.H.
1964 Entropy of simple line drawings. I.E.E.E. Trans. Infor. Theory., IT-10, pp.165-167
- Frijdu, N.H.
1953 The understanding of facial expression of emotion. Acta Psychol. 9, pp.294-362.
- Frijdu, N.H.
1958 Facial expression and situational cues. J. abnorm. soc. Psychol. 57, pp.149-154.
- Frijdu, N.H.
1959 Facial expression and situational cues: a control. Acta Psychologica, 18, pp.239-244.
- Gabor, D., & Hill, P.C.J.
1961 Television band compression by contour interpolation. Proc. I.E.E., 108 part B, p303.
- Gallie, W.E.
1952 Peirce and Pragmatism. Pelican Books Ltd., Hammondsworth, England.
- Gardner, M.B.
1966 Talking levels in loudspeaker communication. J. Acoust. Soc Am., 40, p955.
- Garner, W.R., & Morton, J.
1969 Perceptual independence: definitions, models, and experimental paradigms. Psychol. Bull., 72. No.4, pp.233-259.
- Garratt, G.R.M., & Mumford, A.H.
1952 The history of television. Proc.I.E.E., 99, part IIIA, p25.
- Galyas, K.
1969 Loudspeaking telephone subscribers set. Elec. Comm. USA, 44, No.1.
- Geffe, F.R.
1962 Simplified modern filter design. Iliffe, Guildford.

- Gibson, J.J. & Pick, A.H.
1963 Perception of another person's looking behavior.
Am. J. Psychol., 76, pp.386-394.
- Gleiss, N.
1968 The effect of bandwidth restriction on speech transmission quality in telephony. 4th International Symposium on Human Factors in Telephony, Bad Wiessee, Germany.
- Goffman, E.
1956 Embarrassment and social organisation.
Am. J. Sociol., 62, pp.264-271.
- Goffman, E.
1959 The presentation of self in everyday life. Doubleday Anchor, Golden City, New York.
- Goldman-Eisler, F.
1961A The significance of changes in the rate of articulation.
Lang. & Speech., 4, pp.171-174.
- Goldman-Eisler, F.
1961B Hesitation and information in speech. In C. Cherry(ed) Information Theory. Butterworth, London.
- Graham, R.E.
1962 Snow removal- a noise-stripping process for picture signals. I.R.E. Trans. on Inform. Theory, IT-8, No.2
- Graham, C.H.
1965 Vision and visual perception. John Wiley, New York.
- Gregory, P.
1964 Deafness and public responsibility. The provision of hearing aids. Occasional Papers on Social Administration, No.7, Codicote Press.
- Garneau, M.
-- Dept. of Electrical Engineering, Imperial College, London.
- Hall, A.D.
1964 Experiments with PICTUREPHONE service.
Bell Laboratories Record, 42, No.4, p114
- Hartree, D.R.
1957 Numerical Analysis. (2nd edition) Oxford Press, London.

- Healy
1968 Transportation or communication; some broad considerations. I.E.E. Trans. Comm. Tech. COM-16, No.2.
- Herberle, W.
1968 Loudspeaking telephone versus standard handset. 4th International Symposium on Human Factors in Telephony, Bad Wiessee, Germany.
- Hershenson, R.
1962 Reaction time as a measure of intersensory facilitation. J. Exper. Psychol., 63, pp.289-293.
- Hockey, R.
1969 Noise and efficiency: the visual task. New Scientist, 1st May, pp.244-246.
- Hubel, D.H. & Wiesel, T.N.
1965 Receptive field and functional architecture in two non-striate visual areas (18 & 19) of the cat. J. Neurophysiol. No.2, p.289.
- Ives, H.E.
1927 Television. B.S.T.J. 6, pp.551-559.
- Jacobson, H.
1951 The information capacity of the human eye. Science, 113, p.292.
- Jaffe, J. & Feldstein, S.
1970 Rhythms of Dialogue. Academic Press, N.Y. & London.
- Jesty, L.C.
1952 Television as a communication problem. Proc. I.E.E., 99, Part IIIA, pp.761-769.
- Jesty, L.C.
1958 The relation between picture size, viewing distance and picture quality. Proc. I.E.E., 105B, pp.425-439.
- Kash, S.V. & Mahl, G.F.
1965 The relationship of disturbances and hesitations in spontaneous speech to anxiety. J. pers. soc. Psych., 1, pp.425-433.
- Kendon, G.S.
1967 Some functions of gaze direction in social interaction. Acta. Psychol., 26, pp.22-63.

- Kendon, G.S. & Cook, M.
1969 The consistency of gaze patterns in social interaction.
Br. J. Psychol., 60, pp.481-494.
- Kelly, D.H.
1962 Information capacity of a single retinal channel.
I.R.E. Trans. Inform. Theory, IT-8, No.3, p.221.
- Kleck, R.E. & Nuessle, W.
1968 Congruence between the indicative and communicative
functions of eye contact in interpersonal relations.
Br. J. soc. clin. Psych., 7, pp.241-246.
- Kretzmer, E.R.
1956 Reduced alphabet representation of television signals.
I.R.E. Convention Record, Part 4, pp. 140-147.
- Kris, E.
1952 Psychoanalytic Explorations in Art.
International Universities Press, Inc., New York.
- Laemmle, A.E.
1953 Discussion following the LOEB (1953) Paper.
- Lieberman, P.
1965 On the acoustic basis of the perception of intonation
by linguists. Word, 21, pp.40-54.
- Loeb, J.
1953 Communication theory of transmission of simple drawings.
In Communication Theory (ed. W. Jackson).
Butterworth, London.
- Loveless, N.E., Brebner, J. & Hamilton, P.
1970 Bisensory presentation of information.
Psychol. Bull. 73, No.3, pp.161-199.
- Lunn, G.J. & Mumford, H.
1965 Television and the telephone.
A.T.E. Journal, 21, No.2, pp. 74-80.
- MacLay, H. & Osgood, C.E.
1959 Hesitation phenomena in spontaneous English speech.
Word, 15, pp.19-44.
- Marsh, A.H.
1967 Perception of visual form in intensely noisy television
images. Ph.D. Thesis, University of London.

- McGrath, J.J. & Hatcher, J.F.
1961 The accuracy of judgements of eye fixations: a methodological note on vigilance. Studies of Human Vigilance: An omnibus of technical papers. Human Factors Research Incorporated, Los Angeles, California 90019. Technical memorandum 206-13.
- Martine, W.H.
1931 Rating the transmission performance of telephone circuits. B.S.T.J., 10, p.116.
- Head, G.H.
1934 Mind, Self and Society. University of Chicago Press.
- Mehrabian, A.
1963 The inference of attitudes from the posture, orientation, and distance of a communication. J. Consult. Psychol., 32, pp.296-308.
- Mehrabian, A., & Williams, M.
1969 Nonverbal concomitants of perceived and intended persuasiveness. J. Person. Soc. Psychol. 13, No.1, pp.37-58.
- Molnar, J.P.
1969 Tomorrow. Bell Lab. Record., 47, No.5, pp.186-187.
- Morley, I.E. & Stephenson, G.M.
1969 Interpersonal and interparty exchange; a laboratory simulation of an industrial negotiation at the plant level. Br. J. Psychol., 60, No.4, pp.543-544.
- Moscovici, S.
1963 Communication processes and the properties of language. In - Advances in Experimental Social Psychology, (ed. L. Berkowitz)., Vol.III, Academic Press, New York.
- Mounts, F.W.
1969 A video encoding system with conditional picture element replenishment. B.S.T.J., 48, p.2545.
- Newcomb, T.M.
1953 An approach to the study of communication acts. Psychological Review, 60, pp.393-404.
- Nast, D.W. & Welber, I.
1969 Transmission across town or across the country. Bell Lab. Record., 47, No.5, pp.186-187.

- Norwine, A.C. & Murphy, O.J.
1938 Characteristic time intervals in telephonic conversation.
B.S.T.J., 17, pp.281-291.
- Osgood, C.E.
1966 Dimensionality of the semantic space for communication
via facial expression. Scand. J. Psychol., 7, pp.1-30.
- Pan, J.
1962 Reduction of information redundancy in pictures.
Sc.D. Thesis, M.I.T. Elec. Engrg. Dept., Cambridge, Mass.
- Pearman, D. & Iddles, H.A.
1968 Factors affecting satisfactory sound-with- vision
conference facilities. 4th International Symposium
on Human Factors in Telephony, Bad Wiessee, Germany.
- Pearson, D.
1965 Fidelity criteria for visual image transmission over
noisy communication channels.
Ph.D. Thesis, Imperial College, University of London.
- Pickett, J.M.
1963 Transmitting speech sounds by a Tactual Vocoder and
lip reading. Report No.27, The Royal Inst. of
Technology, Stockholm, Sweden.
- Pierce, J.R. & Karlin, J.E.
1957 Reading rates and the information rate of a human
channel. B.S.T.J., 36, pp.
- Pollack, I.
1964 Interaction between auditory and visual information
sources in word identification.
Lang. & Speech., 7, pp.76-84.
- Porter, E., Argyle, M. & Salter, V.
1970 What is signalled by proximity.
Perceptual and Motor Skills, 30, pp.39-42.
- Pratt, W.K.
1967 A bibliography on television bandwidth reduction studies.
I.E.E.E. Trans. Info. Theory. IT-13, No.1, p.114.
- Purton, R.F. & Lunn, G.J.
1964 An experimental 1.6 Mb/s PCM video telephone system for
use on audio cables. I.E.E. Conf. Transmission
aspects of communication networks. 24th February.

- Reid, A.A.L.
1970 Comparisons between telephone and face-to-face conversation. 5th International Symposium on Human Factors in Telecommunications. London.
- Reid, A.A.L.
1971 Communications Studies Group, University College, London.
- Richards, D.L.
1964 Statistical properties of speech signals. Proc. I.E.E., III, Part 3, pp.941-949.
- Richards, D.L.
1967 Transmission performance of telephone connexions having long propagation times. HET PTT-Bedrijf, English Edition, 15, pp.12-24.
- Richards, D.L.
1969 Characteristics of speech level meters of the indicating instrument type. Internal Report, P.O. Research Station, Dollis Hill, London.
- Richards, D.L.
1970 Communication by speech - the important factors in telephony. The Inst. of Post Office Elec. Engs.
- Richards, D.L. & Hutter, J.
1968 Communication efficiencies of speech links : Possibilities of automatic measurement. 4th International Symposium on Human Factors in Telephony. Bad Weissee, Germany.
- Robinson, A.H.
1969 Automatic digital encoding for bandwidth reduction in visual image transmission systems. Ph.D. Thesis, Imperial College, University of London.
- Rosenfield, A.
1968 Bandwidth reduction bibliography. I.E.E.E. Trans. Info. Theory, IT-14, p.601.
- Rosenthal, R.
1963 On the social psychology of the psychological experiment. Amer. Scientist, 51, No.2, p.268.
- Ruesch, J.
1966 Nonverbal language and therapy. Psychiatry, 18, pp.323-330.
- Sakai, T., Nagao, M. & Fujibayashi, S.
1969 Line extraction and pattern detection in a photograph. Pattern Recognition, 1, pp.233-248.

- Sarbin, T.R. & Hardyck, C.D.
1955 Conformance in role perception as a personality variable. J. consult. Psychol., 19, pp.236-241.
- Saraga, P. & Wollons, D.J.
1968 The design of operators for processing. I.E.E. N.P.L. Conference on Pattern Recognition. London.
- Sartre, J-P.
1943 L'etre et le Neant. (1956) Being and Nothingness, trans. H. Barnes, London : Methuen.
- Schefflen, A.E.
1963 Communication and regulation in psychotherapy. Psychiatry, 26, pp.126-136.
- Schefflen, A.E.
1964 The significance of posture in communication systems. Psychiatry, 27, pp. 316-321.
- Schefflen, A.E.
1965 Quasi-courtship behavior in psychotherapy. Psychiat., 28, pp.245-257.
- Schlegloff, E.
1967 The first five seconds : the order of conversational openings. Ph.D. Thesis, Columbia, New York.
- Schlosberg, H.
1952 The description of facial expression in terms of two dimensions. J. exp. Psychol., 44, pp.229-237.
- Schlosberg, H.
1954 Three dimensions of emotion. Psychol. Rev., 61, pp.81-88.
- Schreiber, W.F.
1956 The measurement of third order probability distribution of television signals. I.R.E. Trans. on Info. Theory, IT-2, pp.94-105.
- Schreiber, W.F.
1967 Picture Coding. Proc. I.E.E.E., 55, No.3, pp.320-330.
- Sezaki, N. & Katagiri, H.
1965 Character recognition by follow method. Proc. I.E.E.E., 53, p.510.
- Siegel, S.
1956 Nonparametric statistics. McGraw-Hill, New York.

- Smith, D.H.
1969 Communication and negotiation outcome.
J. of Communication., 19, pp.248-255.
- Sodolski, J.
1969 Comments on docket No. 18397 Part 5. Electronic Industries Association. Federal Communications Commission : Inquiry into the development of communications technology and services to formulate regulatory policy and rulemaking and/or legislative proposals.
- Sommer, S.
1959 Studies in personal space. Sociometry, 22, pp.247-260.
- Sommer, S.
1961 Leadership and group geography.
Sociometry, 24, pp. 99-109.
- Sommer, S.
1962 The distance for comfortable conversation.
Sociometry, 25, pp.111-116.
- Sommer, S.
1965 Further studies of small group ecology.
Sociometry, 28, pp. 337-348.
- Stafford, E. & Hansel, C.E.M.
1953 The impression of depth in projected photographs.
Brit. J. Psychol., 44, pp.127-131.
- Starkweather, J.A.
1956 The communication value of content free speech.
Am. J. Psychol., 69, pp. 121-123.
- Stephenson, G.M. and Rutter, D.R.
1970 Eye-contact, distance and affiliation : a re-evaluation.
Brit. J. Psychol., 61, No.3, pp.385-393.
- Stokes, R.R.
1969 Human factors and appearance design considerations of the Mod II Picturephone station set.
I.E.E.E. Trans. Comm. Tech., COM-17, No.2, pp.318-323.
- Strongman, K.T. & Champness, B.G.
1968 Dominance hierarchies and conflict in eye contact.
Acta. Psychologica, 28, pp.376-386.

- Sunby, W.H. & Pollack, I.
1954 Visual contribution to speech intelligibility in noise.
J. Acoust. Soc. Am., 26, No.2, pp.212-215.
- Sziklai, G.C.
1956 Some studies of the speed of visual perception.
I.R.E. Trans. on Info. Theory, IT-2, p.125.
- Taylor, W., Pearson, J.C.G., Kell, R. & Mair, A.
1967 A pilot study of hearing loss and social handicap in female jute weavers.
Proc. of the Royal Soc. of Medicine, 60, pp.1117-1121.
- Thayer, S.
1969 The effect of interpersonal looking duration on dominance judgements.
J. soc. Psychol., 79, pp.285-286.
- Thompson, J.E.
1968 Reduction of perceptual redundancy for data compression of television signals.
Ph.D. Thesis, Imperial College, University of London.
- Thorton, G.R.
1944 The effect of wearing glasses upon judgements of personality traits of persons seen briefly.
J. appl. Psychol., 28, pp.203-207.
- Treisman, M. & Howarth, C.J.
1959 Changes in threshold level produced by a signal proceeding or following the threshold stimulus.
Quart. J. Exper. Psychol., 11, pp.129-142.
- Turling, E., Mandler, G. & Bauml, R.
1965 Interaction of two sources of information in tacistoscopic word recognition.
Canadian J. Psychol., 18, p.62.
- Valensi, G.
1969 Colour videophone.
Telecommunications J. (I.T.U.), 36, No.3, pp.
- Van de Geer, J.P. & Levelt, W.J.
1963 Detection of visual patterns disturbed by noise.
Quart. J. of Exp. Psychol., 15, No.3, pp.192-204.
- Van Doorn, A.C.
1966 The plumbicon compared with other television camera tubes.
Philips Tech. Rev., 27, No.1, pp.1-14.

- Veniar, F.A.
1948 Difference thresholds for shape distortion of geometrical squares. J. Psychol., 26, pp.461-476.
- Vernon, M.D.
1953 Perception and understanding of instructional television programmes. Brit. J. of Psychol., 44, pp.116-126.
- Vernon, M.D.
1962 Psychology of Perception. Penguin, London.
- Vine, I.
1970 Communication by facial-visual signals. In - Social Behaviour in Birds and Mammals. (ed. J.H. Cook). Academic Press, 1970.
- Watson, O.M. & Graves, T.D.
1966 Quantitative research in proxemic behavior. Am. Anthropol., 68, pp.971-985.
- Weaver,
1958 The design of a linear phase-shift low-pass filter. B.B.C. Eng. Division Monograph, No.27.
- Wetherill, G.B. & Levitt, H:
1965 Sequential estimation of points of a psychometric function. Brit. J. of Math. and Statist. Psychol., 18, No.1, pp.1-10.
- Winer, B.J.
1962 Statistical Principles in experimental design. McGraw-Hill, New York.
- Woodworth, R.S.
1938 Experimental Psychology. Holt, New York.

Further References

- Dornbush, R.L.
1968 Shadowing in bisensory memory. Quart. J. of Exp. Psychol. 20, pp.225-231
- Exline, R.V.
1963 Explorations in the process of Person Perception: visual interaction in relation to competition, sex, and need for affiliation. Journal of Personality, 31, pp.1-20
- Goldberg, G., Kiesler, C.A., & Collins, B.E.
1969 Visual behaviour and face-to-face distance during interaction. Sociometry, 32, pp.45-53.