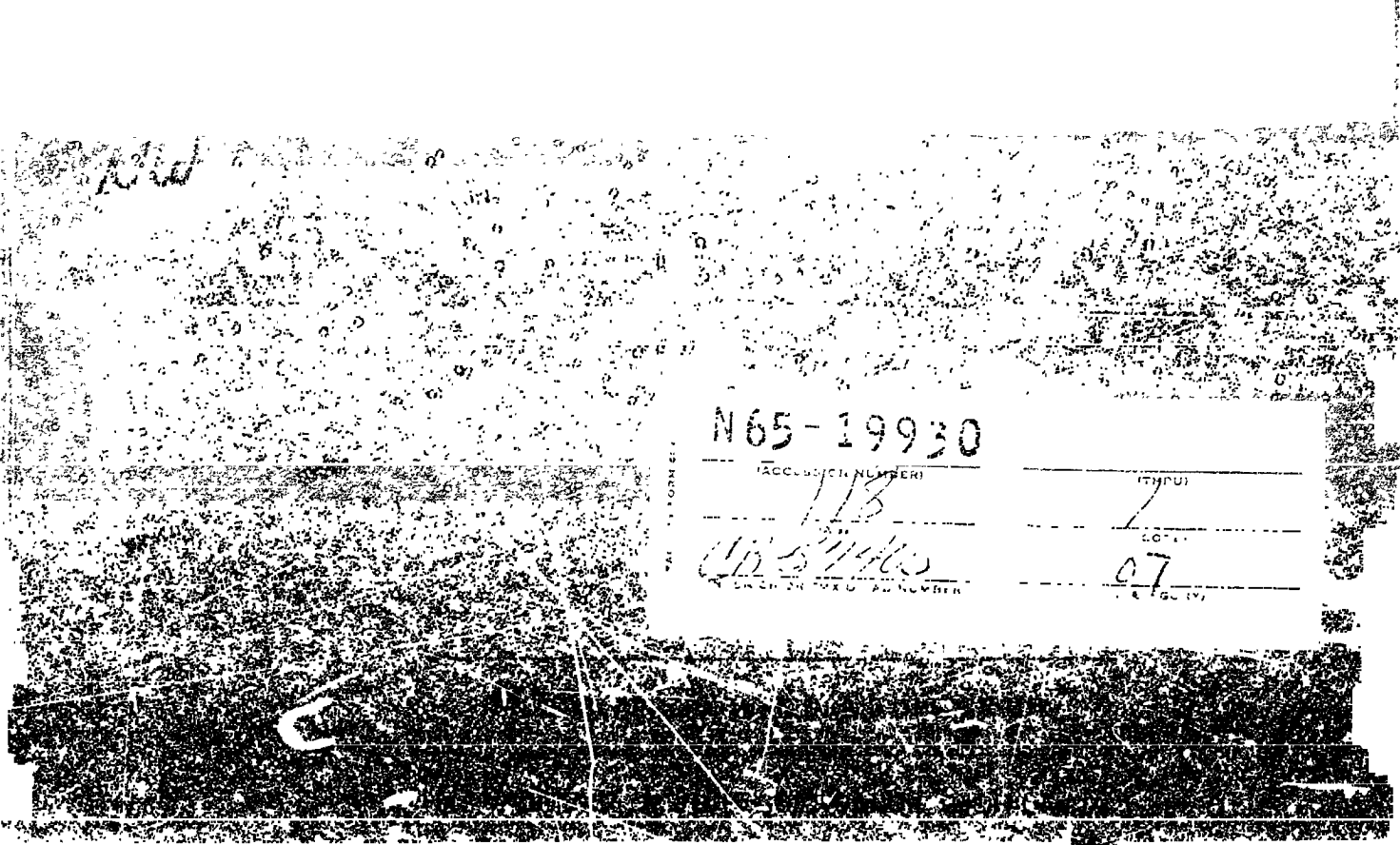


20 March 1964

STUDY ON SPEECH COMPRESSION

FOR

MANNED SPACE VEHICLES

Contract No. NASw - 709

Prepared For
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Washington 25, D.C.

Research and Systems Center
General Precision, Inc.
Librascope Group
Glendale 1, California

CONTENTS

CONTENTS (Continued)

ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

This report is submitted in accordance with the requirements of Contract NASw-700 with National Aeronautics and Space Administration. It is a final technical report, and covers the work carried out during a period of nine months in 1963-4. The program, which lies in the general area of speech processing, was monitored by Dr. John M. Walker, Chief, Communication and Tracking, and Mr. Theodore Tsacoumis, Engineer, Communication and Tracking, Code RET, NASA Headquarters, Washington, D.C.

The bulk of our work was devoted to two tasks: a survey of the present state-of-the-art in the field of speech compression, and a search for new techniques which might lead to improved speech compression capability. We concluded that presently available techniques do not approach the theoretical limits, and several new techniques were uncovered which appear to offer promise.

The problem of speech compression per se is only a part of the larger and more general problem of language processing for the purpose of facilitating communication between two systems. In the case of speech compression, the two systems are (ultimately) two people, and communication is effected by means of accoustic speech signals. Communication is facilitated by making use of (electrical) substitute signals which occupy less band-width than the direct electrical analogues of the original speech signals, and which have a smaller dynamic range. The substitute signals are derived from the original acoustic speech signals, and are of such a nature as to permit synthesis of satisfactory replicas of the original signals. In the more general case of language processing, one, or both, of the systems may be non-human -- or even inanimate (a computer, for example). The original signals include (acoustic) speech signals and may also include (non-acoustic) para-speech signals. Communication may be facilitated, as in the case of speech compression, by minimizing the load, in some sense, on an associated communication sub-system. Alternatively

for example, communication may be facilitated by deriving a set of substitute signals which are chosen so as to be easily recognizable by the receiving system. Throughout the course of the work under this contract, we have attempted to consider the problem of speech compression in the context of the broader problem of language processing.

2. SCOPE OF THE PROGRAM

2.1 Objectives Under the Present Contract

2.1.1 The work called for in Contract NASw-700 was divided into three tasks. Paraphrasing the language used in Article II (Description of Work) of the contract, these are described briefly as follows:

Task A Establish criteria for the evaluation of speech compression techniques.

Task B Analyze the speech compression problem from a technical standpoint.

Task C Survey the current state-of-the-art and make recommendations for future work in the field of speech compression.

In all cases, the speech compression problem was to be considered in the context of a requirement for voice communication from a manned spacecraft to the earth. On the basis of post-contract verbal discussions with NASA, consideration was limited to spacecraft missions within the solar system.

2.1.2 Our first effort was a brief examination of the overall problem with which we were faced in our consideration of speech compression. We discuss this problem, and set forth some of our conclusions, in Subsection 2.2. In the light of these conclusions, we devoted most of our subsequent effort to a survey of the current state-of-the-art (in the field of speech compression) and to a search for possible new techniques. Relatively minor effort was devoted to consideration of the special problems of the communications subsystem, to development of criteria for evaluating speech compression techniques, and to a limited amount of experimental work.

2.1.3 The results we obtained are presented briefly in Section 3. More detailed discussions are set forth in Section 4, Present State-of-the-Art,

Section 5, Current Trends in Research, Section 6, Experimental Work Carried Out Under the Contract, and Section 7, New Avenues for Future Work.

2.2 The Overall Problem

2.2.1 If we postulate the existence, in the near future, of manned spacecraft at interplanetary distances from the earth, it is obvious that there will be a concomitant need for man-to-man communication from such spacecraft to the earth. It is pertinent, then, to ask why speech should be a preferred mode of communication under these conditions. Three rather obvious answers are:

a) The use of speech requires no special training in the use of the communications subsystem. Such training is required, for example, if communication involves the use of a telegraph key or a teletype keyboard.

b) The use of speech is efficient in the sense that the speaker can communicate while simultaneously carrying out other tasks.

c) The availability of an easy mode of communication will contribute to the psychological well-being of the occupant of a spacecraft.

Considerations (a) and (c) imply the desirability of using natural language instead of some artificial language, which might be superior when judged by a limited set of criteria, or some stylized version of natural language. On the basis of the foregoing cursory examination, we take it that the basic problem is one of finding techniques which will permit communication (over interplanetary distances) by means of natural language.

2.2.2 Having posed the basic problem, we now ask what constitutes an acceptable solution. We have, rather arbitrarily, settled on three performance standards which are as follows:

a) Minimum Performance. A minimum performance system is one which delivers the basic information, in the semantic sense, contained in its natural language input. By basic information, we mean the information contained in an unpunctuated written transcript of the spoken input, or the information contained in a spoken version of such a transcript when it is read without expression other than pauses as indicated by spaces in the text. We estimate that the output of a minimum performance system will be a slightly degraded version of good telephone speech. It is to be noted that the degradation occurs at the semantic level, because of the loss of all nuances of stress and intonation which make up expression, rather than at the phonetic level.

b) Intermediate Performance. An intermediate performance system is one which delivers all of the information, again in a semantic sense, contained in its natural language input. This includes all of the nuances which are dropped by a system of minimum performance. The output of an intermediate-performance system will be a high-fidelity reproduction of its acoustic input, and the system will provide a one-way channel of communication equivalent to that available to a blindfolded listener in a face-to-face conversation.

c) Maximum Performance. A maximum performance system is one which delivers, in addition to a high-fidelity reproduction of its acoustic input, a display of its (non-acoustic) para-speech inputs. Such inputs may be obtained from a variety of sensors attached to

the speaker (sensors for skin resistance, pulse rate, potentials associated with facial muscles, facial expression, lip position, and the like). The purpose of the displays associated with the para-speech inputs is to provide, without going as far as a television link, those non-acoustic signals which are available in a face-to-face conversation.

2.2.3 We now ask what minimum requirements the foregoing systems impose on the associated communications subsystem (the radio link). These requirements can be stated in terms of bits/second as shown in Table 2-1.

Table 2-1

System Requirements in Bits/Second

System	Requirements
Minimum Performance	90 bits/second
Intermediate Performance	135 bits/second
Maximum Performance	200 bits/second

The heuristic arguments which justify these numbers are given in Appendix "A" for the minimum performance system and in Appendix "B" for the other two systems. These figures, of course, describe the transmission load which is presented to the communications subsystem, and have nothing to do with the information content (in either the semantic or the information-theoretic sense) of the messages which constitute the transmission load, except that the information content, in the information-theoretic sense, cannot exceed the transmission load.

In each case, the numbers are derived on the basis of brute-force coding. More sophisticated coding, which takes account of the relative frequencies of occurrence of the different input symbols, would probably effect reductions of the order of 30 percent. A further reduction could certainly be achieved by taking account of the inter-symbol constraints, and it is probable that the numbers in Table 2-1 are high by a factor of two. They will serve, nonetheless, as approximate values for

the theoretical lower limits of the transmission loads associated with the three sorts of communications systems described in 2.2.2.

2.2.4 In Table 2-1 in the preceding paragraph, we gave approximations for the theoretical lower limits of the transmission loads associated with communications systems having minimum, intermediate, and maximum performance. Attainment of these limits involves, by definition, the use of optimum speech processing. We now ask what are the transmission loads if the communications system uses speech which has undergone no processing except simple bandlimiting. As a first step toward answering this question, we specify three grades of speech: poor, good, and excellent.

a) Poor speech is about what is delivered by a rather bad intercom system.

b) Good speech is what is delivered by a good telephone system.

c) Excellent speech is what is delivered by a good high-fidelity system adjusted for speech.

The characteristics of these three grades of speech are derived on an heuristic basis in Appendix "C" and are repeated in Table 2.2.

Table 2-2

Characteristics of Three Grades of Speech

Poor	500 - 1500 cps	+ 10 db S/N
clipping at 2.2 times RMS level,	2^2 levels,	4,000 bits/second
Good	300 - 3300 cps	+ 25 db S/N
clipping at 3.1 times RMS level,	2^5 levels,	30,000 bits/second
Excellent	90 - 9180 cps	+ 60 db S/N
clipping at 3.5 times RMS level,	2^{11} levels,	200,000 bits/second

2.2.5 Poor speech is adequate for routine communication under conditions where it is easy to interrupt the speaker with "say again" when something is not understood. Although it is possible that a manned space mission might be considered routine, the possibility of interruption does not exist at interplanetary distances where the transit time for radio signals is measured in minutes or even tens of minutes. Accordingly, we drop this class of speech from further consideration.

Good speech offers something between minimum performance and intermediate performance, as described in paragraph 2.2.2, and excellent speech offers either intermediate performance or something just a little lower. The transmission loads for these grades of speech and for these levels of performance are restated in Table 2-3.

Table 2-3

Transmission Loads for Two Grades of Speech and Performance Levels

Excellent Speech	200,000 bits/second
Good Speech	30,000 bits/second
Intermediate Performance	135 bits/second
Minimum Performance	90 bits/second

We can equate excellent speech with intermediate performance. On the basis of the discussion in this sub-section, we conclude that optimum processing of excellent speech will yield a compression ratio of $\frac{200,000}{135}$, or about 1500, with no loss in performance.

Good speech offers something between minimum performance and intermediate performance. We conclude, therefore, that optimum processing of good speech will yield a compression ratio between $\frac{30,000}{135}$, or 220, and $\frac{30,000}{90}$, or 330, with no loss of performance. Since good speech is far closer to minimum performance than it is to intermediate performance, we suggest 300 as a working value of compression ratio.

2.2.6 Having derived theoretical limits for the compression ratios available through optimum speech processing, we now ask how closely present speech processing systems approach these limits. Present systems handle good speech rather than excellent speech, and the applicable limit is 300 rather than 1500.

Our first major effort was devoted to a survey of the current state-of-the-art. This survey, which is discussed in detail in Section 4, established the fact that present techniques yield compression ratios in the range from 10 to 30.

2.2.7 The basic problem, as developed in paragraph 2.2.1, is that of finding techniques which will permit communication over interplanetary distances by means of natural language. In view of the disparity between presently available speech compression ratios and the theoretical limit attainable by optimum processing, we conclude that the immediate problem is one of discovering new techniques of speech processing which offer some hope of approaching the theoretical limit. On the basis of this conclusion, which we reached about half-way through the program, the second major effort was devoted to a search for such techniques. We turned up a number of techniques which, it is believed, deserve study. These techniques are discussed in detail in Section 7.

3. WORK ACCOMPLISHED UNDER THE PRESENT CONTRACT

3.1 Preliminary Study

3.1.1 Considering only the problem of Speech Compression, we completed a brief study aimed at establishing theoretical limits for the compression ratios available through optimum speech processing.

a) We established the desirability of natural language as a medium of communication from a manned spacecraft to the earth.

b) We formulated qualitative descriptions of the performance of three grades of communications systems, using natural language, and estimated the minimum transmission load, in bits/second, for each.

c) We formulated qualitative descriptions of three grades of unprocessed speech, and estimated the band-width, signal-to-noise ratio, and transmission load for each.

d) We arrived at theoretical limits for the compression ratios attainable through optimum speech processing.

e) We determined, on the basis of a brief survey, that current speech processing techniques do not offer compression ratios which approach the theoretical limit.

3.1.2 The details of this study are discussed at length in sub-section 2.2, and need not be repeated here. The conclusions, however, probably deserve restatement.

a) It appears that optimum processing of high-fidelity speech should yield a compression ratio of about 1500 with no loss of quality.

b) It appears that optimum processing of good telephone speech should yield a compression ratio of about 300 with no loss of quality.

c) Current speech processing techniques yield compression ratios ranging from about 10 to about 30.

d) In view of the wide disparity between current achievements and theoretical limits, it seems entirely reasonable to look for new techniques which might offer drastic improvement in compression ratio.

3.2 The Need for Speech Compression

3.2.1 Having arrived at some idea of how much speech compression is currently available, and having estimated how much can be achieved through optimum processing, we now ask whether or not compression is useful for the missions we have under considerations.

3.2.2 Minimum and maximum operating ranges for the various missions are calculated in Appendix "D". The results, expressed in Astronomical Units (one AU is the mean distance from the earth to the sun) are given, after rounding-off, in Table 3-1.

Table 3-1

Maximum and Minimum Operating Ranges for Various Missions

Mission	Distances to Earth	
	Min.	Max.
Mercury) Venus) Mars)	0.27	2.7
Jupiter) Saturn)	4.0	11.0
Uranus) Neptune) Pluto)	17.0	50.0

3.2.3 We now calculate the power required to send unprocessed speech in analogue form, assuming the use of single-side-band, suppressed-carrier, amplitude-modulation transmission and synchronous detection with image suppression. If no other signal from the spacecraft is available and from which the carrier can be constructed for re-insertion at the receiving end, a vestigial carrier can probably be transmitted for this purpose. The power involved would be small and is omitted from our calculations.

In Appendix "E" we estimate the transmitter output necessary to develop a signal-to-noise ratio of unity (at the receiver) over a band-width of one cps at a distance of one AU. This is

$$20 \text{ milliwatts} = -17 \text{ dbw} \tag{3-1}$$

using current techniques, or

$$10 \text{ microwatts} = -50 \text{ dbw} \tag{3-2}$$

using 1970 techniques. (dbw is db relative to one watt.)

To arrive at the required transmitter output, we multiply the basic power level by the product of the audio band-width, the desired signal-to-noise ratio (with synchronous detection and image suppression, the post-detection and pre-detection signal-to-noise ratios are identical), and the square of the distance in AU.

Band-widths and signal-to-noise ratios are taken from Table 2-2 and repeated in Table 3-2, rounded off to the nearest db.

Table 3-2

Characteristics of Three Grades of Speech

Speech Quality	Band-Width in db Relative to one cps	S/N in db
Poor	+30	+10
Good	+35	+25
Excellent	+40	+60

The squares of the distances given in Table 3-1 are converted to db and given in Table 3-3, rounded off to the nearest db.

Table 3-3

Mission Distances in Decibels

Mission	Distances to Earth	
	Min.	Max.
Mercury)	-11	+9
Venus)		
Mars)		
Jupiter)		
Saturn)	+12	+21
Uranus)		
Neptune)	+25	+34
Pluto)		

The results are summarized, for current techniques, in Table 3-4. The required transmitter outputs are given in dbw.

Table 3-4

Analog Transmission, 1963

Mission	Speech Quality					
	Poor		Good		Excellent	
	Min	Max	Min	Max	Min	Max
Mercury)						
Venus)						
Mars)	+12	+32	+32	+52	+72	+92
Jupiter)						
Saturn)	+35	+44	+55	+64	+95	+104
Uranus)						
Neptune)	+48	+57	+68	+77	+108	+117
Pluto)						

It is clear from this table that if we are restricted to reasonable power levels and the use of current techniques for analogue-transmission of unprocessed speech, the best we can do is to provide poor speech over the shorter distances associated with the Mercury/Venus/Mars mission.

The corresponding power levels using 1970 techniques are given in Table 3-5.

Table 3-5

Analog Transmission, 1970

Mission	Speech Quality					
	Poor		Good		Excellent	
	Min	Max	Min	Max	Min	Max
Mercury)						
Venus)						
Mars)	-21	-1	-1	+19	+39	+59
Jupiter)						
Saturn)	+ 2	+11	+22	+31	+62	+71
Uranus)						
Neptune)						
Pluto)	+15	+24	+35	+44	+75	+84

By 1970, the useful range of a poor-speech system (which we consider to be unsatisfactory) will be extended, but good speech will be limited to the Mercury/Venus/Mars mission. Excellent speech will still be out of range, unless we use excessive power.

3.2.4 Some saving in power can be effected if we digitize the speech signal (with no speech processing). We show, in Appendix "F", that we can transmit a bit-stream at a rate of C bits/second with C times the power required to develop a signal-to-noise ratio of one over a band-width of one cps. Using the basic power figures, multiplying by the square of the distance and by the bit-rate we have the required transmitter power The square of the distance, in db, is given in Table 3-3. The required bit-rate for various grades of unprocessed speech is available in Table 2-2 and is repeated in db relative to one bit/second in Table 3-6.

Table 3-6

Characteristics of Three Grades of Speech

Quality	Bit-rate in db
Poor	+36
Good	+45
Excellent	+53

Using current techniques, the transmitter power in dbw is given in Table 3-7.

Table 3-7

Digital Transmission, 1963

Mission	Speech Quality					
	Poor		Good		Excellent	
	Min	Max	Min	Max	Min	Max
Mercury)						
Venus)						
Mars)	+ 8	+28	+17	+37	+25	+45
Jupiter)						
Saturn)	+31	+40	+40	+49	+48	+57
Uranus)						
Neptune)						
Pluto)	+44	+53	+53	+62	+61	+70

Using 1970 techniques, we have, as shown in Table 3-8:

Table 3-8

Digital Transmission, 1970

Mission	Speech Quality					
	Poor		Good		Excellent	
	Min	Max	Min	Max	Min	Max
Mercury)						
Venus)						
Mars)	-25	- 5	-16	+ 4	- 8	+12
Jupiter)						
Saturn)	- 2	+ 7	+ 7	+16	+15	+24
Uranus)						
Neptune)	+11	+20	+20	+29	+28	+37
Pluto)						

3.2.5 We are now in a position to attempt a justification for a program aimed at perfecting optimum speech processing techniques. We will consider only the case of digital transmission, since this uses less power than analogue transmission, and will use the figures given in Tables 3-7 and 3-8.

In Table 3-9, the power levels for good speech and excellent speech are shown using 1963 communication equipment. The upper figure in each box is based on current speech processing techniques, the lower is based on optimum techniques. It is assumed that current processing techniques offer a power saving of 10 db for good speech and that they are not applicable to excellent speech. We assume that optimum processing techniques offer power savings of 25 db for good speech and 32 db for excellent speech.

Table 3-9

Digital Transmission, 1963

Mission	Current Processing Optimum Processing			
	Speech Quality			
	Good		Excellent	
	Min.	Max.	Min.	Max.
Mercury)	+ 7	+27	+25	+45
Venus)	- 8	+12	- 7	+13
Mars)				
Jupiter)	+30	+39	+48	+57
Saturn)	+15	+24	+16	+25
Uranus)				
Neptune)				
Pluto)	+43	+52	+61	+70
	+28	+37	+29	+38

In calculating the power levels required with 1970 communication equipment, Table 3-10, we assume that current speech processing techniques will (by 1970) yield power savings of 20 db on good speech and 10 db on excellent speech. Optimum processing, of course, is still limited to power savings of 25 db and 32 db.

Table 3-10

Digital Transmission, 1970

Mission	1970 Processing Optimum Processing Speech Quality			
	Good		Excellent	
	Min.	Max.	Min.	Max.
Mercury)				
Venus)	-36	-16	-18	+ 2
Mars)	-41	-21	-40	-20
Jupiter)	-13	- 4	+ 5	+14
Saturn)	-18	- 9	-17	- 8
Uranus)	0	+ 9	+18	+27
Neptune)	- 5	+ 4	- 4	+ 5

In the context of 1963 communications equipment, Table 3-9, the argument for optimum processing is based on power saving. Excellent speech with optimum processing requires less power, by a factor of 14 db, than good speech with current processing techniques.

In the context of 1970 communications equipment and 1970 versions of current speech processing techniques, the power argument applies only to the Uranus/Neptune/Pluto mission. If, however, we ask for excellent speech rather than good speech, and if we have a multi-channel communications system, the 22 db advantage (excellent speech with 1970 processing versus excellent speech with optimum processing) becomes meaningful even for the Jupiter/Saturn mission.

Totally aside from the problem of man-to-man communications, we wish to point out that the development of optimum processing techniques will make it feasible to use natural language as a means of communication across the man-machine interface.

3.3 The Present State of The Art

3.3.1 We examined the present state-of-the-art, primarily through the medium of a literature search. No speech processing equipment is presently available in the operational sense. Current laboratory equipments appear to offer compression ratios of about 10 with reasonable intelligibility as measured on a standard PB (phonetically balanced) word list. The subject is discussed in detail in Section 4, Present State of The Art. The techniques of measuring intelligibility, including the use of PB word lists, are discussed in Appendix "G".

3.3.2 We made a search of recent and current journals in order to reach some idea of present research trends. This search was supplemented by a number of personal contacts with workers who are active in the field of speech processing. Most current work is ultimately oriented toward improvement of current speech processing devices, which are almost all based on the concept of transmitting information about the short-term power spectrum of the speech signal. There is a distinct trend, however, toward the concept of pattern recognition schemes which attempt to discern patterns in the time-varying speech signal. This subject is discussed in detail in Section 5, Current Research.

3.4 Development of Criteria

3.4.1 Performance criteria for speech processing systems seem relatively simple to set up. For a simple analog system, as shown in Figure 3-1, we observe that the transmitter power required is proportional to the product of band-width by signal-to-noise ratio, both measured at the input to the synthesizer. In the absence of the speech processing sub-system, the transmitter power required to achieve equal performance is proportional to the product of band-width by signal-to-noise ratio,

measured this time at the synthesizer output.

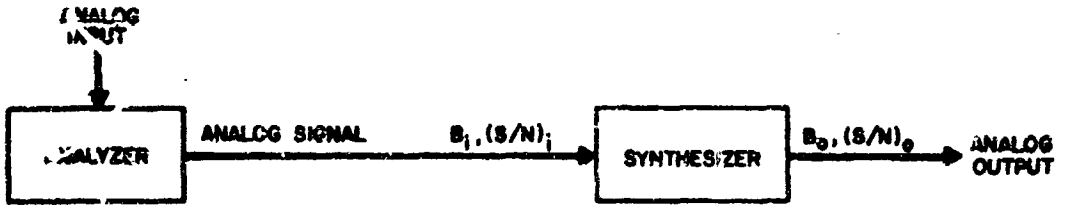


Figure 3-1. Simple Analog System

We have, therefore

$$R = \frac{B_o (S/N)_o}{B_i (S/N)_i} \tag{3-3}$$

where R is the compression ratio.

If the system makes use of a digital transmission link, as shown in Figure 3-2, we still use equation (3-3) as a definition of the compression ratio. B and (S/N) are measured, as before, at the (analog) input and output of the synthesizer.

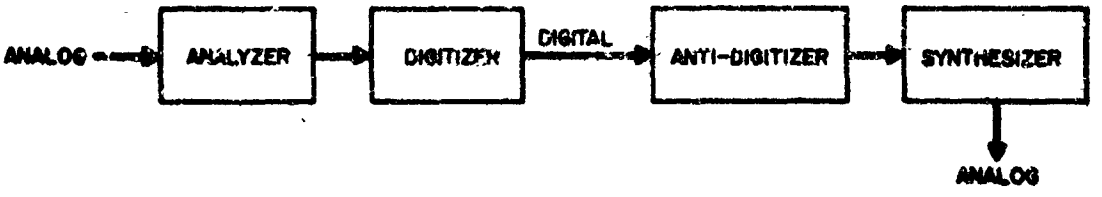


Figure 3-2. System with Digital Transmission Link

For a digital processing system, as shown in Figure 3-3, we define the compression ratio as

$$R = \frac{(\text{bit-rate})_o}{(\text{bit-rate})_i} \tag{3-4}$$

where the measurements are taken at the (digital) input and output of the synthesizer.

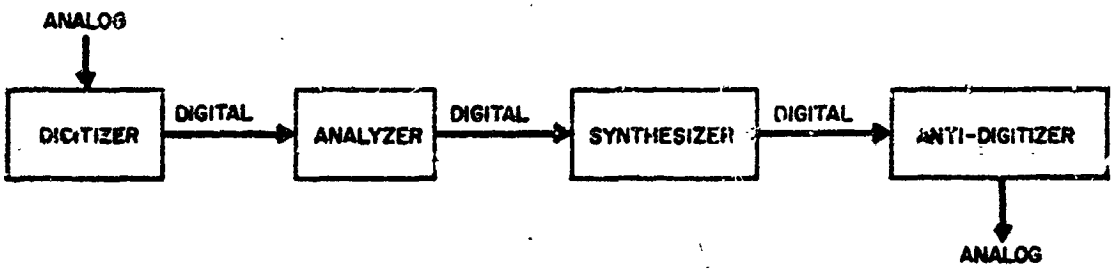


Figure 3-3. Digital Processing System

Finally, for a hybrid system, as shown in Figure 3-4, we use equation (3-4). We derive an equivalent value of $(\text{bit-rate})_o$ by starting with B_o and $(S/N)_o$ and using the method of Appendix "C".

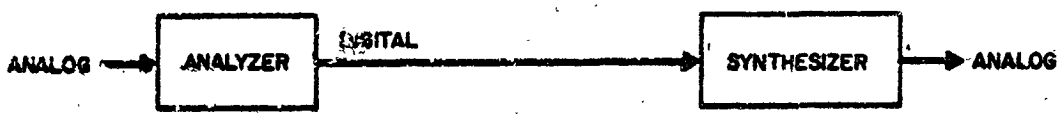


Figure 3-4. Hybrid System

3.4.2 Making use of the theoretical limits for the compression ratio, derived in sub-section 2.2, we can define a compression efficiency

$$\eta = R/R_L \tag{3-5}$$

where R_L is the limiting compression ratio. Finally, we can set up a figure of merit for performance

$$M_p = \eta I_o / I_i \qquad (3\text{-}6)$$

where I is an intelligibility index of some sort and I_o and I_i are measured, respectively, at the system output and input. M_p will then have a value of unity for a perfect system. The question of how I should be measured is discussed in Appendix "G".

3.4.3 M_p, as defined in the previous paragraph, is a combined measure of power reduction and conservation of basic information content (as defined in paragraph 2.2.2a). We have not developed an index for describing the ability of a system to reproduce expression (as defined in the same paragraph). One possibility, however, lies in testing the ability of listeners to identify different speakers. The percentage of correct identifications using the output of the system can be divided by the corresponding percentage using the output of a high-fidelity audio link, or by the percentage achieved in a face-to-face situation with visual contact eliminated by means of a screen or a blindfold. The resulting ratio will serve as an index of conservation of expression.

3.4.4 We have not been successful in developing any useful hardware criteria. It is obvious that the prime desideratum is the reduction in lifted weight, providing that this can be achieved at negligible cost in terms of volume, reliability, maintainability, repairability, operational efficiency, training requirements, etc., etc. Successful speech compression will, to be sure, afford the possibility of a reduction in radiated transmitter power, and reduction in weight can be achieved either by reducing the transmitter output or by reducing the antenna size. It is difficult, however, to develop any meaningful trade-off. A reduction of 900 watts in required transmitter output is an exciting prospect, if it means that a 1000 watt transmitter can be replaced by one delivering only 100 watts. The shift from 2000 watts to 1100 watts is interesting but hardly exciting, and the shift from 10,000

watts to 9, 100 watts or from 50, 000 to 49, 100 is hardly worth mentioning. Accordingly, we suggest that the development of firm hardware criteria and trade-offs should be deferred until more is known about the overall communications requirement of a particular mission.

3.5 Other Investigations

We devoted a small amount of effort to a series of short studies, all of which are described in detail in Section 6, Experimental Work Carried Out Under the Contract. These studies are listed briefly as follows:

a) An examination of the recording and transcript of communications to and from Mercury flight of Walter Schirra, sub-section 6. 1.

b) A brief study of the utility of lip reading as a medium of communication, sub-section 6. 2.

c) A short experimental program, aimed at preliminary verification of the usefulness of a new speech-pattern recognition scheme, sub-section 6. 3.

d) A brief attempt to duplicate the work of G. W. Barton and S. H. Barton, and reported on in 13 December 1963 issue of Science magazine, in producing a new type of speech pattern which may be easier to recognize than a conventional sonogram, sub-section 6. 4.

3.6 New Avenues for Future Work

We feel that we can offer a significant contribution in this area. Briefly, it appears to us that the art of speech processing is now at a

point where it should be considered as a subject for interdisciplinary study. We feel that when scientists from one discipline have pushed a problem to an interdisciplinary boundary, as is the case here, significant advances can be achieved with relative ease by allowing people on the other side of the boundary to do a little pulling. Specifically, we believe that it is time to look at the whole problem of speech processing from the standpoints of information-processing, decision-making and pattern-matching. All of these disciplines are well developed, and each has an extensive body of theory to bolster a broad area of current practice. Our specific recommendations are discussed at length in Section 7, New Areas for Future Work, and are listed briefly as follows:

a) A study of the statistics of quantized sampled speech signals. If each quantizing level is considered as a symbol, in the information-theoretic sense, the speech signal is a message which is expressed as a string of symbols. The set of symbols constitutes a code which has a redundancy, in the information-theoretic sense, in excess of 99%. It should be feasible to devise a translation to a new code having less redundancy, thus affording a significant degree of compression, as measured in bits/second. The problem of devising a new code is almost entirely a problem in information processing, sub-section 7.2.

b) A search for easily-recognizable patterns in quantized sampled speech signals using a technique of pattern-discovery which does not depend on any preconceptions about what the pattern ought to look like, sub-section 7.3.

c) A mechanization of a phoneme recognizer, using a pattern-matching technique which has already been partly tested, sub-section 7.4.

d) A study of the use of non-scanning optical correlation techniques to recognize patterns (sonograms, for example) derived from speech signals, sub-section 7.5.

e) A study of the use of linguistic, para-linguistic, and para-speech cues as aids to pattern recognition in speech processing. This is an application of decision theory, sub-section 7.6.

f) A study of the application of linguistic constraints to the decision-making process in a speech synthesizer whose input has been corrupted by additive noise. To the extent that the synthesizer operates on non-redundant inputs, developed by an optimum analyzer, something of this sort will be mandatory, sub-section 7.7.

g) A study, from the standpoint of linguistics, of the information conveyed by "expression" and by para-speech, sub-section 7.8.

4. THE PRESENT STATE OF THE ART

4.1 General Considerations

In this section we describe present speech compression techniques which are in various stages of research and development. The ultimate speech compression system with a necessary channel capacity of 90 bits/sec is still a research problem with no immediate solution. As each system or compression technique is described, we discuss its limitations and associated problem areas. These problem areas point out an important consideration that must be recognized when we consider bandwidth compression schemes. The development of hardware follows basic speech research and helps to define the current problems of that research. Thus, the basic research performed at Haskin's Laboratory and elsewhere on vowel-consonant transition cues, has helped the formant vocoder to be a workable device in reproducing consonants, although the as yet incomplete knowledge of consonant structure has hindered completely adequate consonant synthesis. For this reason, although vocoders will never be entirely satisfactory until speech research has provided answers to puzzling acoustic problems, the attempt at early hardware developr is justified both by the actual success already achieved and by the problem areas in speech research uncovered and defined by such hardware development.

Table 4-1 indicates the scope of presently available speech compression techniques which are in the near-hardware or hardware stage, based on Stevens et al[1].

TABLE 4-1

Scope of Presently Available Speech Compression Techniques

Group	A	B	C
Range of Information Rate Bits/Sec	18,000 to 2,000	2,000 to 800	Below 800
Approximate Compression Ratio	2:1 to 15:1	15:1 to 30:1	Over 30:1
Techniques	Bandpass Filtering Amplitude Clipping Time Compression Spectrum Sampling Semi-Vocoder Channel Vocoder	Formant Vocoder	Spectrum Pattern Matching + Coding

4.2 Systems with Compression Ratios up to 15:1

4.2.1 The first four systems included under this heading have compression ratios ranging up to 5:1. They depend upon simple operations performed on the electrical speech signal in either the time or the frequency domain. They offer no real hope for future development, but within their limitations of compression ratio and intelligibility they are valid schemes already operational. They will be discussed only briefly.

Bandpass Filtering — Experiments have shown that with a 2000 cycle bandpass filter, intelligibility of P.B. words (see Appendix G) approaches 90%. The naturalness is somewhat less than telephone quality speech but is adequate. This system is used mostly as a reference for testing other systems.

Amplitude Clipping — This technique merely clips the speech above a specified amplitude level. Modest clipping has very little effect on intelligibility and this technique has been used to extend the range of radiotelephone transmitters by reducing the ratio of peak power to average power. It has been found that infinitely clipped speech, which reduces the signal to a succession of zero crossings, can give P. B. word intelligibility of 80% but produces an unnatural voice quality. Clioping can give compression ratios of the order of 3:1 at the expense of voice quality.

Time Compression — This is a general term for schemes involving the extraction of time samples of speech, the division of these samples in frequency and the formation of a continuous signal from the succession of samples. The signals are then replayed at the proper speed in order to reconstitute the speech. P. B. word intelligibility of better than 80% can be achieved, and the voice quality, although not good, is better than that of clipped speech. A compression ratio of 2:1 can be achieved. (It is interesting to note that a variation of this technique was used in some of the Russian manned orbital missions.) The inherent limitations of this technique are the small compression ratio and the somewhat complex timing problems.

4. 2. 2 The following two systems give compression ratios on the order of 3. 1 and provide very good speech intelligibility and fairly good voice quality.

Spectrum Sampling — The basic concept used here to achieve speech compression is that of being able to sample only portions of the frequency spectrum of the speech signal by the use of optimally placed filters. Variations of this simple scheme have been suggested and implemented in prototype form. One system which has been tested uses three 650 cps filters centered about optimum frequencies of 500, 1500 and 2500 cps, and provides P. B. word scores of nearly 90%. A six-filter system having a

total analog bandwidth of about 750 cps has been developed.[2] Other modifications have been suggested, such as moving the center frequencies of the filters in accordance with the short-time energy distribution of the speech spectrum. This technique provides a compression ratio of about 3:1 with good intelligibility.

The Vocoder — The vocoder, which is described below, derives control signals from its input and uses these to program a synthesizer which reproduces speech in a way that is an electrical analogue of human speech reproduction. A few simple perceptive mechanisms are used in vocoder systems. However no operational or projected hardware development has carried the use of either articulatory or perceptive criteria very far. The reason for this is that the basic knowledge of these mechanisms is not fully understood. This problem is treated both in the following discussion and in a more general form in Section 5. Channel Vocoders use compression ratios up to 15:1.

4.3 The Channel Vocoder and Semi-Vocoders

4.3.1 The following is a brief description in a simplified generic fashion of the channel vocoder. Many systems have been attempted with slight modification of the basic premises. These are described briefly here, and more detailed discussions are contained in many of the references in the bibliography. Figure 4-1 is a block diagram of a classical channel vocoder. It provides a bandwidth saving of the order of 10:1. One constraint of production (voiced, voiceless distinction) and one constraint of perception (intelligibility retained by short-time amplitude spectrum) are incorporated in its design. A set of contiguous bandpass filters $BP_1 \longrightarrow BP_N$, with envelope detectors and low pass filters, produces values of the short-time amplitude spectrum at discrete frequency points. This is, to a first approximation, the magnitude of the vocal transfer function plus constant features of excitation and radiation. A voltage proportional to the fundamental frequency is provided by the pitch extractor and is also transmitted

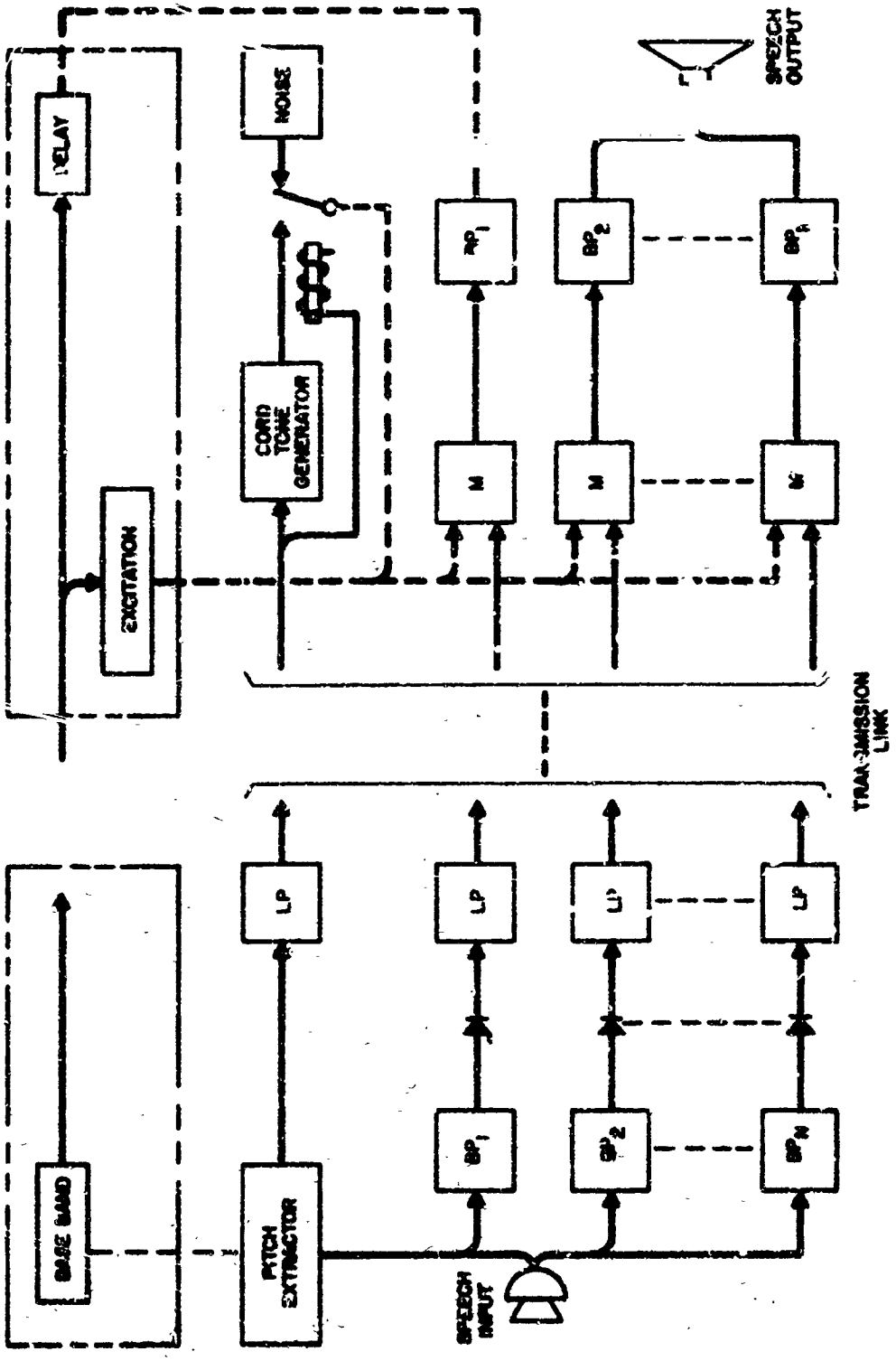


Figure 4-1. Channel Vocoder

via low pass filtering. At the receiver, this signal makes voiced-voiceless distinctions by modulating the frequency of a buzz-source generator or selecting noise for the excitation. (This is essentially a crude analog device for duplicating the human vocal mechanism.) The spectrum signals operate modulators at the respective inputs of a set of bandpass filters identical to those in the transmitter. The outputs are summed and the short-time spectrum is reconstructed.

4.3.2 What do vocoders really measure? Consider how we produce speech. The sound source may be the larynx, which produces a sequence of pulses as it converts the air flow from the lungs into speech-like sounds. The pulse repetition rate and the pulse shape are variable. We can also make sounds by having the air from the lungs impinge upon an obstacle like the teeth, or by forcing air through a constriction in the vocal tract to produce a "hissy" sound. This type of sound is random in character, as opposed to the general formant frequency character of the voiced larynx sound. The term formant frequency is an acoustic-phonetic term that refers to the frequency of the maximum of a gross concentration of energy in a speech sound, as indicated by a peak in the envelope of the spectrum. The formants are indicated by number with the lowest formant being identified as the first, etc. The two types of sound sources may operate singly or as a combination. The sound spectrum from these sources is then modified by the vocal tract cavities and by the articulators (tongue, lips, palate, etc.). These organs do not move rapidly, and hence the spectrum of the speech sound changes slowly.

4.3.3 The basic object of the vocoder (or of any other speech compression device) is to signal, in a more economical manner than by the reproduction of the total acoustic speech output, the significant parameters and information-bearing elements of the speech act. It should be noted that this does not preclude signaling of non-acoustic correlates of the speech act, such as position of the articulators, facial expressions, etc., as

information-bearing elements. As indicated by the foregoing description of the channel vocoder, two basic parameters of the speech act are contained in its implementation. The first is one of production, the voiced-unvoiced distinction, and the other essentially one of perception, and that intelligibility is generally maintained by preserving the short-time amplitude spectrum. Speech sound amplitudes change slowly within the bandpass filter outputs and therefore can be represented by the smoothed outputs of the envelope detectors. The phase spectrum is not transmitted as it appears to carry no usable intelligence to the ear.

4.3.4 The first vocoder was developed in 1939 by Homer Dudley at the Bell Laboratories. Early vocoders had an intelligibility score approaching 70% of P.B. words. In recent years, improvements of the original designs have provided intelligibility ratings of about 85% on P.B. words and a compression ratio of about 10:1.

4.3.5 In its original form, and even now, the channel vocoder is plagued by the "unnaturalness" of its sound output. Essentially, naturalness is a sensory dimension, and as such it is not well defined even today. The basic problem seems to be in tracking at the analyzer and reproducing at the synthesizer the crucial features of the excitation signal. One method of bypassing the problem is shown in Figure 4-1 in the box at the top of the figure. This configuration is called a semi-vocoder. Perhaps 1/3 to 1/2 of the ordinary speech signal is transmitted (the lower part) as an ordinary unprocessed baseband. After delay, it is recombined with the outputs of the higher vocoder channels, which are excited by noise. This eliminates the pitch extraction problem, but no voiced excitation can now be delivered to the higher bands, and the voiced-voiceless distinction remains. In addition, the compression ratio is drastically reduced.

4.3.6 A number of factors contribute to the pitch extraction problem. Among these are the presence of strong harmonics in the signal used to measure the fundamental frequency, and the smoothing inherent in most frequency measuring circuits which wipes out the cycle-by-cycle variations of the fundamental frequency which may be significant. Autocorrelation and cross-correlation techniques[3, 4, 5] have recently shown some promise of providing a reasonably satisfactory solution to these problems. However, research is still needed in this basic area to determine exactly what parameters of the fundamental frequency signal must be transmitted and to what extent these affect inteiligibility and naturalness. Any advances made in the general speech research areas of determining the critical parameters for perception of speech will improve the channel vocoder by indicating what portion of the transmitted speech signal must be emphasized.

4.4 The Formant Tracking Vocoder

4.4.1 Consider the number of output patterns that can be produced by a channel vocoder. If we have a 16 channel vocoder with amplitude quantizing to 8 levels, 2^{48} output patterns are possible. Now consider the fact that vowel-like sounds at least can be represented by the first three formants present in the acoustic output[6], and that the observer cannot perceive these formant frequency differences to less than 3 per cent at best.[7] If each formant frequency has a frequency range of 3.2 to one, it appears at one of 40 possible frequencies and at one of 8 possible levels. Then there are at most $(40 \times 8)^3$ or about 2^{25} possible vowel-sound patterns. This number is an estimate of the ability of the vocal cavity to produce, and of the observer to perceive vowel-like sounds. If one knew how to represent consonantal sounds as a complete set of acoustic output parameters, and if one knew the threshold for discrimination of such parameters,[6, 7] one could estimate the approximate

maximum number of patterns perceived by an observer. It is generally felt this number would be considerably less than the 2^{48} possible patterns produced by a channel vocoder. Furthermore, the actual number of perceived patterns is less than the theoretically possible number determined by this means, as the position of the formants of vowel-like sounds are limited roughly to certain regions and to certain relationships among themselves. At any rate, a speech compression device, more ambitious than the channel vocoder, and called a resonance vocoder or formant tracking vocoder, suggests itself from these principles. A discussion of this device follows.

4.4.2 Most of the recent attempts at speech compression techniques have centered around development of the formant or resonance vocoder. This is a more ambitious attempt at duplicating certain vocal cord functions as well as incorporating some aspects of the channel vocoder. Vowel sounds may be characterized in a general way by nearly constant spectral maxima which are called formants. These are represented in some models as being caused by poles (4 pole representations have proved to be reasonably adequate) and no zeros. The consonants, however, contain zeros as well as poles. It has been possible to develop equipment that can extract the formant frequency and amplitude for the vowels. As yet, it has been impossible to extract the poles and zeros of consonants from continuous speech. Figure 4-2 illustrates the formant vocoder, which can give a compression ratio of 20:1 or even greater. The choice of excitation source is the same as in the classic vocoder, but the speech spectrum is analyzed in order to extract the frequency and amplitude of the formants. It is often divided into three contiguous wide bands corresponding to the range of the formant frequencies, and two voltages are transmitted corresponding to the amplitude and frequency of each of the spectral maxima. At the synthesizer the frequency voltages tune formant resonators and the amplitude voltages modulate the inputs to the resonators. A similar type of hybrid system which transmits the

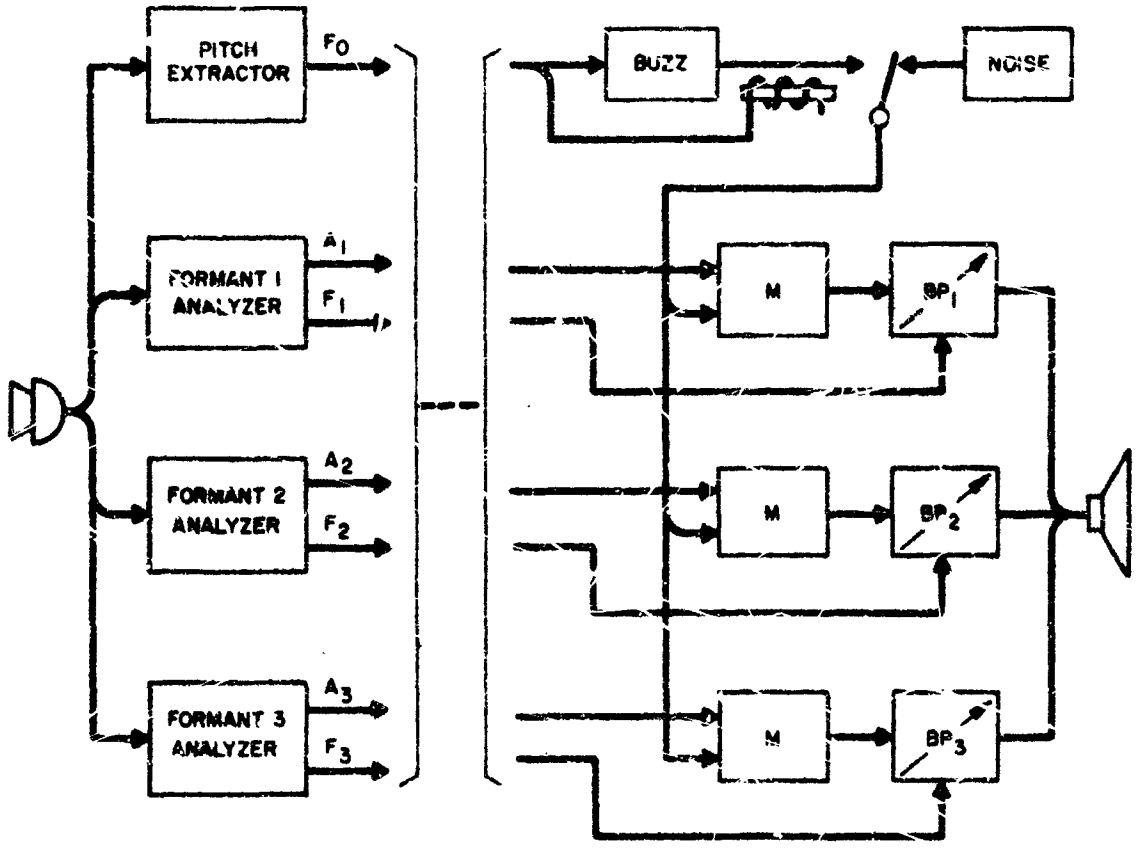


Figure 4-2. Formant or Resonance Vocoder

lower portion of the spectrum as an unprocessed baseband has been investigated and, as with the classical vocoder, improvement is obtained at the expense of bandwidth.

4.4.3 As is obvious from the foregoing discussion, the formant vocoder as well as the channel vocoder, is plagued by the problem of extracting and signaling the excitation function accurately. Beyond this common problem, the formant vocoder is a much more ambitious attempt to signal the information-bearing parameters of speech. Its compression ratio may be as high as 30:1. In order to achieve this much compression, a limited number of parameters must be extracted in the analyzer and then reconstructed into speech at the synthesizer Unfortunately these parameters are not completely defined, especially for most consonents. Vowels are fairly well described by the first three formants, and synthesizing networks have been constructed with can reproduce them to a high degree of intelligibility. In addition, certain fricative consonents such as /S/ and /ʃ/ have been synthesized with different networks than are used for vowels. However, a mathematical model adequate for all speech sounds is not yet available for use in a synthesizer-analyzer system. In fact, the type of synthesizer now used in formant vocoders may not, in general, be adequate to produce consonantal sounds, and it might be more convenient to synthesize these with a vocal tract analogue depending on articulatory constraints. This subject is being heavily researched at present at M.I.T.[8, 9, 10, 11, 12] The work at M.I.T. is part of a general research study in automatic speech recognition and is discussed in Section 5.

4.4.4 The present performance of formant vocoder systems is generally not agreed upon in the literature. What appear to be impartial tests[1, 13, 14] rate the formant vocoder performance very low (P.B. word articulation 33%), but on the other hand Contractor final reports[2, 4, 5]

show test results of the order of 75 to 85% P. B. word intelligibility, (See Appendix G). Further comments on reference (9)[15] indicate that performance may be better than is indicated in that survey. At any rate, certain problems existing in such equipments have not been solved. The chief among these being the basic problem of signaling the significant parameters of certain sounds and the inability to synthesize certain sound patterns. Also, on a hardware development level, no clues and aids from the predictive or past history of the speech signal are used at present, although they obviously contribute to the human production and processing of speech. As the basic study of the frequency spectrum of speech, either on an acoustic basis or on an articulatory basis, reveals further indications and classifications of the speech act, the performance of formant vocoders should reflect these advances and they should become more useful operational devices. The complexity of both the analyzing and synthesizing hardware devices seems destined to increase in large degree as such advances are forthcoming from the research field In addition, the research areas are not ones which promise quick and easy solutions to the existing problems. They have been attacked for a number of years, and although they have yielded sufficient data for the present hardware developments, many problems of both synthesis and analysis of the speech signals on a purely acoustical or articulatory basis remain. It seems that additional help is needed from analysis related to the speech act in slightly different ways than the acoustic or articulatory.

4.5 Miscellaneous Devices

A variety of other compression devices ranging from C.P. Smith's Spectrum Pattern Quantization[16, 17] Technique to perception network patterns have been attempted.[18] In no case has more than limited success with these devices been reported. However, the application of such patterning devices is in its early stages and should prove to be a significant step in speech research.

5. CURRENT TRENDS IN RESEARCH

5.1 General Remarks

Research in speech processing is proceeding in several interrelated areas aimed at finding basic data on the information-bearing elements of the speech signal, on its perceptual attributes, and on the "natural" representation of speech systems by quantitative models. The concrete objectives of these basic programs are many and varied, including automatic speech recognition (non-machine voice communication), information retrieval, bandwidth compression, and others.

In the following discussion, research areas are arbitrarily divided into acoustic-articulatory and perceptive-linguistic categories. These areas are related, and the division into two classes is not meant to indicate that they are divorced from each other. It is merely convenient to make this distinction for the purpose of discussion. The two categories do not exhaust the field, but do cover the major trends.

5.2 Acoustic Speech Analysis and Synthesis

5.2.1 The greatest concentration of effort seems to be centered on analyses of the frequency spectrum of the acoustic output and on models of the underlying source and cavity that produce the acoustic output.[19, 20, 21, 22, 23] Other studies are concentrating on identification of a limited number of patterns (for example, words or syllables).[24, 25, 26]

5.2.2 A recent significant paper indicates one of the key problems. Simply stated, this is the specification of the dynamic properties of speech generation. The static properties of the acoustic output have been defined to a great extent.[6] Specification of the dynamic properties of speech involves determination of the articulatory parameters and study

of how changes in these parameters modify the acoustic speech output. Consideration of the physiological constraints on speech generation may yield invariant and significant cues for speech perception.

The discussion in point[27] deals with the dynamic properties of vowels as determined by a series of controlled experiments providing spectral data on vowels in various consonantal environments. The study seems to indicate that there are definable targets for vowel generation that remain invariant under a variety of consonantal environments as well as under variation of stress and duration. These data are significant, in that they point toward "encoding and decoding in terms of an analysis-by-synthesis procedure[8], or the motor theory of speech perception"...[28]. "The implication is that acoustic and physiological data may provide linguists with criteria that contribute toward elimination of the classical problem of 'nonuniqueness' in phonemic analysis. Moreover, they can serve as a preliminary answer to (the) challenge to the effect that there is no reason to believe that any approach will materially reduce the obvious complexity of phonologic-acoustic correlation." The article concludes with a suggestion for future work in the areas of identification of specific "targets" a speaker aims for (acoustically and physiologically) when he speaks, of the rules which govern the temporal distribution of these targets, and of the properties of the structures that respond to the target instruction.

5.2.3 The significance of these preliminary experimental results, and of the whole series of programs coming out of M.I.T., R.I.T. (Royal Institute of Technology, Stockholm), and other places, is that a concerted attack on an acoustic-articulatory basis shows hope of providing the invariant or significant cues and of identifying the information-bearing parameters of the speech signal. If successful, these studies will provide raw data which will allow automatic speech recognition systems to be built to the highest intelligibility standards. It is obvious that these

programs are long-term endeavors of great complexity, and as yet they have not provided enough data to permit significant use of their findings in hardware development. Many research questions must be answered before practical applications can be attempted — even good consonant models are not yet available in any general form However, it is not too farfetched to envision (as one of the results of this type of research) a speech compression system with the following general characteristics: the system would consist of an analyzer and a synthesizer, each with extensive computational capacity; the analyzer would divide the incoming speech into distinct segmented phonemes, and the synthesizer would then choose stored samples from which it would reconstitute continuous speech.

5.2.4 The area of synthesizing speech from stored segments has occupied the attention of researchers for many years. Among the most recent work to be reported, and a good summary of synthesizing techniques now available, is the recent paper by S. E. Estes et al "Speech Synthesis From Stored Data".[30] An analog synthesizer is used in the reported work. Instead of natural speech segments, the system uses segmented control signals to the synthesizer. This technique bypasses the difficulties of discontinuities in pitch and formant frequencies at the segmental junctions. Control signals are varied until the speech segment being studied is satisfactory, and connected speech may be generated by computer assembly of the synthesizer control signals corresponding to a sequence of speech segments. The system provides for specification of the sequence of segments. This recent paper provides a partial answer to Silvertsens' discussions,[31] which pose the problem that: "It is well known that there are a number of difficulties in segmentation of the speech continuum. Doubts have been raised whether such segmentation is at all possible acoustically or physiologically, except in a purely physical sense. X-ray studies of the physiological mechanism (see reference 6 for a later discussion of X-ray technique), and acoustical analysis of the speech wave show continuously changing formation and

patterns in many cases where an intuitive auditory phonetic analysis finds a segment border."

5.2.5 Additional areas of research on an acoustic level indicate the broad scope of the programs at R.I.T. Stockholm, M.I.T., Bell Labs, and elsewhere. A typical report[23] is a good indication of the acoustic paths being followed. These include:

Spectrum Analysis — Recent work in this area involves the use of a new technique which deserves brief comment. "Inverse filtering", also known as "anti-resonance filtering", attempts to cancel the formant frequencies by the use of filters, thus defining the spectrum of a larynx source function. By applying this technique, one can also determine formant frequencies quite accurately.

Speech Parameter Tracking — In this area, which relates to vocoders, basic studies are being carried out on voice frequency tracking [3, 32] and on the definition of voiced-voiceless distinctions[20].

Speech Production --- Work in this area includes measurements of glottal pulse amplitude and shape, and of the variation of these parameters from speaker to speaker. An understanding of this aspect of the physiology of speech production is important in defining the acoustic source function and in normalizing any generalized machine for speech analysis and synthesis.

Speech Segmentation and Synthesis — This area is discussed briefly above in paragraph 5.2.4. Other work is concerned with analyzing the acoustic cues for segmentation which are contained in the natural speech waveform and with determining their variant and invariant characteristics.

Articulation — An attack is being made on an articulatory level to determine the parameters and significant variations of the acoustic speech act. This research is centered at M.I.T [8, 9, 10, 11] and is based on automatic computer programming. In its early stages, "human-in-the-loop" strategy is used in an attempt at analysis by synthesis by varying the articulatory parameters of a model in order to provide acceptable acoustic outputs. This is in contrast to the technique of varying acoustic parameters. In some ways, it appears that a simpler description of the speech act is available by this method, using fewer parameters than are required by purely acoustic analysis. This research is in its early stages, although significant results have already been reported [10, 11]. As a technique it is complementary to acoustic techniques.

5.3 Speech Perception and Linguistics

5.3.1 Work in speech perception correlates very closely with acoustic and articulatory research. Since normal speech communication is aimed at a human listener, it is tailored to the perceptual ability of man. Therefore any knowledge of perceptual clues will influence the concepts involved in speech processing.

5.3.2 A recent paper [7] summarizes what is known at present about perceptive criteria. In it, difference limens (D.L.) for various sounds and phonemes are given. For formant recognition devices, it is shown that the just-discernable changes in formant bandwidths are much larger than those for formant frequency. Problems in this area are also defined (for example, "Apparently discriminations of simultaneous changes in frequency and bandwidth have not been measured. The manner in which these cues combine would be an interesting question. That is, what is the 'D.L. area' in the complex frequency plane?".) Similarly, pitch changes of 0.3% to 0.5% have been defined as the upper bound to the accuracy necessary in processing pitch data. Many questions exist

regarding the uses of the rate of change of pitch and the detection of inflection contours for higher order processing[33, 34]. There are additional questions of glottal pulse perceptual criteria. Much work remains to be accomplished in the area in terms of models of speech perception[35].

5.3.3 In a recent paper, A. M. Liberman, F. S. Cooper, K. S. Harris and P. F. MacNeilage[28] of Haskins Laboratories define perhaps the most interesting trend in modern speech perception, using linguistic and psychological constraints and inferences. This section will end with a discussion of their Motor Theory of Speech Perception. The first few paragraphs of their discussion offer a precise definition of their problem approach and of the background reasoning that led to their theory. As such, they offer an excellent perspective of the problems of man-machine relationships relative to speech processing. These paragraphs are therefore repeated in their entirety. It is our feeling that this area deserves much attention in future research efforts.

"The accuracy and speed with which speech is perceived must surely rank as one of its most important properties. To appreciate how remarkably good language is in this respect, one need only try to find or to fabricate a set of non-speech sounds that will serve a human being as well. Attempts to do this in connection with reading machines for the blind, for example, have not been notable successful, and we know that Morse code is relatively poor by comparison with speech, even after years of practice.

"Because the sounds of speech are highly distinctive -- that is, absolutely and quickly identifiable -- they are efficient vehicles of information transmission. But the distinctiveness of speech sounds is more than a matter of mere convenience and utility; it is, in fact, necessary if a language as we know it is to exist. An obvious consideration is that

a phonemic system requires by its very nature that the sound elements be identifiable in absolute terms. The phoneme /d/, for example, must be perceived not merely as something which is more or less like /d/ than the last sound heard, but as /d/ itself. Nor does the grammar of our language permit that this or any other phoneme be transformed into a radically different perception according to the context in which it appears Phoneme perception must be absolute, or very nearly so, if language is to be phonemic. We should note, too, that there is a sound psychological reason why languages are phonemic: a non-phonemic code would present difficulties for much the same reason, and to roughly the same extent, that a syllabic or word system of writing and reading is more difficult than one based on an alphabet.

"The requirement that the phonemes be perceived quickly is, perhaps, less obvious, but it is none the less real. If one tries to understand language when it is read to him slowly, letter by painful letter, he sees that the phonemes must come along at a rather high rate else the listener cannot organize them into morphemes, words, and sentences.

"Although speech is highly distinctive in human perception, no machine has yet been designed which finds it so. Indeed, it may well be true -- and, if so, ironic -- that machines will have their greatest difficulty with those very phonemes (e.g., the stop and nasal consonants) which are for human beings most highly distinctive and which probably carry the heaviest load of information. There are of course, various codes, all of them non-linguistic, which are well received by a variety of machines, but these work badly with human beings. That man and machine are so different in this respect is not very remarkable. It merely indicates that there is something psychologically interesting about the efficiency of speech perception, and that we shall increase our understanding of language, and man, by inquiring into the conditions of distinctiveness and the perceptual mechanisms which underlie it. It also

emphasizes, if emphasis were needed, that those who would design a speech recognizer might somehow profit from the same kind of inquiry."

Next they ask what constitutes this distinctiveness. The first point is that distinctiveness may not be ----"inherent in the acoustic speech signal, but is rather added as a consequence of linguistic experience. More important, perhaps, it (the experimental evidence they use) indicates that even with a considerable background of linguistic experience on the part of the listener, the acoustic signal is distinctive only when, being heard as a speech sound, it engages some kind of speech perception system."

They go on to show that stop consonents are heard absolutely and stored as given phonemes. Moreover, they indicate that this catagorical perception is a learned trait. Vowels, it appears, are not perceived catagorical but continuously, although phonemic distinctions like length and tone may lead to distinctive interpretations. (Consider the earlier discussion in Paragraph 5.2.2 which indicates that, articulatorily, one attempts to arrive at a distinct position when saying a vowel although one may never reach the particular position-goal.) At any rate, it is suggested that perception is based on articulatory movements. Evidence is cited that indicates "...the relationship between phoneme and articulation is more nearly one-to-one than between phoneme and acoustic signal." Further, the articulatory movements are not the descriptions of articulatory phonetics which are concerned with changes in the vocal tract positions; rather they are related to the motor commands which initiate these changes. To this end, the report concludes with a discussion of possible experiments to be performed, and of studies presently in process, which will give clues to neural commands or to particular muscle action potentials with which the commands can be associated.

This important paper derives a new approach to speech processing, and the theory opens new avenues in a non-acoustic domain for the processing, perception, and communication of speech signals.

6. EXPERIMENTAL WORK CARRIED OUT UNDER THE PROGRAM

6.1 Processing Mercury Flight Data

6.1.1 One of the most neglected aspects of modern linguistics has been that of "idiolectical linguistics", i.e., the study of speech patterns of individuals rather than the generalized patterns of a linguistic community or "dialect". Since it seems reasonable to assume that for voice communications to and from a spacecraft we shall be interested in a small number of known speakers, the study of individual idiolects could prove to be of great value in designing voice communication systems and speech processing techniques for space travel. To this end, the official NASA transcript of the voice communications of Walter M. Schirra, Jr.'s MA-8 flight was examined for general and idiolectical linguistic information.

6.1.2 Schirra's voice communication was taken from the MA-8 spacecraft's onboard tape recording. This tape was manually transcribed by NASA into conventional message, sentence, and word format and orthography. Our preliminary and necessarily brief examination of the transcribed text, using manual data processing techniques, indicated that valuable linguistic and psycholinguistic information could be gleaned from the text. It was also apparent that the processing of the MA-8 voice data could and should be mechanized for more economic and accurate rendering of the data.

6.1.3 One of our first tasks was to obtain some gross statistics of the message traffic. For example, keeping in mind always that we were not dealing directly with verbal units but rather with transcribed or semi-processed speech, we calculated that Schirra and his ground communicators, in his six-orbit mission lasting some nine hours and twenty-three minutes, uttered at least 30,000 transcribed "words." This means that on the average almost 55 "words" per minute were sent and received by Schirra.

Or to put it another way: assuming that the average speaking rate is about 125 "words" per minute, almost half of Schirra's entire elapsed time was spent in sending and receiving voice communications. This was obviously his main activity and took up most of his time and energy. We also found that, on the average, messages consisted of fifteen words spread over 2.3 sentences - or about seven words per sentence. These and other similar statistics give us some feeling for the size and limits of the space communication problem.

6.1.4 We were also able, in the all-too-brief time spent in examining the transcript, to get some ideas concerning the possibility of gleaning syntactic information from data. For example, we noticed that on occasion, Schirra would stop using personal pronouns. Was this a random or free speech variable or correlated to some physical event? Also, what are the syntactic peculiarities of a man under stress? What syntactic patterns are misunderstood or not received? These and many other questions need to be answered; many of them can be explored by a more careful, controlled examination of the Mercury transcripts.

6.1.5 We were also concerned with the problem of the transcription of voice to printed form By comparing the actual tapes with the official transcript, it was obvious that a faithful transliteration had taken place. That is, from a common-sense point of view, the written text was substantially what Schirra said in terms of gross meaning. However, the non-speech sounds he made were, for the most part, ignored (as they rightly should have been for the purposes of transcription). A closer or finer analysis of the tapes, and possibly a phonetic transcription, might reveal more interesting data about his overall psycholinguistic behavior. We have, for example, a rough estimate of the speed at which the messages are uttered. But this is only an average per message. A phonetic examination of the tapes themselves might reveal (as they did in a cursory inspection) staccato utterances in some places or more subtle pauses, stammers, etc., which were deliberately ignored by the transcriber. These facts, when related with the other events of the flight

and the pilot's physiological records, could be used to correlat guistic with physical and psychological states. In fact, it seems to be feasible to do some psychological processing of the astronaut by a careful examination of the linguistic and paralinguistic data not usually considered. These and the other syntactic and semantic aspects of idiolectical linguistics are fruitful areas for future work. See Section 7.

6.2 The Utility of Lip-Reading

6.2.1 During the study program a brief investigation of lip motion sensing was carried out. Standard references on lip sensing were consulted, and a visit was made to the California School for the Deaf for consultation with teachers and students.

6.2.2 Obviously, visual clues contribute to intelligibility in speech communication. Consider figure 6-1, which indicates words correct vs speech-to-noise ratio and articulation index with and without visual clues. The figure which is derived from reference 39 indicates the increment of intelligibility as a function of signal-to-noise ratio. When this ratio is high, the listener hears clearly and therefore cannot take advantage of the cues given by lip reading; if this ratio is low, the listener needs and uses the visual cues. The "totally deaf" obviously must and do use "lip motion sensing." This technique appears to be successful to the limit of its inherent constraints. Certain sound groups cannot be differentiated by visual clues (e.g. b and p, d and g), and others can only be sensed if a visual clue as to the voiced-voiceless distinction is available (for example, through seeing or even feeling laryngeal action). The differentiations must be made on the basis of "context" or higher order constraints. The remarkable success some deaf people have in comprehension can only be explained by the marvelous adaptive computer they have and the evolution of natural language processing in humans through the eons of time.

6.2.3 This brings us to one of the key problems in speech processing - that of defining the statistical contraints and feedback loops we use to interpret speech. This topic is discussed in section 7.7. There is information available on the physiological level which if sensed may provide pattern data for processing systems. For instance, the recent work of Hillix [36] indicates that for a limited vocabulary at least, information provided by lip-motion sensing can be used together with acoustic data to provide better processing accuracy.

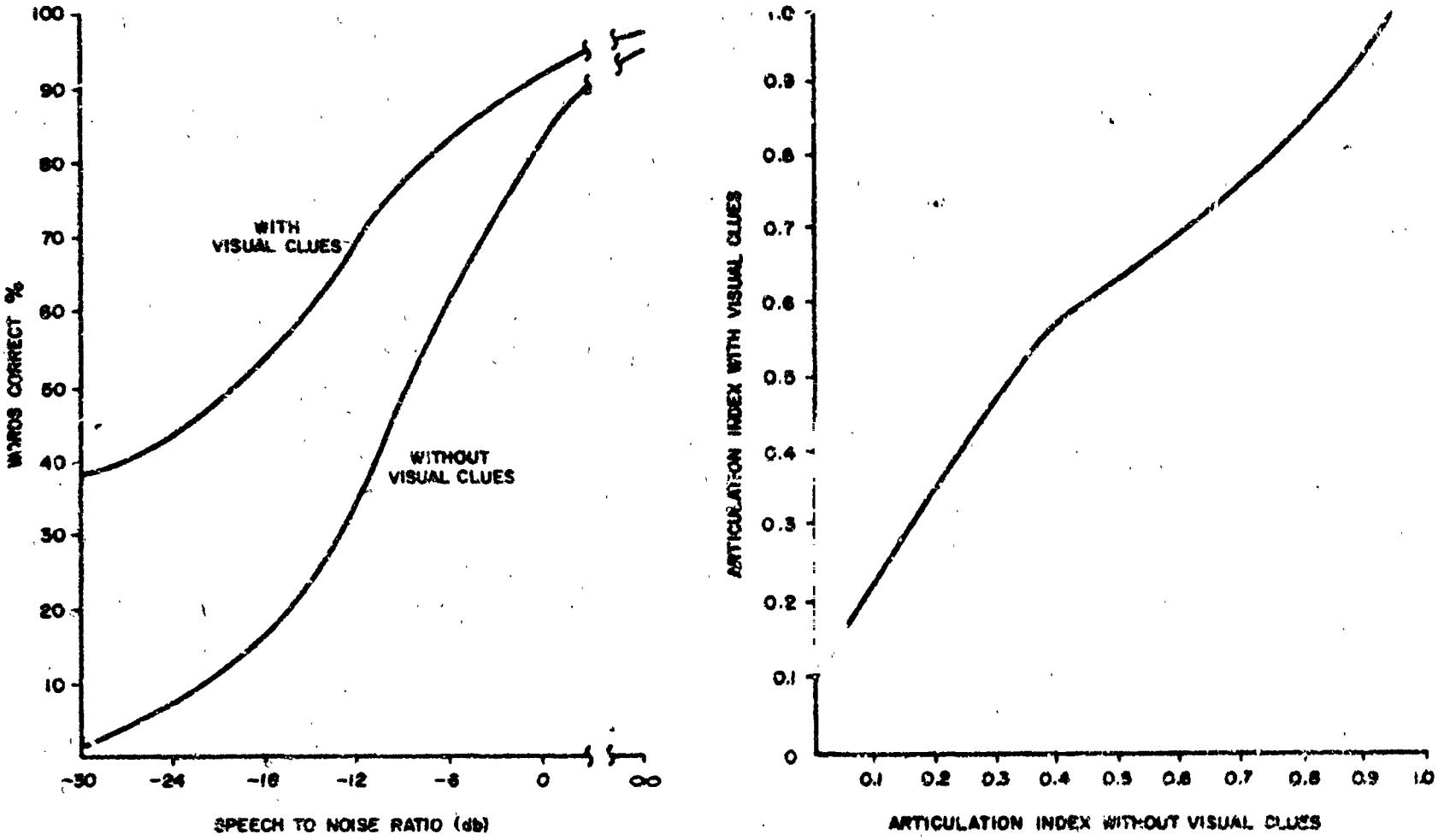


Figure 6-1. Effects of Visual Clues on Intelligibility

6.3 Pattern-Matching of Sampled Speech

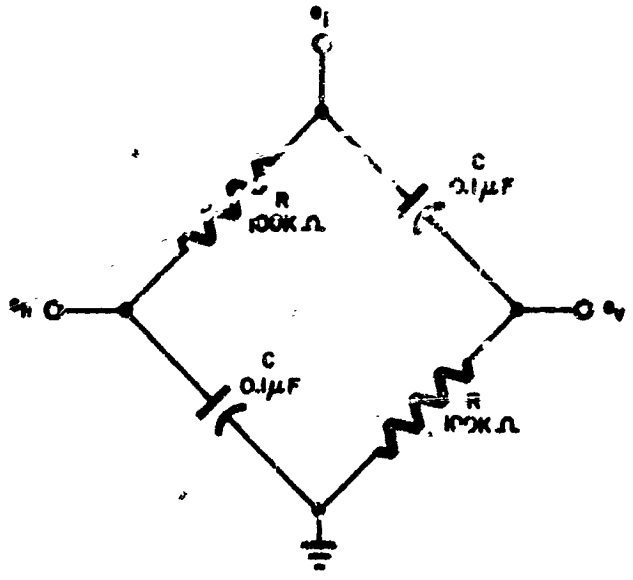
We carried out a brief experiment in which we determined speech signal patterns for three different vowel sounds with two different speakers (three speakers in one case). Cross-correlation between patterns for the same vowel sound as spoken by different speakers was always above 0.5. Cross-correlation between patterns for different vowel sounds, whether spoken by the same speaker or by different speakers, was never above 0.12. This is described in detail in paragraph 7.4.9.

6.4 Two-Dimensional Patterns From Speech Signals

6.4.1 As a small part of our work, we devoted a few hours to a repetition and expansion of the experiments performed by C. W. Barton and S. H. Barton and reported on in 13 December 1963 issue of Science Magazine[37].

6.4.2 The experiment essentially consists of driving an oscilloscope, through a simple network, with speech signals in order to produce recognizable pictures from speech sounds. The actual circuit used is shown in Figure 6-2. Neglecting loading and stray capacitance, the R C leg of the network lags the C R leg of the network by 90°. Speech input (sound pressure from a single voice), when played through a microphone into the circuit, produces a stable oscilloscope trace of sound pressure versus derivative of sound pressure. Pictures of results from five subjects saying the same phonetic material are given in the Science article. Some of the sounds are fairly constant and easily recognizable, especially the vowels, but consonants like th, v, z and mu are quite variable from subject to subject. Moreover only two pictures of unvoiced sounds (f and s) are presented in the Science article and these noisy sounds do not form particularly significant pictures.

6.4.3 The real object of our experiment was to determine if consonants, which are much less energetic and have less harmonic content than vowels, could be usefully displayed by this method and further, to see if we could get vowel representations similar to those in the Barton article. The latter point was verified, and Figures 6-3 and 6-4 are representative line drawings made from polaroid photographs of oscilloscope pictures for two vowels, ē and ah. However, we found it impossible to get displays of plosive consonants and certain other unvoiced consonants. This is not an unexpected result, as it has been shown that the vowel transition in many cases contains the consonantal information and the technique



e_i – INPUT VOLTAGE FROM AUDIO AMPLIFIER

e_h – OUTPUT VOLTAGE TO HORIZONTAL AMPLIFIER OF OSCILLOSCOPE

e_v – OUTPUT VOLTAGE TO VERTICAL AMPLIFIER OF OSCILLOSCOPE

Figure 6-2. Oscilloscope Driving Circuit

Figure 6-3

Line Drawing of Oscilloscope Picture of Vowel ē

Figure 6-4

Line Drawing of Oscilloscope Picture of Vowel ah

used in this experiment cannot sense this. The technique is valid and shows some promise if modified sufficiently to provide as the authors of the Science article state, ..."a natural phonetic alphabet with which one can capture nuances of pronunciation.....". The value of such an alphabet is obvious.

7. NEW AVENUES FOR FUTURE WORK

7.1 General Considerations

7.1.1 We devoted a substantial fraction of our total effort to a search for new ways of looking at the problem of speech processing in general and speech compression in particular. We have, we believe, turned up several approaches which appear to offer the possibility of improvement in the state of the art. All of these are described in this section. Before embarking on detailed discussions, however, we wish to say something about our overall view of the problem.

7.1.2 We feel that the limited problem of speech compression (as opposed to speech processing) can be considered as the problem of minimizing the transmission load required to transmit messages of a particular class — messages which are the electrical equivalent of acoustic speech signals. In order to devise a way of coding such messages so as to minimize the transmission load requirement, we must have considerable detailed knowledge about the messages themselves. We do not, however, need to know anything about where the messages come from, how they are produced, or how they are understood and interpreted.

In effect, to devise a way of coding, we need to know as much as possible about the electrical analogues of acoustic speech signals, but we don't need to know anything about speech per se. This view of the problem results in the first of our approaches, which is described in sub-section 7.2.

7.1.3 When we look at the more general problem of speech processing, we think of speech as a sequence of specialized sound patterns. Since we presumably understand speech by recognizing these patterns, the obvious approach to speech processing is that of devising a machine which will be able to carry out the same recognition process From a linguistic standpoint, the basic units of speech are phonemes. It seems reasonable, therefore, to start by trying to mechanize the process of phoneme recognition. It is not too surprising to discover that this is not an easy solution. Phonemes are, to be sure, the basic structural units of speech. We have

no guarantee, however, that the human nervous system processes speech by recognizing phonemes. Even if phonemes turn out to be the basic units used by the human speech processor, it is always possible that a mechanical processor may deal more efficiently with other units. It is a basic truth that technology does not operate most efficiently when it copies nature — nobody builds four-legged tractors, and the ornithopter is nothing more than an aerodynamic curiosity. The proper approach, we suggest, is to forget all about speech and look for patterns in the electrical speech signals. This is the basis of the approach described in sub-section 7.3.

7.1.4 Our next approach, described in sub-section 7.4, runs counter to everything said in the foregoing paragraph -- it is a technique for mechanizing a phoneme recognizer. We include it because we think that it offers a good chance for a quick payoff. Our remaining approach, described in sub-section 7.5, is an off-beat approach to the problem of pattern recognition.

7.2 Speech Compression as a Problem in Coding

7.2.1 This approach to the problem is aimed primarily at speech compression, to the exclusion of speech processing. It will, however, provide a wealth of interesting data as a by-product. Briefly, we propose to sample and quantize speech signals, as outlined in Appendix "C". We now think of a speech message as consisting of a string of symbols, where the alphabet of symbols is simply the set of numbers which describes the various quantizing levels. Thus, the symbol string

143, 162, 190, 340, 27, -11 (7-1)

might be interpreted to mean that successive samples of the speech signal had voltage levels of

1.43, 1.62, 1.90, 3.40, 0.27, and -0.11 volts (7-2)

7.2.2 We start by determining the statistics of speech messages. Thus, if P_j is the relative frequency of occurrence of the j-th symbol (the symbol which indicates that a sample of the speech signal, after quantizing, was at

the j-th voltage level), we can calculate a value for the information content of speech, neglecting the intersymbol constraints, by using the usual relation

$$H_1 = -\sum_J P_j \log_2 P_j \tag{7-3}$$

where H is the information contents in bits/symbol and the subscript indicates that we have neglected the intersymbol constraints by considering the symbols separately. Similarly, we can calculate

$$H_2 = \frac{1}{2} \sum_{jk} P_{jk} \log_2 P_{jk} \tag{7-4}$$

$$H_3 = \frac{1}{3} \sum_{jkl} P_{jkl} \log_2 P_{jkl} \tag{7-5}$$

etc., where P_{jkl}, for example, is the probability of occurrence of the triplet consisting of the j-th, k-th, and l-th symbols (in that order) and the subscript on H indicates, in equation (7-5), that we have considered intersymbol constraints over groups of 3 symbols.

7.2.3 Using these data, we can attempt to verify some of the heuristic conclusions set forth in sub-section 2.2 above. Suppose that we sample at a rate of R samples/second and quantize to a total of L levels. The bit-rate for unprocessed speech is then

$$R \log_2 L \text{ bits/second} \tag{7-6}$$

The information rate, considering intersymbol constraints over blocks of n symbols, is RH_n. The limiting compression ratio available by optimum coding of blocks of n symbols is then

$$C_n = \frac{R \log_2 L}{RH_n} = \frac{\log_2 L}{H_n} \tag{7-7}$$

Further, since n symbols are generated in a time n/R, we can define

$$\left.\begin{aligned} C(\tau) &= \frac{\log_2 L}{H_n} \\ \tau &= n/R \end{aligned}\right\} \tag{7-8}$$

where $C(\tau)$ is the limiting compression ratio available through optimum coding with a processing time τ.

7.2.4 An interesting series of experiments would be the determination of RH_n (for large n) for various values of R and L. This would give us some idea of the penalty incurred by narrow-banding (reduction of R) and coarse quantizing (reduction of L). Comparing these data with the results of listening tests, we would be able to decide whether or not the human speech processor uses all of the information available to him. It is possible, for example, that narrow-banding starts to reduce information content before the ear can detect any difference. Should this be the case, we would conclude that the high-frequency portion of the speech signal might well contain information that would be useful to a mechanized speech processor even though it is useless to the human listener.

7.2.5 Aside from the utility of this approach, in terms of generating interesting data, it might tell us enough about the statistics of natural speech (from the information-theoretic viewpoint) to enable us to devise a low-redundancy code. Speech compression could then be achieved by simply translating from the "natural" code to the new one. Decompression, or synthesis, involves nothing more than inverse translation followed by digital-to-analogue conversion in order to reproduce the original sampled-and-quantized speech signal. This, in turn, need only be passed through a low-pass filter in order to recover the original signal (as corrupted by quantizing noise). This approach, should it be successful, would therefore eliminate all of the problems of speech analysis and speech synthesis associated with the more ambitious speech processing approach.

7.2.6 The hardware required for this approach is quite straight-forward and involves nothing more than tape recording equipment, sampling circuitry, pulse-height analysis circuitry, and some low-speed logic circuitry.

7.3 A Technique for Discovering Patterns in Speech Signals

7.3.1 While the first approach, sub-section 7.2, considers the speech processing problem as one of coding, this approach considers it as a

decoding problem. Starting as before with sampled quantized speech, we consider a speech message as a string of symbols A_j where each symbol represents a particular voltage level. We now suppose that there exists a set of super-symbols B_K, where a particular B_K consists of a string of symbols A_j. Thus, for example,

$$\left.\begin{aligned} B_1 &= A_{11}A_{43}A_{19}A_5 \\ B_2 &= A_{12}A_2A_{-1}A_{14} \\ &\text{etc.} \end{aligned}\right\} \qquad (7\text{-}9)$$

We set ourselves the problem of discovering the super-symbols B. Hopefully, the number of super-symbols actually used is less than the totality of super-symbols possible, and we can achieve compression by coding the speech message as a string of super-symbols. Further compression will, of course, be possible through a modified coding to eliminate redundancy if the super-symbols do not appear with equal frequency. The merit of this approach lies in the possibility that we can ultimately construct a device for recognizing super-symbols, thus taking a large step forward in solving the man-machine communication problem.

7.3.2 We infer the existence of super-symbols, or patterns in the speech signal, from the fact that speech sounds are a relatively small sub-set of the set of all possible sounds. Speech, for example, does not involve any sustained pure sine-waves.

7.3.3 There are a number of ways of searching for super-symbols, but a description of one way will suffice. Let us consider a particular symbol A_j. This may appear at the beginning of a super-symbol, within a super-symbol, or at the end of a super-symbol. Let us suppose that A_j appears at a particular location, say as the tenth symbol, in a particular super-symbol. There is a certain symbol pattern (of A's) to the left of A_j and another to the right. When A_j appears at another location, say the thirtieth symbol in the same super-symbol, or at some point in another super-symbol, it is accompanied by different symbol patterns on its right and on its left. Roughly speaking, if A_j appears at some number of different internal locations in

the set of super-symbols, the number of different patterns on the left of A_j should be the same as the number of different patterns on its right

If we consider only those cases in which A_j is the last symbol in a super-symbol, the number of patterns to the left of A_j is, at most, just the number of different super-symbols which end in A_j. The number of patterns on the right of A_j, however, is very nearly equal to the total number of super-symbols.

Similarly, if A_j appears at the beginning of a super-symbol, the number of different patterns on its right (the number of super-symbols starting with A_j) is less than the number of different patterns on its left (the total number of super-symbols).

7.3.4 We now initiate our search for super-symbols (patterns) in the following way:

a. We compute a set of probabilities P_j, where P_j is the probability of occurrence of the symbol A_j.

b. We next compute the set of joint probabilities P_{jk}, where P_{jk} is the probability of occurrence of the symbol pair A_jA_k in that order.

c. We next compute the sums

$$\sum_j P_j \log_2 P_j \tag{7-10}$$

$$\sum_k P_{jk} \log_2 P_{jk} \tag{7-11}$$

$$\sum_k P_{kj} \log_2 P_{kj} \tag{7-12}$$

d. We now compute

$$\log_2 P_j - \frac{1}{P_j} \sum_k P_{jk} \log_2 P_{jk} \tag{7-13}$$

If there are no constraints between A_j and the following symbol A_k, we have

$$P_{jk} = P_j P_k \tag{7-14}$$

and equation (7-13) reduces to

$$-\sum_k P_k \log_2 P_k \tag{7-15}$$

If, at the other extreme, there is complete constraint so that A_j is always followed by, say, A_r, then

$$\left.\begin{aligned} P_{jk} &= P_j, \text{ if } k = r \\ &= 0, \text{ if } k \neq r \end{aligned}\right\} \tag{7-16}$$

and equation (7-13) reduces to zero. Expression (7-13) is therefore a measure of the freedom from constraint on the symbol following A_j

e. Similarly, we compute

$$\log_2 P_j - \frac{1}{P_j} \sum_k P_{kj} \log_2 P_{kj} \tag{7-17}$$

which is a measure of the freedom from constraint on the symbol preceding A_j.

f. Finally, we take the ratio

$$\frac{\log_2 P_j - \frac{1}{P_j} \sum_k P_{kj} \log_2 P_{kj}}{\log_2 P_j - \frac{1}{P_j} \sum_k P_{jk} \log_2 P_{jk}} \tag{7-18}$$

and examine this for various values of j. A high ratio indicates low constraint to the left of A_j and high constraint to the right. Hence we suspect that A_j appears at the start of a super-symbol.

If this procedure does not yield a good candidate A_j, we try again with the most likely candidates A_j, forming the ratio (for the likely values of j)

$$\frac{\log_2 P_j - \frac{1}{P_j} \sum_{kl} P_{jkl} \log_2 P_{jkl}}{\log_2 P_j - \frac{1}{P_j} \sum_{kl} P_{klj} \log_2 P_{klj}} \qquad (7\text{-}19)$$

7.3.5 Having found a likely candidate A_j for the start of a super-symbol, we next examine doublets A_jA_k to find out which ones lie at the beginning of a super-symbol starting with A_j and which ones lie within super-symbols. To do this, we calculate the ratio

$$\frac{\log_2 P_{jk} - \frac{1}{P_{jk}} \sum_{l} P_{jkl} \log_2 P_{jkl}}{\log_2 P_{jk} - \frac{1}{P_{jk}} \sum_{l} P_{ljk} \log_2 P_{ljk}} \qquad (7\text{-}20)$$

which gives us A_jA_k as the start of a super-symbol if the ratio is high. As before, if the ratio (7-20) does not enable us to choose A_k, we use the ratio

$$\frac{\log_2 P_{jk} - \frac{1}{P_{jk}} \sum_{lm} P_{jklm} \log_2 P_{jklm}}{\log_2 P_{jk} - \frac{1}{P_{jk}} \sum_{lm} P_{lmjk} \log_2 P_{lmjk}} \qquad (7\text{-}21)$$

The process is continued in the same way. The end of a super-symbol is signalled by a relaxation of constraints, so that there is no obvious next symbol to add to the pattern, or by a simple repetition of the pattern (indicating that the super-symbol is repeated).

7.3.6 Having found a set of super-symbols, we next attempt to derive an even more basic set of symbols by looking for families of super-symbols which differ only, say, in the amplitude of the original speech signals from which they are derived. The speech message can then be coded as a string of basic super-symbols, each with an appropriate modifying symbol attached to it. We would hope, in the end, to have a set of machine-recognizable sub-phonemes which, together with amplitude and pitch modifiers, would represent natural speech with no loss in quality. (The duration of a phoneme will be signalled by the number of times the appropriate sub-phoneme is repeated.)

7.4 A Technique for Instrumenting a Phoneme Recognizer

7.4.1 Our next approach is based on the assumption that the basic speech units recognized by a processor can be the same as the wave forms which, when repeated, form the phonemes used by linguists to describe speech.

7.4.2 The speech signal is first band-limited, adjusted in level by a slow AGC circuit, and clipped in order to avoid excessively high peaks in either direction. This pre-conditioned speech signal is now fed to a level extractor, which determines the short-term level for the purpose of obtaining information about accent. This can be as simple as a rectifier followed by a low-pass filter. The short-term level signal is sampled at a suitable rate, quantized, and transmitted along with other information. At the same time, the pre-conditioned speech is sent to a pitch extractor. Here the speech is low-passed and then infinitely clipped, and a sync pulse (appearing once in each pitch period) is derived, say, from the positive-going zero crossings of the clipped speech. The output of a voltage-controlled oscillator is divided by some suitable number (say 128) in standard counter circuits, and the appropriate sub-multiple of the oscillator output (the 128-th, for example) is matched to the sync pulse by standard techniques. The oscillator frequency is thereby locked to some harmonic of basic pitch frequency. The oscillator control voltage is sampled at an appropriate rate to obtain pitch information. The samples are quantized and the signals are transmitted along with the short-term level signals and the phoneme identification signals. When there is no discernible pitch frequency, the oscillator reverts to some preset standard frequency.

7.4.2 The voltage-controlled oscillator is now used to produce sampling pulses at, for example, 128 times the basic pitch frequency, or at some fixed rate if no pitch frequency can be determined. At each sampling point, the polarity of the speech signal is determined and a corresponding binary signal (for example, a one for plus and a zero for minus) is read into a shift register. At each sync pulse, the shift register is dumped, providing a 128-bit word which describes the speech signal over the most recent sampling period. By this technique, the speech signal is converted to a string

of symbols, one symbol being generated during each pitch period, each symbol consisting of a 128-bit word.

7.4.3 The 128-bit word may be clocked into a high-speed shift register. Alternatively, the one-zero bit-stream may be converted into a sequence of positive (for one) and negative (for zero) pulses and sent to an analogue delay line. In either case, when the shift register or the line is full we have 128 bits available in parallel for cross-correlation against a set of reference patterns. In the analogue delay-line case, which is perhaps simpler to mechanize, the 128 positive or negative voltages are summed with or without sign reversal. We thus have

$$\sum_{j=1}^{128} P_j W_j \qquad (7\text{-}22)$$

where, for example, P_j is plus or minus one volt according to the sign of the j-th pulse, and W_j is plus or minus one according to the wiring of the coupling network from the j-th tap to the summing circuit. We have, in effect, a cross-correlation between the train of 128 pulses and the set of 128 weighting factors.

7.4.4 The outputs of the delay line taps are fed, in parallel, to a set of correlators, one for each sub-phoneme pattern. Each correlator, of course, has its own weighting network between it and the delay line. The correlator outputs are smoothed, and a pattern selector examines the outputs at appropriate intervals, selects the highest output, and transmits a string of digits to indicate which of the correlators shows the highest output.

7.4.5 The system output is, then, a string of symbols, each standing for a particular pattern which, when repeated, forms a particular phoneme. Interleaved with this string, we send level and pitch information.

7.4.6 At the receiver, we store a set of 128-bit words (one for each pattern) in a core memory. We use the pitch information to control an oscillator like the one in the transmitter. At each sync pulse, we note the most recently received pattern identifier and shift the appropriate word

from the core to a shift register. We then clock this word out of the shift register at, for example, the 128-th harmonic of the oscillator. The result is a string of 128 positive or negative pulses, equally spaced over a single pitch period and making the best possible match (within the limited number of pattern-identifying correlators) with the train of pulses which was fed to the correlation matrix in the processor.

7.4.7 The output pulse train can be converted to clipped speech by low-pass filtering. After compensation for the filter characteristics, it is then modulated by the short-term level signal.

7.4.8 This scheme has the great merit that the synthesizer matrix is identical with the analyzer matrix. That is, the words stored in the synthesizer core are identical with the weighting functions used in the analyzer correlators. Further, each element in the analyzer matrix is either plus one or minus one, which allows easy experimentation with slightly-modified phoneme patterns. Clearly, the patterns can be adjusted, if necessary, for a compromise between best phoneme recognition and best reproduction.

7.4.9 The following experiment was carried out using this technique, but using only 16 sample points in each pitch period. The vowel sounds $\bar{o}$ (as in boat), $\bar{u}$ (as in boot), and $\bar{e}$ (as in feet) were spoken by two different people (three for $\bar{e}$), and the best-match correlation patterns were determined. These are given in Table 7-1.

Table 7-1

Best-Match Correlation Patterns

Subject 1																
$\bar{o}_1$	+	-	-	+	+	-	+	+	-	+	+	-	+	-	-	-
$\bar{u}_1$	+	+	-	-	+	+	-	-	+	+	-	+	+	-	-	+
$\bar{e}_1$	-	-	-	+	+	+	-	-	-	+	+	+	+	+	+	+

Table 7-1 (Cont)

Subject 2															
$\bar{o}_2$	+	+	-	+	+	-	+	+	-	+	-	-	+	+	-	+
$\bar{u}_2$	+	+	-	-	+	+	-	-	+	+	-	+	+	-	-	+
$\bar{e}_2$	-	-	+	+	+	-	-	-	+	+	+	+	-	+	+	+
Subject 3																
$\bar{e}_3$	-	-	-	+	+	-	-	-	+	+	+	+	+	+	+	+

The resulting patterns were then cross-correlated, yielding the results given in Table 7-2.

Table 7-2

Cross-Correlated Patterns

	$\bar{o}_1$	$\bar{o}_2$	$\bar{u}_1$	$\bar{u}_2$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_3$
$\bar{o}_1$	1.00	0.50	-.12	-.12	0	-.25	0
$\bar{o}_2$	0.50	1.00	+.12	+.12	0	-.25	0
$\bar{u}_1$	-.12	+.12	1.00	1.00	+.12	-.12	+.12
$\bar{u}_2$	-.12	+.12	1.00	1.00	+.12	-.12	+.12
$\bar{e}_1$	0	0	+.12	+.12	1.00	0.50	0.75
$\bar{e}_2$	-.25	-.25	-.12	-.12	0.50	1.00	0.75
$\bar{e}_3$	0	0	+.12	+.12	0.75	0.75	1.00

Note: Dotted squares include the cross-correlation values for the same phonemes spoken by the various individuals. Values outside the squares are the cross-correlation values for different phonemes.

Those cross-correlations which ought to be high (cross-correlations between patterns for the same phoneme as spoken by different people) are all above +0.50. Those which ought to be low (cross-correlation between different phonemes) are all below +0.12.

7.5 The Use of Optical Correlation for Pattern Recognition

7.5.1 If we instrument some sort of a cross-correlator which is able to recognize a specific pattern (in the time-varying speech signal) when it appears at some standard level, we have no particular trouble in recognizing it at any other level. A change in signal level changes the outputs of all the cross-correlation, but the highest output still appears at the proper place. Pitch variation, however, poses a more complicated problem, because each of the patterns undergoes a change in scale along the time axis when the basic speech frequency is altered. This problem is met by the pitch extractor and variable-rate sampler in one of the approaches described in sub-section 7.4. We now offer an alternative which is, we think, simpler.

7.5.2 Briefly, we make use of optical cross-correlators, which we will describe, to identify patterns in the short-term power spectrum of the speech signal. If the spectrum is properly displayed at the correlator input, a change in basic pitch results in a displacement of the pattern, rather than a change in scale, and a single cross-correlator can recognize a pattern even when the basic pitch is changed. The same cross-correlator also extracts pitch information.

7.5.3 Suppose that a one-dimensional intensity pattern is displayed (on a cathode ray tube, for example) and viewed by a photosensor through a reference transparency which duplicates the pattern (being clear where the intensity pattern is bright, and opaque where the pattern is dark). We have the situation shown in Figure 7-1. The brightness at a is multiplied by the transmissivity at a', the brightness at b is multiplied by the transmissivity at b', and the total amount of light falling on the detector is simply the cross-correlation (un-normalized) of the intensity pattern with the reference

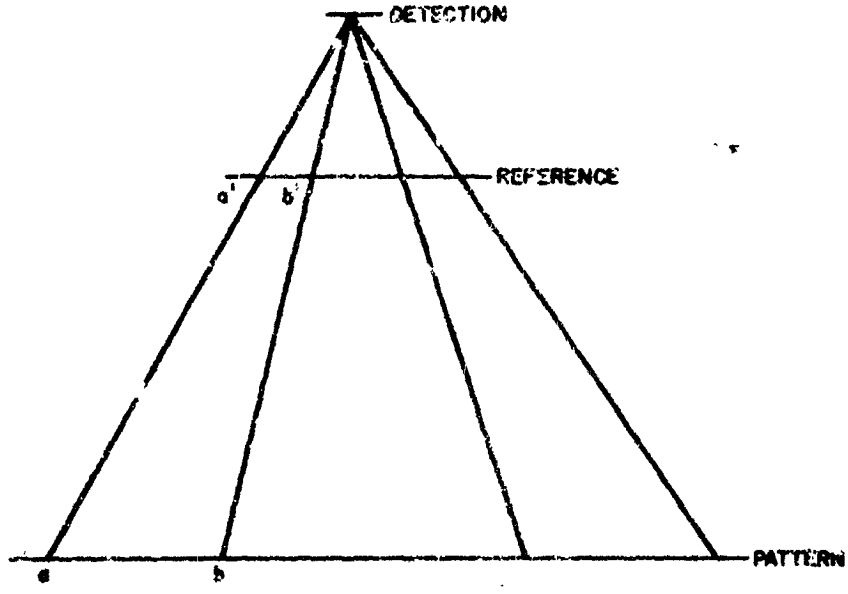


Figure 7-1. Cross-Correlation of Intensity Pattern with Reference Transparency

transparency. In the plane of the detector, we will have a bright spot indicating a high cross-correlation. Suppose, now, that the pattern is shifted to the left. The correlation spot, as can be seen by inspecting the figure, will move to the right.

7.5.4 If the pattern is based on a short-term power spectrum, with intensity proportional to power level at each frequency, and if distance along the pattern is proportional to the logarithm of frequency, then a change in basic pitch merely shifts the pattern to the right or left. A set of detectors, all using the same reference transparency, can then recognize the pattern whenever it appears (whatever the basic pitch period) and, can simultaneously extract pitch information.

7.5.5 In the system we propose, the speech signal is first preconditioned (band-limited, level-adjusted, and peak-clipped, as described in paragraph 7.4.2). It then goes to a filter bank, and the output of each filter is used to light a neon lamp. The set of lamps displays the short-term power spectrum. Simultaneously, a level-extractor determines the short-term speech level. The set of lamps is viewed, through a single set of reference transparencies (one for each spectrum pattern) by a two-dimensional array of photosensors. A pattern identifier circuit now selects the photosensor having the highest output. At the receiver, a broad-band source feeds a similar filter bank. After some smoothing, the pattern identifier signal, which includes pitch information, controls modulators at the filter outputs. The modulator outputs are summed and again modulated as called for by the smoothed short-term level signal. The result should be natural speech.

7.5.6 The source at the receiving end may be either a periodic source rich in harmonics or a white noise source. It may be advantageous to follow vocoder practice and have several sources available, switching from one to another as required for a particular pattern. The number of filters required will not be excessive. Fifty filters, each having a band-width equal to 10% of its center frequency (a Q of 10), will cover a range from

about 80 cps to about 10,500, which exceeds what we specified for "excellent" speech (paragraph 2.2.2). With a Q of 7, fewer than 35 filters are required.

7 5.7 Aside from uncertainty in the filter configuration, this approach is straightforward and well within the state of present hardware art. As a preliminary experiment, we would construct a pair of filter banks to determine a suitable configuration. Each filter in the analyzer bank would be followed by a rectifier and smoothing circuit. The output of each channel would then be used directly to modulate the output of the corresponding filter in the synthesizer bank. The success of this experiment would demonstrate the capability of the system to reproduce natural speech (without compression). Following this, the pattern recognition portions could be built and suitable patterns could be discovered on a trial and error basis. Note that further refinement would be possible (affording further compression) by appropriate coding of the pattern identification signals, or by the use of super-symbols to describe strings of pattern identification signals.

7 5.8 The optical cross-correlation technique is easily extended to two-dimensional patterns, using a two-dimensional display and two-dimensional reference transparencies. This approach might provide a useful way of identifying transitional patterns which appear between sustained speech sounds and which will appear as patterns having relatively short persistence.

Another possible application of two-dimensional pattern matching is in the identification of the patterns described in sub-section 6.4.

7.6 Non-Acoustical Constraints on the Processing of Phonological Data

7.6.1 Often in the encoding, transmitting, and decoding of phonological data by electronic-mechanical means, linguistic information is lost or seriously distorted. At this stage in the history of linguistic research, it

is worth considering whether non-acoustic, or rather non-phonemic, constraints can aid in the processes of: (a) verifying our current interpretation of a piece of phonological data, (b) predicting some phonetic feature, and (c) reconstructing obliterated data.

There are at least three classes of non-phonemic constraints that, to some degree or other, affect the quality of the analysis, the interpretation, and the synthesis of phonological data. These classes are: (a) the statistical constraints of the linguistic codes, (b) the so-called paralinguistic constraints, and (c) the higher level, supra-phonemic, linguistic constraints.

7.6.2 It is believed by most students of language, whether the language is in the written or verbal form, that natural languages are to some extent regular; that is, that they follow certain predictable patterns in their concatenation of sounds, in the structure of their morphemes in the rules of their syntax, etc. However, language being a human institution, alas, follows very complicated rules which are often difficult to perceive (as 2500 years of unfulfilled grammatical research and scholarship suggests). Within the last hundred years, linguists and others interested in language (cryptographers, psychologists, engineers, etc.) have sought more formal methods for dealing with the "irregularities" within the regular system.

7.6.3 The most successful formal method to date has been the application of statistical techniques to language structures. New statistical methods and techniques, plus the digital computer, offer promise that significant headway can be made in this area of linguistic research. Furthermore, there is a growing interest in the statistics of individual idiolects (or personal speech habits) as contrasted with the abstract patterns of a dialect or speech community. For example, it might be more economical or more efficient to tailor a space communication system to one or a few individual astronauts. In this way, the individual peculiarities of idiolectical speech patterns could be taken into consideration rather than being discarded as they usually are when a system is designed for general use.

7.6.4 To be of any practical use, statistical information about language structures must be juxtaposed against some standard or norm. Information about individual speech habits is useful only if contrasted with general expectations about the "code" characteristics of a language. Also, a certain arrangement of symbols is predictable or "normal" only if detailed knowledge about "normal" patterns is available. Thus, it becomes necessary that we continue to acquire information and statistical data about dialectical linguistic behavior: that is, language patterns of the community. At the same time it is obvious that individual speech patterns are essential to the design of specially tailored communications systems. The extra cost of a "custom" system would probably be off-set by overall improvement in the efficiency and reliability of such a system.

Things that should be investigated, then, are the techniques and any special formalisms that might be needed to examine and properly record highly detailed, personal speech patterns. In addition, the speech habits of an individual subject should be examined under various "normal" and "abnormal" states. For example, it is reasonable to assume that the language of a person will reflect his physical and psychological states since his language structures are bound up with his overall behavior patterns and may be the most sensitive indicators of how he "feels" — or what he thinks or fears. Knowing, for example, that an individual pilot or astronaut distorts his regular sentence syntax in certain ways might indicate that he has certain fears before he expresses them overtly.

7.6.5 A knowledge of non-phonemic constraint may be of great assistance in aiding a speech processing device to resolve "difficult" decisions as it carries out its task of speech-unit recognition. To the extent that the processor is tailored to a particular person, or to a small group of people, the value of such assistance will be increased. In addition, a knowledge of such constraints will supply us with para-linguistic information to supplement what is available from physical sensors. A study of such constraints, and of their statistics might well start with a detailed examination of tapes from the various Mercury flights.

7.7 High-Level Linguistic Constraints

7.7.1 There is a relatively safe, explored area of language study which lies between the rocks of phonology, idiolectical linguistics, and acoustical physics, and the whirlpools of paralinguistics and pragmatics. This is the traditional portion of linguistics which includes morphology, syntax, and some aspects of semantics. It is in these areas that modern linguistics has made the greatest headway. It is also in these areas we can expect to obtain sufficient data for constructing special sets of constraints which will aid in the process of mechanically analyzing, verifying, and synthesizing transmitted voice over long distances. Naturally, there is still much we need to know about the structures of morphology, syntax, and semantics in English, especially in light of the singular problems which will attend deep-space voice communications.

7.7.2 A knowledge of such constraints will be of the utmost importance as soon as effective speech compression is achieved at compression ratios near the theoretical limit. In the limiting case, the compressed speech is stripped of all (or nearly all) redundancy at the speech-unit level (whatever speech-units are used). Accordingly, a system using such compressed speech has no way of coping with additive noise at the speech-unit level, just as a teletype system has no way of coping with noise at the single-symbol (letter) level. In both cases, the system will make an error in the presence of excessive noise and will deliver an erroneous output. In the teletype case, errors are corrected by using a knowledge of the inter-symbol constraints (a knowledge of spelling). A knowledge of high-level linguistic constraints will serve the same purpose in correcting errors in the output of a low-redundancy speech transmission system.

7.7.3 Morphological Constraints - Morphology can be characterized as the study of the formation of minimal meaning units. We usually think of two kinds of morphemes: bound and free. In traditional terms the bound morphemes are called "affixes"; the free morphemes are called "stems". The combining of the various morphemes into words follows very strict and fairly well-known rules. The formation of plurals of nouns in English, for

example, is an automatic process in both the spoken and written language. Exluding the dozen or so exceptions, we simply add (-s) to the written form of an English noun to make it plural. In the spoken language we add one of three affixes (-z), (-s), (iz). The choice of the proper one depends in an invariant way on the final sound of the singular noun we are to pluralize. Similar such rules are part of the process for building compound words, participles, verbals, and many other word classes.

We will be able to take advantage of this regularity in constructing our set of constraints for processing language on the word or minimal segment level. For the purposes at hand, there is also much we would like to know about distinctive digraphs and trigraphs and their relation to the affixes of English. This is an area which straddles both English statistics and English morphology. By establishing sets of rules about word formation we will be able to reconstruct garbled word endings and even garbled stems if syntactic constraints are also available to the synthesizer. For example, knowing the grammatical function of a word we will be able to make a reasonable guess about its obliterated sections. This is obvious in the sentence:

He is go _ _ _ to the market.

where the obliterated word must be the present participle, and the affix for the present participle in English is invariably (-ing).

It is inevitable, in such special communications systems as we envisage, that there will also be problems of special languages or cant. That is, the nature of the information to be discussed lends itself to the formation of special words, special phrases, and maybe even a special syntax. This is natural in any specialized professional jargon. We must then be prepared to deal in our statistics and our rules of morphology with problems of cant or jargon. This will add more complexity to our basic morphological constraints, but it cannot be avoided if we are to develop a realistic and workable system. Thus, we must continue to seek out the rules of standard word formation in English, and in addition make allowances for special constructions of cant.

7.7.4 Syntactic Constraints - As was obvious from the discussion of morphological constraints, there is no clear-cut distinction between morphology, syntax, semantics, or even phonology These academic or pedagogical classifications are useful only in teaching elementary concepts of linguistics or in dividing up the linguistic labors. More often than not, they cloud the reality that language is a complex system, with each of its "parts" interdependent on the others. Nevertheless, we will discuss syntactic constraints as if they were independent of the rest of language.

In English the role of syntax, that is, morpheme or word ordering in the speech stream, is much more important than in other, highly inflected languages. In fact, it was only recently that language scholars have acknowledged the enormous richness of English syntax or grammar. This area of language could yield many useful pieces of data with which we could construct powerful constraints, but much work remains to be done, expecially in the field of personal or idiolectical syntax.

One of the places one might start in discovering useful constraints for our special speech problems would be in the sentence types and their variations. It is recognized that there are only about a dozen basic sentence forms. The most common is:

1. Subject - verb - direct object
   The man hit the ball.

Another is:

2. Subject - to be - adjective
   The girl is happy.

There are, however, many variations to these basic types. By adding modifiers to the nouns, verbs, and to the sentence as a whole, these basic types can be extended almost infinitely. The basic sentence structure, however, is usually easily recognized in spite of the variations. There are also transformations one can perform on these basic types to yield different variations for interrogations, inperatives, passive constructions, etc. For example, sentence 1. can easily be converted into a question by adding "did" and changing the intonation of the spoken sentence:

Did the man hit the ball?

We can also put Sentence 1 in the passive voice by a simple transformation:

The ball was hit by the man.

Discovering these basic sentence types and their variations and transformations has been the object of much work in the last few years.

This work should make constructing syntactic constraints a realistic undertaking. One would hope to take advantage of this work and modify it for the special problems of space communications. For example, in our examination of the voice exchanges between Walter Shirra, Jr. in MA-8 and his ground communicators, we noticed many deviations from the standard syntactic patterns. Are these accepted transformations upon the basic sentence types or are they abnormalities that reflect a linguistic "mistake" of an idiolectical form? Maybe neither case is true. It is possible that pilot and astronaut conversations would create special, but, under the circumstances predictable forms. These would have to be discovered and codified in a thorough description of the sentence norms.

Within the sentence types themselves, one may find invariant substructures in the phrases. There are classes of words in English that are very important in keeping our syntax clear and unambiguous. These words, the so-called structural constants, also deserve attention. Examples of these words are the noun markers: the, a, an, etc; the adjective markers: very, rather, quite, etc.; the verb markers: could, may, will to be, to have, etc. and many other groups — usually small in class size. These frequently used words are really syntax markers and rules of their use constitute an important area for study.

Many speakers have their own structural markers in addition to these standard syntax indicators. We would like to know what these are for any particular speaker, how he uses the, how he forms them, and why they are used in place of, or in addition to, the standard markers.

7.7.5 In addition to the morphological and syntactic constraints, it is possible that contextual knowledge could be useful in the processing of language. This represents a higher order of language processing and at

the present little is known about contextual and semantic structures. Again, it is possible that individual speakers could be examined for contextual pecularities that might be built into a communication system as we have suggested. In fact, it may turn out that "context" and "semantics" may have meaning only on an idiolectical level. All these areas of study must somehow be taken into account, if the automatic processing of speech is to be seriously entertained. Again, we suggest that a good point of departure is a detailed examination of the Mercury tape.

7.8 Paralinguistic Information

7.8.1 The recent interest in communications problems by psychologists, mathematicians, engineers, and scientists of all diciplines, has brought new ideas, new conceptual tools, and new ways of viewing old facts to bear on communications problems in general and on linguistics in particular. It is currently popular to look upon man-to-man and man-to-machine communications from a "systems" point of view. It seems reasonable to assume that an exchange of information between two human beings in an extremely complex activity, involving much more than uttering and perceiving patterns of sounds. Each member of the communication system is a member of many environments and we know that these environments interact in complicated fashions. In vis-a-vis speech, for example, the visual gestures and movements of the speakers may turn out to be vital for a complete understanding of human "speech".

7.8.2 "Paralinguistics" is a term usually given for the study of all those accompanying non-auditory, observable, mostly muscular, activities that parallel the acoustic behavior in "speech". Paralinguistics includes such types of behavior as facial expressions, hand and body gestures, body positions and stances, and presumably all other muscular or kinesthetic activity. Other physiological behavior must certainly be relevant also. For example, it is a common, everyday experience to find the meaning of a sentence spoken by someone completely reversed or modified in some subtle way by a smile or other "expression" on the speakers face. The entire activities of teasing, satire, sarcasm, under-statements,

bombast, etc. are almost always accompanied by visual gestures which parallel and modify the interpretation of a "standard" speech pattern.

Whether the speakers are male or female, young or old, have different social, cultural, education backgrounds are also to be included as paralinguistic factors. Even the profession of the speaker; whether he has a cold; a physical impediment; a linguistic tic; is homosexual; under drugs; on alcohol; is fatigued; under psychological stress; and dozens of other factors, are all possible constraints that could influence the interpretation of the speaker's phonological output.

7.8.3 In the problem to which we address ourselves, that of voice communication from deep space, we will find that many of these non-linguistic conditions, states, or activities will play a crucial role in developing an economical communication system.

Because of the nature of the space communication problem, that is, the fact that the astronauts will be few in number and greatly probed and studied, we can take advantage of their exclusiveness and contemplate a communications systems designed to fit their personal speech idiosyncrasies. In such systems, not only do we need to know as much as we can about their personal speech patterns, but, if possible, we would like to correlate their speech with their other individual activities, i.e., pulse, respiration, temperature, and other outputs from sensors attached to each astronaut. These paralinguistic activities could provide us with possible constraints to either interpret garbled data or reinterpret what is spoken but not meant.

7.8.4 Obviously this aspect of linguistics is in its infancy. But the weight of evidence seems to indicate that linguists should begin to consider this aspect of language before a complete description of human communication can be given.

In sub-section 6.1, we indicated how "abnormal" speech patterns could be clues to "abnormal" pilot conditions. If we could also correlate

speech patterns with paralinguistic information, then we might be able to anticipate the astronaut's reactions or distinguish when he is hallucinating, etc. A starting point, here, would be an examination of the Mercury tape together with other records of the Mercury flights in an attempt to demonstrate useful correlation patterns.

8. CONCLUSIONS

8.1 General Remarks

In this section, we shall confine our comments to the more significant aspects of our work under the contract and the conclusions to which we were led.

8.2 Highlights of the Report

8.2.1 From our standpoint, the most important parts of the report are subsection 2.2, subsection 3.2, and the whole of section 7.

8.2.2 In subsection 2.2, we examine the overall problem of voice communications over inter-planetary distances. Using heuristic arguments, we arrive at quantitative descriptions of poor, good, and excellent speech. These descriptions, which are set forth in paragraph 2.2.4, include the band-widths and signal-to-noise ratios required for analogue speech transmission systems of various qualities. In addition, they include the bit-rate required for equivalent digital systems. On the basis of these descriptions, we arrive, by further heuristic arguments, at estimates of the theoretical limits on speech compression ratios. These limits are 300 for speech of good (telephone) quality and 1500 for speech of excellent (high-fidelity) quality. By contrast, present systems offer compression ratios of the order of 10-30:1.

8.2.3 In subsection 3.2, we examine the power required for voice communication over inter-planetary distances with both communication equipment available in 1963 and that which may be expected to be available in the post-1970 period. Our results, set forth in subsection 3.2.5, appear to us to establish the need for a drastic improvement in speech compression techniques if satisfactory voice communication is to be achieved with reasonable power levels.

8.2.4 In section 7, we offer a number of new approaches to the problem of achieving the required break-through in speech processing.

8.3 The New Approaches

8.3.1 We propose a number of (to us) new approaches to the problems of speech compression and speech processing. These are explained in detail in section 7, and are described briefly in subsection 3.6. All of them are interdisciplinary approaches, drawing on the field of coding theory, information theory, pattern matching, and linguistics. Each of them lies within the capability of General Precision, Inc., and each of them is, we think, worthy of support. They are mentioned very briefly below.

8.3.2 In subsection 7.2, we propose an approach to the restricted problem of speech compression. This is based on a coding theory. Considering sampled quantized speech as a message expressed in terms of symbols, where each quantizing level is a distinct symbol, we propose to examine the statistics of the messages, determine the degree of redundancy, and devise a translation which will recode the speech from its existing high-redundancy code to a new one of lower redundancy.

8.3.3 In subsection 7.3, we propose an approach to the more general problem of speech processing. In particular, we propose to draw on information theory in order to find the patterns, in sampled quantized speech, which an optimum speech analyzer will be called on to recognize. We do not assume that these patterns will be identical with the phonemes used by linguistics to describe speech.

8.3.4 In subsection 7.4, we propose work on a mechanical phoneme recognizer. This is a continuation of an initial effort carried out as part of our work under the contract.

8.3.5 In subsection 7.5, we propose an approach which is based on pattern-matching techniques of a special sort. Here, we propose to recognize patterns in the short-term power spectrum of the speech signal. Our instrumentation, however, eliminates the usual complications attendant on pitch-tracking.

8.3.6 In subsection 7.6, 7.7, 7.8, we propose work based on the use of linguistics as an aid to voice-communication systems. As we explained in

subsection 7.7.2, a knowledge of the linguistic aspects of speech will assume increasing importance as we achieve success in developing speech transmission systems of low redundancy; that is, those having high compression ratios. We feel, therefore, that an initial effort in this direction is justified at this time, even though our program in this area is not as well defined as the others mentioned above.

8.4 Concluding Remark

Our feelings about the general problems of speech compression can be summed up very briefly — a breakthrough will be required, and the time is ripe for a number of related attempts based on interdisciplinary approaches, such as those summarized in this section.

APPENDIX "A"

THE MINIMUM TRANSMISSION LOAD FOR A MINIMUM PERFORMANCE SYSTEM

1. We assume that the maximum rate for spoken speech does not exceed 3 words/second.

2. We assume that the average word has a length of 5 letters.

3. If we allocate a character for "space", in addition to the 26 characters used for the letters, the assumptions above imply that spoken speech (when transcribed without punctuation) involves at most 18 characters/second.

4. With a total of 27 characters, we can use brute-force coding and allocate 5 bits to each character.

5. The transmission load for messages of this sort is not greater than 90 bits/second if we neglect timing bits and other housekeeping signals.

APPENDIX "B"

THE MINIMUM TRANSMISSION LOADS FOR INTERMEDIATE AND MAXIMUM PERFORMANCE SYSTEMS

1. We assume that the output of an intermediate performance system is equivalent to that of a minimum performance system with the addition of marks indicating the nuances of expression.

2. Expression is carried by stress and intonation, which may be loosely equated with the lay (from a linguistic standpoint) terms accent and pitch. Linguists use a single stress indicator and a single intonation indicator for each syllable

3. Linguists distinguish 3 or 4 relative levels of stress. We assume that 3 bits/syllable (8 levels) will be adequate for absolute stress indication.

4. Linguists distinguish 3 or 4 relative levels of intonation. We assume that 4 bits/syllable (16 levels) will be adequate for absolute intonation indication.

5. The foregoing assumptions lead to a requirement of 7 bits/syllable for the addition of nuances.

6. Assuming a maximum rate of 6 syllables/second, the addition of nuances requires 42 bits/second.

7. Adding the figure of 42 bits/second to the 90 bits/second derived in Appendix "A" for the minimum transmission load associated with a minimum performance system, we arrive at 132 bits/second for the minimum transmission load associated with an intermediate performance system. We round this figure off to 135 bits/second.

8. To derive the minimum transmission load associated with a maximum performance system, we start with the figure of 135 bits/second and add what seems necessary for the transmission of para-speech inputs.

9. Assuming 10 non-acoustic quantities, and assigning 2 bits (4 levels) to each, we have 20 bits to describe the para-speech inputs at any instant.

10. Supposing that the para-speech inputs change not oftener than 3 times/second, we require 60 bits/second for para-speech signals.

11. Adding 60 bits/second to the figure of 135 bits/second for the intermediate performance system, we arrive at 195 bits/second for the minimum transmission load associated with a maximum performance system. We round this off to 200 bits/second.

APPENDIX "C"

TRANSMISSION LOADS FOR BAND-LIMITED SPEECH

1. In order to arrive at a figure for the transmission load (in bits/second) associated with band-limited speech it is necessary to assume a specific technique for converting the continuous speech signals into a bit stream of some sort. The technique we have assumed is described below, and is essentially similar for all three grades of speech (poor, good and excellent)

2. The acoustic speech signal is converted into an electrical speech signal (the electrical analogue of the acoustic signal) which is a time-varying voltage.

3. The electrical speech signal is band-limited by a suitable filter network.

4. The voltage level of the band-limited speech signal is adjusted, if necessary, so that its long-term RMS voltage level is $\sqrt{S}$ volts. For this purpose, we assume a manual level control for coarse adjustment and a fast-attack slow-decay AGC system for fine adjustment.

5. We now sample the speech signal at a uniform rate. The output of the sampler is then a train of pulses, some positive and some negative, uniformly spaced in time. The height of any given pulse, in volts, is equal to the voltage of the speech signal at the corresponding instant of sampling.

6. If the sampling rate exceeds twice the highest frequency present in the band-limited signal, a filtered version of the original signal can be recovered by passing the pulse train through a low-pass filter whose cut-off frequency is half of the sampling frequency. Subsequent equalization will then yield a reproduction of the band-limited signal as it appeared at the input to the sampler. On a brute-force basis, then, the minimum sampling rate is twice the upper band limit. It is a relatively simple process, however, to shift a band-limited signal along the frequency axis so that its lower band limit, after shifting, lies at a frequency close to zero. The minimum sampling

rate is then twice the band-width of the original band-limited signal. In subsequent calculations, we will assume that this frequency-shifting process, which is described in detail in the following paragraph, has been carried out before sampling. Accordingly, we take the sampling rate to be twice the band-width.

7. We digress, here, to describe the frequency shifting process. The several steps are shown in Figure C-1. The original signal, which extends from f_1 to f_2, is multiplied by a sinusoid of frequency f_o. The original signal is shown at (a) in Figure C-1, and the result after multiplication is shown at (b). We now isolate the upper side band by filtering and multiply it by a sinusoid of frequency $(f_o + f_1)$. The lower side band of the product is shown at (c) and is the frequency-shifted signal we want. The inverse frequency shift must, of course, be carried out after we reconstruct a signal from the pulse train at the receiving end of the system. This involves multiplication by a sinusoid of frequency $(f_o + f_1)$ to obtain the product shown at (d). From this, we isolate the upper side band by filtering, after which we multiply again by a sinusoid of frequency f_o to obtain the original signal, shown at e.

8. Returning to our pulse train, we now quantize the pulses. That is, we choose a finite set of M discrete voltage levels and assign a number to each pulse which identifies the standard voltage level nearest to the pulse height. The pulse train is then represented with some error, which we call quantizing noise, by a string of numbers with one number for each pulse. We can, now, transmit these numbers (in binary form) and reconstruct the quantized pulses at the receiving end. From the train of quantized pulses, we reconstruct a version of the signal as it appeared before sampling. This version differs from the true original signal by virtue of its corruption by additive quantizing noise (and by other noises which we neglect here). We must, of course, carry out a frequency-shift operation after reconstruction in order to recover a version of the original band-limited speech signal.

9. The transmission load, in bits/second, is the product of the sampling rate by the number of bits required to specify the voltage level of a

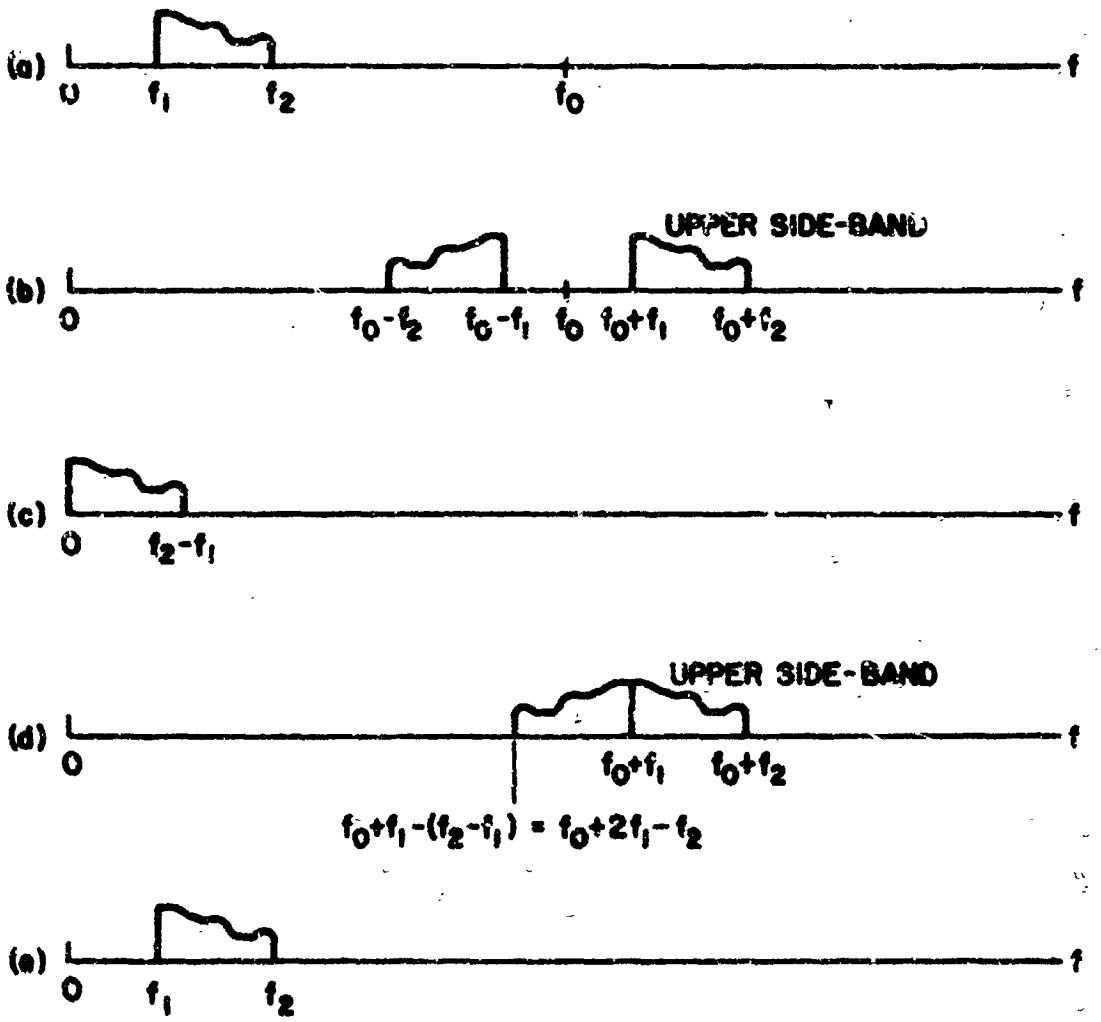


Figure C-1. Frequency Shifting Process

single pulse (after quantization). As we quantize more finely, we have more voltage levels and require more bits to identify a given voltage level. Accordingly, the transmission load goes up. At the same time, the voltage levels are closer together and we have less quantizing noise. We now look for a relation between the number of quantizing levels and the signal-to-noise power ratio of the quantized signal (considering no noise except quantizing noise).

10. Since we will ultimately code the quantized voltage levels in binary form, we start with the assumption that M, the number of levels, is an integral power of 2. We assume, further, that the quantizing levels differ by multiples of some voltage Δ so that the M voltage levels are

$$\left(-\frac{M}{2}+\frac{1}{2}\right)\Delta\ ,\ \left(-\frac{M}{2}+\frac{3}{2}\right)\Delta\ ,\ \ldots\ldots\ \left(+\frac{M}{2}-\frac{1}{2}\right)\Delta \qquad \text{(C-1)}$$

If we choose Δ reasonably small compared to $\sqrt{S}$ (see paragraph 4), the quantizing error will be uniformly distributed over the range from $(-\frac{\Delta}{2})$ to $(+\frac{\Delta}{2})$. In this case, the average error will be zero, and the mean square quantizing error will be $\Delta^2/12$. This results in a signal-to-noise (power) ratio of

$$\frac{S}{N} = \frac{12S}{\Delta^2} \qquad \text{(C-2)}$$

$$\Delta = \sqrt{\frac{12S}{(S/N)}} \qquad \text{(C-3)}$$

which defines Δ in terms of S and (S/N). The largest pulse we can accommodate without peak-clipping in excess of the postulated amount $\Delta/2$ is one whose voltage is $\pm\frac{M}{2}\Delta$. Supposing this peak voltage level to be C times the RMS level S, we have

$$\frac{M}{2}\Delta = C\sqrt{S} \qquad \text{(C-4)}$$

$$\Delta = \frac{2C\sqrt{S}}{M} \qquad \text{(C-5)}$$

and, from (C-3) and (C-5)

$$\frac{2C\sqrt{S}}{M} = \sqrt{\frac{12S}{(S/N)}} \qquad \text{(C-6)}$$

$$M = 2C\sqrt{\frac{(S/N)}{12}} = C\sqrt{\frac{(S/N)}{3}} \qquad \text{(C-7)}$$

However, by our earlier assumption,

$$M = 2^n \tag{C-8}$$

whence, from (C-7) and (C-8)

$$2^n = C\sqrt{\frac{(S/N)}{3}} \tag{C-9}$$

$$2^{2n} = (S/N)\frac{C^2}{3} \tag{C-10}$$

and

$$3 \times 2^{2n} = C^2 (S/N) \tag{C-11}$$

11. In Table C-1, below, we tabulate 3×2^{2n} for some integral values of n.

Table C-1

3×2^{2n} for Some Integral Values of n

n	3×2^{2n}	n	3×2^{2n}
1	12	9	786,432
2	48	10	3,145,728
3	192	11	12,582,912
4	768	12	50,331,648
5	3,072	13	201,326,592
6	12,288	14	805,306,368
7	49,152	15	3,221,225,472
8	196,608	16	12,884,901,888

using the values in the table, we arrive at the following values for n, (S/N), and C

Table C-2

Values for n, (S/N) and C

Quality	n	S/N	C
Poor	2	+10db	2.2
Good (min.)	5	+25db	3.1
Good (max.)	6	+30db	3.5
Excellent (min.)	11	+60db	3.5
Excellent (max.)	13	+70db	4.5

12. We next assign band limits to the various grades of speech

Table C-3

Band Limits for Various Grades of Speech

Quality	Band Limits
Poor	500 - 1,500 cps
Good (min.)	300 - 3,000 cps
Good (max.)	300 - 3,500 cps
Excellent (min.)	100 - 8,000 cps
Excellent (max.)	90 - 10,000 cps

13. Multiplying n by twice the band-width, we obtain the transmission load in bits/second.

Table C-4

Transmission Load in Bits/Second

Poor	4,000 bits/second
Good (min.)	27,000 " "
Good (max.)	38,400 " "
Excellent (min.)	173,800 " "
Excellent (max.)	257,660 " "

14. On the basis of the bit rates in Table C-4, we arrive at estimates of 30,000 bits/second for good speech and 200,000 bits/second for excellent speech. Working backwards to find band-widths and signal-to-noise ratios which correspond to these rates, we have

Table C-5

Band-Widths and Signal-to-Noise Ratios vs. Bits/Second

Quality	n	S/N	C	Band Limits	Bit Rate
Poor	2	+10db	2.2	500-1500 cps	4,000 b/s
Good	5	+25db	3.1	300-3300 cps	30,000 b/s
Excellent	11	+60db	3.5	90-9180 cps	200,000 b/s

APPENDIX "D"

OPERATING RANGES

1. Distances to the planets are conveniently measured in Astronomical Units (abbreviated AU), where one AU is the mean distance from the earth to the sun

$$\text{one AU} = 1.5 \times 10^{13} \text{ cm} \qquad \text{(D-1)}$$

For an electromagnetic signal, the round-trip time over a distance of one AU is 1,000 seconds, or almost 17 minutes.

2. Minimum and maximum distances from each of the planets to the sun are given (in AU) in Table D-1.

Table D-1

Distances from Planets to Sun

	Distance to the Sun	
Planet	Minimum	Maximum
Mercury	0.31	0.47
Venus	0.72	0.73
Earth	0.98	1.01
Mars	1.38	1.67
Jupiter	4.95	5.45
Saturn	9.00	10.07
Uranus	18.28	20.09
Neptune	29.80	30 32
Pluto	29.67	49.34

3. Neglecting orbital inclinations, we now compute the minimum and maximum distances from each planet to the earth. For the inner planets (Mercury and Venus), we use the relations

$$\text{(Min to Earth)} = 1 - \text{(Max to Sun)} \qquad \text{(D-2)}$$

$$\text{(Max to Earth)} = 1 + \text{(Max to Sun)} \qquad \text{(D-3)}$$

For the other planets, we use

$$\text{(Min to Earth)} = \text{(Min to Sun)} - 1 \qquad \text{(D-4)}$$

$$\text{(Max to Earth)} = \text{(Max to Sun)} + 1 \qquad \text{(D-5)}$$

These distances are set forth in Table D-2 below

Table D-2

Distances from Planets to Earth

	Distance to the Sun	
Planet	Minimum	Maximum
Mercury	0.53	1.47
Venus	0.27	1.73
Earth	---	---
Mars	0.33	2.67
Jupiter	3.95	6.45
Saturn	8.00	11.07
Uranus	17.28	21.09
Neptune	28.80	31.32
Pluto	26.67	50.34

4. Using the figures in Table D-2, we now put the planets into three groups. Minimum and Maximum distances for each group are given (with some round-off) in Table D-3,

Table D-3

Rounded-Off Distances of Planets to Earth

	Distance to Earth	
Planet	Minimum	Maximum
Mercury) Venus) Mars)	0.27	2.7
Jupiter) Saturn)	4.0	11.0
Uranus) Neptune) Pluto)	17.0	50.0

APPENDIX "E"

BASIC POWER CALCULATIONS

1. In this appendix, we calculate the transmitter output (in watts) required to develop a signal-to-noise ratio of unity (at the earth-based receiver) over a band-width of one cps at a distance of one Astronomical Unit.

2. The equivalent noise input power in a band-width of one cps at the receiver is kT watts, where k is Boltzmann's constant (1.4×10^{-23} watts/cps per degree Kelvin) and T is the effective noise temperature of the receiver (in degrees Kelvin). (We take a band-width of one cps.) For a signal-to-noise ratio of one, we require that

$$P_R = kT, \tag{E-1}$$

where P_R is the power (in watts) delivered to the receiver.

3. Assuming that the receiving antenna is properly aimed, we can write (for a circular antenna)

$$P_R = 0.5 A_R \phi, \tag{E-2}$$

where A_R is the area of the receiving antenna (in cm^2) and ϕ is the power density (in watts per cm^2) delivered by the transmitter. From (E-1) and (E-2),

$$\phi = \frac{kT}{0.5 A_R}. \tag{E-3}$$

4. For an isotropic radiator at the transmitter (one which radiates uniformly in all directions), we can write

$$\phi = \frac{P_I}{4\pi R^2}, \tag{E-4}$$

where P_I is the radiated power (in watts) and R is the operating range (in cm). From(E-3) and (E-4), we have

$$P_I = \frac{4\pi R^2 kT}{0.5 A_R}. \tag{E-5}$$

5. Assuming the use of a properly aimed directional transmitting antenna, we can write (for a circular antenna)

$$P_I = P_T \frac{2 \cdot A_T}{\lambda^2} \qquad \text{(E-6)}$$

where P_I is the total power input (in watts) to the directional transmitting antenna, A_T is the area (in cm^2) of the antenna, and λ is the operating wavelength (in cm). From (E-5) and (E-6), we have

$$P_T = \frac{4\lambda^2 R^2 kT}{A_T A_R} \qquad \text{(E-7)}$$

6. Finally, we write

$$P_o = QP_T \qquad \text{(E-8)}$$

where P_o is the transmitter output power (in watts) required to deliver an effective radiated power of P_T, and where Q is a dimensionless constant (greater than one) which accounts for everything we have left out:

a) RF losses between the transmitter itself and the transmitting antenna;

b) RF losses between the receiving antenna and the receiver itself;

c) Absorption and scattering by interplanetary debris;

d) Errors due to antenna deformation;

e) etc. (errors in estimating T, for example).

Using (E-7) and (E-8), we have

$$P_o = \frac{4Q\lambda^2 R^2 kT}{A_T A_R} \qquad \text{(E-9)}$$

7. We now insert the following values:

a) k is 1.4×10^{-23} watts/cps per degree Kelvin;

b) R (at one AU) is 1.5×10^{13} cm

On this basis, (D-9) becomes

$$P_o = 1.26 \times 10^4 \frac{Q\lambda^2 T}{A_T A_R} \qquad \text{(E-10)}$$

8. As representative of current practice, we take the following values:

a) λ is 30 cm, corresponding to a frequency of 10^9 cps;

b) T is 27 degrees Kelvin, of which 15 degrees is assignable to space noise and 12 degrees is assignable to the equivalent noise temperature of a Helium maser;

c) A_T is 1.13×10^4 cm^2, corresponding to a circular disk with a a diameter of 4 feet;

d) A_R is 5.1×10^6 cm^2, corresponding to a circular disk with a diameter of 85 feet.

Using these values in (E-10), we have

$$P_o = 3.54 \times 10^{-2} Q \qquad \text{(E-11)}$$

Taking 4 as a reasonable value of Q and rounding off, we have

$$P_o \approx 2 \times 10^{-2} \text{ watts} \qquad \text{(E-12)}$$

for the transmitting output required to develop a signal-to-noise ratio of unity over a band-width of one cps at a distance of one AU, using current techniques.

8. By 1970, we can anticipate that:

a) λ will be 10 cm;

b) the transmitting antenna will have a diameter of 20 feet;

c) the receiving antenna will have a diameter of 250 feet

P_o will therefore become

$$.02 \times \left(\frac{10}{30}\right)^2 \times \left(\frac{4}{20}\right)^2 \times \left(\frac{85}{250}\right)^2 \approx 10^{-5} \text{ watts} \qquad \text{(E-13)}$$

9. It is convenient to express these powers in dbw, or db relative to one watt. On this basis, the transmitter output required to develop a signal to noise ratio of unity (at the receiver) over a band-width of one cps at a distance of one AU is

$$\left.\begin{aligned} P_o &= -17 \text{ dbw, } 1963 \\ &= -50 \text{ dbw, } 1970 \end{aligned}\right\} \qquad \text{(E-14)}$$

APPENDIX "F"

POWER REQUIREMENT FOR A GIVEN BIT-RATE

1. According to Shannon, the capacity of a noisy channel is given by

$$C = B \log_2 (1+S/N) , \qquad (F-1)$$

where C is the channel capacity (for error-free transmission with optimum coding) in bits/second.

B is the band-width of the channel in the single-ended frequency domain, in cps.

S is the signal power spectral density in watts/cps.

N is the noise power spectral density in watts/cps.

2. It follows from (F-1) that

$$\frac{C}{B} = \log_2 (1+S/N) . \qquad (F-2)$$

$$2^{\frac{C}{B}} = 1+S/N . \qquad (F-3)$$

and

$$e^{\frac{C}{B}\ln 2} = 1+S/N . \qquad (F-4)$$

3. If B is much larger than C ln 2, we can write

$$e^{\frac{C}{B}\ln 2} = 1+\frac{C}{B}\ln 2 , \qquad (F-5)$$

and, from (F-4) and (F-5),

$$\frac{C}{B}\ln 2 = \frac{S}{N} \qquad (F-6)$$

4. From (F-6),

$$B S = C N \ln 2 = 0.7 C N , \qquad (F-7)$$

but

$$B S = P ,$$

where P is the transmitter power. Hence

$$P = 0.7\,C\,N, \qquad (F\text{-}9)$$

and the required transmitter power is 0.7 C times that required to develop a signal-to-noise ratio of one over a band-width of one cps.

5. Although Shannon codes exist in principle, they have not, in general, been discovered. It seems reasonable, however, to suppose that we can achieve an acceptably low error rate if

$$P = C\,N. \qquad (F\text{-}10)$$

APPENDIX "G"

THE OBJECTIVE MEASUREMENT OF SPEECH INTELLIGIBILITY

1. There are two general procedures for the measurement of intelligibility in speech communication systems. One, which is used by design engineers, is an analytical procedure which calculates a predictive measure of intelligibility. In general this method is rather complex, and has not been used to any great extent in evaluating speech comprehension systems since it tends to break down when used with complex speech processing systems. A more complete discussion of this method can be found in reference 38.

   The second method involves measurement of intelligibility through direct testing. This procedure will be briefly described here. (Refer to Human Engineering Guide to Equipment Design [39] for more detailed discussions.)

2. Intelligibility testing requires careful control of laboratory conditions. There have been disagreements between contractors test results on their equipment and objective comparative tests such as are reported in reference 9. Much of the disagreement can be attributed to variation in test conditions, size and type of test vocabularies and training of the listener.

3. One of the most generally used measures of intelligibility is PB word intelligibility, which is referred to throughout this report. PB words are drawn from Phonetically Balanced word lists. These are monosyllabic word lists in which the frequencies of occurrence of the various fundamental speech sounds are proportional to their frequencies of occurrence in everyday speech. The following list is typical.

Table G-1

Phonetically Balanced (PB) Word List

1.	gill	11.	gloss	21.	cloud	31.	corpse	41.	bounce
2.	suck	12.	hire	22.	scythe	32.	bait	42.	bud
3.	perk	13.	bought	23.	blush	33.	job	43.	frog
4.	fate	14.	dab	24.	shoe	34.	hit	44.	quart
5.	five	15.	earl	25.	snuff	35.	hock	45.	rap
6.	need	16.	bean	26.	moose	36.	niece	46.	charge
7.	pick	17.	nut	27.	mute	37.	tan	47.	sludge
8.	log	18.	ways	28.	rib	38.	vast	48.	tang
9.	nab	19.	wish	29.	awe	39.	our	49.	them
10.	else	20.	pit	30.	trash	40.	start	50.	vamp

In addition, other test materials can be used to define an intelligibility index. Nonsense syllables, which are random combinations of fundamental speech sounds in the pattern consonant-vowel-consonant, and sentences which are drawn from a large ensemble can be used. Figure G-1 indicates the intelligibility of various forms of test material versus the articulation index (AI). (Unless otherwise noted, 1000 different PB words is the basic index used throughout this report.)

4. AI is a predictive measure of intelligibility, and for face-to-face communication Table G-2 is an approximate representation of the relationship between AI and speech communication. By use of Table G-2 and Figure G-1, one sees that a PB word score of 85% indicates a satisfactory communication system. This relationship is only approximate, however, and the possible unnatural and harsh-sounding outputs of some vocoders may influence the communication acceptability of these systems so that higher AI's are required for satisfactory communication.

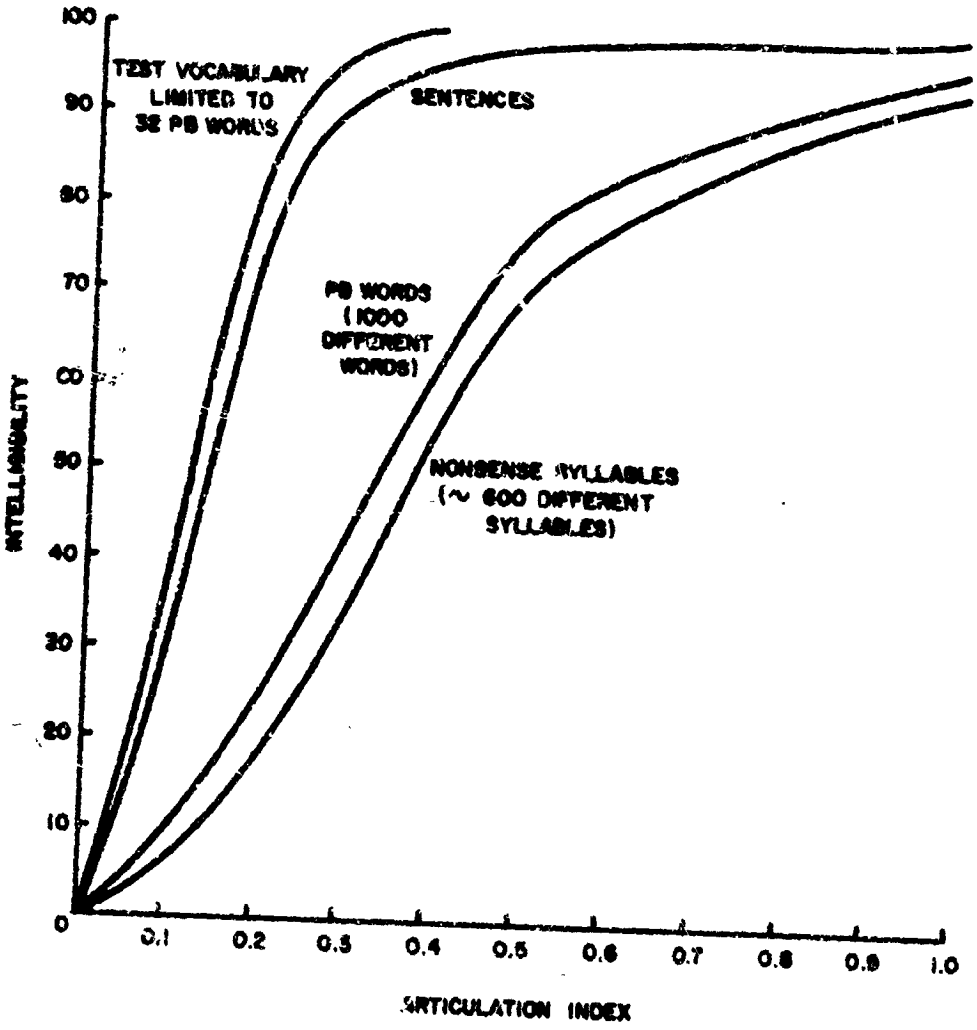


Figure G-1. Intelligibility vs Articulation Index for Various Kinds of Test Materials

Table G-2

Intelligibility Criteria

An AI of.....	Provides communications.....
0.7 to 1.0	Satisfactory to excellent
0.3 to 0.7	Slightly difficult to satisfactory - up to 98% of sentences are heard correctly
0.0 to 0.3	Impossible to difficult - special vocabularies and radio-telephone voice procedures are required

BIBLIOGRAPHY

NOTE: A.D. refers to Astia Document Number

1. Stevens, K.N., Hecker, M.H.L. and Kryter, K., An Evaluation of Speech Compression Systems, (RADC TDR 62-171) A.D. 277,457, 1 March 1962.

2. Bolt, Beranek and Newman, Inc., Study of a Speech Compression System (Spectrum Selection); Quarterly Technical Report No. 8 (Contract DA-36-039-SC-78078)A.D. 274,466, 1961.

3. Stone, R.B and White, G.M., Digital Correlation Defects Voice Fundamental, Electronics, Nov. 22, 1963.

4. Houde, R.A., Larkin, W.D., Stewart, L.C. and Wolff, R.L., Formant Vocoder Speech Compression System, Final Report, General Dynamics Electronics (Contract DA-36-039-SC-87252), 1 July 1961 - 30 April 1963.

5. Coulter, D.C., Early, D.M. and Irons, R.E., A Digitized Speech Compression System, Melpar, Inc. (ASD TR 61-494), (Contract AF 33 (600) 39962), August 1959, Sept. 1961.

6. Gunnar Fant, Acoustic Theory of Speech Production, Mouton and Co., S' Gravenhage, 1960.

7. Flanagan, J.L., Perceptual Criteria in Speech Processing, Proceedings of Speech Communications Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

8. Stevens, K.N., Toward a Model for Speech Recognition, J. Acoust. Soc. Am. 32, 47-51, 1960.

9. Heinz, J.M., An Analysis of Speech Spectra in Terms of a Model of Articulation, MIT, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

10. House, A.S., Paul, A.P., Stevens, K.N. and Arnold, Jane B., Acoustical Description of Syllabic Nuclei, Data Derived by Automatic Analysis Procedures, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

11. House, A.S., Stevens, K.N. and Paul, A.P., Acoustical Description of Syllabic Nuclei, An Interpretation in Terms of a Dynamic Model of Articulation, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

12. Inomata, S., Program for Active Segmentation and Reduction of Phonetic Parameters, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

13. Stevens, K.N., Hecker, M.H.L. and Kryter, K.D., Evaluation of Some Speech Bandwidth Compression Systems, Bolt, Beranek and Newman, Inc., Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

14. Kryter, K.D. and Ball, J.H., An Evaluation of Speech Compression Techniques, Bolt, Beranek and Newman, Inc., Contract No. AF 30 (602) - 2235, A.D. 402604.

15. Schroeder, M. R., Abstracts of the Stockholm Speech Communications Seminar, Acoust. Soc. Am. 35 pp 1116 and 1117, July, 1963.

16. Smith, C P., A Method for Speech Data Processing By Means of a Digital Computer, ERD-CRRS-TM-58-103 AFCRC (1958).

17. Smith, C. P., An Approach to Speech Bandwidth Compression, Proceedings of Seminar on Speech Compression and Processing, AFCRC-TR-59-198, AFCRL (1959).

18. Rosenblatt, F., Applications of Perceptions to Speech Recognition, Cornell University, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

19. Fant, G., Fintoft, K. Liljenciants, J., Linblom, B. and Mártony, J., Formant Amplitude Measurements, J. Acoust. Soc. Am 35-11, November 1963.

20. Liljenciants, J., A Few Experiments on Voiced - Voiceless Identification and Time Segmentation of Speech, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

21. Fujimura, O., Formant - Antiformant Structure of Nasal Murmers, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

22. Mathews, M. V., Miller, Joan E. and David, F. E., Jr., Strategies for Automatic Pole-Zero Analysis of Speech, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

23. Fant, G., Speech analysis and Synthesis, Final Report, Contract AF 61 (052-342) AD 298 309.

24. Edie, J., Sebestyen, G., Voice Identification General Criteria, Litton Systems, Inc. Contract AF 30 (602) - 2499, A.D. 278, 565.

25. Meeker, W.F. and Green, L.S., Voice-to-Teletype Converter Research Program, R.C.A. Contract AF 33-(616) - 6991 AD 272 024.

26. R.C.A., Study of Speech Compression Systems (Syllabic Compression Techniques) Quarterly Reports, No. 1 and No. 2 (Contract DA 36-039-SC-89191), A.D. 284, 811 and 406365, 15 April to 15 October 1962.

27. Linblom, B., Spectrographic Study of Vowel Reduction. J. Acoust. Soc. Am 35-11, November 1963.

28. Liberman, A.M., Cooper, F.S., Harris, K.S. and MacNeilage, P.F., Motor Theory of Speech Perception, Haskins Laboratories, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

29. Liberman, A.M. et al. Minimal Rules for Synthesizing Speech, J. Acoust. Soc. Am 31, 1490, 1959.

30. Estes, S.E., Kerby, H.R., Maxey, H.D., Walker, R.M., Speech Synthesis from Stored Data, IBM Journal, January 1964.

31. Sivertsen, E. and Peterson, G.E., Studies on Speech Synthesis, University of Michigan, Contract NONr 1224 (22) NR 049-122, A.D. 245354, August 1960.

32. Sugimoto, T. and Hashimoto, S. - The Voice Fundamental Pitch and Formant Tracking Computer Program by Short Term Auto-correlation Function, Proceeding of Speech Communication Seminar, Stockholm, Sweden, August 29 - September 1, 1962.

33. Liberman, P., Vowel Intonation Contours, Quarterly Progress Report, Research Laboratory for Electronics, MIT, Cambridge, July 15, 1958.

34. Liberman, P., Perturbations in Vocal Pitch, J. Acoust. Society of Am. 33 597-603 (1961).

35. Flanagan, J. L., Some Influences of the Glottal Wave Upon Vowel Quality, Proc, of IV International Congress on Phonetics Sciences, Helsinki, Finland, September 1961.

36. Hillix, W. A., Use of Two Non-Acoustic Measures in Computer Recognition of Spoken Digits, J. Acoust. Soc. of Am. 35 12, 1978 - 1984, December 1963.

37. Barton, G. W. and Barton, S. H. - Forms of Sounds as Shown on an Oscilloscope by Roulette Figures, Science, December 11, 1963, Vol. 142, 3598.

38. Kryter, K. D., On Predicting the Intelligibility of Speech From Acoustical Measures, J. of Speech and Hearing Disorders, Vol. 21 No. 2, June 1956.

39. Morgan, C. T., Cook, J. S., Chapanis, A. and Lund, M. U. Editors, Human Engineering Guide to Equipment Design, McGraw-Hill, New York, 1963.

40. Gleason, H. A., Jr., An Introduction to Descriptive Linguistics, New York, 1955.

41. Hockett, C. F., A Course in Modern Linguistics, New York, 1958.

42. Saporta, S., Psycholinguistics, Holt Rinehart and Winston, New York, 1962.

43. Pike, K. L., Phonemics, University of Michigan, Ann Arbor, Michigan, 1947.

44. Nida, Morphology, University of Michigan, Ann Arbor, Michigan, 1949.

45. Bloomfield, L., Language, Holt, New York, 1933.

46. Fry, D. B., Experiments in the Perception of Stress, Language and Speech, 1, 126 1958.

47. Peterson, G. E., Some Observations on Speech, Quarterly Journal of Speech, 44, 402, 1958.

48. Peterson, G. E., The Information Bearing Elements of Speech, Bell Telephone Laboratories, Murrey Hill, N. J., AD 122, 685.

49. Delattre, P. C., Liberman, A. M. and Cooper, F. S., Acoustic Loci and Transitional Clues for Consonants, J. Acoust. Soc. Am. 27, 769-773 (1955).

50. Miller, G. A. and Nicely, P. E., Analysis of Perceptual Confusions Among English Consonants, J. Acoust. Soc. Am. 27, 338-352 (1955).

51. Peterson, G.E. and Barney, H.L., Control Methods Used In a Study of the Vowels, J. Acoust. Soc. Am., 24 175-184 (1952).

52. Chomsky, N., Halle, M. and Lakoff, F., On Accent and Juncture in English, For Roman Jakobson, The Hague; Mouton, A.D. 127, 511, 1956.

53. Proceedings of Seminar on Speech Compression and Processing AFCRC-TR-59-198 Vol. 1 and 2, 29 - 30 September 1959.

54. Proceedings of the Speech Communications Seminar, Stockholm, Vol. 1 and 2, August 29 - September 1, 1962.

55. Automatic Speech Recognition, University of Michigan Engineering Summer Conferences, Printed Notes for Course Summer, Summer, 1963.

56. Denes, P., Automatic Speech Recognition: Experiments Using Linguistic Statistics, Contract AF51(514) - 1176, AD 255978, September 1960.

57. Fry, D.B. and Denes, P., On Presenting the Output of a Mechanical Speech Recognizer, J. Acoust. Soc. Am 29 364-7, 1957.

58. Shannon, C.E. and Weaver, W., The Mathematical Theory of Communication, Urbana, University of Illinois Press, 1949.