AD-A055 236 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCH--ETC F/G 5/8
AN ALGORITHM FOR DETERMINING SPEECH INTELLIGIBILITY.(U)
DEC 77 W R BEESON
UNCLASSIFIED AFIT/GE/EE/77-9 NL

| OF |

AD
A055236

END
DATE
FILMED
7-78
DDC

AFIT/GE/EE/77-9

1

6 AN ALGORITHM FOR DETERMINING SPEECH INTELLIGIBILITY,

THESIS

9 Master's thesis,

14 AFIT/GE/EE/77-9

10 Wayne R. Beeson
Captain USAF

11 Dec 77

16 7071

12 63p.

17 ΦΦ

Approved for public release; distribution unlimited

78 06 13 087

012 225

CL

AFIT/GE/EE/77-9

AN ALGORITHM FOR DETERMINING SPEECH INTELLIGIBILITY

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

by

Wayne R. Beeson, B.S.

Captain USAF

Graduate Electronic Engineering

December 1977

Approved for public release; distribution unlimited.

Preface

This topic was selected because of a continuing need in the R&D community to make an intelligibility evaluation of experimental and prototype voice communications systems. Because the Air Force does not have a more automated way to test intelligibility that produces adequate accuracy and is relatively easy to use, they are still using human listener panels to make these determinations. I have the feeling that "there must be a better way" to make these tests.

It seems reasonable that if present state-of-the-art digital computer techniques can synthesize speech, it should be possible to determine the intelligibility of speech using computer processing. Hopefully the approach used in this thesis will provide at least a basis for development of a computerized method for measuring intelligibility that will prove to be sufficiently accurate and simple to replace the human listener method. The ultimate development of this type technique would provide a device which could be hooked to the communications system under test and have a meter which would indicate the intelligibility of the system on a real-time basis.

I am indebted to Major Joe Carl, my advisor, for his guidance, suggestions, advice, and encouragement during the preparation of this thesis. I would like to express my appreciation to Captain Mazzie and Mr. William Hall, Jr. of the Analog/Hybrid Systems Branch of the ASD Computer Center for the many hours they spent on the preliminary processing of the analog speech data. My thanks to Dr. Oestreicher and Richard McKinley of the Aerospace Medical Research Laboratory for allowing me to use their anechoic chamber to prepare the voice data tapes and their computer terminal to develop the computer algorithms and process the data. I would also like to express my appreciation to Captain John Bauer for providing all the subjective listener data used as a basis for comparison in this thesis.

Wayne R. Beeson

Table of Contents

List of Figures

List of Tables

Abstract

A method of predicting speech intelligibility using computer algorithms is presented. Diagnostic Rhyme test number four was used to measure speech intelligibility using a subjective listener test and these results were used as a basis for comparison with the intellibility predictions made by the computer algorithm. An audio recording of a speaker reading the Diagnostic Rhyme test was made. This recording was run through a General Electric radio system and varying amounts of noise were added. The output of the radio system was recorded, providing a copy of the input word corrupted by both additive noise and radio system distortion effects. Both the input recording and the noisy output recording were digitized by sampling the analog waveforms at a 10 kilohertz rate. These digital samples were converted to a frequency format by windowing the time samples with a rectangular window 128 time samples in length and processing them using Fast Fourier transform techniques. This procedure simulated running the analog speech signal through a bank of contiguous narrow bandpass filters covering the range of 0 to 5 KHz, with center frequencies 78 Hz apart. The output of this process was a matrix array, corresponding to each word from the tape, of amplitude values 200 time windows long and divided into 64 frequency bands. These 64 frequency bands were then combined into 1/3 octave groups to model the frequency sensitivity of the average human ear, which reduced the matrix array to 16 frequency bands. This processing of the analog signal was used to model the preprocessing which occurs in the human ear. A comparison between each word from the input tape and the noisy output tape was then made using a weighted mean squared error

calculation. This comparison was conjectured to provide a difference measure which is inversely related to intelligibility. This comparison was used to represent how intelligible the input received from the inner ear is to the brain.

Comparison of the intelligibility results from the human listener tests with the computer processing method outlined above gave a Pearson's Correlation Coefficient value of 0.74 which indicates the computer prediction accounted for 55% of the variance in the listener error scores.

AN ALGORITHM FOR DETERMINING SPEECH INTELLIGIBILITY

I. Introduction

This work is in response to a need identified by Air Force Communications Service (AFCS). There have been numerous studies in the area of machine prediction of speech intelligibility; however, the Air Force planners who requested this work are still using human listeners to determine intelligibility. They either think that available computer methods do not produce sufficiently accurate results or that the computer schemes are too complex and difficult to apply to their specific problems. The intent of this work is to take applicable techniques from work that has already been done and combine them to develop a simplified, accurate method to evaluate voice intelligibility with a computer.

Background

The oldest method for determining the intelligibility of speech is a subjective method that involves trained speakers and listener panels that directly score the percentage of speech that is intelligible. This method is still considered the most reliable way to measure intelligibility because it produces repeatable results. The disadvantages of the subjective method are the considerable cost, large number of manhours, and specialized facilities and equipment required.

An early attempt to simplify the procedure for determining the intelligibility of speech involved calculation of the mean squared error (MSE) between an audio waveform and the same audio signal corrupted by noise. This process uses the procedure given by Equation 1.

$$MSE = \frac{1}{2T} \int_{-T}^{T} |X(t) - Y(t)|^2 dt \qquad (1)$$

Such an approach does not yield acceptable estimates of speech intelligibility. This failure is attributed to the fact that vowels contain more power than consonants in speech, but consonants are more important in determining intelligibility than vowels (Ref 15:277). This method of intelligibility measurement is no longer in use due to these shortcomings.

One of the currently popular methods of automating intelligibility prediction is use of the Articulation Index (AI). One method of calculating the AI is by transforming the speech signal into an electrical signal and then passing it through a set of contiguous bandpass filters each 1/3 octave wide. The voltage output of each of these filters is used to calculate a root mean square (RMS) voltage as shown by Equation 2

$$RMS = \frac{1}{2T} \int_{-T}^{T} X^2(t)dt \qquad (2)$$

The noise that is affecting the system is passed through this same set of filters and a root mean square noise voltage in each filter bandpass is calculated. The value of the noise RMS voltage is subtracted from the speech RMS voltage for each filter. If this difference is 30 or more decibels, it is assigned a value of 30. If the difference falls in the range of 0 to 30 decibels, the actual decibel value is assigned. If the difference is 0 or a negative value, it is assigned a value of 0. These values for each filter are then multiplied by weighting factors for each of the different frequency bands. These products are then added together and their sum is the AI (Ref 1:6-15). This process is illustrated by the block diagram in Figure 1.

The AI can be calculated by programming the previous procedure into a computer routine. These programs are in common use today and are an acceptable predictor of intelligibility as long as the noise present is additive white Gaussian noise. Colored noise and multiplicative noise require a complete recalibration of the AI system to give good results (Ref 7:2). This illustrates that the type of noise present must be known exactly and corrected for to maintain the accuracy of this index. When this method of intelligibility prediction is applied to a digital voice system or a system with quantization noise present, there is no correction of recalibration that will provide acceptable intelligibility estimates (Ref 7:2-3).

A recent development in the field of automated voice intelligibility prediction is the use of linear predictive coding (LPC). LPC derives its name from the predictive process it is based on which states: given P samples of a speech signal, the next sample can be predicted approximately by a linear function of the P known samples (Ref 6: A-3). LPC models the vocal tract as an all-pole digital filter and estimates the filter parameters (predictor coefficients) using the time domain speech waveform. This model of the voice tract assumes the vocal tract model to be a time-varying filter with parameters changing slowly enough so they can be considered fixed over a spcified time interval. It accounts for the glottal volume flow and radiation of sound from the mouth in addition to vocal tract sounds (Ref 7:3-4). The most popular way to estimate linear prediction coefficients (a_i) is the autocorrelation method. This method involves time sampling an analog speech signal and

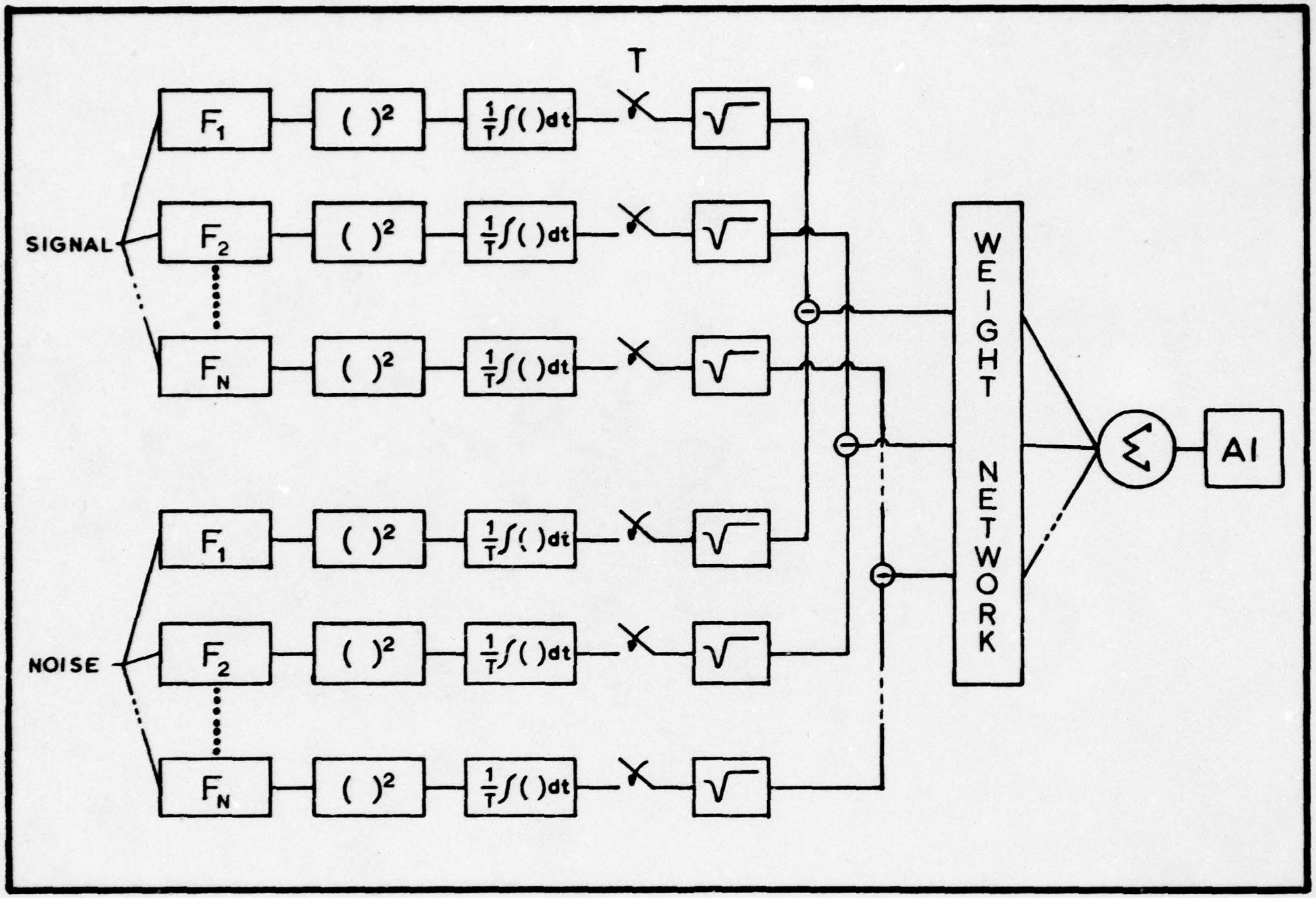

Figure 1. Process Used to Calculate Articulation Index (AI)

windowing these time samples, usually with a Hamming window 256 time samples long. These windowed speech samples are used to calculate a linear prediction of residual energy both for an undistorted speech signal and for the same speech signal after it has been corrupted by additive noise. These residual energy terms are then compared with the actual energy terms and a distance measure is derived from these comparisons. The calculation of these residual energy terms and their comparison is a long and involved mathematical process presented in detail by Hartmann (Ref 6:24-33).

The use of LPC techniques overcomes the disadvantages of sensitivity to the type of noise present that affects AI. LPC intelligibility predictions give a good correlation with listener scores when averaged over 50 or more words (Ref 6:18-19). The disadvantage of LPC is that it involves a large number of computer computations and consequently consumes a great deal of computer time to analyze a small amount of speech. A second disadvantage of this method is that it requires very close synchronization of the words on the undistorted tape with the same words on the tape containing additive noise. This requirement for exacting synchronization makes it necessary to employ very specialized taping equipment to make this process work (Ref 6: 18,20).

Approach

The approach to computer evaluation of speech intelligibility used in this thesis combines some features of the Ariticulation Index calculation, the linear predictive coding method, and the mean squared error calculation.

The human auditory system performs multiple stages of preprocessing on an audio signal before it reaches the brain. Therefore, it seems reasonable to assume that if the processes occurring in the ear and brain can be modeled, it will be possible to make the same type intelligibility determination as the human. The first step in doing this is to model the preprocessing which occurs in the ear.

To model the action of the ear drum in converting sound pressure variations to vibration and the middle ear which transmits these as a varying mechanical vibration to the inner ear, a tape recorder was used. The recorder converts sound pressure variations into an appropriate, continuously varying analog signal.

In the inner ear (cochlea) the mechanical vibration variations undergo the next stage of processing. This process is quite complex, but it appears to involve excitation of the neurons at the base of the hair cells inside the cochlea due to movement of the hair cells. This movement is a result of the mechanical vibration coming from the middle ear causing the fluid in the cochlea to move the hair cells. Since the cochlea is apparently a frequency analyzing device, a model for the inner ear should present the signal in a frequency format (Ref 10). The model for the cochlea used in this thesis consists of sampling the analog waveform from the tape at the Nyquist rate and running these samples through a bank of contigious bandpass filters. The output of these filters are grouped into 1/3 octave bands to simulate the sensitivity of the ear. This changes the analog waveform into a frequency format.

Kabrisky proposed that the cortex of the brain is capable of performing a two-dimensional cross-correlation of a test image with a stored pattern (Ref 9: 47-57). This theory about the visual system was extended to the auditory system by Dailey and Sutton (Ref 3). Assuming the brain performs this two-dimensional cross-correlation between the input signal from the cochlea and phonemes stored in memory and picks the largest correlation value to indicate what phoneme was heard, this process must be modeled. Since the undistorted phoneme stored in memory would have to be in the same format as the incoming signal from the cochlea, a possible model of this process would be to compare the undistorted input word, preprocessed by the models of the ear, with the same word imbedded in noise and run through the same processing. The correlation process will only determine a measure of the difference between a word and its corrupted form, so it appears that a mean squared error calculation can simulate this correlation satisfactorily. This mean squared error will be weighted because of the grouping of filter outputs occurring in the cochlea model.

Objective

The object of this research is to explore the possibility of developing a computer program that will give a reasonably accurate prediction of the intelligibility of speech. This system, if successful, will be used by people with varying degrees of computer support available to them. For this reason the main idea was to keep the procedure simple and automate it as much as possible. Another consideration was to minimize the computer memory and central processor time required for the processing so people can get the program through a busy, time shared computer in a reasonable amount of time. The final goal was to eliminate the need for any elaborate or unique equipment to make or process the audio tape.

Scope

The scope of this project is limited to developing a computer program that will model the actions of the ear on sound waves and apply one possible comparison scheme to model the action of the brain on the processed audio signal. Section II outlines the procedures used to make audio tapes that are used for both human listener and computer intelligibility testing. Section III details the initial computer processing of these audio tapes to sample them at the Nyquist rate and perform a Fast Fourier Transform (FFT) on these time samples. Section IV describes how the original matrix array of amplitude values, produced by the FFT process, was compressed so it would closely approximate the way the ear processes sound data. A method of representing each word by a speech spectrogram and using this to locate the word exactly in a group of time samples of the input wave form is discussed. Section V deals with the cross-correlation method used to locate a word which is imbedded in noise. It evaluates how much the word has been distorted by the additive noise using a weighted mean squared error comparison between the word before the noise is added and the same word plus additive noise. The last two sections show the results of this procedure and make some recommendations for further work in this area.

II. Data Acquisition

The data used in these tests for intelligibility was the Diagnostic Rhyme Test Number IV (DRT-IV). DRT-IV is composed of 58 rhyming word pairs with each word pair designed to test for one of six speech attributes. There are eight rhyming pairs in the list which check for each attribute. The six attributes tested for are voicing, nasality, sustention, sibilation, graveness, and compactness (Ref 15:15-21). The words that test for these attributes are separated by ten pairs of filler words. The DRT-IV used in these tests is shown in Table I and the words which test for each of the speech attributes are identified.

Acquisition Procedure

The data acquisition consisted of a male speaker reading one word of each rhyming word pair from DRT-IV and recording these words on one track of a stereo tape recorder. The other track of the stereo tape was used to record one kilohertz tones which are used for timing references in subsequent processing of the audio tape. The recorder used was a reel-to-reel Sony Model 850 which gave a reasonably high quality of audio reproduction. Recording of the DRT-IV words on tape was used to model the action of the outer ear which converts the pressure variations of sound into an analog signal format.

In recording the test audio tapes, two different male speakers were used to reduce the possible effect of a speaker's regional accent affecting the intelligibility results. The first speaker had a southern accent (Arkansas) and the second had very little regional accent (Minnesota). Four master tapes were made of DRT-IV, two by the first speaker and two by the second speaker. These four master tapes were

Table I

Diagnostic Rhyme Test

DRT IV-(2)

PEST - TEST	-(filler)-	FAN - PAN
VAULT - FAULT	-(voicing)-	CHOCK - JOCK
DUES - NEWS	-(nasality)-	NOTE - DOTE
VEE - BEE	-(sustention)-	TICK - THICK
THANK - SANK	-(sibilation)-	CARE - CHAIR
ROD - WAD	-(graveness)-	DONG - BONG
SO - SHOW	-(compactness)-	YOU - RUE
LID - RID	-(filler)-	REEK - LEAK
DENSE - TENSE	-(voicing)-	GAFF - CALF
BOSS - MOSS	-(nasality)-	BOMB - MOM
FOO - POOH	-(sustention)-	DOUGH - THOUGH
ZEE - THEE	-(sibilation)-	GILT - JILT
FAD - THAD	-(graveness)-	PENT - TENT
HOP - FOP	-(compactness)-	YAWL - WALL
ROW - LOW	-(filler)-	LOOT - ROOT
GIN - CHIN	-(voicing)-	VEAL - FEEL
BEND - MEND	-(nasality)-	NAB - DAB
CHAW - SHAW	-(sustention)-	BON - VON
JUICE - GOOSE	-(sibilation)-	SOLE - THOLE
PEAK - TEAK	-(graveness)-	THIN - FIN
BAT - GAT	-(compactness)-	KEG - PEG
ROCK - LOCK	-(filler)-	LONG - WRONG
GOAT - COAT	-(voicing)-	TUNE - DUNE
MIT - BIT	-(nasality)-	MEAT - BEAT
THEN - DEN	-(sustention)-	SHAD - CHAD
GAUZE - JAWS	-(sibilation)-	GOT - JOT
NOON - MOON	-(graveness)-	DOLE - BOWL
KEY - TEA	-(compactness)-	DILL - GILL
RAMP - LAMP	-(filler)-	LEND - REND

played into the input of a General Electric (GE) preliminary development model, spread-spectrum radio transmitter. The modulated radio signal was transmitted to a hybrid summer where it was mixed with additive noise from a pseudo-random, matched spectrum noise generator. The output of the hybrid summer was then fed into the companion receiver of the GE transmitter and the output audio tapes were recorded at the audio output stage of the receiver. The noise generator was used to simulate intentional jamming of the radio link. Output tapes were made at eight different signal to jammer (S/J) levels. The four input tapes were each used twice as the input when making the different S/J level output tapes. The eighth output tape had the lowest S/J ratio and the signal was too low to be usable for testing, so this tape was discarded. The remaining tapes have S/J levels numbered from one through seven. The highest S/J level is number one and the S/J ratio decreases as the number increases with tape number seven representing the lowest S/J ratio. The actual S/J levels associated with these numbers are classified Secret. If the actual S/J levels are desired, this information is given in the classified portion of Captain Bauer's thesis (Ref 2).

All recordings of the output of this system were made on new Scotch, 1/4 inch tape using a Sony Model 850 recorder. These tapes were recorded at a tape speed of 7½ inches per second.

The GE communications system used in this test was being evaluated by a fellow student for intelligibility using human listener tests (Ref 2). This provided a convenient way to obtain human listener intellibility data to compare to the intellibility predictions made by the computer method presented in this thesis.

III. Analog to Digital Conversion

The initial processing of the four master DRT-IV input tapes and the seven output tapes with different S/J additive noise ratios, was done by the Analog/Hybrid Systems Branch of the Aeronautical Systems Division (ASD) Computer Center.

When the audio tapes were made, the words from DRT-IV were recorded on the right channel of a stereo tape recorder and a one kilohertz (KHz) tone, 1/2 second long, was recorded on the left channel. These one KHz tones were machine generated with the apparatus shown in Figure 2. Every time the tone sounded, the next word from the DRT-IV list was recorded within 2½ seconds after the tone. The tones were spaced seven seconds apart so there was at least 4½ seconds after each word before the next tone. The Analog/Hybrid Branch played each tape back and low pass filtered it to 2.5 KHz and fed this into the Comcor Ci-5000/6 analog computer. The computer sampled the input at the Nyquist rate of 5 KHz. Using 2.5 KHz as the upper cutoff frequency was necessary because the bandwidth of the amplifiers in the analog computer was limited to this value. Since it was desired to analyze the speech input over a range of zero to 5 KHz, it was necessary to analyze the tape by playing it back at a tape speed of 3 3/4 inches per second, half the recording speed, to give the effect of low pass filtering to 5 KHz and sampling at a 10 KHz rate at the original recording speed. This makes it possible to evaluate the speech signal over the desired frequency range in spite of the limitations imposed by the computer's amplifier bandwidth.

The input speech signal was amplified to approximately 100 volts prior to processing to provide a sufficient voltage swing to utilize

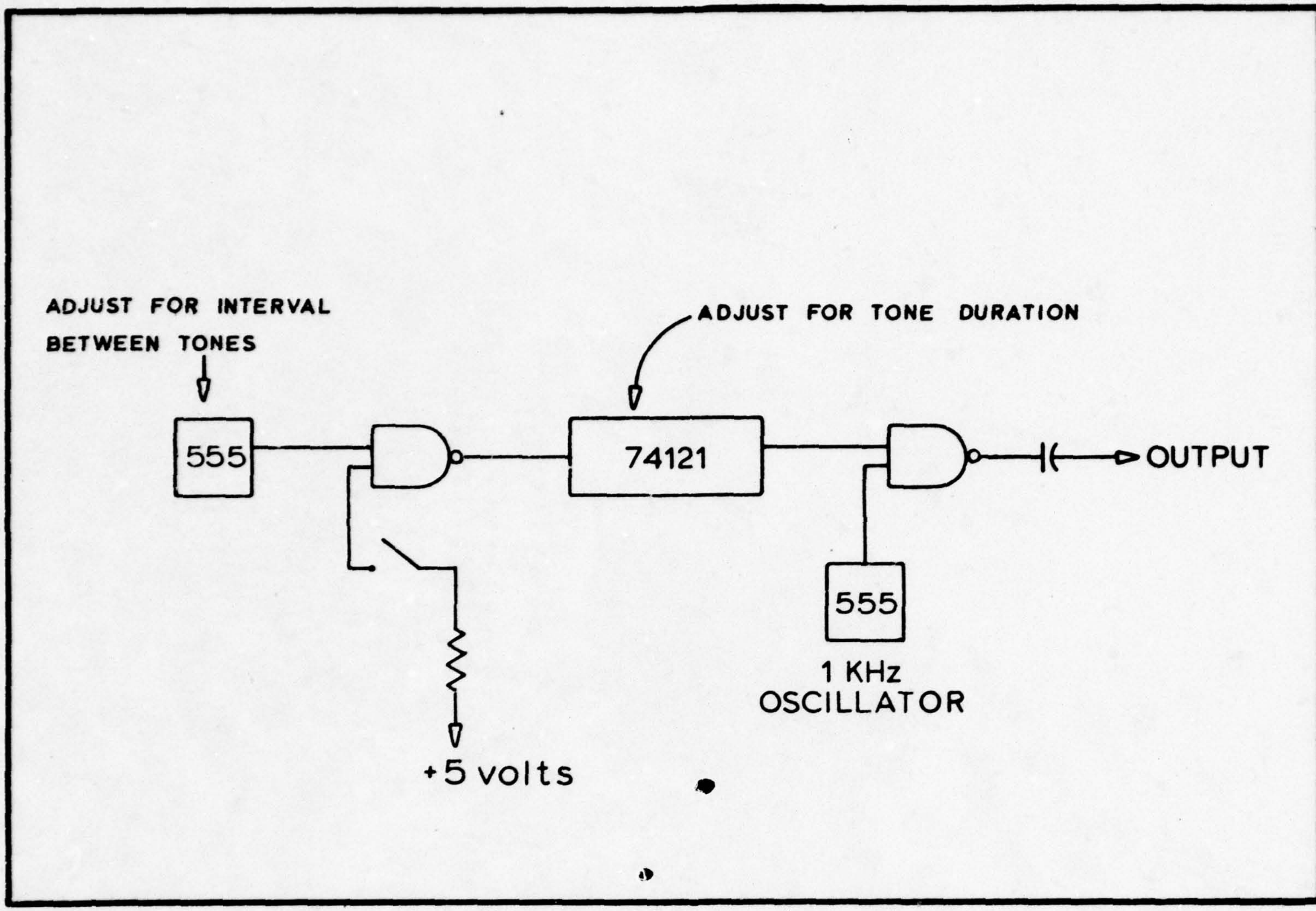

Figure 2. Timing Tone Generator

the accuracy possible with the 11 bit analog to digital converters in the computer. These 11 bit numbers are a binary representation of a 4 digit decimal number. These numbers give the voltage level, between 0 and 100 volts, of the analog waveform each time the waveform is sampled by the digital computer. The 1 KHz tones recorded on the left channel of the audio tapes were used to trigger the sampling equipment in the computer. When the tone occurs, the computer starts sampling the input audio waveform and continues sampling for 2½ seconds, then stops until the next tone occurs. The word from DRT-IV is contained somewhere in this 2½ second sampling interval and will be located exactly by subsequent processing.

Frequency Analysis

The proposed model for the inner ear requires that the digitized analog speech data be represented in the equivalent frequency domain. Fast Fourier Transforms (FFT) techniques were used to convert the digitized data into a frequency representation format (Ref 5:41-52). The actual data conversion involves grouping the digitized time samples into groups of equal length (windowing) and applying FFT techniques to these window groupings to simulate a bank of narrow bandpass filters. The size and shape of the window is based on the desire to have a wide band analysis while retaining reasonable time resolution. The methods for doing this are discussed in detail by Neyman (Ref 12:17-18). The window used is rectangular and 128 time samples long. The digitized data was processed using this window size by an Analog/Hybrid Branch program called AMPSPC. This program gave 64 discrete amplitude values, each corresponding to a 78.125 Hz frequency segment located in the range of 0 to 5 KHz and covering a time window of 12.8 milliseconds

(128 time samples at 10 KHz sample rate). This produced a 64 x 200 matrix array of amplitude values. Each of these matrix arrays contains one word from the DRT-IV audio tape. The matrix arrays were written on a nine track ASD Computer Center library tape (L-tape) and stored for later processing by the CDC-6600 computer.

IV. Digital Signal Processing

Each word of the DRT-IV is now contained in a 64 x 200 matrix array stored on a computer L-tape. Each element of this array represents the signal amplitude to four decimal place accuracy. This signal representation corresponds to an analog speech signal that has been run through a bank of 64 bandpass filters with center frequencies 78.125 Hz apart and an upper cutoff of 5 KHz.

Data Compression

In general the human ear is not a linear receiving device. In order to model the nonlinear frequency response of the ear it was necessary to restructure the digital data so it would app.oximate the ear's unusual sensitivity to frequency change. The six lowest frequency bands of the matrix array were left unchanged. This group has center frequencies of 78.125, 156.250, 234.375, 312.500, 390.625, and 468.750 Hz. All higher frequencies, up to 5 KHz, are grouped into approximately 1/3 octave ranges and the energy content of each group is the sum of the individual array elements contained within that group. This restructuring produced 16 frequency dependent amplitudes from the original 64. The frequency groupings that produce these 16 values are shown in Table II.

Adding the energy of each array element within a 1/3 octave group compensated for the lower amplitude of sound harmonics produced by the vocal chords at high frequencies. This eliminated the need to use the standard preemphasis technique of increasing the signal magnitude by six decibles per octave above 350 Hz (Ref 14:311).

Table II

Speech Frequencies

Center Frequency Original Data	Center Frequency Original Data	Center Frequency Original Data	Center Frequency Original Data
78.125	78.125	2578.125	
156.250	156.250	2656.250	
234.375	234.375	2734.375	
312.500	312.500	2812.500	2812.500
390.625	390.625	2890.625	
468.750	468.750	2968.750	
546.875	585.940	3046.875	
625.000		3125.000	
703.125	742.188	3203.125	
781.250		3281.250	
859.375	898.440	3359.375	
937.500		3437.500	
1015.625		3515.625	3554.690
1093.750	1132.810	3593.750	
1171.875		3671.875	
1250.000		3750.000	
1328.125		2828.375	
1406.250		3906.250	
1448.375	1445.310	3984.375	
1565.500		4062.500	
1640.625		4140.625	
1718.750		4218.750	
1796.875	1793.380	4296.875	
1875.000		4375.000	
1953.125		4453.125	
2031.250		4531.250	4453.125
2109.375		4609.375	
2187.500	2226.560	4687.500	
2265.625		4765.625	
2343.750		4843.750	
2421.875		4921.875	
2500.00		5000.000	

Gray Scale Spectrogram

It is desirable to be able to see a spectrogram of the word when working with the 16 x 200 matrix array that results from the previous compression procedure. This aids in quickly locating where the word occurs in the 200 time windows and determining the length of the word. A convenient method for creating a gray scale spectrogram using computer overprint symbols was developed by Neyman and is used here (Ref 12:22-24). Table III shows the overprint symbols used to create the spectrogram and Figure 3 shows what the actual spectrogram of a word looks like.

Table III

Overprint Symbols for Speech Spectrograms

Number of Overprints	LEVEL OF DARKNESS									
	0	1	2	3	4	5	6	7	8	9
1			+	x	x	x	x	x	x	x
2					-	+	0	0	0	0
3								-	-	$
4									+	+
5										*

Word Location Technique

In order to use the matrix array of amplitude values for comparison purposes it is necessary to locate the word in the 200 time windows and save only the part of the array containing the word. This location process was accomplished by thresholding each value in the matrix array at 1.5 and saving the part of the array where the values exceeded this level. A filtering process was included in the program to eliminate a

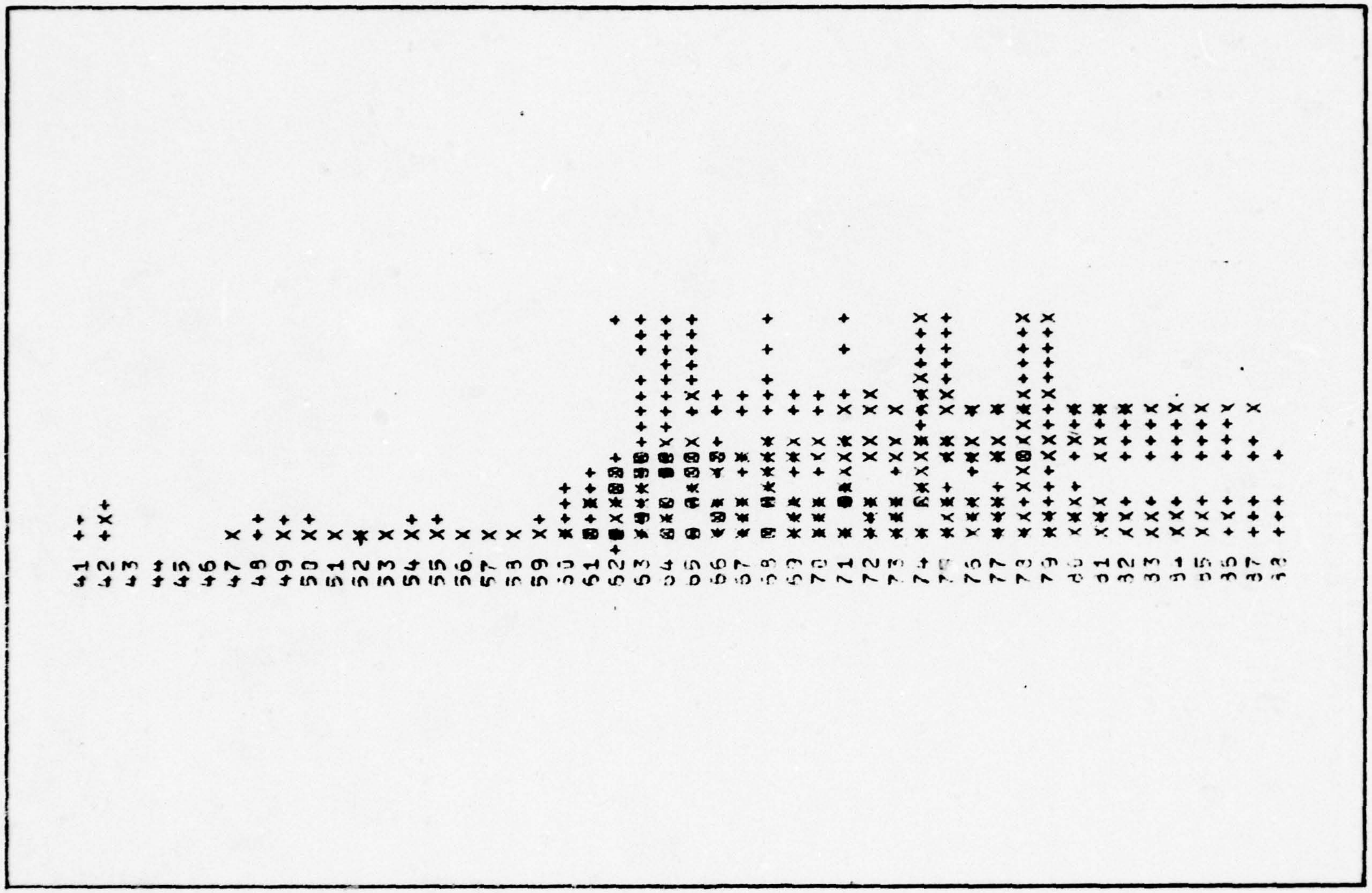

Figure 3. Word Spectrogram

noise spike, occurring outside the word, from being mistaken for part of the word.

Master Tape Processing

The plan was to sample the analog speech data, convert it to a frequency format representation, compress the resulting 64 frequency divisons to 16, locate the word exactly in the 200 time windows, and store only the portion of the 200 column array where the word occurs for use in subsequent processing. These steps were written into a computer program for use in the CDC-6600 computer.

This program was used to process the four master DRT-IV tapes that were used as inputs to the GE radio system. Each master tape was run through this program and the column number where each word started, the number of columns (M) occupied by the word, and the 16 x M array of amplitude values containing the word were recorded on a second L-tape.

An alternate method for locating the word in the 200 time windows was tried, to establish a basis for comparing the effectiveness of the previous procedure. The matrix array was first normalized using Equation 3

$$\hat{a}_{i,j} = \frac{a_{i,j}}{\left[\sum_{i=1}^{L} \sum_{j=1}^{16} a_{i,j}^2 \right]^{1/2}} \tag{3}$$

where $\hat{a}_{i,j}$ = normalized array element. The normalized array was then thresholded at an appropriate level and the computer program predicted where the word was located at in the 200 column matrix. It was easy to see where the word occurred within the 200 column matrix by looking at the accompanying spectrogram, Figure 4, so this was used as a means of evaluating the two computer techniques given above for locating the word. This comparison showed that by normalizing the data in the array prior to doing a computer search for the word, the computer program frequently failed to correctly locate the word. When the computer search for the word was done without normalizing the matrix array, it found the word accurately every time. No explanation can be offered to explain why normalizing the array data caused the word location program to fail.

.50	3.21	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	2.43	2.31	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.14	2.19	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.17	2.22	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.86	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	4.34	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.92	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.92	2.08	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.91	2.03	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.69	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.72	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	3.97	2.30	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	5.34	2.76	2.96	2.81	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.50	5.15	2.90	5.21	2.09	2.47	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
2.32	4.72	3.63	5.15	6.13	6.59	2.92	.50	.50	.50	.50	.50	.50	.50	.50	2.15
.50	5.35	7.46	4.44	4.34	7.57	7.46	2.58	2.34	2.87	2.34	2.16	.50	2.45	2.29	2.76
.50	6.05	5.32	6.13	.50	9.50	7.21	3.26	2.39	2.63	2.16	2.76	2.06	2.39	2.23	2.68
.50	7.36	.50	7.24	5.41	6.45	5.57	3.44	.50	2.39	3.08	2.16	2.06	2.41	2.27	2.69
.50	5.12	6.66	5.34	.50	4.06	6.19	2.59	.50	2.22	2.29	.50	.50	.50	.50	.50
.50	5.71	5.37	5.87	.50	2.84	5.56	.50	.50	2.45	2.39	.50	.50	.50	.50	.50
.50	6.34	.50	7.93	5.12	4.03	4.29	4.14	.50	2.14	2.25	2.10	.50	2.13	.50	2.45
.50	4.82	5.55	5.61	.50	2.55	4.47	3.71	.50	2.77	2.69	.50	.50	.50	.50	.50
.50	5.45	5.07	6.00	.50	2.41	3.74	3.32	.50	2.94	2.26	.50	.50	.50	.50	.50
.50	5.36	.50	6.04	5.34	3.71	3.44	4.07	.50	2.31	2.87	.50	.50	2.14	.50	2.35
.50	4.17	4.91	5.87	.50	.50	3.57	3.57	.50	3.35	3.04	.50	.50	.50	.50	.50
.50	4.16	5.23	5.64	.50	2.54	3.66	3.34	.50	3.33	.50	.50	.50	.50	.50	.50
.50	4.95	.50	6.43	4.33	3.76	3.56	5.66	2.18	4.73	4.03	3.01	2.54	2.33	2.36	3.51
.50	4.93	3.66	5.79	2.73	.50	4.53	4.30	.50	3.77	3.94	2.60	2.18	2.58	2.50	2.93
.50	3.42	5.44	4.57	.50	2.37	4.10	3.16	.50	4.37	.50	.50	.50	.50	.50	.50
.50	4.73	4.45	4.54	2.78	.50	4.20	3.25	.50	4.31	.50	.50	.50	.50	.50	.50
.50	4.78	3.37	2.96	3.77	3.93	6.20	3.20	3.41	5.35	3.35	2.75	2.53	2.93	2.34	3.49
.50	4.99	4.72	2.83	2.37	2.23	3.28	3.57	2.65	2.94	3.74	2.55	2.32	2.71	2.37	3.04
.50	3.41	4.35	3.17	2.05	.50	2.92	3.65	2.17	4.77	.50	.50	.50	.50	.50	.50
.50	3.32	4.42	3.94	.50	.50	3.10	3.41	2.48	4.54	.50	.50	.50	.50	.50	.50
.50	3.22	3.95	2.90	.50	.50	2.80	2.56	2.32	4.15	.50	.50	.50	.50	.50	.50
.50	3.51	3.50	2.76	.50	.50	2.78	2.36	2.47	3.33	.50	.50	.50	.50	.50	.50
.50	3.21	3.56	2.44	.50	.50	2.82	2.40	2.25	3.34	.50	.50	.50	.50	.50	.50
.50	3.13	3.50	2.51	.50	.50	2.77	2.65	2.46	3.31	.50	.50	.50	.50	.50	.50
.50	2.99	3.36	2.52	.50	.50	2.59	2.65	2.27	3.65	.50	.50	.50	.50	.50	.50
.50	2.54	2.46	2.54	.50	.50	2.25	2.69	.50	3.08	.50	.50	.50	.50	.50	.50
.50	2.24	2.37	2.16	.50	.50	2.14	.50	.50	.50	.50	.50	.50	.50	.50	.50

Figure 4. Digital Speech Spectrogram and Associated Matrix Array

V. Cross-Correlation and Mean Squared Error Calculation

The seven noisy tapes made at the audio output of the GE radio under test must now be compared with the master input tape from which each was made. The input tape is compared with the output tape and the mean squared error between each input word and its output plus additive noise is calculated.

To locate where the word occurred on the noisy output tape, it was necessary to perform a cross-correlation between the input word array and the 16 x 200 output array containing the same word imbedded in noise. The length of the word and the part of the array containing the word from each master tape were previously recorded on a computer L-tape. This L-tape is read a word at a time and cross-correlated with the corresponding 16 x 200 array on the L-tape containing the noisy words. The length of the word is read first and that number subtracted from 200 to find the number of cross-correlations that must be performed. Next the arrays are read into the computer core memory and a cross-correlation is performed with the first column of the word from the master tape lined up with the first column of the 16 x 200 noisy array. After each cross-correlation value is determined, the array containing the word from the master list is shifted one column to the right with respect to the noisy array. When all the cross-correlations have been performed for that word, the largest value computed indicates the point where the input word and the noisy output word were aligned. The equation used to compute each of these cross-correlations (P) is

$$P(\tau) = \sum_{i=1}^{L} \sum_{j=1}^{16} A_{i,j} B_{i,j} \qquad (4)$$

where $L = 200$ - length of word read from master tape

$A_{i,j}$ = element of word array from master tape

$B_{i,j}$ = element of word array from noisy tape

When the maximum cross-correlation value has located the word (that is, once the value of τ_o is known such that $P(\tau_o)$ is a maximum) in the 16 x 200 array containing the additive noise, there is enough information available to calculate the mean squared error (MSE) between the two words using the formula

$$MSE = \sum_{i=1}^{L} \sum_{j=1}^{16} (A_{i,j})^2 - 2\,P(\tau_o) + \sum_{i=\tau_o}^{L+\tau_o} \sum_{j=1}^{16} (B_{i,j})^2 \qquad (5)$$

To compute the MSE the maximum cross-correlation value determined in the previous operation is multiplied by two and is the middle term of the MSE equation. Since the exact location of the word in the 16 x 200 noisy array is now known, the last term of the MSE equation can be calculated using this information to square the elements of the array containing the word and sum these squares. The L-tape containing the array of the word from the master input list is used to calculate the first term in the MSE equation by squaring each element of the array and then summing the squares.

The average mean squared error for the DRT-IV list recorded at each of the seven different S/J ratios was determined by summing the MSE for all 58 words in the list and dividing this sum by 58. This gives an average MSE corresponding to each S/J level to be used in deriving an estimate of the intelligibility at that S/J level.

The brain "expects" to see uncorrupted phoneme groups (words) characterized in a certain way. Assuming that the models used here give a reasonably accurate representation of the preprocessing which occurs in the ear, it should now be possible to measure the difference between what the brain expects to hear and what it actually hears. It is conjectured that this difference is inversely related to the intelligibility of what is heard. To model this process the MSE calculation provides a measure for determining the difference between a word and that same word after it has been corrupted by noise. It is assumed this distance measure can now be related to intelligibility.

VI. Results

The first computer program discussed in section IV was designed to locate each word in the 200 time windows. This program worked perfectly as long as the tapes containing the words had a low noise level compared to the signal amplitude of the word. It was also necessary to amplify the peak levels within each word to at least 75 volts prior to digital sampling.

The second program, discussed in section V, was designed to first locate the word in the 200 time windows on the tape with additive noise by a cross-correlation with the same word without noise. This cross-correlation provided a sharp peak with an amplitude well above the other values to indicate when the two words were aligned. This distinct peak occurred even at the lowest S/J ratio. This can be seen by looking at the cross-correlation values for a word at the lowest S/J level, Figure 5.

The second part of the program described in section V is used to calculate the mean squared error between the input word and the same word after it passes through the radio system and is corrupted by noise. Figure 6 shows a plot of mean squared error values versus the seven different S/J ratios. The increase in MSE is approximately linear as the S/J ratio decreases.

Figure 7 shows the average number of errors made by the 10 people who listened to the noisy tapes at each S/J ratio.

A scatter plot of human listener error scores versus mean squared error values is shown in Figure 8. This plot displays the data used to calculate Pearson's Correlation Coefficient.

RECORD NUMBER 7 HAS A MAXIMUM CROSSCORRELATION VALUE OF 3443.45

VALUES OF EACH CROSSCORRELATION

2283.15	2286.36	2307.44	2350.55	2375.35
2430.38	2427.66	2403.39	2442.67	2452.12
2512.39	2466.74	2523.84	2494.79	2477.39
2507.18	2549.63	2541.08	2499.52	2529.81
2528.40	2584.06	2523.57	2485.37	2472.34
2529.14	2622.18	2585.15	2574.93	2561.39
2574.73	2585.51	2636.59	2636.58	2602.56
2639.36	2595.76	2535.39	2552.40	2561.04
2587.32	2620.85	2637.04	2639.65	2648.63
2599.60	2563.31	2521.55	2575.39	2589.16
2583.34	2538.89	2536.94	2525.86	2562.09
2491.09	2501.41	2523.95	2483.81	2463.49
2371.19	2310.93	2310.50	2283.24	2232.59
2231.78	2204.32	2222.35	2197.52	2133.11
2138.78	2117.15	2074.58	2060.74	2067.75
2058.84	2053.27	2074.08	2103.98	2077.40
2065.43	2109.71	2136.01	2225.55	2279.52
2321.64	2415.90	2449.44	2522.58	2561.87
2621.29	2652.15	2687.30	2774.00	2686.51
3006.15	3095.53	3145.38	3219.07	3296.90
3314.19	3363.70	3386.61	3397.87	$\boxed{3443.45}$
3404.21	3322.75	3301.66	3297.31	3224.56
3195.33	3106.08	3029.27	2975.67	2937.41
2847.04	2764.71	2714.32	2674.48	2651.47
2724.31	2729.74	2687.19	2682.50	2622.32
2592.90	2644.09	2633.69	2573.14	2548.18
2508.40	2548.13	2552.30	2513.10	2499.71
2558.47	2573.17	2595.04	2571.22	2568.32
2621.04	2635.42	2636.91	2666.57	2612.94
2619.22	2633.35	2644.85	2659.65	2657.45
2651.23	2627.83	2619.78	2652.16	2625.06
2611.55	2633.13	2595.43	2505.58	2624.91

Figure 5. Matrix of Cross-Correlation Values for a Word at S/J 7

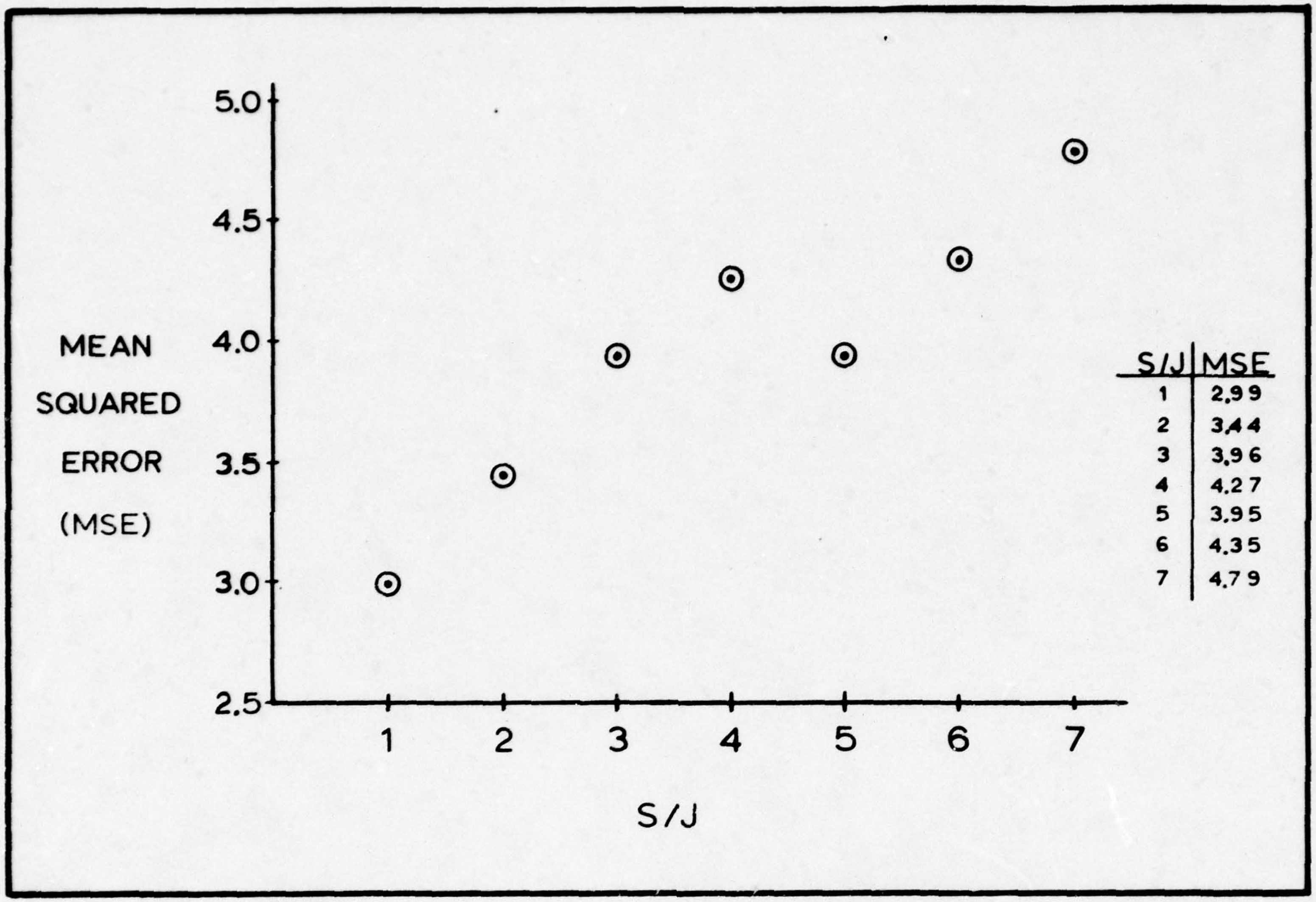

Figure 6. Plot of Mean Squared Error Versus S/J

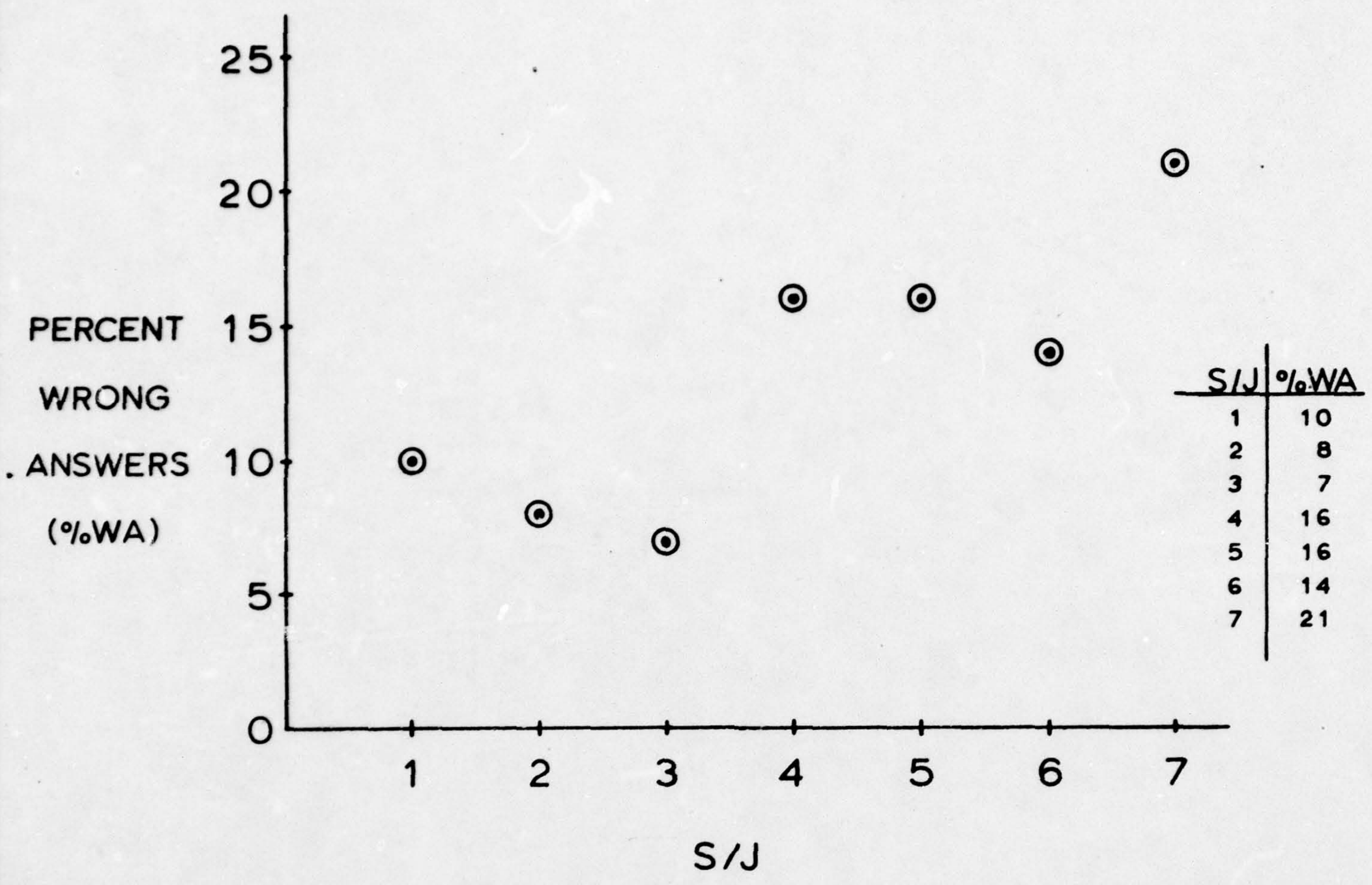

Figure 7. Plot of Percent of Listener Wrong Answers Versus S/J Ratio

Pearson's Correlation Coefficient (P) is used to determine the degree of correlation between the computer calculated mean squared error values and the results of the human listener tests. The formula used to calculate P is

$$P = \frac{\sum_{i=1}^{7} (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{7} (X_i - \overline{X})^2 \sum_{j=1}^{7} (Y_i - \overline{Y})^2\right]^{\frac{1}{2}}} \qquad (6)$$

where

X_i = Mean Squared Error value

Y_i = error score of human listeners

$\overline{X} = 1/7 \sum_{i=1}^{7} X_i$

$\overline{Y} = 1/7 \sum_{i=1}^{7} Y_i$

In this case P was calculated to be 0.74. To find the percentage of the variance in the listener errors accounted for by observing the mean squared error values, under a Gaussian assumption (zero mean Gaussian), P must be squared and multiplied by 100, which gives a value of 55%.

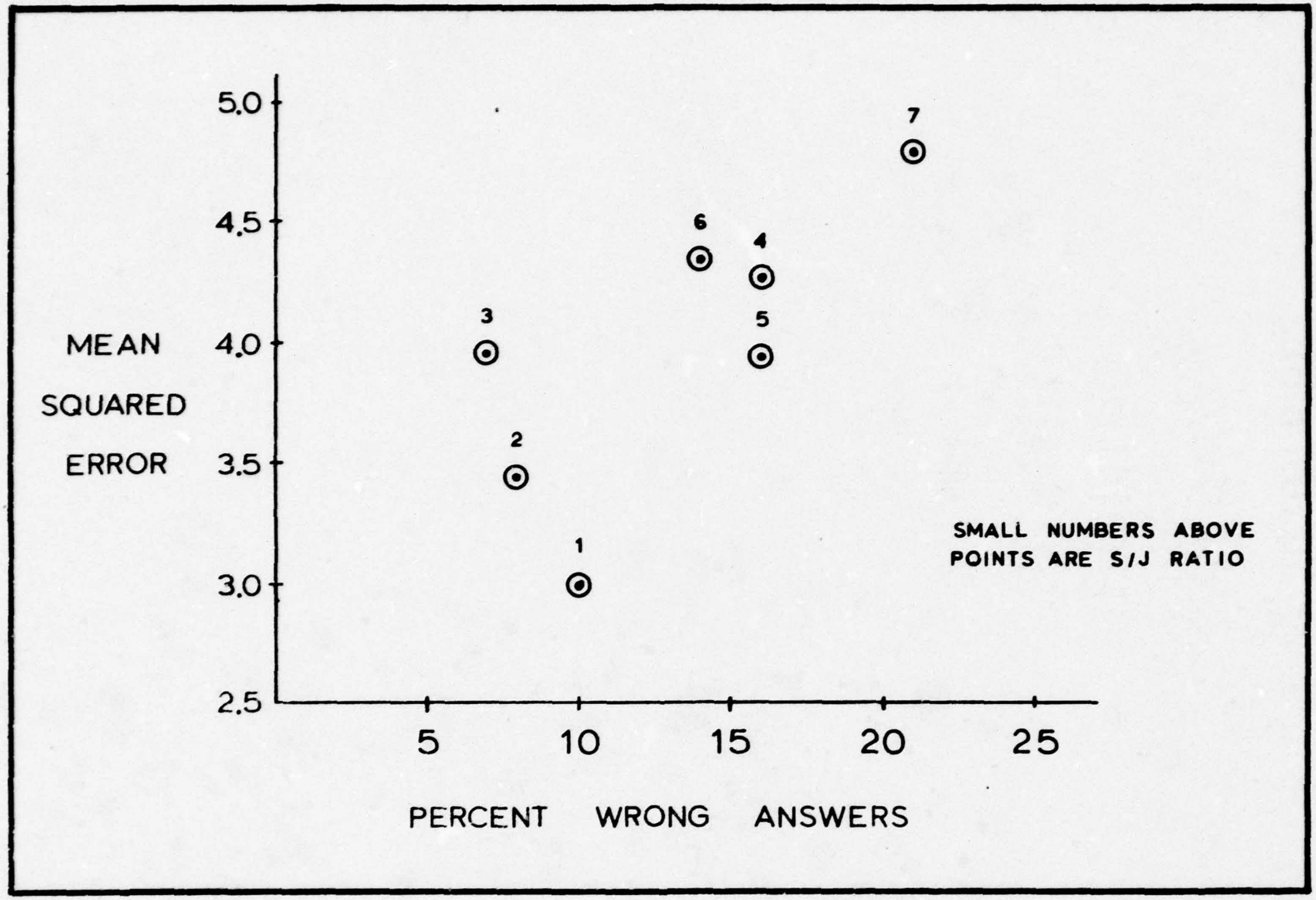

Figure 8. Scatter Plot of Mean Squared Error Versus Percentage of Listener Wrong Answers

VII. Conclusions and Recommendations

The value of 0.74 which was calculated for Pearson's Correlation Coefficient confirms that there is significant correlation between the intelligibility scores of the human listeners and the MSE values calculated by the computer program. This appears to support the ideas set forth in this thesis as a reasonable approach to predicting voice intelligibility. This single comparison is not enough to determine an exact relationship between the mean squared error value and intelligibility, but it seems to provide a starting place for further refinement of the procedure.

The human listener scores used as a comparison were the average of ten listeners. These results showed an unexpectedly high error rate at the best S/J level which decreased as the noise got worse for the first three S/J levels, Figure 7. This trend disappears if a much larger group of listener scores are averaged together and the error rate increases monotonically as the S/J ratio decreases (Ref 2). This listener performance would have given closer agreement with the MSE predictions. This suggests that a listener group considerably larger than ten is required to get a reliable intelligibility figure.

One of the original goals of this thesis was to provide a means of making computerized intelligibility measurement that do not require any special or unique equipment. This goal was met, with the exception of the processing described in section III done by the ASD Analog/Hybrid Branch. A recommendation for further work in this area is to develop a program for one of the more elaborate minicomputers in common use in the Air Force that can perform these operations.

Further investigation in this area could focus on methods of comparing the master word and noisy word, after they have been preprocessed by the ear models, other than MSE. This may yield a better predictor than MSE.

Bibliography

1. Acoustical Society of America. American National Standard Methods for the Calculation of the Articulation Index. ANSI-S3.5-1969. New York: American National Standards Institute, January, 1969.

2. Bauer, John E. Evaluation of Two Sampled Data Communication Systems. M.S. Thesis GE/EE/77-8. Wright-Patterson Air Force Base, Ohio: Air Force Institute of Technology (1977).

3. Dailey, Keith G. and F. Sutton. An Automatic Speech Recognition System Using A Vocoder Input. M.S. Thesis GE/GGC/EE/72-18. Wright-Patterson Air Force Base, Ohio: Air Force Institute of Technology (1972).

4. Electronic Systems Division, U. S. Air Force. Comparative Evaluations of Speech Intelligibility Performance of Three Narrowband Voice Communications Methods: TRIVOC, LPC, and PLPC. ESD-TR-77-131. Hanscom Air Force Base, Massachusetts: ESD, USAF; December, 1976.

5. Hall, William B., Jr. A Digest and Reference Organization of Fast Fourier Transform Literature and Software. VNA, Internal Memo 72-2. Wright-Patterson Air Force Base, Ohio: Analog/Hybrid Systems Branch Computer Center, February, 1972.

6. Hartman, W. J. and S. Boll. Voice Channel Objective Evaluation Using Linear Predictive Coding. FAA-RD-75-189. Washington: Federal Aviation Administration, August, 1976.

7. Hartman, W. J. and K. Gamauf. Objective Evaluation of Voice Communication Channels Using Linear Predictive Coding. unpublished paper. Boulder, Colorado: U. S. Department of Commerce, June, 1976.

8. House, A. S. and G. Hughes. Speech Analysis. AFCRL-69-0371. Bedford, Massachusetts: Air Force Cambridge Research Laboratories August, 1969. AD 696599.

9. Kabrisky, Matthew. A Proposed Model for Visual Information Processing in the Human Brain. Urban, Illinois: University of Illinois Press, 1966.

10. Kiang, Nelson, Y. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. Cambridge, Massachusetts: The MIT Press, 1965.

11. Laefoged, Peter. Elements of Acoustic Phonetics. Chicago, Illinois: The University of Chicago Press, 1962.

12. Neyman, Ralph W. Computer Recognition of Phonemes in Continuous Speech. M.S. Thesis GE/EE/76-10. Wright-Patterson Air Force Base, Ohio: Air Force Institute of Technology (1976).

13. Niederjohn, R. J. and J. Grotelueschen. "The Enhancement of Speech Intelligibility in High Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression." IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-24: 277-282 (August, 1976).

14. Ullman, J. R. Pattern Recognition Techniques. New York: Crane, Russak and Co., Inc. 1973.

15. Voiers, W. D.; A. Shappley; and C. Hehmsoth. Research on Diagnostic Evaluation of Speech Intelligibility. AFCRL-72-0694. Bedford Massachusetts: Air Force Cambridge Research Laboratories, January, 1973.

Appendix A

Sequence Chart for Intelligibility Prediction

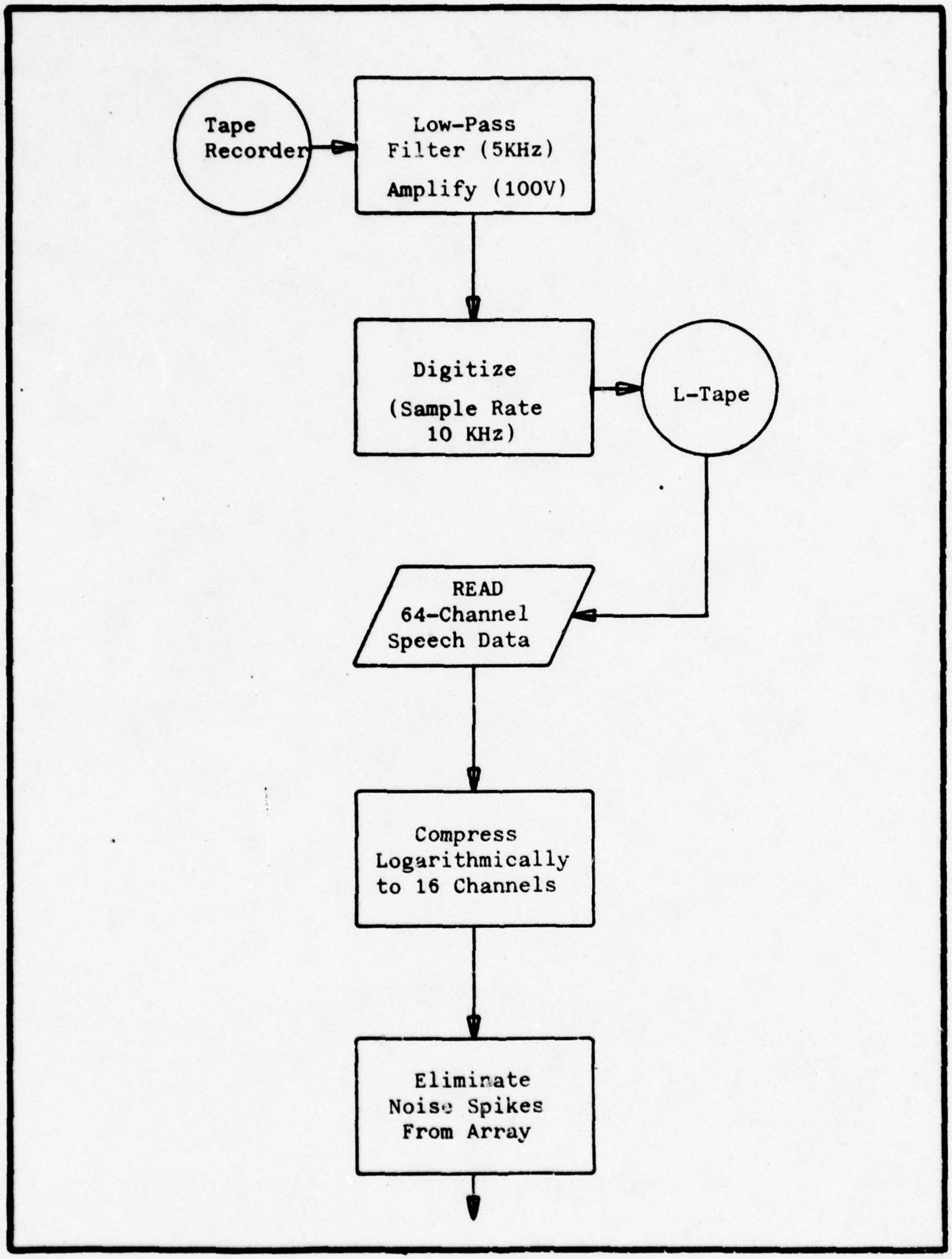

Figure 9. Sequence Chart for Master Tape Processing Program (Plate 1)

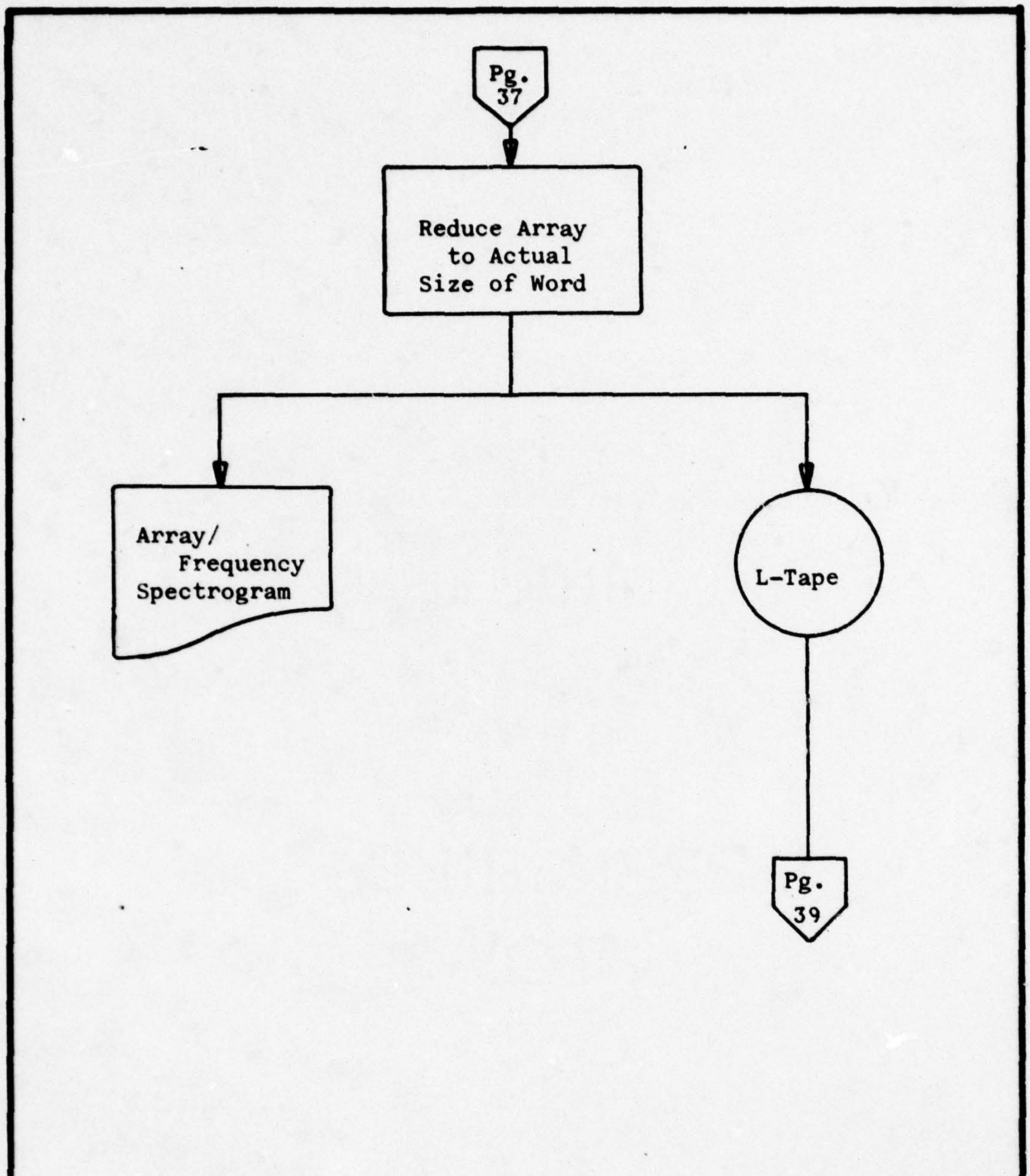

Figure 10. Sequence Chart for Master Tape Processing Program (Plate 2)

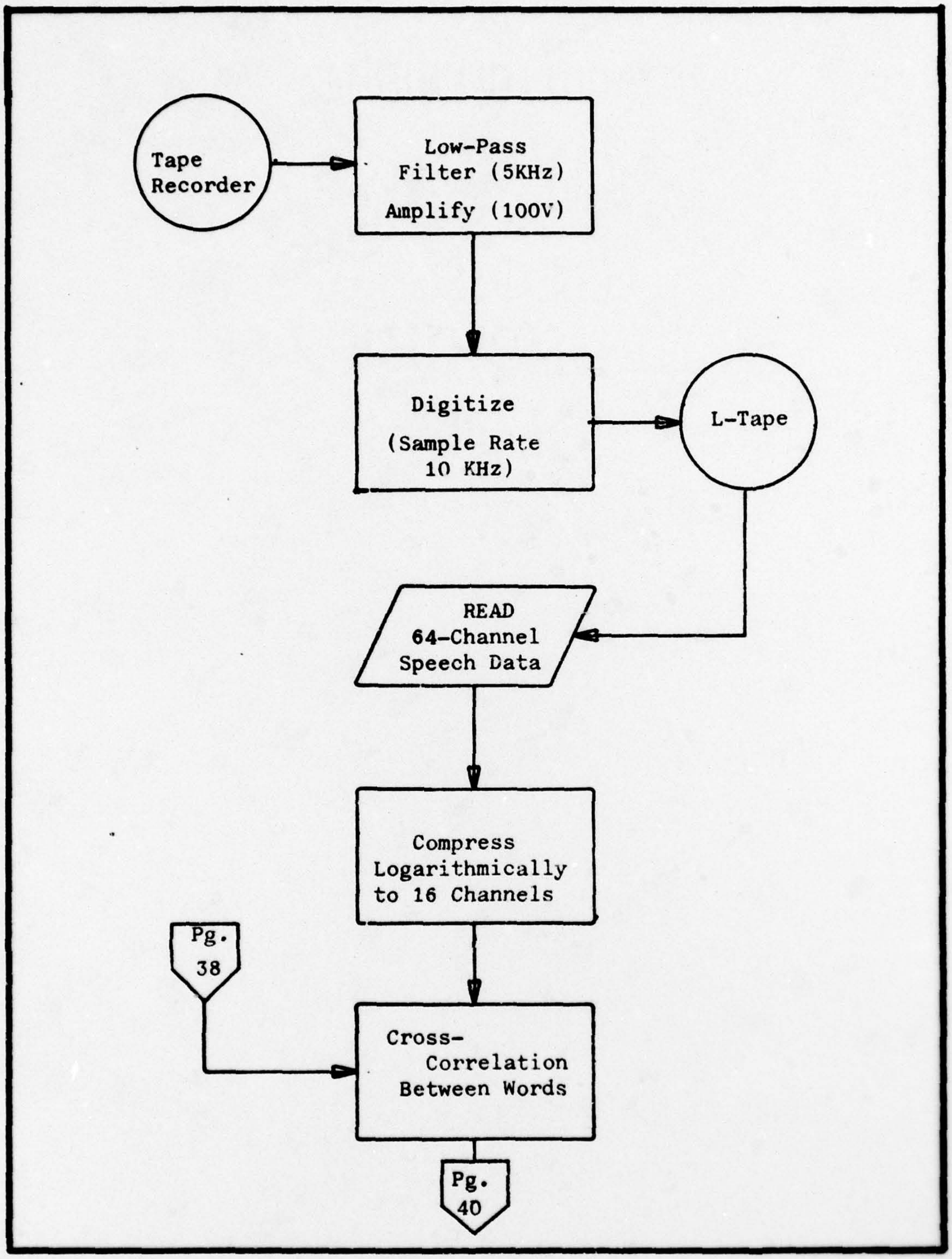

Figure 11. Sequence Chart for Cross-Correlation and MSE Program (Plate 1)

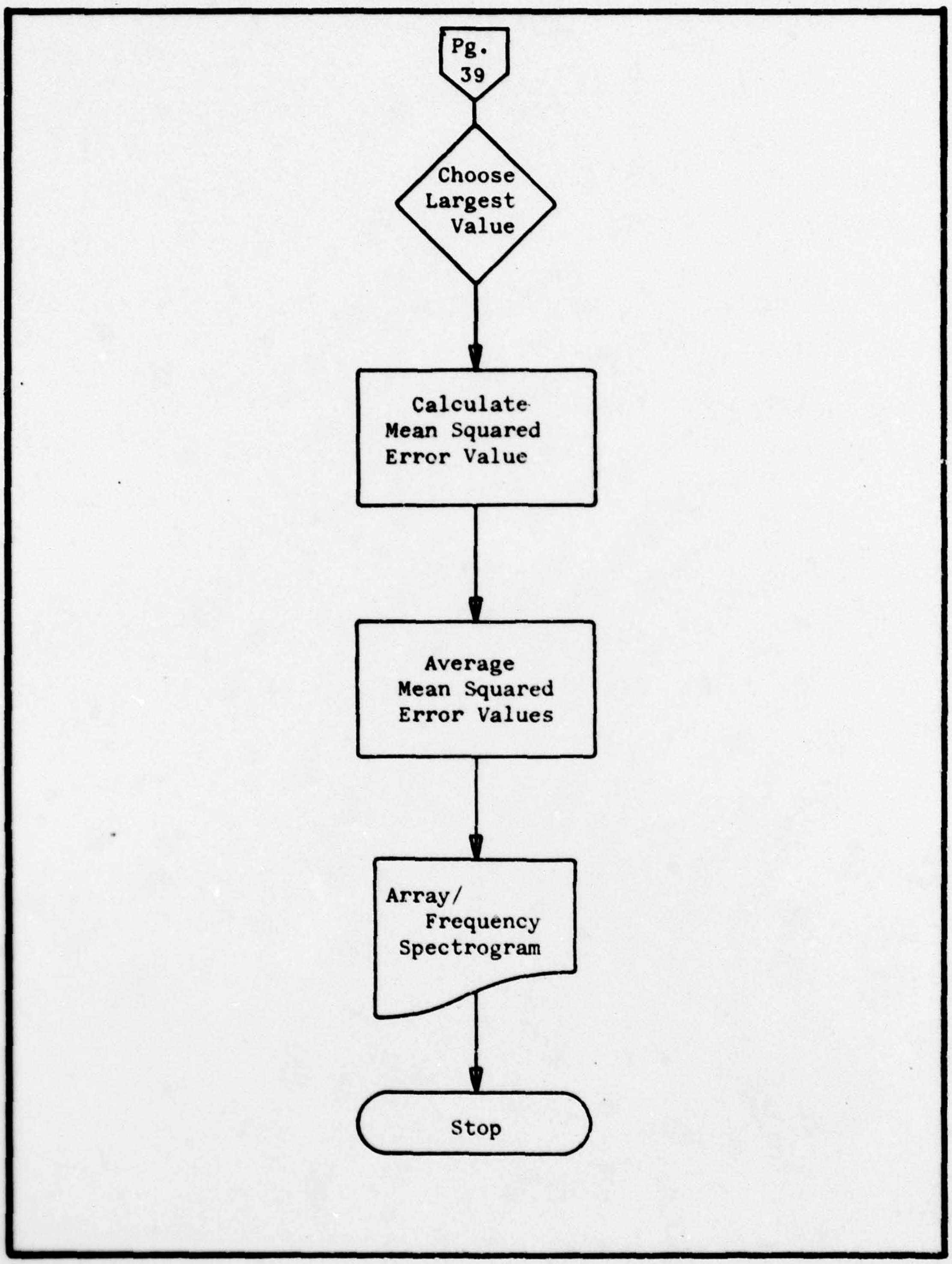

Figure 12. Sequence Chart for Cross-Correlation and MSE Program (Plate 2)

Appendix B

Computer Programs

```
      PROGRAM OCTAVE(INPUT,OUTPUT,TAPE1,TAPE2,TAPE6=OUTPUT)
      DIMENSION SYMBOL2(10),SYMBOL3(10),SYMBOL4(10),SYMBOL5(10)
      DIMENSION A(64),B(19),SYMBOL1(10),IBI(19)
      DIMENSION NCOUNT(200)
      DIMENSION PATRN(16,200)
C*************************************************************************
C**                                                                     **
C**    THIS PROGRAM REDUCES 64 CHANNELS OF DIGITIZED SPEECH DATA TO 16  **
C**    CHANNELS. THE ORIGINAL 64 CHANNELS ARE COMBINED BY ADDING THE ENERGY **
C**    CONTRIBUTIONS OF EACH ELEMENT WITHIN A 1/3 OCTAVE GROUP. THE OUTPUT **
C**    IS THE COMPRESSED ARRAY AND AN ACCOMPANYING SPEECH SPECTROGRAM.  **
C**                                                                     **
C*************************************************************************
C*************************************************************************
C**                                                                     **
C** THIS RUN READS 58 RECORDS OF SINGLE WORD CHOICES FROM LIST 3, ORT #4. **
C**                                                                     **
C*************************************************************************
C*************************************************************************
C**                                                                     **
C**                     SPECTROGRAM OVERPRENT SYMBOLS                   **
C**                                                                     **
C*************************************************************************
      DATA SYMBOL1/1H ,1H ,1H+,1HX,1HX,1HX,1HX,1HX,1HX,1HX/
      DATA SYMBOL2/1H ,1H ,1H ,1H ,1H-,1H+,1HO,1HO,1HO,1HO/
      DATA SYMBOL3/1H ,1H ,1H ,1H ,1H ,1H ,1H ,1H-,1H-,1H*/
      DATA SYMBOL4/1H ,1H ,1H ,1H ,1H ,1H ,1H ,1H ,1H+,1H+/
      DATA SYMBOL5/1H ,1H ,1H ,1H ,1H ,1H ,1H ,1H ,1H ,1H+/
C*************************************************************************
C**                                                                     **
C**                          PROGRAM VARIPLES                           **
C**                                                                     **
C*************************************************************************
C*************************************************************************
C**                                                                     **
C**    NUMBER OF RECORDS TO BE READ                                     **
C**                                                                     **
C*************************************************************************
      NREC=58
C*************************************************************************
C**                                                                     **
C**    MAXIMUM RECORD LENGTH                                            **
C**                                                                     **
C*************************************************************************
      NN2=201
C*************************************************************************
C**                                                                     **
C**                INPUT ARRAY LOGARITHMICALLY COMPRESSED               **
C**                                                                     **
C*************************************************************************
      NGROO=1
      NN1=64
    1 CONTINUE
      DO 30 I=1,NN2
    5 READ(1)(NCHAN,NDIM,(A(J),J=1,64))
      IF(EOF(1))310,30
   30 CONTINUE
```

```
      JJ=1
      DO 40 J=1,6
      B(JJ)=A(J)
      JJ=JJ+1
   40 CONTINUE
      DO 50 J=7,11,2
      B(JJ)=(A(J)+A(J+1))
      JJ=JJ+1
   50 CONTINUE
      DO 60 J=13,17,4
      B(JJ)=(A(J)+A(J+1)+A(J+2)+A(J+3))
      JJ=JJ+1
   60 CONTINUE
      SUM1=0
      DO 70 J=21,25
      SUM1=(SUM1+A(J))
   70 CONTINUE
      B(JJ)=SUM1
      SUM2=0
      DO 80 J=26,31
      SUM2=(SUM2+A(J))
   80 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM2
      SUM3=0
      DO 90 J=32,40
      SUM3=(SUM3+A(J))
   90 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM3
      SUM4=0
      DO 100 J=41,50
      SUM4=(SUM4+A(J))
  100 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM4
      SUM5=0
      DO 110 J=51,64
      SUM5=(SUM5+A(J))
  110 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM5
  500 CONTINUE
C**********************************************************************
C**                                                                  **
C** NOISE SPIKES ELIMINATED FROM ARRAY.                              **
C**                                                                  **
C**********************************************************************
      NSUM=0
      DO 400 JJ=1,16
      IF(B(JJ).LT.1.5) B(JJ)=0
      PATRN(JJ,I)=B(JJ)
      IF(PATRN(JJ,I).GT.0) NSUM=NSUM+1
  400 CONTINUE
      NCOUNT(I)=NSUM
  715 FORMAT(1X,F6.3)
  305 CONTINUE
```

```
  310  CONTINUE
       IF(NCOUNT(1).EQ.1.AND.NCOUNT(2).EQ.0) NCOUNT(1)=0
       IF(NCOUNT(200).EQ.1.AND.NCOUNT(199).EQ.0) NCOUNT(200)=0
       DO 595 I=2,199
       I1=I-1
       I2=I+1
       I3=I+2
       IF(NCOUNT(I).EQ.1.AND.NCOUNT(I1).EQ.0.AND.NCOUNT(I2).EQ.0) NCOUNT
      *(I)=0
       IF(NCOUNT(I).EQ.1.AND.NCOUNT(I1).EQ.0.AND.NCOUNT(I2).EQ.1.AND.
      *NCOUNT(I3).EQ.0) NCOUNT(I)=0
       IF(NCOUNT(I).EQ.2.AND.NCOUNT(I1).EQ.0.AND.NCOUNT(I2).EQ.0) NCOUNT(
      *I)=0
  595  CONTINUE
C*************************************************************************
C**                                                                     **
C** ARRAY REDUCED TO ACTUAL SIZE OF WORD.                               **
C**                                                                     **
C*************************************************************************
       PRINT 561,NRECD
  561  FORMAT(10X,"RECORD NUMBER ",I2,/)
       DO 575 K=1,200
       KSTART=K
       IF(NCOUNT(K).GE.1) GO TO 1000
  575  CONTINUE
 1000  PRINT 565,KSTART
  565  FORMAT(10X,"THIS WORD STARTS WITH TIME SAMPLE ",I3,/)
       DO 585 N=1,200
       J=201-N
       KSTOP=J
       IF(NCOUNT(J).GE.1) GO TO 1100
  585  CONTINUE
 1100  PRINT 566,KSTOP
  566  FORMAT(10X,"THIS WORD ENDS WITH TIME SAMPLE ",I3,/)
       LNGTH=KSTOP-KSTART+1
       PRINT 559,LNGTH
  559  FORMAT(10X,"THIS WORD IS ",I3," TIME SAMPLES LONG",///)
       WRITE(2,750)LNGTH
  750  FORMAT(I3)
       DO 755 I=1,LNGTH
  755  WRITE(2,760)(PATRN(JJ,I+KSTART-1),JJ=1,16)
  760  FORMAT(16F6.3)
       NRECD=NRECD+1
       DO 212 I=KSTART,KSTOP
C*************************************************************************
C**                                                                     **
C** ARRAY VALUES CONVERTED TO INTEGER FORM FOR USE IN PRINTING SPECTRAGRAM. **
C**                                                                     **
C*************************************************************************
       DO 240 JJ=1,16
       PATRN(JJ,I)=(PATRN(JJ,I)+.5)
       IPI(JJ)=IFIX(PATRN(JJ,I))
  240  CONTINUE
       PRINT 211,(PATRN(JJ,I),JJ=1,16),I,(SYMBOL(IPI(JJ)+1),JJ=1,16)
  211  FORMAT(1X,16F5.2,1X,I3,16A1)
       PRINT 211,(SYMBOL(IPI(JJ)+1),JJ=1,16)
       PRINT 211,(SYMBOL(IPI(JJ)+1),JJ=1,16)
```

```
      PRINT 211,(SYMBOL4(IBI(JJ)+1),JJ=1,16)
      PRINT 211,(SYMBOL5(IBI(JJ)+1),JJ=1,16)
 211  FORMAT("+",91X,16A1)
 212  CONTINUE
      ENDFILE2
      PRINT 299
 299  FORMAT(//////)
      NREC=NREC-1
      IF(NREC.GT.0) GO TO 1
      STOP
      END
```

```
C*************************************************************************
C**                                                                     **
C** COMPARISON OF MASTER TAPE 8-5 (KEN ) WITH LIST 1., S/J 1.           **
C**                                                                     **
C*************************************************************************
C*************************************************************************
C**                                                                     **
C** THIS PROGRAM LOCATES A DIGITIZED WORD, IN NOISE, BY USING A CROSS-  **
C** CORRELATION WITH THE SAME WORD OF KNOWN LENGTH WITH NO NOISE.       **
C**                                                                     **
C*************************************************************************
      PROGRAM CORELAT(INPUT,OUTPUT,TAPE1,TAPE2,TAPE6=OUTPUT)
      DIMENSION NOISE(16,5(0),WORD(16,75),MNSQER(56),TEMPL(155)
      DIMENSION A(64),B(19)
      REAL NOISE,LRGST,MEANSQ,MNSQER
      INTEGER ATOLGTH,POSTN
C*************************************************************************
C**                                                                     **
C** NUMBER OF SAMPLES PER RECORD.                                       **
C**                                                                     **
C*************************************************************************
      NREC=56
      NRCD=1
C*************************************************************************
C**                                                                     **
C** NUMBER OF RECORDS TO BE READ                                        **
C**                                                                     **
C*************************************************************************
    1 CONTINUE
C*************************************************************************
C**                                                                     **
C** READS ONE WORD FROM THE TAPE WITH NO NOISE.                         **
C**                                                                     **
C*************************************************************************
C*************************************************************************
C**                                                                     **
C** READS THE LENGTH OF THE WORD (NUMBER OF TIME SAMPLES).              **
C**                                                                     **
C*************************************************************************
      READ(1,750)LNGTH
  750 FORMAT(I2)
  760 FORMAT(16F5.3)
  770 FORMAT(15F5.3)
      I=0
C*************************************************************************
C**                                                                     **
C** READS THE WORD.                                                     **
C**                                                                     **
C*************************************************************************
  752 I=I+1
  753 READ(1,760)(WORD(JJ,I),JJ=1,16)
      IF(EOF(1))752,751
C*************************************************************************
C**                                                                     **
C** READS RECORD CONTAINING THE WORD IN NOISE.                          **
C**                                                                     **
C*************************************************************************
```

```
  752 NN2=301
      NN1=64
      DO 300 I=1,NN2
    5 READ(2) (NCHAN,NDIM,(A(J),J=1,64))
      IF(EOF(2))310,30
   30 CONTINUE
      JJ=1
      DO 40 J=1,6
      B(JJ)=A(J)
      JJ=JJ+1
   40 CONTINUE
      DO 50 J=7,11,2
      B(JJ)=(A(J)+A(J+1))
      JJ=JJ+1
   50 CONTINUE
      DO 60 J=13,17,4
      B(JJ)=(A(J)+A(J+1)+A(J+2)+A(J+3))
      JJ=JJ+1
   60 CONTINUE
      SUM1=0
      DO 70 J=21,25
      SUM1=(SUM1+A(J))
   70 CONTINUE
      B(JJ)=SUM1
      SUM2=0
      DO 80 J=26,31
      SUM2=(SUM2+A(J))
   80 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM2
      SUM3=0
      DO 90 J=32,40
      SUM3=(SUM3+A(J))
   90 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM3
      SUM4=0
      DO 100 J=41,50
      SUM4=(SUM4+A(J))
  100 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM4
      SUM5=0
      DO 110 J=51,64
      SUM5=(SUM5+A(J))
  110 CONTINUE
      JJ=JJ+1
      B(JJ)=SUM5
  500 CONTINUE
      DO 250 JJ=1,16
      NOISE(JJ,I)=B(JJ)
  250 CONTINUE
  300 CONTINUE
  310 CONTINUE
  771 ATLNGTH=2.0-LNGTH
      DO 1. I=1,ATLNGTH
      SUM=0.
```

```
      DO 421 K=1,LNGTH
      KK=K-1+I
      DO 421 JJ=1,16
  421 SUM=SUM+WORD(JJ,K)*NOISE(JJ,KK)
  1   TEMP1(I)=SUM
C***********************************************************************
C**                                                                   **
C** PICK LARGEST VALUE OF CROSSCORRELATION FOR EACH RECORD.           **
C**                                                                   **
C***********************************************************************
      LRGST=TEMP1(1)
      POSTN=1
      DO 17 L=2,ATOLGTH
      IF(TEMP1(L).LE.LRGST) GO TO 17
      LRGST=TEMP1(L)
      POSTN=L
  17  CONTINUE
      PRINT 15,(NB-CD),LRGST,POSTN
  15  FORMAT(1X,"RECORD NUMBER ",I2," HAS A MAXIMUM CROSSCORRELATION VAL
     *UE OF ",G12.6," STARTING AT TIME SAMPLE NUMBER ",I3,///)
      PRINT 12
  12  FORMAT(1X,"VALUES OF EACH CROSSCORRELATION",//)
      PRINT 13,(TEMP1(I),I=1,ATOLGTH)
  13  FORMAT(1X,5(G12.6,3X),/)
      PRINT 56
      SUM=0.
      DO 51 I=1,LNGTH
      DO 505 J=1,16
      SUM=WORD(J,I)**2+NOISE(J,POSTN)**2+SUM
  505 CONTINUE
  51  POSTN=POSTN+1
      MEANSQ=(SUM-2.*LRGST)/(16.*LNGTH)
      MNSQER(NB-CD)=MEANSQ
      NBRCD=NBRCD+1
      NRECD=NRECD-1
      IF(NRECD.GT.0) GO TO 1
      PRINT 56
  56  FORMAT(//////)
      PRINT 52
  52  FORMAT(20X,"RECORD NUMBER",10X,"MEAN SQUARE ERROR",/)
      DO 67 N=1,56
      PRINT 53,N,MNSQER(N)
  53  FORMAT(27X,I2,19X,G12.6,/)
  67  CONTINUE
      SUMMSE=0
      DO 66 II=1,56
  66  SUMMSE=MNSQER(II)+SUMMSE
      AVMSE=SUMMSE/56
      PRINT 67,AVMSE
  67  FORMAT(10X,"THE AVERAGE MEAN SQUARE ERROR FOR THIS RUN IS ",G12.6)
      PRINT 68
  68  FORMAT(////)
      PRINT 69
  69  FORMAT(10X,"THE MEAN SQUARE ERROR OF WORDS TESTING FOR A SINGLE SP
     *EECH ATTRIBUTE IS GROUPED TOGETHER BELOW",////)
      PRINT 66
  66  FORMAT(30X,"***FILLER WORDS***",/)
```

```
      PRINT 52
      DO 71 I=1,29,7
71    PRINT 53,I,MNSQER(I)
      DO 801 I=30,58,7
801   PRINT 53,I,MNSQER(I)
      PRINT 66
      PRINT 72
72    FORMAT(38X,"***VOICING***",/)
      PRINT 52
      DO 73 I=2,29,7
73    PRINT 53,I,MNSQER(I)
      DO 802 I=31,58,7
802   PRINT 53,I,MNSQER(I)
      PRINT 66
      PRINT 74
74    FORMAT(38X,"***NASALITY***",/)
      PRINT 52
      DO 75 I=3,29,7
75    PRINT 53,I,MNSQER(I)
      DO 803 I=32,58,7
803   PRINT 53,I,MNSQER(I)
      PRINT 66
      PRINT 76
76    FORMAT(36X,"***SUSTENTION***",/)
      PRINT 52
      DO 77 I=4,29,7
77    PRINT 53,I,MNSQER(I)
      DO 804 I=33,58,7
804   PRINT 53,I,MNSQER(I)
      PRINT 66
      PRINT 78
78    FORMAT(36X,"***SIBILATION***",/)
      PRINT 52
      DO 79 I=5,29,7
79    PRINT 53,I,MNSQER(I)
      DO 805 I=34,58,7
805   PRINT 53,I,MNSQER(I)
      PRINT 66
      PRINT 64
64    FORMAT(35X,"***GRAVENESS***",/)
      PRINT 52
      DO 81 I=6,29,7
81    PRINT 53,I,MNSQER(I)
      DO 806 I=35,58,7
806   PRINT 53,I,MNSQER(I)
      PRINT 66
      PRINT 82
82    FORMAT(37X,"***COMPACTNESS***",/)
      PRINT 52
      DO 83 I=0,29,7
83    PRINT 53,I,MNSQER(I)
      DO 807 I=36,58,7
807   PRINT 53,I,MNSQER(I)
      STOP
      END
```

Vita

Wayne R. Beeson was born on 29 October 1943 in Dodge City, Kansas. He attended primary and secondary school in Minneola, Kansas, graduating in 1961. In 1966 he received his undergraduate degree in Chemistry and Mathematics from Northwestern State College, Alva, Oklahoma. After graduation from Northwestern, he entered the Air Force and received his commission through the OTS program on 25 November 1966. From December 1966 until October 1967, he attended the Communications Officer Technical School at Keesler AFB, Mississippi. His first assignment was with the 2063rd Communications Group, Lindsey AS, Wiesbaden, Germany. Captain Beeson spent four years in Germany, serving in both the operations and maintenance branches of the 2063rd. In 1971 he was assigned to Air Force Recruiting Service, Detachment 702, Des Moines, Iowa, as the support services officer. In 1975 he entered the Air Force Institute of Technology in the electrical engineering ES Program.

Permanent Address: RFD
Minneola, Kansas 67865

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE — READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFIT/GE/EE/77-9	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): AN ALGORITHM FOR DETERMINING SPEECH INTELLIGIBILITY	5. TYPE OF REPORT & PERIOD COVERED: MS Thesis
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Wayne R. Beeson, Capt, USAF	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Air Force Institute of Technology (AFIT/EN) Wright-Patterson AFB, Ohio 45433	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Project 7071-00-12
11. CONTROLLING OFFICE NAME AND ADDRESS: Mr. Charles Jacobs (AFCS/OA) Hq Air Force Communications Service Scott AFB, Illinois 62225	12. REPORT DATE: December 1977
	13. NUMBER OF PAGES: 52
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Approved for public release; IAW AFR 190-17

JERRAL F. GUESS, Captain, USAF
Director of Information

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Intelligibility
Voice Intelligibility
Algorithm for Determining Speech Intellibility

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A method of predicting speech intelligibility using computer algorithms is presented. Diagnostic Rhyme test number four was used to measure speech intelligibility using a subjective listener test and these results were used as a basis for comparison with the intelligibility predictions made by the computer algorithm. An audio recording of a speaker reading the Diagnostic Rhyme test was made. This recording was run through a General Electric radio system and varying amounts of noise were added. The output of the radio system was recorded, providing a copy of the input word corrupted by both additive noise and

next page

DD FORM 1473, 1 JAN 73 — EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) radio system distortion effects. Both the input recording and the noisy output recording were digitized by sampling the analog waveforms at a 10 kilohertz rate. These digital samples were converted to a frequency format by windowing the time samples with a rectangular window 128 time samples in length and processing them using Fast Fourier transform techniques. This procedure simulated running the analog speech signal through a bank of contiguous narrow bandpass filters covering the range of 0 to 5 KHz, with center frequencies 78 Hz apart. The output of this process was a matrix array, corresponding to each word from the tape, of amplitude values 200 time windows long and divided into 64 frequency bands. These 64 frequency bands were then combined into 1/3 octave groups to model the frequency sensitivity of the average human ear, which reduced the matrix array to 16 frequency bands. This processing of the analog signal was used to model the preprocessing which occurs in the human ear. A comparison between each word from the input tape and the noisy output tape was then made using a weighted mean squared error calculation. This comparison was conjectured to provide a difference measure which is inversely related to intelligibility. This comparison was used to represent how intelligible the input received from the inner ear is to the brain.

Comparison of the intelligibility results from the human listener tests with the computer processing method outlined above gave a Pearson's Correlation Coefficient value of 0.74 which indicates the computer prediction accounted for 55% of the variance in the listener error scores.