

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

12

ADA 128983

Technical Report

647

G. Neben

The Recognition of an Isolated Word
Using Noisy Speech

13 April 1983

Research Department of the Air Force
Contract Number AF33(682)-1-80001-10001
Speech Laboratory
MICHIGAN INSTITUTE OF TECHNOLOGY
LANSING, MICHIGAN 48906



Approved for public release; distribution unlimited.

DTIC
ELECTE
S JUN 6 1983
A

83 06 06 012

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-80-C-0002.

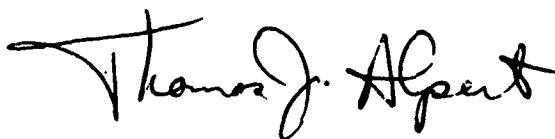
This report may be reproduced to satisfy needs of U.S. Government agencies.

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

A handwritten signature in black ink that reads "Thomas J. Alpert". The signature is written in a cursive style with a large, sweeping initial "T".

Thomas J. Alpert, Major, USAF
Chief, ESD Lincoln Laboratory Project Office

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document
when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

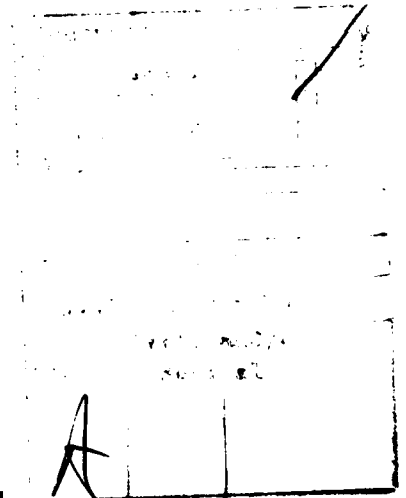
**THE PERFORMANCE OF AN ISOLATED WORD
RECOGNIZER USING NOISY SPEECH**

G. NEBEN

Group 24

TECHNICAL REPORT 647

13 APRIL 1983



Approved for public release; distribution unlimited.


LEXINGTON

MASSACHUSETTS

ABSTRACT*

This report investigates the effects of noise on a speaker dependent, isolated word recognition system. Correct word recognition in a noise-free environment exists in a variety of present-day applications. However, when the acoustic environment includes noise, the problem of correct word recognition becomes more difficult. The noise interferes with the accurate location of the word boundaries and also distorts the spectral representation of the speech waveform.

A series of experiments were performed to determine (1) the effects of using an energy-based endpoint detector and a conventional isolated word recognition system when the input speech is noisy and (2) the effects of placing a noise suppression prefilter in tandem with the word recognizer in an attempt to remove the noise prior to recognition. It was found that the system consisting of the prefilter working in tandem with the word recognizer increased word recognition accuracy.



*This report is based on a thesis of the same title submitted to the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology in February 1983 in partial fulfillment for the degrees of Bachelor of Science and Master of Science.

CONTENTS

| | |
|--|-----|
| ABSTRACT | iii |
| LIST OF ILLUSTRATIONS | vii |
| 1. INTRODUCTION | 1 |
| 2. EXPERIMENTAL SETUP | 5 |
| 2.1 Introduction | 5 |
| 2.2 Format of Speech and Noise Input to the Word Recognition System | 5 |
| 2.3 Recognition Algorithm | 8 |
| 2.4 Endpoint Detector | 10 |
| 2.4.1 Description of the Endpoint Detector Algorithm | 10 |
| 2.4.2 Optimization of the Endpoint Detector for Use in Noise | 14 |
| 2.5 Prefilter | 20 |
| 2.5.1 Description of the Noise Suppression Prefilter | 20 |
| 2.5.2 Optimization of the Prefilter for Use in Noise | 23 |
| 2.6 Signal-to-Noise Specification and Calibration Procedure | 24 |
| 2.7 Electrical Signal Combiner | 28 |
| 2.8 Real-Time Implementation of the System | 29 |
| 3. RESULTS AND CONCLUSIONS | 31 |
| 3.1 Type of Data Collected | 31 |
| 3.2 Performance Evaluation of the Prefilter and the Word Recognizer | 31 |
| 3.3 Evaluation of the Difference in the Endpoints | 38 |
| 3.4 Evaluation of the Best Score | 41 |
| 3.5 Evaluation of the Difference in the Two Best Scores | 41 |

| | |
|------------------------------------|----|
| 4. IDEAS FOR FURTHER INVESTIGATION | 45 |
| ACKNOWLEDGEMENTS | 47 |
| REFERENCES | 48 |

LIST OF FIGURES AND TABLES

FIGURES

| | | |
|-------|--|----|
| 1-1: | Block Diagram of System Configuration..... | 3 |
| 2-1: | Parameterization of Speech Input | 9 |
| 2-2: | Clean Speech "Six"..... | 11 |
| 2-3: | Noisy Speech "Six"..... | 13 |
| 2-4: | Scores for Clean and Noisy Speech "Six"..... | 15 |
| 2-5: | Recognition Accuracy for Different HISTLV Settings | 17 |
| 2-6: | Average Best Score for Different HISTLV Settings..... | 18 |
| 2-7: | Optimized HISTLV Settings for Endpoint Detector..... | 19 |
| 2-8: | Prefiltered Noisy Speech "Six"..... | 21 |
| 2-9: | Scores for Prefiltered Noisy Speech "Six"..... | 22 |
| 2-10: | Optimized SFACTR Settings for Prefilter..... | 25 |
| 2-11: | Configuration for Speech and Noise Input to Recognizer | 27 |
| 2-12: | Schematic of Electrical Signal Combiner..... | 30 |
| 3-1: | Performance Curves for Experiments..... | 33 |
| 3-2: | Performance Curves with Modified Endpoint Detector..... | 37 |
| 3-3: | Average Difference in Endpoints..... | 39 |
| 3-4: | Average Best Score..... | 42 |
| 3-5: | Average Difference in Two Best Scores..... | 43 |

TABLES

| | | |
|------|---|----|
| 2-1: | Random Ordered Lists of Vocabulary..... | 7 |
| 3-1: | Recognition Accuracy for Experiments..... | 32 |
| 3-2: | Recognition Accuracy with Modified Endpoint Detector..... | 36 |
| 3-3: | Average Minimum Energy for Recognizer Alone..... | 40 |

1. INTRODUCTION

Isolated word recognition systems attempt to recognize single words or discrete utterances spoken by a talker. The recognition scheme must be able to pick out the spoken utterance from some recording interval; that is to differentiate the speech sounds from the non-speech sounds that comprise the background noise. Accurately and reliably determining the word boundaries is a critical factor in the performance of a word recognition system [1] and significant research has been devoted to finding acceptable solutions.

The problem becomes more difficult when the acoustic environment includes noise distortion, a situation that is much more realistic. Identifying the word endpoints with background noise (especially when the more troublesome features are involved, such as weak fricatives) requires more sophisticated processing techniques. The use of noise-cancelling microphones may provide some degree of improvement, but they do not completely solve the problem. These microphones fail to sufficiently resolve speech and noise in environments where the signal-to-noise ratio is very low [2].

Background noise creates an additional problem in the form of spectral distortion to the speech waveform. The noise is now coupled with the speech signal and it is this noisy speech that the recognizer must analyze. Depending on the spectral matching techniques that produce word recognition, performance will generally degrade.

This report examines the idea of placing a noise suppression prefilter [3] at the front end of an isolated word recognizer in an attempt to remove the noise prior to recognition. By removing the noise from the speech signal, the recognizer will be able to analyze a cleaner representation of the spoken words. Another benefit of such a system would be that the endpoint detection process could be implemented using existing algorithms, as if it were operating in a noise-free environment.

Three experiments were performed that exploited the use of a flexible prefilter and isolated word recognition system. The experiments used

different combinations of the prefilter and the word recognizer to isolate the effects of endpoint detection and word recognition accuracy in the presence of noise. Figure 1-1 presents a simplified block diagram of the overall system. By controlling the switch settings at A, B, and C, it was possible to configure well-controlled experiments to test the effects of noise on recognition performance with and without the prefilter.

The first experiment was performed with the word recognizer alone. This experiment determined the performance of the recognizer using noisy speech in order to measure the extent to which the recognizer could operate in noise. The next two experiments were conducted with the noise suppressor as part of the system. The effects due to prefiltering the speech for endpoint detection only versus the effects due to prefiltering the speech for endpoints and recognition were examined. The results of the prefilter were then compared with the results of the recognizer alone in the noisy environment to determine what advantages such a system would possess.

For each of the experiments, a new set of reference templates was created. This was necessary since each experiment altered the method in which the recognizer processed the spoken words for recognition. In addition, the reference templates were created from a noise-free environment since this represents the optimum training condition that would be used in practice. The procedure for training the recognizer and generating performance data was identical in each experiment.

To summarize, the following experiments were conducted:

1. Unprocessed endpoints and unprocessed speech. In this case, the recognizer was used alone to select a pair of word endpoints and to analyze the noisy speech input.
2. Prefiltered endpoints and unprocessed speech. In this case, the prefilter was only used to determine a set of word endpoints while the recognizer analyzed the noisy speech input as in (1).
3. Prefiltered endpoints and prefiltered speech. In this case, the prefilter was used to determine a set of word endpoints and to process the noisy speech prior to recognition.

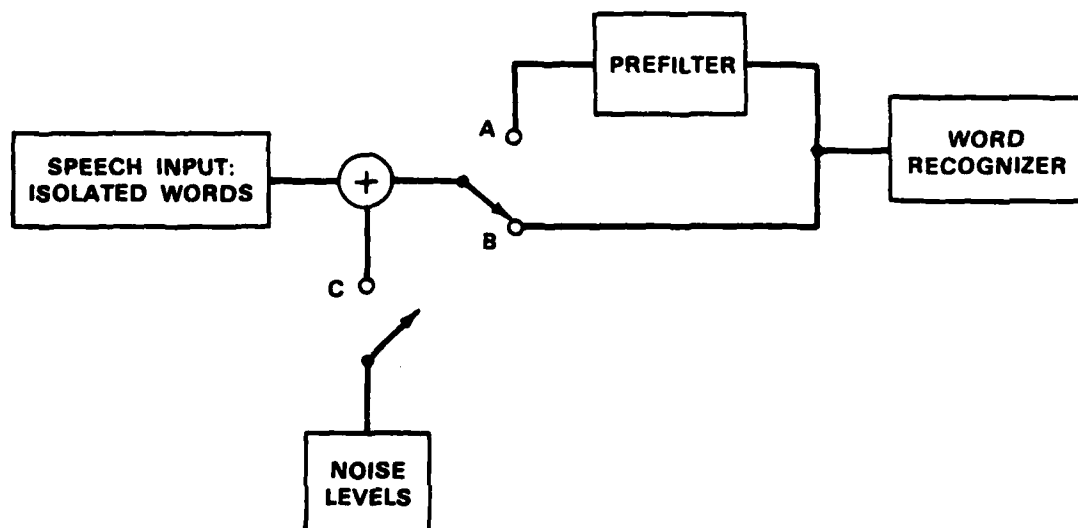


Figure 1-1: Block Diagram of System Configuration.

Chapter 2 details the elements of the system that were used in collecting data for the experiments. The type of speech input to the system, the recognition algorithm, endpoint detector algorithm, calibration and optimization procedures, prefilter, and the details of the real-time system are described. Chapter 3 presents the results of the experiments and the conclusions based on the collected data¹ and Chapter 4 offers ideas for further investigation.

¹A brief summary of some of the research conducted in this thesis will be presented in [4].

2. EXPERIMENTAL SETUP

2.1 Introduction

The following sections detail the components of the prefilter and isolated word recognition system. In addition, a signal-to-noise ratio is defined to measure the different levels of background noise that were coupled to the speech input. A signal-to-noise ratio calibration procedure was then followed at the beginning of every series of experimental runs to insure consistency in evaluating the results from one day to the next.

Two components of the system were optimized to obtain the best possible performance in noise. The endpoint detector was optimized for each noise level when the recognizer was used without the prefilter. This procedure is described in Section 2.4.2. When the prefilter was used, the endpoint detector was not adjusted. Instead, the prefilter was calibrated for the noise according to its normal operating procedure. This procedure is presented in Section 2.5.2. Optimizing the recognition system in this manner allowed the system consisting of the prefilter and the recognizer to be compared with the system using the recognizer alone in the presence of noise.

2.2 Format of Speech and Noise Input to the Word Recognition System

The type of input to the recognition system was high quality speech recorded in a soundproof room using a Sennheiser HMD-224X, noise-cancelling microphone. A typical experiment consisted of processing a pre-recorded training or enrollment session followed by a recognition or test session. Training required the talker to make a pass through the vocabulary so that the recognizer could create the reference templates for the utterances in its dictionary. The words used were from a twenty-word vocabulary used in previous experiments [5], consisting of the digits 0 through 9 and ten command words: start, stop, yes, no, go, help, erase, rubout, repeat, and enter. This vocabulary remained fixed in the experiments. The test run consisted of repetitions or tokens of the same vocabulary from which the recognizer attempted to match the test template against the reference templates.

This format was adhered to during the recording sessions by the talker and was subsequently used to generate real-time data by the recognizer. For the noise experiments, a single tape of F15 aircraft noise was recorded so that it could be combined electrically with the taped speech and applied to the input of the recognizer. Thus, once a tape had been made for the particular talker, it was used as often as required for the different experiments.

The speech tape was produced using a single speaker and the data collected from the experiments are based on this tape. The training portion of the tape was generated by making three passes through random ordered lists of the vocabulary (one pass was used for practice, a second pass was used for training the recognizer, and a third pass was kept as a spare). These lists appear in Table 2-1. Lists A, B, and C were used for training with List C being used for practice. In creating the tape, the male talker was instructed to speak crisply and clearly. Any gross error made in the utterance of one of the training words was re-recorded. Adherence to these instructions was required in order to generate a good training set so that the recognizer could perform reasonably well on the test tokens. Only one template for each word in the vocabulary was stored in the dictionary. The intent was to use an acceptable data base to measure the effects of noise rather than to measure the absolute performance of the word recognizer.

The test portion of the tape was generated on different days by making several passes through random ordered lists of the vocabulary. This part of the tape contains six repetitions of each word represented in the dictionary. Two recording sessions were used, each consisting of three passes through the vocabulary. In Table 2-1, Lists 1-6 comprise the test templates with Lists 1-3 being used during the first recording session and Lists 4-6 being used during the second recording session. The final speech tape contains 140 words: the first 20 representing the reference templates used for training the recognizer and the remaining 120 representing the test tokens used for each recognition run.

TABLE 2-1

RANDOM ORDERED LISTS OF VOCABULARY

| <u>List A</u> | <u>List B</u> | <u>List C</u> | <u>List 1</u> | <u>List 2</u> | <u>List 3</u> | <u>List 4</u> | <u>List 5</u> | <u>List 6</u> |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| erase | help | repeat | 5 | 1 | no | 6 | start | no |
| no | go | 4 | repeat | 8 | 4 | rubout | 0 | 0 |
| yes | stop | erase | 9 | 5 | 3 | yes | yes | start |
| 7 | 8 | 9 | go | 7 | 7 | enter | 8 | 6 |
| go | 2 | 1 | 0 | go | stop | 1 | 9 | go |
| 8 | erase | 2 | enter | repeat | help | 5 | go | yes |
| help | 4 | 3 | 1 | 6 | enter | 8 | enter | help |
| 9 | 6 | 6 | start | 9 | rubout | help | erase | stop |
| start | 3 | 5 | 4 | start | 9 | erase | 3 | 7 |
| stop | yes | help | no | 0 | go | start | 1 | 8 |
| 1 | 9 | 0 | help | no | 5 | no | stop | 1 |
| 4 | rubout | 7 | erase | rubout | 1 | 0 | 4 | repeat |
| rubout | enter | 8 | stop | 4 | 8 | 4 | help | 4 |
| 2 | 1 | stop | 7 | yes | start | repeat | 6 | 3 |
| 0 | 7 | enter | yes | help | repeat | 3 | no | 9 |
| 5 | 5 | go | 6 | 3 | 6 | stop | repeat | 2 |
| 3 | repeat | rubout | 3 | erase | erase | 2 | 5 | rubout |
| 6 | no | start | rubout | stop | 2 | 7 | 2 | 5 |
| enter | 0 | no | 8 | 2 | 0 | go | rubout | erase |
| repeat | start | yes | 2 | enter | yes | 9 | 7 | enter |

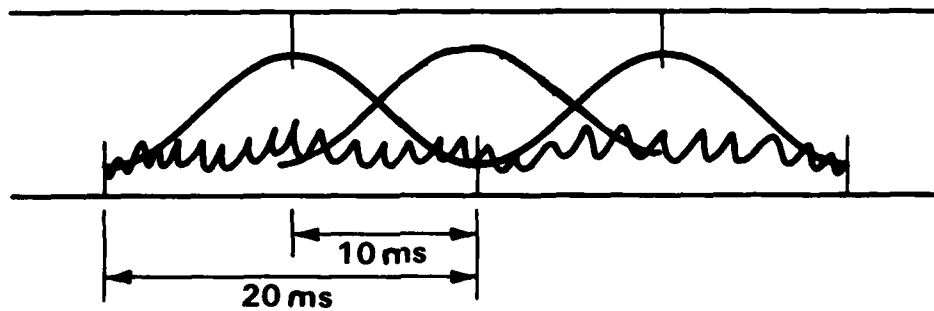
The recognition runs made with the test tokens proceeded automatically once the tape was started. It took approximately fifteen minutes for the recognizer to complete one pass through these utterances. Twenty-five minutes of noise was recorded on the noise tape and this was played simultaneously with the speech input. At the beginning of each recognition run, the noise tape was started at randomly selected locations so that the same noise was not associated with the same words. Each series of recognition runs for a given experiment were repeated as many times as necessary until the results were within 1% to 3% of each other. In general, five or six repetitions of the set of test templates were sufficient to produce very consistent results.

2.3 Recognition Algorithm

The isolated word recognizer uses linear predictive analysis (LPC) to estimate the parameters associated with the all-pole model of the vocal tract. A set of autocorrelation coefficients (r 's) is used to determine the predictor coefficients (a 's) of a 10th order inverse filter that defines the all-pole transfer function.

The parameterization of the speech input is shown in Figure 2-1. The speech signal is sampled at an 8 kHz rate. The parameters are computed with a frame size of 20 ms (160 samples) using a Hamming window and are updated with a frame overlap of 10 ms. When a word is detected, the recognizer processes 150 frames or 1.51 s (150 frames x 10 ms + 10 ms) of speech. Thus, the maximum length of a spoken word to be entered into the recognizer is 1.51 s.

Recognition is achieved using the Itakura distance measure with dynamic time warping implemented using Itakura local constraints and fixed endpoints [6]. The recognizer creates a dictionary by resolving a given set of words into r 's and a 's on a frame-by-frame basis. The test utterance is then compared against each reference template in the dictionary until a best fit is found according to the distance metric.



HAMMING WINDOW

$$w(n) = 0.54 - 0.46 \cos \left(\frac{2\pi n}{N-1} \right),$$

$$0 \leq n \leq N-1$$

WHERE,

N = NUMBER OF SAMPLES PER FRAME

n = nth SAMPLE

Figure 2-1: Parameterization of Speech Input.

2.4 Endpoint Detector

2.4.1 Description of the Endpoint Detector Algorithm

Several approaches to endpoint detection include silence matching algorithms, voiced-unvoiced-silence decisions, and energy level techniques. The purpose of this thesis is not to develop a new endpoint detector for noisy speech, but rather to choose an endpoint detector that has already been implemented, that works relatively well, and that has some provision to handle background noise. The energy-based detector chosen meets these requirements and was used in the word recognition system. This detector is of the explicit type [7] in that a single endpoint pair is chosen and fed forward to the recognition stage. The recognition algorithm then uses these endpoints to make a best guess of the word.

The energy-based endpoint detector that is used in the experiments is based on an algorithm originally described by Rabiner and Sambur [8]. This algorithm used double thresholds to locate the word boundaries. The current detector uses a triple threshold technique to measure the rise and fall of energy levels to determine the word boundaries. For example, Figure 2-2 displays an energy contour of the utterance "six" recorded in a noise-free environment. The beginning of the word is marked by an energy rise from K1 to K2 and the end of the word is marked by an energy decrease from K2 to K3. The gap between the two energy pulses has been smoothed out, thereby correctly identifying the brief silence as part of the word. The important point of this illustration is that the endpoint detector had no difficulty in locating the word boundaries since there was no interference obscuring the energy contour of the word.

The original algorithm by Rabiner and Sambur also used zero crossing information to further refine boundary locations for more difficult features, such as weak fricatives and plosives. There are several reasons why a zero crossing rate is not now being implemented in the detector. According to Wichiencharoen [9], experiments were conducted showing that an energy

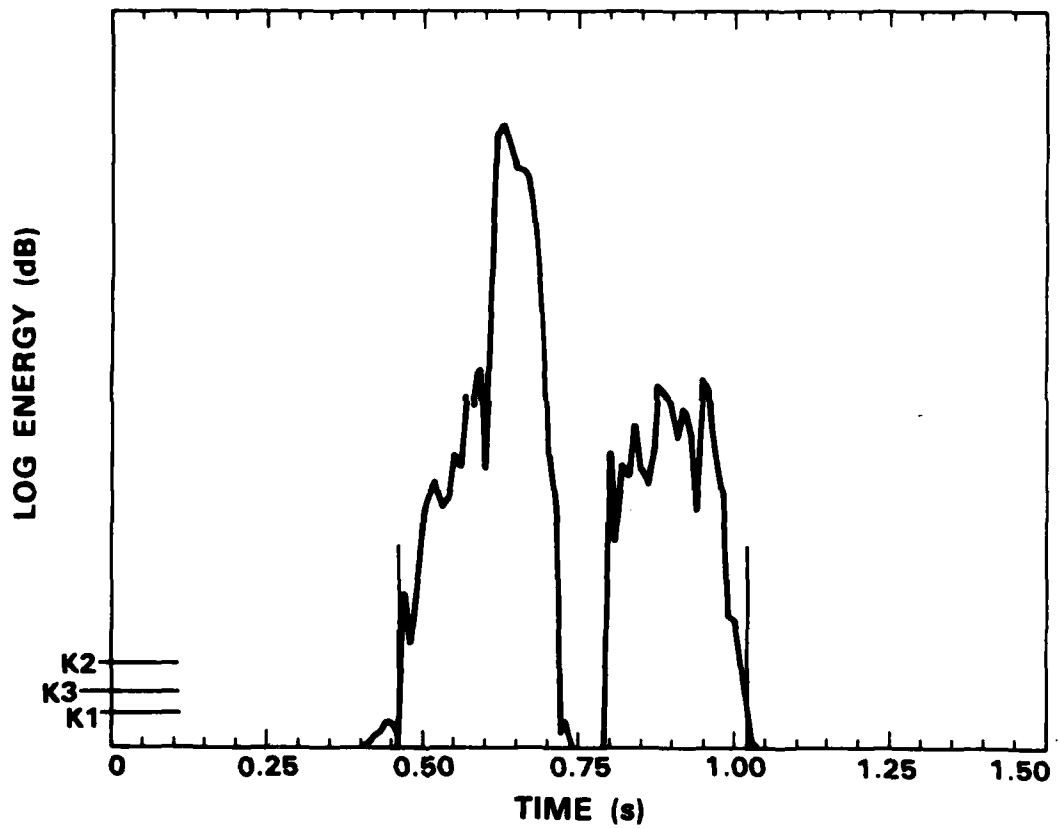


Figure 2-2: Clean Speech "Six."

threshold alone could be used to detect weak fricatives, although determining fricative duration using this method may be suspect. It has also been shown that for narrow-band applications the number of zero crossings is significantly reduced, thereby minimizing its significance [10]. More importantly, there is the observation that a zero crossing rate becomes ineffective in a noisy environment [11].

The addition of noise in the recording environment complicates the word detection process. The energy contour now includes legitimate energy pulses generated by the speech sounds as well as background energy generated by the noise. It was mentioned above that the endpoint detector has a limited capability to adjust to background noise. This is accomplished by first subtracting from the recorded energy interval a minimum energy (MINE) and then forming a histogram of the lower 10 dB points of the energy contour. The mode (MODE) of this histogram is subtracted from the energy contour, giving rise to a final energy display that is processed by the endpoint detector using absolute threshold levels. Thus, this adaptive level equalization procedure [7] normalizes the recorded energy interval by two quantities: MINE, a minimum energy, and MODE, the mode of the histogram. With background noise, this adaptive scheme is necessary in order to compare the energy within the recording interval to the absolute threshold levels used in the endpoint detection process. In the case of low level background noise, the adaptive procedure provides a convenient and acceptable means for locating the word boundaries. However, as shown next for high level background noise, this procedure can no longer discriminate the entire word from the noise. A significant portion of the recorded word is incorrectly identified as noise and is subsequently excluded from the spoken utterance.

Figure 2-3 illustrates the behavior of the endpoint detector when applied to the utterance "six" that was recorded in a low signal-to-noise environment. To better display the effects of noise in Figure 2-3, note that instead of normalizing the energy contour by MINE and MODE, the absolute threshold levels

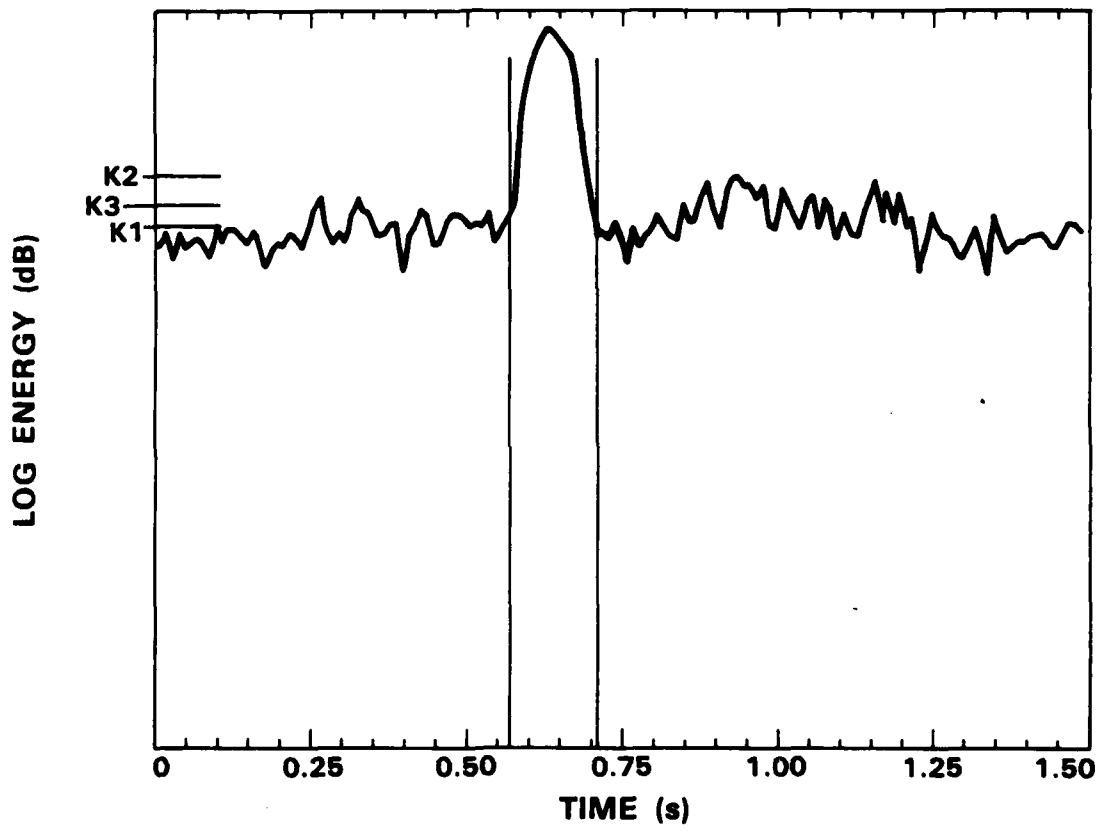


Figure 2-3: Noisy Speech "Six."

are graphically shifted up by the same amount. The endpoint detector attempts to adapt to the noise level by adjusting the energy interval according to the adaptive level equalization procedure. The result is that the endpoint detector fails to correctly locate the utterance "six." Only the peak of the first energy pulse is found and the second energy pulse is completely obscured by noise. The complete picture is seen when this endpoint information is passed to the recognition stage and a best guess is attempted. The four best candidates from the clean speech "six" corresponding to Figure 2-2 appear in Figure 2-4(a), for which the recognition was accurate. However, the noisy speech "six" corresponding to Figure 2-3 was so affected by noise that correct recognition in Figure 2-4(b) was impossible; in fact, the scores exceeded the scale.

The results illustrated in Figures 2-2 to 2-4 indicate how background noise can degrade the accurate location of endpoints and can distort the original speech waveform. This also illustrates the contention that the definition of the word boundaries is a fundamental problem in a noisy environment.

2.4.2 Optimization of the Endpoint Detector for Use in Noise

The endpoint detector adapts to background noise by normalizing the energy contour with respect to a minimum energy and the mode of the lower 10 dB point histogram. This 10 dB value is variable and is defined as a maximum dB histogram level (HISTLV). The HISTLV sets an upper bound on the histogram formed by scanning the 150 frame energy buffer of the recording interval. The **MODE** is then found and is used as the final normalizing quantity for the energy contour.

The HISTLV is an adjustable level for adapting to background noise. To see what effect this level has on recognition, several tests were performed with the system configured as in Experiment 1. The objective of these tests was to set the HISTLV at a value that optimized recognizer performance for a given noise level.

Recognition accuracy was recorded for six sample HISTLV values at seven

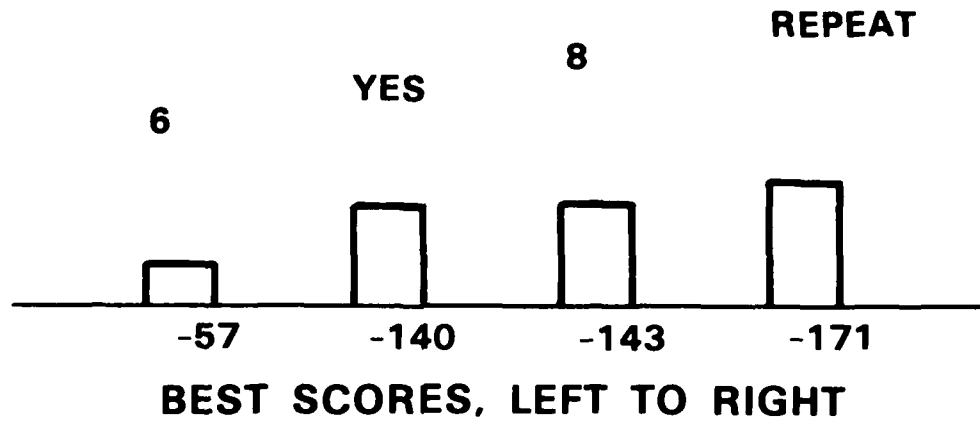


Figure 2-4(a): Scores for Clean Speech "Six."

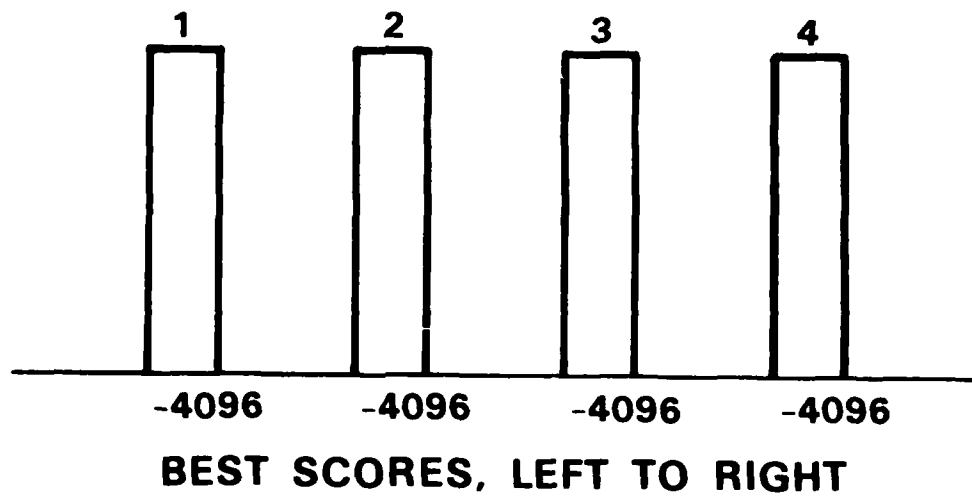


Figure 2-4(b): Scores for Noisy Speech "Six."

different signal-to-noise ratios. This data appears in Figure 2-5. Accuracy was measured by having the recognizer attempt recognition on the identical twenty words that were used for training. The reason for matching the training set against itself was to isolate the effects that noise had on the HISTLV setting and not to include the effects on performance due to repetitions with a larger test vocabulary. As can be seen in Figure 2-5, varying the HISTLV does affect performance for signal-to-noise ratios below 24.6 dB.

To resolve the HISTLV setting at 34.6 dB and 24.6 dB, a second measure was used to provide additional information. The average best score for the recognition runs was examined. With a higher score indicating a better candidate produced by the distance metric matching algorithm, Figure 2-6 illustrates how the two HISTLV settings were further refined. For example, for a signal-to-noise ratio of 34.6 dB, a HISTLV=10 dB should be used to improve performance.

Figure 2-7 shows the optimized HISTLV values as a function of the signal-to-noise ratios. As more noise is coupled to the speech input, one would expect the optimized HISTLV to decrease to maximize recognition accuracy. To see this, consider the case where no histogram is formed and only a MINE normalizes the energy contour. As the noise level increases, less speech energy will be seen by the endpoint detector (as illustrated in Figure 2-3). Consequently, the MINE for the recording interval will increase and the endpoints will move closer together. Now consider the case where a MINE and MODE value normalize the energy contour. As one raises the HISTLV setting, a greater probability exists to normalize the energy contour by a larger MODE value. If the MODE increases, then again the endpoints will move closer together. Thus, as more noise is added to the speech signal, one would expect to see the HISTLV decrease so that more of the valid speech frames will be detected.

Another consideration in evaluating the behavior of the HISTLV value has to

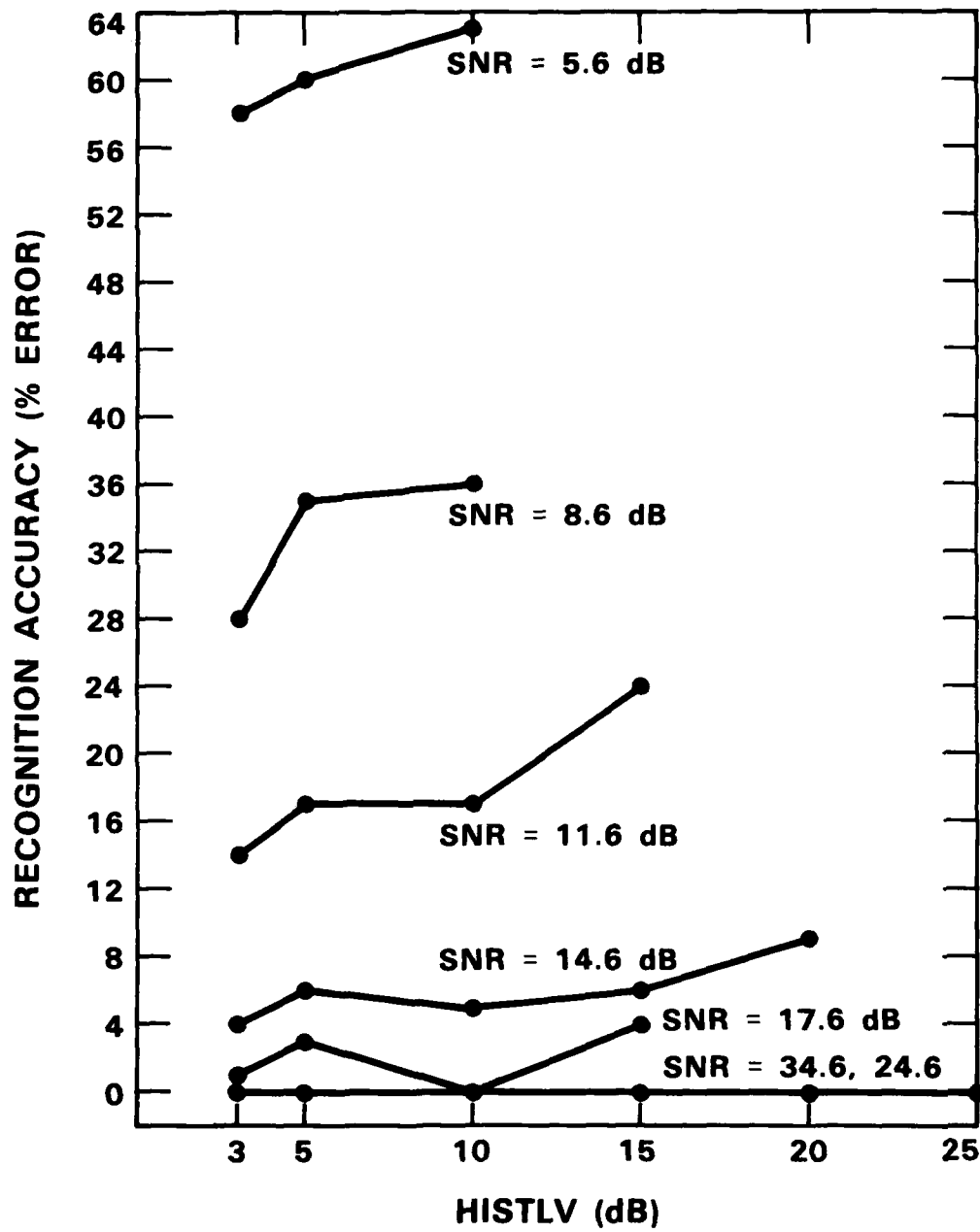


Figure 2-5: Recognition Accuracy for Different HISTLV Settings.

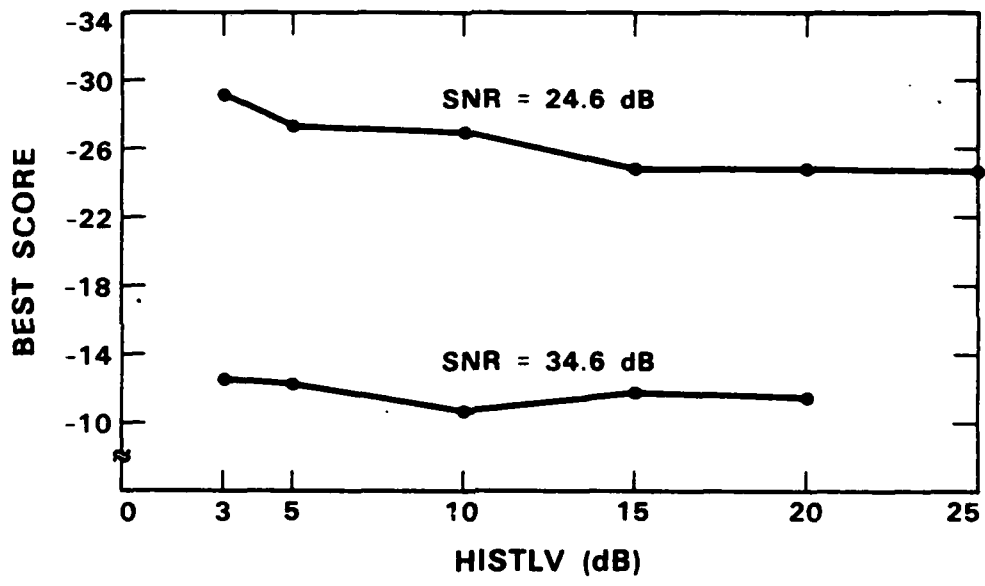


Figure 2-6: Average Best Score for Different HISTLV Settings.

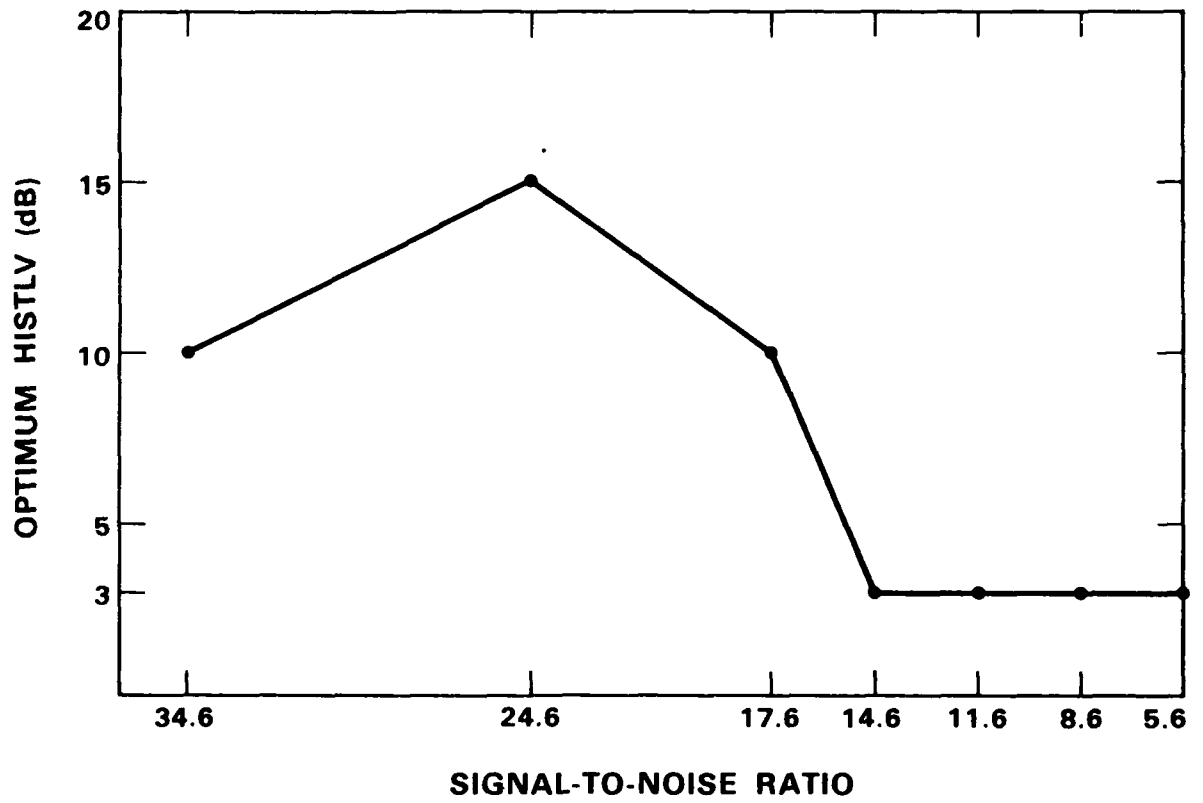


Figure 2-7: Optimized HISTLV Settings for Endpoint Detector.

do with the particular vocabulary that is being used. That is, these HISTLV settings may be vocabulary sensitive (this would explain the slight excursion in the HISTLV value at the 24.6 dB point in Figure 2-7).

The HISTLV settings in Figure 2-7 represent the optimized values for the endpoint detector to achieve the best recognition in noise. Obtaining these values required a laborious procedure and one would not want to repeat it for each new vocabulary and for each new speaker. Moreover, these results are based on a particular type of noise. Noise that exhibits large variations in signal strength during the recording interval would produce a different behavior in the optimized HISTLV values. The prefilter may provide an advantage in useability by allowing the endpoint detector to be preset to one specific HISTLV value for any noise level.

2.5 Prefilter

2.5.1 Description of the Noise Suppression Prefilter

One possible approach to the problem of operating in a noisy environment is to remove the noise from the signal prior to recognition. If the noise were removed, then the speech waveform could be processed in a conventional manner, simply by using the energy-based endpoint detector. This thesis explores the idea of placing the noise suppression prefilter [3] in tandem with the word recognizer. The prefilter would essentially strip the noise from the signal and pass only legitimate speech sounds to the endpoint detector and recognition algorithm. To test this hypothesis, a preliminary experiment was performed using the noisy speech utterance of "six." The same level of noise as in Figure 2-3 was used, but the speech and noise were first passed through the prefilter. The result of the endpoint detection stage is shown in Figure 2-8. Not only is it apparent that a more acceptable set of endpoints was found, but it is also evident that much of the noise had been filtered out. As shown in Figure 2-9, when these endpoints were passed to the recognition stage, the correct word was identified. Thus, the potential for using the prefilter to enhance recognition in noise is worth exploring.

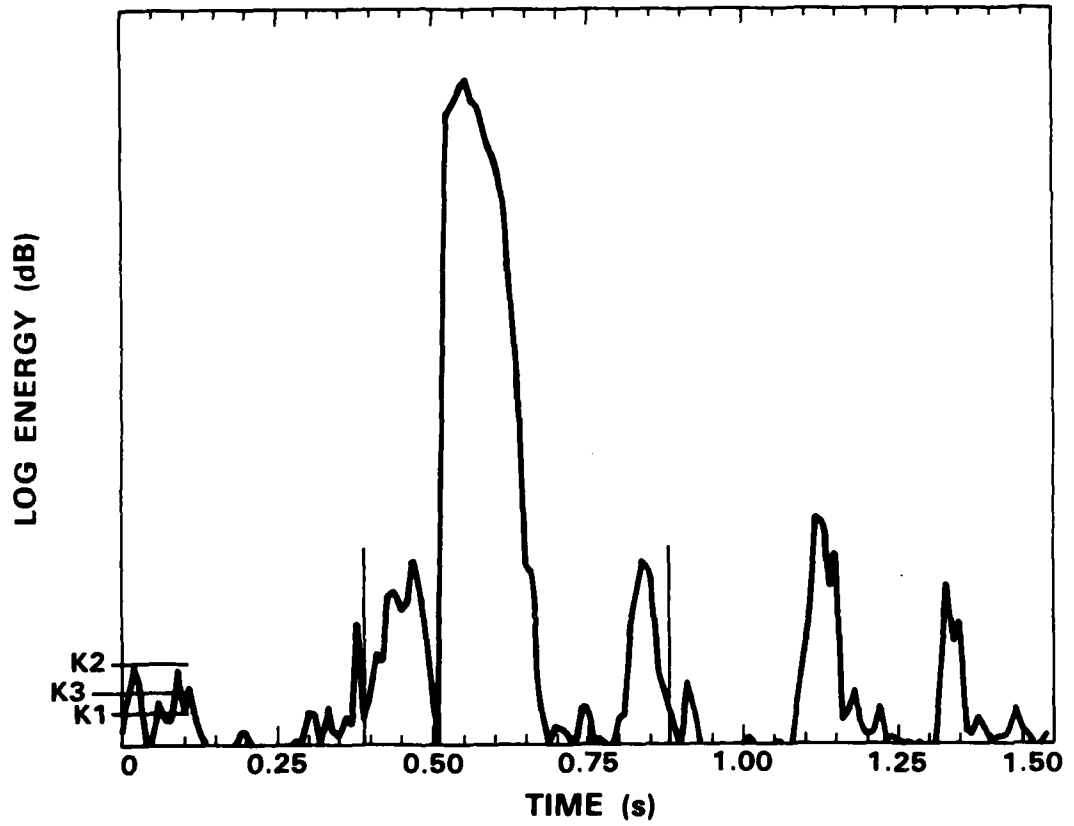


Figure 2-8: Prefiltered Noisy Speech "Six."

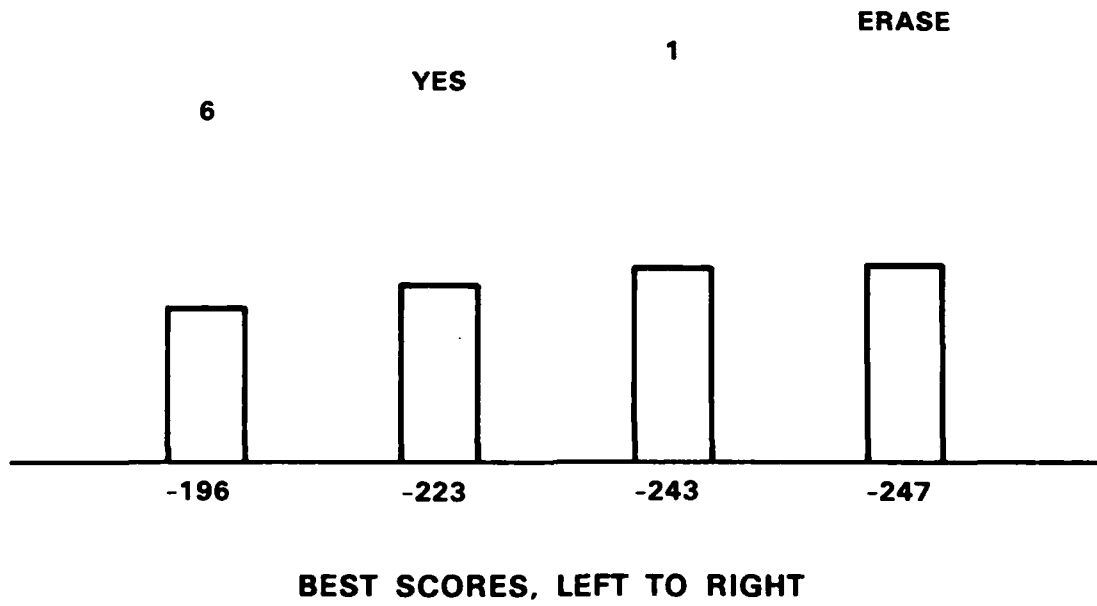


Figure 2-9: Scores for Prefiltered Noisy Speech "Six."

Further experiments used a much larger set of words to assess the performance of the prefilter.

The additional energy pulses in Figure 2-8 are due to the residual noise that remains after the prefiltering process. To remove significant levels of noise from the input speech, penalties are exacted in the form of new distortions to the waveform. This effect must be considered in evaluating the recognition process.

2.5.2 Optimization of the Prefilter for Use in Noise

The prefilter can be adjusted or optimized in the presence of noise. However, the procedure is much simpler and more predictable than adjusting the HISTLV in the endpoint detector. One of fifteen (1-15) noise suppression factors (SFACTR) can be chosen to limit the amount of noise output from the prefilter. For example, a SFACTR=1 will pass the speech and noise to the output of the prefilter unaltered, while a SFACTR=15 will attenuate the noise as much as possible. As the SFACTR is increased, however, the speech signal becomes increasingly distorted. One effect is that the additional energy pulses noted in Figure 2-8 translate into a gurgling type of sound. This residual noise or energy can be mistakenly included as part of the word by the endpoint detector. A second effect due to increasing the SFACTR value is that more of the speech is attenuated. This effect can also occur within the word when multiple energy pulses make up the utterance.

Consequently, there is an optimum SFACTR setting that reduces the processed noise and enhances recognition. Three criteria were used for selecting this value: (1) recognition accuracy, (2) the average best score computed from the distance metric, and (3) listening to the speech output from the prefilter. (The human ear performs a remarkable job in selecting and confirming the choice of SFACTR.) These criteria were used to examine data with the recognition system configured as in Experiment 2. In a manner similar to that of optimizing the HISTLV in the endpoint detector, recognition was based on matching the training set against itself. If recognition accuracy could not

resolve a SFACTR setting for a particular noise level, then the average best score was examined. Likewise, if both recognition accuracy and the average best score proved to be inadequate in choosing a SFACTR value, then the output of the prefilter was monitored. The results appear in Figure 2-10. Plotted are the optimized SFACTR settings as a function of the signal-to-noise ratios. When the prefilter is used in conjunction with the word recognizer, these SFACTR values will be employed to collect performance data.

A final calibration was required to use the prefilter with the word recognizer. The HISTLV in the endpoint detector had to be fixed at some value in order to operate the prefilter independently of the recognizer. Examining the output data at a signal-to-noise ratio of 34.6 dB revealed that the highest MODE in the tested set of words was equal to one. A HISTLV=3 dB was chosen as the fixed, preset value for the endpoint detector. Thus, in Experiments 2 and 3 using the prefilter, only the SFACTR was varied according to its optimized settings.

2.6 Signal-to-Noise Specification and Calibration Procedure

The signal-to-noise ratio is defined on an average frame energy basis. The twenty-word vocabulary used for training the recognizer is the control set used in this energy calculation. The average frame energy enables the user to accurately determine the start-up signal-to-noise level prior to the daily experiments. Once the calibration level is set, data could then be collected at different signal-to-noise ratios.

The average frame energy is computed in the following manner. The autocorrelation value $r(0)$ represents the energy in a particular speech frame. The total energy in a given word is found by summing each $r(0)$ corresponding to the speech frames of the word. The energy in each word is then summed over the entire twenty-word vocabulary. The average frame energy (AFE) is computed by dividing the total energy in this control vocabulary by its corresponding total number of speech frames. Expressed in mathematical terms, the AFE is given by

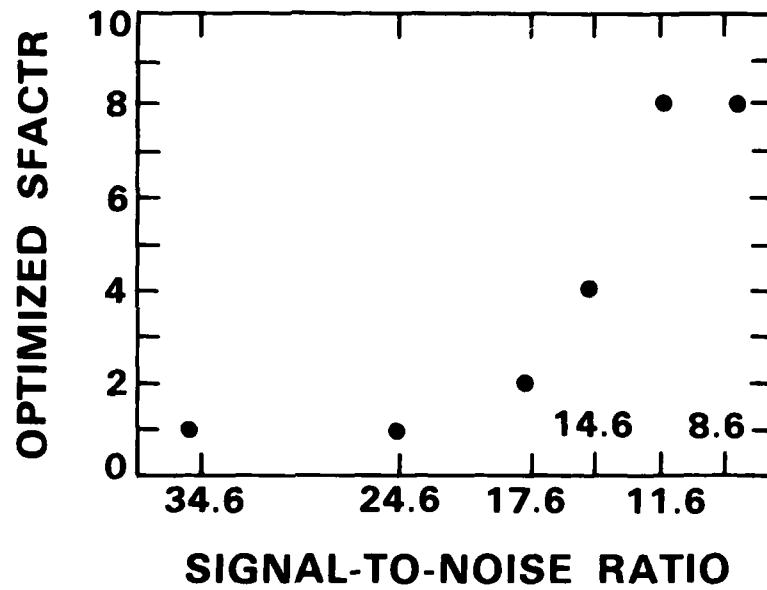


Figure 2-10: Optimized SFACTR Settings for Prefilter.

$$\text{Average Frame Energy} = \text{AFE} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} r_{ij}(0)}{\sum_{i=1}^n m_i} .$$

where,

n = the number of words = 20

m_i = the number of speech frames in the i^{th} word

$r_{ij}(0)$ = the energy in the j^{th} frame of the i^{th} word

Using this procedure, an average frame energy can be computed for the speech signal ($\text{AFE}_{\text{speech}}$) and for the noise signal ($\text{AFE}_{\text{noise}}$). Thus, the signal-to-noise ratio is defined as follows:

$$\text{SNR} = 10 \log_{10} \frac{\text{AFE}_{\text{speech}}}{\text{AFE}_{\text{noise}}} \quad (2-1) .$$

To calibrate the system according to these definitions, it is necessary to examine the electrical connections to the input of the recognizer. Figure 2-11 shows the configuration for the speech and noise input to the recognizer. Basically, the speech and noise are passed through two isolation amplifiers, providing gain and impedance matching, and are then combined electrically before being input to the recognizer.

The noise input level is calibrated by using this configuration with the speech tape turned off. In this case, the endpoint detector forces a "word" detection of length 50 frames so that an energy calculation can be made for a hypothetical twenty-word vocabulary of noise.² The only criterion used in setting the gain levels of the system devices was that there be a wide enough range of noise available at the input of the recognizer to simulate a low signal-to-noise ratio environment as well as a high signal-to-noise ratio environment. Using the noise tape, a 50 dB calibration setting was chosen for the HP-350D attenuator which, when one listens to the tape output, produces a

²In calibrating the noise and speech inputs, a HISTLV=10 dB was used.

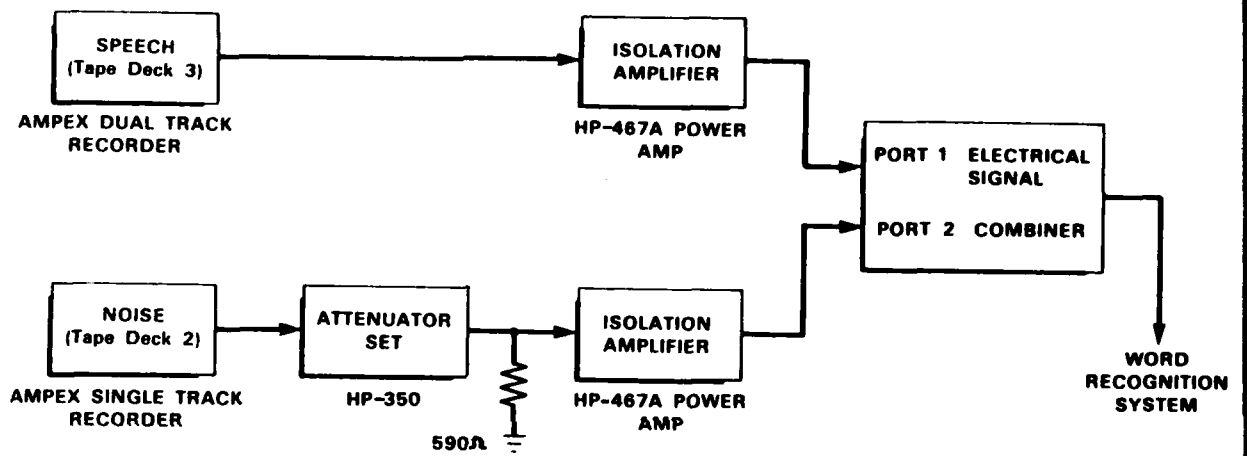


Figure 2-11: Configuration for Speech and Noise Input to Recognizer.

low noise level. The average frame energy for noise was computed to be

$$AFE_{\text{noise}} = 2.637e-06 .$$

Thus, a 40 dB attenuator setting, for example, produces a 10 dB increase in noise from the calibration setting.

In a similar manner, the speech tape is calibrated with the noise source turned off. The criterion used in setting the gain controls was that the speech have a maximum gain at the input to the recognizer without overdriving the analog-to-digital converter. The average frame energy for speech was found to be

$$AFE_{\text{speech}} = 7.564e-03 .$$

Thus, according to equation 2-1,

$$SNR_{\text{cal}} = 10 \log_{10} \frac{7.564e-03}{2.637e-06} = 34.6 \text{ dB} .$$

This value represents the calibrated signal-to-noise ratio used at start-up. The different signal-to-noise environments are simulated by varying only the noise level of the attenuator from the calibration setting.

As a consistency check on this procedure, one can examine the maximum signal-to-noise ratio obtained with the noise attenuated as much as possible. In this case,

$$SNR_{\text{max}} = 10 \log_{10} \frac{7.322e-02}{1.254e-07} = 47.7 \text{ dB} .$$

The analog-to-digital converter produces sixteen bit samples. At 3 dB/bit, one would expect a maximum accuracy of about 48 dB. This agrees with the experimentally determined value.

2.7 Electrical Signal Combiner

The electrical signal combiner is used to combine the speech signal with the noise signal for input to the word recognition system. The schematic for

this device appears in Figure 2-12. It is a passive circuit which weights the inputs equally by the formula

$$v_{out} = .33(v_1 + v_2).$$

Impedances are matched such that the recognizer sees a 600 ohm source.

2.8 Real-Time Implementation of the System

The recognition algorithm, endpoint detection scheme, and the prefilter exist completely in software and are run on a Lincoln Digital Signal Processor (LDSP). An outboard memory providing up to 128K is accessed by the LDSP and is used for storing and retrieving the dictionary required during recognition.

To permit the collection of a large amount of data, the system is capable of running in real-time. Utterances need only be separated by a few seconds of silence before the recognizer begins scanning for a new word. As mentioned in Section 2.5.2, a port is accessible for listening to the output of the prefilter as it is being input to the recognizer. Similarly, one can also listen to the output of the word recognizer, which reproduces the input signal until a word has been detected. Thus, the user can acoustically monitor the processing of the spoken words.

The LDSP is connected to a host PDP-11/45 computer through an I/O port. This connection allows continuous and real-time monitoring of the performance of the word recognizer. The output of the endpoint detection stage, including endpoints and energy normalizations, as well as the best four candidates from the recognition stage are monitored. This data is displayed visually on a VT11 graphics terminal and a VT52 data entry terminal. The prefilter software is run in a second LDSP. Using coax cables, the prefilter is connected to the front end of the recognizer, enabling the data collection facilities to operate exactly as before. All of the information is automatically stored in files for future hard copy and processing.

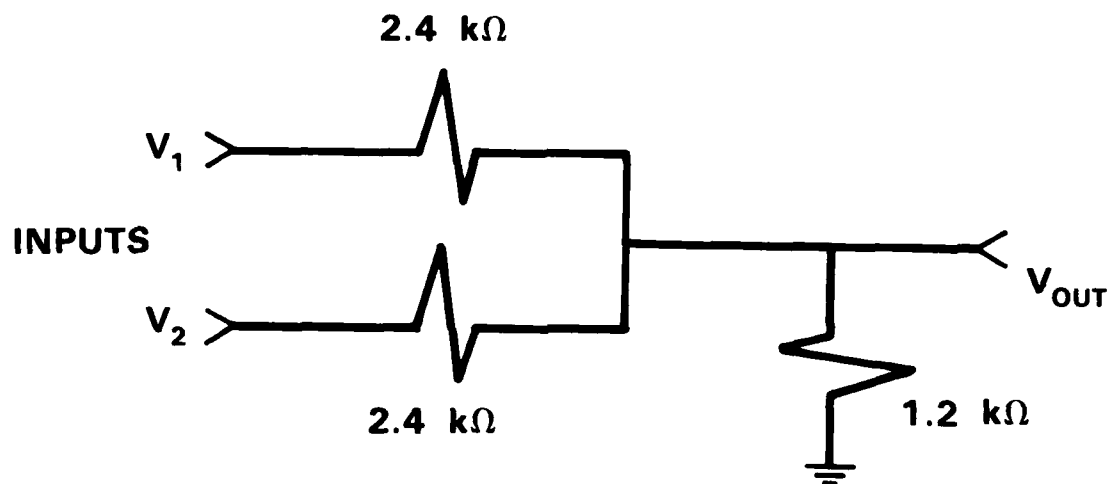


Figure 2-12: Schematic of Electrical Signal Combiner.

3. RESULTS AND CONCLUSIONS

3.1 Type of Data Collected

Four statistics were measured during the experiments: performance, the difference in the endpoints (word length), the best score, and the difference in the two best scores. The performance of the recognizer, with and without the prefilter, is expressed as a percentage of the words recognized correctly from the 120 test tokens used during recognition. The difference in the endpoints, as determined by the endpoint detector, is measured in speech frames. The best score measures the accuracy of the match between the test token and the best choice from the dictionary of the recognizer. A higher score indicates a better match. The difference in the two best scores can be looked upon as a type of quality measure for performance. The greater the difference between the first candidate and the second candidate, the less likely the recognizer will confuse two words.

For each of these statistics, an average for the entire 120 word recognition run was taken. Since five or six repetitions of this run were performed to complete a portion of the experiment, a final average was computed over the repetitions. All of the data were collected at the six signal-to-noise ratio points of 34.6 dB, 24.6 dB, 17.6 dB, 14.6 dB, 11.6 dB, and 8.6 dB. In the experiment using the recognizer alone, an additional data point at 5.6 dB was collected.

3.2 Performance Evaluation of the Prefilter and the Word Recognizer

Table 3-1 lists the performance results for the three experiments defined in Chapter 1. These data are plotted in Figure 3-1 as performance curves for the different signal-to-noise ratios. The curve representing the prefiltered endpoints and prefiltered speech experiment begins at a noticeably lower accuracy than the other curves for the 34.6 dB calibration point. The reason for this is that only one template for each word in the vocabulary was stored in the dictionary of the recognizer. When unprocessed speech was used, this method was acceptable. However, when prefiltered speech was used, generating

TABLE 3-1
 RECOGNITION ACCURACY FOR EXPERIMENTS

| Signal-to-Noise Ratio (dB) | Unprocessed Endpoints and Unprocessed Speech | Prefiltered Endpoints and Unprocessed Speech | Prefiltered Endpoints and Prefiltered Speech |
|----------------------------------|---|---|---|
| | (%) | (%) | (%) |
| 34.6 | 98.3 | 99.2 | 96.6 |
| 24.6 | 96.5 | 97.1 | 96.9 |
| 17.6 | 94.9 | 95.6 | 97.1 |
| 14.6 | 90.3 | 91.8 | 93.0 |
| 11.6 | 78.0 | 80.5 | 81.4 |
| 8.6 | 53.8 | 62.5 | 67.1 |
| 5.6 | 34.4 | | |

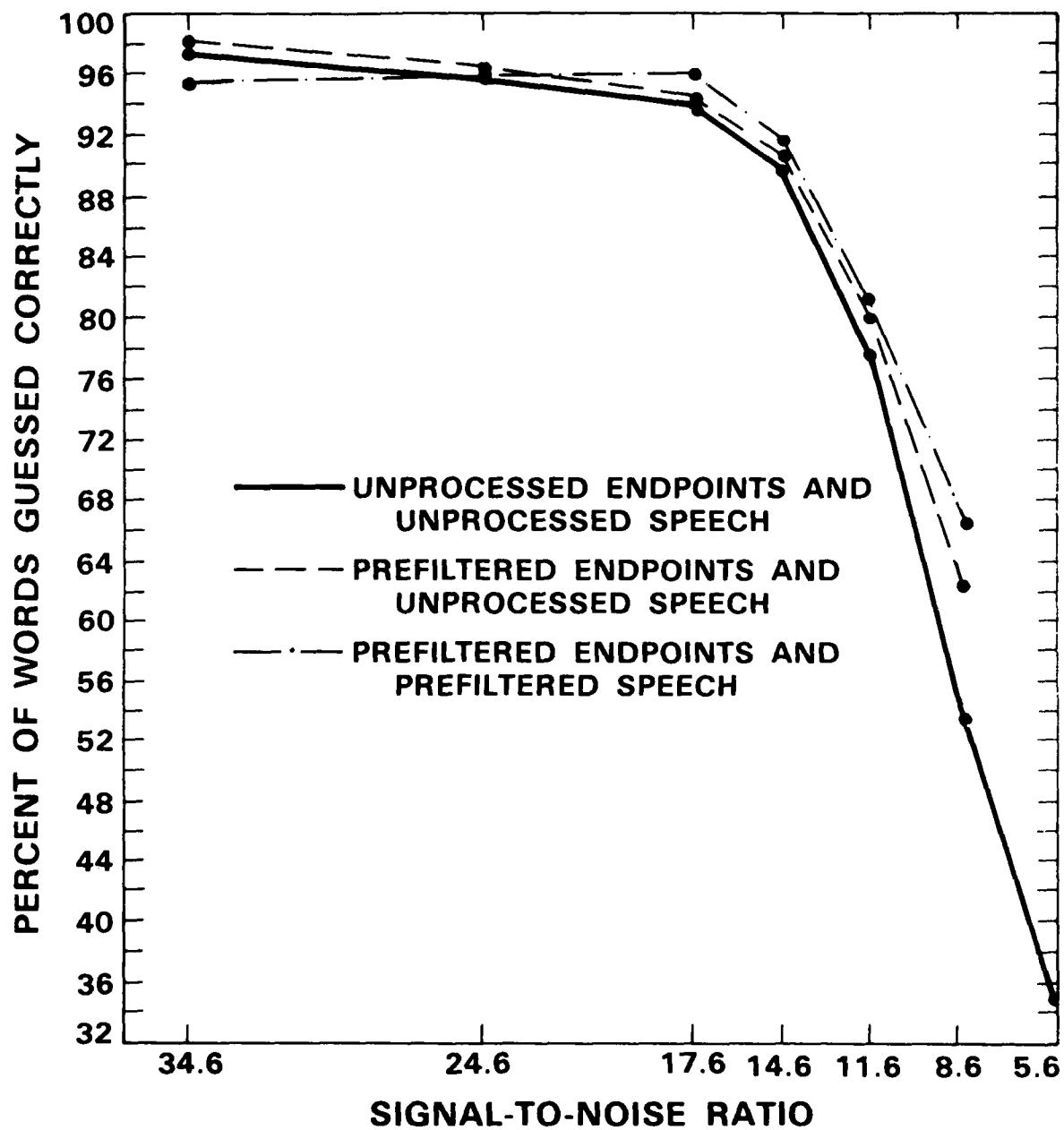


Figure 3-1: Performance Curves for Experiments.

a good dictionary became more critical since a few of the words were distorted by the prefiltering process. The recognizer may have found it more difficult to match some of the test tokens with only a single representation of this word in its dictionary. This could cause recognition performance to be lower in the absence of noise. When a small amount of noise was added to the speech signal, the noise actually smoothed out some of the utterances. At the signal-to-noise ratio of 24.6 dB, this smoothing may have improved performance to the point where the results were again consistent with the other experiments. For the performance results described below, the first data point at 34.6 dB is excluded from the calculations.

Following are several conclusions which can be drawn from the data in Figure 3-1.

1. Given the three experiments conducted, the best possible performance from the recognizer is achieved when prefiltered endpoints and prefiltered speech are used. By placing the prefilter in tandem with the recognizer and allowing it to process the noisy speech prior to recognition, recognition accuracy improved over that of using the recognizer alone or using the prefilter just to find the endpoints. The average improvement in performance over the recognizer alone, taken over five signal-to-noise test points (24.6 dB - 8.6 dB), is 4.4%.

This improvement was attained with no attempt at modifying the original prefilter or recognizer (other than optimizing the SFACTR in the prefilter and the HISTLV in the recognizer). The distortion to the speech waveform introduced by the prefiltering process was still inherent in the system. Particularly, it was noted that the prefilter produced additional energy pulses surrounding the word or embedded within the word. These pulses became more visible in terms of frequency of occurrence and greater amplitude at higher suppression factor settings. This type of distortion may have negative affects on recognition accuracy. The pulses surrounding the word interfere with the accurate location of the word boundaries while, within the word,

there are distortions to the spectral representation of the speech.

An attempt was made to remove these extraneous energy pulses from the endpoint detection process by setting a level which the peak in each detected pulse must exceed in order for it to be declared a legal pulse. This modification was made in the endpoint detector in the recognizer. While the pulses generated by the prefilter were not actually removed from the system, it was hoped that the endpoint detector would not include these pulses as part of the word.

Using the modified endpoint detector, a fourth experiment was performed and a substantial improvement in performance over Experiment 3 was observed. The new data is listed in Table 3-2 and plotted in Figure 3-2 with the previous performance results. The average improvement in performance over the recognizer alone, taken over the same five signal-to-noise test points, is 7.0%. It is also interesting to note that performance remained essentially constant down to a signal-to-noise ratio of 14.6 dB before dropping off. Apparently, the additional energy pulses adversely affects the selection of the word boundaries and, subsequently, recognition accuracy.

One must take care in concluding that the system using prefiltered endpoints and prefiltered speech is the best possible system. Of the three principal experiments conducted, this is true, but the experiment using unprocessed endpoints and prefiltered speech was not performed. This experiment would need to be performed to draw the general conclusion of an overall best system.

2. Given that the recognition system is operating with unprocessed noisy speech, it is better to use prefiltered endpoints rather than unprocessed endpoints. Experiment 2 used the prefilter to process the input speech to only determine a set of word endpoints. The recognizer then used these endpoints to extract the word from the original noisy speech waveform. This proved to be a better approach than allowing the recognizer to select its own

TABLE 3-2

RECOGNITION ACCURACY WITH MODIFIED ENDPOINT DETECTOR

| <u>Signal-to-Noise Ratio (dB)</u> | <u>Prefiltered Endpoints and Prefiltered Speech (%)</u> |
|---------------------------------------|---|
| 34.6 | 94.7 |
| 24.6 | 96.6 |
| 17.6 | 96.0 |
| 14.6 | 96.2 |
| 11.6 | 88.1 |
| 8.6 | 71.8 |

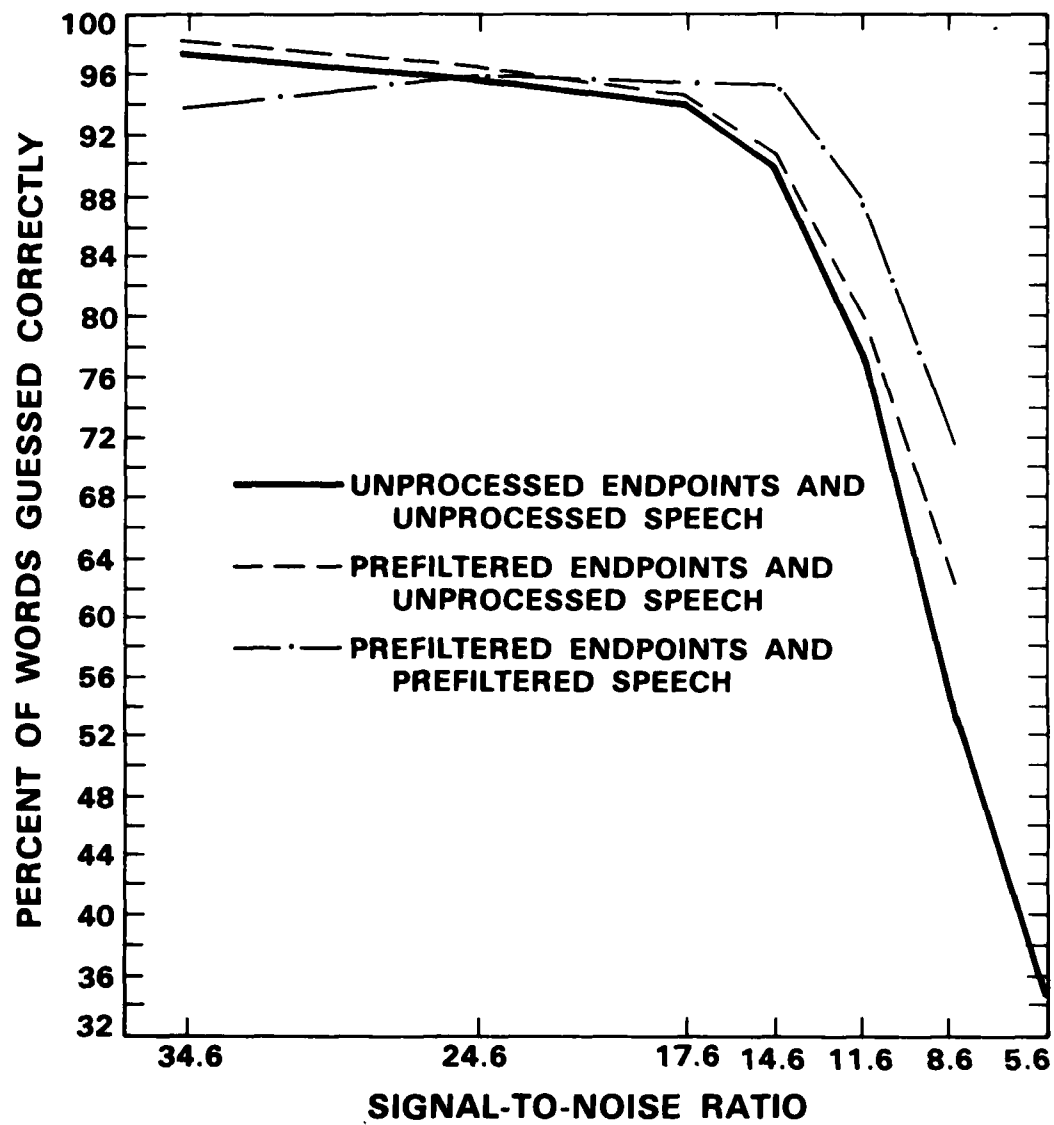


Figure 3-2: Performance Curves with Modified Endpoint Detector.

endpoints as in Experiment 1. The improvement in recognition accuracy over five signal-to-noise test points is 2.8%.

3. Given that the recognition system is operating with prefiltered endpoints, it is better to use prefiltered speech rather than unprocessed speech. Experiment 3 used the prefilter to not only select the word endpoints but to process the noisy speech as well. The recognizer then used the prefiltered speech in its spectral matching algorithm. It was found that this approach worked better than allowing the recognizer to analyze the unprocessed speech as in Experiment 2. The improvement in recognition accuracy over five signal-to-noise test points is 1.6%. For Experiment 4 using the modified endpoint detector, this improvement is 4.2%.

3.3 Evaluation of the Difference in the Endpoints

The results of the variations in endpoint locations due to the additive noise are displayed in Figure 3-3. As predicted in Section 2.4.2 for the experiment using the recognizer alone, the addition of noise caused a reduction in the difference between the endpoints. As the noise increased, the energy contour was normalized by a greater minimum energy. Table 3-3 shows this effect on MINE in Experiment 1 for the different signal-to-noise ratios. Since more of the valid speech frames were blanketed by noise, the word boundaries shifted closer together.

The prefiltered endpoints react quite differently to the increased noise levels. For the prefilter, the difference in endpoints remains essentially constant down to 14.6 dB. The curve characterizing the prefilter and the modified endpoint detector remains extremely flat down to 11.6 dB before dropping off. The fluctuations in the prefiltered endpoints are most likely due to the tradeoff between the suppression factor setting and the resulting residual noise and attenuation that a higher setting produces. For example, consider Experiments 3 and 4 using prefiltered endpoints and prefiltered speech. Between 34.6 dB and 14.6 dB, the residual noise produces additional energy pulses that the endpoint detector locates and includes as part of the

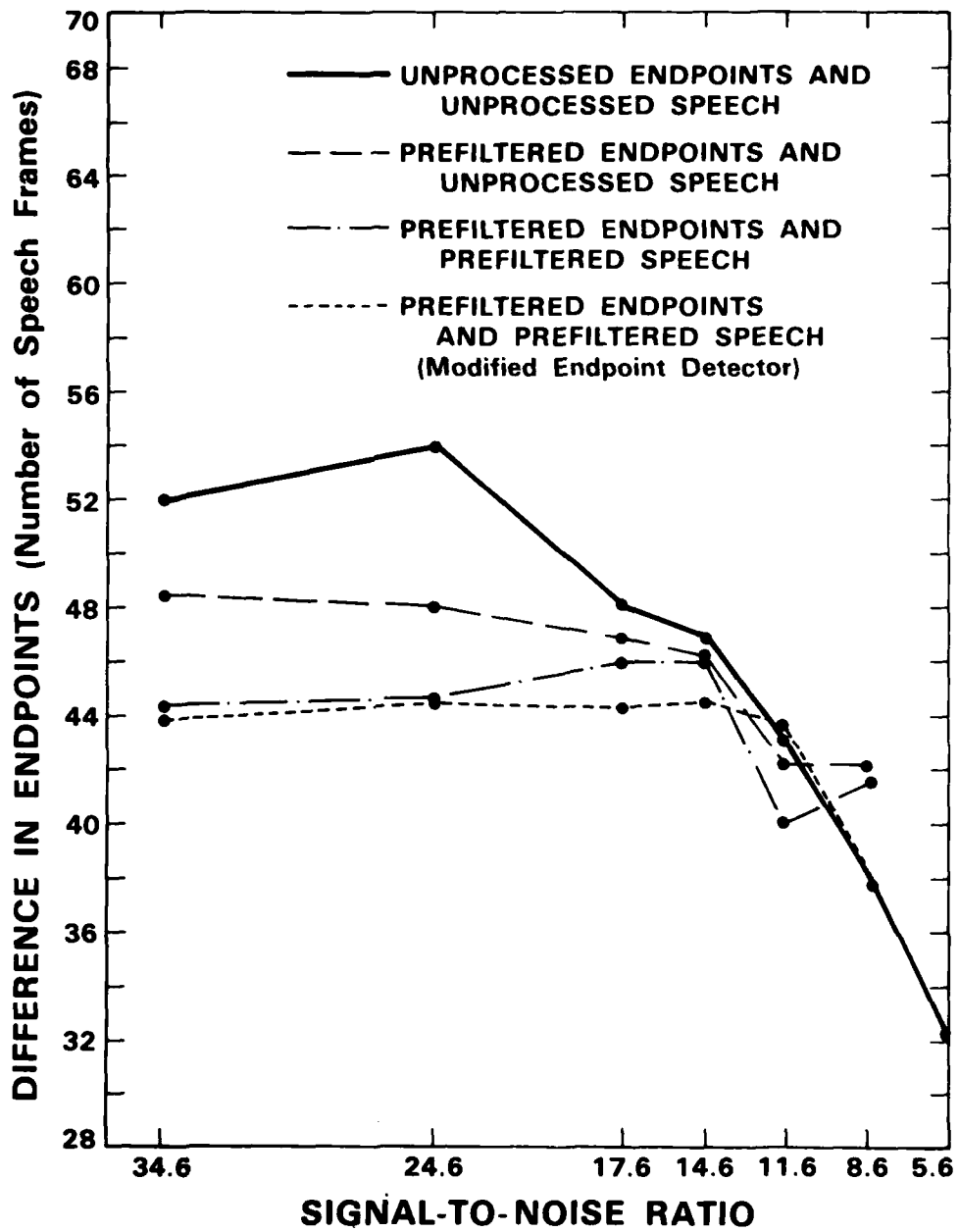


Figure 3-3: Average Difference in Endpoints.

TABLE 3-3

AVERAGE MINIMUM ENERGY FOR RECOGNIZER ALONE

| <u>Signal-to-Noise Ratio (dB)</u> | <u>Average MINE (dB)</u> |
|-----------------------------------|--------------------------|
| 34.6 | 33.1 |
| 24.6 | 36.0 |
| 17.6 | 100.9 |
| 14.6 | 127.9 |
| 11.6 | 155.7 |
| 8.6 | 190.7 |

word. The result is that the word boundaries move further apart as the noise increases and higher suppression factor settings are used. Notice that the modified endpoint detector performs a much better job in eliminating these extra energy pulses. Beginning at 14.6 dB, however, the prefilter begins to noticeably attenuate the speech signal as well as the noise input. Despite the fact that additional energy pulses are present, more of the speech signal is suppressed and, thus, the word boundaries again move closer together.

Ideally, the desired result would be no change in the endpoints as the noise is increased. This would indicate that the additional noise is having no affect on the endpoint detection process. Any degradation in recognizer performance would then be due to the spectral distortion of the speech waveform. The prefilter, when used in conjunction with the modified endpoint detector, comes very close to realizing this goal.

3.4 Evaluation of the Best Score

The results of the best score as a function of the signal-to noise levels are presented in Figure 3-4. No one curve exhibits a clear advantage over the others in terms of having a better or higher score for all of the test points. The only exception would be with Experiment 4, using the prefilter and the modified endpoint detector, where the curve does seem to offer a slight improvement in the best score. In general, all four curves produce increasingly worse scores as additional noise levels are added to the speech signal.

The merits for using this data may be in setting a threshold for false alarms. That is, if the guesses made by the recognizer begin to exceed this threshold, one would reject the input and request another repetition. This would have the effect of maintaining a desired recognition performance, but at the expense of increased repetitions.

3.5 Evaluation of the Difference in the Two Best Scores

The results of the difference in the two best scores as a function of the signal-to-noise ratios are plotted in Figure 3-5. As mentioned in

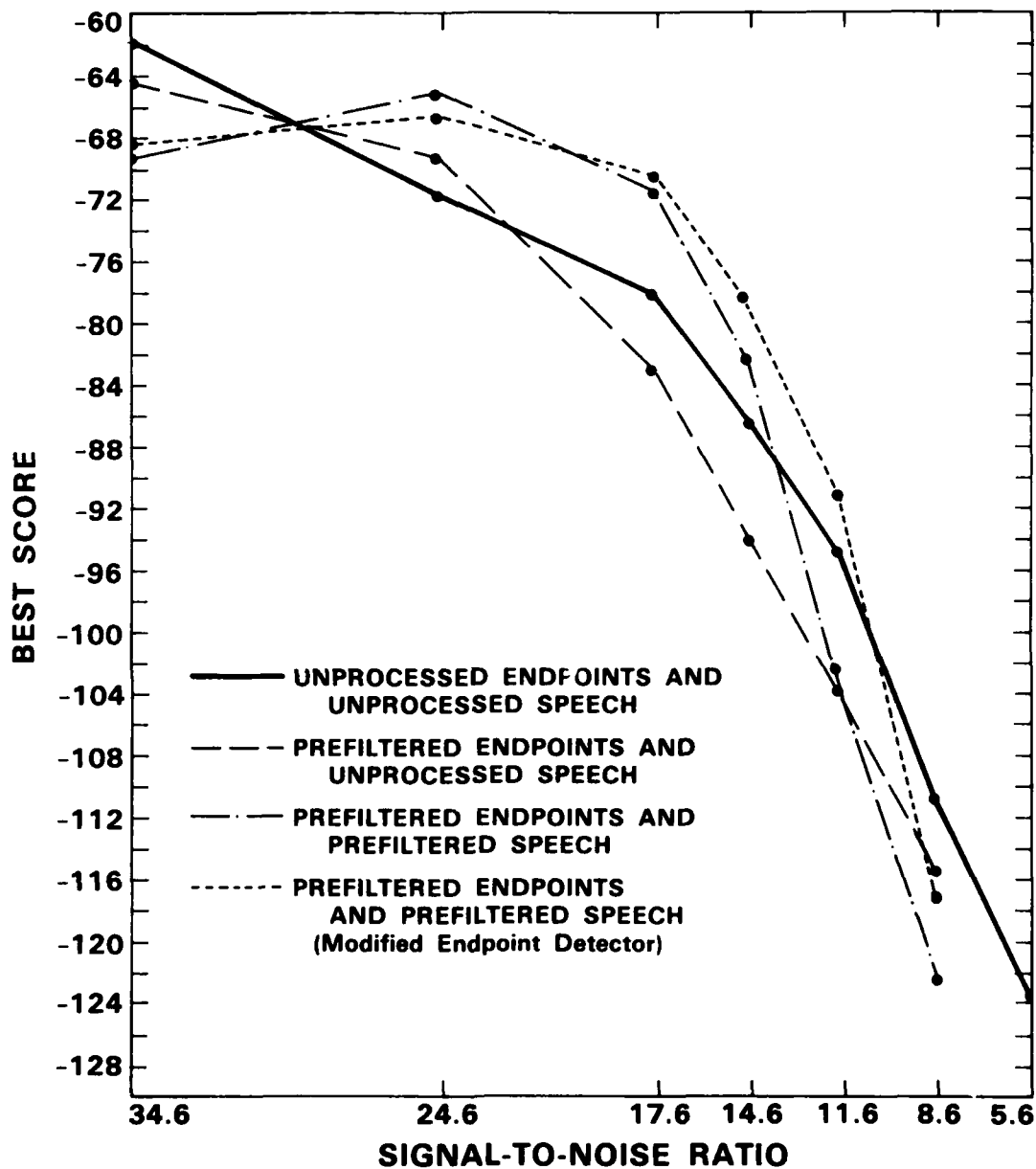


Figure 3-4: Average Best Score.

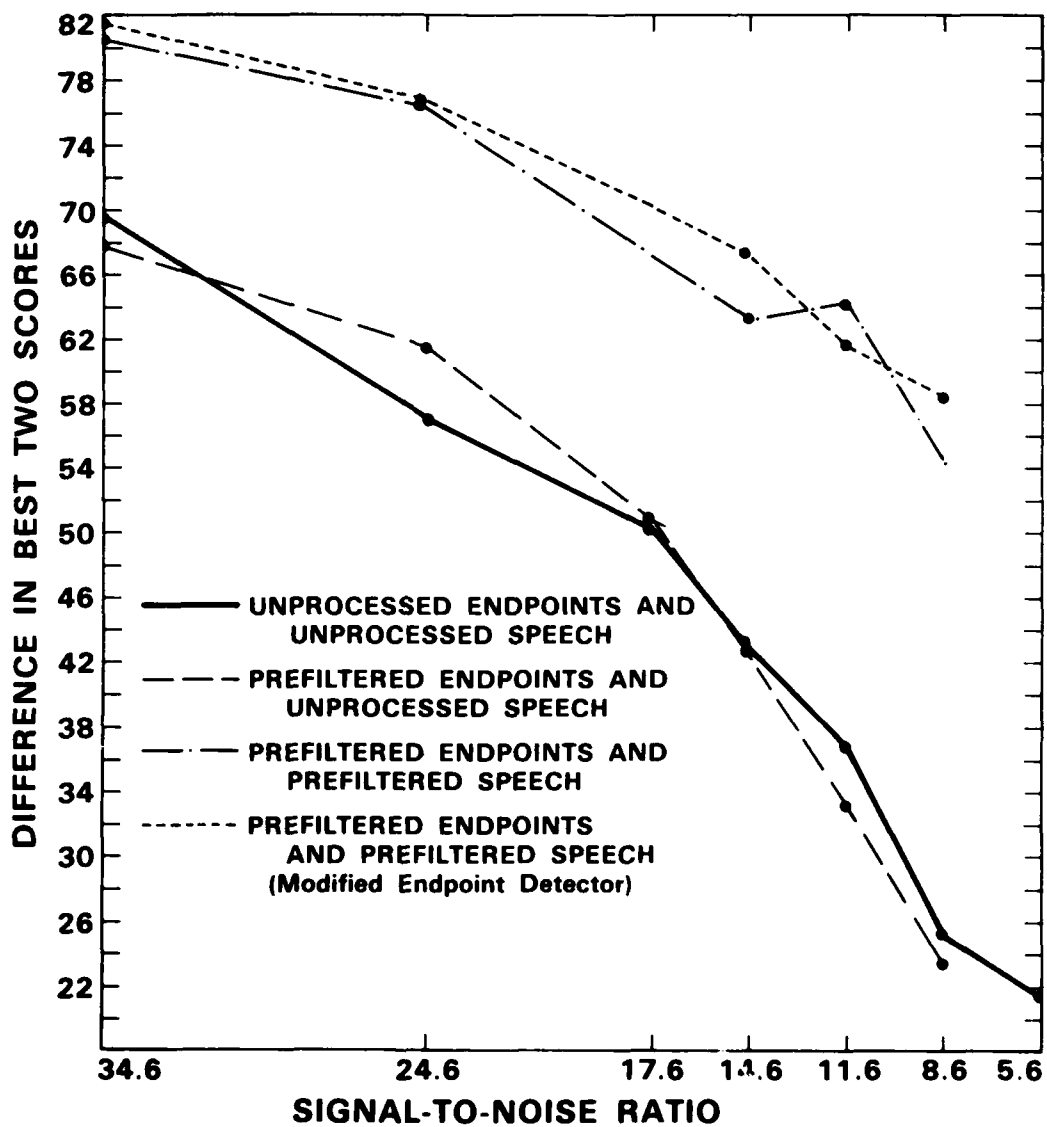


Figure 3-5: Average Difference in Two Best Scores.

Section 3.1, one would ideally want this difference to be as great as possible so that the recognizer would be less likely to confuse two words. This appears to be true in comparing Experiment 3 with Experiment 1 which shows that the difference in the two best scores is much greater when prefiltered endpoints and prefiltered speech are used rather than when the recognizer is used alone. The average improvement, taken over six signal-to-noise test points (34.6 dB - 8.6 dB), is 20.8 scoring points. The average improvement for the prefilter and the modified endpoint detector over the recognizer alone is 22.5 scoring points.

The increase in this difference is reflected in the improved performance of the recognizer. The performance in Experiments 3 and 4 using prefiltered endpoints and prefiltered speech was substantially better than that observed in Experiment 1 using unprocessed endpoints and unprocessed speech. Care should be taken in interpreting this quality measure. The results show that an increase in the difference between the two best scores also corresponds to an improvement in performance. However, the converse is not necessarily true, as Experiment 2 demonstrates. An improvement in performance may not correspond to an increase in the difference between the two best scores.

4. IDEAS FOR FURTHER INVESTIGATION

Working with the noise suppression prefilter and the word recognizer has suggested new ways in which the two systems could be linked together to provide better recognition performance and ease of use. With minimal work, a software system similar to the one used in this thesis could be configured to explore new ideas. Following are additional ideas for further research.

1. One idea would be to apply some weighting function to emphasize frames in higher signal-to-noise areas over those frames in lower signal-to-noise areas. Weighting the frame scores could be a first-cut approach to this idea. Frames with little signal energy and an equal or greater amount of noise energy would be scored lower than frames with a large amount of signal energy. A signal-to-noise ratio would have to be determined for each frame, perhaps by using simple energy calculations as in the endpoint detector. The weighting function could correspond to a vertical energy scale in much the same way the absolute energy thresholds for the endpoint detector are set.

This approach might yield better performance for two reasons. First, assuming that the word endpoints are not perfect and are off by some number of frames, the frame scores near the word boundaries would not contribute a significant error to the overall word score. The frames near the word boundaries would naturally be located in the lower signal-to-noise areas. Second, it is anticipated that in the frames where the signal energy is much greater than the noise energy, the recognition analysis and spectral matching process will perform better and result in more useful scores. The first step in gaining a better understanding for this research idea would be to trace through typically recognized words, frame by frame, at various noise levels and see what kind of scores are generated.

2. In conjunction with (1) and to improve the location of the word endpoints, it might be a good idea to average the frame energy among several neighboring frames. This would present a smoother energy contour to the endpoint detector. If (1) were implemented, this smoothing might produce an

improvement in performance by affecting the way in which the signal-to-noise ratio is determined for each frame. Likewise, the beginning point and the ending point of the word would change slightly since the energy rise and fall would be more gradual. In general, the energy pulses detected in the word would be smoothed.

3. Another research idea would be to use a filter bank front-end in the recognition analysis instead of the present Itakura-based LPC technique. This would allow many features of the prefilter to be incorporated directly into the recognition scheme. A much simpler prefilter and recognizer could be produced since much of the analysis would now overlap. For example, the method the prefilter uses in determining the signal-to-noise level in each filter by applying suppression curves is directly applicable to an endpoint detection process. The combined signal energy in all of the filters would be used as a basis for making an endpoint decision on that frame. Signal-to-noise frame weighting as described in (1) could also be easily implemented.

Another consideration is the new type of spectral matching for the distance measure that would be employed. It might be that this measure will be more robust in the presence of noise than the linear predictive analysis and Itakura distance metric.

ACKNOWLEDGEMENTS

I wish to thank my thesis advisor at M.I.T. Lincoln Laboratory, Dr. Robert J. McAulay, for his enthusiasm and creative suggestions throughout the course of this work and in the preparation of this report. I am also thankful to Dr. Clifford J. Weinstein, Group Leader of Speech Systems Technology at M.I.T. Lincoln Laboratory, for providing me with the necessary facilities and support to perform this research. I also wish to thank my academic thesis advisor, Professor Victor W. Zue, for his helpful suggestions and for providing the facilities to produce the documentation of this thesis.

Special thanks are due Joel A. Feldman, Joe Tierney, Marilyn L. Malpass, Francis Bonifanti, and the other members of the Speech Systems Technology Group at M.I.T. Lincoln Laboratory for their many helpful suggestions. I am also grateful to Sharon Kennedy and Linda Nessman for their dedication in producing the thesis proposal and soon-to-be-published paper. Special thanks are also due the Publications Division at M.I.T. Lincoln Laboratory for their help in producing the figures for this thesis.

Finally, thanks are also due Stephanie Seneff for her invaluable contribution of the recognizer software at the beginning of the project and Lori F. Lamel for her generous assistance with the endpoint detector.

REFERENCES

- [1] T.B. Martin, "Practical Applications of Voice Input to Machines," Automatic Speech and Speaker Recognition, N.R. Dixon and T.B. Martin (ed.) (New York: IEEE Press, 1979), p.174.
- [2] C.R. Coler, "Helicopter Speech-Command Systems: Recent Noise Tests Are Encouraging," Speech Technology, (September/October 1982), pp. 76-81.
- [3] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-28 (April 1980), pp. 137-145.
- [4] G. Neben, R.J. McAulay, and C.J. Weinstein, "Experiments in Isolated Word Recognition Using Noisy Speech," IEEE International Conference on Acoustics, Speech and Signal Processing, (April 1983).
- [5] G.R. Doddington and T.B. Schalk, "Speech Recognition: Turning Theory to Practice," IEEE Spectrum, Vol. 18 (September 1981), pp. 26-32.
- [6] F. Itakura "Minimum Prediction Residual Principle Applied to Speech Recognition," Automatic Speech and Speaker Recognition, N.R. Dixon and T.B. Martin (ed.) (New York: IEEE Press, 1979), pp. 145-150.
- [7] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-29 (August 1981), pp. 777-785.
- [8] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," The Bell System Technical Journal Vol. 54 (February 1975), pp. 297-315.
- [9] A. Wichiencharoen, "An Investigation for the Design of a Microcomputer Based Speech Recognition System." Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA., February 1981.
- [10] L.R. Rabiner, C.E. Schmidt, and B.S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech," The Bell System Technical Journal, Vol. 56 (March 1977), pp. 455-482.
- [11] R.J. McAulay, private correspondence.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|--|-----------------------|--|
| 1. REPORT NUMBER ESD-TR-83-007 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle) The Performance of an Isolated Word Recognizer Using Noisy Speech | | 5. TYPE OF REPORT & PERIOD COVERED Technical Report |
| | | 6. PERFORMING ORG. REPORT NUMBER Technical Report 647 |
| 7. AUTHOR(s) Gary Neben | | 8. CONTRACT OR GRANT NUMBER(s) F19628-80-C-0002 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173-0073 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element No. 33401F Project No. 7820 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20331 | | 12. REPORT DATE 13 April 1983 |
| | | 13. NUMBER OF PAGES 58 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Electronic Systems Division Hanscom AFB, MA 01731 | | 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 None | | |
| 19. KEY WORDS (Continue on reverse side if necessary and identify by block number) speech recognition word recognition isolated word recognition recognition and noise prefiltering noisy speech | | |
| 20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This report investigates the effects of noise on a speaker dependent, isolated word recognition system. Correct word recognition in a noise-free environment exists in a variety of present-day applications. However, when the acoustic environment includes noise, the problem of correct word recognition becomes more difficult. The noise interferes with the accurate location of the word boundaries and also distorts the spectral representation of the speech waveform. A series of experiments were performed to determine (1) the effects of using an energy-based endpoint detector and a conventional isolated word recognition system when the input speech is noisy and (2) the effects of placing a noise suppression prefilter in tandem with the word recognizer in an attempt to remove the noise prior to recognition. It was found that the system consisting of the prefilter working in tandem with the word recognizer increased word recognition accuracy. | | |

END

FILMED

6-83

DTIC