Report 3731

NAVAL SHIP RESEARCH AND DEVELOPMENT CENTER

Bethesda, Md. 20034

SLOPE-INTENSITY DETECTION SYSTEM FOR SPEECH RECOGNITION

Dr. S. Berkowitz

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

COMPUTATION AND MATHEMATICS DEPARTMENT
RESEARCH AND DEVELOPMENT REPORT

August 1971

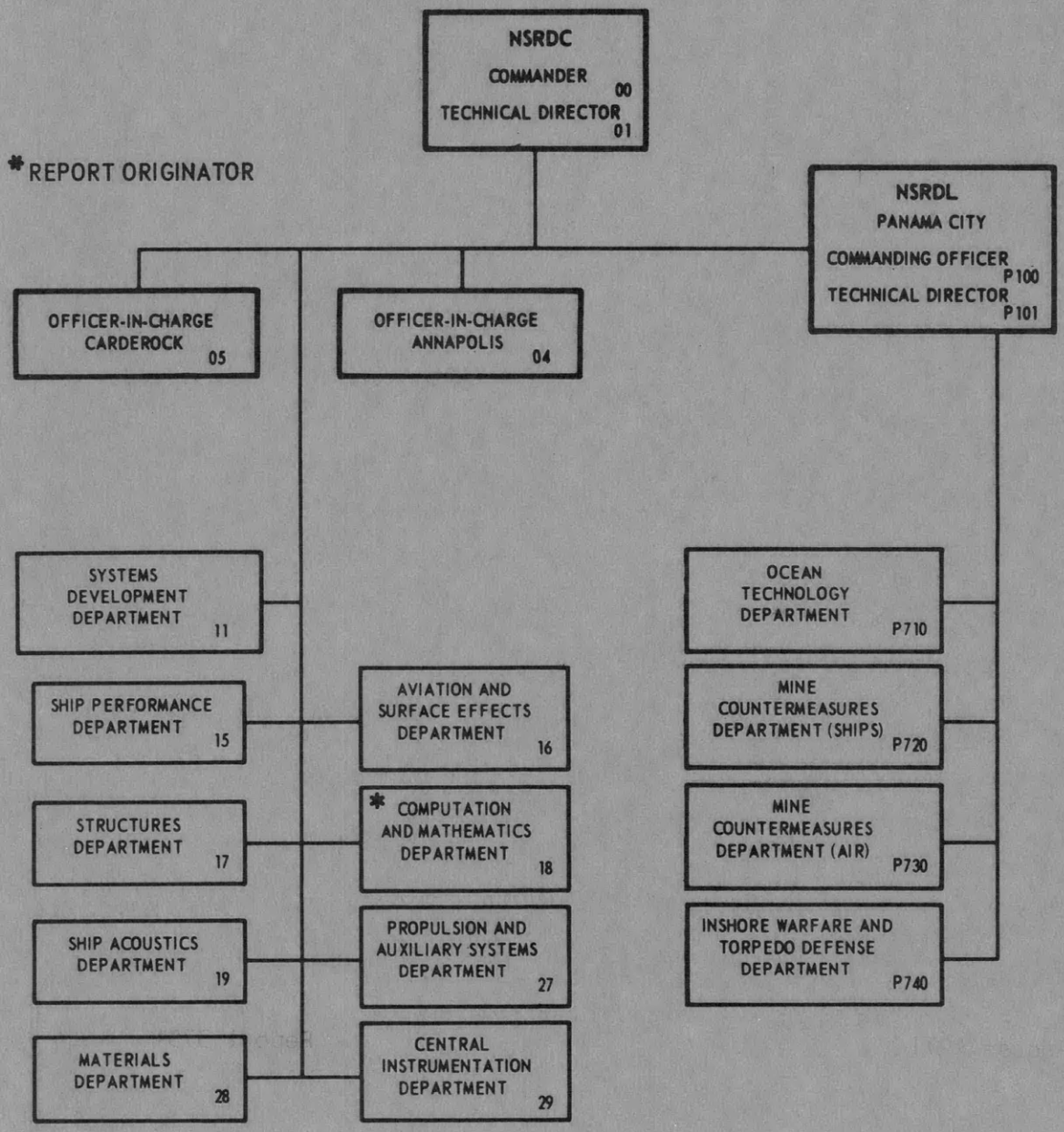
Report 3731

SLOPE-INTENSITY DETECTION SYSTEM FOR SPEECH RECOGNITION

The Naval Ship Research and Development Center is a U. S. Navy center for laboratory effort directed at achieving improved sea and air vehicles. It was formed in March 1967 by merging the David Taylor Model Basin at Carderock, Maryland with the Marine Engineering Laboratory at Annapolis, Maryland. The Mine Defense Laboratory (now Naval Ship R & D Laboratory) Panama City, Florida became part of the Center in November 1967.

Naval Ship Research and Development Center
Bethesda, Md. 20034

MAJOR NSRDC ORGANIZATIONAL COMPONENTS



DEPARTMENT OF THE NAVY

NAVAL SHIP RESEARCH AND DEVELOPMENT CENTER

Bethesda, Maryland 20034

SLOPE-INTENSITY DETECTION SYSTEM FOR SPEECH RECOGNITION

Dr. S. Berkowitz

August 1971

Report 3731

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

This report describes the design and operation of a class of speech recognition devices which

(1) extract slope and intensity information from a short-term spectral analysis of a spoken word (one of a fixed but arbitrary vocabulary),

and

(2) determine which vocabulary word has been spoken, by means of programmable decision logic using the thresholded slope intensity parameters just mentioned as inputs.

Application has been made for a patent concerning the class of devices described in this report.

ADMINISTRATIVE INFORMATION

The work described here was done in the Computer Sciences Division under Task Area SF14532107, Task 15329, Speech Recognition Project.

PURPOSE

The system described is designed to sense and recognize words from a finite, spoken vocabulary, and to display the recognized words.

This report describes the design and operation of a class of speech recognition devices which

(1) extract slope and intensity information from a short-term spectral analysis of a spoken word (one of a fixed but arbitrary vocabulary), and

(2) determine which vocabulary word has been spoken, by means of programmable decision logic using the thresholded slope intensity parameters just mentioned as inputs.

BACKGROUND

A human being produces vowels by exciting the structural resonances of his vocal tract with a periodic, strictly positive, sawtooth-like acoustic wave from his larynx. If the analyzing filters have a sufficiently fine resolution (a 1/3-octave filter-bank would suffice), the resonances can be exhibited as intensity peaks, or formants, in the 250-5000 Hz region of a short-term spectral analysis of the vowel waveform. Certain features of speech other than vowels can be detected regardless of the formant structure by the presence or absence of acoustical energy in certain portions of the spectrum. Thus, the unvoiced fricative /s/ can be characterized both by high-intensity noise energy across a band in the 4-5 kHz region of the spectrum and by silence in the lower frequency bands. The voiced fricative /z/ has characteristics similar to those of /s/ in the 4-5 kHz band, but also formant structure in the lower frequency bands. Other low-intensity

fricatives such as /f th/ are classifiable by their similarity to /z/, by their lower thresholds, and by their particular durations. Still other non-vowel sounds cannot be determined in isolation, but must be analyzed by the phonemic context in which they reside. For example, stop consonants such as /p k b t/ occurring in the middle of a word have characteristic durations of silence, characteristic post-silence intensity thresholds, and—most important—distinctive formant slopes leading into and/or out of the silent periods (when approached or succeeded by a phoneme-bearing formant structure), all depending on the particular phonemic context of the stop consonant.

Thus, there are several features that one can feasibly extract from the spoken word and use to identify elements of a fixed vocabulary.

- Silence
- Formant intensity and slope structure
- Noisy frequency bands
- Duration of sound or silence

In passing, we note that, even when the analyzing filters do not provide sufficient resolution to display a formant structure—either because of the overlap of the bands or because of overly large bandwidths, the slope of the filter outputs may offer enough invariance in context to permit the identification of certain non-vowel phonemes.

A CLASS OF SPEECH RECOGNITION DEVICES— DESCRIPTION AND OPERATION

Conventional speech-recognition feature-extraction techniques such as filtering and voice/unvoice detection have not heretofore taken advantage of the slope-intensity product. The block diagram on the next page represents a class of speech recognition devices which use slope and intensity characteristics of rectified, smoothed, band-pass filtered representations of a spoken word in the recognition of isolated words of a given vocabulary. The flow of information is as follows:

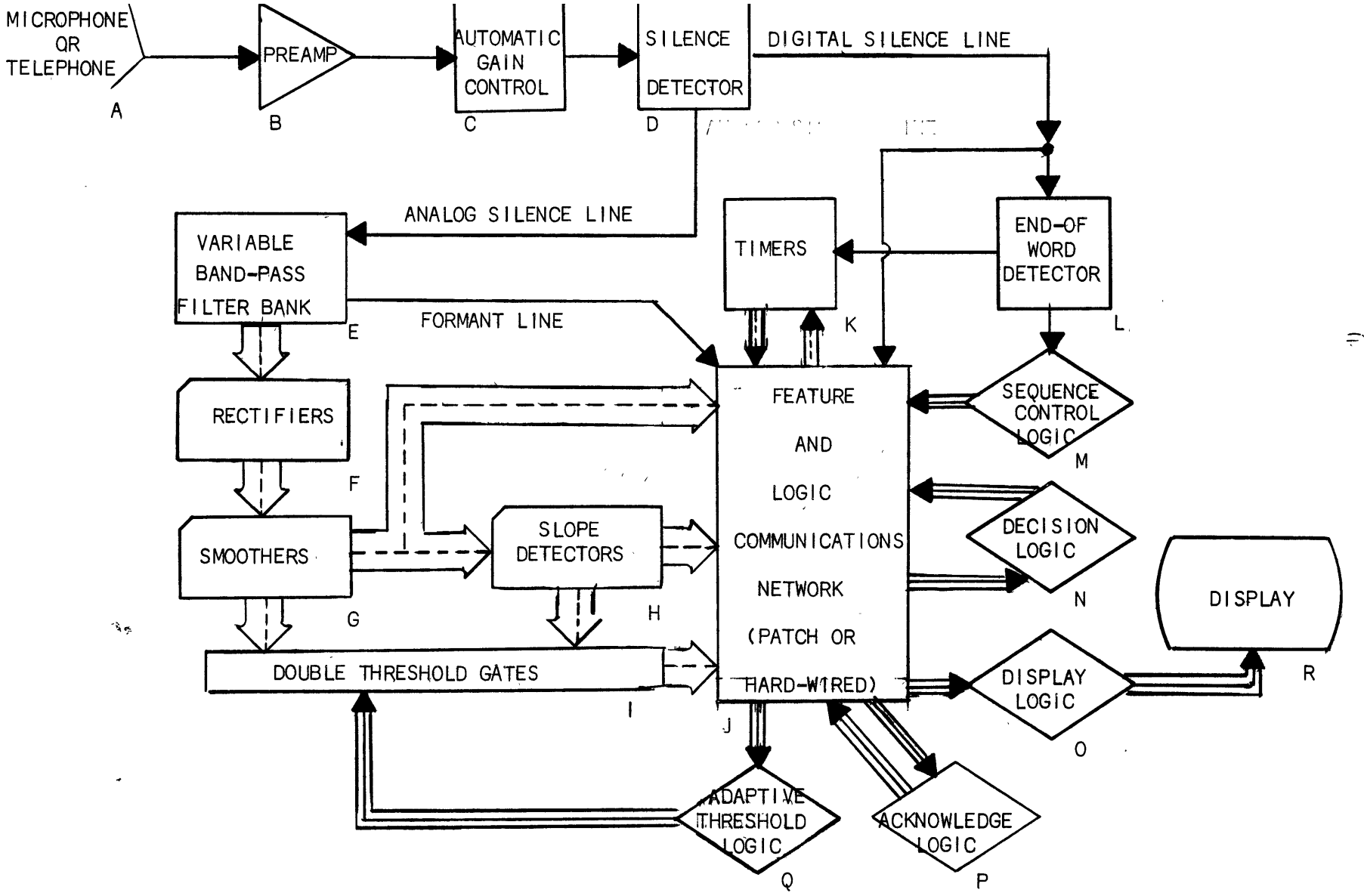


Figure 1 - Slope/Intensity Detection System for Speech Recognition

Acoustical waves from a person or tape recorder are transduced—by means of a microphone or a telephone, or some similar transducer, A—into a signal of electrical energy. This signal is then amplified, optionally with special emphasis on preselected regions of the frequency spectrum, by means of the preamplifier B. The preamplified signal then passes (optionally) through an automatic gain control C which maintains the average intensity at a constant level, thus normalizing the average voice intensity for all speakers. The normalized signal is then operated on by a silence detector D which produces two signals: (1) a digital (binary) signal on the digital silence line which is in one state when a silence is detected and in the opposite state when a non-silence is detected; and (2) an analog signal on the analog silence line which passes the normalized signal untransformed (except for an optional amplification or attenuation) during those intervals when silence is detected on the digital line. The silence that is "detected" by the silence detector D is actually any signal intensity below a non-zero signal level threshold, the magnitude of which is controllable externally by means of an adjustable potentiometer.

The analog silence line carries the signal to a bank of band-pass filters E whose bandwidths and center frequencies may (optionally) be externally adjustable. Alternatively, each filter of the bank E may self-adjust its center frequency within a particular frequency band in such a way as to automatically track the frequency of the input signal at its peak intensity. The tracked frequency—which, for a well-chosen band of frequencies, is a formant—may be communicated directly to the feature and logic communications network J, the function of which is to be explained later.

The bank of filters decomposes the signal into several channels, each of which holds (approximately) only that portion of the signal resident in a specific band of frequencies of the audio spectrum. The filter channel lines pass to a rectifier bank F where the signal amplitude in each channel is rendered positive, whether or not originally so. The rectified signals flow to a bank of smoothing circuits G which produce, for each input channel, a signal proportional to the root mean square of the input with low (less than 10 percent) ripple. The resulting set of filtered, rectified, smoothed signals is called a short-term power spectrum of the originally normalized signal. The spectrum serves as input to three portions of the system:

1. The bank of slope detectors H, which produces a signal that is the product of an approximate first derivative and the intensity of each channel of the short-term power spectrum;

2. The double bank of threshold gates I; each input channel enters two threshold gates, each of which produces a digital (binary) signal indicating whether or not the input signal has exceeded a threshold intensity which may be varied by means of an externally adjustable potentiometer; the two gates in effect detect three levels of intensities of a given channel;

3. The feature and logic communications network (FLCN) J, which is either a patch plug-board or hard-wired system of connections between features (formants, channel intensities and slopes, thresholded channel intensities and slopes, durations, and silence—the last two explained later) and the control circuitry (sequence, decision, display, achknowledge, and adaptive threshold logical circuitry, all to be explained later). As implied by the description of the FLCN J, the output of the bank of slope detectors H—the first derivative of the short-term power spectrum—enters not only its own double bank of threshold gates I but also the FLCN J; moreover, the outputs of all threshold gates are entered into the FLCN J.

The FJCN J interfaces several elements:

1. The timing circuits K, each of which accepts features as inputs and produces a digital (binary) indication as to the duration of the feature relative to an internal clock (i.e., whether slower or faster than a given unit of time);

2. The digital silence line, mentioned previously;

3. The sequence control logic circuitry M which sequences the decision, display, acknowledge, and threshold update functions so that first, a recognition decision is made in the decision logic N; second, a display reflecting the consequences of the decision is updated by the display logic O; third, a reception of the acknowledgement (an acceptance or rejection) of the decision is allowed in an optional mode of operation by the acknowledge logic P; fourth, the thresholds are adjusted in another optional mode of operation by the adaptive threshold logic Q; and last, all temporary storage is cleared in preparation for the next input word;

4. The decision logic circuitry N, which logically combines features of a digital character so that a decision as to the input spoken word is displayed. The decision logic may be adaptive and externally trainable;

5. The display logic circuitry O, which transforms the decision into an action (e.g., addition on a calculator) so that a display R of the consequences of the decision is effected;

6. The acknowledge logic circuitry P, which interprets a spoken word as acceptance or rejection of the decision, by means of a logical combination of features;

7. The adaptive threshold logic circuitry Q, which reflects a decision acknowledged with acceptance by adjusting the feature thresholds closer to the feature amplitudes received.

The digital silence line (1) activates an end-of-word detector L which is simply a monostable one-shot device that emits a pulse at a fixed time after detecting a sufficiently long silence; and (2) enters the FLCN J as a digital (binary) feature. The end-of-word detector L activates the sequence control logic M so that a decision is not fixed until a complete word is uttered.

The system can be tuned from the outside by

- Adjusting the amplification of the preamplifier B;
- Adjusting the rate of response and the normalized level of the automatic gain control C;
- Adjusting the silence threshold of the silence detector D;
- Adjusting the center frequencies and bandwidths of the filter bank E; overlapping the bands may introduce desirable redundancy;
- Adjusting the threshold gate levels in I;
- Adjusting the clock pulse widths in the timers K and changing input features to the timers in order to find a more invariant duration discrimination;
- Adjusting the delay of the end-of-word detector to allow more or less time for word segmentation;
- Adding redundancy logic (i.e., several means of recognizing one word) to the decision logic N to facilitate invariance in vocabulary discrimination among different speakers.

ADVANTAGES AND NEW FEATURES

The slope-intensity detection feature employed by the system described should enhance the accuracy of the machine recognition of spoken words by utilizing the information contained in the vowel-to-vowel, vowel-to-silence, or silence-to-vowel transitions, a technique not previously employed. Also, use of the double threshold gates I provides a means of dinstinguishing low-intensity channel information from high-intensity information, in effect forming a three-level gate.

ALTERNATIVE COMPONENT COMBINATIONS

Alternatives and options to the scheme shown in Figure 1 are presented here.

1. Any acoustic wave to electrical signal transducer may replace the microphone or telephone A.

2. The automatic gain control C is optional.

3. The band-pass filter bank E, rectifiers F and smoothers G, all may be replaced by any device which produces a short-term power spectrum; the formant line is optional and implies a replacement or addition to the band-pass filter bank E, which produces frequencies at which local high-intensity peaks occur as a function of time.

4. The double threshold gates I may be replaced by single gates if the low-intensity information is to be disregarded.

5. The timers K are an optional feature.

6. The acknowledge logic P is optional.

7. The adaptive threshold logic Q is optional.

8. The feature and logic communications network J may be a patch-wire board for manually interconnecting the features and the logic by wires, and even for interconnecting elementary logic elements to form arbitrary configurations of the complex logic elements M, N, O, P, Q and display R. Alternatively, the connections may be fixed by hard-wiring. Alternatively, the feature extraction elements, E, F, G, H, I, or K all or in part may be simulated on a digital computer and entered into the FLCN J; or, indeed, the logic elements M, N, O, P, Q and the FLCN J all or in part may be simulated on a digital computer. The system of elements E, F, G, H, I, J, K, L, M, N, O, P, Q, all or in part may be simulated on a digital computer.

AN EXAMPLE OF THE CLASS OF MACHINES DESCRIBED

A machine of the type just described, named PROFESR (PROgrammable Feature Extractor and Speech Recognizer), has been built by the author and J. Carlberg. This device has four variable filters but as yet no adaptive logic. We contemplate interfacing the device with an electronic calculator.

Application has been made for a patent concerning the class of devices described in this report.

INITIAL DISTRIBUTION

Copies	
1	NAVSHIPSYSCOM B. Orleans (0311)
2	ONR 1 Dr. R. Ryan (434) 1 M. Denicoff (437)
1	NRL Dr. P. Richards (7800)
1	CDR, NAVSHIPYD BSN
1	CDR, NAVSHIPYD CHASN
1	CDR, NAVSHIPYD HUNTERS PT
1	CDR, NAVSHIPYD LBEACH
1	CDR, NAVSHIPYD MARE ISLAND
1	CDR, NAVSHIPYD NORVA
1	CDR, NAVSHIPYD PEARL
1	CDR, NAVSHIPYD PHILA
1	CDR, NAVSHIPYD PTSMH
1	CDR, NAVSHIPYD BREM
12	DDC
2	AFOSR 1 CDR 1 Lt. Col. R. Ives
1	CDR RADC
1	Supt., U.S. Naval Academy Annapolis, Maryland
1	Supt., U.S. Naval Postgraduate School Monterey, California
1	CO. U.S. Naval ROTC and Administrative Unit, MIT Cambridge, Massachusetts
1	University of Louisville Prof. W. H. Pierce
1	Johns Hopkins University Prof. W. Huggins

CENTER DISTRIBUTION

Copies	Code
1	18
1	1808
1	1805
1	183
25	1834
1	184
1	185
1	1854 (H. Sheridan)
1	186
1	009

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Naval Ship Research and Development Center Bethesda, Maryland 20034	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

Slope-Intensity Detection System for Speech Recognition

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

Dr. S. Berkowitz

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 1971	14	

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO. SF 14 532 107	Report 3731
c. 15329 830-918	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

This report describes the design and operation of a class of speech recognition devices which

(1) extract slope and intensity information from a short-term spectral analysis of a spoken word (one of a fixed but arbitrary vocabulary), and

(2) determine which vocabulary word has been spoken, by means of programmable decision logic using the thresholded slope intensity parameters just mentioned as inputs.

Application has been made for patent concerning the class of devices described in this report.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
speech analysis pattern recognition formants spectral analysis						