RICE UNIVERSITY

Acoustic Segmentation of Speech

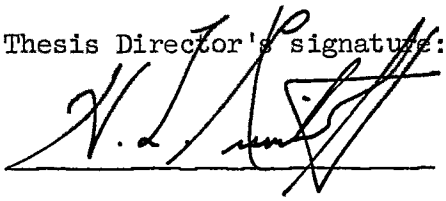
by

Gary Arthur Sitton

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
IN
ELECTRICAL ENGINEERING

Thesis Director's signature:



Houston, Texas

June 1969

Gary A. Sitton
"Acoustic Segmentation of Speech"

ABSTRACT

A brief history of speech research is given along with the current state of the art in acoustic speech recognition. The problem of speech segmentation in the acoustic domain using a digital computer is specifically addressed, i.e. determining an acoustic partition in time which has linguistic relevance. This problem is viewed, in more general terms, as that of detecting transitions, in a globally nonstationary process, from one local stationary state to another. Nonstationary analyses are approximated by considering short fixed length time series sections as seen through a window which moves by a fixed increment.

Various nonstationary signal representations are explored in order to establish a feature space suitable for segmentation applications. Spectral representations are only generated as a reference space used to compare any mechanical segmentation procedure with the linguistically determined segmentation of any given speech sample. Temporal representations of the zero crossings of speech signals are explored in detail. In particular the central sample moments of the reciprocal zero crossings as a function of time are used as input to a simple segmentation algorithm. The results of a demonstration of this algorithm show that speech segmentation as defined is possible by nonhuman means.

CONTENTS

FIGURES

TABLES

I INTRODUCTION

There have been three definite turning points in acoustic speech research history: (1) the invention of the vocoder by Homer Dudley in 1932, (2) development of the sound spectrograph by Bell Laboratories in 1948, and (3) utilization of the digital computer after 1960. It has been only in the last few years, however, that the digital computer has really been used effectively in this area. Prior to 1965, most of the research work was carried out on specially built devices or on analog computers. One of the ultimate goals of this research in the last thirty years has been that of automatic speech recognition, i.e., to build a device which accurately converts an acoustic speech signal in some given language into a sequence of symbols drawn from a finite inventory (such as the set of phonemes for that language). The complexity of the problem becomes clear when one considers that some of the nation's top research laboratories have to date failed to solve this problem; some experts believe it to be unsolvable.

The current state of the art can best be summed up as follows. For isolated single words from an unrestricted vocabulary spoken by a variety of male and female speakers, recognition accuracies of about 80% for vowels and 50% for consonants have been achieved [1,2,3]. Under the same conditions but using a limited specially selected vocabulary of 10-20 words, e.g. the digits, the words "yes", "no", "go", "stop", etc., top scores of about 98% are reported. The reasons for these failures appear to be twofold: dedication to the "black box" approach for research, and the practice of attempting recognition without segmentation.

The "black box" approach is that of deciding on some model or set of rules for recognition and then building an electromechanical analog of

that model. The folly of this procedure is that having spent months in designing and constructing such a "black box", any changes or alterations after the fait accompli are extremely difficult. It is impossible to build into such a device the generality necessary to effect the changes which are always required. The trend toward software simulation on a general purpose high speed digital computer has vastly improved model testing.

Segmentation is the process of partitioning (not necessarily disjointly) the speech stream into linguistically significant subunits which are hopefully atomic. It has been greatly disputed whether this is possible or even necessary. These primitive speech units are known as phonemes. Although one can with little or no training reliably write down the series of phonemic symbols which correspond one to one to a series of phonemes which were heard (see Table I), it does not follow that such a procedure actually occurs in the covert real time recognition process in humans. In practice the acoustic phonetic "boundaries" in speech are subtle and may give rise to an overlapped temporal partition. That is to say, the actual effect of a particular phoneme may begin during the first or second previous phonemes and/or terminate several phonemes later than its peak. Such supra-segmental phonemes are still, however, easily identified and localized by the human recognition process [4].

With very few exceptions, all past recognition research has ignored or avoided the problem of segmentation as a processing corequisite. Typical recognition devices had an output indicator which continuously showed the device's current decision as to which phoneme was currently being processed. Thus segmentation was achieved after the fact by

TABLE I

Part I--Conversational

Rank	Phoneme	Example	%		Type
1.	t	(take)	9.8		$\overline{v}$-p
2.	*n*	(not)	8.1		n
3.	I	(tip)	6.3		V
4.	r	(row)	6.1		l
5.	ʌ (up) / ə (herd)		6.0		V
6.	d	(dill)	4.6		v-p
7.	*l*	(like)	4.6		l
8.	s	(set)	4.0	(50%)	$\overline{v}$-f
9.	w	(win)	3.7		s
10.	m	(me)	3.6		n
11.	k	(kit)	3.6		$\overline{v}$-p
12.	ɛ	(ten)	2.7		V
13.	ð	(then)	2.5		v-f
14.	ai	(dike)	2.4		V(d)
15.	h	(hat)	2.2		$\overline{v}$-f
16.	z	(zip)	2.2		v-f
17.	a	(top)	2.1		V
18.	ae	(tap)	2.1		V
19.	j	(you)	2.1		s
20	i	(eve)	2.1	(80%)	V
21.	u	(boot)	2.0		V
22.	f	(for)	2.0		$\overline{v}$-f
23.	e	(mate)	1.9		V
24.	v	(vote)	1.8		v-f
25.	p	(pit)	1.7		$\overline{v}$-p
26.	o	(tone)	1.5		V
27.	g	(get)	1.5		v-p
28.	ɔ	(all)	1.3		V
29.	ŋ	(sing)	1.1		n
30.	U	(took)	1.0		V
31.	θ	(thin)	0.7		$\overline{v}$-f
32.	ʃ	(she)	0.7		$\overline{v}$-f
33.	b	(bit)	0.6		v-p
34.	au	(out)	0.6		V(d)
35.	dʒ	(jar)	0.3		v-p-f
36.	tʃ	(chew)	0.3		$\overline{v}$-p-f
37.	oi	(oil)	0.1		V(d)
38.	iu	(few)	0.1		V(d)
39.	ʒ	(azure)	0.01		v-f

Legend: V - vowel, (d) dipthong; v - voiced; $\overline{v}$ - not voiced; p - plosive; f - fricative; n - nasal; l - liquid; s - semivowel

Derived from Fletcher's data for edited telephone speech.

TABLE I

Part II - Written

Rank	Phoneme	Example	%		Type
1.	I	(tip)	7.9		V
2.	*n*	(not)	7.2		n
3.	t	(take)	7.1		$\overline{v}$-p
4.	r	(row)	6.9		l
5.	ʌ / ə	(up) / (herd)	5.0		V
6.	s	(set)	4.6		$\overline{v}$-f
7.	d	(dill)	4.3		v-p
8.	ae	(tap)	4.2		V
9.	i	(eve)	3.9	(50%)	V
10.	ℓ	(like)	3.7		l
11.	z	(zip)	3.6		v-f
12.	ɛ	(ten)	3.4		V
13.	ð	(then)	3.4		v-f
14.	a	(top)	3.3		V
15.	m	(me)	2.8		n
16.	k	(kit)	2.7		$\overline{v}$-p
17.	e	(mate)	2.4		V
18.	v	(vote)	2.3		v-f
19.	w	(win)	2.1	(80%)	s
20.	p	(pit)	2.1		$\overline{v}$-p
21.	h	(hat)	1.8		$\overline{v}$-f
22.	f	(for)	1.8		$\overline{v}$-f
23.	b	(bit)	1.8		v-p
24.	u	(boot)	1.6		V
25.	o	(tone)	1.6		V
26.	ai	(dike)	1.6		V(d)
27.	ŋ	(sing)	1.6		n
28.	ɔ	(all)	1.3		V
29.	g	(get)	0.8		v-p
30.	ʃ	(she)	0.8		$\overline{v}$-f
31.	U	(took)	0.7		V
32.	au	(out)	0.6		V(d)
33.	j	(you)	0.6		s
34.	dʒ	(jar)	0.5		v-p-f
35.	tʃ	(chew)	0.5		$\overline{v}$-p-f
36.	θ	(thin)	0.4		$\overline{v}$-p
37.	iu	(few)	0.3		V(d)
38.	oi	(oil)	0.1		V(d)
39.	ʒ	(azure)	0.05		v-f

Legend: V - vowel, (d) dipthong; v - voiced; $\overline{v}$ - not voiced; p - plosive; f - fricative; n - nasal; l - liquid; s - semivowel

Derived from canonical phonemic representations for alphabetic characters.

detecting a change in the output. Such an approach is expensive since recognition must be performed continuously. Pre-segmentation would appear to be easier because the recognition procedure needs only to be applied at several places within a phoneme's "boundaries" in order to identify that phoneme. It is also very possible that the recognition information could be applied to refine or correct the boundaries. In other words, the two processes should most likely be reciprocally interactive and iterative or self-correcting in nature.

II PURPOSE

Recognizing the theoretical and practical importance of demonstrating a mechanical segmentation procedure, the author has investigated several methods of speech signal representation which might be useful in attacking this problem. The goals set forth for this thesis are then to (1) develop a flexible and sophisticated software-hardware time series analysis system, (2) examine both spectral and nonspectral representations of the speech signal, (3) demonstrate the existence of physical "events" in speech time series that can be reasonably identified as phoneme "boundary" phenomena, and (4) design and test a naive segmentation algorithm on the basis of the work done.

The first goal was set forth in order to assure maximum flexibility and speed in implementation of the system [3].

Two basic techniques exist in acoustic speech signal analysis: spectral versus nonspectral (temporal). Spectral techniques consider transformations of the waveforms which result in functions of a frequency-like (reciprocal time) variable. Temporal techniques, however, involve transformations or manipulations directly in the time domain. In the past, spectral methods, e.g., Fourier analysis, have dominated speech research. It was felt that other areas could bear investigation, e.g., zero crossings (OX) analysis and nonorthogonal transformations.

A large number of signal representations will be investigated in order to define a feature space suited for segmentation of speech. This requires that significant "events" in some representation be highly correlated with the occurrence of classical heuristically determined linguistic phoneme boundaries or transition regions. It will thus also be necessary to generate representations of speech suitable for use by a

trained person in conjunction with the actual speech sounds in order to have this heuristic linguistic base for comparison purposes.

Finally, a crude segmentation algorithm is to be tested not to show the merit of any particular algorithm but simply to demonstrate in an objective mechanical fashion what will be fairly obvious by inspection. It is a regrettable human weakness to assume that a given human ability to easily recognize some class of patterns can always be as easily stated in a formal fashion in order to accomplish a nonhuman implementation. Since only a demonstration and not an iron-clad proof is intended, the algorithm will not be applied to a large amount of speech from a variety of speakers. This is even more reasonable when one considers the fact that the algorithm is to be arbitrarily chosen in the sense that it represents one person's (the author's) interpretation of how he would use a particular feature (segmentation) space to determine phoneme "boundaries". The algorithm will in any case supply an objective definition of "event" and "boundary" in speech signals or functions thereof.

III APPROACH

The most important consequence of using a digital computer for analysis is that only functions of discrete variables can easily be handled. After some preprocessing the continuous time series is digitized, and thus discretized in both amplitude and time, and then stored in the computer for processing later. Rather than use approximations to continuous analytic techniques, e.g., integration or differentiation, it was decided to use exact discrete methods, e.g., summation and differencing. This is perfectly acceptable for every continuous method which has a discrete equivalent [5]. Thus the only errors which are involved are the negligible round-off errors involved in computation.

In this work only well known standard mathematical and numerical techniques have been used. Heuristic or complicated nonlinear operations have been avoided where possible for the sake of simplicity, ease of analysis, and physical interpretation. Discrete techniques have been taken from standard time series, statistical, and orthogonal transformation methods.

The spectral class of analyses will consist of discrete Fourier transforms (DFT) and discrete Haar transforms (DHT) of the time series. As a result, two types of digital spectrograms are obtained. The DFT spectrogram is the digital counterpart of the multiple band pass filter techniques which have been used for years in speech analysis [6]. The Haar transforms, however, are new and were suggested in [7] as a means of analyzing transient phenomena, e.g., certain consonants in speech.

The temporal class of analyses will be made up of linear transformations of the zero crossing and reciprocal zero crossing distance distributions of small sections of the time series. Central sample

moments of these distributions will be examined as an example of non-orthogonal transformations. This approach has only been used in a very limited sense in the past [3,8]. Two different discrete orthogonal functions will be used to compute orthogonal transformations of the zero crossing distribution functions. These are the Gram polynomials and the Krawtchouk functions which will be described in detail later. This approach, to my knowledge, is completely new.

The most important consequence of working with nonstationary signals like speech is the necessity of using nonstationary techniques in the analysis of such signals. For reasons of simplicity, all of the time series were assumed to be locally stationary over a period of less than 15 msec. This is reasonable because the vocal excitation function, a smooth periodic sawtooth function, has about this period for the average person. An examination of the frequency spectrum as a function of time for speech shows that the generating mechanisms, i.e., the vocal cords, tongue, lips, etc., are essentially motionless over short durations of time. For these reasons, all statistics and functions generated from the speech data will always be dependent on time. This is necessary in order to examine any temporally evolving process like speech.

By considering only that portion of speech visible in a time window of fixed width, and moving that window by a fixed discrete increment to the next window position, a nonstationary analysis can be approximated. The functions or numbers that result from the analysis of each sequential window can be concatenated to form new functions which are discretely time dependent. In order to establish some continuity and dependence between adjacent windows, the window will usually be shifted by less than its own width, i.e., some fixed overlap

is included. To further stabilize these new time series, a small amount of digital low pass filtering is usually performed witn respect to time.

IV EQUIPMENT

There were two basic steps used in the analysis of the speech data: (1) the preprocessing step done by the external input equipment, and (2) the actual numerical processing and input/output done using the Rice University computer. The differentiation of these two phases of analysis is made to emphasize the minimal role that nondigital processing played in the time series analysis system that was used. All equipment used was chosen as a matter of convenience and any system with equivalent components should serve as well.

The preprocessing system used is shown schematically below.

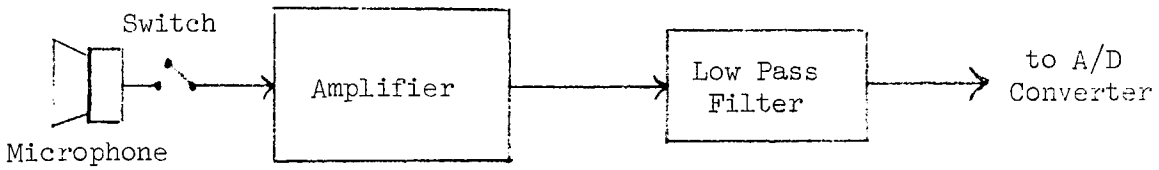


FIGURE #1

The microphone was an Electrovoice model 664, a highly directional dynamic cardioid type with 3 db cut-off points at 40 Hz and 10 kHz. The amplifier was constructed with FETs and had a voltage gain of about 30 db. The 3 db cut-off points for the amplifier were 50 Hz and 10 kHz; the low pass filter used was a second order cascaded RC with the 6 db point set at 4 kHz. The high frequency response curve for the amplifier-filter system is shown below.

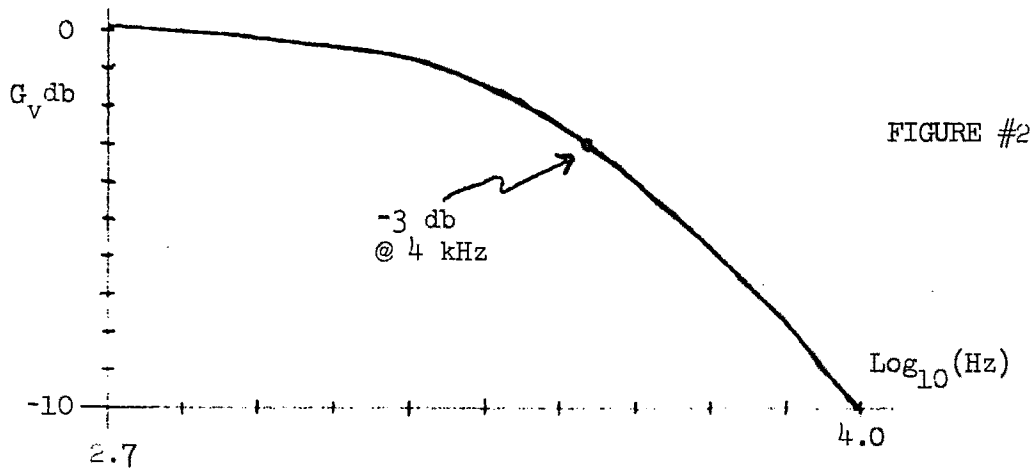


FIGURE #2

The average RMS signal to the computer was about 0.5 volts which was less than 2 volts peak to peak; this assures no clipping at ±1.28 volts which is the range of the A/D converter.

The computer system for A/D and D/A conversion, numerical processing and output is shown schematically below.

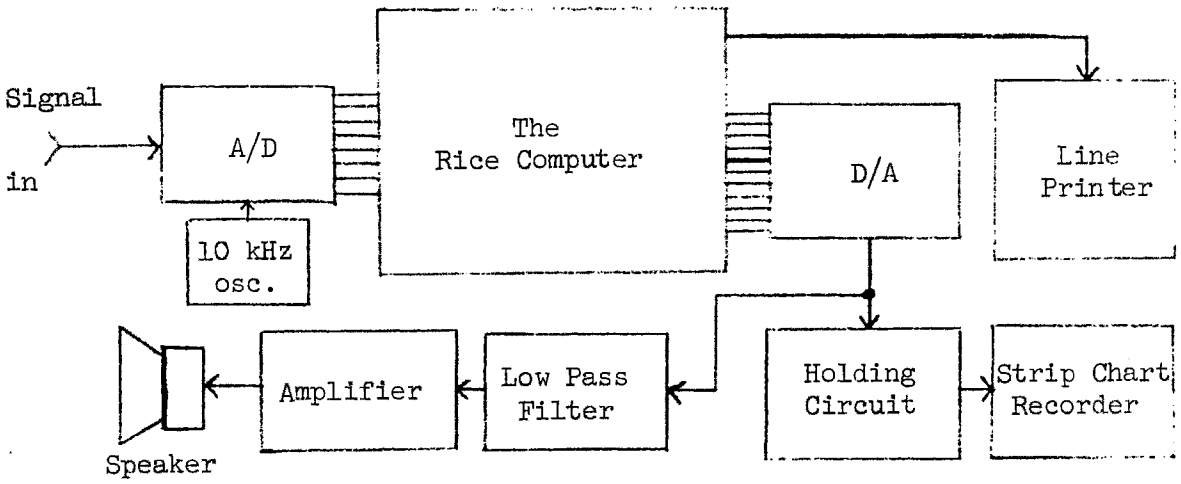


FIGURE #3

The variable sample rate A/D converter was set for a 10 kHz sample rate. This imposes a Nyquist frequency of 5 kHz which is larger than the bandwidth of the data after filtering. The conversion is essentially instantaneous, circa 4 μsec., and yields a sequence of 8 bit 1's complement numbers which are automatically packed 6/computer word and stored in core memory. The

Rice computer has a 24K core memory of 54 bit words and a rather slow arithmetic section. The flexibility and open shop availability of this machine far outweigh the minor speed problems in processing. A 600 line/minute line printer was used to produce all of the original plots. All of the input speech data was saved on punched paper tape and the signal quality was verified by using the D/A converter. The stored speech was either plotted at a lower rate on the strip chart recorder or listened to directly.

All of the programs for input, output, manipulation, and numerical processing were written by the author in either the assembly languages "AP1" and "AP2", or in the compiler language, "GENIE". Where appropriate, care and effort was taken to assure a high degree of efficiency for low running times. This is sometimes necessary when one considers the quantity of numbers to be processed in, say, one second of speech, e.g., at a 10 kHz sample rate, that is 10,000 data points!

IV DATA ACQUISITION

The corpus of words used as a data base for this work was obtained by selecting the most frequently used words which contain internal allophones of the most frequent phonemes. An allophone is one of the contextual acoustic variants of a given phoneme which is usually not consciously differentiated from any other variant of that phoneme. One does not perceive the difference in the vowel in "cat" as compared to that in "pack", but the linguistic context effects the acoustic properties of the vowel to some extent.

The ranked frequencies of phonemes in large samples of both written and spoken American English are given in Table I [9]. The spoken phoneme table was constructed from an examination of edited telephone conversations. Words peculiar to telephone speech, e.g., "hello", "goodbye", and profane words were not included in the tabulations. The written phoneme table was compiled from newspaper articles, written speeches, and novels. Both samples are considered to be statistically stable for this application.

The ranked frequency of words list used was that compiled at Brown University [10]. It is based on written English data but the differences (See Table I), at least as regards phonemes, are not important for this purpose. The twenty words selected for this work and their phonemic equivalents are shown in Table II. In addition, an auxiliary group of words was used in this research which was not based on frequency of occurrence. These are shown in Table III; the word "sunless" is of particular interest because it contains an acoustically subtle syllable break.

The speaker who supplied the speech input to the analysis system was the author, a native speaker of mid-western American English having

TABLE II

	Rank	Orthographic	Phonemic
1.	23	not	/nat/
2.	28	have	/haev/
3.	31	which*	/wIt∫/
4.	32	one	/wΛn/
5.	37	she	/∫i/
6.	42	him	/hIm/
7.	46	who	/hu/
8.	58	than	/ðaen/
9.	59	into*	/In'tu/
10.	62	only*	/on'li/
11.	63	other	/Λðər/
12.	66	some*	/sΛm/
13.	69	two	/tu/
14.	71	first*	/fərst/
15.	72	then	/ðIn/
16.	75	like	/laik/
17.	81	man	/maen/
18.	86	after	/aef'tər/
19.	89	did*	/dId/
20.	90	many*	/mɛ'ni/

* Words used to test segmentation algorithm

TABLE III

	Orthographic	Phonemic
1.	sunless*	/sʌn'lɛs/
2.	monday*	/mʌn'di/
3.	zero*	/zi'ro/
4.	speakers*	/spik'ərz/
5.	himself*	/hIm'sɛlf/
6.	speechless*	/spitʃ'lɛs/
7.	win	/win/
8.	ten	/tɛn/
9.	vote	/vot/
10.	jar	/dʒɔr/
11.	chew	/tʃu/
12.	thin	/θIn/
13.	see	/si/
14.	zoo	/zu/
15.	nation	/neʃan/
16.	vision	/vIʒan/

* Words used to test segmentation algorithm

a slight Texas urban accent. This speaker has had no formal speech training and made no conscious attempt to control the structure of the input speech to bias the results. All utterances were made with as little inflection as possible and at a normal conversational level. The speaker sat in a normal position with the microphone about three inches from his lips. An examination of some of the speech data containing the plosive consonants, (/k/, /p/, /t/, /b/, /d/, and /g/) showed no transient distortion in the waveforms due to puffing of the breath.

All words were uttered in the computer room, which contains noisy equipment, without the use of an acoustic shield. The signal-to-noise ratio averaged circa 25 db. This environment was selected because it was both convenient and realistic. It was desired that any methods discovered for segmentation or recognition be relatively insensitive to a reasonable ambient background noise level.

VI ANALYSIS

The data having been filtered, converted, packed, and stored, the numerical processing phase is entered. When a portion of the data is needed, it first must be unpacked, converted to a floating point form, and stored in a working vector. The data is always analysed by short "sections", each 50 to 200 sample points long. Each section represents that portion of the total signal presently being viewed through the moving time window. A schematic illustration of this moving window technique is shown below.

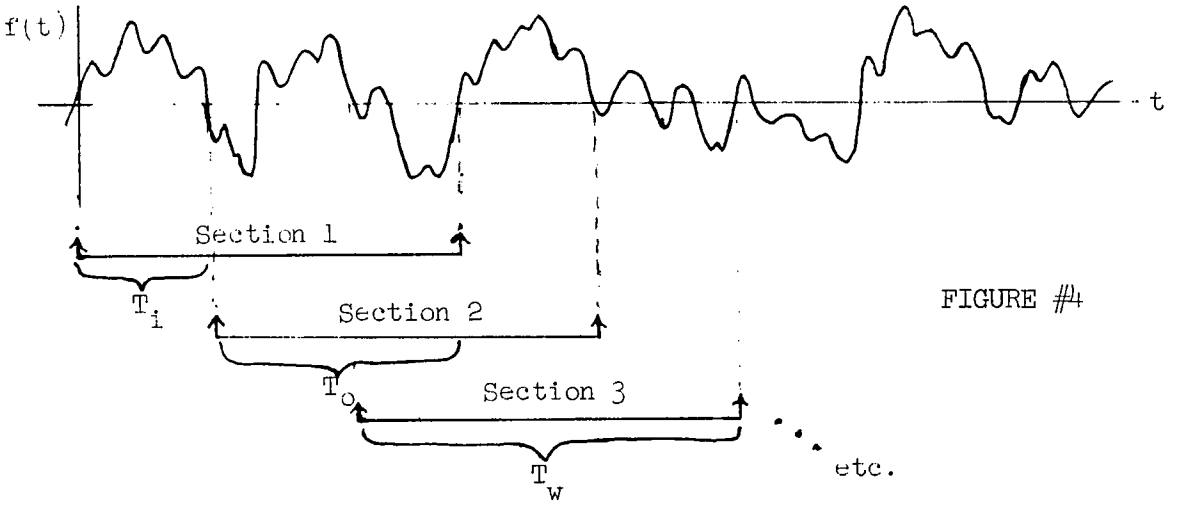


FIGURE #4

The time variable is shown as being continuous for the sake of clarity and simplicity. The section duration is T_w, the window increment is T_i, and the overlap between adjacent sections is $T_w - T_i = T_o$. Each of these time durations is an integer multiple of the sampling period 0.1 msec. No analysis "between the points" using interpolation was ever used since this would lead to unnecessary approximations and analytical complications.

One of the most vivid transformations of a speech signal is the time varying Fourier transform displayed as a variable density, two

dimensional plot, i.e., the sound spectrogram [11]. Using the time saving Cooley-Tukey DFT, one of the first digital sonographs was implemented. The basic methods used were taken from classical time series analysis and applied to nonstationary, discrete, signals.

The first step when using short term Fourier transforms, i.e., computing spectral estimates, is to remove any prominent uninteresting spectral components before analysis [12]. This is required due to the fact that the unavoidable convolution of the transform of the infinite time series and the transform of the finite time window, the sinc function

$$\text{sinc}(y) = \frac{\sin(y)}{y} \tag{1}$$

will smear any large spectral components and obliterate neighboring components which may be of interest. The moving window technique applied to a long time series whose mean is zero will give rise to a sequence of shorter time series whose means are generally not zero. Therefore, as a first step in computing the DFT of any section, its mean is computed and subtracted. A large mean will give rise to distortion of the first few coefficients of the DFT. We thus define

$$f_j(t_i)^* = f_j(t_i) - \frac{1}{N}\sum_{k=1}^{N} f_j(t_k) , \tag{2}$$

where j is the index of the section or window position, i or k is the index of the points in the section, and N is the number of points in a section.

The next step prior to the DFT itself was multiplication by an approximate time window taper [12]. The default taper is unity but its

transform (1) has large poorly convergent side lobes. A smoother window taper due to Arsac [13] was chosen, which is given by

$$w(x) = (1-x^2)^2, \qquad x \in [-1, 1], \tag{3}$$

and has the Fourier transform

$$W(y) = -\frac{\sin(y)}{y^3} - \frac{3\cos(y)}{y^4} + \frac{3\sin(y)}{y^5} . \tag{4}$$

The Arsac taper and its transform are compared to the sinc transform pair below.

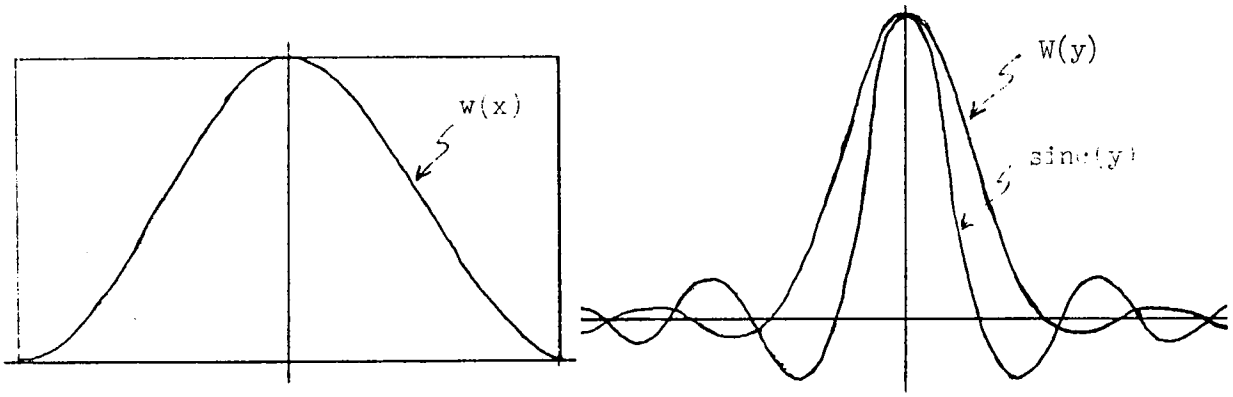


FIGURE #5

This window taper is much more suitable and its discrete equivalent, $w(x)$ sampled at equal intervals, was used. Thus we compute

$$f_j(t_i)^{**} = f_j^*(t_i)\cdot w(t_i) . \tag{5}$$

The DFT of (5) was next computed using the "power of two" Cooley-Tukey method [5,14]. The DFT is obtained by

$$A_{j,\nu} = \frac{1}{N}\sum_{k=0}^{N-1} f_j(t_k)e^{-i2\pi k\nu/N} , \tag{6}$$

$$\nu = 0,1,\ldots,N/2, \quad N = 2^m$$

where $A_{j,\nu}$ is the ν^{th} complex spectral coefficient for j^{th} time series section, and N is tne length of that section (128 points in

this work, i.e., m = 7 and thus T_w = 12.8 msec.). The coefficients for $\nu < 0$ were not computed since the DFT has conjugate symmetry for $f_j(t_i)^{**}$ real, i.e., $A_{j,\nu} = \overline{A_{j,\nu}}$.

In order to reduce the complex valued representation of the spectra to a more physically tractable form, only the moduli of the coefficients were used. Furthermore some form of amplitude compression was desired in order to observe the spectral detail of certain lower power consonants, e.g., the fricative consonant /s/. Therefore the final transformation was a form of logarithmic compression of the amplitude spectrum given by

$$a_{j,\nu} = \log(1 + c|A_{j,\nu}|), \tag{7}$$

where c is a constant which varies the compression effect. For $c|A_{j,\nu}| \ll 1$, essentially no compression is observed since

$$\log(1 + \epsilon) \approx \epsilon, \qquad \epsilon \ll 1 . \tag{8}$$

This transformation (7) was felt to be superior to a pure logarithm because the exaggerated diminuation of small amplitudes was avoided and the compression effect can be controlled.

$a_{j,\nu}$ is a real function of two discrete variables ν and j. The actual frequency range spanned by ν is 5 kHz and thus each value of ν corresponds to an increment of 78.125 Hz. T_i was chosen to be $T_w/2$ or 6.4 msec. which then corresponds to each increment of j. An overlap of one half the window width was chosen in order to obtain equally weighted statistical samples of the entire time series. This is true since the area of the Arsac taper relative to a unit window is given exactly by

$$\frac{1}{2}\int_{-1}^{1} (1-x^2)^2 dx = \frac{8}{15}, \tag{9}$$

which is $\approx 1/2$. The bandwidth of this spectral technique based on the half power point of the spectral window #5 given by (4) is circa 112 Hz. This represents a fairly "narrow band" analysis for speech.

In order to observe some measure of the total instantaneous (averaged over 12.8 msec.) perceived intensity of the speech signal as a function of time, the function

$$S_j = \frac{1}{N} \sum_{\nu=0}^{N} a_{j,\nu}, \qquad N = 64 . \tag{10}$$

was computed; this is itself a discrete time series. It is useful in determining the background noise level and thus gives a reference threshold for silence which can be used to eliminate ambiguities in the onset and termination times of speech sounds.

As an example of another spectral technique, a little used set of discrete ortnogonal functions was chosen for use in a transformation. These are the discrete Haar functions [15] described in Appendix I. The transformation was the standard linear form

$$\alpha^k_{j,n} = \sum_{i=1}^{N} f_j(t_i)\varphi^k_n(t_i) , \tag{11}$$

$$N = 2^m = 64 .$$

The restriction of requiring the time series section to be a power of two in length is a peculiarity of the discrete Haar basis and is unrelated to the same restrictions imposed by the particular DFT algorithm employed earlier. The unit window was used here because no convolution theorem exists for the Haar transform and no justification could be

found for using another window. Values of $T_w = T_i = 6.4$ msec. were chosen as a matter of convenience for this analysis.

Because of the strange nature of the Haar basis, e.g., the multiple indexing scheme required, representation and interpretation of the transformation is a serious problem. The index n is related to some exponential form of a frequency variable and the index k is related to a time position in the interval. Although the transient nature of the Haar basis might suggest an application to transient frequency analysis, it also has the ability to extract temporal information. As a final step the same form of logarithmic compression used on the DFT was applied to the DHT for the same reasons. Thus

$$\beta^k_{j,n} = \log(1 + c|\alpha^k_{j,n}|) \qquad (12)$$

was computed for plotting and visual examination.

Zero, or axis, crossing (OX) data has been used for many years in various schemes to track the first few formants (principle spectral peaks) of speech for recognition and compression applications [16,17]. These early investigations were no doubt prompted by the discovery and subsequent analysis of the statistical relationship between OX measurements of signals and their frequency spectra [18,19,20]. The use of OX data for speech compression and transmission has been shown both in theory and practice [21,22,23]. All of the approaches referenced above used the first moment of the OX distribution, i.e., the mean OX count. Most earlier work involved analog techniques to find OX rates but more recently digital computers have been used for this purpose [24,25].

Since speech is generated by a stochastic process, i.e., the speech signal is a function of a random variable as well as time, transformations

of the speech signal will be functions of a random variable. This fact suggested the use of standard statistical techniques to extract probabilistic information from such signals or functions of them. A standard tool used in the analysis of a discrete distribution function is the computation of its central sample moments, i.e.,

$$\mu_q = \frac{1}{M} \sum_{i=1}^{M} h(x_i)(x_i - \mu_1)^q, \qquad q \geq 2 \tag{13}$$

where $h(x_i)$ is the discrete distribution function. The definition of $n(x_i)$ includes the normalization

$$\frac{1}{M} \sum_{i=1}^{M} h(x_i) = 1 . \tag{14}$$

Here μ_1 is taken to be the first noncentral moment, i.e.,

$$\mu_1 = \frac{1}{M} \sum_{i=1}^{M} h(x_i)x_i . \tag{15}$$

Moment analyses of the OX distance and reciprocal OX distance distributions were carried out. As before, the time series section currently in the moving window was considered. In order to establish a meaningful zero reference line, the best straight line fitting the section in the least squares sense was used for that purpose as shown below.

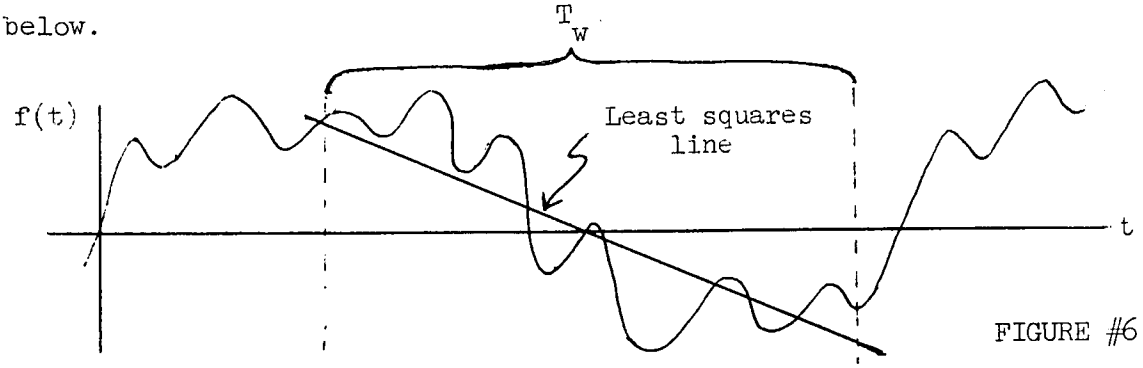


FIGURE #6

Linearly detrending in this manner has the effect of a mild high pass filter. The computation of the reference line is explained in Appendix II.

As a computational convenience the moments were computed directly from the sample data $\{d_i\}$ shown in #6 using the equations

$$\mu_{j,q} = \frac{1}{P_j} \sum_{i=1}^{P_j} (d_{j,i} - \bar{d}_j)^q , \tag{16}$$

and

$$\bar{d}_j = \mu_{j,1} = \frac{1}{P_j} \sum_{i=1}^{P_j} d_{j,i} \tag{17}$$

rather than computing the sample distribution $h(d_{j,i})$ beforehand and then using (13). The subscript j is as before the index of the time series section and the number of zeroes (P_j+1) in a section is a function of j.

The central moments for the Reciprocal OX distances were also computed by replacing $d_{j,i}$ in (16) and (17) with $d_{j,i}^{-1}$. The $d_{j,i}^{-1}$ have the dimensions of frequency and thus $h(d_{j,i}^{-1})$ is a type of crude frequency spectrum. In order to compare these two moment sets, $\{\mu_q(d_j)\}$ and $\{\mu_q(d_j^{-1})\} \triangleq \mu_q^-(d_j)$, the first group was replaced by $\{[\mu_q(d_j)]^{-1}\}$ purely for dimensional consistency. The relationship between these two different approaches, i.e., reciprocal arithmetic means versus harmonic means, is examined in detail in Appendix III.

Sample moments may be regarded as the coefficients of an expansion of a function of a discrete variable in terms of a nonorthogonal, but independent, power basis. A generalized expansion would be written as

$$\Phi_k = \frac{1}{M} \sum_{i=1}^{M} h(x_i)\varphi_k(x_i) \quad . \tag{18}$$

$$k = 0,1,\ldots,M-1$$

The relation of (13) to the above equations is seen by making the substitution

$$\varphi_k(x_i) = (x_i - \bar{x})^k \quad . \tag{19}$$

Expansions, using the linear transformation shown in (18), of the time dependent OX distributions $h(d_{j,i})$ were computed using two different discrete orthonormal bases. These two bases and discrete orthogonal functions in general are discussed in Appendix IV.

The first discrete basis to be examined was the Krawtchouk functions [26,27]. These functions are the discrete equivalent of the Hermites [28] which are related to derivatives of the normal Gaussian distribution. An expansion using these functions is then the discrete equivalent of the Gram-Charlier A series [29] . Series of this kind are useful in representing distributions which are approximately normal Gaussian. The testing of such a hypothesis is unfortunately complicated by the normal variable restriction which is implied for the distribution in question, i.e., one must set $\mu_1 = 0$, and $\mu_2 = 1$. The analytical complications involved in normalizing a discrete variable were felt to be beyond the scope of this present work and thus the normal Gaussian hypothesis was never tested.

The second discrete basis used was the Gram functions [28]. These functions have no continuous equivalent but strongly resemble Chebychev functions [28] for orders higher than two. (See Appendix IV.) The Gram functions are roughly sinusoidal and thus give approximately even

weighting over the interval in question, i.e., the domain of $g(d_{j,i})$. In contrast, the Krawtchouk functions are more centrally dense and are very small at the extremes of the domain for the lower order functions, i.e., $k \leq 10$. It was therefore felt that the Gram expansion might give a more economical series, i.e., require fewer terms to approximate any given function, than would the Krawtchouk. Examination of the OX distributions for speech showed them to be highly variant in form and thus more easily represented by a series in some basis like the Gram functions.

Since OX distributions are virtually independent of the amplitude of the signals from which they are derived, differentiation between silence (background noise) and the speech signal on a basis of these distributions is difficult. Therefore associated with an analysis of OX functions is the computation of time varying RMS values of the speech signal. This was found by computing

$$I_j = \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{f}_j(t_i)^2\right]^{1/2} \tag{20}$$

where $\tilde{f}_j(t_i)$ is the linearly detrended time series in the j^{th} section. This is simply the deviation of $\tilde{f}_j(t_i)$ given by

$$\sigma_j = \left[\frac{1}{N}\sum_{i=1}^{N} (\tilde{f}_j(t_i) - \bar{\tilde{f}}_j)^2\right]^{1/2} \tag{21}$$

since (as shown in Appendix II) $\bar{\tilde{f}}_j \equiv 0$. A characteristic function for the presence of speech can now be obtained by examination of I_j.

Most of the analyses just described involve the reduction of a very long time series to a few short time series, e.g., I_j. The reduction

of the number of points is rougnly by a factor of 100 depending on the window width and overlap. It was usually found that the new shorter time series had a small amount of high frequency noise superimposed on them. This is mainly an artifact of the approximations used for nonstationary analysis, i.e., the fixed increment and width moving window technique. In order to reduce this effect and smooth the short time series, a very mild low pass digital filter [12] was used. The filtering was accomplished by averaging sequential groups of s points and replacing the short time series by these moving averages by computing

$$\tilde{g}_j(t_i) = \frac{1}{s} \sum_{k=0}^{s-1} g_j(t_{i+k}) , \qquad (22)$$

$$i = 1,2,\ldots,N-s+1 .$$

Notice that filtering in tnis manner has the effect of shortening the time series by s-1 points and causing a time shift of (s-1)/2 points relative to the unfiltered function. A further explanation and analysis of this procedure is given in [12]. In the work reported herein, s was either two or three.

The exact effect of tne summing filter (22) can be found by computing the DFT of the filter function which is the unit discrete boxcar function of s points. We have then

$$\left| \hat{s}_k \right| = \frac{1}{N} \left| \sum_{j=0}^{s-1} e^{i2\pi kj/N} \right| = \frac{\sin[\pi sk/N]}{N \sin[\pi k/N]} \qquad (23)$$

which is the amplitude transfer function. For small k this function resembles the sinc function (1) since $\sin x \approx x$ for $|x| \ll 1$. This

filter function has zeroes for k = multiples of N/s but can be smoothed out if multiple filtering is performed with s chosen such that maxima of one filter are cancelled by the minima of another. The scheme for s = 3 followed by s = 2 gives such results. In all, tnis method of low pass filtering was found to be very efficient and useful.

VII DESCRIPTION OF RESULTS

It will be impossible to display all the results that were obtained in this work. The examples that are cited are typical and were chosen for that reason. Some of the less fruitful methods will be illustrated by one example while some methods warrant many diverse examples and extensive discussion. Due to the large amount of information contained in some plots, display problems have been serious. The plots have been reproduced and displayed within the limitations of the medium. Each type of display will be explained in detail. One particular word, "sunless," has many interesting properties and will be shown in many examples.

Figure #7 (foldout) shows a digital spectrogram of a typical utterance of the two syllable word "sunless" by the author. This plot strongly resembles the sound spectrograms of speech shown in [11], where a variable density is used to indicate intensity. The horizontal scale is time (marked in centi-seconds), and the vertical scale is frequency (in Hertz). The numbers shown (zeros are omitted for clarity) are the $a_{j,\nu}$ from (7), i.e., the compressed discrete spectral estimates. The time scale ranges from 0 to 1.5 seconds in increments of 6.4 msec. and the frequency scale ranges from 0 to 5 KHz in increments of 78.125 Hz. The rounded numbers are printed in the range from 1 to 9 on this type of plot although sometimes the upper limit is a, b, ..., f. This is due to the necessity of predicting $\max_{j,\nu}\{a_{j,\nu}\}$ from $\max_{j}\{I_j\}$, which is given by (20), for scaling the $a_{j,\nu}$ for plotting. This simply means that the maximum number printed is determined by an approximate method to avoid storing and scanning the entire time variant transform before plotting.

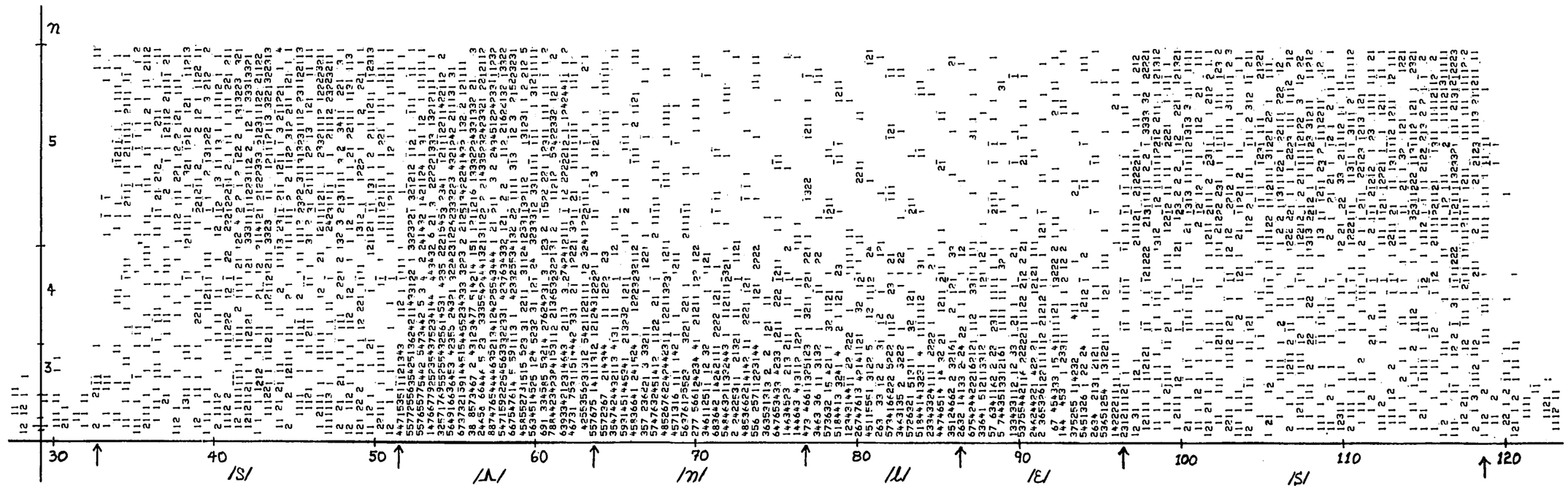


Figure # 9

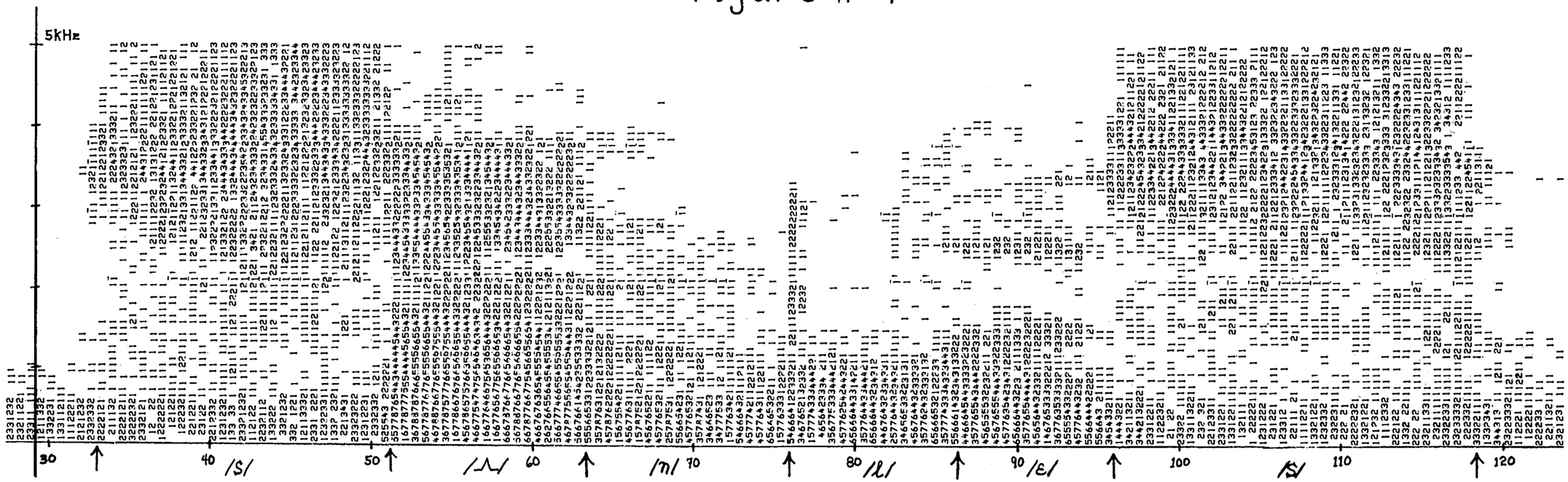


Figure # 7

The actual word begins at 33 on the time scale and ends at 119. This represents a total duration of ~ .86 seconds for this utterance of "sunless." The energy at lower frequencies (< 1 KHz) indicated before and after these points represent background noise. The phonemic representation of this word is /sʌnlɛs/. All of these symbols, drawn from the international phonetic alphabet, and examples of their pronunciation are to be found in Table I. Due to the known spectral structure of these phonemes, they can be easily located in this plot.

The initial /s/ is seen to start at 33 and terminate at 51. /s/ is typified by a relatively flat spectrum from ~ 2.5 KHz up to ~ 5 KHz. The final /s/ is also easily located from 96 to 119. The roll-off in the higher frequencies of the spectrum is due to the aliasing filter. The /s/ - /ʌ/ boundary is very clear in this plot as is the /ʌ/ - /n/ transition. Careful examination of the /ʌ/ region, 51 to 63, shows several spectral ridges called "formants" which are typical of vowels. The /n/ - /l/ boundary can be provisionally located by noticing the sudden population of slightly higher frequencies and the transient event both occurring at ~ 76. This phenomenon will be examined later in greater detail and identified with the syllable boundary which occurs there. The nasal consonants /n/ and /m/ have very weak formants and a strong voice band in the lower frequencies. The second vowel /ɛ/ in this example is betrayed by the reappearance of formants at 87. The formants of /ɛ/ are weaker and displaced relative to those in the /ʌ/.

The total (compressed) spectral intensity given by (10) is plotted in Figure #8. The time span and increment is identical for this function except for a slight effect due to digital smoothing. This function was filtered by running sums of 3 points followed by running

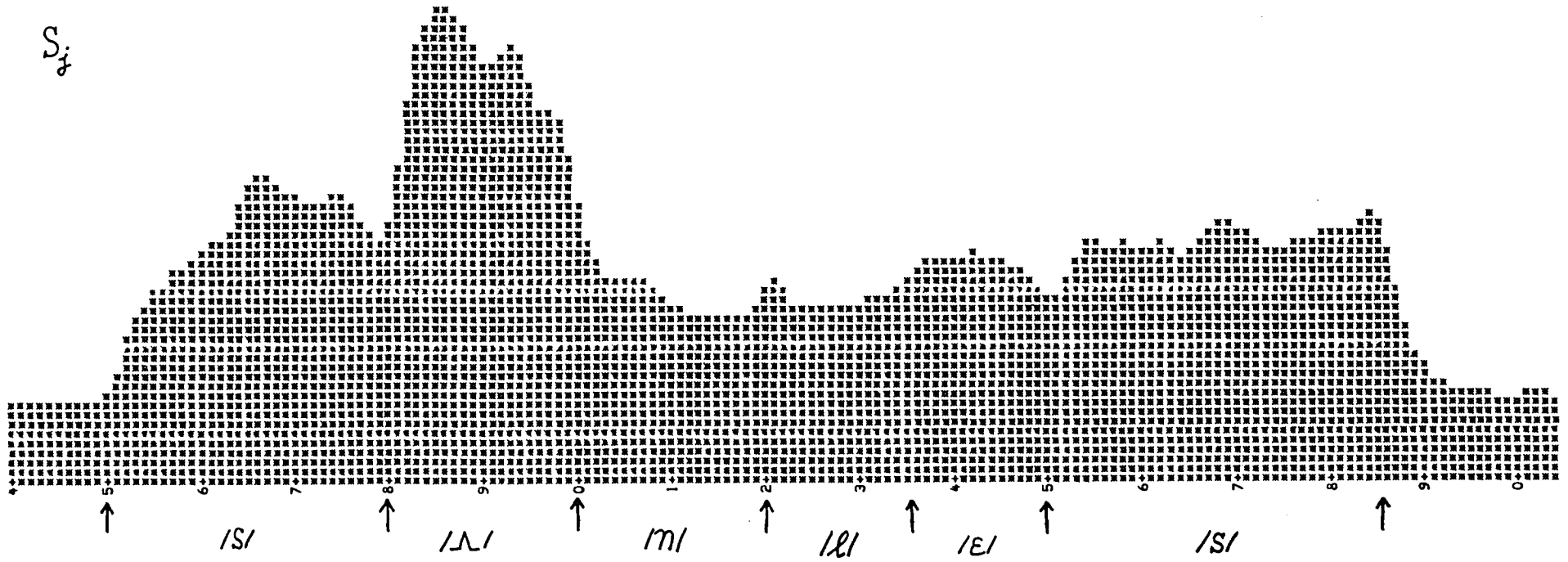


Figure #8

sums of 2 points. The result of this process is a shortening of the time series S_j by 3 points and a time shift of 1-1/2 points. Taking this into account, the plot has been marked at the same points as the spectrogram. Notice that each marked point corresponds to a region of change in perceived intensity. The "loudest" part of the word occurs during the vowel /ʌ/. This is because vowels are typically more intense than consonants and also /ʌ/ occurs during the accented syllable. It is obvious that even knowing the word beforehand, one could hardly have located the phoneme boundaries accurately using this intensity plot alone.

Figure #9 shows a discrete Haar spectrogram for the same word. This represents the compressed coefficients given by (12). In both the DFT and DHT spectrograms, the compression constant c was chosen by trial and error to show the greatest detail in the plots. The horizontal scale (time) is the same for both spectrograms. The vertical scale, however, is not simply related to frequency as before but is a linearized form of the multiple indices k and n. The index k was allowed to run over its range for each value of n. The boundaries for each value of $n = 0,1,\ldots,5$ are shown on the vertical scale.

The DHT spectrogram and the DFT spectrogram bear a strong resemblance to one another. Some of the boundaries, e.g., /s/-/ʌ/, /ʌ/-/n/, and /ɛ/-/s/, can readily be seen. The effect of changing from one value of n, e.g., 5 to 4 to 3, can be noticed in the region 60 to 100 on the time scale. The formant structure in /ʌ/ can be seen here as rising bands rather than horizontal bands as before. One can roughly associate n with $\log_2$ of the frequency and k as the time shift in the section. This strange frequency and time dependence of the DHT coefficients help to explain the appearance of the spectrogram.

The actual time series for "sunless" is shown in Figure #10 as plotted on a rectilinear hot pen recorder receiving its input from the D/A converter of the computer. The time scale here is 20 msec/large division (1 cm on the original plot). Some of the phoneme boundaries are obvious from this plot, e.g., /s/-/ʌ/, /ʌ/-/n/, and /ɛ/-/s/. It would, however, have been impossible to apply any heuristic segmentation procedure to this function as was done using the DFT spectrogram. The raw time series shows the complex nature of the signals under investigation here.

The RMS as a function of time for the linearly detrended signal given by (20) is shown in Figure #11. The time scale for this function is 10 msec/point which is also the case for all of the OX functions to be discussed. This function tends to show absolute intensities rather than the perceived (logarithmically compressed) intensities shown in Figure #8. The RMS function was low passed filtered with a minimal "sum by 2" filter resulting in a time shift of $\sim$ 1 point. Relative intensities of the various phonemes in "sunless" can readily be seen in Figure #11.

The remaining functions to be discussed are concerned with the short time OX distributions of the speech waveform. Figure #12 shows the time varying OX histogram for the word "sunless." The horizontal scale is time, 10 msec/point, and the vertical scale is distance between zeros in each detrended section. The actual frequencies are scaled between 0 and f (interstitial hexadecimal) and rounded. As in the spectrograms, the 0's are omitted for clarity. Small OX distances correspond to high frequencies since time is the reciprocal of frequency. Each histogram, i.e., one column of numbers, is normalized according to (14). The

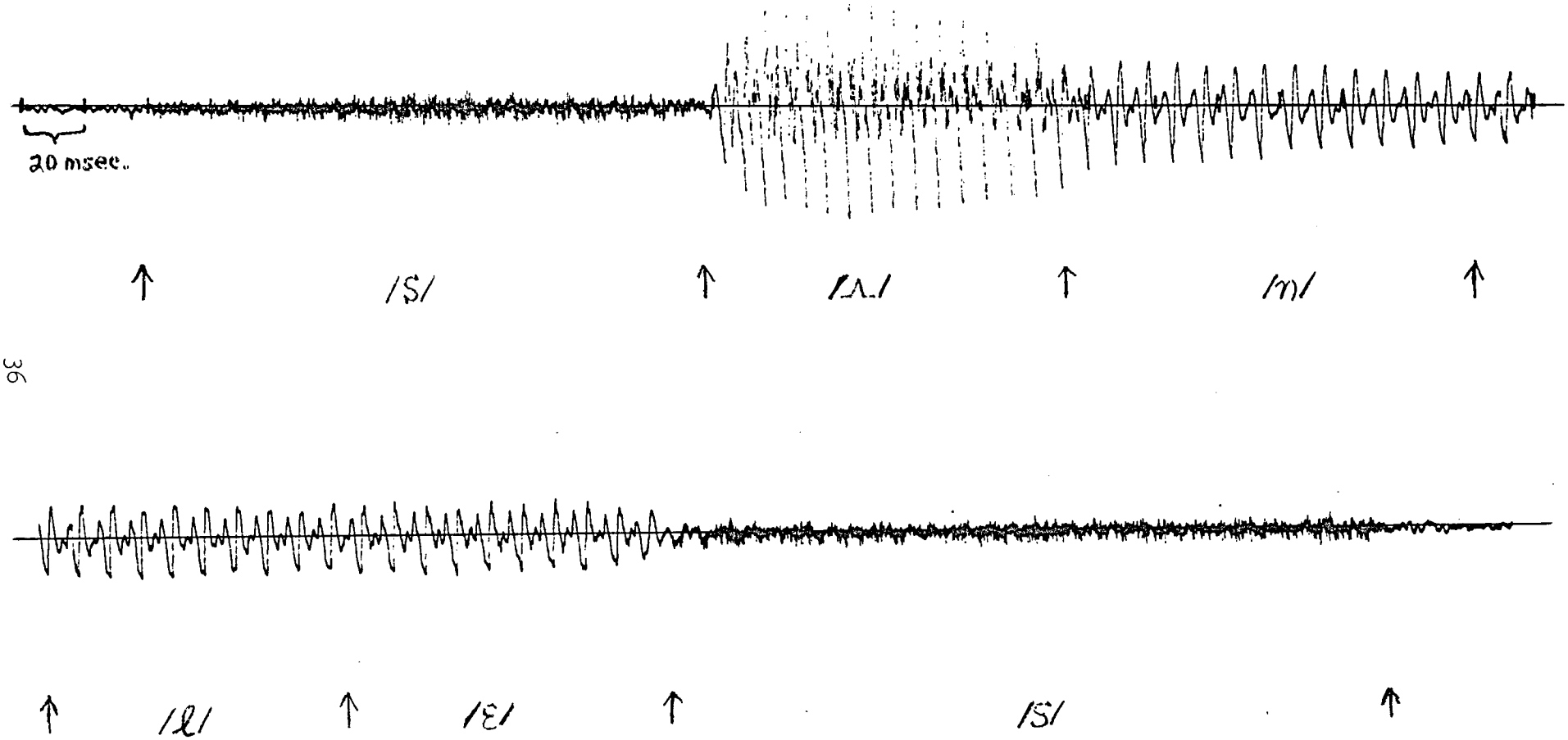


Figure # 10

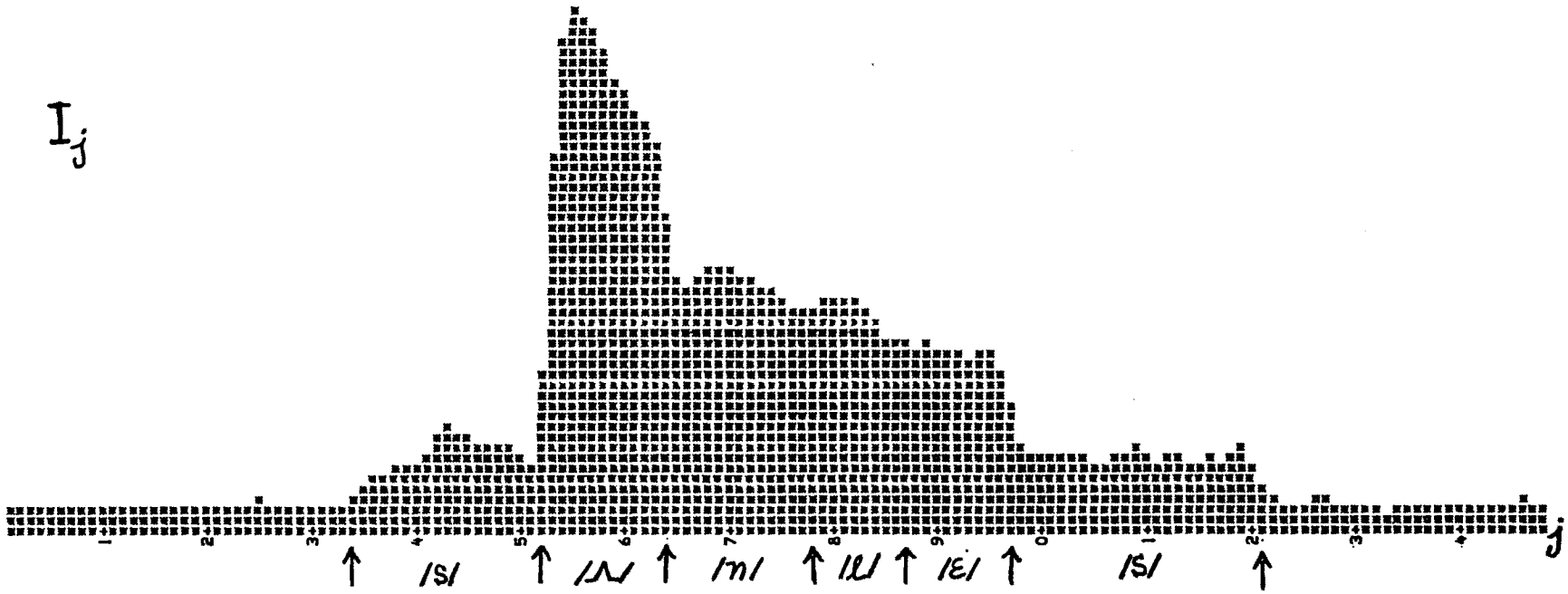


Figure # 11

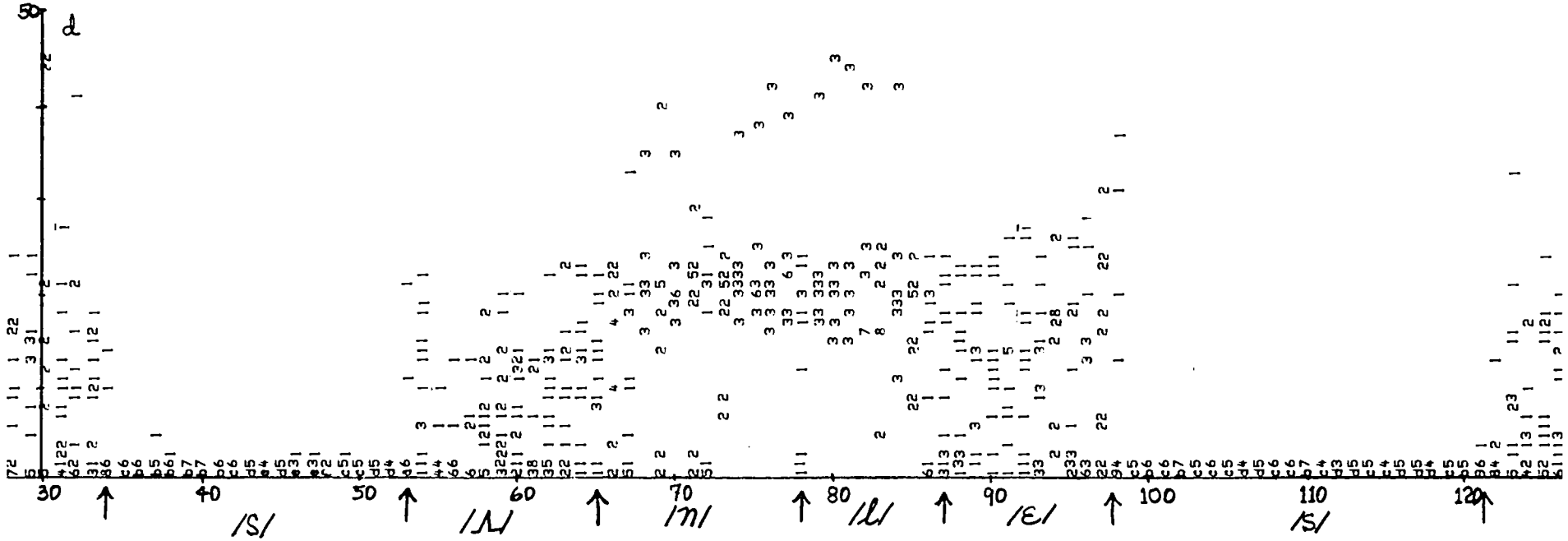


Figure #12

distributions have an upper limit of 51 points between zeros, which is $\sim 1/3$ of the window length of 150. This limit was determined empirically from a large amount of data and results from the low frequency rejection of the amplifier system and the detrending process. The range of d, [1, 51], is roughly equivalent to a frequency range of [5 KHz, 100 Hz] [18].

The first transformation of the OX distributions we will consider is $[\mu_{j,q}]^{-1}$ from (16) and (17), displayed in Figure #13 for q = 1,2. Higher moments, q > 2, are not shown due to their very low information content and poor quality. $[\mu_1]^{-1}$ is proportional to the mean frequency, i.e., half the number of OX/second of the signal [18]. The initial and final /s/ phonemes stand out very noticeably. These functions and the $\bar{\mu}_{j,q} \triangleq \mu_q(d_j^{-1})$ in Figures #14-16 have been filtered by summing by 2. $\bar{\mu}_{j,q}$ for q = 1,2 is shown in Figure #14, for q = 3,4 in figure #15, and for q = 5 and $\bar{\mu}_3/(\bar{\mu}_2)^{3/2}$ in Figure #16. The last function was computed to determine the effect of normalizing the third moment $\bar{\mu}_3$ by the variance $\bar{\mu}_2$. These moment functions will be discussed in more detail later.

It is appropriate here to point out the similarity between $\bar{\mu}_1$ and $[\mu_1]^{-1}$ in Figures #14 and #13 respectively. It is apparent that $\bar{\mu}_1$ is more sensitive to signal variations, e.g., phoneme transitions, than is $[\mu_1]^{-1}$ which has been used exclusively in speech research in the past. This difference is easily explained in Appendix III where the approximate relation

$$\bar{\mu}_1 \simeq [\mu_1]^{-1}\left[1 + \frac{\mu_2}{\mu_1^2}\right] \qquad (24)$$

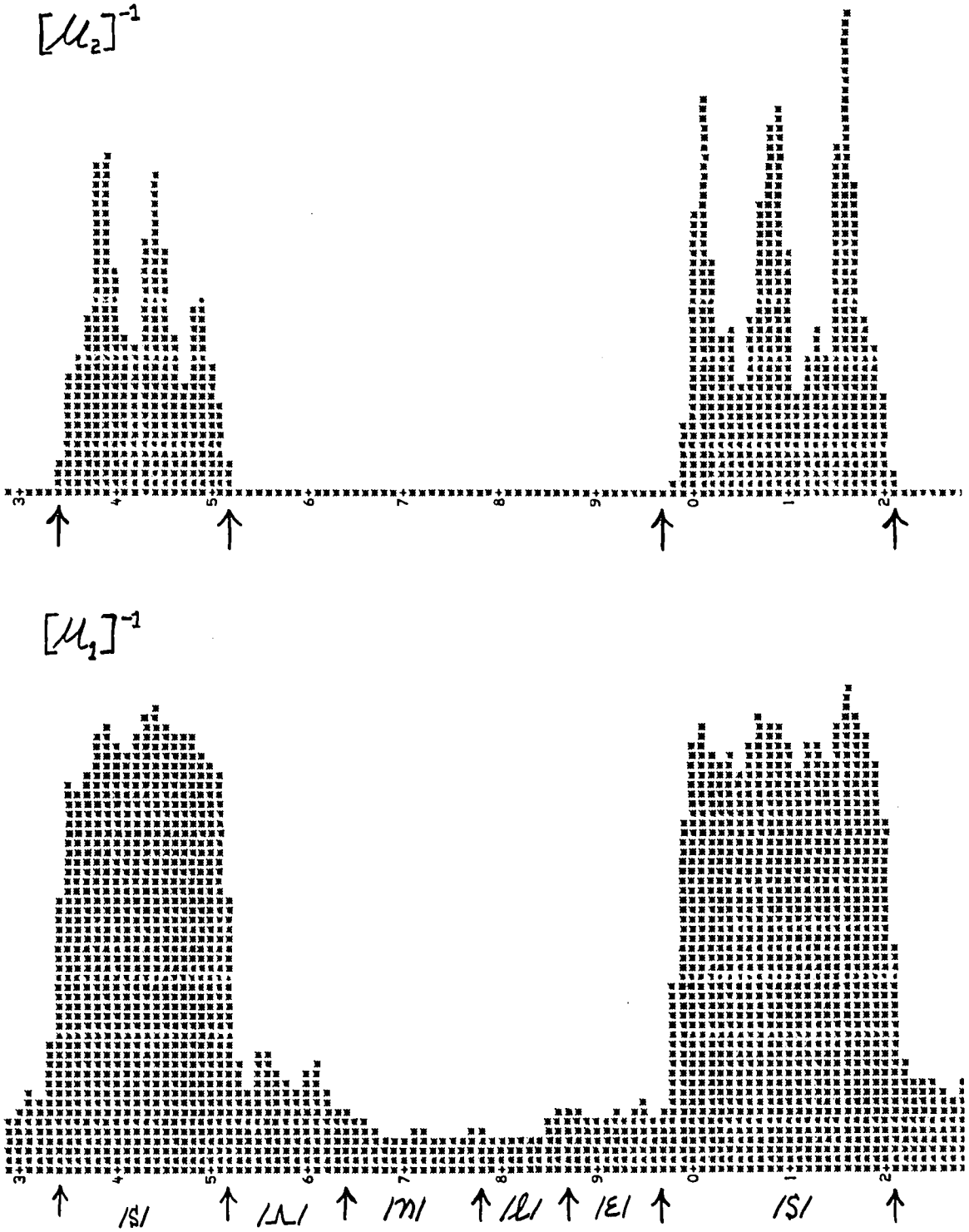


Figure #13

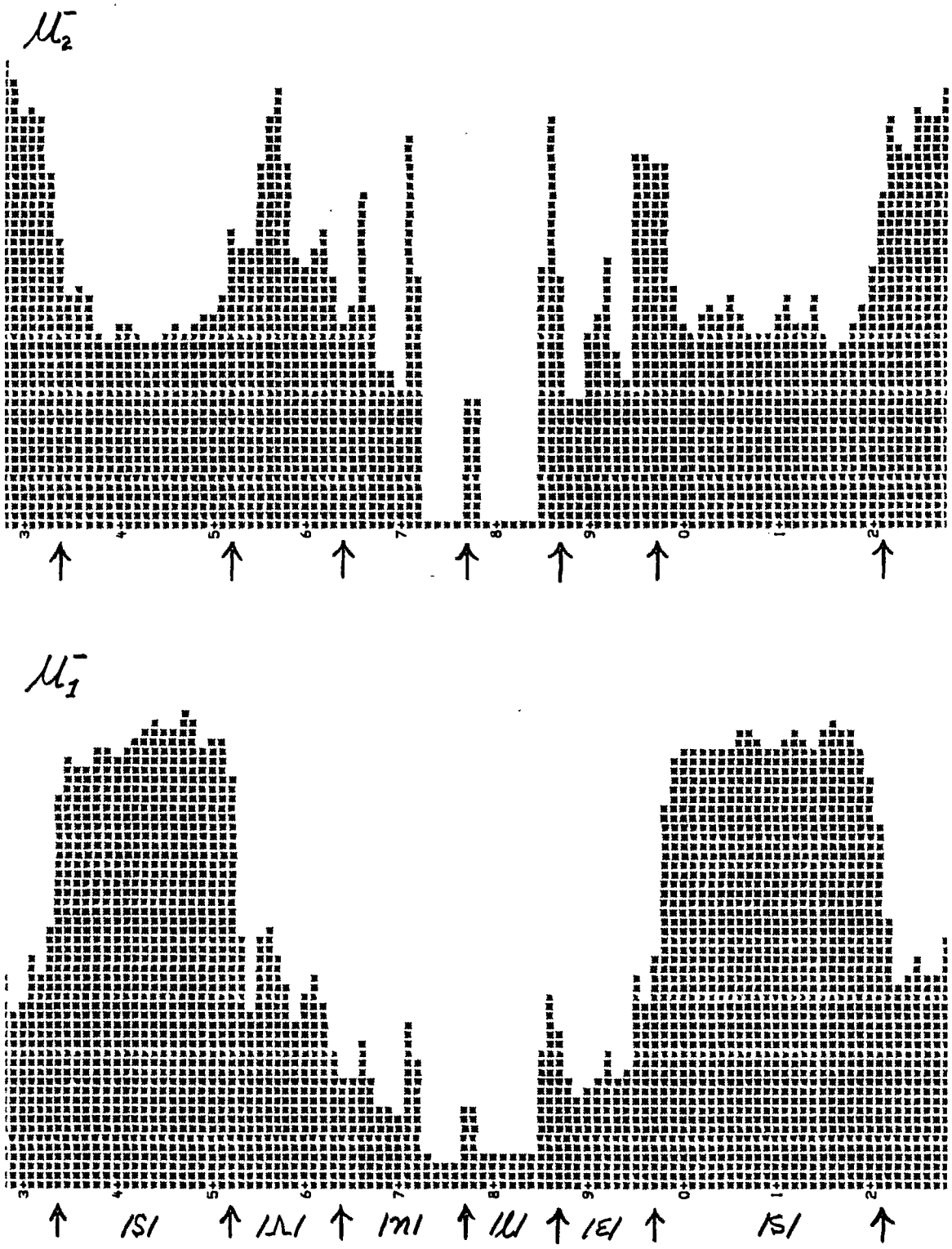


Figure #14

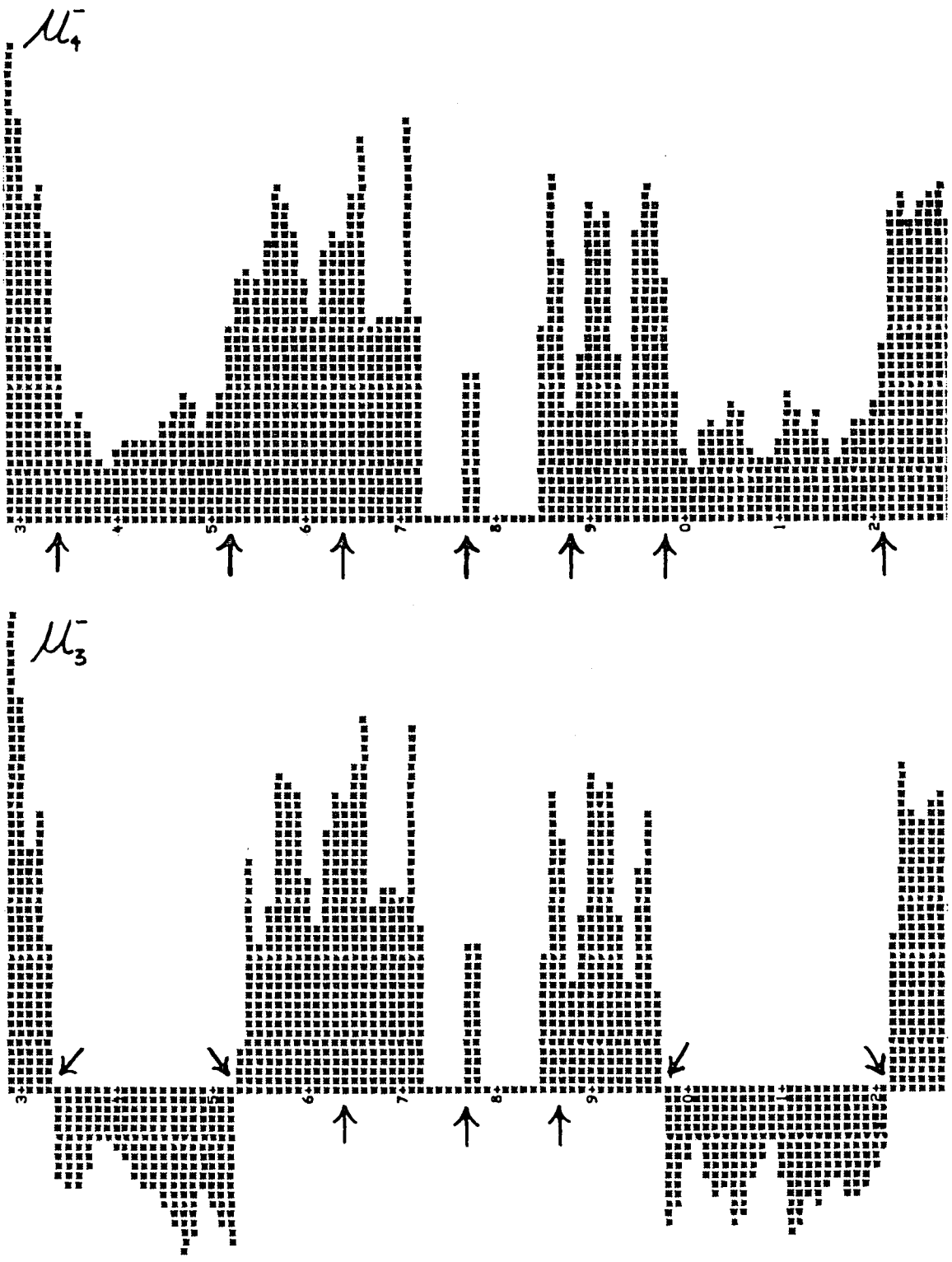


Figure # 15

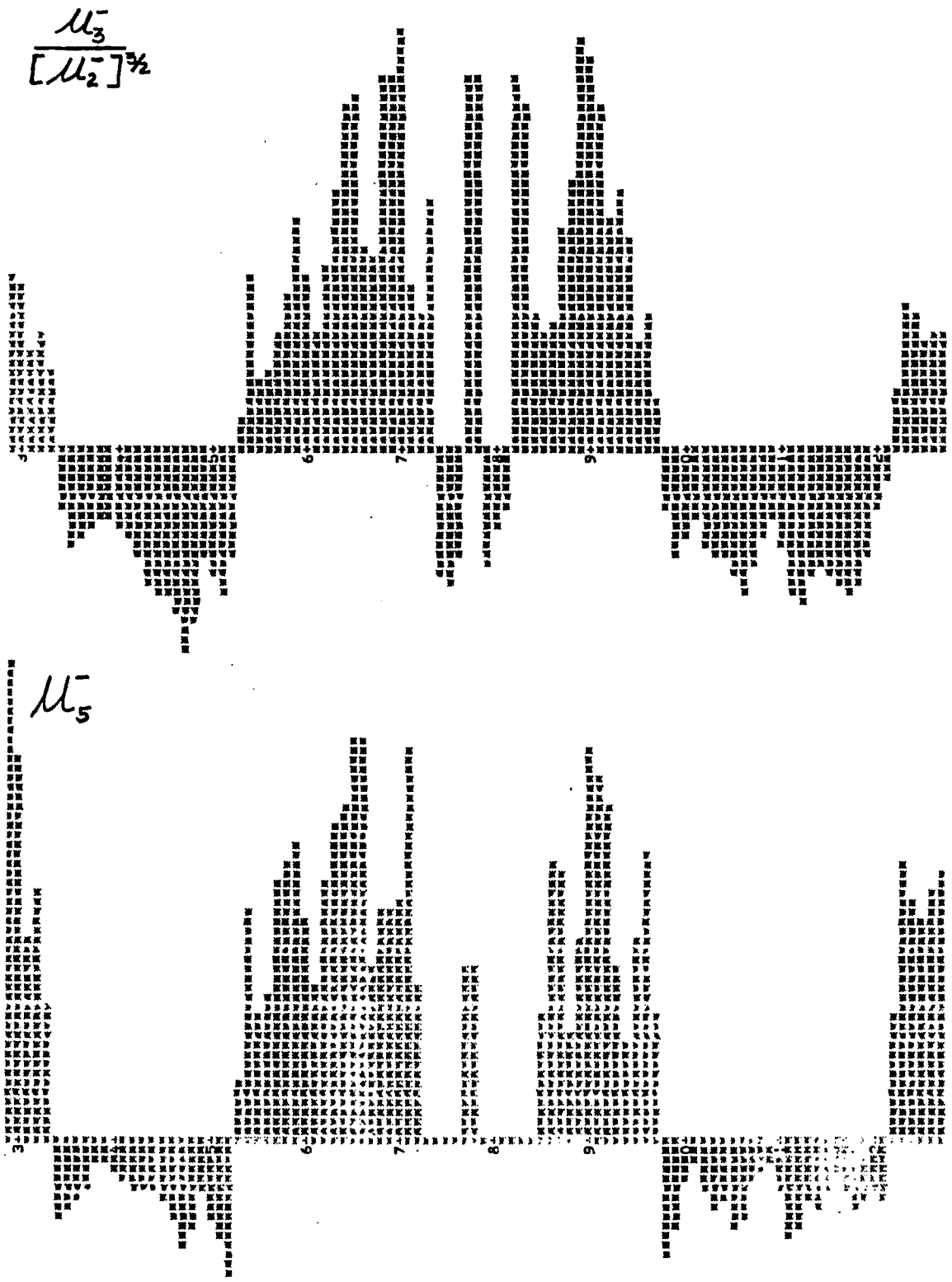


Figure # 16

is derived. This simply means that μ_1^- is a function of both the OX count and the variance of the OX distances. This approximation is only valid for $\mu_1 > \Delta d_i$ which was later shown not to hold for all portions of the speech signal, e.g., the /s/ phoneme where μ_1 is small. In any case the results in Appendix III are useful at least qualitatively in explaining the behavior of these new moments.

Figures #17 and #18 show the time varying Gram transform coefficients g_1, g_2, g_3, and g_4 from (18) for the normalized OX distributions by using $\varphi_i(x) = G_i(x)$ and thus $\Phi_i = g_i$. The same distributions were expressed in terms of Krawtchouk functions, $\varphi_i(x) = \mathcal{K}_i(x)$, and their coefficients as a function of time k_1, k_2, k_3 and k_4 are shown in Figures #19 and #20. As before, these functions have been slightly smoothed and were derived from the same utterance of "sunless." Both occurrences of /s/ are very obvious in these figures for both transformations. In both cases the 0^{th} coefficient is not shown because g_0 is just the mean OX distance, μ_1, since $G_0 = 1$, and k_0 gave very poor results due to the concentration of K_0 at the center of the domain of OX distributions, d_i (see Appendix IV.)

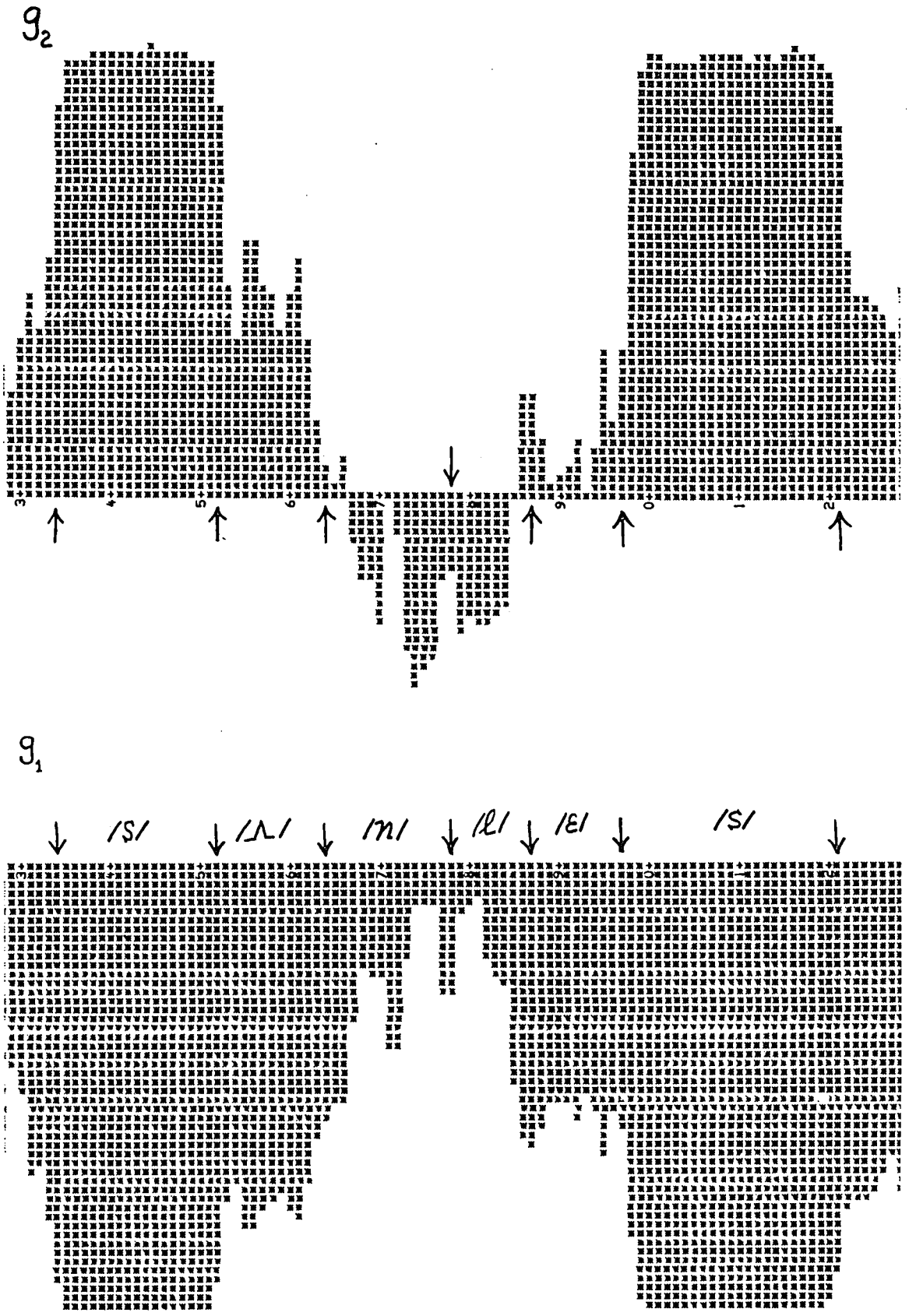


Figure # 17

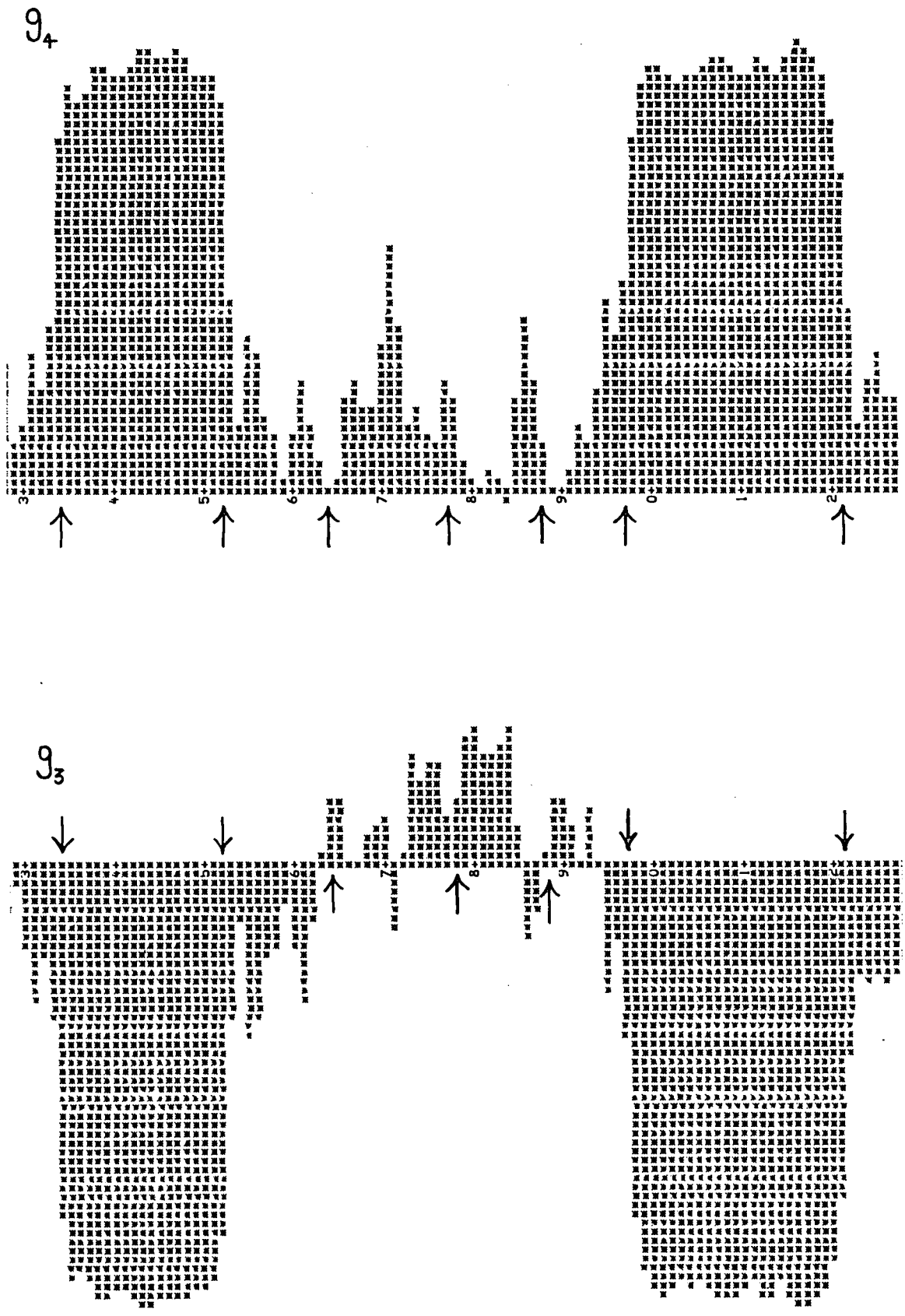


Figure # 18

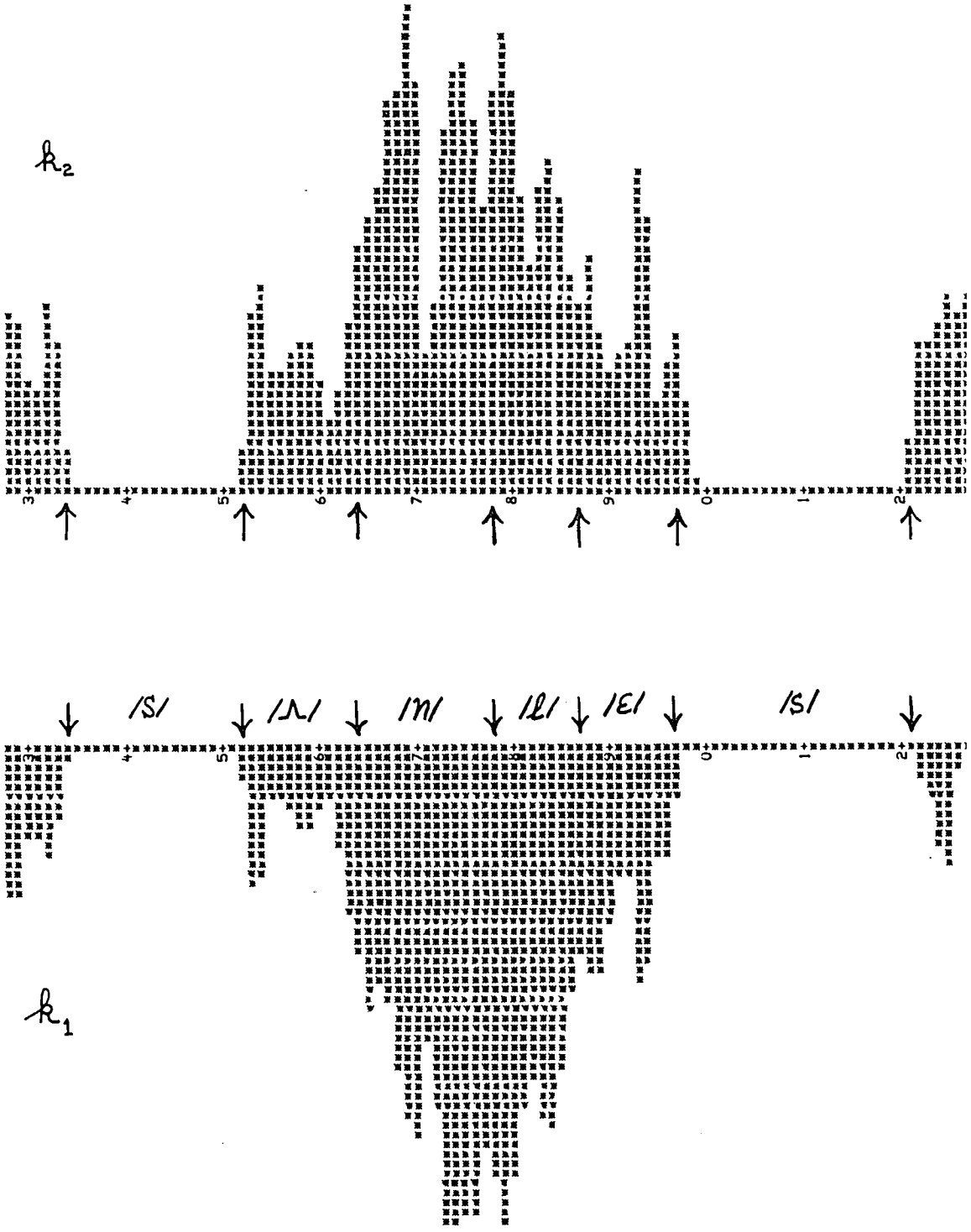


Figure # 19

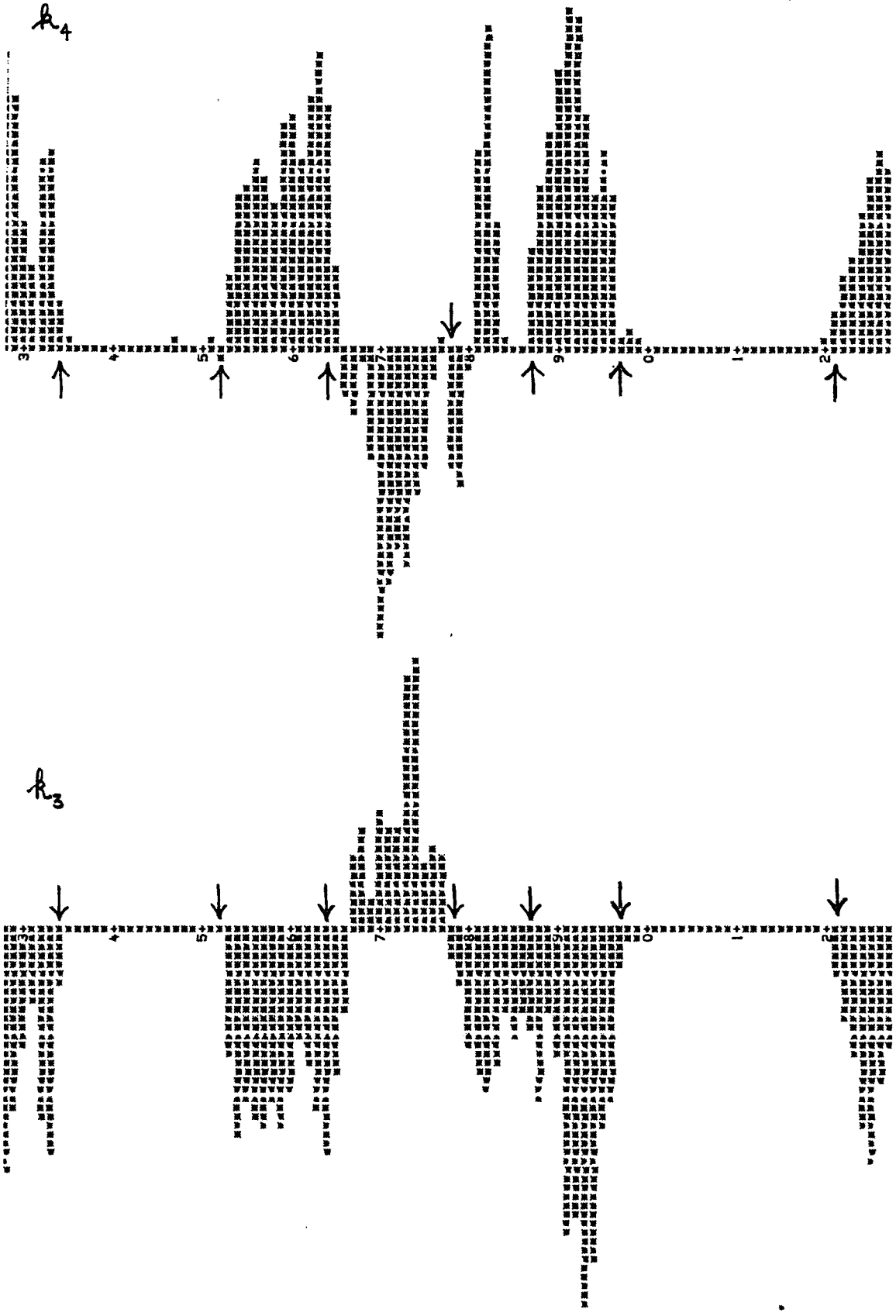


Figure # 20

VIII DISCUSSION OF RESULTS

The purpose of the previous section has been to show how to read and interpret the plots of the various functions of the speech signal that have been computed. The location of the phonemes in the word "sunless" was determined heuristically on a linguistic basis mostly from the DFT spectrogram and also using the a priori knowledge of the sound of the word spoken. The DHT spectrogram confirmed these findings but was not found to be as useful as the DFT for this purpose. An examination of the various transformations of the OX data for the "sunless" example (Figures #12-20) and the remainder of the word corpus (Tables II and III) showed strong correlations between classically determined phoneme boundaries and transients or "events" in these OX functions during the same time intervals.

The real problem in speech analysis is the reduction of a large amount of data, $\sim 5 \times 10^4$ bits/second, to a more tractable quantity comparable with the theoretical information rate of ~ 50 bits/second [30]. This can be achieved only if a reduction of data by at least a factor of 100 is first obtained, otherwise any practical attempt at machine analysis is hopeless. For these reasons, the DFT and DHT spectrographic techniques described here will not be exploited for segmentation procedures. The reason for this is that these spectra are basically one to one transformations which are only amenable to visual human analysis unless further reduced. The OX techniques do however give a substantial data reduction (100:1) and provide a set of functions which are in a vector space of low enough dimension for direct machine analysis in searching for phoneme boundaries.

Of the two types of OX transformations, moments and orthogonal

functions, one of each kind seemed worthy of further investigation. These are the central sample moments of the reciprocal OX distances and the Gram transforms of the OX distributions. The Gram transform is an orthogonal transformation in the L^2 sense, i.e., using the inner product defined by (18) (the Euclidean norm) and therefore generates representations in terms of the Gram polynomials, having errors minimized with respect to sums of squares or power. It is rarely the case in speech signals or OX distributions of them that the information (with respect to segmentation or recognition) is proportional to the power of some component of the speech signal or functions of it. But since most physical functions are L^2 and L^2 transforms are mathematically tractable, this approach to signal representation is very attractive.

In the case of OX distributions, it appears that the nonorthogonal moment analysis of the reciprocal distributions holds the most promise for the problems under consideration. This is believed since the $\{\bar{\mu}_{j,q}\}$ had the most consistent transient activity of all the functions at classical phoneme boundaries for all of the words tested. Moment analysis was originally chosen because of the simplicity of computation and interpretation. Such analyses have long been used for feature extraction of distribution functions. The physical interpretations of the first moment and the second through fourth central moments are well known; μ_1 = mean (center of gravity), μ_2 = variance (spread), μ_3 = skew (symmetry), and μ_4 = kertosity (peakedness). Because of their attractiveness and unusual nature, the $\{\bar{\mu}_{j,q}\}$ were chosen as input to a segmentation algorithm. Since these functions are essentially independent of amplitudes the RMS intensity function I_j was also chosen in order to establish a

characteristic function for the speech signal to determine regions in time of relative silence.

The principal elements of the segmentation algorithm which was finally chosen can be justified by examination of the next example, the word "into" which is written /In'tu/ phonemically. In fact the phonetic form of the instance of "into" which was spoken by the author was /$\tilde{\text{I}}\tilde{\text{I}}$n'tuUΛ/.† This shows the transitory nasal sound /$\tilde{\text{I}}$/ and the dipthongnization and closure of the final vowel /u/. The onset of nasalization during the final portion of a prenasal vowel is well known and has been demonstrated physiologically [31]. The termination of a final vowel with a weak neutral /Λ/ is a common regional linguistic phenomenon. The complex phonetic structure of this word was verified when the final algorithm was applied to it.

The first step necessary for segmentation, using the moment functions, is the delimiting of the null phoneme /Φ/. This was done by marking those places where I_j, Figure #21 for "into", crossed a threshold line and did not recross it for at least 3 points (30 msec.). The threshold line was 2.5 times the average of the first 10 points of I_j. This average thus represents a high estimate of the ambient noise level prior to the onset of speech. The speech boundaries found by this technique for the example "into" are shown as ↑ on Figure #21. The horizontal line drawn is the computed threshold level. The ↑ indicate speech between 41 and 67 and between 72 and 115. The 30 msec. hysteresis on the threshold was chosen because the shortest phonemes, the plosives /p/, /k/, and /t/, last at least 40 msec. This can be seen here as the /t / (after the short /Φ/ in the middle of the word) in "into" which lasts about 70 msec.

---

† Verified by a person with linguistic training

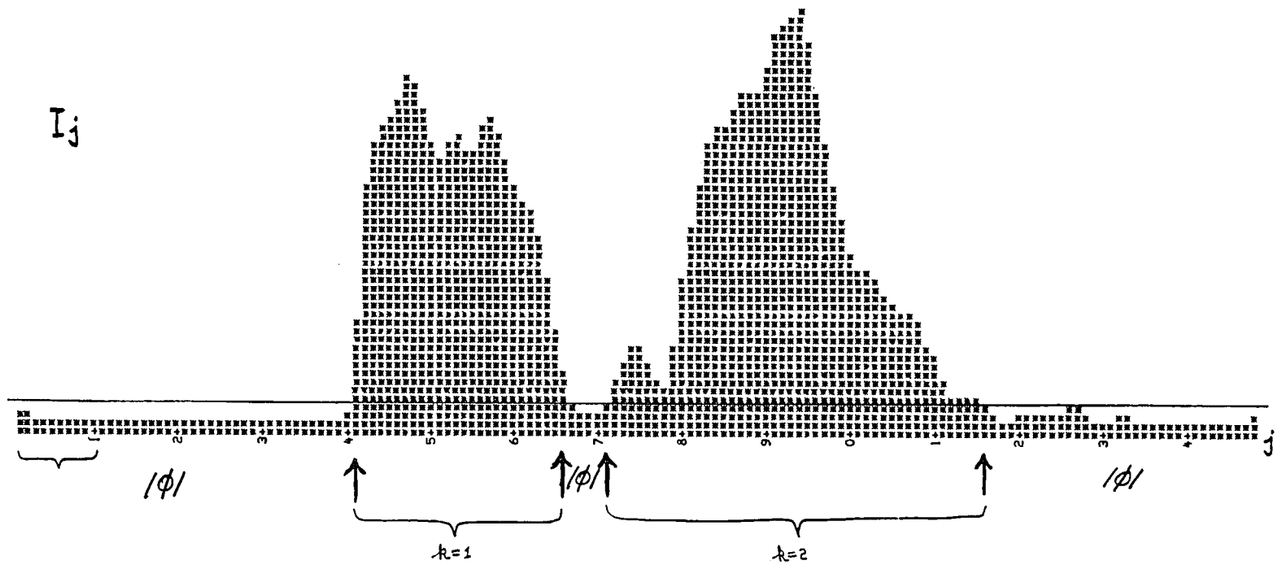


Figure #21

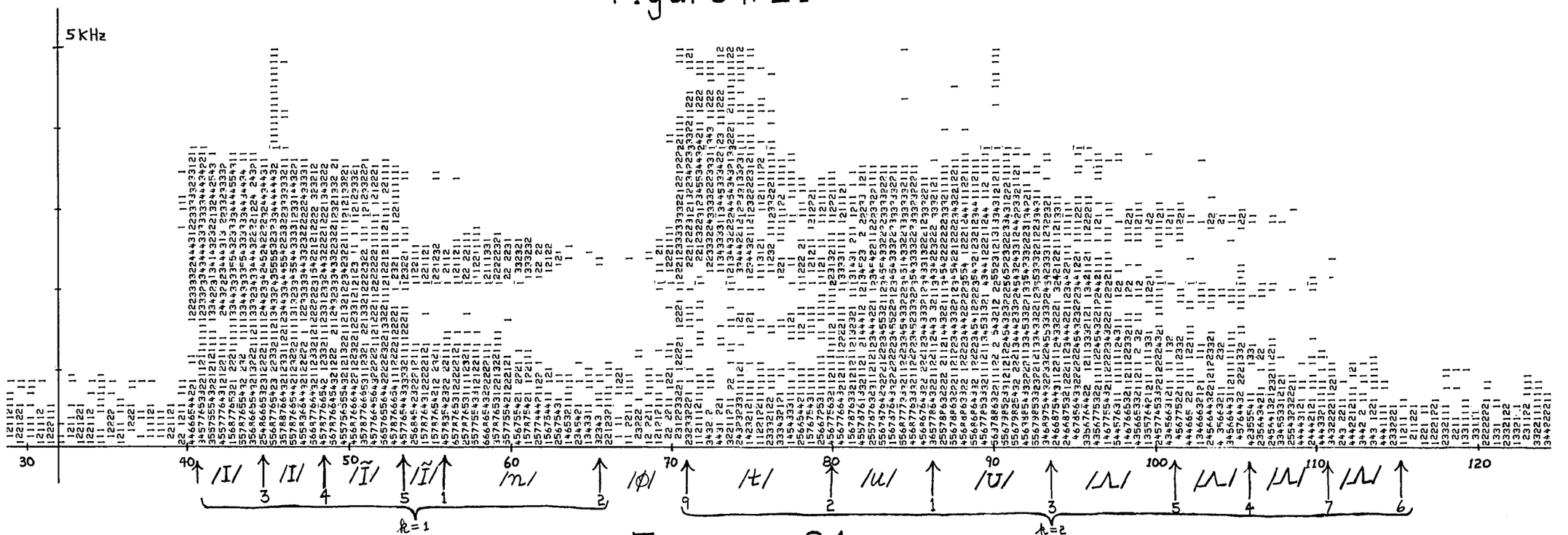


Figure #24

Figures #22 and #23 show $\bar{\mu}_{j,1}$, $\bar{\mu}_{j,2}$ and $\bar{\mu}_{j,3}$, $\bar{\mu}_{j,4}$ respectively with the delimiting ↑ for /Φ/ and other interesting "events". Event may be defined for the present as activity in $\{\bar{\mu}_{j,q}\}$, where q = 1,2,3,4, not in /Φ/ regions, and which is indicated by relatively large concurrent changes in the magnitude of the first differences of $\{\bar{\mu}_{j,q}\}$. The latter quantity is defined generally as

$$\Delta\bar{\mu}_{j,q} = \bar{\mu}_{j+1,q} - \bar{\mu}_{j,q} \tag{25}$$

In order to detect such events it was necessary to compute $\{|\Delta\bar{\mu}_{j,q}|\}$ and then order this set over each interval of time (the index j) not in regions of /Φ/. In order to relativize all measurements, the normalized quantity

$$\mathcal{E}^{k}_{j,q} = \frac{|\Delta\bar{\mu}_{j,q}|}{\max_{j}|\Delta\bar{\mu}_{j,q}|}, \quad j \in J_k \tag{26}$$

is computed where J_k is the k^{th} disjoint range set of j for the k^{th} speech interval, i.e., not in a /Φ/ region.

It should be obvious that the number of meaningful acousto-linguistic events in a given interval of speech is limited by the rate of speech and the language itself. For the problem at hand,

$$n_k = \text{card}\{J_k\}/4 \tag{27}$$

was chosen as the maximum number of events to be considered in the k^{th} speech interval. This simply means that a maximum rate of one phoneme/40 msec. is allowed, which is reasonable for this demonstration since the shortest phonemes in English are of this duration when spoken at a normal rate. For each k, the first n_k elements of $\mathcal{E}^{k}_{j,q}$ are defined as

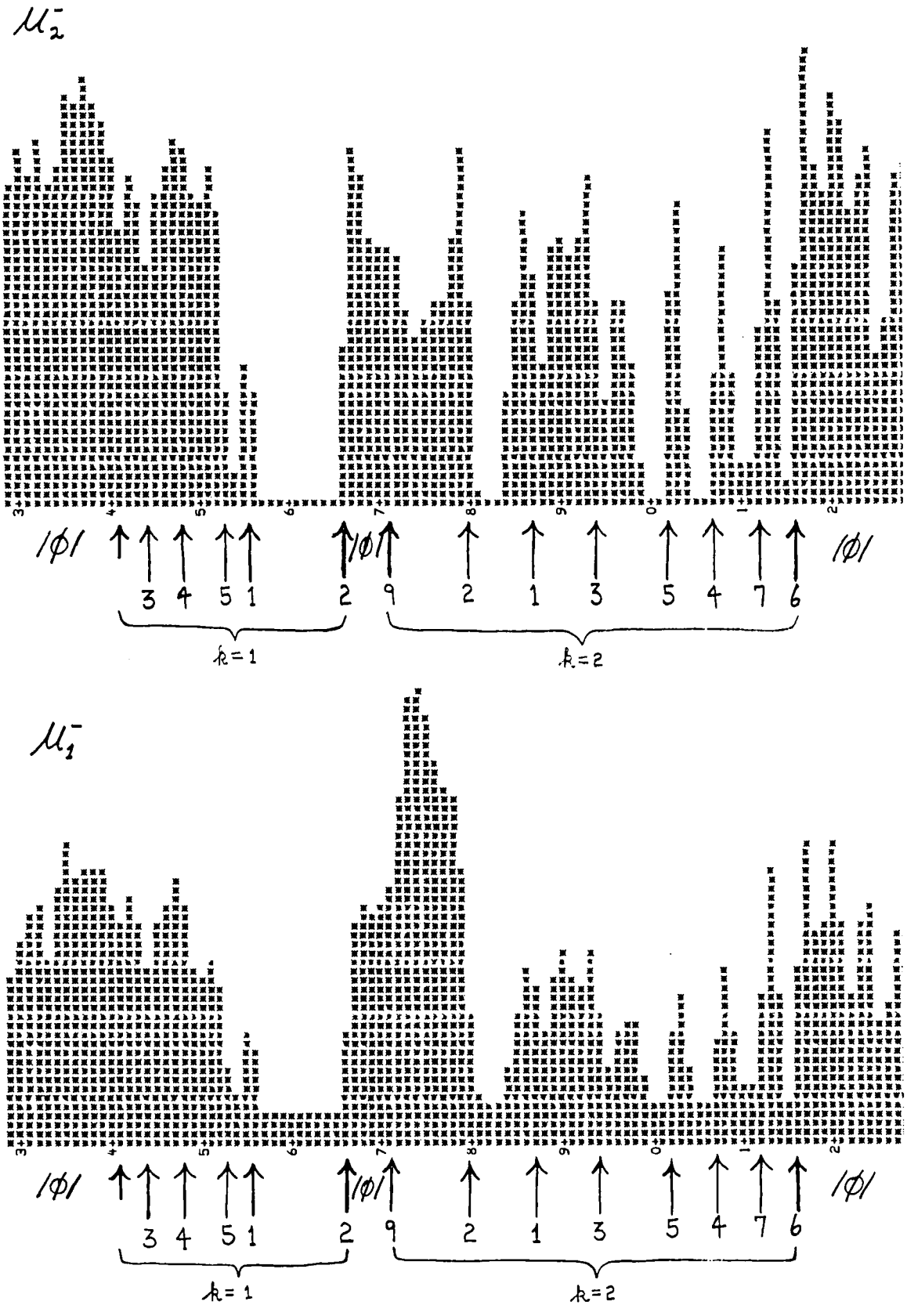

μ_2^-
3+
4+
5+
6+
7+
8+
9+
0+
1+
2+
|φ|
3 4 5 1 2 9 2 1 3 5 4 7 6
k=1
k=2
μ_1^-
3+
4+
5+
6+
7+
8+
9+
0+
1+
2+
|φ|
3 4 5 1 2 9 2 1 3 5 4 7 6
k=1
k=2

Figure #22

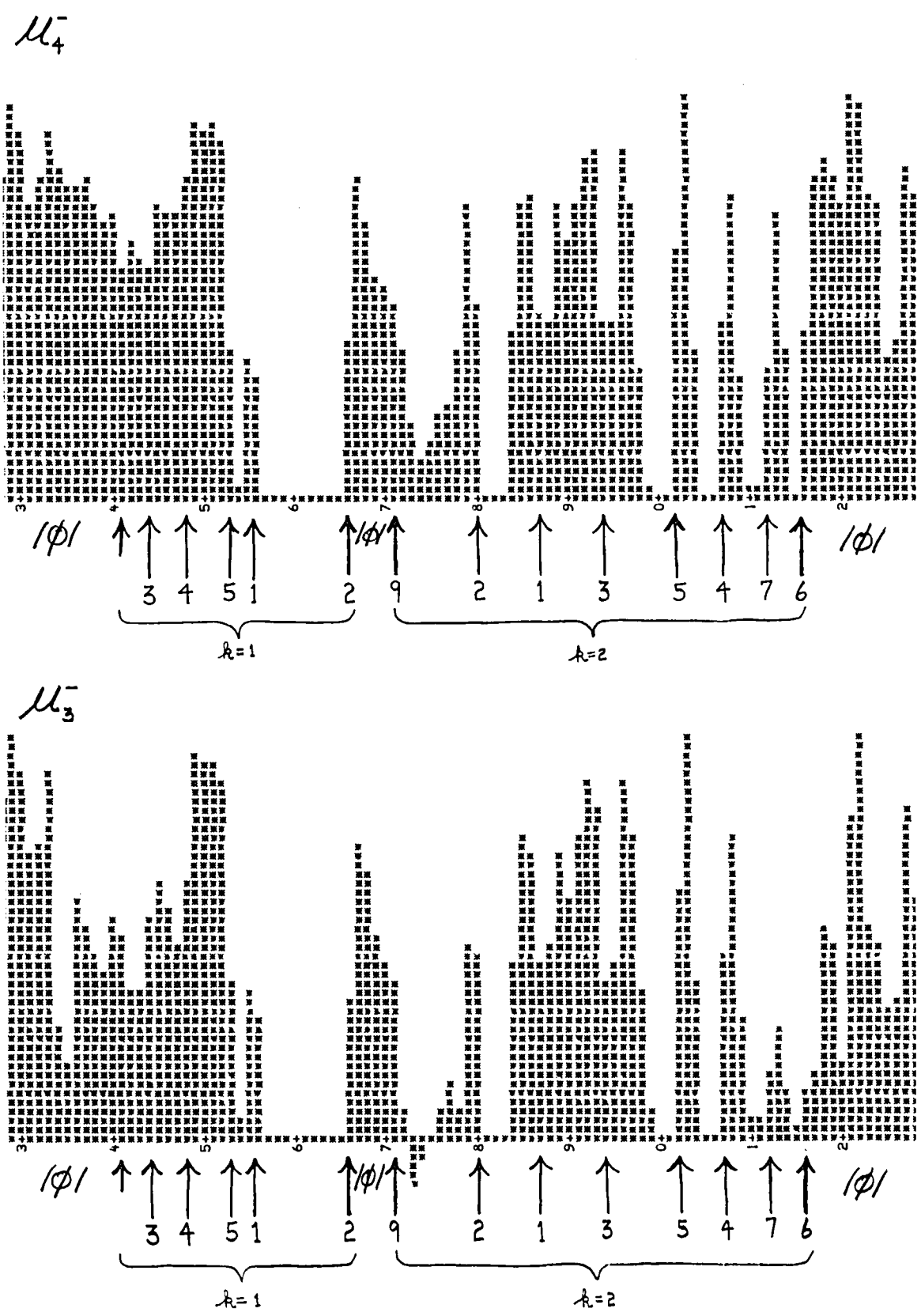

μ_4^-
3+
4+
5+
6+
7+
8+
9+
0+
1+
2+
|φ|
3 4 5 1 2 9 2 1 3 5 4 7 6
k=1
k=2
μ_3^-
3+
4+
5+
6+
7+
8+
9+
0+
1+
2+
|φ|
3 4 5 1 2 9 2 1 3 5 4 7 6
k=1
k=2

Figure #23

$\epsilon^k_{j,q}$. Since the $\mathcal{E}^k_{j,q}$ were ordered over j this means that given k and q $\epsilon^k_{j,q}$ is a sequence of the n_k largest first difference magnitudes in $\bar{\mu}_{j,q}$ for $j \in J_k$. For "into" k = 1,2 and $n_1 = (41-67)/4 = 6$ and $n_2 = (115-72)/4 = 10$. The number of possible events in any word containing T speech intervals is simply

$$n_T = \sum_{k=1}^{T} n_k \tag{28}$$

which is deliberately set high in order not to miss any subtle events.

The segmentation algorithm is then based on $\{\epsilon^k_{j,q}\}$ whose elements will be considered as weights of the linguistic significance of an event; computationally the respective values of j for each $\epsilon^k_{j,q}$ are stored along with the weight itself in the computer. The remaining information for k and q is implicit in the storage sequence of the weights. It should be clear that $\forall$ k,j,q, $\epsilon^k_{j,q} \in [0,1]$ due to (26). It now remains to measure concurrence for any k over q for $\epsilon^k_{j,q}$, i.e. which weights for different q have the same or nearly the same j or time or occurrence. Experience with a large number of words showed that there was a trend of less and less information in $\{\bar{\mu}_{j,q}\}$ as q increased. For this reason $\{\epsilon^k_{j,q}\}$ was used for q = 1,2,3,4 and concurrence in time was always determined with respect to $\epsilon^k_{j,1}$.

The algorithm then consists of three steps on $\{\epsilon^k_{j,q}\}$ which finally yields the segmentation markers Σ^k_j which are really pairs of weights and locations (in time). Σ^k_j can be thought of as an accumulator for the algorithm over the three steps which is initially set to zero. The first step is given by

$$\Sigma_j^k = \epsilon_{j,1}^k + \sum_{q=2}^{4} \epsilon_{i,q}^k, \quad \forall\ i = j. \tag{29}$$

This amounts to summing over q all weights which have exactly the same temporal index as those where q = 1, i.e. in the first moment. The second step is given by

$$\Sigma_j^k = \Sigma_j^k + \epsilon_{j,1}^k + \sum_{q=2}^{4} \epsilon_{i,q}^k, \tag{30}$$

$$\forall\ i,j \ni |i-j| = 1.$$

This then is summing all weights which are nearly (± 1 point) concurrent temporally to the weights from the first moment. An added complication in this step is that if $\epsilon_{i,q}^k > \epsilon_{j,1}^k$ for $|i-j| = 1$, then j in Σ_j^k is replaced for that weight by i. This simply associates the location which corresponds to the larger weight for more accurate determination of boundaries. The final step removes all markers which are too close (± 1) together by computing

$$\Sigma_j^k = \Sigma_j^k + \Sigma_i^k, \tag{31}$$

$$\forall\ i,j \ni |i-j| = 1.$$

Here again the index of the larger weight is used as a final marker location and the index of the smaller weight is discarded. The remaining indices and their accumulated weights define the set of segmentation markers for a word.

For the word "into" the segmentation markers found by the algorithm are indicated by the large ↑ in Figures #22 and #23. The ordinal under each ↑ is the rank of the weight associated with that marker. Some of

the markers coincide exactly with the markers of /Φ/ found from I_j in Figure #21. It is clear that all of the markers are at or very near transition regions in the moments and to a lesser extent in the RMS intensity. Due to the discrete nature of all of the functions used in this analysis, the markers are really only determined within ± 1 point of their computed location, i.e., ± 10 msec. This uncertainty of position is still well within the limits necessary for practical application of the technique for recognition purposes.

It is appropriate now to indicate what legitimate linguistic phenomena if any can be associated with these markers. This is done by comparing the location of classical linguistic activity, heuristically determined from the DFT spectrogram, to the position of the markers. The spectrogram for "into" can be seen in Figure #24 with the ↑. Recall that the phonetic form of the sample was /ɪĩn'tuʊʌ/. With this in mind let us examine the first speech interval for k = 1 in the spectrogram. The most obvious change in this interval can be seen where marker 5 is at 53. This is the vowel-nasal boundary for the first syllable, i.e., the point when a majority of the acoustic power is nasal rather than oral. This is true since the two formant structure of the /n/ starts at that time. The onset of nasalization, when the velum begins to open the path from the larynx to the nasal cavity, is indicated at 48 by marker 4. This can be seen as a dropping off of the second formant in the vowel, and an overall loss of power as seen in Figure #21.

The remaining markers 3 and 1 in this interval can also be explained. Marker 1 at 56 is undoubtedly due to the completion of the vowel-to-nasal transition. Physiologically this occurs when the apex (tip) of the

tongue touches the hard palate just behind the teeth and thus cuts off any sound from the mouth. The region between markers 4 and 1 is the transitional sound /$\tilde{I}$/ indicated earlier. Marker 3 could only have been caused by the initial vowel /I/ reaching near full power from silence in about 50 msec. This sudden change in the rate of buildup in air velocity in the vocal tract carries no apparent linguistic information but caused a sufficient change in the moments to be detected. Marker 2 coincides with the stop interval before the plosive /t/ which had been already detected by the threshold level detector.

The second interval and syllable, k = 2, are easier to explain with reference to the markers found there. Marker 9 corresponds to the explosion of the /t/ and the beginning of the interval. It, as well as marker 2 in the previous interval and 6 in this interval, coincides with points found by the threshold detector. The typical formant structure of a vowel can be seen starting at marker 2. From marker 2 through 5 we can see the top three formants move steadily downward to lower frequencies. This is the triphthong /uUʌ/ found in this instance of "into". From marker 5 to 6 lies a region where the word is essentially dying. The power in the sound is very low after 100 csec. as can be seen from the small values in Figure #24 or graphically on Figure #21. The markers in this region correspond to a very low frequency component in the vocal cord vibrations due to instability caused by decreasing air pressure from the lungs. This is seen on the spectrogram as an increase in the value of the spectral components just before markers 4 and 7. This must be considered as noise but is typical of this speaker for words ending in a vowel.

It is felt that all the markers found for this word are linguistically and acoustically meaningful. There is a definite correlation between linguistic significance and the weights that were assigned automatically to the markers. For this word and almost all others used, a marker was always placed very near (± 10 msec.) a classically determined phoneme boundary. The remaining markers could be interpreted as being associated with transitional sounds or phonemes, or as high level linguistic noise, e.g., lip noise, and vocal cord instability. The only mismarking or omissions were due to known shortcomings in the marking process. A more involved threshold algorithm perhaps including the use of $\dot{I}_j$ would have prevented improper delimiting of /Φ/, which in some cases resulted in the application of the segmentation algorithm to regions of noise generating an inordinate number of highly weighted markers. Because of this, the quota n_k for an interval was used up on spurious events causing other real events not in the noise interval to be completely overlooked. A more sophisticated segmentation algorithm could possibly contend with: (1) very low power or dying vowels where vocal instabilities arose, and (2) highly transitional phonemes such as the liquids /ℓ/ and especially /r/.

The subset of the corpus used to actually test the segmentation algorithm was the set of words on Tables II and III which are followed by an asterisk. A consolidated list of these words and the segmentation scores are found in Table IV. Each word is followed by two fractions. The <u>first</u> is the ratio of markers which corresponded to actual classical phonemic boundaries to the total number of such boundaries. The <u>second</u> is the ratio of markers which corresponded to significant linguistic or

TABLE IV

Test Word	Scores
1. Sunless	6/7, 3/3
2. Monday	6/6, 4/4
3. Zero	5/5, 4/4
4. Speakers	8/8, 4/4
5. Himself	8/8, 3/4
6. Speechless	8/8, 4/4
7. Which	4/4, 3/3
8. Into	5/5, 5/5
9. Only	4/5, 2/5
10. Some	5/5, 2/4
11. First	6/6, 3/3
12. Did	4/4, 2/3
13. Many	5/5, 1/3
Totals	97%, 82%

acoustic phenomena. The latter class of events are made up of: (1) noisy phonemes, e.g., /r/ and final vowels, (2) high level articulatory noise, and (3) phonetic boundaries which do not coincide with phonemic boundaries, e.g., nondisjoint pairs like vowel-nasal combinations. The most significant fact, other than the consistently high scores, is that none of the ratios are greater than unity. This attests to the validity of even so simple a segmentation algorithm. It was also the case that every instance of a syllable boundary was detected by this algorithm. It seems clear that the algorithm could be improved to eliminate the problems previously stated and would give highly accurate results. The high performance of this algorithm considering its simplicity indicates that the reciprocal OX moments give a reliable measure of changes in stationarity indicating phoneme transitions in speech.

IX CONCLUSIONS

As far as the original goals are concerned, this research was a success. The flexibility of the time series analysis system which was developed played a major role in allowing a large number of different approaches to be explored in an efficient manner. Although the spectral representations (DFT and DHT) investigated were used only as a fiducial base for phoneme boundaries generated linguistically by the author, the work done in that area was for the most part new and could bear further probing for other applications, e.g., speech recognition. The new techniques for OX analysis of nonstationary time series proved to be very fruitful for speech analysis and have opened new avenues for future research. These methods are currently being used for signal analysis research under an ONR contract and have been suggested as a basis for a new type of hearing aid [32].

The segmentation of speech can be considered in three domains: linguistic, acoustic, and articulatory (physiological). The methods explored have shown the existence of acoustic cues which correspond to certain linguistic and spectral features denoting phonetic transitions. In fact earlier work has also revealed certain physiological correlates to acousto-linguistic boundaries [33]. In practice this segmentation procedures gives only phonetic boundaries which are not paired as to onset and termination. This results in a subphonetic partition which then must be filtered using linguistic information via recognition in order to achieve phonemic segmentation. The nondisjoint nature of this partition seems to have been borne out in this experiment. An objective definition of a segmentation has thus been created in the form

of an algorithm and has led to a deeper insight into the possible nature of the human speech generation process.

APPENDIX I

Haar Functions

The Haar basis [15] is defined as

$$\mathcal{H}_n^k(x) = \begin{cases} 2^{n-1/2}, & x \in [(2k-2)\lambda, (2k-1)\lambda] \\ -2^{n-1/2}, & x \in [(2k-1)\lambda, 2k\lambda] \\ 0, & \text{elsewhere} \end{cases}$$

where

$\lambda = 2^{-(n+1)}$,

$n = 0, 1, 2, \ldots, \infty$

$k = 1, 2, 3, \ldots, 2^{n-1}$, and

$\mathcal{H}_0(x) = 1, \quad x \in [0,1]$.

This is an orthonormal set which is complete in L^2. The first few Haar functions are shown below.

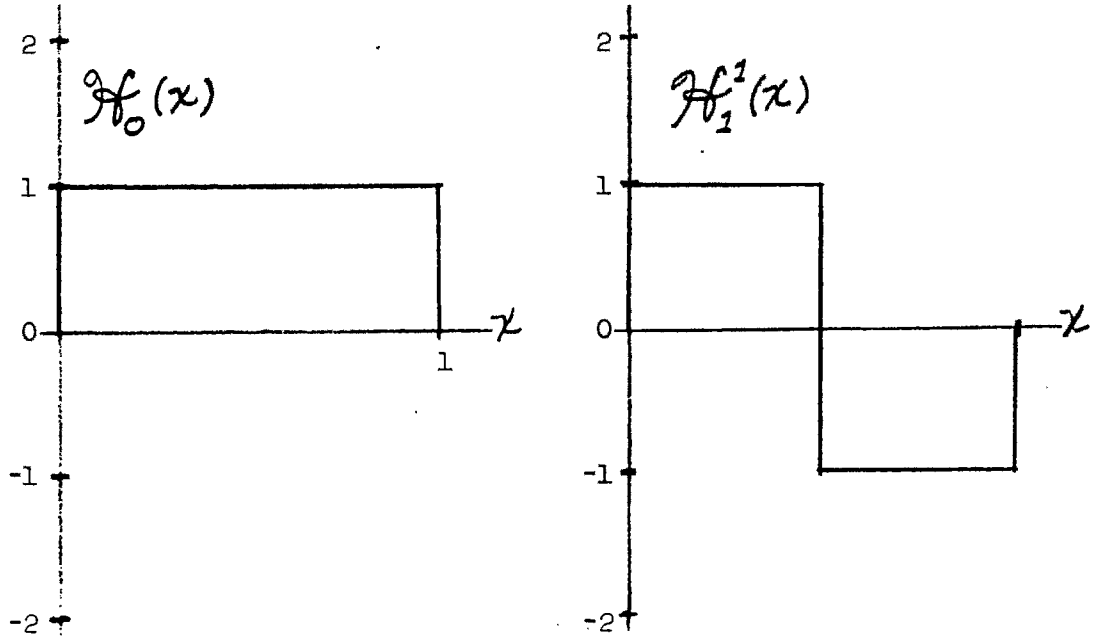


FIGURE #I-1

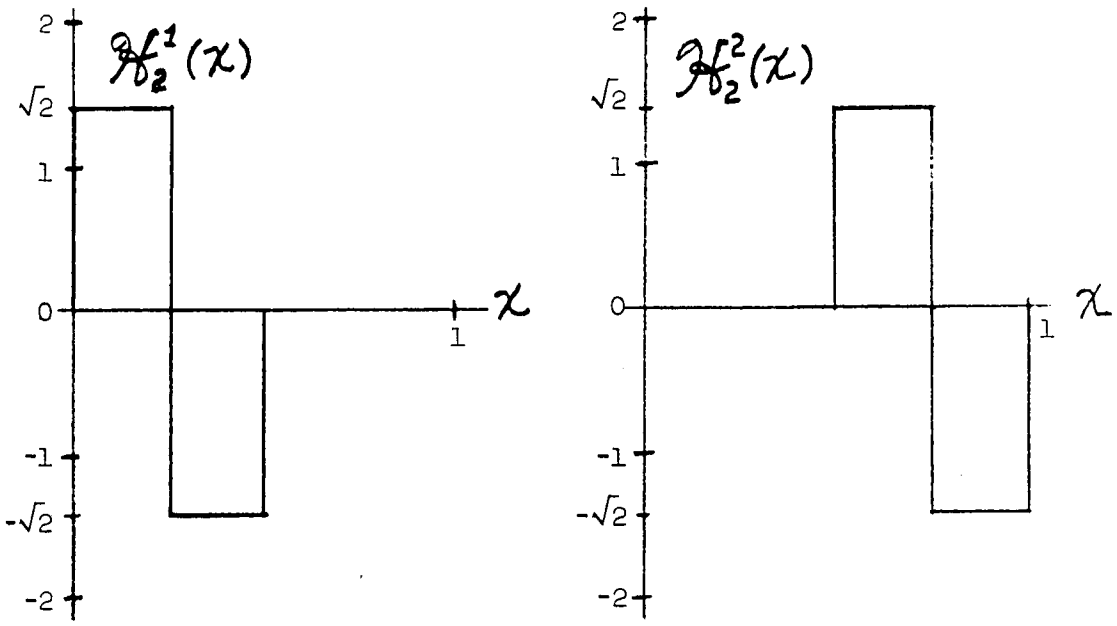


FIGURE #I-2

Haar functions may therefore be constructed from combinations of step functions. Given a function $f(x) \in L^2[0,1]$, we may therefore expand $f(x)$ in terms of Haar functions by computing

$$\alpha_n^k = \int_0^1 f(x)\mathcal{H}_n^k(x)dx,$$

where

$$f(x) = \sum_{n,k}^{\infty} \alpha_n^k \mathcal{H}_n^k(x).$$

In the case where $x' = 0,1,\ldots,N-1$ we may define the discrete Haar basis by restricting $N = 2^m$. The original definition needs only the following modifications:

$$\lambda' = \lambda(N-1),$$

$$n = 0,1,2,\ldots,m-1, \text{ and}$$

$$k = 1,2,3,\ldots,2^{n-1}.$$

Any bounded function f(x') may be expanded in terms of the discrete Haar basis by using

$$\alpha_n^k = \frac{1}{N} \sum_{x'=0}^{N-1} f(x') \mathcal{H}_n^k(x'),$$

where

$$f(x') = \sum_{n=0}^{m-1} \sum_{k=1}^{2^{n-1}} \alpha_n^k \mathcal{H}_n^k(x').$$

Further properties of orthogonal bases composed of functions of a discrete, bounded variable are discussed in Appendix III. The actual computation of the α_n^k's are trivial numerically since the basis functions are simply constants. One merely computes sums of those portions of f(x') where $\mathcal{H}_n^k(x') \neq 0$, and then multiplies by $\mathcal{H}_n^k(x')/N$.

APPENDIX II

Least Squares Line

Given a set of observed values $\{y_i^o\}$ it is desired to find the function $y_i^t = a+bi$ such that the

$$\sum_{i=1}^{N} \left(y_i^o - y_i^t\right)^2$$

is minimum. This is the least squares problem for the straight line and the coefficients a and b may be found by solving the system of equations

$$a \sum_{i=1}^{N} 1 + b \sum_{i=1}^{N} i = \sum_{i=1}^{N} y_i^o$$

$$a \sum_{i=1}^{N} i + b \sum_{i=1}^{N} i^2 = \sum_{i=1}^{N} y_i^o i$$

if one can make the restriction that the y_i^o are equispaced samples. In the case of most discrete time series this is true and therefore leads to simplifications. The above system can be rewritten by replacing the sums on the left hand sides giving

$$aN + b\,\frac{N(N+1)}{2} = \sum_{i=1}^{N} y_i^o$$

$$a\,\frac{N(N+1)}{2} + b\,\frac{N(2N^2+3N+1)}{6} = \sum_{i=1}^{N} y_i^o i.$$

In order to facilitate manipulation, the above is rewritten as

$$an_1 + bn_2 = n_3$$

$$an_2 + bn_4 = n_5,$$

which may be solved analytically. The solution was found to be

$$a = \frac{n_3}{n_1} - \frac{n_2}{n_1}\left(\frac{n_1 n_5 - n_2 n_3}{n_4 n_1 - n_2^2}\right),$$

$$b = \frac{n_1 n_5 - n_2 n_3}{n_4 n_1 - n_2^2} .$$

Before substituting the values of the n's, another simplification can be found. Recall that

$$n_3 = \sum_{i=1}^{N} y_i^o, \text{ or}$$

$$n_3 = N\overline{y^o} .$$

If the mean value of the time series is removed beforehand, i.e.

$$y_i^{o*} = y_i^o - \overline{y^o}, \text{ or}$$

$$\overline{y^{o*}} = 0 ,$$

then after resubstitution and reduction we have

$$a = -\frac{(N+1)b}{2}, \text{ and}$$

$$b = \frac{\sum_{i=1}^{N} y_i^{o*} i}{N/12(N^2-1)} .$$

In terms of b, the least squares straight line is then

$$y_i^t = -\frac{(N+1)b}{2} + bi \text{ , or}$$

$$y_i^t = b\left[i - \frac{(N+1)}{2}\right] .$$

This represents a very compact and efficient means for removing linear trends in equispaced data.

A convenient property of linearly detrended data is that its mean is zero. This can be easily proved by writing first

$$\tilde{y}_i^o = y_i^{o*} - y_i^t ,$$

and computing the mean

$$\overline{\tilde{y}^o} = \overline{y^{o*}} - \overline{y^t} .$$

By definition we have set $\overline{y^{o*}} = 0$ and

$$\overline{y^t} = \overline{b\left[i - \frac{(N+1)}{2}\right]} \text{ , or}$$

$$\overline{y^t} = b\left[\bar{i} - \overline{\left(\frac{N+1}{2}\right)}\right] \text{ , where}$$

$$\bar{i} = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N(N+1)}{N2} = \frac{(N+1)}{2} .$$

Therefore we find that

$$\overline{\tilde{y}^o} = 0 - b\left[\frac{(N+1)}{2} - \frac{(N+1)}{2}\right] \equiv 0 .$$

Q.E.D.

APPENDIX III

Arithmetic and Harmonic Moments

Let d_i be the distance or time, i.e., number of points, between the i^{th} and $i + 1^{st}$ zero for any given time series section. We may then write

$$d_i = \bar{d} + \Delta d_i,$$

where Δd_i is the variation of d_i and

$$\bar{d} = \frac{1}{P} \sum_{i=1}^{P} d_i$$

is the mean. P+1 is the total number of zeros in a segment D points long, therefore

$$\bar{d} \simeq \frac{D}{P},$$

i.e., the mean of the ox distances is approximately proportional to the reciprocal number of zeros. By definition then we have $\overline{\Delta d_i} = 0$ and thus $\bar{d} \neq f(\Delta d_i)$; i.e., the mean is independent of the variation.

For the case of reciprocal ox distances d_i^{-1} we can write

$$\overline{d^{-1}} = \overline{(\bar{d}+\Delta d_i)^{-1}} = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{(\bar{d}+\Delta d_i)}$$

Assuming that $\Delta d_i / \bar{d} < 1 \; \forall i$, then

$$\overline{d^{-1}} = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{\bar{d}} \left[1 - \frac{\Delta d_i}{\bar{d}} + \left(\frac{\Delta d_i}{\bar{d}} \right)^2 - \ldots \right],$$

or

$$\overline{d^{-1}} \approx \frac{1}{P\bar{d}^2} \sum_{i=1}^{P} \left[\bar{d} - \Delta d_i + \frac{\Delta d_i^2}{\bar{d}}\right].$$

This reduces to

$$\overline{d^{-1}} \approx \frac{1}{\bar{d}} \left[1 + \overline{\left(\frac{\Delta d}{\bar{d}}\right)^2}\right] = g(\Delta d),$$

and thus $\overline{d^{-1}}$ is dependent on the variation.

Let us define for $q \geq 2$:

$$\mu_q \triangleq \overline{(\Delta d_i)^q}, \text{ and}$$

$$\mu_q^- \triangleq \overline{[\Delta(d_i^{-1})]^q} = \overline{\left(\frac{1}{\bar{d}+\Delta d_i} - \overline{d^{-1}}\right)^q}.$$

Let $<\quad>$ be used for $\overline{\qquad\qquad}$, then

$$\mu_q^- = \left\langle\left[(\bar{d}+\Delta d_i)^{-1} - \langle(\bar{d}+\Delta d_i)^{-1}\rangle\right]^q\right\rangle$$

$$\mu_q^- = \frac{1}{\bar{d}^q}\left[\left(1 - \frac{\Delta d_i}{\bar{d}} + \left(\frac{\Delta d_i}{\bar{d}}\right)^2 - \cdots\right) - \left\langle\left(1 - \frac{\Delta d_i}{\bar{d}} + \left(\frac{\Delta d_i}{\bar{d}}\right)^2 - \cdots\right)\right\rangle\right]^q,$$

for $\frac{\Delta d_i}{\bar{d}} < 1$

$$\mu_q^- \approx \frac{1}{\bar{d}^q}\left\langle\left[-\frac{\Delta d_i}{\bar{d}} + \frac{(\Delta d_i)^2 - \overline{(\Delta d_i)^2}}{\bar{d}^2}\right]^q\right\rangle.$$

$$\mu_q^- \approx \frac{1}{\bar{d}^q}\left\langle\left(\frac{\Delta d_i}{\bar{d}}\right)^q + q\left(-\frac{\Delta d_i}{\bar{d}}\right)^{q-1}\left[\frac{(\Delta d_i)^2 - \overline{(\Delta d_i)^2}}{\bar{d}^2}\right] + \cdots\right\rangle,$$

$$\mu_q^- \approx \frac{(-\Delta d_i)^q}{\bar{d}^{2q}} = \frac{(-1)^q}{\mu_1^{2q}}\mu_q, \quad \text{and}$$

$$\bar{\mu}_1 \approx \frac{1}{\mu_1}\left[1 + \frac{\mu_2}{\mu_1^2}\right], \text{ where } \mu_1 \triangleq \bar{d}.$$

The above relations between $\bar{\mu}_q$ and μ_q hold only if $\Delta d_i < \bar{d}$ but are at least of interest for qualitative comparison.

APPENDIX IV

Discrete, Finite Orthogonal Functions

The most general form of the L^2 orthogonality property is given by

$$\int_{-\infty}^{\infty} \psi_n(x)\psi_m(x)d\alpha(x) = c_n\delta_{n,m}$$

which if $\alpha(x) \in C^1$ can be written

$$\int_{-\infty}^{\infty} \psi_n(x)\psi_m(x)\alpha'(x)dx = c_n\delta_{n,m}$$

where

$$\delta_{n,m} = \begin{cases} 1, & n = m \\ 0, & n \neq m \end{cases}$$

This means that the set $\{\psi_i(x)\}$ is orthogonal in L^2 with respect to the weighting function $W(x) = \alpha'(x)$. If $\alpha(x)$ is a step function with jumps at $x = 0,1,2,\ldots,N-1$ (where N is a fixed positive integer), then we may write

$$\sum_{x=0}^{N-1} \psi_n(x)\psi_m(x)W(x) = c_n\delta_{n,m}$$

$$n,m = 0,1,\ldots,N-1$$

Here $W(x)$ are the values of the jumps at the grid points $x \in [0,N-1]$ as described above. This now defines orthogonality for functions of a

discrete bounded variable with respect to a weighting function of the same variable.

It is usually most convenient to deal with functions that are orthonormal with respect to a unity weighting function, i.e., $c_n = 1$ and $W(x) = 1$. This can be easily done by defining a new set of functions

$$\varphi_n(x) = \frac{\psi_n(x)}{\sqrt{c_n W(x)}}$$

which will have the property

$$\sum_{x=0}^{N-1} \varphi_n(x)\varphi_m(x) = \delta_{m,n}.$$

Functions of this type are very useful in the computation of linear transformations on arbitrary bounded functions of a discrete bounded variables, e.g.

$$\Phi_n = \sum_{x=0}^{N-1} f(x)\varphi_n(x), \text{ where}$$

$$f(x) = \sum_{n=0}^{N-1} \Phi_n\varphi_n(x).$$

Completeness is guaranteed if there exist N mutually orthogonal functions in a vector space of dimension N.

One of the most interesting classes of discrete orthonormal functions are those where

$$\psi_n(x) = P_n(x,N) = \sum_{i=0}^{n} a_i x^i$$

$$x = 0,1,\ldots,N-1$$

i.e., $\{\psi_n(x)\}$ are polynomials evaluated on an equispaced grid of N points [26,28]. If one chooses the discrete weighting functions $W(x,N) = 1$, then the resulting finite set is given by

$$G_n(x,N+1) = c_{n,N} \sum_{k=0}^{n} \frac{(-1)^k (n+k)^{(n)}}{k!} \binom{N-k}{n-k} x^{(k)},$$

and are known as the Gram or Chebychev polynomials. The form $a^{(b)}$ in in the above equation is defined as

$$a^{(b)} = \frac{a!}{(a-b)!} = \prod_{j=0}^{b-1} (a-j) = \frac{\binom{a}{b}}{b!}.$$

The normalizing constant can be found to be

$$c_{n,N} = (-1)^n [(2n+1)N^{(n)}/(N+n+1)^{(n+1)}]^{1/2}.$$

The more standard form for these polynomials can be derived and is

$$G_n(x,N+1) = c_{n,N} \sum_{k=0}^{n} \frac{(-1)^k (n+k)^{(2k)}}{(k!)^2 N^{(k)}} x^{(k)}$$

The polynomials were evaluated for $N = 74$, and $n = 1,2,3,4,5,6$ and are shown in Figure #IV-1 to be compared to Chebychev functions

$$\mathcal{T}_n(x) = \frac{\cos(n \cos^{-1} x)}{(1-x^2)^{1/4}},$$

$$x \in [-1,1]$$

which they resemble for higher orders.

Another very interesting set of polynomials is associated with the weighting function

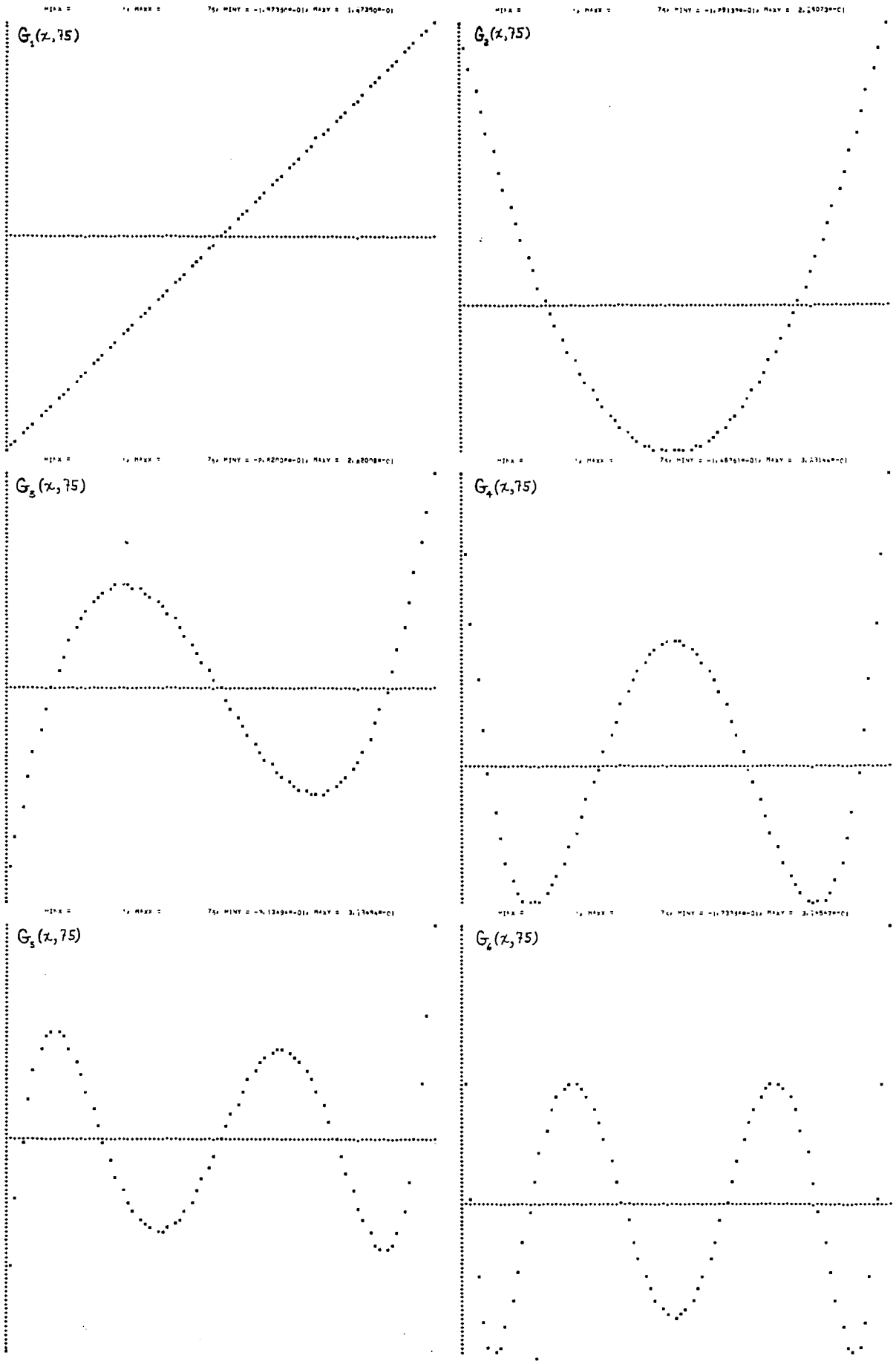


Figure # IV-1

$$W(x,N+1) = \binom{N}{x},$$

and are given by

$$K_n(x,N+1) = c_{n,N} \sum_{k=0}^{n} (-1)^{n-k} \binom{N-x}{n-k} \binom{x}{k}.$$

These are the Krawtchouk polynomials [26,27] and the orthonormal basis is

$$\mathcal{K}_n(x,N+1) = K_n(x,N+1)[W(x,N+1)]^{1/2}$$

where

$$c_{n,N} = [(\tfrac{1}{2})^N / \binom{N}{n}]^{1/2}.$$

These functions have the interesting property that

$$\lim_{N\to\infty} \mathcal{K}_n(x,N) = \mathcal{H}_n(z),$$

where $\mathcal{H}_n(z)$ are the Hermite functions [27,28] given by

$$\mathcal{H}_n(z) = \frac{H_n(z)e^{-z^2/2}}{(n!2^n\sqrt{\pi})^{1/2}}$$

$H_n(z)$ are the Hermite polynomials which are defined as

$$H_n(z) = (-1)^n e^{z^2} \frac{d(e^{-z^2})}{dz^n},$$

and

$$x = \frac{N}{2} + z(\frac{N}{2})^{1/2}.$$

The Krawtchouk functions are therefore the discrete equivalent of the Hermite functions. These two functions were computed for N = 74.

and n = 0,1,2,3,4,5 and are shown in Figure #IV-2 and Figure #IV-3. The difference in these two functions only becomes evident for higher order n.

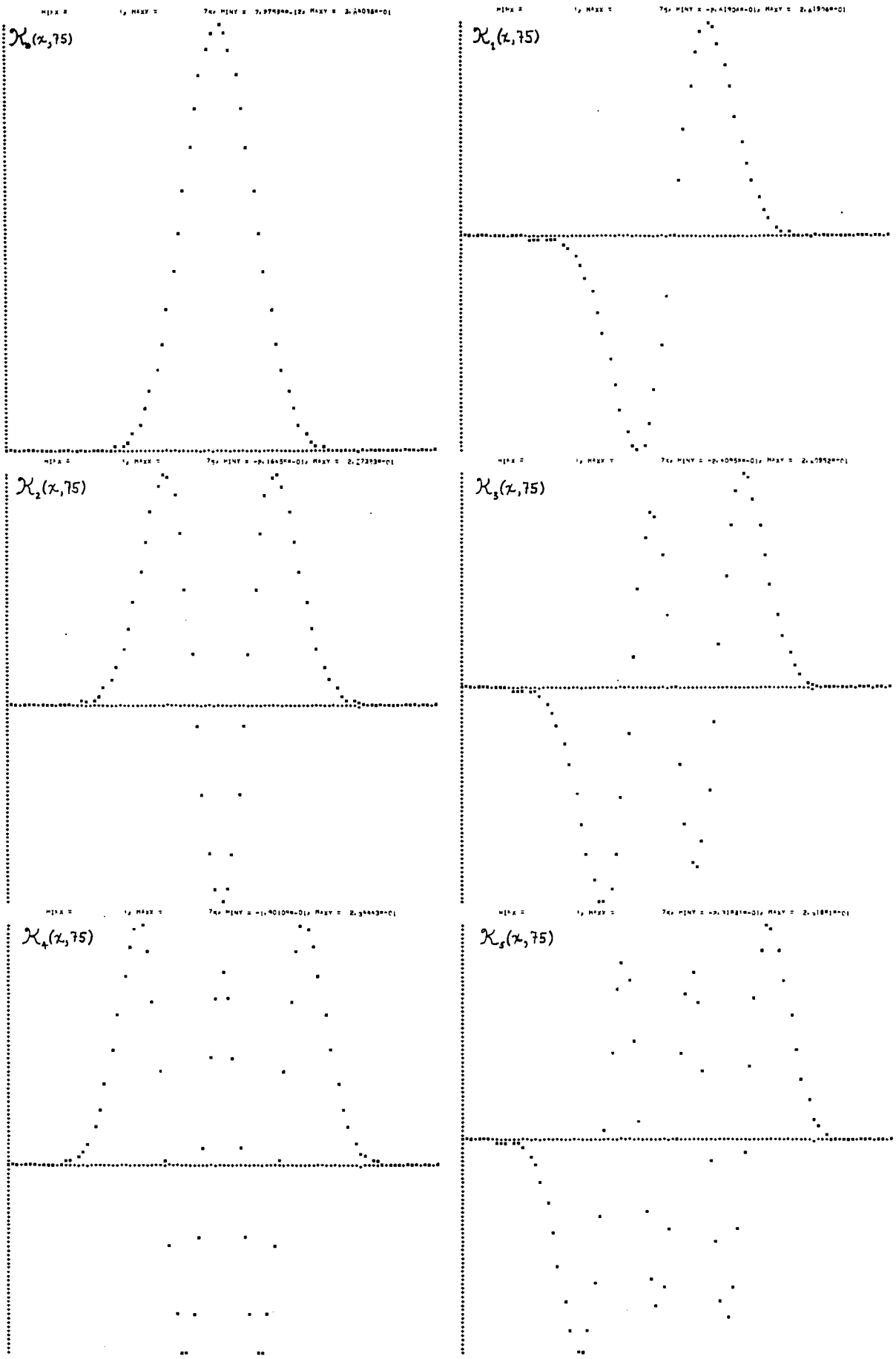

$\mathcal{K}_0(x,75)$
$\mathcal{K}_1(x,75)$
$\mathcal{K}_2(x,75)$
$\mathcal{K}_3(x,75)$
$\mathcal{K}_4(x,75)$
$\mathcal{K}_5(x,75)$

Figure # IV-2

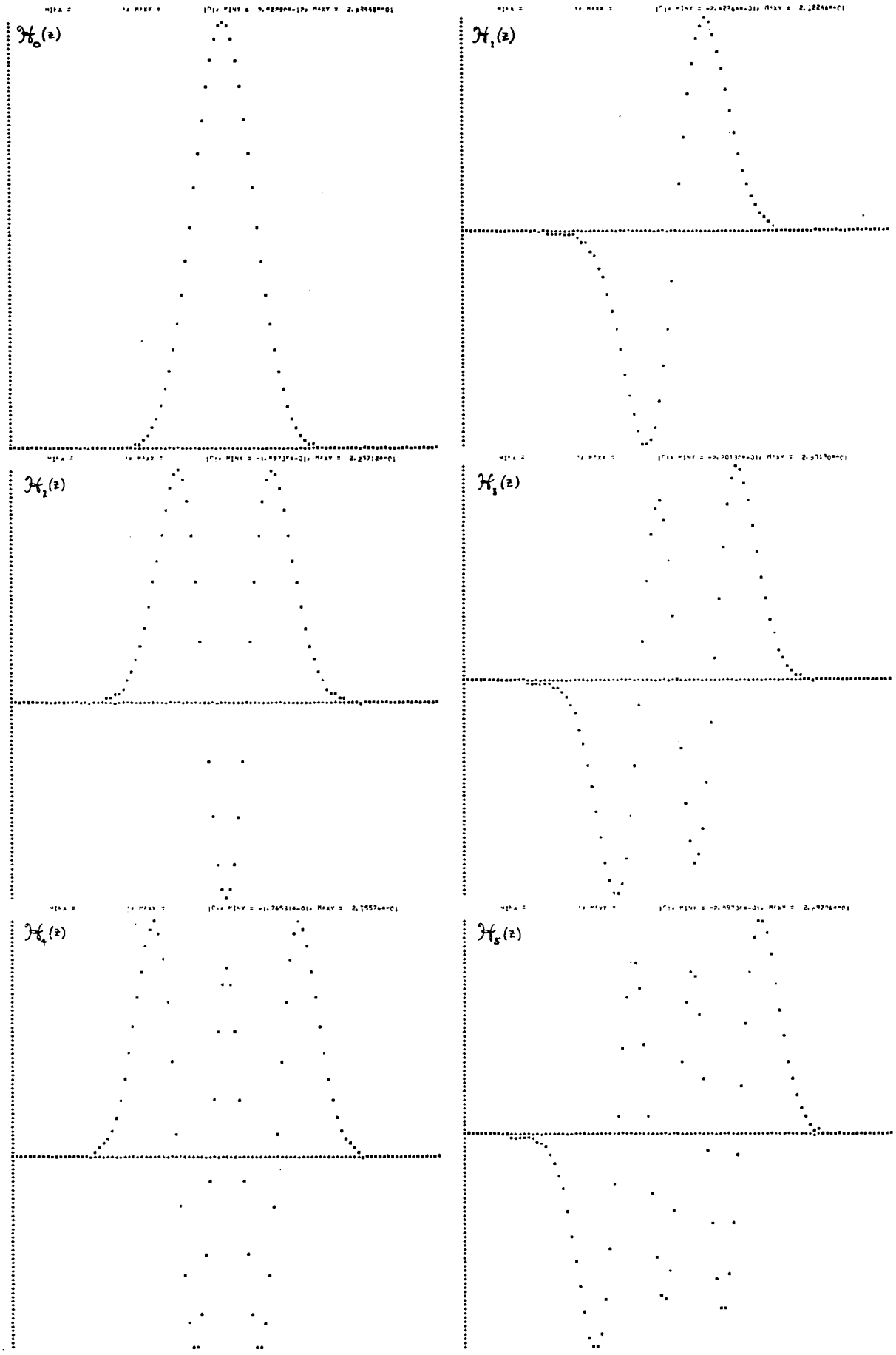

$\mathcal{H}_0(z)$
$\mathcal{H}_1(z)$
$\mathcal{H}_2(z)$
$\mathcal{H}_3(z)$
$\mathcal{H}_4(z)$
$\mathcal{H}_5(z)$

Figure # IV-3

ACKNOWLEDGEMENTS

This work was supported by the United States Atomic Energy Commission under contract AT-(40-1)-2572 and by the Stanford Research Institute. I am deeply indebted to the staff of the Rice University Computer Project and its Director, Walter Orvedahl for all the support which was given during this research. For his advice and consultation on linguistic problems I give thanks to Dr. Joseph Wilson. To the patient members of my thesis committee, Dr. Frank Huband, Dr. Tom Parks, and Dr. Howard Resnikoff my tireless advisor, I give my sincere thanks for their counsel and encouragement.

REFERENCES

1. Lindgren, Nilo, "Machine Recognition of Human Language, Part I," IEEE Spectrum, March 1965, pp. 114-136.

2. King, J. H. Jr., and C. J. Tunis, "Some Experiments in Spoken Word Recognition," IBM Journal, January 1966, pp. 65-79.

3. Reddy, D. R., "An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave," Technical Report No. CS47 (Ph.D. Thesis) Computer Science Department, Stanford University, September 1966.

4. Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," Psychological Review, Vol. 74, No. 6, 1967, pp. 431-461.

5. Gentleman, W. W., and G. Sande, "Fast Fourier Transforms for Fun and Profit," 1966 Fall Joint Computer Conference, AFIPS Proceedings, Vol. 29, 1966, pp. 563-578.

6. Schroeder, M. R., "Vocoders: Analysis and Synthesis of Speech," Proc. IEEE, Vol. 54, No. 5, May 1966, pp. 720-734.

7. Bremermann, H. J., "Pattern Recognition, Functionals, and Entropy," Technical Report, Department of Mathematics, University of California at Berkeley, August 1967.

8. Hurd, W. J., "Statistical Properties of Speech, Music, and Noises, and the Detection of Speech in Noise," USCEE Report No. 185 (Ph.D. Thesis) University of Southern California, December 1966.

9. Fletcher, H., Speech and Hearing in Communication, Van Nostrand Co. Inc., New York, 1953.

10. Kučera, Henry K., and Francis W. Nelson, Computational Analysis of Present-Day American English, Brown University Press, Providence, R. I., 1967.

11. Potter, R., G. Kopp, and H. Kopp, Visible Speech, Dover Inc., New York, 1966.

12. Blackmann, R. B., and J. W. Tukey, The Measurement of Power Spectra, Dover Inc., New York, 1958.

13. Arsac, J., Transformation de Fourier et Théorie des Distributions, Dunod, Paris, 1961.

14. Cooley, J. W., and J. W. Tukey, "An Algorithm for Machine Calculation of Complex Fourier Series," Mathematics of Computation, Vol. 19, No. 90, April 1965, pp. 297-301.

15. Collatz, L., Functional Analysis and Numerical Mathematics, Academic Press, New York, 1966.

16. Peterson, E., "Frequency Detection and Speech Formants," Journal Acous. Soc. Am., Vol. 23, No. 6, November 1951, pp. 668-674.

17. Chang, S., "Two Schemes of Speech Compression System," Journal Acous. Soc. Am., Vol. 28, No. 4, July 1956, pp. 565-572.

18. Rice, S. O., "Mathematical Analysis of Random Noise," Bell System Tech. Journal, Vol. 23, 1944, pp. 287-292 and Vol. 24, 1945, pp. 51-67.

19. Chang, S., "Representations of Speech Sounds and Some of Their Statistical Properties," Proc. IRE, Vol. 39, February 1951, pp. 147-153.

20. Bond, F. E., and C. R. Cahn, "On Sampling the Zeros of Bandwidth Limited Signals," IRE Trans. on Inf. Theory, IT-4, September 1958, pp. 110-113.

21. Cherry, E. C., and V. J. Phillips, "Some Possible Uses of SSB Signals in Formant Tracking Systems," Journal Acous. Soc. Am., Vol. 33, No. 8, August 1961, pp. 1067-1077.

22. Licklider, J.C.R., and I. Pollack, "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech," Journal Acous. Soc. Am., Vol. 20, No. 1, January 1948, pp. 42-51.

23. Licklider, J.C.R., "The Intelligibility of Amplitude-Dichotomized Time-Quantized Speech Waves," Journal Acous. Soc. Am., Vol. 22, No. 6, November 1950, pp. 820-823.

24. Scarr, R.W.A., "Zero Crossings as a Means of Obtaining Spectral Information in Speech," Proc. 1967 Conference on Speech Comm. and Processing, MIT pp. 232-238.

25. Reddy, D.R., "Segmentation of Speech Sounds," Journal Acous. Soc. Am., Vol. 40, No. 2, August 1966, pp. 307-312.

26. Szegö, G., Orthogonal Polynomials, Am. Math. Soc., Spauling-Moss, Boston, 1959.

27. Krawtchouk, M., "Sur une Généralization des Polynomes d' Hermite," Comptes Rendus de l'académie des Sciences Paris, Vol. 189, 1929, pp. 620-622.

28. Hildebrand, F. B., Introduction to Numerical Analysis, McGraw Hill, New York, 1956.

29. Burington, R. S., and D. C. May, Handbook of Probability and Statistics with Tables, Handbook Publishers Inc., Ohio, 1953.

30. Flanagan, J. L., Speech: Analysis, Synthesis, and Perception, Springer, New York, 1965.

31. Speaks, C., and L. Benitez, "Temporal Characteristics of Palatal Movement," Journal Acous. Soc. Am., Vol. 44, No. 1 (A), July 1968, p. 354.

32. Resnikoff, H. L., and G. A. Sitton, "A New Type of Hearing Aid," The Rice Review, November 1968.

33. Resnikoff, H. L., and G. A. Sitton, "Linguistic Segmentation of Acoustic Speech Waveforms," Journal Acous. Soc. Am., Vol. 44, No. 1 p. 366 (abstract) July 1968, and AEC report, ORO 2572-15, May 1968.