PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification [4] : G10L 5/00	A1	(11) International Publication Number: WO 87/ 02816
		(43) International Publication Date: 7 May 1987 (07.05.87)

(21) International Application Number: PCT/US86/02313

(22) International Filing Date: 29 October 1986 (29.10.86)

(31) Priority Application Number: 792,965

(32) Priority Date: 30 October 1985 (30.10.85)

(33) Priority Country: US

(71) Applicant: CENTRAL INSTITUTE FOR THE DEAF [US/US]; 818 South Euclid Avenue, Saint Louis, MO 63110 (US).

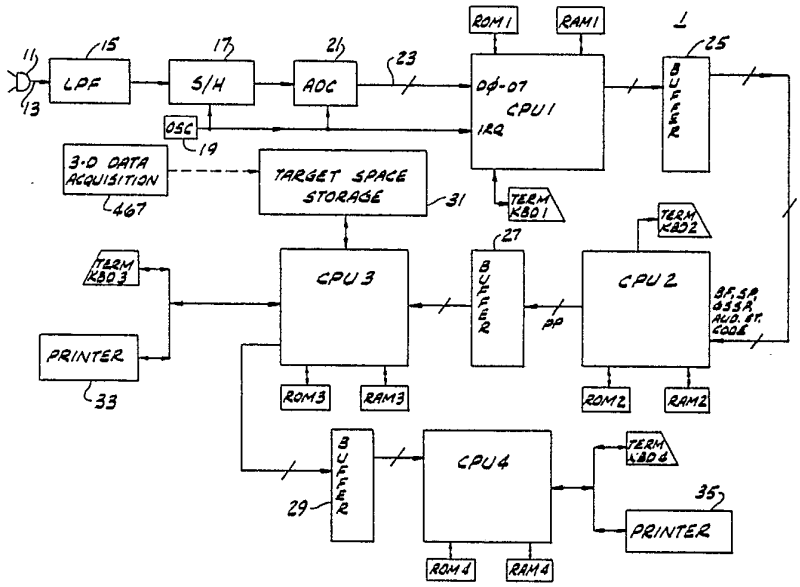
(72) Inventor: MILLER, James, D. ; 818 South Euclid Avenue, Saint Louis, MO 63110 (US).

(74) Agents: SENNIGER, Stuart, N. et al.; Senniger, Powers, Leavitt and Roedel, 611 Olive Street, Saint Louis, MO 63101 (US).

(81) Designated States: DE (European patent), FR (European patent), GB (European patent), JP, NL (European patent), SE (European patent).

Published
With international search report.

(54) Title: SPEECH PROCESSING APPARATUS AND METHODS



(57) Abstract

Speech processing apparatus (1) including a memory (31) for holding prestored information (PHE) indicative of different phonetic representations corresponding to respective sets of addresses (ADR) in the memory (31). Circuitry (CPU1, CPU2 and CPU3) in the apparatus (1) electrically derives a series of coordinate values (Xp, Yp, Zp) of points on a path in a mathematical space from frequency spectra (D(K)) of the speech occuring in successive time intervals respectively, identifies coordinate values (Xp, Yp, Zp) approximating at least one position along the path of a peak (455) in magnitude of acceleration, generates a memory address (ADR) as a function of the position coordinate values (Xp, Yp, Zp) and obtains from the memory (31) the phonetic representation information (PHE) prestored at that memory address (ADR). Methods and other apparatus for speech processing are also disclosed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**FR**	France	**ML**	Mali
AU	Australia	**GA**	Gabon	**MR**	Mauritania
BB	Barbados	**GB**	United Kingdom	**MW**	Malawi
BE	Belgium	**HU**	Hungary	**NL**	Netherlands
BG	Bulgaria	**IT**	Italy	**NO**	Norway
BJ	Benin	**JP**	Japan	**RO**	Romania
BR	Brazil	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SN**	Senegal
CH	Switzerland	**LI**	Liechtenstein	**SU**	Soviet Union
CM	Cameroon	**LK**	Sri Lanka	**TD**	Chad
DE	Germany, Federal Republic of	**LU**	Luxembourg	**TG**	Togo
DK	Denmark	**MC**	Monaco	**US**	United States of America
FI	Finland	**MG**	Madagascar		

SPEECH PROCESSING APPARATUS AND METHODS

Background of the Invention

The present invention relates to speech processing apparatus and methods. More particularly, the present invention relates to apparatus and methods for use in automatic speech recognition applications and research.

Speech, as it is perceived, can be thought of as being made up of segments or speech sounds. These are the phonetic elements, the phonemes, of a spoken language and they can be represented by a set of symbols, such as International Phonetic Association symbols.

These segments are linguistic units and have their bases in speech as it is perceived and spoken. All of the syllables and words of a language are made up of a relatively small number of phonetic elements. For example, in the case of English, textbooks in phonetics may list as few as 25 consonants and 12 vowels for a total of 37 phonemes. If the finer phonetic distinctions are included, then the list of distinguishable speech sounds or phones may lengthen to as high as 50 or 60.

It has been proposed that the phonemes of a spoken language can be understood in terms of a small set of distinctive features numbering about 12. These features have their bases in articulatory, perceptual, and linguistic analyses. A feature approach is often used in textbooks on phonetics as the phones and phonemes are described in terms of place of articulation and manner of articulation.

There are several theories of how the human listener processes an incoming acoustic waveform of speech and translates that waveform into a series of linguistic elements such as phonemes or words. The exact mechanisms and processes involved in the perception of speech are not yet fully understood. Finding simple and reliable acoustic-auditory correlates of the phones, phonemes and presumed features has proved elusive.

Research on speech perception has led to complicated, highly conditioned statements of relations between acoustic-auditory patterns and perception of phonemes, and even these statements are often of narrowly circumscribed generality. For example, the problem of how the listener can divide the acoustic input into segments relevant to linguistic perception is not understood. Even if a solution of this segmentation problem were available, the auditory-acoustic expression of a phoneme or feature seems to depend on the phonetic context, the particular talker, and the rate of speaking.

As a result of these problems there are several viable theories of speech perception. All of the current theories can be cast into a generic three-stage model, with the acoustic input undergoing three stages of processing in a bottom-up sequence. Stage 1 is an auditory-sensory analysis of the incoming acoustic waveform whereby representation of the signal is achieved in auditory-sensory terms. Stage 2 is an auditory-perceptual transformation whereby the spectral output of stage 1 is transformed into a perceptual form relevant to phonetic recognition. Here the spectral descriptions are transformed into dimensions more directly relevant to perception. For example, in various theories the perceptual form may be related to articulatory correlates of speech production or auditory features or pattern sequences. Finally, there is stage 3 in which the perceptual dimensions of stage 2 are transformed by a phonetic-linguistic transformation into strings of phonemes, syllables, or words. Stages 2 and 3 also are influenced by top-down processing wherein stored knowledge of language and events and recent inputs, including those from other senses as well as language, are brought into play.

Some work in automatic speech recognition has involved a narrow-band spectral analysis performed on a time-windowed speech waveform. In one system described in "Recognizing continuous speech remains an elusive goal" by R. Reddy et al., IEEE

Spectrum, Nov., 1983, pp. 84-87, incoming digitized signals are broken into centisecond slices and spectrally analyzed. Each slice is compared with a collection of sound prototypes and the prototype closest to each slice is entered into a sequence. The prototype sequence is then used to roughly categorize the initial sound of the word, which in turn is used to produce word hypotheses. Each word is then tested by creating a probability matrix and a cycle of operation repeats for the next word until an entire sentence is identified.

Summary of the Invention

Among the objects of the present invention are to provide improved speech processing apparatus and methods which process speech occurring at different rates; to provide improved speech processing apparatus and methods which usefully process speech from different talkers; to provide improved speech processing apparatus and methods which segment the speech to distinguish phonetic elements therein; to provide improved speech processing apparatus and methods which recognize phonetic elements in speech sounds in which there is apparent acoustic overlap of phones; to provide improved speech processing apparatus and methods which integrate cues such as silence and transitions; to provide improved speech processing apparatus and methods which recognize burst and stop phonetic elements; to provide improved speech processing apparatus and methods which can accomplish phonemic restoration to recognize a phoneme as if it occurred when a speech sound usually associated with the phoneme has not actually occurred; to provide improved speech processing apparatus and methods which recognize phonetic elements in the speech of talkers of different dialects and languages; to provide improved speech processing apparatus and methods for converting speech into symbols for storage and display; and to provide improved speech processing apparatus and methods for

converting speech into a string of phonetic elements for use in generating written text material corresponding to the speech.

Other objects and features will be in part apparent and in part pointed out hereinafter.

In a form of the present invention, speech processing apparatus includes a memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory and circuitry for electrically deriving a series of coordinate values of points on a path in a mathematical space from frequency spectra of the speech occurring in successive time intervals respectively, for identifying coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the position coordinate values and obtaining from said memory means the phonetic representation information prestored at that memory address.

In another form of the present invention, speech processing apparatus includes circuitry for producing samples of an analog waveform of speech and converting the samples into digital form. Also included is circuitry for deriving sets of digital values representative of frequency spectra of the speech from the samples in digital form, for generating one of a plurality of auditory state codes for each of the sets of digital values and producing a series of pointer values in a mathematical space, which pointer values are determined from the respective sets of digital values, and for computing from the series of pointer values a series of coordinate values of points on a path in the mathematical space by one of a plurality of different computational processes depending on which auditory state code is generated. A further circuit temporarily stores in digital form the computed coordinate values of the points on the path.

In a method form of the present invention, a method of processing speech includes the steps of electrically deriving a series of coordinate values of points in a mathematical space from the frequency spectra of the speech occurring in successive time intervals respectively, the series of coordinate values defining a path of the points in the mathematical space, and electrically identifying coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the coordinate values of the position on the path and obtaining from a memory means, having prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory, the phonetic representation information prestored at that memory address.

In a further form of the invention, speech processing apparatus includes a memory circuit and circuitry for producing samples of an analog waveform of speech and converting the samples into digital form. Further circuitry derives a set of digital values representative of a frequency spectrum of the speech from the samples in digital form, selectively stores in distinct locations in the memory the values of frequency of one or more frequency peaks in the spectrum wherein a selected one or more of the distinct memory locations in which the frequency value of a given peak is stored depends on whether the peak lies in a first predetermined band of frequencies and on whether or not any other peak lies both in the first band and a second band overlapping the first band, and generates a set of digital values corresponding to coordinate values in a mathematical space depending both on the stored values of frequency and on the distinct locations of the stored values of frequency.

Brief Description of the Drawings

Fig. 1 is a block diagram of a speech processing apparatus of the present invention;

Fig. 2 is a graph of voltage versus time of a typical speech waveform;

Fig. 3 is a diagram of operations of an interrupt routine of a unit CPU1 of Fig. 1;

Fig. 4 is a diagram of operations of a main routine of CPU1 of Fig. 1 according to a method of the invention;

Fig. 5 is a graph of amplitude versus log-frequency of a ten-millisecond sample of the speech waveform of Fig. 2, showing a frequency spectrum thereof;

Fig. 5A is a diagram of a table in a memory for CPU1 for holding a set of spectral values corresponding to multiples K of a basic frequency;

Figs. 6, 7, 8 and 9 are a set of graphs of spectral envelopes in decibels versus log-frequency for illustrating a method of the invention for analyzing different frequency spectra of speech;

Fig. 10 is a diagram of three spectral envelopes in decibels versus log-frequency for showing how a quantity called speech goodness depends on shapes of spectra;

Fig. 11 is a graph of speech goodness versus width of one or more peaks in a spectrum for illustrating operations of apparatus of the invention according to a method of the invention;

Fig. 12 is a graph of a quantity called speech loudness versus a decibel sum resulting from operations of apparatus of the invention according to a method of the invention;

Figs. 13A and 13B are two parts of a diagram further detailing operations in the main routine of Fig. 4 according to inventive methods for analyzing spectra by the apparatus of the invention;

Fig. 14 is a diagram of operations according to a method of the invention for generating a spectral reference value;

Fig. 15 is a diagram of operations according to an inventive method in a unit CPU2 of Fig. 1 for converting from sensory pointer coordinate values to coordinate values on a path having perceptual significance;

Fig. 15A is a diagram of a table for use by CPU2 in the operations of Fig. 15;

Fig. 16 shows an illustration of a mathematical model for converting from sensory pointer coordinates to coordinates X_p, Y_p and Z_p of a perceptual pointer in a three-dimensional mathematical space;

Fig. 17 is a simplified diagram of the mathematical space of Fig. 16, showing target zones for two phonetic elements, and showing a trajectory or path traced out by the perceptual pointer in the mathematical space;

Fig. 18 shows an X,Y,Z coordinate system and an X',Y',Z' coordinate system in the mathematical space;

Figs. 19 and 20 show two different views of a vowel slab with target zones for the vowels in the mathematical space relative to the X',Y',Z' coordinate system of Fig. 18 and viewing along the X' axis in Fig. 19 and along the Z' axis in Fig. 20;

Fig. 21 depicts target zones in the mathematical space for voiceless stops as viewed along the Y axis of Fig. 18;

Fig. 22 depicts target zones in the mathematical space for voiced stops and unaspirated voiceless stops and nasal consonants as viewed along the Y axis of Fig. 18;

Fig. 23 depicts target zones in the mathematical space for voiceless fricatives of American English as viewed along the Y axis of Fig. 18;

Fig. 24 depicts target zones in the mathematical space for voiced fricatives and the phonetic approximates as viewed along the Z' axis of the X', Y', Z' coordinate system of Fig. 18;

Fig. 25 depicts target zones in the mathematical space for the voiced fricatives and the phonetic approximates of Fig. 24 as viewed along the X' axis of the X', Y', Z' coordinate system of Fig. 18;

Fig. 26 is a diagram of inventive operations of a CPU3 of Fig. 1 of the inventive apparatus in analyzing the path in the mathematical space and obtaining phonetic elements when phonetically significant events occur; and

Fig. 27 is a diagram of a table for use in the operations of Fig. 26.

Corresponding reference characters indicate corresponding parts throughout the several views of the drawings.

Detailed Description of Preferred Embodiments

In Fig. 1 a speech processing system 1 of the invention has a microphone 11 for converting sound pressure variations of an acoustic waveform of speech to an analog electrical signal on a line 13. System 1 performs a short-term analysis on the speech waveform that allows it to represent, every few milliseconds, the spectral shape and the auditory state of the incoming speech. This sensory processing serves as an input to a higher level perceptual electronic system portion. The perceptual electronic system portion integrates sensory information over time, identifies auditory-perceptual events (or "sounds"), and converts the sensory input into a string of symbols or category codes corresponding to the phonetic elements of a human language.

The electrical signal on line 13 is filtered by an antialiasing low pass filter 15 and fed to a sample-and-hold (S/H) circuit 17. S/H circuit 17 is enabled by an oscillator 19 at a sampling frequency such as 20 KHz. and supplies samples of the analog electrical signal to an analog-to-digital converter

(ADC) 21 where the samples are converted in response to oscillator 19 to parallel digital form on a set of digital lines 23 connected to data inputs of a first central processing unit CPU1. CPU1 reads in the latest sample in digital form upon interrupt by oscillator 19 at interrupt pin IRQ every 50 microseconds.

CPU1 is one of four central processing units CPU1, CPU2, CPU3 and CPU4 in Fig. 1, which respectively have programmable read only memory (ROM1, ROM2, ROM3 and ROM4), random access memory (RAM1, RAM2, RAM3 and RAM4), and a video terminal-keyboard unit (TERMKBD1, TERMKBD2, TERMKBD3, and TERMKBD4). CPU1 generates data for CPU2 which is buffered by a data buffer 25. CPU2 generates data for CPU3 which is buffered by a data buffer 27, and CPU3 generates data for CPU4 which is buffered by a data buffer 29. CPU3 has a memory 31 of approximately 2 megabyte capacity that holds prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory. CPU3 is provided with a printer 33 for recording phonetic element information in the order obtained by it from memory 31. CPU4 is a lexical access processor for converting the phonetic element information into plain-text and printing it out on a printer 35 to accomplish automatic dictation.

Fig. 2 shows a portion of an electrical waveform 51 of speech. The waveform 51 generally has several peaks and troughs over a time interval, or window, of about ten milliseconds, as well as higher frequency behavior. CPU1 is interrupted 20000 times per second so that in each ten millisecond time interval a set of 200 samples is obtained from ADC 21.

In Fig. 3 operations of an interrupt routine 70 of CPU1 commence upon interrupt at pin IRQ with a BEGIN 71 and proceed to a step 73 to read the latest sample into an address location in a section of N1 (e.g. 80) addresses in RAM1. Then

in a step 75 both the address and a sample count N are incremented by one. In a decision step 77, the count N is compared with the number N1 to determine if the latest set of samples is complete. If so, then in a step 79 the sample count N is returned to zero and a flag FLG is set to 1 as a signal that the latest set of samples is complete. Also, the address location for the next sample is reset to a predetermined location ADR0 at the beginning of the section of N1 addresses, whence a RETURN 81 is reached. If the latest set of samples is not complete, the operations branch from step 77 to RETURN 81 whence a main program resumes in CPU1 at an operation where the interrupt occurred.

The operations of CPU1 according to its main program are shown in Fig. 4 commencing with a START 101 and input-output housekeeping and initialization at a step 103. Also in step 103, sample set size N1 is set to 80, and a sample flag FLG and a variable FIL are set to zero. A sensory reference frequency SR and a variable GMTF0 are both initialized to 168 Hertz, a constant close to the geometric mean pitch of the human speaking voice. A variable N2 is set to 100. Then at a step 105, a set of variables or quantities herein called an auditory state code, as well as a set of frequency values SF1L, SF1H, SF2 and SF3, are all initialized to zero. The variables in the auditory state code of the present embodiment are: burst-friction BF, glottal-source GS, nasality NS, loudness indices LIBF and LIGS for burst-friction and glottal-source sounds respectively, and speech goodness values GBF and GGS for burst-friction and glottal-source sounds respectively. In other embodiments, variables are included in the auditory state code for some or all of a variety of source characteristics of speech including nasality, voicing, frication, aspiration, whisper, loudness and goodness.

Next in a step 107, the flag FLG is checked to confirm that a full set of N1 samples is available. The interrupt operations of Fig. 3 are collecting the next set of N1 samples as the operations of Fig. 4 are executed. If the system 1 has just been turned on, CPU1 will wait until the first set of samples has been obtained and FLG has been set to 1 in the interrupt routine, which wait occurs by a branch from step 107 back to itself. When FLG becomes one, a full set of samples is present and FLG is reset to zero in a step 109. Next in a step 111 a set of digital values representing a frequency spectrum corresponding to the latest N1 samples from ADC 21 is computed according to a Discrete Fourier Transform (DFT) procedure. In other words each such set of digital values represents the frequency spectrum of the speech in each successive ten millisecond interval.

An example frequency spectrum is depicted by vertical lines 113 of Fig. 5. The frequency spectrum is computed as follows. The digital values of the samples are designated S(N) where N goes from zero to N1-1. Their DFT is given by the equation

$$D(Kf) = \sum_{N=0}^{N1-1} S(N)e^{-j2\pi KN/N1} \qquad (1)$$

where e is the base of natural logarithms, j is the square root of minus one, and pi is the ratio of circumference to diameter of a circle. f is a basic frequency equal to the reciprocal of the time required to collect a set of N1 samples (when time is 10 milliseconds, f is 100 Hertz) and Kf is an integer multiple of the frequency f at which one of the lines 113 in the spectrum is to be computed. CPU1 computes the DFT by the Fast Fourier Transform algorithm familiar to the art for frequency multiples K from 1 to a number M. The number M is half the sampling rate

times the time required to collect a set of the N1 samples (20000Hz.x0.5x0.01sec.=100).

The values of D(Kf) are stored as illustrated in Fig. 5A in a spectrum table in RAM at successive addresses corresponding to the K values respectively.

In another method for deriving the spectral envelope of the speech waveform, the speech waveform is multiplied by time-window weighting functions of 5-40 millisecond duration but shifted in 1.0-2.5 millisecond steps. Thus the successive time intervals defining the windows can be either overlapping or distinct. The window duration and step size as related to bursts, transitions and relatively steady-state segments are adjusted for best performance. The short-term spectrum is calculated for each segment by either DFT or linear prediction analysis (LPA). The DFT, of course, produces a line spectrum with components at integral multiples of the reciprocal of the window length while the LPA produces a smoothed spectral envelope--transfer function--with detail dependent on the number of LP-parameters selected. Either spectrum is represented in log-magnitude by log-frequency dimensions. Operations accomplish or approximate the following. The spectrum is "windowed" in the log frequency domain so that the amplitudes are represented in sensation levels or loudness levels. The spectrum is subjected smoothing filters one of which is similar to the critical-band. Another minimizes confusing minor spectral peaks. Finally, the spectral envelope is subjected to high-pass filtering in the log-frequency domain to eliminate spectral tilt. The resulting spectra preferably have formant peaks of nearly uniform height--tilt having been removed and have minor irregularities removed by the smoothing filters. A nasal wave can be detected in the lower half of the speech spectrum by looking for a weakened and broadened first formant, or to window the processed spectral envelope in the appropriate range of log frequency units and band-pass filter that segment in search of the

nasal wave, or to use correlational signal processing techniques.

In an another alternative embodiment, a real time filter bank circuit is used to produce the spectrum for CPU1. Such a filter bank advantageously reduces the computing required of CPU1, and in such embodiment the spectrum table is updated from the real time filter bank at regular intervals such as every ten milliseconds or even more frequently, for example every 1-2.5 milliseconds. Also, signal processing chips for inexpensively and rapidly computing spectra are available such as the Texas Instruments TMS 320.

In Fig. 5 the spectrum has several peaks 115, 116 and 117 which decline in height or "tilt" with increasing frequency. To facilitate the description an envelope 119 is drawn on Fig. 5 which envelope has the same peaks 115, 116 and 117. Envelope 119 is redrawn dashed in Fig. 6 with the spectral lines 113 being understood but suppressed in Fig. 6 for clarity. CPU1 in a step 121 of Fig. 4 converts the spectrum to decibels (dB) of sensation level according to the equation

$$D(Kf)_{dB} = 20 \log_{10} D(Kf/ref) \qquad (2)$$

where D(Kf) is each spectral value at frequency Kf, and ref is normal human threshold for that frequency in sound pressure.

The spectrum is smoothed by sliding a critical-band-like weighting function along the log-frequency or pitch-like axis, and spectral tilt or "combing" is also eliminated by passing the smoothed spectrum through a high-pass lifter defined in the log-frequency or pitch-like domain. The resulting smooth envelope is rectified (straightened) to eliminate low-level excursions, including those some fixed number of decibels below the highest spectral peaks as well as those below the threshold of hearing, since these are irrelevant to phonetic perception.

The processed spectral envelope is tested for the presence, location and strength of the nasal wave. After determination of nasalization, which can be removed by further spectrum processing in some embodiments, the spectral envelope is examined for low and high frequency cutoffs and significant spectral prominences.

In a step 123 the tilt suggested by a dashed line 125 in Fig. 6 is eliminated from the spectrum by adding values to the spectrum that increase with frequency at a rate of C dB per ten-fold increase in frequency. The value of the constant C is determined using a linear regression analysis of the spectrum. By eliminating the tilt from the spectrum, a relatively flat spectral representation is achieved wherein energy prominences have approximately the same amplitude. The values for high-pass lifting to eliminate tilt are determined from the equation

$$D1 = C \times \log K \qquad (3)$$

In words, to eliminate the tilt, each of the M values (where M is illustratively 40) of the spectrum in decibels is respectively added to a corresponding value computed according to equation (3) for each K from 1 to M. The resulting spectrum is suggested by an envelope 127 of Fig. 6 having three peaks P1, P2 and P3 in order of increasing frequency.

The above-described short-term spectral analysis of the time-windowed speech waveform identifies the amplitudes and frequencies of tonal components in the speech waveform and at the same time produces a power spectrum of any significant aperiodic energy or other unresolved high-frequency components in the speech waveform. This information is used to distinguish aperiodic, periodic, and mixed segments and to establish an effective lower frequency F0 or low pitch, of the periodic and mixed segments. This same short-term spectral information

undergoes further processing to generate auditory-spectral patterns that can be called sensory-excitation patterns, auditory-sensory spectra, or auditory-spectral envelopes.

Voice pitch plays a role in the identification of voiced phonetic segments such as vowels like a, e, i, o and u. Detection of aperiodic energy in speech is very important for the recognition of aspiration sounds as in /h/, /p/, /k/ and /t/ and of fricatives such as /s/ and /f/ and so on. Voiced fricatives such as /z/, /zh/ and /v/ have a mixture of periodic and aperiodic energy and are a combination of both glottal-source and burst-friction spectra.

Figs. 7, 8 and 9 show envelopes illustrating different types of spectra associated with different types of speech sounds. These spectra have different numbers and shapes of prominences, or peaks, at different frequencies compared with envelope 127 of Fig. 6. Clearly the spectra resulting from steps 111, 121 and 123 of Fig. 4 can vary widely as different sets of speech samples are processed by CPU1.

To characterize these spectra with relatively few variables, each latest spectrum is analyzed in a step 131. In this step, three spectral frequencies SF1, SF2 and SF3 are computed. The spectral frequencies SF1, SF2 and SF3 are in some cases the frequencies at which peaks occur such as P1, P2 and P3 in Fig. 6, and the manner of determining them is described more specifically in connection with Figs. 13A and 13B hereinafter. Distinct lower and higher values SF1L and SF1H are computed for SF1 when nasality is present. A spectral frequency reference SR is also computed to indicate the overall general pitch (timbre) of the speech so that voices with high pitch (timbre) and voices with low pitch (timbre) are readily processed by the system 1. Also in step 131 auditory state code quantities BF, GS, NS, LIGS, LIBF, GGS and GBF are determined from the spectrum.

Next in a decision step 133 the speech goodness values GGS and GBF are tested and the loudness index values LIGS and

LIBF are tested, and if none is positive, operations branch to a step 135. In step 135 a set of registers in CPU1 (corresponding to a set of three coordinates called sensory pointer coordinates X_s, Y_s and Z_s) are loaded with a code "*" indicating that the coordinates are undefined. Then in a step 137 the contents of the registers for X_s, Y_s and Z_s are sent to CPU2 through buffer 25 of Fig. 1. If in decision step 133 the speech goodness is positive, operations proceed to a step 143 where sensory pointer coordinate value X_s is set equal to the logarithm of the ratio of SF3 to SF2, pointer value Y_s is set equal to the logarithm of the ratio of SF1L to SR, and pointer value Z_s is set equal to the logarithm of the ratio of SF2 to SF1H, whence step 137 is reached. The equations of step 143 are computed once except when glottal source and burst friction spectra are simultaneously present, as in voiced fricatives, in which case step 143 is executed twice to compute sensory pointer coordinates X_{gs}, Y_{gs}, Z_{gs} for the glottal source spectrum and X_{bf}, Y_{bf}, Z_{bf} for the burst-friction spectrum.

After sensory pointer coordinate values X_s, Y_s and Z_s are sent to CPU2 in step 137, the auditory state code quantities BF, GS, NS, LIGS, LIBF, GGS and GBF are also sent in a step 145 to CPU2 through buffer 25. Then in a step 147, a test is made to determine if an OFF-ON switch is on, and if not, operations terminate at END 149. If the switch is on, as is normal, operations loop back to step 105 for obtaining the next spectrum, analyzing it and sending information to CPU2 as described above. CPU1 thus executes operations continually to obtain spectral information about the samples of speech as they arrive in real time.

The auditory-spectral pattern at any moment in time is given by the auditory-spectral envelope in dB (Phons or Sensation Level or equivalent) against log frequency, as shown in Fig. 5. After appropriate processing of this envelope, the frequency values of SR, SF1, SF2 and SF3 are found for the vocalic portions of speech. Vocalic portions are those segments or

spectral components that ordinarily result from an acoustic source at the glottis and have the vocal tract, with or without the nasal tract, as a transmission path to the external air. Thus, voiced speech, which has periodic spectra, and whispers or aspirated sounds, which have aperiodic spectra, are the vocalic components of speech and have spectra called glottal-source (GS) spectra herein. One sign of the presence of a vocalic segment is a low-frequency prominence (P1) that can be associated with a first resonance of the vocal tract.

A sensory pointer for vocalic portions of speech has a position in a mathematical space, or phonetically relevant auditory-perceptual space, computed in step 143 of Fig. 4. This pointer is called a glottal-source sensory pointer (GSSP). Usually SF1, SF2 and SF3 are the center frequencies of the first three spectral prominences in the auditory-spectral envelope 127 of Fig. 6. Sometimes, however, SF3 is interpreted as the upper edge of the spectral envelope when no clear peak P3 can be observed, such as when peaks P2 and P3 merge during a velar segment or is taken as being a fixed logarithmic distance over SR when P3 is absent. Spectral frequency SF1 generally corresponds to the center frequency of the first significant resonance of the vocal tract. However, during nasalization two peaks, or one broadened peak, appear near the first significant resonance, as in Figs. 9 and 8 respectively. To take account of such spectral differences steps 131 and 143 of Fig. 4 are made sufficiently flexible to compute the sensory pointer position differently for nasalization spectra than for other spectra.

In another major class of spectra suggested by the envelope of Fig. 9, there is no major prominence in the area of peak P1 of Fig. 6. In other words, the latter two of the three prominences of Fig. 6 may occur without the first prominence in this class of spectra. Such spectra are associated with burst sounds and sustained friction sounds and are produced by a talker with supraglottal sources such as when the tongue meets or

approximates the velum, palate, or teeth or at the teeth and lips, themselves. These spectra are referred to as burst-friction (BF) spectra herein. A BF spectrum is analyzed differently from a GS spectrum by CPU1 in order to produce the spectral frequency values SF1, SF2 and SF3 and sensory reference value SR, and the position of the resulting sensory pointer values computed in step 143 of Fig. 4 is generally in the X_s, Z_s plane. These pointer values are regarded as defining the position of a pointer called the burst-friction sensory pointer (BFSP) which is distinct from the GSSP.

As the incoming speech is analyzed in step 131 of Fig. 4, the glottal-source GS value is set to 1 in the auditory state code whenever a glottal-source spectrum is above the auditory threshold. As the values of SR, SF1, SF2, and SF3 change, the GSSP is regarded as moving through a mathematical space, or auditory-perceptual space. The path of the GSSP is interrupted by silences and by burst-friction spectra. Then the GS value is set to zero and the BF value is set to 1 in the auditory state code. In such case, the GSSP is replaced by the BFSP. The GSSP can be regarded as moving through the mathematical space as the glottal-source spectrum changes shape and sometimes this movement is nearly continuous as in the case of the sentence, "Where were you a year ago?", where the only interruption would occur during the friction burst of "g" in "ago." In other words the quantity GS in the auditory state code can remain at a value of one (1) through many spectra in various examples of speech, but the quantity BF in the auditory state code when set to one is generally reset to zero very shortly thereafter, because spectra which are not of the burst-friction type occur so soon thereafter. In terms of the mathematical space, burst-friction sensory pointer BFSP will usually appear and disappear shortly thereafter as friction sounds are inserted in the speech stream. As burst-friction spectra are unstable, the BFSP may

exhibit considerable jitter, and it usually will not move in any smooth, continuous way in the mathematical space.

Often the quantity BF in the auditory state code is 1 when the quantity GS is zero, and vice versa. However, in the case of voiced fricatives, both BF and GS are equal to one simultaneously. In terms of the mathematical space, both of the sensory pointers are simultaneously present as one is associated with the glottal-source spectrum of the voiced part of the voiced fricative speech sound and the other is associated with the burst-friction spectrum of the friction part of the sound.

CPU1 computes goodness values and loudness values in the auditory state code for the GS and BF spectra. The speech goodness is a measure of the degree to which the sound represented by the latest spectrum is like a sound of speech, and is regarded as the cross-correlation between an ideal spectrum for a given speech sound and the latest actual spectrum of that sound. Since calculation of the cross-correlation itself represents a significant computer burden, the goodness value is estimated in the preferred embodiment.

As shown in Fig. 10, the speech goodness value is low when an actual spectrum consists of a few pure tones showing up as very narrow peaks 171, 173 and 175; and the goodness value is also low when the spectrum is very broad-band with tiny bumps for peaks as in envelope 177. On the other hand, the goodness value is high for carefully produced natural speech of high fidelity, which has distinct moderately-wide prominences 181, 183 and 185 with distinct valleys between them.

The goodness value is estimated, for instance, by determining when the width of at least one of the peaks in the frequency spectrum, such as P2, is within a predetermined range. The width is illustratively defined as the difference of the nearest two frequencies higher and lower than the center frequency of the peak at which the DFT value in decibels is at least a predetermined number of decibels (e.g. 15 dB) below the

maximum decibel level of the peak itself. When more than one peak is used in the calculation, an average or weighted average peak width is suitably determined.

Then as shown in Fig. 11, the goodness value is set to zero if the width is outside the range. The goodness value when the width is in range, is a triangular function 191 which peaks at unity for a best width value and illustratively declines linearly on either side of the best value to a value of 0.25 at a width of zero and to a value of zero at an upper limit of the range.

The loudness index is estimated from the sum of the decibel levels (or total power) of the lines of a spectrum within the width of at least one (and preferably all) of the prominences or peaks, wherein the width is defined as in the previous paragraph. As illustrated by the graph of Fig. 12, this decibel sum is then compared with a value T indicative of a hearing threshold, and if the sum is less than T, the loudness index L is zero. The decibel sum is compared with a value U indicative of adequate loudness as for everyday conversational speech, and if the sum exceeds U, the loudness index L is 1. Between the levels T and U the decibel sum is converted into loudness index L by the function

$$L = (\text{dB Sum}-T)/(U-T) \qquad (4)$$

The operations performed by CPU1 in analyzing each spectrum in step 131 of Fig. 4 are now described in sequence with reference to Figs. 13A and 13B.

After a BEGIN 201, CPU1 in a step 203 finds the maximum value MAX, or highest peak, of the spectrum. This is illustratively accomplished by first setting to zero all spectral values which are less than a predetermined threshold decibel level, so that low sound levels, noise and periods of silence will not have apparent peaks. The nonzero values remaining, if

any, are checked to find the highest value among them to find the value MAX.

Then in a step 205 a loudness L is computed as discussed above in connection with Fig. 12. Next, in a step 207 a value of 15 dB is subtracted from the maximum value MAX to yield a reference level REF. In a following step 209 the level REF is subtracted from all of the M values in the DFT spectrum and all of the resulting negative values are set to zero to normalize the spectrum so that the reference line is zero dB and spectral values that fall below the reference are set to zero dB. The values in the spectrum at this point in operations are called normalized spectral values and are suggested in Fig. 6 by the portions of envelope 127 lying above the dashed horizontal line marked REF.

In a step 211 following step 209 the fundamental frequency is found by a pitch-extraction algorithm such as that of Scheffers, M.T.M. (1983). Simulation of auditory analysis of pitch; An elaboration of the DWS pitch meter." J. Acoustic Soc. Am. 74, 1716-25. (see Fig. 6) and stored as a spectral frequency SF0. Next the spectrum is analyzed in each of three frequency bands B1, B2 and B3, if the spectrum is a glottal-source spectrum, as suggested beneath Fig. 8; and otherwise analyzed in two frequency bands B2 and B3 with different numerical limits, as suggested beneath Fig. 9. These frequency bands are used as a way of discriminating the P1, P2 and P3 peaks and the frequency values selected to define each band are adjusted for best results with a variety of speaking voices.

In a decision step 213, CPU1 determines whether there are any positive normalized spectral values lying in the band B1 which is defined as 0 less than or equal to $\log_{10}(f/SR)$ less than or equal to 0.80, where SR is the spectral reference and f is frequency in Hertz. If there are no such positive normalized spectral values, it is concluded that the spectrum is a burst-friction spectrum (although this may also be a period of silence) and a branch is made to a step 215 where quantity BF is set to 1 in the auditory state code and the spectral higher and

lower frequency values SF1L and SF1H are both set equal to SR. The burst-friction loudness index LIBF is set equal to the loudness L computed in step 205. (During silence the loudness is zero, and there is no harm in subsequent operations in having BF equal 1.) The frequency band B2 is established as 0.6 less than or equal to $\log_{10}(f/SR)$ less than or equal to 1.45, and frequency band B3 is established as 1.0 less than or equal to $\log_{10}(f/SR)$ less than or equal to 1.65.

On the other hand, if in step 213 there is any positive normalized spectral value in band B1 then operations proceed to a step 217 in which CPU1 scans the normalized spectral values in order of increasing address values corresponding to frequency multiplier K until the first normalized spectral value is found which is succeeded by a lower normalized spectral value at the next higher value of K. That first modified spectral value is regarded as the lowest-frequency peak in frequency band B1 and the spectral frequency values SF1 and SF1L are set equal to the K value representing the frequency of this peak. Also in step 217 the glottal-source quantity GS is set to one in the auditory state code. The glottal-source loudness index LIGS is set equal to the loudness L computed in step 205. The frequency band B2 is established as 0.6 less than or equal to $\log_{10}$-(f/SR) less than or equal to 1.18, and frequency band B3 is established as 1.0 less than or equal to $\log_{10}(f/SR)$ less than or equal to 1.30.

Following step 217 a decision step 219 determines whether there is a second peak at a higher frequency than SF1L in frequency band B1. If so, operations branch to a step 221 where nasality NS is set to one in the auditory state code, and proceed to a step 223 where the frequency of the second peak is determined and stored at a location SF1H.

If in decision step 219 no second peak is found in band B1, operations proceed to another decision step 225 where the width of the peak is compared with a predetermined width W1

(such as 300 Hz. at 10 db down) to determine whether the peak is wider than a typical GS peak would be without nasality. If the predetermined width is exceeded, a branch is made to a step 227 where nasality NS is set to one. Also in step 227 the edges of the nasally broadened P1 peak are defined by setting the lower frequency SF1L to SF0 and the higher frequency SF1H to the frequency at the upper edge of the P1 peak where a normalized spectral value again is zero. If the predetermined width W1 is not exceeded in step 225, however, operations proceed to a step 229 where the value SF1H is set equal to SF1L because there is only P1 peak and no nasality.

Operations of CPU1 proceed from any of the steps 215, 223, 227 or 229 in Fig. 13A through a point X to a decision step 231 of Fig. 13B. In step 231 CPU1 tests the normalized spectral values to determine whether there is a peak P2 in band B2 above the peak having value SF1H. Band B2 is already established to correspond with the BF or GS nature of the spectrum. The testing begins above value SF1H if SF1H lies in band B2, to avoid confusing the peak sought with a peak found earlier. If a peak P2 exists, then operations proceed to a step 233 where second spectral frequency value SF2 is set to the frequency K value of the first peak above frequency SF1H in band B2, and a decision step 237 is reached. If there is no peak found in step 231, operations branch from step 231 to a decision step 238 where the value of SF1H is tested to determine whether it is in the band B2. If not, operations branch to a step 239 where the value SF2 is set equal to SF1H and SF1H is not affected, whence operations reach step 237. If in decision step 238, the value of SF1H is in band B2 then operations proceed to a step 240 where the value SF2 is set equal to SF1H. Also, in step 240 SF1H is set equal to value SF1L and the nasality NS is reset to zero because nasality is not regarded as being present after all. Operations then pass from step 240 to step 237.

In this way, means are provided for deriving a set of digital values representative of a frequency spectrum of the speech from the samples in digital form, for selectively storing in distinct locations in the memory the values of frequency of one or more frequency peaks in the spectrum wherein a selected one or more of the distinct memory locations in which the frequency value of a given peak is stored depends on whether the peak lies in a first predetermined band of frequencies and on whether or not any other peak lies both in the first band and a second band overlapping the first band, and for generating a set of digital values corresponding to coordinate values in a mathematical space depending both on the stored values of frequency and on the distinct locations of the stored values of frequency.

In addition, means are thus provided for selecting values of end frequencies for both the second band and a third band overlapping the second band, the selected values depending on whether or not a peak exists in the first predetermined band of frequencies. Moreover, means are in this way provided for selecting values of end frequencies for both the second band and a third higher band overlapping the second band and for determining whether or not one of the peaks is the only peak in the third band and lies in both the second and third bands, and if so, storing in one of the distinct locations another frequency value corresponding to an upper frequency edge of the one peak. In another aspect means are thus provided for determining whether or not one of the peaks lies in a third band which is generally higher in frequency than the second band and overlaps it, and if none of the peaks lies in the third band, storing another frequency value in one of the distinct locations, the other frequency value lying in the third band and being a function of a reference frequency value determined from at least two of the spectra.

Also, means are thus provided for storing as a lower first frequency the value of frequency of any lowest frequency peak in the first predetermined band of frequencies and as a higher first frequency the value of frequency of any next higher frequency peak in the first band, and for storing as a second frequency the value of frequency of any peak in the second band higher in frequency than the higher first frequency if the higher first frequency is also in the second band, and if there is no peak in the second band higher in frequency than the higher first frequency when it is in the second band then storing as the second frequency the value of frequency originally stored as the higher first frequency and storing as the higher first frequency the value of frequency stored as the lower first frequency also. Also provided thus is means for identifying lower and higher first frequencies descriptive of a peak which is widened or split upon at least one occurrence of nasality and for producing a signal indicative of the occurrence of nasality.

In step 237, CPU1 tests the normalized spectral values over increasing frequency K values to determine whether there is a peak P3 above any peak having value SF2 in band B3. Band B3 is already established to correspond with the BF or GS nature of the spectrum. The testing begins above value SF2 if SF2 lies in band B3, to avoid confusing the peak sought with any peak P2 found earlier. If a peak P3 is found, then operations proceed to a step 241 where third spectral frequency value SF3 is set to the frequency K value of the first peak above frequency SF2 in band B3. Next in a step 243, the speech goodness from step 235 is calculated based on a weighted average of the width of both peaks P2 and P3 using the function of Fig. 11 in the manner described hereinabove, and a calculation step 245 for SR is reached.

If there is no P3 peak found in step 237, operations branch to a step 247 where spectral frequency SF2 is tested to determine if it is in band B3. If so, operations proceed to a

step 249 where SF3 is set at the upper edge of the spectral envelope, whence step 243 is reached. If SF2 is not in band B3, operations branch to a step 251 where value SF3 is set to a value equal to reference SR multiplied by ten-to-the-1.18-power, whence step 243 is reached.

In step 245 the spectral reference value SR is illustratively set equal to the frequency of the first non-zero spectral value SF0 determined in step 211 if the spectrum is a GS spectrum and SF0 is greater than zero. A more sophisticated alternative calculation of value SR for step 245 is described in more detail later herein with reference to Fig. 14. After step 245 operations proceed to a RETURN 257.

In Fig. 14, CPU1 automatically computes spectral reference value SR (step 245 of Fig. 13B). The value SR is so defined that it is influenced by the geometric means of SF0 across the adult population (approximately 168 Hertz), by the geometric means of the pitch of the current talker, and by modulations in pitch of current talker filtered so as to eliminate the slow pitch changes such as those associated with pitch declination and so as to eliminate the very rapid transients at voice onset and offset. Specifically,

$$SR = (K1)(GMTF0/K1)^{a} + FIL(SF0_{i}) \qquad (5)$$

where K1 is a constant of about 168, GMTF0 is the goemetric mean of the current talker's pitch, a is a constant equal to about 1/3, and $FIL(SF0_{i})$ is the instantaneous value of the filtered modulations in the talker's SF0 for GS spectra. These parameters are chosen so as to make the average value of $Y_{s} = \log_{10}(SF1L/SR)$ constant across talkers, thus eliminating differences between talkers and so as to allow those SF0 modulations, which are believed to have phonetic significance, to influence the position of the sensory pointer. Only pitch modulations between about 1.5 Hertz and 50 Hertz are passed by a

software bandpass filter. More exactly, values for the filter band are selected so that slow variations of the pitch declination and the very rapid variations at pitch onset and termination are advantageously eliminated.

In Fig. 14 operation commences with BEGIN 301 and proceeds to a decision step 309 in which the spectrum is tested to determine whether it includes a periodic component. This test is performed according to any appropriate procedure such as the spectral analysis disclosed in L.J. Siegel et al. Voiced/unvoiced/mixed excitation classification of speech, IEEE Trans. Acoust. Speech Signal Processing, 1982, ASSP-30, pp. 451-460. If there is not a component that is periodic, then operations proceed to a RETURN 311 directly from step 309. If GS is 1, then in a step 315 a recalculation of the value of SR commences according to the formulas

$$GMTF0 = EXP((\ln SF0 + N2 \ln GMTF0)/(N2+1)) \quad (6A)$$

$$SR = 168(GMTF0/168)^{1/3} \quad (6B)$$

(EXP is the exponential function e^x, and ln is the natural logarithm function.) In words, the GMTF0 is based on the last N2 values of SF0 and gradually adapts from its initialized value of 168 Hertz to the talker's pitch. Then the reference value SR (unadjusted as yet for pitch modulation) is calculated by the empirical formula (6B) from the updated geometric mean GMTF0. Operations proceed from step 315 to a step 319.

In step 319 the software bandpass filter for pitch modulation is illustratively implemented by maintaining a table of the values SF0 of periodic spectra of glottal-source type. This table is analyzed for any discernible pitch modulation in the frequency range between 1.5 Hertz and 50 Hz. Then a value FIL which is originally initialized to zero is updated with the size of the pitch modulation determined from the output of the

pitch modulation software filter. Each pass through the operations of Fig. 4 accesses step 245 so the table has an entry added regularly when a glottal-source speech sound is in progress.

After step 319, the value of SR is increased in a step 321 by the value of FIL, whence a RETURN 323 is reached.

In this way CPU1 constitutes means for computing at least one of the values in the sets of first-named coordinate values (e.g. sensory pointer values) as a function of a reference frequency value which is a function of frequency values (e.g. values of SF0) determined from at least two of the spectra. CPU1 also constitutes means for computing at least one of the values in the sets of first-named coordinate values as a function of a reference frequency value which is a function of a geometric mean of frequency values determined from at least some glottal-source spectra over time. CPU1 additionally constitutes means for computing at least one of the values in the sets of first-named coordinate values as a function of a reference frequency which is a function of A) a frequency of pitch modulation of the speech and B) a mean of frequency values determined from at least some of the spectra of the speech over time.

Depending on the hardware used to implement CPU1, one or more processors are needed to accomplish the operations described for CPU1. Where a single processor is fast enough to accomplish the operations, it is contemplated that the block of Fig. 1 marked CPU1 represents a single processor. When the skilled worker uses a slower type of processor, then it is contemplated that several such processors are used in a multiprocessing arrangement to compute several spectra at the same time and then to analyze the spectra so obtained in order to accomplish real time analysis of the speech waveform. In such an arrangement, several microprocessors are multiplexed to line 23 from ADC 21 of Fig. 1 so that they take turns inputting the latest set of N1 samples in overlapping manner, for instance.

With a number P of microprocessors, each microprocessor need only input and compute the spectrum of every Pth set of N1 samples. Then the spectra can be supplied to one or more additional processors to analyze and output the auditory state code and the sensory pointer values X_S, Y_S and Z_S.

In Fig. 15, the flow of operations in CPU2 for converting from sensory to perceptual coordinates is detailed. In this process a vector difference equation, or set of three difference equations for the coordinates respectively, is solved by CPU2 point by point by executing a loop continually. The difference equations are the numerical versions of three differential equations discussed hereinbelow.

Solving the difference equations is regarded as a sensory-perceptual transformation, or transformation from sensory coordinates to perceptual coordinates as an integrative-predictive function. The fundamental concept of the sensory-perceptual transformation is that sensory pointers GSSP and BFSP as illustrated in Fig. 16 attract a perceptual pointer PP in the three dimensional mathematical space, or auditory-perceptual space having a coordinate system defined by three mutually perpendicular axes X, Y and Z, and induce the perceptual pointer to move through the auditory-perceptual space and trace out a perceptual path. Perceptual pointer PP has coordinate values X_p, Y_p and Z_p. The perceptual pointer PP almost instantaneously, that is within a few milliseconds, takes on the summed loudnesses of the sensory pointers GSSP and BFSP. However, when the sensory pointers disappear, the loudness of the perceptual pointer decays slowly over a period of 100 to 200 milliseconds. In this way the perceptual response is maintained during brief silences in the acoustic input.

The perceptual pointer, like the sensory pointer, is regarded as having at any moment an auditory state, for which a perceptual auditory state code is computed. The auditory state code of the perceptual pointer matches that of the sensory

pointer, except that a certain amount of time is required for state switching. For example, if both the sensory pointer and perceptual pointer are in the frication state (BF) and the sensory pointer suddenly switches to the voiced, nonnasal state (GS=1, NS=0), then a period of time is required before the perceptual pointer switches to the new state.

Also, fixed pointers called neutral points NPGS and NPBF affect the motion of the perceptual pointer PP in the absence of the sensory pointers. The use of at least one neutral point advantageously provides a home location for the perceptual pointer when a lengthy period of silence occurs. During such a period of silence, an attractive force from the neutral point NPGS causes the perceptual pointer PP to migrate toward it. Moreover, the use of at least one neutral point also remarkably allows the system to interpret even periods of silence in phonetically relevant ways in a manner similar to human speech perception. (For instance, many listeners hear "split" when a talker says "s" followed by brief silence followed by "lit.")

In Fig. 16 the neutral point NPGS attracts the perceptual pointer immediately upon GS changing from one to zero in the auditory state code if BF is already zero. The attraction by NPGS lasts as long as the period of silence does, and the neutral point NPBF does not attract pointer PP at all. On the other hand, if GS is already zero and BF in the auditory state code is one and changes to zero, then the neutral point NPBF attracts the perceptual pointer immediately upon BF changing from one to zero. The attraction by NPBF lasts about 120 milliseconds and is replaced upon the expiration of the 120 milliseconds by an attraction from the neutral point NPGS which lasts for the remainder of the period of silence until either GS or BF become one again.

The sensory pointers GSSP and BFSP are conceived as being attached by springs to the perceptual pointer PP which is

regarded as having mass and inertia. The stiffness of a spring depends on the goodness value and the loudness value of its associated sensory pointer. In this way, near-threshold spectra with little resemblance to speech have almost no influence on the perceptual response while moderately loud speech-like spectra have a strong influence on the perceptual response. The analogy to a spring is used because the attractive force of a sensory pointer or neutral point increases with the distance from the perceptual pointer PP. Unlike a physical system, however, the position of any sensory pointer or neutral point is not influenced by the spring, and all of the force acts on the perceptual pointer PP. In addition, the auditory-perceptual space is regarded as being a viscous medium and the perceptual pointer encounters resistance which not only varies with velocity but varies with the location of the perceptual pointer in a remarkable way. It is emphasized that the particular mathematical model of the sensory-perceptual transformation is illustrative and can be modified in the practice of the invention by the skilled worker as additional experimental information about the process of auditory perception is obtained.

The foregoing concepts are expressed in mathematical form by the difference equations which are solved by CPU2 to accomplish the sensory-perceptual transformation. In a further aspect, the difference equations are expressed in terms of variables which are coordinate values exponentiated. Since the sensory pointers of Fig. 16 have coordinates which are expressed in terms of logarithmic functions of frequency ratios in step 143 of Fig. 4, the mathematical space of Fig. 16 is called a "log space" herein. Because the coordinates are exponentiated in the difference equations, only the frequency ratios remain and the expression "ratio space" is adopted herein to refer to the domain in which the difference equations are expressed. It is contemplated that in some embodiments, no logarithms are

calculated in step 143 of Fig. 4 to avoid subsequently exponentiating in CPU2 to recover the ratios themselves. Subsequent analysis by CPU3 occurs in log space, however.

The following chart states the nomenclature for the variables in ratio space and log space:

Equations For Conversion To Ratio Space From Log Space

Ratio Space	Log Space	Remarks
Perceptual Pointer Coordinates:		
$XRP = 10^{X_p}$	X_p	X_p = PF3/PF2
$YRP = 10^{Y_p}$	Y_p	Y_p = PF1L/PR
$ZRP = 10^{Z_p}$	Z_p	Z_p = PF2/PF1H
(Where applicable numeral suffixes of 0, 1, or 2 are appended to XRP, YRP, ZRP, X_p, Y_p and Z_p to denote values for the same variable at different times.)		
Burst-Friction Sensory Pointer BFSP Coordinates:		
$XRSBF = 10^{X_{sBF}}$	X_{sBF}	
$YRSBF = 10^{Y_{sBF}}$	Y_{sBF}	
$ZRSBF = 10^{Z_{sBF}}$	Z_{sBF}	
Glottal-Source Sensory Pointer GSSP Coordinates:		
$XRSGS = 10^{X_{sGS}}$	X_{sGS}	
$YRSGS = 10^{Y_{sGS}}$	Y_{sGS}	
$ZRSGS = 10^{Z_{sGS}}$	Z_{sGS}	
Burst-Friction Neutral Point (NPBF) Coordinates:		
$XRNBF = 10^{X_{NBF}}$	X_{NBF}	$X_{NBF} = 0.6$
$YRNBF = 10^{Y_{NBF}}$	Y_{NBF}	$Y_{NBF} = 0$
$ZRNBF = 10^{Z_{NBF}}$	Z_{NBF}	$Z_{NBF} = 0.6$

Glottal-Source Neutral Point (GSSP) Coordinates:

XRNGS	$= 10^{X_{NGS}}$	X_{NGS}	$X_{NGS} = 0.4$
YRNGS	$= 10^{Y_{NGS}}$	Y_{NGS}	$Y_{NGS} = 0.4$
ZRNGS	$= 10^{Z_{NGS}}$	Z_{NGS}	$Z_{NGS} = 0.4$

CPU1 and CPU2 together electrically derive a series of coordinate values of points on a path in the mathematical space from frequency spectra of the speech occurring in successive time intervals respectively.

In Fig. 15 the operations of CPU2 commence with a START 401 and proceed to a step 403 to initialize a table 405 of Fig. 15A with two triplets of initial values XRP0, YRP0, ZRP0, XRP1, YRP1, ZRP1, for the set of coordinates XRP, YRP, ZRP in ratio space. In table 405, row zero (suffix zero on the variables) is regarded as earliest in time, row one as next in time, and row 2 as latest in time and to be solved for. The initial position coordinates are in row zero and are 10 raised to the power of the respective coordinates of the neutral pointer NPGS or $10^{\cdot 4}$. The initial velocity is assumed to be zero in both ratio space and log space so all the entries in row one are $10^{\cdot 4}$ too, because there is no change in position initially.

Next in a step 407, CPU2 reads the sensory pointer values X_s, Y_s and Z_s for either the BF sensory pointer or the GS sensory pointer or both, and the auditory state code values BF, GS, LIBF, LIGS, GBF, GGS and NS from CPU1. Then a computation step 413 occurs in which the difference equations involving the sensory pointer values in ratio space are solved to obtain the next in a series of coordinate values X_p, Y_p and Z_p on a path in the mathematical space. More specifically, the difference equations are solved for the entries for row 2 of table 405, and subsequently the logs of the entries in row 2 are computed in order to obtain perceptual pointer coordinates X_p, Y_p and Z_p in log space. The perceptual pointer coordinates X_p, Y_p and Z_p are regarded as tracing out a path

in the mathematical log space of Fig. 16 which path has a perceptual significance.

The difference equations solved in step 413 are now described.

Let a differential equation for each ratio space component of the position vector (XRP, YRP, ZRP) of pointer PP be first written as a force summation-to-zero of the pointer mass m times its acceleration (second derivative of ratio space position) plus the viscous drag as a function of velocity (first derivative of ratio space position) plus forces due to the sensory pointer(s) and neutral point(s) acting through springs.

Solving the difference equations numerically by CPU2 utilizes values of the coordinates XRP, YRP and ZRP from the two next-previous time intervals represented by rows zero and one of table 405, as well as quantities from the auditory state code and the sensory pointer coordinates in ratio space. Row two (2) of the table of Fig. 15A represents the unknown latest coordinate values on the path of the perceptual pointer in the ratio space which are to be obtained by solving the difference equations. Row one (1) of table 405 in general represents the next-previous coordinate values of the perceptual pointer which were found in the next previous pass through the computation loop of Fig. 15 by CPU2. Row zero (0) of the table generally represents the second-next-previous coordinate values of the perceptual pointer which were found in the second-next-previous pass through the computation loop of Fig. 15 by CPU2.

The derivative of XRP is approximated by

$$dXRP/dt = H(XRP2-XRP1) \qquad (7)$$

where H is the reciprocal of the time interval between spectra, e.g. 1/(2 milliseconds) or 500 Hertz. XRP2 is the latest X-coordinate value in ratio space to be solved for, and XRP1 is the next previous such X-coordinate value. These coordinate

values are derived by CPU1 from spectra that are generated periodically so the factor H is included in the Equation (7).

The second derivative of X_p is approximated by

$$d^2XRP/dt^2 = H^2(XRP2-2XRP1+XRP0) \quad (8)$$

The quantity H is the same as in Equation (7). XRP2 (table 405, row 2, column XRP) is the latest X coordinate value to be solved for and XRP1 is the next previous X coordinate value (table 405, row 1, column XRP). XRP0 is the second-next-previous X coordinate value (row 0, column XRP). The factor H-square occurs in Equation (8) because the second derivative is the derivative of the first derivative.

Based on the foregoing conceptual description and using the relationships of Equations (7) and (8), a set of difference equations to be solved by CPU2 is as follows:

Equation (9A)

$$\begin{aligned} 0 = {} & H^2(XRP2-2XRP1+XRP0) \\ & + rH(XRP2-XRP1)/B^{ABS(XRP2-XRNGS)} \\ & + LIGSxGGSxKGS(XRP2-XRSGS) \\ & + LIBFxGBFxKBF(XRP2-XRSBF) \\ & + NFx((1-GS)x(1-BF))xKNGSx(XRP2-XRNGS)^A \\ & + (1-NF)(1-GS)(1-BF))xKNBFx(XRP2-XRNBF)^A \end{aligned}$$

Equation (9B)

$$0 = H^2(YRP2-2YRP1+YRP0) + rH(YRP2-YRP1)/B^{ABS(YRP2-YRNGS)} + LIGSxGGSxKGS(YRP2-YRSGS) + LIBFxGBFxKBF(YRP2-YRSBF) + NFx((1-GS)x(1-BF))xKNGSx(YRP2-YRNGS)^A + (1-NF)(1-GS)(1-BF))xKNBFx(YRP2-YRNBF)^A$$

Equation (9C)

$$0 = H^2(ZRP2-2ZRP1+ZRP0) + rH(ZRP2-ZRP1)/B^{ABS(ZRP2-ZRNGS)} + LIGSxGGSxKGS(ZRP2-ZRSGS) + LIBFxGBFxKBF(ZRP2-ZRSBF) + NFx((1-GS)x(1-BF))xKNGSx(ZRP2-ZRNGS)^A + (1-NF)(1-GS)(1-BF))xKNBFx(ZRP2-ZRNBF)^A$$

CPU2 is programmed to perform an iterative or other suitable computation method to solve each of the three equations 9A, 9B and 9C for the latest coordinate values XRP2, YRP2 and ZRP2 of the perceptual pointer PP in the mathematical space.

The absolute value function is represented by ABS. Coordinate values XRP1, YRP1, ZRP1 and XRP0, YRP0, ZRP0 are previously calculated from the equations 9A, 9B and 9C and are available in the table 405 of Fig. 15A. Values of constants are illustratively set forth as follows:

Constant	Value
r	465
H	500
KGS	3000
KBF	6000
KNGS	3000
KNBF	3000
A	0
B	1.5

The viscous drag term is typified by the term $rH(YRP2-YRP1)/B^{ABS(YRP2-YRNGS)}$ in Equation 9B, which amounts to velocity times $r/B^{ABS(YRP2-YRNGS)}$. B is a base for the exponentiation, and the viscous drag factor is about equal to constant r near the neutral point NPGS (which has a Y coordinate of YRNGS in ratio space) because the exponent for B is about zero. The value of B is selected so that when the perceptual pointer PP moves over to the plane Y=0 in log space, then the viscous drag factor falls somewhat, e.g., to roughly half of constant r. When YRP2 is $10^0=1$, the denominator is $B^{ABS(1-10^{\cdot 4})}$ or very roughly B^2.

The variables LIGS, GGS, GS, LIBF, GBF, and BF are in the auditory state code supplied by CPU1. These variables activate or deactivate (state switch) appropriate terms in Equations 9A, 9B and 9C depending on which sensory pointer(s) or neutral point is exerting an attraction on perceptual pointer PP. Then

since the burst-friction flag BF and glottal-source flag GS are each either 0 or 1 and the loudness and goodness are zero during silence, the appropriate terms of the equations 9A, 9B and 9C figure in the solution computations or are advantageously neglected as circumstances require.

A neutral flag NF is included in the neutral point terms (the last two terms in each of the difference equations). Neutral flag NF is controlled by a timer in CPU2 which monitors the states of BF and GS in the auditory state code. If either BF or GS is 1, flag NF is 0. If BF is zero and GS makes a transition from 1 to zero, flag NF becomes 1 until either GS or BF becomes 1. If BF is 1 and GS is 0, and then BF makes a transition from 1 to zero as detected by step 407, then a 120 millisecond timer in CPU2 is activated to keep flag NF zero until the 120 milliseconds expires, whence flag NF is set to 1. In this way, the last term (for neutral point NPBF) in each difference equation is activated for 120 milliseconds and then is replaced by the second to last term (for neutral point NPGS) in each difference equation. Each term for a sensory pointer or neutral point is regarded as providing a contribution to the position of the perceptual pointer PP.

In this way means are provided for deriving sets of digital values representative of frequency spectra of the speech from the samples in digital form, for generating one of a plurality of auditory state codes for each of the sets of digital values and supplying at least two sets of coordinate values in a mathematical space, and for computing a series of other coordinate values of points defining a path with selected contributions from one or more of the sets of first-named coordinate values depending on which auditory state code is generated.

CPU1 is also advantageously programmed to perform operations to compute different loudnesses and goodnesses specific to the glottal-source and burst-friction portions of the same spectrum of a voiced fricative or other speech sound, which

values LIBF, LIGS, GGS and GBF are transmitted from CPU1 to CPU2, and two sets of sensory pointer values X_{sGS}, Y_{sGS}, Z_{sGS}, X_{sBF}, Y_{sBF} and Z_{sBF} are sent for the glottal-source pointer GSSP and the burst-friction pointer BFSP, instead of one triplet X_s, Y_s and Z_s. In this way means are provided for producing a first of the two sets of first-named coordinate values from one of the sets of digital values representing spectra when the auditory state code indicates a glottal-source auditory state and for also producing the second of the two sets of first-named coordinate values from the same one set of digital values when the auditory state code simultaneously indicates a burst-friction auditory state.

The use of at least one neutral point as well as at least one sensory pointer in CPU2 provides means of producing a first of two sets of first-named coordinate values from the sets of digital values representing spectra and wherein the second set (e.g. neutral point values) of the first-named coordinate values is independent of the sets of digital values representing spectra.

In Equations 9A, 9B and 9C, the value A is an exponent, illustratively 0, indicating that a neutral point attracts the perceptual pointer PP with a force that does not vary with distance. The value of A is made positive if experimental observations suggest that the force should increase with distance, or A is made negative if the force should decrease with distance. It is presently believed that the best value of A is zero.

For purposes of description, the equations 9A, 9B and 9C are collectively regarded as expressing one vector difference equation for the vector position of the perceptual pointer PP. Advantageously, all sensory inputs to microphone 11 of Fig. 1, including bursts, transitions, steady-states, and silences are

all integrated into a single perceptual response by the sensory-perceptual transformation. In a further advantage, the perceptual pointer PP position depends not only on the position of the sensory pointers but also their dynamics. When the equations correspond to an underdamped system, a sensory pointer may rapidly approach and veer away from a target location, and yet it induces the perceptual pointer to overshoot and reach that desired location in the mathematical space. Operations by CPU2 in solving the difference equations are advantageously arranged to be analogous to such overshooting behavior, particularly in the cases of stop consonants and very rapid speech.

In step 415 of Fig. 15, the latest values XRP2, YRP2, ZRP2 resulting from solution of Equations 9A, 9B and 9C are stored in row 2 of table 405 of Fig. 15A. Then in a step 417 common logarithms of these latest values are sent as X_p, Y_p, Z_p to CPU3. Operations proceed to a decision step 419 to determine if CPU3 is to remain ON. If ON, then a loop is made back to step 407. A new set of sensory pointer coordinates and auditory state code information is received in step 407. Table 405 is maintained in a cyclic manner to prepare for the next pass through the computation step 413, so that in table 405 the values in row 2 become the first-previous values and the values in row 1 become the second-next-previous values for purposes of XRP1, YRP1, ZRP1 and XRP0, YRP0, ZRP0 respectively. Equations 9A, 9B and 9C are solved again in step 413 and operations continue in the loop of Fig. 15 until CPU3 is not ON at decision step 419 whence operations terminate at an END 421.

The operations of CPU3 are first discussed conceptually in connection with Fig. 17. Auditory-perceptual events, or perceived sounds occur when the behavior of the perceptual pointer PP meets certain criteria. These are (a) an auditory-perceptual event occurs when the perceptual pointer undergoes a period of low velocity; (b) an auditory-perceptual event occurs when the perceptual pointer undergoes sharp deceleration; and

(c) an auditory-perceptual event occurs when the path of the perceptual pointer has high curvature. CPU3 is appropriately programmed to determine such events. The computations can involve any one or more of the criteria, and time constraints can be added such that a velocity must be maintained for a predetermined number of milliseconds, or that a path or a certain locus and curvature have to be traversed within certain time limits.

In these various cases the auditory-perceptual event is regarded as associated with a position along the path in log space of a peak in magnitude of acceleration (determined now in log space and not ratio space in the preferred embodiment) of the perceptual pointer PP. The position of the perceptual pointer PP in log space is a vector defined by the coordinate values X_p, Y_p and Z_p. Its velocity is a vector quantity equal to speed in a particular direction relative to the X, Y, Z frame of reference. The velocity has the components dX_p/dt, dY_p/dt and dZ_p/dt, which are the time derivatives of X_p, Y_p and Z_p. Speed is the magnitude, or length, of the velocity vector at any given time and is equal to the square root of the sum of the squares of the velocity components dX_p/dt, dY_p/dt and dZ_p/dt. In general, the magnitude, or length, of any vector is equal to the square root of the sum of the squares of its components. Acceleration is a vector which represents change of velocity or rate of such change, as regards either speed or direction or both. The components of acceleration are the time derivatives of the components of the velocity vector respectively. In mathematical terms, the acceleration has components d^2X_p/dt^2, d^2Y_p/dt^2 and d^2Z_p/dt^2, which are the time derivatives of dX_p/dt, dY_p/dt and dZ_p/dt.

Even when deceleration is involved in an auditory-perceptual event, the event is associated with a position along of the path of a peak in magnitude of acceleration of the perceptual pointer PP because a period of low velocity results from

a deceleration which amounts to a peak in magnitude of acceleration. Also, a sharp deceleration is a peak in magnitude of acceleration because deceleration is negative acceleration and a negative sign does not affect the magnitude which involves sums of squares. When the path of the perceptual pointer has high curvature, the acceleration is a vector peaking in magnitude and pointing centripetally from the path.

CPU3 acts as at least one or more of the following: A) means for identifying coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the position coordinate values and obtaining from said memory means the phonetic representation information prestored at that memory address; B) means for computing a parameter approximating the curvature of the path and, when the parameter exceeds a predetermined magnitude at a point on the path, identifying the coordinate values of that point to approximate the position of a peak in magnitude of acceleration; C) means for computing a speed along the path and identifying the coordinate values of a position where the speed decreases by at least a predetermined amount within a predetermined time, to approximate the position of a peak in magnitude of acceleration; or D) means for computing a speed along the path and identifying the coordinate values of a position where a decrease in speed occurs that is both preceded and succeeded by increases in speed within a predetermined time, to approximate the position of a peak in the magnitude of acceleration.

Each auditory-perceptual event is said to leave a trace or tick mark that fades in time. When a cloud of ticks occurs, that is when a region of high density of ticks surrounded by a region of lower density is formed, as would be the case for an oft-repeated speech sound, it is postulated that in human beings, the nervous system automatically places an envelope around the cloud of tick marks and creates a target zone capable

of issuing a neural symbol or a category code. Under most circumstances such target zones are temporary and dissolve with time. Other target zones, such as those for the phones of one's native language and dialect, are formed during infancy and childhood under certain circumstances, such that they are nearly permanent and difficult to modify.

The concept of the target zone is perceptual. In the preferred embodiment the large memory 31 for target space storage is a memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory. CPU1, CPU2, and CPU3 together constitute means for electrically deriving a series of coordinate values of points on a path in a mathematical space from frequency spectra of the speech occurring in successive time intervals respectively, for identifying coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the position coordinate values and obtaining from said memory means the phonetic representation information prestored at that memory address.

The target zones for stop phonemes such as /b/, /d/, /g/, /k/, /p/ and /t/ are associated with respective sets of addresses in the memory corresponding to a region of the mathematical space which cannot be entered by sensory pointer values X_s, Y_s and Z_s but which can be entered by the coordinate values X_p, Y_p and Z_p because of underdamping in the sensory-perceptual transformation.

CPU3 finds a peak in the magnitude of acceleration. The coordinates on the path at which a latest peak occurs are converted to integer values along each axis X, Y and Z. In terms of the coordinate values for the sensory pointer which can be expected to result from the step 143 of Fig. 4, the target zones lie within ranges for X between 0 and 2, Y between -.5 and 1.5 and Z between 0 and 2. In the preferred embodiment each

axis is regarded as having 200 divisions which, for example include 150 divisions along the positive Y axis and 50 divisions along the negative Y axis. In this way, the shape of each target zone is definable with considerable precision. Therefore, the X_p, Y_p and Z_p values at which the latest peak occurs are multiplied by 100 and rounded to the nearest integer by a function INT. Since a peak can occur anywhere within the ranges, a number of memory addresses equal to the cube of 200, or 8 megabytes, is used. In other words 23 bits are used to express each memory address in binary form, since 2^{23} is about 8 million. The coordinates are converted to a memory address by the equation

$$ADR = INT(100X) + 200xINT(100Y+50) + 40000xINT(100Z) \quad (10)$$

In other words when CPU3 finds a peak in the magnitude of acceleration by velocity analysis, curvature analysis, or acceleration analysis, it then generates memory address ADR according to the above equation or the equivalent and obtains from the memory 31 the phonetic representation information prestored at that address. A binary code representing each phoneme, or phonetic element generally, of a language is stored at each of a set of addresses in the memory. The 8 bits in a byte provide ample flexibility to provide distinct arbitrary binary designations for the different phonemes in a given human language. When CPU3 asserts memory address ADR, memory 31 supplies the binary code stored at that address. CPU3 then converts the binary code to a letter or other symbol representation of the phoneme and displays it on the video screen of its terminal and prints it out on printer 33.

The targets for the nonsustainable speech sounds are placed outside of the octant of positive X, Y, and Z. In Fig. 17 the sensory pointer BFSP can only approach a target zone such as 451 for a sound such as "p" and must do with appropriate

dynamics such that the perceptual pointer actually reaches the target zone in the negative Y region. For example, suppose a talker is just finishing saying the word "Stop." The perceptual pointer has just made a sharp curve while passing through a target zone 453 for the vowel sound in "stop" under the influence of the glottal-source sensory pointer GSSP, now absent, and the suddenly appearing burst-friction sensory pointer BFSP. Because of the sharp curve, a memory lookup occurs for the coordinates of a point 455 and a phonetic element /a/ (as in "father") is obtained from memory 31. The burst-friction sensory pointer BFSP appears in the X-Z plane because of the "p" sound, attracting the perceptual pointer PP toward BFSP. Perceptual pointer PP overshoots the plane Y=0, in which BFSP occurs, and reaches target zone 451 for "p". Because of the attractive force of BFSP followed in succession by the neutral points NPBF and then NPGS, perceptual pointer PP reverses its direction of motion at a point 457 in the target zone 451, resulting in another peak in magnitude of acceleration. A memory lookup again occurs, this time for the coordinates of point 457 and a phonetic element for "p" is obtained from memory 31. The sensory pointers thus can in some cases only go to approach zones in such a way as to induce the perceptual pointer PP to reach the more distant perceptual target zone. However, the target zones such as 453 for the vowels are able to be entered by both the sensory and perceptual pointers. The perceptual response should reach vowel target zones when starting from neutral point NPGS in about 50 milliseconds.

Fig. 18 shows the axes X, Y and Z of a coordinate system for the mathematical space. In describing the target zones for the vowels, it is useful to define additional axes X', Y' and Z' which intersect at a point in the first octant of the X, Y, Z system and are inclined relative to the axes X, Y and Z. The equations defining the X', Y', Z' coordinates are as follows:

$$X' = 0.70711*(Y-X) \quad (11A)$$

$$Y' = 0.8162*Z - 0.4081*(X+Y) \quad (11B)$$

$$Z' = 0.5772*(X+Y+Z) \quad (11C)$$

Fig. 19 is a view of an approximately planar slab 465 in the X', Y', Z' coordinate which has been found to hold the target zones for the vowels. Fig. 19 shows the slab 465 edge as viewed along the X' axis. The neutral point NPGS is approximately centered in the vowel slab. Even though the vowel slab is thin, lip-rounding moves the vowel to the back of the slab, while retroflection as in r-coloring moves the position far back toward the origin so that even with the vowels alone, the use of three-dimensions is beneficial. The consonants fall in or near the vowel slab or in another slab that is orthogonal to the vowel slab, further supporting the use of a three dimensional space. It is contemplated that in some embodiments of the invention, however, that the slabs can be unfolded and unwrapped in such a way that a two dimensional space can be used. Also, it is contemplated that the slabs be mapped into the memory 31 addresses in such a way that the available memory capacity is efficiently used only for the slabs.

Fig. 20 is a view of the slab 465 face-on and viewed along the Z' axis in the X', Y', Z' coordinate system. Outlines for the target zones for the vowels are shown, from which outlines sets of addresses are derived for prestoring codes representing each of the vowel symbols in memory 31 of Fig. 1. These codes are prestored by manually entering them for each of the addresses corresponding to a point within each of the target zones. Also the codes can be prestored by preparing 3-dimensional position acquisition equipment 467 such as a Perceptor unit from Micro Control Systems, Inc., of Vernon

Connecticut. The unit has a teflon-coated, precision-ground aluminum reference plate, on which is mounted a precision-machined digitizing arm. A circuit that performs electrical data acquisition functions is housed beneath the reference plate. Dual RS-232 ports let the unit transmit data. The digitizing arm has five pre-loaded ball bearing supported joints which allow the arm to move. Potentiometers housed in the joints transmit electrical information about the angles of rotation of each segment of the arm. Then a Z-80A microprocessor in the unit computes the x, y, and z coordinates of the position of the arm's pointer tip. In this way the shapes of the target zones are recorded relatively rapidly for use in automatically programming the memory 31 of Fig. 1.

Fig. 21 shows target zones in the mathematical space for voiceless stops as viewed along the Y axis of Fig. 18. The legend for this Figure is found in Table 1. The shapes of the target zones defined by Fig. 21 are projected onto the X-Z plane but they actually only occupy a negative Y region between y=-0.10 and y=-0.04.

Fig. 22 depicts target zones in the mathematical space for voiced stops and unaspirated voiceless stops and nasal consonants as viewed along the Y axis of Fig. 18. The legend for this Figure is found in Table 2. The shapes of the target zones for the four voiced stops and unaspirated voiceless stops defined in the upper part of Fig. 22 are projected onto the X-Z plane but they actually only occupy a negative Y region between y=-0.04 and y=-0.02. Similarly, the shapes of the target zones for the three nasal consonants defined in the lower part of Fig. 22 are projected onto the X-Z plane but they actually only occupy a positive Y region between y=+0.05 and y=+0.34.

Fig. 23 depicts target zones in the mathematical space for voiceless fricatives of American English as viewed along the Y axis of Fig. 18. The legend for this Figure is found in Table

3. The shapes of the target zones defined by Fig. 21 are projected onto the X-Z plane but they actually occupy a Y region between y=-0.02 and y=+0.02.

Fig. 24 depicts target zones in the mathematical space for voiced fricatives and the phonetic approximates as viewed along the Z' axis of the X', Y', Z' coordinate system of Fig. 18. Fig. 25 depicts target zones in the mathematical space for the voiced fricatives and the phonetic approximates of Fig. 24 as viewed along the X' axis of the X', Y', Z' coordinate system of Fig. 18. The legend for Figs. 24 and 25 is found in Table 4. These target zones are generally juxtaposed in or near the vowels, so the the X', Y', Z' coordinate system is used. The Figs. 24 and 25 are interpreted in the manner of an orthographic projection to define the three dimensional shapes of the target zones. A superficial comparison of Figs. 20 and 24 might suggest that the target zones for /er/ and /r/ in Fig. 24 conflict with the target zones of some of the vowels of Fig. 20, but this is not the case. Fig. 25 makes it clear that /er/ and /r/ fall behind the vowels in the log space. In general target zones do not overlap. A legend for the vowel Figs. 19 and 20 is found in Table 5.

TABLE I

VOICELESS (ASPIRATED) STOPS (PLOSIVES)

$k_v \equiv |k^h|$ - velar

$k_p \equiv |k^h|$ - palatal

$t \equiv |t^h|$

$p \equiv |p^h|$

TABLE 2

VOICED PLOSIVES (STOPS) AND UNASPIRATED VOICELESS PLOSIVES (STOPS)

$q_v \equiv /g/$ - velar

$q_p \equiv /g/$ - palatal

$d \equiv /d/$

$b \equiv /b/$

Note: These include unaspirated k, t, p of American-English

NASAL CONSONANTS

$m \equiv /m/$

$n \equiv /n/$

$\eta \equiv /\eta/ \equiv$ ng in sing

TABLE 3

(VOICELESS FRICATIVES (AMERICAN-ENGLISH)

s ≡ /s/

h ≡ /h/

sh ≡ /ʃ/

th$\bar{v}$ ≡ /θ/

f ≡ /f/

wh ≡ /ʍ/

TABLE 4

VOICED FRICATIVES AND THE APPROXIMATES

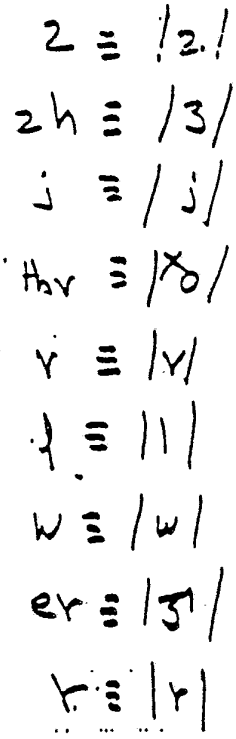
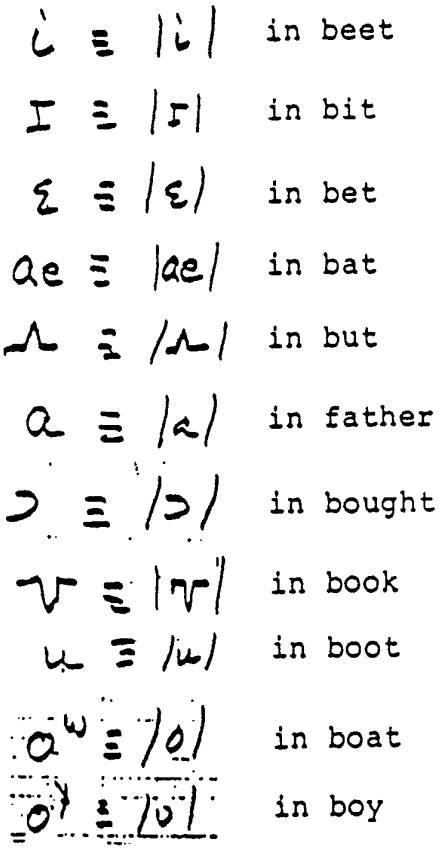


TABLE 5

VOWELS



It is contemplated that the skilled worker use the shapes and coordinate information contained in Figs. 19-25 for loading memory 31 in constructing the preferred embodiment.

In Fig. 26 operations of CPU3 of Fig. 1 commence with a START 501 and proceed to a step 503 where the coordinate values X_p, Y_p and Z_p of the latest point on the path in the mathematical space are input from CPU2 and stored in a table 504 of Fig. 27. Next in step 505, the significant parameters of the trajectory are computed, so that it can be subsequently determined when a significant speech event occurs. The coordinate values result from sampling by S/H 17 at equal intervals in time and analyzing the spectra at a repetition rate expressed by the quantity H hereinabove. Therefore, the magnitude of acceleration is computed from the latest coordinate values and the two previous triplets of coordinate values from table 504. The subscripts zero (0), one (1) and two (2) are used to indicate the latest triplet, the next previous triplet, and the triplet before that. The magnitude of acceleration is illustratively computed according to the equation

$$\begin{aligned} \text{MAGACCEL} = H^2 \text{SQRT}((&X_{p0}-2X_{p1}+X_{p2})^2 \\ &+ (Y_{p0}-2Y_{p1}+Y_{p2})^2 \\ &+ (Z_{p0}-2Z_{p1}+Z_{p2})^2) \end{aligned} \qquad (12)$$

In some embodiments the curvature CURV is also calculated from MAGACCEL just given and from a velocity-squared quantity according to the equations

$$\text{CURV} = \text{MAGACCEL}/\text{VELSQ}$$

$$\text{where VELSQ} = (X_{p0}-X_{p1})^2 + (Y_{p0}-Y_{p1})^2 + (Z_{p0}-Z_{p1})^2 \qquad (13)$$

Each latest value of the magnitude of acceleration MAGACCEL is stored during step 505 in table 504 holding it and four previous values of MAGACCEL. Similar tabular analysis of the curvature CURV is applied where curvature is used. The argument in the square root SQRT function is sufficient for use as a parameter related to the magnitude of acceleration also. It is emphasized that there are many ways of calculating significant trajectory parameters to accomplish an equivalent analysis of the path of the perceptual pointer in the mathematical space.

Next in step 507, the table 504 holding five values of MAGACCEL is tested to determine if a significant peak has occurred. A suitable test is that a peak has occurred if the table has a tested value entered therein which exceeds a predetermined level and is is preceded and succeeded in the table 504 by values less than the tested value. If this test is not passed operations pass from step 507 to a decision step 509 where CPU3 checks an ON/OFF switch to determine whether it is to continue, and if so operations loop back to step 503. Eventually, a phonetically significant event occurs at step 507 and operations proceed to a step 511 to generate an address ADR according to ADR Equation (10) hereinabove.

Then in a step 513 the address ADR is asserted by CPU3 to memory 31 of Fig. 1 to obtain a prestored phonetic element code PHE byte identifying the phonetic element of the target zone where the significant X_p, Y_p, Z_p coordinate values lie. In a step 515, this PHE value is stored in a memory space holding PHE values in the order in which they are obtained. In a step 517, the PHE value, or byte is looked up in a table providing instructions for writing the corresponding phonetic symbol or category code corresponding to a phone of the language to printer 33 of Fig. 1.

In a next step 519 all PHE values stored in the order in which they are obtained are sent to CPU4 of Fig. 1. CPU4 is

a lexical access processor which converts the PHE values to a series of words spelled according to a language chosen.

When step 519 is completed, operations proceed to ON decision step 509, and loop back to step 503 unless the CPU3 is to be no longer ON, whence operations terminate at an END 521.

A system (not shown) for studying target zones for refining system 1 from examples of talkers' speech displays and analyzes the target zones in three-dimensional display of the mathematical space. Such a system has an Evans and Sutherland PS300 Graphic System and a VAX-750 or uVAX-II computer, a special purpose "coordinate transformer" and appropriate peripherals that allow three-dimensional viewing of line figures. Features of the display include knob control of "zoom", and knob control of rotation or translation relative to the system's axes.

The mathematical space, or auditory-perceptual space is displayed with axes. Three-dimensional target zones are created with a programs in the system. A target zone can be located in the space with a specified color, orientation and size as well as with a phonetic symbol located near it as desired.

To display the path of the sensory pointer, a quadruple set of values F0, F1, F2, F3 is entered for each time t at which time the fundamental and the first three spectral prominences are estimated using current speech-analysis techniques. These quadruples comprise a file. Next a value of a constant a is selected and quadruples (t, log(F3/F2), log (F1/R), log(F2/F1) are formed, where is a reference. These are the logarithms of the formant ratios and comprise a second file. When F1 is not defined log(F1/R) is arbitrarily set to zero. Next, linear interpolation is performed by the computer to provide a file of the quadruples spaced at 5 or 10 millisecond intervals. A line segment connecting each set of coordinates can be displayed at user option. On the tip of each such

segment a pyramid, appropriate ly oriented is displayed to represent the sensory pointer. The line segments and pyramids are stored in a third file. The mathematical space is displayed with appropriate selection of target zones. The user selects a sensory path, e.g. the syllable "dud" as spoken by a particular speaker. Then a rate of display, such as five times real time, is selected and the run is started. The displays shows the sensory pointer moving through the mathematical space, and its path is shown by the segments.

The interpolated log ratio file is converted into a table representing perceptual coordinates by applying the sensory-perceptual transformation to the sensory coordinates. n-resonators (second order) serve as the transformation. In this way, certain rates of spectral modulation are emphasized and others attenuated. These are stored in a fourth file. The perceptual path is displayed in the same way as the sensory path.

Further programs enable the study of the magnitudes of velocity v, acceleration a, and curvature k as either the sensory pointer or perceptual pointer moves through the space. Appropriately scaled displays permit viewing of x, y, x, v, a, and k as a function of time or to view similarly log(F3), log(F2), log(F1), log(F0), v, a, and k as a function of time. In this way, one can study sensory and perceptual paths to discover the correlates of the phoneme and syllable. Knob control of a cursor permits marking points of interest and determination of the values of the coordinates and dynamic parameters at those points. Modeling of the sensory-perceptual transformation as a single second-order resonator with a center frequency of 55 Hz. and a damping factor of 0.6 results in perceptual paths that are orderly and reasonable, although experimental refinements can be made.

Further features involving top-down processing are now discussed. The importance of top-down processing in a great many listening situations is significant, and the separation of

the perceptual and sensory aspects of phonetic processing advantageously permits top-down processing by CPU2, CPU3 and/or CPU4. For example, information derived by the system by pattern recognition apparatus, prestorage or other means is suitably used to generate additional contributions in Equations 9A, 9B and 9C that attract the perceptual pointer toward particular target zones. In this way, the perceptual pointer is driven not only by the sensory pointer(s) and the other factors previously mentioned, but also by the other information derived by the system as they are controlled by context, knowledge of the language, and so on. Another form of top-down processing involves information such as visual cues and information from other senses resulting in attractive or repulsive forces on the perceptual pointer. For example, mouth movements can be observed by pattern recognition apparatus and used to add forces that attract the perceptual pointer PP to various target zones and thus influence phonetic perception. Even more complicated forms of top-down processing are contemplated. For example, the sizes and shapes of the target zones are changed depending on the speech characteristics of the talker such as having a foreign accent, deaf speech, and so on.

Additional kinds of top-down processing are introduced as the output of the auditory-perceptual space undergoes additional processing such as that required for the identification of words and meanings. For instance, in CPU3 memory 31 in such embodiments, the PHE information prestored in the memory is accompanied by confidence level information bits representing a confidence between 0 and 1. PHE information for volume elements deep in the interior of a target zone has a high confidence, and PHE information for volume elements near the surface of a target zone has a low confidence. The confidence information derived from the target zones when a peak in acceleration magnitude occurs is compared with confidence information derived from the pattern recognition apparatus and a decision is made as to the

most probable interpretation of the speech. Similar analyses are executed in embodiments of the invention at the lexical access level by CPU4 to identify words and meanings.

In other embodiments of the invention, CPU3 forms and refines the target zones in memory 31 automatically. Streams of speech are fed to the system 1 and phonetically significant events identify addresses in memory 31. CPU3 tabulates the frequencies of events in regions of the memory and assigns distinctive binary category codes to regions having clusters of events. The category codes are listed in a table, and the skilled worker assigns conventional phonetic symbols to the tabulated category codes generated by the system, so that the system prints out the conventional symbols needed for human interpretation of the category codes generated by the system in a manner analogous to teaching the system to spell at a phonetic element level.

In view of the above, it will be seen that the several objects of the invention are achieved and other advantageous results attained.

As various changes could be made in the above constructions without departing from the scope of the invention, it is intended that all matter contained in the above description or shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

Claims

WHAT IS CLAIMED IS:

1. Speech processing apparatus comprising:

memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory; and

means for electrically deriving a series of coordinate values of points on a path in a mathematical space from frequency spectra of the speech occurring in successive time intervals respectively, for identifying coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the position coordinate values and obtaining from said memory means the phonetic representation information prestored at that memory address.

2. Speech processing apparatus as set forth in claim 1 wherein said deriving means includes means for computing a parameter approximating the curvature of the path and, when the parameter exceeds a predetermined magnitude at a point on the path, identifying the coordinate values of that point to approximate the position of a peak in magnitude of acceleration.

3. Speech processing apparatus as set forth in claim 1 wherein said deriving means includes means for computing a speed along the path and identifying the coordinate values of a position where the speed decreases by at least a predetermined amount within a predetermined time, to approximate the position of a peak in magnitude of acceleration.

4. Speech processing apparatus as set forth in claim 1 wherein said deriving means includes means for computing a speed along the path and identifying the coordinate values of a position where a decrease in speed occurs that is succeeded by an increase in speed within a predetermined time, to approximate the position of a peak in the magnitude of acceleration.

5. Speech processing apparatus as set forth in claim 1 wherein said deriving means includes means for producing sets of digital values representative of the frequency spectra of the speech, for generating one of a plurality of auditory state codes for each of said sets of digital values and supplying at least two sets of coordinate values in a mathematical space, and for computing the series of derived coordinate values of points defining the path with selected contributions from one or more of said sets of coordinate values depending on which auditory state code is generated.

6. Speech processing apparatus as set forth in claim 1 wherein said deriving means includes means for producing sets of digital values representative of the frequency spectra of the speech, for producing sets of sensory pointer values in the mathematical space, which sets of sensory pointer values are determined from the respective sets of digital values, and for computing from the sets of sensory pointer values the series of derived coordinate values of points defining the path.

7. Speech processing apparatus as set forth in claim 1 wherein said memory means includes means for holding prestored information representative of at least one stop phoneme at addresses corresponding to a region of the mathematical space which cannot be entered by the sets of sensory pointer values.

8. Speech processing apparatus comprising:

memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory; and

means for electrically deriving a series of coordinate values of points in a mathematical space from frequency spectra of the speech occurring in successive time intervals respectively, the series of coordinate values defining a path of the points in the mathematical space, for electrically computing a parameter approximating the curvature of the path and, when the parameter exceeds a predetermined magnitude at a point on the path, generating a memory address as a function of the coordinate values of the point on the path and obtaining from said memory means the phonetic representation information prestored at that memory address.

9. Speech processing apparatus as set forth in claim 8 further comprising means connected to said deriving means for generating the frequency spectra of the speech occurring in the successive time intervals.

10. Speech processing apparatus as set forth in claim 9 further comprising a microphone for converting an acoustic waveform of the speech to electrical form for use by said spectra generating means.

11. Speech processing apparatus as set forth in claim 9 wherein at least one of the frequency spectra has a plurality of spectral peaks and said deriving means includes means for computing the series of coordinate values of the points on the path in the mathematical space from frequency values of the spectral peaks.

12. Speech processing apparatus as set forth in claim 8 further comprising means connected to said deriving means for recording the phonetic representation information in the order obtained from the memory means.

13. Speech processing apparatus comprising:

means for producing samples of an analog waveform of speech and converting the samples into digital form;

means for deriving sets of digital values representative of frequency spectra of the speech from the samples in digital form, for generating one of a plurality of auditory state codes for each of the sets of digital values and supplying at least two sets of coordinate values in a mathematical space, and for computing a series of other coordinate values of points defining a path with selected contributions from one or more of the sets of first-named coordinate values depending on which auditory state code is generated; and

means for temporarily storing in digital form the computed coordinate values of the points on the path.

14. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for producing a first of the two sets of first-named coordinate values from the sets of digital values representing spectra and wherein the second set of the first-named coordinate values is independent of the sets of digital values representing spectra.

15. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for producing a first of the two sets of first-named coordinate values from one of the sets of digital values representing spectra when the

auditory state code indicates a glottal-source auditory state and for also producing the second of the two sets of first-named coordinate values from the same one set of digital values when the auditory state code simultaneously indicates a burst-friction auditory state.

16. Speech processing apparatus as set forth in claim 13 wherein one of the frequency spectra has a set of three spectral prominences with first, second, and third frequencies in order of increasing frequency and said deriving means includes means for generating an auditory state code indicative of a glottal-source state for said one spectrum and for producing the first of the two sets of coordinate values to have respective values involving the ratio of the third to the second frequencies, the ratio of the first frequency to a reference frequency value, and the ratio of the second to the first frequencies.

17. Speech processing apparatus as set forth in claim 13 wherein one of the frequency spectra has a set of three spectral prominences with first, second, and third frequencies in order of increasing frequency and another of the frequency spectra lacks the first one of the three prominences, and said deriving means includes means for generating an auditory state code indicative of a burst-friction state for the other frequency spectrum and for producing one of the two sets of coordinate values with a first value involving the ratio of the third to the second frequencies, a second value which is substantially constant, and a third value involving the ratio of the second frequency to a reference frequency value.

18. Speech processing apparatus as set forth in claim 13 wherein the frequency spectra have peaks and said deriving means includes means for increasing at least one of the contributions when the width of at least one of the peaks is within a predetermined range indicative of speech goodness.

19. Speech processing apparatus as set forth in claim 13 wherein the frequency spectra have peaks and said deriving means includes means for varying at least one of the contributions as a function of the total power of at least one of the peaks, which total power is indicative of speech loudness.

20. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for computing at least one of the values in the sets of first-named coordinate values as a function of a reference frequency value which is a function of frequency values determined from at least two of the spectra.

21. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for computing at least one of the values in the sets of first-named coordinate values as a function of a reference frequency value which is a function of a geometric mean of frequency values determined from at least some periodic glottal-source spectra over time.

22. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for computing at least one of the values in the sets of first-named coordinate values as a function of a reference frequency which is a function of A) a frequency of pitch modulation of the speech and B) a mean of frequency values determined from at least some of the spectra of the speech over time.

23. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for computing coordinate values of the points defining the path so that upon an occurrence of at least one burst sound the path passes through a region of the mathematical space which said two sets of coordinate values cannot enter.

24. Speech processing apparatus as set forth in claim 13 wherein said deriving means includes means for digitally solving a set of difference equations involving said two sets of coordinate values to obtain the series of other coordinate values of the points defining the path.

25. Speech processing apparatus comprising:

means for deriving sets of digital values representative of frequency spectra of the speech from samples of the speech in digital form, for generating one of a plurality of auditory state codes for each of the sets of digital values and producing glottal-source sensory pointer values and burst-friction sensory pointer values, which pointer values are determined from the respective sets of digital values;

means for computing from the glottal-source sensory pointer values and burst-friction sensory pointer values a series of coordinate values defining a path of a perceptual pointer in a mathematical space by digitally solving a set of difference equations involving selected contributions from the glottal-source sensory pointer values, from the burst-friction sensory pointer values and from coordinate values of a neutral point independent of the spectra, the selected contributions depending on which auditory state code is generated;

memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory; and

means for identifying coordinate values approximating at least one position along the path where magnitude of acceleration is pronounced, generating a memory address as a function of the identified position coordinate values and obtaining from said

memory means the phonetic representation information prestored at that memory address.

26. Speech processing apparatus as set forth in claim 25 wherein one of the frequency spectra has a set of three spectral prominences with first, second, and third frequencies in order of increasing frequency and another of the frequency spectra lacks the first one of the three prominences, and said deriving means includes means for generating an auditory state code indicative of a burst-friction state for the other spectrum and for producing the burst-friction sensory pointer values to have a first value involving the ratio of the third to the second frequencies, a second value which is substantially constant, and a third value involving the ratio of the second frequency to a reference frequency value.

27. Speech processing apparatus as set forth in claim 26 wherein the reference frequency value is a function of a geometric mean of frequency values determined by the deriving means from at least some of the glottal-source spectra over time.

28. Speech processing apparatus as set forth in claim 25 wherein one of the frequency spectra has a set of three spectral prominences with first, second, and third frequencies in order of increasing frequency and a second one of the frequency spectra lacks the first one of the three prominences, and said deriving means includes means for generating an auditory state code indicative of a glottal-source state for the one spectrum and a burst-friction state for the second spectrum.

29. Speech processing apparatus as set forth in claim 25 wherein said computing means includes means for computing coordinate values of the perceptual pointer so that upon an

occurrence of at least one burst sound the perceptual pointer passes through a region of the mathematical space which the glottal-source sensory pointer values and burst-friction sensory pointer values do not enter.

30. A method of processing speech comprising the steps of:

electrically deriving a series of coordinate values of points in a mathematical space from the frequency spectra of the speech occurring in successive time intervals respectively, the series of coordinate values defining a path of the points in the mathematical space; and

electrically identifying coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the coordinate values of the position on the path and obtaining from a memory means, having prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the memory, the phonetic representation information prestored at that memory address.

31. A method of processing speech comprising the steps of:

deriving sets of digital values representative of frequency spectra of the speech from samples of the speech in digital form;

generating one of a plurality of auditory state codes for each of the sets of digital values and supplying at least two sets of coordinate values in a mathematical space; and

electrically computing a series of other coordinate values defining a path with selected contributions from one or more of the sets of first named coordinate values depending on which auditory state code is generated.

32. Speech processing apparatus comprising:

memory means;

means for producing samples of an analog waveform of speech and converting the samples into digital form; and

means for deriving a set of digital values representative of a frequency spectrum of the speech from the samples in digital form, for selectively storing in distinct locations in the memory the values of frequency of one or more frequency peaks in the spectrum wherein a selected one or more of the distinct memory locations in which the frequency value of a given peak is stored depends on whether the peak lies in a first predetermined band of frequencies and on whether or not any other peak lies both in the first band and a second band overlapping the first band, and for generating a set of digital values corresponding to coordinate values in a mathematical space depending both on the stored values of frequency and on the distinct locations of the stored values of frequency.

33. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for storing as a lower first frequency the value of frequency of any lowest frequency peak in the first predetermined band of frequencies and as a higher first frequency the value of frequency of any next higher frequency peak in the first band, and for storing as a second frequency the value of frequency of any peak in the second band higher in frequency than the higher first frequency if

the higher first frequency is also in the second band, and if there is no peak in the second band higher in frequency than the higher first frequency when it is in the second band then storing as the second frequency the value of frequency originally stored as the higher first frequency and storing as the higher first frequency the value of frequency stored as the lower first frequency also.

34. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually deriving sets of digital values representative of frequency spectra of the speech wherein one of the frequency spectra has a set of three spectral prominences with first, second, and third frequencies in order of increasing frequency, for generating an auditory state code indicative of a glottal-source state for said one spectrum, and for producing the set of coordinate values to have respective values involving the ratio of the third to the second frequencies, the ratio of the first frequency to a reference frequency value, and the ratio of the second to the first frequencies.

35. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for identifying lower and higher first frequencies descriptive of a peak which is widened or split upon at least one occurrence of nasality and for producing a signal indicative of the occurrence of nasality.

36. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually deriving sets of digital values representative of frequency spectra of the speech wherein one of the frequency spectra has a set of three spectral prominences with first, second, and third frequencies in order of increasing frequency and another of the frequency spectra lacks the first one of the three prominences,

and for generating an auditory state code indicative of a burst-friction state for the other frequency spectrum and for producing the set of coordinate values with a first value involving the ratio of the third to the second frequencies, a second value which is substantially constant, and a third value involving the ratio of the second frequency to a reference frequency value.

37. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually deriving sets of digital values representative of frequency spectra of the speech, and for computing at least one of the values in the set of coordinate values as a function of a reference frequency value which is a function of frequency values determined from at least two of the spectra.

38. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually deriving sets of digital values representative of frequency spectra of the speech, and for computing at least one of the values in the set of coordinate values as a function of a reference frequency value which is a function of a geometric mean of frequency values determined from at least some periodic glottal-source spectra over time.

39. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for computing at least one of the values in the set of coordinate values as a function of a reference frequency which is a function of A) a frequency of pitch modulation of the speech and B) a mean of frequency values determined from the speech over time.

40. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for selecting values of end frequencies for the second band, the selected

values depending on whether or not a peak exists in the first predetermined band of frequencies.

41. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for selecting values of end frequencies for both the second band and a third band overlapping the second band, the selected values depending on whether or not a peak exists in the first predetermined band of frequencies.

42. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for selecting values of end frequencies for both the second band and a third higher band overlapping the second band and for determining whether or not one of the peaks is the only peak in the third band and lies in both the second and third bands, and if so, storing in one of the distinct locations another frequency value corresponding to an upper frequency edge of the one peak.

43. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually deriving sets of digital values representative of frequency spectra of the speech, and for selecting values of end frequencies for the second band as a function of a reference frequency value determined from at least two of the spectra.

44. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually deriving sets of digital values representative of frequency spectra of the speech, and for determining whether or not one of the peaks lies in a third band which is generally higher in frequency than the second band and overlaps it, and if none of the peaks lies in the third band, storing another frequency value in one of the distinct locations, the other frequency

value lying in the third band and being a function of a reference frequency value determined from at least two of the spectra.

45. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually generating additional sets of the digital values corresponding to coordinate values in the mathematical space as the speech continues over time, and the apparatus further comprises:

second memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the second memory, and

means for electrically generating a series of second coordinate values of points on a path in the mathematical space from the first named coordinate values, for identifying those second coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the position coordinate values and obtaining from said memory means the phonetic representation information prestored at that memory address.

46. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually generating additional sets of the digital values corresponding to coordinate values in the mathematical space as the speech continues over time, for generating one of a plurality of auditory state codes for each of said sets of digital values, and for computing a series of second coordinate values of points defining a path in the mathematical space with selected contributions from one or more of said first-named sets of digital coordinate values depending on which auditory state code is generated.

47. Speech processing apparatus as set forth in claim 32 wherein said deriving means includes means for continually generating additional sets of the digital values corresponding to coordinate values in the mathematical space as the speech continues over time and for generating one of a plurality of auditory state codes for each of said sets of digital values, the apparatus further comprising:

second memory means for holding prestored information indicative of different phonetic representations corresponding to respective sets of addresses in the second memory, and

means for electrically generating a series of second coordinate values of points on a path in the mathematical space with selected contributions from one or more of said first-named sets of digital coordinate values depending on which auditory state code is generated, for identifying those second coordinate values approximating at least one position along the path of a peak in magnitude of acceleration, generating a memory address as a function of the position coordinate values and obtaining from said memory means the phonetic representation information prestored at that memory address.

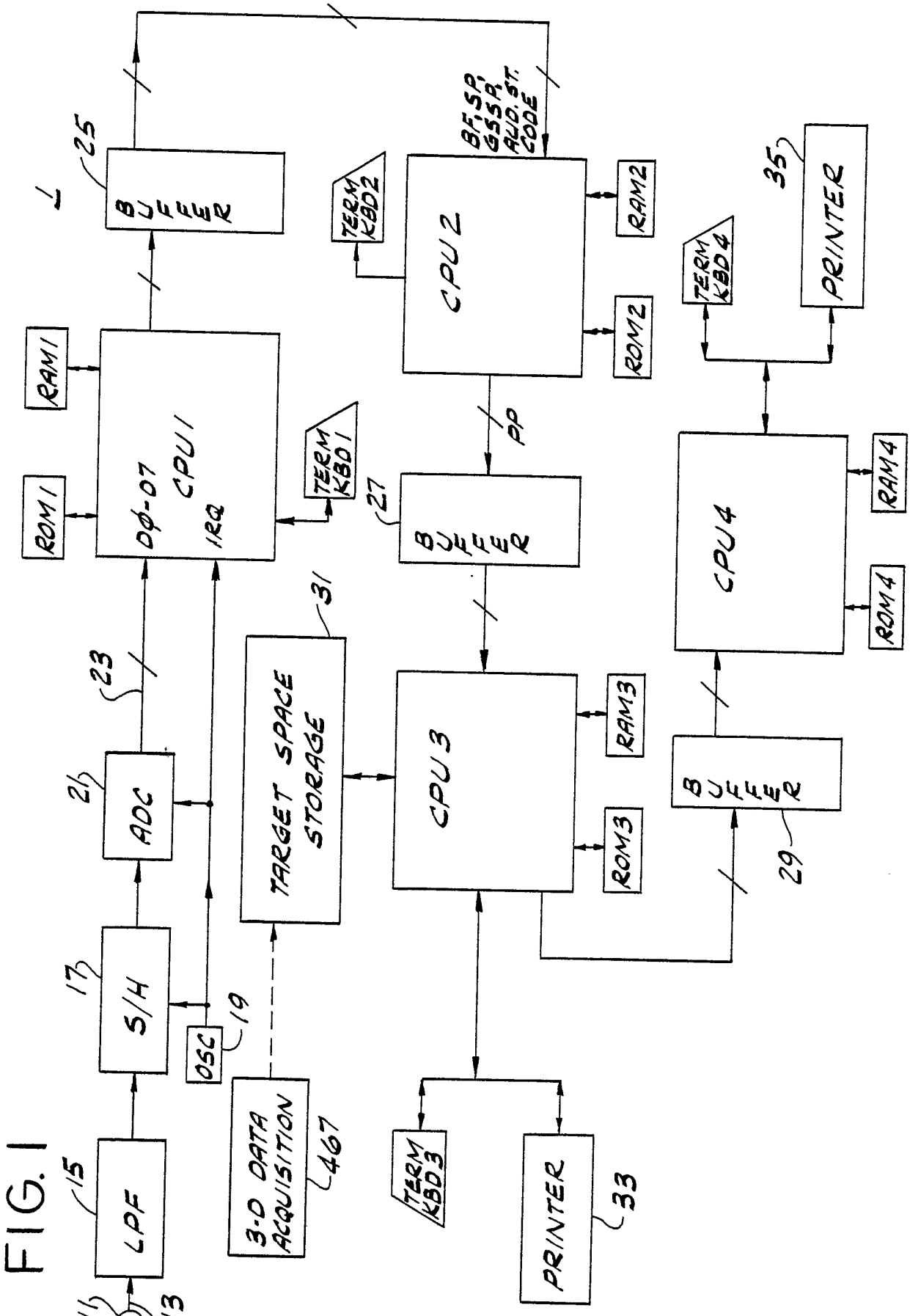

1
25
BUFFER
RAM1
ROM1
DØ-D7
CPU1
IRQ
TERM KBD1
23
21
ADC
17
S/H
OSC
19
15
LPF
11
13
3-D DATA ACQUISITION
467
TARGET SPACE STORAGE
31
TERM KBD2
CPU2
BF, SP, GSSR, AUD. ST. CODE
RAM2
ROM2
PD
27
BUFFER
CPU3
RAM3
ROM3
TERM KBD3
PRINTER
33
TERM KBD4
35
PRINTER
CPU4
RAM4
ROM4
BUFFER
29

FIG. 1

FIG. 2

FIG. 5

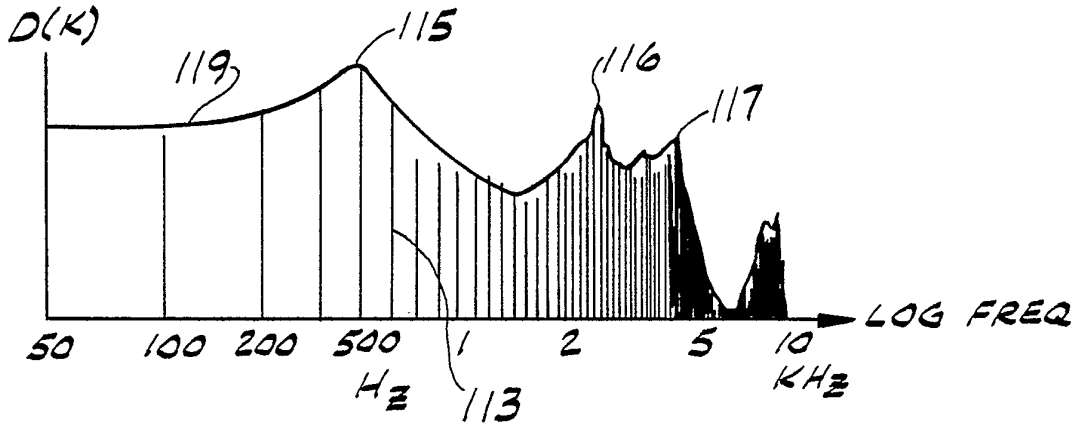


FIG. 5A

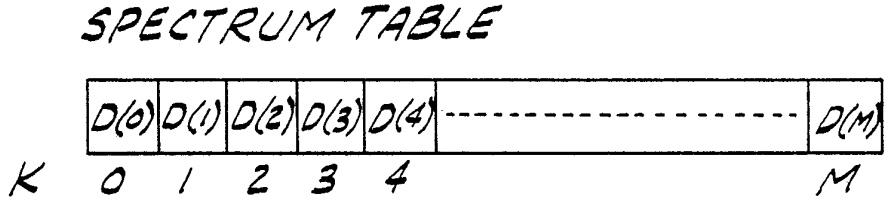


FIG. 3

CPU I INTERRUPT ROUTINE

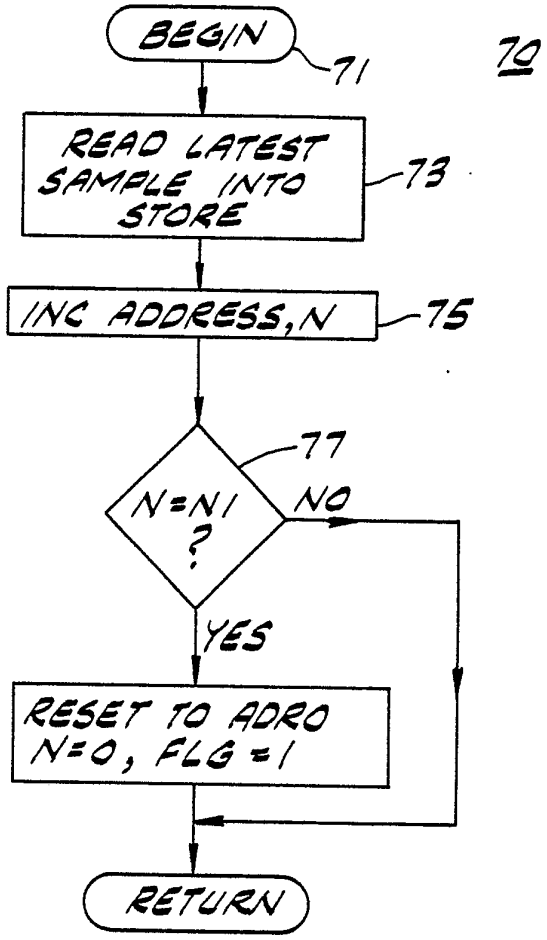


FIG.4

AUDITORY-SENSORY PROGRAM (CPU1)

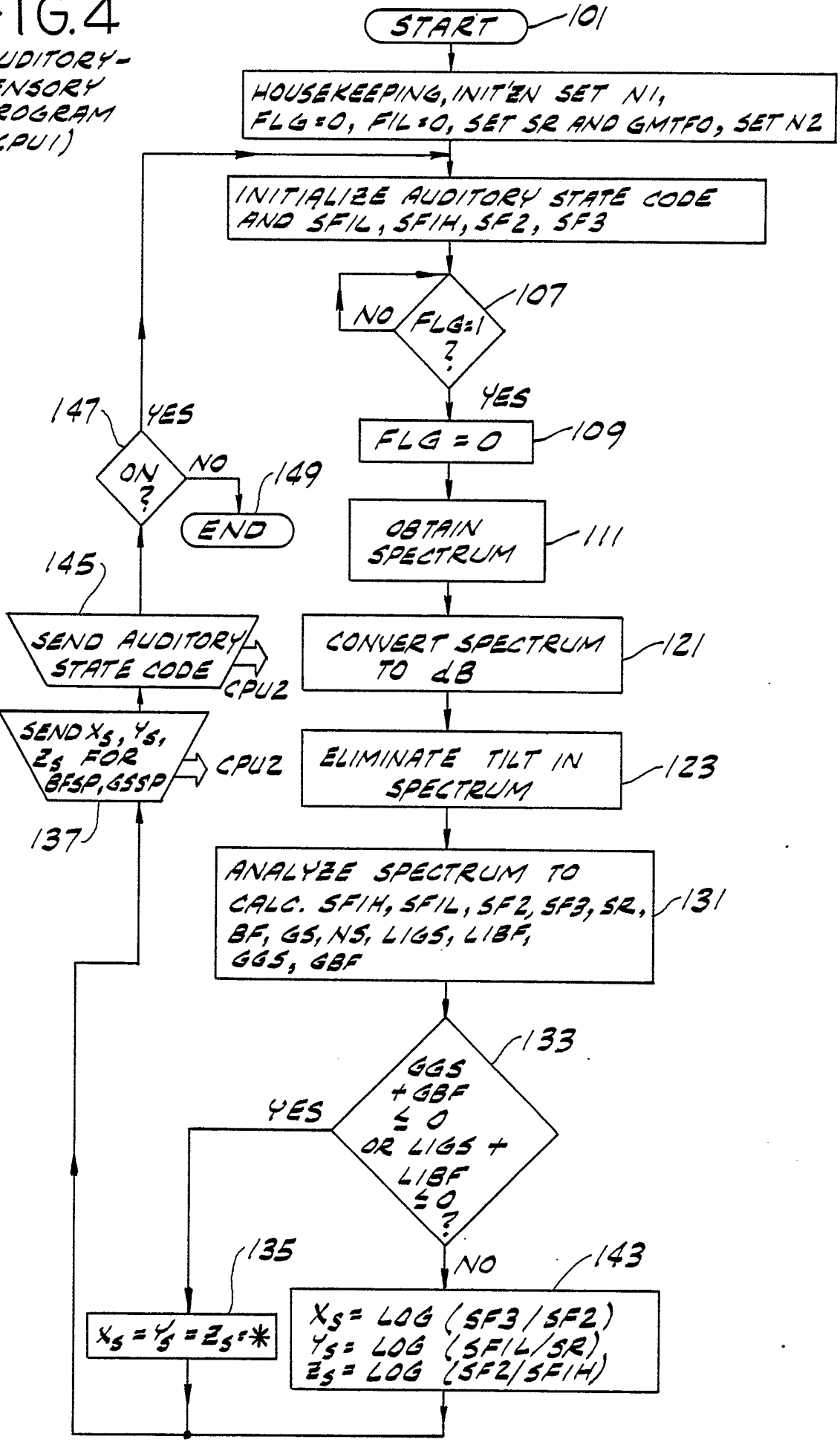


FIG. 6

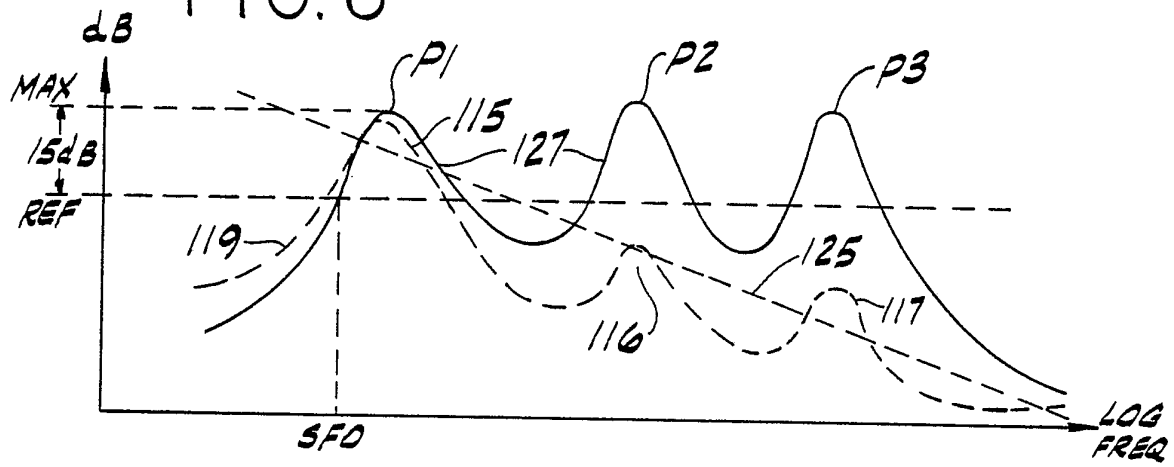


FIG. 7

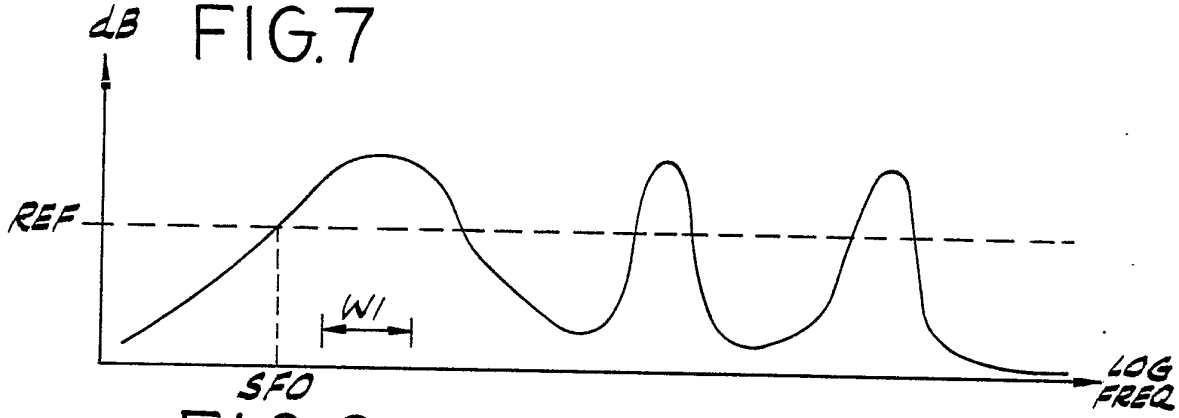


FIG. 8

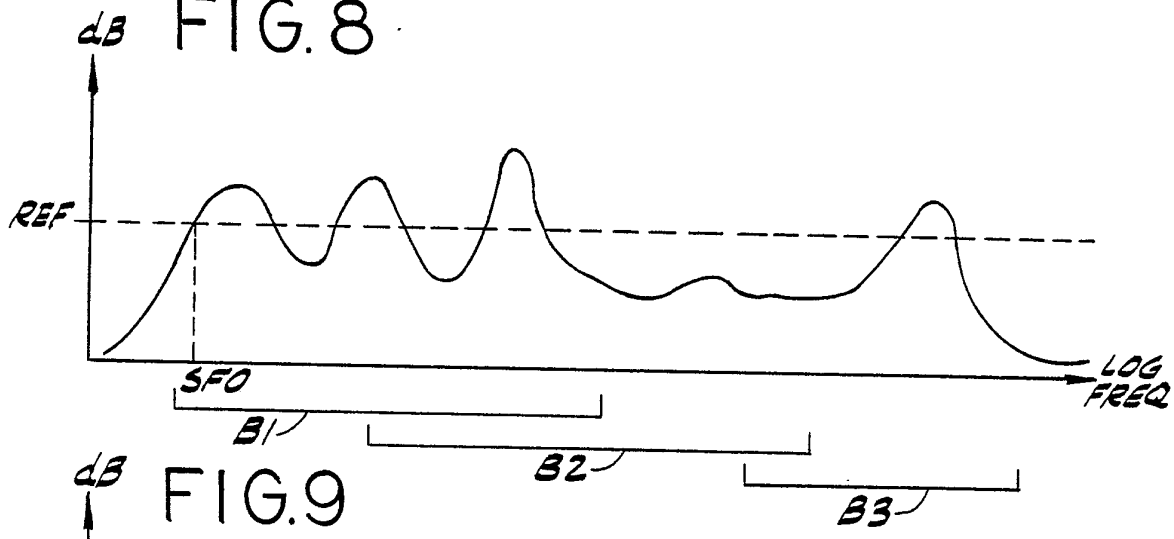


FIG. 9

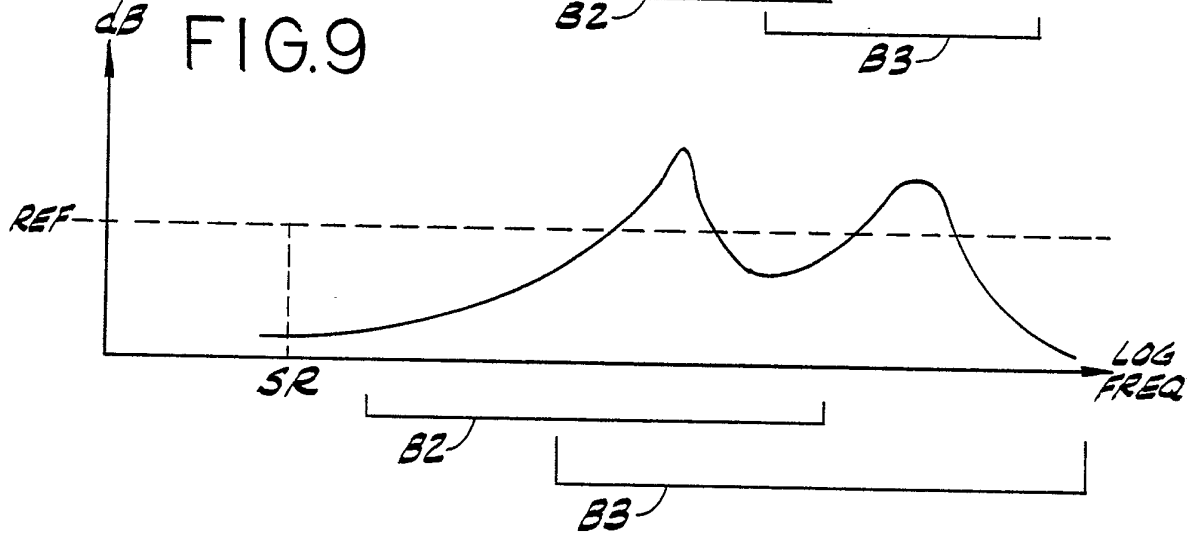


FIG. 10

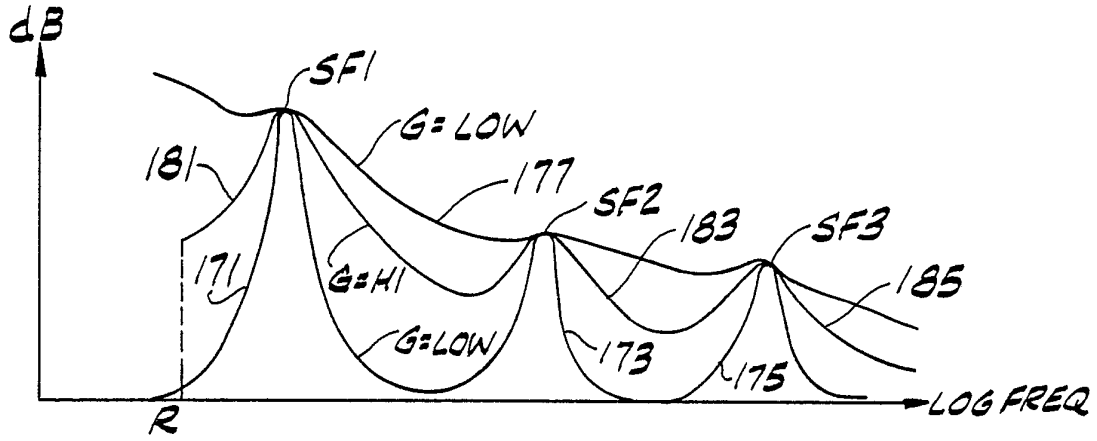


FIG. 12

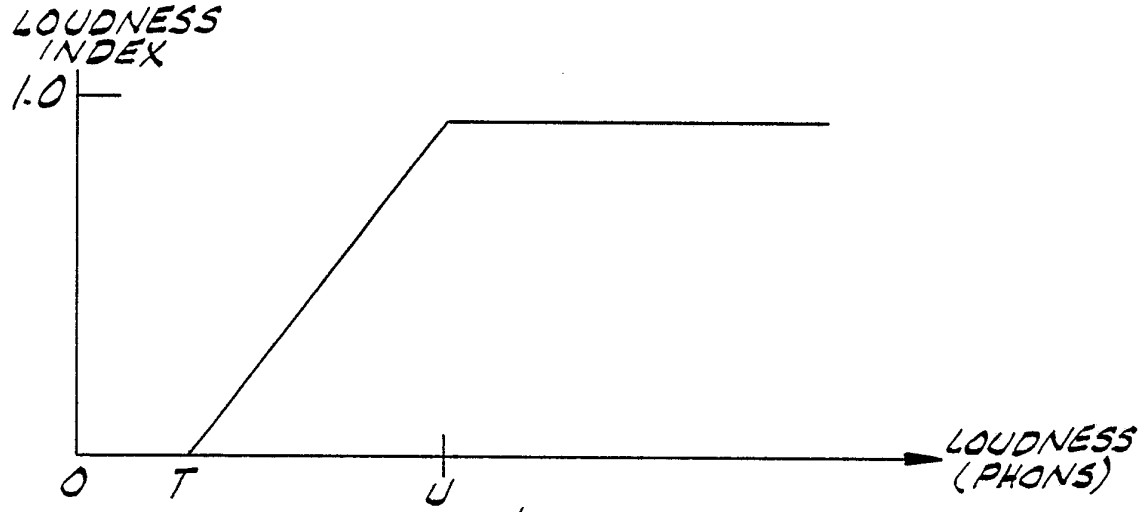


FIG. 18

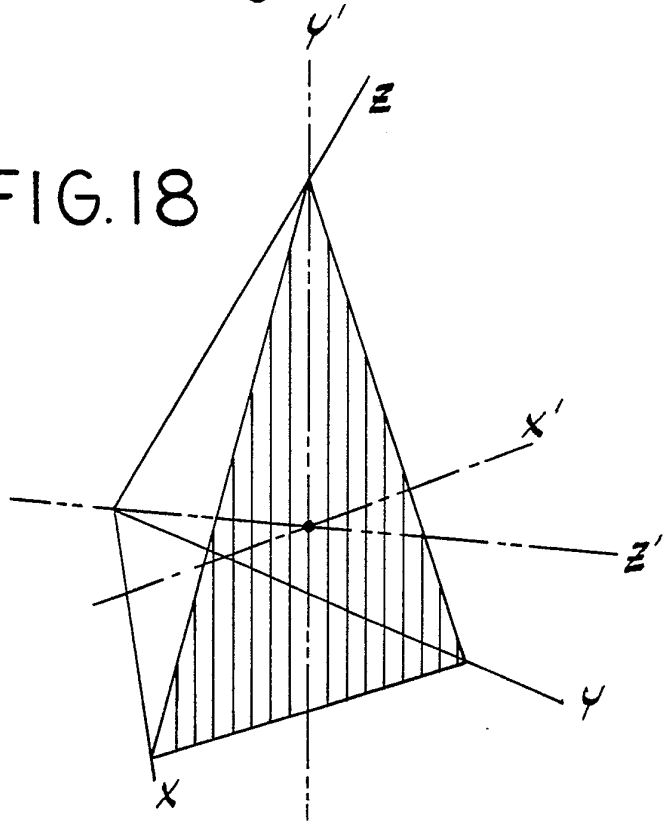


FIG.13A

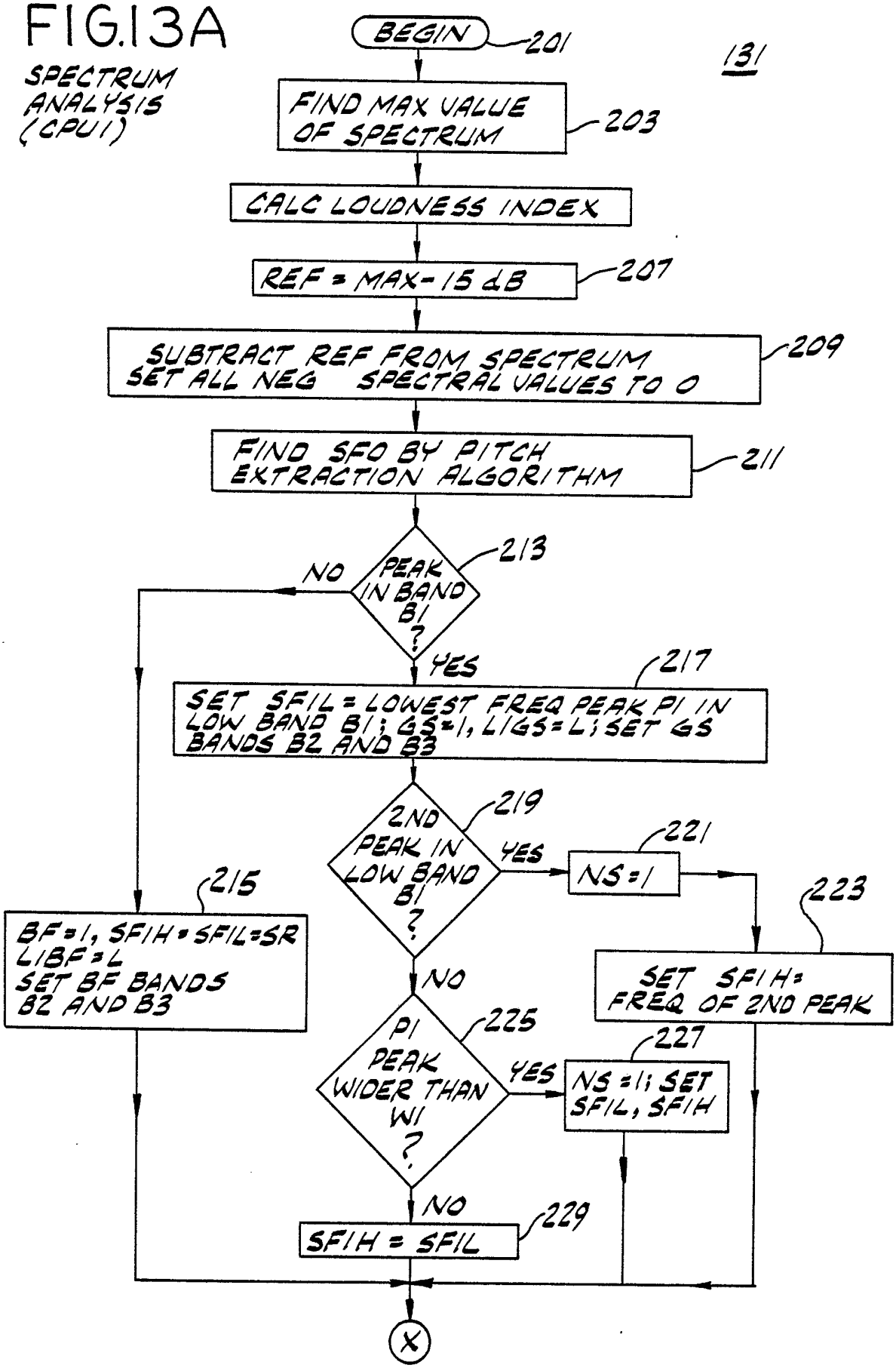

SPECTRUM ANALYSIS (CPU1)
131
BEGIN
201
FIND MAX VALUE OF SPECTRUM
203
CALC LOUDNESS INDEX
REF = MAX - 15 dB
207
SUBTRACT REF FROM SPECTRUM SET ALL NEG SPECTRAL VALUES TO 0
209
FIND SFO BY PITCH EXTRACTION ALGORITHM
211
PEAK IN BAND B1 ?
213
NO
YES
SET SFIL = LOWEST FREQ PEAK PI IN LOW BAND B1; GS=1, LIGS=L; SET GS BANDS B2 AND B3
217
2ND PEAK IN LOW BAND B1 ?
219
YES
NS=1
221
NO
BF=1, SFIH=SFIL=SR LIBF=L SET BF BANDS B2 AND B3
215
SET SFIH= FREQ OF 2ND PEAK
223
PI PEAK WIDER THAN W1 ?
225
YES
NS=1; SET SFIL, SFIH
227
NO
SFIH = SFIL
229
X

FIG.13B
(CPU1)

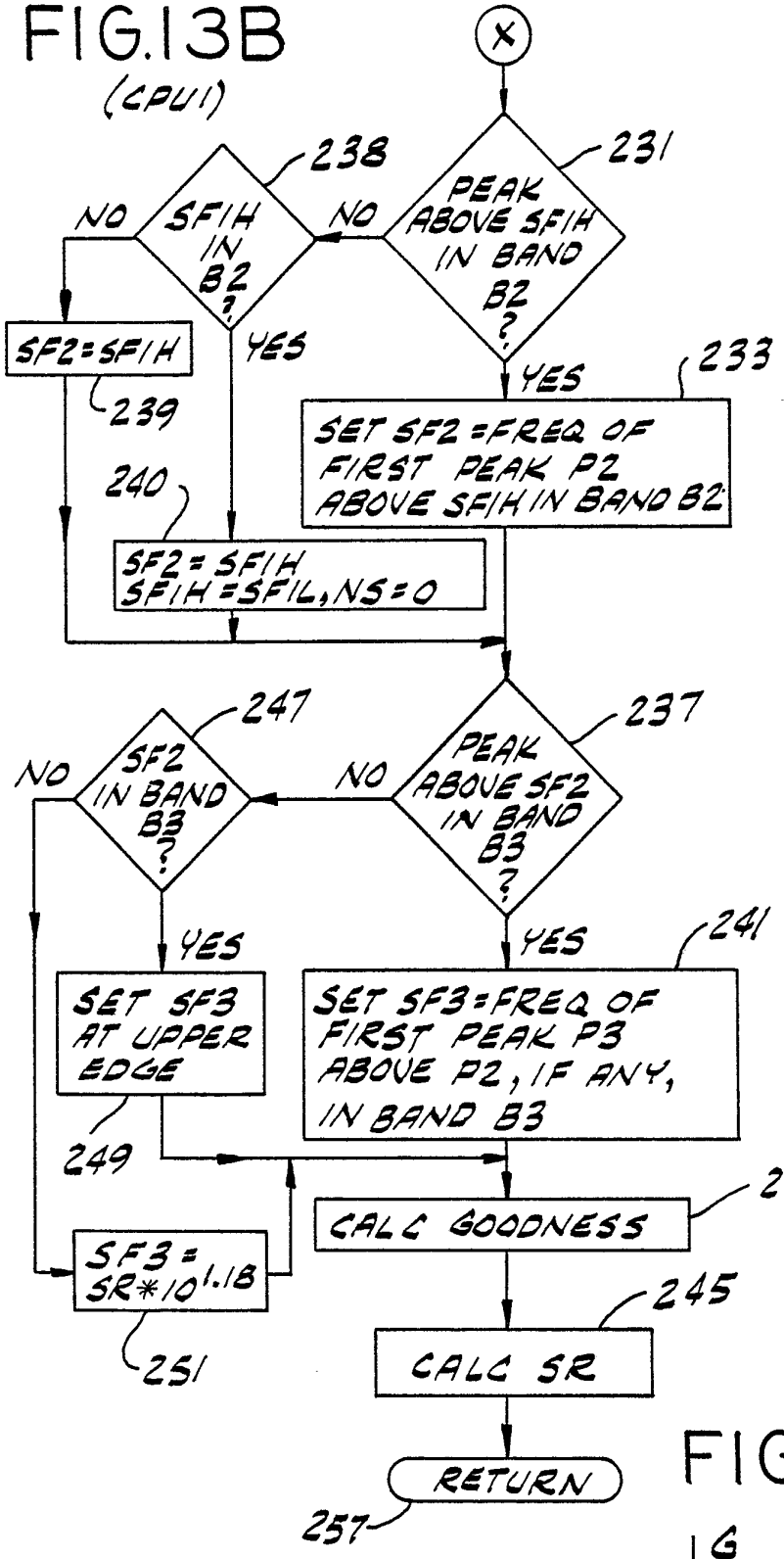


FIG.11

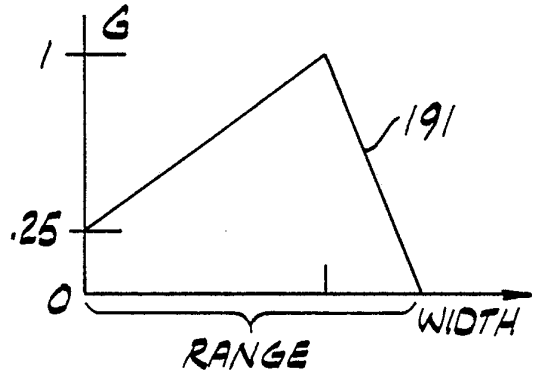


FIG. 14

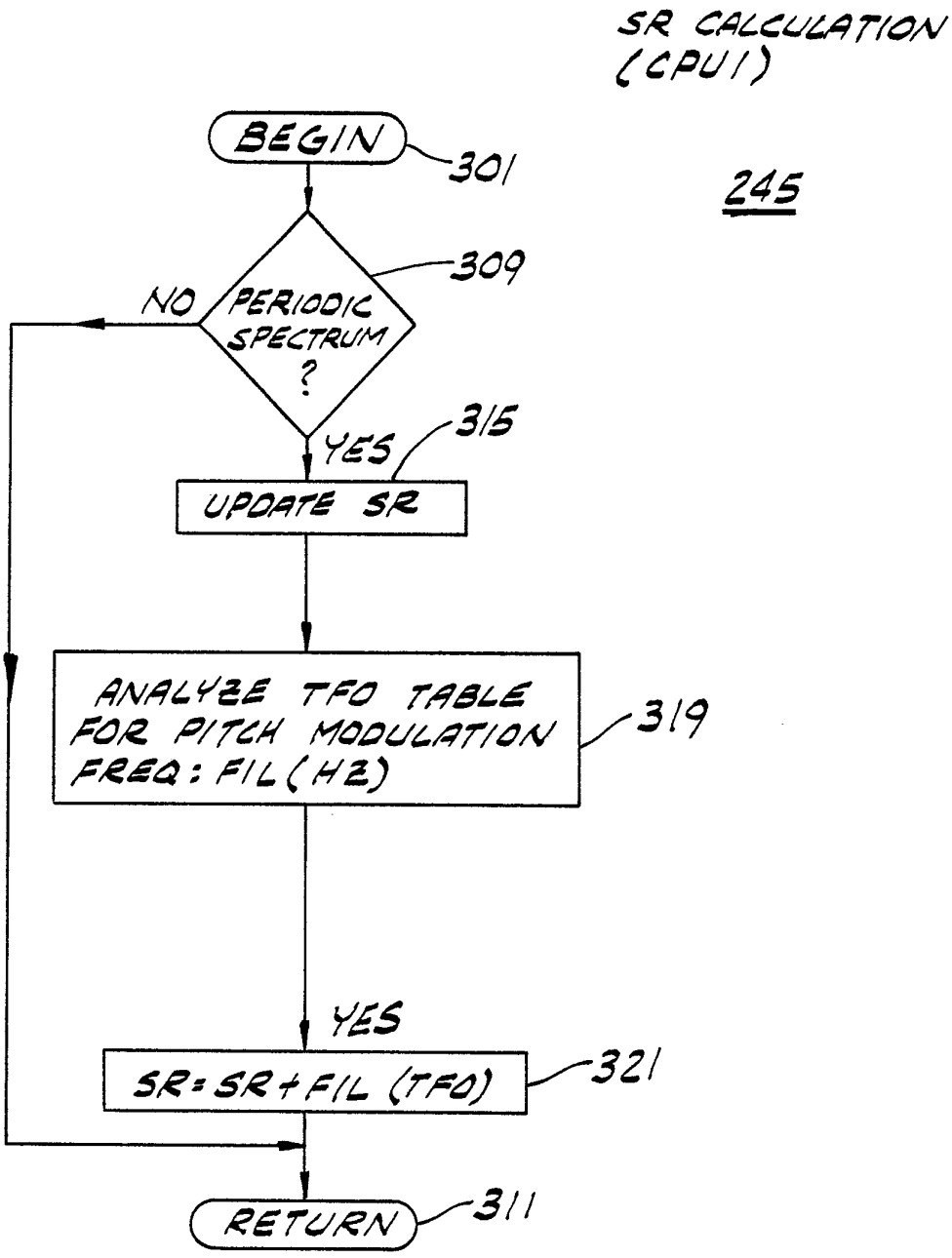

SR CALCULATION (CPU1)
245
BEGIN
301
PERIODIC SPECTRUM ?
309
NO
YES
315
UPDATE SR
ANALYZE TFO TABLE FOR PITCH MODULATION FREQ: FIL (HZ)
319
YES
SR = SR + FIL (TFO)
321
RETURN
311

FIG.15

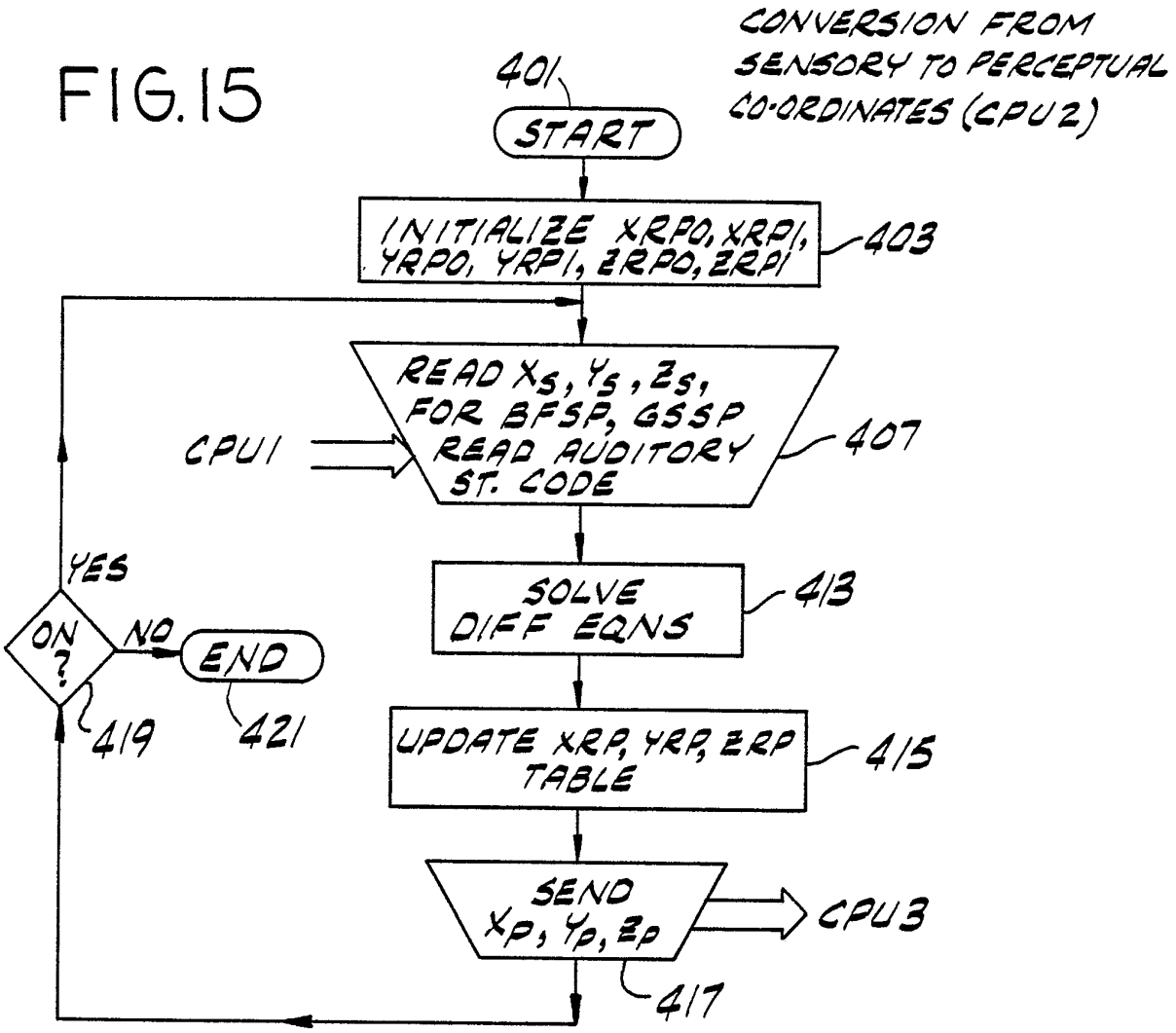


FIG.15A

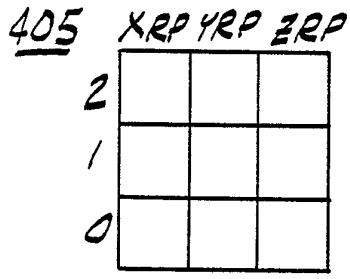


FIG.27

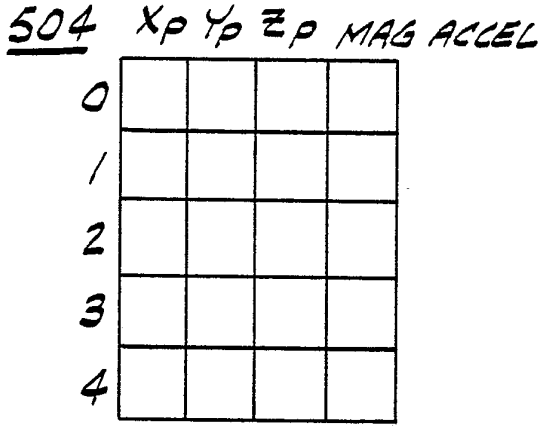


FIG. 16

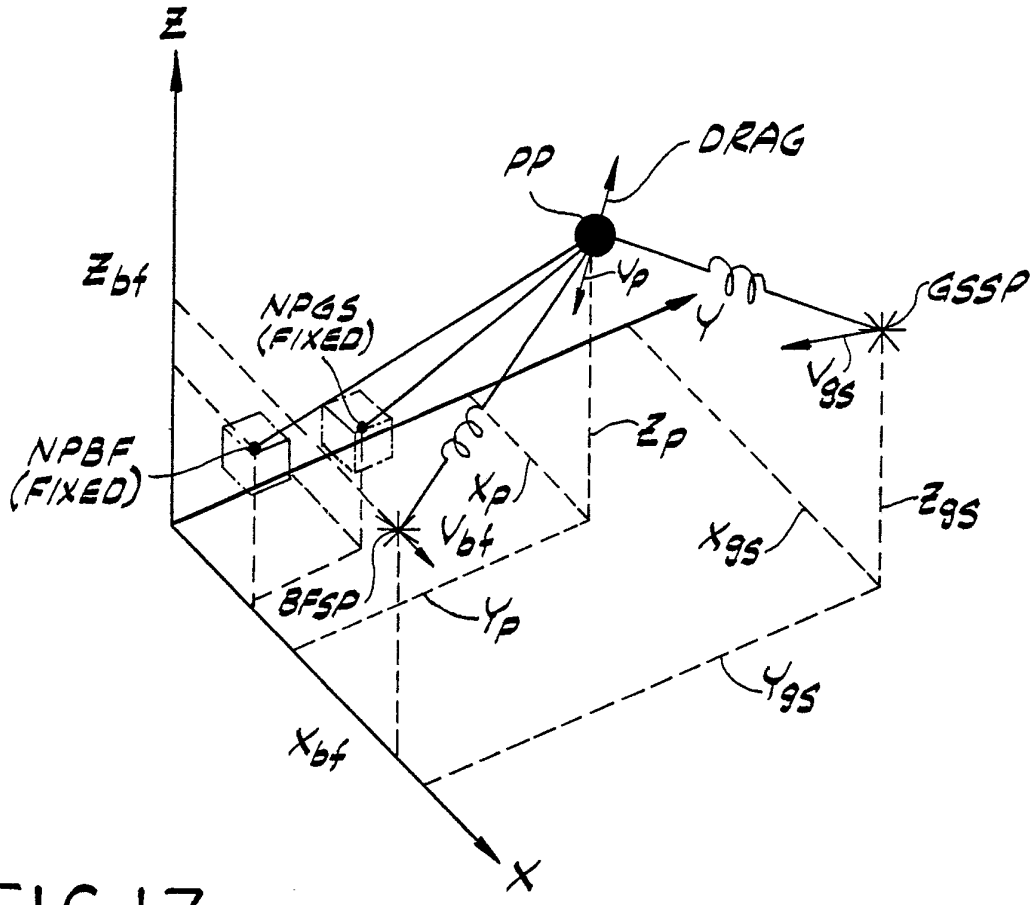

Z
PP
DRAG
Z_{bf}
V_p
Y
GSSP
NPGS
(FIXED)
V_{gs}
NPBF
(FIXED)
Z_p
X_p
Z_{gs}
V_{bf}
X_{gs}
BFSP
Y_p
Y_{gs}
X_{bf}
X

FIG. 17

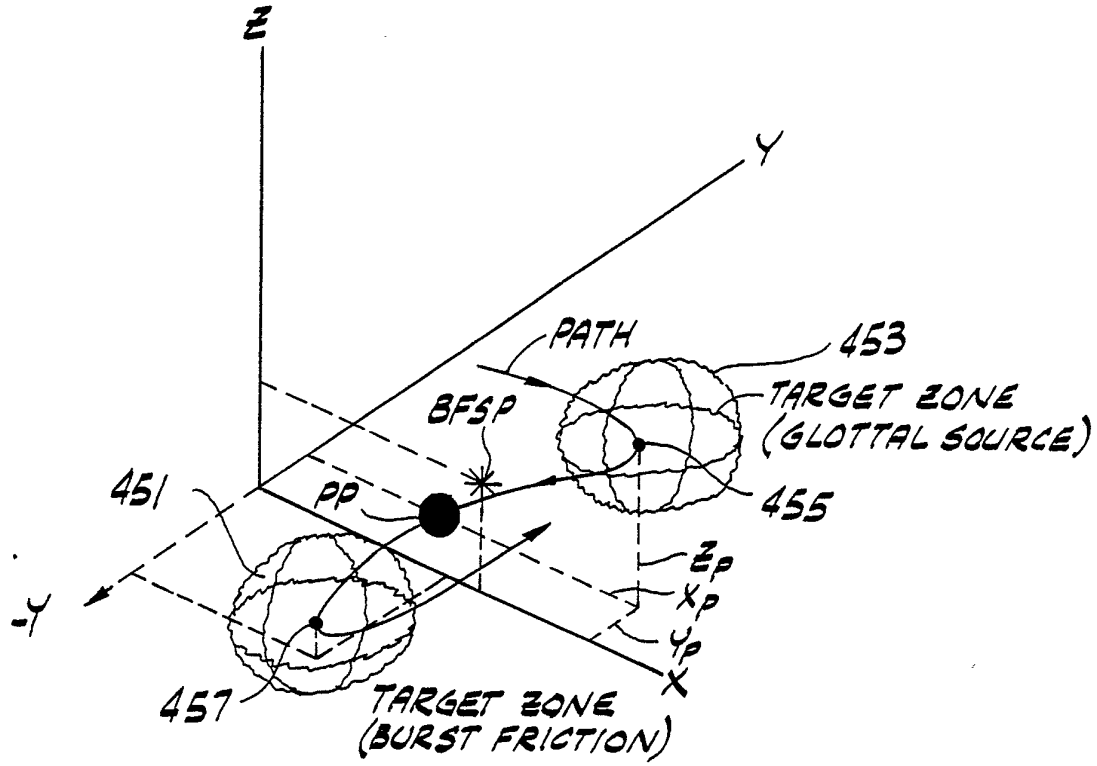

Z
Y
PATH
453
BFSP
TARGET ZONE
(GLOTTAL SOURCE)
451
PP
455
Z_p
X_p
-Y
Y_p
X
457
TARGET ZONE
(BURST FRICTION)

FIG. 20

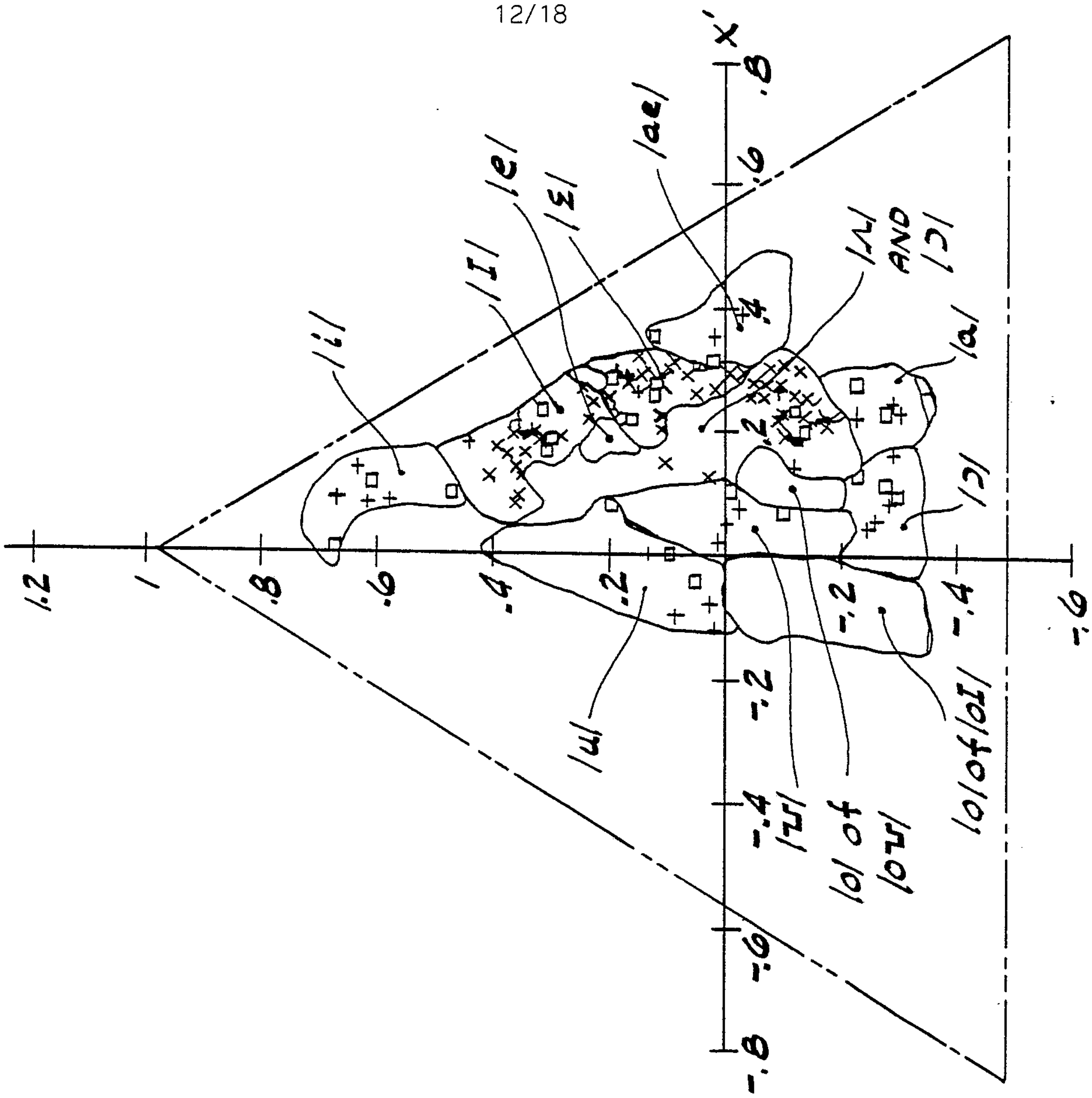


FIG.19

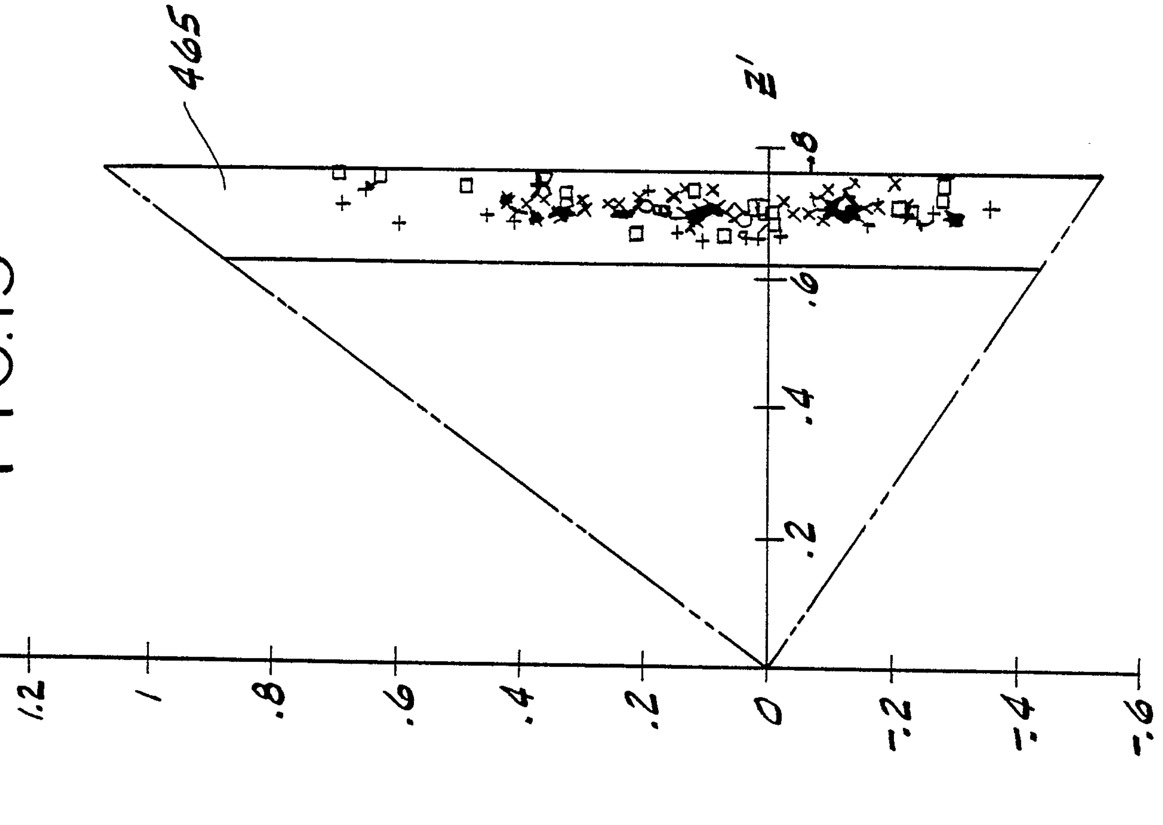


FIG. 21

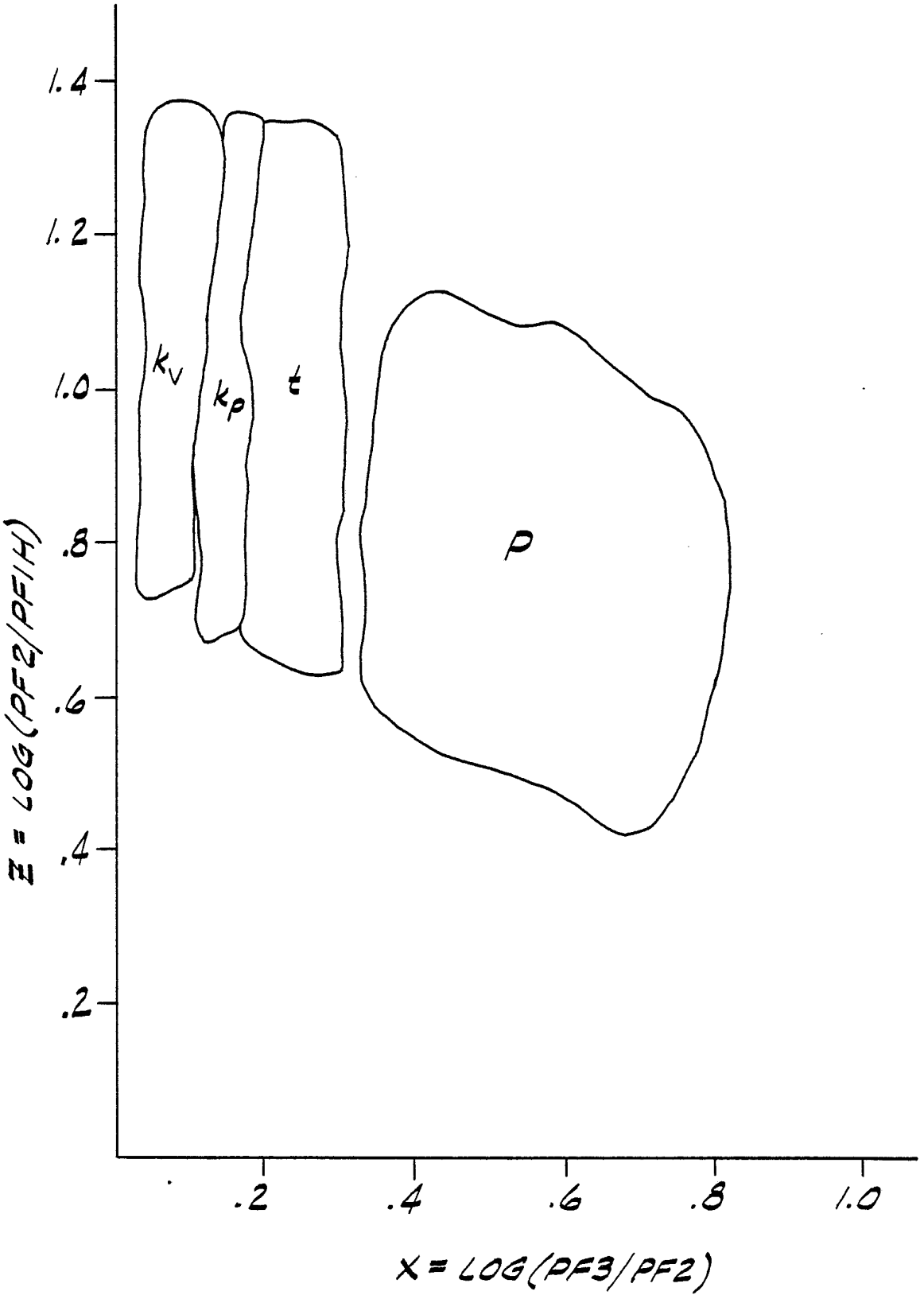


FIG. 22

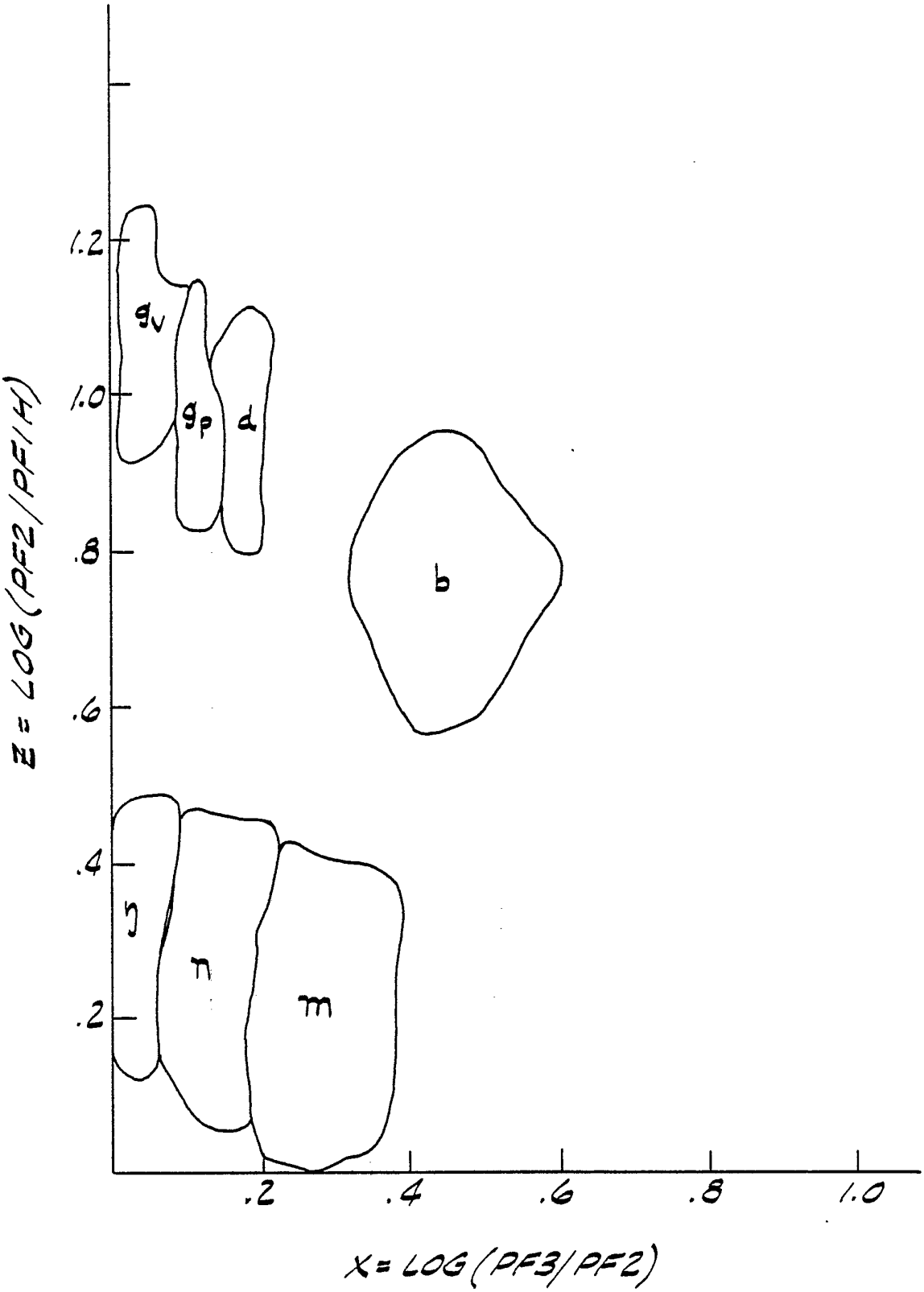


FIG.23

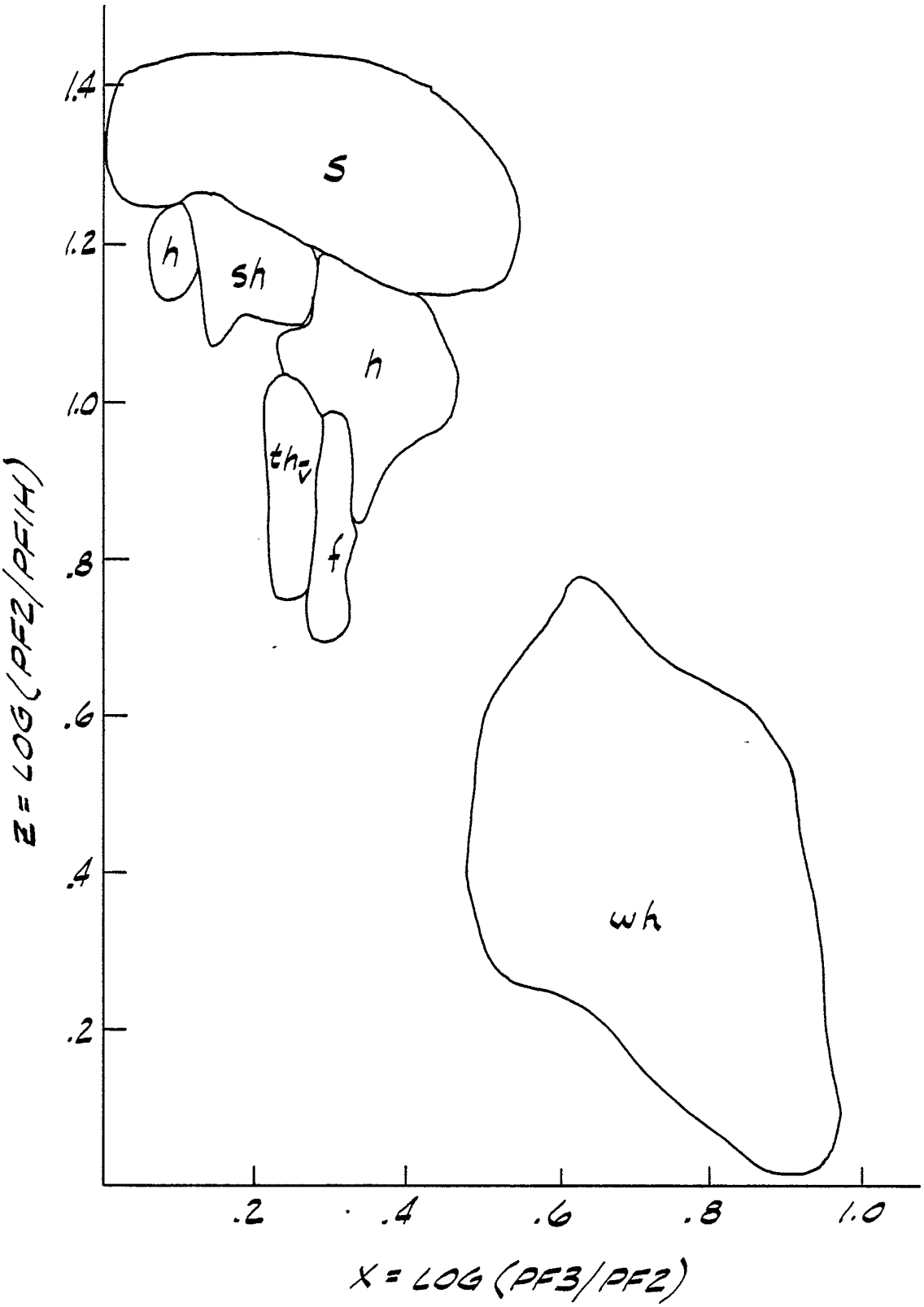


FIG. 24

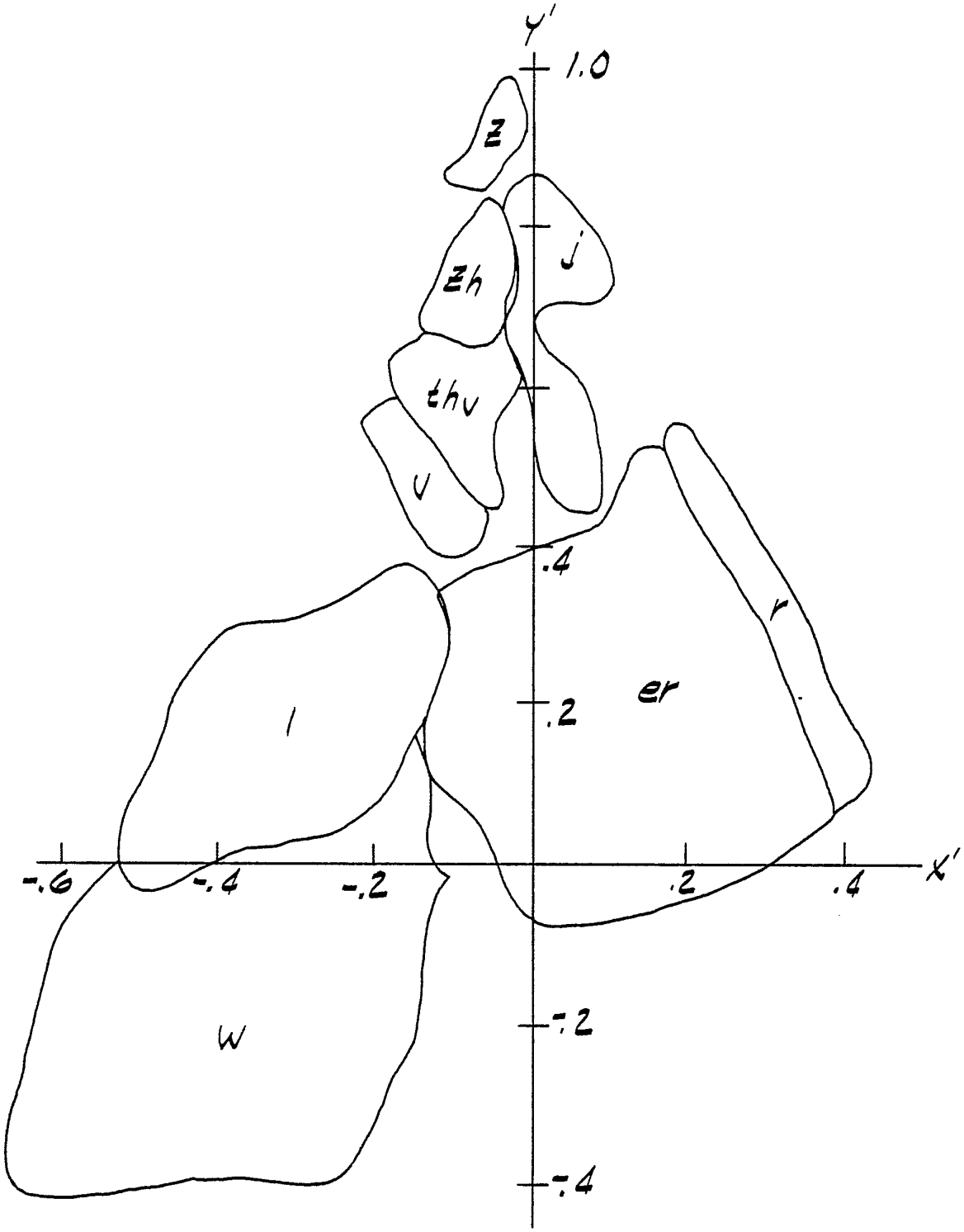


FIG. 25

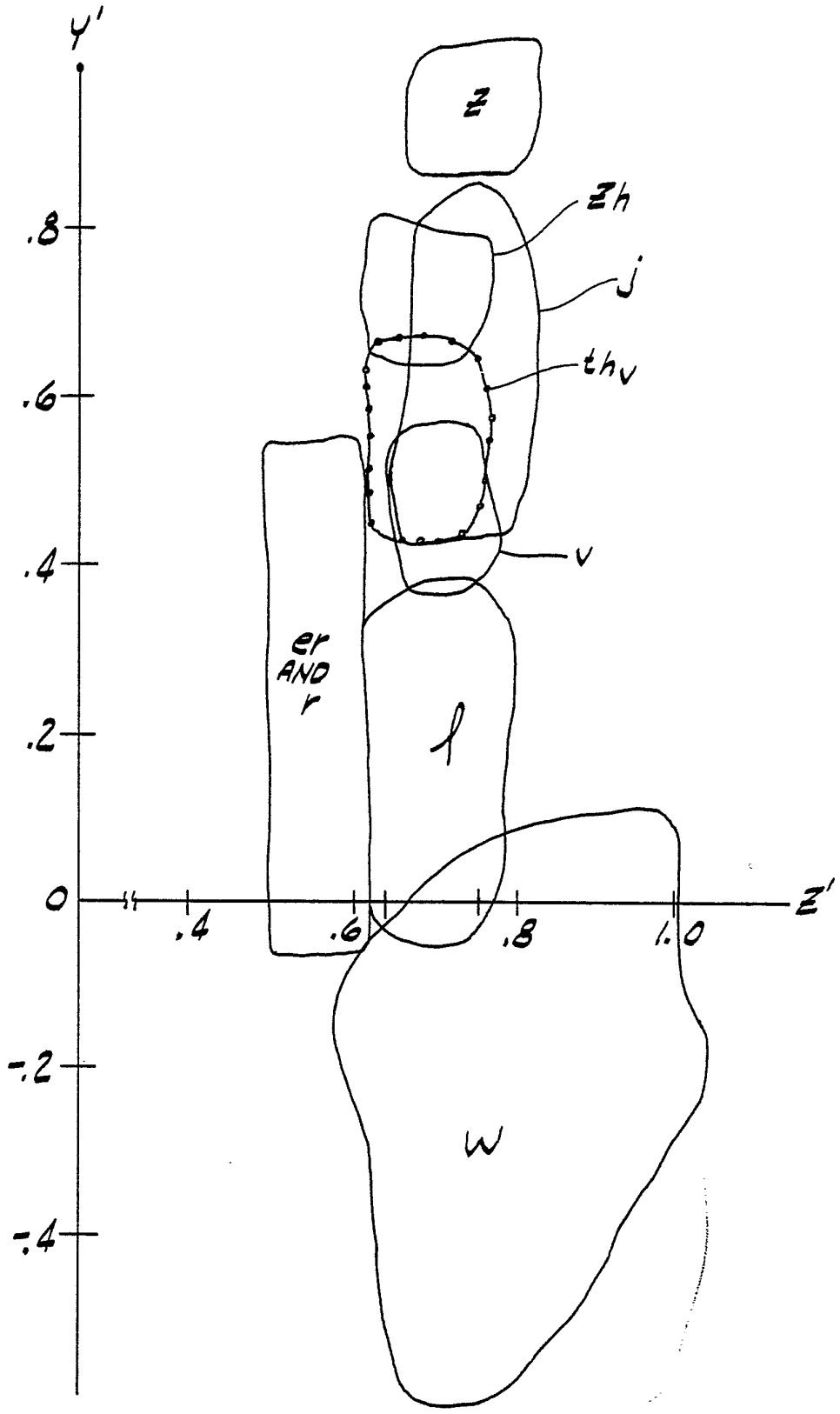
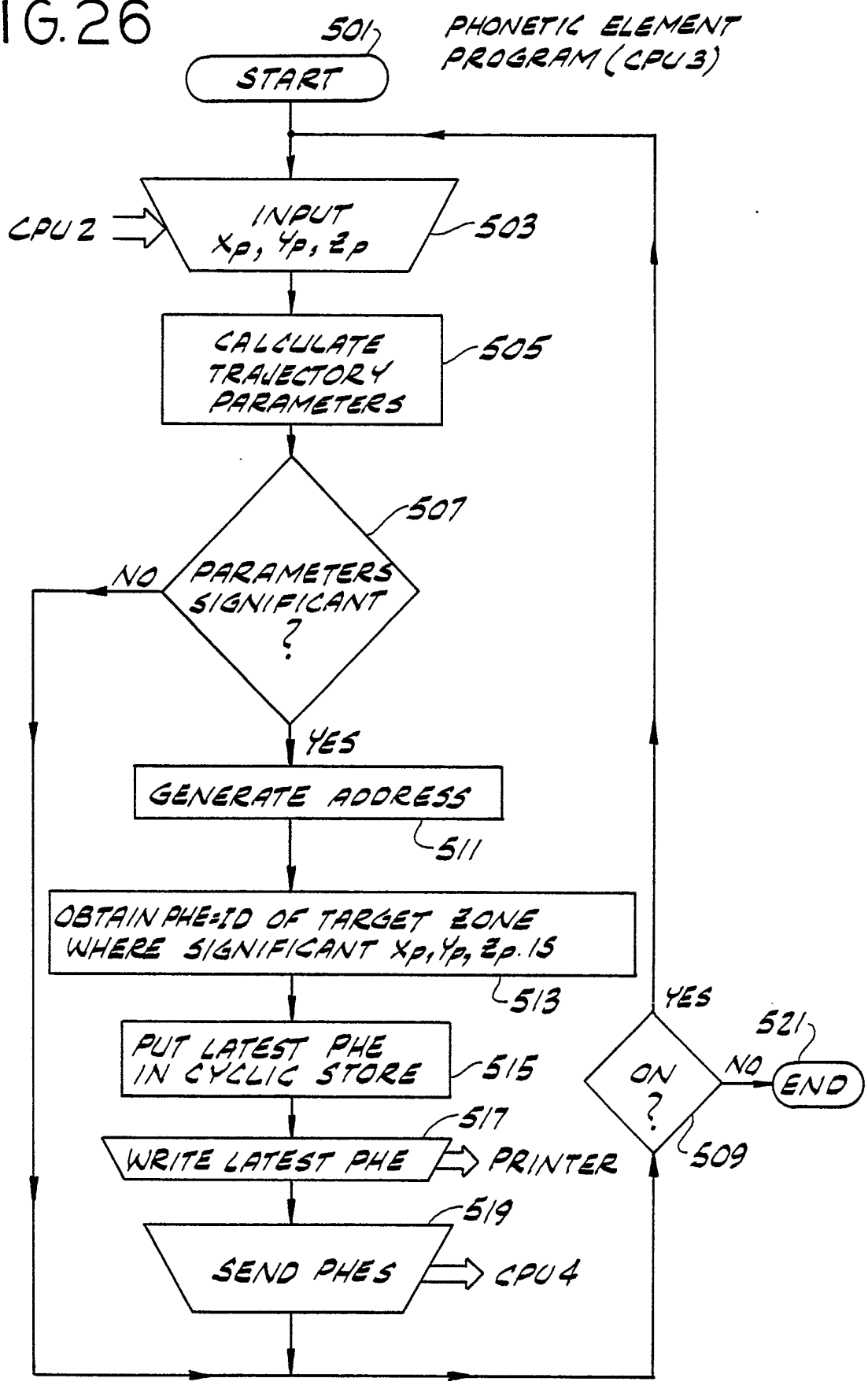


FIG. 26


501
PHONETIC ELEMENT PROGRAM (CPU 3)
START
CPU 2
INPUT Xp, Yp, Zp
503
CALCULATE TRAJECTORY PARAMETERS
505
507
NO
PARAMETERS SIGNIFICANT ?
YES
GENERATE ADDRESS
511
OBTAIN PHE-ID OF TARGET ZONE WHERE SIGNIFICANT Xp, Yp, Zp. IS
513
PUT LATEST PHE IN CYCLIC STORE
515
517
WRITE LATEST PHE
PRINTER
519
SEND PHES
CPU 4
YES
ON ?
509
NO
521
END

INTERNATIONAL SEARCH REPORT

International Application No PCT/US86/02313

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) [3]

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC (4): G10L 5/00

U.S. CL. 364/513.5; 381/43

II. FIELDS SEARCHED

Minimum Documentation Searched [4]

Classification System	Classification Symbols
U.S.	364/513.5; 381/39-53

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched [5]

III. DOCUMENTS CONSIDERED TO BE RELEVANT [14]

Category *	Citation of Document, [16] with indication, where appropriate, of the relevant passages [17]	Relevant to Claim No. [18]
Y	US, A, 4,087,632, (HAFER), 02 MAY 1978 SEE COL. 1, lines 44-60.	1-47
Y	US, A, 3,679,830, (UFFELMAN), 25 JULY 1972 SEE COL. 1, LINES 15-21.	16, 34
Y	J. FLANAGAN, "SPEECH ANALYSIS, SYNTHESIS AND PERCEPTION", SECOND EDITION, PUBLISHED 1972 BY SPRINGER-VERLAG (BERLIN), SEE PAGES 194, 196, 202.	1-47

* Special categories of cited documents: [15]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search [2]	Date of Mailing of this International Search Report [2]
17 DECEMBER 1986	05 JAN 1987
International Searching Authority [1]	**Signature of Authorized Officer [20]**
ISA/US	S S Kemeny E. S. KEMENY