(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 143 161**
A1

(12) **EUROPEAN PATENT APPLICATION**

(21) Application number: **84107846.2**

(22) Date of filing: **05.07.84**

(51) Int. Cl.⁴: **G 10 K 15/04**

(30) Priority: **08.07.83 US 512068**

(43) Date of publication of application: **05.06.85**
Bulletin 85/23

(84) Designated Contracting States: **AT DE FR GB**

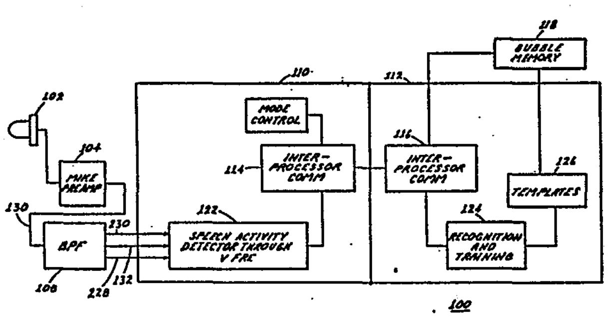
(71) Applicant: **International Standard Electric Corporation, 320 Park Avenue, New York New York 10022 (US)**

(72) Inventor: **Hutchins, Sandra E., 13674 Boquita Drive, Del Mar California (US)**
Inventor: **Boll, Steven F., 10284 Gray Fox Drive, San Diego California (US)**
Inventor: **Vensko, George, 16927 Handlebar Road, Ramona California (US)**
Inventor: **Carlin, Lawrence, 1940 East Duell Street, Glen Dora, California 91740 (US)**
Inventor: **Smith, Allen R., 27 Buttercup Lane, Huntington Connecticut (US)**

(74) Representative: **Graf, Georg Hugo, Dipl.-Ing. et al, c/o Standard Elektrik Lorenz AG Patent- und Lizenzwesen Postfach 300 929 Kurze Strasse 8, D-7000 Stuttgart 30 (DE)**

(54) Apparatus for automatic speech activity detection.

(57) An apparatus and method for automatic detection of speech signals in the presence of noise including noise events occurring when speech is not present and having signals whose signal strengths are substantially equal to or greater than the speech signals. Frames of data representing digitized output signals from a plurality of frequency filters are operated on by a linear feature vector to create a scalar feature for each frame which is indicative of whether the frame is to be associated with speech signals or noise event signals. The scalar features are compared with a detection threshold value which is created and updated from a plurality of previously stored scalar features. A plurality of the results of the comparison for a succession of frames is stored and the stored results combined in a predetermined way to obtain an indication of when speech signals are present. In automatic speech recognizers employing the above-described speech detections, when such indication is given, frames are further preprocessed and then compared with stored templates in accordance with the dynamic programming algorithm in order to recognize which word was spoken.



EP 0 143 161 A1

0143161

S.E. Hutchins et al 1-1-4-3-3

APPARATUS FOR AUTOMATIC SPEECH ACTIVITY DETECTION

This invention relates to an apparatus and method for speaker independent speech activity detection in an environment of relatively high level noise, and to automatic speech recognizers which use such speaker independent speech activity detection. This invention also relates to U.S application Serial No. 473,422, filed March 9, 1983, entitled "Apparatus and Method for Automatic Speech Recognition", assigned along with this application to a common assignee, and hereby incorporated by reference as if specifically set forth herein.

Automatic speech recognition systems provide a means for man to interface with communication equipment, computers and other machines in a human's most natural and convenient mode of communication. Where required, this will enable operators of telephones, computers, etc. to call others, enter data, request information and control systems when their hands and eyes are busy, when they are in the dark, or when they are unable to be stationary at a terminal.

S.E.Hutchins et al 1-1-4-3-3

One known approach to automatic speech recognition involves the following: periodically sampling a bandpass filtered (BPF) audio speech input signal to create frames of data and then preprocessing the data to convert them to processed frames of parametric values which are more suitable for speech processing; storing a plurality of templates (each template is a plurality of previously created processed frames of parametric values representing a word, which when taken together form the reference vocabulary of the automatic speech recognizer); and comparing the processed frames of speech with the templates in accordance with a predetermined algorithm, such as the dynamic programming algorithm (DPA) described in an article by F. Itakura, entitled "Minimum prediction residual principle applied to speech recognition", IEEE Trans. Acoustics, Speech and Signal Processing, Vo. ASSP-23, pp. 67-72, February 1975, to find the best time alignment path or match between a given template and the spoken word.

Automatic Speech Recognizers depend on detecting the end points of speech based on measurements of energy. Prior art speech activity detectors discriminate between energy, assumed to be speech, and lack of energy, assumed to be silence. Therefore, prior art Automatic Speech Recognizers require a relatively quiet environment in which to operate, otherwise, performance in terms of recognition accuracy drops drastically. Requiring a quiet environment restricts the uses to which a Speech Recognizer can be put, for example, prior art recognizers would have difficulty operating on a noisy factory floor or in a cockpit of a tactical aircraft, etc. Such noisy environments as these can be characterized as having background noise present whether or not speech is present and noise events occurring when speech is not present, the noise events sometimes having signal levels equal to or greater than the speech signal levels.

It is the object of the invention, therefore, to provide an apparatus for speaker independent speech activity detection and for such speech activity detection for use in automatic speech recognizers which must operate in an environment wherein noise events with relatively high signal levels occur when speech is nor present.

This object is achieved as set forth in claim 1. Further embodiements are set forth in the subclaims.

The invention can be summarized as follows:

The input signals are digitized and frames of digital signal values associated with said digitized signals are repeatedly formed. The speech signals and noise event signals are automatically separated. In the preferred embodiment, this is done with a speaker independent predefined, fixed operation or transformation performed on the frames.

Also, in the preferred embodiment, the input signals are frequency filtered to provide a plurality of filter output signals which are then digitized. The frames are created from the digitized filter output signals. A linear transformation is applied to the frames of digital signal values to create a scalar feature for each frame whose magnitude will be larger for speech signals than for noise event signals.

A detection threshold value is created for the scalar feature magnitudes and repeatedly updated. Scalar features are compared with the detection threshold value, and the results of a plurality of successive comparisons are stored. The stored results are combined in a predetermined manner to obtain an indication of when speech signals are present.

When an indication that speech signals are present is given, frames are further preprocessed before being compared with stored templates representing the vocabulary of recognizable words. The comparison is based on the dynamic programming algorithm (DPA).

Objects, features and advantages of the present invention will become more fully apparent from the following detailed description of the preferred embodiment, the appended claims and the accompanying drawings, in which:

Fig. 1 is a preferred embodiment block diagram of the automatic speech recognition apparatus of the present invention.

Fig. 2 is a more detailed block diagram of the bandpass filter portion of the invention of Fig. 1.

Fig. 3 is a table giving the filter characteristics of the bandpass filter portion of Fig. 2.

Fig. 4 is a preferred embodiment block diagram of the operation of the speech recognition algorithm of the present invention.

Fig. 5 is a graph summarizing the time alignment and matching of the recognition portion of the speech recognition algorithm of Fig. 4.

Fig. 6 shows three graphs of amplitude vs. frequency for voice, jet noise and oxygen regulator noise.

Fig. 7 is a more detailed block diagram of the speech activity detector portion of the speech recognition algorithm of Fig. 4.

Fig. 1 is a block diagram of an automatic speech recognizer apparatus designated generally 100. It comprises a microphone 102; a microphone preamplifier circuit 104; a bandpass filter bank circuit 108 for providing a digital spectrum sampling of the audio output of circuit 104; a pair of processors 110 and 112 interconnected by inter-processor communication circuits 114 and 116; and an external non-volatile memory device 118. In the preferred embodiment, processors 110 and 112 are Motorola MC68000 microprocessors and inter-processor communication circuits 114 and 116 are conventionally designed circuits for handling interrupts and data transfers between MC68000 microprocessors. Interrupt procedures for the MC68000 are adequately described in the MC68000 specification.

The speech recognition algorithm is stored in the EPROM memory portions 122 and 124 of the processors 110 and 112, respectively, while the predefined vocabulary is stored as previously created templates in the external non-volatile memory device 118 which in the preferred embodiment is an Intel bubble memory, Model No. 7110, capable of storing one million bits. In the preferred embodiment, there are only 36 words in the vocabulary, and, hence, 36 templates with 4000 bits required per template on the average. Hence, the bubble memory is capable of storing approximately 250 templates. When templates are needed for comparison with incoming frames of speech data from BPF circuit 108, they are brought from memory 118 into working memory 126 in processor 112.

Referring now to Fig. 2, a more detailed block diagram of the bandpass filter bank circuit 108 is shown. The output from preamp 104 on lead 130 from Fig. 1 is transmitted to an input amplifier stage 200 which has a 3 db bandwidth of 10kHz. This is followed by a 6 db per octave preemphasis amplifier 202 having selectable cut in frequencies of 500 or 5000 Hz. This is conventional practice to provide more gain at the higher frequencies than at the lower frequencies since the higher frequencies are generally lower in amplitude in speech data. At the output of amplifier 202 the signal splits and is provided to the inputs of anti-aliasing filters 204 (with a cutoff frequency of 1.4 kHz) and 206 (with a cutoff frequency of 10.5 KHz). These are provided to eliminate aliasing which may result because of subsequent sampling.

The outputs of filters 204 and 206 are provided to bandpass filter circuits (BPF) 208 and 210, respectively. BPF 208 includes channels 1-9 while BPF 210 includes channels 10-19. Each of channels 1-18 contains a one/third octave filter. Channel 19 contains a full octave filter. The channel filters are implemented in a conventional manner using Reticon Model Numbers R5604 and R56606 switched-capacitor devices. Fig. 3 gives the clock input frequency, center frequency and 3 db bandwidth of the 19 channels of the BPF circuits 208 and 210. The bandpass filter clock frequency inputs required for the BPF circuits 208 and 210 are generated in a conventional manner from a clock generator circuit 212 driven by a 1.632 MHz clock 213.

The outputs of BPF circuits 208 and 210 are rectified, low pass filtered (cutoff frequency = 30 Hz) and sampled simultaneously in 19 sample and hold circuits (National Semiconductor Model No. LF398) in sampling circuitry 214. The 19 channel samples are then multiplexed through multiplexers 216 and 218 (Siliconix Model No. DG506) and converted from analog to digital signals in log A/D converter 220, a Siliconix device, Model No. DF331. The converter 220 has an 8 bit serial output which is converted to a parallel format in serial to parallel register 222 (National Semiconductor Model No. DM86LS62) for input to processor 110 via bus 132.

A 2 MHz clock 224 generates various timing signals for the circuitry 214, multiplexers 216 and 218 and for A/D converter 220. A sample and hold command is sent to circuitry 214 once every 10 milliseconds over lead 215. Then each of the sample and hold circuits is multiplexed sequentially (one every 500 microseconds) in response to a five bit selection signal transmitted via bus 217 to circuits 216 and 218 from timing circuit 226. Four bits are used by each circuit while one bit is used to select which circuit. It therefore takes 10 milliseconds to A/D convert 19 sampled channels plus a ground reference sample. These 20 8-bit digital signals are called a frame of data and they are transmitted over bus 132 at appropriate times to microprocessor 110. Once every frame a status signal is generated from timing generator circuit 226 and provided to processor 110 via lead 228. This signal serves to sync the filter circuit 108 timing to the processor 110 input. Timing generator circuiit 226 further provides a 2 kHz data ready strobe via lead 230 to processor 110. This provides 20 interrupt signals per frame to processor 110.

Referring now to Fig. 4, a block diagram of the automatic speech recognition algorithm 400 of the present invention is presented. It can be divided into four subtasks: bandpass filter data transformation 402; speech activity detection 404; variable frame rate encoding and normalized mel-cepstral transformation 406; and recognition 408. The speech activity detection subtask 404 has been implemented in C language for use on a VAX 11/780 and in assembly language for use on an MC68000. C language is a higher order language commonly used in the technical community and available from Western Electric. The C language version of subtask 404 will be described in more detail in connection with a description of Fig. 7.

As discussed earlier, every 500 microseconds the microprocessor 110 is interrupted by the circuit 108 via lead 230. The software which handles that interrupt is the BPF transformation subtask 402. Usually, the new 8-bit filter value from bus 132 is stored into a buffer, but every 10 millisecond (the 20th interrupt) a new frame signal is sent via lead 228. The BPF transformation subtask 402 takes the 19 8-bit filter values that were buffered, combines the first three values as the first coefficient and the next two values as the second coefficient, and discards the 19th value because it has been found to contain little if any useful information, especially in a noisy environment. The resulting 15 coefficients characterize one 10 ms frame of the input signal

The transformed frame of speech is passed onto buffer 410 and then to the VFRE and mel-cepstral transformation subtask 406 if the speech activity detector substask 404 has indicated that speech is present. The speech activity detector subtask 404 will be explained in more detail later. Assuming for the moment that subtask 404 indicates that speech is present, then in subtask 406, the Euclidean distance between a previously stored frame and the current frame in buffer 410 is determined. If the distance is small (large similarly)

- 9 -

and not more than two frames of data have been skipped, the current frame is passed over, otherwise it is stored for future comparison and passed onto the next step of normalized mel-cepstral transformation. On the average, one-half of the data frames from the circuit 108 are passed on (i.e. 50 frames per second).

To reduce the data to be processed, the 15 filter coefficients are reduced to 5 coefficients by a linear transformation matrix. A commonly used matrix comprises a family of 5 "mel-cosine" vectors that transform the bandpass filter data into an approximation of "mel-cepstral" coefficients. Mel-cosine linear transformations are discussed in (1) Davis, S.B. and Mermelstein, P. "Evaluation of Acoustic Parameters for Monosyllable Word Identification", Journal Acoust. Soc. Am., Vol. 64, Suppl. 1, pp. S180-181, Fall 1978 (Abstract) and (2) S. Davis and P. Mermelstein "Comparison of Parameter Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 357-366, both of which are hereby incorporated by reference as if specifically set forth herein. However, in the preferred embodiment of the present invention, a variation on "mel-cosine" linear transformation is used called normalized mel-cepstral transformation, i.e., the raw BPF data is normalized to zero mean, normalized to zero net slope above 500 H_Z and mel-cosine transformed in one step. The first mel-cepstral coefficient (which is very sensitive to spectral slope) is not used.

Each frame which has undergone mel-cepstral transformation is then compared with each of the templates representing the vocabulary which are now stored in the processor's working memory 126. The comparison is done in accordance with a recognition portion 408 of the algorithm described in the above-mentioned patent application, Serial No. 473,422, filed March 9, 1983 and based on the well-known dynamic programming algorithm (DPA) which is described in an article by F. Itakura entitled "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, pp. 67-72, February 1975. In the above-mentioned patent application, a modified version of the DPA is used, called a windowed DPA with path boundary control. A summary of the DPA is provided in connection with a description of Fig. 5. A template is placed on the y-axis 502 and the input word to be recognized is placed on the x-axis 504 to form a DPA matrix 500. Every cell in the matrix corresponds to a one-to-one mapping of a template frame with a word frame. Any time alignment between the frames of these patterns can be represented by a path through the matrix from the lower-left corner to the upper-right corner. A typical alignment path 506 is shown. The DPA function finds the locally optimal path through the matrix by progressively finding the best path to each cell, D, in the matrix by extending the best path ending in the three adjacent cells labeled by variables, A, B, and C. The path that has the minimum score is selected to be extended to D subject to the local path constraint: every horizontal or vertical step must be followed by a diagonal step. For example, if a vertical step was made into cell C, the path at cell C cannot be chosen as the best path to cell D. The path score at cell D is updated with the previous path score (from A, B, or C) plus the frame-to-frame distance at cell D. This distance is doubled before adding if a diagonal step was chosen to aid in path score normalization. The movement of the DPA function is along the template axis for each utterance frame. The function just described is repeated in the innermost loop of the recognition algorithm by resetting the B variable to cell D's score, the A variable to cell C's score and retrieving from storage a new value for C.

However, before the subtasks 406 and 408 can operate, the beginning and end of speech must be detected. Where speech recognition is taking place in a quiet environment with little or no noise present, endpoint detection based on energy measurement can be used. However, in the environment of tactical fighters, for example, there are present two types of noise which render traditional speech activity detectors useless. Background noise from engines and wind is added to the speech signal and results in the classical detection problem of separating signal and additive noise. See curve 602 in Fig. 6. The use of an oxygen regulator with a mask introduces noise from inhales and exhales which are not concurrent with speech but resemble speech in spectral shape and can cause spurious detection. See Curves 604 and 606, respectively. The amplitudes of the signals associated with these noise events often exceed the speech signal amplitudes in many cockpit conditions.

Referring now to Fig. 7, a more detailed description of the speech activity detection subtask 404 is given. A large number of frames of data from subtask 402 representing both speech and noise event sounds from a variety of speakers and oxygen regulators were studied to determine a fixed transformation which when applied to the frames would provide a good separation between speech and noise events over a range of speakers. It was determined that a single 15 parameter feature vector 702 could be found whose inner product 703 with modified frames 704 derived from the bandpass filter frame 705 would provide a scalar feature 706 giving good separation of speech from noise events. The frames coming from the BPF transformation subtask 402 are logarithmically encoded frames due to the action of the log A/D Converter 220. Better results are achieved, however, if frames proportional to the energy of the noise event signals and speech signals are formed. This is accomplished by modifying the BPF frames from 705 via the operation of squaring the inverse log of the frame components 707. This step enhances speech activity detection by increasing the dynamic range of the features, thus providing greater separation between the peaks of the speech spectra and the relatively broad band noise and non-speech spectra.

To derive a good feature vector F, a collection of frames of BPF data from a plurality of speakers and noise events occurring when speech is not present are collected and modified as described above. The data is divided into sets of speech frames [S] and noise event frames [N]. By inspection, a good intuitive guess at F is made and then in accordance with the equation below, the inner products of F with all of [S] and all of [N] is formed, and the statistical overlap of the resulting two classes of scalar features, [F·S] and [F·N] is measured to form a separation figure of merit. (· represents forming the inner product of the two vectors.)

$$\text{Separation} = \frac{\text{Mean}\ ([F \cdot S]) - \text{Mean}\ ([F \cdot N])}{\text{Std Dev}\ ([F \cdot S]) + \text{Std Dev}\ ([F \cdot N])}$$

Small changes in each of the feature vector components, f_j is made, for example, the first component, f_1, of F is made a little larger and then a little smaller, then the same is done for f_2 and so on. For each small change F.S and F.N is recomputed for all the frames [S] and [N] and the separation remeasured. This identifies the direction to take to change F for better separation. F is changed accordingly, obtaining a new vector for a starting point and then the process is repeated. This approach is known as a gradient search.

When a feature vector F is formed which appears to be a significant improvement, it is tried in the recognizer algorithm to see how it works. If certain types of noise events are found to still trigger the detection, or if certain speech sounds are consistently missed, samples of them are taken and added to the data base [S] and [N]. Then a new feature vector is searched for that handles the new data as well as the old.

To assist in carrying out all the inner product and separation computations required during the gradient search, a program was created in C language for a VAX computer.

The preferred embodiment, 15 parameter feature vector found by the gradient search as substantially described above is,

1	0.0
2	13.9
3	5.9
4	1.2
5	1.4
6	1.4
7	1.5
8	1.6
9	2.4
10	1.3
11	2.0
12	1.2
13	4.8
14	-13.6
15	0.0

Once the optimum feature vector is determined, the resultant scalar features formed by the inner product operation with the modified frames are collected and formed into a histogram designated generally 710 in Fig. 7. The x-axis 712 is the magnitude of the scalar feature while the y-axis 714 is the number of times a particular magnitude occurs. Jet noise 716 and regulator sounds 718 occur below a threshold 720 while voice 722 occurs above the threshold 720.

When the speech recognizer is being used, e.g., in flight in an aircraft cockpit, the speech activity detection subtask 404 initially selects a detection threshold but thereafter continually gathers statistics and updates the histogram on the feature 726. Every 1000 frames, the detection threshold is adjusted based on the statistics in the histogram. For example, the peak 750 is located in the histogram 710, and a search is conducted forward from the peak 750 to locate the low point 720. The threshold is set to the low point value plus some bias such as one or two. Finally, each histogram entry is divided by two to keep the histogram values from growing too large.

The magnitude of the detection threshold 708 is subtracted from the magnitude of the scalar feature 706 at block 730 for each frame. A weighting function 732 is applied to the output value of block 730 to smooth out the values before they are filtered and clamped at 734. The weighting function reduces large negative values from block 730 and reduces small positive values. Large positive values are left substantially unaffected. The weighting function cooperates with the integration process performed by the filter and clamp function 734 to provide sharp cutoff points between the beginning and end of speech detection. Large negative values provide no better indication of non-speech than smaller values, but will distort and delay the integration process from indicating when speech is present. Small positive values create uncertainty as to whether speech is present and are better left undetected. An example of the preferred embodiment weighting function and filter and clamping functions are provided in C language on page 19 of the specification.

Four values from filter and clamp 734 corresponding to four successive frames from subtask 402 are stored in buffers 736. Then multi-frame decision logic 738 is employed to make a decision whether speech is present. For example, if no speech were present and if all four buffers provide a positive indication, then a decision is made that speech is present, and this is passed on to block 410 in Fig. 4, otherwise a decision is made that speech still is not present. On the other hand, if speech is currently present, a decision is made that speech is still present if any one of the buffers indicates that a speech signal is present. Only if all four buffers indicate no speech signals present will a decision be made that speech is now over. The above-described decoding is provided in C language at pages 19 and 20 of the specification

It should be noted that in the preferred embodiment, subtasks 402, 404 and 406 are performed in processor 110 while subtask 408 is performed in processor 112. However, there is no reason why the two processors could not be combined as one. Although the present invention relates to a 36 word vocabulary with isolated word recognition, there is no reason why the speech activity detector could not be used with larger vocabulary continuous speech recognition machines. Also, speech activity detection through the use of the inner product between a predefined feature vector and frames of speech can be performed on frames of speech provided directly from the bandpass filter transformation subtask 402 even though this frame is proportional to the log of the value of the digital signals. Similarly, the inner product could be performed using frames whose digital signals are proportional to the magnitude of the digital signals and not the magnitude squared.

Results to date on the performance of the recognizer indicate recognition accuracy of 85 to 95% for worst cases of cockpit sound pressure level of 115 dB and acceleration forces of 5G. In fact, the system shows no degradation from low level ambient noise performance (95+% accuracy) to noise levels of approximately 106 dB. It should be pointed out, however, that the 115 dB sound levels at 5G acceleration forces are often simulated. The pilot is speaking into an oxygen regulator which partially seals off the ambient cockpit noise. However, the stress of the noise and acceleration forces causes the pilot to speak in a less than normal speaking manner. Also, the noise events caused by the stressed breathing of the pilot into the oxygen regulator are also present.

S.E. Hutchins et al 1-1-4-3-3

Claims:

1. An apparatus for speech activity detection of speech in the presence of noise including noise events occuring when speech is not present characterized in that it comprises:

means (108) for digitizing signals associated with said speech signals and signals associated with said noise events and for forming frames of digital signal values associated with said speech and noise event signals; and

separation means (110, 112) coupled to said digitizing means for automatically separating said speech signals from said noise event signals.

2. The apparatus as claimed in claim 1, characterized in that said separation means further comprises means for applying a speaker independent, predetermined, fixed transformation to said digital signal values of said frames whereby frames associated with said speech signals are separated from frames associated with said noise event signals.

ZT/P1-Kg/B
July 3, 1984

S.E. Hutchins et al 1-1-4-3-3

3. The apparatus as claimed in claim 2, characterized in that said means for applying said speaker independent, predetermined, fixed transformation comprises:

means for creating scalar features from said frames; and

that said separation means further comprises:

means for extablishing and updating a detection threshold value wherein frames associated with scalar features having a magnitude less than said detection threshold value are considered as associated with noise event signals while frames associated with scalar features having magnitudes greater than said detection threshold values are considered as associated with speech signals.

4. The apparatus as claimed in claim 3, characterized in that said apparatus further comprises:

means for comparing said scalar features with said detection threshold value;

means for storing the results of a plurality of said comparisions for a plurality of successive frames; and

means for combining said stored results to obtain an indication of when speech signals are present.

5. The apparatus as claimed in any one of the preceding claims, characterized in that said separation means (110,112) comprises:

speech activity means (404) coupled to said digitizing means (108, 408) for automatically separating said speech signals from said noise event signals to determine when said speech signals are present;

S.E. Hutchins et al 1-1-4-3-3

speech recongnition means (406, 408) coupled to said digitizing means (402) and said speech activity means (404) for converting said frames into frames of parametric data more suitable for further recognition processing when said speech activity means determines that speech signals are present; and

means coupled to said recognition means for comparing selected ones of said frames of parametric data with a plurality of templates which are representative of said speech to be recognized whereby said speech signals are recognized.

6. The apparatus as claimed in claim 5, characterized in that said speech activity means (404) further comprises:

means (706) for creating scalar features from said frames;

means (708, 728) for extablishing and updating a detection threshold value wherein frames associated with scalar features having a magnitude less than said detection threshold value are considered as associated with noise event signals while frames associated with scalar features having magnitudes greater than said detection threshold value are considered as associated with speech signals;

means (730) for comparing said scalar features with said detection threshold values;

means (732, 736) for storing the results of a plurality of said comparisions for a plurality of successive frames; and

means (738) for combining said stored results to obtain an indication of when speech signals are present.

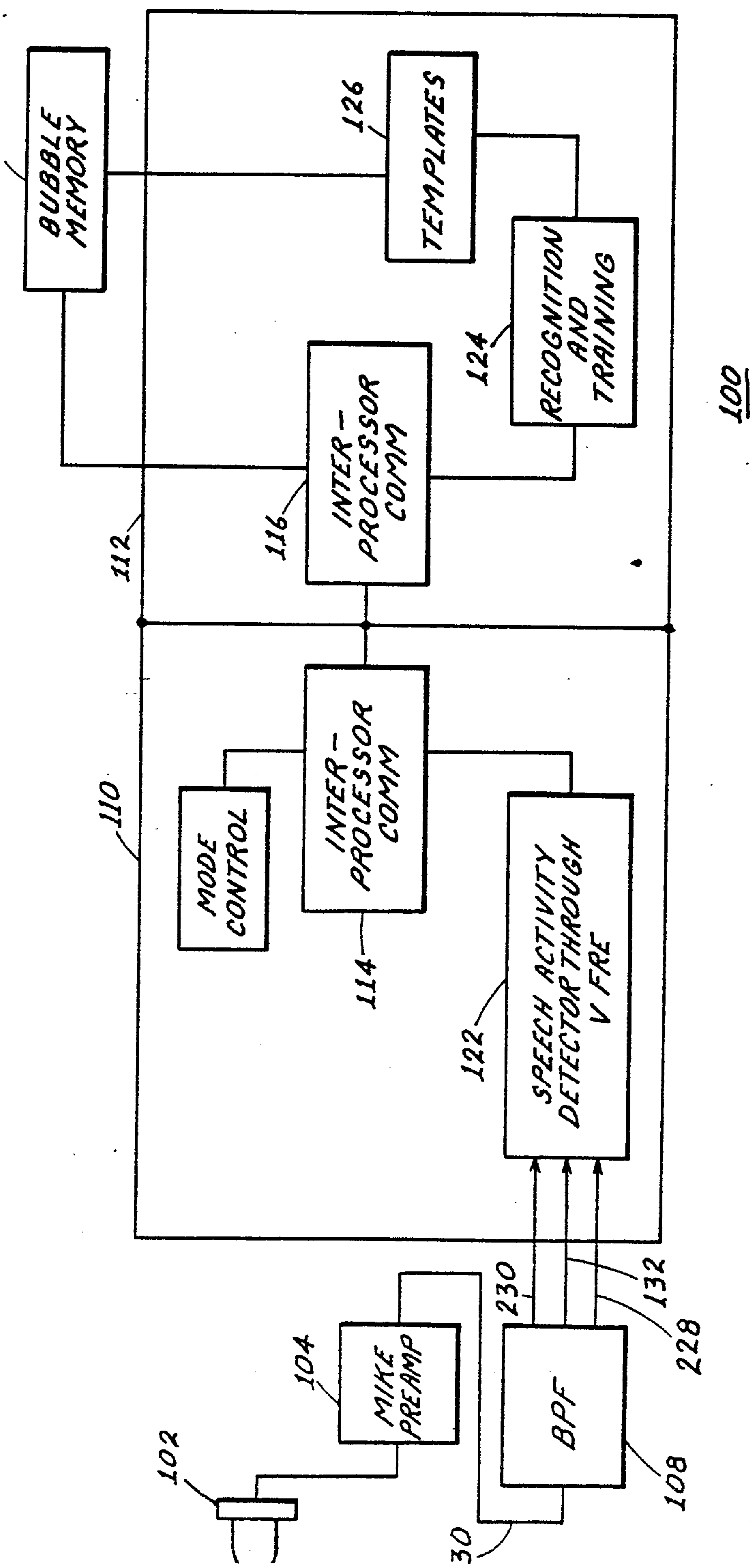
7. The apparatus as claimed in claim 6, characterized in that the magnitude of said noise event signals is equal to or greater than the magnitude of said speech event signals.

8. The apparatus as claimed in claim 6, characterized in that said apparatus further comprises means (707,)o4) for modifiying said frames of digital signals coupled to said speech activity means to form modified frames of digital signals wherein said digital signal values are related to the square of the magnitude of said speech and noise event signals.

0143161

S.E. Hutchins 1-1-4-3-3

Fig. 1



0143161

S.E. Hutchins 1-1-4-3-3

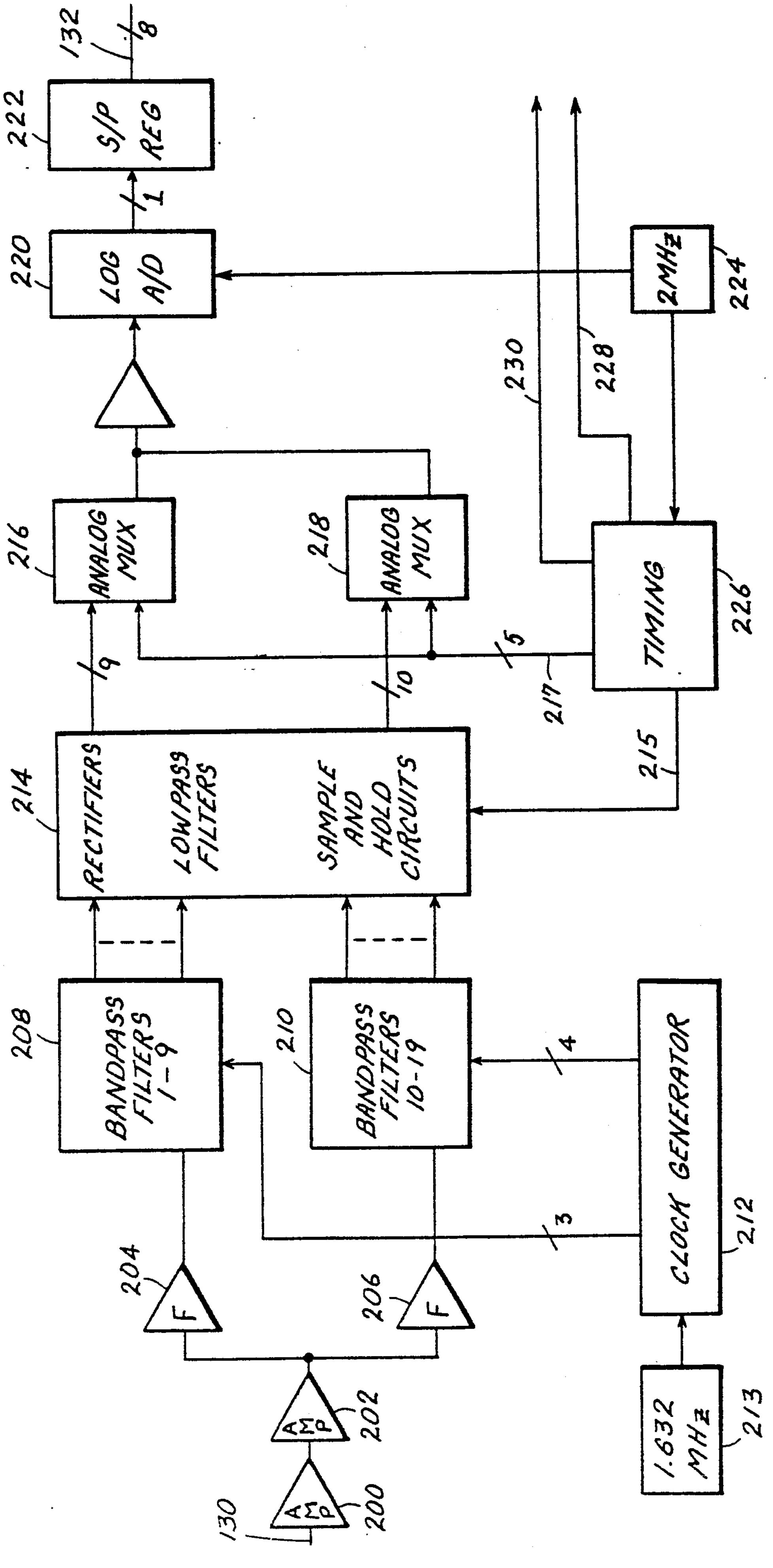


Fig. 2

0143161

FILTER CHARACTERISTICS

CHANNEL	CLOCK INPUT KHz	CENTER FREQUENCY	3db BANDWIDTH
1.		125	25
2.	17	157	33
3.		196	41
4.		250	53
5.	34	313	66
6.		391	82
7.		501	105
8.	68.2	626	131
9.		783	164
10.		1002	210
11.	136.5	1253	263
12.		1566	329
13.		2003	421
14.	273	2505	526
15.		3131	656
16.		4006	841
17.	545.1	5010	1052
18.		6263	1315
19.	819.2	7515	4359

NOTE: THESE ARE NOMINAL FREQUENCIES WITH ABOUT ± 5% TOLERANCE

Fig. 3

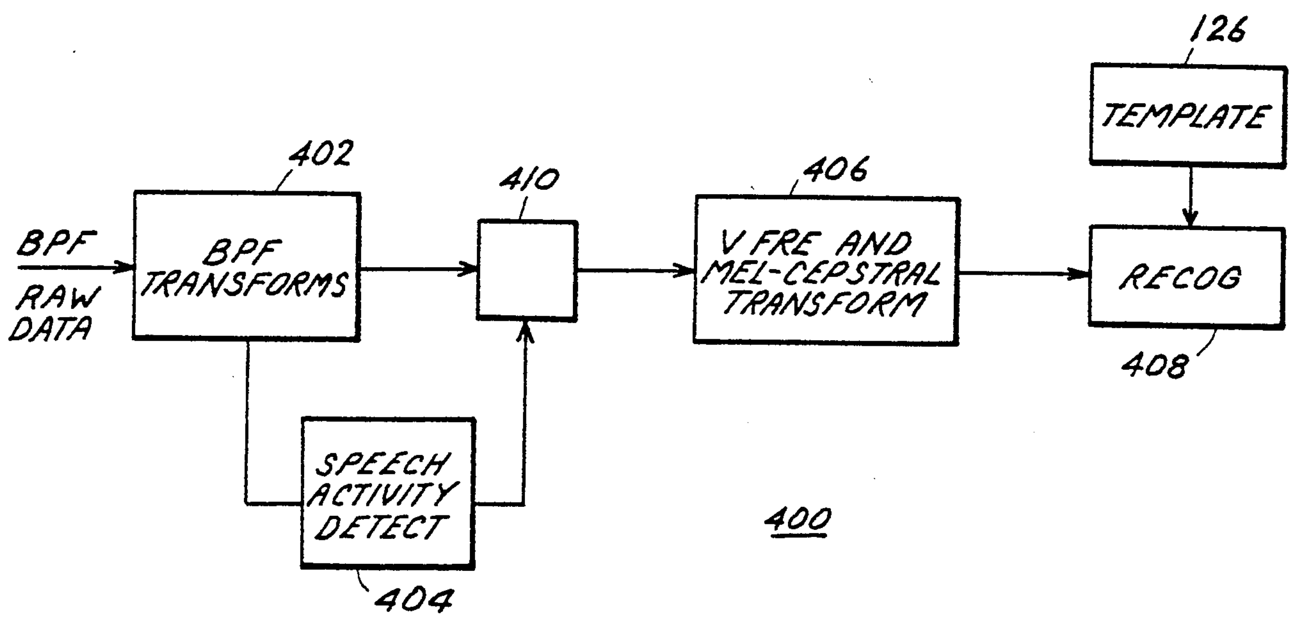


Fig. 4

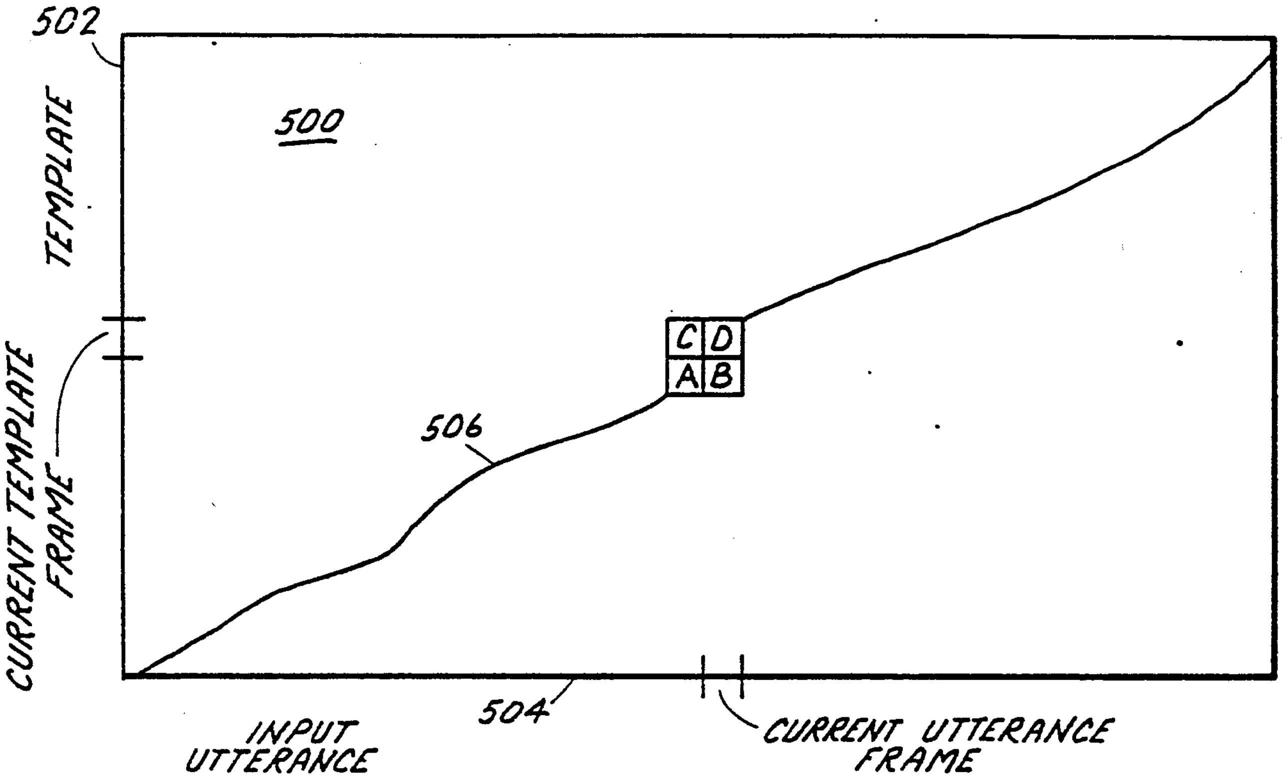

502
500
TEMPLATE
CURRENT TEMPLATE FRAME
C D
A B
506
504
INPUT UTTERANCE
CURRENT UTTERANCE FRAME

Fig. 5

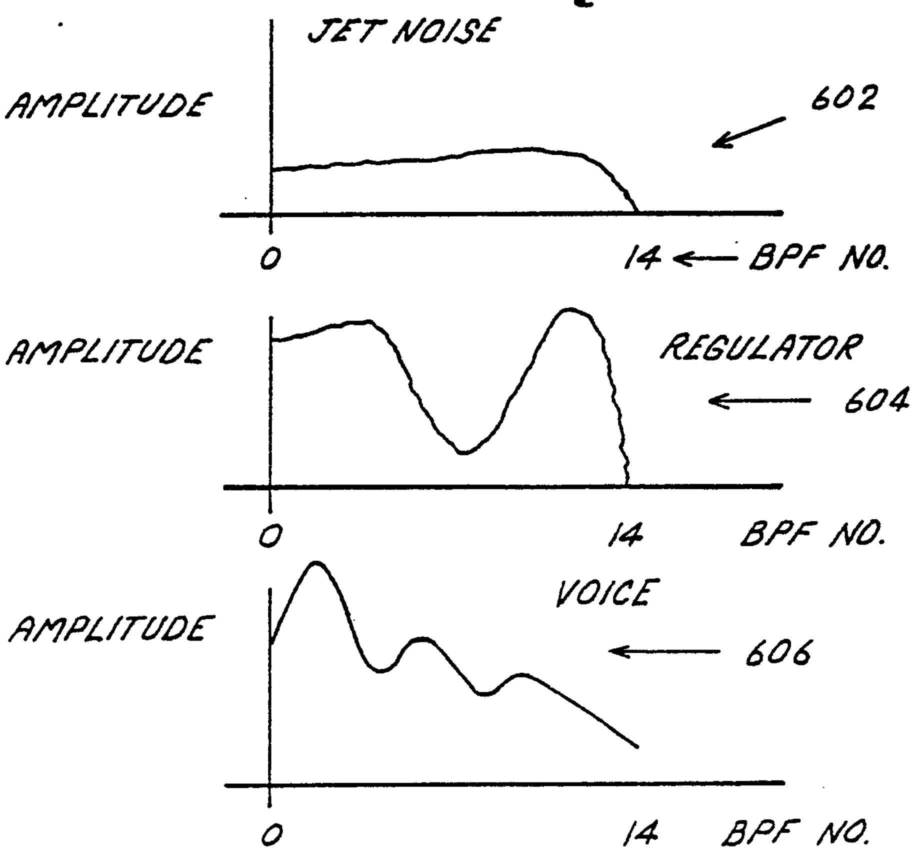

JET NOISE
AMPLITUDE
602
0
14
BPF NO.
AMPLITUDE
REGULATOR
604
0
14
BPF NO.
VOICE
AMPLITUDE
606
0
14
BPF NO.

Fig. 6

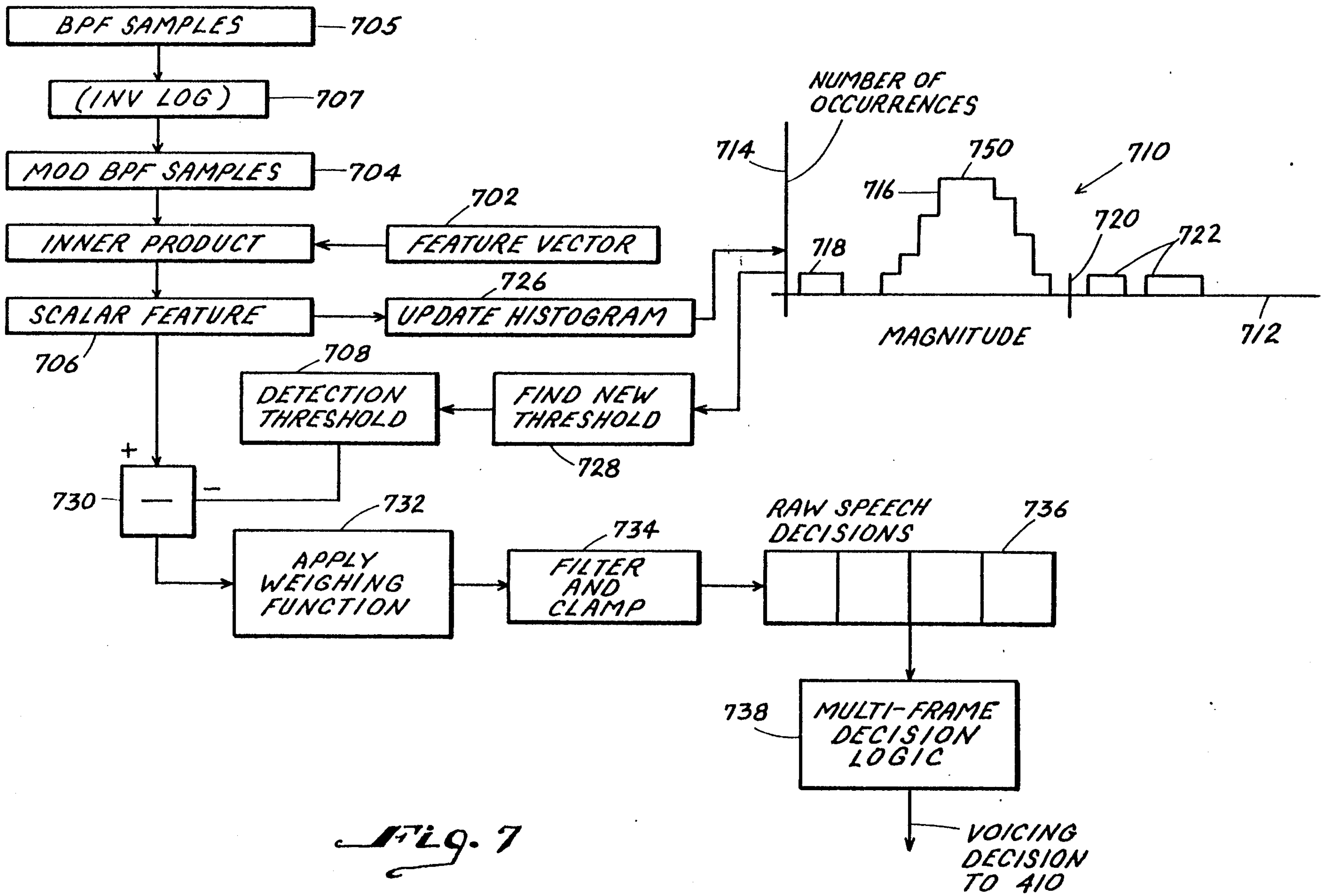


Fig. 7

S.E. Hutchins 1-1-4-3-3

European Patent Office

EUROPEAN SEARCH REPORT

0143161 Application number

EP 84 10 7846

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.4)
X	EP-A-0 008 551 (THOMSON-CSF) * Claim 1 * ---	1,2	G 10 K 15/04
X	GB-A-2 109 205 (TOKYO SHIBAURA DENKI K.K.) * Abstract * ---	1-4,6	
A	IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, vol. IE-30, no. 2, May 1983, pages 150-155, IEEE, New York, USA; N. KISHI et al.: "A voice input system for automobiles using a microprocessor" * Paragraph III.B. "Recognition method" * -----	5	

TECHNICAL FIELDS SEARCHED (Int. Cl.4)
G 10 K 15/04

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	22-10-1984	ARMSPACH J.F.A.M.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document