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

(11) Publication number: **0 179 280**
A2

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **85111905.7**

(51) Int. Cl.[4]: **G10L 9/10**

(22) Date of filing: **20.09.85**

(30) Priority: **26.10.84 US 665401**

(43) Date of publication of application:
30.04.86 Bulletin 86/18

(84) Designated Contracting States:
CH DE FR GB IT LI NL

(71) Applicant: **International Business Machines Corporation**
Old Orchard Road
Armonk, N.Y. 10504(US)

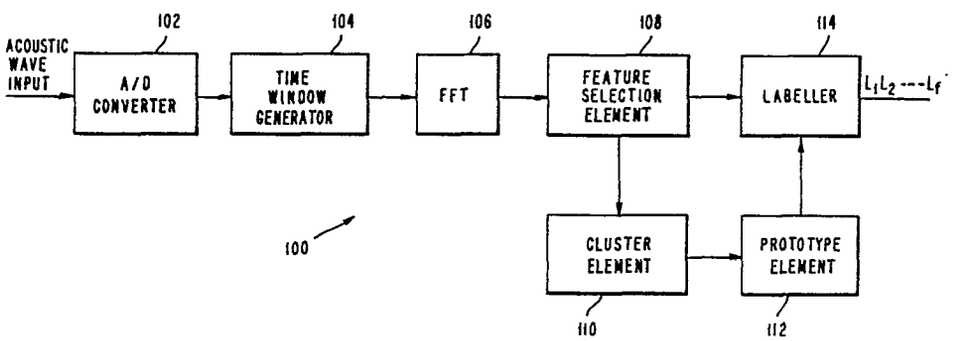
(72) Inventor: **Bakis, Raimo**
2274 Hunterbrook Road
Yorktown Heights N. Y. 10598(US)
Inventor: **Cohen, Jordan Rian**
1863 Hanover Street
Yorktown Heights N. Y. 10598(US)

(74) Representative: **Gasslander, Sten et al**
IBM Svenska AB Box 962
S-181 09 Lidingö(SE)

(54) Nonlinear signal processing in a speech recognition system.

(57) An acoustic processor and a method of processing an acoustic wave input which includes a non-linear auditory model of the neural firing rate in the ear. The firing rate is determined by the replenishment of neurotransmitter and loss of neurotransmitter due to spontaneous decay, spontaneous firing, and acoustic wave inputs defined, preferably, in sones. The number of free parameters in the acoustic processor is reducible to one, namely the ratio R of two steady-state firing rates each resulting from a different loudness. Preferably, the free parameter is adjusted to minimize steady-state effects which are adversely impacted by speaker differences, background noise, distortion, and the like. The present invention addresses the general problem of processing an acoustic wave input in a speech recognition system and addresses the specific problem of adjusting acoustic processor performance to reduce adverse effects.

FIG. 1



EP 0 179 280 A2

NONLINEAR SIGNAL PROCESSING IN A SPEECH RECOGNITION SYSTEM

The present invention relates primarily to speech recognition, and specifically to of selecting features, and parameters which affect values of features in the front end of a speech recognition system.

Speech recognition systems or machines are generally aimed at automatically transforming natural speech into some other form, for example, written form. In achieving this aim, Bahl et al, in "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume PAMI-5, No. 2, pp. 179-190 (1983), discuss several approaches to speech recognition. In each approach, one can hypothesize a text generator which determines what it is to be said. The text generator is followed by a speaker, or talker, which produces a natural speech waveform that provides input to an acoustic processor. The acoustic processor output enters a linguistic decoder.

According to the Bahl et al article, the elements in the system may be associated in various ways. For example, the speaker and acoustic processor may be combined to form an acoustic channel wherein the speaker transforms the text into a speech waveform and wherein the acoustic processor acts as a data transducer and compressor which provides a string of labels to the linguistic decoder. The linguistic decoder recovers the original text from the string of labels.

More specifically, an acoustic wave input enters an analog-to-digital converter, which samples at a prescribed rate. The digital signals are then transformed to frequency spectrum outputs to be processed to produce characteristic labels representing the speech wave input. The selection of appropriate features is a key factor in deriving these labels and the present invention relates to improved feature selection means as well as the front end processor and the speech-recognition system in which such feature selection means is included.

The feature selection element of the present invention responds so as to model the peripheral auditory system, that is, considers the auditory nerve firing rates at selected frequencies as the features which define the acoustic input. While the ear has, in the past, been modelled by others (see "Model for Mechanical to Neural Transduction In the Auditory Receptor" by Hall and Schroeder, Journal of the Acoustics Society of America, Volume 55, No. 5, May 1974), the present invention incorporates a model based on neural firings in the ear for use as a feature selection element, which results in notably enhanced speech recognition relative to prior art systems.

Schroeder and Hall, in the above-noted article, suggest a model for the ear which relates to the transduction of mechanical motion or vibration of the basilar membrane into action potentials or "spikes" in the auditory nerve. The Schroeder and Hall model is based on the generation and depletion of electrochemical quanta in a hypothetical hair cell. The Schroeder and Hall model involves three features: the fixed rate of generation of quanta of an electrochemical agent, the rate of disappearance of quanta without any neural firing, and the firing probability with no signal.

The auditory model of the present invention, as in the Schroeder and Hall model, seeks closer conformance with neurophysiological data than conventional threshold models of the ear. However, unlike Schroeder and Hall, the present model as implemented employs both a different time scale and a different compressive non-linearity preceding the firing rate computation. This new formulation allows macroscopic neural data to be used in setting parameter values, and its output is appropriate for use directly in the front end of a speech recognition system. Also, the present model unlike Schroeder and Hall is used in a speech recognition system. Moreover, the implemented model accounts for factors that are not addressed, or are resolved differently, by Schroeder and Hall. For example, with a large dynamic range of speech amplitude inputs, the Schroeder and Hall model provides firing rates that do not appear accurate. The model implemented in the present invention overcomes this problem.

In tests involving a number of subjects, the word error rate in each instance improved when the present parameter selection element replaced an existing element. Accordingly, the present invention has as an object improvement in performance of a speech recognition system, by employing in the parameter selection element, an auditory model based on neural firings.

To further conform to the ear, the present invention employs critical band filtering to reflect the action of the basilar membrane of the ear as a frequency analyzer. That is, like the basilar membrane which experiences increased loudness as two components of an audio input spread to reside in different critical bands, the present invention also preferably provides a response that filters the acoustic wave input according to similar bands.

It is yet another object of the invention to define the features of the feature selection element as a function of loudness, preferably in compressed amplitude form, such as sones. Moreover, to account for inequalities in loudness (in sones) at different frequencies and to account for inequalities in loudness (in sones) relative to variations in loudness level in phons, the present invention includes an equal loudness adjustment element and a loudness scaling element to achieve normalization.

The feature selection element achieves the above objects in a speech recognition system and contributes to the realization of large vocabulary recognition in a real-time system that is preferably speaker-trained and preferably of the isolated word variety.

The present model which achieves the above-noted objects processes the acoustic (speech) wave input by initially digitizing the waveform and then determining waveform magnitude as a function of frequency for successive discrete periods of time. The magnitudes are preferably grouped according to critical bands (as with the basilar membrane). In accordance with the model, it is presumed that there are (modelled) neural firings at a rate, f, in the ear (for each critical frequency) and that the neural firings depend on the amount n of a modelled neurotransmitter in the ear, among other factors. The rate of change of neurotransmitter for each critical band is viewed as a function of neurotransmitter replenishment -- which is considered to be at a rate Ao -- and neurotransmitter loss.

The loss in neurotransmitter over time is viewed as having several components: (1) (Sh x n), Sh corresponding to the natural decay or disappearance of neurotransmitter over time independent of acoustic wave input; (2) (So x n) So corresponding to the rate of spontaneous neural firings which occur regardless of acoustic wave input, and (3) DLn corresponding to neural firings as a function of loudness L

scaled by a factor D. The model is represented by the equations:

$$dn/dt = Ao - (So + Sh + DL) \quad (1)$$

$$f = (So + DL)\, n \quad (2)$$

Equations (1) and (2) are defined for each critical frequency band, where t is time.

The present invention is also concerned with determining the "next state" of the neurotransmitter amount that is to be used in the next determination of firing rate f. In a general sense, the next state may be defined by the following equation:

$$n(t+\Delta t) = n(t) + (dn/dt)\,\Delta t \quad (3)$$

The next state equation (3) and neurotransmitter change equation (1) help define the next value of the firing rate f. In this regard, the firing rate f (for each frequency band) is nonlinear in that it depends multiplicatively on the previous state. This -- as noted above -- closely tracks the time adaptive nature of the auditory system.

The firing rates for the various respective frequency bands together provide the features for speech recognition labelling. For twenty bands, for example, twenty firing rates -- one for each band -- together provide a vector in 20-dimension space that can be entered into the labeller 114 so that vectors corresponding to the acoustic wave input can be matched against stored data and labels generated.

It is noted that both f and n in equations (1) and (2) tend to have large DC pedestals. Where the dynamic range of the terms in the equations is to be broad, a series of equations are provided to decrease pedestal height. In this regard, the invention separates n into a steady state component $\bar{n}$ and a varying component $\tilde{n}(t)$ so that equation (2) becomes:

$$\tilde{f}(t) = (So + D \times L)\,\frac{(\tilde{n}(t) + D\ L\ Ao)}{(So + Sh)} \quad (4)$$

Similarly, by defining n as $\bar{n} + \tilde{n}(t)$ and ignoring constant terms, equation (3) becomes:

$$\tilde{n}(t + \Delta t) = \tilde{n}(1 - So\Delta t) - \tilde{f}(t) \quad (5)$$

Equations (4) and (5) constitute a special case output equation and state update equation, respectively, applied to the signal of each critical frequency band during successive frames in time. Equation (4) for each frequency band, defines a vector dimension for each time frame that is improved over the basic output from equations (1) through (3).

The performance of a speech recognition system can be improved by adjusting or modifying the values of parameters which affect the feature values. However, testing the system for improvement after each adjustment or modification is a time-consuming process, especially where there are a number of parameters which can be adjusted or modified. It is thus another object of the invention to provide a functional auditory model for use as a feature selection element with as few free parameters as possible. By use of empir ical data to specify certain of the terms in the above equations, the number of free parameters is reduced to as few as one.

The invention thereby permits the model to be adjusted by altering a single parameter to determine how system performance may be changed or improved. In particular, the single parameter is a ratio defined as:

$$R = \frac{f_{\text{steady state}} \,\big|\, L = L_{max}}{f_{\text{steady state}} \,\big|\, L = \phi} \quad (6)$$

R represents the ratio of (a) the steady firing rate when the loudness is at a maximum (e.g. the threshold of feeling) to (b) the steady firing rate when the loudness is at the minimum (e.g. zero). According to the invention, R is preferably the only variable of the system which is varied to adjust or modify the parameter.

By providing for equal loudness relative to frequency and for loudness scaling in the loudness included in the above-discussed model, the present invention is able to reduce the production of inconsistent output patterns for similar acoustic wave inputs. This is achieved by emphasizing transient portions of the acoustic (speech) input which are not affected by factors such as differences in frequency response of the acoustic channel, speaker differences, background noise, and distortion.

Finally, with respect to defining equal loudness, a further improvement is proposed wherein the relationship between loudness and intensity is derived from the acoustic input. Specifically, histograms are maintained at each critical frequency band. When a predefined number of filters (at critical frequency bands) have outputs which exceed a given value for a prescribed time, speech is presumed. A threshold-of-feeling and a threshold-of-hearing are then determined for use in loudness normalization based on the histograms during the prescribed time of presumed speech.

The present invention thus provides an enhanced auditory model and employs it in a speech recognition system. In a specific embodiment, the invention relates to a method of processing acoustic wave input in a speech recognition system, the method comprising the steps of: measuring the sound of the acoustic wave input in each of at least one

frequency band; determining, in an auditory model, a neural firing rate for and as a function of the measured sound level at each frequency band; representing the acoustic wave input as the neural firing rates determined for the respective frequency bands; determining, for each frequency band, the current amount of neurotransmitter available for neural firing; and determining, for each frequency band, a rate of change of neurotransmitter based on (a) a replenishment constant that represents the rate at which neurotransmitter is produced and (b) the determined neural firing rate for the respective frequency band; the neural firing rate being dependent on the amount of neurotransmitter available for neural firing, the amount of neurotransmitter available for neural firing in the next state being based on the amount of neurotransmitter available in the current state and the rate of change of neurotransmitter.

Preferably, the sound measuring step includes measuring the loudness of the acoustic wave input at each of a plurality of frequency bands, each frequency band corresponding to a critical frequency band associated with the human ear and includes defining loudness in a compressed amplitude form.

The present invention will now be more closely explained with reference to the accompanying drawings, where

Fig. 1 illustrates a specific embodiment of an acoustic processor.

Fig. 2 shows part of the inner human ear.

Fig. 3 shows the filtering means for filtering the outputs from fourier transform element 106 in fig. 1.

Fig. 4 shows relationship intensity level/frequency.

Fig. 5 shows relationship between sones and plans.

Fig. 6 is a flowchart of the present acoustic processor.

Fig. 7 is a flowchart of how the power density is transformed from log magnitude to loudness level.

In FIG. 1 a specific embodiment of an acoustic processor 100 is illustrated. An acoustic wave input (e.g., natural speech) enters an analog-to-digital converter 102 which samples at a prescribed rate. A typical sampling rate is one sample every 50 microseconds. To shape the edges of the digital signal, a time window generator 104 is provided. The output of the window 104 enters a fast fourier transform (FFT) element 106 which provides a frequency spectrum output for each time window.

The output of the FFT element 106 is then processed to produce labels L_1L_2---L_f . Four elements -- a feature selection element 108, a cluster element 110, a prototype element 112, and a labeller 114 -- coact to generate the labels. In generating the labels, prototypes are defined as points (or vectors) in the space based on selected features and acoustic inputs and are then characterized by the same selected features to provide corresponding points (or vectors), in space that can be compared to the prototypes.

Specifically, in defining the prototypes, sets of points are grouped together as respective cluster by cluster element 110. A prototype of each cluster -- relating to the centroid or other characteristic of the cluster -- is generated by the prototype element 112. The generated prototypes and acoustic input -- both characterized by the same selected features -- enter the labeller 114. The labeller 114 performs a matching procedure.

It is noted that the conventional audio channel typically provides a plurality of parameters which may be adjusted in value to alter performance. To examine changes in performance in response to parameter variations requires that the entire acoustic processor 100 be run which typically takes a day. Hence, the more parameters there are to vary, the more difficult and time-consuming is the task of examining performance changes.

The design philosophy of the present invention is to provide an acoustic processor 100 that has a minimal number of adjustable parameters to facilitate performance improvement.

In accordance with the invention, an auditory model is derived and applied in an acoustic processor of a speech recognition system. In explaining the auditory model, reference is made to FIG. 2, which shows part of the inner human ear. Specifically, an inner hair cell 200 is shown with end portions 202 extending therefrom into a fluidcontaining channel 204. Upstream from inner hair cells are outer hair cells 206 also shown with end portions extending into the channel 204. Associated with the inner hair cell 200 and outer hair cells 206 are nerves which convey information to the brain. Specifically, nerve neurons undergo electrochemical changes which result in electrical impulses being conveyed along a nerve to the brain for processing. Effectuation of the electrochemical changes, is stimulated by the mechanical motion of the basilar membrane 210.

It has been recognized in prior teachings, that the basilar membrane 210 serves as a frequency analyzer for acoustic waveform inputs and that portions along the basilar membrane 210 respond to respective critical frequency bands. That different portions of the basilar membrane 210 respond to corresponding frequency bands has an impact on the loudness perceived for an acoustic waveform input. That is, the loudness of tones is perceived to be greater when two tones are in different critical frequency bands than when two tones of similar power intensity occupy the same frequency band. It has been found that there are on the order of twenty-two critical frequency bands defined by the basilar membrane 210.

Conforming to the frequency-response of the basilar membrane 210, the present invention in its preferred form physically defines the acoustic waveform input into some or all of the critical frequency bands and then examines the signal component for each defined critical frequency band separately. This function is achieved by appropriately filtering the signal from the FFT element 106 (see FIG. 1) to provide a separate signal in the feature selection element 108 for each examined critical frequency band.

The separate inputs, it is noted, have also been blocked into time frames (of preferably 25.6 msec) by the time window generator 104. Hence, the feature selection element 108 preferably includes twenty-two signals -- each of which represents sound intensity in a given frequency band for one frame in time after another.

The filtering is preferably performed by a conventional critical band filter 300 of FIG. 3. The separate signals are then processed by an equal loudness converter 302 which accounts for perceived loudness variations as a function of frequency. In this regard, it is noted that a first tone at a given dB level at one frequency may differ in perceived loudness from a second tone at the same given dB level at a second frequency. The converter 302 can be based on empirical data, converting the signals in the various fre-

quency bands so that each is measured by a similar loudness scale. For example, the converter 302 can map from acoustic power to equal loudness based on studies of Fletcher and Munson in 1933, subject to certain modifications. The modified results of these studies are depicted in FIG. 4. In accordance with FIG. 4, a 1KHz tone at 40dB is comparable in loudness level to a 100Hz tone at 60dB as shown by the X in the figure.

The converter 302 adjusts loudness preferably in accordance with the contours of FIG. 4 to effect equal loudness regardless of frequency.

In addition to dependence on frequency, power changes and loudness changes do not correspond as one looks at a single frequency in FIG. 4. That is, variations in the sound intensity, or amplitude, are not at all points reflected by similar changes in perceived loudness. For example, at 100 Hz, the perceived change in loudness of a 10dB change at about 110dB is much larger than the perceived change in loudness of a 10dB change at 20dB. This difference is addressed by a loudness scaling element 304 which compresses loudness in a predefined fashion. Preferably, the loudness scaling element compresses power P by a cube-root factor to $p^{1/3}$ by replacing loudness amplitude measure in phons by sones.

FIG. 5 illustrates a known representation of phons versus sones determined empirically. By employing sones, the present model remains substantially accurate at large speech signal amplitudes. One sone, it should be recognized, has been defined as the loudness of a 1KHz tone at 40dB.

Referring again to FIG. 3, a novel time varying response element 306 is shown which acts on the equal loudness, loudness scaled signals associated with each critical frequency band. Specifically, for each frequency band examined, a neural firing rate f is determined at each time frame. The firing rate f is defined in accordance with the invention as:

$$f = (So + DL)n \quad (7)$$

where n is an amount of neurotransmitter; So is a spontaneous firing constant which relates to neural firings independent of acoustic waveform input; L is a measurement of loudness; and D is a displacement constant. So x n corresponds to the spontaneous neural firing rate which occurs whether or not there is an acoustic wave input and DLn corresponds to the firing rate due to the acoustic wave input.

Significantly, the value of n is characterized by the present invention as changing over time according to the relationship:

$$dn/dt = Ao - (So + Sh + DL)n \quad (8)$$

where Ao is a replenishment constant and Sh is a spontaneous neurotransmitter decay constant. The novel relationship set forth in equation (8) takes into account that neurotransmitter is being produced at a certain rate (Ao) and is lost (a) through decay (Sh x n), (b) through spontaneous firing (So x n), and (c) through neural firing due to acoustic wave input (DL x n) The presumed locations of these modelled phenomena are illustrated in FIG. 2.

Equation (8) also reflects the fact that the present invention is non-linear in that the next amount of neurotransmitter and the next firing rate are dependent multiplicatively on the current conditions of at least the neurotransmitter amount. That is, the amount of neurotransmitter at a state $(t+\Delta t)$ is equal to the amount of neurotransmitter at a state t plus dn/dt, or:

$$n(t + \Delta t) = n(t) + dn/dt\ \Delta t \quad (9)$$

Equations (7), (8), and (9) describe a time varying signal analyzer which, it is suggested, addresses the fact that the auditory system appears to be adaptive over time, causing signals on the auditory nerve to be non-linearly related to acoustic wave input. In this regard, the present invention provides the first model which embodies non-linear signal processing in a speech recognition system, so as to better conform to apparent time variations in the nervous system.

In order to reduce the number of unknowns in equations (7) and (8), the present invention uses the following equation (10) which applies to fixed loudness L:

$$So + Sh + DL = 1/\tau \quad (10)$$

τ is a measure of the time it takes for an auditory response to drop to 37% of its maximum after an audio wave input is generated. τ, it is noted, is a function of loudness and is, according to the invention, derived from existing graphs which display the decay of the response for various loudness levels. That is, when a tone of fixed loudness is generated, it generates a response at a first high level after which the response decays toward a steady condition level with a time constant τ. With no acoustic wave input, $\tau = \tau_o$ which is on the order of 50 msec. For a loudness of L_{max}, $\tau = \tau_{max}$ which is on the order of 30 msec. By setting Ao = 1, 1/So + Sh is determined to be 5 csec, when L = 0. When L is L_{max} and L_{max} = 20 sones, equation (11) results:

$$So + Sn + D(20) = 1/30 \quad (11)$$

With the above data and equations, So and Sh are defined by equations (12) and (13) as:

$$So = DL_{max}/[R + (DL_{max}\ \tau_o\ R) - 1] \quad (12)$$

$$Sh = 1/\tau_o - So \quad (13)$$

where

$$R = \frac{f_{steady\ state}\ |\ L_{max}}{f_{steady\ state}\ |\ L = 0} \quad (14)$$

$f_{steady\ state}|$ represents the firing rate at a given loudness when dn/dt is zero.

R, it is noted, is the only variable left in the acoustic processor. Hence, to alter the performance of the processor, only R is changed. R, that is, is a single parameter which may be adjusted to alter performance which, normally, means minimizing steady state effects relative to

transient effects. It is desired to minimize steady state effects because inconsistent output patterns for similar speech inputs generally result from differences in frequency response, speaker differences, background noise, and distortion which affect the steady state portions of the speech signal but not the transient portions. The value of R is preferably set by optimizing the error rate of the complete speech recognition system. A suitable value found in this way is R = 1.5. Values of So and Sh are then 0.0888 and 0.11111 respectively, with D being derived as 0.00666.

Referring to FIG. 6, a flowchart of the present acoustic processor is depicted. Digitized speech in a 25.6 msec time frame, sampled at preferably 20KHz passes through a Hanning Window the output from which is subject to a Fourier Transform, taken at preferably 10 msec intervals. The transform output is filtered to provide a power density output for each of at least one frequency band -- preferably all the critical frequency bands or at least twenty thereof. The power density is then transformed from log magnitude to loudness level. This is performed either by the modified graph of FIG. 4 or by the process outlined hereafter and depicted in FIG. 7.

In FIG. 7, a threshold-of-feeling T_f and a threshold-of-hearing T_h are initially defined for each filtered frequency band m to be 120dB and 0dB respectively. Thereafter, a speech counter, total frames register, and a histogram register are reset.

Each histogram includes bins, each of which indicates the number of samples or counts during which power or some similar measure -- in a given frequency band -- is in a respective range. A histogram in the present instance preferably represents -- for each given frequency band -- the number of centiseconds during which loudness is in each of a plurality of loudness ranges. For example, in the third frequency band, there may be twenty centiseconds between 10dB and 20dB in power. Similarly, in the twentieth frequency band, there may be one hundred fifty out of a total of one thousand centiseconds between 50dB and 60dB. From the total number of samples (or centiseconds) and the counts contained in the bins, percentiles are derived.

A frame from the filter output of a respective frequency band is examined and bins in the appropriate histograms -- one per filter -- are incremented. The total number of bins in which the amplitude exceeds 55dB are summed for each filter (i.e. frequency band) and the number of filters indicating the presence of speech is determined. If there is not a minimum of filters (e.g. six of twenty) to suggest speech, the next frame is examined. If there are enough filters to indicate speech, a speech counter is incremented. The speech counter is incremented until 10 seconds of speech have occurred whereupon new values for T_f and T_h are defined for each filter.

The new T_f and T_h values are determined for a given filter as follows. For T_f, the dB value of the bin holding the 35th sample from the top of 1000 bins (i.e. the 96.5th percentile of speech) is defined as BIN_H. T_f is then set as: $T_f = BIN_H + 40dB$. For T_h, the dB value of the bin holding the (.01) (TOTAL BINS - SPEECH COUNT) th value from the lowest bin is defined as BIN_L. That is, BIN_L is the bin in the histogram which is 1% of the number of samples in the histogram excluding the number of samples classified as speech. T_h is then defined as: $T_h = BIN_L - 30dB$.

Returning to FIG. 6, the sound amplitudes are converted to sones and scaled based on the updated thresholds as described hereinbefore. An alternative method of deriving sones and scaling is by taking the filter amplitudes "a" (after the bins have been incremented) and converting to dB according to the expression:

$$dB = 20 \log_{10} (a) - 10 \quad (15)$$

Each filter amplitude is then scaled to a range between 0 and 120 to provide equal loudness according to the expression:

$$a^{eql} = 120 \, (dB = T_h)/(T_f - T_h) \quad (16)$$

a^{eql} is then preferably converted from a loudness level (phons) to an approximation of loudness in sones (with a 1KHz signal at 40dB mapping to 1) by the expression:

$$L^{dB} = \frac{(a^{eql} - 30)}{4} \quad (17)$$

Loudness in sones is then approximated as:

$$L_s \text{ (appr)} = 10 \, (L^{dB})/20 \quad (18)$$

The loudness in sones L_s is then provided as input to the equations (7) and (8) to determine the output firing rate f for each frequency band. With twenty-two frequency bands, a twenty-two dimension vector characterizes the acoustic wave inputs over successive time frames. Generally, however, twenty frequency bands are examined by employing a mel-scaled filter bank defined by FIG. 8.

Prior to processing the next time frame, the next state of n is determined in accordance with equation (9).

The acoustic processor hereinbefore described is subject to improvement in applications where the firing rate f and neurotransmiter amount n have large DC pedestals. That is, where the dynamic range of the terms of the f and n equations is important, the following equations are derived to reduce the pedestal height.

In the steady state, and in the absence of an acoustic wave input signal (L = 0), equation (8) can be solved for a steady-state internal state $\bar{n}$:

$$\bar{n} = A/(So + Sh) \quad (19)$$

The internal state of the neurotransmitter amount n(t) can be represented as a steady state portion and a varying portion

$$n(t) = \bar{n} + \tilde{n}(t) \quad (20)$$

Combining equations (7) and (20), the following expression for the firing rate results:

$$f(t) = (So + D \times L)(\bar{n} + \tilde{n}(t)) \quad (21)$$

The term So x $\bar{n}$ is a constant, while all other terms include either the varying part of n or the input signal represented by (D x L). Future processing will involve only the squared differ ences between output vectors, so that constant terms may be disregarded. Including equation (19) for $\bar{n}$, we get

$$f(\tilde{t}) = (So + D \times L) \times \frac{(\tilde{\bar{n}}(t) + D \times L \times A)}{(So + Sh} \qquad (22)$$

Considering equation (9), the next state becomes:

$$n(t + \Delta t) = \bar{n}(t + \Delta t) + \tilde{\tilde{n}}(t + \Delta t) \quad (23)$$

$$= \tilde{\tilde{n}}(t) + A - (So + Sh + D \times L) \times (\bar{n} + \tilde{\tilde{n}}(t)) \quad (24)$$

$$= \tilde{\tilde{n}}(t) - (Sh \times \tilde{\tilde{n}}(t) - (So + Ao \times L^A)\tilde{\tilde{n}}(t)$$

$$- (Ao \times L^A \times D)/(So + Sh) + Ao - (So \times Ao)$$

$$+ (Sn \times Ao)/(So + Sh) \quad (25)$$

This equation (25) may be rewritten, ignoring all constant terms, as:

$$\tilde{\tilde{n}}(t + \Delta t) = \tilde{\tilde{n}}(t)(1 - So\,\Delta t) - \tilde{\tilde{f}}(t) \quad (26)$$

Equations (21) and (26) now constitute the output equations and state-update equations applied to each filter during each 10 millisecond time frame. The result of applying these equations is a 20 element vector each 10 milliseconds, each element of the vector corresponding to a firing rate for a respective frequency band in the mel-scaled filter bank.

With respect to the embodiment here described, the flowchart of FIG. 6 applies except that the equations for f, dn/dt, and n(t+1) are replaced by equations (17) and (22) which define special case expressions for firing rate f and next state n (t+Δt) respectively.

In accordance with the invention as described above, the auditory model of the invention embodies, in preferred form, the following characteristics:

1. Auditory nerves respond to acoustic signals as though they were looking through critical-band-width filters.

2. In response to a zero acoustic wave input (silence) the nerves fire at some spontaneous rate.

3. The step response to a loud sound is a large firing rate which decreases with a time constant of about 30 milliseconds.

4. The neural firing rate in response to turning off a loud signal is a decrease in firing, with a recovery constant of about 50 milliseconds.

5. The steady-state responses to soft and loud sounds are a simple function of loudness, as defined psychophysically.

6. The balance between the transient response and the steady-state response is adjusted to emphasize the transient response of the system.

7. The model acts semi-independently for each critical band.

It should be realized that the preferred embodiment may be varied without departing from the scope of the invention as claimed hereinafter. First, although loudness is preferably in sones or some other compressed form, it is also possible to provide other measurements of loudness or power intensity into the equations -- at, perhaps, the expense of some of the benefits realized by using sones. Second, defining the frequency bands as the critical bands of the basilar membrane 210 is preferable but not required. Hence, although a mel-scaled filter bank of twenty or more channels may be preferred such is not required. Third, the values attributed to the terms in the various equations (namely τ_o = 5 csec, τ_{Lmax} = 3csec, Ao = 1, R = 1.5, and L_{max} = 20) may be set otherwise and the terms So, Sh, and D may differ from the preferable derived values of 0.0888, 0.11111, and 0.00666, respectively, as other terms are set differently.

The invention has been practiced using the PL/I programming language, however may be practiced by various other software or hardware approaches.

Claims

1. A method of characterizing, in the front end of a speech recognition system, an acoustic wave input by a limited number of parameters indicative of speech elements, the method comprising the steps of:

forming a model of the human ear in which the state of the system is characterized by the amount of neurotransmitter available for neural firing, the amount of neurotransmitter being variable over time; and

generating neural firing rate data which depends at least in part on

(a) the previous state of the amount of neurotransmitter and

(b) the rate of change of neurotransmitter.

2. A method as in claim 1 comprising the further step of:

storing speech recognition prototypes each of which is defined by data that is matchable against the neural firing

rate data;

the values of features which define the neural rate firing being matchable against the values of features which define the stored data for the prototypes.

3. A method as in claim 1 of processing acoustic wave input in a speech recognition system, the method comprising the steps of:

making a measurement of the loudness of the acoustic wave input for each of at least one frequency band;

determining, in an auditory model, a modelled neural firing rate for and as a function of the value of the loudness measurement at each frequency band; and

representing the acoustic wave input as the neural firing rates determined for the respective frequency bands.

4. A method as in claim 3 further comprising the steps of:

defining the neural firing rates as feature values; and

performing a matching between the values of the neural firing features and stored feature data for at least one of a plurality of prototypes.

5. A method as in claim 3 wherein said firing rate determining step for each respective frequency band includes the steps of:

determining an amount of modelled neurotransmitter which varies over time;

generating a first value which corresponds to the level of neural firing independent of acoustic wave input, the first value varying with the amount of neurotransmitter;

generating a second value which varies with the value of the loudness measurement and the amount of neurotransmitter; and

evaluating the neural firing rate as a function of the first generated value and the second generated value.

6. A method as in claim 3 wherein said firing rate determining step for a subject frequency band includes the step of multiplying the amount of time-varying neurotransmitter together with the value of the loudness measurement measured for the subject frequency band.

7. A method as in claim 3 including the further steps of:

forming a ratio between

(a) the neural firing rate for a first fixed loudness when there is no change in the amount of neurotransmitter and

(b) the neural firing rate for a second fixed loudness when there is no change in the amount of neurotransmitter over time, the first and second fixed loudness differing in magnitude;

said ratio defining a parameter which is adjustable to alter system performance.

8. Apparatus for matching an acoustic wave input to stored feature values of prototypes in a speech recognition system, the apparatus comprising:

a nonlinear processor which models the standard human ear, the processor including:

a) means for determining for at least one frequency band the respective rate of change of neurotransmitter available for neural firing;

(b) means for determining the next state associated with each frequency band as the current amount of neurotransmitter adjusted by the current rate of change of neurotransmitter; and

(c) means for generating a next neural firing rate output for each frequency band as a function of the next state of the neurotransmitter.

9. Apparatus as in claim 8 further comprising:

means for producing, for each frequency band, a measurement of loudness in sones;

wherein said firing rate generating means includes means for defining the neural firing rate for a subject frequency band as dependent at least in part on

(a) the loudness measurement taken in the respective subject frequency band and

(b) the level of neurotransmitter available for neural firing in the respective subject frequency band.

10. Apparatus as in claim 9 wherein the loudness measurement producing means includes means for converting power intensity derived from the acoustic wave input into sones;

the sones providing the loudness measure upon which said neural firing rate depends.

11. Apparatus as in claim 8 further comprising:

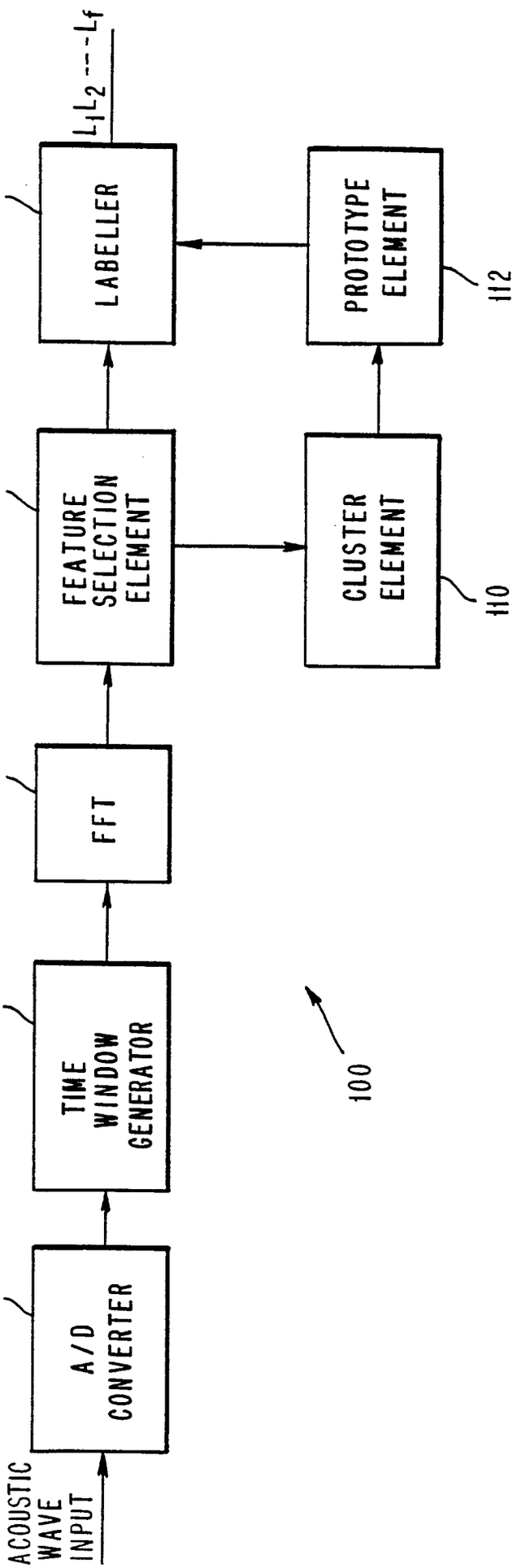
means for storing a set of values for each of a plurality of prototypes; and

means for performing matching between the generated neural firing rate values and the stored sets of values.

12. Apparatus as in claim 9 further comprising:

means for normalizing the loudness measurements between two threshold levels.

FIG. 1


ACOUSTIC WAVE INPUT
A/D CONVERTER
102
TIME WINDOW GENERATOR
104
FFT
106
FEATURE SELECTION ELEMENT
108
LABELLER
114
$L_1 L_2 \cdots L_f$
CLUSTER ELEMENT
110
PROTOTYPE ELEMENT
112
100

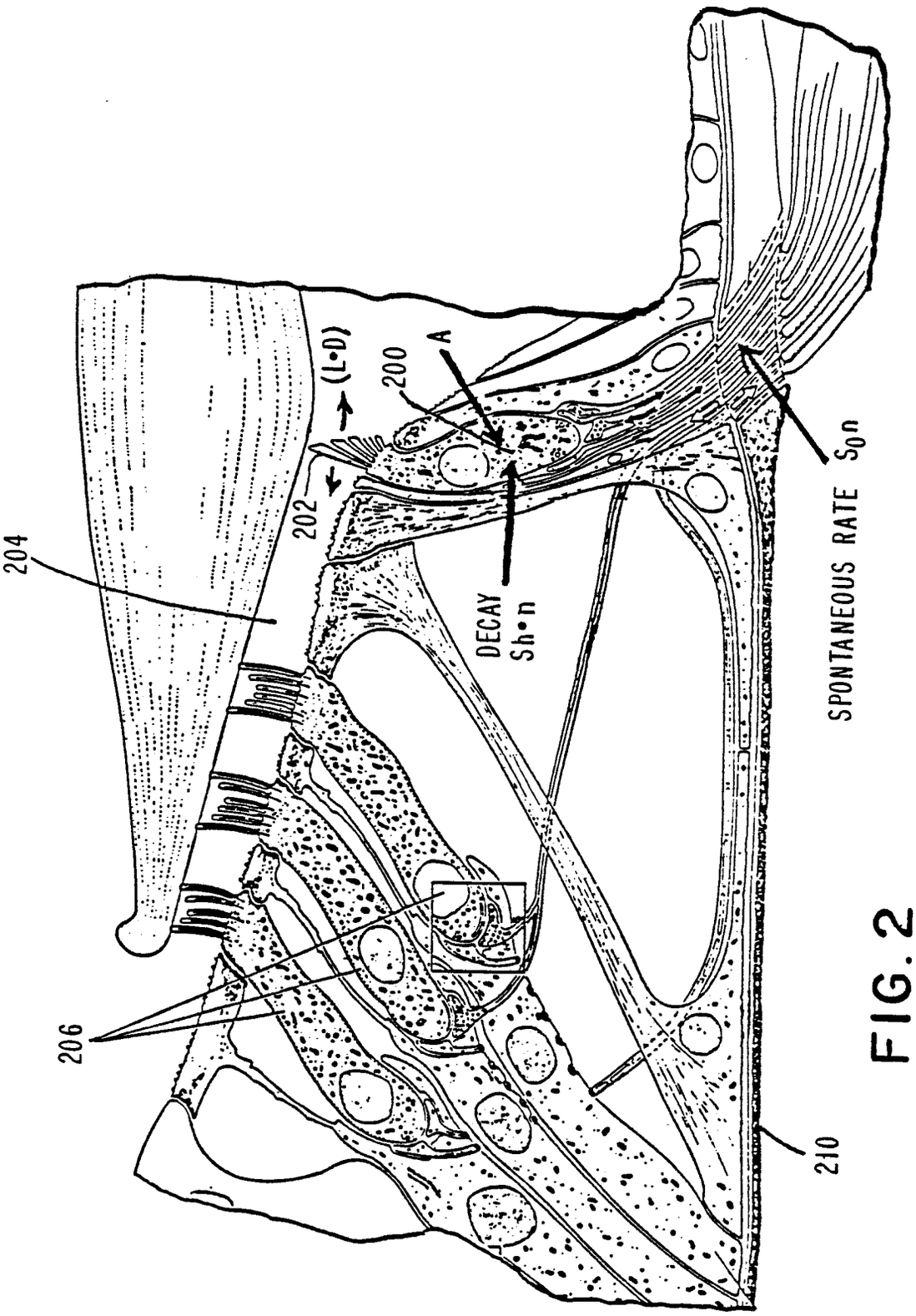


FIG. 2

FIG. 3

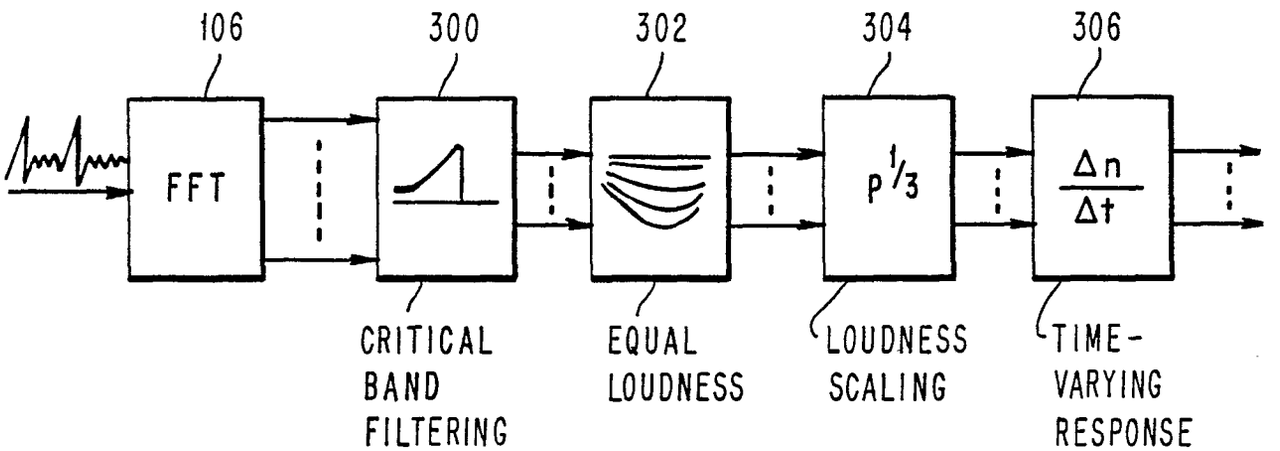

106
300
302
304
306
FFT
$p^{1/3}$
$\frac{\Delta n}{\Delta t}$
CRITICAL BAND FILTERING
EQUAL LOUDNESS
LOUDNESS SCALING
TIME-VARYING RESPONSE

FIG. 4

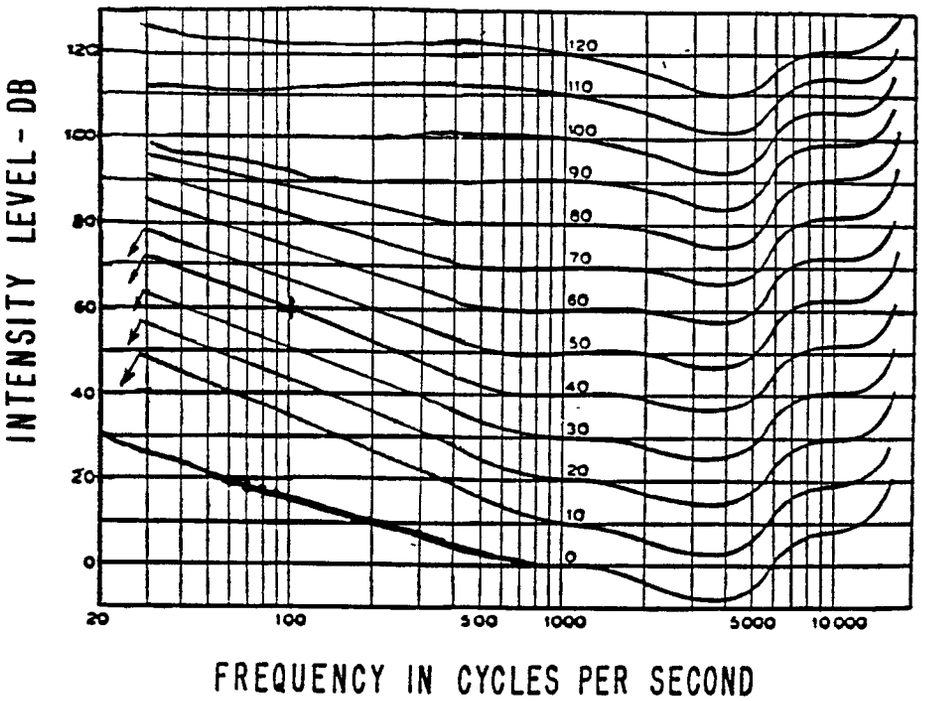

INTENSITY LEVEL - DB
120
100
80
60
40
20
0
20
100
500
1000
5000
10 000
120
110
100
90
80
70
60
50
40
30
20
10
0
FREQUENCY IN CYCLES PER SECOND

FIG. 5

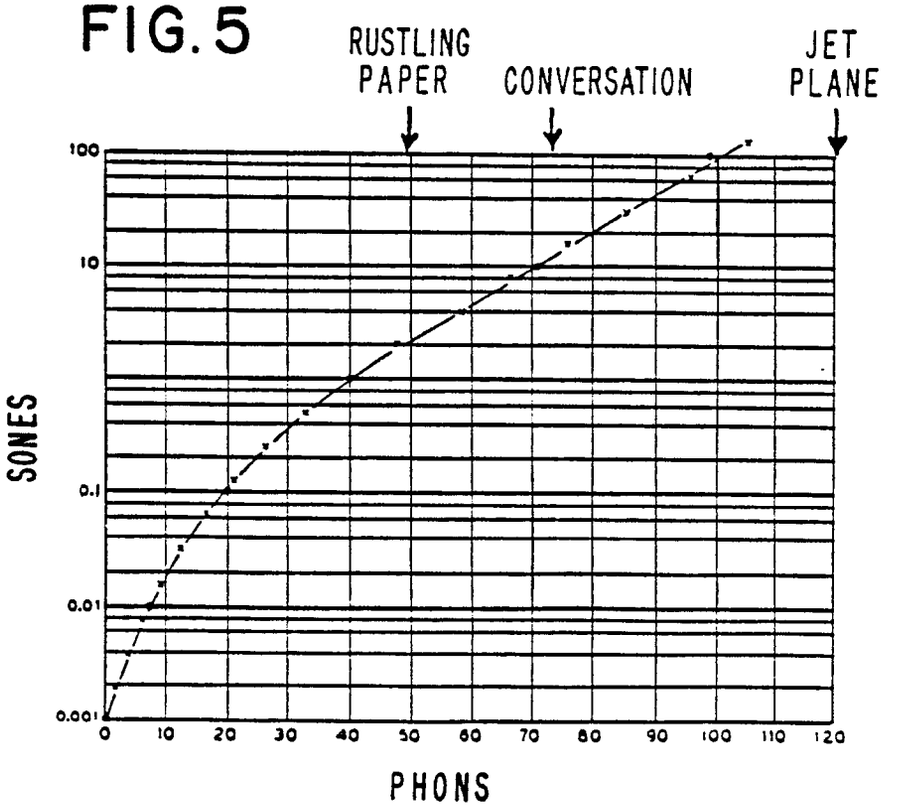

RUSTLING PAPER
CONVERSATION
JET PLANE
SONES
100
10
0.1
0.01
0.001
0
10
20
30
40
50
60
70
80
90
100
110
120
PHONS

FIG. 6

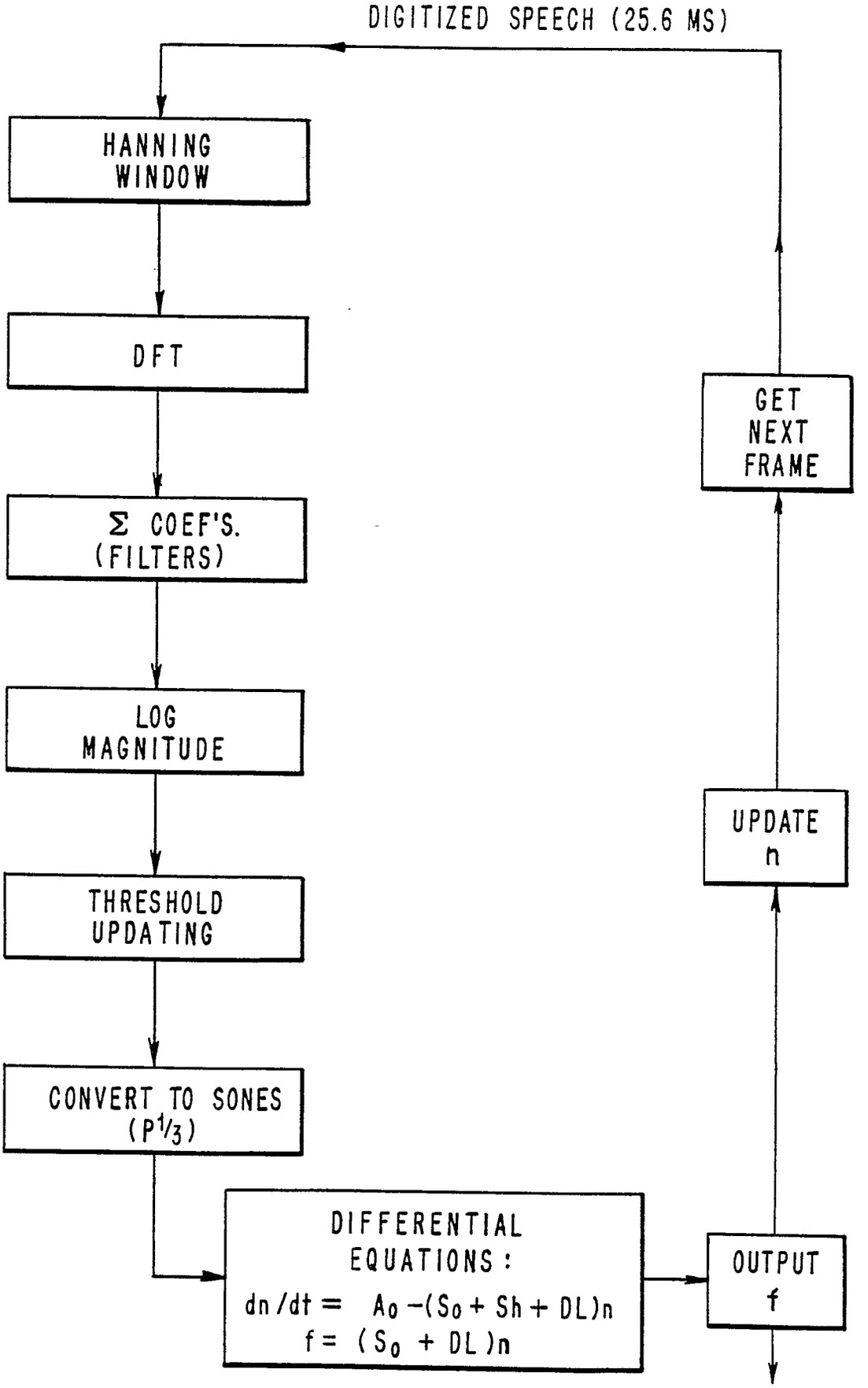
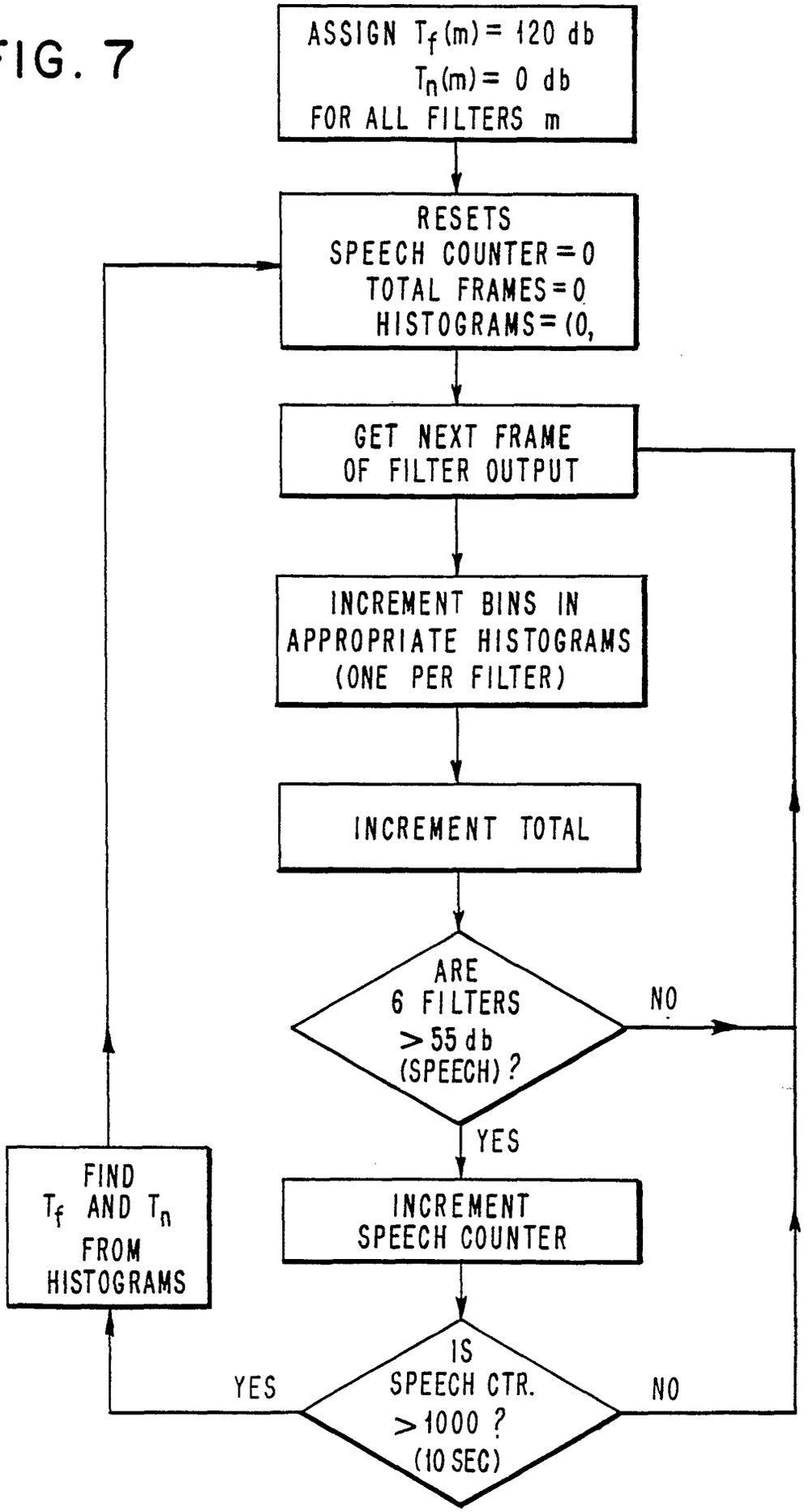

DIGITIZED SPEECH (25.6 MS)
HANNING WINDOW
DFT
Σ COEF'S. (FILTERS)
LOG MAGNITUDE
THRESHOLD UPDATING
CONVERT TO SONES ($P^{1/3}$)
DIFFERENTIAL EQUATIONS:
$dn/dt = A_0 - (S_0 + Sh + DL)n$
$f = (S_0 + DL)n$
OUTPUT f
UPDATE n
GET NEXT FRAME

FIG. 7


ASSIGN $T_f(m) = 120$ db
$T_n(m) = 0$ db
FOR ALL FILTERS m
RESETS
SPEECH COUNTER = 0
TOTAL FRAMES = 0
HISTOGRAMS = (0,
GET NEXT FRAME
OF FILTER OUTPUT
INCREMENT BINS IN
APPROPRIATE HISTOGRAMS
(ONE PER FILTER)
INCREMENT TOTAL
ARE
6 FILTERS
> 55 db
(SPEECH) ?
NO
YES
INCREMENT
SPEECH COUNTER
IS
SPEECH CTR.
> 1000 ?
(10 SEC)
YES
NO
FIND
T_f AND T_n
FROM
HISTOGRAMS