OPIC
Office de la propriété intellectuelle du Canada

CIPO
Canadian Intellectual Property Office

(12) (19) (CA) **Brevet-Patent**

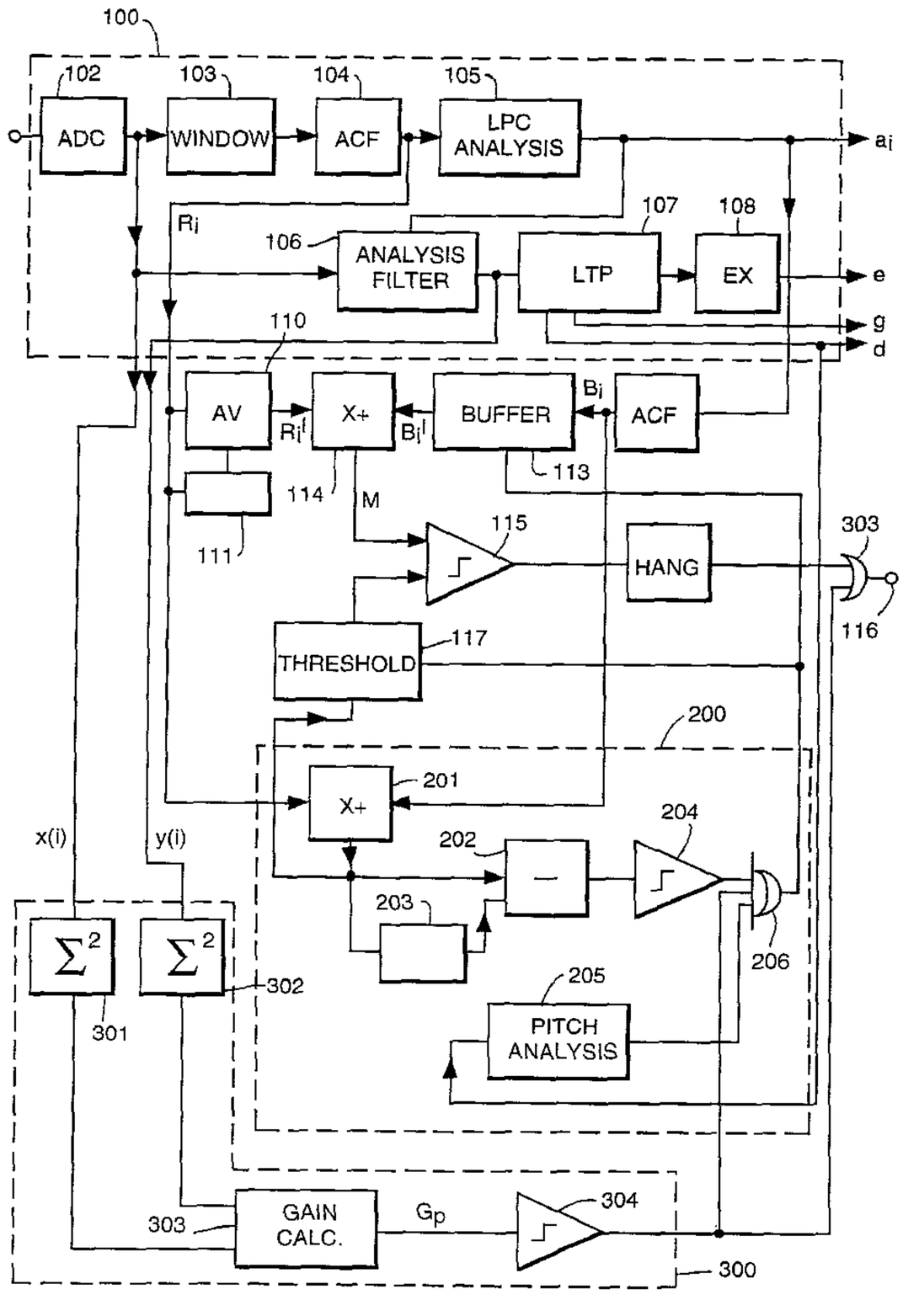
(11) (21) (C) **2,169,745**
(86) 1994/09/14
(87) 1995/03/23
(45) 2000/05/16

(72) Barrett, Paul Alexander, GB
(73) British Telecommunications public limited company, GB

(51) Int.Cl.[6] G10L 9/14
(30) 1993/09/14 (93307211.8) EP
(30) 1993/12/06 (9324967.0) GB
(30) 1994/06/21 (9412451.8) GB

(54) **DETECTEUR D'ACTIVITE VOCALE**

(54) **VOICE ACTIVITY DETECTOR**



(57) Des signaux vocaux sont distingués d'un bruit par une comparaison spectrale (114, 115) d'un signal d'entrée et d'une estimation de bruit mémorisée (113). La mise à jour de l'estimation de bruit (dans une mémoire tampon (113)) peut être effectuée au cours de périodes ne contenant pas de signaux vocaux sous la

(57) Speech is distinguished from noise by a spectral comparison (114, 115) of an input signal with a stored noise estimate (113). Updating of the noise estimate (in a buffer (113)) is permitted during periods when speech is absent under control of an auxiliary detector (200). In order to improve operation in the presence of signals

OPIC
Office de la propriété intellectuelle du Canada

CIPO
Canadian Intellectual Property Office

(11) (21) (C) **2,169,745**
(86) 1994/09/14
(87) 1995/03/23
(45) 2000/05/16

commande d'un détecteur auxiliaire (200). Afin d'améliorer le fonctionnement du détecteur en présence de signaux à composantes harmoniques puissantes (tels que des tonalités de signalisation), un gain de prédiction du codage à prédiction linéaire (LPC) est calculé à partir du signal d'entrée (x(i)) et d'un signal résiduel (y(i)) obtenu du signal d'entrée après le filtrage par un filtre (105) présentant une réponse complémentaire au spectre de fréquence du signal d'entrée, et, si le gain dépasse le seuil, la mise à jour de la mémoire tampon est supprimée.

with strong harmonic components (e.g. signalling tones) an LPC prediction gain is computed from the input (x(i)) and a residual (y(i)) obtained from the input following filtering by a filter (105) having a response complementary to the frequency spectrum of the input, and if the gain exceeds a threshold, buffer updating is suppressed.

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G10L 3/00, H04Q 1/46

A1

(11) International Publication Number: WO 95/08170

(43) International Publication Date: 23 March 1995 (23.03.95)

(21) International Application Number: PCT/GB94/01999

(22) International Filing Date: 14 September 1994 (14.09.94)

(30) Priority Data:

93307211.8	14 September 1993 (14.09.93)	EP
(34) Countries for which the regional or international application was filed:		GB et al.
9324967.0	6 December 1993 (06.12.93)	GB
9412451.8	21 June 1994 (21.06.94)	GB

(60) Parent Applications or Grants
(63) Related by Continuation

US	08/158,852 (CIP)
Filed on	29 November 1993 (29.11.93)
US	08/232,475 (CIP)
Filed on	25 April 1994 (25.04.94)

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: BARRETT, Paul, Alexander [GB/GB]; 18 Fletchers Lane, Kesgrave, Ipswich, Suffolk IP5 7XY (GB).

(74) Agent: LLOYD, Barry, George, William; BT Group Legal Services, Intellectual Property Dept., 13th floor, 151 Gower Street, London WC1E 6BA (GB).

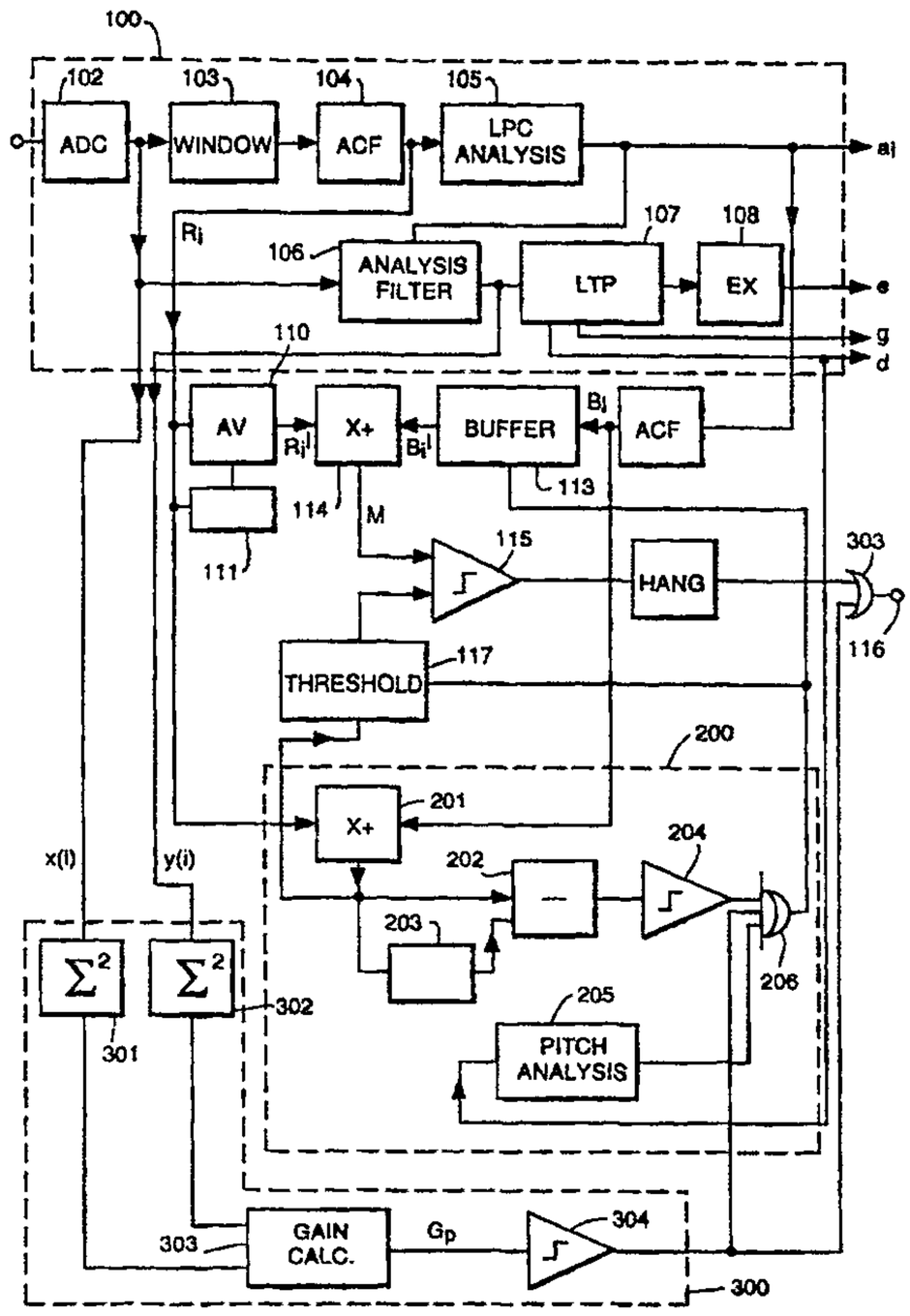
(81) Designated States: AU, BG, BR, CA, CN, CZ, FI, HU, JP, KR, NO, NZ, PL, RO, RU, SK, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: VOICE ACTIVITY DETECTOR

(57) Abstract

Speech is distinguished from noise by a spectral comparison (114, 115) of an input signal with a stored noise estimate (113). Updating of the noise estimate (in a buffer (113)) is permitted during periods when speech is absent under control of an auxiliary detector (200). In order to improve operation in the presence of signals with strong harmonic components (e.g. signalling tones) an LPC prediction gain is computed from the input (x(i)) and a residual (y(i)) obtained from the input following filtering by a filter (105) having a response complementary to the frequency spectrum of the input, and if the gain exceeds a threshold, buffer updating is suppressed.



VOICE ACTIVITY DETECTOR

A voice activity detector is a device which is supplied with a signal with the object of detecting periods of speech, or periods containing only noise. Although the present invention is not limited thereto, one application of particular interest for such detectors is in mobile radio telephone systems where the knowledge as to the presence or otherwise of speech can be exploited to reduce power consumption and interference by turning off a transmitter during periods of silence. Here also the noise level (from a vehicle-mounted unit) is likely to be high. Another possible use in radio systems is to improve the efficient utilisation of radio spectrum.

Figure 1 shows a voice activity detector as described in our International Patent Application WO89/08910.

Noisy speech signals are received at an input 1. A store 2 contains data defining an estimate or model of the frequency spectrum of the noise; a comparison is made (3) between this and the spectrum of the current signal to obtain a measure of similarity which is compared (4) with a threshold value. In order to track changes in the noise component, the noise model is updated from the input only when speech is absent. Also, the threshold can be adapted (adaptor 6).

In order to ensure that adaptation occurs only during noise-only periods, without the danger of progressive incorrect adaptation following a wrong decision, adaptation is performed under the control of an auxiliary detector 7, which comprises an unvoiced speech detector 8 and a voiced speech detector 9: the detector 7 deems speech to be present if either of the detectors recognises speech, and suppresses updating and threshold adaptation of the main detector. Typically the unvoiced speech detector 8 obtains a set of LPC

coefficients for the signal and compares the autocorrelation function of these coefficients between successive frame periods, whilst the voiced speech detector 9 examines variations in the autocorrelation of the LPC residual.

This arrangement is very successful in distinguishing between periods of speech and periods during which only noise is received. However, a problem arises in that signalling tones are often assumed by the auxiliary detector to be simply noise (i.e. it does not recognise them as speech) so that the main detector adapts to the tones as if they were noise, and transmission of the tones is prevented, or at least terminated prematurely.

This problem could be overcome by provision of tone detectors each tuned to the frequency(s) of a particular signalling tone; however, the diversity of different signalling tones throughout the world is considerable, so that a large number of individual detectors would be needed in order, for example, that a mobile telephone user making an international call may be able to hear the 'engaged' tone reliably, irrespective of the country from which it originates.

According to the present invention, there is provided a voice activity detector for detecting the presence of speech in an input signal, comprising

(a) means for storing an estimate of the noise component of an input signal;

(b) means for recognising the spectral similarity of the input signal and the stored estimate to produce an output decision signal;

(c) means for updating the stored estimate;

(d) an auxiliary detector arranged to control the updating means so that updating occurs only when speech is indicated by the auxiliary detector to be absent from the input signal;

characterised by means operable to calculate a prediction gain parameter for the input signal, and modifying means

arranged to suppress updating in the event that the prediction gain exceeds a threshold value.

Some embodiments of the invention will now be described, by way of example, with reference to the accompanying drawings, in which:

Figure 2 is a block diagram of a speech coder with a voice activity detector in accordance with one aspect of the present invention;

Figures 3 and 4 show graphically prediction gain values from various input signals;

Figures 5, 6 and 7 are block diagrams of further embodiments of the inventor.

In figure 2, a conventional speech coder 100 has a speech input 101, the speech signal being sampled at 8kHz and converted into digital form by an analogue-to-digital converter 102. A windowing unit 103 divides the speech samples into frames of (for example) 160 samples (i.e. a 20ms frame) and multiplies it by a Hamming window or other function which reduces the contribution of samples at the beginning and end of the frame. A correlator 104 receives the digitised speech samples and produces the autocorrelation coefficients R_i for each frame. An LPC analysis unit 105 calculates the coefficients a_i of a filter (sometimes referred to as a synthesis filter) having a frequency response corresponding to the frequency spectrum of the input speech signal using a known method e.g. a Levinson-Durbin or Schurr-algorithm.

The digitised input signal is also passed through an inverse filter (or analysis filter) 106 controlled by the coefficients, to produce a residual signal which is further analysed by a long term predictor analysis unit 107 which computes the optimum delay for predicting the LPC residual from its previous values, and a corresponding gain value for the prediction. The analysis unit 106 also forms a second residual (i.e. the difference between the current LPC residual and the LPC residual when delayed and scaled by the parameters obtained). An excitation unit 108 derives

excitation parameters for transmission to a decoder, by simply quantisising the LTP residual, or by other conventional means.

The LPC coefficients a_i, the long term predictor delay $\hat{d}$ and gain g, and excitation parameters e are transmitted to a decoder.

A main voice activity detector in accordance with our earlier patent application averages the autocorrelation coefficients R_i by means of an averager 110 which produces a weighted sum R_i' of the current coefficients and those from previous frames stored in a buffer 111. A further autocorrelator 112 forms the autocorrelation coefficients B_i of the LPC coefficients a_i which are passed to a buffer 113. The contents of the buffer are updated only during periods deemed by an auxiliary detector (to be described below) to contain only noise, so that the contents of the buffer 113 B_i' represent an estimate of the noise spectrum of the input signal. A multiplication/addition unit 114 forms a measure M of the spectral similarity between the input signal and the noise model defined as

$$M = B_0' + 2\sum_{i=1}^{n} \frac{R_i' B_i'}{R_0'}$$

Where a zero suffix signifies the zero order autocorrelation coefficient and n is the number of samples in a speech frame.

The measure M is compared in a comparator 115 against a threshold level and produces at an output 116 a signal indicating the presence of absence of speech. The threshold may be adaptively adjusted (117) according to the current noise power level.

The updating of the noise estimate in the buffer store 113 is not controlled by the output 116 of the detector just described, since failure to recognise speech would result in updating of the buffer with speech information and consequent

further recognition failures - a "lock" situation. Therefore updating is controlled by an auxiliary detector 200. In order to distinguish between noise and unvoiced speech, this forms (201) a sum of products of the (unaveraged) autocorrelation coefficients Ri of the input and the (unbuffered) autocorrelation coefficients Bi of the LPC coefficients. A subtractor 202 compares this sum with the corresponding sum for a previous speech frame, delayed in a buffer 203. This difference representing the spectral similarity between successive frames of the input signal is thresholded (204) to produce a decision signal.

For recognising voiced speech, the long term predictor delay d is measured by a pitch analysis unit 205. The outputs of this is combined with that of the thresholding stage 204 in an OR gate 206 - i.e. speech is deemed by the auxiliary detector 200 to be present if either (or both) of the units 204 or 205 products an output indicating that speech is present. As discussed in the introduction, if a system is to pass signalling tones, these must be recognised as speech rather than as noise, and the auxiliary detector just described is not very effective at achieving this. Although it recognises some tones others (generally those with a relatively pure spectral content) are not recognised. Once the auxiliary detector 200 has failed, the main detector also fails since the noise estimate in the buffer 113 is then "trained" on the signalling tone.

Accordingly, a further auxiliary detector is provided for the detection of signalling tones. Preferably this makes use of the observation that signalling tones, being artificially generated, contain a small number of frequency components (which may be modulated). The performance of an LPC predictor is exceptionally high for such signals, and this is made use of to discriminate between tone-based signals (including multi-tone signals) and background or environmental noise signals.

The LPC prediction gain Gp is defined as the ratio of the input signal power to the output signal power for a frame of speech viz is

$$Gp = \frac{\sum_{i=0}^{n-1} x^2(i)}{\sum_{i=0}^{n-1} y^2(i)}$$

where x_i is the filter input and y_i is the output of the inverse filter:

$$y(t) = x(t) + \sum_{i=1}^{m} y(t-i)a_i$$

(where m is the number of filter coefficients, typically 8 or 10). Signals x(i) and y(i) are available from the LPC coder 100, at the outputs of converter 102 and filter 106 respectively. These values are squared (301, 302) and the prediction gain is obtained by an arithmetic unit 303 which calculates Gp according to the above equation. Its output is compared by a comparator 304 with a fixed threshold value T; if the gain exceeds the threshold (typically T = 63 or 18 dB), a tone is considered to be recognised. There are several possible responses to tone recognition:

(a) to override the main detector output by means of an OR gate 303

(b) to override the auxiliary detector by means of a third input to the OR gate 206

(c) both of these (as shown)

Of course, instead of calculating the quotient, the Σx^2 term can be compared with the Σy^2 multiplied by the threshold value. Figure 3 shows histograms of prediction gains in dB obtained from background environmental noise, speech,

background noise in signalling tones, and the signalling tones themselves, whilst Figure 4 shows plots of prediction gain against time for different UK signalling tones, viz.

'Subscriber Engaged' tone

Dial tone

Ring tone

'Number Unobtainable' tone

'Equipment engaged' tone

In practice, subscriber engaged tone, dial tone and 'number unobtainable' tone are successfully recognised by the further detector, as indeed are multifrequency tones (e.g. from a keypad). Ring tone and 'equipment engaged' tone are recognised by the pitch analysis unit 205.

The further detector 300 may be considered as a detector for certain types of tone; alternatively (in the embodiment of figure 2) it may be viewed as detecting a situation where the residual y_1 is small, so that operation of the long term predictor 107 (and hence of the pitch analysis 205) is not robust.

An alternative option for detecting voiced speech is to replace the pitch detector 205 with items analogous to 301, 302, 303 and 304 to form (and threshold) a prediction gain based on the longterm predictor analysis 107.

Two further modifications to the apparatus of Figure 2 will now be described with reference to Figure 5. Firstly, in the embodiment showing in Figure 2, the prediction gain calculated is that of the LPC analysis of the speech coder 100, which might typically employ an 8th or even 10th order predictor. However, noting that the basis of this part of the analysis is that information tones result in higher prediction gains than does environmental noise, and that the higher the order of the analysis the higher is the ability of the predictor to model the noise environment, it is found that, by limiting the gain calculation to a fourth order analysis, information signals consisting of one or two tones

give a high prediction gain whilst the prediction gain for environmental noise can be reduced.

In principle this could be achieved by providing a fourth order analysis and filter alongside the eighth-order units 105, 106, to feed the auxiliary detector. However it is simpler to compute the prediction gain from reflection coefficients (sometimes referred to as Parcor coefficients). In Figure 5 these are calculated in known manner by a unit 400 from the autocorrelation coefficients R_i (though, depending on the design of the speech coder it might be possible to pick them up from an intermediate point with the LPC analysis unit 105). A measure of the prediction gain can be obtained by computing from the first four reflection coefficients Rc_i a prediction error Pe, as follows.

$$Pe = \prod_{1=1}^{4} (1-Rc_i^2)$$

this being performed at 401. A high prediction error corresponds to a low prediction gain and vice versa, so that a signalling tone is deemed to be present if Pe is less than a threshold value Pth. This comparison 403 replaces comparison 304 of Figure 2.

Secondly, noise in a mobile radio environment contain very strong resonances at low frequencies, and a further test is made to determine whether the "tone" is below a threshold frequency. Selection of a threshold involves a degree of compromise but, since most signalling tones lie above 400Hz, 385 Hz is suggested.

This further test operates by determining the frequencies of the poles of the LPC filter. A low order filter is preferred to reduce the complexity of analysis. Again, a further LPC analysis could be performed but it is easier to proceed as in Figure 5, by computing the LPC coefficients from the reflection coefficients. Supposing that only the first two reflection coefficients from unit 400 are used, then the LPC coefficients a_i are calculated in

conventional manner by a unit 404, being defined such that the synthesis filter response is

$$H(z) = 1/ \{a_0 + a_1 z^{-1} + a_2 z^{-2}\}$$

Then the positions of the poles in the z-plane are given by the solution to the quadratic equation:

$$a_0 z^2 + a_1 z + a_2 = 0 \qquad a_0 = 1$$

$$z = \frac{-a_1}{2} \pm j\sqrt{\frac{4a_2 - a_1^2}{4}}$$

i.e.

If the term inside the square root is negative then the pole lies on the real axis and the signal is not a tone. If it is positive, but the real part of the pole position is negative (i.e. $a_1<0$) then the pole is in the left-hand half of the z-plane. This necessarily implies that the frequency is more than 25% of the sampling rate - i.e. above 2000Hz for a sampling frequency f_s of 8kHz, in which case the frequency calculation is unnecessary and a ">385" signal can be generated right away.

The pole frequency is given by:

$$f = \arctan\left\{\frac{-\sqrt{4a_2 - a_1^2}}{a_1}\right\} \times \frac{f_s}{2\pi}$$

The condition that f<385 Hz can be written (avoiding square roots) as:

$$(4a_2 - a_1^2) / a_1^2 < \tan^2 \left\{\frac{2\pi \times 385}{f_s}\right\}$$

OR

$$(4a_2 - a_1^2) / a_1^2 < 0.0973 \qquad at\ f_s = 8kHz$$

This calculation is performed by unit 405.

Its output is combined in an and-gate 406 with that of the comparator 403 so that a 'tone' decision is produced only when both the prediction gain is high and the pole frequency is greater than 385Hz.

If desired, pole frequencies above 2000Hz (or some other upper limit) may also be trapped so that high-frequencies above the expected signalling tone range may not be recognised as tones.

If the extra computation in solving a quartic equation can be accommodated, then it is possible to use the third and fourth reflection coefficients too; in this case two complex conjugate pairs of poles - with two associated frequencies could potentially be identified, it being envisaged that a tone would not be considered to be present if both frequencies were below the threshold.

It has already been mentioned that the embodiments of Figures 2 and 5 employ a Hamming window prior to the autocorrelation calculation 103 (as is common with autocorrelation-based LPC analysis). If it is desired not to perform such windowing in the speech coder, then a possible alternative is in the case of Figure 5 to omit the windowing 103 and to replace the reflection coefficient calculation 400 by a conversion of autocorrelation values into covariance values, units 401, 404 being modified to use covariance values rather than reflection coefficients. Alternatively, as shown in Figure 6 (which shows only those parts which have change relating to Figure 5), the initial processing may be by means of a covariance analysis 109, the output of which is supplied to a reflection coefficient calculator 400' and a modified autocorrelation coefficient unit 104'. The LPC analysis unit 105 may be connected as before to the autocorrelation unit 104' or - as shown - directly to the covariance analysis unit 109.

The above-described 'tone-detection' embodiments produce good results; they may, however, fail on mechanically generated tones employed in some territories, as these tend to have a high harmonic content resulting in low prediction gain. Simply filtering out the higher harmonics is not a solution since the insertion of a filter tends to increase the autocorrelation of all signals and hence higher prediction gains for other signals too. It is found that the predictor tends to model the filter poles rather than the characteristics of the input signal. We have however discovered that good results can be obtained using filtering if the prediction gain analysis can be constrained to assess the predictability of the signal only within a frequency range corresponding to the passband of the harmonic filter. This can be achieved by subsampling the signal at a frequency of twice the filter bandwidth prior to the prediction gain analysis.

Thus the embodiment of Figure 7, similar in other respects to Figure 5, employs filter 450, this is a low pass equiripple FIR filter having zeros on the unit circle having a passband up to 600 (3dB point) and having a stopband attenuation of 20dB at 1200 Hz. It is thought preferable that the stopband attenuation not be too great. The filter output is subsampled at 1200 Hz in subsampling unit 451.

With this filtering applied, the opportunities for the tone detection to share components with the speech coder 100 are of course much reduced; thus the filter 450 is fed directly with the digitised input signal from the analogue-to-digital converter 102, and feeds a reflection coefficient analysis unit 400", or covariance or autocorrelation analysis as discussed earlier. The autocorrelation option will require windowing as explained above.

Another embodiment alleviates the "harmonics" problem without unduly limiting the frequency range of prediction gain analysis; this is achieved by using filters to divide the signal into two or more frequency bands each of which is narrow enough that it cannot contain both the fundamental and

the third harmonic of a tone. Each channel is then subsampled and subjected to a separate prediction gain analysis.

Thus in figure 8, the signal is divided into frequency bands 400-1200 Hz and 1200-2000 Hz by filters 450a, 450b, and subsampled at 1.6 kHz (451a, 451b). Reflection coefficient computation 400" a,b, prediction error analysis 401a,b and thresholding 403a,b are performed separately for the two bands. The two outputs from comparators 403a, 403b are conducted to separate inputs of the OR gate 206, so that a high prediction gain in either of the channels is considered to indicate the presence of a tone. The other items 100-303 of Figure 7 are not shown in figure 8 as they are unchanged.

CLAIMS

1. A voice activity detector for detecting the presence of speech in an input signal, comprising

(a) means for storing an estimate of the noise component of an input signal;

(b) means for recognising the spectral similarity of the input signal and the stored estimate to produce an output decision signal;

(c) means for updating the stored estimate;

(d) an auxiliary detector arranged to control the updating means so that updating occurs only when speech is indicated by the auxiliary detector to be absent from the input signal;

characterised by means operable to calculate a prediction gain parameter for the input signal, and modifying means arranged to suppress updating in the event that the prediction gain exceeds a threshold value.

2. A voice activity detector according to claim 1 in which the auxiliary detector includes a voiced speech detector responsive to signals derived from an LPC residual signal.

3. A voice activity detector according to claim 1 or 2 in which the gain parameter represents the prediction gain of an LPC prediction of 6th order or less.

4. A voice activity detector according to claim 3 in which the gain parameter represents the prediction gain of an LPC prediction of 4th order or less.

5. A voice activity detector according to any one of the preceding claims further including means for detecting one or more primary frequency components of the input signal, and means operable to compare the frequencies with a

predetermined threshold and to permit suppression of updating only when a said primary component exceeds the threshold.

6. A voice activity detector according to any one of the preceding claims in which the gain calculation means is preceded by a filter to remove an upper portion of the frequency range of the input signal, the gain calculation being performed only for frequency components lying within the passband of the filter.

7. A voice activity detector according to Claim 6 having two filters having respective passbands and respective gain calculation means to calculate gain parameters for the respective passbands, the modifying means being arranged to suppress updating in the event of the prediction gain within either passband exceeds a threshold value.

8. A voice activity detector according to Claim 6 or 7 including means for subsampling the filtered signal(s).

Fig.1.

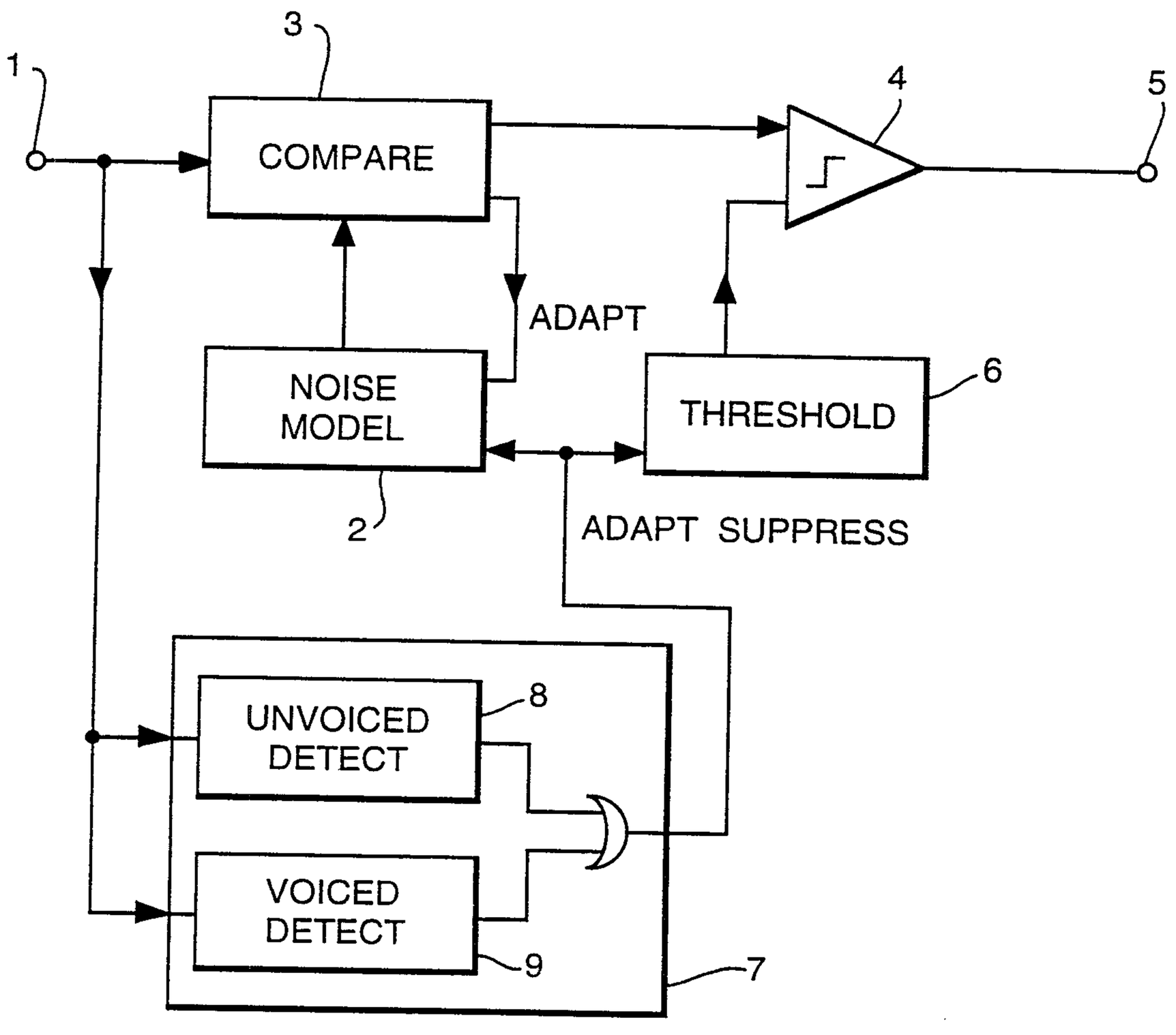
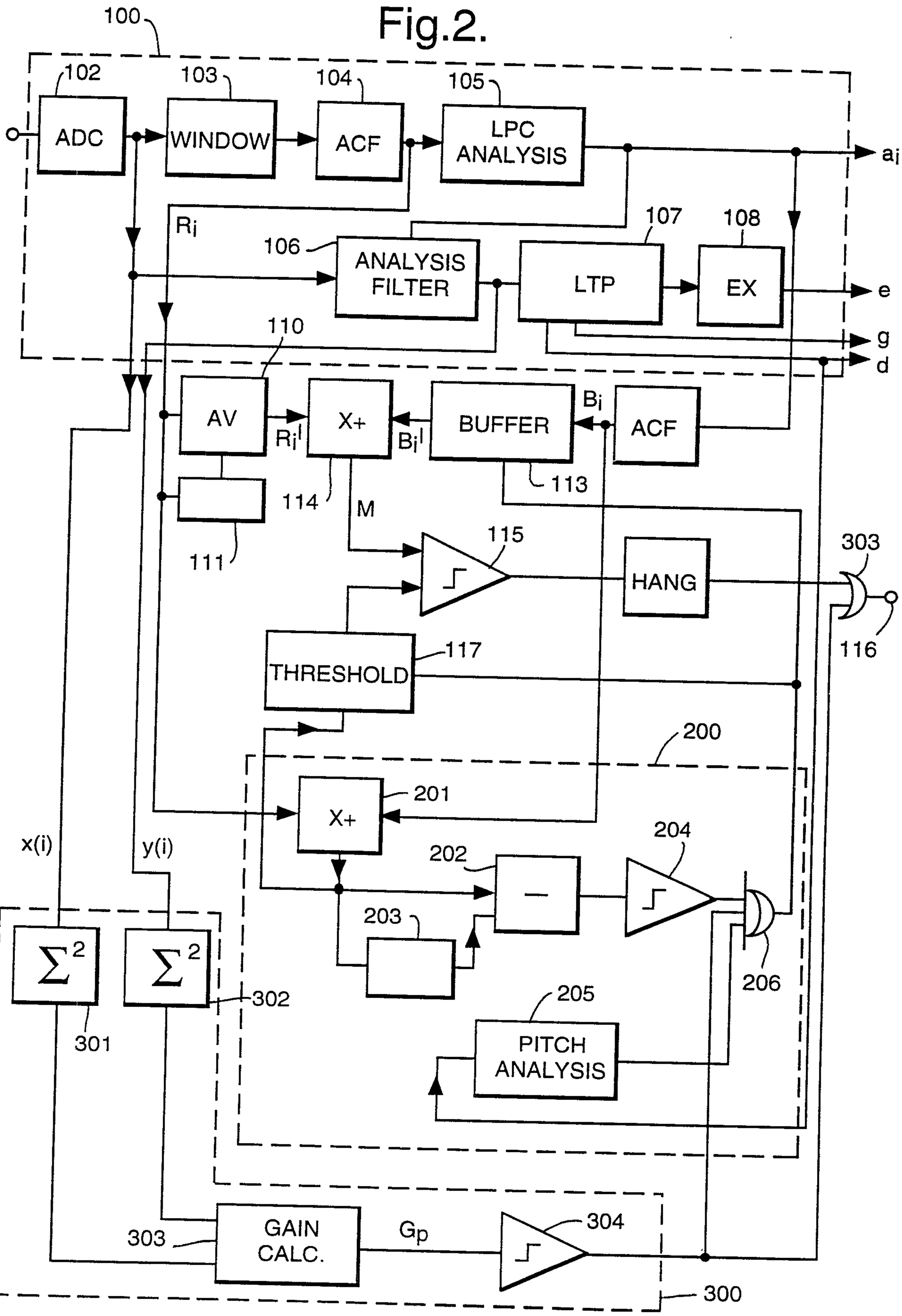


Fig.2.


100
102
ADC
103
WINDOW
104
ACF
105
LPC ANALYSIS
ai
Ri
106
ANALYSIS FILTER
107
LTP
108
EX
e
g
d
110
AV
Ri'
X+
Bi'
BUFFER
Bi
ACF
113
114
M
111
115
HANG
303
116
THRESHOLD
117
200
201
X+
x(i)
y(i)
202
204
203
206
205
PITCH ANALYSIS
302
301
303
GAIN CALC.
Gp
304
300

2169745

Fig.3.

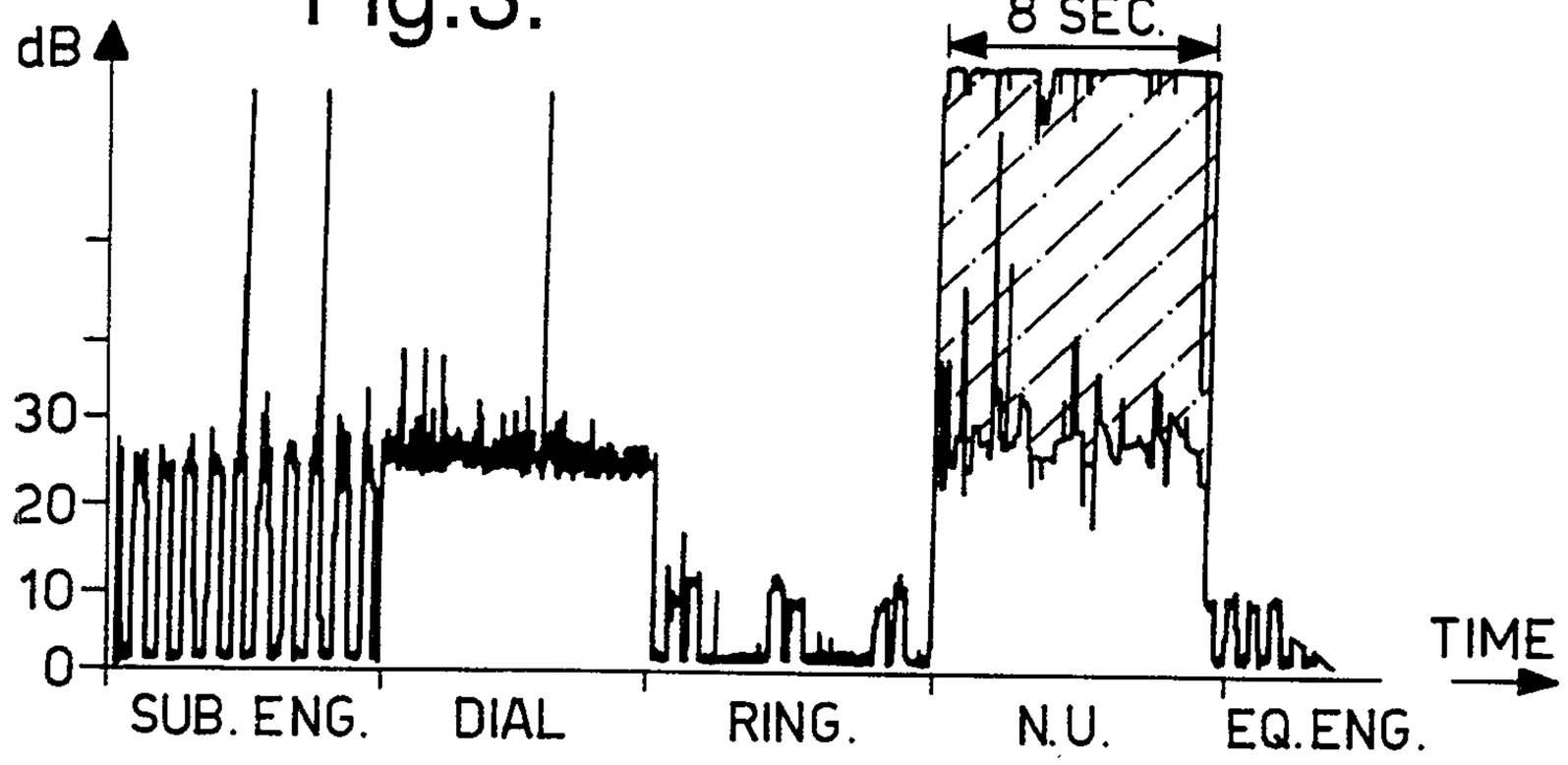
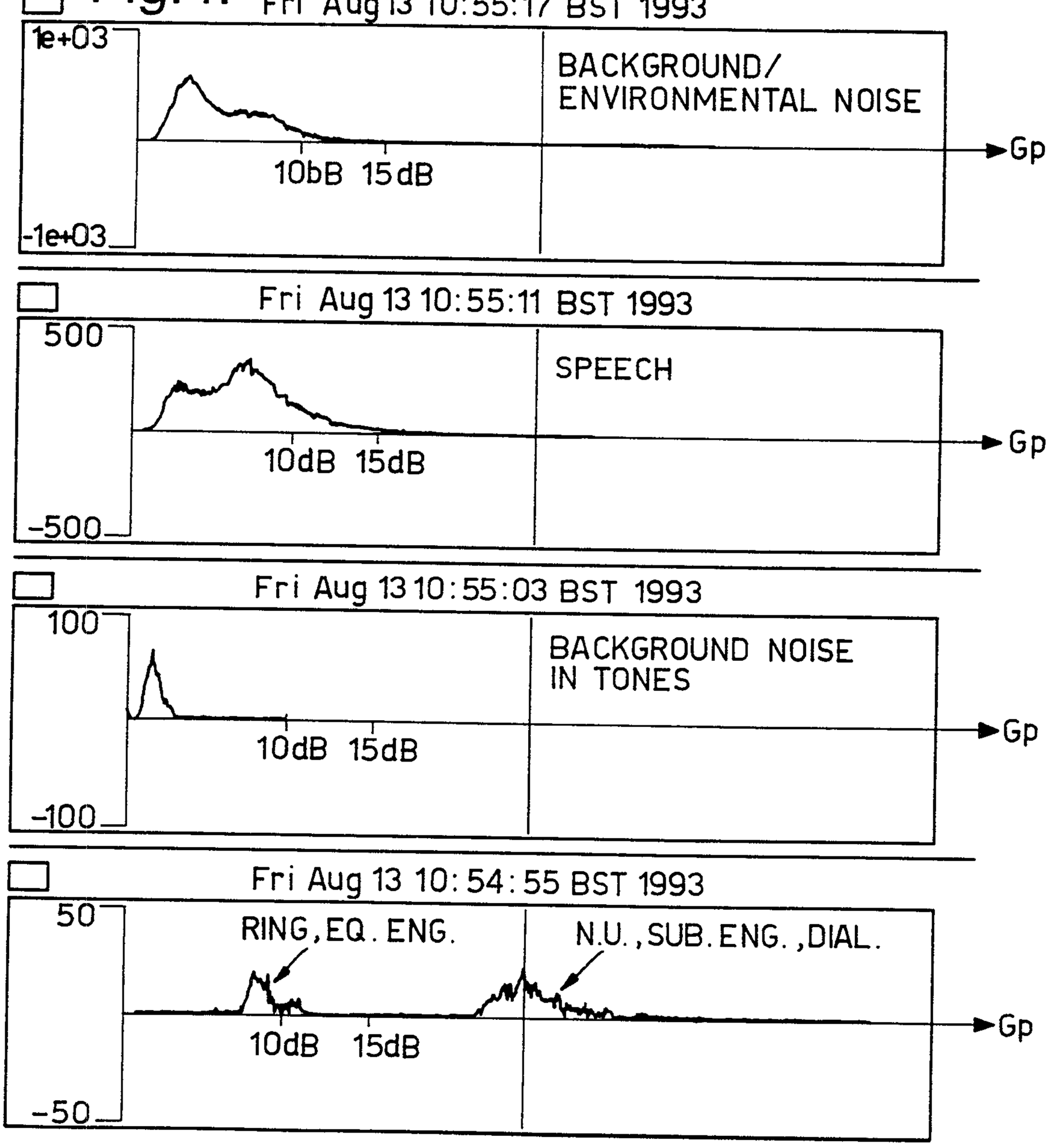


Fig.4.



2169745

Fig.5.

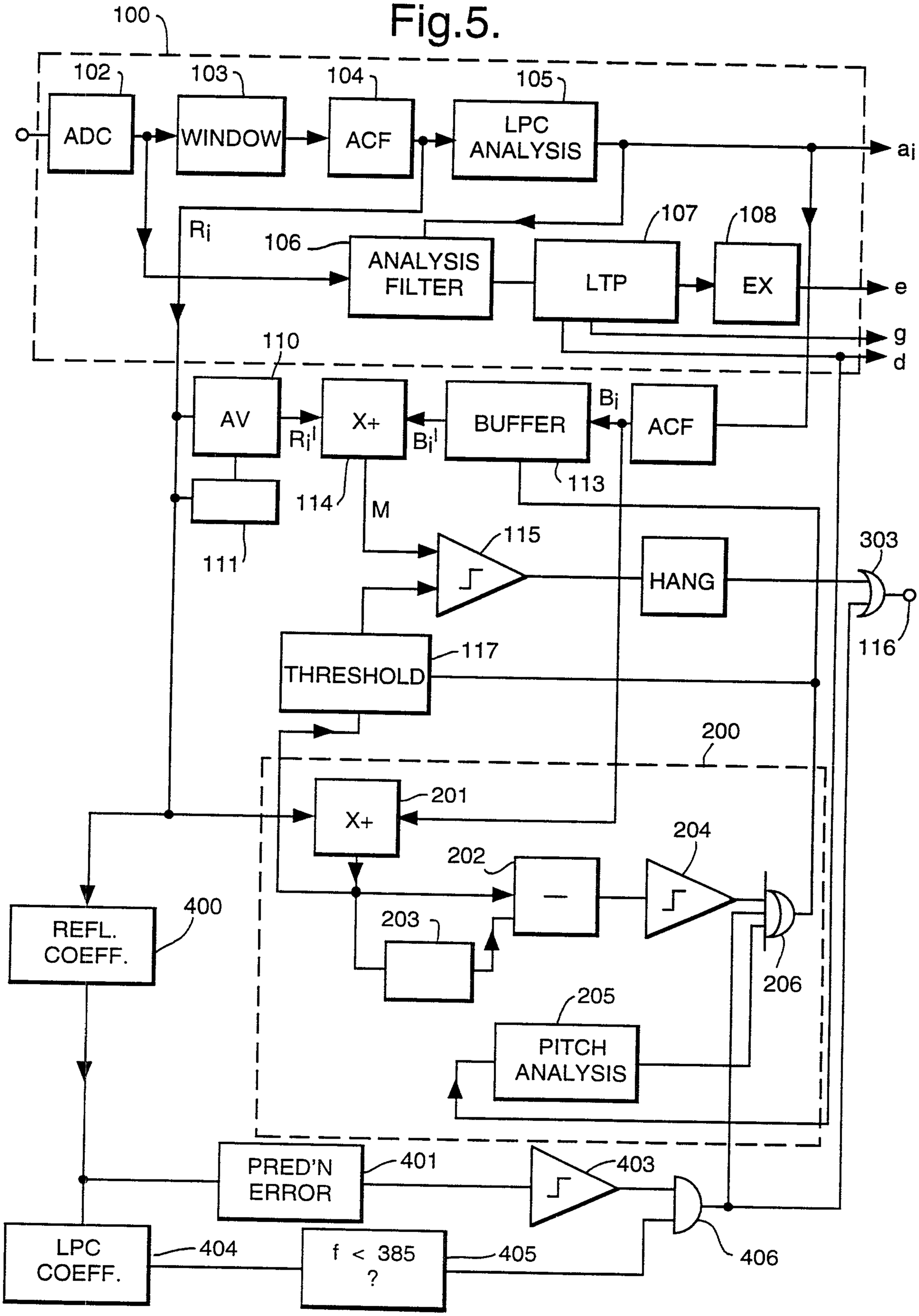
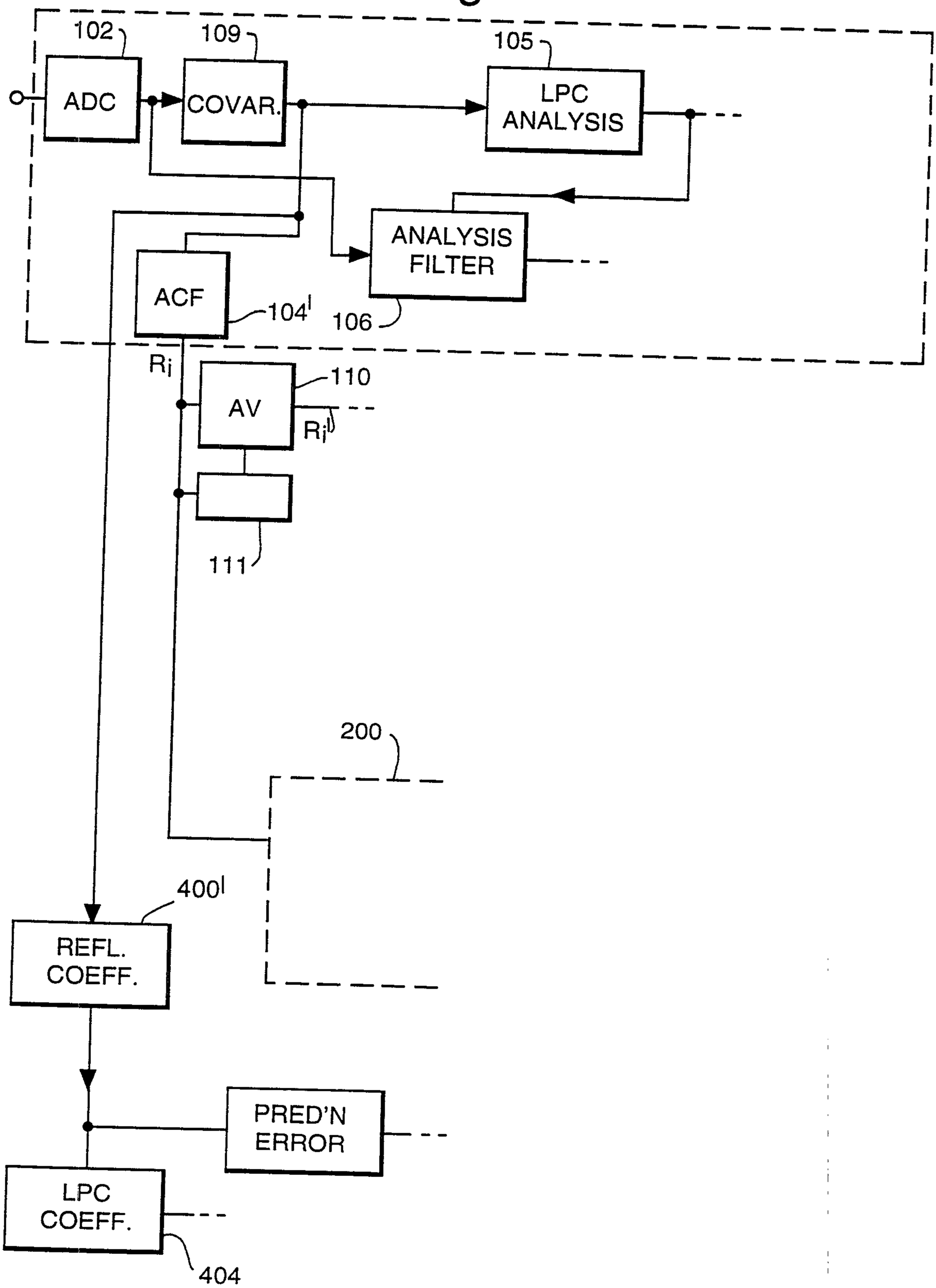


Fig.6.



2169745

Fig.7.

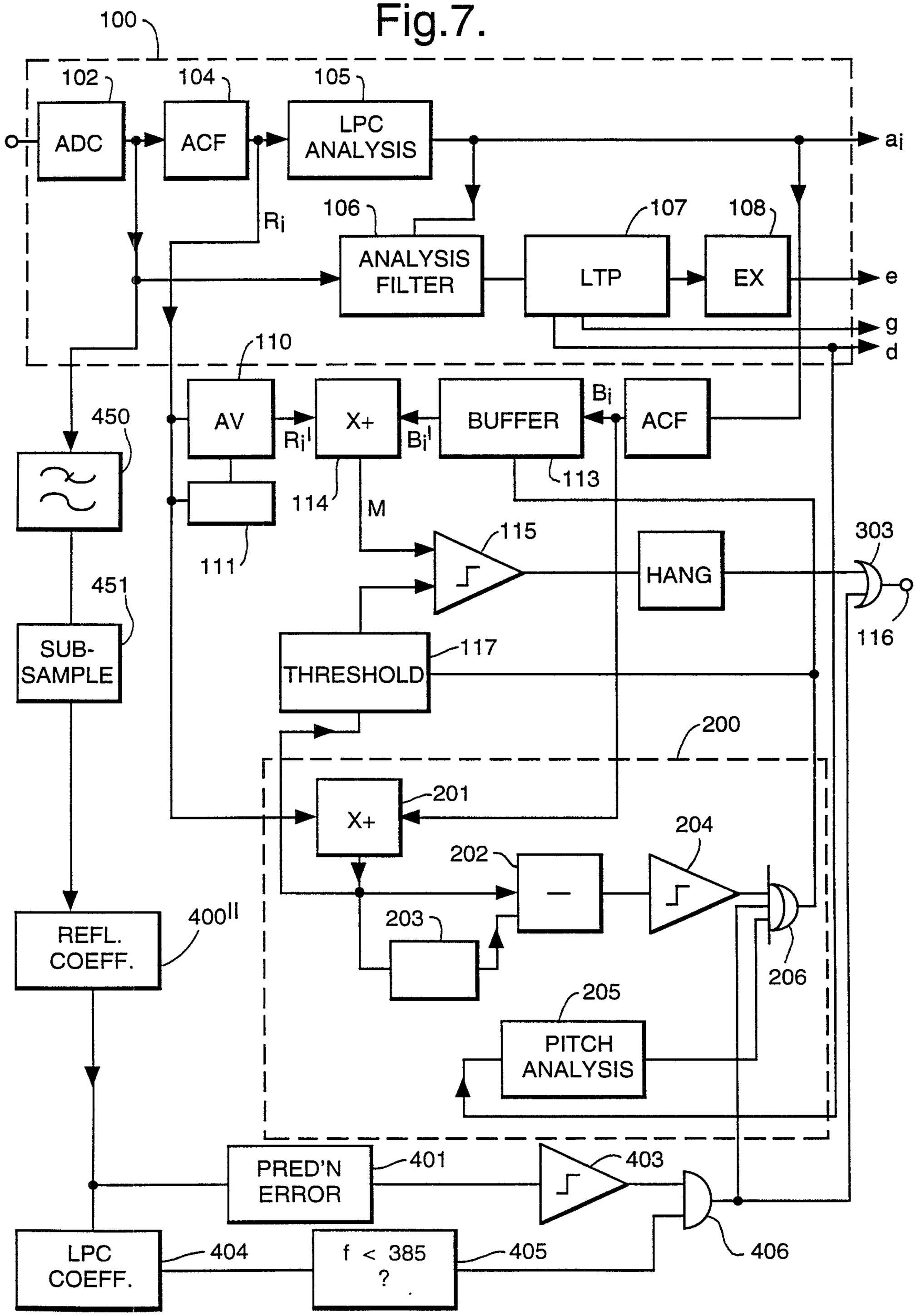


Fig.8.