Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2399706 C 2006/01/24

(11)(21) 2 399 706

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2001/02/12
(87) Date publication PCT/PCT Publication Date: 2001/08/16
(45) Date de délivrance/Issue Date: 2006/01/24
(85) Entrée phase nationale/National Entry: 2002/08/09
(86) N° demande PCT/PCT Application No.: US 2001/004526
(87) N° publication PCT/PCT Publication No.: 2001/059766
(30) Priorité/Priority: 2000/02/11 (60/181,734) US

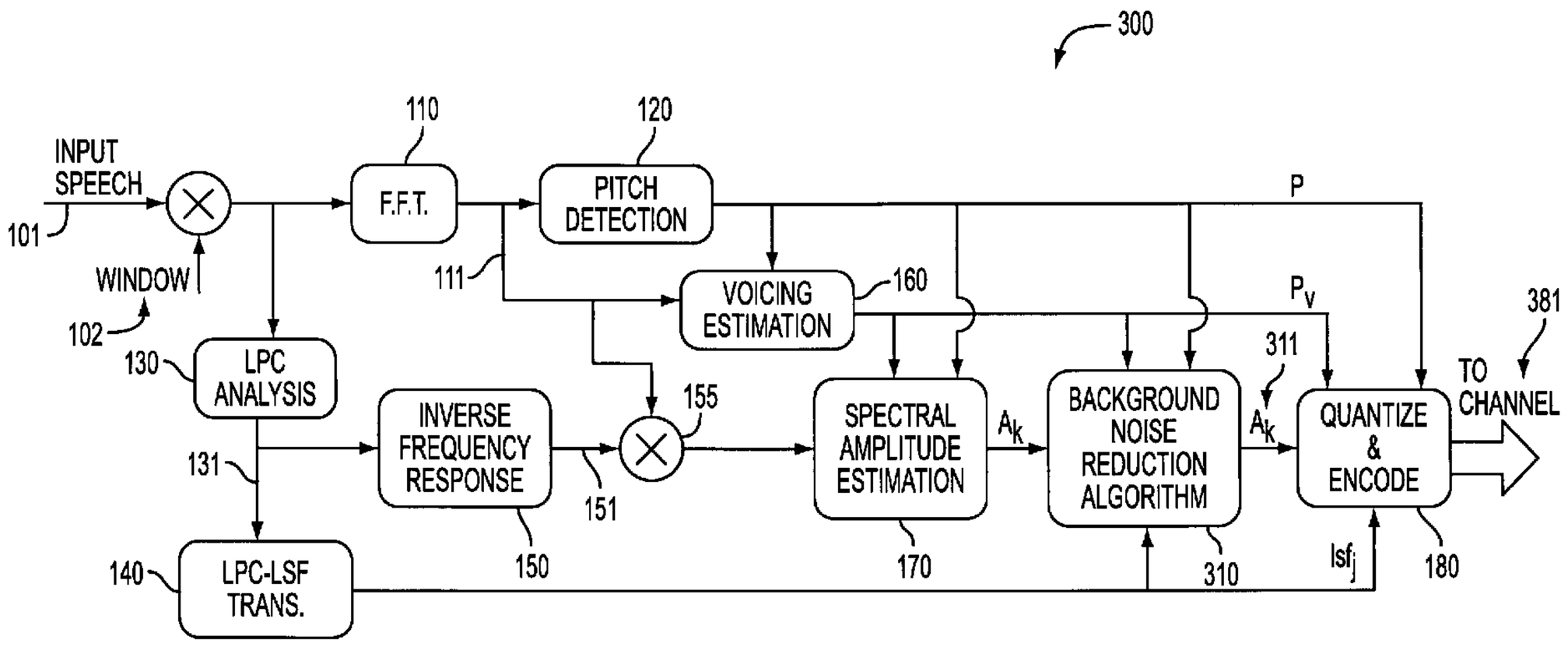
(51) Cl.Int.^{7}/Int.Cl.^{7} G10L 21/02

(72) Inventeur/Inventor:
YELDENER, SUAT, US

(73) Propriétaire/Owner:
COMSAT CORPORATION, US

(74) Agent: GOWLING LAFLEUR HENDERSON LLP

(54) Titre : REDUCTION DU BRUIT DE FOND DANS DES SYSTEMES DE CODAGE VOCAL SINUSOIDAUX
(54) Title: BACKGROUND NOISE REDUCTION IN SINUSOIDAL BASED SPEECH CODING SYSTEMS



(57) Abrégé/Abstract:
A method and apparatus to reduce background noise in speech signals in order to improve the quality and intelligibility of processed speech. In mobile communications environment, speech signals are degraded by additive random noise. A randomness of the noise, which is often described in terms of its first and second order statistics, make it difficult to remove much of the noise without introducing background artifacts. This is particularly true for lower signal to background noise ratios. The method and apparatus provides noise reduction without any knowledge of the signal to background noise ratio.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 August 2001 (16.08.2001)

PCT

(10) International Publication Number
WO 01/59766 A1

(51) International Patent Classification[7]: G10L 21/02

(21) International Application Number: PCT/US01/04526

(22) International Filing Date: 12 February 2001 (12.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/181,734 11 February 2000 (11.02.2000) US

(71) Applicant *(for all designated States except US)*: **COMSAT CORPORATION** [US/US]; 6560 Rock Spring Drive, Bethesda, MD 20817 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **YELDENER, Suat** [CY/US]; 19606 Crystal Rock Drive, #14, Germantown, MD 20874 (US).

(74) Agents: **KASPER, Alan, J.** et al.; Sughrue, Mion, Zinn, Macpeak & Seas, PLLC, 2100 Pennsylvania Ave., N.W., Suite 800, Washington, DC 20037-3213 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/59766 A1

(54) Title: BACKGROUND NOISE REDUCTION IN SINUSOIDAL BASED SPEECH CODING SYSTEMS

(57) Abstract: A method and apparatus to reduce background noise in speech signals in order to improve the quality and intelligibility of processed speech. In mobile communications environment, speech signals are degraded by additive random noise. A randomness of the noise, which is often described in terms of its first and second order statistics, make it difficult to remove much of the noise without introducing background artifacts. This is particularly true for lower signal to background noise ratios. The method and apparatus provides noise reduction without any knowledge of the signal to background noise ratio.

BACKGROUND NOISE REDUCTION IN SINUSOIDAL BASED SPEECH CODING SYSTEMS

Background of the Invention

Speech enhancement involves processing either degraded speech signals or clean speech that is expected to be degraded in the future, where the goal of processing is to improve the quality and intelligibility of speech for the human listener. Though it is possible to enhance speech that is not degraded, such as by high pass filtering to increase perceived crispness and clarity, some of the most significant contributions that can be made by speech enhancement techniques is in reducing noise degradation of the signal. The applications of speech enhancement are numerous. Examples include correction for room reverberation effects, reduction of noise in speech to improve vocoder performance and improvement of un-degraded speech for people with impaired hearing. The degradation can be as different as room echoes, additive random noise, multiplicative or convolutional noise, and competing speakers. Approaches differ, depending on the context of the problem. One significant problem is that of speech degraded by additive random noise, particularly in the context of a Harmonic Excitation Linear Predictive Speech Coder (HE-LPC).

The selection of an error criteria by which speech enhancement systems are optimized and compared is of central importance, but there is no absolute best set of criteria. Ultimately, the selected criteria must relate to the subjective evaluation by a human listener, and should take into account traits of auditory perception. An example of a system that exploits certain perceptual aspects of speech is that developed by Drucker, as described in "Speech Processing in a High Ambient Noise Environment", IEEE Trans. On AudioElecrtoacoustics, Vol.: Au-16, pp: 165-168, June 1968. Based on experimental findings, Drucker concluded that a primary cause for intelligibility loss in speech degraded by wide-band noise is confusion between fricatives and plosive sounds, which is partially due to a loss of short pauses immediately before the plosive sounds. Drucker reports a significant improvement in intelligibility after high pass filtering the /s/ fricative and inserting short pauses before the plosive sounds. However, Drucker's assumption that the plosive sounds can be accurately determined limits the usefulness of the system.

Many speech enhancement techniques take a more mathematical approach, which are empirically matched to human perception. An example of a mathematical criterion that is useful in matching short time spectral magnitudes, a perceptually important characterization of speech, is the mean squared error (MSE). A computational advantage to using this criteria is that the minimum MSE reduces to a linear set of equations. Other factors, however, can make an "optimally small" MSE misleading. In the case of speech degraded by narrow-band noise, which is considerably less comfortable to listen to than wide-band noise, wide-band noise can be added to mask the more unpleasant narrow-band noise. This technique makes the mean squared error larger.

The enhancement of speech degraded by additive noise has led to diverse approaches and systems. Some systems, like Drucker's, exploit certain perceptual aspects of speech. Others have focused on improving the estimate of the short time Fourier transform magnitude (STFTM), which is perceptually important in characterizing speech. The phase, on the other hand, may be considered as relatively unimportant.

Because the STFTM of speech is perceptually very important, one approach has been to estimate the STFTM of clean speech, given information about the noise source. Two classes of techniques have evolved out of this approach. In the first, the short time spectral amplitude is estimated from the spectrum of degraded speech and information about the noise source. Usually, the processed spectrum adopts the phase of the spectrum of the noisy speech because phase information is not as important perceptually. This first class includes spectral subtraction, correlation subtraction and maximum likelihood estimation techniques. The second class of techniques, which includes Wiener filtering, uses the degraded speech and noise information to create a zero-phase filter that is then applied to the noisy speech. As reported by H. L. Van Trees in "Detection, Estimation and Modulation Theory", Pt. 1, John Wiley and Sons, New York, N.Y. 1968, with Wiener filtering the goal is to develop a filter which can be applied to noisy speech to form the enhanced speech.

Turning first to the class concerned with estimation of short time spectral amplitude, particularly where spectral subtraction is used, statistical information is obtained about the noise source to estimate the STFTM of clean speech. This

technique is also known as power spectrum subtraction. Variations of these techniques included the more general relation identified by Lim et al in "Enhancement and Bandwidth Compression of Noisy Speech", Proc. of the IEEE, Vol.: 67, No.: 12, December 1979, as:

$$|\hat{S}(\omega)|^{\alpha} = |Y(\omega)|^{\alpha} - \beta E[|N(\omega)|^{\alpha}] \quad (1)$$

where α and β are parameters that can be chosen. Magnitude spectral subtraction is the case where $\alpha = 1$, and $\beta = 1$. A different subtractive speech enhancement algorithm was presented by McAulay and Malpass in "Speech Enhancement Using Soft Decision Noise Suppression Filter", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.: ASSP-28, No.: 2, pp: 137-145, April 1980. Their method uses a maximum-likelihood estimate of the noisy speech signal assuming that the noise is gaussian. When the enhanced magnitude yields a value smaller than an attenuation threshold, however, the spectral magnitude is automatically set to the defined threshold.

Spectral subtraction is generally considered to be effective at reducing the apparent noise power in degraded speech. Lim has shown however that this noise reduction is achieved at the price of lower speech intelligibility (8). Moderate amounts of noise reduction can be achieved without significant intelligibility loss, however, large amount of noise reduction can seriously degrade the intelligibility of the speech. Other researchers have also drawn attention to other distortions which are introduced by spectral subtraction (5). Moderate to high amounts of spectral subtraction often introduce "tonal noise" into the speech.

Another class of speech enhancement methods exploits the periodicity of voiced speech to reduce the amount of background noise. These methods average the speech over successive pitch periods, which is equivalent to passing the speech through an adaptive comb filter. In these techniques, harmonic frequencies are passed by the filter while other frequencies are attenuated. This leads to a reduction in the noise between the harmonics of voiced speech. One problem with this technique is that it severely distorts any unvoiced spectral regions. Typically this problem is

handled by classifying each segment as either voiced or unvoiced and then only applying the comb filter to voiced regions. Unfortunately, this approach does not account for the fact that even at modest noise levels many voiced segments have large frequency regions which are dominated by noise. Comb filtering these noise dominated frequency regions severely changes the perceived characteristics of the noise.

These known problems with current speech enhancement methods have generated considerable interest in developing new or improved speech enhancement methods which are capable of reducing the substantial amount of noise without adding noticeable artifacts into the speech signal. A particular application for such technique is the Harmonic Excitation Linear Predictive Coding (HE-LPC), although it is desirable for such technique to be applicable to any sinusoidal based speech coding algorithm.

The conventional Harmonic Excitation Linear Predictive Coder (HE-LPC) is disclosed in disclosed in S. Yeldener " A 4 kb/s Toll Quality Harmonic Excitation Linear Predictive Speech Coder", Proc. of ICASSP-1999, Phoenix, Arizona, pp: 481-484, March 1999. A simplified block diagram of the conventional HE-LPC coder is shown in Figure 1. In the illustrated HE-LPC speech coder 100, the basic approach for representation of speech signals is to use a speech synthesis model where speech is formed as the result of passing an excitation signal through a linear time varying LPC filter that models the characteristics of the speech spectrum. In particular, input speech 101 is applied to a mixer 105 along with a signal defining a window 102. The mixer output 106 is applied to a fast Fourier transform FFT 110, which produces an output 111, and an LPC analysis circuit 130, which itself produces an output 131 to an LPC-LSF transform circuit 140. The LPC-LSF transform circuit 140 combines to act as a linear time-varying LPC filter that models the resonant characteristics of the speech spectral envelope. The LPC filter is represented by a plurality of LPC coefficients (14 in a preferred embodiment) that are quantized in the form of Line Spectral Frequency (LSF) parameters. The output 131 of the LPC analysis is provided to an inverse frequency response unit 150, whose output 151 is applied to mixer 155 along with the

output 111 of the FFT circuit 110. The same output 111 is applied to a pitch detection circuit 120 and a voicing estimation circuit 160.

In the HE-LPC speech coder, the pitch detection circuit 120 uses a pitch estimation algorithm that takes advantage of the most important frequency components to synthesize speech and then estimate the pitch based on a mean squared error approach. The pitch search range is first partitioned into various sub-ranges, and then a computationally simple pitch cost function is computed. The computed pitch cost function is then evaluated and a pitch candidate for each sub-range is obtained. After pitch candidates are selected, an analysis by synthesis error minimization procedure is applied to choose the most optimal pitch estimate. In this case, the LPC residual signal is low pass filtered first and then the low pass filter excitation signal is passed through an LPC synthesis filter to obtain the reference speech signal. For each candidate of pitch, the LPC residual spectrum is sampled at the harmonics of the corresponding pitch candidate to get the harmonic amplitude and phases. These harmonic components are used to generated a synthetic excitation signal based on the assumption that the speech is purely voiced. This synthetic excitation signal is then passed through the LPC synthesis filter to obtain the synthesized speech signal. The perceptually weighted mean squared error (PWMSE) in between the reference and synthesized signal is then computed and repeated for each candidate of pitch. The candidate pitch period having the least PWMSE is then chosen as the most optimal pitch estimate P.

Also significant to the operation of the HE-LPC is the computation of the voicing probability that defines a cut-off frequency in voicing estimation circuit 160. First, a synthetic speech spectrum is computed based on the assumption that speech signal is fully voiced. The original and synthetic speech signals are then compared and a voicing probability is computed on a harmonic-by-harmonic basis, and the speech spectrum is assigned as either voiced or unvoiced, depending on the magnitude of the error between the original and reconstructed spectra for the corresponding harmonic. The computed voicing probability Pv is then applied to a spectral amplitude estimation circuit 170 for an estimation of spectral amplitude A_k for the k^{th} harmonic. A quantize and encoder unit 180 receives the pitch detection signal P, the noise residual in the amplitude, the voicing probability Pv and the

spectral amplitude A_k, along with the output lsf_j of the LPC-LCF transform 140 to generate an encoded output speech signal for application to the output channel 181.

In other coders to which the invention would apply, the excitation signal would also be specified by a consideration of the fundamental frequency, spectral amplitudes of the excitation spectrum and the voicing information.

At the decoder 200, as illustrated in Fig. 2, the transmitted signal is deconstructed into its components lsf_j, P and Pv. Specifically, signal 201 from the channel is input to a decoder 210, which generates a signal lsf_j for input to a LSF-LPC transform circuit 220, a pitch estimate P for input to voiced speech synthesis circuit 240 and a voicing probability P_v, which is applied to voicing control circuit 250. The voicing control circuit provides signals to synthesis circuits 240 and 260 via inputs 251 and 252. The two synthesis circuits 240 and 260 also receive the output 231of an amplitude enhancing circuit 230, which receives an amplitude signal A_k from the decoder 210 at its input.

The voiced part of the excitation signal is determined as the sum of the sinusoidal harmonics. The unvoiced part of the excitation signal is generated by weighting the random noise spectrum with the original excitation spectrum for the frequency regions determined as unvoiced. The voiced and unvoiced excitation signals are then added together at mixer 270 and passed through an LPC synthesis filter 280, which responds to an input from the LPC-LSF transform 220 to form the final synthesized speech. At the output, a post-filter 290, which also receives an input from the LSF-LPC transform circuit 220 via an amplifier 225 with a constant gain α is used to further enhance the output speech quality. This arrangement produces high quality speech.

However, the conventional arrangement of HE-LPC encoder and decoder does not provide the desired performance for a variety of input signal and background noise conditions. Accordingly, there is a need for a further way to improve speech quality significantly in background noise conditions.

Summary of the Invention

The present invention comprises the reduction of background noise in a processed speech signal prior to quantization and encoding for transmission on an output channel.

More specifically, the present invention comprises the application of an algorithm to the spectral amplitude estimation signal generated in a speech codec on the basis of detected pitch and voicing information for reduction of background noise.

The present invention further concerns the application of a background noise algorithm on the basis of individual harmonics k in a spectral amplitude estimated signal A_k in a speech codec.

The present invention more specifically concerns the application of a background noise elimination algorithm to any sinusoidal based speech coding algorithm, and in particular, an algorithm based on harmonic excitation linear predictive encoding.

Brief Description of the Drawings

Figure 1 is a block diagram of a conventional HE-LPC speech encoder.

Figure 2 is a block diagram of a conventional HE-LPC speech decoder.

Figure 3 is a block diagram of a HE-LPC speech encoder in accordance with the present invention.

Figure 4 is a block diagram detailing an implementation of a preferred embodiment of the invention.

Figure 5 is a flow chart illustrating a method for achieving background noise reduction in accordance with the present invention.

Description of The Preferred Embodiment

The preferred embodiment of the present invention can be best appreciated by considering in Figure 3 the modifications that are made to the HE-LPC encoder that was illustrated in Figure 1. The same reference numbers from Figure 1 are used for those components in Figure 3 that are identical to those utilized in the basic block diagram of the conventional circuit illustrated in Figure 1. The operation of the components, as described therein, are identical. The notable addition in the improved HE-LPC encoder 300 circuit over the encoder 100 of Figure 1 is the background noise

reduction algorithm 310. The pitch signal P from the pitch detection circuit 120, the voicing probability signal Pv from the voicing estimation circuit 160, the spectral amplitude estimation signal A_k from the spectral amplitude estimation circuit 170 as well as the output of the LPC-LSF circuit 140 are all received by the background noise reduction algorithm 310. The output of that algorithm A_k (hat) 311 is input to the quantize and encode circuit 180, along with signals P, Pv and A_k for generation of the output signal 381 for transmission on the output channel. The processing of the signal A_k in order to reduce the effect of background noise provides a significantly improved and enhanced output onto the channel, which can then be received and processed in the conventional HE-LPC decoder of Figure 2, in a manner already described.

In considering the detailed operation of the background noise-compensating encoder of the present invention, reference is made to Figures 4 and 5, which illustrate the functional block diagram and flowchart of the algorithm that provides the enhanced performance. The algorithm processes the pitch P_0, as computed during the encoding process, and an auto-correlation function ACF, which is a function of the energy of the incoming speech as is well known in the art.

The first step S1 of the speech enhancement process is to have a voice activity detection (VAD) decision for each frame of speech signal. The VAD decision in block 410 is based on the periodicity P_0 and the auto-correlation function ACF of the speech signal, which appear as inputs on lines 401 and 405, respectively, of Fig. 4. The VAD decision is a 1 if a voice signal is over a given threshold (speech is present) and 0 if it is not over the threshold (speech is absent). If speech is present, there is noise gain control implemented in step S7 , as subsequently discussed.

If the VAD decision is that there is no speech, in step S2, the noise spectrum is updated every speech segment where speech is not active, and a long term noise spectrum is estimated in noise spectrum estimation unit 420. The long term average noise spectrum is formulated as (2):

$$|N_m(\omega)| = \begin{cases} \alpha|N_{m-1}(\omega)| + (1-\alpha)|U(\omega)|, & \text{if } VAD = 0; \\ |N_{m-1}(\omega)|, & \text{otherwise.} \end{cases}$$

where $0 \leq \omega \leq \pi$, $|N_m(\omega)|$ is the long term noise spectrum magnitude, α is a constant that is can be set to 0.95, and VAD = 0 means that speech is not active. In this formulation $|U(\omega)|$ can be formed by two ways. In the first way, $|U(\omega)|$ can be considered to be directly the current signal spectrum. In the second case, harmonic spectral amplitudes are first estimated according to equation (3) as:

$$A_k = \sqrt{\frac{1}{\omega_0} \sum_{\omega=(k-0.5)\omega_0}^{(k+0.5)\omega_0} |S(\omega)|^2} \qquad (3)$$

where A_k is the k^{th} harmonic spectral amplitude, and ω_0 is the fundamental frequency of the current signal, $|S(\omega)|$, which is an input to the noise spectrum estimation circuit 320 along with the pitch P_0 . Notably, $S(\omega)$ and P_0 are inputs to each of the VAD decision circuit 410, noise spectrum estimation unit 420, harmonic-by harmonic noise-signal ratio unit 430 and the harmonic noise attenuation factor unit 460, as subsequently discussed.

In step S3, the Estimated Noise to Signal Ratio (ENSR) for each harmonic lobe is calculated on the basis of S(w), excitation spectrum and ptich input.. In this case, the ENSR for the k^{th} harmonic is computed as:

$$\gamma_k = \frac{\sum_{\omega=B_L^k}^{B_U^k} [N_m(\omega)W_k(\omega)]^2}{\sum_{\omega=B_L^k}^{B_U^k} [S(\omega)W_k(\omega)]^2} \qquad (7)$$

where γ_k is the k^{th} ENSR, N_m (m}(ω) is the estimated noise spectrum, S(ω) is the speech spectrum and $W_k(\omega)$ is the window function computed as:

$$W_k(\omega) = 0.52 - (0.48 \cos\left(\frac{2\pi[\omega - B_L^k]}{[B_U^k - B_L^k]}\right) \quad ; \quad B_L^k \leq \omega < B_U^k. \qquad (8)$$

where $B^k{}_L$ and $B^k{}_U$ are the lower and upper limits for the k^{th} harmonic and computed as:

$$B_L^k = \left(k - \frac{1}{2}\right)\omega_0 \qquad (9)$$

$$B_U^k = \left(k + \frac{1}{2}\right)\omega_0 \qquad (10)$$

In step S4, long term average ACF is calculated section 440, using an ACF - autocorrelation function, and on the basis of an input of the VAD decision in section 410, an input is provided to noise reduction control circuit 450, which in step S5 is used to control the noise reduction gain, β_m, from one frame to the next one:

$$\beta_m = \begin{cases} \beta_{m-1} + \Delta, & \text{if } VAD = 1; \\ \beta_{m-1} - \Delta, & \text{otherwise.} \end{cases} \qquad (5)$$

where Δ is a constant (typically $\Delta = 0.1$) and

$$\beta_m = \begin{cases} 1.0, & \text{if } \beta_m > 1.0; \\ min, & \text{if } \beta_m < min; \end{cases} \qquad (6)$$

where *min* is the lowest noise attenuation factor (typically, *min* = 0.5).

In step S5, a harmonic-by-harmonic noise-signal ratio is calculated in section 430 and the harmonic spectral amplitudes are interpolated according to equation (4) to have a fixed dimension spectrum as:

$$U(\omega) = A_k + [A_{k+1} - A_k(i)] \frac{(\omega - k\omega_0)}{\omega_0} \quad ; \quad k\omega_0 \le \omega \le (k+1)\omega_0. \qquad (4)$$

where $1 \leq k \leq L$ and L is the total number of harmonics within the 4 kHz speech band. The noise gain control that is calculated in step S7, on the basis of the VAD decision output 1 and 0, and as represented in the block 450 of Fig. 4, is used as an input to the computation of the noise attenuation factor in step S5. Specifically, in step S5, the noise attenuation factor for each harmonic is calculated as:

$$\alpha_k = \beta_m \sqrt{(1.0 - \mu\gamma_k)} \qquad (11)$$

In this case, if $\alpha_k < 0.1$, then α_k is set to 0.1. Here, μ is a constant factor that can be set as:

$$\mu = \begin{cases} 4.0, & \text{if } E_m > 10000.0; \\ 3.0, & \text{if } E_m > 3700.0; \\ 2.5, & \text{otherwise.} \end{cases} \quad (12)$$

where E_m is the long term average energy that can be computed as:

$$E_m = \alpha E_{m-1} + (1.0 - \alpha) E_0 \quad (13)$$

where α is a constant factor (typically $\alpha = 0.95$) and E_0 is the average energy of the current frame of the speech signal.

The noise attenuation factor for each harmonic that was computed in step S5 is used in step S6 to scale the harmonic amplitudes that are computed during the encoding process of HE-LPC coder, and to attenuate noise in the residual spectral amplitudes A_k, and produce the modified spectral amplitudes A_k (hat).

The background noise reduction algorithm discussed above may be incorporated into the Harmonic Excitation Linear Predictive Coder (HE-LPC), or any other coder for a sinusoidal based speech coding algorithm.

The decoder as illustrated in Fig. 2, may be used to decode a signal encoded according to the principles of the present invention, as for decoding a signal processed by the conventional encoder, the voiced part of the excitation signal is determined as the sum of the sinusoidal harmonics. The unvoiced part of the excitation signal is generated by weighting the random noise spectrum with the original excitation spectrum for the frequency regions determined as unvoiced. The voiced and unvoiced excitation signals are then added together to form the final synthesized speech. At the output, a post-filter is used to further enhance the output speech quality.

While the present invention is described with respect to certain preferred embodiments, the invention is not limited thereto. The full scope of the invention is to be determined on the basis of the claims.

What is claimed is:

1. A speech codec comprising:

an input for receiving a speech signal having a speech spectrum with a plurality of harmonics defined by harmonic lobes, a periodicity and an auto-correlation function;

a linear time varying LPC filter that models the characteristics of the speech spectrum;

a pitch detection section for generating an estimate of optimal pitch in the received speech;

a voicing estimation section for computing a voicing probability that defines a cutoff frequency;

a spectral amplitude estimation section, responsive to the output of the pitch detection section and the voicing estimation section for generating an amplitude estimation for each of said harmonics; and

a background noise generation section responsive to the output of said pitch detection section and voicing estimation section for modifying the amplitude estimation for each of said harmonics from said spectral amplitude estimation section.

2. The speech codec as claimed in claim 1, wherein said background noise generation section comprises:

a voice activity detection section responsive to said periodicity and said auto-correlation function;

a noise spectrum estimation section, respective to the detection of voice activity and said pitch detection section for estimating the noise spectrum of said speech signal;

a section responsive to said estimated noise spectrum and said pitch detection section and being operative to calculate a harmonic by harmonic noise-signal ratio;

a noise reduction control section for generating a noise control signal in response to said auto-correlation function; and

a harmonic noise attenuation factor section, responsive to said pitch detection section, said noise reduction control section and said auto-correlation function for modifying said speech spectrum to provide a noise reduced output.

3. The speech codec as claimed in claim 2, wherein said noise spectrum estimation section is operative to generate a long term average noise spectrum as:

$$|N_m(\omega)| = \begin{cases} \alpha|N_{m-1}(\omega)| + (1-\alpha)|U(\omega)|; & \text{if } VAD = 0; \\ |N_{m-1}(\omega)|, & \text{otherwise.} \end{cases}$$

where $0 \leq \omega \leq \pi$, $|N_m(\omega)|$ is the long term noise spectrum magnitude, α is a constant that can be set to 0.95, and VAD=0 means that speech is not active.

4. The speech codec as claimed in claim 3, wherein $U(\omega)$ is one of the current signal spectrum and a harmonic spectral amplitude calculated as:

$$A_k = \sqrt{\frac{1}{\omega_0} \sum_{\omega=(k-0.5)\omega_0}^{(k+0.5)\omega_0} |S(\omega)|^2}$$

where A_k is the k^{th} harmonic spectral amplitude, and ω_0 is the fundamental frequency of the current signal, $|S(\omega)|$,

and interpolated to have a fixed dimension spectrum as:

$$U(\omega) = A_k + [A_{k+1} - A_k(i)] \frac{(\omega - k\omega_0)}{\omega_0} \quad ; \quad k\omega_0 \leq \omega \leq (k+1)\omega_0.$$

where $1 \leq k \leq L$ and L is the total number of harmonics within a speech band.

5. The speech codec as claimed in claim 2 wherein said voice activity detection section controls noise reduction gain frame by frame.

6. The speech codec as claimed in claim 2 wherein an attenuation factor for each harmonic is computed on the basis of estimated noise to signal ration (ENSR) for each harmonic lobe.

7. The speech codec as claimed in claim 6, wherein the ENSR for the kth harmonic is computed as:

$$\gamma_k = \frac{\sum_{\omega=B_L^k}^{B_U^k} [N_m(\omega)W_k(\omega)]^2}{\sum_{\omega=B_L^k}^{B_U^k} [S(\omega)W_k(\omega)]^2}$$

where γ_k is the k^{th} ENSR, $N_m(\omega)$ is the estimated noise spectrum, $S(\omega)$ is the speech spectrum and $W_k(\omega)$ is the window function computed as:

$$W_k(\omega) = 0.52 - (0.48 \cos\left(\frac{2\pi[\omega - B_L^k]}{[B_U^k - B_L^k]}\right) \quad ; \quad B_L^k \leq \omega < B_U^k.$$

where B_L^k and B_U^k are the lower and upper limits for the k^{th} harmonic and computed as:

$$B_L^k = \left(k - \frac{1}{2}\right)\omega_0$$
$$B_U^k = \left(k + \frac{1}{2}\right)\omega_0$$

where ω_0 is the fundamental frequency of the corresponding speech sequence.

8. The speech codec as claimed in claim 6, wherein the noise attenuation factor for each harmonic is used to scale computed harmonic amplitudes.

9. The speech codec as claimed in claim 2, further comprising a LPC filter that models the characteristics of the speech spectrum, said filter being represented by a plurality of line spectral frequency parameters.

10. A method of correcting for background noise in a speech codec comprising the steps of:

detecting voice activity for each frame of a speech signal, having a speech spectrum with a plurality of harmonics defined by harmonic lobes, a periodicity P_0 and an auto-correlation function ACF, based on the periodicity P_0 and the auto-correlation function ACF of the speech signal;

updating the noise spectrum every speech segment where speech is not active, and estimating a long term noise spectrum;

calculating a harmonic-by-harmonic noise-signal ratio and interpolating harmonic spectral amplitude;

calculating long term average ACF, and on the basis of an input of the detected voice activity, providing an input to control the noise reduction gain, β_m from one frame to the next one;

computing an attenuation factor for each harmonic based on Estimated Noise to Signal Ratio (ENSR) for each harmonic lobe;

calculating a noise attenuation factor for each harmonic; and

applying the noise attenuation factor to scale the harmonic amplitudes that are computed during the encoding process.

11. The method of claim 10 wherein the updating step is performed on the basis of an estimation of the spectral amplitudes as:

$$A_k = \sqrt{\frac{1}{\omega_0} \sum_{\omega=(k-0.5)\omega_0}^{(k+0.5)\omega_0} |S(\omega)|^2}$$

where A_k is the k^{th} harmonic spectral amplitude, and ω_0 is the fundamental frequency of the current signal $|S(\omega)|$.

12. The method of claim 11 wherein the harmonic spectral amplitudes are interpolated to have a fixed dimension spectrum.

13. The method of claim 10 wherein the harmonic spectral amplitudes are interpolated to have a fixed dimension spectrum.

14. The method of claim 13 wherein the fixed dimension spectrum is defined as:

$$U(\omega) = A_k + [A_{k+1} - A_k(i)] \frac{(\omega - k\omega_0)}{\omega_0} \quad ; \quad k\omega_0 \leq \omega \leq (k+1)\omega_0.$$

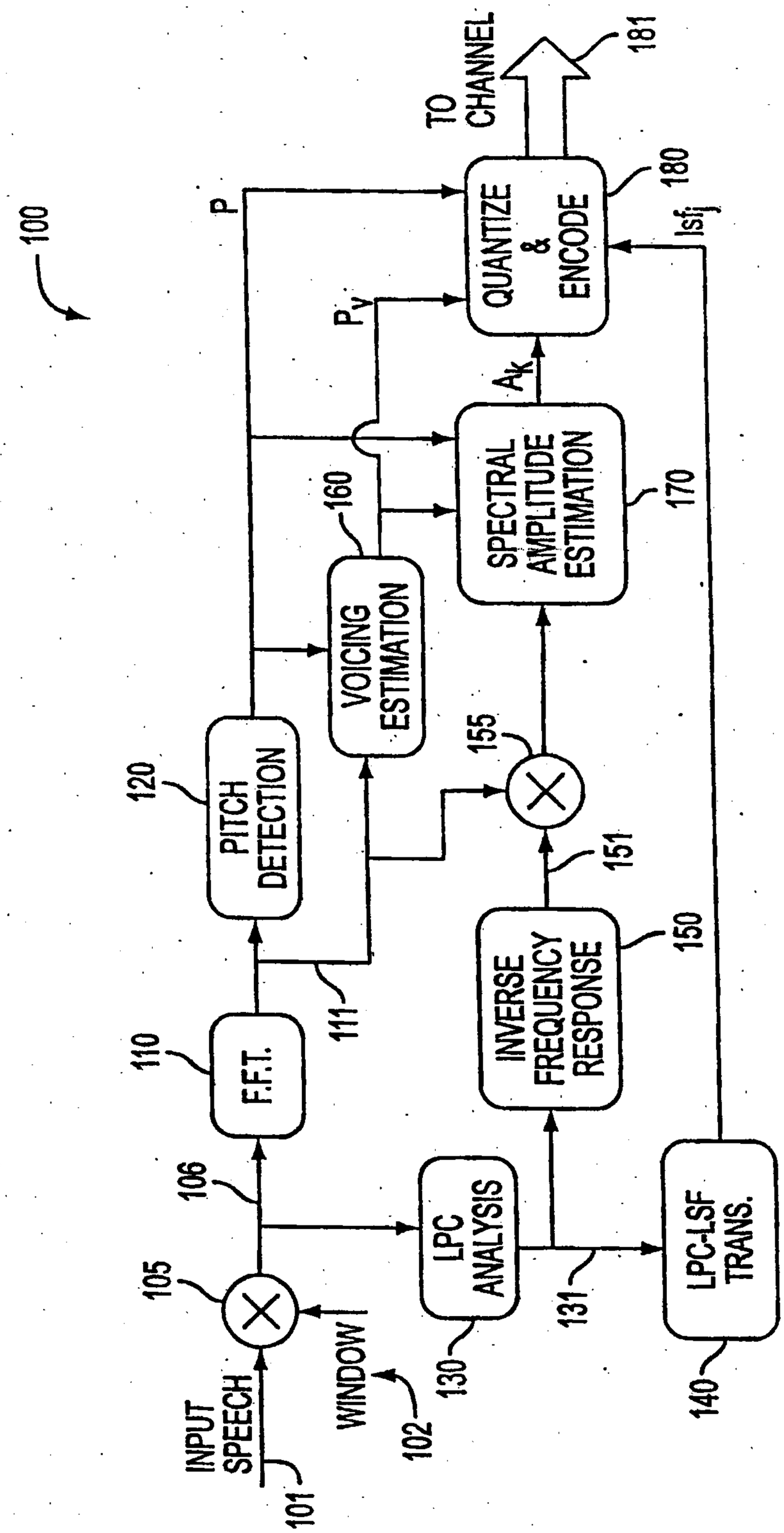
where A_k is the k^{th} harmonic spectral amplitude, ω_0 is the fundamental frequency of a current signal, $1 < k < L$, and L is the total number of harmonics within a speech band.

15. The method of claim 14 wherein the updating step is performed on the basis of $U(\omega)$ being the current signal spectrum.

16. The method of claim 12 wherein the fixed dimension spectrum is defined as:

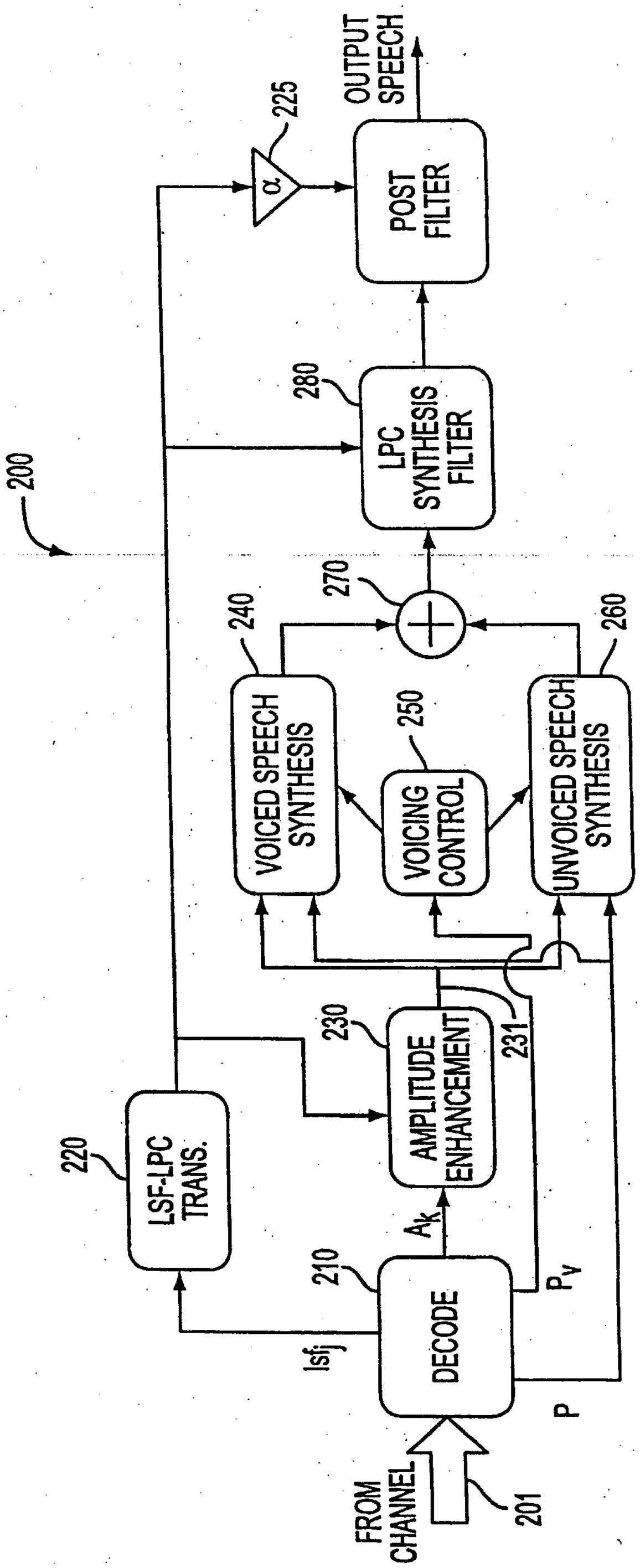
$$U(\omega) = A_k + [A_{k+1} - A_k(i)] \frac{(\omega - k\omega_0)}{\omega_0} \quad ; \quad k\omega_0 \leq \omega \leq (k+1)\omega_0.$$

where $1 < k < L$, and L is the total number of harmonics within a speech band.



Prior Art

FIG. 1



Prior Art

FIG. 2

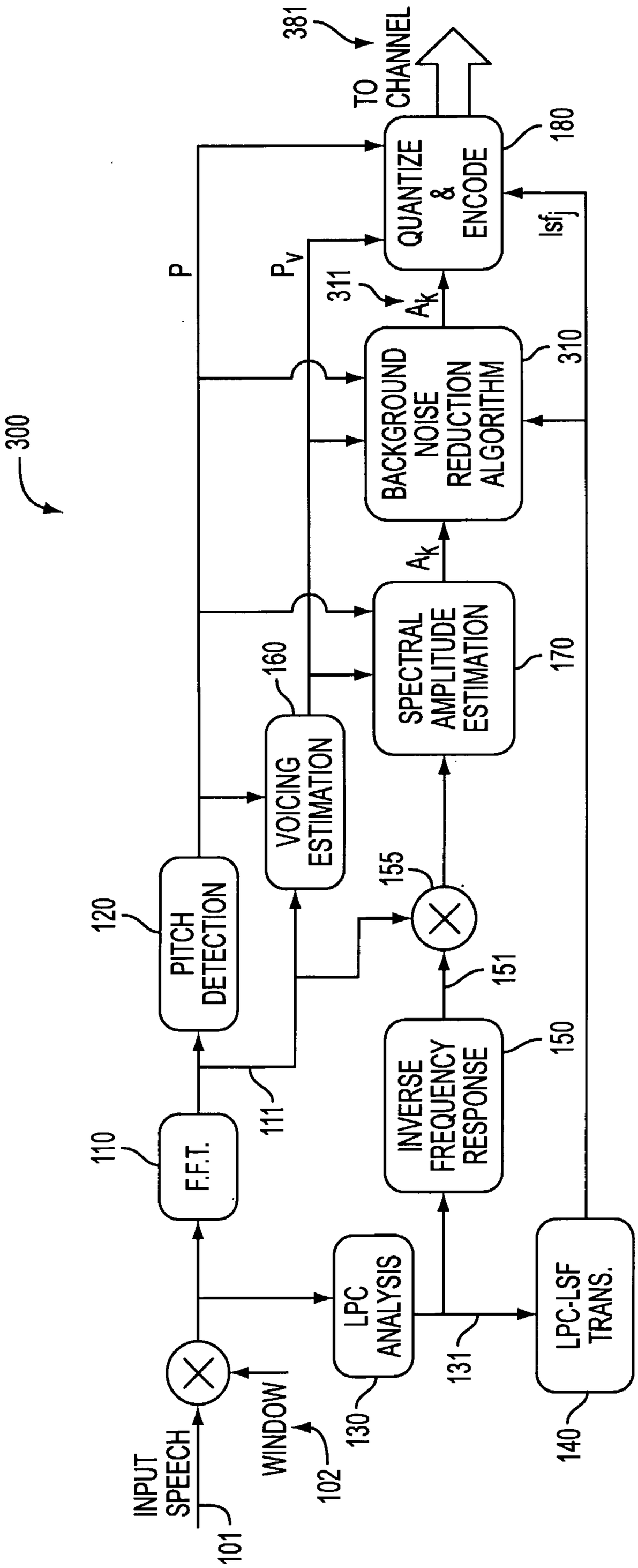

300
INPUT SPEECH
101
WINDOW
102
F.F.T.
110
111
PITCH DETECTION
120
LPC ANALYSIS
130
131
LPC-LSF TRANS.
140
INVERSE FREQUENCY RESPONSE
150
151
155
VOICING ESTIMATION
160
SPECTRAL AMPLITUDE ESTIMATION
170
BACKGROUND NOISE REDUCTION ALGORITHM
310
311
A_k
P_v
P
QUANTIZE & ENCODE
180
lsf_j
TO CHANNEL
381

FIG. 3

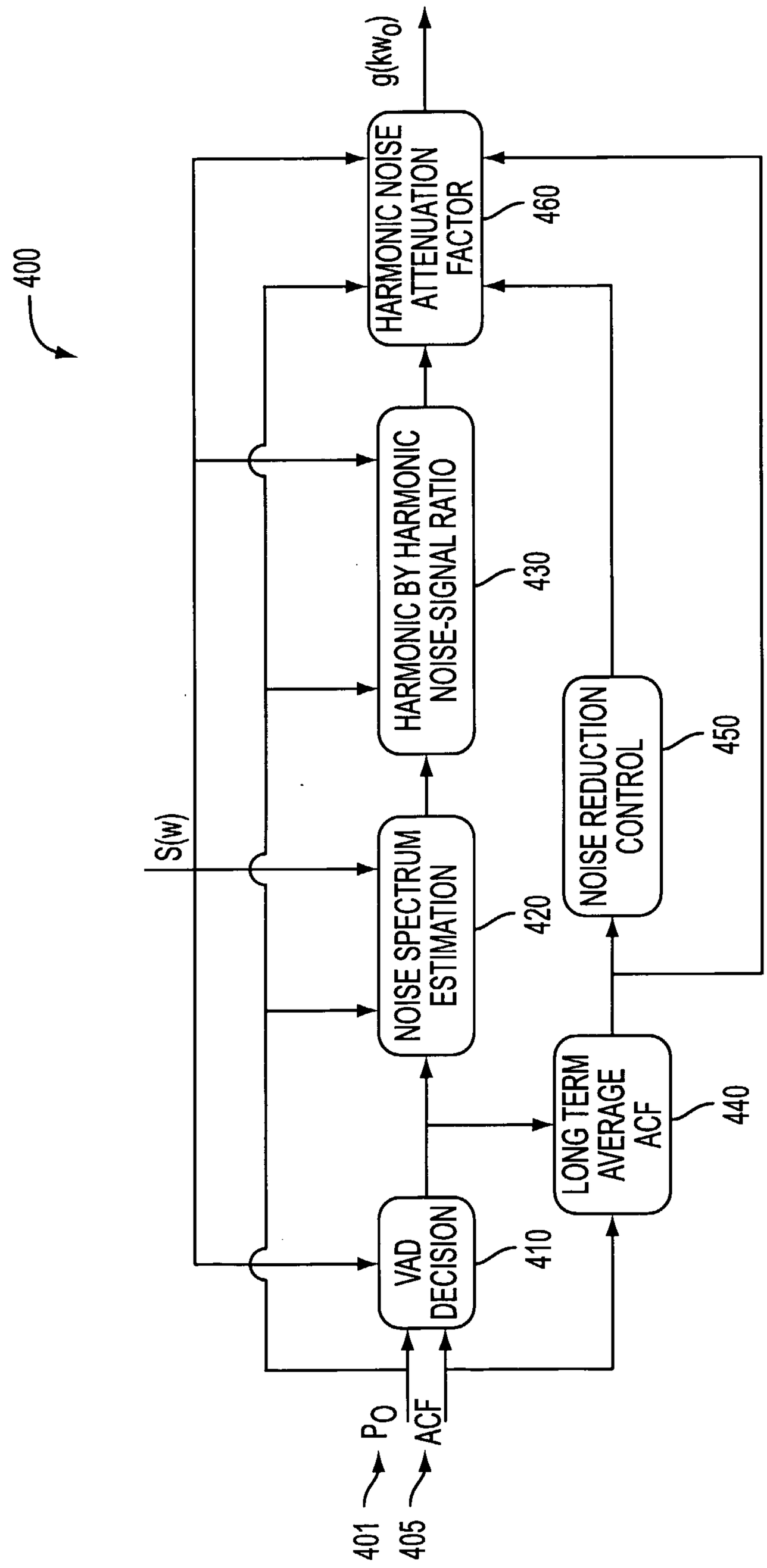

400
S(w)
VAD DECISION
410
NOISE SPECTRUM ESTIMATION
420
HARMONIC BY HARMONIC NOISE-SIGNAL RATIO
430
LONG TERM AVERAGE ACF
440
NOISE REDUCTION CONTROL
450
HARMONIC NOISE ATTENUATION FACTOR
460
g(kw0)
P0
ACF
401
405

FIG. 4

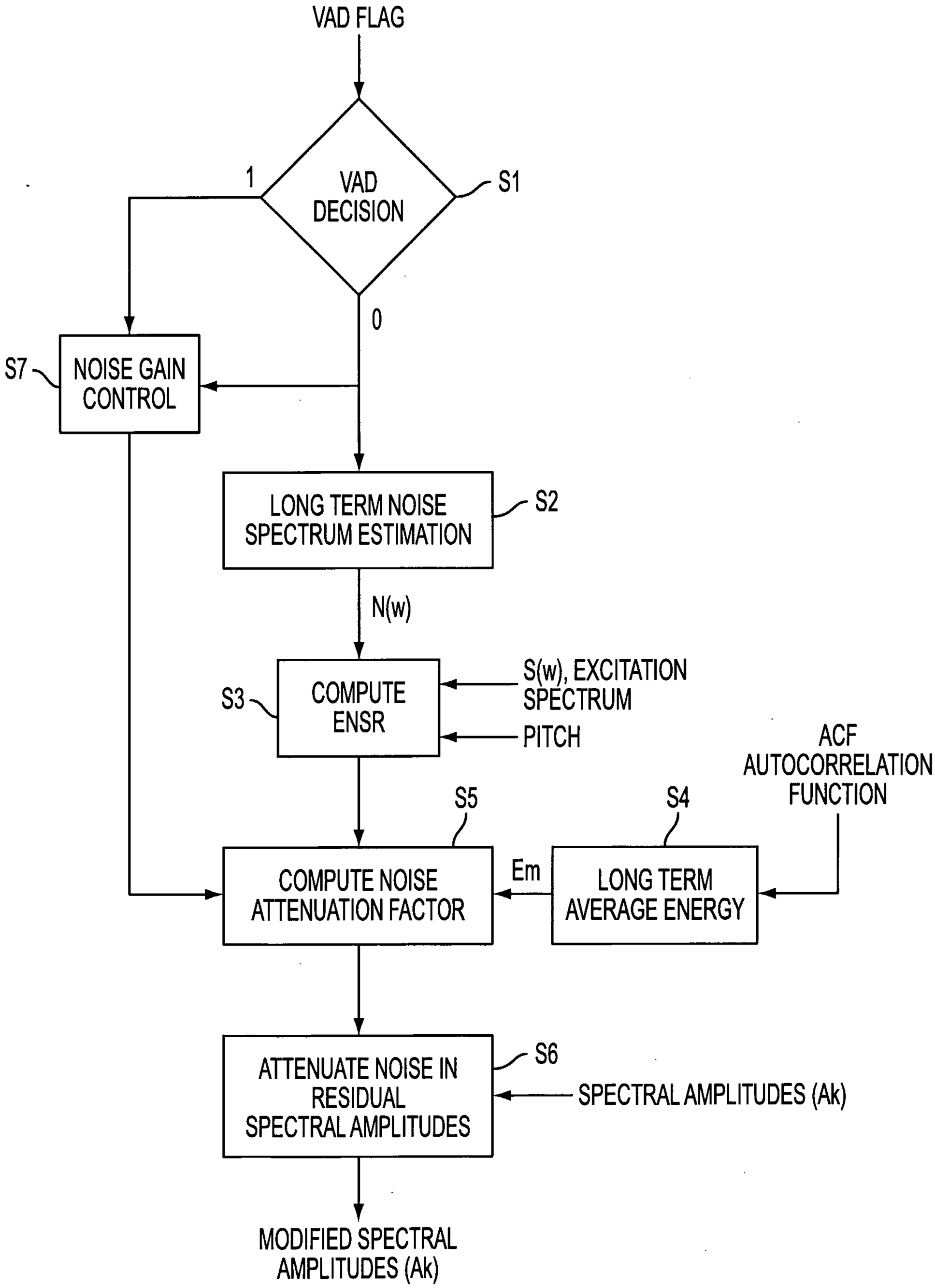
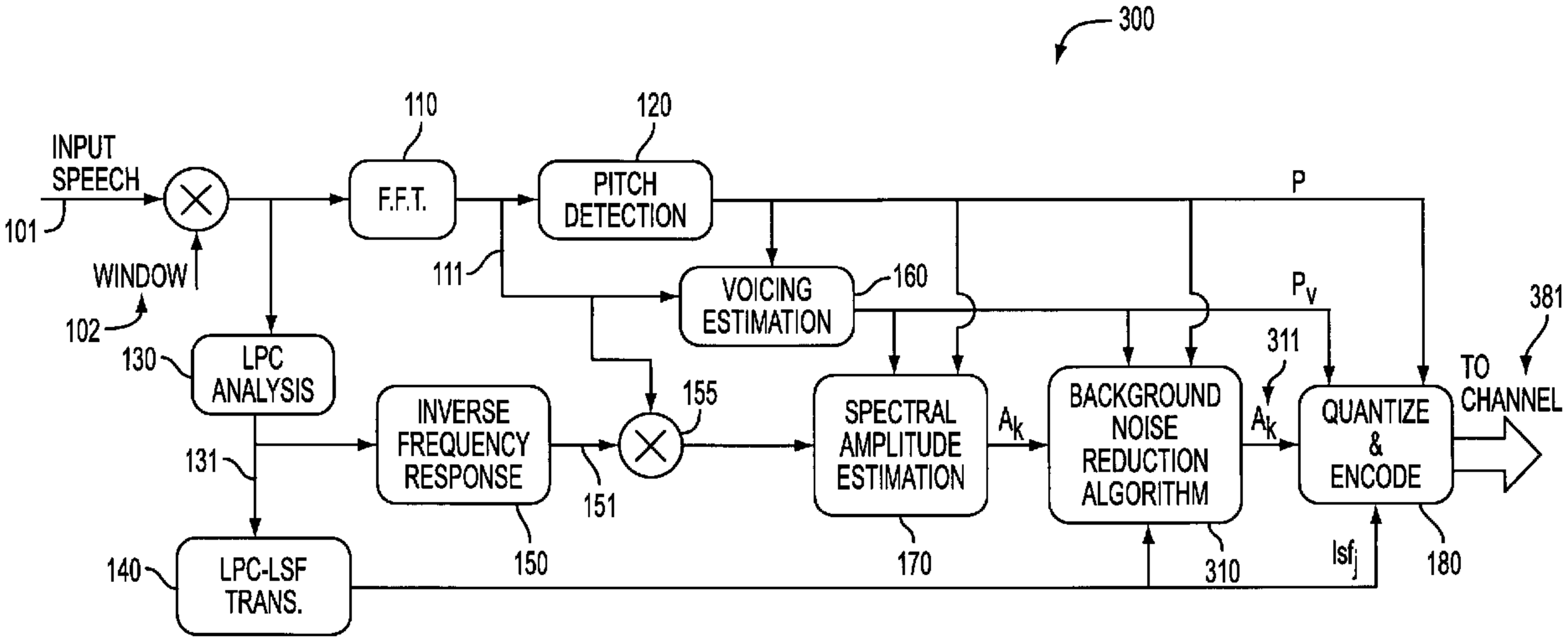

VAD FLAG
VAD DECISION
1
0
S1
S7
NOISE GAIN CONTROL
LONG TERM NOISE SPECTRUM ESTIMATION
S2
N(w)
COMPUTE ENSR
S3
S(w), EXCITATION SPECTRUM
PITCH
ACF AUTOCORRELATION FUNCTION
S5
S4
COMPUTE NOISE ATTENUATION FACTOR
Em
LONG TERM AVERAGE ENERGY
ATTENUATE NOISE IN RESIDUAL SPECTRAL AMPLITUDES
S6
SPECTRAL AMPLITUDES (Ak)
MODIFIED SPECTRAL AMPLITUDES (Ak)

FIG. 5


300
INPUT SPEECH
101
WINDOW
102
110
F.F.T.
120
PITCH DETECTION
P
111
VOICING ESTIMATION
160
P_V
381
TO CHANNEL
130
LPC ANALYSIS
311
155
INVERSE FREQUENCY RESPONSE
SPECTRAL AMPLITUDE ESTIMATION
A_k
BACKGROUND NOISE REDUCTION ALGORITHM
A_k
QUANTIZE & ENCODE
131
151
140
LPC-LSF TRANS.
150
170
310
lsf_j
180