

## Exhibit 8



US006385573B1

(12) **United States Patent**  
Gao et al.

(10) **Patent No.:** US 6,385,573 B1  
(45) **Date of Patent:** May 7, 2002

(54) **ADAPTIVE TILT COMPENSATION FOR SYNTHESIZED SPEECH RESIDUAL**

(75) Inventors: **Yang Gao**, Mission Viejo; **Huan-Yu Su**, San Clemente, both of CA (US)

(73) Assignee: **Conexant Systems, Inc.**, Newport Beach, CA (US)

(\* ) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/156,826**

(22) Filed: **Sep. 18, 1998**

**Related U.S. Application Data**

(60) Provisional application No. 60/097,569, filed on Aug. 24, 1998.

(51) **Int. Cl.<sup>7</sup>** ..... **G10L 19/04**

(52) **U.S. Cl.** ..... **704/220; 704/219**

(58) **Field of Search** ..... **704/220, 219**

(56) **References Cited**

**U.S. PATENT DOCUMENTS**

4,969,192 A	*	11/1990	Chen et al.	704/222
5,491,771 A		2/1996	Gupta et al.	
5,615,298 A	*	3/1997	Chen	704/228
5,664,055 A	*	9/1997	Kroon	704/223
5,682,404 A	*	10/1997	Miller	375/222
5,778,335 A	*	7/1998	Ubale et al.	704/219
5,778,338 A	*	7/1998	Jacobs et al.	704/223
5,864,798 A	*	1/1999	Miseki et al.	704/225
5,884,010 A	*	3/1999	Chen et al.	704/228
5,995,539 A	*	11/1999	Miller	375/222
6,073,092 A	*	6/2000	Kwon	704/219

**FOREIGN PATENT DOCUMENTS**

EP 0421360 A2 2/1990

**OTHER PUBLICATIONS**

W. Bastiaan Kleijn and Peter Kroon, "The RCELP Speech-Coding Algorithm," vol. 5, No. 5, Sep.-Oct. 1994, pp. 339/573-47/581.

(List continued on next page.)

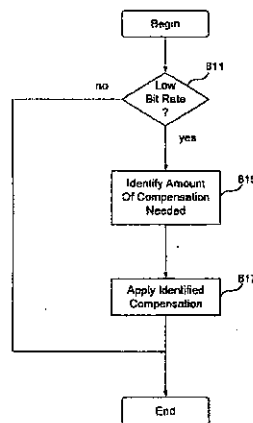
*Primary Examiner*—Fan Tsang  
*Assistant Examiner*—Michael N. Opsasnick  
(74) *Attorney, Agent, or Firm*—Farjami & Farjami LLP

(57) **ABSTRACT**

A multi-rate speech codec supports a plurality of encoding bit rate modes by adaptively selecting encoding bit rate modes to match communication channel restrictions. In higher bit rate encoding modes, an accurate representation of speech through CELP (code excited linear prediction) and other associated modeling parameters are generated for higher quality decoding and reproduction. To achieve high quality in lower bit rate encoding modes, the speech encoder departs from the strict waveform matching criteria of regular CELP coders and strives to identify significant perceptual features of the input signal. To support lower bit rate encoding modes, a variety of techniques are applied many of which involve the classification of the input signal. For each bit rate mode selected, pluralities of fixed or innovation subcodebooks are selected for use in generating innovation vectors. At lower encoding bit rates, a decoder utilizes adaptive compensation to attempt to correct for spectral variations in the weighted synthesized residual. Although many approaches are possible, a long asymmetric window is applied to the synthesized residual to generate a reflection coefficient that is smoothed, scaled and used in a first order filter. Because the content of the window varies over time, the coefficient and therefore the filter varies (or adapts) to remove at least a portion of the spectral tilt. As a result, the synthesized speech signal sounds brighter without having introduced significant coding noise.

**6 Claims, 10 Drawing Sheets**

Microfiche Appendix Included  
(1 Microfiche, 45 Pages)



## OTHER PUBLICATIONS

- C. Laflamme, J-P. Adoul, H.Y. Su, and S. Morissette, "On Reducing Computational Complexity of Codebook Search in CELP Coder Through the Use of Algebraic Codes," 1990, pp. 177-180.
- Chih-Chung Kuo, Fu-Rong Jean, and Hsiao-Chuan Wang, "Speech Classification Embedded in Adaptive Codebook Search for Low Bit-Rate CELP Coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, No. 1, Jan. 1995, pp. 1-5.
- Erdal Paksoy, Alan McCree, and Vish Viswanathan, "A Variable-Rate Multimodal Speech Coder with Gain-Matched Analysis-By-Synthesis," 1997, pp. 751-754.
- Gerhard Schroeder, "International Telecommunication Union Telecommunications Standardization Sector," Jun. 1995, pp. i-iv, 1-42.
- "Digital Cellular Telecommunications System; Comfort Noise Aspects for Enhanced Full Rate (EFR) Speech Traffic Channels (GSM 06.62)," May 1996, pp. 1-16.
- W. B. Kleijn and K.K. Paliwal (Editors), *Speech Coding and Synthesis*, Elsevier Science B.V.; Kroon and W.B. Kleijn (Authors), Chapter 3: "Linear-Prediction Based on Analysis-by-Synthesis Coding", 1995, pp. 81-113.
- W. B. Kleijn and K.K. Paliwal (Editors), *Speech Coding and Synthesis*, Elsevier Science B.V.; A. Das, E. Paskoy and A. Gersho (Authors), Chapter 7: "Multimode and Variable-Rate Coding of Speech," 1995, pp. 257-288.
- B.S. Atal, V. Cuperman, and A. Gersho (Editors), *Speech and Audio Coding for Wireless and Network Applications*, Kluwer Academic Publishers; T. Taniguchi, Y. Tanaka and Y. Ohta (Authors), Chapter 27: "Structured Stochastic Codebook and Codebook Adaptation for CELP," 1993, pp. 217-224.
- B.S. Atal, V. Cuperman, and A. Gersho (Editors), *Advances in Speech Coding*, Kluwer Academic Publishers; I. A. Gerson and M.A. Jasiuk (Authors), Chapter 7: "Vector Sum Excited Linear Prediction (VSELP)," 1991, pp. 69-79.
- B.S. Atal, V. Cuperman, and A. Gersho (Editors), *Advances in Speech Coding*, Kluwer Academic Publishers; J.P. Campbell, Jr., T.E. Tremain, and V.C. Welch (Authors), Chapter 12: "The DOD 4.8 KBPS Standard (Proposed Federal Standard 1016)," 1991, pp. 121-133.
- B.S. Atal, V. Cuperman, and A. Gersho (Editors), *Advances in Speech Coding*, Kluwer Academic Publishers; R.A. Salami (Author), Chapter 14: "Binary Pulse Excitation: A Novel Approach to Low Complexity CELP Coding," 1991, pp. 145-157.
- Patent Abstracts of Japan, vol. 1997, No. 11, Nov. 1997 (one page).
- Vien V. Nguyen, Vladimir Goncharoff, and John Damlakis, "Correcting Spectral Envelope Shifts in Linear Predictive Speech Compression Systems", *Proceedings of the Military Communications Conference (Milcom '90)*, vol. 1, 1990, pp. 354-358.
- Masaaki Honda, "Speech Coding Using Waveform Matching Based on LPC Residual Phase Equalization", *International Conference on Acoustics, Speech & Signal Processing (ICASSP '90)*, vol. 1, pp. 213-216.

\* cited by examiner

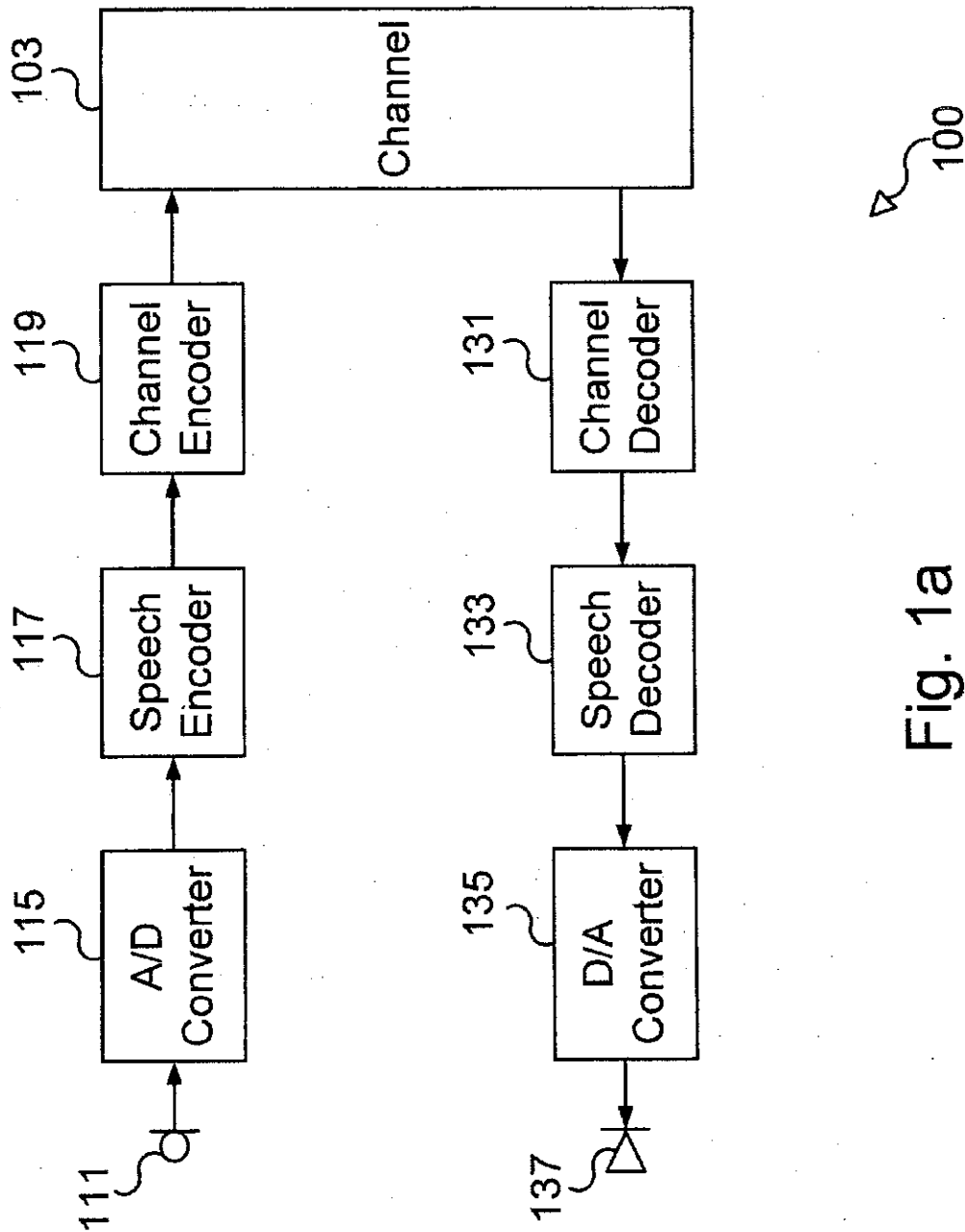


Fig. 1a

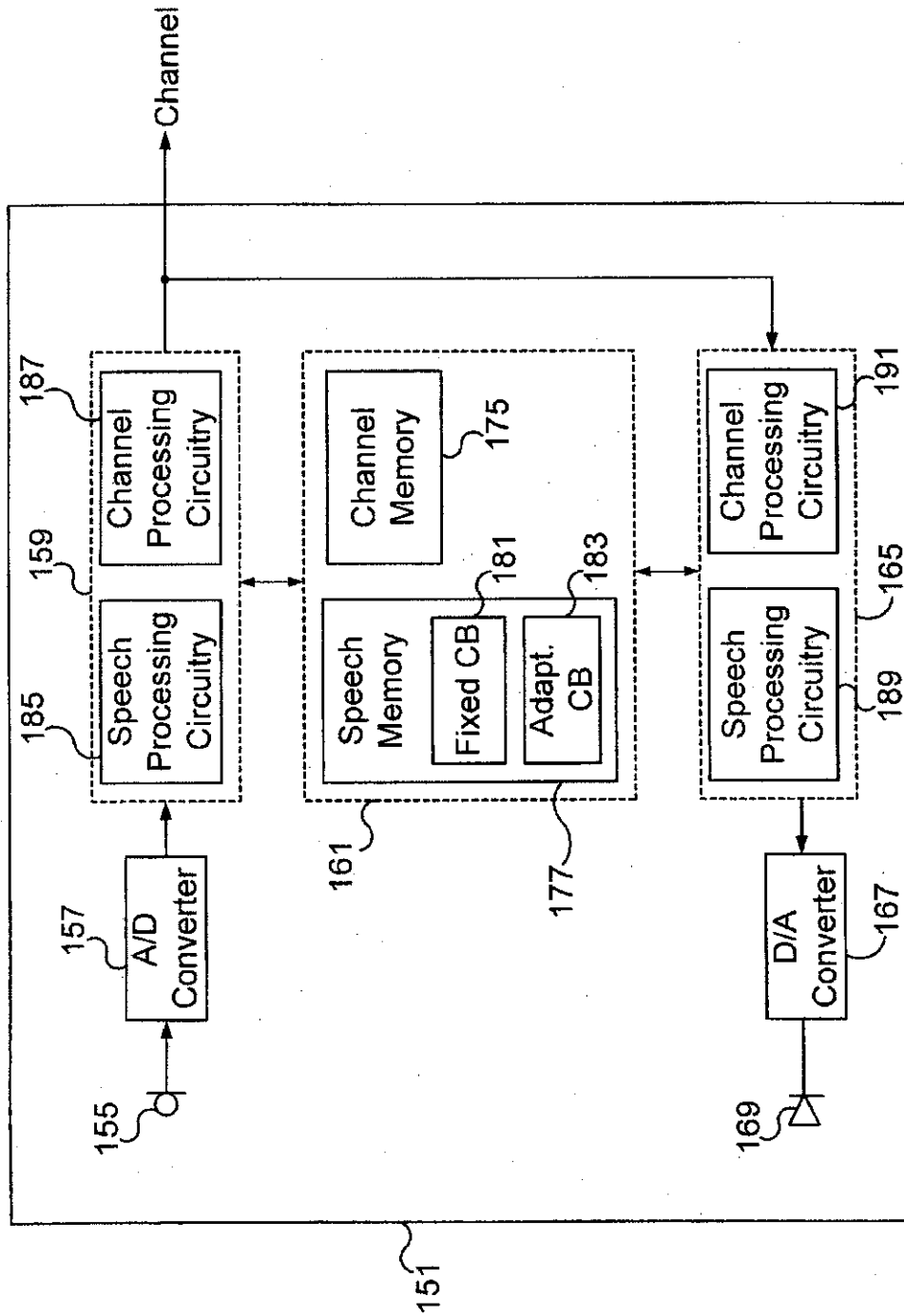


Fig. 1b

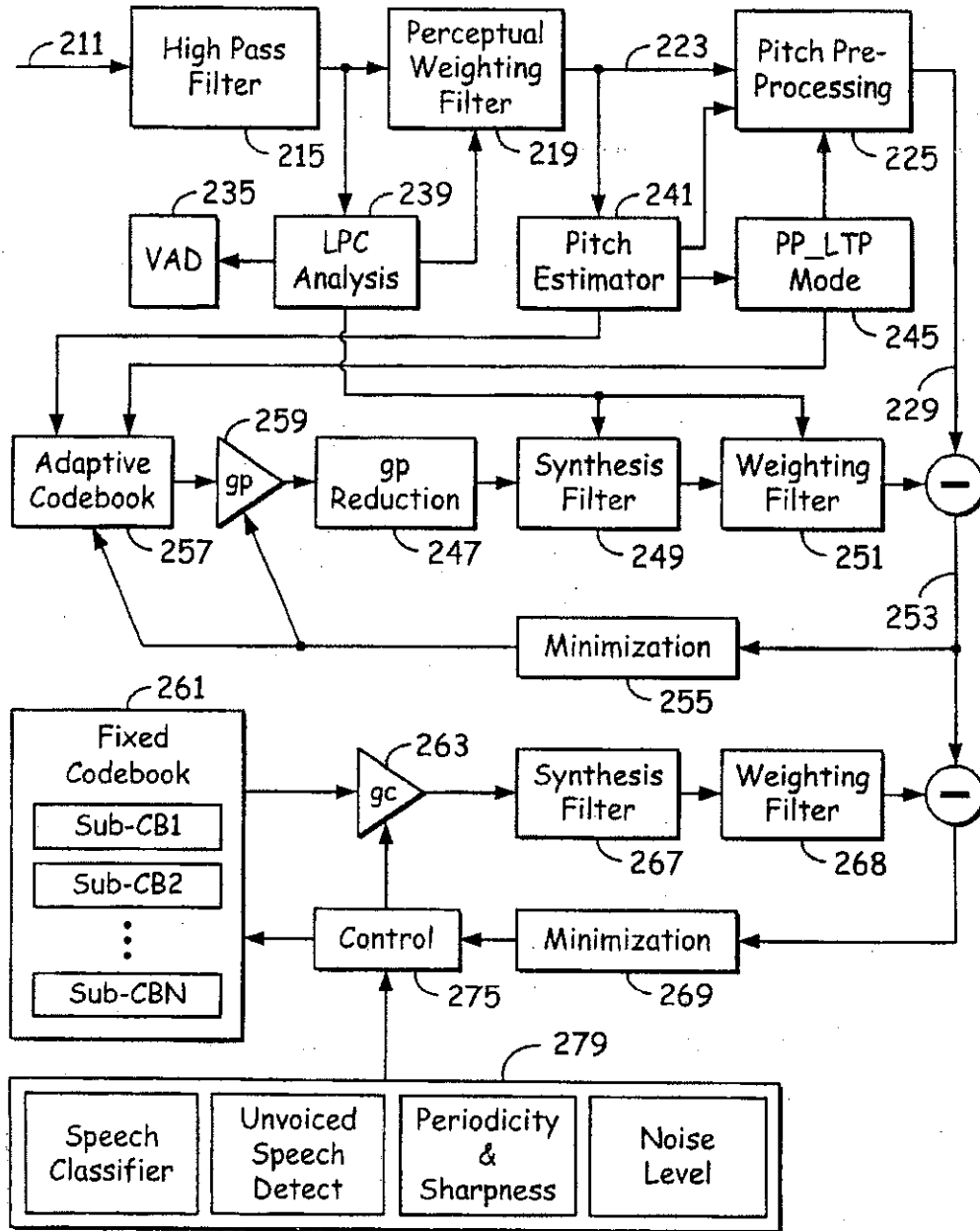


Fig. 2

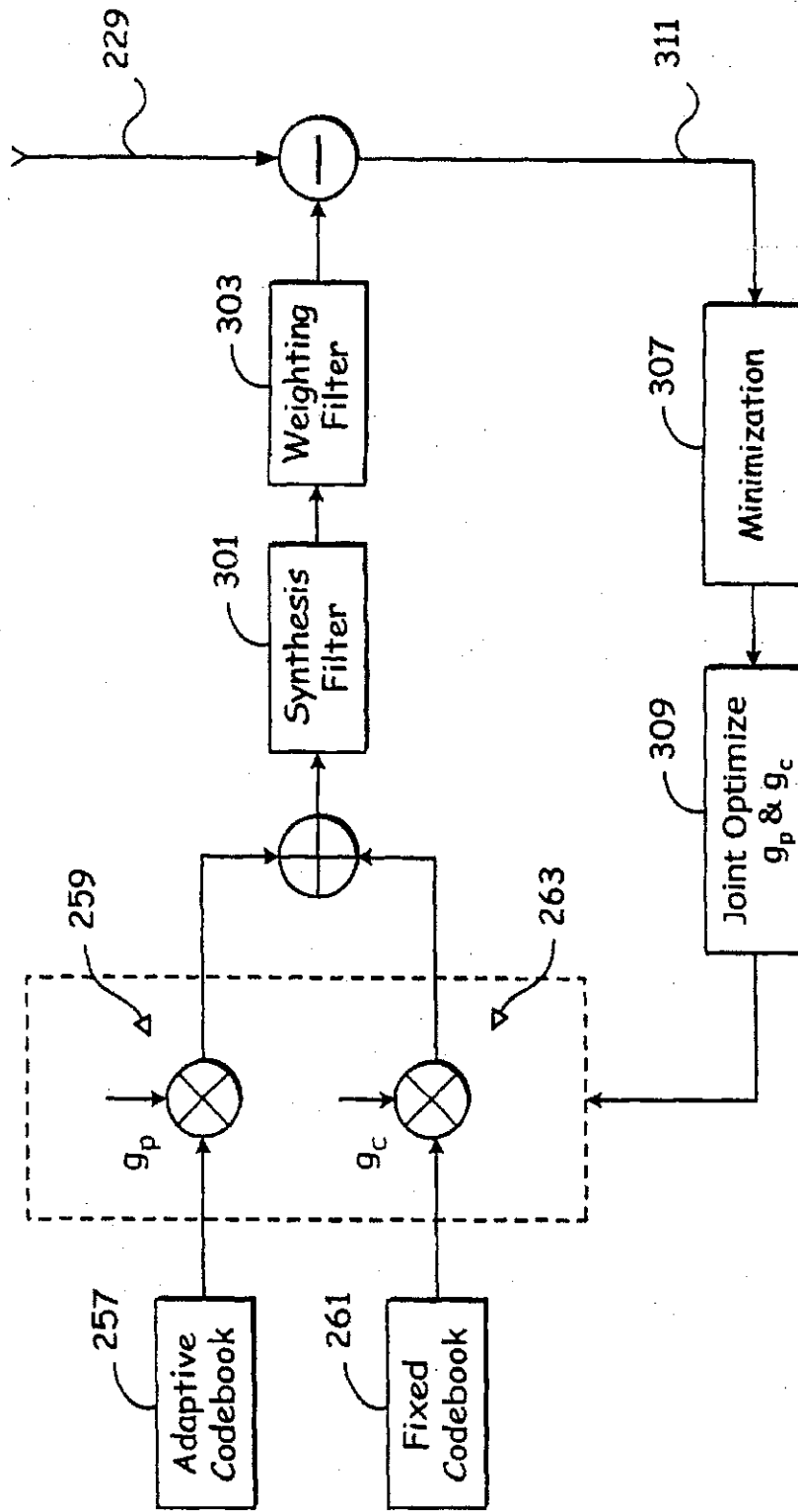


Fig. 3

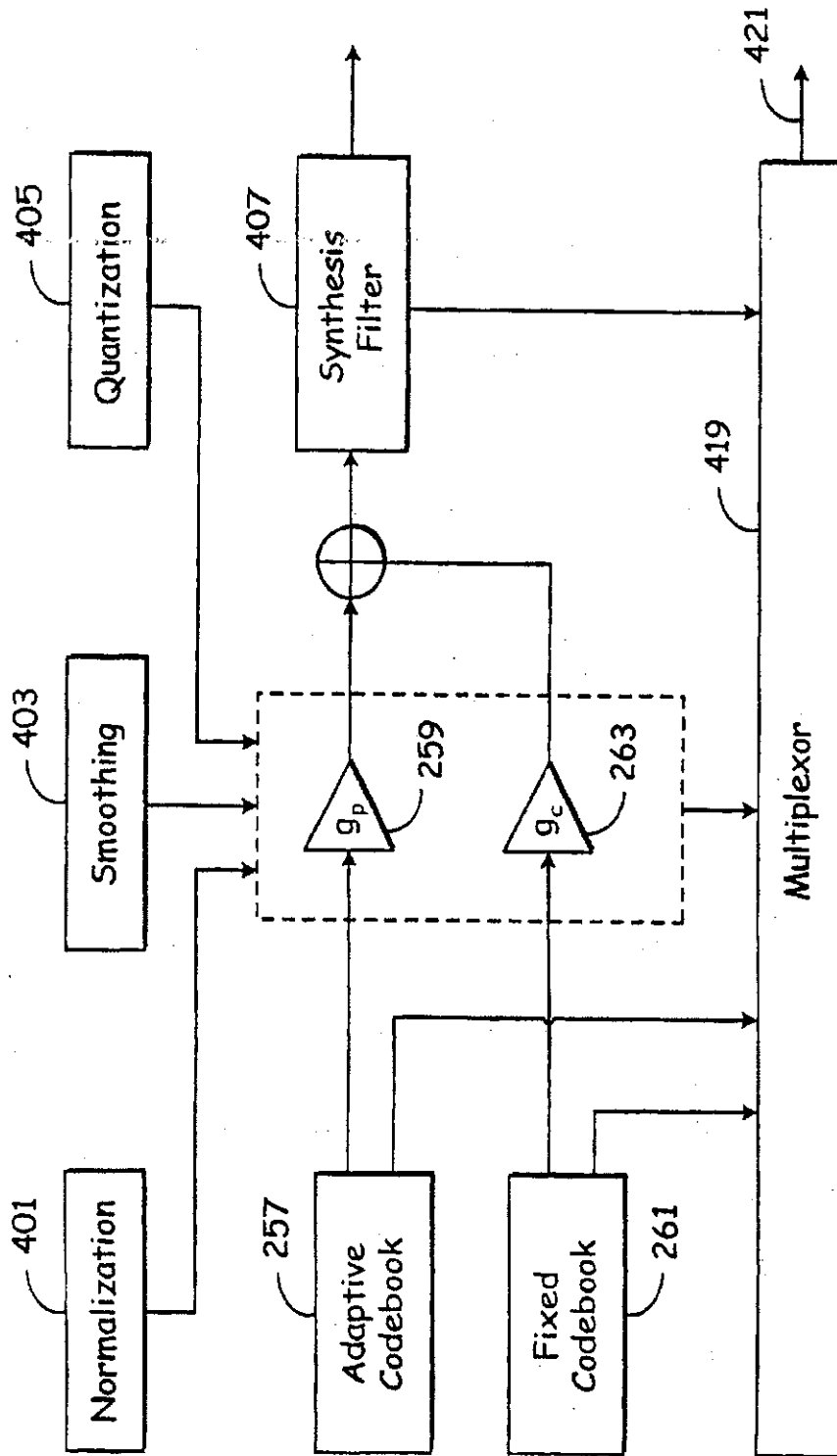


Fig. 4



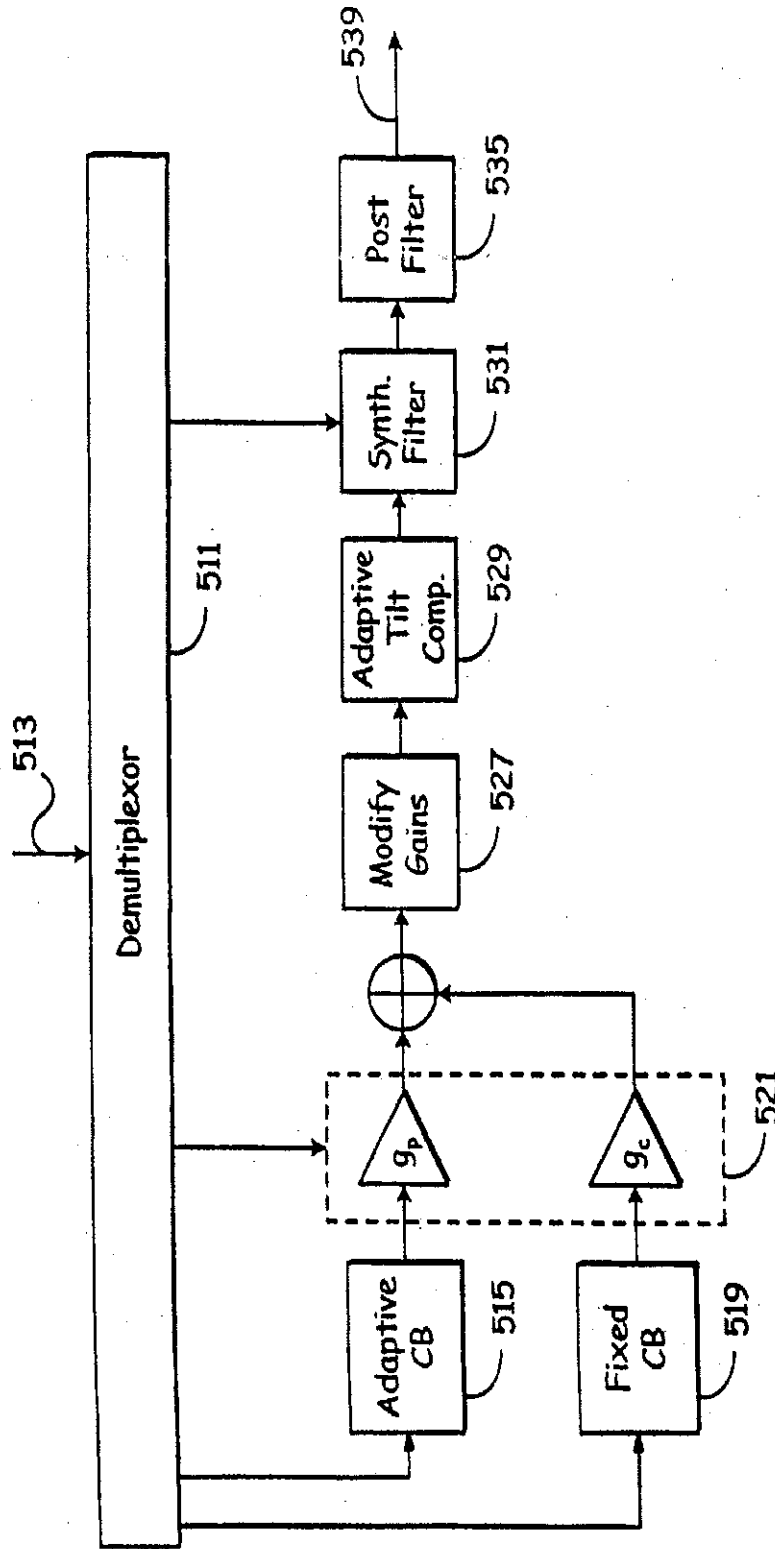


Fig. 5

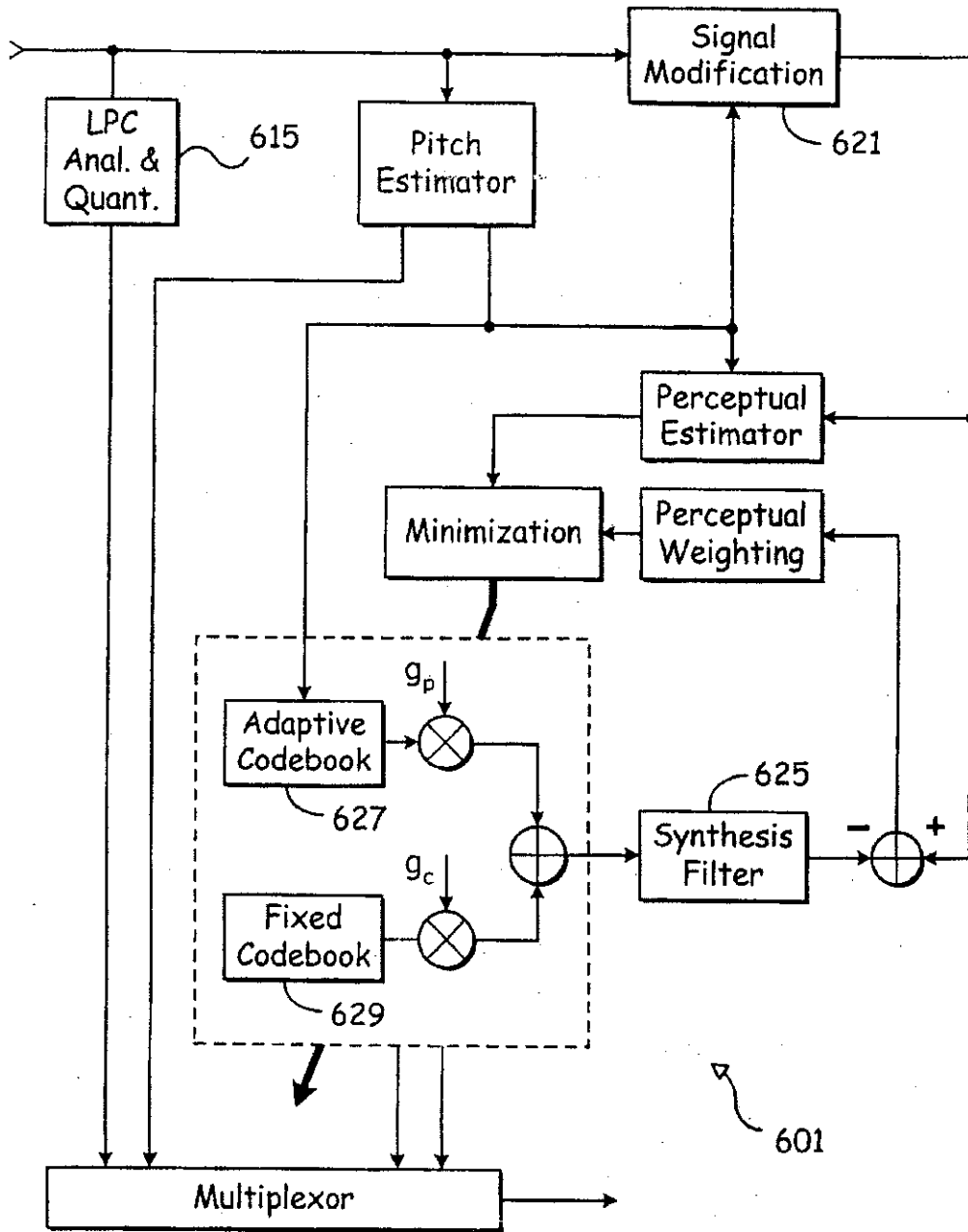


Fig. 6

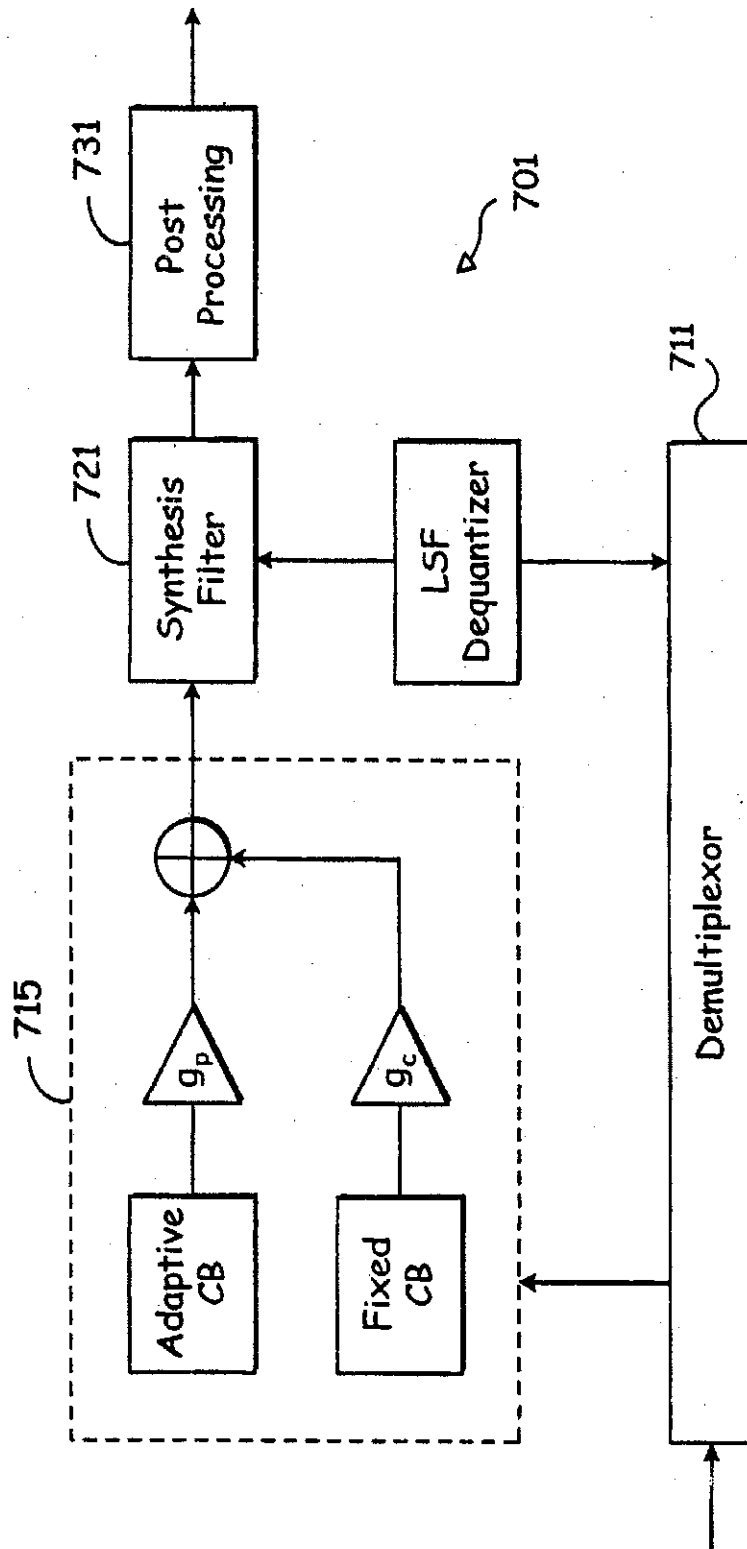


Fig. 7

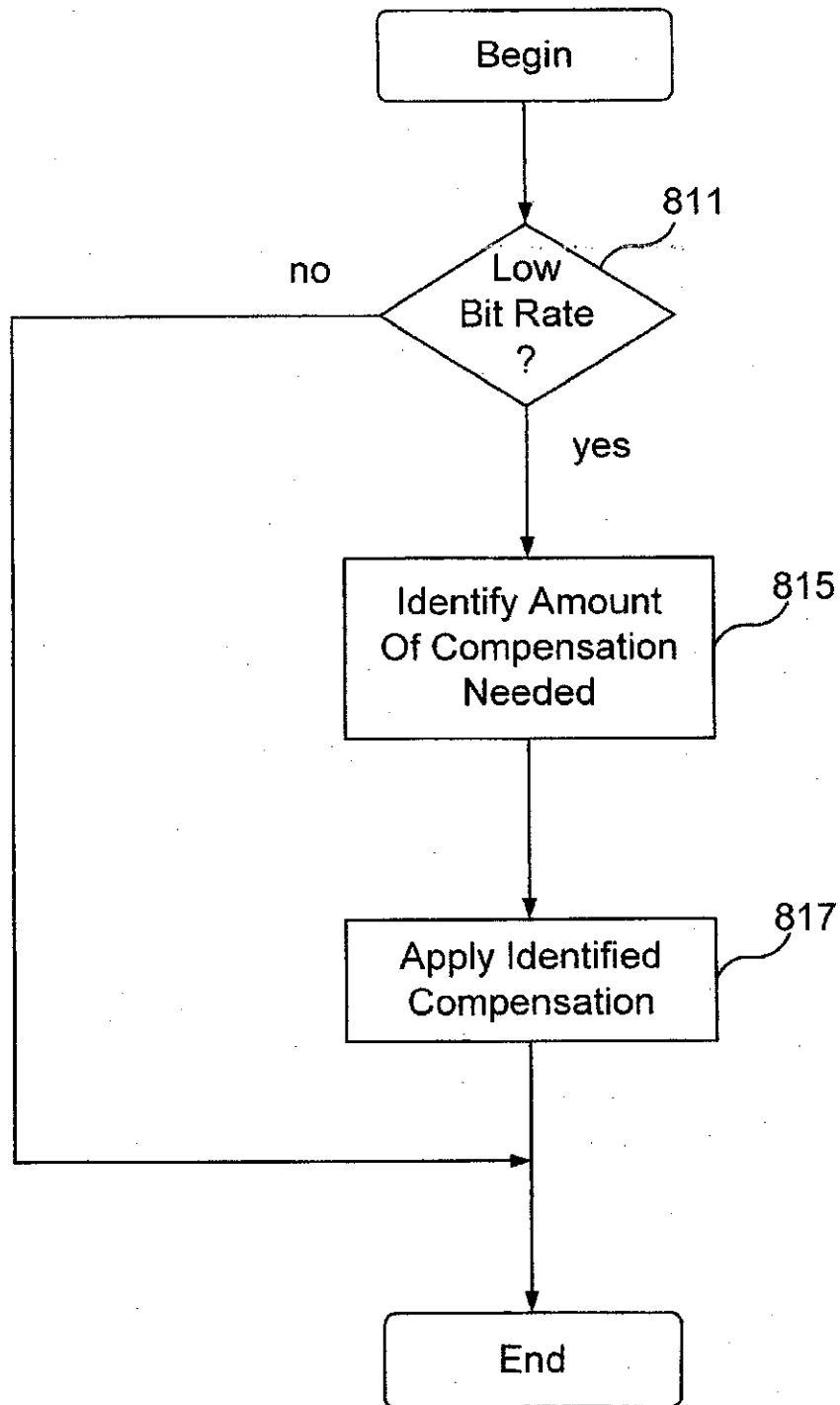


Fig. 8

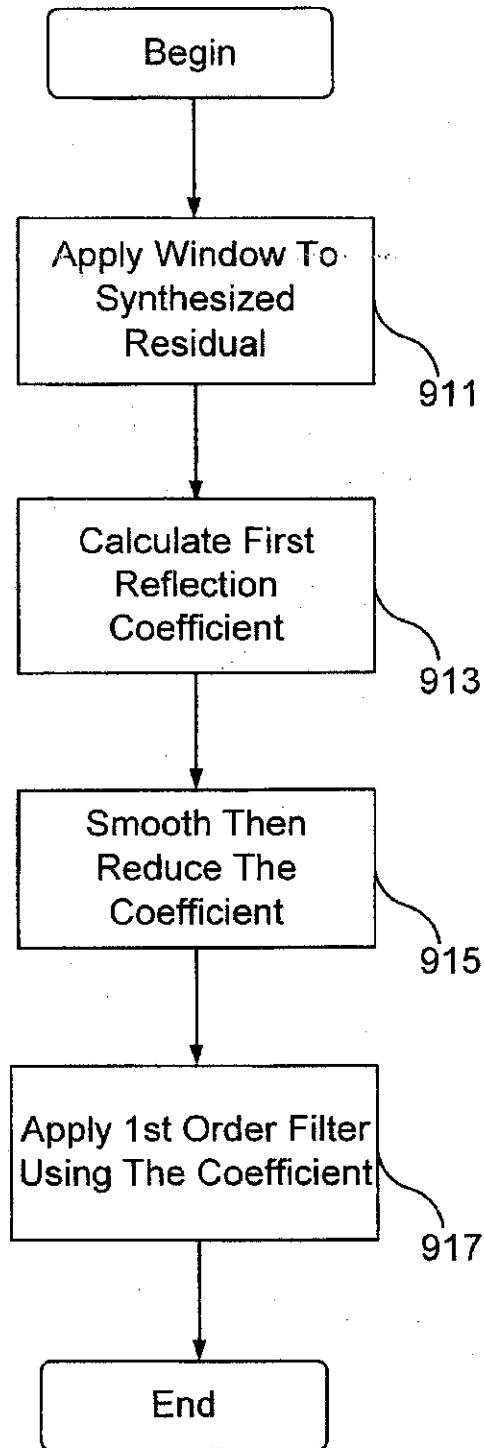


Fig. 9

## ADAPTIVE TILT COMPENSATION FOR SYNTHESIZED SPEECH RESIDUAL

### CROSS-REFERENCE TO RELATED APPLICATIONS

The present application is based on U.S. Provisional Application Serial No. 60/097,569, filed Aug. 24, 1998.

### MICROFICHE/COPYRIGHT REFERENCE

A microfiche Appendix of the code is attached and comprises one (1) sheet having a total of forty five (45) frames.

### INCORPORATION BY REFERENCE

The following applications are hereby incorporated by reference in their entirety and made part of the present application:

1. U.S. Provisional Application Serial No. 60/097,569, entitled "Adaptive Rate Speech Codec," filed Aug. 24, 1998;
2. U.S. patent application Ser. No. 09/154,675, entitled "Speech Encoder Using Continuous Warping In Long Term Preprocessing," filed Sep. 18, 1998;
3. U.S. patent application Ser. No. 09/156,814, entitled "Selectable Mode Vocoder System," filed Sep. 18, 1998;
4. U.S. patent application Ser. No. 09/156,649, entitled "Comb Codebook Structure," filed Sep. 18, 1998;
5. U.S. patent application Ser. No. 09/156,648, entitled "Low Complexity Random Codebook Structure," filed Sep. 18, 1998;
6. U.S. patent application Ser. No. 09/156,650, entitled "Speech Encoder Using Gain Normalization that Combines Open and Closed Loop Gains," filed Sep. 18, 1998;
7. U.S. patent application Ser. No. 09/156,832, entitled "Speech Encoder Using Voice Activity Detection in Coding Noise," filed Sep. 18, 1998;
8. U.S. patent application Ser. No. 09/154,660, entitled "Speech Encoder Adaptively Applying Pitch Processing with Continuous Warping," filed Sep. 18, 1998;
9. U.S. patent application Ser. No. 09/154,654, entitled "Pitch Determination Using Speech Classification and Prior Pitch Estimation," filed Sep. 18, 1998;
10. U.S. patent application Ser. No. 09/154,657, entitled "Speech Encoder Using A Classifier For Smoothing Noise Coding," filed Sep. 18, 1998;
11. U.S. patent application Ser. No. 09/154,663, entitled "Adaptive Gain Reduction To Produce Fixed Codebook Target Signal," filed Sep. 18, 1998;
12. U.S. patent application Ser. No. 09/154,662, entitled "Speech Classification and Parameter Weighting Used in Codebook Search," filed Sep. 18, 1998;
13. U.S. patent application Ser. No. 09/154,653, entitled "Synchronized Encoder-Decoder Frame Concealment Using Speech Coding Parameters," filed Sep. 18, 1998;
14. U.S. patent application Ser. No. 09/157,083, entitled "Robust Fast Search For Two-Dimensional Gain Vector Quantizer," filed Sep. 18, 1998;
15. U.S. patent application Ser. No. 09/156,416, entitled "Method and Apparatus for Detecting Voice Activity in a Speech Signal," filed Sep. 18, 1998.

## BACKGROUND

### 1. Technical Field

The present invention relates generally to speech encoding and decoding in voice communication systems; and, more particularly, it relates to various techniques used with code-excited linear prediction coding to obtain high quality speech reproduction through a limited bit rate communication channel.

### 2. Related Art

Signal modeling and parameter estimation play significant roles in communicating voice information with limited bandwidth constraints. To model basic speech sounds, speech signals are sampled as a discrete waveform to be digitally processed. In one type of signal coding technique called LPC (linear predictive coding), the signal value at any particular time index is modeled as a linear function of previous values. A subsequent signal is thus linearly predictable according to an earlier value. As a result, efficient signal representations can be determined by estimating and applying certain prediction parameters to represent the signal.

Applying LPC techniques, a conventional source encoder operates on speech signals to extract modeling and parameter information for communication to a conventional source decoder via a communication channel. Once received, the decoder attempts to reconstruct a counterpart signal for playback that sounds to a human ear like the original speech.

A certain amount of communication channel bandwidth is required to communicate the modeling and parameter information to the decoder. In embodiments, for example where the channel bandwidth is shared and real-time reconstruction is necessary, a reduction in the required bandwidth proves beneficial. However, using conventional modeling techniques, the quality requirements in the reproduced speech limit the reduction of such bandwidth below certain levels.

In conventional code-excited linear predictive coding, waveform matching in the high frequency region proves more difficult than matching in the low frequency region. Thus, the energy of the high frequency region of a synthesized speech signal drops more than in the low frequency region, especially for low bit rate coding. Moreover, the amount of high frequency energy drop is not consistent. As a result, with conventional, lower bit rate speech codecs, reproduced speech signals exhibit poor (dull) sound quality.

Further limitations and disadvantages of conventional systems will become apparent to one of skill in the art after reviewing the remainder of the present application with reference to the drawings.

### SUMMARY OF THE INVENTION

Various aspects of the present invention can be found in a speech system using an analysis by synthesis approach on a speech signal. The speech system comprises at least one codebook, containing at least one code vector, and processing circuitry. Using the at least one codebook, the processing circuitry generates a synthesized residual signal. The processing circuitry applies adaptive tilt compensation to the synthesized residual signal. The processing circuitry may also comprise both an encoder processing circuit that generates the synthesized residual signal, and a decoder processing circuit that applies the adaptive tilt compensation.

In other variations, the synthesized residual signal is a weighted synthesized residual signal. The adaptive tilt compensation may involve identification of a filter coefficient for use in a compensating filter, e.g., a first order filter. Such identification can be carried out by applying a window to the synthesized residual.

Further aspects of the present invention may be found in a speech system that also uses an analysis by synthesis approach on a speech signal. Therein, in addition to a codebook, a first processing circuit and second processing circuit can be found. The first processing circuit generates both a residual signal and, using the codebook, a synthesized residual signal. Both of these signals may be weighted. The residual signal has a first spectral envelope, while the synthesized residual has a second spectral envelope that exhibits variations from the first. The second processing circuit adaptively attempts to minimize such variations. In at least some embodiments, the attempt is made without having access to the residual signal. Of course, at least most of the aforementioned variations are equally applicable to the present speech system.

Other aspects, advantages and novel features of the present invention will become apparent from the following detailed description of the invention when considered in conjunction with the accompanying drawings.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1a is a schematic block diagram of a speech communication system illustrating the use of source encoding and decoding in accordance with the present invention.

FIG. 1b is a schematic block diagram illustrating an exemplary communication device utilizing the source encoding and decoding functionality of FIG. 1a.

FIGS. 2-4 are functional block diagrams illustrating a multi-step encoding approach used by one embodiment of the speech encoder illustrated in FIGS. 1a and 1b. In particular, FIG. 2 is a functional block diagram illustrating of a first stage of operations performed by one embodiment of the speech encoder of FIGS. 1a and 1b. FIG. 3 is a functional block diagram of a second stage of operations, while FIG. 4 illustrates a third stage.

FIG. 5 is a block diagram of one embodiment of the speech decoder shown in FIGS. 1a and 1b having corresponding functionality to that illustrated in FIGS. 2-4.

FIG. 6 is a block diagram of an alternate embodiment of a speech encoder that is built in accordance with the present invention.

FIG. 7 is a block diagram of an embodiment of a speech decoder having corresponding gas functionality to that of the speech encoder of FIG. 6.

FIG. 8 is a flow diagram illustrating use of adaptive tilt compensation in an exemplary decoder built in accordance with the present invention.

FIG. 9 is a flow diagram illustrating a specific embodiment of a decoder that illustrates and exemplary approach for performing the identification and compensation processing of FIG. 8.

#### DETAILED DESCRIPTION

FIG. 1a is a schematic block diagram of a speech communication system illustrating the use of source encoding and decoding in accordance with the present invention. Therein, a speech communication system 100 supports communication and reproduction of speech across a communication channel 103. Although it may comprise for example a wire, fiber or optical link, the communication channel 103 typically comprises, at least in part, a radio frequency link that often must support multiple, simultaneous speech exchanges requiring shared bandwidth resources such as may be found with cellular telephony embodiments.

Although not shown, a storage device may be coupled to the communication channel 103 to temporarily store speech information for delayed reproduction or playback, e.g., to perform answering machine functionality, voiced email, etc. Likewise, the communication channel 103 might be replaced by such a storage device in a single device embodiment of the communication system 100 that, for example, merely records and stores speech for subsequent playback.

In particular, a microphone 111 produces a speech signal in real time. The microphone 111 delivers the speech signal to an A/D (analog to digital) converter 115. The A/D converter 115 converts the speech signal to a digital form then delivers the digitized speech signal to a speech encoder 117.

The speech encoder 117 encodes the digitized speech by using a selected one of a plurality of encoding modes. Each of the plurality of encoding modes utilizes particular techniques that attempt to optimize quality of resultant reproduced speech. While operating in any of the plurality of modes, the speech encoder 117 produces a series of modeling and parameter information (hereinafter "speech indices"), and delivers the speech indices to a channel encoder 119.

The channel encoder 119 coordinates with a channel decoder 131 to deliver the speech indices across the communication channel 103. The channel decoder 131 forwards the speech indices to a speech decoder 133. While operating in a mode that corresponds to that of the speech encoder 117, the speech decoder 133 attempts to recreate the original speech from the speech indices as accurately as possible at a speaker 137 via a D/A (digital to analog) converter 135.

The speech encoder 117 adaptively selects one of the plurality of operating modes based on the data rate restrictions through the communication channel 103. The communication channel 103 comprises a bandwidth allocation between the channel encoder 119 and the channel decoder 131. The allocation is established, for example, by telephone switching networks wherein many such channels are allocated and reallocated as need arises. In one such embodiment, either a 22.8 kbps (kilobits per second) channel bandwidth, i.e., a full rate channel, or a 11.4 kbps channel bandwidth, i.e., a half rate channel, may be allocated.

With the full rate channel bandwidth allocation, the speech encoder 117 may adaptively select an encoding mode that supports a bit rate of 11.0, 8.0, 6.65 or 5.8 kbps. The speech encoder 117 adaptively selects an either 8.0, 6.65, 5.8 or 4.5 kbps encoding bit rate mode when only the half rate channel has been allocated. Of course these encoding bit rates and the aforementioned channel allocations are only representative of the present embodiment. Other variations to meet the goals of alternate embodiments are contemplated.

With either the full or half rate allocation, the speech encoder 117 attempts to communicate using the highest encoding bit rate mode that the allocated channel will support. If the allocated channel is or becomes noisy or otherwise restrictive to the highest or higher encoding bit rates, the speech encoder 117 adapts by selecting a lower bit rate encoding mode. Similarly, when the communication channel 103 becomes more favorable, the speech encoder 117 adapts by switching to a higher bit rate encoding mode.

With lower bit rate encoding, the speech encoder 117 incorporates various techniques to generate better low bit rate speech reproduction. Many of the techniques applied are based on characteristics of the speech itself. For example,

with lower bit rate encoding, the speech encoder 117 classifies noise, unvoiced speech, and voiced speech so that an appropriate modeling scheme corresponding to a particular classification can be selected and implemented. Thus, the speech encoder 117 adaptively selects from among a plurality of modeling schemes those most suited for the current speech. The speech encoder 117 also applies various other techniques to optimize the modeling as set forth in more detail below.

FIG. 1b is a schematic block diagram illustrating several variations of an exemplary communication device employing the functionality of FIG. 1a. A communication device 151 comprises both a speech encoder and decoder for simultaneous capture and reproduction of speech. Typically within a single housing, the communication device 151 might, for example, comprise a cellular telephone, portable telephone, computing system, etc. Alternatively, with some modification to include for example a memory element to store encoded speech information the communication device 151 might comprise an answering machine, a recorder, voice mail system, etc.

A microphone 155 and an A/D converter 157 coordinate to deliver a digital voice signal to an encoding system 159. The encoding system 159 performs speech and channel encoding and delivers resultant speech information to the channel. The delivered speech information may be destined for another communication device (not shown) at a remote location.

As speech information is received, a decoding system 165 performs channel and speech decoding then coordinates with a D/A converter 167 and a speaker 169 to reproduce something that sounds like the originally captured speech.

The encoding system 159 comprises both a speech processing circuit 185 that performs speech encoding, and a channel processing circuit 187 that performs channel encoding. Similarly, the decoding system 165 comprises a speech processing circuit 189 that performs speech decoding, and a channel processing circuit 191 that performs channel decoding.

Although the speech processing circuit 185 and the channel processing circuit 187 are separately illustrated, they might be combined in part or in total into a single unit. For example, the speech processing circuit 185 and the channel processing circuit 187 might share a single DSP (digital signal processor) and/or other processing circuitry. Similarly, the speech processing circuit 189 and the channel processing circuit 191 might be entirely separate or combined in part or in whole. Moreover, combinations in whole or in part might be applied to the speech processing circuits 185 and 189, the channel processing circuits 187 and 191, the processing circuits 185, 187, 189 and 191, or otherwise.

The encoding system 159 and the decoding system 165 both utilize a memory 161. The speech processing circuit 185 utilizes a fixed codebook 181 and an adaptive codebook 183 of a speech memory 177 in the source encoding process. The channel processing circuit 187 utilizes a channel memory 175 to perform channel encoding. Similarly, the speech processing circuit 189 utilizes the fixed codebook 181 and the adaptive codebook 183 in the source decoding process. The channel processing circuit 187 utilizes the channel memory 175 to perform channel decoding.

Although the speech memory 177 is shared as illustrated, separate copies thereof can be assigned for the processing circuits 185 and 189. Likewise, separate channel memory can be allocated to both the processing circuits 187 and 191. The memory 161 also contains software utilized by the

processing circuits 185, 187, 189 and 191 to perform various functionality required in the source and channel encoding and decoding processes.

FIGS. 2-4 are functional block diagrams illustrating a multi-step encoding approach used by one embodiment of the speech encoder illustrated in FIGS. 1a and 1b. In particular, FIG. 2 is a functional block diagram illustrating of a first stage of operations performed by one embodiment of the speech encoder shown in FIGS. 1a and 1b. The speech encoder, which comprises encoder processing circuitry, typically operates pursuant to software instruction carrying out the following functionality.

At a block 215, source encoder processing circuitry performs high pass filtering of a speech signal 211. The filter uses a cutoff frequency of around 80 Hz to remove, for example, 60 Hz power line noise and other lower frequency signals. After such filtering, the source encoder processing circuitry applies a perceptual weighting filter as represented by a block 219. The perceptual weighting filter operates to emphasize the valley areas of the filtered speech signal.

If the encoder processing circuitry selects operation in a pitch preprocessing (PP) mode as indicated at a control block 245, a pitch preprocessing operation is performed on the weighted speech signal at a block 225. The pitch preprocessing operation involves warping the weighted speech signal to match interpolated pitch values that will be generated by the decoder processing circuitry. When pitch preprocessing is applied, the warped speech signal is designated a first target signal 229. If pitch preprocessing is not selected by the control block 245, the weighted speech signal passes through the block 225 without pitch preprocessing and is designated the first target signal 229.

As represented by a block 255, the encoder processing circuitry applies a process wherein a contribution from an adaptive codebook 257 is selected along with a corresponding gain 257 which minimize a first error signal 253. The first error signal 253 comprises the difference between the first target signal 229 and a weighted, synthesized contribution from the adaptive codebook 257.

At blocks 247, 249 and 251, the resultant excitation vector is applied after adaptive gain reduction to both a synthesis and a weighting filter to generate a modeled signal that best matches the first target signal 229. The encoder processing circuitry uses LPC (linear predictive coding) analysis, as indicated by a block 239, to generate filter parameters for the synthesis and weighting filters. The weighting filters 219 and 251 are equivalent in functionality.

Next, the encoder processing circuitry designates the first error signal 253 as a second target signal for matching using contributions from a fixed codebook 261. The encoder processing circuitry searches through at least one of the plurality of subcodebooks within the fixed codebook 261 in an attempt to select a most appropriate contribution while generally attempting to match the second target signal.

More specifically, the encoder processing circuitry selects an excitation vector, its corresponding subcodebook and gain based on a variety of factors. For example, the encoding bit rate, the degree of minimization, and characteristics of the speech itself as represented by a block 279 are considered by the encoder processing circuitry at control block 275. Although many other factors may be considered, exemplary characteristics include speech classification, noise level, sharpness, periodicity, etc. Thus, by considering other such factors, a first subcodebook with its best excitation vector may be selected rather than a second subcodebook's best excitation vector even though the second subcodebook's better minimizes the second target signal 253.



FIG. 3 is a functional block diagram depicting a second stage of operations performed by the embodiment of the speech encoder illustrated in FIG. 2. In the second stage, the speech encoding circuitry simultaneously uses both the adaptive and the fixed codebook vectors found in the first stage of operations to minimize a third error signal 311.

The speech encoding circuitry searches for optimum gain values for the previously identified excitation vectors (in the first stage) from both the adaptive and fixed codebooks 257 and 261. As indicated by blocks 307 and 309, the speech encoding circuitry identifies the optimum gain by generating a synthesized and weighted signal, i.e., via a block 301 and 303, that best matches the first target signal 229 (which minimizes the third error signal 311). Of course if processing capabilities permit, the first and second stages could be combined wherein joint optimization of both gain and adaptive and fixed codebook vector selection could be used.

FIG. 4 is a functional block diagram depicting a third stage of operations performed by the embodiment of the speech encoder illustrated in FIGS. 2 and 3. The encoder processing circuitry applies gain normalization, smoothing and quantization, as represented by blocks 401, 403 and 405, respectively, to the jointly optimized gains identified in the second stage of encoder processing. Again, the adaptive and fixed codebook vectors used are those identified in the first stage processing.

With normalization, smoothing and quantization functionally applied, the encoder processing circuitry has completed the modeling process. Therefore, the modeling parameters identified are communicated to the decoder. In particular, the encoder processing circuitry delivers an index to the selected adaptive codebook vector to the channel encoder via a multiplexor 419. Similarly, the encoder processing circuitry delivers the index to the selected fixed codebook vector, resultant gains, synthesis filter parameters, etc., to the multiplexor 419. The multiplexor 419 generates a bit stream 421 of such information for delivery to the channel encoder for communication to the channel and speech decoder of receiving device.

FIG. 5 is a block diagram of an embodiment illustrating functionality of speech decoder having corresponding functionality to that illustrated in FIGS. 2-4. As with the speech encoder, the speech decoder, which comprises decoder processing circuitry, typically operates pursuant to software instruction carrying out the following functionality.

A demultiplexor 511 receives a bit stream 513 of speech modeling indices from an often remote encoder via a channel decoder. As previously discussed, the encoder selected each index value during the multi-stage encoding process described above in reference to FIGS. 2-4. The decoder processing circuitry utilizes indices, for example, to select excitation vectors from an adaptive codebook 515 and a fixed codebook 519, set the adaptive and fixed codebook gains at a block 521, and set the parameters for a synthesis filter 531.

With such parameters and vectors selected or set, the decoder processing circuitry generates a reproduced speech signal 539. In particular, the codebooks 515 and 519 generate excitation vectors identified by the indices from the demultiplexor 511. The decoder processing circuitry applies the indexed gains at the block 521 to the vectors which are summed. At a block 527, the decoder processing circuitry modifies the gains to emphasize the contribution of vector from the adaptive codebook 515. At a block 529, adaptive tilt compensation is applied to the combined vectors with a goal of flattening the excitation spectrum. The decoder

processing circuitry performs synthesis filtering at the block 531 using the flattened excitation signal. Finally, to generate the reproduced speech signal 539, post filtering is applied at a block 535 deemphasizing the valley areas of the reproduced speech signal 539 to reduce the effect of distortion.

In the exemplary cellular telephony embodiment of the present invention, the A/D converter 115 (FIG. 1a) will generally involve analog to uniform digital PCM including: 1) an input level adjustment device; 2) an input anti-aliasing filter; 3) a sample-and-hold device sampling at 8 kHz; and 4) analog to uniform digital conversion to 13-bit representation.

Similarly, the D/A converter 135 will generally involve uniform digital PCM to analog including: 1) conversion from 13-bit/8 kHz uniform PCM to analog; 2) a hold device; 3) reconstruction filter including  $x/\sin(x)$  correction; and 4) an output level adjustment device.

In terminal equipment, the A/D function may be achieved by direct conversion to 13-bit uniform PCM format, or by conversion to 8-bit/A-law compounded format. For the D/A operation, the inverse operations take place.

The encoder 117 receives data samples with a resolution of 13 bits left justified in a 16-bit word. The three least significant bits are set to zero. The decoder 133 outputs data in the same format. Outside the speech codec, further processing can be applied to accommodate traffic data having a different representation.

A specific embodiment of an AMR (adaptive multi-rate) codec with the operational functionality illustrated in FIGS. 2-5 uses five source codecs with bit-rates 11.0, 8.0, 6.65, 5.8 and 4.55 kbps. Four of the highest source coding bit-rates are used in the full rate channel and the four lowest bit-rates in the half rate channel.

All five source codecs within the AMR codec are generally based on a code-excited linear predictive (CELP) coding model. A 10th order linear prediction (LP), or short-term, synthesis filter, e.g., used at the blocks 249, 267, 301, 407 and 531 (of FIGS. 2-5), is used which is given by:

$$H(z) = \frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^m \hat{a}_i z^{-i}} \quad (1)$$

where  $\hat{a}_i$ ,  $i=1, \dots, m$ , are the (quantized) linear prediction (LP) parameters.

A long-term filter, i.e., the pitch synthesis filter, is implemented using the either an adaptive codebook approach or a pitch pre-processing approach. The pitch synthesis filter is given by:

$$\frac{1}{B(z)} = \frac{1}{1 - g_p z^{-T}} \quad (2)$$

where  $T$  is the pitch delay and  $g_p$  is the pitch gain.

With reference to FIG. 2, the excitation signal at the input of the short-term LP synthesis filter at the block 249 is constructed by adding two excitation vectors from the adaptive and the fixed codebooks 257 and 261, respectively. The speech is synthesized by feeding the two properly chosen vectors from these codebooks through the short-term synthesis filter at the block 249 and 267, respectively.

The optimum excitation sequence in a codebook is chosen using an analysis-by-synthesis search procedure in which the error between the original and synthesized speech is minimized according to a perceptually weighted distortion

measure. The perceptual weighting filter, e.g., at the blocks 251 and 268, used in the analysis-by-synthesis search technique is given by:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{3}$$

where  $A(z)$  is the unquantized LP filter and  $0 < \gamma_2 < \gamma_1 \leq 1$  are the perceptual weighting factors. The values  $\gamma_1 = [0.9, 0.94]$  and  $\gamma_2 = 0.6$  are used. The weighting filter, e.g., at the blocks 251 and 268, uses the unquantized LP parameters while the formant synthesis filter, e.g., at the blocks 249 and 267, uses the quantized LP parameters. Both the unquantized and quantized LP parameters are generated at the block 239.

The present encoder embodiment operates on 20 ms (millisecond) speech frames corresponding to 160 samples at the sampling frequency of 8000 samples per second. At each 160 speech samples, the speech signal is analyzed to extract the parameters of the CELP model, i.e., the LP filter coefficients, adaptive and fixed codebook indices and gains. These parameters are encoded and transmitted. At the decoder, these parameters are decoded and speech is synthesized by filtering the reconstructed excitation signal through the LP synthesis filter.

More specifically, LP analysis at the block 239 is performed twice per frame but only a single set of LP parameters is converted to line spectrum frequencies (LSF) and vector quantized using predictive multi-stage quantization (PMVQ). The speech frame is divided into subframes.

Third, in the LTP mode, closed-loop pitch analysis is performed to find the pitch lag and gain, using the first target signal 229,  $x(n)$ , and impulse response,  $h(n)$ , by searching around the open-loop pitch lag. Fractional pitch with various sample resolutions are used.

In the PP mode, the input original signal has been pitch-preprocessed to match the interpolated pitch contour, so no closed-loop search is needed. The LTP excitation vector is computed using the interpolated pitch contour and the past synthesized excitation.

Fourth, the encoder processing circuitry generates a new target signal  $x_2(n)$ , the second target signal 253, by removing the adaptive codebook contribution (filtered adaptive code vector) from  $x(n)$ . The encoder processing circuitry uses the second target signal 253 in the fixed codebook search to find the optimum innovation.

Fifth, for the 11.0 kbps bit rate mode, the gains of the adaptive and fixed codebook are scalar quantized with 4 and 5 bits respectively (with moving average prediction applied to the fixed codebook gain). For the other modes the gains of the adaptive and fixed codebook are vector quantized (with moving average prediction applied to the fixed codebook gain).

Finally, the filter memories are updated using the determined excitation signal for finding the first target signal in the next subframe.

The bit allocation of the AMR codec modes is shown in table 1. For example, for each 20 ms speech frame, 220, 160, 133, 116 or 91 bits are produced, corresponding to bit rates of 11.0, 8.0, 6.65, 5.8 or 4.55 kbps, respectively.

TABLE 1

Bit allocation of the AMR coding algorithm for 20 ms frame						
CODING RATE						
	11.0 KBPS	8.0 KBPS	6.65 KBPS	5.80 KBPS	4.55 KBPS	
Frame size	20 ms					
Look ahead	5 ms					
LPC order	10 <sup>th</sup> -order					
Predictor for LSF	1 predictor:			2 predictors:		
Quantization	0 bit/frame			1 bit/frame		
LSF Quantization	28 bit/frame		24 bit/frame			18
LPC interpolation	2 bits/frame	2 bits/f	0	2 bits/f	0	0
Coding mode bit	0 bit	0 bit		1 bit/frame	0 bit	0 bit
Pitch mode	LTP	LTP	LTP	PP	PP	PP
Subframe size	5 ms					
Pitch Lag	30 bits/frame (9696)	8585	8585	0008	0008	0008
Fixed excitation	31 bits/subframe	20	13	18	14 bits/subframe	10 bits/subframe
Gain quantization	9 bits (scalar)		7 bits/subframe			6 bits/subframe
Total	220 bits/frame	160	133	133	116	91

Parameters from the adaptive and fixed codebooks 257 and 261 are transmitted every subframe. The quantized and unquantized LP parameters or their interpolated versions are used depending on the subframe. An open-loop pitch lag is estimated at the block 241 once or twice per frame for PP mode or LTP mode, respectively.

Each subframe, at least the following operations are repeated. First, the encoder processing circuitry (operating pursuant to software instruction) computes  $x(n)$ , the first target signal 229, by filtering the LP residual through the weighted synthesis filter  $W(z)H(z)$  with the initial states of the filters having been updated by filtering the error between LP residual and excitation. This is equivalent to an alternate approach of subtracting the zero input response of the weighted synthesis filter from the weighted speech signal.

Second, the encoder processing circuitry computes the impulse response,  $h(n)$ , of the weighted synthesis filter.

With reference to FIG. 5, the decoder processing circuitry, pursuant to software control, reconstructs the speech signal using the transmitted modeling indices extracted from the received bit stream by the demultiplexor 511. The decoder processing circuitry decodes the indices to obtain the coder parameters at each transmission frame. These parameters are the LSF vectors, the fractional pitch lags, the innovative code vectors, and the two gains.

The LSF vectors are converted to the LP filter coefficients and interpolated to obtain LP filters at each subframe. At each subframe, the decoder processing circuitry constructs the excitation signal by: 1) identifying the adaptive and innovative code vectors from the codebooks 515 and 519; 2) scaling the contributions by their respective gains at the block 521; 3) summing the scaled contributions; and 3) modifying and applying adaptive tilt compensation at the blocks 527 and 529. The speech signal is also reconstructed

on a subframe basis by filtering the excitation through the LP synthesis at the block 531. Finally, the speech signal is passed through an adaptive post filter at the block 535 to generate the reproduced speech signal 539.

The AMR encoder will produce the speech modeling information in a unique sequence and format, and the AMR decoder receives the same information in the same way. The different parameters of the encoded speech and their individual bits have unequal importance with respect to subjective quality. Before being submitted to the channel encoding function the bits are rearranged in the sequence of importance.

Two pre-processing functions are applied prior to the encoding process: high-pass filtering and signal down-scaling. Down-scaling consists of dividing the input by a factor of 2 to reduce the possibility of overflows in the fixed point implementation. The high-pass filtering at the block 215 (FIG. 2) serves as a precaution against undesired low frequency components. A filter with cut off frequency of 80 Hz is used, and it is given by:

$$H_{hp}(z) = \frac{0.92727435 - 1.8544941z^{-1} + 0.92727435z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-2}}$$

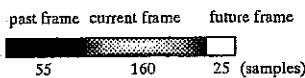
Down scaling and high-pass filtering are combined by dividing the coefficients of the numerator of  $H_{hp}(z)$  by 2.

Short-term prediction, or linear prediction (LP) analysis is performed twice per speech frame using the autocorrelation approach with 30 ms windows. Specifically, two LP analyses are performed twice per frame using two different windows. In the first LP analysis (LP\_analysis\_1), a hybrid window is used which has its weight concentrated at the fourth subframe. The hybrid window consists of two parts. The first part is half a Hamming window, and the second part is a quarter of a cosine cycle. The window is given by:

$$w_1(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{\pi n}{L}\right), & n = 0 \text{ to } 214, L = 215 \\ \cos\left(\frac{0.49(n-L)\pi}{25}\right), & n = 215 \text{ to } 239 \end{cases}$$

In the second LP analysis (LP\_analysis\_2), a symmetric Hamming window is used.

$$w_2(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{\pi n}{L}\right), & n = 0 \text{ to } 119, L = 120 \\ 0.54 + 0.46\cos\left(\frac{(n-L)\pi}{120}\right), & n = 120 \text{ to } 239 \end{cases}$$



In either LP analysis, the autocorrelations of the windowed speech  $s'(n)$ ,  $n=0,239$  are computed by:

$$r(k) = \sum_{n=k}^{239} s'(n)s'(n-k), k = 0, 10.$$

A 60 Hz bandwidth expansion is used by lag windowing, the autocorrelations using the window:

$$w_{lag}(i) = \exp\left[-\frac{1}{2}\left(\frac{2\pi 60i}{8000}\right)^2\right], i = 1, 10.$$

Moreover,  $r(0)$  is multiplied by a white noise correction factor 1.0001 which is equivalent to adding a noise floor at -40 dB.

The modified autocorrelations  $r'(0)=1.0001r(0)$  and  $r'(k)=r(k)w_{lag}(k)$ ,  $k=1,10$  are used to obtain the reflection coefficients  $k_i$  and LP filter coefficients  $a_i$ ,  $i=1,10$  using the Levinson-Durbin algorithm. Furthermore, the LP filter coefficients  $a_i$  are used to obtain the Line Spectral Frequencies (LSFs).

The interpolated unquantized LP parameters are obtained by interpolating the LSF coefficients obtained from the LP analysis\_1 and those from LP\_analysis\_2 as:

$$q_1(n) = 0.5q_4(n-1) + 0.5q_2(n)$$

$$q_3(n) = 0.5q_2(n) + 0.5q_4(n)$$

where  $q_1(n)$  is the interpolated LSF for subframe 1,  $q_2(n)$  is the LSF of subframe 2 obtained from LP\_analysis\_2 of current frame,  $q_3(n)$  is the interpolated LSF for subframe 3,  $q_4(n-1)$  is the LSF (cosine domain) from LP\_analysis\_1 of previous frame, and  $q_4(n)$  is the LSF for subframe 4 obtained from LP\_analysis\_1 of current frame. The interpolation is carried out in the cosine domain.

A VAD (Voice Activity Detection) algorithm is used to classify input speech frames into either active voice or inactive voice frame (background noise or silence) at a block 235 (FIG. 2).

The input speech  $s(n)$  is used to obtain a weighted speech signal  $s_w(n)$  by passing  $s(n)$  through a filter:

$$W(z) = \frac{A\left(\frac{z}{\gamma^2}\right)}{A\left(\frac{z}{\gamma}\right)}$$

That is, in a subframe of size  $L\_SF$ , the weighted speech is given by:

$$s_w(n) = s(n) + \sum_{i=1}^{10} a_i \gamma^i s(n-i) - \sum_{i=1}^{10} a_i \gamma^i s_w(n-i), n = 0, L\_SF - 1.$$

A voiced/unvoiced classification and mode decision within the block 279 using the input speech  $s(n)$  and the residual  $r_w(n)$  is derived where:

$$r_w(n) = s(n) + \sum_{i=1}^{10} a_i \gamma^i s(n-i), n = 0, L\_SF - 1.$$

The classification is based on four measures: 1) speech sharpness P1\_SHP; 2) normalized one delay correlation P2\_R1; 3) normalized zero-crossing rate P3\_ZC; and 4) normalized LP residual energy P4\_RE.

The speech sharpness is given by:

$$P1\_SHP = \frac{\sum_{n=0}^{L-1} \text{abs}(r_w(n))}{\text{MaxL}}$$

where Max is the maximum of abs( $r_w(n)$ ) over the specified interval of length L. The normalized one delay correlation and normalized zero-crossing rate are given by:

$$P2\_R1 = \frac{\sum_{n=0}^{L-1} s(n)s(n+1)}{\sqrt{\sum_{n=0}^{L-1} s(n)s(n) \sum_{n=0}^{L-1} s(n+1)s(n+1)}}$$

$$P3\_ZC = \frac{1}{2L} \sum_{i=0}^{L-1} [|\text{sgn}[s(i)] - \text{sgn}[s(i-1)]|]$$

where sgn is the sign function whose output is either 1 or -1 depending that the input sample is positive or negative. Finally, the normalized LP residual energy is given by:

$$P4\_RE = 1 - \sqrt{\text{lpc\_gain}}$$

where

$$\text{lpc\_gain} = \prod_{i=1}^{10} (1 - k_i^2)$$

where  $k_i$  are the reflection coefficients obtained from LP analysis\_1.

The voiced/unvoiced decision is derived if the following conditions are met:

- if P2\_R1<0.6 and P1\_SHP>0.2 set mode=2,
- if P3\_ZC>0.4 and P1\_SHP>0.18 set mode=2,
- if P4\_RE<0.4 and P1\_SHP>0.2 set mode=2,
- if (P2\_R1<-1.2+3.2P1\_SHP) set VUV=-3
- if (P4\_RE<-0.21+1.4286P1\_SHP) set VUV=-3
- if (P3\_ZC>0.8-0.6P1\_SHP) set VUV=-3
- if (P4\_RE<0.1) set VUV=-3

Open loop pitch analysis is performed once or twice (each 10 ms) per frame depending on the coding rate in order to find estimates of the pitch lag at the block 241 (FIG. 2). It is based on the weighted speech signal  $s_w(n+n_m), n=0, 1, \dots, 79$ , in which  $n_m$  defines the location of this signal on the first half frame or the last half frame. In the first step, four maxima of the correlation:

$$C_k = \sum_{n=0}^{79} s_w(n_m+n)s_w(n_m+n-k)$$

are found in the four ranges 17...33, 34...67, 68...135, 136...145, respectively. The retained maxima  $C_{k_i}, i=1,2,3,4$ , are normalized by dividing by:

$$\sqrt{\sum_n s_w^2(n_m+n-k)}$$

$i=1, \dots, 4$ , respectively.

The normalized maxima and corresponding delays are denoted by  $(R_i, k_i), i=1,2,3,4$ .

In the second step, a delay,  $k_i$ , among the four candidates, is selected by maximizing the four normalized correlations. In the third step,  $k_i$  is probably corrected to  $k_i(i<4)$  by favoring the lower ranges. That is,  $k_i(i<4)$  is selected if  $k_i$  is within  $[k_i/m-4, k_i/m+4], m=2,3,4,5$ , and if  $k_i > k_j, 0.95^{i-j} D, i<4$ , where D is 1.0, 0.85, or 0.65, depending on whether the previous frame is unvoiced, the previous frame is voiced and  $k_j$  is in the neighborhood (specified by  $\pm 8$ ) of the previous pitch lag, or the previous two frames are voiced and  $k_i$  is in the neighborhood of the previous two pitch lags. The final selected pitch lag is denoted by  $T_{op}$ .

A decision is made every frame to either operate the LTP (long-term prediction) as the traditional CELP approach (LTP\_mode=1), or as a modified time warping approach (LTP\_mode=0) herein referred to as PP (pitch preprocessing). For 4.55 and 5.8 kbps encoding bit rates, LTP\_mode is set to 0 at all times. For 8.0 and 11.0 kbps, LTP\_mode is set to 1 all of the time. Whereas, for a 6.65 kbps encoding bit rate, the encoder decides whether to operate in the LTP or PP mode. During the PP mode, only one pitch lag is transmitted per coding frame.

For 6.65 kbps, the decision algorithm is as follows. First, at the block 241, a prediction of the pitch lag pit for the current frame is determined as follows:

if (LTP\_MODE\_m=1)

$$\text{pit} = \text{lag1} + 2.4 * (\text{lag}_f[3] - \text{lag1});$$

else

$$\text{pit} = \text{lag}_f[1] + 2.75 * (\text{lag}_f[3] - \text{lag}_f[1]);$$

where LTP\_mode\_m is previous frame LTP\_mode,  $\text{lag}_f[1], \text{lag}_f[3]$  are the past closed loop pitch lags for second and fourth subframes respectively,  $\text{lag1}$  is the current frame open-loop pitch lag at the second half of the frame, and  $\text{lag1}$  is the previous frame open-loop pitch lag at the first half of the frame.

Second, a normalized spectrum difference between the Line Spectrum Frequencies (LSF) of current and previous frame is computed as:

$$e\_lsf = \frac{1}{10} \sum_{i=0}^9 \text{abs}(LSF(i) - LSF_{m(i)}).$$

if ( $\text{abs}(\text{pit} - \text{lag1}) < \text{TH}$  and  $\text{abs}(\text{lag}_f[3] - \text{lag1}) < \text{lag1} * 0.2$ )  
if ( $\text{Rp} > 0.5$  &&  $\text{pgain\_past} > 0.7$  and  $e\_lsf < 0.5/30$ )  
LTP\_mode=0;

else LTP\_mode=1;

where  $\text{Rp}$  is current frame normalized pitch correlation,  $\text{pgain\_past}$  is the quantized pitch gain from the fourth subframe of the past frame,  $\text{TH} = \text{MIN}(\text{lag1} * 0.1, 5)$ , and  $\text{TH} = \text{MAX}(2.0, \text{TH})$ .

The estimation of the precise pitch lag at the end of the frame is based on the normalized correlation:

$$R_k = \frac{\sum_{n=0}^L s_w(n+n1)s_w(n+n1-k)}{\sqrt{\sum_{n=0}^L s_w^2(n+n1-k)}}$$

where  $s_w(n+n1)$ ,  $n=0,1,\dots,L-1$ , represents the last segment of the weighted speech signal including the look-ahead (the look-ahead length is 25 samples), and the size  $L$  is defined according to the open-loop pitch lag  $T_{op}$  with the corresponding normalized correlation  $C_{T_{op}}$ :

$$\begin{aligned} &\text{if}(C_{T_{op}} > 0.6) \\ &L = \max\{50, T_{op}\} \\ &L = \min\{80, L\} \\ &\text{else} \\ &L = 80 \end{aligned}$$

In the first step, one integer lag  $k$  is selected maximizing the  $R_k$  in the range  $k \in [T_{op}-10, T_{op}+10]$  bounded by [17, 145]. Then, the precise pitch lag  $P_m$  and the corresponding index  $I_m$  for the current frame is searched around the integer lag,  $[k-1, k+1]$ , by up-sampling  $R_k$ .

The possible candidates of the precise pitch lag are obtained from the table named as PitLagTab8b[i],  $i=0,1,\dots,127$ . In the last step, the precise pitch lag  $P_m = \text{PitLagTab8b}[I_m]$  is possibly modified by checking the accumulated delay  $\tau_{acc}$  due to the modification of the speech signal:

$$\begin{aligned} &\text{if}(\tau_{acc} > 5) I_m \leftarrow \min\{I_m+1, 127\}, \text{ and} \\ &\text{if}(\tau_{acc} < -5) I_m \leftarrow \max\{I_m-1, 0\}. \end{aligned}$$

The precise pitch lag could be modified again:

$$\begin{aligned} &\text{if}(\tau_{acc} > 10) I_m \leftarrow \min\{I_m+1, 127\}, \text{ and} \\ &\text{if}(\tau_{acc} < -10) I_m \leftarrow \max\{I_m-1, 0\}. \end{aligned}$$

The obtained index  $I_m$  will be sent to the decoder.

The pitch lag contour,  $\tau_c(n)$ , is defined using both the current lag  $P_m$  and the previous lag  $P_{m-1}$ :

$$\text{if} ( |P_m - P_{m-1}| < 0.2 \min\{P_m, P_{m-1}\} )$$

$$\tau_c(n) = P_{m-1} + n(P_m - P_{m-1})/L_p, n=0,1,\dots,L_p-1$$

$$\tau_c(n) = P_m, n=L_p, \dots, 170$$

else

$$\tau_c(n) = P_{m-1}, n=0,1,\dots,39;$$

$$\tau_c(n) = P_m, n=40, \dots, 170$$

where  $L_p=160$  is the frame size.

One frame is divided into 3 subframes for the long-term preprocessing. For the first two subframes, the subframe size,  $L_s$ , is 53, and the subframe size for searching,  $L_{sr}$ , is 70. For the last subframe,  $L_s$  is 54 and  $L_{sr}$  is:

$$L_{sr} = \min\{70, L_s + L_{hd} - 10 - \tau_{acc}\},$$

where  $L_{hd}=25$  is the look-ahead and the maximum of the accumulated delay  $\tau_{acc}$  is limited to 14.

The target for the modification process of the weighted speech temporally memorized in  $\{\hat{s}_w(m0+n), n=0,1,\dots,L_{sr}-1\}$  is calculated by warping the past modified weighted speech buffer,  $\hat{s}_w(m0+n)$ ,  $n < 0$ , with the pitch lag contour,  $\tau_c(n+mL_s)$ ,  $m=0,1,2$ ,

$$\hat{s}_w(m0+n) = \sum_{i=-f_t}^{f_t} \hat{s}_w(m0+n-T_c(n)+i)I_s(i, T_{IC}(n)),$$

$$n = 0, 1, \dots, L_{sr}-1,$$

where  $T_c(n)$  and  $T_{IC}(n)$  are calculated by:

$$T_c(n) = \text{trunc}\{\tau_c(n+mL_s)\},$$

$$T_{IC}(n) = \tau_c(n) - T_c(n).$$

$m$  is subframe number,  $I_s(i, T_{IC}(n))$  is a set of interpolation coefficients, and  $f_t$  is 10. Then, the target for matching,  $\hat{s}_t(n)$ ,  $n=0,1,\dots,L_{sr}-1$ , is calculated by weighting  $\hat{s}_w(m0+n)$ ,  $n=0,1,\dots,L_{sr}-1$ , in the time domain:

$$\hat{s}_t(n) = n\hat{s}_w(m0+n)/L_s, n=0,1,\dots,L_s-1,$$

$$\hat{s}_t(n) = \hat{s}_w(m0+n), n=L_s, \dots, L_{sr}-1$$

The local integer shifting range [SR0, SR1] for searching for the best local delay is computed as the following:

if speech is unvoiced

$$\begin{aligned} \text{SR0} &= -1, \\ \text{SR1} &= 1, \end{aligned}$$

else

$$\text{SR0} = \text{round}\{ -4 \min\{1.0, \max\{0.0, 1-0.4(P_{sh}-0.2)\}\} \},$$

SR1 = round{ 4 min{1.0, max{0.0, 1-0.4(P<sub>sh</sub>-0.2)} } }, where  $P_{sh} = \max\{P_{sh1}, P_{sh2}\}$ ,  $P_{sh1}$  is the average to peak ratio (i.e., sharpness) from the target signal:

$$P_{sh1} = \frac{\sum_{n=0}^{L_{sr}-1} |\hat{s}_w(m0+n)|}{L_{sr} \max\{|\hat{s}_w(m0+n)|, n=0,1,\dots,L_{sr}-1\}}$$

and  $P_{sh2}$  is the sharpness from the weighted speech signal:

$$P_{sh2} = \frac{\sum_{n=0}^{L_{sr}-L_s/2-1} |s_w(n+n0+L_s/2)|}{(L_{sr}-L_s/2) \max\{|s_w(n+n0+L_s/2)|, n=0,1,\dots,L_{sr}-L_s/2-1\}}$$

where  $n0 = \text{trunc}\{m0 + \tau_{acc} + 0.5\}$  (here,  $m$  is subframe number and  $\tau_{acc}$  is the previous accumulated delay).

In order to find the best local delay,  $\tau_{opp}$ , at the end of the current processing subframe, a normalized correlation vector between the original weighted speech signal and the modified matching target is defined as:

$$R_t(k) = \frac{\sum_{n=0}^{L_{sr}-1} s_w(n0+n+k)\hat{s}_t(n)}{\sqrt{\sum_{n=0}^{L_{sr}-1} s_w^2(n0+n+k) \sum_{n=0}^{L_{sr}-1} \hat{s}_t^2(n)}}$$

A best local delay in the integer domain,  $k_{opp}$ , is selected by maximizing  $R_t(k)$  in the range of  $k \in [\text{SR0}, \text{SR1}]$ , which is corresponding to the real delay:

$$k_r = k_{opp} + n0 - m0 - \tau_{acc}$$

If  $R_f(k_{opt}) < 0.5$ ,  $k_r$  is set to zero.

In order to get a more precise local delay in the range  $\{k_r - 0.75 + 0.1j, j=0, 1, \dots, 15\}$  around  $k_r$ ,  $R_f(k)$  is interpolated to obtain the fractional correlation vector,  $R_f(j)$ , by:

$$R_f(j) = \sum_{i=0}^6 R_f(k_{opt} + I_j + i) I_f(i, j), \quad j = 0, 1, \dots, 15,$$

where  $\{I_f(i, j)\}$  is a set of interpolation coefficients. The optimal fractional delay index,  $j_{opt}$ , is selected by maximizing  $R_f(j)$ . Finally, the best local delay,  $\tau_{opt}$ , at the end of the current processing subframe, is given by,

$$\tau_{opt} = k_r - 0.75 + 0.1 j_{opt}$$

The local delay is then adjusted by:

$$\tau_{opt} = \begin{cases} 0, & \text{if } \tau_{acc} + \tau_{opt} > 14 \\ \tau_{opt}, & \text{otherwise} \end{cases}$$

The modified weighted speech of the current subframe, memorized in  $\{\hat{s}_w(m0+n), n=0, 1, \dots, L_s-1\}$  to update the buffer and produce the second target signal 253 for searching the fixed codebook 261, is generated by warping the original weighted speech  $\{s_w(n)\}$  from the original time region,

$$[m0 + \tau_{acc}, m0 + \tau_{acc} + L_s + \tau_{opt}]$$

to the modified time region,

$$[m0, m0 + L_s]$$

$$\hat{s}_w(m0+n) = \sum_{i=-j_{opt}}^{j_i} s_w(m0+n+T_w(n)+\delta) I_s(i, T_w(n)),$$

$$n = 0, 1, \dots, L_s - 1,$$

where  $T_w(n)$  and  $T_{fw}(n)$  are calculated by:

$$T_w(n) = \text{trunc}\{\tau_{acc} + n\tau_{opt}/L_s\},$$

$$T_{fw}(n) = \tau_{acc} + n\tau_{opt}/L_s - T_w(n),$$

$\{I_s(i, T_w(n))\}$  is a set of interpolation coefficients.

After having completed the modification of the weighted speech for the current subframe, the modified target weighted speech buffer is updated as follows:

$$\hat{s}_w(n) \leftarrow \hat{s}_w(n+L_s), \quad n=0, 1, \dots, n_m-1.$$

The accumulated delay at the end of the current subframe is renewed by:

$$\tau_{acc} \leftarrow \tau_{acc} + \tau_{opt}$$

Prior to quantization the LSFs are smoothed in order to improve the perceptual quality. In principle, no smoothing is applied during speech and segments with rapid variations in the spectral envelope. During non-speech with slow variations in the spectral envelope, smoothing is applied to reduce unwanted spectral variations. Unwanted spectral variations could typically occur due to the estimation of the LPC parameters and LSF quantization. As an example, in stationary noise-like signals with constant spectral envelope introducing even very small variations in the spectral envelope is picked up easily by the human ear and perceived as an annoying modulation.

The smoothing of the LSFs is done as a running mean according to:

$$lsf_i(n) = \beta(n) \cdot lsf_i(n-1) + (1-\beta(n)) \cdot lsf\_est_i(n), \quad i=1, \dots, 10$$

where  $lsf\_est_i(n)$  is the  $i^{th}$  estimated LSF of frame  $n$ , and  $lsf_i(n)$  is the  $i^{th}$  LSF for quantization of frame  $n$ . The parameter  $\beta(n)$  controls the amount of smoothing, e.g. if  $\beta(n)$  is zero no smoothing is applied.

$\beta(n)$  is calculated from the VAD information (generated at the block 235) and two estimates of the evolution of the spectral envelope. The two estimates of the evolution are defined as:

$$\Delta SP = \sum_{i=1}^{10} (lsf\_est_i(n) - lsf\_est_i(n-1))^2$$

$$\Delta SP_{int} = \sum_{i=1}^{10} (lsf\_est_i(n) - ma\_lsf_i(n-1))^2$$

$$ma\_lsf_i(n) = \beta(n) \cdot ma\_lsf_i(n-1) + (1-\beta(n)) \cdot lsf\_est_i(n), \quad i=1, \dots, 10$$

The parameter  $\beta(n)$  is controlled by the following logic:

Step 1:

if (Vad=1 | PastVad=1 |  $k_1 > 0.5$ )

$$N_{mode\_frm}(n-1) = 0$$

$$\beta(n) = 0.0$$

elseif ( $N_{mode\_frm}(n-1) > 0$  & ( $\Delta SP > 0.0015$  |  $\Delta SP_{int} > 0.0024$ ))

$$N_{mode\_frm}(n-1) = 0$$

$$\beta(n) = 0.0$$

elseif ( $N_{mode\_frm}(n-1) > 1$  &  $\Delta SP > 0.0025$ )

$$N_{mode\_frm}(n-1) = 1$$

endif

Step 2:

if (Vad=0 & PastVad=0)

$$N_{mode\_frm}(n) = N_{mode\_frm}(n-1) + 1$$

if ( $N_{mode\_frm}(n) > 5$ )

$$N_{mode\_frm}(n) = 5$$

endif

$$\beta(n) = 0.9^{1/5} \cdot (N_{mode\_frm}(n) - 1)^2$$

else

$$N_{mode\_frm}(n) = N_{mode\_frm}(n-1)$$

endif

where  $k_1$  is the first reflection coefficient.

In step 1, the encoder processing circuitry checks the VAD and the evolution of the spectral envelope, and performs a full or partial reset of the smoothing if required. In step 2, the encoder processing circuitry updates the counter,  $N_{mode\_frm}(n)$ , and calculates the smoothing parameter,  $\beta(n)$ . The parameter  $\beta(n)$  varies between 0.0 and 0.9, being 0.0 for speech, music, tonal-like signals, and non-stationary background noise and ramping up towards 0.9 when stationary background noise occurs.

The LSFs are quantized once per 20 ms frame using a predictive multi-stage vector quantization. A minimal spacing of 50 Hz is ensured between each two neighboring LSFs before quantization. A set of weights is calculated from the LSFs, given by  $w_i = K |P(f_i)|^{0.4}$  where  $f_i$  is the  $i^{th}$  LSF value and  $P(f_i)$  is the LPC power spectrum at  $f_i$  ( $K$  is an irrelevant multiplicative constant). The reciprocal of the power spectrum is obtained by (up to a multiplicative constant):

$$P(f_i)^{-1} = \begin{cases} (1 - \cos(2\pi f_i)) \prod_{\text{odd } j} [\cos(2\pi f_j) - \cos(2\pi f_j)]^2 & \text{even } i \\ (1 + \cos(2\pi f_i)) \prod_{\text{even } j} [\cos(2\pi f_j) - \cos(2\pi f_j)]^2 & \text{odd } i \end{cases}$$

and the power of -0.4 is then calculated using a lookup table and cubic-spline interpolation between table entries.

A vector of mean values is subtracted from the LSFs, and a vector of prediction error vector fe is calculated from the mean removed LSFs vector, using a full-matrix AR(2) predictor. A single predictor is used for the rates 5.8, 6.65, 8.0, and 11.0 kbps coders, and two sets of prediction coefficients are tested as possible predictors for the 4.55 kbps coder.

The vector of prediction error is quantized using a multi-stage VQ, with multi-surviving candidates from each stage to the next stage. The two possible sets of prediction error vectors generated for the 4.55 kbps coder are considered as surviving candidates for the first stage.

The first 4 stages have 64 entries each, and the fifth and last table have 16 entries. The first 3 stages are used for the 4.55 kbps coder, the first 4 stages are used for the 5.8, 6.65 and 8.0 kbps coders, and all 5 stages are used for the 11.0 kbps coder. The following table summarizes the number of bits used for the quantization of the LSFs for each rate.

	prediction	1 <sup>st</sup> stage	2 <sup>nd</sup> stage	3 <sup>rd</sup> stage	4 <sup>th</sup> stage	5 <sup>th</sup> stage	total
4.55 kbps	1	6	6	6			19
5.8 kbps	0	6	6	6	6		24
6.65 kbps	0	6	6	6	6		24
8.0 kbps	0	6	6	6	6		24
11.0 kbps	0	6	6	6	6	4	28

The number of surviving candidates for each stage is summarized in the following table.

	prediction candidates into the 1 <sup>st</sup> stage	Surviving candidates from the 1 <sup>st</sup> stage	surviving candidates from the 2 <sup>nd</sup> stage	surviving candidates from the 3 <sup>rd</sup> stage	surviving candidates from the 4 <sup>th</sup> stage	
4.55 kbps	2	10	6	4		
5.8 kbps	1	8	6	4		
6.65 kbps	1	8	8	4		
8.0 kbps	1	8	8	4		
11.0 kbps	1	8	6	4	4	

The quantization in each stage is done by minimizing the weighted distortion measure given by:

$$e_k = \sum_{i=0}^9 (w_i (f_{e_i} - C_i^k))^2$$

The code vector with index  $k_{min}$  which minimizes  $e_k$  such that  $e_{k_{min}} < e_k$  for all  $k$ , is chosen to represent the prediction/quantization error (fe represents in this equation both the initial prediction error to the first stage and the successive quantization error from each stage to the next one).

The final choice of vectors from all of the surviving candidates (and for the 4.55 kbps coder—also the predictor) is done at the end, after the last stage is searched, by

choosing a combined set of vectors (and predictor) which minimizes the total error. The contribution from all of the stages is summed to form the quantized prediction error vector, and the quantized prediction error is added to the prediction states and the mean LSFs value to generate the quantized LSFs vector.

For the 4.55 kbps coder, the number of order flips of the LSFs as the result of the quantization if counted, and if the number of flips is more than 1, the LSFs vector is replaced with  $0.9 \cdot (\text{LSFs of previous frame}) + 0.1 \cdot (\text{mean LSFs value})$ . For all the rates, the quantized LSFs are ordered and spaced with a minimal spacing of 50 Hz.

The interpolation of the quantized LSF is performed in the cosine domain in two ways depending on the LTP\_mode. If the LTP\_mode is 0, a linear interpolation between the quantized LSF set of the current frame and the quantized LSF set of the previous frame is performed to get the LSF set for the first, second and third subframes as:

$$\begin{aligned} \bar{q}_1(n) &= 0.75\bar{q}_a(n-1) + 0.25\bar{q}_a(n) \\ \bar{q}_2(n) &= 0.5\bar{q}_a(n-1) + 0.5\bar{q}_a(n) \\ \bar{q}_3(n) &= 0.25\bar{q}_a(n-1) + 0.75\bar{q}_a(n) \end{aligned}$$

where  $\bar{q}_a(n-1)$  and  $\bar{q}_a(n)$  are the cosines of the quantized LSF sets of the previous and current frames, respectively, and  $\bar{q}_1(n)$ ,  $\bar{q}_2(n)$  and  $\bar{q}_3(n)$  are the interpolated LSF sets in cosine domain for the first, second and third subframes respectively.

If the LTP\_mode is 1, a search of the best interpolation path is performed in order to get the interpolated LSF sets. The search is based on a weighted mean absolute difference between a reference LSF set  $\bar{r}(n)$  and the LSF set obtained from LP analysis  $\bar{l}(n)$ . The weights  $\bar{w}$  are computed as follows:

$$\begin{aligned} w(0) &= (1-l(0))(1-l(1)+l(0)) \\ w(9) &= (1-l(9))(1-l(9)+l(8)) \end{aligned}$$

for  $i=1$  to 9

$$w(i) = (1-l(i))(1 - \text{Min}(l(i+1)-l(i), l(i)-l(i-1)))$$

where  $\text{Min}(a,b)$  returns the smallest of  $a$  and  $b$ .

There are four different interpolation paths. For each path, a reference LSF set  $\bar{r}q(n)$  in cosine domain is obtained as follows:

$$\bar{r}q(n) = \alpha(k)\bar{q}_a(n) + (1-\alpha(k))\bar{q}_a(n-1), k=1 \text{ to } 4$$

$\bar{\alpha} = \{0.4, 0.5, 0.6, 0.7\}$  for each path respectively. Then the following distance measure is computed for each path as:

$$D = |\bar{r}q(n) - \bar{l}(n)|^T \bar{w}$$

The path leading to the minimum distance  $D$  is chosen and the corresponding reference LSF set  $\bar{r}q(n)$  is obtained as:

$$\bar{r}q(n) = \alpha_{opt}\bar{q}_a(n) + (1-\alpha_{opt})\bar{q}_a(n-1)$$

The interpolated LSF sets in the cosine domain are then given by:

$$\begin{aligned} \bar{q}_1(n) &= 0.5\bar{q}_a(n-1) + 0.5\bar{r}q(n) \\ \bar{q}_2(n) &= \bar{r}q(n) \\ \bar{q}_3(n) &= 0.5\bar{r}q(n) + 0.5\bar{q}_a(n) \end{aligned}$$

The impulse response,  $h(n)$ , of the weighted synthesis filter  $H(z)W(z) = A(z/\gamma_1)/[A(z)A(z/\gamma_2)]$  is computed each

subframe. This impulse response is needed for the search of adaptive and fixed codebooks 257 and 261. The impulse response  $b(n)$  is computed by filtering the vector of coefficients of the filter  $A(z/\gamma_1)$  extended by zeros through the two filters  $1/\bar{A}(z)$  and  $1/A(z/\gamma_2)$ . The target signal for the search of the adaptive codebook 257 is usually computed by subtracting the zero input response of the weighted synthesis filter  $H(z)W(z)$  from the weighted speech signal  $s_w(n)$ . This operation is performed on a frame basis. An equivalent procedure for computing the target signal is the filtering of the LP residual signal  $r(n)$  through the combination of the synthesis filter  $1/\bar{A}(z)$  and the weighting filter  $W(z)$ .

After determining the excitation for the subframe, the initial states of these filters are updated by filtering the difference between the LP residual and the excitation. The LP residual is given by:

$$r(n) = s(n) + \sum_{i=1}^{10} \bar{a}_i s(n-i), n = 0, L\_SF - 1$$

The residual signal  $r(n)$  which is needed for finding the target vector is also used in the adaptive codebook search to extend the past excitation buffer. This simplifies the adaptive codebook search procedure for delays less than the subframe size of 40 samples.

In the present embodiment, there are two ways to produce an LTP contribution. One uses pitch preprocessing (PP) when the PP-mode is selected, and another is computed like the traditional LTP when the LTP-mode is chosen. With the PP-mode, there is no need to do the adaptive codebook search, and LTP excitation is directly computed according to past synthesized excitation because the interpolated pitch contour is set for each frame. When the AMR coder operates with LTP-mode, the pitch lag is constant within one subframe, and searched and coded on a subframe basis.

Suppose the past synthesized excitation is memorized in  $\{ext(MAX\_LAG+n), n < 0\}$ , which is also called adaptive codebook. The LTP excitation codevector, temporally memorized in  $\{ext(MAX\_LAG+n), 0 \leq n < L\_SF\}$ , is calculated by interpolating the past excitation (adaptive codebook) with the pitch lag contour,  $\tau_c(n+m \cdot L\_SF)$ ,  $m=0, 1, 2, 3$ . The interpolation is performed using an FIR filter (Hamming windowed sinc functions):

$$ext(MAX\_LAG+n) = \sum_{i=-1}^4 ext(MAX\_LAG+n - T_c(n) + i) \cdot f_i(i, T_c(n)), n = 0, 1, \dots, L\_SF - 1, \dots$$

where  $T_c(n)$  and  $T_{IC}(n)$  are calculated by

$$T_c(n) = \text{trunc}\{\tau_c(n+m \cdot L\_SF)\},$$

$$T_{IC}(n) = \tau_c(n) - T_c(n),$$

$m$  is subframe number,  $\{f_i(i, T_{IC}(n))\}$  is a set of interpolation coefficients,  $f_1$  is 10,  $MAX\_LAG$  is 145+11, and  $L\_SF=40$  is the subframe size. Note that the interpolated values  $\{ext(MAX\_LAG+n), 0 \leq n < L\_SF-17+11\}$  might be used again to do the interpolation when the pitch lag is small. Once the interpolation is finished, the adaptive codevector  $V_a = \{v_a(n), n=0$  to 39 $\}$  is obtained by copying the interpolated values:

$$v_a(n) = ext(MAX\_LAG+n), 0 \leq n < L\_SF$$

Adaptive codebook searching is performed on a subframe basis. It consists of performing closed-loop pitch lag search,

and then computing the adaptive code vector by interpolating the so past excitation at the selected fractional pitch lag. The LTP parameters (or the adaptive codebook parameters) are the pitch lag (or the delay) and gain of the pitch filter. In the search stage, the excitation is extended by the LP residual to simplify the closed-loop search.

For the bit rate of 11.0 kbps, the pitch delay is encoded with 9 bits for the 1<sup>st</sup> and 3<sup>rd</sup> subframes and the relative delay of the other subframes is encoded with 6 bits. A fractional pitch delay is used in the first and third subframes with resolutions:  $\frac{1}{6}$  in the range

$$\left[17, 93 \frac{4}{6}\right]$$

and integers only in the range [95,145]. For the second and fourth subframes, a pitch resolution of  $\frac{1}{6}$  is always used for the rate 11.0 kbps in the range

$$\left[T_1 - 5 \frac{3}{6}, T_1 + 4 \frac{3}{6}\right]$$

where  $T_1$  is the pitch lag of the previous (1<sup>st</sup> or 3<sup>rd</sup>) subframe.

The close-loop pitch search is performed by minimizing the mean-square weighted error between the original and synthesized speech. This is achieved by maximizing the term:

$$R(k) = \frac{\sum_{n=0}^{39} T_{gs}(n) y_k(n)}{\sqrt{\sum_{n=0}^{39} y_k(n) y_k(n)}}$$

where  $T_{gs}(n)$  is the target signal and  $y_k(n)$  is the past filtered excitation at delay  $k$  (past excitation convoluted with  $h(n)$ ). The convolution  $y_k(n)$  is computed for the first delay  $t_{min}$  in the search range, and for the other delays in the search range  $k=t_{min}+1, \dots, t_{max}$  it is updated using the recursive relation:

$$y_k(n) = y_{k-1}(n-1) + u(n)h(n),$$

where  $u(n), n=-143+11$  to 39 is the excitation buffer.

Note that in the search stage, the samples  $u(n), n=0$  to 39, are not available and are needed for pitch delays less than 40. To simplify the search, the LP residual is copied to  $u(n)$  to make the relation in the calculations valid for all delays. Once the optimum integer pitch delay is determined, the fractions, as defined above, around that integer are tested. The fractional pitch search is performed by interpolating the normalized correlation and searching for its maximum.

Once the fractional pitch lag is determined, the adaptive codebook vector,  $v(n)$ , is computed by interpolating the past excitation  $u(n)$  at the given phase (fraction). The interpolations are performed using two FIR filters (Hamming windowed sinc functions), one for interpolating the term in the calculations to find the fractional pitch lag and the other for interpolating the past excitation as previously described. The adaptive codebook gain,  $g_p$ , is temporally given then by:



$$g_p = \frac{\sum_{n=0}^{39} T_{g_p}(n)y(n)}{\sum_{n=0}^{39} y(n)y(n)}$$

bounded by  $0 < g_p < 1.2$ , where  $y(n) = v(n) * h(n)$  is the filtered adaptive codebook vector (zero state response of  $H(z)W(z)$  to  $v(n)$ ). The adaptive codebook gain could be modified again due to joint optimization of the gains, gain normalization and smoothing. The term  $y(n)$  is also referred to herein as  $C_p(n)$ .

With conventional approaches, pitch lag maximizing correlation might result in two or more times the correct one. Thus, with such conventional approaches, the candidate of shorter pitch lag is favored by weighting the correlations of different candidates with constant weighting coefficients. At times this approach does not correct the double or treble pitch lag because the weighting coefficients are not aggressive enough or could result in halving the pitch lag due to the strong weighting coefficients.

In the present embodiment, these weighting coefficients become adaptive by checking if the present candidate is in the neighborhood of the previous pitch lags (when the previous frames are voiced) and if the candidate of shorter lag is in the neighborhood of the value obtained by dividing the longer lag (which maximizes the correlation) with an integer.

In order to improve the perceptual quality, a speech classifier is used to direct the searching procedure of the fixed codebook (as indicated by the blocks 275 and 279) and to-control gain normalization (as indicated in the block 401 of FIG. 4). The speech classifier serves to improve the background noise performance for the lower rate coders, and to get a quick start-up of the noise level estimation. The speech classifier distinguishes stationary noise-like segments from segments of speech, music, tonal-like signals, non-stationary noise, etc.

The speech classification is performed in two steps. An initial classification (`speech_mode`) is obtained based on the modified input signal. The final classification (`exc_mode`) is obtained from the initial classification and the residual signal after the pitch contribution has been removed. The two outputs from the speech classification are the excitation mode, `exc_mode`, and the parameter  $\beta_{sub}(n)$ , used to control the subframe based smoothing of the gains.

The speech classification is used to direct the encoder according to the characteristics of the input signal and need not be transmitted to the decoder. Thus, the bit allocation, codebooks, and decoding remain the same regardless of the classification. The encoder emphasizes the perceptually important features of the input signal on a subframe basis by adapting the encoding in response to such features. It is important to notice that misclassification will not result in disastrous speech quality degradations. Thus, as opposed to the VAD 235, the speech classifier identified within the block 279 (FIG. 2) is designed to be somewhat more aggressive for optimal perceptual quality.

The initial classifier (`speech_classifier`) has adaptive thresholds and is performed in six steps:

1. Adapt thresholds:

if (`updates_noise`  $\geq 30$  & `updates_speech`  $\geq 30$ )

$$SNR\_max = \min\left(\frac{ma\_max\_speech}{ma\_max\_noise}, 32\right)$$

else  
`SNR_max` = 3.5  
 endif

if (`SNR_max` < 1.75)  
`dec_i_max_mes` = 1.30  
`dec_i_ma_cp` = 0.70  
`update_max_mes` = 1.10  
`update_ma_cp_speech` = 0.72  
 elseif (`SNR_max` < 2.50)  
`dec_i_max_mes` = 1.65  
`dec_i_ma_cp` = 0.73  
`update_max_mes` = 1.30  
`update_ma_cp_speech` = 0.72

else  
`dec_i_max_mes` = 1.75  
`dec_i_ma_cp` = 0.77  
`update_max_mes` = 1.30  
`update_ma_cp_speech` = 0.77  
 endif

2. Calculate parameters:

Pitch correlation:

$$cp = \frac{\sum_{i=0}^{L\_SF-1} \tilde{s}(i) \cdot \tilde{s}(i-lag)}{\sqrt{\left(\sum_{i=0}^{L\_SF-1} \tilde{s}(i) \cdot \tilde{s}(i)\right) \cdot \left(\sum_{i=0}^{L\_SF-1} \tilde{s}(i-lag) \cdot \tilde{s}(i-lag)\right)}}$$

Running mean of pitch correlation:

$$ma\_cp(n) = 0.9 \cdot ma\_cp(n-1) + 0.1 \cdot cp$$

Maximum of signal amplitude in current pitch cycle:

$$max(n) = \max\{|\tilde{s}(i)|, i=start, \dots, L\_SF-1\}$$

where:

$$start = \min\{L\_SF-lag, 0\}$$

Sum of signal amplitudes in current pitch cycle:

$$mean(n) = \sum_{i=start}^{L\_SF-1} |\tilde{s}(i)|$$

Measure of relative maximum:

$$max\_mes = \frac{max(n)}{ma\_max\_noise(n-1)}$$

Maximum to long-term sum:

$$\text{max2sum} = \frac{\text{max}(n)}{\sum_{k=1}^{14} \text{mean}(n-k)}$$

Maximum in groups of 3 subframes for past 15 subframes:

$$\text{max\_group}(n,k) = \text{max}\{\text{max}(n-3-(4-k)-j), j=0, \dots, 2\}, k=0, \dots, 4$$

Group-maximum to minimum of previous 4 group-maxima:

$$\text{endmax2minmax} = \frac{\text{max\_group}(n,4)}{\min\{\text{max\_group}(n,k), k=0, \dots, 3\}}$$

Slope of 5 group maxima:

$$\text{slope} = 0.1 \cdot \sum_{k=0}^4 (k-2) \cdot \text{max\_group}(n, k)$$

3. Classify subframe:

if (((max\_mes < deci\_max\_mes & ma\_cp < deci\_ma\_cp) & (VAD=0)) & (LTP\_MODE=1|5.8 kbit/s|4.55 kbit/s))  
 speech\_mode=0/\* class 1 \*/

else  
 speech\_mode=1/\* class 2 \*/  
 endif

4. Check for change in background noise level, i.e. reset required:

Check for decrease in level:  
 if (updates\_noise=31 & max\_mes <= 0.3)  
 if (consec\_low < 15)  
 consec\_low++  
 endif

else  
 consec\_low=0  
 endif

if (consec\_low=15)  
 updates\_noise=0  
 lev\_reset=-1/\* low level reset \*/  
 endif

Check for increase in level:

if ((updates\_noise >= 30 | lev\_reset=-1) & max\_mes > 1.5 & ma\_cp < 0.70 & cp < 0.85 & k1 < -0.4 & endmax2minmax < 50 & max2sum < 35 & slope > -100 & slope < 120)  
 if (consec\_high < 15)  
 consec\_high++  
 endif

else  
 consec\_high=0  
 endif

if (consec\_high=15 & endmax2minmax < 6 & max2sum < 5)  
 updates\_noise=30  
 lev\_reset=1/\* high level reset \*/  
 endif

5. Update running mean of maximum of class 1 segments, i.e. stationary noise: if (

/\* 1. condition: regular update \*/  
 (max\_mes < update\_max\_mes & ma\_cp < 0.6 & cp < 0.65 & max\_mes > 0.3) 5  
 /\* 2. condition: VAD continued update \*/  
 (consec\_vad\_0=8) 5  
 /\* 3. condition: start-up/reset update \*/  
 (updates\_noise <= 30 & ma\_cp < 0.7 & cp < 0.75 & k1 < -0.4 & endmax2minmax < 5 & (lev\_reset=-1 | (lev\_reset=-1 & max\_mes < 2)))

ma\_max\_noise(n) = 0.9 · ma\_max\_noise(n-1) + 0.1 · max(n)

if (updates\_noise <= 30)  
 updates\_noise++  
 else  
 lev\_reset=0  
 endif 20

where k<sub>1</sub> is the first reflection coefficient.

6. Update running mean of maximum of class 2 segments, i.e. speech, music, tonal-like signals, non-stationary noise, etc, continued from above:

elseif (ma\_cp > update\_ma\_cp\_speech)  
 if (updates\_speech <= 80)  
 α<sub>speech</sub> = 0.95  
 else  
 α<sub>speech</sub> = 0.999  
 endif 30

ma\_max\_speech(n) = α<sub>speech</sub> · ma\_max\_speech(n-1) + (1 - α<sub>speech</sub>) · max(n) 35

if (updates\_speech <= 80)  
 updates\_speech++  
 endif

40 The final classifier (exc\_preselect) provides the final class, exc\_mode, and the subframe based smoothing parameter, β<sub>sub</sub>(n). It has three steps:

1. Calculate parameters:

45 Maximum amplitude of ideal excitation in current subframe:

$$\text{max}_{res2}(n) = \text{max}\{|\text{res2}(i)|, i=0, \dots, L_{SF}-1\}$$

Measure of relative maximum:

$$\text{max\_mes}_{res2} = \text{max}_{res2}(n) / \text{ma\_max}_{res2}(n-1)$$

2. Classify subframe and calculate smoothing:

if (speech\_mode=1 | max\_mes<sub>res2</sub> >= 1.75)  
 exc\_mode=1/\* class 2 \*/  
 β<sub>sub</sub>(n)=0  
 N\_mode\_sub(n)=-4 55

else  
 exc\_mode=0/\* class 1 \*/  
 N\_mode\_sub(n) = N\_mode\_sub(n-1) + 1  
 if (N\_mode\_sub(n) > 4)  
 N\_mode\_sub(n) = 4  
 endif  
 if (N\_mode\_sub(n) > 0)  
 β<sub>sub</sub>(n) = 0.7/9 · (N\_mode\_sub(n)-1)<sup>2</sup>  
 else  
 β<sub>sub</sub>(n) = 0 65

```

endif
endif
3. Update running mean of maximum:
if (max_mesres2 ≤ 0.5)
  if (consec < 51)
    consec++
  endif
else
  consec=0
endif

if ((exc_mode=0 & (max_mesres2>0.5[consec>50]))
  (updates ≤ 30 & ma_cp < 0.6 & cp < 0.65))
  ma_max(n)=0.9*ma_max(n-1)+0.1*maxres2(n)
  if (updates ≤ 30)
    updates++
  endif
endif
endif
    
```

When this process is completed, the final subframe based classification, exc\_mode, and the smoothing parameter, β<sub>sm</sub>(n), are available.

To enhance the quality of the search of the fixed codebook 261, the target signal, T<sub>g</sub>(n), is produced by temporally reducing the LTP contribution with a gain factor, G<sub>r</sub>:

$$T_g(n) = T_{gs}(n) - G_r * g_p * Y_a(n), \quad n=0,1, \dots, 39$$

where T<sub>gs</sub>(n) is the original target signal 253, Y<sub>a</sub>(n) is the filtered signal from the adaptive codebook, g<sub>p</sub> is the LTP gain for the selected adaptive codebook vector, and the gain factor is determined according to the normalized LTP gain, R<sub>p</sub>, and the bit rate:

```

if (rate <= 0) /* for 4.45 kbps and 5.8 kbps */
  G_r = 0.7 R_p + 0.3;
if (rate == 1) /* for 6.65 kbps */
  G_r = 0.6 R_p + 0.4;
if (rate == 2) /* for 8.0 kbps */
  G_r = 0.3 R_p + 0.7;

if (rate == 3) /* for 11.0 kbps */
  G_r = 0.95;

if (T_op > L_SF & g_p > 0.5 & rate <= 2)
  G_r = G_r * (0.3 * R_p + 0.7); and
    
```

where normalized LTP gain, R<sub>p</sub>, is defined as:

$$R_p = \frac{\sum_{n=0}^{39} T_{gs}(n) Y_a(n)}{\sqrt{\sum_{n=0}^{39} T_{gs}(n) T_{gs}(n)} \sqrt{\sum_{n=0}^{39} Y_a(n) Y_a(n)}}$$

Another factor considered at the control block 275 in conducting the fixed codebook search and at the block 401 (FIG. 4) during gain normalization is the noise level+“” which is given by:

$$P_{NSR} = \sqrt{\frac{\max(E_n - 100, 0.0)}{E_s}}$$

where E<sub>s</sub> is the energy of the current input signal including background noise, and E<sub>n</sub> is a running average energy of the

background noise. E<sub>n</sub> is updated only when the input signal is detected to be background noise as follows:

if (first background noise frame is true)

$$E_n = 0.75 E_s;$$

else if (background noise frame is true)

$$E_n = 0.75 E_{n-m} + 0.25 E_s;$$

where E<sub>n-m</sub> is the last estimation of the background noise energy.

For each bit rate mode, the fixed codebook 261 (FIG. 2) consists of two or more subcodebooks which are constructed with different structure. For example, in the present embodiment at higher rates, all the subcodebooks only contain pulses. At lower bit rates, one of the subcodebooks is populated with Gaussian noise. For the lower bit-rates (e.g., 6.65, 5.8, 4.55 kbps), the speech classifier forces the encoder to choose from the Gaussian subcodebook in case of stationary noise-like subframes, exc\_mode=0. For exc\_mode=1 all subcodebooks are searched using adaptive weighting.

For the pulse subcodebooks, a fast searching approach is used to choose a subcodebook and select the code word for the current subframe. The same searching routine is used for all the bit rate modes with different input parameters.

In particular, the long-term enhancement filter, F<sub>p</sub>(z), is used to filter through the selected pulse excitation. The filter is defined as F<sub>p</sub>(z)=1/(1-βz<sup>-T</sup>), where T is the integer part of pitch lag at the center of the current subframe, and β is the pitch gain of previous subframe, bounded by [0.2, 1.0]. Prior to the codebook search, the impulsive response h(n) includes the filter F<sub>p</sub>(z).

For the Gaussian subcodebooks, a special structure is used in order to bring down the storage requirement and the computational complexity. Furthermore, no pitch enhancement is applied to the Gaussian subcodebooks.

There are two kinds of pulse subcodebooks in the present AMR coder embodiment. All pulses have the amplitudes of +1 or -1. Each pulse has 0, 1, 2, 3 or 4 bits to code the pulse position. The signs of some pulses are transmitted to the decoder with one bit coding one sign. The signs of other pulses are determined in a way related to the coded signs and their pulse positions.

In the first kind of pulse subcodebook, each pulse has 3 or 4 bits to code the pulse position. The possible locations of individual pulses are defined by two basic non-regular tracks and initial phases:

$$POS(n_p, i) = TRACK(m_p, i) + PHAS(n_p, \text{phas\_mode}),$$

where i=0,1, . . . , 7 or 15 (corresponding to 3 or 4 bits to code the position), is the possible position index, n<sub>p</sub>=0, . . . , N<sub>p</sub>-1 (N<sub>p</sub> is the total number of pulses), distinguishes different pulses, m<sub>p</sub>=0 or 1, defines two tracks, and phas\_mode=0 or 1, specifies two phase modes.

For 3 bits to code the pulse position, the two basic tracks are:

$$\{ TRACK(0, i) \} = \{ 0, 4, 8, 12, 18, 24, 30, 36 \},$$

and

$$\{ TRACK(1, i) \} = \{ 0, 6, 12, 18, 22, 26, 30, 34 \}.$$

If the position of each pulse is coded with 4 bits, the basic tracks are:

$$\{ \text{TRACK}(0, l) \} = \{ 0, 2, 4, 6, 8, 10, 12, 14, 17, 20, 23, 26, 29, 32, 35, 38 \},$$

and

$$\{ \text{TRACK}(1, l) \} = \{ 0, 3, 6, 9, 12, 15, 18, 21, 23, 25, 27, 29, 31, 33, 35, 37 \}.$$

The initial phase of each pulse is fixed as:

$$\text{PHAS}(n_p, 0) = \text{modulus}(n_p / \text{MAXPHAS})$$

$$\text{PHAS}(n_p, 1) = \text{PHAS}(N_p - 1 - n_p, 0)$$

where MAXPHAS is the maximum phase value.

For any pulse subcodebook, at least the first sign for the first pulse,  $\text{SIGN}(n_p)$ ,  $n_p=0$ , is encoded because the gain sign is embedded. Suppose  $N_{\text{sign}}$  is the number of pulses with encoded signs; that is,  $\text{SIGN}(n_p)$ , for  $n_p < N_{\text{sign}}$ , is encoded while  $\text{SIGN}(n_p)$ , for  $n_p \geq N_{\text{sign}}$ , is not encoded. Generally, all the signs can be determined in the following way:

$$\text{SIGN}(n_p) = -\text{SIGN}(n_p - 1), \text{ for } n_p \geq N_{\text{sign}}$$

due to that the pulse positions are sequentially searched from  $n_p=0$  to  $n_p=N_p-1$  using an iteration approach. If two pulses are located in the same track while only the sign of the first pulse in the track is encoded, the sign of the second pulse depends on its position relative to the first pulse. If the position of the second pulse is smaller, then it has opposite sign, otherwise it has the same sign as the first pulse.

In the second kind of pulse subcodebook, the innovation vector contains 10 signed pulses. Each pulse has 0, 1, or 2 bits to code the pulse position. One subframe with the size of 40 samples is divided into 10 small segments with the length of 4 samples. 10 pulses are respectively located into 10 segments. Since the position of each pulse is limited into one segment, the possible locations for the pulse numbered with  $n_p$  are,  $\{4n_p\}$ ,  $\{4n_p, 4n_p+256\}$ , or  $\{4n_p, 4n_p+1, 4n_p+2, 4n_p+3\}$ , respectively for 0, 1, or 2 bits to code the pulse position. All the signs for all the 10 pulses are encoded.

The fixed codebook 261 is searched by minimizing the mean square error between the weighted input speech and the weighted synthesized speech. The target signal used for the LTP excitation is updated by subtracting the adaptive codebook contribution. That is:

$$x_2(n) = x(n) - \hat{g}_p y(n), \quad n=0, \dots, 39,$$

where  $y(n) = v(n) * b(n)$  is the filtered adaptive codebook vector and  $\hat{g}_p$  is the modified (reduced) LTP gain.

If  $c_k$  is the code vector at index  $k$  from the fixed codebook, then the pulse codebook is searched by maximizing the term:

$$A_k = \frac{(C_k)^2}{E_{D_k}} = \frac{(d^T c_k)^2}{c_k^T \Phi c_k},$$

where  $d = H^T x_2$  is the correlation between the target signal  $x_2(n)$  and the impulse response  $h(n)$ ,  $H$  is a the lower triangular Toeplitz convolution matrix with diagonal  $h(0)$  and lower diagonals  $h(1), \dots, h(39)$ , and  $\Phi = H^T H$  is the matrix of correlations of  $h(n)$ . The vector  $d$  (backward filtered target) and the matrix  $\Phi$  are computed prior to the codebook search. The elements of the vector  $d$  are computed by:

$$d(n) = \sum_{i=0}^{39} x_2(i) h(i-n), \quad n=0, \dots, 39,$$

and the elements of the symmetric matrix  $\Phi$  are computed by:

$$\phi(i, j) = \sum_{n=j}^{39} h(n-i) h(n-j), \quad (j \geq i).$$

The correlation in the numerator is given by:

$$C = \sum_{i=0}^{N_p-1} \delta_i d(m_i),$$

where  $m_i$  is the position of the  $i$ th pulse and  $v_i$  is its amplitude. For the complexity reason, all the amplitudes  $\{v_i\}$  are set to +1 or -1; that is,

$$v_i = \text{SIGN}(i), \quad i=0, \dots, N_p-1.$$

The energy in the denominator is given by:

$$E_D = \sum_{i=0}^{N_p-1} \phi(m_i, m_i) + 2 \sum_{i=0}^{N_p-2} \sum_{j=i+1}^{N_p-1} \delta_i \delta_j \phi(m_i, m_j).$$

To simplify the search procedure, the pulse signs are preset by using the signal  $b(n)$ , which is a weighted sum of the normalized  $d(n)$  vector and the normalized target signal of  $x_2(n)$  in the residual domain  $\text{res}_2(n)$ :

$$b(n) = \frac{\text{res}_2(n)}{\sqrt{\sum_{i=0}^{39} \text{res}_2(i) \text{res}_2(i)}} + \frac{2d(n)}{\sqrt{\sum_{i=0}^{39} d(i) d(i)}}, \quad n=0, 1, \dots, 39$$

If the sign of the  $i$ th ( $i=n_p$ ) pulse located at  $m_i$  is encoded, it is set to the sign of signal  $b(n)$  at that position, i.e.,  $\text{SIGN}(i) = \text{sign}[b(m_i)]$ .

In the present embodiment, the fixed codebook 261 has 2 or 3 subcodebooks for each of the encoding bit rates. Of course many more might be used in other embodiments. Even with several subcodebooks, however, the searching of the fixed codebook 261 is very fast using the following procedure. In a first searching turn, the encoder processing circuitry searches the pulse positions sequentially from the first pulse ( $n_p=0$ ) to the last pulse ( $n_p=N_p-1$ ) by considering the influence of all the existing pulses.

In a second searching turn, the encoder processing circuitry corrects each pulse position sequentially from the first pulse to the last pulse by checking the criterion value  $A_k$  contributed from all the pulses for all possible locations of the current pulse. In a third turn, the functionality of the second searching turn is repeated a final time. Of course further turns may be utilized if the added complexity is not prohibitive.

The above searching approach proves very efficient, because only one position of one pulse is changed leading to changes in only one term in the criterion numerator  $C$  and few terms in the criterion denominator  $E_D$  for each computation of the  $A_k$ . As an example, suppose a pulse subcodebook is constructed with 4 pulses and 3 bits per pulse to encode the position. Only 96 ( $4 \text{ pulses} \times 2^3 \text{ positions per}$

pulse×3turns=96) simplified computations of the criterion  $A_k$  need be performed.

Moreover, to save the complexity, usually one of the subcodebooks in the fixed codebook 261 is chosen after finishing the first searching turn. Further searching turns are done only with the chosen subcodebook. In other embodiments, one of the subcodebooks might be chosen only after the second searching turn or thereafter should processing resources so permit.

The Gaussian codebook is structured to reduce the storage requirement and the computational complexity. A comb-structure with two basis vectors is used. In the comb-structure, the basis vectors are orthogonal, facilitating a low complexity search. In the AMR coder, the first basis vector occupies the even sample positions, (0,2, . . . ,38), and the second basis vector occupies the odd sample positions, (1,3, . . . ,39).

The same codebook is used for both basis vectors, and the length of the codebook vectors is 20 samples (half the subframe size).

All rates (6.65, 5.8 and 4.55 kbps) use the same Gaussian codebook. The Gaussian codebook,  $CB_{Gauss}$ , has only 10 entries, and thus the storage requirement is 10·20=200 16-bit words. From the 10 entries, as many as 32 code vectors are generated. An index,  $idx_g$ , to one basis vector 22 populates the corresponding part of a code vector,  $c_{idx_g}$ , in the following way:

$$c_{idx_g}(2 \cdot (i-\tau)+\delta) = CB_{Gauss}(l, i) \quad i = \tau, \tau+1, \dots, 19$$

$$c_{idx_g}(2 \cdot (i+20-\tau)+\delta) = CB_{Gauss}(l, i) \quad i = 0, 1, \dots, \tau-1$$

where the table entry,  $l$ , and the shift,  $\tau$ , are calculated from the index,  $idx_g$ , according to:

$$\tau = \text{trunc}\{idx_g/10\}$$

$$l = idx_g - 10\tau$$

and  $\delta$  is 0 for the first basis vector and 1 for the second basis vector. In addition, a sign is applied to each basis vector.

Basically, each entry in the Gaussian table can produce as many as 20 unique vectors, all with the same energy due to the circular shift. The 10 entries are all normalized to have identical energy of 0.5, i.e.,

$$\sum_{i=0}^{19} CB_{Gauss}(l, i)^2 = 0.5, \quad l = 0, 1, \dots, 9$$

That means that when both basis vectors have been selected, the combined code vector,  $c_{idx_0 idx_1}$ , will have unity energy, and thus the final excitation vector from the Gaussian subcodebook will have unity energy since no pitch enhancement is applied to candidate vectors from the Gaussian subcodebook.

The search of the Gaussian codebook utilizes the structure of the codebook to facilitate a low complexity search. Initially, the candidates for the two basis vectors are searched independently based on the ideal excitation,  $res_2$ . For each basis vector, the two best candidates, along with the respective signs, are found according to the mean squared error. This is exemplified by the equations to find the best candidate, index  $idx_g$ , and its sign,  $s_{idx_g}$ :

$$idx_g = \max_{k=0, \dots, N_{Gauss}} \left\{ \sum_{i=0}^{19} res_2(2 \cdot i + \delta) \cdot c_k(2 \cdot i + \delta) \right\}$$

$$s_{idx_g} = \text{sign} \left( \sum_{i=0}^{19} res_2(2 \cdot i + \delta) \cdot c_{idx_g}(2 \cdot i + \delta) \right)$$

where  $N_{Gauss}$  is the number of candidate entries for the basis vector. The remaining parameters are explained above. The total number of entries in the Gaussian codebook is  $2 \cdot 2 \cdot N_{Gauss}^2$ . The fine search minimizes the error between the weighted speech and the weighted synthesized speech considering the possible combination of candidates for the two basis vectors from the pre-selection. If  $c_{k_0, k_1}$  is the Gaussian code vector from the candidate vectors represented by the indices  $k_0$  and  $k_1$  and the respective signs for the two basis vectors, then the final Gaussian code vector is selected by maximizing the term:

$$A_{k_0, k_1} = \frac{(C_{k_0, k_1})^2}{E_{Dk_0, k_1}} = \frac{(d^T c_{k_0, k_1})^2}{c_{k_0, k_1}^T \Phi c_{k_0, k_1}}$$

over the candidate vectors.  $d = H^T x_2$  is the correlation between the target signal  $x_2(n)$  and the impulse response  $h(n)$  (without the pitch enhancement), and  $H$  is a lower triangular Toeplitz convolution matrix with diagonal  $h(0)$  and lower diagonals  $h(1), \dots, h(39)$ , and  $\Phi = H^T H$  is the matrix of correlations of  $h(n)$ .

More particularly, in the present embodiment, two subcodebooks are included (or utilized) in the fixed codebook 261 with 31 bits in the 11 kbps encoding mode. In the first subcodebook, the innovation vector contains 8 pulses. Each pulse has 3 bits to code the pulse position. The signs of 6 pulses are transmitted to the decoder with 6 bits. The second subcodebook contains innovation vectors comprising 10 pulses. Two bits for each pulse are assigned to code the pulse position which is limited in one of the 10 segments. Ten bits are spent for 10 signs of the 10 pulses. The bit allocation for the subcodebooks used in the fixed codebook 261 can be summarized as follows:

Subcodebook1: 8 pulses×3 bits/pulse+6 signs=30 bits

Subcodebook2: 10 pulses×2 bits/pulse+10 signs=30 bits

One of the two subcodebooks is chosen at the block 275 (FIG. 2) by favoring the second subcodebook using adaptive weighting applied when comparing the criterion value F1 from the first subcodebook to the criterion value F2 from the second subcodebook:

if  $(W_c \cdot F1 > F2)$ , the first subcodebook is chosen,

else, the second subcodebook is chosen,

where the weighting,  $0 < W_c < 1$ , is defined as:

$$W_c = \begin{cases} 1.0, & \text{if } P_{NSR} < 0.5, \\ 1.0 - 0.3 P_{NSR} (1.0 - 0.5 R_p) \cdot \min(P_{sharp} + 0.5, 1.0), & \text{otherwise} \end{cases}$$

$P_{NSR}$  is the background noise to speech signal ratio (i.e., the "noise level" in the block 279),  $R_p$  is the normalized LTP gain, and  $P_{sharp}$  is the sharpness parameter of the ideal excitation  $res_2(n)$  (i.e., the "sharpness" in the block 279).

In the 8 kbps mode, two subcodebooks are included in the fixed codebook 261 with 20 bits. In the first subcodebook, the innovation vector contains 4 pulses. Each pulse has 4 bits to code the pulse position. The signs of 3 pulses are transmitted to the decoder with 3 bits. The second subcode-

book contains innovation vectors having 10 pulses. One bit for each of 9 pulses is assigned to code the pulse position which is limited in one of the 10 segments. Ten bits are spent for 10 signs of the 10 pulses. The bit allocation for the subcodebook can be summarized as the following:

- Subcodebook1: 4 pulses×4 bits/pulse+3 signs=19 bits
- Subcodebook2: 9 pulses×1 bits/pulse+1 pulse×0 bit+10 signs=19 bits

One of the two subcodebooks is chosen by favoring the second subcodebook using adaptive weighting applied when comparing the criterion value F1 from the first subcodebook to the criterion value F2 from the second subcodebook as in the 11 kbps mode. The weighting,  $0 < W_c \leq 1$ , is defined as:

$$W_c = 1.0 - 0.6 P_{NSR}(1.0 - 0.5 R_p) \cdot \min\{P_{sharp} + 0.5, 1.0\}.$$

The 6.65 kbps mode operates using the long-term preprocessing (PP) or the traditional LTP. A pulse subcodebook of 18 bits is used when in the PP-mode. A total of 13 bits are allocated for three subcodebooks when operating in the LTP-mode. The bit allocation for the subcodebooks can be summarized as follows:

PP-mode:

- Subcodebook: 5 pulses×3 bits/pulse+3 signs=18 bits

LTP-mode:

- Subcodebook1: 3 pulses×3 bits/pulse+3 signs=12 bits, phase\_mode=1,
- Subcodebook2: 3 pulses×3 bits/pulse+2 signs=11 bits, phase\_mode=0,
- Subcodebook3: Gaussian subcodebook of 11 bits.

One of the 3 subcodebooks is chosen by favoring the Gaussian subcodebook when searching with LTP-mode. Adaptive weighting is applied when comparing the criterion value from the two pulse subcodebooks to the criterion value from the Gaussian subcodebook. The weighting,  $0 < W_c \leq 1$ , is defined as:

$$W_c = 1.0 - 0.9 P_{NSR}(1.0 - 0.5 R_p) \cdot \min\{P_{sharp} + 0.5, 1.0\},$$

$$\text{if (noise-like unvoiced), } W_c = W_c \cdot (0.2 R_p(1.0 - P_{sharp}) + 0.8).$$

The 5.8 kbps encoding mode works only with the long-term preprocessing (PP). Total 14 bits are allocated for three subcodebooks. The bit allocation for the subcodebooks can be summarized as the following:

- Subcodebook1: 4 pulses×3 bits/pulse+1 signs=13 bits, phase\_mode=1,
- Subcodebook2: 3 pulses×3 bits/pulse+3 signs=12 bits, phase\_mode=0,
- Subcodebook3: Gaussian subcodebook of 12 bits.

One of the 3 subcodebooks is chosen favoring the Gaussian subcodebook with adaptive weighting applied when comparing the criterion value from the two pulse subcodebooks to the criterion value from the Gaussian subcodebook. The weighting,  $0 < W_c \leq 1$ , is defined as:

$$W_c = 1.0 - P_{NSR}(1.0 - 0.5 R_p) \cdot \min\{P_{sharp} + 0.6, 1.0\},$$

$$\text{if (noise-like unvoiced), } W_c = W_c \cdot (0.3 R_p(1.0 - P_{sharp}) + 0.7).$$

The 4.55 kbps bit rate mode works only with the long-term preprocessing (PP). Total 10 bits are allocated for three subcodebooks. The bit allocation for the subcodebooks can be summarized as the following:

- Subcodebook1: 2 pulses×4 bits/pulse+1 signs=9 bits, phasemode=1,

Subcodebook2: 2 pulses×3 bits/pulse+2 signs=8 bits, phasemode=0,

Subcodebook3: Gaussian subcodebook of 8 bits.

One of the 3 subcodebooks is chosen by favoring the Gaussian subcodebook with weighting applied when comparing the criterion value from the two pulse subcodebooks to the criterion value from the Gaussian subcodebook. The weighting,  $0 < W_c \leq 1$ , is defined as:

$$W_c = 1.0 - 1.2 P_{NSR}(1.0 - 0.5 R_p) \cdot \min\{P_{sharp} + 0.6, 1.0\},$$

$$\text{if (noise-like unvoiced), } W_c = W_c \cdot (0.5 R_p(1.0 - P_{sharp}) + 0.4).$$

For 4.55, 5.8, 6.65 and 8.0 kbps bit rate encoding modes, a gain re-optimization procedure is performed to jointly optimize the adaptive and fixed codebook gains,  $g_p$  and  $g_c$ , respectively, as indicated in FIG. 3. The optimal gains are obtained from the following correlations given by:

$$g_p = \frac{R_1 R_2 - R_3 R_4}{R_3 R_2 - R_1 R_3}$$

$$g_c = \frac{R_4 - g_p R_3}{R_2}$$

where  $R_1 = \langle \bar{C}_p, \bar{T}_{gs} \rangle$ ,  $R_2 = \langle \bar{C}_c, \bar{C}_c \rangle$ ,  $R_3 = \langle \bar{C}_p, \bar{C}_c \rangle$ ,  $R_4 = \langle \bar{C}_p, \bar{C}_p \rangle$ , and  $R_5 = \langle \bar{C}_p, \bar{C}_p \rangle \cdot \bar{C}_c, \bar{C}_p$  and  $\bar{T}_{gs}$  are filtered fixed codebook excitation, filtered adaptive codebook excitation and the target signal for the adaptive codebook search.

For 11 kbps bit rate encoding, the adaptive codebook gain,  $g_p$ , remains the same as that computed in the close-loop pitch search. The fixed codebook gain,  $g_c$ , is obtained as:

$$g_c = \frac{R_5}{R_2}$$

where  $R_5 = \langle \bar{C}_c, \bar{T}_{gs} \rangle$  and  $\bar{T}_{gs} = \bar{T}_{gs} - g_p \bar{C}_p$ .

Original CELP algorithm is based on the concept of analysis by synthesis (waveform matching). At low bit rate or when coding noisy speech, the waveform matching becomes difficult so that the gains are up-down, frequently resulting in unnatural sounds. To compensate for this problem, the gains obtained in the analysis by synthesis close-loop sometimes need to be modified or normalized.

There are two basic gain normalization approaches. One is called open-loop approach which normalizes the energy of the synthesized excitation to the energy of the unquantized residual signal. Another one is close-loop approach with which the normalization is done considering the perceptual weighting. The gain normalization factor is a linear combination of the one from the close-loop approach and the one from the open-loop approach; the weighting coefficients used for the combination are controlled according to the LPC gain.

The decision to do the gain normalization is made if one of the following conditions is met: (a) the bit rate is 8.0 or 6.65 kbps, and noise-like unvoiced speech is true; (b) the noise level  $P_{NSR}$  is larger than 0.5; (c) the bit rate is 6.65 kbps, and the noise level  $P_{NSR}$  is larger than 0.2; and (d) the bit rate is 5.8 or 4.45 kbps.

The residual energy,  $E_{res}$ , and the target signal energy,  $E_{T_{gs}}$ , are defined respectively as:

$$E_{res} = \sum_{n=0}^{L_{SF}-1} r25^2(n)$$

-continued

$$E_{Tgs} = \sum_{n=0}^{L\_SF-1} T_{gs}^2(n)$$

Then the smoothed open-loop energy and the smoothed closed-loop energy are evaluated by:  
if (first subframe is true)

$$OL\_Eg = E_{Tgs}$$

else

$$OL\_Eg \leftarrow \beta_{sub} OL\_Eg + (1 - \beta_{sub}) E_{Tgs}$$

if (first subframe is true)

$$CL\_Eg = E_{Tgs}$$

else

$$CL\_Eg \leftarrow \beta_{sub} CL\_Eg + (1 - \beta_{sub}) E_{Tgs}$$

where  $\beta_{sub}$  is the smoothing coefficient which is determined according to the classification. After having the reference energy, the open-loop gain normalization factor is calculated:

$$ol\_g = \text{MIN} \left\{ C_{ol} \sqrt{\frac{OL\_Eg}{\sum_{n=0}^{L\_SF-1} v^2(n)}}, \frac{1.2}{g_p} \right\}$$

where  $C_{ol}$  is 0.8 for the bit rate 11.0 kbps, for the other rates  $C_{ol}$  is 0.7, and  $v(n)$  is the excitation:

$$v(n) = v_a(n)g_p + v_c(n)g_c, \quad n=0,1, \dots, L\_SF-1.$$

where  $g_p$  and  $g_c$  are unquantized gains. Similarly, the closed-loop gain normalization factor is:

$$CL\_g = \text{MIN} \left\{ C_{cl} \sqrt{\frac{CL\_Eg}{\sum_{n=0}^{L\_SF-1} y^2(n)}}, \frac{1.2}{g_p} \right\}$$

where  $C_{cl}$  is 0.9 for the bit rate 11.0 kbps, for the other rates  $C_{cl}$  is 0.8, and  $y(n)$  is the filtered signal ( $y(n) = v(n) * h(n)$ ):

$$y(n) = y_a(n)g_p + y_c(n)g_c, \quad n=0,1, \dots, L\_SF-1.$$

The final gain normalization factor,  $g_p$  is a combination of  $CL\_g$  and  $OL\_g$ , controlled in terms of an LPC gain parameter,  $C_{LPC}$   
if (speech is true or the rate is 11 kbps)

$$g_p = C_{LPC} OL\_g + (1 - C_{LPC}) CL\_g$$

$$g_p = \text{MAX}(1.0, g_p)$$

$$g_p = \text{MIN}(g_p, 1 + C_{LPC})$$

if (background noise is true and the rate is smaller than 11 kbps)

$$g_p = 1.2 \text{ MIN}\{CL\_g, OL\_g\}$$

where  $C_{LPC}$  is defined as:

$$C_{LPC} = \text{MIN}\{\text{sqrt}(E_{res}/E_{Tgs}), 0.8\}/0.8$$

Once the gain normalization factor is determined, the unquantized gains are modified:

$$g_p \leftarrow g_p \bar{g}_l$$

For 4.55, 5.8, 6.65 and 8.0 kbps bit rate encoding, the adaptive codebook gain and the fixed codebook gain are vector quantized using 6 bits for rate 4.55 kbps and 7 bits for the other rates. The gain codebook search is done by minimizing the mean squared weighted error,  $\text{Err}$ , between the original and reconstructed speech signals:

$$\text{Err} = \|\bar{T}_{gs} - g_p \bar{C}_p - g_c \bar{C}_c\|^2.$$

For rate 11.0 kbps, scalar quantization is performed to quantize both the adaptive codebook gain,  $g_p$ , using 4 bits and the fixed codebook gain,  $g_c$ , using 5 bits each.

The fixed codebook gain,  $g_c$ , is obtained by MA prediction of the energy of the scaled fixed codebook excitation in the following manner. Let  $\bar{E}(n)$  be the mean removed energy of the scaled fixed codebook excitation in (dB) at subframe  $n$  be given by:

$$\bar{E}(n) = 10 \log \left( \frac{1}{40} \sum_{i=0}^{39} c^2(i) \right) - \bar{E}$$

where  $c(i)$  is the unscaled fixed codebook excitation, and  $\bar{E} = 30$  dB is the mean energy of scaled fixed codebook excitation.

The predicted energy is given by:

$$\bar{E}(n) = \sum_{i=1}^4 b_i \hat{R}(n-i)$$

where  $[b_1, b_2, b_3, b_4] = [0.68, 0.58, 0.34, 0.19]$  are the MA prediction coefficients and  $\hat{R}(n)$  is the quantized prediction error at subframe  $n$ .

The predicted energy is used to compute a predicted fixed codebook gain  $g'_c$  (by substituting  $\bar{E}(n)$  by  $\hat{E}(n)$  and  $g_c$  by  $g'_c$ ). This is done as follows. First, the mean energy of the unscaled fixed codebook excitation is computed as:

$$E_c = 10 \log \left( \frac{1}{40} \sum_{i=0}^{39} c^2(i) \right)$$

and then the predicted gain  $g'_c$  is obtained as:

$$g'_c = 10^{(0.05(\bar{E}(n) - E_c))}$$

A correction factor between the gain,  $g_c$ , and the estimated one,  $g'_c$ , is given by:

$$\gamma = g_c / g'_c$$

It is also related to the prediction error as:

$$R(n) = E(n) - \bar{E}(n) = 20 \log \gamma$$

The codebook search for 4.55, 5.8, 6.65 and 8.0 kbps encoding bit rates consists of two steps. In the first step, a binary search of a single entry table representing the quantized prediction error is performed. In the second step, the index  $\text{Index\_1}$  of the optimum entry that is closest to the unquantized prediction error in mean square error sense is used to limit the search of the two-dimensional VQ table representing the adaptive codebook gain and the prediction

error. Taking advantage of the particular arrangement and ordering of the VQ table, a fast search using few candidates around the entry pointed by Index\_1 is performed. In fact, only about half of the VQ table entries are tested to lead to the optimum entry with Index\_2. Only Index\_2 is transmitted.

For 11.0 kbps bit rate encoding mode, a full search of both scalar gain codebooks are used to quantize  $g_p$  and  $g_c$ . For  $g_p$ , the search is performed by minimizing the error  $Err = \text{abs}(\bar{g}_p - \bar{g}_p)$ . Whereas for  $g_c$ , the search is performed by minimizing the error  $Err = |\bar{g}_c - \bar{g}_c|$ .

An update of the states of the synthesis and weighting filters is needed in order to compute the target signal for the next subframe. After the two gains are quantized, the excitation signal,  $u(n)$ , in the present subframe is computed as:

$$u(n) = \bar{g}_p v(n) + \bar{g}_c c(n), n=0,39,$$

where  $\bar{g}_p$  and  $\bar{g}_c$  are the quantized adaptive and fixed codebook gains respectively,  $v(n)$  the adaptive codebook excitation (interpolated past excitation), and  $c(n)$  is the fixed codebook excitation. The state of the filters can be updated by filtering the signal  $r(n) - u(n)$  through the filters  $1/\bar{A}(z)$  and  $W(z)$  for the 40-sample subframe and saving the states of the filters. This would normally require 3 filterings.

A simpler approach which requires only one filtering is as follows. The local synthesized speech at the encoder,  $\hat{s}(n)$ , is computed by filtering the excitation signal through  $1/\bar{A}(z)$ . The output of the filter due to the input  $r(n) - u(n)$  is equivalent to  $e(n) = s(n) - \hat{s}(n)$ , so the states of the synthesis filter  $1/\bar{A}(z)$  are given by  $e(n), n=0,39$ . Updating the states of the filter  $W(z)$  can be done by filtering the error signal  $e(n)$  through this filter to find the perceptually weighted error  $e_w(n)$ . However, the signal  $e_w(n)$  can be equivalently found by:

$$e_w(n) = T_{ps}(n) - \bar{g}_p C_p(n) - \bar{g}_c C_c(n).$$

The states of the weighting filter are updated by computing  $e_w(n)$  for  $n=30$  to 39.

The function of the decoder consists of decoding the transmitted parameters (dLP parameters, adaptive codebook vector and its gain, fixed codebook vector and its gain) and performing synthesis to obtain the reconstructed speech. The reconstructed speech is then postfiltered and upsampled.

The decoding process is performed in the following order. First, the LP filter parameters are encoded. The received indices of LSF quantization are used to reconstruct the quantized LSF vector. Interpolation is performed to obtain 4 interpolated LSF vectors (corresponding to 4 subframes). For each subframe, the interpolated LSF vector is converted to LP filter coefficient domain,  $a_p$ , which is used for synthesizing the reconstructed speech in the subframe.

For rates 4.55, 5.8 and 6.65 (during PP\_mode) kbps bit rate encoding modes, the received pitch index is used to interpolate the pitch lag across the entire subframe. The following three steps are repeated for each subframe:

- 1) Decoding of the gains: for bit rates of 4.55, 5.8, 6.65 and 8.0 kbps, the received index is used to find the quantized adaptive codebook gain,  $\bar{g}_p$ , from the 2-dimensional VQ table. The same index is used to get the fixed codebook gain correction factor  $\gamma$  from the same quantization table. The quantized fixed codebook gain,  $\bar{g}_c$ , is obtained following these steps:

the predicted energy is computed

$$\bar{E}(n) = \sum_{i=1}^4 b_i \hat{r}(n-i);$$

the energy of the unscaled fixed codebook excitation is calculated as

$$E_i = 10 \log \left( \frac{1}{40} \sum_{i=0}^{39} c^2(i) \right);$$

and

the predicted gain  $g'_c$  is obtained as  $g'_c = 10^{0.05(\bar{E}(n) + \bar{E} - \bar{E})}$ .

The quantized fixed codebook gain is given as  $\bar{g}_c = \gamma g'_c$ .

For 11 kbps bit rate, the received adaptive codebook gain index is used to readily find the quantized adaptive gain,  $\bar{g}_p$  from the quantization table. The received fixed codebook gain index gives the fixed codebook gain correction factor  $\gamma$ . The calculation of the quantized fixed codebook gain,  $\bar{g}_c$ , follows the same steps as the other rates.

- 2) Decoding of adaptive codebook vector: for 8.0, 11.0 and 6.65 (during LTP\_mode=1) kbps bit rate encoding modes, the received pitch index (adaptive codebook index) is used to find the integer and fractional parts of the pitch lag. The adaptive codebook  $v(n)$  is found by interpolating the past excitation  $u(n)$  (at the pitch delay) using the FIR filters.

- 3) Decoding of fixed codebook vector: the received codebook indices are used to extract the type of the codebook (pulse or Gaussian) and either the amplitudes and positions of the excitation pulses or the bases and signs of the Gaussian excitation. In either case, the reconstructed fixed codebook excitation is given as  $c(n)$ . If the integer part of the pitch lag is less than the subframe size 40 and the chosen excitation is pulse type, the pitch sharpening is applied. This translates into modifying  $c(n)$  as  $c(n) = c(n) + \beta c(n-T)$ , where  $\beta$  is the decoded pitch gain  $\bar{g}_p$  from the previous subframe bounded by  $[0.2, 1.0]$ .

The excitation at the input of the synthesis filter is given by  $u(n) = \bar{g}_p v(n) + \bar{g}_c c(n), n=0,39$ . Before the speech synthesis, a post-processing of the excitation elements is performed. This means that the total excitation is modified by emphasizing the contribution of the adaptive codebook vector:

$$\bar{u}(n) = \begin{cases} u(n) + 0.25\beta\bar{g}_p v(n), & \bar{g}_p > 0.5 \\ u(n), & \bar{g}_p \leq 0.5 \end{cases}$$

Adaptive gain control (AGC) is used to compensate for the gain difference between the unemphasized excitation  $u(n)$  and emphasized excitation  $\bar{u}(n)$ . The gain scaling factor  $\eta$  for the emphasized excitation is computed by:

$$\eta = \begin{cases} \sqrt{\frac{\sum_{n=0}^{39} u^2(n)}{\sum_{n=0}^{39} \bar{u}^2(n)}} & \bar{g}_p > 0.5 \\ 1.0 & \bar{g}_p \leq 0.5 \end{cases}$$

The gain-scaled emphasized excitation  $\bar{u}(n)$  is given by:

$$\bar{u}'(n) = \eta \bar{u}(n).$$



The reconstructed speech is given by:

$$\hat{s}(n) = \hat{r}'(n) - \sum_{i=1}^{10} \bar{a}_i \hat{s}(n-i), n = 0 \text{ to } 39,$$

where  $\bar{a}_i$  are the interpolated LP filter coefficients. The synthesized speech  $\hat{s}(n)$  is then passed through an adaptive postfilter.

Post-processing consists of two functions: adaptive post-filtering and signal up-scaling. The adaptive postfilter is the cascade of three filters: a formant postfilter and two tilt-compensation filters. The postfilter is updated every sub-frame of 5 ms. The formant postfilter is given by:

$$H_f(z) = \frac{\bar{A}(z/\gamma_n)}{\bar{A}(z/\gamma_d)}$$

where  $\bar{A}(z)$  is the received quantized and interpolated LP inverse filter and  $\gamma_n$  and  $\gamma_d$  control the amount of the formant postfiltering.

The first tilt compensation filter  $H_{r1}(z)$  compensates for the tilt in the formant postfilter  $H_f(z)$  and is given by:

$$H_{r1}(z) = (1 - \mu z^{-2})$$

where  $\mu = \gamma_{r1} k_1$  is a tilt factor, with  $k_1$  being the first reflection coefficient calculated on the truncated impulse response  $h_f(n)$ , of the formant postfilter

$$k_1 = \frac{r_h(1)}{r_h(0)}$$

with:

$$r_h(i) = \sum_{j=0}^{L_h-i-1} h_f(j)h_f(j+i), (L_h = 22).$$

The postfiltering process is performed as follows. First, the synthesized speech  $\hat{s}(n)$  is inverse filtered through  $\bar{A}(z/\gamma_n)$  to produce the residual signal  $\hat{r}(n)$ . The signal  $\hat{r}(n)$  is filtered by the synthesis filter  $1/\bar{A}(z/\gamma_d)$  is passed to the first tilt compensation filter  $h_{r1}(z)$  resulting in the postfiltered speech signal  $\hat{s}_f(n)$ .

Adaptive gain control (AGC) is used to compensate for the gain difference between the synthesized speech signal  $\hat{s}(n)$  and the postfiltered signal  $\hat{s}_f(n)$ . The gain scaling factor  $\gamma$  for the present subframe is computed by:

$$\gamma = \sqrt{\frac{\sum_{n=0}^{39} \hat{s}^2(n)}{\sum_{n=0}^{39} \hat{s}_f^2(n)}}$$

The gain-scaled postfiltered signal  $\bar{s}(n)$  is given by:

$$\bar{s}(n) = \beta(n) \hat{s}_f(n)$$

where  $\beta(n)$  is updated in sample by sample basis and given by:

$$\beta(n) = \alpha \beta(n-1) + (1-\alpha) \gamma$$

where  $\alpha$  is an AGC factor with value 0.9. Finally, up-scaling consists of multiplying the postfiltered speech by a factor 2 to undo the down scaling by 2 which is applied to the input signal.

FIGS. 6 and 7 are drawings of an alternate embodiment of a 4 kbps speech codec that also illustrates various aspects of the present invention. In particular, FIG. 6 is a block diagram of a speech encoder 601 that is built in accordance with the present invention. The speech encoder 601 is based on the analysis-by-synthesis principle. To achieve toll quality at 4 kbps, the speech encoder 601 departs from the strict waveform-matching criterion of regular CELP coders and strives to catch the perceptual important features of the input signal.

The speech encoder 601 operates on a frame size of 20 ms with three subframes (two of 6.625-ms and one of 6.75 ms). A look-ahead of 15 ms is used. The one-way coding delay of the codec adds up to 55 ms.

At a block 615, the spectral envelope is represented by a 10<sup>th</sup> order LPC analysis for each frame. The prediction coefficients are transformed to the Line Spectrum Frequencies (LSFs) for quantization. The input signal is modified to better fit the coding model without loss of quality. This processing is denoted "signal modification" as indicated by a block 621. In order to improve the quality of the reconstructed signal, perceptual important features are estimated and emphasized during encoding.

The excitation signal for an LPC synthesis filter 625 is build from the two traditional components: 1) the pitch contribution; and 2) the innovation contribution. The pitch contribution is provided through use of an adaptive codebook 627. An innovation codebook 629 has several sub-codebooks in order to provide robustness against a wide range of input signals. To each of the two contributions a gain is applied which, multiplied with their respective codebook vectors and summed, provide the excitation signal.

The LSFs and pitch lag are coded on a frame basis, and the remaining parameters (the innovation codebook index, the pitch gain, and the innovation codebook gain) are coded for every subframe. The LSF vector is coded using predictive vector quantization. The pitch lag has an integer part and a fractional part constituting the pitch period. The quantized pitch period has a non-uniform resolution with higher density of quantized values at lower delays. The bit allocation for the parameters is shown in the following table.

Table of Bit Allocation

Parameter	Bits per 20 ms
LSFs	21
Pitch lag (adaptive codebook)	8
Gains	12
Innovation codebook	3 × 13 = 39
Total	80

When the quantization of all parameters for a frame is complete the indices are multiplexed to form the 80 bits for the serial bit-stream.

FIG. 7 is a block diagram of a decoder 701 with corresponding functionality to that of the encoder of FIG. 6. The decoder 701 receives the 80 bits on a frame basis from a demultiplexor 711. Upon receipt of the bits, the decoder 701 checks the sync-word for a bad frame indication, and decides whether the entire 80 bits should be disregarded and frame erasure concealment applied. If the frame is not declared a frame erasure, the 80 bits are mapped to the parameter indices of the codec, and the parameters are decoded from the indices using the inverse quantization schemes of the encoder of FIG. 6.

When the LSFs, pitch lag, pitch gains, innovation vectors, and gains for the innovation vectors are decoded, the excitation signal is reconstructed via a block 715. The output signal is synthesized by passing the reconstructed excitation signal through an LPC synthesis filter 721. To enhance the perceptual quality of the reconstructed signal both short-term and long-term post-processing are applied at a block 731.

Regarding the bit allocation of the 4 kbps codec (as shown in the prior table), the LSFs and pitch lag are quantized with 21 and 8 bits per 20 ms, respectively. Although the three subframes are of different size the remaining bits are allocated evenly among them. Thus, the innovation vector is quantized with 13 bits per subframe. This adds up to a total of 80 bits per 20 ms, equivalent to 4 kbps.

The estimated complexity numbers for the proposed 4 kbps codec are listed in the following table. All numbers are under the assumption that the codec is implemented on commercially available 16-bit fixed point DSPs in full duplex mode. All storage numbers are under the assumption of 16-bit words, and the complexity estimates are based on the floating point C-source code of the codec.

Table of Complexity Estimates

Computational complexity	30 MIPS
Program and data ROM	18 kwords
RAM	3 kwords

The decoder 701 comprises decode processing circuitry that generally operates pursuant to software control. Similarly, the encoder 601 (FIG. 6) comprises encoder processing circuitry also operating pursuant to software control. Such processing circuitry may coexist, at least in part, within a single processing unit such as a single DSP.

FIG. 8 is a flow diagram illustrating use of adaptive tilt compensation in an exemplary decoder built in accordance with the present invention. Especially inherent with lower bit rate encoding, waveform matching of lower frequency regions proves easier than higher frequency regions. As a result, for example, a codec might produce a synthesized residual that has greater high frequency energy and lesser low frequency energy than would otherwise be desired. In other words, the resultant synthesized residual would exhibit an unwanted spectral tilt.

Although a preset mechanism for readjusting the synthesized residual might in general help counter such tilt, in the present embodiment an adaptive mechanism is employed. The adaptive mechanism (herein adaptive correction or adaptive compensation) provides superior performance in at least most circumstances because the amount of spectral tilt is inconsistent either from one encoding bit rate to another or from one synthesized residual portion to the next using a single encoding bit rate.

A first mechanism for adaptation comprises selecting a predetermined amount of compensation to apply, for example by filtering, based on the encoding bit rate selected in an adaptive multi-rate codec. The amount of compensation increases as the encoding bit rate decreases, and visa versa.

A second mechanism comprises adaptively selecting more or less compensation to apply to track the actual tilt from one synthesized residual portion to the next. Lastly, the first and second mechanisms might be combined. For example, the first mechanism might be used to select a tilt compensation range and/or a tilt weighting factor based on

the encoding bit rate, while the second might fine tune the compensation within the range and/or employing the weighting factor. Clearly, many variations are possible including those identified with reference to FIGS. 8 and 9.

Although such adaptive compensation may occur at any time after the initial generation of the synthesized residual (for example in the encoder), in the present embodiment, it is applied at the decoder as illustrated in FIG. 5. The decoder applies adaptive compensation to the summed component parts of the synthesized residual, i.e., to the resultant sum of the fixed and adaptive codebook contributions. Alternatively, adaptive compensation might be applied prior to combining the fixed and the adaptive codebook contributions, e.g., to each contribution separately, or at any point prior to synthesis.

In particular, with reference to FIG. 8, at a block 811, a decoder processing circuit first considers the encoding bit rate to determine whether to apply adaptive compensation. If a relatively high bit rate is selected, the decoder processing circuit (although it may anyway in some embodiments) need not apply adaptive compensation. Otherwise, at a block 815, the decoder processing circuit identifies the amount of compensation needed. Thereafter, the identified amount of compensation needed is applied at a block 817.

Although the identification and compensation at the blocks 815 and 817 comprises two independent steps, alternatively, they might be combined into a single process or broken into many further steps. The identification and compensation process together constitutes adaptive compensation.

FIG. 9 is a flow diagram illustrating a specific embodiment of a decoder that illustrates an exemplary approach for performing the identification and compensation processing of FIG. 8. First, at a block 911, the decoder applies a long asymmetric window to the synthesized residual. The window is typically 240 samples in length, and centered at a current subframe having a typical size of 40 samples. A first reflection coefficient, the normalized first order correlation, of the windowed synthesized residual is calculated, smoothed and weighted by a constant factor at blocks 913 and 915. The resultant coefficient value comprises a compensation factor, which, of course, adapts based on the windowed content.

After identifying the adaptive compensation factor, i.e., the smoothed and weighted reflection coefficient, the decoder compensates for the spectral tilt at a block 917. Specifically, the decoder constructs a first order filter using the reflection coefficient, and applies the filter to the synthesized residual to remove at least part of the spectral tilt. Further, at least in some embodiments, the filtering is actually applied to the weighted synthesized residual.

As with the embodiment illustrated by FIG. 8, the decoder of FIG. 9 might also only apply such adaptive compensation at lower encoding bit rates. Similarly, other of the aforementioned variations might also be applied.

Of course, many other modifications and variations are also possible. In view of the above detailed description of the present invention and associated drawings, such other modifications and variations will now become apparent to those skilled in the art. It should also be apparent that such other modifications and variations may be effected without departing from the spirit and scope of the present invention.

In addition, the following Appendix A provides a list of many of the definitions, symbols and abbreviations used in this application. Appendices B and C respectively provide source and channel bit ordering information at various encoding bit rates used in one embodiment of the present

invention. Appendices A, B and C comprise part of the detailed description of the present application, and, otherwise, are hereby incorporated herein by reference in its entirety.

We claim:

1. A speech system using an analysis by synthesis approach on a speech signal, the speech system comprising:  
at least one codebook containing at least one code vector;  
processing circuitry that generates a synthesized residual signal using the at least one codebook; and  
the processing circuitry applying adaptive tilt compensation to the synthesized residual signal based in part on an encoding bit rate of the speech system and a flatness of the synthesized residual signal.
2. The speech system of claim 1 wherein the processing circuitry comprises an encoder processing circuit that gen-

erates the synthesized residual signal, and a decoder processing circuit that applies the adaptive tilt compensation.

3. The speech system of claim 1 wherein the synthesized residual signal comprises a weighted synthesized residual signal.
4. The speech system of claim 1 wherein the adaptive tilt compensation comprises identifying a filter coefficient for use in a compensating filter.
5. The speech system of claim 4 wherein the compensating filter comprises a first order filter.
6. The speech system of claim 4 wherein the identification of the filter coefficient comprises application of a window to the synthesized residual.

\* \* \* \* \*