



☒	Description	: paged ...1.... to paged ...3....
☒	Claim(s)	: paged ...14... to paged ...18....
☒	Drawing(s)	: paged ...1/4... to paged ...1/6....
☒	Abstract	: page 1 ...13... to paged ...-....
DATE OF RECEIPT		

AN EXCITATION ANALYSIS AND SYNTHESIS METHOD FOR CODING OF AUDIO AND SPEECH SIGNALS

Field of the Invention

The present invention relates to digital coding systems and, in particular, discloses an arrangement that can be used in very low bitrate compression coding of audio and speech signals.

Background Art

Digital speech, wideband speech and low-bandwidth audio compression coding algorithms have been standardised internationally by various bodies, such as the Telecommunications Standardisation Sector of International Telecommunication Union (ITU-T) and European Technical Standards Institute. Such speech compression algorithms are currently being used in the coding of speech or audio signals of 7 kHz bandwidth and below. Various standardised algorithms which have emerged included the ITU-T recommendations e.g. G.723, G.729 and G.728. These algorithms are useful for coding of speech signals at the bitrate range of 5.3 kbits/s to 16 kbits/s. The quality of coded outputs from such algorithms range from communication quality to toll quality.

While various techniques such as Multipulse, Code Excited Linear Prediction (CELP), and Multiband Excitation have been used for speech coding, CELP has proven to be very popular. As a result, different enhancements and features have been added to the CELP scheme. The ITU-T G.723 standard is highly representative of the present state of the art. Details are reported in the ITU-T document "Draft Recommendation G.723- Dual rate speech coder for multimedia telecommunication transmitting at 5.3 kbits/s & 6.3 kbits/s" dated October 14, 1995.

Figure 3 shows a speech encoder block diagram of the prior art. The coder first buffers the input consecutive speech samples in which the input signal is sampled at 8000 Hz and converted to a 16 bit linear PCM format. These consecutive input samples, denoted by S3.1, are buffered into frames of 240 samples by the framer 3.1, and a further partitioning into two parts of 120 samples each and a subdivision into four subframes of 60 samples is performed. The resultant output signal S3.2 is then passed



into a high pass filter 3.2, whereby the dc component is removed. The filtered output signal S3.3, is subjected to a tenth order linear predictive (LP) analysis. In the LP analysis, a hamming window is first applied to a 180-sample window centred on a current subframe. Eleven autocorrelation coefficients are then computed from the windowed signal, and a correction factor follows for the first autocorrelation coefficient while the remaining coefficients are multiplied by values from a table. Altogether, four linear predictive coefficient sets are computed for the 240-sample frame. The output from the LPC analysis, S3.4, is then quantized using a line spectral pair (LSP) predictive split vector quantizer 3.4. The resultant three sub-vector output signal S3.5 is then passed through an LSP decoder 3.5, to obtain a decoded LSP vector S3.6. Linear interpolation, performed by an interpolator 3.6, is executed on a subframe basis, between a decoded LSP vector and the previous LSP vector. The four interpolated LSP vectors are then converted to LPC vectors, and a decoded speech signal is obtained through an LPC synthesis filter, also done within the interpolator.

For each subframe, the unquantized LPC coefficients signal S3.4, are used to construct a formant perceptual weighting filter 3.10, and each high pass filtered input speech subframe of signal S3.3 is filtered by the constructed filter. The formant perceptual weighting filter coefficients S3.8 are passed to the impulse response calculator 3.7. The weighted output speech signal S3.9 of the formant perceptual weighting filter 3.10 is passed to a pitch estimate 3.12 and a harmonic noise shaping filter 3.11. The harmonic noise shaping filter 3.11, provides a mechanism to improve the quality of the encoded speech. Computation of the noise shaping filter coefficients S3.10 involve the estimation of the optimum filter gain from an optimal lag. The noise shaping filter coefficients S3.10, are passed to the impulse response calculator 3.7. At the impulse response calculator 3.7, a combination filter is synthesised using the inputs from the LPC synthesis filter coefficients S3.7, the formant perceptual weighting filter coefficients S3.8, and the harmonic noise shaping filter coefficients S3.10. The coefficients of the resultant filter S3.11, are saved in a memory update unit 3.8, for processing a reconstructed excitation sequence. The memory output S3.13 of the



combination filter is passed to a zero input response module 3.9, for computing the output of the combination filter when the input signal is all zero-valued samples. The output zero-input response S3.14, is then subtracted from an output S3.16 of the harmonic noise shaping filter 3.11, producing a resulting target vector S3.17.

For every two subframes, one pitch estimate is performed in the pitch estimator 3.12, using the formant perceptually weighted speech signal S3.9. In the pitch estimator 3.12, an index which maximises a defined cross-correlation factor is selected as the estimate of the open loop pitch period.

In a pitch predictor 3.13, a fifth order pitch predictor uses the open loop pitch period indices S3.15 from the pitch estimator 3.12, to determine the closed loop pitch lags of subframes 0 and 2. These pitch lags are coded using 7 bits, and for subframes 1 and 3, the closed loop pitch lag is coded differentially using 2 bits, relative to the pitch lags of subframes 0 and 2 respectively. The pitch predictor gains are vector quantized using two codebooks. These two codebooks have 85 or 170 entries for the higher bitrate and 170 entries for the low bitrate. An output S3.19 of the pitch predictor 3.3 is subtracted from the target vector S3.17, to produce the residual signal S3.20, which is an input to the MP-MLQ/ACELP module are the pitch predictor lags S3.18, and the impulse response S3.12, of the impulse response calculator 3.7.

In the MP-MLQ/ACELP module, the quantisation of the residual signal is performed. MP-MLQ is used for quantisation at the higher bitrate and ACELP for the lower bitrate. The quantization process is such that the residual signal is approximated by an excitation vector consisting of gain factor and pulse position lag indices. These indices serve as an input S3.21. to the excitation decoder, 3.16. In the excitation decoder 3.16, the decoding of pulses is performed by first decoding a series of parameters such as a maximum gain index, position of pulses, grid position, pulse sign and in the case of high bitrate decoding, the pulse train bit. The pitch predictor lag and gain vector are decoded in a pitch decoder 3.15, and are used to generate the pitch contribution S3.23. The pitch contribution S3.23 and the pulse contribution S3.22, are summed together to form the excitation vector S3.24, which is used to update the



memory unit 3.6 of the combination filter, provided by a signal S3.11. This is achieved by passing the signal S3.24 through the filter and the memory of the combination filter is saved in the memory update module 3.8.

The bit allocations for the coder are used for the LPC indices, adaptive code-book lags, all the gains, pulse positions, pulse signs and grid index. The total number of bits allocated per frame amount to 192 bits and 158 bits for coding at 6.3 kbit/s and 5.3 kbit/s respectively. The bits are arranged in a defined bitstream syntax for transmission.

In the prior art decoder shown in Figure 4, the received bitstream signal is first analysed and unpacked into the respective signal inputs, namely an input S4.1 for the LSP decoder 4.1, an input S4.4 for the pitch decoder 4.3, and an input S4.6 for the excitation decoder 4.4.

For the decoder of Fig. 4, the LSP decoder 4.1, the LSP interpolator 4.2, the pitch decoder 4.3, and the excitation decoder 4.4, all operate identically to their corresponding modules in the encoder of Fig. 3, (i.e. LSP decoder 3.5, interpolator 3.6, pitch decoder 3.15 and excitation decoder 3.16 respectively). The output signals from the decoder modules, namely, S4.2, S4.3, S4.5, S4.7 and S4.8 are similar to those of S3.6, S3.7, S3.23, S3.22 and S3.24 respectively.

A pitch postfilter 4.5 takes in a complete frame excitation signal S4.8, and performs the postfiltering on a subframe basis. The pitch postfilter 4.5 serves to enhance the quality of the synthesized signal. The postfiltered signal S4.9 is fed into a tenth order LPC synthesis filter 4.6, where a speech signal S4.10, is synthesised. A formant postfilter 4.7, then performs a conventional ARMA filtering to obtain an output S4.11. This signal is gain-scaled in a gain scaling unit 4.8 to obtain the decoder output vector S4.12.

Even though the ITU-T G.723 has achieved good communication quality for speech coding at 5.3 kbit/s and 6.3kbit/s, as a result of increasing demand for lower bitrate speech coding services, similar or better speech and low-bandwidth audio quality



is sought at bitrates of 4.8 kbit/s and below, with no or marginal corresponding increase in algorithmic complexity.

Synthesis of high quality speech and low-bandwidth audio at low bitrates using linear predictive coding techniques thus requires efficient coding of both the LPC filter parameters and input excitation parameters.

It is also desirable that means to encode the excitation at low bitrates is provided to generate a representation that provides an accurate and compact representation of the excitation for speech synthesis.

As the excitation part constitutes more than 80 percent of the codec's coding bits, improvement can be obtained by a more efficient excitation analysis and synthesis method and coding.

Summary of the Invention

Various aspects of the present invention are defined in the claims appended hereto.

Specific embodiments provide a means of further bitrate reduction using a pitch synchronous concept and classifying the signal into three modes for which different coding techniques are applied. It enables the secondary residual to be coded at a lower bitrate. One of the techniques used to code the secondary residual is the extended excitation of the previous excitation, which is efficient in reducing the bit rate for the voiced frame. In order to overcome the over-periodic problem for extended excitation, a joint optimization of the adaptive codebook and the extended excitation is used.

Brief Description of the Drawings

A number of embodiments of the present invention will now be described with reference to the drawings in which:

Figure 1 shows an illustration of an encoder of the present invention;

Figure 2 shows an illustration of a decoder of the invention;

Figure 3 shows an illustration of an encoder of the prior art;

Figure 4 shows an illustration of a decoder of the prior art;



Figure 5 shows an illustration of the derivation of extended excitation from a previous ACELP excitation of 4 pulses; and

Figure 6 shows an illustration of the derivation of extended excitation from a previous ACELP excitation of 8 pulses.

Detailed Description of the Preferred Embodiments

This coder of the embodiment operates on the principles of linear prediction analysis-by-synthesis coding and uses different excitation techniques to code the residual signal. The block diagram of an exemplary encoder is shown in Figure 1. The encoder operates on blocks (frames) of 240 samples each, which is equal to 30 msec at an 8 kHz sampling rate. An input speech signal S1.1 is first buffered into frames of 240 samples. Each frame is then high-pass filtered to remove the DC component and then divided into four subframes of 60 samples each, denoted by signal S1.2. For every subframe, a tenth order linear prediction coder (LPC) filter 1.2 is computed using the pre-processed signal S1.2. The LPC filter 1.2 for the last subframe is quantized using a Predictive Split Vector Quantizer (PSVQ) 1.4. The other subframes of the LSP coefficients are linearly interpolated. The unquantized LPC coefficients are used to construct a short term perceptual weighting filter 1.3, which is used to filter the entire frame and to obtain the perceptually weighted speech signal S1.5.

The open loop pitch period signal S1.7 is computed by a pitch estimator 1.6 using a formant weighted speech signal S1.5 output from the filter 1.3, for every two subframes (120 samples). The pitch period is searched in the range from 18 to 142 samples.

From this point the speech is processed on a 60 sample per subframe basis.

Using a estimated open-loop pitch period signal S1.7 computed by the pitch estimator 1.6, a harmonic noise shaping filter 1.7, is constructed. The combination of the LPC synthesis filter 1.2, the formant perceptual weighting filter 1.3, and the harmonic noise shaping filter 1.7 is used to create an impulse response signal S1.9.



The impulse response signal S1.9 is then convolved with an excitation signal S1.13, and a zero state response signal S1.14, is obtained.

Using the pitch period estimation S1.7, and the impulse response S1.9, an adaptive codebook excitation S1.21 is computed. In adaptive codebook excitation, a fifth order pitch predictor is used. The pitch period is computed as a small differential value around the open loop pitch estimate. The contribution of the adaptive codebook S1.14, (when S1.22 is zero) is then subtracted from the initial target vector S1.12. Both the pitch period of subframes 1 and 2 and the differential values for subframes 1 and 3 are transmitted to the decoder.

Finally, the secondary residual signals are encoded by either a voiced or unvoiced excitation module according to the signal classifier's output. For the voiced case, a further classification is conducted according to the closed-loop pitch lag, which determines whether an algebraic codebook excitation (ACELP) S1.18, or gain-scaled extended excitation S1.19, is used for encoding this voiced residual accordingly. When using extended excitation, both the adaptive and extended excitation are jointly optimized. For the unvoiced case, a Gaussian stochastic codebook S1.20, is used.

The transmitted information to the receiver includes the LSP vector index, adaptive codebook lag and tap indices, 1-bit voiced/unvoiced flag, and secondary excitation information.

The decoder diagram is shown in Figure 2. The decoder operation is also performed on a frame-by-frame basis. First the quantized LPC indices S2.1, are decoded, then the decoder constructs the LPC synthesis filter 2.8. For every subframe, both the adaptive codebook excitation indices S2.6 and fixed (secondary) codebook excitation indices S2.4, are decoded and the decoded excitation is input S2.14 to the synthesis filter 2.8. For the fixed codebook excitation, the excitation selector selects the ACELP excitation 2.4, extended excitation 2.5, or stochastic excitation 2.6, according to a signal mode flag, S2.5. The adaptive postfilter consists of a formant postfilter 2.9, and a forward-backward pitch postfilter 2.7. The excitation signal, S2.13, is input to the pitch postfilter 2.7, which in turn is input to the synthesis filter



2.8 whose output S2.15, is input to the formant postfilter 2.9. A gain scaling unit 2.10, maintains the energy at the input level of the formant postfilter 2.9.

Operation of the preferred embodiments of the present invention can now be explained using Figure 1, Figure 2, Figure 5 and Figure 6 which provide a method for speech coding to reduce the coding bits for voiced speech or audio.

In the encoder of Figure 1 and decoder of Figure 2, the shaded modules depict additions to the prior art. In particular, pre-processes module 1.1 in Figure 1 is the combination of both the framer 3.1 and high pass filter 3.2 of Figure 3. The LPC analysis module 1.2 of Figure 1 is similar to module 3.3 in Figure 3. Modules indicated by literals 1.3, 1.5, 1.6, 1.7 and 1.8 in Figure 1 are the same as the modules indicated by literals 3.10, 3.7, 3.12, 3.11 and 3.9 in Figure 3. Module indicated by literal 1.4 in Figure 1 is the combination of modules indicated by literals 3.4, 3.5 and 3.6 in Figure 3. In Figure 2, the modules denoted by literals 2.1, 2.2, 2.7, 2.8, 2.9 and 2.10 are the same as the modules denoted by literals 4.1, 4.2, 4.5, 4.6, 4.7 and 4.8 respectively. The algebraic codebook excitation module 1.11 of Figure 1 and 2.4 in Figure 2 are similar to the ACELP module 3.14 of Figure 3 and the excitation decoder module 4.4, in Figure 4.

Adaptive codebook excitation module 1.10 of Figure 1 is similar to pitch predictor module 3.13 of Figure 3, however the optimization procedure is slightly different between them. For the preferred embodiment, the adaptive codebook excitation optimization can be jointly conducted with the extended excitation according to the decision of the mode classifier module 1.9.

The signal mode classifier 1.9 of Figure 1 is input with the open pitch lag S1.7, from which a voiced/unvoiced indicator β_i ($i=0,1,2,3$) can be estimated by analyzing formant perceptual weighted signal S1.5. The different possible combinations of voiced and unvoiced decision of each set of consecutive subframes are shown in Table 1. In Table 1, V means voiced and U means unvoiced subframes. A final classification for the entire frame is given. Based on the extended excitation



strategy being used, only frames with subframe 0 identified as voiced and which also has a majority of voiced subframes may be classified as voiced frames.



Table 1
Voiced/unvoiced classification

Case	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Subframe 0	V	V	V	V	V	V	V	V	U	U	U	U	U	U	U	U
Subframe 1	V	V	V	V	U	U	U	U	V	V	V	V	U	U	U	U
Subframe 2	V	V	U	U	V	V	U	U	V	V	U	U	V	V	U	U
Subframe 3	V	U	V	U	V	U	V	U	V	U	V	U	V	U	V	U
Frame classification V/U	V	V	V	U	V	U	U	U	U	U	U	U	U	U	U	U

After the classification of voiced/unvoiced, a further classification, signal mode classification, is conducted for the voiced with respect to the pitch lag of subframe 0. The signal mode classification of the frame consists of three modes. Mode 1 is assigned if a frame is voiced and the pitch lag is less than 60 samples. Mode 2 is assigned if a frame is voiced and the pitch lag is between 60 and 120 samples. Mode 3 is assigned if a frame is unvoiced or the pitch lag is greater than 120 samples. A summary of signal mode classification is shown in Table 2.

Table 2
Signal mode classification

Cases		Signal Mode
Voiced	Pitch $\leq$ 60 (Samples)	Mode 1
Voiced	$60 \leq$ Pitch $\leq$ 120	Mode 2
Others (unvoiced & pitch $>$ 120)		Mode 3

As soon as the signal classification is finished, the corresponding techniques to encode the secondary residual signals for each mode are used. The coding techniques are shown in Table 3. If the current frame is classified as Mode 1, for instance, the secondary residuals (target signal S1.12, minus the adaptive codebook contribution) of subframes 0 and 2 are encoded by ACELP techniques 1.11, the secondary residuals of subframes 1 and 3 are encoded by extended excitation 1.12. At Mode 2, the secondary residuals of subframes 0 and 1 are encoded by ACELP techniques, the secondary residuals of subframes 2 and 3 are encoded by extended excitation. At Mode 3, the



secondary residuals of every subframe are encoded by stochastic excitation 1.13. It will be shown during extended excitation subframes, that there is no requirement to encode secondary residual parameters as these are derivable from parameters already coded as joint optimization techniques are used.

Table 3
Techniques for coding the secondary residual at each mode

Signal mode	Subframe 0	Subframe 1	Subframe 2	Subframe 3
Mode 1	ACELP	Extended	ACELP	Extended
Mode 2	ACELP	ACELP	Extended	Extended
Mode 3	Stochastic	Stochastic	Stochastic	Stochastic

The secondary residual signal S1.16, after subtracting the adaptive codebook contribution is coded by the stochastic codebook 1.13, if the signal is classified into Mode 3. The secondary residual signal is transferred as secondary target vector to the random codebook excitation module, 1.13. This module performs the quantization of this vector. The quantization process is approximately the secondary target vector by the gain G_s , scaled version of signal S1.20. The best gain G_s and excitation vector S1.20 which can minimize S1.15 are found.

According to the output of the classifier 1.9, as shown in Table 2, the residuals of some subframes are coded by extending the previous ACELP encoded excitations. A further classification is made according to the pitch period, i.e. Mode 1 or Mode 2. The reason for using extended excitation 1.12 is that by forcing the secondary excitation to have a pitch structure, it is possible to match voiced speech more accurately.

The extended excitation of a current subframe S1.19, is obtained by extending the best excitation of the previous subframe with an interval equal to the current closed loop pitch, and being scaled by a gain G_e .

An illustration of the derivation of an extended excitation from a previous ACELP excitation (e.g. 4 pulses) is shown in Figure 5. The solid-line pulses are the original ACELP excitation pulses of previous subframe. The broken-line pulses are the



extended ones with an interval equal to pitch. The intervals between any two identically numbered pulses are equal to current subframe's pitch.

The extended excitation of subframe 2 is obtained by extending the previous two subframes' best excitations with multiple intervals of the current closed loop pitch, and being scaled by a gain G_e .

An illustration of the derivation of extended excitation from previous ACELP excitation (e.g. 8 pulses) is shown in Figure 6. The solid-line pulses are the original ACELP excitation pulses of previous subframes 0 and 1. The broken-line pulses are the extended ones with an interval equal to pitch.

For the subframe 3, the excitation of subframes 1 and 2 are regarded as the previous subframes' excitation. The same extension techniques which are applied to subframe 2 to obtain the extended excitation are similarly applied for subframe 3.

In the prior arts, the adaptive codebook excitation and the secondary (ACELP, extended, or stochastic) excitations are computed sequentially. When the adaptive excitation is optimized, the effect of the secondary excitation is assumed to be zero. Thus, after the secondary excitation is calculated, the adaptive excitation becomes sub-optimal. The reason for sequential optimisation is that the number of computations is reduced significantly.

In the preferred embodiments described herein, for the subframes with extended excitations, the secondary excitation is actually decided already except for the gain levels. Therefore, employing joint optimization does not result in a severe increase in computation load.

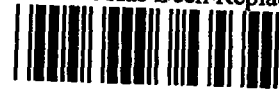
The shape of this extended excitation is almost decided. The shape is only decided by the closed-loop pitch lag. For example, let G_e denote the gain level for the extended excitation S1.19, S1.7 is the open loop pitch lag, b_k is the adaptive codebook gain vector, then the joint mean square error E_{joint} is minimized jointly by selecting G_e , the shift of open pitch lag S1.7 (also referred as closed loop pitch lag) and b_k , by which the adaptive codebook excitation and extended excitation can be determined.



If sequential optimization is used instead of joint optimization, the processed speech sounds slightly "vibrating". The reason is that the direct extension of the previous excitation by the interval of adaptive codebook pitch lag amplifies the periodic effects. By using this joint scheme, the "vibrating" effect can be avoided.

This present invention uses a more efficient representation of the input excitation of a Linear Prediction Analysis-by-Synthesis Coder, thereby resulting in the high quality speech and low-bandwidth audio coding being attainable at total bitrate of 4.8 kbit/s and below. The new excitation coding techniques outlined here are highly useful for very low bitrate compression coding of speech and low-bandwidth audio.

The foregoing describes only a number of embodiments of the present invention and modifications, obvious to those skilled in the art, can be made thereto without departing from the scope of the present invention.



1. A method of encoding audio or speech signals, said method comprising the steps of:
 - a) sampling an input audio or speech signal and grouping said samples into frames each of an equal number of said samples;
 - b) passing said frames through a high pass filter to remove any DC component, thereby producing pre-processed signal frames;
 - c) dividing each of said pre-processed signal frames into a number of subframes of equal size;
 - d) computing a set of tenth order linear predictive (LP) filter coefficients on a three subframe window centred on each subframe;
 - e) converting the LP filter coefficients into line spectral pair (LSP) coefficients to remove any long-term DC component and then vector quantizing the resultant dc-removed LSP coefficients;
 - f) decoding the quantized LSP vectors using a LSP decoder and performing linear interpolation between a decoded LSP vector and the previous LSP vector to produce 4 interpolated LSP vectors;
 - g) constructing a formant perceptual weighting filter using said set of LP coefficients generated for each subframe;
 - h) filtering each said subframe of an input frame using the perceptual weighting filter resulting in a perceptually weighted audio or speech frame;
 - i) computing at least two pitch estimates for every frame, one pitch estimate being for the first two subframes, and one other pitch estimate for the last two subframes;
 - j) subjecting the formant perceptually weighted frame to harmonic noise shaping filtering to obtain a target vector frame;
 - k) generating an impulse response filter with component inputs from the quantized LP filter coefficients, harmonic noise shaping filter coefficients and the formant perceptual weighting filter coefficients;



- l) establishing a zero input response for the impulse response filter when the input signals are all zero-valued samples;
- m) computing an open-loop pitch period using the formant weighted speech signal for every two subframes;
- n) using the open-loop pitch period and formant perceptually weighted speech or voice as inputs to a mode classifier to obtain a voiced/unvoiced classification;
- o) generating an excitation signal using excitation analysis for convolution with the impulse response to produce a zero state response; and
- p) adjusting said target vector frame by subtraction of the zero state response and the zero input response and feedback of the resultant adjusted target vector through a least-mean-square algorithm for excitation analysis.

2. A method for decoding audio or speech signals encoded according to the method of claim 1, said decoding method comprising the steps of:

- a) decoding linear predictive coder (LPC) indices using an LSP decoder and LSP interpolation process corresponding to those of the encoding method and constructing a LPC synthesis filter for each subframe of an input signal;
- b) decoding an adaptive codebook excitation and a fixed codebook excitation and inputting the decoded excitation to the LPC synthesis filter;
- c) selecting one of an algebraic codebook excitation (ACELP), extended excitation or stochastic excitation depending on a signal mode flag set;
- d) subjecting the decoded excitation vector to a forward-backward pitch postfilter, the output of which is fed into the LPC synthesis filter; and
- e) subjecting the decoded LPC synthesis output to a formant postfilter and a gain scaling unit to provide a decoded output signal.

3. A signal mode classification method comprising the steps of:

- a) subdividing a signal frame into at least four subframes each of an equal number of samples;



- b) analysing the amplitude of samples in each said subframe and classifying the subframe as voiced or unvoiced according to a threshold criteria;
- c) classifying the signal frame as an unvoiced frame when two or more of the four subframes are classified as unvoiced, otherwise the signal frame is classified as voiced;
- d) computing a pitch lag of the signal frame;
- e) defining two pitch lag thresholds, n_1 and n_2 , such that n_1 is less than n_2 and n_2 is less than the total number of samples per frame; and
- f) designating the signal frame as mode 1 when the frame has been identified as a voiced frame and the pitch lag is n_1 samples or less;
- g) designating the frame as mode 2 when the frame has been identified as a voiced frame and the pitch lag is greater than n_1 samples but less than or equal to n_2 samples; or
- h) designating the frame as mode 3 when the frame has been identified as a voiced frame and the pitch lag exceeds n_2 samples or when the frame has been identified as an unvoiced frame.

4. A multi-mode excitation coding method comprising the steps of:

- a) selection of one of three excitation modes, namely, algebraic codebook excitation (ACELP), extended excitation and stochastic excitation, based on the mode decision of a voiced/unvoiced signal mode classification;
- b) performing a further classification of said frame when said frame is classified as a voiced frame, said further classification being performed according to the pitch lag which determines whether an ACELP or gain-scaled extended excitation is used for encoding the voiced residual;
- c) repeating the ACELP excitation of the previous subframe at an interval of pitch for a voiced frame; and
- d) using a joint optimization of adaptive codebook excitation and extended excitation to optimize the speech or audio quality attained.



5. A stochastic excitation coding method comprising the steps of:

- a) partitioning a signal frame into four subframes each having a residual signal;
- b) coding the residual signal of each subframe after adaptive codebook coding by treating the residual signal as a secondary target vector input to a stochastic codebook excitation module;
- c) obtaining a best gain and excitation vector for each subframe by minimizing a least mean square error signal;
- d) linearly quantizing the gain of the initial subframe according to a predetermined number of levels, the gains of the subsequent three subframes being differentially encoded using a lesser number of levels; and
- e) sequentially optimizing the shape vector for each of these quantized gain levels.

6. An extended excitation coding method for coding of a voiced signal, said method comprising the steps of:

- a) checking the pitch period of said voiced signal, wherein said voiced signal is classified as mode 1 if the pitch period is less than or equal to one-quarter of the samples in a frame, or is classified as mode 2 if the pitch period exceeds one-quarter of the samples but less than or equal to half of the total number of samples per frame;
- b) identifying the best excitation vector of the previous subframe when the signal is classified as of mode 1, wherein the fixed excitation of the current subframe being composed by extension of the previous excitation at multiples of current subframe's pitch interval;
- c) identifying the best vectors of the two subframes, subframe 0 and subframe 1, when the signal is classified as of mode 2, the fixed excitation of subframe 2 being obtained by extension of the previous excitation at the current subframe's pitch



interval and the fixed excitation of subframe 3 is obtained by a corresponding procedure by operating on subframe 1 and subframe 2; and

d) identifying optimal gain levels for the extended fixed excitations through a sequential optimisation procedure.

7. A method for joint optimization for adaptive and extended excitation of audio or speech signals, said method comprising the steps of:

a) sequentially computing an adaptive codebook excitation and a secondary excitation, and

b) performing joint optimization only when the secondary excitation is by extended excitation, said joint optimization including defining a joint mean square error, wherein the optimized gain level for extended excitation, pitch lag and adaptive codebook gain vector can be obtained such as to minimise the said joint mean square error.

8. A method according to claim 7, wherein the secondary excitation comprises one of algebraic codebook excitation, extended excitation or stochastic excitation.

9. A transmission syntax for a codebook encoded signal decoder, said syntax comprising LSP vector index, adaptive codebook lag and tap indices, a voiced/unvoiced flag and secondary excitation information.

**ABSTRACT****AN EXCITATION ANALYSIS AND SYNTHESIS METHOD
FOR CODING OF AUDIO AND SPEECH SIGNALS**

This disclosure presents a multi-mode CELP coder (Fig. 1) which exploits a pitch synchronous concept for bitrate reduction. By classifying the signal into three modes, different coding techniques may be applied to code the secondary residual (S1.16). The coding techniques are algebraic codebook excited linear prediction (1.11), extended (1.12) and stochastic (1.13) excitation coding. Among these three coding techniques for the secondary residual (S1.16), the extended excitation (1.12) of the previous excitation is important for reduction of bitrate requirements during voiced frames. While the bitrate is reduced by extended excitation (1.12), the speech or audio quality can be maintained at a high level through using joint optimization of the adaptive codebook (1.10) and the extended excitation (1.12). The computational load only increases marginally for this joint optimization process.

Fig. 1

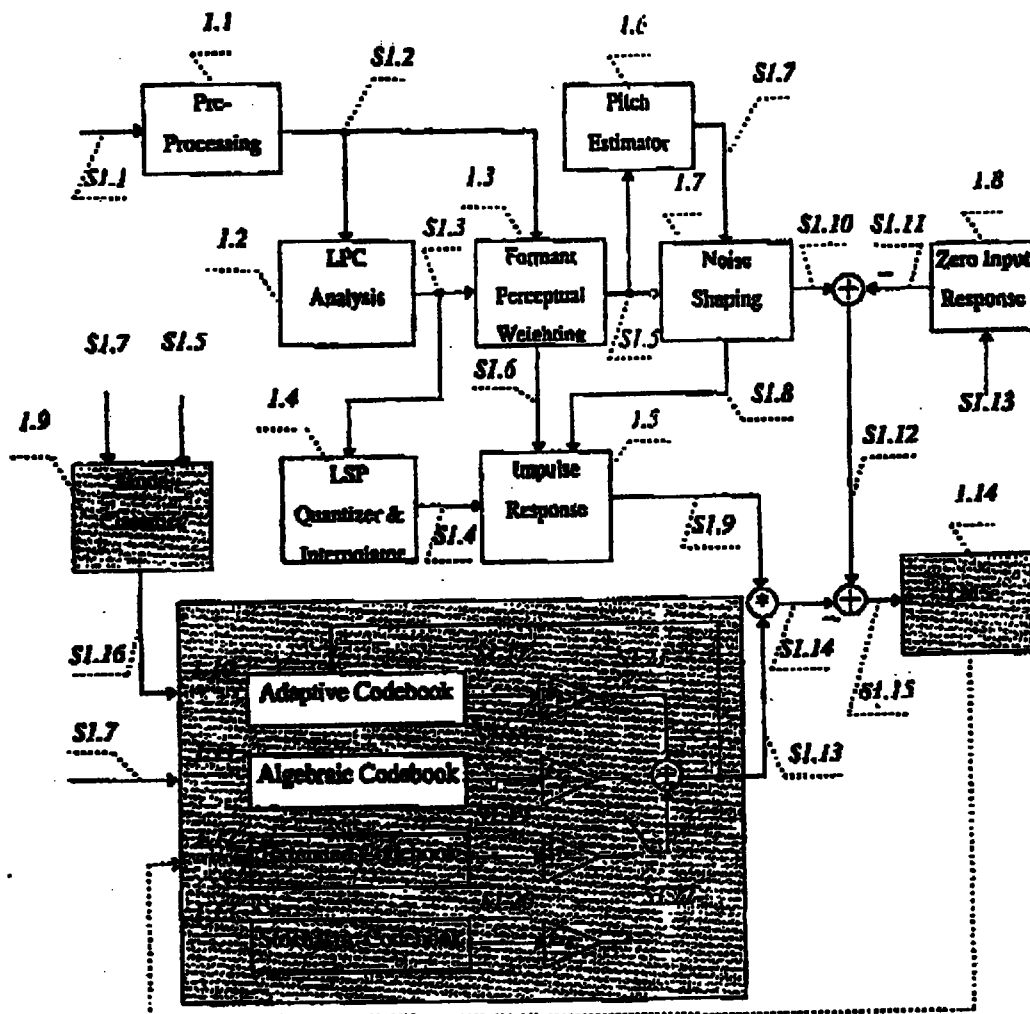


Figure 1 Block diagram of the speech encoder for current invention



2/6

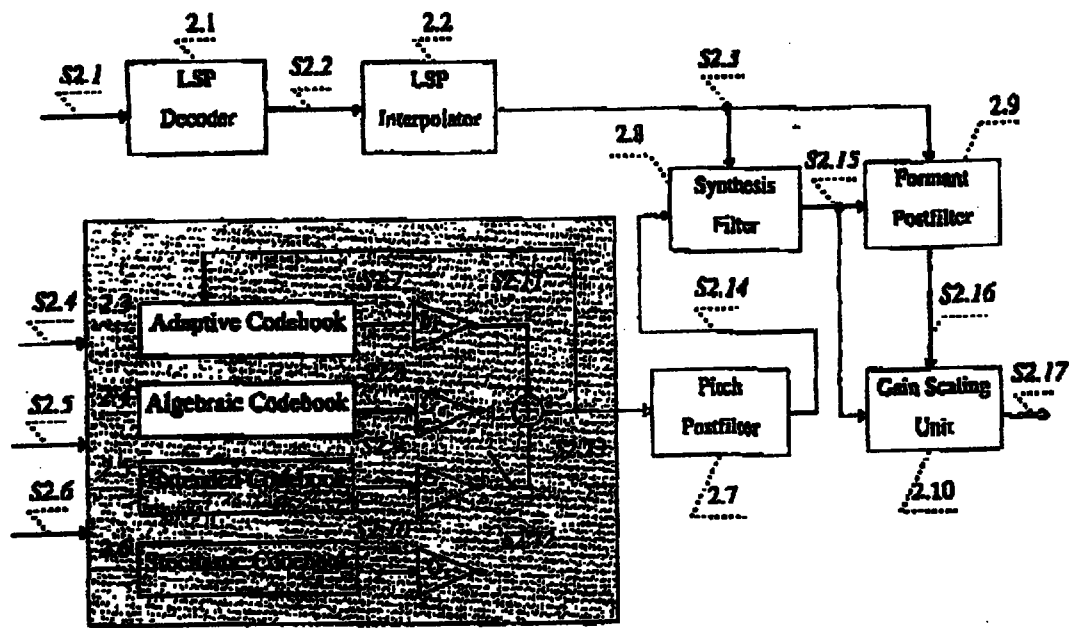
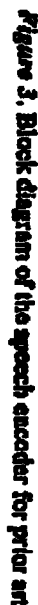


Figure 2. Block diagram of the speech decoder for current invention





4/6

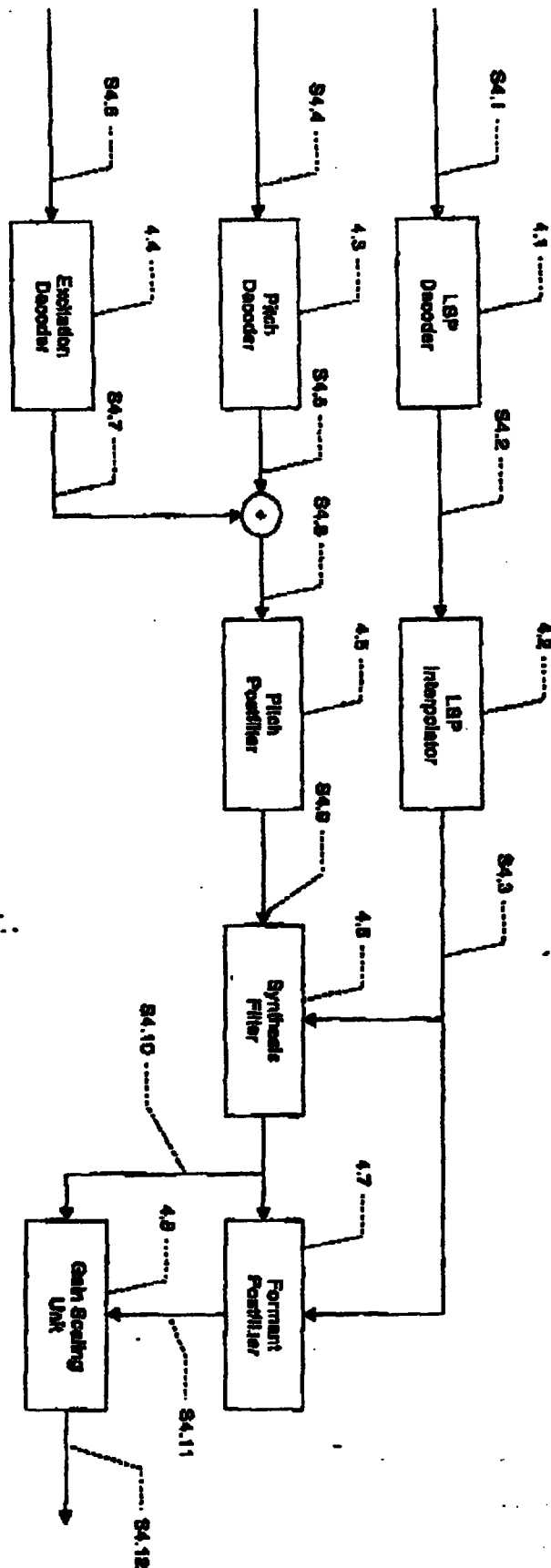


Figure 4. Block diagram of the speech decoder for prior art



5/6

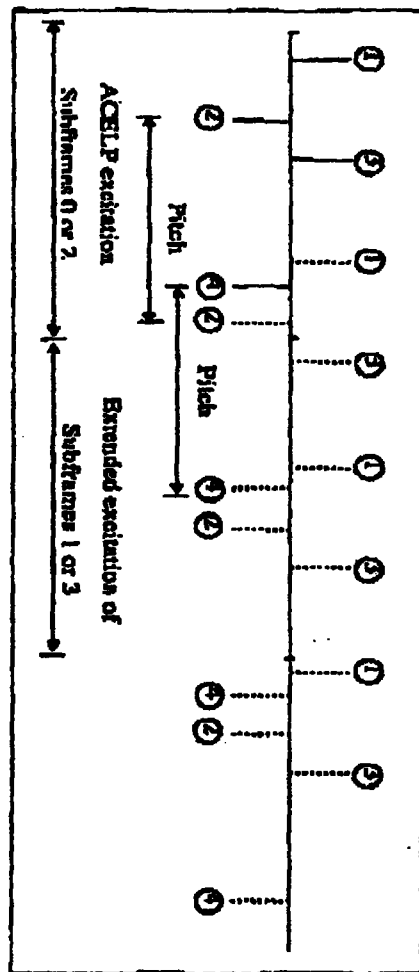


Figure-5. An illustration of the extended excitation for signal mode 1

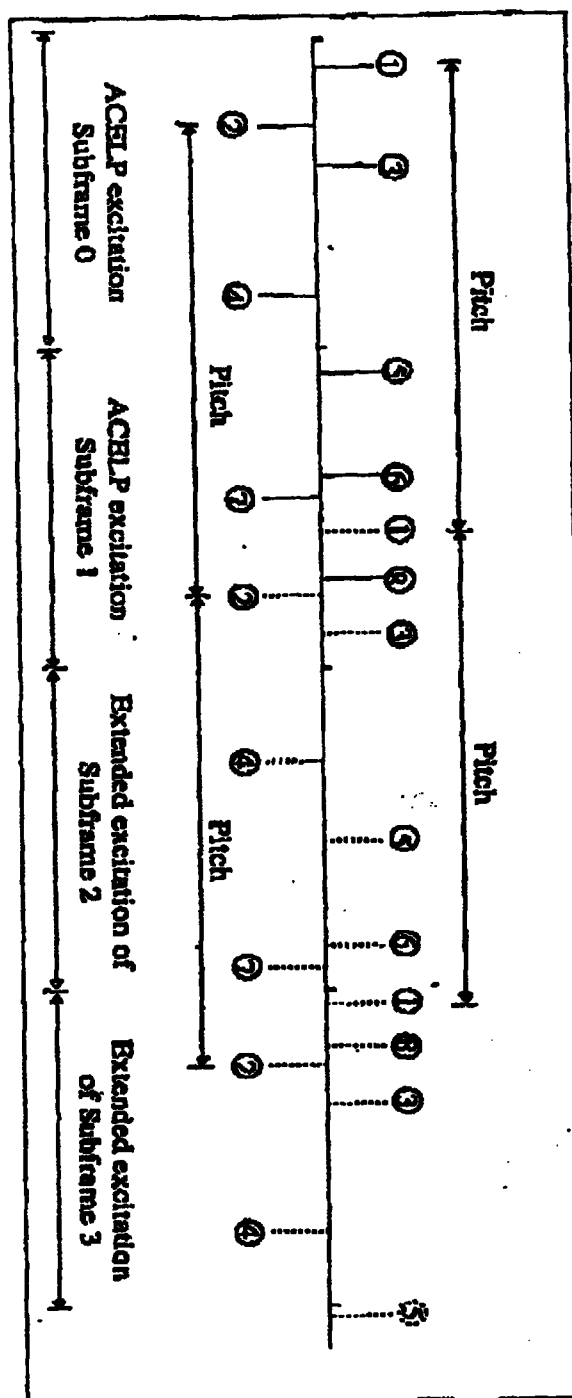
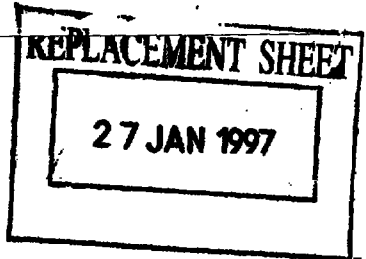


Figure 6. An illustration of the extended excitation for signal mode 2



REPLACEMENT SHEETS		SG
☒	Description	: paged .1. to paged .12.
☒	Claim(s)	: paged .13. to paged .14.
☒	Drawing(s)	: paged .15. to paged .16.
☒	Abstract	: paged .17. to paged .18.
DATE OF RECEIPT		27/1/97

**AN EXCITATION ANALYSIS AND SYNTHESIS METHOD
FOR CODING OF AUDIO AND SPEECH SIGNALS**

Field of the Invention

The present invention relates to digital coding systems and, in particular, discloses an arrangement that can be used in very low bitrate compression coding of audio and speech signals.

Background Art

Digital speech, wideband speech and low-bandwidth audio compression coding algorithms have been standardised internationally by various bodies, such as the Telecommunications Standardisation Sector of International Telecommunication Union (ITU-T) and European Technical Standards Institute. Such speech compression algorithms are currently being used in the coding of speech or audio signals of 7 kHz bandwidth and below. Various standardised algorithms which have emerged included the ITU-T recommendations e.g. G.723, G.729 and G.728. These algorithms are useful for coding of speech signals at the bitrate range of 5.3 kbits/s to 16 kbits/s. The quality of coded outputs from such algorithms range from communication quality to toll quality.

While various techniques such as Multipulse, Code Excited Linear Prediction (CELP), and Multiband Excitation have been used for speech coding, CELP has proven to be very popular. As a result, different enhancements and features have been added to the CELP scheme. The ITU-T G.723 standard is highly representative of the present state of the art. Details are reported in the ITU-T document "Draft Recommendation G.723- Dual rate speech coder for multimedia telecommunication transmitting at 5.3 kbits/s & 6.3 kbits/s" dated October 14, 1995.

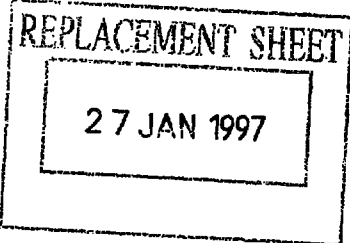
Figure 3 shows a speech encoder block diagram of the prior art. The coder first buffers the input consecutive speech samples in which the input signal is sampled at 8000 Hz and converted to a 16 bit linear PCM format. These consecutive input samples, denoted by S3.1, are buffered into frames of 240 samples by the framer 3.1, and a further partitioning into two parts of 120 samples each and a subdivision into four subframes of 60 samples is performed. The resultant output signal S3.2 is then passed

27 JAN 1997

- 2 -

into a high pass filter 3.2, whereby the dc component is removed. The filtered output signal S3.3, is subjected to a tenth order linear predictive (LP) analysis. In the LP analysis, a hamming window is first applied to a 180-sample window centred on a current subframe. Eleven autocorrelation coefficients are then computed from the windowed signal, and a correction factor follows for the first autocorrelation coefficient while the remaining coefficients are multiplied by values from a table. Altogether, four linear predictive coefficient sets are computed for the 240-sample frame. The output from the LPC analysis, S3.4, is then quantized using a line spectral pair (LSP) predictive split vector quantizer 3.4. The resultant three sub-vector output signal S3.5 is then passed through an LSP decoder 3.5, to obtain a decoded LSP vector S3.6. Linear interpolation, performed by an interpolator 3.6, is executed on a subframe basis, between a decoded LSP vector and the previous LSP vector. The four interpolated LSP vectors are then converted to LPC vectors, and a decoded speech signal is obtained through an LPC synthesis filter, also done within the interpolator.

For each subframe, the unquantized LPC coefficients signal S3.4, are used to construct a formant perceptual weighting filter 3.10, and each high pass filtered input speech subframe of signal S3.3 is filtered by the constructed filter. The formant perceptual weighting filter coefficients S3.8 are passed to the impulse response calculator 3.7. The weighted output speech signal S3.9 of the formant perceptual weighting filter 3.10 is passed to a pitch estimate 3.12 and a harmonic noise shaping filter 3.11. The harmonic noise shaping filter 3.11, provides a mechanism to improve the quality of the encoded speech. Computation of the noise shaping filter coefficients S3.10 involve the estimation of the optimum filter gain from an optimal lag. The noise shaping filter coefficients S3.10, are passed to the impulse response calculator 3.7. At the impulse response calculator 3.7, a combination filter is synthesised using the inputs from the LPC synthesis filter coefficients S3.7, the formant perceptual weighting filter coefficients S3.8, and the harmonic noise shaping filter coefficients S3.10. The coefficients of the resultant filter S3.11, are saved in a memory update unit 3.8, for processing a



reconstructed excitation sequence. The memory output S3.13 of the combination filter is passed to a zero input response module 3.9, for computing the output of the combination filter when the input signal is all zero-valued samples. The output zero-input response S3.14, is then subtracted from an output S3.16 of the harmonic noise shaping filter 3.11, producing a resulting target vector S3.17.

For every two subframes, one pitch estimate is performed in the pitch estimator 3.12, using the formant perceptually weighted speech signal S3.9. In the pitch estimator 3.12, an index which maximises a defined cross-correlation factor is selected as the estimate of the open loop pitch period.

In a pitch predictor 3.13, a fifth order pitch predictor uses the open loop pitch period indices S3.15 from the pitch estimator 3.12, to determine the closed loop pitch lags of subframes 0 and 2. These pitch lags are coded using 7 bits, and for subframes 1 and 3, the closed loop pitch lag is coded differentially using 2 bits, relative to the pitch lags of subframes 0 and 2 respectively. The pitch predictor gains are vector quantized using two codebooks. These two codebooks have 85 or 170 entries for the higher bitrate and 170 entries for the low bitrate. An output S3.19 of the pitch predictor 3.3 is subtracted from the target vector S3.17, to produce the residual signal S3.20, which is an input to the MP-MLQ/ACELP module. The pitch predictor lags S3.18, and the impulse response S3.12, of the impulse response calculator 3.7.

In the MP-MLQ/ACELP module, the quantisation of the residual signal is performed. MP-MLQ is used for quantisation at the higher bitrate and ACELP for the lower bitrate. The quantization process is such that the residual signal is approximated by an excitation vector consisting of gain factor and pulse position lag indices. These indices serve as an input S3.21. to the excitation decoder, 3.16. In the excitation decoder 3.16, the decoding of pulses is performed by first decoding a series of parameters such as a maximum gain index, position of pulses, grid position, pulse sign and in the case of high bitrate decoding, the pulse train bit. The pitch predictor lag and gain vector are decoded in a pitch decoder 3.15, and are used to generate the pitch

contribution S3.23. The pitch contribution S3.23 and the pulse contribution S3.22, are summed together to form the excitation vector S3.24, which is used to update the memory unit 3.6 of the combination filter, provided by a signal S3.11. This is achieved by passing the signal S3.24 through the filter and the memory of the combination filter is saved in the memory update module 3.8.

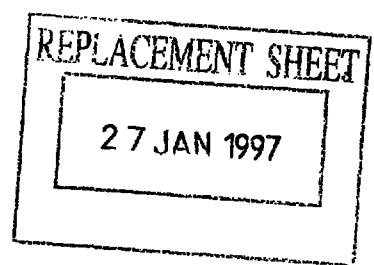
The bit allocations for the coder are used for the LPC indices, adaptive code-book lags, all the gains, pulse positions, pulse signs and grid index. The total number of bits allocated per frame amount to 192 bits and 158 bits for coding at 6.3 kbit/s and 5.3 kbit/s respectively. The bits are arranged in a defined bitstream syntax for transmission.

In the prior art decoder shown in Figure 4, the received bitstream signal is first analysed and unpacked into the respective signal inputs, namely an input S4.1 for the LSP decoder 4.1, an input S4.4 for the pitch decoder 4.3, and an input S4.6 for the excitation decoder 4.4.

For the decoder of Fig. 4, the LSP decoder 4.1, the LSP interpolator 4.2, the pitch decoder 4.3, and the excitation decoder 4.4, all operate identically to their corresponding modules in the encoder of Fig. 3, (i.e. LSP decoder 3.5, interpolator 3.6, pitch decoder 3.15 and excitation decoder 3.16 respectively). The output signals from the decoder modules, namely, S4.2, S4.3, S4.5, S4.7 and S4.8 are similar to those of S3.6, S3.7, S3.23, S3.22 and S3.24 respectively.

A pitch postfilter 4.5 takes in a complete frame excitation signal S4.8, and performs the postfiltering on a subframe basis. The pitch postfilter 4.5 serves to enhance the quality of the synthesized signal. The postfiltered signal S4.9 is fed into a tenth order LPC synthesis filter 4.6, where a speech signal S4.10, is synthesised. A formant postfilter 4.7, then performs a conventional ARMA filtering to obtain an output S4.11. This signal is gain-scaled in a gain scaling unit 4.8 to obtain the decoder output vector S4.12.

Even though the ITU-T G.723 has achieved good communication quality for speech coding at 5.3 kbit/s and 6.3kbit/s, as a result of increasing demand for lower



bitrate speech coding services, similar or better speech and low-bandwidth audio quality is sought at bitrates of 4.8 kbit/s and below, with no or marginal corresponding increase in algorithmic complexity.

Synthesis of high quality speech and low-bandwidth audio at low bitrates using linear predictive coding techniques thus requires efficient coding of both the LPC filter parameters and input excitation parameters.

It is also desirable that means to encode the excitation at low bitrates is provided to generate a representation that provides an accurate and compact representation of the excitation for speech synthesis.

As the excitation part constitutes more than 80 percent of the codec's coding bits, improvement can be obtained by a more efficient excitation analysis and synthesis method and coding.

Summary of the Invention

Various aspects of the present invention are defined in the claims appended hereto.

Specific embodiments provide a means of further bitrate reduction using a pitch synchronous concept and classifying the signal into three modes for which different coding techniques are applied. It enables the secondary residual to be coded at a lower bitrate. One of the techniques used to code the secondary residual is the extended excitation of the previous excitation, which is efficient in reducing the bit rate for the voiced frame. In order to overcome the over-periodic problem for extended excitation, a joint optimization of the adaptive codebook and the extended excitation is used.

Brief Description of the Drawings

A number of embodiments of the present invention will now be described with reference to the drawings in which:

Figure 1 shows an illustration of an encoder of the present invention;

Figure 2 shows an illustration of a decoder of the invention;

Figure 3 shows an illustration of an encoder of the prior art;

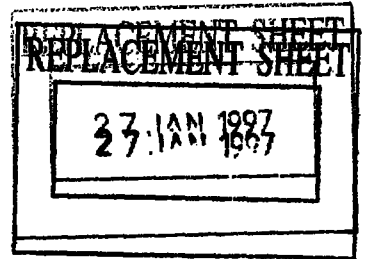


Figure 4 shows an illustration of a decoder of the prior art;

Figure 5 shows an illustration of the derivation of extended excitation from a previous ACELP excitation of 4 pulses; and

Figure 6 shows an illustration of the derivation of extended excitation from a previous ACELP excitation of 8 pulses.

Detailed Description of the Preferred Embodiments

This coder of the embodiment operates on the principles of linear prediction analysis-by-synthesis coding and uses different excitation techniques to code the residual signal. The block diagram of an exemplary encoder is shown in Figure 1. The encoder operates on blocks (frames) of 240 samples each, which is equal to 30 msec at an 8 kHz sampling rate. An input speech signal S1.1 is first buffered into frames of 240 samples. Each frame is then high-pass filtered to remove the DC component and then divided into four subframes of 60 samples each, denoted by signal S1.2. For every subframe, a tenth order linear prediction coder (LPC) filter 1.2 is computed using the pre-processed signal S1.2. The LPC filter 1.2 for the last subframe is quantized using a Predictive Split Vector Quantizer (PSVQ) 1.4. The other subframes of the LSP coefficients are linearly interpolated. The unquantized LPC coefficients are used to construct a short term perceptual weighting filter 1.3, which is used to filter the entire frame and to obtain the perceptually weighted speech signal S1.5.

The open loop pitch period signal S1.7 is computed by a pitch estimator 1.6 using a formant weighted speech signal S1.5 output from the filter 1.3, for every two subframes (120 samples). The pitch period is searched in the range from 18 to 142 samples.

From this point the speech is processed on a 60 sample per subframe basis.

Using a estimated open-loop pitch period signal S1.7 computed by the pitch estimator 1.6, a harmonic noise shaping filter 1.7, is constructed. The combination of the LPC synthesis filter 1.2, the formant perceptual weighting filter 1.3, and the harmonic noise shaping filter 1.7 is used to create an impulse response signal S1.9. The

impulse response signal S1.9 is then convolved with an excitation signal S1.13, and a zero state response signal S1.14, is obtained.

Using the pitch period estimation S1.7, and the impulse response S1.9, an adaptive codebook excitation S1.21 is computed. In adaptive codebook excitation, a fifth order pitch predictor is used. The pitch period is computed as a small differential value around the open loop pitch estimate. The contribution of the adaptive codebook S1.14, (when S1.22 is zero) is then subtracted from the initial target vector S1.12. Both the pitch period of subframes 1 and 2 and the differential values for subframes 1 and 3 are transmitted to the decoder.

Finally, the secondary residual signals are encoded by either a voiced or unvoiced excitation module according to the signal classifier's output. For the voiced case, a further classification is conducted according to the closed-loop pitch lag, which determines whether an algebraic codebook excitation (ACELP) S1.18, or gain-scaled extended excitation S1.19, is used for encoding this voiced residual accordingly. When using extended excitation, both the adaptive and extended excitation are jointly optimized. For the unvoiced case, a Gaussian stochastic codebook S1.20, is used.

The transmitted information to the receiver includes the LSP vector index, adaptive codebook lag and tap indices, 1-bit voiced/unvoiced flag, and secondary excitation information.

The decoder diagram is shown in Figure 2. The decoder operation is also performed on a frame-by-frame basis. First the quantized LPC indices S2.1, are decoded, then the decoder constructs the LPC synthesis filter 2.8. For every subframe, both the adaptive codebook excitation indices S2.6 and fixed (secondary) codebook excitation indices S2.4, are decoded and the decoded excitation is input S2.14 to the synthesis filter 2.8. For the fixed codebook excitation, the excitation selector selects the ACELP excitation 2.4, extended excitation 2.5, or stochastic excitation 2.6, according to a signal mode flag, S2.5. The adaptive postfilter consists of a formant postfilter 2.9, and a forward-backward pitch postfilter 2.7. The excitation signal, S2.13, is input to the

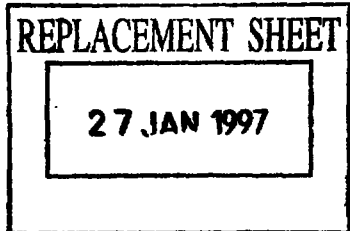
pitch postfilter 2.7, which in turn is input to the synthesis filter 2.8 whose output S2.15, is input to the formant postfilter 2.9. A gain scaling unit 2.10, maintains the energy at the input level of the formant postfilter 2.9.

Operation of the preferred embodiments of the present invention can now be explained using Figure 1, Figure 2, Figure 5 and Figure 6 which provide a method for speech coding to reduce the coding bits for voiced speech or audio.

In the encoder of Figure 1 and decoder of Figure 2, the shaded modules depict additions to the prior art. In particular, pre-processes module 1.1 in Figure 1 is the combination of both the framer 3.1 and high pass filter 3.2 of Figure 3. The LPC analysis module 1.2 of Figure 1 is similar to module 3.3 in Figure 3. Modules indicated by literals 1.3, 1.5, 1.6, 1.7 and 1.8 in Figure 1 are the same as the modules indicated by literals 3.10, 3.7, 3.12, 3.11 and 3.9 in Figure 3. Module indicated by literal 1.4 in Figure 1 is the combination of modules indicated by literals 3.4, 3.5 and 3.6 in Figure 3. In Figure 2, the modules denoted by literals 2.1, 2.2, 2.7, 2.8, 2.9 and 2.10 are the same as the modules denoted by literals 4.1, 4.2, 4.5, 4.6, 4.7 and 4.8 respectively. The algebraic codebook excitation module 1.11 of Figure 1 and 2.4 in Figure 2 are similar to the ACELP module 3.14 of Figure 3 and the excitation decoder module 4.4, in Figure 4.

Adaptive codebook excitation module 1.10 of Figure 1 is similar to pitch predictor module 3.13 of Figure 3, however the optimization procedure is slightly different between them. For the preferred embodiment, the adaptive codebook excitation optimization can be jointly conducted with the extended excitation according to the decision of the mode classifier module 1.9.

The signal mode classifier 1.9 of Figure 1 is input with the open pitch lag S1.7, from which a voiced/unvoiced indicator β_i ($i=0,1,2,3$) can be estimated by analyzing formant perceptual weighted signal S1.5. The different possible combinations of voiced and unvoiced decision of each set of consecutive subframes are shown in Table 1. In Table 1, V means voiced and U means unvoiced subframes. A final classification for



- 9 -

the entire frame is given. Based on the extended excitation strategy being used, only frames with subframe 0 identified as voiced and which also has a majority of voiced subframes may be classified as voiced frames.

Table 1
Voiced/unvoiced classification

Case	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Subframe 0	V	V	V	V	V	V	V	V	U	U	U	U	U	U	U	U
Subframe 1	V	V	V	V	U	U	U	U	V	V	V	V	U	U	U	U
Subframe 2	V	V	U	U	V	V	U	U	V	V	U	U	V	V	U	U
Subframe 3	V	U	V	U	V	U	V	U	V	U	V	U	V	U	V	U
Frame classification V/U	V	V	V	U	V	U	U	U	U	U	U	U	U	U	U	U

After the classification of voiced/unvoiced, a further classification, signal mode classification, is conducted for the voiced with respect to the pitch lag of subframe 0. The signal mode classification of the frame consists of three modes. Mode 1 is assigned if a frame is voiced and the pitch lag is less than 60 samples. Mode 2 is assigned if a frame is voiced and the pitch lag is between 60 and 120 samples. Mode 3 is assigned if a frame is unvoiced or the pitch lag is greater than 120 samples. A summary of signal mode classification is shown in Table 2.

Table 2
Signal mode classification

Cases		Signal Mode
Voiced	Pitch $\leq$ 60 (Samples)	Mode 1
Voiced	$60 \leq$ Pitch $\leq$ 120	Mode 2
Others (unvoiced & pitch $>$ 120)		Mode 3

As soon as the signal classification is finished, the corresponding techniques to encode the secondary residual signals for each mode are used. The coding techniques are shown in Table 3. If the current frame is classified as Mode 1, for instance, the secondary residuals (target signal S1.12, minus the adaptive codebook contribution) of subframes 0 and 2 are encoded by ACELP techniques 1.11, the secondary residuals of subframes 1 and 3 are encoded by extended excitation 1.12. At Mode 2, the secondary residuals of subframes 0 and 1 are encoded by ACELP techniques, the secondary

residuals of subframes 2 and 3 are encoded by extended excitation. At Mode 3, the secondary residuals of every subframe are encoded by stochastic excitation 1.13. It will be shown during extended excitation subframes, that there is no requirement to encode secondary residual parameters as these are derivable from parameters already coded as joint optimization techniques are used.

Table 3
Techniques for coding the secondary residual at each mode

Signal mode	Subframe 0	Subframe 1	Subframe 2	Subframe 3
Mode 1	ACELP	Extended	ACELP	Extended
Mode 2	ACELP	ACELP	Extended	Extended
Mode 3	Stochastic	Stochastic	Stochastic	Stochastic

The secondary residual signal S1.16, after subtracting the adaptive codebook contribution is coded by the stochastic codebook 1.13, if the signal is classified into Mode 3. The secondary residual signal is transferred as secondary target vector to the random codebook excitation module, 1.13. This module performs the quantization of this vector. The quantization process is approximately the secondary target vector by the gain G_s scaled version of signal S1.20. The best gain G_s and excitation vector S1.20 which can minimize S1.15 are found.

According to the output of the classifier 1.9, as shown in Table 2, the residuals of some subframes are coded by extending the previous ACELP encoded excitations. A further classification is made according to the pitch period, i.e. Mode 1 or Mode 2. The reason for using extended excitation 1.12 is that by forcing the secondary excitation to have a pitch structure, it is possible to match voiced speech more accurately.

The extended excitation of a current subframe S1.19, is obtained by extending the best excitation of the previous subframe with an interval equal to the current closed loop pitch, and being scaled by a gain G_e .

An illustration of the derivation of an extended excitation from a previous ACELP excitation (e.g. 4 pulses) is shown in Figure 5. The solid-line pulses are the

original ACELP excitation pulses of previous subframe. The broken-line pulses are the extended ones with an interval equal to pitch. The intervals between any two identically numbered pulses are equal to current subframe's pitch.

The extended excitation of subframe 2 is obtained by extending the previous two subframes' best excitations with multiple intervals of the current closed loop pitch, and being scaled by a gain G_e .

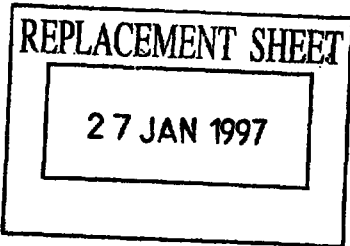
An illustration of the derivation of extended excitation from previous ACELP excitation (e.g. 8 pulses) is shown in Figure 6. The solid-line pulses are the original ACELP excitation pulses of previous subframes 0 and 1. The broken-line pulses are the extended ones with an interval equal to pitch.

For the subframe 3, the excitation of subframes 1 and 2 are regarded as the previous subframes' excitation. The same extension techniques which are applied to subframe 2 to obtain the extended excitation are similarly applied for subframe 3.

In the prior arts, the adaptive codebook excitation and the secondary (ACELP, extended, or stochastic) excitations are computed sequentially. When the adaptive excitation is optimized, the effect of the secondary excitation is assumed to be zero. Thus, after the secondary excitation is calculated, the adaptive excitation becomes sub-optimal. The reason for sequential optimisation is that the number of computations is reduced significantly.

In the preferred embodiments described herein, for the subframes with extended excitations, the secondary excitation is actually decided already except for the gain levels. Therefore, employing joint optimization does not result in a severe increase in computation load.

The shape of this extended excitation is almost decided. The shape is only decided by the closed-loop pitch lag. For example, let G_e denote the gain level for the extended excitation S1.19, S1.7 is the open loop pitch lag, b_k is the adaptive codebook gain vector, then the joint mean square error E_{joint} is minimized jointly by selecting G_e ,



- 13 -

the shift of open pitch lag $S_{l.7}$ (also referred as closed loop pitch lag) and b_k , by which the adaptive codebook excitation and extended excitation can be determined.

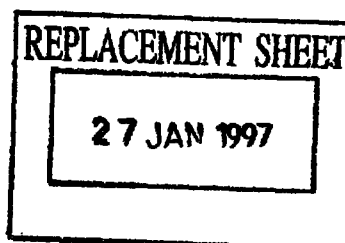
If sequential optimization is used instead of joint optimization, the processed speech sounds slightly "vibrating". The reason is that the direct extension of the previous excitation by the interval of adaptive codebook pitch lag amplifies the periodic effects. By using this joint scheme, the "vibrating" effect can be avoided.

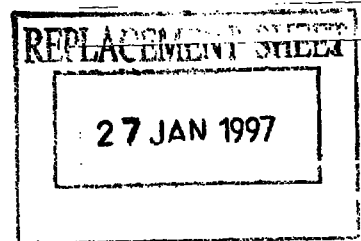
This present invention uses a more efficient representation of the input excitation of a Linear Prediction Analysis-by-Synthesis Coder, thereby resulting in the high quality speech and low-bandwidth audio coding being attainable at total bitrate of 4.8 kbit/s and below. The new excitation coding techniques outlined here are highly useful for very low bitrate compression coding of speech and low-bandwidth audio.

The foregoing describes only a number of embodiments of the present invention and modifications, obvious to those skilled in the art, can be made thereto without departing from the scope of the present invention.



1. A method of encoding audio or speech signals, said method comprising the steps of:
 - a) sampling an input audio or speech signal and grouping said samples into frames each of an equal number of said samples;
 - b) passing said frames through a high pass filter to remove any DC component, thereby producing pre-processed signal frames;
 - c) dividing each of said pre-processed signal frames into a number of subframes of equal size;
 - d) computing a set of tenth order linear predictive (LP) filter coefficients on a three subframe window centred on each subframe;
 - e) converting the LP filter coefficients into line spectral pair (LSP) coefficients to remove any long-term DC component and then vector quantizing the resultant dc-removed LSP coefficients;
 - f) decoding the quantized LSP vectors using a LSP decoder and performing linear interpolation between a decoded LSP vector and the previous LSP vector to produce 4 interpolated LSP vectors;
 - g) constructing a formant perceptual weighting filter using said set of LP coefficients generated for each subframe;
 - h) filtering each said subframe of an input frame using the perceptual weighting filter resulting in a perceptually weighted audio or speech frame;
 - i) computing at least two pitch estimates for every frame, one pitch estimate being for the first two subframes, and one other pitch estimate for the last two subframes;
 - j) subjecting the formant perceptually weighted frame to harmonic noise shaping filtering to obtain a target vector frame;





- 15 -

k) generating an impulse response filter with component inputs from the quantized LP filter coefficients, harmonic noise shaping filter coefficients and the formant perceptual weighting filter coefficients;

l) establishing a zero input response for the impulse response filter when the input signals are all zero-valued samples;

m) computing an open-loop pitch period using the formant weighted speech signal for every two subframes;

n) using the open-loop pitch period and formant perceptually weighted speech or voice as inputs to a mode classifier to obtain a voiced/unvoiced classification;

o) generating an excitation signal using excitation analysis for convolution with the impulse response to produce a zero state response; and

p) adjusting said target vector frame by subtraction of the zero state response and the zero input response and feedback of the resultant adjusted target vector through a least-mean-square algorithm for excitation analysis.

2. A method for decoding audio or speech signals encoded according to the method of claim 1, said decoding method comprising the steps of:

a) decoding linear predictive coder (LPC) indices using an LSP decoder and LSP interpolation process corresponding to those of the encoding method and constructing a LPC synthesis filter for each subframe of an input signal;

b) decoding an adaptive codebook excitation and a fixed codebook excitation and inputting the decoded excitation to the LPC synthesis filter;

c) selecting one of an algebraic codebook excitation (ACELP), extended excitation or stochastic excitation depending on a signal mode flag set;

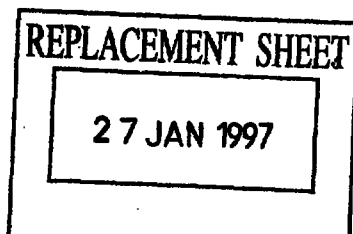
d) subjecting the decoded excitation vector to a forward-backward pitch postfilter, the output of which is fed into the LPC synthesis filter; and

e) subjecting the decoded LPC synthesis output to a formant postfilter and a gain scaling unit to provide a decoded output signal.



THIS SHEET HAS BEEN
REPLACED

3. A signal mode classification method comprising the steps of:
- a) subdividing a signal frame into at least four subframes each of an equal number of samples;
 - b) analysing the amplitude of samples in each said subframe and classifying the subframe as voiced or unvoiced according to a threshold criteria;
 - c) classifying the signal frame as an unvoiced frame when two or more of the four subframes are classified as unvoiced, otherwise the signal frame is classified as voiced;
 - d) computing a pitch lag of the signal frame;
 - e) defining two pitch lag thresholds, n_1 and n_2 , such that n_1 is less than n_2 and n_2 is less than the total number of samples per frame; and
 - f) designating the signal frame as mode 1 when the frame has been identified as a voiced frame and the pitch lag is n_1 samples or less;
 - g) designating the frame as mode 2 when the frame has been identified as a voiced frame and the pitch lag is greater than n_1 samples but less than or equal to n_2 samples; or
 - h) designating the frame as mode 3 when the frame has been identified as a voiced frame and the pitch lag exceeds n_2 samples or when the frame has been identified as an unvoiced frame.
4. A multi-mode excitation coding method comprising the steps of:
- a) selection of one of three excitation modes, namely, algebraic codebook excitation (ACELP), extended excitation and stochastic excitation, based on the mode decision of a voiced/unvoiced signal mode classification;
 - b) performing a further classification of said frame when said frame is classified as a voiced frame, said further classification being performed according to the





This Sheet Has Been Replaced



THIS SHEET HAS BEEN
REPLACED

- 17 -

pitch lag which determines whether an ACELP or gain-scaled extended excitation is used for encoding the voiced residual;

- c) repeating the ACELP excitation of the previous subframe at an interval of pitch for a voiced frame; and
- d) using a joint optimization of adaptive codebook excitation and extended excitation to optimize the speech or audio quality attained.

5. A stochastic excitation coding method comprising the steps of:

- a) partitioning a signal frame into four subframes each having a residual signal;
- b) coding the residual signal of each subframe after adaptive codebook coding by treating the residual signal as a secondary target vector input to a stochastic codebook excitation module;
- c) obtaining a best gain and excitation vector for each subframe by minimizing a least mean square error signal;
- d) linearly quantizing the gain of the initial subframe according to a predetermined number of levels, the gains of the subsequent three subframes being differentially encoded using a lesser number of levels; and
- e) sequentially optimizing the shape vector for each of these quantized gain levels.

6. An extended excitation coding method for coding of a voiced signal, said method comprising the steps of:

- a) checking the pitch period of said voiced signal, wherein said voiced signal is classified as mode 1 if the pitch period is less than or equal to one-quarter of the samples in a frame, or is classified as mode 2 if the pitch period exceeds one-quarter of the samples but less than or equal to half of the total number of samples per frame;

REPLACEMENT SHEET

27 JAN 1997



THIS SHEET HAS BEEN
REPLACED

b) identifying the best excitation vector of the previous subframe when the signal is classified as of mode 1, wherein the fixed excitation of the current subframe being composed by extension of the previous excitation at multiples of current subframe's pitch interval;

c) identifying the best vectors of the two subframes, subframe 0 and subframe 1, when the signal is classified as of mode 2, the fixed excitation of subframe 2 being obtained by extension of the previous excitation at the current subframe's pitch interval and the fixed excitation of subframe 3 is obtained by a corresponding procedure by operating on subframe 1 and subframe 2; and

d) identifying optimal gain levels for the extended fixed excitations through a sequential optimisation procedure.

7. A method for joint optimization for adaptive and extended excitation of audio or speech signals, said method comprising the steps of:

a) sequentially computing an adaptive codebook excitation and a secondary excitation, and

b) performing joint optimization only when the secondary excitation is by extended excitation, said joint optimization including defining a joint mean square error, wherein the optimized gain level for extended excitation, pitch lag and adaptive codebook gain vector can be obtained such as to minimise the said joint mean square error.

8. A method according to claim 7, wherein the secondary excitation comprises one of algebraic codebook excitation, extended excitation or stochastic excitation.

9. A transmission syntax for a codebook encoded signal decoder, said syntax comprising LSP vector index, adaptive codebook lag and tap indices, a voiced/unvoiced flag and secondary excitation information.

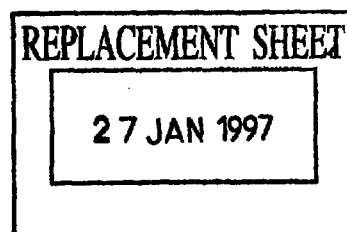
REPLACEMENT SHEET
27 JAN 1997

**AN EXCITATION ANALYSIS AND SYNTHESIS METHOD
FOR CODING OF AUDIO AND SPEECH SIGNALS**

ABSTRACT

This disclosure presents a multi-mode CELP coder (Fig. 1) which exploits a pitch synchronous concept for bitrate reduction. By classifying the signal into three modes, different coding techniques may be applied to code the secondary residual (S1.16). The coding techniques are algebraic codebook excited linear prediction (1.11), extended (1.12) and stochastic (1.13) excitation coding. Among these three coding techniques for the secondary residual (S1.16), the extended excitation (1.12) of the previous excitation is important for reduction of bitrate requirements during voiced frames. While the bitrate is reduced by extended excitation (1.12), the speech or audio quality can be maintained at a high level through using joint optimization of the adaptive codebook (1.10) and the extended excitation (1.12). The computational load only increases marginally for this joint optimization process.

Fig. 1



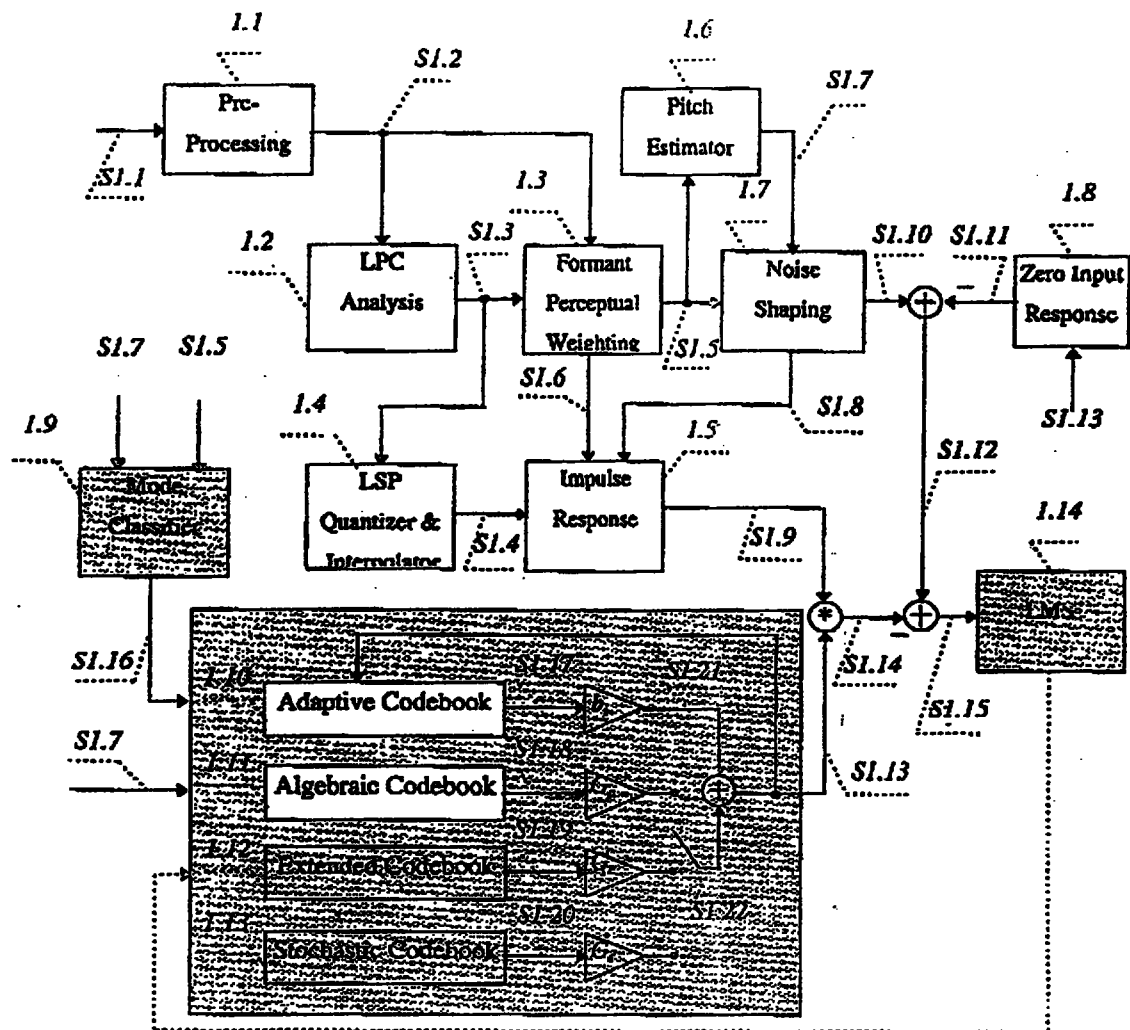
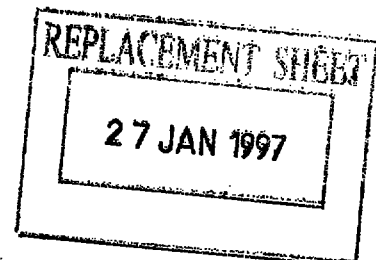


Figure 1 Block diagram of the speech encoder for current invention



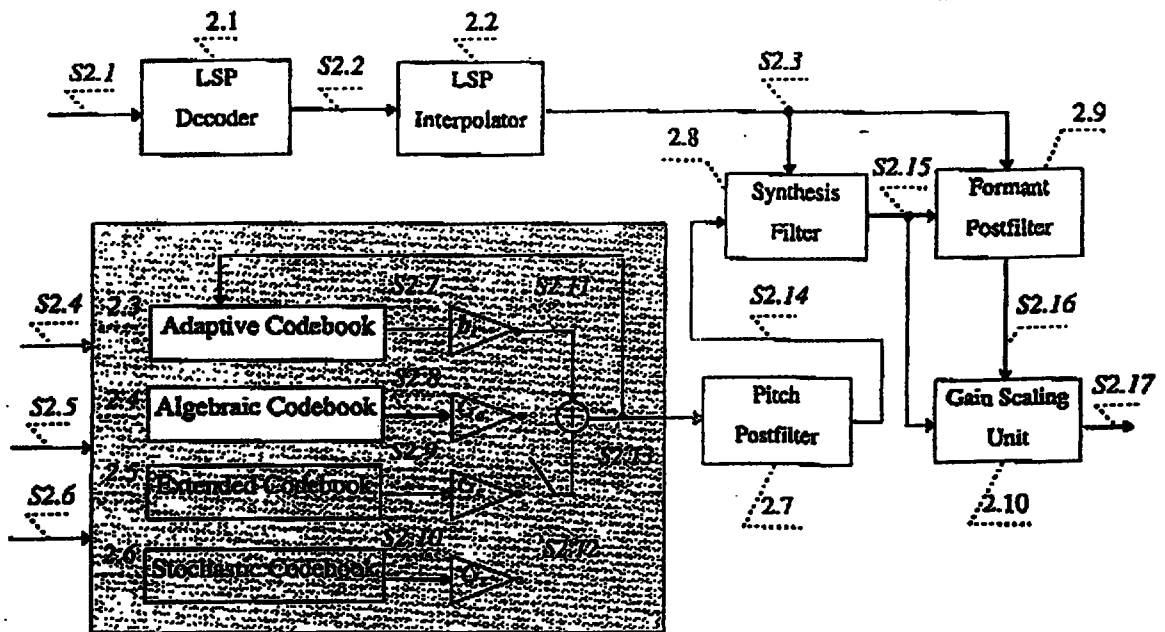
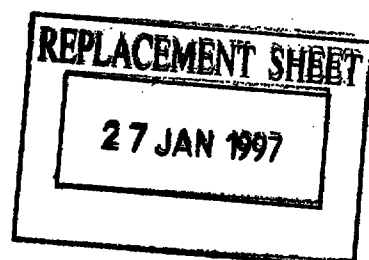


Figure 2. Block diagram of the speech decoder for current invention



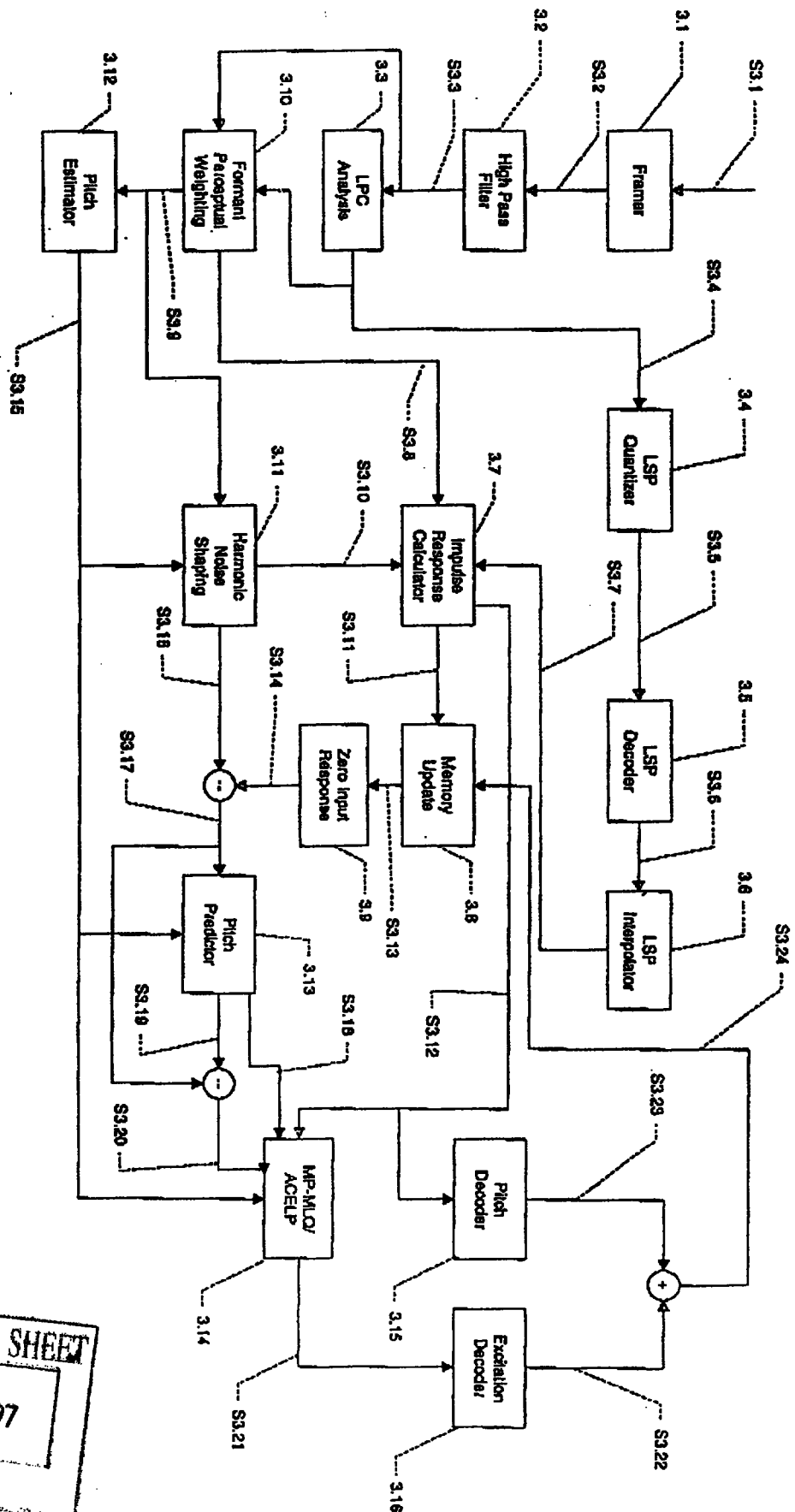
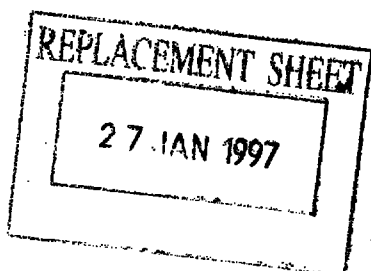


Figure 3. Block diagram of the speech encoder for prior art



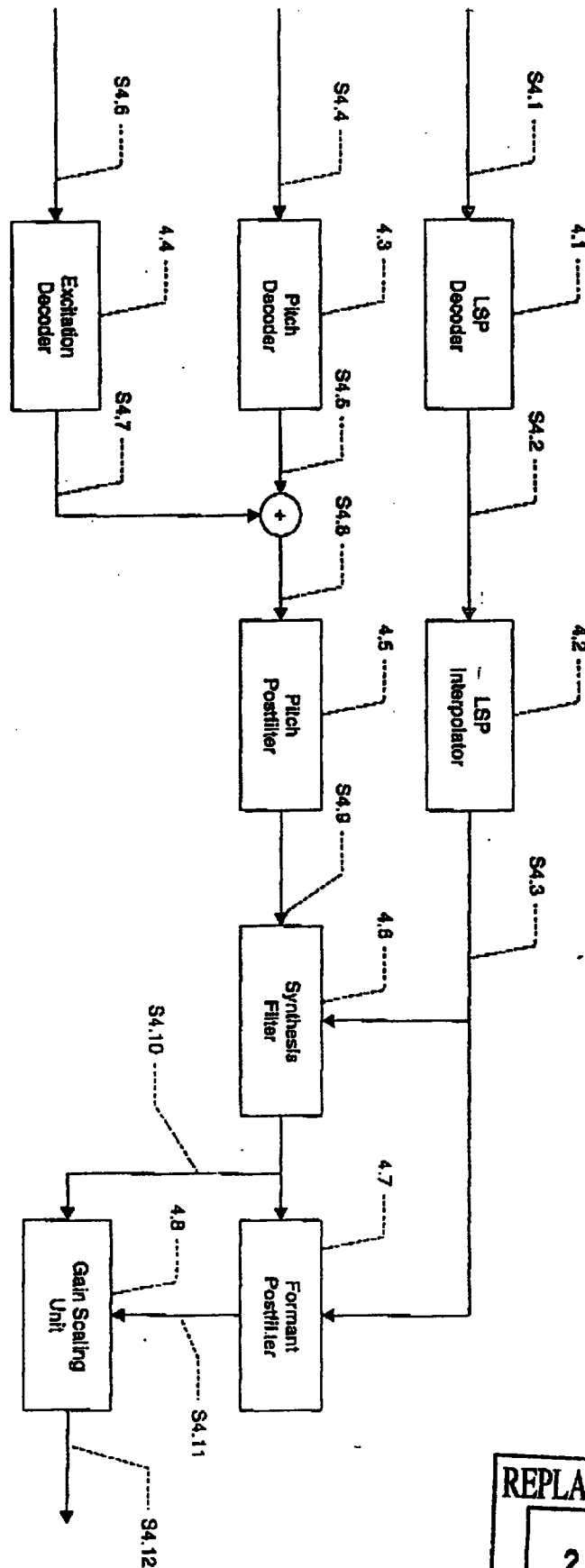


Figure 4. Block diagram of the speech decoder for prior art

REPLACEMENT SHEET

27 JAN 1997

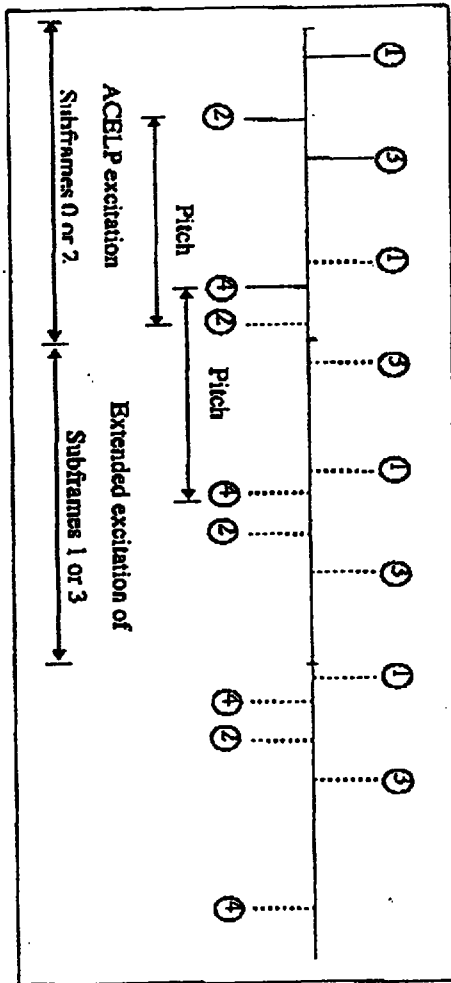


Figure-5. An illustration of the extended excitation for signal mode 1

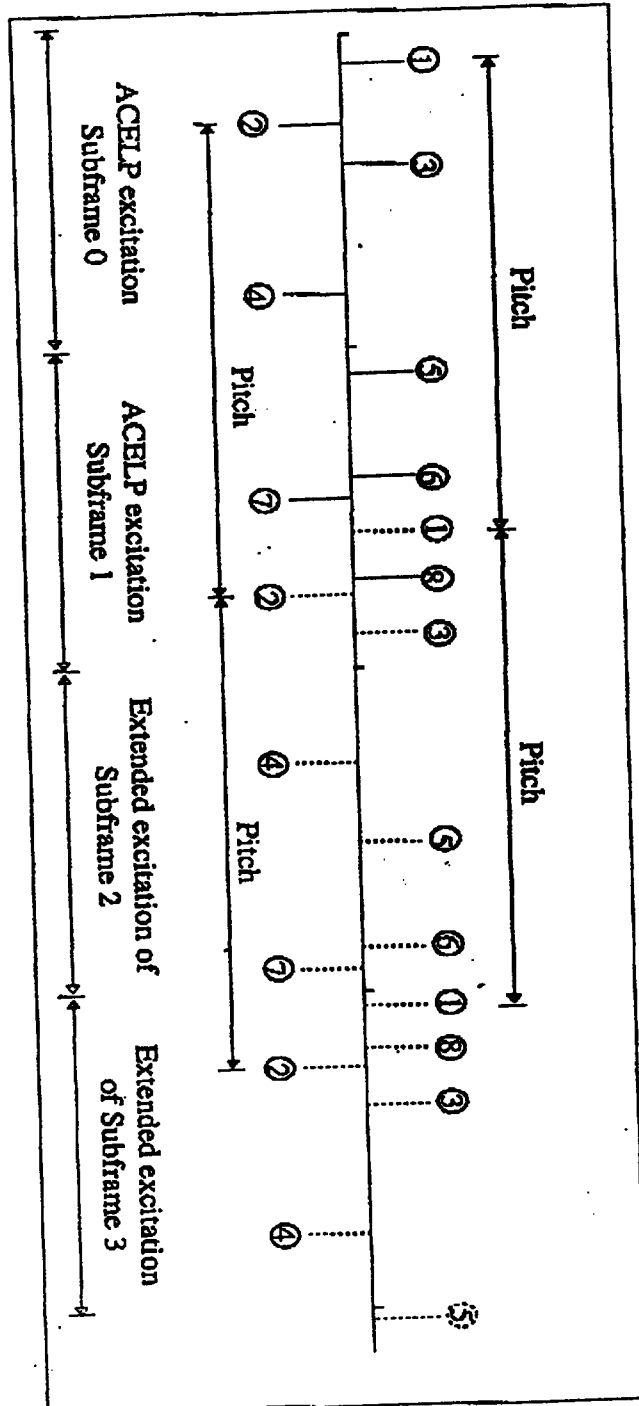
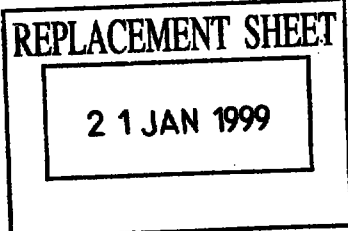


Figure-6. An illustration of the extended excitation for signal mode 2

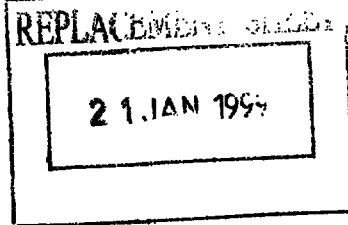


9110658-8

16

3. A multi-mode excitation coding method based on a signal classification criteria comprising the steps of:
- (a) subdividing a signal frame into at least four subframes each of an equal number of samples;
 - (b) analysing the amplitude of samples in each said subframe and classifying the subframe as voiced or unvoiced according to a threshold criteria;
 - (c) classifying the signal frame as an unvoiced frame when two or more of the four subframes are classified as unvoiced, otherwise the signal frame is classified as voiced;
 - (d) computing a pitch lag of the signal frame;
 - (e) selection of one of three excitation modes, namely, algebraic codebook excitation (ACELP), extended excitation and stochastic excitation, based on the mode decision of a voiced/unvoiced signal mode of step (c);
 - (f) performing a further classification of said frame when said frame is classified as a voiced frame, said further classification being performed according to the pitch lag obtained in step d) and which determines whether an ACELP or gain scaled extended excitation is used for encoding the voiced residual;
 - (g) repeating the ACELP excitation of the previous subframe at an interval of pitch for a voiced frame; and
 - (h) using a joint optimization of adaptive codebook excitation and extended excitation to optimize the speech or audio quality attained.
4. A stochastic excitation coding method comprising the steps of:
- (a) partitioning a signal frame into four subframes each having a residual signal;
 - (b) coding the residual signal of each subframe after adaptive codebook coding by treating the residual signal as a secondary target vector input to a stochastic codebook excitation module;
 - (c) obtaining a best gain and excitation vector for each subframe by minimizing a least mean square error signal;

REPLACEMENT SHEETS		SG
☐	Description	: paged to paged
☒	Claim(s)	: paged 11.. to paged 11..
☐	Drawing(s)	: paged to paged
☐	Abstract	: paged to paged
DATE OF RECEIPT 21/1/99.....		



- (d) linearly quantizing the gain of the initial subframe according to a predetermined number of levels, the gains of the subsequent three subframes being differentially encoded using a lesser number of levels; and
 - (e) sequentially optimizing the shape vector for each of these quantized gain levels.
5. An extended excitation coding method for coding of a voiced signal, said method comprising the steps of:
- (a) checking the pitch period of said voiced signal, wherein said voiced signal is classified as mode 1 if the pitch period is less than or equal to one-quarter of the samples in a frame, or is classified as mode 2 if the pitch period exceeds one-quarter of the samples but less than or equal to half of the total number of samples per frame;
 - (b) identifying the best excitation vector of the previous subframe when the signal is classified as of mode 1, wherein the fixed excitation of the current subframe being composed by extension of the previous excitation at multiples of current subframe's pitch interval;
 - (c) identifying the best vectors of the two subframes, subframe 0 and subframe 1, when the signal is classified as of mode 2, the fixed excitation at the current subframe's pitch interval and the fixed excitation of subframe 3 is obtained by a corresponding procedure by operating on subframe 1 and subframe 2; and
 - (d) identifying optimal gain levels for the extended fixed excitations through a sequential optimisation procedure.
6. A method for joint optimization for adaptive and extended excitation of audio or speech signals, said method comprising the steps of:
- (a) sequentially computing an adaptive codebook excitation and a secondary excitation, and
 - (b) performing joint optimization only when the secondary excitation is by extended excitation, said joint optimization including defining a joint mean square error, wherein the optimized gain level for extended excitation, pitch lag and adaptive codebook gain vector can be obtained such as to minimise the said joint mean square error.



7. A method according to claim 7, wherein the secondary excitation comprises one of algebraic codebook excitation, extended excitation or stochastic excitation.

