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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 October 2004 (21.10.2004)

PCT

(10) International Publication Number
WO 2004/090864 A2

(51) International Patent Classification[7]: G10L

(21) International Application Number:
PCT/IN2004/000060

(22) International Filing Date: 12 March 2004 (12.03.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
273/MUM/2003 12 March 2003 (12.03.2003) IN

(71) Applicant *(for all designated States except US)*: **THE INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY** [IN/IN]; Poway, Mumbay 400 076, Maharashtra (IN).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **RAO, Preeti** [IN/IN]; C-129, IITB Campus, Powai, Mumbai 400 076, Maharashtra (IN).

(74) Agent: **GANGULI, Prabuddha**; Vision - IPR, 103B Senate, Lokhandwala Township, Akurli Road, Kandivali (E), Mumbai 400 101 (IN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *without international search report and to be republished upon receipt of that report*

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

(54) Title: METHOD AND APPARATUS FOR THE ENCODING AND DECODING OF SPEECH

(57) Abstract: Methods and apparatus for encoding speech for communication to a decoder for reproduction of the speech signal where the speech signal is represented by the parameters of a speech model, and a specific quantisation' scheme is used for each parameter, with novel quantisation schemes for the spectral amplitudes. The spectral amplitudes are represented by line spectral frequencies (LSFs) and gain. The LSF vector is split into sub-vectors for quantisation by SNPVQ and frame-fill interpolation. The low-frequency split vector is quantised by an SN-PVQ scheme, and the high frequency split vector by SN-PVQ in the even-numbered frames and frame-fill interpolation in the odd-numbered frames. Optionally all LSF sub-vectors can be quantised by SN-PVQ. Further, the gain parameters of two frames are jointly quantised. These result in a system of encoder and decoder for speech coding with communication quality output speech at bit rates below 2 kbps.

WO 2004/090864 A2

METHOD AND APPARATUS FOR THE ENCODING AND DECODING OF SPEECH

Field of the Invention

This invention relates to speech coding techniques, and specifically to harmonic coding of narrowband speech at fixed bit rates less than 2 kbps. More particularly, the invention relates to the efficient quantisation of the spectral amplitude parameters of the speech signal.

Background of the Invention

A speech compression system comprises of an encoder and a decoder, each of which is typically a hardware unit such as integrated circuit or digital signal processor which is designed/programmed to realise a specific set of signal processing operations (the method or algorithm) on the incoming sampled speech signal. Speech coding research has been an active field for over three decades resulting in a number of internationally standardized methods for the compression of narrowband speech at various bit rates from 64 kbps down to 2.4 kbps [1]. Many of the standardized speech coding methods have been implemented on several different hardware platforms. The choice of a particular speech coding method in any application is influenced mainly by the desired speech quality and the bit-rate. The lower the bit rate, the higher is the compression. In the low (below 8 kbps) to very low (below 2 kbps) bit rate region, there is a distinct compromise of speech quality for a reduction in bit rate. However there are a number of applications where voice compression at very low bit rates is essential for the efficient digital transmission of speech over wireless channels with poor signal-to-noise ratios (e.g. HF radio channels). Such channels require the insertion of error-correcting codes thereby reducing significantly the number of bits available for the coding of speech. In summary, there is a strong, currently unfilled need for good quality speech coding methods in the 1 to 2 kbps range.

At high bit rates, waveform approximating coders such as PCM and ADPCM provide transparent quality coding of narrowband speech at rates at and above 32 kbps. At all lower bit rates, the speech compression method known as Code Excited Linear Prediction (CELP) has been the accepted standard for achieving toll quality speech at bit rates as low as 6 kbps. CELP-based coders are based on the source-filter model of speech production, but otherwise similar in approach to waveform coders. They use an analysis-by-synthesis procedure to obtain the best match to the waveform of the target signal. However, if the bit rate is reduced to below 4 kbps, the quality of CELP coded speech drops very steeply. In this realm, "vocoders", i.e. speech coders based not on waveform matching but rather on a purely parametric description of the signal (usually derived from a speech production model) have been adopted in various applications. Vocoders are based on a model of speech production, and represent the speech signal in terms of the parameters of a chosen model. They typically use the periodic characteristics of voiced speech and the noise-like characteristics of unvoiced speech to

achieve compact parametric representations of the speech signal. The most popular vocoders today are: Harmonic coders (includes STC, MBE, HNM), Prototype Waveform Interpolation (PWI) coders and LPC-based vocoders (includes MELP). There is extensive research activity world over in developing high quality codecs based on one or other of these three main approaches at rates in the region below 4 kbps. Recently a 2.4 kbps MELP speech-coding method for embedding in appropriate hardware was selected as the U.S. Department of Defense standard for secure communications. Apart from this, there is no standardized coding method for bit rates at and below 4 kbps. There are no standardized codecs available at the bit rates below 2 kbps. However there have been various methods proposed in the literature for coding speech at such low bit rates. All these are based on one or another of the three vocoder categories mentioned above. The MBE vocoder uses a flexible voicing structure which allows it to produce natural sounding speech, and which makes it more robust to the presence of acoustic background noise. These properties have caused the MBE speech model to be employed in a number of commercial mobile communication applications. Although MBE based speech coders have been used in various applications, a number of problems have been identified with the speech quality at very low bit rates. These problems can be attributed chiefly to the large quantisation errors of the spectral amplitudes due to an insufficient number of bits. In the present invention we use the MBE speech model to design a good quality codec at low bit rates.

Figure 1 shows the main functional blocks of a speech coding system comprising the Encoder device and the Decoder device. The functional modules of the encoder are (1) analysis of speech to estimate the parameters and (2) quantisation and encoding of the parameters to get the bit stream. The functional modules of the decoder are (3) decoding and dequantisation of the parameters from the bit stream, and (4) synthesis of reconstructed speech from the dequantised parameters. In the case of the MBE vocoder, methods for *analysis i.e. module 1 and synthesis i.e. module 4 are known. However methods described in the prior art do not meet the specific target bit rate below 2 kbps. The present invention addresses the weaknesses in the prior art and provides novel effective quantisation methods i.e. modules 2 and 3 thereby providing a novel and efficient comprehensive process* for Multi-Band Excitation (MBE) coding of speech at very low bit rates. This process is generally applicable to any speech coding method which uses LPCs to represent the speech spectrum.

MBE Analysis and Synthesis of Speech

The MBE coder is essentially a frame-based harmonic coder with a voice/unvoiced decisions over a set of harmonic frequency bands. The unvoiced regions are synthesized using spectrally shaped random noise. A mixture of voiced and unvoiced excitation represents each frame. The publication of Griffin and Lim in 1988 and the U.S. Patent nos. 5715365, 5754974, 5701390

[2,3,4] disclose the MBE method which involves an analysis process to determine the pitch, spectral magnitudes and voicing information for each frame of input speech. The parameters are encoded and speech is synthesized from decoded parameters using regenerated phase information. MBE coders use the harmonic representation for the voiced part of the speech spectrum and noise to model the unvoiced region. Figure 2 shows a typical frame of speech and its MBE parameters' representation. The excitation parameters namely pitch and voicing, are determined for each frame of input speech. U.S. Patent no. 5,754,974 describes the procedure to estimate the spectral magnitudes at the harmonics based on the pitch and voicing information. It also presents a method to quantise and encode the voicing and spectral magnitudes and represent these in a digital bit stream for an overall speech coding rate of 3.6 kbps. U.S. Patent no. 5,701,390 presents a method to decode the parameters from the bit stream and to synthesise speech from the decoded parameters and regenerated phase. Harmonic coding may be applied directly to the harmonic speech spectrum or to the LPC residual. At low bit rates, the former approach is to be preferred since the latter involves dividing already scarce bits between the LPCs and the harmonic amplitudes of the residual.

The Multi-Band Excitation (MBE) coding method of Griffin and Lim [5] uses multiple harmonic and noise bands. The first MBE method was developed for use in an 8 kbps vocoder. Subsequently, version of an Improved MBE (IMBE) with a speech coding bit rate of 4.15 kbps was selected as a standard for the INMARSAT satellite communication system [6]. Although the original MBE coder of Griffin and Lim [5] uses a multi-band, and therefore detailed, description of the harmonic and non-harmonic regions of the spectrum, recent studies have suggested that it is sufficient to divide the spectrum into only two bands: a low harmonic band and a high non-harmonic band [7] requiring the specification of only a single voicing cut-off frequency. This simplified representation of voicing information coupled with interpolative vector quantisation of a spectral envelope has led to a communication-quality codec at 3 kbps [7]. The methods described in the prior art provide analysis, quantisation and synthesis procedures for MBE codes. However the actual bit rate achieved depends critically on the quantisation module.

In the next section, we review various quantisation methods that have been proposed in the past to reduce the bit rate of available coding systems including MBE coding.

Quantisation of MBE Parameters

Of the three types of parameters of the MBE representation (namely pitch, voicing cut-off and spectral amplitudes), the spectral amplitudes consume the largest proportion of bits. Any effort to reduce the bit rate then needs to be directed toward improving the efficiency of spectral magnitudes' coding. Since the adoption of the IMBE coder in the INMARSAT standard [6], there have been a number of research efforts to reduce the bit rate of the coder while maintaining the speech quality. The IMBE coder has a speech coding bit rate of 4.15 kbps of which the majority

of the bits (over 75%) are used to encode the spectral magnitudes by a combination of scalar and vector quantisation. The U.S. Patent no. 5,754,974 presents a method of spectral parameter quantisation that uses 57 bits per 20 ms frame to quantise the spectral amplitudes for the overall bit rate of 3.6 kbps. However to achieve our targeted bit rate of less than 2 kbps places an upper limit of 25 bits/frame for quantisation of the spectral amplitudes.

An accepted approach for the efficient low rate representation of spectral amplitudes is by the samples of a compactly modeled spectral envelope obtained by the log linear interpolation of the spectral amplitudes [8]. An LPC model is used to represent the spectral envelope. It is desirable to keep the model order as low as possible to reduce the final bit rate. Various model orders ranging from 10 to 22 have been reported in the literature as necessary to achieve various levels of perceptual accuracy in the output speech quality of harmonic coders, with good quality being possible only at the higher model orders. McAulay [8] has stated that the use of spectral warping according to a prescribed perceptual frequency scale reduces the required LP model order for good quality speech from 22 to 14. Patwardhan and Rao have found that when accompanied by frequency warping according to a mild version of the Bark scale, an all-pole model order of 12 is adequate to preserve speech quality [9].

There are many methods available for the quantisation of LPCs and gain. A traditional method of encoding the gain is by the scalar quantisation of the first-order prediction error in the logarithm of the frame gain [8]. The prediction error is adequately quantised at 5 bits/frame using a trained scalar quantisation codebook.

Methods for LPC quantisation have evolved over the past two decades to improve the efficiency of transmitting the (typically 10^{th}-order) LSF vector in LPC-based coders. Many of these methods for LSF quantisation use memoryless vector quantisation to exploit intra-frame correlation of LSF parameters. The vector quantisers themselves are designed to achieve specific trade-offs between bit rate, quality and complexity. Since the only important difference between the LSF vectors in LPC-based coders and those of harmonic coders is the length of the vector, this body of knowledge is easily extended to the problem of LSF quantisation in harmonic coders. Split-VQ and multi-stage VQ methods are widely used in standard codecs, and considered to provide good performance at 22-26 bits per frame for 10^{th} order LSFs [1]. In the context of harmonic coders, Kondoz [6] proposes a 2.4 kbps MBE coder which quantises a 10^{th}-order LSF vector representing the harmonic envelope using 26-bit split VQ. Smith et al [10] have studied a variety of approaches for the split VQ of LSFs for model orders ranging from 10 to 18, and bit allocation upwards of 25 bits/frame. However, to achieve an overall bit rate of below 2 kbps for an MBE vocoder constrains the bit allocation for LSFs to less than 20 bits/frame. *Thus none of the methods for direct quantisation of LSFs in the prior art are suitable for the quantisation of the LSF vector at our targeted bit rates for the following 2 reasons: the bit*

allocation is above our upper limit of 20 bits.frame, and the methods are all based on 10th order LSF vectors which is inadequate for use in the MBE vocoder.

It is necessary to provide methods that can be used to lower the bit rate of parametric speech coders.

Known approaches to bringing down the bit rate of a frame-based, fixed-rate parametric speech coder are (a) increasing the frame size, (b) exploiting interframe correlation of parameters via predictive quantisation, and (c) encoding parameters in selected frames only and reconstructing those of the remaining frames by interpolation.

The approach (a) is advocated by Brandstein [11] and McAulay [8]. Instead of the typical frame size of 20 ms, a longer frame size of 30 ms is used in order to bring down the codec bit rate. However, larger frame sizes lead to a lower quality due to the smoothing of parameter estimates over larger windows, as well as greater changes in parameter values across frames. The resulting output speech quality is limited even without quantisation of the parameters. Brandstein [11] claims to achieve a poor-to-fair quality (comparable to LPC-10e with its quality of 2.2 MOS) with his 1.5 kbps MBE vocoder thereby making it unsuitable in the working range i.e. < 2kbps at the desired quality level.

As for the approach of (b): LSF parameters are typically highly correlated from frame to frame. This has prompted much research to be directed toward evolving quantisation schemes that exploit this interframe correlation effectively. Interframe coding typically involves the prediction of LSF parameters based on previous frames. Predictive coding, in which the error between an input vector and its predicted value from a previous encoded vector is quantised, shows good performance for highly correlated frames but performs worse than regular vector quantisation (VQ) for the occasional low correlation frames. This leads to perceptible degradation even when the average spectral distortion is low. Further, interframe coding suffers from the drawback that speech quality degrades sharply in noisy channels due to the associated error propagation. Based on a study of different interframe LSF coding schemes, Eriksson and Linden [12] proposed the Safety Net – Predictive VQ (SN-PVQ) scheme which combines the benefits of bit rate reduction due to interframe prediction with the ability to encode outliers and channel error robustness of regular (memoryless) VQ of LSFs. They have shown via experiments on an LPC coder that a savings of 4-5 bits per frame is possible choosing SN-PVQ over memoryless VQ for the quantisation of 10^{th}-order LSF vectors. The principle of SN-PVQ is illustrated in Figure 3. For each input frame, both the SN and PVQ codebooks are searched. A flag bit is set to indicate to the decoder which mode is chosen depending on whether SN or PVQ minimizes the distortion with respect to the input LSF vector. In the context of harmonic coders, Cho et al [13] have applied an SN-PVQ type scheme on a LPC-residual spectral amplitudes' vector in a sinusoidal speech coding scheme. This however is the direct

quantisation of the spectral amplitudes rather than of LSFs resulting in the overall relatively high bit rate of 4 kbps.

The approach (c) to lowering the bit rate significantly is to drop the transmission of the parameters of alternate frames and use interpolation to reconstruct these at the decoder from the available previous and next frame parameters together with control information known as "frame-fill" bits [8]. The basic idea of frame-fill is that every alternate frame is not transmitted but interpolated as a weighted combination of the information contained in the two neighbouring frames. The dropping of spectral parameters, however, is known to result in the loss of voiced-unvoiced transitions and transient sounds such as stops and voice onsets. Increasing the bits allotted to the frame-fill helps to alleviate this to an extent. The frame-fill bits specify a particular weighting scheme, and are selected at the encoder based on a chosen distance metric. This approach essentially trades off speech quality and delay for gains in terms of lower bit rate, and has been applied in different forms to harmonic coders. Ahmadi [14] applies interpolation to all parameters of the frame in a sinusoidal coder: spectral envelope, gain, pitch, voicing. However it is accepted that to minimize the degradation in speech quality, the pitch parameter should be transmitted in every frame. Kondoz [6] has evolved a 1.2 kbps coder from an available 2.4 kbps harmonic coder by dropping the transmission of LSFs, gain and voicing parameters in alternate frames. The dropping of voicing information and/or gain and its subsequent interpolation from neighbouring frames leads to a significant degradation in speech quality. MacAulay and Quatieri [8] proposed a 4.8 kbps sinusoidal coder which quantises a 16^{th} order LSF vector using 6-bit per LSF scalar quantisation in alternate 15-ms frames. The remaining frames' LSFs are interpolated using interpolation between neighbouring frames together with frame-fill information bits. The remaining parameters are transmitted in every frame. While this method is the least detrimental to quality since frame interpolation is limited to the LSFs only, it does not achieve the needed levels of compression to be useful for a below-2 kbps speech codec. Further, it has been noted that in the context of a split VQ scheme it is best to apply interpolation to the higher LSFs rather than to the entire LSF vector or just the lower LSFs in the interest of maximizing speech quality [15].

In summary, the prior art illustrates that:

- SN-PVQ, can lead to a reduction in the bit rate while preserving the quality but that this reduction is not sufficient for the case of a very low rate coder
- There is no method of transferring the design of SN-PVQ quantisers for 10^{th}-order LSFs to the MBE coder context
- The 10^{th}-order LSF representation of the spectral amplitudes of an MBE coder would result in serious degradation of speech quality due to the high degree of modeling error.

- The actual configuration and the bit allocation for the case of frequency-warped 12th order LSF vectors by SN-PVQ are not known and need to be specifically designed for applications involving bit rate of less than 2 kbps.
- There is no example where predictive VQ has been applied to the coding of LSFs representing the harmonic magnitudes in a MBE speech coder.
- The frame-interpolation approach (c) can give an added small reduction to the bit rate without much degradation in quality if its use is limited to the higher frequency LSFs.
- There has been no attempt in prior art to develop methods combining approaches (b) and (c) so as to enable their synergistic (cooperative or combined) application. The difficulty lies in using a prediction based (i.e. quantisation of a frame requires knowledge of the past quantised frames) with an interpolation approach (quantisation of a frame requires knowledge of the next future quantised frame).

SUMMARY OF THE INVENTION

There is a strong, and currently unfilled, need for good quality speech coding methods especially in the 1.2 to 2.0 kbps range. At bit rates close to but above the upper limit of this range, the family of sinusoidal speech coders including harmonic and MBE model coders offer good quality output speech. However the *approaches in the prior art to lowering the bit rate to less than 2 kbps have resulted in degradation in speech quality.* The degradation is attributed to the inaccurate quantisation of the spectral amplitudes of the MBE model. The efficient quantisation of the spectral amplitudes represented by 12th-order frequency-warped LSFs and gain is the main object of the present invention.

In this invention, an SN-PVQ scheme is designed for the 12th-order frequency-warped LSF representation of the spectral amplitudes in an MBE vocoder. Also, interframe interpolation of higher-frequency LSFs is introduced to achieve a further lowering of the bit rate without a corresponding loss in quality. The present invention judiciously combines distinct techniques of bit rate reduction to achieve a synergistic harmonic speech coding method with communication quality performance at rates below 2 kbps.

The main object of the invention is to provide a good "communication quality", fixed-rate speech compression method operating at bit rates below 2 kbps that is capable of being embedded/ported to an appropriate hardware platform to obtain an improved speech compression system. More specifically, the present invention is directed toward reducing the bit rate required for quantisation of spectral parameters in a sinusoidal speech coder so that the overall bit rate is below 2 kbps while minimizing the loss in speech quality.

Another object of the invention is to apply SN-PVQ to the quantisation of frequency-warped 12th-orderLSFs split into two equal subvectors to achieve a reduction in the bits required for the coding of LSFs at the same spectral distortion.

Another object of the invention is to apply interframe interpolation to the high-frequency split LSF vector to achieve a reduction in bit rate without a corresponding loss in speech quality.

Yet another object of the invention is to provide a judicious combination of methods such as SN-PVQ and interframe interpolation for the quantisation of LSFs to obtain a speech quality versus bit-rate reduction that is superior to the results obtained from the independent use of these methods.

Yet another object of the invention is to jointly code the gain parameters of a pair of adjacent frames to reduce the bit rate required for the coding of the gain.

Further objects and advantages of the invention will be brought out in the following portions of the specification, wherein the detailed description is for the purpose of fully disclosing preferred embodiments of the invention without placing limitations thereon.

Thus in accordance of this invention the method of encoding comprises the steps of:

1. Processing the speech signal to divide it into speech frames each representing a fixed time interval of speech
2. Processing the speech frames to obtain the parameters of a speech model including spectral parameters;
3. Representing the spectral parameters by means of LPCs and gain
4. Converting the LPCs to LSFs
5. Quantising and encoding the LSFs by a combination of SN-PVQ and frame-fill interpolation
6. Joint quantisation of the gains of a pair of frames
7. Qantising the remaining parameters of the model

The method of decoding in accordance of this invention comprises the steps of:

a) Reconstructing the quantised LSFs from the flag and codebook indices using SN-PVQ reconstruction

b) Reconstructing the interpolated LSFs from the interpolation index and the neighbouring frames' quantised LSFs

c) Converting the LSFs to LPCs after optionally correcting for stability

d) Reconstructing the gains of two frames from the indices and the gain codebook

e) Reconstructing the remaining model parameters

f) Synthesizing a speech signal from the decoded parameters

Further the method of encoding as applied to a harmonic/sinusoidal speech coder comprises the steps of:

a) Processing the speech signal to divide it into speech frames each representing a fixed interval of speech

b) Processing the speech frames to obtain the parameters of a harmonic model of speech, namely the pitch, voicing information and spectral amplitudes

c) Qantising the pitch and voicing information by regular or differential quantisation

d) Interpolating a spectral envelope through the spectral amplitudes

e) Determining the LPCs and gain of the spectral envelope, and converting the LPCs to LSFs

f) Quantisation of the LSFs and gain

The invention generally comprises a method for quantisation of the LPCs and gain. The quantisation method of the present invention may be applied to the quantisation of spectral parameters in any sinusoidal or hybrid coding scheme that uses that the LPC representation for the spectral amplitudes.

At the Encoder, the quantisation of spectral parameters in accordance with the present invention generally comprises the following steps.

- Frequency-scale warping of the harmonic spectral amplitudes based on a mild version of the Bark scale, and interpolation to a fixed frequency interval of 20 Hz to obtain a frequency-warped spectral envelope
- Modeling the spectral envelope by gain and 12-th order LPCs converted to LSFs and split into two equal sub-vectors
- Quantisation of the low-frequency LSF split vector using a safety-net predictive VQ (SN-PVQ) scheme
- Quantisation of the high-frequency split vector either by SN-PVQ or by using a combination of frame-fill interpolation and SN-PVQ with lag=2, i.e. every even frame is quantised and every odd frame is interpolated from its two neighbouring frames using frame-fill information
- Jointly coding the gains of a pair of frames using 2-dimensional VQ with the LSF and gain codebook indices and flag bits being embedded in the bit stream.

At the Decoder, the corresponding steps for the recovery of the spectral amplitudes from the bit stream of a frame pair are:

- Reconstruction of the LSFs of the lower split vector from the flag and codebook indices of each of the frames using the SN-PVQ reconstruction with the previous frame's reconstructed LSFs

- Reconstruction of the LSFs of the higher split vector of the even frame from the corresponding flag and codebook indices using the SN-PVQ with the previous even frame's reconstructed LSFs
- Reconstruction of the LSFs of the higher split vector of the odd frame by applying the frame-fill weighting to interpolate from the corresponding LSFs of the previous and next frames
- Conversion of the LSFs to LPCs
- Obtaining the gains of the two frames from the indices and the 2-dimensional gain codebook
- Computing the spectral envelope from the LPCs and gain with the spectral amplitudes obtained by sampling the spectral envelope at the warped pitch harmonic frequencies.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be more fully understood by reference to the attached drawings, which are for illustrative purposes only. The terms in the figures are to be interpreted using the Table below.

No.	Name	Meaning
1	*Lsf*	Mean-removed input LSF vector of frame "n"
2	*LSF[n]*	Input LSF vector of frame "n"
3	$\overline{lsf[n]}$	Quantised mean-removed LSF vector of frame "n"
4	$\overline{lsf[n-1]}$	Quantised mean-removed LSF vector of frame "n-1"
5	$\overline{lsf[n-2]}$	Quantised mean-removed LSF vector of frame "n-2"
6	$\overline{LSF[n]}$	Quantised LSF vector of frame "n"
7	*Mean*	Mean LSF vector
8	*err [n]*	Prediction error vector of frame "n"
9	$\overline{err[n]}$	Quantised prediction error vector of frame "n"
10	W.E.D	Weighted Euclidean Distance
11	*epvq*	Minimum error from PVQ C.B search
12	Esnvq	Minimum error from SNVQ C.B search
13	Flag	Flag to indicate which of SN & PVQ selected
14	(a_i, b_i)	Frame-fill weighting vector

Figure 1: basic functional modules of the Encoder and Decoder devices of the MBE vocoder

Figure 2: MBE model parameters as obtained from the analysis of a frame of speech samples

Figure 3: principle of SN-PVQ quantisation of an input LSF vector.

Figure 4: flow chart for the SN-PVQ quantisation of an input LSF vector of frame "n".

Figure 5: flow chart for the frame-fill interpolative quantisation of the LSF-split 2 (i.e. higher-split) vector of an odd-numbered frame "n".

Figure 6: flow chart for the SN-PVQ quantisation with lag=2 prediction, as applied to the quantisation of the LSF-split 2 vector of an even-numbered frame "n".

Figure 7: quantisation of LSF-Split2 vector

DETAILED DESCRIPTION OF THE INVENTION

For illustrative purposes the present invention is described with reference to the Figures 1-7. However, the details of the configuration may vary without departing from the basic method as disclosed herein.

The Encoder Device

The encoder device embeds the analysis and quantisation modules. The analysis module estimates the MBE model parameters shown in Fig. 2 for each 20 ms input frame of speech as detailed in [5,6]. The parameters of the MBE speech model for each analysis frame are: the fundamental frequency, voicing decisions and harmonic amplitudes. The quantisation module directly controls the bit rate of the codec. The 3 distinct categories of parameters must each be quantised as efficiently as possible to achieve the very low target bit rate. To quantise the voicing and pitch, we use available approaches for efficient and robust quantisation. The 2-band voicing information is represented by a 3-bit index corresponding to the highest voiced frequency. The pitch is quantised using a combination of differential and regular scalar quantisation of the logarithm of the pitch [16] at 5 bits per frame.

The quantisation of the spectral amplitudes is done by the methods of this invention.

LPC modeling of spectral amplitudes

The set of spectral amplitudes of each input speech frame obtained by MBE analysis constitute a discrete harmonic spectrum which can be represented compactly by a set of linear prediction coefficients (LPCs) fitted to a smooth spectral envelope. The LPCs are computed from the spectral envelope, which is derived by the log linear interpolation of the estimated spectral amplitudes. A fixed 20 Hz interpolation interval is found to be adequate to provide a smooth spectral envelope. The frequency samples before the first harmonic and after the last harmonic are extrapolated using the slope of the neighbouring harmonic interval.

The "modeled" spectral amplitudes are then obtained by sampling the all-pole envelope at the pitch harmonics and used to reconstruct the speech frame in the MBE speech synthesis. To obtain perceptually accurate LP modeling at a low model order, frequency-scale warping is applied to the harmonic frequencies before interpolation and LP analysis. Using an analytical expression permits the realization of different degrees of warping by varying a warping parameter. Subjective listening tests have shown us that an LP model order=12 computed from a spectral envelope that has been warped according to a mild version of the Bark scale provides overall good quality speech [9]. Next the 12 LP coefficients are converted to LSFs using the standard Kabal-Ramachandran numerical method [17].

The LSFs are split into 2 equal sub-vectors, and together with the gain, are quantised to the bit stream by the methods of this invention which comprises of 3 parts as follows:

- In the first part, the LSFs are quantised by an SN-PVQ method;
- In the second part the bit allocation for the LSFs is further lowered by introducing frame interpolation of the high-frequency LSFs;
- In the third part, the gains of a pair of frames are jointly coded by 2-dim. VQ.

Quantisation of LSFs by SN-PVQ

Fig. 4 shows a flow-chart of the SN-PVQ method of coding LSFs as applied in this invention. A training set of over 86,000 LSF vectors derived from a total of 30 minutes of speech drawn from various sources was used to estimate the statistical means and first-order correlation coefficients of the LSFs and to train all the SN and PVQ codebooks. The codebooks are obtained by a training procedure in which the safety-net codebook is trained on the full database, and the memory VQ only on a selected subset of the training database consisting of vectors with high interframe correlation. Since the higher-frequency LSFs are perceptually less important than the lower-frequency LSFs, the codebook sizes for the two split vectors can be different.

The input LSF vector of dimension 12 is split into two equal subvectors. The flow-chart of Fig. 4 shows the method employed for the SN-PVQ quantisation of each of the 6-dim split LSF vectors. The pre-determined mean (mean) of the LSF vector is subtracted to get a zero-mean vector ($\overline{lsf[n]}$). The SN codebook (SNVQ C.B.) is searched to find the best matched codevector for the mean-removed LSF vector ($\overline{lsf[n]}$) based on a weighted Euclidean distance (W.E.D.) metric. The weights are chosen so that the error in the highest 3 LSFs is low relative to the remaining LSFs, i.e. the weights for LSFs 10,11 and 12 are given by 0.81, 0.64 and 0.16 respectively. In the parallel branch, an error vector, $err[n]$, is determined as the difference between the mean-removed LSF vector and its first-order predicted value as determined by multiplication of the correlation matrix (A) with the previous frame's quantised mean-removed

LSF vector. The PVQ codebook is searched to find the best matched error codevector ($\widetilde{err[n]}$). The quantised LSF vector ($\widetilde{LSF[n]}$) corresponding to each of the modes (SN and PVQ) is reconstructed, and compared with the input LSF vector. The mode that yields the minimum distortion in terms of the W.E.D. is selected and the appropriate flag bit and codebook indices are encoded in the bit stream.

This W.E.D. measure is used to decide encoding of LSF quantised by SN and PVQ and also in deciding which of SN or PVQ is selected.

The SN-PVQ scheme can be used to quantise both the LSF split vectors. The search for the best matching second split LSF vector is constrained so as to maintain the ordering property of the LSFs between the first and second split vectors. Typically a single flag bit is used to signal which mode of the two, SN and PVQ, is chosen. The selected mode is the one that minimizes the overall W.E.D. between the input LSF vector and the corresponding fully-SN quantised LSF and the fully-PVQ quantised LSF vectors.

Depending on the size of the VQ codebooks, a bit allocation as low as 20 bits/frame can be achieved by the SN-PVQ method of this invention. The flag and codebook indices are embedded in the bit stream, and are decoded to reconstruct the LSF vector by the exact inverse operations of the encoder. Checks and corrections to ensure the stability of the decoded LSFs are implemented in the decoder.

Interframe interpolation of the high-frequency LSFs

In the second step of the invention, a further reduction in the bits allotted to the coding of LSFs is achieved by incorporating frame-fill interpolation within the framework of SN-PVQ. The goal is to lower the bit rate while minimizing degradation in speech quality. The pitch, voicing and gain parameters are more important in terms of coding accuracy than the LSFs.

To reduce the bits required for the coding of the LSFs, the LSFs of every alternate frame are not encoded but interpolated from the quantised LSFs of the previous and the next frame using a frame-fill interpolation method [8]. The frame-fill interpolation is done for alternate frames (i.e. all odd numbered frames) and only for the higher-frequency split vector to keep any loss of speech quality to the minimum. The method is detailed in flow-chart of Fig. 5. The mean-removed LSF vector of frame "n" is formed. Next the weighted sum of the mean-removed quantised LSFs of the previous (n-1) and next (n+1) is computed for the entire set of frame-fill weighting options available in the interpolation codebook. In this invention, we use a 3-bit interpolation index with the codebook of weighting vectors (a_i,b_i) as follows.

[(0.125,0.875),(0.25,0.75),(0.375,0.625),(0.5,0.5),(0.625,0.375),(0.75,0.25),(0.875,0.125), (1,0)]

W.E.D. is used to determine the best matched interpolated LSF vector. The codebook index (k) of the corresponding weighting function (a_k,b_k) is selected for coding the frame-fill

information. The use of frame-fill interpolation is based on the availability of quantised LSFs of the **previous and next** frames. Hence the SN-PVQ method is no longer applicable, and any bit rate reduction achieved by frame-fill interpolation will be considerably lower in significance due to the loss of the SN-PVQ benefit. To retain both the methods: *SN-PVQ* **and** *frame-fill interpolation* for the coding of the high-frequency split LSF vector requires a significant modification in the way the SN-PVQ scheme is applied. This is detailed next.

The SN-PVQ scheme as applied to the higher split LSF vector in the even frames is converted to use prediction based on the *previous quantised* vector which implies a prediction with lag=2. The prediction VQ codebook and correlation vector are retrained accordingly. Fig. 6 shows the method of SN-PVQ with lag=2 prediction. We see that the method is similar to that of SN-PVQ with lag=1 (Fig. 4), with the important difference that now the previous quantised vector is not that of the previous frame but of the previous-to-previous frame (i.e. $\overline{lsf[n-2]}$). Further, now it is necessary to use separate flag bits for the lower- and higher-LSF split vectors.

Since correlation between LSFs decreases with increasing lag, there is a slight loss in speech quality as the lag is increased from 1 to 2 without a change in the number of bits allocated for the prediction VQ codebook. However we find that this expected loss in quality is compensated for by the introduction of separate flag bits for the two LSF split vectors.

Quantisation of frame gain

This invention implements the joint quantisation of two frame gains to more fully exploit the correlation between frame gains and bring down the bit rate. The 2-dim vector of logarithm of the gain of a frame pair (odd-even) is quantised by an 8-bit vector quantiser. The VQ codebook is obtained by prior training on the non-silence-frame gains of the training set used in the LSF training. The 8-bit VQ of frame gain pairs is found to provide the same output speech quality as the predictive quantisation of the error at 10 bits/frame-pair. The 2-dim VQ is also provides better robustness to bit errors than the predictive gain quantisation with its inherent property of error propagation.

The Decoder Device

The decoder device accepts the incoming digital bit stream, decodes and dequantises the bits corresponding to each frame pair, and reconstructs the speech samples. As seen in Fig. 1, there are 2 functional modules, namely the decoding-dequantisation module and the synthesis module, embedded in the decoder device. The voicing decisions and the pitch are reconstructed from the corresponding bits of the bit stream.

The decoding-dequantisation module reconstructs the spectral amplitudes by the following steps.

1. Decoding the digital bitstream to obtain the digital bits corresponding to the indices of SN-PVQ codebooks, interpolation and gain codebooks.
2. The LSFs of the lower split vector are reconstructed from the flag and codebook indices of each of the frames using the SN-PVQ reconstruction.
3. The LSFs of the higher split vector of the even frame are reconstructed from the corresponding flag and codebook indices using the SN-PVQ reconstruction method based on the quantised previous even frame.
4. The LSFs of the higher split vector of the odd frame are reconstructed by applying the frame-fill weighting to interpolate from the corresponding LSFs of the previous and next frames.
5. The LSFs are converted to LPCs after correcting for stability if necessary.
6. The gains of the two frames are obtained from the indices and the 2-dimensional gain codebook.
7. The spectral envelope is computed from the LPCs and gain.
8. The spectral amplitudes are obtained by sampling the spectral envelope at the frequencies corresponding to the frequency-warped pitch harmonics.

The spectral magnitudes are enhanced by a frequency-domain postfilter [18], and synthesis of speech is achieved from the reconstructed parameters by MBE synthesis.

EXAMPLES

We provide three examples of the realization of the present invention in a low bit-rate MBE speech compression system. MBE model based coders are popular in mobile communication due to their ability to produce natural sounding speech and robustness to background noise. However, at bit rates near and below 2 kbps, MBE coders suffer a serious degradation in quality due to the insufficiency of available bits for the quantisation of the spectral amplitudes. The present invention uses innovative methods to reduce the bit allocation for spectral parameters in an MBE vocoder without seriously impairing the speech quality. The baseline system in which the present invention is embedded is as follows. For each 20 ms frame of speech input, the MBE parameters are: a single voicing cut-off frequency, pitch period and a 12^{th}-order frequency-warped LSF-gain representation for the set of harmonic spectral amplitudes. The voicing cut-off frequency is represented by a 3-bit frequency band number. The pitch is quantised using a combination of differential and regular scalar quantisation of the logarithm of the pitch at 5 bits per frame. For the coding of LSFs and gain, different configurations of the methods of the present invention are presented in the form of the three examples below.

Case A:

Applying SN-PVQ to quantise the LSFs which model the spectral envelope in a harmonic coder so as to achieve a lower bit rate relative to that achievable by conventional memoryless VQ schemes without a loss in speech quality. The 12-dim LSF vector is split into two 6-dim sub-vectors, which are each quantised by an SN-PVQ scheme. The gain is quantised using log-gain prediction at 5 bits/frame. The Table A shows the resulting bit allocation scheme. The LSFs are quantised at a total of 20 bits/frame with first split vector (lower 6 LSFs) allocated 10 bits and the second split vector (higher 6 LSFs) allocated 9 bits; 1 bit is used as a flag to signal SN or PVQ depending on which mode gives the best overall distortion. The overall codec delay of this scheme is 40 ms.

Parameter	Number of bits
VUV	3
Pitch	5
Gain	5
LSF Split1	10
LSF Split2	9
LSF Flag	1

Table A. Bit allocation for Example A.

Total: 33 bits + 1 synchronisation bit => 34 bits/frame (20 ms) => 1.7 kbps

Case B:

Starting with the configuration of Example A, the bit rate is lowered further by dropping the higher split LSF vector of alternate (i.e. odd-numbered) frames. These LSFs are reconstructed using a frame-fill interpolation scheme. Now since the frames are encoded in pairs (see Fig. 7), the overall codec delay is increased to 60 ms. In order to retain the advantages of SN-PVQ for the LSFs of the non-interpolated frames, it is necessary to modify the PVQ to base it on lag=2 prediction rather than lag=1. Of the 9 bits saved in each alternate frame due to non-transmission of the second split LSF vector, 3 are allotted to frame-fill information by choosing the best matching index out of 8 possibilities, and 1 additional bit to flag SN-PVQ, thus obtaining a separate flag bit for each split of the LSF vector. The Table B show the details of the bit allocation.

Parameter	Number of bits for Odd Frame	Number of Bits for Even Frame
VUV	3	3
Pitch	5	5
Gain	5	5
LSF Split1	10	10
LSF Split2	3	9
LSF Flag	1	2

Table B. Bit allocation for Example B.

Total: 61 bits + 1 synchronisation bit => 62 bits/frame-pair (40 ms) => 1.55 kbps

Case C:

From the configuration of Example B, the bit-rate is lowered even further by combining the log gain values for 2 frames, and using 2-D VQ to quantise the pair. Table C shows the details of the bit allocation.

Parameter	Number of bits for Odd Frame	Number of Bits for Even Frame
VUV	3	3
Pitch	5	5
Gain	8	
LSF Split1	10	10
LSF Split2	3	9
LSF Flag	1	2

Table C. Bit allocation for Example C.

Total: 59 bits + 1 synchronisation bit => 60 bits/frame-pair (40 ms) => 1.5 kbps

PERFORMANCE STUDIES

To assess the significance of the present invention in the LSF and gain quantisation of the baseline MBE model speech codec, it is necessary to evaluate the obtained speech quality relative to the speech quality prior to the bit-rate reducing innovations of this invention.

Speech quality is best evaluated by formal listening tests such as Mean Opinion Score (MOS) testing. It is a subjective listening test where a large set of listeners are asked to rate the quality of speech output of the codec for a large, varied sentence set on a 5-point scale. Since subjective testing is a slow and expensive process, objective models based on human auditory

perception have been developed to predict the results of subjective testing. The most recent and advanced objective model is the ITU-T (International Telecommunications Union) recommendation P.862 known as PESQ (for Perceptual Evaluation of Speech Quality) which was standardized in 2001 [19]. The PESQ model is able to predict subjective quality with good correlation in a wide range of conditions including coding distortion. We next present some objective test results based on the current configuration of our low bit rate speech codec.

In order to evaluate the performance of our speech codec including the impact of the innovations of the present invention, we coded and decoded a large number of sentences from various sources: recorded in our laboratory, standard speech sentence databases, internet sites demonstrating commercially available codecs. All the sentences used in the testing were outside the training set for our speech codec. We carried out informal listening tests as well PESQ MOS prediction on the test sentences. A selected sample of the results is presented in Table 1.

We note that the 1.7 kbps system of Example A provides a speech quality at an average of 3.0 MOS, close to that of the U.S. Federal Standard MELP codec at the much higher rate of 2.4 kbps. It must be kept in mind that a considerable amount of testing and tuning is typically needed to bring any new coding scheme to its full capability. The MELP codec has already been through this exercise while the low rate MBE coder of this work has not. Also, we note that the lower rate systems of Examples B and C provide a speech quality close to that of the 1.7 kbps codec of Example A. This can be attributed to the fact that the degradation in quality due to lag=2 prediction (reduced correlation) is made up for by the additional flag bit. Also the 2-dim VQ of the gains achieves the same quality as that of first-order prediction at a savings of 1 bit/frame.

We observe that in all cases the PESQ MOS decrease due to bit rate reduction, if any, is within 0.1 MOS. Informal listening tests indicate that there is no perceived difference between the speech qualities of the same sample under different processing methods of Examples A, B and C.

No.	Item	Duration (sec)	Case A (1.7 kbps)	Case B (1.55 kbps)	Case C (1.5 kbps)	MELP (2.4 kbps)
1.	mixmf	97	3.02	2.99	2.98	3.07
2.	conv1	26	3.14	3.10	3.10	3.13
3.	si1367	4.5	3.28	3.22	3.20	3.42
4.	cybm1	4.2	2.88	2.90	2.80	2.88
5.	cybf2	6.8	2.82	2.72	2.76	2.94

Table 1. PESQ MOS for a set of test speech items as obtained by the three different coding techniques: A. LSFs at 20 bits/frame, gain at 5 bits/frame; B. LSFs at 17.5 bits/frame, gain at 5 bits per frame; C. LSFs at 17.5 bits/frame, gain at 4 bits/frame. (Also shown are PESQ MOS as obtained by the U.S. Federal Standard 2.4 kbps MELP codec to give a rough understanding of MOS values. MELP codec has a published subjective MOS of about 3.2.

We have shown, by example, a MBE model based speech compression method in accordance with the present invention using for each 20 ms frame of speech input: a single voicing cut-off frequency, a hybrid pitch quantiser and a 12^{th}-order frequency-warped LPC-gain representation for the set of spectral amplitudes. The invention achieves an MBE coder with a very low bit rate due to the efficient quantisation of the spectral amplitudes to less than 25 bits/20 ms frame of speech. The resulting speech quality is rated at about 3.0 PESQ MOS. The examples illustrate how memory-based VQ by way of SN-PVQ is applied to reduce the bit allocation to the LSFs to get the advantage of reduced bit rate with no accompanying loss in quality. Additionally, the SN feature provides a robustness to channel errors over that available from PVQ alone. A further reduction in the bit rate is achieved by combining frame-fill interpolation of higher-frequency LSFs with the SN-PVQ method. A better quantisation of the individual LSF split vectors due to the introduction of an additional flag bit is achieved and also a smoother time-evolution of the high-frequency LSFs is obtained which is better for the speech quality. An even further reduction in the bit rate is obtained by the joint VQ of the frame gains of two frames.

References

1. Hanzo, Sommerville and Woodard, Voice Compression and Communications, IEEE Press, 2001.
2. U.S. Patent 5,715,365: Griffin et al, Estimation of excitation parameters, Feb 1998.
3. U.S. Patent 5,754,974: Griffin et al, Spectral magnitude representation for MBE speech coders, May 1998.
4. U.S. Patent 5,701,390: Griffin et al, Synthesis of MBE-based coded speech using regenerated phase information, Dec 1997.
5. Griffin D.W., Lim J.S., "Multiband excitation vocoder," in *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 36, No 8., August 1988.
6. A.M.Kondoz, "Digital Speech For Coding Low Bit-Rate Communication Systems", Chapter 8, John Wiley, 1991
7. Nishiguchi et al, "Vector quantised MBE with simplified V/UV decision at 3.0 kbps," Proc. of ICASSP, 1993.
8. MacAulay, R.J. and Quatieri T.F, Chapter 8, in *Speech Coding and Synthesis*, Editors Kleijn W.B, Paliwal K.K., Elsevier 1995.

9. P.Patwardhan and P.Rao, "Controlling perceived degradation in spectral envelope modeling via predistortion," Proc. of the Int. Conf. on Spoken Lang. Processing, Denver, Sep., 2002.

10. Smith A.M., Rambadran T., McLaughlin M.J., "Modeling and quantization of speech magnitude spectra at low data rates-Evaluating design trade-offs," *Proc. IEEE Workshop on Speech Coding for Telecommunications Proceedings.* Sep 1997.

11. M.S. Brandstein, "A 1.5 kbps MBE Speech Coder", M.S. Thesis, Dept. of ECE, M.I.T., 1990.

12. Eriksson, Linden and Skoglund, "Exploiting interframe correlation in spectral quantisation: A study of different memory schemes," Proc. of ICASSP, 1995.

13. Cho, Villette and Kondoz, "Efficient spectral magnitude quantisation for high-quality sinusoidal speech coders," Proc. of VTC-2001 Spring, Rhodes, Greece, May 2001.

14. Ahmadi and Spanias, "Low bit rate speech coding based on an improved sinusoidal model, Speech Communication 34 (2001), pp.369-390.

15. Zinser et al, "TDVC: A high-quality, low-complexity 1.3-2.0 kbps vocoder", *Proc. IEEE Workshop on Speech Coding for Telecommunications Proceedings.* Sep 1997.

16. Eriksson and Kang, "Pitch quantisation in low bit-rate speech coding," Proc. of ICASSP, 1999.

17. Kabal and Ramachandran, "The computation of LSFs using Chebyshev polynomials," IEEE Trans. ASSP, Dec., 1986.

18. Chan and Yu, "Frequency-domain postfiltering for MBE-LPC coding of speech," Electronics Letters, vol. 32, 1996.

19. Beerends et al, "PESQ, The new ITU Standard", J.A.E.S., October 2002.

WE CLAIM

1. A novel method of coding speech signals to achieve communication quality at bit rates less than 2 kbps involving effective quantisation of spectral parameters of speech signals as obtained from frame-based analysis of speech in accordance with a speech model.
2. The method of encoding as claimed in 1. comprising the steps of:
   (a) Processing the speech signal to divide it into speech frames each representing a fixed time interval of speech
   (b) Processing the speech frames to obtain the parameters of a speech model including spectral parameters
   (c) Representing the spectral parameters by means of LPCs and gain
   (d) Converting the LPCs to LSFs
   (e) Quantising and encoding the LSFs by a combination of SN-PVQ and frame-fill interpolation
   (f) Joint quantisation of the gains of a pair of frames
   (g) Quantising the remaining parameters of the model
3. The method of decoding as claimed in 1 comprising the steps of:
   (a) Reconstructing the quantised LSFs from the flag and codebook indices using SN-PVQ reconstruction
   (b) Reconstructing the interpolated LSFs from the interpolation index and the neighbouring frames' reconstructed LSFs
   (c) Converting the LSFs to LPCs after optionally correcting for stability
   (d) Reconstructing the gains of two frames from the indices and the gain codebook
   (e) Reconstructing the remaining model parameters
   (f) Synthesizing a speech signal from the decoded parameters
4. The method of encoding and decoding as claimed in Claims 1-3 wherein the vector of LSFs is divided into sub-vectors, each of which is quantised independently, either by SN-PVQ or, by a combination of SN-PVQ and frame-fill interpolation
5. A quantisation method for the split LSF sub-vectors as claimed in Claims 1-4 comprising the steps of:
   (a) Forming the corresponding mean-removed vector
   (b) Searching the SN codebook for the best matched codevector and associated index based on a weighted Euclidean distance metric
   (c) Forming the error vector as the difference between the mean-removed vector and its first-order predicted value from the previous quantised frame
   (d) Searching the PVQ codebook to find the best matched error codevector and associated index based on a weighted Euclidean distance metric
   (e) Determining the mode that yields the minimum distortion and setting the flag bit accordingly

6. A quantisation method for the split LSF sub-vectors as claimed in Claims 1-4, for a pair of odd and even numbered frames wherein the LSFs of the even numbered frame are quantised in steps comprising of:
   (a) Forming the corresponding mean-removed vector
   (b) Searching the SN codebook for the best matched codevector and associated index based on a weighted Euclidean distance metric
   (c) Forming the error vector as the difference between the mean-removed vector and its first-order predicted value from the previous quantised even frame
   (d) Searching the PVQ codebook to find the best matched error codevector and associated index based on a weighted Euclidean distance metric
   (e) Determining the mode that yields the minimum distortion and setting the flag bit accordingly

   Followed by quantisation of the LSFs of the odd numbered frame by the steps comprising of:
   (1) Forming the weighted sum of the corresponding quantised vectors of the previous and next frames where the weights are given in the interpolation index codebook
   (2) Finding the codebook index of the coefficients that provide the best match to the odd frame's LSF vector
7. A method of quantisation of the speech signal parameters in any harmonic/sinusoidal coder comprising the steps of:
   (a) Processing the speech signal to divide it into speech frames each representing a fixed time interval of speech
   (b) Processing the speech frames to obtain the parameters of a harmonic model of speech, namely the pitch, voicing information and spectral amplitudes
   (c) Quantising the pitch and voicing information by regular or differential quantisation
   (d) Interpolating a spectral envelope through the spectral amplitudes
   (e) Determining the LPCs and gain of the spectral envelope, and converting the LPCs to LSFs
   (f) Quantisation of the LSFs and gain by SN-PVQ or by a combination of SN-PVQ and frame interpolation
8. The spectral amplitudes as claimed in Claim 7 are frequency warped prior to the interpolation of the spectral envelope
9. A method of coding speech signals as claimed in Claims 1-8 wherein a MBE speech model is used comprising the steps of:
   (a) Analysis of input speech frames with fixed frame duration in the range 20-30 ms to estimate the MBE model parameters of pitch, band voicing and spectral amplitudes
   (b) Interpolating a spectral envelope through the spectral amplitudes warped to some selected frequency scale in the range from linear scale to Bark scale

(c) Representing the spectral envelope by a gain and LPCs of order ranging from 10-22

(d) Using a split VQ scheme to quantise each of two sub-vectors of the LSF vector

(e) Using a SN-PVQ scheme for the coding of each LSF split vector with a shared single flag bit

(f) Quantisation of the pitch parameter by regular and/or differential quantisation

(g) Using a voicing cut-off band number to encode voicing

(h) Using first-order predictive coding for the frame gain

10. A method of coding speech signals as claimed in Claims 1-8 wherein a MBE speech model is used comprising the steps of:

(a) Analysis of input speech frames with frame durations in the range 20-30 ms to estimate the MBE model parameters of pitch, band voicing and spectral amplitudes

(b) Interpolating a spectral envelope through the spectral amplitudes warped to some selected frequency scale in the range from linear scale to Bark scale

(c) Representing the spectral envelope by a gain and LPCs of order ranging from 10-22

(d) Using a split VQ scheme to quantise each of two sub-vectors of the LSF vector

(e) Using an SN-PVQ scheme for the coding of the low-frequency LSF split vector

(f) Using SN-PVQ with lag-2 predictive coding and an additional flag bit for the high-frequency split LSF vector of the even frames

(g) Using frame-fill interpolation for the high-frequency split LSF vector of the odd frames

(h) Quantisation of the pitch parameter by differential and/or regular quantisation

(i) Using a voicing cut-off band number to encode voicing

(j) Using first-order predictive coding for the frame gain

11. A method of coding speech signals as claimed in Claims 1-8 wherein a MBE speech model is used comprising the steps of:

(a) Analysis of input speech frames with frame durations in the range 20-30 ms to estimate the MBE model parameters of pitch, band voicing and spectral amplitudes

(b) Interpolating a spectral envelope through the spectral amplitudes warped to some selected frequency scale in the range from linear scale to Bark scale

(c) Representing the spectral envelope by a gain and LPCs of order ranging from 10-22

(d) Using a split VQ scheme to quantise each of two sub-vectors of the LSF vector

(e) Using an SN-PVQ scheme for the coding of the low-frequency LSF split vector

(f) Using SN-PVQ with lag-2 predictive coding and an additional flag bit for the high-frequency split LSF vector of the even frames

(g) Using frame-fill interpolation for the high-frequency split LSF vector of the odd frames

(h) Quantisation of the pitch parameter by differential and/or regular quantisation

(i) Using a voicing cut-off band number to encode voicing

(j) Using 2-dimensional VQ for the joint coding of a pair of frame gains of odd and even frames

12. A method of coding speech signals as claimed in Claims 1-11 to achieve a speech codec with communication quality speech at 1.7 kbps wherein a MBE speech model is used comprising the steps of:

(a) Analysis of input speech frames with frame duration of 20 ms to obtain the MBE model parameters of pitch, band voicing and spectral magnitudes

(b) Interpolating a spectral envelope at frequency intervals of 20 Hz through the spectral amplitudes warped to a mild-Bark frequency scale

(c) Representing the spectral envelope by a gain and LPCs of order 12

(d) Using a split VQ scheme to quantise each of two equal sub-vectors of the LSF vector

(e) Using a SN-PVQ scheme for the coding of each LSF split vector with a shared single flag bit

(f) Quantisation of the pitch parameter by combined differential and regular quantisation

(g) Using a voicing cut-off band number to encode voicing

(h) Using first-order predictive coding for the frame gain

13. A method of coding speech signals as claimed in Claims 1-11 to achieve a speech codec with communication quality speech at 1.55 kbps wherein a MBE speech model is used comprising the steps of:

(a) Analysis of input speech frames with frame duration 20 ms to obtain the MBE model parameters of pitch, band voicing and spectral magnitudes

(b) Interpolating a spectral envelope at frequency intervals of 20 Hz through the spectral amplitudes warped to a mild-Bark frequency scale

(c) Representing the spectral envelope by a gain and LPCs of order 12

(d) Using a split VQ scheme to quantise each of two equal sub-vectors of the LSF vector

(e) Using an SN-PVQ scheme for the coding of the low-frequency LSF split vector

(f) Using SN-PVQ with lag-2 predictive coding and an additional flag bit for the high-frequency split LSF vector of the even frames

(g) Using frame-fill interpolation for the high-frequency split LSF vector of the odd frames

(h) Quantisation of the pitch parameter by combined differential and regular quantisation

(i) Using a voicing cut-off band number to encode voicing

(j) Using first-order predictive coding for the frame gain

14. A method of coding speech signals as claimed in Claims 1-11 to achieve a speech codec with communication quality speech at 1.5 kbps wherein a MBE speech model is used comprising the steps of:

(a) Analysis of input speech frames with frame duration 20 ms to estimate the MBE model parameters of pitch, band voicing and spectral amplitudes

(b) Interpolating a spectral envelope at frequency intervals of 20 Hz through the spectral amplitudes warped to a mild-Bark frequency scale

(c) Representing the spectral envelope by a gain and LPCs of order 12

(d) Using a split VQ scheme to quantise each of two equal sub-vectors of the LSF vector

(e) Using an SN-PVQ scheme for the coding of the low-frequency LSF split vector

(f) Using SN-PVQ with lag-2 predictive coding and an additional flag bit for the high-frequency split LSF vector of the even frames

(g) Using frame-fill interpolation for the high-frequency split LSF vector of the odd frames

(h) Quantisation of the pitch parameter by combined differential and regular quantisation

(i) Using a voicing cut-off band number to encode voicing

(j) Using 2-dimensional VQ of the frame gains of the pair of odd and even frames

15. An encoder device to achieve a digital bitstream at 1.7 kbps with embedded modules for analysis, quantisation and encoding MBE model parameters functioning in the steps comprising

(a) Analysis of input speech frames with frame duration 20 ms to estimate the MBE model parameters of pitch, band voicing and spectral magnitudes

(b) Interpolating a spectral envelope at frequency intervals of 20 Hz through the spectral amplitudes warped to a mild-Bark frequency scale

(c) Representing the spectral envelope by a gain and LPCs of order 12

(d) Using a split VQ scheme to quantise each of two equal sub-vectors of the LSF vector

(e) Using a SN-PVQ scheme for the coding of each LSF split vector with a shared single flag bit

(f) Quantisation of the pitch parameter by combined differential and regular quantisation

(g) Using a voicing cut-off band number to encode voicing

(h) Using first-order predictive coding for the frame gain

16. A decoder device to achieve communication quality reconstructed speech with embedded modules for decoding and dequantisation from the digital bitstream of Claim 15, and synthesis from MBE model parameters functioning in the steps comprising
    (a) Decoding the bitstream to obtain the digital bits corresponding to the quantised parameters of each frame
    (b) Dequantising the pitch and voicing parameters
    (c) Reconstructing the quantised LSFs by retrieving the indexed codevectors from the SN-PVQ codebooks and the flag bit
    (d) Reconstructing the gain from the previous frame's reconstructed gain and dequantised gain error
    (e) Converting the LSFs to LPCs and computing the spectral envelope
    (f) Obtaining the spectral amplitudes by sampling the spectral envelope at the frequency-warped pitch harmonics, and optionally enhancing these by postfiltering
    (g) Synthesizing speech from the reconstructed MBE model parameters
17. An encoder device to achieve a digital bitstream at 1.55 kbps with embedded modules for analysis, quantization and encoding of MBE model parameters functioning in the steps comprising
    (a) Analysis of input speech frames with frame duration 20 ms to estimate the MBE model parameters of pitch, band voicing and spectral amplitudes
    (b) Interpolating a spectral envelope at frequency intervals of 20 Hz through the spectral amplitudes warped to a mild-Bark frequency scale
    (c) Representing the spectral envelope by a gain and LPCs of order 12
    (d) Using a split VQ scheme to quantise each of two equal sub-vectors of the LSF vector
    (e) Using an SN-PVQ scheme for the coding of the low-frequency LSF split vector
    (f) Using SN-PVQ with lag-2 predictive coding and an additional flag bit for the high-frequency split LSF vector of the even frames
    (g) Using frame-fill interpolation for the high-frequency split LSF vector of the odd frames
    (h) Quantisation of the pitch parameter by combined differential and regular quantisation
    (i) Using a voicing cut-off band number to encode voicing
    (j) Using first-order predictive coding of the frame gain
18. A decoder device to achieve communication quality reconstructed speech with embedded modules for decoding and dequantisation from the digital bitstream of Claim 17, and synthesis from MBE model parameters functioning in the steps comprising

(a) Decoding the bitstream to obtain the digital bits corresponding to the quantised parameters of each frame

(b) Dequantising the pitch and voicing parameters

(c) Reconstructing the quantised LSFs of the odd and even frames by retrieving the indexed codevectors from the SN-PVQ codebooks and the flag bit

(d) Reconstructing the interpolated LSFs of the odd frames by the weighted addition of the reconstructed LSFs of the previous and next frames as given by the interpolation index

(e) Reconstructing the gain from the previous frame's reconstructed gain and dequantised gain error

(f) Converting the LSFs to LPCs and computing the spectral envelope

(g) Obtaining the spectral amplitudes by sampling the spectral envelope at the frequency-warped pitch harmonics, and optionally enhancing these by postfiltering

(h) Synthesizing speech from the reconstructed MBE model parameters

19. An encoder device to achieve a digital bitstream at 1.5 kbps with embedded modules for analysis, quantisation and encoding MBE model parameters functioning in the steps comprising

(a) Analysis of input speech frames with frame duration 20 ms to estimate the MBE model parameters of pitch, band voicing and spectral magnitudes

(b) Interpolating a spectral envelope at frequency intervals of 20 Hz through the spectral amplitudes warped to a mild-Bark frequency scale

(c) Representing the spectral envelope by a gain and LPCs of order 12

(d) Using a split VQ scheme to quantise each of two equal sub-vectors of the LSF vector

(e) Using an SN-PVQ scheme for the coding of the low-frequency LSF split vector

(f) Using SN-PVQ with lag-2 predictive coding and an additional flag bit for the high-frequency split LSF vector of the even frames

(g) Using frame-fill interpolation for the high-frequency split LSF vector of the odd frames

(h) Quantisation of the pitch parameter by combined differential and regular quantisation

(i) Using a voicing cut-off band number to encode voicing

(j) Using 2-dimensional VQ of the frame gains of the pair of odd and even frames

20. A decoder device to achieve communication quality reconstructed speech with embedded modules for decoding and dequantisation from the digital bitstream of Claim 19, and synthesis from MBE model parameters functioning in the steps comprising

(a) Decoding the bitstream to obtain the digital bits corresponding to the quantised parameters

(b) Dequantising the pitch and voicing parameters

(c) Reconstructing the quantised LSFs of the odd and even frames by retrieving the indexed codevectors from the SN-PVQ codebooks and the flag bit

(d) Reconstructing the interpolated LSFs of the odd frames by the weighted addition of the reconstructed LSFs of the previous and next frames as given by the interpolation index

(e) Reconstructing the gains of a pair of frames by retrieving the indexed codevector

(f) Converting the LSFs to LPCs and computing the spectral envelope

(g) Obtaining the spectral amplitudes by sampling the spectral envelope at the frequency-warped pitch harmonics, and optionally enhancing these by postfiltering

(h) Synthesizing speech from the reconstructed MBE model parameters

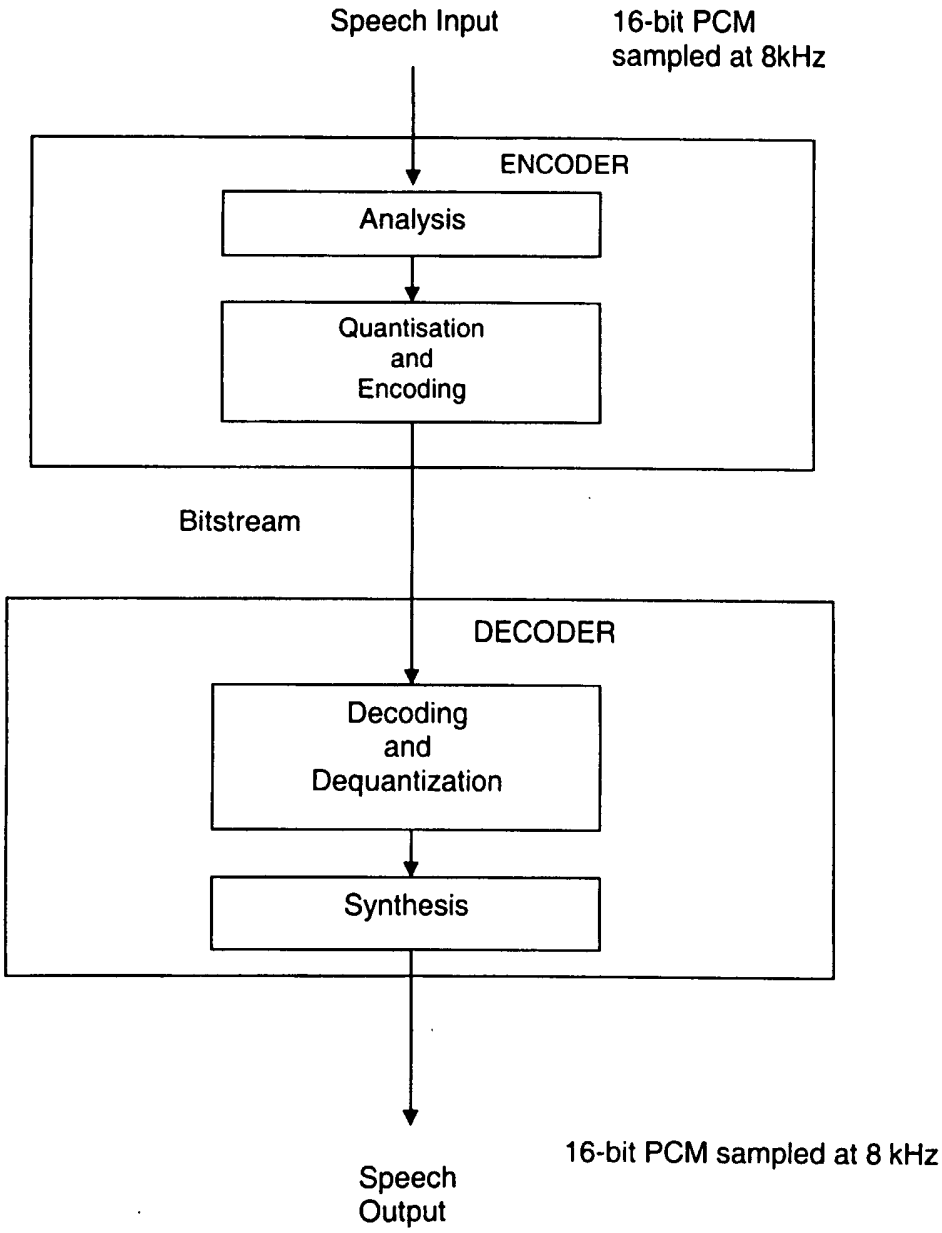
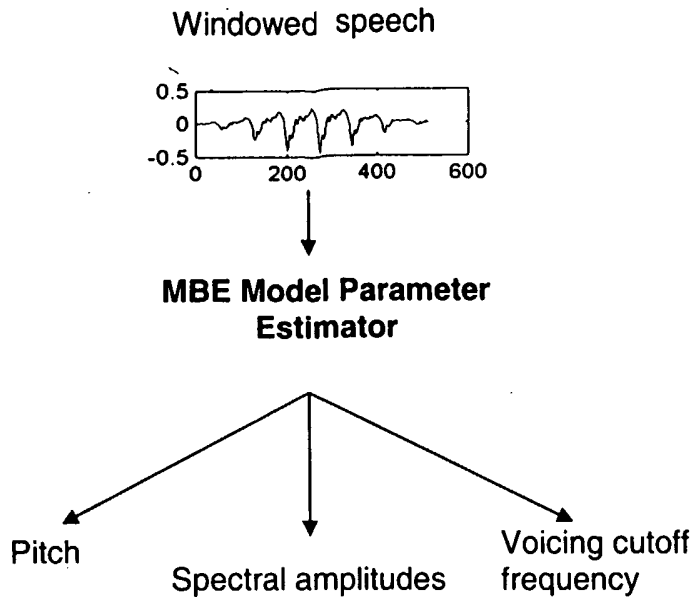

Speech Input
16-bit PCM
sampled at 8kHz
ENCODER
Analysis
Quantisation
and
Encoding
Bitstream
DECODER
Decoding
and
Dequantization
Synthesis
16-bit PCM sampled at 8 kHz
Speech
Output

Figure 1


Windowed speech
0.5
0
-0.5
0
200
400
600
MBE Model Parameter Estimator
Pitch
Spectral amplitudes
Voicing cutoff frequency

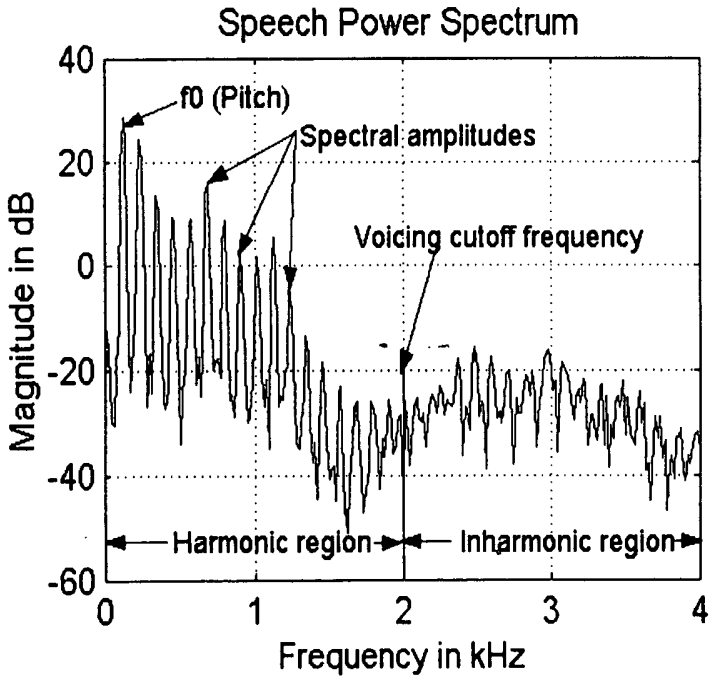

Speech Power Spectrum
f0 (Pitch)
Spectral amplitudes
Voicing cutoff frequency
Harmonic region
Inharmonic region
Magnitude in dB
Frequency in kHz
40
20
0
-20
-40
-60
0
1
2
3
4

Figure 2

Principle of SN-PVQ

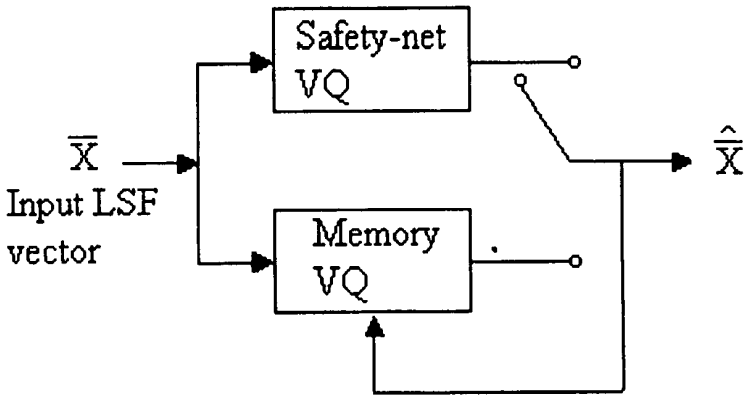


Figure 3

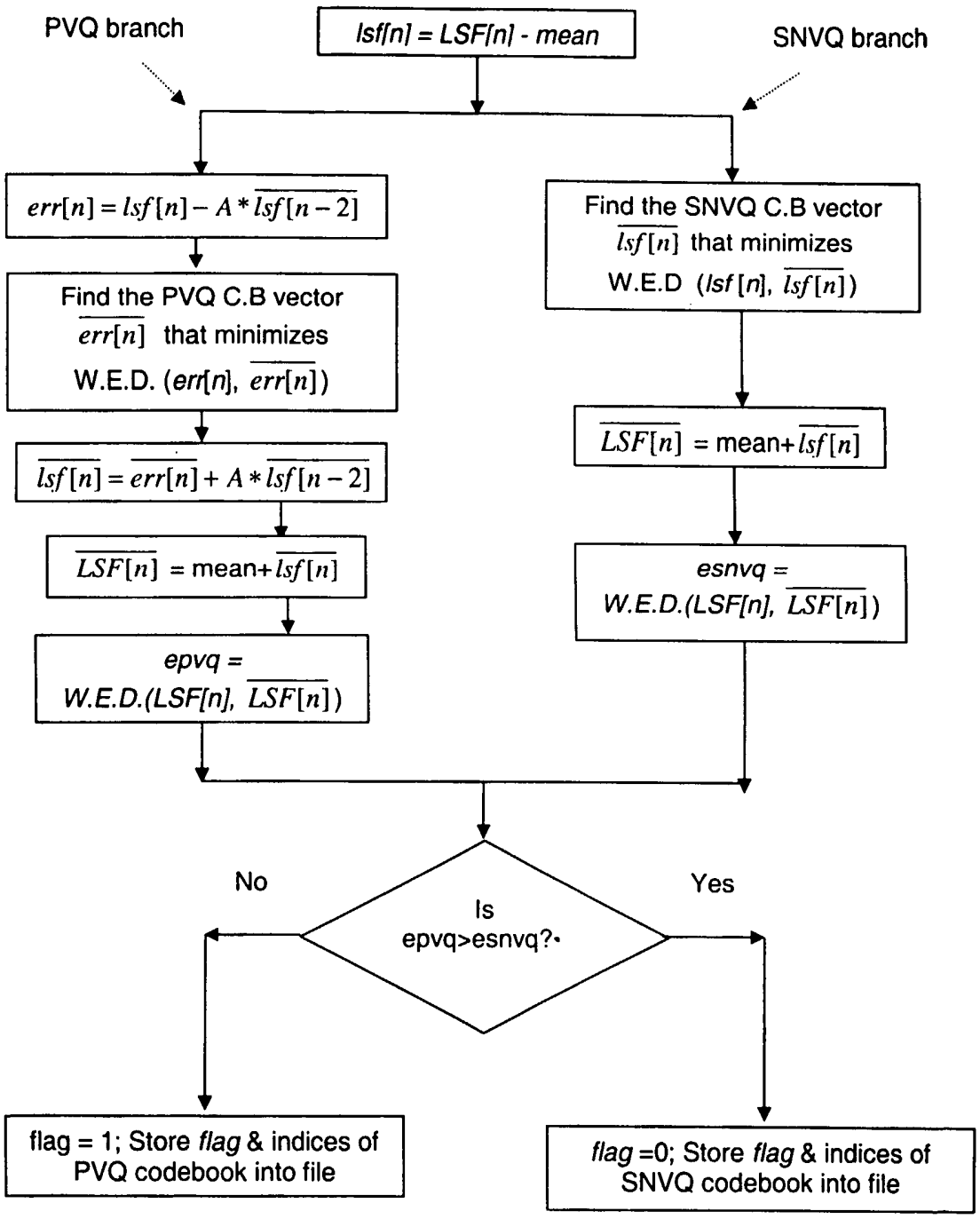

PVQ branch
lsf[n] = LSF[n] - mean
SNVQ branch
$err[n] = lsf[n] - A * \overline{lsf[n-2]}$
Find the PVQ C.B vector $\overline{err[n]}$ that minimizes W.E.D. (err[n], $\overline{err[n]}$)
$\overline{lsf[n]} = \overline{err[n]} + A * \overline{lsf[n-2]}$
$\overline{LSF[n]}$ = mean+$\overline{lsf[n]}$
epvq = W.E.D.(LSF[n], $\overline{LSF[n]}$)
Find the SNVQ C.B vector $\overline{lsf[n]}$ that minimizes W.E.D (lsf[n], $\overline{lsf[n]}$)
$\overline{LSF[n]}$ = mean+$\overline{lsf[n]}$
esnvq = W.E.D.(LSF[n], $\overline{LSF[n]}$)
No
Is epvq>esnvq?
Yes
flag = 1; Store flag & indices of PVQ codebook into file
flag =0; Store flag & indices of SNVQ codebook into file

Figure 4

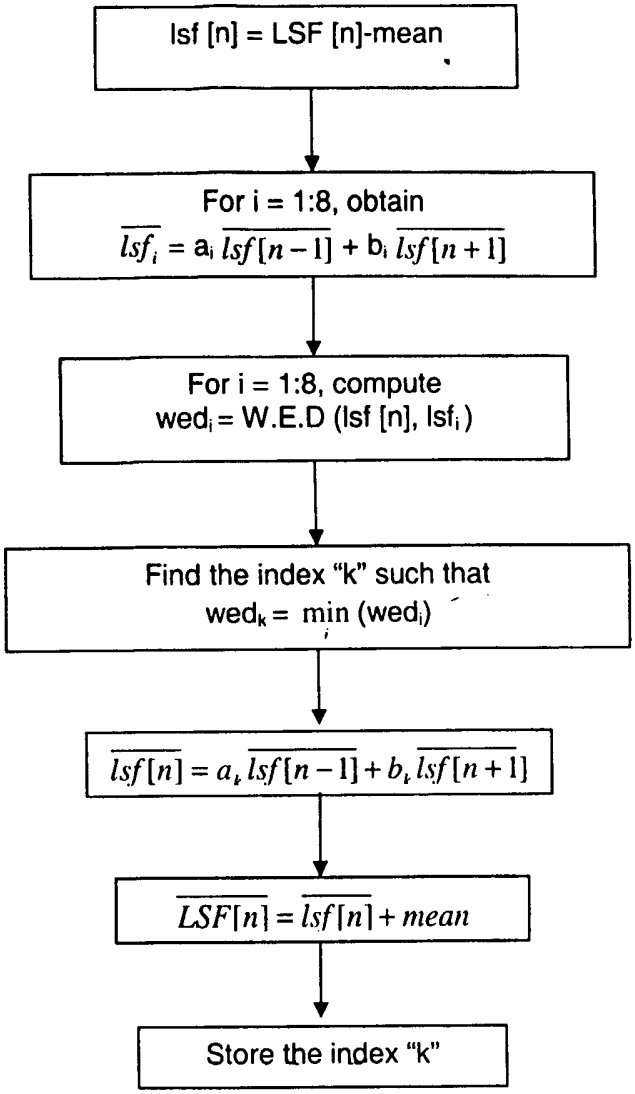

lsf [n] = LSF [n]-mean
For i = 1:8, obtain
$\overline{lsf_i} = a_i \overline{lsf[n-1]} + b_i \overline{lsf[n+1]}$
For i = 1:8, compute
wed_i = W.E.D (lsf [n], lsf_i)
Find the index "k" such that
$wed_k = \min_i (wed_i)$
$\overline{lsf[n]} = a_k \overline{lsf[n-1]} + b_k \overline{lsf[n+1]}$
$\overline{LSF[n]} = \overline{lsf[n]} + mean$
Store the index "k"

Figure 5

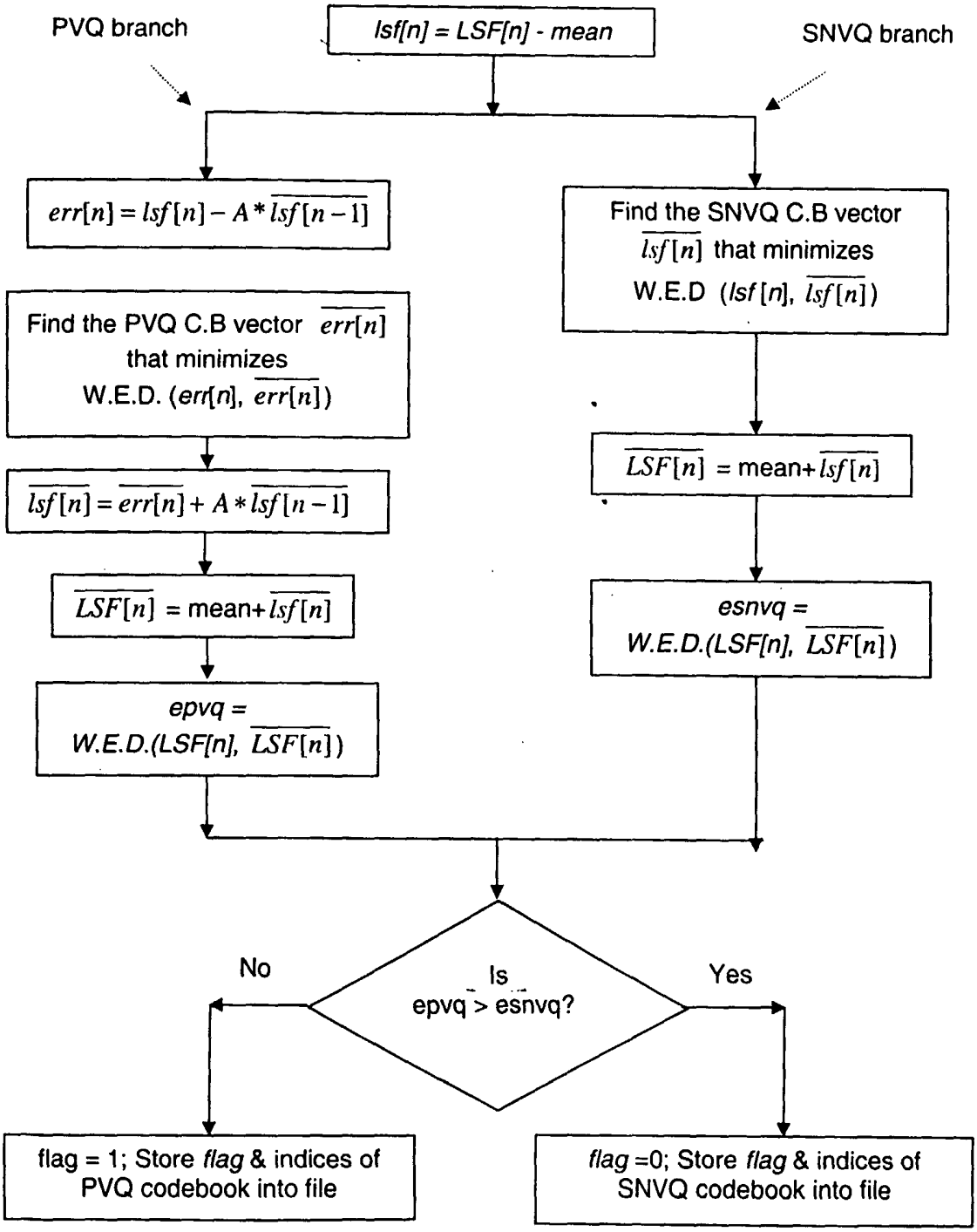

PVQ branch
lsf[n] = LSF[n] - mean
SNVQ branch
$err[n] = lsf[n] - A * \overline{lsf[n-1]}$
Find the SNVQ C.B vector $\overline{lsf[n]}$ that minimizes W.E.D ($lsf[n]$, $\overline{lsf[n]}$)
Find the PVQ C.B vector $\overline{err[n]}$ that minimizes W.E.D. ($err[n]$, $\overline{err[n]}$)
$\overline{LSF[n]}$ = mean+$\overline{lsf[n]}$
$\overline{lsf[n]} = \overline{err[n]} + A * \overline{lsf[n-1]}$
$\overline{LSF[n]}$ = mean+$\overline{lsf[n]}$
esnvq = W.E.D.(LSF[n], $\overline{LSF[n]}$)
epvq = W.E.D.(LSF[n], $\overline{LSF[n]}$)
No
Is epvq > esnvq?
Yes
flag = 1; Store flag & indices of PVQ codebook into file
flag =0; Store flag & indices of SNVQ codebook into file

Figure 6

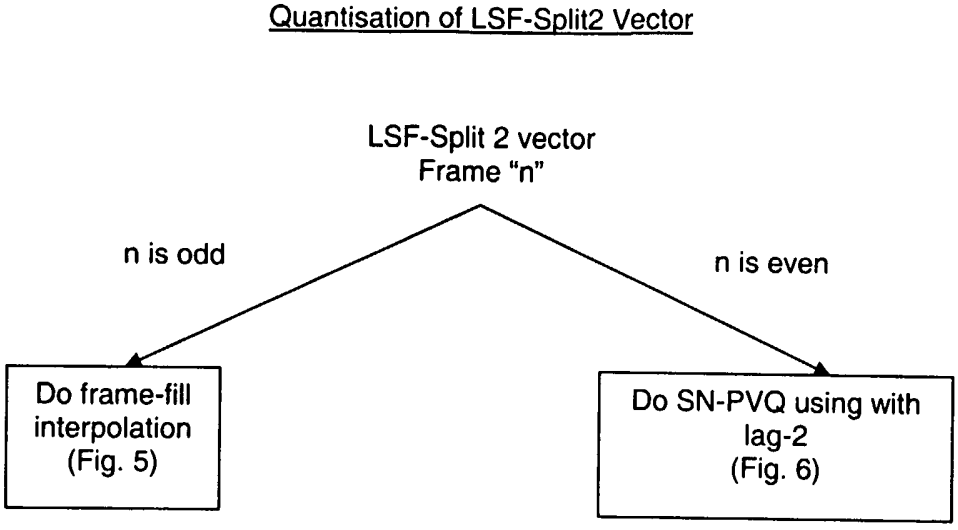

Quantisation of LSF-Split2 Vector
LSF-Split 2 vector
Frame "n"
n is odd
n is even
Do frame-fill interpolation (Fig. 5)
Do SN-PVQ using with lag-2 (Fig. 6)

Figure 7