An 800 bps MBE vocoder with low delay

Neelesh Tamrakar, Deepak M. and Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai-400076
{nilesh, deepu, prao}@ee.iitb.ac.in

***Abstract*— Segment based speech coders exploit the inter-frame redundancies of slowly varying speech segments to achieve low bit rates. The Multiband Excitation (MBE) speech model, known to provide natural sounding speech and robustness to acoustic background noise, is considered for very low bit rate coding based on speech segmentation. An algorithm is proposed for the fixed segment based coding of MBE model parameters at a fixed bit rate of 800 bps with low delay. Parameter quantization methods are presented. The codec is evaluated for quality and intelligibility on clean speech and on speech in background noise. The results indicate that the proposed parameter quantization methods preserve the effectiveness and robustness of the MBE model.**

I. Introduction

Speech coding is an important component of voice based digital communications. While the bit rate achieved is a critical factor in meeting the channel bandwidth constraints, received speech quality and robustness to channel errors, as well as to acoustic background noise, greatly influence the practical acceptability of a speech coding algorithm. Yet another significant issue in two-way communication is the delay. The algorithmic delay introduced by the speech codec contributes to the overall delay that occurs between the speech uttered by the sender and that received by the recipient. Large latencies (greater than 100 ms) lead to unnaturalness in conversational speech quality and need to be avoided. Algorithmic delays arise from the coding of speech based on large segments in which the redundancies in the speech signal are more easily exploited. Thus there is typically a trade off between obtained bit rate and delay. In this work, we address the coding of speech at the very low bit rate of 800 bps, a standard rate for secure voice in HF military communications.

Low bit rates can be obtained by model based coders that represent the speech signal as fixed duration frames each modeled by a compact set of parameters. Intelligible speech at 600 bps based on requantization of the parameters of the standard 2400 LPC vocoder has been achieved [1]. No measurements of quality are available but it well known that the 2400 LPC vocoder was unnatural sounding. In the present work, the Multiband Excitation (MBE) speech model is considered for the very low bit rate coding. The MBE model is known to provide natural sounding speech and has an inherent robustness to acoustic background noise. For low bit rates, the challenge is to investigate efficient parameter quantization techniques that also preserve the robustness of the model parameters.

We propose an algorithm for the segment based coding of MBE model parameters at a fixed bit rate of 800 bps. Parameter quantization methods are described. A significant aspect of the segment level parameter quantization methods is that frame-wise spectral distortion is considered together with distortion in the spectral dynamics across frames. With fixed segment duration of 80 ms, the coding delay is well within acceptable limits. The coder is evaluated for speech intelligibility and quality using subjective listening and objective measures.

II. Segment based coding

Model based coders represent speech by a fixed set of parameters per frame, typically at a frame rate of 50 frames per second. This limits the achievable compression with parameter transmission needed at 20 ms intervals. Speech, however, is known to be relatively stationary over much longer intervals depending on the underlying phone class. This redundancy can be exploited by encoding larger segments of speech where speech parameters vary only slowly across frames thus facilitating more efficient quantization.

A segment is collection of fixed or variable number of frames. If we try to find variable length segments such that the speech parameters vary in a smooth manner within each segment, then very efficient quantization of the segments is possible by exploiting much of the redundancy present in speech. However, coding based on variable length segments has some disadvantages, esp. if a fixed bit rate is desired. Segment boundaries must be optimally derived over sufficiently long blocks by a search based method adding to the computational complexity [2]. Further, the maximum block duration directly impacts the achievable coding delay. Longer this duration, the more correlation can be exploited, and the performance of the coder improves. Fixed segments on the order of few frames only, however, provide for low delay but present challenges in the accurate encoding of rapidly changing sounds like stops and affricates. For a frame duration of 20 ms, a segment comprising of four frames gives a rate of 12.5 segments per second, comparable with the average phone rate in speech. Based on this consideration, we present quantization methods for model parameters estimated over four frames. This work presents an improvement over the block based coder of [2] where a bit rate of 940 bps was attained at the delay of 400 ms.

III. MBE SPEECH MODEL

In the MBE speech model, voiced regions are represented by harmonics of a fundamental frequency, and unvoiced regions by spectrally shaped random noise [3]. The voicing information allows the mixing of the harmonic spectrum with a random noise spectrum in a frequency dependent manner in the synthesized speech output. The phase of harmonics is not transmitted but predicted during synthesis in most low rate coders. The parameters of the MBE speech model thus consist (for each analysis frame) of the fundamental frequency, voicing decisions (one for each group of 3 harmonics) and the harmonic amplitudes. Figure 1 depicts the MBE parameters and their relationship to the speech signal power spectrum.

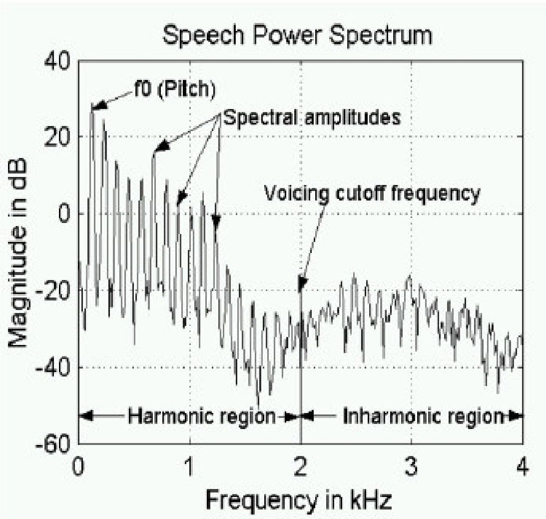

Figure 1: MBE speech model parameters

Speech analysis involves the estimation of the MBE parameters for each input speech frame of 20 ms duration. The MBE analysis algorithm estimates the model parameters (pitch and spectral amplitudes) by fitting an ideal harmonic spectrum to the actual spectrum [3]. For estimating the binary voicing decisions, the spectral mismatch between the actual and modeled spectrum is compared with a voicing threshold (which is dynamically updated) in each voicing band. In the interest of obtaining a compact representation, the band voicing pattern is simplified to a single frequency value, the highest voiced frequency. The region above this frequency is assumed to be unvoiced.

Of the MBE speech model parameters the spectral amplitudes are the most demanding on the bit allocation. The spectral amplitudes can be more compactly represented by a spectral envelope obtained by suitable interpolation. Frequency-warped LP modeling of the spectral envelope provides a compact and perceptually accurate representation. In the present work, an LP model of order 10 is chosen. The equivalent representation in the form of 10 line spectral frequencies (LSF) is finally obtained for its superior quantization properties.

IV. PARAMETER QUANTIZATION

To obtain low bit rate, we have to quantize efficiently the parameters obtained from the MBE analysis. Segment-level quantization methods are proposed for LSFs, gain, pitch and voicing parameters at a fixed number of bits per block. The methods for LSFs and gain parameters closely follow the work of [4] on LPC vocoding. New methods are presented for the pitch and band voicing parameters. All parameter quantizers are trained on a large database comprising of 418 English sentences spoken by 49 speakers (both male and female). The samples for native English speakers were taken directly from TIMIT database, and for non-native, were recorded in our lab. Those from TIMIT database were down sampled to 8000 Hz, while the others were recorded with 16-bit sample width at 8000 Hz sampling frequency. All the speech sentences used in this are noise free. This database contains a total of 89702 frames of 20 ms duration.

A. LSF vector

The 10-dim LSF vector represents the spectral envelope of the corresponding frame. Encoding LSF vectors across the segment at a low bit rate calls for an understanding of the trade-off between preserving temporal accuracy and spectral accuracy. It is known that for some cases (e.g. plosive sounds) it is perceptually more significant to capture the characteristic temporal evolution with the use of high temporal resolution of transmitted parameters, but in the case of slowly varying sounds (e.g. vowels) it is perceptually more important to provide high spectral accuracy. One solution to this problem is to adapt the quantization method to the characteristics of the underlying sound. A pre-determined set of alternative quantization schemes is searched for the best one to encode a given input LSF segment matrix [4]. The distortion measure used in the search must reflect the requirements of frame-level (static) distortion as well as the preservation of temporal evolution of the spectral parameters (dynamic distortion).

Following the approach of [4], 8 distinct quantization schemes are designed for the joint quantization of the 4 frame LSF vectors. At a bit allocation of 32 bits/segment, we have the flexibility of choosing to quantize all frames coarsely at 8 bits each, or, of selecting one or two frames to quantize more finely while interpolating the remaining. Table 1 provides the detailed bit allocation for each scheme. When 16 bits (or 24

Table I
QUANTIZATION SCHEMES OF LSFS QUANTIZATION OF A SEGMENT

Sr. No	Type of sound	Frames chosen	VQ method for chosen frames
1	consonants	All frames	8 bit coarse VQ
2	mixed	1st and 2nd frame	8 bit coarse VQ + 8 bit residual VQ
3		1st and 3rd frame	8 bit coarse VQ + 8 bit residual VQ
4		1st and 4th frame	8 bit coarse VQ + 8 bit residual VQ
5		2nd and 3rd frame	8 bit coarse VQ + 8 bit residual VQ
6		2nd and 3rd frame	8 bit coarse VQ + 8 bit residual VQ
7		3rd and 4th frame	8 bit coarse VQ + 8 bit residual VQ
8	vowels	2nd frame	12 bit coarse VQ + 12 bit residual VQ

bits) are available for the VQ of a single frame, multi-stage VQ is implemented in order to keep the search complexity within reasonable bounds. The quantization method selected for a given segment of input speech is the one that minimizes the sum of the static and dynamic distortions across the segment.

Since the receiver requires the knowledge of the selected quantization scheme for proper decoding, an overhead of 3 bits/segment is incurred by the transmission of the corresponding index. All the LSF vector codebooks are trained on non-silence frame LSFs of the training database.

The distortion criterion used for the frame-level quantization is the perceptually weighted Euclidean distance criterion with weights w_i that are product of two types of weights c_i and p_i, where c_i are constant perceptual weights chosen such that lower LSFs are given more weight compared to higher LSFs, based on the fact that the human ear cannot resolve differences at higher frequencies. The p_i are perceptual weights which are calculated for each frame, by calculating spectral sensitivity for each LSF, which is based on closeness of LSFs. The value of distortion or error found by distortion criterion with weights w_i is denoted by spectral error. For proper encoding and to choose the segment quantization scheme (from Table I), we need to take dynamic distortion into account. This is done via the spectral change computed as Euclidean distance between two consecutive frames as

$$SC^n = \sum_{i=1}^{M}\left[f_i^n - f_i^{n-1}\right]^2 \qquad - (1)$$

In the above equation f_i^n and f_i^{n-1} are i^{th} LSF of LSF vector for n^{th} and $(n-1)^{th}$ frames respectively. After finding the spectral change for quantized and unquantized frames, smoothness error is calculated as the ratio of unquantized and quantized spectral errors minus 1. A pictorial representation of this is given in Figure 2.

For the calculation of smoothness error we need to have knowledge of one future frame. In case of segment boundary (last frame of segment) this leads to an extra 20 ms delay in the processing. Therefore in the present work, spectral change is restricted to calculation only with respect to past frame in case of segment boundary frame. For remaining three frames of the segment it is calculated by considering both past and future frames. The smoothness error over the segment must be calculated for each of the eight schemes. This requires the quantized LSFs of all the frames of the segment. Depending on

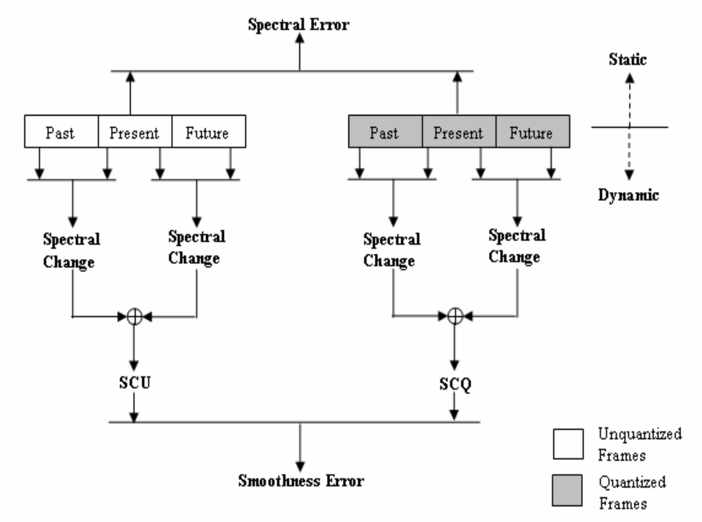

Figure 2: Pictorial representation of errors

the chosen scheme of Table 1, a frame's quantized LSFs may be obtained by codebook search or by interpolation/extrapolation.

To select the best method for a given category of sound, we need to consider both spectral error and smoothness error. Therefore, spectral error and smoothness error are calculated for each frame of the segment and a perceptual error is calculated by summing both of them. Finally, total perceptual error is calculated by summing perceptual error of each frame of the segment. All the eight schemes of the quantization are applied to given segment and total perceptual distortion is calculated in each case. The spectral quantization method having the minimum total perceptual error is the best available method for quantizing the given category of sound. The index of the identified method along with the encoded spectral information is then transmitted to decoder.

B. Gain

The gain shows high correlation across frames making it suitable for vector quantization. The four element of the vector corresponds to the log gain values of four consecutive frames. For codebook search weighted Euclidean distance criterion is applied for unquantized and quantized log gain values and weights are used to emphasize accurate coding of energy transitions at the expense of energy accuracy during the steady state. The weights are given as

$$w_i = 1 + A_o\,(\,|\,g_i - g_{i-1}\,| + |\,g_{i+1} - g_i\,|\,) \qquad \text{- (2)}$$

In the above equation, g_i is unquantized gain for i^{th} frame and value of i varies from 1 to 4. Further g_o = (previous segment's g_4) and $|\,g_5 - g_4\,| = 0$. The value of A_o controls the weights for desired rise and fall. Exhaustive search of codebook is performed and index of code vector (9 bit) which gives the minimum error is transmitted to the decoder. At the decoder, the values corresponding to the index are read from the codebook.

C. Voicing

The highest voiced band is obtained from the 12 bit voicing decisions, by marking the last voiced band as highest voicing band. Frames which have bands less than or equal to seven are represented with 3 bits directly, and for frames with more than 7 bands, the highest voiced band is restricted to set of {0, 1, 2, 3, 4, 6, 8, 12} to encode the index in 3 bits. Thus 12 bit voicing decisions are encoded into 3 bit frame voicing index. To encode the frame voicing indices for the segment efficiently, a study was conducted to observe the frequency of occurrence of a particular voicing index pattern (frame voicing indices of four consecutive frames). A total of 4096 distinct patterns are possible. The study was conducted on the training database mentioned earlier. Of the possible patterns, it was found that 218 patterns contributed to nearly 80% of total segments. Based on this study, a codebook is designed by listing the most frequently occurring pattern in decreasing order of frequency. During the codebook search, weighting for band voicing decisions is given to emphasize the low frequency region, as well as weighting for the frame voicing decision index of each frame in the segment in order to prevent the change of voicing index of frames with zero value. At the decoder, the index (9 bit) transmitted is interpreted correctly based on the knowledge of the number of bands for each frame.

D. Pitch

Since it is important to preserve the pitch contour, a method based on vector quantization of normalized log pitch is proposed. The normalization of each frame log pitch is with respect to the quantized log pitch of the last frame of the previous segment. If last frame of past segment is found to be unvoiced then quantized pitch of last known voiced frame is considered. In the present 800 bps coder, three codebooks are used for vector quantization of segments having 4, 3 and 2 voiced frames; these codebooks will have 4D, 3D and 2D code vectors respectively. Codebook search is done based on the weighted Euclidean distance criterion calculated between quantized and unquantized log pitch value and weights are calculated based on the voicing index of the frame such that more weight is given to frames which are highly voiced. Because pitch information is used to synthesize the voiced bands of the frame in MBE synthesis, pitch error in the highly voiced frames will have more effect compared to that of less voiced frames. An appropriate codebook is selected based on the number of voiced frames in the segment; the index (10 bit) of the code vector which gives the minimum error is transmitted to the decoder. For segments having only one voiced frame, log pitch of voiced frame is quantized using uniform scalar quantizer.

E. Postfiltering

Postfiltering is applied to reduce the coding noise introduced during the quantization process based on the fact that noise in the spectral valleys is more audible than that near the peaks. A short-term cepstrum based postfilter has been found to perform well and is incorporated in the decoder [2].

The overall bit allocation for the 800 bps coder is shown in Table II.

V. Performance Evaluation

The reconstructed speech quality of coder is judged on the dimensions of intelligibility and naturalness. These are evaluated by standard subjective and objective measures. The PESQ-MOS scores [5] (maximum 4.5) is obtained on a test set containing six female and six male speech sentences (outside the training set). In order to examine the robustness of the quantization methods, PESQ-MOS is also obtained for noisy speech (additive white noise at SNRs of 30 dB, 20 dB, 16 dB and 10 dB). The PESQ score is computed with respect to the corresponding reference speech input (i.e. the noisy speech) in each case. The scores appear in Table III. In order to better appreciate the extent of degradation due to quantization, scores for speech reconstructed after MBE-LP modeling (i.e. before quantization) are also provided.

In order to measure intelligibility, a listening test is necessary. The Modified Rhyme Test (MRT) is such a test that measures the distinguishability of confusable words. The MRT comprises of lists of single-syllable rhyming words that differ in one consonant [6]. One word randomly chosen of a set of 6 rhyming words is played out and the listener marks which word he/she thinks he/she hears, on a multiple choice answer sheet. The 800 bps coder intelligibility was measured for two listeners; each subject has gone through the test (51 sets of 6 one-syllable words) twice at different times. Results, as shown in Table IV, are obtained for clean speech and noisy speech signals for reference, modeled and quantized speech.

VI. Discussion

From the performance results of Tables III and IV, we note that the 800 bps coder obtains an average PESQ score of 2.8 on our test set. This average score is close to that of the segment based coder of [2] and can be considered very reasonable for a coder at 800 bps. A significant fraction of the speech quality loss is observed to occur in the LSF quantization and pitch quantization stages. This is also confirmed by an examination of the average spectral distortion due to LSF quantization which shows the same trend as PESQ scores in Table III. More sophisticated methods of quantization would help to improve the quality.

On adding noise, the relative degradation in quality is maintained indicating that the proposed parameter quantization methods are robust. It is observed from listening that the noise background is not significantly distorted up to 10 dB SNR even though the codec is based on a speech model. This may be attributed to the power of the MBE model's multiband voicing feature which helps to represent noisy speech better. In fact, the perceived noise in the reconstructed speech has a slightly reduced audibility compared to that before quantization. MRT scores indicate that intelligibility decreases with coding but the relative decrease with added noise is constant. All PESQ scores in Table III are with respect to the corresponding reference speech (noisy or clean).

In summary, a fixed segment based coder is developed based on the MBE speech model. A fixed bit rate of 800 bps is obtained with delay of less than 100 ms. The low bit rate coder achieves acceptable speech quality and intelligibility. Further, the robustness of the model is preserved after quantization of parameters. Future work is directed towards further improvements in quality and on reducing computational complexity in modeling and quantization.

Audio demos are available at:
http://www.ee.iitb.ac.in/daplab/demos/mbe/

References

[1] F. Zou, Y. Guo, X. Chen and Y. Liu, "Design and description of a 600 bps speech coder based on MELPe ," in *Proc. IEEE workshop on information theory*, pp. 356-359, October 2006.

[2] R. S. Kumar, N. Tamrakar and P. Rao, "Segment based MBE speech coding at 1000 bps," in Proc. *National Conference on Communication*, pp. 446-450, February 2008.

[3] D. W. Griffin and J. S. Lim, "Multi-band excitation vocoder," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223-1235, August 1988.

[4] B. A. Fette and C. A. Jaskie, "Low bit rate vocoder means and method," United States Patent, 5,255,339, October 1993.

[5] ITU-T, "Rec P.862, Perceptual evaluation of speech quality (PESQ) an objective assessment of narrowband networks and speech codecs," ITU, 2002.

[6] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*, Ch. 2, Prentice Hall, 1988.

Table II
BIT ALLOCATION TABLE

SR.NO	PARAMETER	BITS REQUIRED PER SEGMENT	BIT RATE(BPS)
1	Voicing	9	112.5
2	LSFs	(8+8)+(8+8)=32	400
3	Gain	9	112.5
4	Pitch	10	125
5	Index of LSF quantization method	3	37.5
6	Sync bit	1	12.5
	Overall bit rate		800

Table III
AVERAGE PESQ SCORE

SNR	LP modeled version	After quantization
No noise	3.2	2.8
30 dB	3.3	2.9
20 dB	3.3	3.0
16 dB	3.3	2.9
10 dB	3.3	2.8

Table IV
RESULTS OF MODIFIED RHYME TEST

SNR	Percentage Intelligibility: Original	LP modeled version	After quantization
No noise	94.5	86.0	78.0
30 dB	94.5	85.5	75.5
20 dB	91.0	80.0	64.5
16 dB	88.0	75.5	61.5
10 dB	80.5	65.0	55.5