Segment Based MBE Speech Coding at 1000 bps

R. Siva Kumar, Neelesh Tamrakar and Preeti Rao
Department of Electrical Engineering,
Indian Institute of Technology, Bombay
Powai, Mumbai-400076
{rskumar,nilesh,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. Segment based coding generally leads to variable bit rate, which is unacceptable in certain communication applications. An algorithm is proposed for the segment based coding of MBE model parameters at a fixed bit rate below 1 kbps. Speech segmentation, as well as parameter quantisation methods suited to the segmental behaviour of each parameter, are described. The coder is evaluated for speech intelligibility and quality using subjective listening and objective measures.**

I. Introduction

The objective of speech coding is to represent the speech signal with a minimum number of bits while retaining its intelligibility and perceived quality. Low bit-rate speech coding is a key component in the new digital wireless communication services. It continues to be of research importance due to the increasing demands on band limited media such as radio and satellite links, and storage limited media such as CD-ROMs and silicon memories. While a number of voice compression standards are available for bit rates above 4 kbps, there is a crucial need for very low bit rate operation in the case of military communications and storage applications such as voice recorders. Low bit rates can be obtained by model based coders that represent the speech signal as fixed duration frames with each modeled compactly by a set of estimated parameter. However the typical frame rate of 50 frames per second limits the achievable compression with parameter transmission required at 20 ms intervals. Speech, however, is known to be relatively stationary over much longer intervals depending on the underlying phone class. Bit rates below 1 kbps make it essential to exploit this redundancy by encoding larger segments of speech where speech parameters vary only slowly across frames thus facilitating more efficient quantisation.

In this work, the Multiband Excitation (MBE) speech model is considered for very low bit rate coding based on speech segmentation [1]. The MBE model is known to provide natural sounding speech and has an inherent robustness to acoustic background noise. Segment based coding generally leads to variable bit rate, which is unacceptable in certain communication applications involving low bandwidth channels with their limit on maximum rate. We propose an algorithm for the segment based coding of MBE model parameters at a fixed bit rate below 1 kbps. Speech segmentation, as well as parameter quantisation methods suited to the segmental characteristics of each parameter, are described. The coder is evaluated for speech intelligibility and quality using subjective listening and objective measures.

II. 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 [1]. 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. Fig.1 depicts the MBE parameters and their relationship to the speech signal power spectrum.

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 [1]. 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 compactly represented by a spectral envelope obtained by suitable interpolation. Adaptive

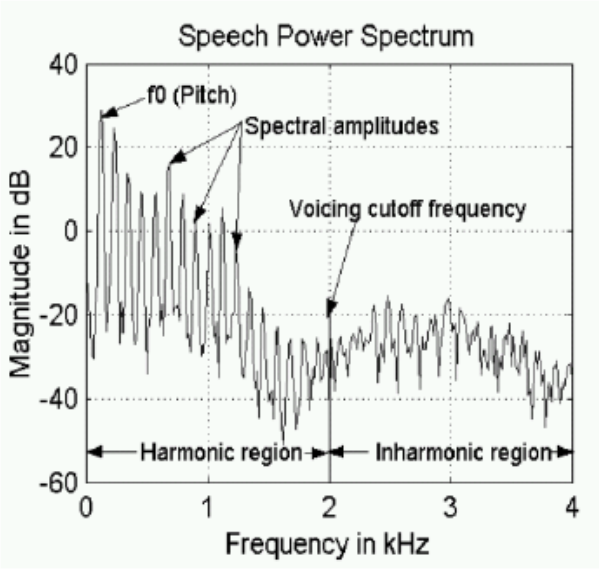

Fig. 1. MBE speech model parameters marked on the spectrum of a voiced speech frame

frequency-warped LP modeling of the spectral envelope has been shown provide a compact and perceptually accurate representation [2]. In the present work, an LP model of order 12 is chosen. The equivalent representation in the form of 12 line spectral frequencies (LSF) is finally obtained for its superior quantisation properties. This ensures that (1) the stability of the synthesis filters is guaranteed upon quantization provided the natural ordering of the LSF's is maintained, and (2) the spectrum distortion due to quantization noise of a particular LSF, remains local to the corresponding frequency region.

III. Segmentation of Speech

The speech spectrum is typically slowly varying with the relatively abrupt spectral transitions marking the phone boundaries. Phones such as vowels have nearly constant spectra over several tens of milliseconds. A few phones such as plosives, however, are characterised by rapid variations within the phone duration. Segmentation of speech, into relatively homogenous regions of variable length is based on detecting spectral transitions. The slow variation of spectral parameters across the frames belonging to a single segment facilitates a simple parametric representation of the segment. If the frame-level spectrum is represented by line spectral frequencies (LSF), a simple measure of intra-segment distortion is provided by the mean squared error between the input LSF vectors and the corresponding segment representative vector. This intra-segment distortion is minimized by the centroid of the spectral vectors spanning the frame as given by

$$r = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad (1)$$

where 'k' is number of frames within the segment and x_i is the spectral vector (e.g. LSF) of the ith frame. Svendsen [3] has shown that such centroid-based representation of segment LSFs obtains a spectral distortion within 1 dB.

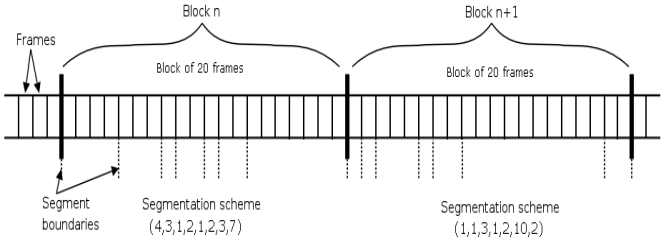

Fig. 2. Segmentation Scheme

The actual segmentation of the speech spectral vector sequence can be carried out in a sequential manner by examining each subsequent frame for a possible boundary, based on its error with respect to an updated centroid. Due to the varying durations of the segments, the resultant bit rate is variable. In order to achieve a fixed bit rate, it is necessary to introduce constraints on the segment-based coding. For instance, fixing the number of segments over a specified large block of frames helps to obtain a fixed bit rate at the block level at the cost of the block-length delay. Coding fixed duration blocks (each comprised of a fixed number of variable length segments) also offers the opportunity of optimizing the segmentation over the block by the approach of maximum likelihood. Pictorial view of this segmentation scheme is shown in Fig. 2 for two blocks.

With block-based maximum likelihood (ML) segmentation, the segment boundaries obtained minimize the average intra-segment distortion across the block [4]. Let a block duration of speech be represented by $X_1^T = (x_1, x_2, \ldots\ldots x_T)$, a spectral parameter vector sequence of T speech frames, where x_n is a p-dimensional parameter vector at frame 'n'. The segmentation problem is to find 'm' consecutive segments in the observation sequence X_1^T. Let the segment boundaries be denoted by the sequence of integers $B=(b_0, b_1,\ldots..b_m)$. The i^{th} segment starts at frame $b_{i-1}+1$ and ends at frame b_i; with $b_0=0$ and $b_m=T$. The optimal segmentation $B=(b_0, b_1,\ldots..b_m)$ is obtained so as to minimize the sum of intra-segment distortion over all possible sets of segment boundaries, i.e. to minimize

$$D(m,T) = \sum_{i=1}^{m}\sum_{n=b_{i-1}+1}^{b_i} d(x_n, \mu_i) \qquad (2)$$

where $D(m,T)$ is the total distortion of m-segment, segmentation of $X_1^T = (x_1, x_2, \ldots\ldots x_T)$; μ_i is the centroid of the i^{th} segment and $d(x_n, \mu_i)$ is the Euclidean distance between the vector x_n and the centroid μ_i. The optimal segment boundaries can be obtained efficiently by a dynamic programming procedure using recursion [4].

An important design decision involves the choice of the block duration and the number of segments per block. Longer

block durations improve coding efficiency but adversely impact the codec delay. Communication systems utilizing low bandwidth channels such as HF can often tolerate codec delay of up to 400 ms. In view of this, we choose a block duration of 400 ms, or 20 frames. The number of segments per block can be decided based on the expected average segment rate. A segment rate of around 18 per second is found to achieve high segmentation performance [4]. This is also consistent with the normal human speaking rate of 12 phones per second. For the 20-frame block of speech, this translates to 7 segments.

IV. Parameter Quantisation

The segmentation discussed above is based on homogeneity as determined by frame LSF vectors, which represent the spectral envelope. A study of the segments so obtained reveals that the large majority is comprised of similarly voiced frames i.e. all (highly) voiced or all unvoiced. When there are changes in the voicing characteristics of frames within a segment, it usually occurs at the segment edges. Another related observation is that the estimated pitches across voiced frames constituting a segment vary slowly and smoothly. These observations suggest the possibility of efficient coding of the parameters by exploiting inter-frame redundancies within segments. We next describe the segment-level quantisation of the LSFs, gain, pitch and voicing parameters at a fixed number of bits per block.

LSF vector and gain

The LSF vectors of the set of frames belonging to a segment are represented by the centroid LSF vector as in Eq (1). The 12-dimensional centroid vector is quantised by split-vector quantisation. The independent quantisation of the 6-dimensional upper and lower splits reduces the codebook search complexity significantly at the cost of a slight loss of optimality. Due to the greater importance of the low frequency region, the low split is allocated 12 bits and upper split 11 bits. The gain parameter (or rather, the log of the gain) shows high correlation across frames making it suitable for vector quantisation. However there was no obviously preferred way to encode the variable-dimension vector of frame gains across the segment with a fixed number of bits. The 20-frame block is instead divided into 5 equal sub-blocks of 4 frames each. The 4-dimensional log gain vector of each sub-block is vector quantised with a 10-bit codebook. The database used for codebook generation for both LSF and gain vectors is the TIFR Hindi database [5]. The database contains a total of 1000 Hindi sentences made up of sets of 10 sentences read by each of 100 speakers (male and female).

TABLE I
Bit Allocation

Parameters	Bits Required		Bit Rate (bps)
	Per Segment	Per Block	
LSFs	12+11 =23	23 × 7 = 161	402.5
Gain	-	10 × 5 = 50	125
Pitch	-	7 × 6 + 13 × 3 = 81	202.5
Voicing	-	3 × 20 = 60	150
Segment Duration	4	4 × 6 = 24	60
Total bit rate of the coder			940

Pitch

Since the frame-level pitch estimates vary smoothly across the segment, a method based on intra-segmental differential pitch is proposed. The center frame pitch is fully quantised with a 6-bit scalar quantiser (uniform intervals in log pitch, based on perceptual frequency resolution). Next, the remaining frames' pitches are differentially quantised with respect to the center frame and each other. Since the direction of pitch change from frame to frame, is perceptually more significant than the magnitude of the change a 3-bit quantiser is found sufficient to encode the pitch difference.

Voicing

Although the general nature of the frequency-dependent voicing is similar across the frames in a segment, there are fluctuations in the estimated highest voiced frequency. These arise from minor shifts in voicing nature and are compounded by the frame to frame pitch variation. Further, in the case of a few segments the parameter makes abrupt jumps at the boundaries. Considering that voicing information is crucial to speech intelligibility, an attempt is made to preserve the individual frame voicing in a compact form by restricting highest voiced band to a limited set of values (8 levels or 3 bits/frame).

At the decoder end, the dequantised segment parameters are converted to frame parameters for speech synthesis. The required segment duration information is obtained from a 4-bit transmitted code (indicating segment length in the range 1-14 frames). The overall bit allocation is shown in Table I.

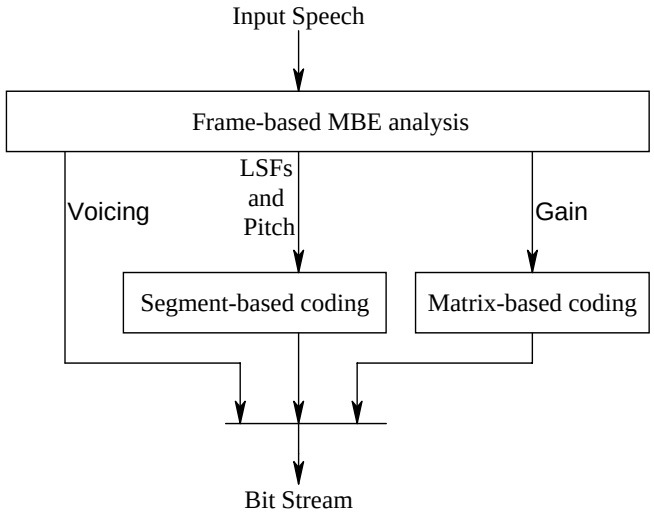

Fig. 3. Encoder block diagram

Fig. 3 summarizes the quantisation approach for each of the

parameters over the fixed block of 20 frames of input speech.

V. POSTFILTERING

In low bit rate speech coders, it is common to use a short-term postfilter at the decoder to improve the perceived quality of reconstructed speech [6]. The postfilter is derived from the spectral envelope of the speech synthesized from the decoded parameters. The postfilter spectral shaping serves to attenuate the spectral valleys with respect to the spectral peaks thus reducing the coding noise in regions of high perceptual sensitivity. The postfilter also introduces a spectral tilt which needs to be compensated. This is typically achieved by including a low-order adaptive all-zero component. In this work, we incorporate the adaptive short-term postfilter proposed in [7], based on combining weighted LPC coefficients and pseudo-cepstral coefficients. This postfilter does not require explicit tilt compensation. The postfilter has an efficient implementation. Further, it is easily incorporated within the MBE decoder via the direct shaping of the reconstructed spectral magnitudes by the postfilter frequency response.

Fig. 4 shows the postfilter response together with the speech and LPC spectra before postfiltering. The postfilter parameters include the scaling factors (α_1, α_2) for roots of symmetric and asymmetric polynomials derived from LPC coefficients. α_1, α_2 moves the zeroes postfilter inside the unit circle. Third parameter (β) moves the poles inside unit circle. Phase distortion will be small if $\alpha_1 + \alpha_2 = 2\beta$. In present work these parameters are tuned to the segment-based MBE coder as determined by subjective listening. The values of parameters chosen are α_1=0.8, α_2 = 0.1 and β = 0.45.

VI. PERFORMANCE EVALUATION

Subjective listening and objective measures can be used to judge the quality and intelligibility of the codec output speech. Speech quality refers to the naturalness of the output speech while intelligibility refers to the ease of understanding the message content. A standard scale for speech quality assessment is the Mean Opinion Score (MOS) scale ranging from 1.0 (poor) to 4.0 (transparent). An objective prediction of MOS is provided by the ITU-T standard P.862 PESQ (Perceptual Evaluation of Speech Quality) [8]. In the PESQ standard, the referenced and degraded signals are mapped on to internal auditory representations for comparison within a cognitive framework. The PESQ provides a score between -0.5 to 4.5, although most cases it is in the range of 1.0 to 4.5. A PESQ score of 4.5 indicates there is no perceivable distortion between reference and degraded signal, while 1.0 indicates "bad" degradation. Subjective listening is used for measuring speech intelligibility via the distinguishability of confusable words. The Modified Rhyme Test (MRT)

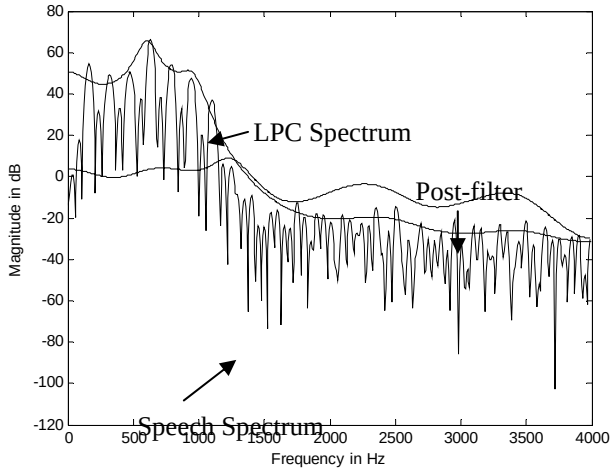

Fig. 4. Frequency response of postfilter

comprises of lists of single-syllable rhyming words that differ in one consonant [9]. 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. Codec intelligibility is measured by average scores obtained for correctly recognized words over a number of listeners.

We use the above methods to evaluate the speech quality and intelligibility obtained at 940 bps. We would also like to evaluate the specific role of the new postfilter. In addition to the PESQ MOS ratings and MRT, a subjective listening test known as the A-B comparison test is carried out to compare speech before and after postfiltering. Each pair, A-B, comprises of a specific sentence utterance after decoding, but before and after postfiltering. To avoid bias, the order A and B are interchanged and randomized across sentences and listeners. Thus a total of 120 subjective judgments are obtained with three listeners and 20 test sentences.

In present work, speech utterances chosen for PESQ MOS measurement are from TIMIT [10] and TIFR databases, from different speakers (male and female). A total 40 speech utterances are chosen, 20 from each database. Speech utterances from TIMIT are considered as outside the training set because training is performed using TIFR database. Utterances from TIMIT database are down sampled to 8000

TABLE II
RESULTS OF MODIFIED RHYME TEST

		Percentage Intelligibility	
Sr. No.	Subject	Before Postfilter	After Postfilter
1	Subject -1	86.2	90.2
2	Subject -2	86.2	94.2
3	Subject -3	78.4	80.4

TABLE III
RESULTS OF A-B TEST

		Percentage Preferred		
Sr. No.	Subject	Before Postfilter	After Postfilter	Neutral
1	Subject -1	15	60	25
2	Subject -2	7.5	52.5	40
3	Subject -3	25	57.5	17.5

TABLE IV
AVERAGE PESQ SCORE

Sr. No.	Test Item	Before Postfilter	After Postfilter
1	TIFR	2.65	2.73
2	TIMIT	2.57	2.64

samples/second before coding.

Results of the Modified Rhyme Test are presented in Table II. The average percentage intelligibility is 83.6 without postfiltering and 88.3 with postfiltering. It is clear from results of MRT that the postfilter spectral shaping does not decrease the intelligibility of the output speech. Results of the A-B test are presented in Table III. Average percentage preference for speech before postfiltering is 15.8 and with postfiltering it is 56.7 and remaining 27.5 is neutral. We observe that there is a clear listener preference for the postfiltered speech over the before postfiltering.

Average PESQ scores of TIMIT and TIFR sentences are given in Table IV. The average PESQ score of the coder is found to be 2.61 before postfiltering and 2.69 after postfiltering.

VII. SUMMARY

A segment based coder is developed based on the MBE speech model. A fixed bit rate is achieved by constraining the number of segments to a fixed value over a 20-frame block of speech. An optimised segmentation of the block is obtained based on the homogeneity of LSF vectors. Centroid modeling of the segment LSF vectors followed by split vector quantization contributes to the efficiency of the spectrum encoding. The gain parameter is quantized with vector quantization of sub-blocks. The pitch parameter is quantized via an intra segmental differential quantization method. An adaptive short-term postfilter serves to reduce the coding noise in the reconstructed speech. The low bit rate coder operating at 940 bps achieves acceptable speech quality and intelligibility.

It is noted that a significant fraction of the speech quality loss occurs in the LSF segmentation stage. A systematic study of segment rate versus quality, under the given constraints of delay and fixed rate, as well as more sophisticated methods of segmentation would help to reduce the distortion. Further reduction in bit rate is possible with segment based coding of the voicing information, which currently, is transmitted for every frame.

REFERENCES

[1] D. W. Griffin and J. S. Lim, "Multi-Band Excitation Vocoder," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 36, no. 8, pp. 1223-1235, August 1988.

[2] P. Patwardhan and P. Rao, "Effect of voice quality on frequency-warped modeling of vowel spectra," *Speech Communication*, vol. 8, no. 8, pp. 1009–1023, August 2006.

[3] T. Svendsen, "Segmental quantization of speech spectral information", *Proc. IEEE Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 19-22, April 1994.

[4] V. Ramasubramanian and T. V. Sreenivas, "Automatically derived units for segment vocoders", *Proc. IEEE Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 473-476, May 2004.

[5] K. Samudravijaya, P. V. S. Rao and S.S.Agrawal, "Hindi Speech Database", *Proc. Int. Conf. on Spoken Language Processing*, 2000.

[6] J. H. Chen and A. Gersho, "Adaptive Postfiltering for Quality Enhancement of Coded Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 3, issue 1, pp. 59-71, January 1995.

[7] H. K. Kim and H. G. Kang, "A Pseudo Cepstrum based Short Term Postfilter," *Proceedings of the IEEE workshop on speech coding*, pp. 99-101, September 2000.

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

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

[10] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, "The DARPA speech recognition research database: specification and status." *Proc. DARPA Speech Recognition Workshop*, 1986.