A low-bit rate segment vocoder using minimum residual energy criteria

Abhijit Pradhan, Sadhana Chevireddy, Kamakoti Veezhinathan, Hema Murthy
Department of Computer Science
Indian Institute of Technology Madras, Chennai, India
Email: {abhijit,sadhana}@lantana.tenet.res.in, {kama,hema}@cse.iitm.ac.in

***Abstract*—In speech coding, segment vocoders offer good intelligibility at low bit rates. A segment vocoder has four basic components 1)Segmentation of input speech 2)Segment quantization 3)Residual quantization 4)Synthesis of speech. Most segment vocoders use a recognition approach to segment quantization. In this paper, we assume a different approach to segment quantization. The segmental unit is a syllable and the segment codebook stores the sequence of LPC vectors. During the encoding process the speech segment is quantized using the sequence of LPC vectors that result in the smallest residual energy. PESQ scores indicate that this vocoder achieves better quality compared to that of a corresponding vocoder that uses a speech recognition framework.**

I. Introduction

Representing digital speech signals using as few bits as possible is a challenging problem. While achieving lower bit rates for transmission is one concern, preserving the intelligibility of the synthesized speech is the other concern. After successfully employing MELP (Mixed-Excitation Linear Prediction) [3] vocoder algorithm for defense communications, research in speech coding has focused on segment vocoders to deliver intelligible speech at even lower bit rates [4]. This has brought in a new class of speech coders called segment vocoders which encode speech at the segment level. The segment here is a well defined unit of speech, ex: phonemes/diphones/triphones etc. The encoder encodes, the system and source characteristics of the signal separately. The previous paper [1] presents a segment vocoder that uses syllable as the segmental unit of compression. The vocoder [1] delivered intelligible speech at bit rates close to 1400 bps. The system encoding is done at the syllable level and source compression is carried out at the frame level. LP analysis is used to obtain system(LPC) and source(residual) characteristics of the input speech signal that are later encoded separately. LPCs are encoded using a system codebook. Source characteristics is encoded using a standard algorithm as in MELP [3]. At the decoder, LP synthesis is used to synthesize back the speech signal using the decoded LPCs and residual.

The vocoder [1] uses a group delay based segmentation [5] to get syllable like units. The segment codebook (HMM codebook) is obtained using an unsupervised HMM (Hidden Markov Model) clustering algorithm. LPCC/MFCC features are used to represent the system characteristics of the segments. Each cluster is defined by an HMM. A *representative syllable(aka cluster centroid)* is chosen from each cluster. The LPC vector sequences of *representative syllables* from all clusters are used to form the segment codebook. During encoding, the input speech signal is first segmented into syllable like units using group delay based segmentation [5]. Each syllable segment is recognized against the HMM codebook to find the best matching cluster. The indices of the recognized clusters are encoded and transmitted along with the duration information of the syllable segments. The syllable segment is inverse filtered using the sequence of LPCs to obtain the residual. Duration mismatches are addressed by appropriate repetition/deletion of frames. The source information of the input signal is encoded using MELP residual coding algorithm. This resulted in an average bit rate of 1400 bps when a syllable rate of 7 syllables/s is assumed. The synthesized speech at the decoder has good intelligibility with PESQ scores comparable to that of MELP [3].

In the work [1], the synthesized speech quality deteriorates when there is a mismatch between the input and recognised unit. The reason for this low quality is due to the poor encoding of the residual using MELP [3]. During encoding, the residual is compressed and coded using MELP residual coding algorithm. When the representative syllable segments are very different from the input syllable segments the residual energy is very high; almost resembling a speech signal. The compression offered by MELP residual encoding (at 1.2Kbps) results in significant loss of information. This results in a mismatched residual at the decoder. The speech synthesized using the modeled residual lacks intelligibility and produces buzziness. This suggests that when a segment is misrecognised, the system characteristics are incorrect, representing the source characteristics with less number of bits is difficult.

Earlier work [4], [10], [13], [14], [15], [16] suggest techniques to improve the recognition performance in terms of modeling a better system codebook. The methods mostly focus on improving the segmentation performance to obtain better segments. This enables better clustering by reducing errors in recognition [4]. But as the codebook is limited by its size, there are always recognition errors. In the present work wherein a novel idea is proposed to identify the system characteristics from the codebook that reduces the residual energy thereby enabling better source encoding.

In the segment vocoder discussed in [1], the LPC sequences of representative syllables from the recognized clusters are used to form an inverse filter through which the input speech signal is passed. As discussed earlier, the resultant residual does not have minimum energy and thus cannot be coded properly with a few bits. The representative syllable should be chosen in such a way that it gives a residual which has minimum energy. Two different criteria are proposed for selecting the representative syllable:

1) Method I: Recognition followed by minimum residual energy.
2) Method II: Only minimum residual energy.

The new segment vocoder is identical to the previous vocoder [1], except for *system encoding*. The two methods proposed above are employed for encoding the system characteristics of the input syllable segment. Section II, the syllable based segment vocoder using the two proposed methods of system encoding are explained in detail. Section III presents the results and a detailed discussion on the performance of the vocoder (section III) with conclusions in section V.

II. Overview of the new segment vocoder

The following subsections give an overview of the new segment vocoder. We describe encoding and decoding process and the bit allocation details. A codebook of syllable HMMs is obtained as in [1]. LPC sequence vectors for ALL syllable segments are also stored.

A. *Encoder*

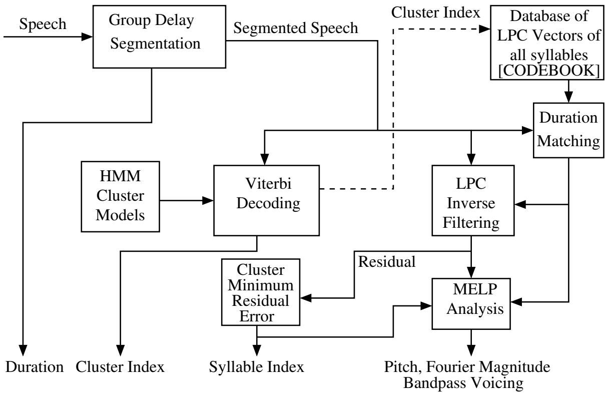

Fig. 1. *Speech Encoder in Method I.*

Method I:- Figure 1 details the encoding part of the segment vocoder using recognition followed by the minimum residual energy criterion. Raw speech signal at 8000 samples/second and 16bit/sample (128Kbps), is input to the encoder. Group Delay Segmentation [5] segments the speech into syllable like units and provides the segment/syllable boundaries. As in [1] the best index in the HMM codebook is determined. The cluster index is used to encode the particular cluster. Among all the syllables in the best matching cluster, the syllable which gives the minimum residual error is chosen and the syllable index is stored/transmitted (after performing a duration match). The index of the minimum residual energy syllable is used by MELP [3] analysis block. This calculates all non-LPC parameters: pitch, bandpass voicing, Fourier magnitude etc. The duration of the original speech segment is also encoded. Excitation and non-LPC parameters are calculated frame by frame(22.5ms) and encoded the same way as done in 2.4Kbps FS_MELP [6] with a modification in LPC calculation. In case of FS_MELP [6] analysis the LPCs for a frame are calculated from the original speech frame. In the segment vocoder the stored LPCs corresponding to the minimum residual error are used.

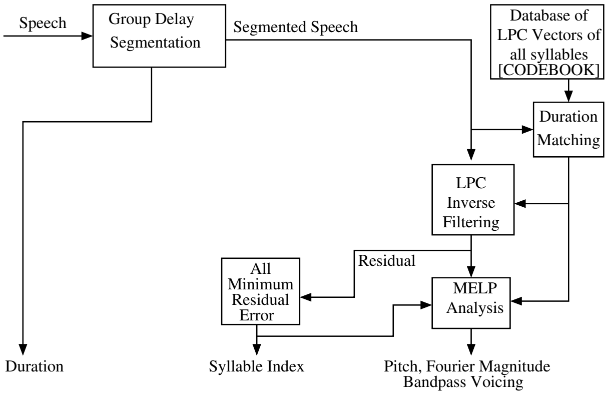

Fig. 2. *Speech Encoder in Method II.*

Let the best matching HMM cluster be C. Let P be the number of syllables in the cluster C. The j th syllable in cluster C is denoted by s_j. The residual error corresponding to each s_j for the current speech segment is calculated as follows. Let f_k be the raw speech frames(22.5ms) in the current input speech segment and $H_k(z)$ is the corresponding LPC filter obtained after duration matching. The residual error e_k for the current frame f_k is obtained by inverse filtering f_k using $H_k(z)$. The residual error $RErr_j$ for the entire frame corresponding to the stored syllable s_j in cluster C is.

$$RErr_j = \sum_{k=1}^{M}\sum_{n=1}^{N}(e_k^2[n]) \tag{1}$$

where N is number of samples in a speech frame(22.5ms). M is number of speech frames in the current speech segment. The syllable index corresponding to minimum residual error is

$$MinSylIndex = \arg\min_j(RErr_j) \quad j = 1..P \tag{2}$$

Method II:- Figure 2 details the encoder part of the segment vocoder that *does not* perform recognition but only uses residual energy. In this method *Viterbi decoding (recognition)* is omitted. The *Minimum residual block* in the figure, calculates the minimum error energy of all the syllables in the training data.

Let T be the total number of syllables in the training data. The j th syllable is denoted by s_j. The residual error $RErr_j$ for the entire frame corresponding to the stored syllable s_j

$$RErr_j = \sum_{k=1}^{M}\sum_{n=1}^{N}(e_k^2[n]) \tag{3}$$

where N is number of samples in a speech frame(22.5ms). And M is number of speech frames in the current speech segment. The syllable index corresponding to minimum residual error is

$$MinSylIndex = \arg\min_{j}(RErr_j) \quad j = 1..T \tag{4}$$

B. Decoder

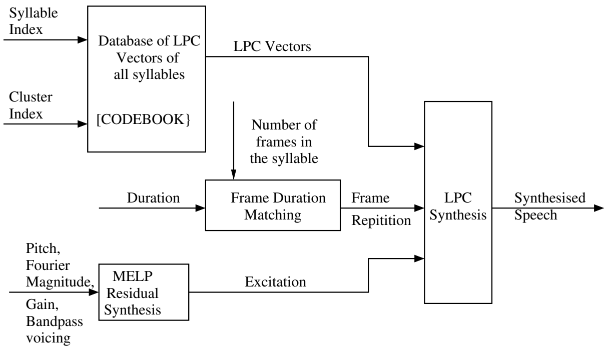

Fig. 3. *Speech Decoder in Method I*

Figure 3 shows the decoder for the segment vocoder for Method I. Together, the cluster index and syllable index identify the appropriate syllable in the codebook. All the parameter that were transmitted/stored during analysis are obtained. The cluster index and syllable index identify the closest syllable. Duration matching is performed. Using the corresponding LPC and other non-LPC parameters, frame by frame MELP synthesis is performed to produce the synthesized speech. The decoding in Method II is same as in Method I except that no cluster index is available and thus syllable index is the only input to the codebook.

C. Bit Allocation

The new vocoder requires only 49bps more than its previous counterpart [1] for transmission or storage. In the previous work[1] only the cluster centroid was used to encode a syllable. Since there is only one cluster centroid per cluster, use of cluster Index was sufficient. In this work we need to find the best syllable in a cluster and therefore we need an index(we call it syllable index) to identify the best syllable in the cluster. It is to be noted that a syllable index does not identify a syllable in the entire set of syllables. It merely identifies the best syllable in the best cluster. Assuming a maximum of 128 syllables per cluster we have added 7bits for syllable index within a cluster. And assuming a syllable rate of 7syllable/sec this adds 49 extra bits per second. In FS MELP [6] unit of transmission is a frame. In segment vocoder the unit of transmission is a syllable. An entire syllable is encoded as a unit.

TABLE I
Syllable encoding format in Method I

Duration (N)	Cluster Index	Syl Index	Residual Frame1	.....	Residual Frame N

TABLE II
Syllable encoding format in Method II

Duration (N)	Syl Index	Residual Frame1	.....	Residual Frame N

Encoding format of an entire syllable is shown in Table I for Method I and in Table II for Method II. Bit allocation for the syllable encoding is given in Table III for Method I and Table IV for Method II. Table V gives the bit allocation for each frame residual. Assuming a syllable rate of 7 syllables/second, the bit rate becomes (29*44.44 + (11+4+7)*7) ∼=1289+77+28+49=1443.

TABLE III
Bit allocation per syllable(Duration + Cluster Index + Syllable Index) in Method I.

Paramerters	**Number of bits**
Cluster Index	11
Syllable Index (within cluster)	7
Syllable Duration (Multiple of 22.5ms)	4

TABLE IV
Bit allocation per syllable(Duration + Cluster Index + Syllable Index) in Method II.

Paramerters	**Number of bits**
Syllable Index (In entire training set)	18
Syllable Duration (Multiple of 22.5ms)	4

TABLE V
Bit Allocation for excitation + other non-LPC parameters per 22.5ms frame(Same as 2.4Kbps Federal Standard MELP).

Paramerters	**VOICED**	**UNVOICED**
FS Magnitude	8	-
Gain(2 per frame)	8	8
Pitch, overall voicing	7	7
Bandpass voicing	4	-
Aperiodic flag	1	-
Error protection	-	13
Sync bit	1	1

III. Experiments, results and discussion

The experiments are conducted on DBIL(database for Indian Languages)[7] Tamil news database. DBIL consists of news bulletins in 8 Indian languages sampled at 16KHz and quantized with 16bits/sample. For this work, 1564 Tamil sentences uttered by 19 female speakers are considered for preparing the segment codebook. The test database consists of 174 sentences. The speech data is downsampled to 8Khz for the experiments. Group delay based segmentation [5] on 1564 sentences resulted in 15742 syllables which are clustered into 1780 clusters using unsupervised HMM based clustering algorithm. The segment codebooks are prepared separately for Method I and Method II.

IV. Performance Evaluation

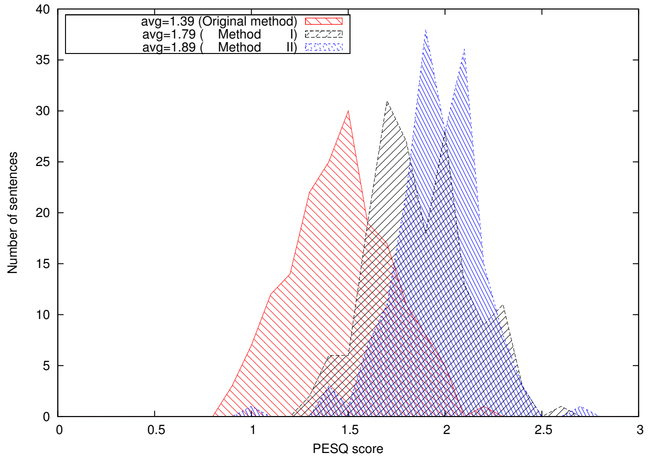

Fig. 4. *Histogram of PESQ scores for Original method, Method I, Method II*

Informal listening tests on the synthesized speech confirms improvement for all the test sentences except for a few. Since obtaining subjective scores like MOS for such large set of test sentences is laborious and time consuming, objective measure like PESQ [8] is used to evaluate the speech quality of the synthesized speech. Also since we have the original sound waveforms PESQ is more relevant in this particular context. The PESQ scores for the test sentences are obtained for three methods viz original method [1], vocoder with Method I type of encoding, and vocoder with Method II type of encoding. To capture the variation in the results for all the three methods, the results are plotted as histograms. Figure 4 shows the histograms of the three methods. The following observations are made from the plot.

1) The average PESQ score has improved from 1.39 [1] to 1.79 for Method I and 1.89 for Method II.
2) The scores for most test sentences (in the proposed framework) result in higher PESQ scores. While the highest PESQ scores achieved for the original vocoder is 2.2, the PESQ scores for the proposed methods achieves 2.7.
3) The narrower histograms do indicate most sentences are synthesised with PESQ around the average value.

The above PESQ scores are in comparison with MELP. MELP(2400bps) gives an average PESQ of 2.45. And we are able to achieve ∼40% reduction (((2400-1443)/2400)* 100%=39.87%) in bitrate. It has been achieved by reducing the bits required for the system encoding. No attempt has been made to reduce the bits required for the source encoding. The source used in this work is same as MELP source encoding. From the experiments, it is obvious that the quality of the speech signal significantly improves when minimum residual error energy is used for segment quantization. Method I for system encoding shows that even if only minimum residual is considered for selecting the representative syllable, PESQ quality can be improved. But, for Method II, the time taken for encoding is large because the codebook contains a large number of syllable segments and since no clustering is performed. Further, the technique is computationally intense as a linear search is performed to obtain the appropriate syllable. This problem can be addressed by designing a hierarchical codebook based on minimum residual energy.

V. Conclusions

The new segment vocoder synthesizes better quality speech compared to [1]. The improvement is achieved using the minimum-error segment quantization in tandem with recognition. The improvement comes at the cost of 49 extra bits per second. Method I is easy to compute whereas Method II is computationally intensive. A hierarchical arrangement of the syllables in the codebook could be used to reduce the searching time complexity in Method II. Although Method II averages a PESQ score of 1.89, MELP gives an average of 2.45. There is still room for improvement. Reducing spectral distortion between consecutive syllables could be explored to achieve further improvement.

References

[1] Sadhana Chevireddy, Hema A. Murthy and C. Chandra Sekhar, “Signal processing based segmentation and HMM based acoustic clustering for a syllable based segment vocoder at 1.4Kbps”, 16th European conference on Signal Processing and Communication, EUSIPCO-2008, Lausanne, Switzerland.

[3] Alan V. McCree and Thomas P. Barnwell, “A Mixed excitation LPC Vocoder Model for Low Bit Rate Speech Coding”, *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, p242-250, July 1995.

[4] V. Ramasubramanian and T. V. Sreenivas, “Automatically Derived Units for Segment Vocoders”, *Proc. ICASSP*, p473-476, 2004.

[5] T. Nagarajan and Hema A Murthy, “Group Delay based segmentation of spontaneous speech into syllable-like units”, *EURASIP Journal of Applied Signal Processing*, vol-17, p2614-2625, 2004.

[6] Supplee L.M., Cohn R.P., Collura J.S., McCree A.V., “MELP: the new Federal Standard at 2400 bps”, *Proc. ICASSP*, p1591-1594, April 1997.

[7] “Database for Indian Languages”, Speech and Vision Lab, IIT Madras, India, 2001.

[8] PESQ, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, http://www.itu.int/rec/T-REC-P.862/en.

[10] J. Cernocky, G. Baudoin, and G. Chollet. “Segmental vocoder - going beyond the phonetic approach”, *Proc. ICASSP*, vol-2, p605-608, 1998.

[11] K. Tokuda et al. “A very low bit rate speech coder using HMM-based speech recognition / synthesis techniques”, *Proc. ICASSP*, vol-2, p609-612, May 1998.

[12] G.L. Sarada, N. Hemalatha, T. Nagarajan, and Hema A Murthy, “Automatic transcription of continuous speech using unsupervised and incremental training”, *Proc. INTERSPEECH*, p405-408, 2004.

[13] J. Picone, and G. Doddington. “A phonetic vocoder”. *Proceedings of ICASSP*, vol-2, p580-583, May 1989.

[14] M. Felici, M. Borgatti, and R. Guerrieri. “Very low bit rate speech coding using a diphone-based recognition and synthesis approach”. *IEE electronic letters*, vol-34, no.9, p859-860,1998.

[15] Y. Shiraki, and M. Honda. “LPC speech coding based on variable-length segment quantization”. *IEEE transactions on Acoustics, Speech and signal Processing*, vol-36, p1437-1444,1988.

[16] S. T. Brandenhagen, K. L. Brown, and R. D. Braun. “Low bit rate speech compression using Hidden Markov Models”. *Proc. of MILCOM*, vol-1, p507-511,1997.