COMPARISON OF SEGMENT QUANTIZERS: VQ, MQ, VLSQ AND UNIT-SELECTION ALGORITHMS FOR ULTRA LOW BIT-RATE SPEECH CODING

D. Harish V. Ramasubramanian

Siemens Corporate Technology - India, Bangalore, India
D.Harish@siemens.com, V.Ramasubramanian@siemens.com

ABSTRACT

We consider the class of segment quantizers used for low to ultra-low rate speech coding, namely, vector quantization (VQ), matrix quantization (MQ), variable-length segment quantization (VLSQ) and two more recent unit-selection based segment quantization algorithms which represent an important shift in using large unclustered continuous codebooks in contrast to the conventional clustered codebooks of VQ, MQ and VLSQ. We examine the advantage, if any, in this shift from small clustered codebooks of VQ, MQ and VLSQ (10-14 bits/segment), to the larger continuous unit databases (16-18 bits/segment) in the unit-selection framework, by comparing their rate-distortion curves. We show that while the conventional VQ, MQ have higher distortions and VLSQ satuarates in its distortion reduction, unit-selection algorithms provide lower distortions and steeper reductions at marginally low increase in bit-rates and justifies exploring their potential further.

Index Terms: Speech coding, segment quantization, vocoders, unit-selection

1. INTRODUCTION

Segment vocoders based on variable-length segment quantization has provided the means of achieving low to ultra low bit-rates in the range of 800 to 150 bits/sec while offering intelligible speech quality [1], [2], [3], [4]. The basic functioning of a segment vocoder can be given as follows:

1. Segmentation of input speech (a sequence of LP parameter vectors) into a sequence of variable length segments.
2. Segment quantization of each of these segments using a segment codebook and transmission of the best-match code - segment index and input segment duration.
3. Synthesis of speech by LP synthesis using the code-segment time-normalized to match input segment duration.
4. The residual obtained by LP analysis is parameterized and quantized; the residual decoder reconstructs the residual to be used for synthesis in step (3).

The main issues in the above segment vocoder framework are, i) The definition of segmental units used for segment quantization, ii) How segmentation (step-1) and segment quantization (step-2) are realized and, iii) The type of segment codebook.

The definition of unit is implicitly tied to the manner in which segmentation and segment quantization are performed. Use of segments of fixed length (l) obviates an explicit segmentation step and reduces to vector ($l = 1$) and matrix quantization ($l > 1$). The VQ-LPC coder [1] marked an important milestone in low bit-rate coding by applying the then emerging concept of vector quantization (VQ) to quantize the LP parameters as a vector for each frame of speech as against the conventional scalar quantization of the parameters as in the standard LPC-10 vocoder. This brought about a remarkable reduction of the bit-rate from 2.4 Kbits/s to 800 bits/s while preserving the quality of the LPC-10 vocoder. Matrix quantization further extended the notion of VQ to deal with 'fixed-length segments' at a time and saw the emergence of matrix-quantization based LPC-10 system, which reduced the bit-rate to 300 bits/s while preserving the quality of coded speech as same as that of LPC-10 [2].

With respect to variable-length segments, segment vocoders have explored a variety of units such as diphone units, phonetic units, automatically derived units, etc. These techniques emerged to deal with 'variable-length' segments to exploit the variable durations of speech units (typically, phones) and then quantize them efficiently using structured or unstructured 'segment' codebooks. Much of the basic architecture in this framework was laid by [3], and further by Shiraki and Honda [4]. With respect to how segmentation and segment quantization is performed, Shiraki and Honda [4] proposed an important framework wherein segmentation and segment quantization were performed in a single step, using the 2-level dynamic programming algorithm with a segment codebook designed by an iterative joint-segmentation and clustering procedure. The segment quantization essentially performs a 'connected segment recognition' and determines the optimal segment boundaries (and hence the segment lengths) and the segment labels which are transmitted and used for reconstructing speech at the decoder after length normalization.

With respect to the segment codebook, it can be noted that almost all the segment vocoders used 'clustered codebooks' of the corresponding 'units' (i.e., VQ or MQ codebooks or VLSQ codebooks). However, in what can be considered a very significant convergence of recognition, synthesis and coding, Lee and Cox [5], [6] proposed a sub-1000 bits/s coder which operated on the principles of 'unit selection' that is normally employed in text-to-speech synthesis using the concatenative synthesis methodology. Here, a large codebook (actually a continuous speech database) is used for selecting the appropriate segments that best match the input speech using a modified Viterbi decoding principle that incorporates the costs of both the segment quantization and the segment-to-segment continuity.

In a further development in the unit-selection based segment quantization approach [7], we analyzed the algorithm of Lee and Cox, 2002 [6], to show how it intrinsically suffers from several sub - optimalities, such as due to pre-quantization of the test utterance using an intermediate Shiraki-Honda clustered segment codebook, and resulting fixing of unit labels and segment boundaries in test speech as well as the use of only a sub-set of units from the unit-database for concatenative unit-selection. In this recent work [7], we proposed a unified and generalized framework for segment quantization of speech at ultra low bit-rates of 150 bits/sec based on unit-selection principle using a modified one-pass dynamic programming algorithm [7] and showed how it is optimal for both fixed and variable-

length segments and how it solves the sub-optimality of the Lee and Cox, 2002 [6] algorithm by performing unit-selection based quantization 'directly' using the units of a continuous codebook without pre-quantizing the input speech.

In this paper, our objective is to benchmark the performances of all these segment quantization algorithms using rate-distortion curves. This has more or less not been attempted at all, though [4] provides the early comparisons between VQ, MQ and VLSQ. However, here we put these early algorithms in perspective with respect to the recent unit-selection algorithms cited above. By this, our main intention is to bring out the important differences between the classical segment quantization schemes (VQ, MQ and VLSQ) and the current unit-selection based segment quantization algorithms, and provide insights into these differences and the causative factors. Primarily, as noted earlier, the difference comes about in terms of the classical quantizers using clustered segment codebooks (fixed and variable length segments) and the use of large (long) continuous unit databases as in concatenative TTS by the unit selection algorithms. The question that arises is regarding what particular advantage does the use of very large continuous unit databases bring about (in the range of 16-18 bits/segment), in comparison to the much smaller clustered codebook sizes that VQ, MQ and VLSQ use (in the range of 8-10 bits/segment). Moreover, the early work of Lee and Cox did not also concern itself with quantifying the segment quantizaton performances in terms of rate-distortion curves, or answer the above question of what particular advantage has been gained by resorting to the unit-selection principles using large continuous codebook sizes. This paper essentialy attempts to answer this.

2. PRINCIPLES OF UNIT-SELECTION BASED SEGMENT QUANTIZATION

Fig. 1 shows a schematic of the optimal unit-selection framework proposed us [7].

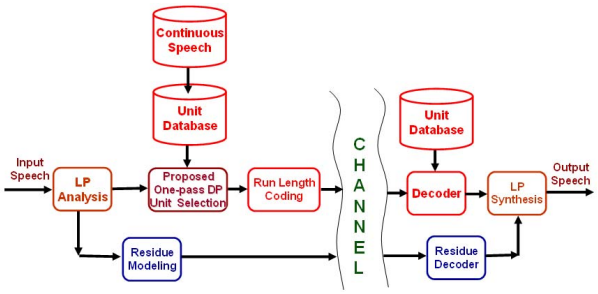

Fig. 1. *Optimal unit-selection based segment quantization [7]*

Consider a 'continuous codebook' which is essentially a sequence of MFCC or linear-prediction (LP) vectors as occurring in continuous speech. Let this codebook be viewed as being composed of N variable length segments $(u_1, u_2, \ldots, u_N)$, where a unit u_n is of length l_n frames, given by $u_n = (u_n(1), u_n(2), \ldots, u_n(l_n))$. The codebook is said to be made of 'fixed length' units, if $l_n = l, \forall n = 1, \ldots, N$, i.e., each unit has l frames (when $l = 1$, the codebook is said to be a 'single-frame' codebook). The codebook is said to be made of 'variable length' units if l_n is variable over n.

Let the input speech utterance which is to be quantized using the above codebook be a sequence of vectors (MFCC or LP parameters) $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$. Segment quantization, in its most general form involves segmenting and labeling this sequence of vectors $\mathbf{O}$ by a 'decoding' or 'connected segment recognition' algorithm which optimally segments the sequence and quantizes each segment by an appropriate label or index from the codebook. The segment indices and segment lengths together constitute the information to be transmitted to the decoder at the receiver, which then reconstructs a sequence of vectors by concatenating the segments of the received indices after normalizing the original segments in the codebook to the received segment lengths.

Consider an arbitrary sequence of K segments $S = (s_1, s_2, \ldots, s_{k-1}, s_k, \ldots, s_K)$ with corresponding segment lengths $(L_1, L_2, \ldots, L_k, \ldots, L_K)$. This segmentation can be specified by the segment boundaries $B = ((b_0 = 0), b_1, b_2, \ldots, b_{k-1}, b_k, \ldots, (b_K = T))$, such that the k^{th} segment s_k is given by $s_k = (\mathbf{o}_{b_{k-1}+1}, \ldots, \mathbf{o}_{b_k})$. Let each segment be associated with a label from the codebook, with each index having a value from 1 to N; let this index sequence be $Q = q_1, q_2, \ldots, q_{k-1}, q_k, \ldots, q_K$.

We propose here a constrained one-pass dynamic-programming algorithm which performs an optimal segment quantization by employing 'concatenation costs' in order to constrain the resultant decoding by a measure of how 'good' is the sequence Q with respect to ease of run-length coding (described in Sec. 2.1).

The optimal decoding algorithm solves for K^*, B^*, Q^* so as to minimize an overall decoding distortion (quantization error) given by

$$D^* = \arg \min_{K,B,Q} [\alpha \sum_{k=1}^{K} D_u(s_k, u_{q_k}) + (1-\alpha) \sum_{k=2}^{K} D_c(q_{k-1}, q_k)] \quad (1)$$

Here, $D_u(s_k, u_{q_k})$ is the unit-cost (or distortion) in quantizing segment s_k using unit u_{q_k}. This is as measured along the optimal warping path between s_k and u_{q_k} in the case of the one-pass DP based decoding which is described in Sec. 4. $D_c(q_{k-1}, q_k)$ is the concatenation-cost (or distortion) when unit $u_{q_{k-1}}$ is followed by unit u_{q_k}, which is given by

$$D_c(q_{k-1}, q_k) = \beta_{k-1,k} \cdot d(u_{q_{k-1}}(l_{q_{k-1}}), u_{q_k}(1)) \quad (2)$$

where, $d(.,.)$ is the Euclidean distance between the last frame of unit q_{k-1} and the first frame of unit q_k. $\beta_{k-1,k} = 0$, if $q_k = q_{k-1} + 1$ and $\beta_{k-1,k} = 1$ otherwise. This favors quantizing two consecutive segments (s_{k-1}, s_k) with two units which are consecutive in the codebook; run-length coding (Sec. 2.1) further exploits such 'contiguous' unit sequences to achieve lowered bit-rates.

2.1. Run-length coding and effective bit-rate

Run length coding refers to the following coding scheme applied on the decoded label sequence obtained as a solution to Eqn. (1). Let a partial sequence of labels in Q^* be $(\ldots, q_{i-1}, q_i, q_{i+1}, q_{i+2}, \ldots, q_{i+m-1}, q_{i+m}, \ldots)$ which are such that $q_{i-1} \neq q_i$, $q_{i+j} = q_i + j, j = 1, \ldots, m-1$ and $q_{i+m-1} \neq q_{i+m}$. The partial sequence $(q_i, q_{i+1}, q_{i+2}, \ldots, q_{i+m-1})$ is referred to as a 'contiguous group' with a 'contiguity' of m, i.e., a group of m segments whose labels are consecutive in the unit codebook. Run-length coding exploits this contiguity in coding the above contiguous group by transmitting the address of unit q_i first (henceforth referred to as the base-index), followed by the value $m-1$ (quantized using an appropriate number of bits). At the decoder, this indicates that q_i is to be followed by its $m-1$ successive units in the codebook, which the decoder retrieves for reconstruction. Naturally, all the m segment lengths $l_{i+j}, j = 1, \ldots, m-1$ are quantized and transmitted as in a normal segment vocoder.

Use of an appropriate concatenation cost favors the optimal label sequence to be 'contiguous' thereby aiding run-length coding and decreasing the bit-rate effectively. The unit-cost represents the spectral distortion and the concatenation cost (indirectly) the bit-rate; a

trade-off between the two costs allows for obtaining different rate-distortion points for the above algorithm. This is achieved by the factor α (which takes values from 0 to 1).

The effective bit-rate with the run-length coding depends entirely on the specific contiguity pattern for a given data being quantized. For a given input utterance $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, let $Q^* = q_1^*, q_2^*, \ldots, q_{k-1}^*, q_k^*, \ldots, q_{K^*}^*$ be the optimal labels obtained by the one-pass DP algorithm as above. Let there be P 'contiguous groups' in this K-segment label sequence, given by $g_1, g_2, \ldots, g_p, \ldots, g_P$, where the group g_p has a 'contiguity' c_p, i.e., c_p segments whose labels are contiguous in the unit codebook. Then the total number of bits $\mathbf{B}$ for quantization of the input utterance $\mathbf{O}$ with run-length coding is given by,

$$\mathbf{B} = P \cdot \log_2 N + P \log_2 c_{max} + K^* \log_2 L_{max} \qquad (3)$$

where, the first term is the total number of bits for the base-indices for the P contiguous groups, each being quantized to the address of the size N continuous codebook. The second term is the number of bits for the 'contiguity' information (providing for a maximum contiguity of c_{max} units) and the third term is the number of bits for the individual segment lengths in the K^* segment solution (providing for a maximum length of L_{max} frames). The effective bit-rate in bits/second is obtained by dividing this total number of bits $\mathbf{B}$ by the duration of the speech utterance Tf, for an input of T frames with a frame-size of f ms (20ms in this paper).

3. LEE AND COX UNIT-SELECTION ALGORITHM

Fig. 2 gives the system proposed by Lee and Cox [6] for unit-selection based segment quantization for variable-length segmental units, i.e., to essentially realize a solution as specified by Eqn. (1).

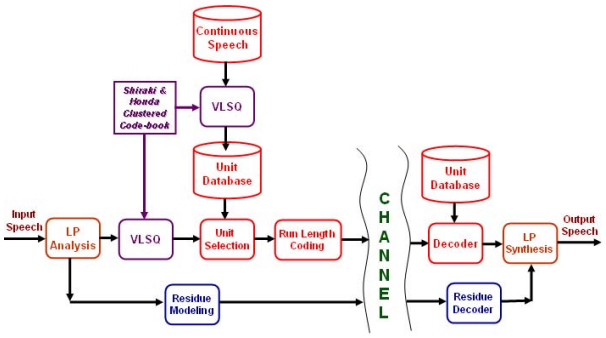

Fig. 2. *Schematic of the segmental unit-selection algorithm proposed by Lee and Cox, 2002 [6]*

Here, they used a continuous codebook, i.e. an 'unit database' of continuous sequences of mel-frequency cepstral coefficient (MFCC) vectors as obtained from continuous speech. Here, this 'unit-database' is derived from continuous speech by segmenting and quantizing (i.e., labeling) the continuous speech using a 'clustered' codebook designed by the joint-segmentation quantization algorithm of Shiraki and Honda [4]. By this, the database now becomes a codebook of variable-length segments with each segment having an index from the clustered codebook. Lee and Cox [6] use this segmented and labeled database for a second stage quantization of the input speech, which is also segmented and quantized by the same clustered codebook. Here, they apply a Viterbi decoding based unit selection procedure on a trellis of segment distortion values for segment quantization. The Viterbi decoding uses concatenation costs which favor quantizing consecutive segments of input speech using consecutive units in the 'continuous codebook'. The system then exploited this 'index-contiguity' to perform a run-length coding and achieving low effective bit-rates though the codebook sizes used could be large.

4. OPTIMAL UNIT-SELECTION DECODING

In contrast to the above 2-stage quantization solution for Eqn. (1), we had proposed a modified one-pass dynamic programming algorithm to solve the above decoding problem of Eqn. (1) optimally; the details of this algorithm can be found in [7].

The Viterbi algorithm used by Lee and Cox [5] with a 'single-frame' continuous codebook is a special case of this one-pass DP algorithm in [7] when the units in the continuous codebook are of fixed length one. For variable length units, this optimal algorithm performs a decoding of the input utterance 'directly' using the units of the unit codebook, unlike the two-stage procedure of Lee and Cox [6] which uses an intermediate segmentation (and labeling) using a clustered codebook (of size 64) followed by a forced-alignment Viterbi decoding. As a result, we do not incur any of the sub optimalities that the algorithm in [6] in Sec. 3 suffers from. Thus, the above algorithm handles fixed-length segments of any size as well as variable length segments in a unified and optimal manner without taking recourse to two different ways of decoding as was done in [5] and [6] for single-frame and variable-length units respectively.

5. EXPERIMENTS AND RESULTS

We now present results comparing the following segment quantizers, namely, vector quantization (VQ) [1], matrix quantization (MQ) [2], variable-length segment quantization (VLSQ) [4] and the unit-selection algorithms of [6] and [7]. The comparison is mainly in terms of quantization accuracy using rate-distortion curves between spectral distortion and the effective bit-rate (as appropriate in each case). We measure the segment quantization performance in terms of the average spectral distortion between the original sequence of linear-prediction vectors and the sequence obtained after segment quantization and length renormalization. The average spectral distortion is the average of the single frame spectral distortion over the number of frames in the input speech; the single frame spectral distortion is the squared difference between the log of the linear-prediction power spectra of the original frame and the quantized frame, averaged over frequency. The effective bit-rate for segment quantization for the two unit selection algorithms is measured as given in Eqn. (3) in Sec. 2.1 using the run-length coding. We have used the TIMIT database for all the experiments.

In Fig. 3, we show the rate-distortion performance of these 5 quantizers/algorithms, obtained through different frame / matrix / segment codebook sizes for VQ, MQ and VLSQ and unit-datbase sizes for the unit-selection algorithms. For vector quantization, the VQ codebooks of size 16 to 2048 (4 to 11 bits/frame) were designed from 48000 frames of training data (320 sentences from 32 speakers, 16 male and 16 female) using the LBG algorithm and used for spectral quantization as given in [1]. For matrix quantization, MQ codebooks of size 16 to 2048 (4 to 11 bits/matrix) were designed for matrix block-sizes of 2 and 4 from the same training data as for VQ and used for quantization as in [2]. The VLSQ codebooks of size 32 to 16384 (5 to 14 bits/segment) were designed by the joint segmentation and quantization algorithm of [4] from 90000 frames of training data (600 sentences from 60 speakers, 30 male and 30 female) and used for segment quantization as described in Sec. 1.

For both the unit-selection algorithms used here [6] and [7], we use the same continuous speech codebook as the 'unit database' which is a continuous sequence of linear-prediction vectors (log-area ratios) of continuous speech utterances in the TIMIT database,

treated as being made of variable sized units as defined by the manually defined phonetic units. We have used 'unit databases' of size ranging from 512 to 131072 corresponding to bit-rates of 9 to 17 bits. These are the first 131072 phonetic segments of TIMIT sentences with male and female sentences interleaved, from 200 sentences from 20 speakers of nearly 2 hours of continuous speech.

The test data used for obtaining the R-D curves for all the quantizers was the same set of 8 sentences with 4 male and 4 female speakers from outside the speakers used in the codebook design for VQ, MQ and VLSQ and outside the unit-database for the unit-selection algorithms. In the rate-distortion curves in Fig. 3, the number along side each point in the curves is the effective bits/frame (which is essentially the codebook size in bits/segment divided by the average length of a segment in the codebook, i.e., frames / segments); this yields the effective bit-rate in bits/sec when multiplied by the frame-rate of frames/sec, which in this case is 50 frames/sec for a framesize of 20ms). The numbers shown alongside each point within paranthesis is the codebook size (in bits/segment or bits/unit as appropriate). Both these are given to facilatate a quick comparison of the R-D performance of the different quantizers, either with respect to a given codebook size (which is appropriate when comparing VLSQ and unit-selection) or with respect to bits/frame which is more appropriate when comparing VQ, MQ and VLSQ, since these quantizers differ in the segment size in their codebooks.

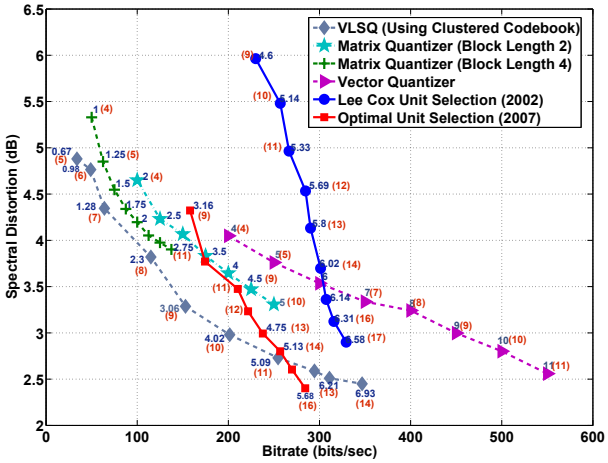

Fig. 3. *Rate-distortion curves for VQ [1], MQ [2], VLSQ [4] and the two unit-selection algorithms: i) Optimal algorithm [7] and ii) Lee-Cox'02 2-stage algorithm [6])*

We observe the following from this set of R-D curves: **i)** The VQ and MQ family of curves are expected, with MQ of larger block sizes providing a left and downward shift the R-D curve; the reduction in spectral distortion (SD) for increase in block-size from 1 (VQ) to MQ(2) and MQ(4) for the same bits/frame is quite evident. **ii)** VLSQ offers improvement over VQ and MQ though only marginally with respect to MQ of block size 4. **iii)** When we shift to the unit-selection algorithms of Lee and Cox [6] or the optimal algorithm of [7], the codebook is unclustered and therefore these R-D curves have higher distortion for a given codebook size when compared to the clustered codebook performances of VLSQ, at least up to the maximum of 14 bit codebooks of VLSQ we have used. **iv)** However, the unit-selection algorithm reduce the spectral distortion more rapidly for every doubling of the codebook size, thanks largely to the run-length advantage of unit-selection and the associated reduction in the effective bit-rate which does not increase in proportion to the base-index bit-rate of the full codebook size. This results in a steep fall in spectral distortion even within 400 bits/sec, while in contrast, VLSQ saturates at a SD of 2.5 dB for codebook sizes of size 16384. **v)** The advantage of the optimal unit-selection algorithm over the 2-stage sub-optimal unit-selection of Lee and Cox can also be noted. **vii)** This optimal algorithm starts offering spectral distortions lower than VLSQ for considerably smaller unit database sizes than the sub-optimal unit-selection algorithm, and at a significantly smaller effective bits/frame than both VLSQ and the sub-optimal unit-selection.

In summary, we note that the unit-selection framework does offer an interesting rate-distortion trend of rapidly decreasing the spectral distortion for increase in the unit-database size, i.e., a steeper rate-distortion curve when compared to the VQ, MQ and VLSQ algorithms which tend to saturate in their spectral distortion reductions around codebook sizes of 10 to 14 bits/segment. This alone would be the distinctive factor that would allow unit-selection frameworks to offer distortion even in the range of 2 dB and less even though with use of very large continuous codebook sizes (perhaps exceeding even 18 bits / segment). More importantly, we believe issues related to computational complexity and memory and decoding latency time in the unit-selection algorithms will have to be addressed to take advantage of this rate-distortion trend and establish this class of segment quantizers as truly applicable for real ultra low bit-rate applications, in keeping with its seeming potential to offer low distortions with only marginal bit-rate increases, thanks to the run-length coding principles and advantages underlying the unit-selection framework.

6. CONCLUSIONS

We have considered the class of segment quantizers used for low to ultra-low rate speech coding, ranging from vector quantization (VQ), matrix quantization (MQ), variable-length segment quantization (VLSQ) and two more recent unit-selection based segment quantization algorithms. We have examined the advantage, if any, in using large unclustered continuous unit databases by the unit-selection algorithms, in comparison to the smaller clustered codebook sizes that VQ, MQ and VLSQ use, by comparing the rate-distortion curves of these quantizers. We have shown that while unlike VQ, MQ and VLSQ, the unit-selection algorithms tend to provide lower distortions and steeper reductions at marginally low increase in bit-rates and justify exploring their potential further.

7. REFERENCES

[1] D. Y. Wong et al. An 800 b/s vector quantization LPC vocoder. IEEE Trans. on ASSP, vol. 30, no. 6, pp. 770-780, Oct. 1982.

[2] C. Tsao and R. M. Gray. Matrix quantizer design for LPC speech using the generalized Lloyd algorithm. IEEE Trans. on ASSP, vol.33, no. 3, pp. 537-545, Jun 1985.

[3] S. Roucos, R. M. Schwartz, and J. Makhoul. A segment vocoder at 150 b/s. In *Proc. ICASSP'83*, pages 61–64, 1983.

[4] Y. Shiraki and M. Honda. LPC speech coding based on variable-length segment quantization. *IEEE Trans. on Acoust., Speech and Signal Proc.*, 36(9):1437–1444, Sept. 1988.

[5] K. S. Lee and R. V. Cox. A very low bit rate speech coder based on a recognition/synthesis paradigm. *IEEE Trans. on Speech and Audio Proc.*, 9(5):482–491, Jul 2001.

[6] K. S. Lee and R. V. Cox. A segmental speech coder based on a concatenative TTS. *Speech Commun.*, 38:89–100, 2002.

[7] V. Ramasubramanian and D. Harish. An optimal unit-selection algorithm for ultra low bit-rate speech coding. In ICASSP '07, April 2007, Hawaai.