# An unified unit-selection framework for ultra low bit-rate speech coding

V. Ramasubramanian D. Harish*

Siemens Corporate Technology - India
Siemens Information Systems Ltd., Bangalore - 560100, India
V.Ramasubramanian@siemens.com, Harish.D@iiitb.ac.in

10.21437/Interspeech.2006-61

## Abstract

We propose a unified 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. The algorithm handles both fixed- and variable- length units in a unified manner, thereby providing a generalization over two existing unit selection methods, which deal with 'single-frame' and 'segmental' units differently. The proposed algorithm performs unit-selection based quantization directly on the units of a continuous codebook, thereby not incurring any of the sub - optimalities of the existing 'segmental' algorithm. Moreover, the existing 'single-frame' algorithm becomes a special case of the proposed algorithm. Based on the rate-distortion performance on a multi-speaker database, we show that fixed length units of 6-8 frames perform significantly better than single-frame units and offer similar spectral distortions as variable-length phonetic units, thereby circumventing expensive segmentation and labeling of a continuous database for unit selection based low bit-rate coding.
**Index Terms:** Speech coding, ultra low bit-rate, segment vocoder, unit selection, one-pass DP algorithm

## 1. Introduction

Segment vocoders based on variable-length segment quantization provide the means of achieving ultra low bit-rates as low as 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 a 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 ($l > 1$). With respect to variable-length segments, segment vocoders have explored a variety of units such as diphone units [1], phonetic-segments [3], and automatically derived units [4].

With respect to how segmentation and segment quantization is performed, Shiraki and Honda [2] 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 the variable length segment codebooks). More recently, in what can be considered as a major paradigm shift in segment-quantization for very low bit-rate speech coding, Lee and Cox [5] proposed a system based on a recognition-synthesis paradigm. This has several important distinctions from the conventional segment vocoder structure sketched above. Firstly, they used a 'continuous codebook', which is a sequence of mel-frequency cepstral coefficient (MFCC) vectors as obtained from continuous speech; the codebook is thus a 'single-frame codebook', i.e. a codebook of single-frame vectors like a vector quantizer, but obtained without any clustering. Secondly, they employed a Viterbi decoding algorithm to perform segmentation and segment quantization using the 'unit selection' principle. Here, the Viterbi decoding uses concatenation costs which favor quantizing consecutive frames of input speech using consecutive frames in the 'continuous codebook'. The system then exploited this 'index-contiguity' to perform a run-length coding thereby achieving low effective bit-rates though the codebook sizes used were significantly large (19 bits/frame).

Subsequently, Lee and Cox also came up with a 'segmental' version of the above system in [6]. Here they used a similar large size 'continuous codebook' (called 'database', henceforth), but now segmented and quantized (i.e., labeled) by a 'clustered' codebook designed by the joint-segmentation quantization algorithm of Shiraki and Honda [2]. 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 again, they apply a Viterbi decoding based unit selection procedure, but now to aid run-length

* On internship from International Institute of Information Technology, Bangalore, India

coding on the unit indices of the database.

However, in extending the 'single-frame' unit-selection principle to a 'segmental codebook', Lee and Cox [6] had introduced several sub-optimalities in the segment quantization procedure, arising in three ways: i) Pre-quantization of the input speech before Viterbi decoding produces a segmentation that is sub-optimal with respect to the units of the actual units in the database, ii) Using only those segments from the database which have the labels of the pre-quantized input speech restricts the units available for quantization to a small sub-set of units, and iii) The unit selection Viterbi decoding essentially works only on segments defined by pre-quantization, and hence incurs a sub-optimality with respect to the overall spectral distortion of the final segmentation and quantization of the input speech with respect to the database units, which after all are the actual units used for synthesis at the receiver.

In this paper, we propose a unified framework for segmenting and quantizing the input speech using a constrained one-pass dynamic programming algorithm for performing unit-selection on continuous codebooks as used by Lee and Cox [5], [6]. Unlike the above sub-optimal algorithm in [6], the algorithm proposed here provides an optimal segment quantization of the input speech with respect to the 'units' of the continuous codebook. Moreover, unlike the very disparate ways in which Lee and Cox realized the single-frame unit selection [5] and segmental unit selection [6], our proposed framework provides a 'unified' approach to treating the continuous codebook as made up of segmental units which can be of two kinds: i) fixed lengths of arbitrary length (such as 1, 2, 3, 4, etc.) or, ii) variable lengths such as phone-like units or units as derived after segmenting (and labeling) the continuous speech database using a 'clustered codebook' (as done in [6]). By this, we achieve several advantages over the methods of [5] and [6]:

1. The framework is based on a single elegant algorithm which is a generalization of both the single-frame system [5] and segmental system [6],
2. This allows evaluation of the unit-selection based system for fixed unit sizes greater than 1; this was not attempted in [5] or [6]. We show that using long fixed sized units of 6-8 frames offers significantly improved performance over 'single-frame' units [5]
3. We also show that using fixed size units of 6-8 frames (that approximate phone-like segments) offers performance comparable to variable sized units (such as phonetic units), thereby completely obviating the need to segment and label the continuous speech database manually or automatically using phonetic or clustered codebooks.

## 2. Unit-selection framework

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. 3.1).

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. 3. $D_c(q_{k-1}, q_k)$ 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. 3.1) further exploits such 'contiguous' unit sequences to acheive lowered bit-rates.

## 3. Proposed one-pass DP algorithm

We propose a modified one-pass dynamic programming algorithm to solve the above optimal decoding problem of Eqn. (1). We first state the dynamic program recursions of our modified one-pass DP algorithm based unit-selection. The recursions are in two parts: within-unit recursion and cross-unit recursions.

*Within-unit recursion*

$$D(i,j,n) = \min_{k \in (j, j-1, j-2)} [D(i-1,k,n) + \alpha \cdot d(i,j,n)] \quad (3)$$

*Cross-unit recursion*

$$D(i,1,n) = \min(a,b) + \alpha \cdot d(i,1,n) \quad (4)$$

where,

$$a = D(i-1,1,n) \quad (5)$$

$$b = \min_{r \in (1,\ldots,N)} [D(i-1, l_r, r) + (1-\alpha) \cdot D_c(r,n)] \quad (6)$$

Here, the above two recursions are applied over all frames of all the units in the codebook for every frame $i$ of the input utterance. The within-unit recursion is applied to all frames in a unit which are not the starting frame, i.e., for $j \neq 1$; the cross-unit recursion is applied only for the starting frames of all units, i.e., for $j = 1$, to account for a potential entry into unit $n$ from the last frame $l_r$ of any of the other units $r = 1, \ldots, N$ in the codebook.

$D(i, j, n)$ is the minimum accumulated distortion by any path reaching the grid point defined by frame '$i$' of the input utterance and frame '$j$' of unit $u_n$ in the codebook. $d(i, j, n)$ is the local distance between frame '$i$' of the input utterance and frame '$j$' of unit $u_n$. $\alpha$ and $1 - \alpha$, respectively weigh the unit-cost and concatenation cost, thereby realizing Eqn. (1) and providing a parameter for controlling the relative importance of the two costs in determining the optimal path (this is described further in the next section on run-length coding). The final optimal distortion is given by,

$$D^* = \min_{n=1,\ldots,N} D(T, l_n, n) \tag{7}$$

The optimal number of segments $K^*$, segment boundaries $B^*$ and segment labels $Q^*$ (corresponding to this optimal $D^*$ in Eqn. (1)) are retrieved by back-tracking as in the conventional one-pass DP algorithm [7].

The Viterbi algorithm used by Lee and Cox [5] with a 'single-frame' continuous codebook is a special case of the above one-pass DP algorithm when the units in the continuous codebook are of fixed length one. For variable length units, the above 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 conventional forced-alignment Viterbi decoding. As a result, we do not incur any of the sub-optimalities that the algorithm in [6] suffers from, as pointed in Sec. 1. Thus, the above algorithm handles fixed-length segments of any size as well as variable length segments in a unified and optimal manner.

### 3.1. Run-length coding and effective bit-rate

Run length coding refers to the following coding scheme applied on the decoded label sequence obtained by the above one-pass DP algorithm that solves 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 $\alpha$ (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,

$$\mathbf{B} = P \cdot \log_2 N + \sum_{p=1}^{P} \log_2 c_p + \sum_{k=1}^{K^*} \log_2 l_{q_k} \tag{8}$$

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 and the third term is the number of bits for the individual segment lengths in the $K^*$ segment solution. 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).

## 4. Experiments and Results

We now present results of the proposed unit-selection based segment quantization algorithm with respect to its quantization accuracy in terms of rate-distortion curves between spectral distortion and the effective bit-rate with run-length coding. 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 (i.e., steps 1 to 3 in Sec. 1). 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 bit-rate for segment quantization is measured as given in Eqn. (8) in Sec. 3.1 using the run-length coding. We have used the TIMIT database for all the experiments. We have used a value of $\alpha = 0.5$ (Eqns. (1), (3), (4), (6)) in all the experiments, giving equal weightage to both unit-cost and concatenation cost.

In Fig. 1, we show the rate-distortion performance of the unit-selection algorithm for two kinds of unit sizes: i) fixed length units with lengths ranging from 1 to 8 and ii) variable-length phonetic units. In both cases, the codebook is a continuous sequence of linear-prediction vectors (log-area ratios) of continuous speech utterances in the TIMIT database, but treated as being made of fixed length units or variable sized units. Since TIMIT is phonetically segmented, we used this phonetic segmentation to define the variable-length units. This represents the best performance achievable for variable length units, such as when the automatic segmentation used to obtain the units is as good as manual segmentation that defines phonetic segments. In both cases, we have used codebooks of sizes 32 to 4096 which are essentially the first 32 (or 4096) vectors of the TIMIT sentences ordered with male and female sentences interleaved. The numbers along side each curve is the codebook size (in bits/unit). The number of sentences used to form these codebooks range from 1 to 128 sentences. The test data used was 8 sentences with 4 male and 4 female speakers from outside the speakers used in the codebook.

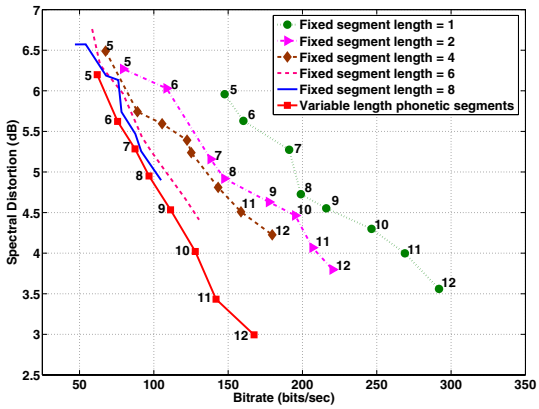


Figure 1: *Rate-distortion curves for different fixed length units and variable-length phonetic units*

From this figure, it can be observed that the effective bit-rate reduces significantly (nearly halves, such as from 200 bits/s to 100 bits/s), with increase in the fixed length unit-size from 1 to 4 to 8 frames. This is largely due to the fact that with a larger unit, the segment rate (number of segments per second) is reduced, and even without run-length coding, the number of bits used for base-index quantization would decrease proportionately. In addition, the use of run-length coding further reduces the effective bit-rate; the contiguity of larger length units implies that more frames are quantized with the same run-length base indices resulting in improved run-length advantage for longer fixed length units.

It can be noted that the variable length phonetic units performs the best, offering an halving of the bit-rate from 300 bits/s (for single-frame units) to 150 bits/s for the same distortion, clearly validating the potential of the unit-selection algorithm to gain rate-distortion with larger unit sizes that approximate phone-like units. However, fixed length units of length 6 and 8 (shown up to codebook sizes 4096 and 2048) also provide a performance comparable to that of variable length phonetic units. This circumvents the need for defining variable length units in a continuous codebook by automatically segmenting it or by other means. It would be sufficient to simply use a large continuous speech data and define fixed length units of lengths comparable to phonetic units.

The effect of increasing the fixed length on the run-length based bit-rate advantage is brought out clearly from the distribution of contiguity in Fig. 2 which plots the number of times a contiguity group of contiguity '$m$' occurs. As can be expected, the contiguity is high even for units of length one. With increase in the unit lengths from 1 to 2, 4, 6, and 8, and finally to the variable length phonetic units, the largest contiguity tends to come down, since each unit now already spans multiple frames. However, the effective number of frames grouped by a contiguity has increased considerably even with limited contiguity for longer units. For instance, for unit lengths of 4, a contiguity of 4 performs an effective run-length coding over 16 frames in comparison to the maximum of 9 frames of single-frame units.

It is important to note that, in Fig. 1, for longer fixed length units and variable length units, there is a sharper decrease in spectral distortion (SD) for a given bit-rate increase, in comparison to single-frame units. This steep trend in the rate-distortion curves of the proposed unit-selection algorithm even with fixed-length units, indicates that large reductions in spectral distortion can be

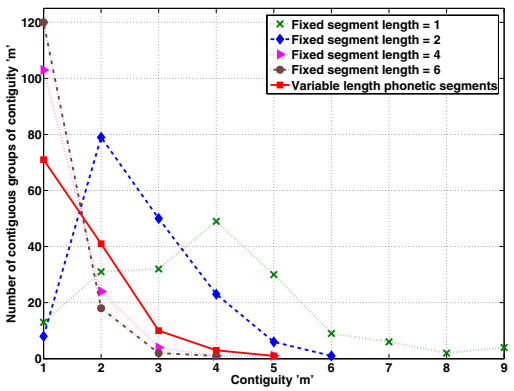


Figure 2: *Contiguity distribution for different fixed length units and variable length phonetic units*

achieved by using suitably large codebook sizes. This is particularly appealing since the codebook need not be 'designed' by any complex algorithms and nor does it have to be segmented (phonetically or otherwise) prior to use. Solutions to make the one-pass DP unit-selection algorithm perform with low computational complexities at very large codebook sizes, will enable it to achieve close to 1-2 dB average spectral distortions, which make this ultra low bit rate quantization as good as high rate spectral quantizers.

## 5. Conclusions

We have proposed a unified framework for segment quantization of speech at ultra low bit-rates of 150 bits/sec based on unit-selection principle using a constrained one-pass dynamic programming algorithm. The algorithm handles both fixed- and variable- length units in a unified manner, thereby providing a generalization over two existing unit selection methods, which deal with 'single-frame' and 'segmental' units in different ways. We show that fixed length units of 6-8 frames perform significantly better than single-frame units and offer the same spectral distortions as variable-length phonetic units, thereby circumventing expensive segmentation and labeling of a continuous database for unit selection based low bit-rate coding

## 6. References

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

[2] 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.

[3] J. Picone and G. Doddington. A phonetic vocoder. In *Proc. ICASSP'89*, pages 580–583, 1989.

[4] V. Ramasubramanian and T. V. Sreenivas. Automatically derived units for segment vocoders. In *Proc. ICASSP'04*, pages I-473–I-476, Montreal, Canada, 2004.

[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] H. Ney. The use of one-stage dynamic programming algorithm for connected word recognition. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, 32(2):263–271, Apr 1984.