AN OPTIMAL UNIT-SELECTION ALGORITHM FOR ULTRA LOW BIT-RATE SPEECH CODING

V. Ramasubramanian D. Harish

Siemens Corporate Technology - India, Bangalore, India

V.Ramasubramanian@siemens.com, D.Harish@siemens.com

ABSTRACT

In this paper, we first analyze an algorithm proposed recently by Lee and Cox, 2002, which attempts to perform a segment unit-selection for segmental quantization and show how it intrinsically suffers from several sub-optimalities. We then propose a generalized unit-selection algorithm for ultra low bit-rate segment quantization based on a modified one-pass dynamic programming algorithm. We show that this proposed algorithm is exactly optimal for both fixed and variable-length segments and also how it solves the sub-optimalities of the Lee-Cox-2002 algorithm. From rate-distortion curves from a very large continuous speech multi-speaker database, we show that our algorithm has a significantly superior performance than the algorithm of Lee and Cox by achieving considerably lower spectral distortions (up to 3 dB lower distortions) as well as much lower bit-rates for a given distortion over a range of unit database sizes.

Index Terms: Speech coding, vocoders, speech codecs, speech communication

1. INTRODUCTION

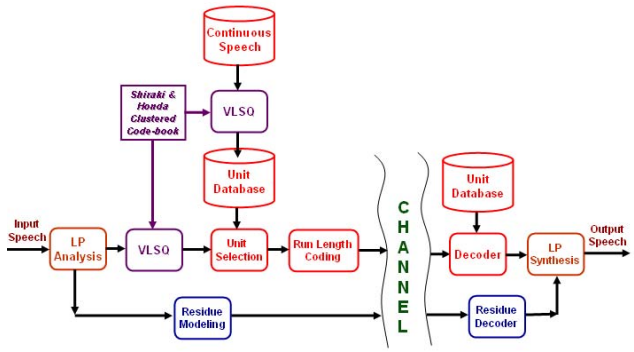

Fig. 1. *Schematic of the segmental unit-selection algorithm proposed by Lee and Cox, 2002 [4]*

Recently, in what can be considered as a major paradigm shift in segment-quantization for very low bit-rate speech coding, Lee and Cox [4] proposed a system based on a recognition-synthesis paradigm. This has several important distinctions from the conventional segment vocoder structure used so far for low bit-rate spectral quantization [1], [2], [3].

Fig. 1 gives a schematic of this system. 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 [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 [4] 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 thereby achieving low effective bit-rates though the codebook sizes used could be significantly large.

However, this algorithm by Lee and Cox [4] has several sub-optimalities in the segment quantization procedure, arising in the following ways:

1. 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
2. The use of a unit-selection quantization on such pre-segmented and pre-quantized input utterance results in further loss of optimality as the optimal segmentation (and the corresponding quantization) would be significantly different, particularly with respect to the overall spectral distortion
3. 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,
4. 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. Fig. 2 shows a schematic of this optimal unit-selection algorithm proposed here. Unlike the above sub-optimal algorithm in [4], the algorithm proposed by us here provides an optimal segment quantization of the input speech 'directly' with respect to the actual 'units' of the continuous codebook. By this, we achieve a significantly superior performance than the algorithm of Lee and Cox [4] with the proposed algorithm having considerably lower spectral distortions (up to 3 dB lower distortions) for a given bit-rate as well as much lower bit-rates for a given distortion than the algorithm by Lee and Cox [4] over a range of database sizes.

The algorithm proposed here was originally conceived as a 'unified' framework towards dealing with continuous codebooks of both fixed-length units of arbitrary length as well as variable lengths such as manually (or automatically) defined phone-like units [6]. In this earlier work, we showed that the unified algorithm allows achieving performance of variable length phonetic units using fixed length units of sufficient lengths (such as 6 to 8 frames). Here, we extend this algorithm to the case of dealing with a 'unit database' consisting of units derived after segmenting (and labeling) the continuous speech database using a 'clustered codebook' (as done in [4]) and show the advantages of the unified optimality of the proposed algorithm over the sub-optimal realization of the algorithm by Lee and Cox, 2002.

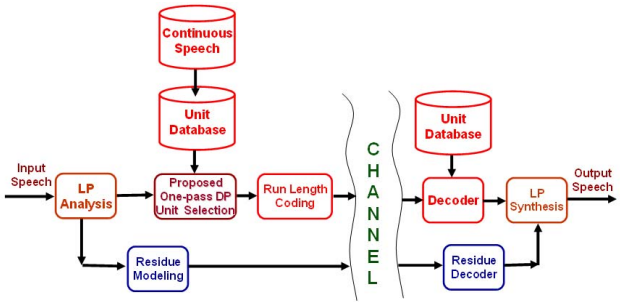

Fig. 2. *Schematic of the proposed optimal unified unit-selection algorithm based on a constrained one-pass dynamic programming (DP)*

2. PROPOSED OPTIMAL UNIT-SELECTION

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 achieve lowered bit-rates.

3. PROPOSED OPTIMAL 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. α 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,

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

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 [5].

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 [4] 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 [4] suffers from, as pointed out in Sec.1.

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 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,

$$\mathbf{B} = P \cdot \log_2 N + P \log_2 c_{max} + K^* \log_2 L_{max} \qquad (8)$$

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 (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).

4. EXPERIMENTS AND RESULTS

We now present results of the proposed unit-selection algorithm for segment quantization and compare it with the algorithm of Lee and Cox, 2002, [4] in terms of quantization accuracy using 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. 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. 3, we show the rate-distortion performance of the unit-selection algorithm proposed by us here and the algorithm of Lee and Cox, 2002, [4]. For both the algorithms, 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. Since TIMIT is phonetically segmented, we have used this phonetic segmentation to define the variable-length units for both the algorithms. This represents the best performance achievable for variable length units, and can be expected to provide an optimal baseline performance to the case when automatic segmentation is used to obtain the units such as using a clustered codebook (by the variable-length segment quantization (VLSQ) technique) as used in Lee and Cox, 2002 [4].

We have used 'unit databases' of size ranging from 512 to 65536 corresponding bit-rates of 9 to 17 bits. These are essentially the first 65536 phonetic segments of the TIMIT sentences ordered with male and female sentences interleaved, from about 200 sentences from 20 speakers constituting nearly 2 hours of continuous speech. The number along side each point in the curves is the codebook size (in bits/unit). In the case of the proposed algorithm we have used database size up to 8192, as this achieves the same spectral distortions as the Lee and Cox, 2002 algorithm [4] with a database size of 65536 and was hence adequate to bring out the performance advatange achievable (at significantly lower bit-rates), due to optimality of the proposed algorithm. The test data used was 10 sentences with 5 male and 5 female speakers from outside the speakers used in the codebook.

From this figure, the following important differences between the proposed optimal unified unit-selection algorithm and the sub-optimal algorithm of Lee and Cox' 02 can be noted:

1. In general, the rate-distortion curve of the proposed algorithm has the ideal shift towards left-bottom with a significant distortion and rate margins over the rate-distortion curve of Lee and Cox'02. This is as would be expected for an enhanced quantization scheme with both rate and distortion advantages.
2. Specifically, it can be seen that the proposed algorithm has significantly lower distortions for a given database size (given in bits alongside) and final effective bit-rate. For instance, for a database size of size 512 (9 bits), the spectral distortion of

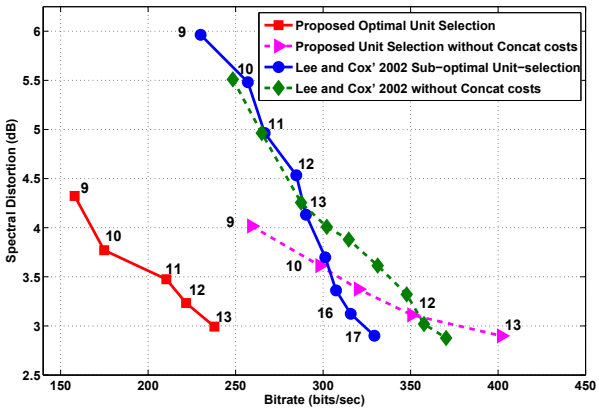

Fig. 3. *Rate-distortion curves for proposed optimal unified unit-selection algorithm and the 2-stage sub-optimal algorithm of Lee and Cox, 2002 [4]*

the proposed algorithm is about 1.5 dB less than that of Lee-Cox'02 algorithm with a corresponding effective bit-rate that is 75 bits/sec less.

3. For a size of 13 bits, the proposed algorithm is able to provide a much lower spectral distortions (as much as 3 dB less) than the Lee-Cox'02 algorithm at the same effective bit-rate. It should be noted that this 3 dB difference is highly significant for the ultra low-rate ranges being dealt with here.

4. It can be further noted that the 13 bit database with the proposed algorithm gives about 1.5 dB performance improvement over that of Lee-Cox'02 and at a much lower bit-rate.

5. Further, it can be noted that the Lee-Cox'02 algorithm needs extremely large database sizes (of the order of 17 bits which is 65536 segmental units or approximately 2 hours of continuous speech), to achieve distortions comparable to that achievable by the proposed algorithm with a much smaller database size 13 bits (8192 segmental units, or about 13 minutes of continuous speech, which is nearly 8 times less than that needed by Lee-Cox'02 algorithm).

We also show another important performance advantage of the proposed algorithm due to its optimality when compared to the sub-optimal algorithm of Lee and Cox, 2002 [4]:

1. Fig. 3 shows the rate-distortion curves of the two algorithms when the concatenation cost is not used; i.e., the Viterbi decoding in Lee and Cox, 2002 as well as the proposed one-pass DP constraints in this paper does not have the second term $\sum_{k=2}^{K} D_c(q_{k-1}, q_k)$ in Eqn. 1. By this, the two algorithms have better (i.e., lower) spectral distortions, since not using the concatenation constraint leads to more optimal decoding with respect to the unit-cost of term 1 in this equation. It can be observed that the proposed algorithm has significantly lower spectral distortions than Lee and Cox'02 for a given unit database size. This clearly brings out the effect of gain of optimality resulting from quantizing the input utterance 'directly' using 'all' the units of the database, unlike the two step procedure of Lee and Cox' 02 which uses an intermediate quantization (with a separate clustered codebook using the Shiraki and Honda VLSQ algorithm) and a subsequent unit-selection using a highly 'reduced' choice of units for each segment of the pre-segmented input utterance.

2. Further, it can be noted that the proposed algorithm gains significantly in achieving much lower effective bit-rates, once the concatenation-constraints are restored, again indicating another performance advantage of the optimality of the proposed algorithm: The proposed algorithm is able to produce more contiguous decoding, which in turn reduces the effective bit-rate with run-length coding. Again, this is due to the fact, the decoding is done with the entire unit-database, in comparison to the highly 'reduced' unit choices available in Lee and Cox'02 algorithm due to the pre-quantization using the clustered intermediate codebook.

While we have used phonetically defined variable length units as available in the TIMIT database, the above algorithms should in principle be used with a unit database defined automatically, i.e. with units defined by automatic methods such as the VLSQ method of Shiraki and Honda [2]. However, in an interesting result shown by us in an earlier work [6], it turns out it is possible to completely avoid such expensive segmentation and labeling (either manually or by automatic methods), by using fixed-length units of sufficient lengths (comparable to average phonetic units) such as 6 to 8 frames and still get rate-distortion performances comparable to what is possible with variable-length units. This leads to the conclusion that the 'optimal' algorithm proposed here is able to firstly overcome the sub-optimalities of the Lee and Cox' 2002 algorithm with a consequent improved rate-distortion performance and in addition, completely circumvent the need to have pre-defined variable-length units, as was obtained by using clustered codebooks in the Lee and Cox' 2002 algorithm.

5. CONCLUSIONS

We have brought out the intrinsic sub-optimalities of an algorithm proposed recently by Lee and Cox, 2002 for segmental unit-selection. We have proposed an alternative generalized unit-selection algorithm for segment quantization based on a modified one-pass dynamic programming algorithm. We have shown that the proposed algorithm is exactly optimal for variable length unit-selection based segment quantization and that it solves the sub-optimalities of the Lee-Cox-2002 algorithm. Based on rate-distortion curves from a very large continuous speech multi-speaker database, we have shown that our algorithm has a significantly superior performance than the algorithm of Lee and Cox with considerably lower spectral distortions (up to 3 dB lower distortions) as well as much lower bit-rates for a given distortion over a range of unit database sizes.

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

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

[5] 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.

[6] V. Ramasubramanian and D. Harish. An unified unit-selection framework for ultra low bit-rate speech coding. *Proc. Interspeech - 2006, ICSLP*, pp. 217-220, Pittsburgh, Sept. 2006.