Variable Frame Rate Speech Coding Using Optimal Interpolation

Chii-Jen Chung and Sin-Horng Chen

Abstract— **A VFR LPC vocoder using optimal interpolation is presented in this paper. In the encoder, some representative frames of an utterance are selected for transmission. In the decoder, LPC parameters of all untransmitted frames are restored by optimal interpolation. Simulation results show that this coding scheme outperforms the conventional VFR vocoder using linear interpolation. By incorporating contour quantization of gain and pitch information, a low variable-rate LPC vocoder is realized. An informal listening test shows that very high intelligible reconstructed speech was obtained at an average data rate of 300 bps.**

I. Introduction

IN MOST speech communication applications, it is desirable to compress the transmission data rate as much as possible while still retaining a reasonable level of speech quality. One interesting approach to achieve very low data rate is by variable-frame-rate (VFR) coding of spectral parameters [1]–[5]. The basic idea is to segment speech signal into phoneme-like segments and then to efficiently encode each segment using techniques such as matrix quantization [1], [2] or decimation/interpolation [3], [4]. In the approach using matrix quantization, a computation intensive nonlinear mapping between variable length vector sequences and fixed length vector sequences is usually required for obtaining high-quality reconstructed speech. Determining segment boundaries is also not a trivial problem. On the contrary, the approach using decimation/interpolation is much simpler. In the encoder, parameters of representative frames are selected and transmitted only when the spectral property has changed significantly. Parameters of untransmitted frames are then restored in the decoder by interpolation. Usually, linear interpolation is used for simplicity.

In this paper, a novel VFR coding of spectral information based on decimation/interpolation is proposed. It differs from a conventional linear interpolation based coding scheme in the method of selecting representative frames for transmission as well as the method of restoring untransmitted frames. In this coding scheme, an optimal interpolation algorithm is employed in the decoder to generate the best estimates of all untransmitted frames from the received information. In the encoder, best representative frames are determined to achieve the goal of maximum coding efficiency subject to the constraint of obtaining best interpolated parameters of untransmitted frames in the decoder. Two advantages of this coding method can be found. First, a better reconstructed speech signal can be obtained because the best representative frames to transmit are selected in the encoder using a novel algorithm (discussed in Section II) and the best estimates of parameters of untransmitted frames are generated in the decoder by an optimal interpolation algorithm. Second, higher coding efficiency can be achieved due to the use of the optimization procedures in the algorithm of finding representative frames in the encoder.

This paper is organized as follows. Section II describes the proposed coding scheme. In Section III, computer simulations were done to examine its performance. Conclusions are given in the last section

Paper approved by S. Dimolitas, the Editor for Speech Processing of the IEEE Communications Society. Manuscript received May 15, 1991; revised May 18, 1992 and September 21, 1992. This work was supported in part by the National Science Council, Republic of China, under Contract NSC81-0404-E009-021. This paper was presented in part at the 1991 International Symposium on Communications, Taninan, Taiwan, Republic of China, December 9–13, 1991.

The authors are with the Department of Electrical Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan 300, Republic of China.

IEEE Log Number 9401046.

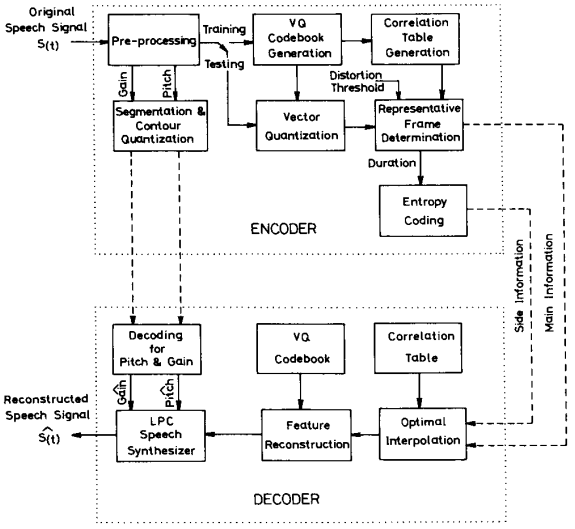

Fig. 1. The block diagram of the proposed VFR LPC vocoder.

II. The Proposed VFR Coding Scheme

The block diagram of the proposed VFR coding scheme is shown in Fig. 1. In the encoder, input speech signal is first preprocessed to extract parameters of linear prediction. Preprocessing consists of LP-filtering at 3.5 kHz, sampling at 8 kHz, A/D conversion into 12 bit digital form, and feature extraction for every 20 ms frame. Three kinds of features, namely 13 gain-normalized autocorrelation coefficients, a pitch period, and a gain factor, are extracted. These features are then separately encoded. Here, we consider the encoding of the autocorrelation coefficients only.

After feature extraction, autocorrelation coefficients of each frame are vector-quantized and labelled into a symbol sequence. Then, the symbol sequence is nonuniformly down sampled to extract some representative frames for transmission. Coding efficiency is achieved by suppressing those frames which are unnecessary to transmit. In the decoder, reproduction feature vectors of transmitted frames are first reconstructed. Then, those of untransmitted frames are restored from the transmitted frames and additional timing information by an optimal interpolation algorithm. The reconstructed speech signal is finally obtained by an LPC synthesis. In the following, the optimal interpolation algorithm and the algorithm for finding the best representative frames for transmission are described in detail.

The task of the optimal interpolation algorithm is to generate the best estimates of the reproduction feature vectors of an untransmitted frame sequence given the preceding and the following feature vectors, obtained from the main transmitted information and the length known from the timing information. By defining an objective function to be maximized or minimized, this becomes an optimization problem. Two types of objective functions are considered in this paper. They are the cumulative distance of successive frame pairs and the joint probability of interpolated frames, conditioned on the feature vectors of the preceding and following frames. The criterion is hence to minimize the cumulative distance or to maximize the joint probability of interpolated frames conditioned on the features vectors of the preceding and following frames. In this paper, the problem is further simplified by finding interpolated feature vectors from the subspace of vector quantization (i.e., codebook). This is motivated by the fact that vector quantization only results in little degradation in speech quality if the codebook size is large enough (say 512 and above). So the criterion becomes

$$\hat{X}_1\hat{X}_2\cdots\hat{X}_n = \arg \min_{C_{j_1}\cdots C_{j_n}} \left(d(C(X_0), C_{j_1}) + \sum_{i=1}^{n-1} d(C_{j_i}, C_{j_{i+1}}) + d(C_{j_n}, C(X_{n+1})) \right) \quad (1)$$

for minimizing the cumulative distance and becomes

$$\hat{X}_1\hat{X}_2\cdots\hat{X}_n = \arg \max_{C_{j_1}\cdots C_{j_n}} P(C_{j_1}\cdots C_{j_n} | C(X_0)C(X_{n+1})) \quad (2)$$

for maximizing the jointly conditional probability. In (1), the sequence of parametric vectors X_1 to X_n are represented by the codewords from the codebook $CB(j)$ and j may represent 512 codewords, such that the cumulation of the likelihood ratio distance $d(C_{j_i}, C_{j_{i+1}})$ is the minimum. Similarly, in (2), X_1 to X_n are also represented by the codewords from $CB(j)$ such that the joint probability of $C_{j_1}\cdots C_{j_n}$ conditioned on $C(X_0)$ and $C(X_{n+1})$ is maximum. Here $C(X)$ is the encoded codeword of the feature vector X. In the case of minimum cumulative distance, the interpolation algorithm can be simply realized by dynamic programming (DP) technique. A distance table of codework pairs can be established in advance and taken as a look-up table in order to make the DP procedures computationally efficient.

For the case of maximum jointly conditional probability, (2) must be further simplified to make it mathematically tractable. A first-order Markov chain is employed in this paper to model the sequence of encoded feature vectors. It uses a transition probability to describe the relationship between encoded feature vectors of a successive frame pair. All transition probabilities form a matrix to be used as the look-up table in the DP-based interpolation algorithm. These transition probabilities can be empirically estimated using a large set of training utterances.

The algorithm for finding the best representative frames in the encoder is stated as follows. Given the preceding frame which is a representative frame to transmit, the task is to find the next representative frame with the length of intermediate suppressed frames being maximized subject to the constraint that all frames suppressed in the encoder must be optimally interpolated in the decoder with reconstruction distortions less than a predetermined threshold. This is a constrained optimization problem with the goal of achieving maximum coding efficiency. The constraint is set for guaranteeing the quality of the reconstructed speech. A DP-based algorithm is proposed for accomplishing the task. Procedures of the algorithm for the case of minimum cumulative distance are listed as follows.

Algorithm of Finding the Next Representative Frame: 1) *Initialization:* Let X_0 be the feature vector of the preceding representative frame, Dth be the predetermined distortion threshold, M be the codebook size. Set $i = 1$.

2) *Processing the First Frame:* For $j = 1, \ldots, M$, calculate the likelihood distance $d_j = d(C_j, X_1)$ and set $\phi_1(j) = d(C(X_0), C_j)$ as the initial cumulative distance. If $d_j < D$th, set $\zeta_1(j) = 1$ to indicate C_j is an acceptable codeword; otherwise set $\zeta_1(j) = 0$. Set $i = i + 1$.

3) *Processing the ith Frame and Identifying all Survival Paths:* For $j = 1, \ldots, M$, calculate $d_j = d(C_j, X_i)$ and the cumulative distance of the best path reaching codeword C_j at frame i by $\phi_i(j) = \min_l [d(C_l, C_j + \phi_{i-1}(l)]$, and find the backtracking codeword index at frame $i-1$ by $l^*(j) = \arg\min_l [d(C_l, C_j) + \phi_{i-+1}(l)]$ where l is the codeword index at frame $i-1$. If $d_j < D$th and $\zeta_{i-1}(l^*(j)) = 1$, set $\zeta_i(j) = 1$ to indicate the path reaching C_j is survival; otherwise, set $\zeta_i(j) = 0$.

4) *Termination Test:* If all paths at frame i are dead, i.e., $\zeta_i(j) = 0$ for all j, set the next frame to be transmitted as $i^* = i - 1$ and find the best survival path at frame $i-1$ as $k = \arg\min_{l,\, \zeta_{i-1}(l)=1} \phi_{i-1}(l)$ and goto Step 5; otherwise let $i = i + 1$ and goto Step 3.

5) *Termination:* Regard C_k as the codeword of the next representative frame to be transmitted and send k as main information and i^* as side information.

For the case of maximum jointly conditional probability, substitute the likelihood ratio distance, $d(\cdot, \cdot)$, in the calculation of the objective function, $\phi_i(j)$, with the transition probability, $P(\cdot|\cdot)$, and change all the operators of min to max. It is noted that the duration of suppressed frame sequence is efficiently entropy-encoded.

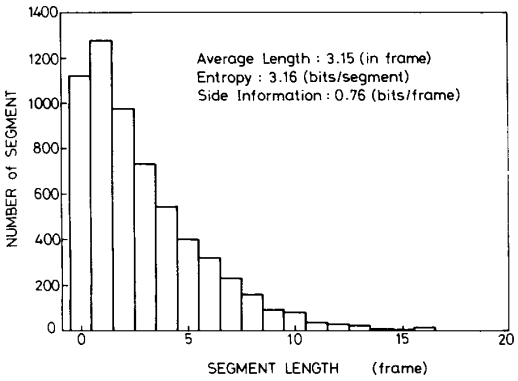

Fig. 2. The histogram of the length of untransitted segment.

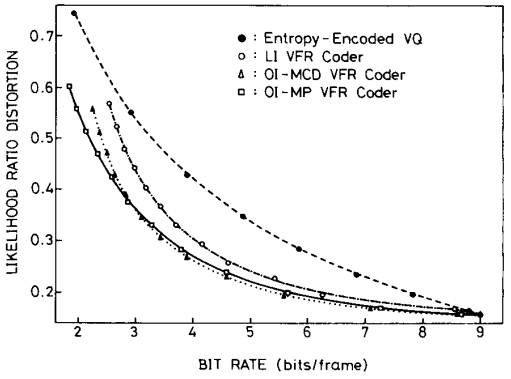

Fig. 3. Rate versus distortion curves of encoding spectral information using the proposed coder with minimum cumulative distance (OI-MCD VFR coder) and maximum conditional probability (OI-MP VFR coder) and the conventional linear interpolated based coder (LI VFR coder).

III. Simulations

The efficiency of this VFR coding scheme was examined by simulations. A database composed of 600 sentential utterances in Mandarin Chinese was used in both the training and testing phases. It was generated by 30 male speakers. Each speaker uttered 20 sentences randomly selected from an ensemble of 112 phonetically balanced sentences [9]. An endpoint detection [6] was firstly employed to discard nonspeech signals in the beginning and the ending parts of all utterances. Then, feature vectors in the form of gain-normalized autocorrelation coefficients were extracted. There are total 89224 feature vectors. These vectors were then used to generate codebooks for vector quantization by the LBG training algorithm. Here, a likelihood ratio (LR) distance is used in both the training and the encoding phases of VQ. The distance between a frame X of input speech and a codeword C_i is [8]

$$d_{\mathrm{LR}}(C_i, X) = \frac{r_x(0)}{\alpha_p} a_i(0) + 2\sum_{j=1}^{p} \frac{r_x(j)}{\alpha_p} a_i(j) - 1 \quad (3)$$

where $r_x(j)$ and $a_i(j)$ denote the autocorrelation sequences of X and the coefficients of the inverse filter of C_i, respectively; α_p is the residual energy of X and p is the order of LPC. After generating the codebook, all utterances in the database were vector-quantized for empirically estimating the transition probability $p(C_{j_{i+1}}|C_{j_i})$.

Now, the efficiency of the proposed VFR coding scheme is examined. Fig. 2 shows the histogram of the length of an untransmitted segment for the case of $D\mathrm{th} = 0.4$ with the length being constrained to be less than 20 frames. From this figure, we find that the average length of an untransmitted segment is 3.15 frames (i.e., 60.3 ms). By entropy encoding [10], about 0.76 b/frame is needed to encode this durational information. Performance comparisons of the proposed coding scheme using optimum interpolation with two different criteria and the conventional method using linear interpolation are showed in Fig. 3. Here, the codebook size is 512 and the distortion threshold $D\mathrm{th}$ is varied from 0.1 to 1.2. As it can be seen from this figure, the proposed coding scheme outperforms the conventional method by about 0.6 b/frame (i.e., 30 b/s).

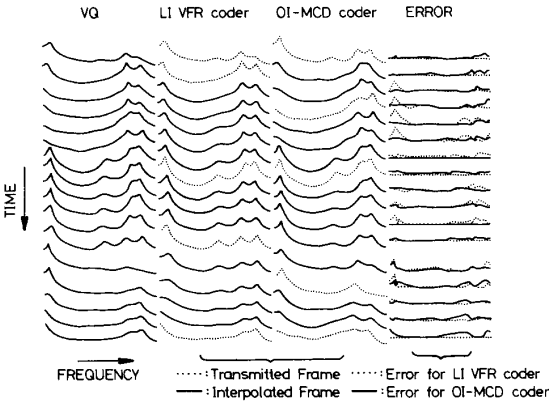

Fig. 4. The reconstructed spectra of a typical speech segment encoded by the proposed OI-MCD VFR and the conventional LI VFR coders.

Fig. 4 shows the reconstructed spectra of a typical speech encoded by these two coding schemes at a bit rate of 4 b/frame. We found that optimal interpolation usually results in smaller spectral error.

Finally, a complete LPC vocoder was implemented. It was necessary to additionally consider the encoding of pitch and gain information. An efficient contour quantization based coding scheme was proposed recently for encoding pitch information [7]. It first segments the pitch contour of an utterance into syllable-like segments. Then, each segment is encoded by vector-quantizing the first four coefficients of its orthogonal polynomial transform. A similar approach is used here. First, pitch and energy contours are simultaneously segmented by finding the local maximum of the norm of the regression coefficient vector. Then, they are separately vector quantized using 7 bit and 6 bit codebooks, respectively. The resultant average data rate is 2.19 b/frame. Durational information is also entropy-encoded [10] with an average data of 0.59 bits/frame. Combining the optimal interpolation based coding for spectral parameters and the contour quantization of pitch and gain information, a variable-bit-rate LPC vocoder is

realized. An average data rate of 300 bits/sec. with average distortion of 0.3 was obtained for $D\text{th} = 0.6$. As Dth decreases, the average data rate will increase and the average distortion will decrease accordingly. For $D\text{th} = 0.3$, an average data rate of 400 b/s with average distortion of 0.2 was achieved. An informal listening test confirmed that the quality of the reconstructed speech is reasonably good with very high intelligibility at an average data rate of 300 b/s. The quality is very similar to that of a linear-interpolated LPC vocoder with the same average distortion of 0.3 and is almost comparable to the standard LPC vocoder with major difference occurred at some consonants in which a slightly perceptual degradation can be heard by careful listeners.

IV. Conclusion

An optimal interpolation based VFR LPC vocoder has been presented for very low bit rate speech coding. It outperforms the conventional linear interpolation based method while achieving higher coding efficiency. A reasonably good quality of reconstructed speech with very high intelligibility was obtained at an average rate of 300 b/s.

References

[1] D. Y. Wong, B. H. Juang, and D. Y. Cheng, "Very low data rate speech compression with LPC vector and matrix quantization," in *ICASSP'83*, pp. 65–68.

[2] Y. Shriraki and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. Acoust. Speech. Signal Processing*, vol. 36, pp. 1437–1444, Sept. 1988.

[3] P. E. Papamichalis and T. P. Barnwell, "Variable rate speech communication by coding subsets of the PARCOR coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 706–712, June 1983.

[4] S. R. Lee, H. S. Lee, and C. K. Un, "A low rate speech coding algorithm with variable transmission frame length," in *Proc. Int. Conf. Spoken Language Processing 1990*, pp. 15.17.1–15.17.4.

[5] J. Picone and G. R. Doddington, "A phonetic VOCODER," in *ICASSP'89*, pp. 580–583.

[6] L. R. Rabiner and B. W. Schafer, *Digital processing of speech signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978, pp. 130–135.

[7] S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin Speech," *IEEE Trans. Commun.*, vol. 38, pp. 1317–1320, Sept. 1990.

[8] B.-H. Juang, D. Y. Wong, and A. H. Gray, "Distortion performance of vector quantization for LPC voice coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 294–304, Apr. 1982.

[9] S. M. Yu and C. S. Lin, "The construction of phonetically balanced Chinese sentences," Taiwan, ROC, Telecommun. Lab. Tech. Rep. 1989.

[10] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 668–674, Nov. 1978.