COMPLEXITY REDUCED SHAPE VQ OF SPECTRAL ENVELOPE WITH PERCEPTION CONSIDERATION

Mu-Liang Wang and Jar-Ferr Yang

Department of Electrical Engineering, National Cheng Kung University
1 University Road, Tainan, Taiwan 70101
jfyang@ee.ncku.edu.tw

ABSTRACT[+]

The parametric coders provide a good communication quality at low bit rate. Efficient encoding of variable dimension harmonic spectral envelope is an essential task in parametric speech coders. In this paper, we propose an efficient vector quantization (VQ) scheme with perception consideration to improve the performance of parametric speech coders. With the benefit of reduction in dimension, the computational complexity of spectral envelope VQ (SEVQ) has been reduced while the speech quality is retained. Experimental results show that the proposed perceptual SEVQ method significantly reduces the computational complexity of SEVQ by a factor of 9.

1. INTRODUCTION

The parametric coders such as the sinusoidal transform coder (STC) [1], the multi-band excitation (MBE) coder [2] and the mixed excitation linear prediction (MELP) vocoder [3] provide good communication quality below 4.8kbps. The harmonic vector excitation coder (HVXC) [4] suggested in the MPEG-4 achieves communication quality with bit rate at 2.0kbps[5]. It will be suitable for applications in speech communication, speech data storage, etc.

The parametric speech coder extracts the harmonic spectral envelope vector (SEV) from the power spectrum and then encodes the SEV with shape vector quantization (SEVQ). The subjective quality of parametric speech coders highly depends on the quantization efficiency of SEVQ. The dimension of SEV depends on the period of voiced speech. There is much research proposed for efficient vector quantization of variable-dimension SEV [6-9]. The SEVQ can be performed with/without the dimension conversion. The converted-dimension SEVQ scheme [9] has been adopted in the MPEG-4 HVXC coder, which successfully achieves the communication quality at the rate of 2kbps.

In this paper, we propose a nonlinear split band SEVQ (NSBSEVQ) scheme to fit the characteristics of human auditory sensation. The proposed SEVQ scheme can achieve better representation of spectral envelope for human perception. With benefit of reduced dimension, the computational complexity of SEVQ can also be reduced.

This paper is organized as follows. In section 2, the traditional SEVQ scheme suggested by the HVXC coder is briefly reviewed. The coding procedure of the proposed scheme is addressed in Section 3. In Section 4, experimental results for the traditional and proposed SEVQ scheme are then presented. Finally, the conclusions are addressed in Section 5.

2. THE PARAMETRIC SPEECH CODER

The MPEG-4 parametric speech coder (HVXC) encodes the harmonics of LPC residual signal for voiced segments. The spectrum of the LPC residual signal, which is weighted by Hamming window, is computed by DFT. The spectral magnitude, $r(p)$, at the p^{th} pitch harmonic is estimated. The SEV is defined as a set of spectral magnitudes, $\underline{\boldsymbol{r}} = \{r(p), p = 1,2,...,P\}$, estimated at each harmonic. The dimension of SEV, *P,* is the integer part of one half of pitch lag Γ, i.e., $P = \lfloor \Gamma/2 \rfloor$. The SEVQ schemes suggested in [9] convert the SEV into a fixed *M*-dimension shape vector $\underline{\boldsymbol{r}}^{(F)}$, $\underline{\boldsymbol{r}}^{(F)} = \{r^{(F)}(m), m = 1,2,...,M\}$. Fig. 1 shows the structure of the SEVQ scheme suggested in MPEG-4 HVXC base-layer coder. After the dimension conversion, the fixed dimension SE shape vector is then vector quantized with two-stage VQ. With the output of two 4-bits codebooks, $\underline{\boldsymbol{A}}^{(F)} = \{\underline{\boldsymbol{a}}_{j_0}^{(F)}\}$ and $\underline{\boldsymbol{B}}^{(F)} = \{\underline{\boldsymbol{b}}_{j_1}^{(F)}\}$, the

[+] This research was supported by National Science Council under Contract #NSC-91-22193-E006-017, Taiwan, ROC.

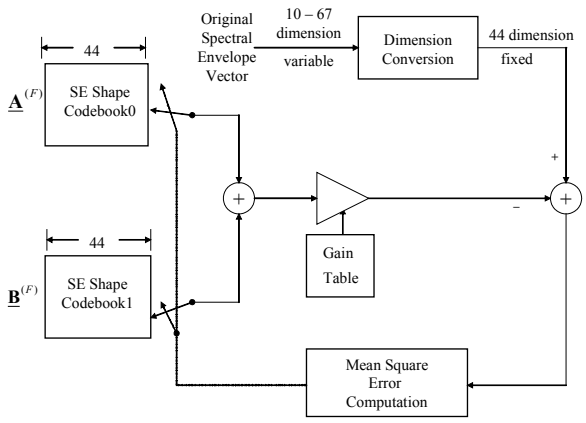

Figure 1. The structure of HVXC base layer SEVQ scheme (after [4])

possible codewords, $\underline{\boldsymbol{c}}_j^{(F)} = \{c_j^{(F)}(m), m = 1,2,...M\}$, for the SEVQ are obtained as

$$\underline{\boldsymbol{c}}_j^{(F)} = \underline{\boldsymbol{a}}_{j_0}^{(F)} + \underline{\boldsymbol{b}}_{j_1}^{(F)}, j = 0,1,2,\ldots,L-1, \quad (1)$$

where j is the combined 8-bits index as $j = (16 * j_0 + j_1)$.

The optimal codeword of SEVQ is searched with weighted minimum mean square error criterion. Let $\{\alpha_i, i = 0,1,...,10\}$ denote the LPC coefficients of the current frame signal, the perceptual weighting filter is given by

$$W(z) = \frac{\sum_{i=0}^{10} \alpha_i \gamma_1{}^i z^{-i}}{\sum_{i=0}^{10} \alpha_i \gamma_2{}^i z^{-i}} \cdot \frac{1}{\sum_{i=0}^{10} \alpha_i z^{-i}}, \quad (2)$$

Let $\underline{\boldsymbol{w}}^{(F)} = \{w^{(F)}(m), m = 1,2,...,M\}$ denotes the fixed dimension weighting vector obtained from the spectrum of $W(z)$. The optimal index j^* of the codeword will be the one which maximize the term described as

$$j^* = \arg\max_j \left[\frac{\left\| \left(\underline{\boldsymbol{r}}^{(W)}\right)^T \cdot \underline{\boldsymbol{c}}_j^{(W)} \right\|^2}{\left\| \underline{\boldsymbol{c}}_j^{(W)} \right\|^2}, j = 0,1,\ldots,L-1 \right], \quad (3)$$

where the elements of vector $\underline{\boldsymbol{r}}^{(W)}$ and $\underline{\boldsymbol{c}}_j^{(W)}$ are defined as

$$r^{(W)}(m) = r^{(F)}(m) w^{(F)}(m) \quad (4)$$

and

$$c_j^{(W)}(m) = c_j^{(F)}(m) w^{(F)}(m), \quad (5)$$

respectively. The superscript "T" denotes the transpose operation. After the optimal index j^* is found, the shape gain g^* for the optimal codeword is given by

$$g^* = \frac{\left(\underline{\boldsymbol{r}}^{(W)}\right)^T \cdot \underline{\boldsymbol{c}}_{j^*}^{(W)}}{\left\| \underline{\boldsymbol{c}}_{j^*}^{(W)} \right\|^2} \quad (6)$$

Table 1. Split band coefficient present in Hertz mapping table

Bark Index i	Included Discrete Frequency Indices	Corresponding Center Frequency (Hz)	Corresponding Bandwidth (Hz)
1	1-1	91	91
2	2-2	182	91
3	3-3	273	91
4	4-4	364	91
5	5-6	500	182
6	7-8	681	182
7	9-10	864	182
8	11-12	1055	182
9	13-15	1273	273
10	16-18	1546	273
11	19-21	1818	273
12	22-25	2136	364
13	26-30	2546	455
14	31-36	3046	546
15	37-44	3636	727

The optimal gain g^* is then scalar quantized with Q-bits quantizer. The total of computation in (4) and (5) acquires $(L+1)M$ multiplications. The computation of j^* totally acquires $(2M+1)L$ multiplications, L divisions and $2ML$ additions.

3. THE PROPOSED NSBSEVQ SCHEME

It is well known that the human perception is nonlinear to the frequency resolution. In the human auditory system, the sound intensities within one critical band characterized by critical band filters are added to form the sum of intensity. The width of one critical band has been defined as one bark [10]. The nonlinear frequency transformation from natural frequency, f, to the bark scale, b, can be approximated as [11]:

$$b = 6\sinh^{-1}\left(\frac{f}{600}\right), \quad 0 \le f \le 4k(\text{Hz}), \quad (7)$$

In this section, a nonlinear split-band SEVQ (NSBSEVQ) scheme was proposed to quantize the SEV with perception consideration. The SEV is nonlinear spitted into N sub-bands. For easily performing nonlinear band splitting of variable dimension SEV, the fixed M-dimension SEV is used. This arrangement will be helpful to make the proposed scheme to be compatible with the formal decoder. The proposed split band parameters as shown in Table 1 were properly selected to fit the frequency-bark transfer curve in (7). The fitting curves depicted in Fig. 2 compare the difference of two different mapping methods.

According to the lower and upper edges of the band shown in Table 1, the N-dimensional split-band SEV can be obtained by

$$r_B(n) = 10 * \log_{10}\left[\frac{1}{q_n} \sum_{m \in band\ n} [r^{(F)}(m)]^2 \right], \quad n = 1,2,...,N, \quad (8)$$

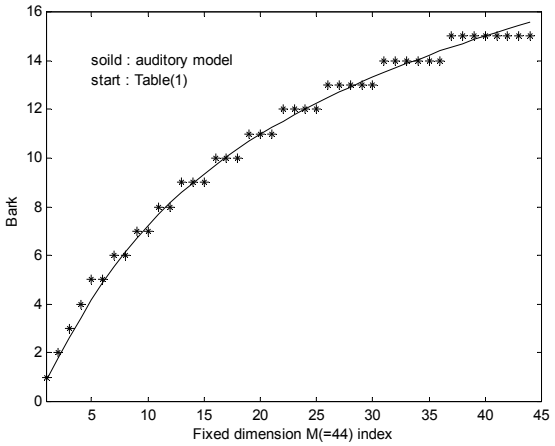

Figure 2. Curves of two different mapping functions

where q_n denotes the number of indices belong to band n. Utilizing the set of $\{r_B(n)\}$, the zero mean version of the harmonic SE shape vector $\underline{\tilde{r}}_B$ is computed as the target vector of codebook search. By the similar way, the codewords, $\{\underline{c}_j^{(F)}, j=0,1,...,L-1\}$ are also converted into *N*-dimensional split-band codeword $\{\underline{\tilde{c}}_j^{(B)}, j=0,1,...,L-1\}$. Then, the zero-mean split-band codebook $\boldsymbol{C}^{(B)}=\{\underline{\tilde{c}}_j^{(B)}, j=0,1,...,L-1\}$ and the mean of $\{\underline{c}_j^{(B)}\}$ can be pre-computed and stored before encoding processing. Hence we only need to compute $\underline{\tilde{r}}_B$ and the mean of $\underline{r}_B$ for each segment with *N* additions, *N* subtractions and 1 division. The optimal codeword will be the one that has the most similar shape to $\underline{\tilde{r}}_B$ with the MSE criterion as

$$j^* = \arg\min_j \left[\left\| \underline{\tilde{r}} - \underline{\tilde{c}}_j^{(B)} \right\|^2, \quad j=0,1,...,L-1 \right] \quad (9)$$

It acquires *N* additions, *N* subtractions and *N* multiplications for each codeword. The block diagram of the NSBSEVQ scheme is shown in Fig. 3. Once the optimal index of shape codebook j^* is found, the SE shape gain g_{dBj^*} can be found:

$$g_{dBj^*} = \frac{1}{N}\sum_{n=1}^{N}\left[r_B(n) - c_{j^*}^{(B)}(n) \right] \quad (10)$$

The optimal SE shape gain g_{dBj^*} can be quantized with the *Q*-bits shape gain table G_0 suggested in HVXC coder. The gain index i_g is determined by

$$i_g = \arg\min_i \left\{ \left[g_{dBj^*} - 2*\log_{10}[G_0(i)] \right]^2, i=0\sim 2^Q-1 \right\} \quad (11)$$

Finally, by utilizing the quantized optimal gain $\hat{g}^*=G_0(i_g^*)$ and optimal index of shape codebook, j^*, the quantized harmonic residual SEV can be reconstructed. So,

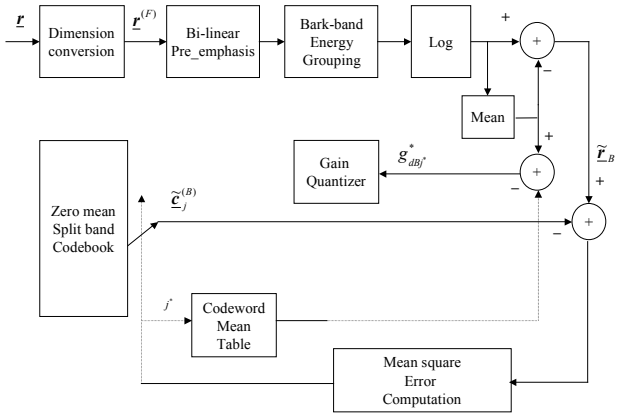

Figure 3. The structure of proposed NSBSEVQ scheme

the NSBSEVQ scheme is compatible with the formal decoding process suggested in MPEG-4 HVXC speech coder.

4. EXPERIMENTAL RESULTS

To evaluate the performance, the traditional spectral distances (SD) of the existing and the proposed schemes are measured. For the i^{th} segment, the spectral distortion, $SD^{(i)}$ is defined as:

$$SD^{(i)} = \left[\frac{1}{P_i}\sum_{p=1}^{P_i} \left\{ 20\log_{10}[r(p)] - 20\log_{10}[\hat{r}(p)] \right\}^2 \right]^{1/2}, \quad (12)$$

where P_i denotes the dimension of SEV for the i^{th} segment. The final SD is the average over N_s voiced frames as

$$SD = \frac{1}{N_s}\sum_{i=1}^{N_s}\left[SD^{(i)} \right]. \quad (13)$$

In simulations, we choose 4 sentences speech uttered in Chinese by 2 males and 2 females for the testing. Table 2 shows the experimental results of SD measures for different SEVQ methods. It exhibits that the proposed NSBSEVQ scheme achieves the smaller SD than the original SEVQ method suggested in the MPEG-4 HVXC coder. By informal listening test, the speech quality of the proposed NSBSEVQ scheme is distinguishably better than the original SEVQ method in MPEG-4 HVXC coder.

The complexity analyses of the SEVQ schemes are enlisted in Table 3, where the "A/S" and "M/D" denote the addition/subtraction and multiplication/division operations, respectively. Comparing to the original SEVQ scheme in the HVXC coder, we reduce the number of multiplications by a factor of 9 and save the number of additions by a factor of 3.

Table 2. SD measurement for different SEVQ methods (unit: dB)

Search Methods	Male #1	Male #2	Female #1	Female #2
HVXC	3.87	3.46	4.11	3.20
NSBSEVQ	3.51	3.25	3.74	2.90

Table 3. Complexity of codebook search of SEVQ schemes

Search Methods	Addition/ Subtraction	Multiplication/ Division	Examples*	
			A/S	M/D
HVXC	$2ML$	$(3M+2)L+M$	22528	34348
NSBSEVQ	$2NL+2N+1$	$NL+2$	7711	3842

* for $M = 44$, $N = 15$, $L = 256$

5. CONCLUSIONS

We propose an effective SEVQ search scheme to reduce the computational complexity of SE shape vector. Simulation results have shown that the SD of the proposed methods is smaller than the traditional scheme. Furthermore, the computational complexity of codebook searching is significantly reduced. It will greatly help general HVXC coders for real-time applications while the speech quality is retained. Although we evaluate the performance of the proposed schemes in the MPEG-4 HVXC coders, without loss of generality, the proposed schemes can be also suitable for the speech coders, which perform the vector quantization of the harmonic SEV of the speech or LPC residual signal.

REFERENCES

[1] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Trans. on ASSP, vol. 34, no.4, pp.744-754, Aug. 1986.

[2] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder", IEEE Trans. on Acoustic, Speech and Signal Processing, vol. ASSP-36, no. 8, pp.1223-1235, Aug. 1988.

[3] A. V. McCree and T. P. Barnwell III, "A Mixed excitation LPC vocoder model for low bit rate speech coding", IEEE Trans. on ASSP, vol. 3, no. 4, pp. 242-250, July 1995.

[4] ISO/JTC 1/SC29/WG11, Information technology - coding of audio visual objects, N2503-2C, Nov, 1998.

[5] ISO/IEC JTC1/SC29/WG11, Coding of moving pictures and audio, N2424, October 1998.

[6] B. Lupini and V. Cuperman, "Vector Quantization of Harmonic Magnitudes for Low Rate Speech Coders", Proc. IEEE Globecom Conf., vol 2, pp.858-862, 1994.

[7] A. Das, A. V. Rao and A. Gersho, "Variable-Dimension Vector Quantization", IEEE Signal Processing Letters, vol. 3, no. 7, pp.200-202, July 1996.

[8] E. Shlomot, V. Cuperman, and A. Gersho, "Hybrid Coding: Combined Harmonic and Waveform Coding of Speech at 4kb/s", IEEE Trans. on Speech and Audio Processing, vol. 9, no. 6, pp. 632-646, Sep. 2001.

[9] M. Nishiguchi, J. Matsumoto and, R. Wakatsuki and S. Ono, "Vector Quantized MBE with Simplied V/UV Division at 3.0kb/s", Proc. IEEE ICASSP, pp151-154, 1993.

[10] E. Z. Wicker, and H. Fastl, "Psychoacoustic-Facts and Models", Hirzel-verlag, Berlin, Germany, 1990.

[11] Fourcin, "Speech processing by man and machine – Group report", Recognition of Complex Acoustic Signals, ed. T. Bulllock, 1977.