VARIABLE BIT RATE MBELP SPEECH CODING VIA V/UV DISTRIBUTION DEPENDENT SPECTRAL QUANTIZATION

Eric W.M. Yu and Cheung-Fat Chan

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong.
Email: eecfchan@cityu.edu.hk

ABSTRACT

A variable bit rate multiband excited linear predictive speech coder is proposed in this paper. Speech signal is compressed in different bit rates ranging from 0.88 kbps to 2.6 kbps according to the mode of operation and the optimum V/UV transition frequency. An average bit rate of 1.24 kbps is achieved. The proposed speech coder improves the speech quality by splitting the non-stationary speech segments for analysis. The V/UV distribution of a short-time speech spectrum is represented efficiently by using a closed-loop minimised V/UV transition frequency. Depending on the V/UV transition frequency, the spectrum envelope is quantized in variable bit rate through embedded differential predictive scalar and vector quantizations of the LSP parameters. The proposed spectral quantization scheme results in a spectral distortion comparable to a fixed 24-bit 2-dimensional differential scalar quantization scheme.

1. INTRODUCTION

Variable bit rate speech coders operating below 2 kbps were proposed recently in [1] and [2]. A common approach in variable rate speech coding is to transmit each speech segment with different bit rate according to the speech characteristics. A voice activity detector (VAD) is usually incorporated to distinguish between the active and inactive speech. Research effect has been concentrated on the approaches for coding of active speech. As proposed in [1], phonetic speech classification and variable dimension vector quantization (VQ) are used. A different approach was proposed in [2] where the variable bit allocations for spectral parameters and excitation vectors are derived from the spectral flatness measures.

In this paper, a variable bit rate speech coder is proposed based on the multiband excited linear predictive (MBELP) speech coder in [3]. In the proposed speech coder, speech signal is classified into silence, transient or steady state for each segment. For the efficient coding of V/UV information in MBE coders, a technique was developed where a voiced/unvoiced (V/UV) transition frequency was determined through closed-loop analysis of speech in the frequency domain. The V/UV transition frequency is then applied for splitting of a set of line spectrum pair (LSP) parameters. The LSP prediction residuals associated with the lower bands are quantized through scalar quantization (SQ) while the set of prediction residuals associated with the higher bands are quantized through variable dimension VQ. Thus, the number of scalar quantizers and the dimension of vector quantizer vary adaptively according to the V/UV transition frequency. Different codebooks and codebook sizes are adopted to vector quantizers of different dimensions. The number of bits required for spectral quantization is adaptive.

2. DETECTION OF ABRUPT TRANSITION

For the classification of transient, a technique is developed to detect the abrupt transition in a speech segment. Our concentration is placed on the abrupt increase in signal energy of the speech segment. In the windowing of speech signal $s(n)$, we may denote the v^{th} segment as $\{s(Nv+m)\}_{m=-L/2}^{L/2-1}$ where N is the window shift and L is the window length. For each speech segment that precedes a low energy segment and has an abrupt increase in the signal energy, the occurrence of the abrupt increase is detected and the time instant of the occurrence is determined. For L=256 samples while the sampling frequency is 8 kHz, the speech segment $\{s(Nv+m)\}_{m=-L/2}^{L/2-1}$ is divided into overlapped small segments by using a rectangular window of 96 samples in length. The rectangular window is sliding through the time with an interval of 20 samples. In the marked intervals, an energy ratio is computed. The abrupt transition in the speech segment can be detected by observing the energy ratio $\sigma(v,n_i)$ with respect to the time sample n_i, where

$$n_i = 20i + 48 \tag{1}$$

for $i = 0,1,\cdots,\lfloor (L-96)/20 \rfloor$. The symbol $\lfloor x \rfloor$ denotes the lower closest integer value of x. The energy ratio is defined as

$$\sigma(v,n_i) = \frac{\sum_{j=n_i}^{n_i+47} s^2(Nv - L/2 + j)}{\sum_{j=n_i-48}^{n_i-1} s^2(Nv - L/2 + j)} \tag{2}$$

for $i = 0,1,\cdots,\lfloor (L-96)/20 \rfloor$. The time instant t_v of the occurrence of abrupt transition is estimated as

$$t_v = \arg\max_{\forall n_i}\{\sigma(v,n_i)\} \tag{3}$$

provided that the following 2 inequalities are also satisfied:

$$\sigma(v,t_v) > \zeta_1 \tag{4}$$

and $$\frac{1}{L-t_v}\sum_{j=t_v}^{L-1} s^2(Nv-L/2+j) > \zeta_2 \tag{5}$$

The constants ζ_1 and ζ_2 are the pre-defined thresholds. The constraint in (4) ensures the existence of abrupt increase in signal energy while the constraint in (5) is used to reduce false detection owing to noise. In practice, for $\zeta_1 = 6$ and $\zeta_2 = 0.03$ reliable detection of the abrupt increase in speech signal energy and accurate estimation of the transition time index t_v can be obtained.

3. DETERMINATION OF V/UV TRANSITION FREQUENCY

In MBE speech coding, the V/UV decisions are made after the optimum pitch period τ_o is estimated from the spectrum $|S(k)|$, $0 \le k < \pi$, of the speech signal. Conventionally, the V/UV decision for each harmonic band is made by comparing the normalised error in each harmonic band to a threshold [4]. The normalised error for band *m* is defined as

$$\xi_m = \frac{\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)} \left[|S(k)| - A_m(\tau_o)|P(\tau_o,k)|\right]^2}{(1-W\tau_o)\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)}|S(k)|^2} \tag{6}$$

$$m = 1,2,\cdots,M(\tau_o)$$

where W is a weighting factor for unbiasing the pitch-dependent error and $M(\tau_o)$ is the total number of harmonic bands. The band magnitude $A_m(\tau_o)$ is computed by

$$A_m(\tau_o) = \frac{\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)}|S(k)||P(\tau_o,k)|}{\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)}|P(\tau_o,k)|^2} \tag{7}$$

$$m = 1,2,\cdots,M(\tau_o)$$

where $a_m(\tau_o)$ and $b_m(\tau_o)$ are the lower and upper band limits, respectively. The pitch-dependent periodic spectrum is denoted by $|P(\tau_o,k)|$. If ξ_m is below a pre-defined threshold, band *m* is marked as voiced. Otherwise band *m* is assumed to contain noise-like energy and is then marked as unvoiced. From the V/UV decisions on all the harmonic bands, a binary sequence is constructed. The binary sequence is then transmitted to the decoder as the V/UV mixture function for speech synthesis.

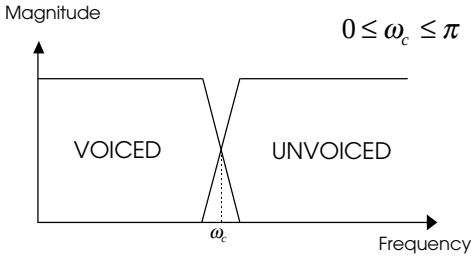

Figure 1 Simplified V/UV distribution.

For bit rate reduction in V/UV information coding, a simplified V/UV mixture function which is based on the simplified V/UV distribution as shown in Figure 1, was proposed. A procedure is needed to determine the V/UV transition frequency ω_c. Since the error measure as defined in (6) is based on using a voiced synthetic spectrum, we need to define a new error measure for unvoiced synthetic spectrum in order to determine the V/UV transition using an analysis-by-synthesis procedure. We now define the band errors $\{\varsigma_m\}_{m=1}^{M(\tau_o)}$ which are computed with the assumption that only unvoiced excitation exists in the synthetic spectrum. In this case, the excitation spectrum is the spectrum of random noise. By assuming that the spectrum magnitudes of random noise are unity over the entire signal spectrum, the band magnitude $A_m(\tau_o)$ of band *m* would equal to the average magnitude spectrum of input speech $|S(k)|$ with lower and upper limits at $a_m(\tau_o)$ and $b_m(\tau_o)$, respectively. Hence, the band error for unvoiced synthetic spectrum is computed as

$$\varsigma_m = \frac{\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)}\left(|S(k)| - \mathcal{U}_m\right)^2}{(1-W\tau_o)\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)}|S(k)|^2} \qquad m = 1,2,\cdots,M(\tau_o) \tag{8}$$

where $\mathcal{U}_m = \frac{1}{\left[b_m(\tau_o)-a_m(\tau_o)+1\right]}\sum_{k=a_m(\tau_o)}^{b_m(\tau_o)}|S(k)|$. The optimum transition frequency ω_c is determined by

$$\omega_c = \arg\min_{\forall x_i}\{C(x_i)\} \tag{9}$$

where $x_i = \frac{2\pi(i+0.5)}{\tau_o}$ for $i = 0,1,\cdots,M(\tau_o)-1$. The cost function $C(x_i)$ is defined as

$$C(x_i) = \frac{1}{M(\tau_o)}\left[\sum_{m=1}^{\lfloor \tau_o x_i/2\pi \rfloor}\xi_m + \sum_{m=\lfloor(\tau_o x_i/2\pi)+1\rfloor}^{M}\varsigma_m\right] \tag{10}$$

In this simulation, an artificial spectrum is used to check the performance of the proposed V/UV detection procedure. The artificial spectrum is characterised by 10 voiced harmonics in the low band and the high band is filled with unvoiced signal as shown in Figure 2(a). The corresponding cost function for the determination of ω_c is shown in Figure 2(b). A spectrogram of

real speech signal and the corresponding V/UV transitions are shown in Figure 3.

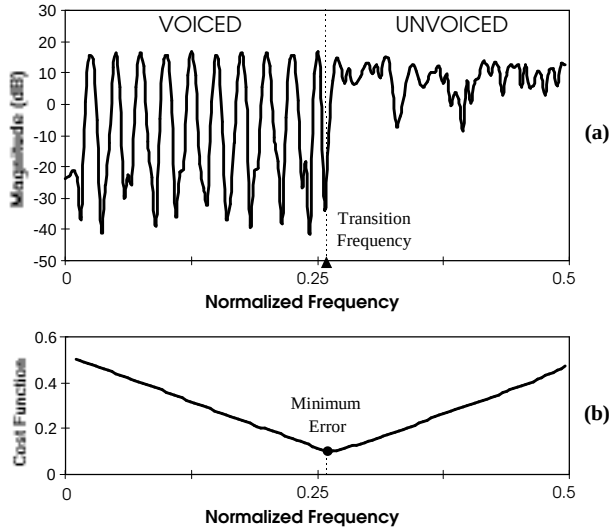

Figure 2 An illustration of the V/UV transition frequency: (a) an artificial spectrum and its V/UV transition frequency; (b) the cost function $C(x_i)$ corresponding to the artificial spectrum.

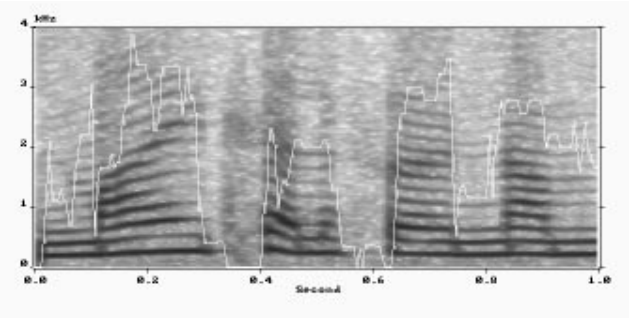

Figure 3 A speech spectrogram and the corresponding V/UV transitions (white curve).

4. VARIABLE BIT RATE SPECTRAL QUANTIZATION

In [3], the 1.6 kbps fixed bit rate MBELP coder employed an embedded 2-dimensional differential LSP (2DDLSP) quantization scheme for the quantization of the 10th-order spectrum envelope. The prediction residuals are scalar quantized by using only 24 bits/frame. If the number of bits required for SQ of the i^{th} prediction residual is denoted by B_i, the 24 bits for spectral quantization are allocated as $\{\mathrm{B}_i\}_{i=1}^{10} = \{2,3,3,3,3,2,2,2,2,2\}$. By using the spectral distortion (SD) measure defined in [5], the SQ scheme can achieve an average SD of 1.26 dB. In this paper, we propose a spectral quantization scheme where the required number of bits is changed adaptively according to the V/UV transition frequency ω_c. The technique is based on the observation that the magnitudes of unvoiced bands are usually low and relatively flat. So, without obvious degradation in subjective speech quality and average SD, we can quantize the prediction residuals of unvoiced bands by using fewer number of bits. In addition, the quantizer structure is simple since the unvoiced bands are located in the higher frequency region of a speech spectrum.

The prediction residuals associated with the line spectral frequencies in unvoiced bands are quantized with fewer number of bits through a variable dimension vector quantizer. Other prediction residuals associated with the line spectral frequencies in voiced bands are scalar quantized. With an optimum V/UV transition frequency ω_c and a set of LSP parameters $\{\theta_i\}_{i=1}^{10}$, the dimension P_v of the vector quantizer is determined as

$$P_v = 10 - P_s \tag{11}$$

where P_s is the number of the scalar quantizers. By utilising the ordering property of LSP parameters, i.e. $\theta_0 < \theta_1 < \theta_2 < \cdots < \theta_{10} < \theta_{11}$ for a 10th-order system, P_s is determined as

$$P_s = \arg\min_{\forall i}\{d_i : d_i > 0\} \tag{12}$$

where $d_i = \theta_i - \omega_c$ for $i = 2,3,\cdots,10$. By denoting the k^{th} codevector of a codebook of size K as $\mathbf{r}_k = \{r_j(k)\}_{j=P_s+1}^{10}$ for $k = 0,1,\cdots,K-1$, the best codebook index k_o is determined as

$$k_o = \arg\min_{\forall k}\left\{\sum_{i=P_s+1}^{10}\left(\theta_i - \tilde{\theta}_i(k)\right)^2\right\} \tag{13}$$

where

$$\tilde{\theta}_i(k) = \begin{cases} \mathrm{a}_i\hat{\theta}_{i-1} + \mathrm{b}_i\dot{\hat{\theta}}_i + r_i(k) & \text{when } i = P_s+1 \\ \mathrm{a}_i\tilde{\theta}_{i-1}(k) + \mathrm{b}_i\dot{\hat{\theta}}_i + r_i(k) & \text{otherwise} \end{cases} \tag{14}$$

is the estimated i^{th} LSP parameter with respect to the k^{th} codevector. The parameter $\hat{\theta}_{i-1}$ is the $(i\text{-}1)^{\text{th}}$ quantized LSP parameter of the present frame and $\dot{\hat{\theta}}_i$ is the i^{th} quantized LSP parameter of the previous frame. The coefficients $\{\mathrm{a}_i\}$ and $\{\mathrm{b}_i\}$ are the intra-frame and the inter-frame prediction coefficients, respectively [6]. The codebook sizes for different values of vector quantizer dimension P_v are shown in Table 1.

Table 1 Vector Quantizer Dimensions and the Corresponding Sizes of Codebooks

P_v	2	3	4	5	6	7	8
K	8	8	16	32	64	128	256

For $P_v = 1$ and $P_v \geq 9$, all the prediction residuals are quantized by using the 24-bit 2DDLSP SQ. For silent and very noisy speech, the proposed variable bit rate quantizer may not provide satisfactory performance without a large codebook. In order to avoid the exhaustive search on a large codebook, in the silence mode, only 5 of the prediction residuals are quantized by SQ and the LSPs of a 10th-order system are reconstructed by

interpolations. The proposed variable bit rate spectral quantization scheme working at 19 bits/frame in average is comparable with the fixed 24-bit 2DDLSP SQ scheme used in [3] in term of average SD. The results are given in Table 2.

Table 2 Average SD of the Fixed 24-bit 2DDLSP SQ Scheme and the Proposed Variable Rate Spectral Quantization Scheme

Scheme	Avg. SD (dB)	Outliers (%)	
		2-4 dB	>4 dB
24-bit Fixed Rate	1.26	5.37	0.32
19-bit (avg.) Variable Rate	1.31	5.61	0.53

5. THE VARIABLE BIT RATE MBELP CODER

Speech signal is bandlimited to 4 kHz and then sampled in 8 kHz. The digitised speech signal is segmented by Hamming window with a rate of 40 frames/sec. The window size is 256 samples in time. Speech is classified into silence, transient or steady for each segment. A specify mode of operation will be selected according to the speech classification. A block diagram of the variable bit rate MBELP coder is shown in Figure 4. The bit allocations of the proposed speech coder is shown in Table 3. The number of bits for spectrum envelope quantization varies from 13 to 24 bits. Average bit rate of the proposed MBELP coder is 1.24 kbps.

Table 3 Bit Allocations

	Silence	Transient		Steady
		Low Energy	High Energy	
Mode	2	2		2
Gain	6	6	6	6
Spectrum Envelope	14	14	13-24	13-24
V/UV	-	3		3
Pitch	-	8		8
Transition Time Index	-	2		-
Total	22	54-65		32-43

In the transient mode, 2 bits are required for the time index of the occurrence of abrupt transition. With the abrupt transition time t_v, the speech segment is split for analysis. A variable-length window is used to capture the low energy portion and a fixed-length window (256 samples in length) is used to capture the high energy portion of the speech segment. The low energy portion is quantized in the same way as the silent speech while the high energy portion is quantized in the same way as the steady speech segments. For silent speech and the low energy portion of transient speech, the V/UV transition frequency is assumed to be zero. Therefore, no V/UV information is required to be sent. On the other hand, as a mean to improve the speech quality, the candidates of V/UV transition frequency are increased to twice of that in [3]. Informal listening tests shown that the speech quality of the proposed MBELP coder is comparable with the fixed rate 1.6 kbps MBELP coder and is sufficient for communication purposes.

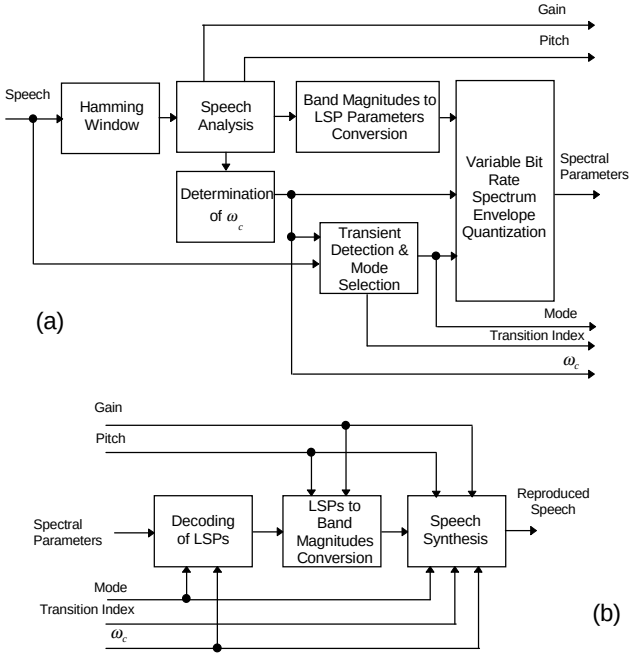

Figure 4 Block diagram of the variable bit rate MBELP coder: (a) encoder; (b) decoder.

6. CONCLUSION

A spectral quantization scheme is proposed to quantize the spectrum envelope in variable bit rates. By referring to the V/UV transition frequency, the LSP parameters are split for quantization. A high quality variable rate MBELP coder was developed for coding of speech.

REFERENCES

[1] A. Das and A. Gersho, "A variable-rate natural-quality parametric speech coder," *Proc. ICC*, pp. 216-220, 1994.

[2] S.A. McClellan and J.D. Gibson, "Variable rate CELP based on subband flatness," *Proc. ICC*, pp. 1409-1413, 1995.

[3] W.M.E. Yu and C.F. Chan, "Efficient multiband excitation linear predictive coding of speech at 1.6 kbps," *Proc. 4th Eurospeech*, pp. 685-688, 1995.

[4] D.W. Griffin and J.S. Lim, "Multi-band excitation vocoder," *IEEE Trans. on ASSP*, vol.36, pp. 1223-1235, August 1988.

[5] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. on Speech & Audio Proc.*, vol.1, pp. 3-14, January 1993.

[6] C.C. Kuo, F.R. Jean and H.C. Wang, "Low bit-rate quantization of LSP parameters using two-dimensional differential coding", *Proc. ICASSP*, pp. 97-100, 1992.