HARMONIC+NOISE CODING USING IMPROVED V/UV MIXING AND EFFICIENT 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.
E-mail: 96421256@plink.cityu.edu.hk, eecfchan@cityu.edu.hk

ABSTRACT

This paper presents a harmonic+noise speech coder which uses an efficient spectral quantization technique and a novel voiced/unvoiced (V/UV) mixing model. The harmonic magnitudes are coded at 23 bits/frame using the magnitude response of a linear predictive coding (LPC) system. The difference between the harmonic magnitudes and the sampled magnitude response is minimized by the closed-loop approach. The V/UV mixing is modeled by a smooth function which is derived from the speech spectrum envelope based on the flatness measure. The V/UV mixing model allows noise to be added in the harmonic portion of speech spectrum so that buzzyness is reduced. The V/UV mixing information is determined from the spectral parameters available in the decoder, no bits are needed for transmitting the V/UV information. A 1.4 kbps harmonic coder is developed. The speech quality of the coder is comparable to other harmonic coders operating at higher rates.

1. INTRODUCTION

Harmonic+noise model has been extensively used in coding speech signals at bit rates around 2.4 kbps [1][2]. In this model, speech spectrum is divided into voiced and unvoiced regions. In the voiced region, speech signal is modeled as the sum of harmonics of the fundamental frequency (pitch). In the unvoiced region, speech signal is modeled as random noise. The parameters required by the harmonic coder are the voicing information, the pitch and the short-term spectral information. For low-rate harmonic coding, the harmonic magnitudes are usually coded to a fixed rate by using all-pole linear predictive coding (LPC) technique [3]. However, all-pole LPC system would introduce spectral distortion and is partially responsible for producing speech of synthetic quality. In this paper, we propose a technique which improves the coding performance by closed-loop minimization of the difference between the harmonic magnitudes and the LPC magnitude spectrum. The line spectrum pair (LSP) parameters are used to represent the LPC spectrum. The 10^{th}-order LSP parameters are quantized efficiently by exploring both the intra- and inter-frame correlation using 2-dimensional differential LSP (2DDLSP) coding [4]. The prediction residuals of the 2DDLSP predictor are quantized using the split vector quantization (VQ) technique. Another factor that also affects the robustness of harmonic coders is the incorrect estimation of voiced/unvoiced (V/UV) information for rapid time-varying signal, for example, a speech onset. For efficient coding of V/UV information, speech spectrum is usually divided into low- and high-frequency regions. The low-frequency region is assumed to contain only voiced signal while the high-frequency region is assumed to contain only unvoiced signal. A transition from voiced region to unvoiced region is determined during speech analysis and then sent together with other information to the decoder for synthesis. Typically, at least 3 bits are required to code the V/UV transition. Although rigid V/UV division into low- and high-frequency regions is very efficient for coding, it introduces many artifacts in the synthetic signal. A noticeable distortion is the buzzyness quality as a result of very "clean" harmonic spectrum in the voiced region. In this paper, a V/UV mixing model is proposed to improve the mixing of the voiced and the unvoiced signals. Additionally, based on the proposed V/UV mixing model, a technique is developed to determine the V/UV mixing from the LSP parameters available in the decoder. Therefore, no bits are required to transmit the V/UV information.

2. QUANTIZATION OF HARMONIC MAGNITUDES

The number of harmonic magnitudes of a short-time speech spectrum varies according to the fundamental frequency of the speech segment. In order to code the variable-length harmonic magnitude vector using fixed-rate coding technique, we use the magnitude response of a 10^{th}-order all-pole LPC system to approximate the harmonic magnitudes. If the fundamental frequency of the speech frame is known, a close approximation of the harmonic magnitudes can be evaluated from the all-pole filter parameters and the filter gain provided the difference between the harmonic magnitudes and the LPC magnitude spectrum is minimized. In this paper, we propose a quantization technique which minimizes the quantization distortion in the mean-square sense. In this technique, the optimum codewords are selected by minimizing the error function

$$\varepsilon(\ell,\mathbf{k}) = \sum_{m=1}^{M}\left[B_m - \frac{g(\ell)}{\left|A(m\omega_o,\mathbf{k})\right|}\right]^2 \qquad (1)$$

where ω_o is the fundamental frequency, ℓ is the index of the differential gain candidate, and $\mathbf{k}$ is a vector of the indices of the filter parameter codebooks. The m^{th} harmonic magnitude is denoted by B_m and the total number of harmonic magnitudes is M. The gain and the sampled magnitude response of the all-pole filter are denoted by $g(\ell)$ and $g(\ell)/\left|A(m\omega_o,\mathbf{k})\right|$, respectively. The filter parameters are represented by the 10^{th}-order LSP parameters because of their desirable properties in spectral quantization. The all-pole filter response can be obtained from the LSP parameters by noting that

$$\left|A(m\omega_o,\mathbf{k})\right| = \frac{\left|P(m\omega_o,\mathbf{k}) + Q(m\omega_o,\mathbf{k})\right|}{2}$$

where

$$P(m\omega_o,\mathbf{k}) = 2^6 e^{-j\frac{11m\omega_o}{2}} \cos\frac{m\omega_o}{2} \prod_{i\ \text{odd}}^{9} \left[\cos m\omega_o - \cos\theta_i(\mathbf{k})\right],$$

$$Q(m\omega_o,\mathbf{k}) = -j2^6 e^{-j\frac{11m\omega_o}{2}} \sin\frac{m\omega_o}{2} \prod_{i\ \text{even}}^{10} \left[\cos m\omega_o - \cos\theta_i(\mathbf{k})\right],$$

and $\{\theta_i(\mathbf{k})\}_{i=1}^{10}$ are the LSP parameters with respect to the codebook indices **k**. For the sake of efficient quantization of the LSP parameters, a 2DDLSP predictor is employed to predict the values of the LSP parameters and the prediction residuals are quantized using split VQ. The prediction residual vector is split into 3 subvectors and their dimensions are (3,3,4). By using k_1, k_2, and k_3 to denote the codebook indices, the vector $\mathbf{k} = \{k_1, k_2, k_3\}$ and the vectors $\{r_i(k_1)\}_{i=1}^{3}$, $\{r_i(k_2)\}_{i=4}^{6}$, and $\{r_i(k_3)\}_{i=7}^{10}$ are the elements of the prediction residual codevector **r**(**k**). According to the codevector **r**(**k**), an estimation of the i^{th} LSP parameter with respect to the codebook indices **k** is

$$\theta_i(\mathbf{k}) = a_i\tilde{\theta}_{i-1} + b_i\dot{\tilde{\theta}}_i + r_i(\mathbf{k})$$

where $\tilde{\theta}_{i-1}$ is the $(i\text{-}1)^{\text{th}}$ quantized LSP parameter of the present frame and $\dot{\tilde{\theta}}_i$ is the i^{th} quantized LSP parameter of the previous frame. The coefficients $\{a_i\}$ and $\{b_i\}$ are the intra- and the inter-frame prediction coefficients of the 2DDLSP predictor [4], respectively. The difference between the gains of the previous frame and the present frame is quantized by scalar quantization (SQ) technique. The filter gain $g(\ell)$ in (1) is obtained by

$$g(\ell) = \dot{\tilde{g}} + \Delta g(\ell)$$

where $\Delta g(\ell)$ is the ℓ^{th} codeword of the differential gain table and $\dot{\tilde{g}}$ is the quantized filter gain of the previous frame. The error function $\varepsilon(\ell,\mathbf{k})$ in (1) is used for joint optimization of the filter gain and the filter parameters. The table index of the optimum differential gain and the codebook indices of the optimum LSP prediction residuals are determined by closed-loop minimization of the error function $\varepsilon(\ell,\mathbf{k})$. A block diagram of the proposed quantization technique is shown in Figure 1. The size of the differential gain table is 8. The sizes of the 3 prediction residual codebooks are 128, 128, and 64, respectively. Therefore the harmonic magnitudes are quantized using 23 bits only.

The proposed quantization technique is compared with that employed by a 1.6 kbps harmonic coder developed earlier [3]. In the 1.6 kbps harmonic coder, the harmonic magnitudes are also coded to a fixed rate by using all-pole LPC technique. The filter gain and the LSP prediction residuals are quantized using SQ technique. The selection of optimum codewords is performed in an open-loop fashion. The total number of bits required for the quantization of harmonic magnitudes is 30. The techniques are assessed objectively by evaluating their average spectral distortion. A spectral distortion (SD) measure defined in [5] is used for evaluation. The SD measure D_n for the n^{th} frame is defined as

$$D_n^{\,2} = \frac{100}{M}\sum_{m=1}^{M}\left(\log_{10} B_m^{\,2} - \log_{10}\tilde{B}_m^{\,2}\right)^2$$

where $\tilde{B}_m$ is the quantized harmonic magnitude of the m^{th} band.

The techniques are evaluated over 7600 speech frames contributed by both male and female speakers. The average SDs of the 2 quantization techniques are tabulated in Table 1. It is observed that the proposed closed-loop quantization technique has a better performance at lower bit rate.

Table 1 Comparison of the Average SDs of the 2 Difference Harmonic Magnitude Quantization Schemes

	Bit Rate (bits/frame)		Average SD (dB)
	Gain Parameter	Spectral Parameters	
Open-Loop 2DDLSP-SQ	6	24	3.104
Proposed Technique	3	20	2.976

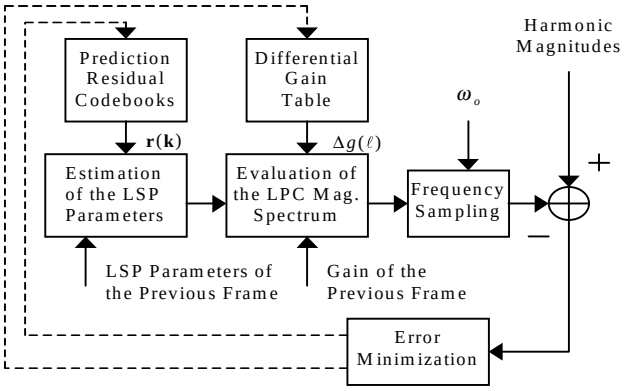

Figure 1 A block diagram of the proposed closed-loop quantization technique.

3. MODEL OF V/UV MIXING

For low-rate harmonic coders, a simplified V/UV mixing strategic based on the division of speech spectrum into low-frequency voiced region and high-frequency unvoiced region is always used. This model creates very "clean" harmonics in the voiced region and hence introduces buzzyness into the synthetic speech. Besides, at least 3 bits are needed to code the V/UV transition. In this paper, we propose a method to improve the V/UV mixing while simultaneously reduce the bits required to carry this information to the decoder. It has been shown that the speech spectrum envelope inherently correlates to the degree of V/UV mixing [6]. We can observe from speech spectra that high-frequency region of speech spectrum contains mostly noisy signal and spectral region with strong formants contains mostly voiced harmonics. Based on these observations, we can assume that signal in the high-frequency region with high spectral flatness can be classified as unvoiced while signal in low-frequency region with small spectral flatness can be classified as voiced. Now, let us define a spectral flatness measure which covers the region from θ to π in the speech spectrum $S(\omega)$ as

$$f(\theta) = \frac{1}{\pi-\theta}\int_{\theta}^{\pi}\left[\log|S(\omega)| - \bar{S}(\theta)\right]^2 d\omega \qquad (2)$$

where $\bar{S}(\theta) = \frac{1}{\pi-\theta}\int_{\theta}^{\pi}\log|S(\omega)|d\omega$ is the mean of the log

magnitude spectrum over the region from θ to π. This flatness measure will be low if the high-frequency region from θ to π contains only noisy signal. By comparing $f(\theta)$ to a predefined threshold T_{uv}, we can locate a transition point θ_t where signal in the spectrum region from θ_t to π is classified as unvoiced and signal in the spectrum region from 0 to θ_t is classified as voiced. If $S(\omega)$ is an unity gain LPC spectrum, we can experimentally determine the threshold T_{uv} to be 0.025. Note that this V/UV transition can be completely determined from the spectral parameters available in the speech decoder, therefore no bits are needed to carry the V/UV information to the decoder. However, we have to point out that this V/UV analysis is based on an assumption that the excitation contains at least some pitch harmonics. In case there is a lack of periodic excitation and the spectrum is completely unvoiced, then by using the spectral flatness measure to determine the voicing information is irrelevant. Therefore we still need to send a parameter for indicating the overall voicing decision. In practice, this can be easily done by encoding a special code, say 0, on the pitch which will be sent to the decoder. The accuracy of the proposed method is compared to a method based on closed-loop V/UV analysis proposed by us in [7]. Fig. 2 shows the spectrogram of a speech sentence from a female speaker. The white curve on the diagram is the V/UV transition curve obtained by closed-loop V/UV analysis and the dark curve is the V/UV transition curve obtained by using the proposed method based on spectral flatness. We can observe from the figure that the proposed method achieves very accurate estimation of V/UV transition.

The V/UV information obtained from the proposed method has been applied to a 1.6 kbps harmonic coder developed earlier [3]. We found that there is no drop in speech quality by using the new V/UV information. In fact, the speech quality is slightly improved. Also, since it is not necessary to transmit the V/UV information, the bit rate is reduced. The next major improvement in speech quality is based on using a smooth V/UV mixing rather than a rigid V/UV transition. It has been observed from many speech synthesis experiments that the quality of synthetic speech from a harmonic coder is always very buzzy. This is mainly due to the use of a "pure" harmonic excitation in the voiced speech spectrum. Fig. 3(a) shows a synthetic speech spectrum (dotted curve) derived from the original speech spectrum (solid curve) by using a rigid V/UV transition model. By observing the synthetic spectra, we find that there are strong dips in the inter-harmonic regions since no signal or noise exists there. On the other hand, by observing the original speech spectrum there is apparently a noise floor in the inter-harmonic regions and the strength of this noise floor seems to vary locally with the spectrum magnitudes. From physical point of view, as the air waves travel through the vocal tract, the noise generated as a result of turbulent air should be spread over the whole spectrum and be weighted by the vocal tract response. Based on this observation, we propose a smooth V/UV mixing function defined from the spectral flatness measure of (2) as:

$$v(\theta) = \begin{cases} 1 - \dfrac{f(\theta)}{2T_{uv}} & f(\theta) < T_{uv} \\ \dfrac{T_{uv}}{2f(\theta)} & f(\theta) > T_{uv} \end{cases}$$

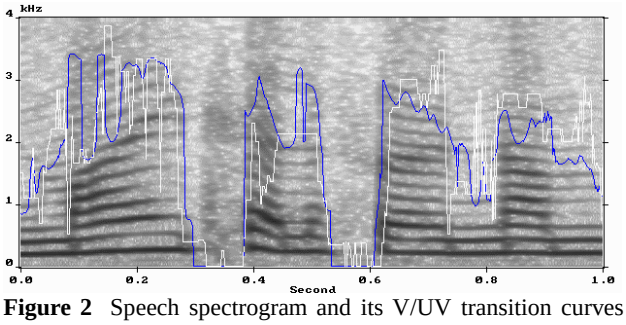

Figure 2 Speech spectrogram and its V/UV transition curves obtained from closed-loop analysis (white curve) and proposed method (dark curve).

Note that the V/UV transition is at $v(\theta_t) = 0.5$, and the voiced and unvoiced regions are corresponding to $v(\theta) < 0.5$ and $v(\theta) > 0.5$, respectively. Fig. 3(b) shows the mixing function $v(\theta)$ for the spectrum shown in Fig. 3(a). This V/UV mixing function can be combined with the speech spectrum envelope to generate the unvoiced spectrum. Practically, there are two approaches to generate the unvoiced spectrum. The first approach is the frequency-domain approach as similar to the one that is used in multiband excitation coder [2]. The magnitudes of the unvoiced bands are weighed by the V/UV mixing function prior to the unvoiced speech synthesis via IFFT. The second approach is the time-domain approach where the LPC spectrum is weighted by the V/UV mixing function. The weighted power spectrum is converted to autocorrelation data and then an all-pole LPC model is fitted to autocorrelation data to compute the synthesis filter's parameters for the unvoiced signal. Based on the experimental results achieved, we found that the second approach is slightly better than the first approach. Fig. 3(c) shows the speech spectrum of Fig. 3(a) overlay with plots of the LPC spectrum (dotted curve) and the response of the derived unvoiced synthesis filter (dashed curve). These plots reaffirm that the proposed V/UV mixing strategy models the noise floor of the speech spectrum accurately. Fig. 3(d) shows the combined voiced/unvoiced synthetic speech spectrum using the improved V/UV mixing model. Listening tests for comparing the quality of synthetic speech from the rigid and smooth V/UV mixing models have confirmed that buzzyness in the synthetic speech is significantly reduced by using the proposed smooth V/UV mixing function.

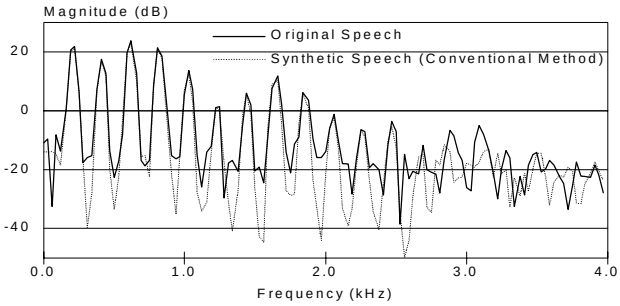

Figure 3(a) A typical speech spectrum (solid curve) and the corresponding synthetic spectrum (dotted curve) obtained using conventional V/UV mixing model.

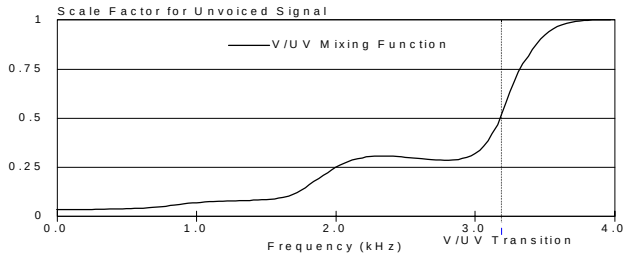

Figure 3(b) The V/UV mixing function $v(\omega)$ derived from speech spectrum shown in Fig. 3(a).

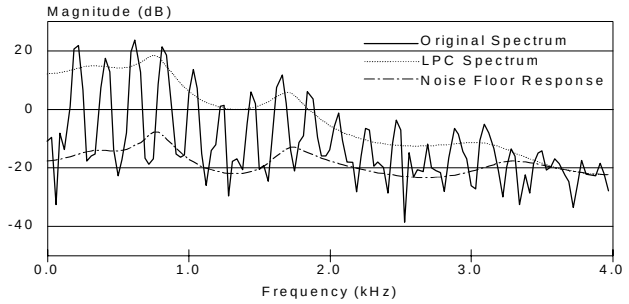

Figure 3(c) Response of the unvoiced synthesis filter (dashed curve).

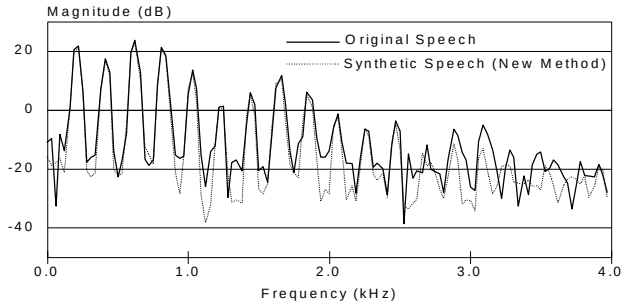

Figure 3(d) A typical speech spectrum (solid curve) and the corresponding synthetic spectrum (dotted curve) obtained using the improved V/UV mixing model.

4. A 1.4 KBPS HARMONIC CODER

The proposed techniques are applied to develop a 1.4 kbps harmonic coder. The coder operates at 8 kHz sampling frequency and the frame size is 160 samples. The parameters to be transmitted to the decoder are the differential gain, the LSP prediction residuals, and the pitch. Table 2 lists the bit allocation scheme. The coder has 3 modes of operation: 1) the steady voiced mode (V→V) indicates that the previous frame and the current frame are both voiced, 2) the transition mode (UV→V) indicates that the previous frame is unvoiced and the current frame is voiced, and 3) the unvoiced mode (UV) indicates that the current frame is unvoiced. The pitch code is set to 0 in unvoiced mode. In transition mode, the pitch is scalar quantized to 5 bits and, in steady voiced mode, it is differential quantized to 5 bits. The harmonic magnitudes and the unvoiced spectrum envelope are coded by SQ of the differential gain of the all-pole filter and split VQ of the 10^{th}-order LSP prediction residuals. The quantization error is minimized by the proposed closed-loop quantization technique. The differential gain is quantized to 3 bits and the LSP prediction residuals are quantized to 20 bits. The V/UV mixing function is obtained from the LSP parameters at the decoder. Therefore no extra bit is required for the transmission of V/UV information. The smooth V/UV mixing function is applied to synthesize unvoiced signal in the time-domain.

A pair-wise comparison test between the 1.6 kbps MBELP coder [3] and the proposed coder is performed. Ten listeners are involved in this test. The results show that all listeners prefer the 1.4 kbps coder. We also perform the subjective quality tests and the results show that the proposed 1.4 kbps harmonic coder achieves an MOS score of 3.3.

Table 2 Bit Allocation of the 1.4 kbps Harmonic Coder

Mode	V→V	UV→V	UV
Gain Parameter	3 bits		
Spectral Parameters	20 bits		
Pitch	5 bits (0xxxx) xxxx=1→15 differential	5 bits (0xxxx) xxxx=1→15 differential	5 bits (00000)

5. CONCLUSION

Techniques for improving the harmonic+noise model for speech coding at 1.4 kbps are proposed. The harmonic magnitudes are quantized efficiently by split VQ of the LSP prediction residuals based on closed-loop analysis. Significant improvements on voice quality are achieved by using a smooth V/UV mixing model where the model parameters are derived from spectrum envelope. A 1.4 kbps coder incorporating the techniques proposed is successfully developed and the speech quality achieved is sufficient for communications.

6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, eds., pp. 148-150, Elsevier Science B. V., 1995.

[2] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223-1235, August 1988.

[3] Eric W. M. Yu and C. F. Chan, “Efficient multiband excited linear predictive coding of speech at 1.6 kbps,” in *Proc. EUROSPEECH-95*, pp. 685-688, 1995.

[4] C. C. Kuo, F. R. Jean, and H. C. Wang, "Low bit-rate quantization of LSP parameters using two-dimensional differential coding," in *Proc. IEEE ICASSP-92*, pp. 97–100, 1992.

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

[6] C. F. Chan, “High-quality synthesis of LPC speech using multiband excitation model,” in *Proc. EUROSPEECH-93*, pp. 535-538, 1993.

[7] Eric W. M. Yu and C. F. Chan, “Variable bit rate MBELP speech coding via V/UV distribution dependent spectral quantization,” in *Proc. IEEE ICASSP-97*, pp. 1607-1610, 1997.