SPLIT BAND CELP (SB-CELP) SPEECH CODER

Mohammad R. Nakhai and Farokh A. Marvasti

Department of Electronic Engineering
King's College London, Strand, London WC2R 2LS, U.K.

ABSTRACT

In this paper, we discuss the split band code-excited linear prediction (SB-CELP) speech coder which employs an iterative version of the harmonic sinusoidal coding algorithm to encode the periodic contents of speech signal. Speech spectrum is split into two frequency regions of harmonic and random components and a reliable fundamental frequency is estimated for the harmonic region using both speech and its linear predictive (LP) residual spectrum. The resulting sinusoidal parameters are interpolated to reconstruct the periodicity in speech waveform. The level of periodicity is controlled by computing a cutoff frequency between the harmonic and random regions of spectrum. The random part of spectrum and unvoiced speech are processed using the CELP coding algorithm. The SB-CELP speech coder which combines the powerful features of the sinusoidal and CELP coding algorithms yields a high quality synthetic speech at 4.05 kb/s.

1. INTRODUCTION

In sinusoidal coding, the excitation signal is interpreted in terms of sum of sinusoids whose parameters are estimated from speech spectrum. This interpretation models well the voiced segments of speech but can impart an undesirable and tonal character to the noiselike unvoised segments of speech and arises objectionable artifacts in the synthetic speech [1]. Good quality of synthetic speech at rates above 4.8 kb/s has been obtained by CELP algorithm [2]. The CELP algorithm basically maximizes the spectrally weighted signal-to-noise ratio (SW-SNR) on a subframe-by-subframe basis to achieve a good match between the original and the reconstructed speech signals. However, it has been reported in [3] that in spite of decreasing SW-SNR, the perceptual quality of synthetic speech could be improved in CELP by increasing the periodicity of the voiced speech. Therefore, SW-SNR is not an ideal measure of the perceptual quality and instead preserving the correct degree of periodicity of speech signal is of great importance to its perceptual quality. At low bit rates, i.e., under 4.8 kb/s, the conventional CELP structure [2] does not reproduce the essential level of periodicity of voiced speech signals. A mixture of CELP and prototype-waveform interpolation (PWI) introduced in [4] achieves a high level of periodicity in voiced speech at low bit rates. In this method, a downsampled version of a representative pitch cycle, called prototype waveform, is transmitted. At the receiver, the prototype waveforms are linearly interpolated between their regularly spaced update points to reconstruct the voiced speech and the CELP algorithm is used to reconstruct the unvoiced speech. In this paper, we employ a modified version of the harmonic sinusoidal coding algorithm to achieve a high level of periodicity in voiced speech. In our approach, sinusoidal parameters are extracted from the magnitude spectrum of speech at regular intervals of 20 msec at the encoder and then interpolated between these update points at the decoder to reconstruct the voiced portions of speech. In this way, we smoothly evolve the periodic structure of the synthetic speech and, hence, obtain a high level of periodicity in reconstruction of the voiced parts of speech. In PWI method, the time domain pitch-cycle waveforms are measured and interpolated directly. In our method, the pitch frequencies and the amplitudes of sinusoids are estimated in the frequency domain and interpolated. We use the CELP algorithm to encode and decode the unvoiced parts of speech.

The paper is organized as follows. In the next section, the basic principles of the SB-CELP coder are discussed. In section 3, we introduce our revised version of the harmonic sinusoidal analysis method employed in the analysis part of the SB-CELP. In section 4, the synthesis part of the SB-CELP is discussed. Some experimental results are given in section 5. Finally, we conclude this paper with remarks in section 6.

2. BASIS FOR THE PROPOSED CODER

In mixed-voicing frames, the speech spectrum is modelled as the concatenation of a harmonic component and a random component. Throughout this paper, the harmonic component is termed as the periodic component and the random component as the aperiodic component. The border between periodic and aperiodic regions is specified by a cutoff frequency. Based on the location of the cutoff frequency, a fixed number of harmonic sinusoids are fitted to the periodic part of spectrum. The periodic part of speech is synthesized at the encoder and subtracted from the original speech to obtain the random component. The CELP algorithm [2] is tailored to encode the unvoiced sounds and the random component of the mixed-voicing speech. In this paper, we revise the harmonic sinusoidal analysis [5] in terms of fundamental frequency estimation and decomposition of speech spectrum into periodic and aperiodic components. The metric $\rho(\omega_0)$ and the signal-to-noise ratio $\eta(\omega_0)$ given in the following are two basic relations in the conventional harmonic sinusoidal coding algorithm [5].

This work was supported by Fujitsu Europe Telecom.

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} B(k\omega_0)\{\max_{\omega_l \epsilon \Lambda(k\omega_0)} [A_l D(\omega_l - k\omega_0)] - \frac{1}{2} B(k\omega_0)\} \quad (1)$$

$$\eta(\omega_0) = \frac{P_s}{P_s - 2\rho(\omega_0)} \quad (2)$$

where

$$D(\omega_l - k\omega_0) = \begin{cases} \frac{\sin[2\pi(\frac{\omega_l - k\omega_0}{\omega_0})]}{2\pi(\frac{\omega_l - k\omega_0}{\omega_0})} & \text{if } \mid \omega_l - k\omega_0 \mid \leq \frac{\omega_0}{2} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$\Lambda(k\omega_0) = \{\omega : k\omega_0 - \frac{\omega_0}{2} \leq \omega < k\omega_0 + \frac{\omega_0}{2}\}$,
$P_s = \sum_{l=1}^{L} A_l^2$,
A_l and ω_l are the magnitude and the corresponding frequency location of a measured peak of spectrum, respectively, L is the number of measured peaks, ω_0 is a pitch candidate and $K(\omega_0)$ is the number of harmonics of ω_0 in the speech bandwidth. The function $B(\omega)$ is typically an interpolating function fitted to the measured set of spectral peaks $\{A_l\}$. The function $D(\omega_l - k\omega_0)$ is centered at the k^{th} harmonic of pitch candidate ω_0 as defined in (3). The pitch frequency candidate which maximizes the metric in (1) is chosen as the pitch frequency estimate. The accuracy measure of fitness between an estimated harmonic set of sine-wave data and the sine-wave data measured from the original spectrum is given by $\eta(\omega_0)$ in (2).
In section 3, we introduce a method based on the iterative versions of (1) and (2) to estimate the pitch and the corresponding cutoff frequency which splits the speech spectrum into periodic and aperiodic regions.

3. ANALYSIS

The SB-CELP speech coder is operating in two modes, i.e., sinusoidal mode for voiced speech and periodic part of the mixed-voicing frames and CELP mode for unvoiced speech and random part of the mixed-voicing frames. In sinusoidal analysis, the pitch frequency is estimated in two stages. In the first stage, a coarse estimate of pitch is obtained over the baseband of 1500 Hz of spectrum where the harmonic structure is rich and reliable. In the second stage, the coarse pitch is used to estimate the accurate pitch and the cutoff frequency. In our approach, in contrast to the conventional sinusoidal analysis [5], no initial average pitch is required to determine the peaks of spectrum in order to estimate the coarse pitch. In the first stage, the set of spectral peaks $\{A_l\}$ are measured by determining the location of points at which the slope changes from positive to negative. Then, all small peaks falling more than 25dB below the adjacent peaks are removed. The metric $\rho(\omega_0)$ is evaluated for all possible pitch condidates ω_0. The candidate which results in the largest $\rho(\omega_0)$ is chosen as the coarse pitch estimate if the corresponding $\eta(\omega_0)$ exceeds a voicing threshold T_l. If $\eta(\omega_0)$ is less than T_l, the analysed spectrum belongs either to an unvoiced frame or to a voiced/mixed-voicing frame whose periodic energy in the baseband is concentrated under the first few harmonic peaks. In the latter case, as shown in Fig. 1, removing low energy harmonic peaks results in $\max_{\omega_l \epsilon \Lambda(k\omega_0)} [A_l D(\omega_l - k\omega_0)] = 0$ in (1) and boosts the value of interpolating function $B(\omega)$ over these harmonics. Hence, the term $\frac{1}{2} B(k\omega_0)$ in (1) can be considerable

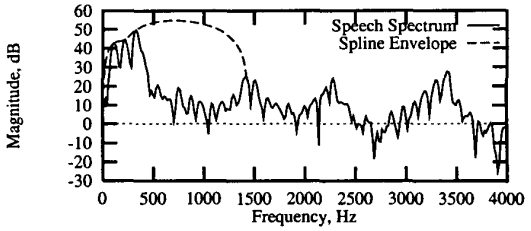

Figure 1: An example of speech spectrum with the periodic energy concentrated under the first 3 harmonics.

enough to significantly decrease the value of $\rho(\omega_0)$ over the low energy harmonics in the baseband and lead to the classification of a voiced/mixed-voicing frame as an unvoiced frame. This error is basically due to a sharp peak of the LP envelope at low frequencies. To improve the calssification accuracy, the frame whose $\eta(\omega_0)$ is less than the threshold T_l is passed through a LP inverse filter to form a residual signal. Then, the peak-picking and coarse pitch estimation procedures, as mentioned above, are repeated for the spectrum of the resulting residual. If the resulting $\eta(\omega_0)$ is still less than the threshold T_l, the frame is classified as unvoiced, otherwise it is passed to the second stage as a voiced/mixed-voicing frame. Inverse filtering whitens the speech spectrum and boosts the low energy regions of the spectrum which tend to be dominated by noiselike energy. This could result in a reduced periodic signal-to-noise ratio. That is why, the process of pitch estimation is mainly carried out over the speech spectrum rather than the residual signal. The objectives of the first stage are to obtain a coarse estimate of pitch and to identify the unvoiced frames. In the second stage, the locations of the frequencies are refined using the quadratic interpolation. The knowledge of coarse pitch is used in the method described in [6] to measure the underlying set of sine-wave amplitudes of spectrum $\{A_l\}$. A spline interoplating function is fitted to the sine-wave amplitudes to form the spectral peak envelope $B(\omega)$. Then a narrowed-down interval of pitch range centered at the coarse pitch is searched for the refined pitch. In the following, we introduce an iterative algorithm based on (1) and (2) to refine the coarse pitch and to estimate the cutoff frequency. Let $\rho_m(\omega_0)$ and $\eta_m(\omega_0)$ denote the overall values of $\rho(\omega_0)$ and $\eta(\omega_0)$ upto m^{th} harmonic for a refined pitch candidate ω_0, respectively. Using (1), we can write

$$\rho_m(\omega_0) = \rho_{m-1}(\omega_0) + \Delta\rho_m(\omega_0), \qquad m > b \quad (4)$$

where
$\Delta\rho_m(\omega_0) =$
$$B(m\omega_0)\{\max_{l:\omega_l \epsilon \Lambda(m\omega_0)} [A_l D(\omega_l - m\omega_0)] - \frac{1}{2} B(m\omega_0)\},$$

$\rho_{m-1}(\omega_0) = \rho_b(\omega_0) + \sum_{i=b+1}^{m-1} \Delta\rho_i(\omega_0)$,
$\rho_b(\omega_0)$ is the value of (1) evaluated over the frequency range from 0 to $\hat{f}$ and b is the number of harmonics in this range. We use $\hat{f} = 500Hz$ if the coarse pitch is less than 150 Hz, and $\hat{f} = 1000Hz$ if the coarse pitch exceeds 150 Hz to accommodate at least three harmonics of spectrum between 0 and $\hat{f}$. Using (2), we can also write

$$\eta_m(\omega_0) = \frac{P_s^m(\omega_0)}{P_s^m(\omega_0) - 2\rho_m(\omega_0)}, \qquad m > b. \quad (5)$$

where
$P_s^m(\omega_0) = P_s^{m-1}(\omega_0) + \Delta P_s^m(\omega_0)$,
$P_s^{m-1}(\omega_0) = P_s^b + \sum_{i=b+1}^{m-1} \Delta P_s^i(\omega_0)$,
$P_s^b = \sum_{[l:\omega_l \in \lambda(\hat{f})]} A_l^2$,
$\Delta P_s^i(\omega_0) = \sum_{[l:\omega_l \in \Lambda(i\omega_0)]} A_l^2, \qquad b < i \leq m,$
$\lambda(\hat{f})$ indicates the frequency range from 0 to $\hat{f}$. The iterative algorithm starts by computing $\rho_b(\omega_0)$ and the corresponding $\eta_b(\omega_0)$ over the region of spectrum specified by $\lambda(\hat{f})$. These computations are carried out for all refined pitch candidates ω_0 within a short frequency range centered at the coarse pitch and denoted by $\Re$. The iteration stops for the candidates whose corresponding $\eta_b(\omega_0)$ is less than a threshold denoted by T_h. If $\eta_b(\omega_0)$ falls below T_h for all candidates in $\Re$, then the candidate yielding the largest $\rho_b(\omega_0)$ is chosen as the refined pitch and the cutoff frequency is set at its minimum level given by $\hat{f}$. In fact, these frames are classified as mixed-voicing frames with poor harmonic structure. The iteration continues beyond the region $\lambda(\hat{f})$ for each candidate ω_0 within $\Re$ which results in $\eta_b(\omega_0) > T_h$, as follows:

- m^{th} Iteration: Starting with $m = b+1$, $\rho_{m-1}(\omega_0) = \rho_b(\omega_0)$, $P_s^{m-1}(\omega_0) = P_s^b$ compute $\rho_m(\omega_0)$ in (4) and $\eta_m(\omega_0)$ in (5).
- Iteration Stopping Condition: Stop the iteration when $\eta_m(\omega_0)$ falls under the threshold T_h and record the corresponding value of m.

The pitch candidate $\omega_0 = \omega^*$ maximizing $\rho_m(\omega_0)$ is declared as the refined pitch. Let m^* denote the value of m recorded for ω^*. The frequency range from zero to the end of the harmonic region $\Lambda(m^*\omega^*)$ is declared as the periodic part of spectrum, and the remaining frequency range is declared as the aperiodic part of spectrum. The low voicing threshold T_l is used in the first stage to classify the unvoiced frames and pass the voiced/mixed-voicing frames to the second stage. The high voicing threshold T_h is used in the second stage to determine the cutoff frequency. By listening to the synthesized speech good values for T_l and T_h were determined to be 6dB and 11dB, respectively. Fig. 2 represents the sinusoidal analysis stages of SB-CELP after coarse pitch estimation for the speech spectrum depicted in Fig. 2(a). The spline peak envelope fitted to the measured sine-wave amplitudes is also shown in Fig. 2(a). Fig. 2(b) represents different values of $\rho_m(\omega_0)$ computed in the pitch frequency range. Fig. 2(c) compares $\eta_m(\omega_0)$ for ω^* and 4 pitch candidates immediately before and after ω^*. In Fig. 2(c), the frequency at which the accuracy of fitness $\eta_m(\omega_0)$ falls down the threshold T_h indicates the upper cutoff frequency limit for the pitch candidate ω_0. The computations of $\rho_m(\omega_0)$ in (4) and $\eta_m(\omega_0)$ in (5) are only carried out for pitch candidates in a narrow range centered at the coarse pitch. This range is indicated by the bold line in Fig. 2(b). The refined pitch ω^* corresponds to the peak value of $\rho_m(\omega_0)$ in Fig. 2(b) and results in the highest values for $\eta_m(\omega^*)$ as depicted by the bold line in Fig. 2(c). The spline peak envelope can be encoded with high accuracy using a high order LP all-pole filter. The resulting high order coefficients can be transferred into the line spectral frequency (LSF) domain by employing the algorithm described in [7]. However a 10-12 order all-pole filter is enough to prevent the

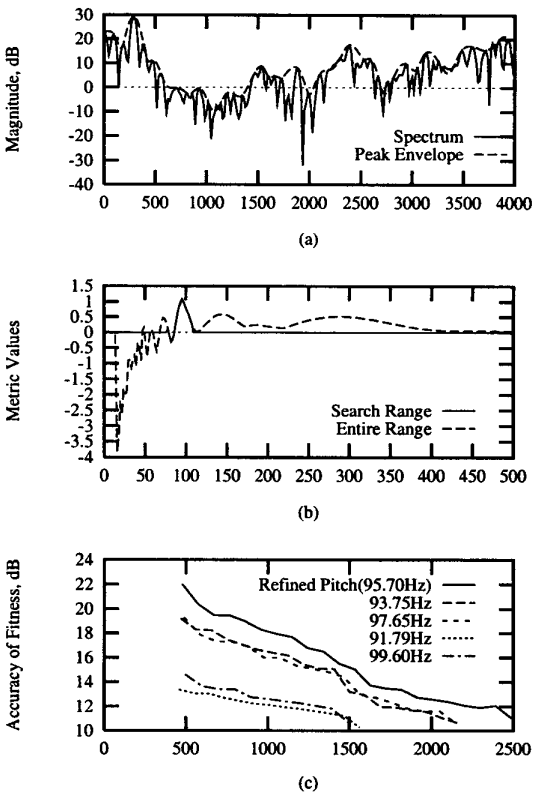

Figure 2: Illustration of pitch refining and spectral decomposition stages; the x-axis is the frequency axis in Hz.

serious degradations in the encoded spectral envelope. The CELP algorithm uses a fixed random codebook and the LP peak envelope parameters, obtained in the sinusoidal analysis mode, to encode the random part of speech waveform. The frames classified as unvoiced are totally analysed in the CELP mode. These frames are analysed in the time domain to extract the LPC parameters which are then used by the CELP algorithm to encode the unvoiced frames.

4. SYNTHESIS

The decoder of SB-CELP reconstructs the periodic part of the synthetic speech $s_p^i(n)$ in frame i as

$$s_p^i(n) = \sum_{l=1}^{L} a_l^i(n) \cos[\psi_l^i(n)] \qquad (6)$$

where L is the greater of the number of harmonics in frames $i-1$ and i, $a_l^i(n)$ and $\psi_l^i(n)$ are, respectively, the amplitude and the phase evolution functions between the parameter update points at frames $i-1$ and i. Matched harmonic pairs in frames $i-1$ and i are defined as the harmonic frequencies whose absolute difference is less than a threshold.

Parameters	Subframe	Frame
Fundamental Frequency	–	7
Voicing & Cutoff Frequency	–	4
Spectral Peak Envelope	–	30
Codebook Index (CELP)	7	28
Codebook Gain (CELP)	3	12
Total		81

Table 1: Bit allocation for the SB-CELP speech coder.

An unmatched harmonic in one frame is matched to a hypothetical harmonic in the other frame which has the same frequency but zero amplitude. Amplitudes of the matched harmonic pairs are linearly interpolated in time and denoted by $a_l^i(n)$. The linear interpolation function of the matched harmonic frequencies at frames $i-1$ and i is integrated between the update points and added to the initial phase measured at the end of the frame $i-2$ to form the phase evolution function $\psi_l^i(n)$. The initial phase for the harmonic frequency in frame i which is matched to itself with zero amplitude in frame $i-1$ is set to zero. In transition from an unvoiced frame $i-1$ to a voiced/mixed-voicing frame i, hypothetical harmonics are assumed in frame $i-1$ with frequencies equal to the harmonic frequencies in frame i but with zero amplitudes and with zero initial phases. In transition from a voiced/mixed-voicing frame $i-1$ to an unvoiced frame i, hypothetical harmonics are assumed in frame i with frequencies equal to the harmonic frequencies in frame $i-1$ but with zero amplitudes. The CELP decoding algorithm is used to synthesize the unvoiced frames and the aperiodic part of the mixed-voicing frames. The individual synthetic periodic and random components are combined to obtain the synthetic speech.

5. RESULTS

The original and the synthetic spectrograms of a dialogue spoken by a female and a male speakers are shown in Fig. (3). Table 1 shows the bit allocations for 20 msec frames and 5 msec subframes. A 4 kb/s CELP coder with only a fixed codebook was used to code the unvoiced frames. Uniform quantization was used to code the fundamental frequency. The cutoff frequency was uniformly quantized and coded together with the unvoiced classification information by 4 bits. For this purpose, starting from 500 Hz, the spectrum was divided into 14 equal frequency bands. Subjective listening tests were carried out on female and male speech using twenty subjects and under error-free conditions for the proposed SB-CELP coder operating at 4.05 kb/s, FS1016 DoD CELP operating at 4.8 kb/s and Inmarsat IMBE operating at 4.15 kb/s. The resulting MOS scores indicated the superior synthetic speech quality of SB-CELP compared to that of CELP and IMBE.

6. CONCLUSIONS

A unified algorithm was introduced to estimate the pitch and the cutoff frequencies without requiring the knowledge of an initial average pitch. The proposed algorithm, first, estimates a coarse pitch for voiced/mixed-voicing frames

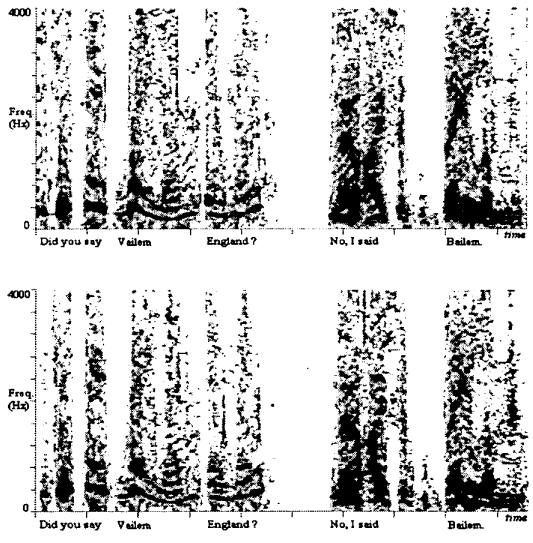

Figure 3: Spectrogram of 4 sec of (female voice) "Did you say Vailem England?" followed by (male voice) "No, I said Bailem". Top: original. Bottom: Synthetic.

and, then, refines the coarse pitch and splits the speech spectrum into periodic and aperiodic regions. Based on this method and the CELP coding algorithm, the basic rules of the SB-CELP speech coder was introduced.

7. REFERENCES

[1] M. W. Macon and M. A. Clements, "Sinusoidal Modeling and Modification of Unvoiced Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6 pp. 557-560, November 1997.

[2] M. Schroeder and B. Atal, "Code Excited Linear Prediction (CELP): High Quality Speech at Low Bit Rates," *Proc. IEEE Int. Conf. on Acoustic, Speech, and Signal Processing*, pp. 937-940, March 1985.

[3] Y. Shoham, "Constrained-Excitaion Coding of Speech at 4.8 kb/s," *Advances in Speech Coding*, pp. 339-348, B. S. Atal, V. Cuperman and A. Gersho, Editors, Kluwer Academic Press, Holland, 1991.

[4] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 4, pp. 386-399, October 1993.

[5] R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Voicing Decision Based on a Sinusoidal Speech Model," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 249-252, 1990.

[6] D. B. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. ASSP-29, pp. 786-794, 1981.

[7] M. R. Nakhai and F. A. Marvasti, "A Novel Algorithm to Estimate the Line Spectral Frequencies from LPC Coefficients," *Proc. IEEE Int. Symp. on Circuts and Systems* , Monterey,California, TAA4-3 (pp. 1-4), June 1998.