A 4.1 KB/S HYBRID SPEECH CODER

Mohammad R. Nakhai and Farokh A. Marvasti

King's College, University of London,
Department of Electronic Engineering, Strand, London WC2R 2LS, UK.

ABSTRACT

In this paper, a hybrid speech coder based on a mixture of sinusoidal and CELP coding algorithms is introduced. The conventional harmonic sinusoidal coding algorithm is revised in terms of fundamental frequency estimation, unvoiced classification and decomposition of speech spectrum into two regions of harmonic and non-harmonic components. Sinusoidal and CELP coding algorithms are used to code the harmonic and non-harmonic components, respectively. Peaks of speech spectrum are, first, interpolated and, then, represented by a linear predictive all-pole filter parameters. A harmonic tracking algorithm interpolates the sinusoidal parameters between adjacent frames to introduce an essential level of periodicity in the synthetic speech waveform. Pure unvoiced frames are only coded by CELP algorithm. Spectrograms of the original speech and its synthetic versions synthesised by the proposed hybrid speech coder operating at 4.1 kb/s and the Motorola GSM half-rate speech coder operating at 5.6 kb/s are compared in both noise-free and noisy environments.

1. INTRODUCTION

Good quality of synthetic speech at rates above 4.8 kb/s has been obtained by various speech coders such as VSELP [1] and CS-CELP [2] which are basically based on the conventional CELP algorithm [3]. However, at low bit rates, i.e., under 4.8 kb/s, the conventional CELP structure does not reproduce the essential level of periodicity of voiced speech signals [4]. It has been shown in [5] that the perceptual quality of synthetic speech can be improved in CELP even at low bit rates by increasing the level of periodicity in the reconstructed voiced speech. Harmonic sinusoidal method [6] models well the voiced parts of speech at low bit rates but can impart an undesirable and tonal character to the noiselike unvoiced parts of speech and arise objectionable artifacts in the synthetic speech.

In this paper, we replace the long-term prediction part of the CELP speech coder, which is usually based on an adaptive codebook, with a revised harmonic sinusoidal scheme to achieve a high level of periodicity in voiced speech, as is common in natural speech, at low bit rates. Sinusoidal parameters are extracted from the magnitude spectrum of speech at regular intervals of 20 msec at encoder and then interpolated between these update points at 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 the reconstructed voiced speech. We use the CELP algorithm , with only a fixed codebook, to code the unvoiced parts of speech.

This paper is organised as follows. In Section 2, we introduce the proposed hybrid encoder. Section 3 describes the principles of our hybrid decoder. The quantization methods used in the proposed coder are discussed in Section 4. Some illustrations are given in Section 5 to demonstrate the performance of our hybrid speech coder. Finally, we conclude this paper with remarks in Section 6.

This work was supported by Fujitsu Europe Telecom.

2. HYBRID ENCODER

In general, for a speech segment $s(t)$ composed of voiced and unvoiced sounds, we assume the following model in the spectral domain:

$$S(\omega) = S_p(\omega) + S_a(\omega) \tag{1}$$

where $S(\omega)$ is the Fourier transform of $s(t)$, $S_p(\omega)$ is the harmonic part or the lowpass component of $S(\omega)$, and $S_a(\omega)$ is the non-harmonic part or the highpass component of $S(\omega)$. The hybrid encoder decomposes speech spectrum $S(\omega)$ into two parts of harmonic and non-harmonic components, estimates a fundamental frequency in the harmonic region and classifies pure unvoiced frames. We have already introduced the principles of our hybrid encoder in [7]. The proposed encoder is operating in two modes, i.e., sinusoidal mode for periodic part of speech waveform and CELP mode for noise-like part of speech waveform. In sinusoidal mode, a pitch/fundamental frequency, a cut-off frequency which splits the speech spectrum into harmonic and non-harmonic parts, and a set of linear predictive (LP) coefficients which models the peak-envelope of magnitude spectrum of speech are estimated for each voiced frame. In voiced frames, these LP coefficients, which are estimated in the frequency domain, are also used to model the vocal tract synthesis filter for processing the noise-like component of speech waveform by CELP. In pure unvoiced frames, time domain linear predictive coding (LPC) parameters are used to represent the vocal tract synthesis filter. Throughout this paper, pure voiced frames and also frames which are a mixture of voiced and unvoiced sounds will be termed as voiced. In particular, our approach differs from earlier methods in the following ways:

1. The method described in [6] requires an initial average pitch to estimate pitch. In this method, an experimental function is used to estimate the cut-off frequency and sinusoids with equally spaced frequencies

0-7803-5471-0/99/$10.00©1999 IEEE

are assumed for the region beyond the cut-off frequency to model the non-harmonic part of spectrum. Our approach [7] estimates pitch without requiring an initial average pitch and determines the cut-off frequency from speech spectrum using a threshold for the required accuracy of fitness between the original and the synthetic spectrum. In our approach [7], the non-harmonic part of spectrum is modelled by CELP algorithm.

2. In [5], pitch cycle prototype waveforms are measured in the time domain at the encoder and directly interpolated between their update points at the decoder to reconstruct the voiced sounds. In our approach, amplitudes and frequencies of sinusoids are measured in the frequency domain at the encoder and interpolated between their update points at the decoder to synthesise the voiced sounds.

3. HYBRID DECODER

Fundamental frequency variation between adjacent frames in the original speech spectrum is fairly continuous with time, when small analysis/synthesis window shifts in time is used. Small analysis/synthesis window shift results in a smooth time-evolution of the sinusoidal parameters at the expense of a high transmission bit rate, which is not useful in even medium rate speech coding applications such as GSM. Hence, provisions should be taken to avoid the discontinuities in model parameters due to a large window shift and achieve a high level of periodicity of a voiced speech even at low bit rates.

Speech synthesis in hybrid decoder is performed in two modes, i.e., harmonic sinusoidal mode to reproduce the periodic portions of speech and CELP mode to reproduce the noise-like portions of speech. We have used time domain approach in both modes to obtain the synthetic speech. Periodic speech is synthesised as sum of sinusoids with frequencies at harmonics of a fundamental frequency and amplitudes taken from the all-pole model of the spectral peak envelope. This technique allows us to implement an efficient and effective method for interpolating the harmonic frequencies and amplitudes across the frame boundaries and, hence, achieve a high level of periodicity in the synthetic voiced speech waveform. Interpolating process, also, prevents the quality degradation in synthetic speech, which is due to the abrupt changes in harmonic sinusoidal parameters in transition from one frame to the next frame. Noise-like part of speech waveform and also pure unvoiced frames are synthesised using the CELP decoding algorithm.

Periodic component of synthetic speech on the i-th frame (i.e., $\tilde{s}_p^i$) can be synthesised in time domain as sum of sinusoids given by

$$\tilde{s}_p^i = \sum_{l=1}^{M} a_l^i(n) cos[\psi_l^i(n)] \tag{2}$$

where M is the total number of sinusoidal components evolving from frame $i-1$ to frame i, $a_l^i(n)$ and $\psi_l^i(n)$ are the l-th amplitude track and the l-th phase track at time n, respectively, which evolve the corresponding sinusoid from frame $i-1$ to frame i. Let the y-th harmonic sinusoidal component in frame i be denoted by (A_y^i, ω_y^i) where A_y^i and ω_y^i represent the amplitude and frequency of the sinusoid, respectively. We define the function

$$d(\omega_x^{i-1}, \omega_y^i) = \mid \omega_x^{i-1} - \omega_y^i \mid \tag{3}$$

as the distance between ω_x^{i-1} and ω_y^i where the index x denotes a parameter (amplitude or frequency) in frame $i-1$. We say, harmonic frequencies ω_x^{i-1} and ω_y^i in frames $i-1$ and i are matched if

$$d(\omega_x^{i-1}, \omega_y^i) \leq \delta, \tag{4}$$

where δ is defined as the matching interval.

In two adjacent frames, a harmonic in one frame is matched either to a real harmonic or a hypothetical harmonic in the other frame. In the latter case, a harmonic with zero amplitude and with the same frequency as that of the real harmonic in the opposite frame is assumed. If this hypothetical harmonic component is assumed in frame $i-1$, i.e., $(0, \omega_y^i)$, and the corresponding real harmonic is located in frame i, i.e., (A_y^i, ω_y^i), then, a new harmonic is smoothly created in transition from frame $i-1$ to frame i by the corresponding sinusoid in (2). If the hypothetical harmonic component is assumed in frame i, i.e., $(0, \omega_x^{i-1})$, and the corresponding real harmonic is located in frame $i-1$, i.e., $(A_x^{i-1}, \omega_x^{i-1})$, then, the real harmonic is smoothly decayed in transition from frame $i-1$ to frame i by the corresponding sinusoid in (2). Therefore, for any harmonic component in frames i or $i-1$, we can define a matched harmonic component in the opposite frame. The ℓ-th amplitude track $a_\ell^i(n)$ and the ℓ-th phase track $\psi_\ell^i(n)$ are derived for a harmonic component pair $(A_x^{i-1}, \omega_x^{i-1})$ and (A_y^i, ω_y^i) in frames $i-1$ and i, respectively, as

$$a_l^i(n) = A_x^{i-1} + \frac{A_y^i - A_x^{i-1}}{N} n \tag{5}$$

$$\psi_\ell^i(n) = \psi_\ell^{i-1}(0) + \sum_{\sigma=1}^{n} \omega_\ell^i(\sigma) \tag{6}$$

where N is the number of speech samples in one frame, $\psi_\ell^{i-1}(0)$ indicates the initial value of the phase track at the update point at frame $i-1$ and $\omega_\ell^i(\sigma)$ is the instantaneous frequency given by

$$\omega_\ell^i(\sigma) = \omega_x^{i-1} + \frac{\omega_y^i - \omega_x^{i-1}}{N} \sigma. \tag{7}$$

The above formulation constitutes the basic principle of our harmonic tracking algorithm. This algorithm smoothly evolves the sinusoidal components from one frame to the next frame and generates an essential level of periodicity in synthetic voiced speech.

4. QUANTIZATION AND BIT ALLOCATION

Peaks of the magnitude spectrum of voiced speech are interpolated by a spline function to form a peak envelope function. This function is, then, represented by a number of LP all-pole filter coefficients. The resulting coefficients are transferred into the line spectral frequency (LSF) domain to obtain a suitable representation for quantization purpose.

We have used ten coefficients to model the peak envelope function. Split vector quantization [8] is used to quantize these coefficients in the LSF domain. The LSF vector representation is split up in three parts with (3,3,4) splitting scheme, i.e., first three line spectral frequencies (LSF's), the following three LSF's (i.e., forth, fifth and sixth) and the last four LSF's. Each part is quantized independently using vector quantization. We have used 9 bits for the first-part codebook, 8 bits for the second-part codebook and 8 bits for the third-part codebook, i.e., a (9,8,8) bit allocation scheme. This results in 25 bits/frame to code the peak envelope information. In order to preserve the ascending order of LSF's after quantization which ensures the stability of the peak-envelope reconstructing filter, only those vectors from the next-part codebook whose first LSF's are greater than the quantized value of the last LSF in the previous part are considered in the next-part codebook search process. The same quantization scheme is used to quantize the spectral envelope information in pure unvoiced speech in CELP algorithm.

The peak envelope function can be modelled more accurately using a higher order all-pole filter. The resulting higher order LP coefficients can be transferred into the LSF domain using the algorithm described in [9].

To quantize the gain factor (G) of the all-pole filter modelling the peak-envelope function of magnitude spectrum of speech, we obtained the distribution of the logarithm of G^2 using 8000 voiced frames. Since the resulting distribution was fairly uniform, we uniformly quantized the log-square of the all-pole model gain factor using 6 bits.

The fundamental frequency is coded using the uniform quantization with 7 bits.

Starting from baseband of 500 Hz, the spectrum is divided into 14 equal frequency bands with 250 Hz bandwidth for each band. The 500 Hz baseband and the next 14 bands are coded using 15 quantization levels of a 4 bit quantizer. The quantization level of zero is used to indicate pure unvoiced frames. Let f_h represent the upper limit of the frequency band where the estimated cut-off frequency falls. Let $\hat{f}_0$ denote the decoded fundamental frequency. Then, the decoded number of harmonics $\hat{n}$ in the harmonic region of spectrum is given by

$$\hat{n} = ceil(\frac{f_h}{\hat{f}_0}) \tag{8}$$

where $ceil(.)$ operator returns the smallest integer not less than $\frac{f_h}{\hat{f}_0}$. In this method of decoding, we decode the number of harmonics in favour of expanding the periodic region of spectrum. In fact, $\hat{n}$ determines the number of sinusoidal components which take part in constructing the periodic part of speech at decoder.

In voiced frames, the periodic component is coded with 42 bits/frame or 2.1 kb/s, assuming 50 Hz frame rate, and the noise-like component is coded using a 2 kb/s CELP excluding LSF coding. In pure unvoiced frames, only a 3.25 kb/s CELP including LSF coding is required. Therefore, the maximum transmission bit rate of the proposed hybrid speech coder is 4.1 kb/s. Notice that, the CELP algorithm operating in sinusoidal mode codes the noise-like part of voiced frames using the LP parameters representing the peak-envelope function of speech spectrum. While the CELP algorithm which codes pure unvoiced frames uses the conventional LPC parameters.

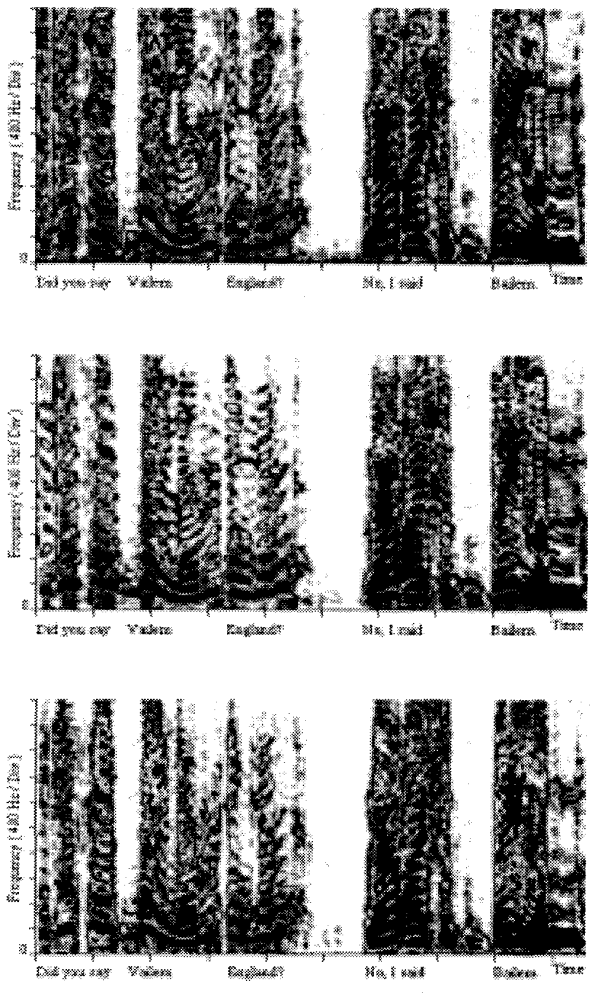

Figure 1: Spectrograms of 4 sec of clean speech; Top: Original, Middle: Synthetic (by the proposed hybrid speech coder), Bottom: Synthetic (by the Motorola half-rate GSM).

5. EXPERIMENTAL RESULTS

The original and synthetic spectrograms of a 4 second dialogue: "Did you say Vailem England?" (female voice) followed by "No, I said Bailem." (male voice) spoken in a noise-free environment are shown in Fig 1. In this Figure, the spectrogram on top belongs to the original speech, the middle spectrogram belongs to the synthetic speech synthesised by the proposed hybrid coder operating at 4.1 kb/s and the bottom spectrogram belongs to the synthetic speech synthesised by the Motorola GSM half-rate speech coder which is a VSELP operating at 5.6 kb/s. To evaluate the performance of the proposed coder in the presence of a severe

background noise, the original speech was corrupted by additive random noise. The top, middle and bottom spectrograms in Fig 2 show the original spectrogram, the synthetic spectrogram synthesised by the proposed hybrid coder at 4.1 kb/s, and the synthetic spectrogram synthesised by the Motorola GSM half-rate speech coder at 5.6 kb/s, respectively. Comparison of the spectrograms in Figures 1 and 2 demonstrates the high performance of the proposed hybrid coder in reconstructing the details of the original speech at low bit rates, in both clean and noisy environments.

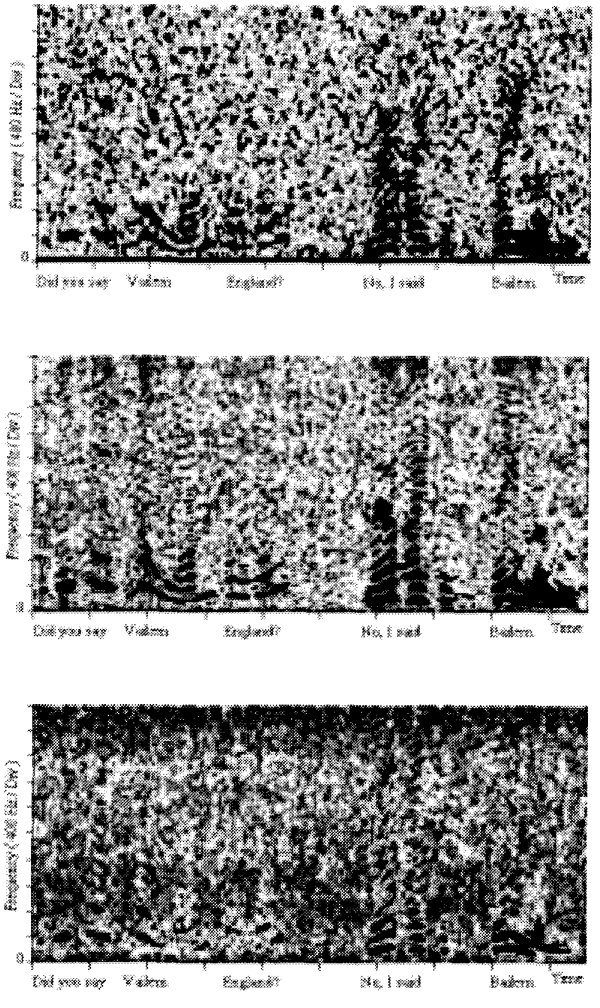

Figure 2: Spectrograms of 4 sec of speech with background noise; Top: Original, Middle: Synthetic (by the proposed hybrid speech coder), Bottom: Synthetic (by the Motorola half-rate GSM).

6. CONCLUSIONS

A hybrid speech coder at 4.1 kb/s was introduced. The proposed coder is operating in two modes, i.e., the harmonic sinusoidal coding mode for voiced part of speech and the CELP coding mode for unvoiced part of speech. In fact, the hybrid coder replaces the long-term prediction part of the CELP algorithm, which is usually based on an adaptive codebook, with a harmonic sinusoidal algorithm. A harmonic tracking algorithm interpolates the sinusoidal parameters in adjacent frames and generates a high degree of periodicity in voiced sounds, as is common in natural voiced speech. A spline interpolating function is fitted to the peaks of magnitude spectrum of speech and then modelled by the coefficients of a LP all-pole filter. These coefficients are quantized in the LSF domain to code the sine-wave amplitude information. Since, the sinusoidal parameters are updated frame by frame, the proposed hybrid coder can operate at low bit rates at the cost of more complex decoder, as compared to the CELP decoder, due to the harmonic tracking algorithm.

7. REFERENCES

[1] I. Gerson and M. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbps," *Proc. ICASSP-90*, pp. 461-464, 1990.

[2] A. Kataoka, T. Moriya, S. Hayashi, "An 8-kb/s Conjugate Structure CELP (CS-CELP) Speech Coder," *in IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 6, pp. 401-411, Nov. 1996.

[3] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Low Bit Rates," *Proc. ICASSP-85*, pp. 937-940, 1985.

[4] W. B. Kleijn, "On the Periodicity of Speech Coded with Linear-Predictive Based Analysis by Synthesis Coders," *in IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 539-542, October 1994.

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

[6] R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Voicing Decision Based upon a Sinusoidal Speech Model," *Proc. ICASSP-90*, vol. 1, pp 249-252, 1990.

[7] M. R. Nakhai, F. A. Marvasti, "Split Band CELP (SB-CELP) Speech Coder," *Proc. ICASSP-99*, Phoenix, Arizona, March 1999.

[8] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 1, pp. 3-14, January 1993.

[9] 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.