A 4 KB/S TOLL QUALITY HARMONIC EXCITATION LINEAR PREDICTIVE SPEECH CODER

Suat Yeldener

COMSAT Laboratories, 22300 Comsat Drive, Clarksburg, Maryland, USA
E-mail: yeldener@ctd.comsat.com

ABSTRACT

The Harmonic Excitation Linear Predictive Speech Coder (HE-LPC) is a technique derived from MBE [1] and MB-LPC [2] type of speech coding algorithms. The HE-LPC coder has the potential of producing high quality speech at 4.8 kb/s and below. This coder employs a new pitch estimation and voicing technique. In addition, new DCT based LPC and residual amplitude quantization techniques have been developed. The 4 kb/s HE-LPC coder with a 14th order LPC filter was found to produce much better speech quality than the various low rate speech coding standards, including 3.6 kb/s INMARSAT Mini-M AMBE vocoder. During formal ITU ACR test [3], the 4 kb/s HE-LPC vocoder was found to produced equivalent performance to 32 kb/s ADPCM and G.729 for both flat and modified IRS filtered clean input speech conditions. The HE-LPC algorithm can also be extended to cover bit rates between 1.2 and 8 kb/s range depending on the application.

1. INTRODUCTION

The most current speech coders operating at bit rates of 6.0 kb/s and below fall into one of two categories: the linear prediction based techniques such as Code Excited Linear Prediction (CELP) [4], Mixed Excited Linear Prediction (MELP) [5] and LPC-10 vocoder [6], and frequency domain techniques such as multi band excitation (MBE) vocoder [1], Sinusoidal Transform Coding (STC) [7] and the channel vocoder [8]. Both CELP and MBE vocoders are capable of producing good quality speech at around 4.8 kb/s. Below 4.8 kb/s however, these coders suffer from distortions introduced by coarse quantization of model parameters due to the limited number of bits. LPC-10 and Channel vocoders, on the other hand, although their model parameters are very efficiently quantizeable at lower bit rates, they suffer from their speech modeling techniques and as a result they produce synthetic and unnatural speech. Other currently popular speech coding algorithms are STC [7] and MELP [5] vocoders. STC produces good quality speech at low bit rates for mainly voiced speech signals, but its extension to model unvoiced or noisy type speech signals fails to produce good speech quality. The MELP vocoder on the other hand was mainly designed to produce high quality speech at around 2.4 kb/s. As a result of this, a 2.4 kb/s MELP vocoder was chosen as the new DoD standard [5].

An alternative speech coding algorithm, termed Harmonic Excitation Linear Predictive Speech Codec (HE-LPC), has the potential of producing good quality speech at very low bit rates (4.8 kb/s and below). This coding scheme uses the advantages of both time domain (LPC based) and frequency domain techniques to improve the speech quality. In this paper, we are reporting on a 4 kb/s HE-LPC coder that provides toll quality speech for clean input speech conditions.

2. HE-LPC SPEECH CODER

The simplified block diagram of the HE-LPC speech coder is shown in Fig. 1.

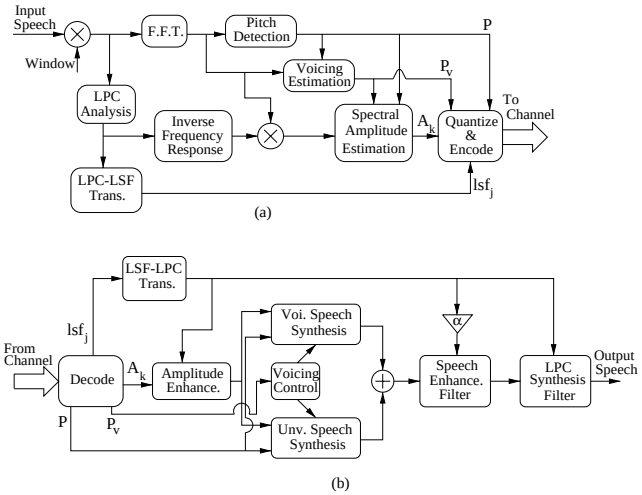

Figure 1: Simplified block diagram of HE-LPC speech coder (a) encoder (b) decoder.

In the HE-LPC coder, the approach to represent the speech signals $s(n)$ is to use the speech production model in which

speech is viewed as the result of passing an excitation, $e(n)$ through a linear time-varying filter (LPC), $h(n)$, that models the resonant characteristics of the speech spectral envelope [9]. The $h(n)$ is represented by 14 LPC coefficients which are quantized in the form of Line Spectral Frequency (LSF) parameters. In the HE-LPC speech coder, the excitation signal $e(n)$ is specified by a fundamental frequency or pitch, its spectral amplitudes, and a voicing probability. The voicing probability defines a cut-off frequency that separates low frequency components as voiced and high frequency components as unvoiced [10]. The techniques for estimating the model parameters will be addressed later in this paper. The computed model parameters are quantized and encoded for transmission.

At the receiving end, the information bits are decoded and hence, the model parameters are recovered. At the decoder, the voiced part of the excitation spectrum is determined as the sum of harmonic sine waves. The harmonic phases of sine waves are predicted using the phase information of the previous frames. For the unvoiced part of the excitation spectrum, a white random noise spectrum, normalized to unvoiced excitation spectral harmonic amplitudes, is used. The voiced and unvoiced excitation signals are then added together to form the overall synthesized excitation signal. The resultant excitation is then shaped by the linear time-varying filter $h(n)$ to form the final synthesized speech. In order to enhance the output speech quality and make it cleaner, a frequency domain post-filter is used [2].

2.1. Pitch Estimation

One of the most prevalent features in speech signals is the periodicity of voiced speech known as pitch. The pitch contribution is very significant in terms of the natural quality of speech. Many pitch estimation algorithms have been developed over the past few decades, however, it still remains one of the most difficult problems in speech processing. As a result of this, we have developed a perception based analysis by synthesis pitch estimation algorithm, that takes advantage of the most important frequency components to synthesize speech and then estimate the pitch based on a mean squared error approach. The block digram of the perception based analysis by synthesis algorithm is shown in Fig. 2. The pitch search range is first partitioned into various sub-ranges, and then a computationally simple pitch cost function is computed. The computed pitch cost function is then evaluated and a pitch candidate for each sub-range is obtained. After pitch candidates are selected, an Analysis By Synthesis error minimization procedure is applied to chose the most optimal pitch estimate. In this case, the LPC residual signal is low pass filtered first. The low pass filtered excitation is then passed through an LPC synthesis filter to obtain the reference speech signal. For each candidate of pitch, the LPC residual spectrum is sampled at the harmonics of the corresponding pitch candidate to get the harmonic amplitudes, and phases. These harmonic components are used to generate a synthetic excitation signal based on the assumption that the speech is purely voiced. This synthetic excitation signal is then passed through the LPC synthesis filter to obtained the synthesized speech signal. The perceptually weighted mean squared error (PWMSE) in between the reference and synthesized signals is then computed. This process is repeated for each candidate of pitch. The candidate pitch period having the least PWMSE is then chosen as the most optimal pitch estimate. This pitch estimation algorithm was found to be very robust for a variety of input speech conditions.

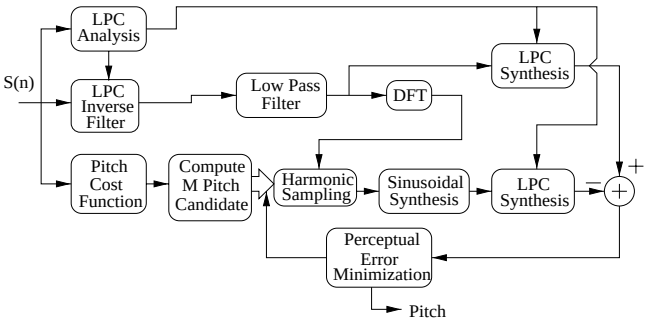

Figure 2: Perception Based Analysis By Synthesis Pitch Estimation Algorithm.

2.2. Voicing Determination

There can be various ways of computing the voicing probability that defines a cut-off frequency [10]. The basic block diagram of the voicing estimation is shown in Fig. 3. Firstly,

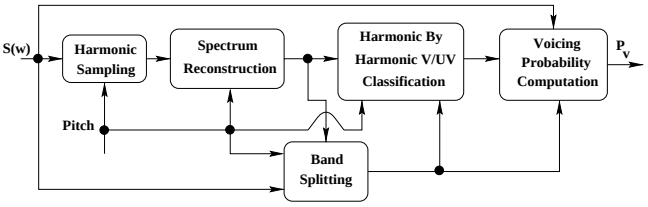

Figure 3: Voicing Probability Computation

a synthetic speech spectrum is computed based on the assumption that the speech signal is fully voiced. The Original and Synthetic speech spectra are then compared and a voicing probability is computed as follows: The original and reconstructed spectra are compared harmonic by harmonic, and each harmonic of the speech spectrum is then declared as either voiced ($V(k) = 1$) or unvoiced ($V(k) = 0, \quad 1 \leq k \leq L$) depending on the magnitude of the error between original and reconstructed spectra for the cor-

responding harmonic. Here, L is the total number of harmonics within 4 kHz speech band. The voicing probability for the entire speech frame is then computed as the energy ratio between voiced and all harmonics within the 4 kHz speech band as:

$$P_v = \sqrt{\frac{\sum_{k=1}^{L} V(k)A(k)^2}{\sum_{k=1}^{L} A(k)^2}} \quad (1)$$

where $V(k)$ and $A(k)$ are the binary voicing decision and spectral amplitudes respectively for the k^{th} harmonic. The computation of voicing probability with this way was found to improve the naturalness of synthesized speech signals compared to the voicing technique reported in [10].

3. 4 KB/S CODER CONFIGURATION

For operation at 4 kb/s, a frame length of 20 ms (160 samples at 8 kHz sampling rate) is used. Each frame is divided into 2 sub-frames each having a length of 10 ms. Therefore, 80 bits/frame are available for coding the model parameters at 4 kb/s. These bits are allocated for each parameter as tabulated in Table 1.

Parameters	*No of Bits/Frame*	*Bit Rate (b/s)*
Pitch	5+7	600
14 LSF Coef's.	3+40	2150
Spectral Amplitudes	3+19	1100
Voicing Info.	0+3	150
Total	80	4000

Table 1: Bit allocation for 4 kb/s HE-LPC vocoder.

The pitch period for the second sub-frame is directly quantized using 7 bits and the pitch period for the first sub-frame is differentially quantized using 5 bits. Voicing information for the second sub-frame is quantized using 3 bits and the voicing for the first sub-frame is recovered at the decoder by linear interpolating voicing information for the adjacent sub-frames. The 14 LSF coefficients for the second sub-frame are split vector quantized in the LOG and DCT domains. For this purpose, the LSF coefficients are split as $\{3,3,4,4\}$, and the coefficients for each split are then transformed into first the LOG and then the DCT domain. The DCT coefficients for each split are then vector quantized using 10 bits each ($4 \times 10 = 40$ bits). During the vector quantization of DCT coefficients, a well known weighting is used that gives more emphasis on low order DCT coefficients, since the low order DCT coefficients are more important than the higher ones. The LSF coefficients for the first sub-frame are quantized using the concept of optimal linear interpolation as reported in [2]. The index for the best linear interpolated LSF coefficients, which minimizes the mean square error between the original and linear interpolated LSF coefficients, is then coded and transmitted using 3 bits [2]. The residual spectral amplitudes for the first sub-frame are quantized in a similar way to the LSF coefficients of the first sub-frame as described above again using 3 bits to code the optimal interpolation index. The residual gain for the second sub-frame is quantized using 5 bits, and the shape of the residual spectral harmonic amplitudes are split into odd and even harmonic amplitude vectors. The shape of the odd and even harmonic amplitude vectors are then converted into the DCT domain. The DCT coefficients for the odd harmonic amplitude vector are then vector quantized using 8 bits, and the error vector in between the quantized odd and original even harmonic amplitude vectors are then vector quantized using 6 bits only. Since the vector quantization for spectral amplitudes are done in the DCT domain, a weighting is used that gives more emphasis to the low order DCT coefficients than the higher order ones.

4. SUBJECTIVE LISTENING TESTS

Subjective listening tests were used to compare a number of speech sentences processed by the 4 kb/s HE-LPC Coder and various other standard coders (ITU G726 (32 kb/s ADPCM), ITU G729 (8 kb/s CS-ACELP), 8 kb/s IS-54 VSELP, and 3.6 kb/s INMARSAT Mini-M AMBE vocoder). For speech quality assessment, ITU ACR and/or Mean Opinion Score (MOS) tests were used [3]. The speech signals used as the input to all coders were subject to the same analog conditions. In the ACR test, the performance with input level variation (high (-16 dB), nominal (-26 dB) and low (-36 dB) levels), tandem codecs (2 tandems in both 4 kb/s and G.729 coders and 4 tandems in the G.726 coder), random bit error rate (BER) of 0.1% and random frame erasure rate (FER) of 3% was assessed for modified-IRS speech signals. A total of 10 different sentence pairs for each of two male and two female talkers were processed for the 20 test conditions as defined in [3]. Source speech was selected from the NTT CD-ROM speech database. A total of 24 non-expert listeners were arranged in six groups of four listeners who used handsets for monophonic listening at -15 dBpa. The randomization sequences were generated by COMSAT Laboratories. The summary of the ACR MOS test results are given in Table 2.

Coder	**Test Conditions**					
	-16	*-26*	*-36*	*Tandem*	*Fer*	*Ber*
G.726	3.50	3.54	3.34	2.76	-	-
G.729	-	3.66	-	3.36	3.35	3.61
4 kb/s Coder	3.38	3.41	3.36	2.42	3.25	2.51

Table 2: ACR MOS Test Results

It can be seen that the 4 kb/s HE-LPC coder produces

similar performance to G.726 and G.729, and hence, passes the ITU 4 kb/s standard performance requirement in error-free conditions with input level variation, and in the presence of random frame erasures (FER). The codec fails only the tandem requirements, and its random bit error performance is worse than that of G729 under the same error condition. The 4 kb/s coder performance was also assessed using the 7-point scale CCR method defined in [3] for three different types of background noise (30 dB Babble, 20 dB Interfering Talkers and 15 dB Car noise) and one instance of tandem for modified-IRS weighted speech and noise. In this test, although the 4 kb/s HE-LPC coder failed to pass the ITU requirements for all background noise conditions, in fact, the performance level, for these conditions, was outside the critical distance measure between the reference and 4 kb/s coders.

Another subjective test was also done in COMSAT Laboratories, using the 4 kb/s HE-LPC coder, 8 kb/s IS-54 VSELP [11], G729 and 3.6 kb/s INMARSAT Mini-M AMBE vocoder. For this test, only flat weighted -26 dB input level speech conditions were used. The results for this test are summarized in Table 3.

Coder	*MOS Score*
4 kb/s Coder	3.68
8 kb/s IS-54 VSELP	3.66
G.729	3.40
3.6 kb/s Mini-M	3.35

Table 3: MOS Scores for various coders

From these test results, it was very clear that the 4 kb/s HE-LPC coder produced similar performance to the 8 kb/s IS-54 VSELP coder, and better than ITU G.729 and 3.6 kb/s INMARSAT Mini-M AMBE coders. Both formal and informal listening tests indicated that the HE-LPC vocoder produces almost toll quality speech at 4 kb/s. Another advantage of the HE-LPC speech coding algorithm is that it can be extended to cover bit rates between 1.2 and 8 kb/s depending on the application.

5. CONCLUSIONS

In this paper, HE-LPC speech coder operating at 4 kb/s was presented. New robust techniques for pitch estimation based on the perception based analysis by synthesis concept and voicing probability determination of speech signals were also described. The DCT based techniques were used to quantize both 14 LSF coefficients and excitation spectral amplitudes. Both formal and informal subjective listening tests were conducted and the results indicate that the 4 kb/s HE-LPC speech coder produces toll quality speech that is equivalent to 32 kb/s ADPCM performance under clean input speech conditions. Since the HE-LPC model parameters are more efficiently quantizeable at low bit rates (4 kb/s and below) than other coding systems such as CELP, HE-LPC is a very promising low bit rate speech coding technique.

6. REFERENCES

[1] D. W. Griffin, J. S. Lim "Multi-Band Excitation Vocoder" IEEE Trans. ASSP, 1988, Vol:36 No:8 pp:664-678.

[2] S. Yeldener, A. M. Kondoz, B. G. Evans "Multi-Band Linear Predictive speech coding at very low bit rates" IEE Proc. Vis. Image and Signal Processing, October 1994, Vol:141 No:5 pp:289-295.

[3] ITU-T SG 16 "Subjective Qualification Test Plan for the ITU-T 4 kb/s Speech Coding Algorithm", Version 2.2, September 1998.

[4] M. Schroeder, B. Atal "Code Excited Linear Prediction: High Quality Speech at Low Bit Rates" Proc. ICASSP, 1985, pp:937-940.

[5] A.V. McCree et al., "A 2.4 Kb/s MELP Coder Candidate for The New U.S. Federal Standard," Proc. ICASSP, 1996, p. 200–203.

[6] T. Tremain "The Government Standard Linear Predictive Coding Algorithm (LPC-10)" Speech Technology, 1982, Vol:1 pp:40-49.

[7] R. J. McAulay, T. F. Quatieri "Speech Analysis/Synthesis Based On a Sinusoidal Representation" IEEE Trans. ASSP, 1986, Vol:34 pp:744-754.

[8] J. N. Holmes "The JSRU Channel Vocoder" IEE Proc., 1980, Vol:127 pp:53-60.

[9] S. Yeldener, A. M. Kondoz, B. G. Evans "Sine Wave Excited Linear Predictive Coding of Speech" Proc. Int. Conf. on Spoken Language Processing, Kobe, Japan, November 1990, pp. 4.2.1 - 4.2.4.

[10] S. Yeldener, A.M. Kondoz, B.G. Evans, "A High Quality Speech Coding Algorithm Suitable for Future Inmarsat Systems," Proc. 7th European Signal Processing Conf. (EUSIPCO-94), Edinburgh, September 1994, p. 407–410.

[11] EIA/TIA - IS-54 North American Digital Mobile Radio Specification for 8 kb/s VSELP Speech Coder, 1989.