Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 49(63), Fascicola 1, 2004

Excitation modeling in CELP speech coders

Cornel Balint[1]

Abstract – The introduction of CELP (Codebook Excited Linear Prediction) speech coding provided an efficient way to compress speech data to 4.8 kbps with high quality, but with the disadvantage of great computational complexity required for real-time processing. In this paper, we investigate the overall computing complexity of the CELP speech coders and we propose some particular structure of codebook involved in residual speech coding, in order to obtain a significant reduction of the complexity.

Keywords: CELP, residual coding, codebook, vector quantization.

I. INTRODUCTION

During the last several years, there has been a spectacular growth of digital services, such as digital wireless and wire communications, satellite communications and digital voice storage systems that covers a vide range of applications, including Internet and multimedia. . Such services require the use of high-quality low bit-rate coders to efficiently code the speech signal before transmission or storage.

The majority of such coders employ algorithms that are based on Code-Excited Linear Prediction (CELP) [1], [2]. Numerous CELP based coders have been standardized for various applications and are widely used in the commercial world.

This family of CELP techniques exploits the models of human speech production and human auditory perception which provide a high quality at relative low bit-rate, which outperforms most existing compression techniques at these rates.

Speech can be classified into two general categories:

- Voiced speech characterized by quasi-periodic and in general high energy segments of sounds such as vowels.
- Unvoiced speech which generally describes the low energy segments such as consonants.

Voiced speech is produced when the air flow is interrupted by a periodic opening and closing of the vocal cords, generating a periodic glottal excitation for the vocal tract.

Unvoiced speech is produced when the vocal cords do not vibrate and the vocal tract is excited by a turbulent noise generated when the air passes through a narrow constriction in the vocal tract.

Speech production can be alternatively viewed as a filtering operation in which a sound source excites a vocal tract filter [6]. The sound source represents the noise generated at a constriction of the vocal tract during unvoiced sounds or the glottal pulses during voicing, or a combination of these two. The spectrum of the sound source during voiced sounds contains harmonics spaced by fundamental frequency with most of the energy concentrated at low frequencies, whereas during unvoiced sounds the spectrum is approximately flat and without harmonic structure.

The vocal tract will finally modify the distribution of energy in the spectrum of the sound source. Representing the vocal tract as a time-varying filter, the resonances and anti-resonances are due to the poles and zeros of the vocal tract frequency response. Low bit-rate coders try to reduce the bit rate, while preserving speech quality, by taking advantage of redundancies in the speech signal and perceptual limitations of the human ear [6]. The former arises from the following observations: (a) in general the speech spectrum changes relatively slowly (except during the articulation of stops), (b) successive pitch periods are generally similar and (c) the spectral envelope is relatively smooth, with most of the energy concentrated at low frequencies. These are attributed to the mechanical limitations of the speech organs, i.e. vocal tract and vocal cords. The redundancy in the speech signal led to the conclusion that speech samples are correlated. The spectral envelope corresponds to the short-term correlations and the harmonic structure corresponds to the long-term correlations. These correlations can be exploited to yield a lower bit rate by using linear prediction.

II. CELP CODERS

Linear prediction is one of the most important tools in speech analysis. Its relative simplicity of computation and its ability to provide accurate estimates of the speech parameters, make this method predominant in low bit-rate coding of speech [6].

[1] Department of Communications, University "Politehnica" of Timisoara, Blvd. V. Parvan 2, 300223 Timisoara, Romania, Phone +40256-403310, E-mail: cbalint@etc.utt.ro

The idea of linear prediction is that a speech sample can be approximated as a linear combination of past samples. Then, by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones over a finite interval, a unique set of predictor coefficients can be determined.

Prediction can be used to either remove redundancies from the speech signal, or to create a model for the vocal tract. The redundancy removal is performed with a linear prediction filter.

Figure 1 show the CELP coder based on analysis-by-synthesis principle and speech production mechanism that can be viewed as a filtering operation in which a sound source excites a vocal tract filter

The LPC analysis filter removes the formant structure of the speech signal and the result is the output prediction error called LPC residual or excitation signal. Transfer function of LPC analysis filter is:

$$A(z) = 1 + \sum_{i=1}^{P} a_i z^{-i} \tag{1}$$

The inverse LPC analysis filter, i.e. the LPC synthesis filter, models the vocal tract and its transfer function describes the spectral envelope of the speech signal:

$$H(z) = \frac{1}{A(z)} \tag{2}$$

Improvements can be obtained by considering the long-term correlations of voiced speech using long-term prediction. In this case another LP filter can be used to remove far-sample redundancies:

$$P_a(z) = 1 - \beta z^{-L} \tag{3}$$

This filter is usually called the pitch predictor and exploits the periodicity of the signal. The inverse of the pitch predictor, often called the pitch filter, models the effect of the glottis and its transfer function describes the harmonic structure of the speech signal. Pitch prediction will have no useful effect for unvoiced speech since the unvoiced excitation is random and its spectrum is flat.

Pitch resolution is very important, especially for high pitched speakers. However, the pitch resolution is bounded by the sampling rate (typical 8 kHz). Increasing the pitch resolution without increasing the sampling rate is possible using an artificial increasing of internal sampling rate by interpolation.

Perceptual criteria are introduced in fig.1 by the perceptual weighting filter with transfer function:

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{k=1}^{P} a_k z^{-k}}{1 - \sum_{k=1}^{P} a_k \gamma^k z^{-k}} \tag{4}$$

were $0 < \gamma < 1$.

Perceptual weighting is based on the observation that in spectrum where signal levels are high, the noise is masked by signal and a little contribution to the audible distortion than where signal levels are low. This suggests that we can weigh the noise according to speech spectrum in order to obtain the best perceptual results.

Perceptual criteria yield to a higher computation complexity, due to filtering operation. Some simplifications can be obtained by rearranging the coder structure, in order to dispose the weighting filter to weighing the residual before LPC synthesis filter and compare the result of synthesis filter with the weighted speech signal.

The speech residual signal, after short and long term predictions filter, are Gaussian distributed, so a vector quantization [4] using a stochastic codebooks, generated by a Gaussian process can be used to predict speech residuals. Due to extensive codebook search involved, vector quantization of speech residual signal request a great computational complexity [5].

However, since the stochastic codebooks are generated randomly, there are no special structures to organize them, so we need to use exhaustive search to find the optimum codebook vector. Although some CELP coders, like FS1016 or G729, use overlapped codebook to reduce the complexity of convolution in perceptual weighting by end-point correction techniques, the computational complexity is still high and, furthermore, the use of overlapped codebook for coding the speech residual is an approximation that degrades the speech quality.

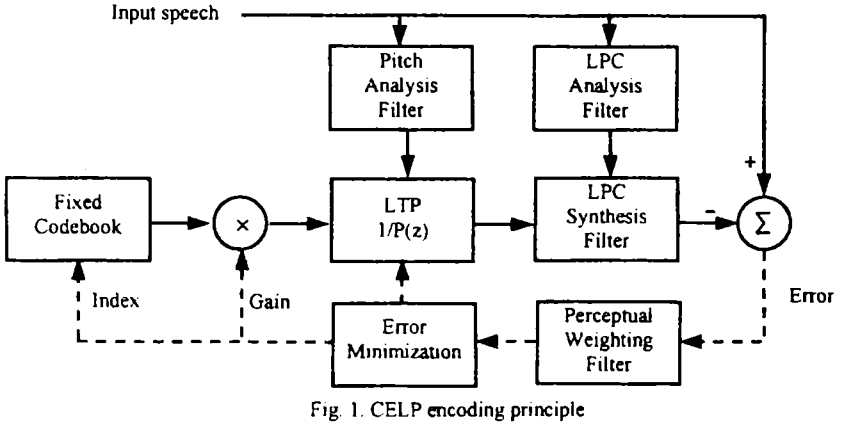

Fig. 1. CELP encoding principle

II. COMPLEXITY COMPUTATION IN CELP

In order to analyze the computational complexity of the CELP coder, table 1 present briefly some important data about a typical CELP coders, related to bit allocation and computational complexity. According to fig. 1 and table 1, CELP coders involves 3 categories of operations: LPC analysis for spectrum coding, pitch coding and residual coding.

Table 1 CELP bit allocation and complexity

	Spectrum	Pitch	Codebook
Update rate	30 ms 240 samples	30/4 = 7.5 ms 60 samples	30/4 = 7.5 ms 60 samples
Order	LPC 10	256 delay 1 gain	512 vectors 1 gain
Analysis	Open loop Correlation Hamming window	Closed loop VQ Delay: 20 - 147	Closed loop VQ 512 vectors
Bits/frame	34 bits / 10 LSF [3444433333]	Index 8/6/8/6 gain: (-1, 2) 5/subframe	Index 9/subframe gain 5/subframe
Bit rate	1133.3 bps	1600 bps	1866.67 bps
Computational complexity	0.08 MIPS	4.5 MIPS	87.7 MIPS
Total bit rate			4600 bps
Total computational compexity			92.28 MIPS

According to fig. 1 CELP analysis consist of 3 steps:

a) short term prediction that extracts the spectrum (envelope) information of speech signal. The result of LPC analysis is an all-zero predictor filter or a corresponding all-pole synthesis filter. The parameters of this filter can be transmitted directly as LPC coefficients or in equivalent forms like reflexion coefficients or line spectrum pairs (LSP). CELP analysis involves:

- windowing the input speech signal
- computing the LPC parameters
- convert LPC parameters to alternative parameters to be transmitted (if necessary).

According table 1, the complexity of LPC analysis is negligible compared to pitch search and codebook search.

b) pitch search involves more complexity, but some observation can reduce the computing complexity. In the proposed implementation we do not search the whole range of delay at once, using a two stage search strategy. First search integer delay and find the best integer delay. Second, fine tune this integer delay searching its neighboring fractional delay up and down to first integer delay. In this way, the complexity for first stage search is reduced according to the number of integer delay (typically 128) and computational complexity for second stage fractional delay (max 6 fractional delays) is negligible compared to integer delay.

Because the pitch codebook is overlapped, each vector is just a shift of the previous vector and contain only 1 new element and the end point correction technique can be used to reduce the amount of operations.

Suppose the first codebook vector is $\{c(0), c(1), \ldots c(59)\}$, perceptual weighting impulse response is $\{h(0), h(1), \ldots h(9)\}$ and vector after perceptual weighting is $\{y_0(0), y_0(1), \ldots, y_0(59)\}$. Then the convolution can be computed using recurrence :

$$
\begin{aligned}
&y_0(0) = h(0) * c(0) \\
&y_0(1) = y_1(0) + h(1) * c(0) \\
&y_0(2) = y_1(1) + h(2) * c(0) \\
&y_0(3) = y_1(2) + h(3) * c(0) \\
&\ldots\ldots\ldots\ldots \\
&y_0(9) = y_1(8) + h(9) * c(0) \\
&y_0(10) = y_1(9) \\
&\ldots\ldots\ldots \\
&y_0(59) = y_1(58)
\end{aligned}
\qquad (5)
$$

when $y_1(i)$ denote the next response vector. In this way, the iteration start with computing $\{y_{127}(i)\}$, then $\{y_{126}(i)\}$ to $\{y_0(i)\}$.

In this way, the complexity for pitch search is:

Convolution: for the first vector it needs 1+2+...+10+10+...+10 (60 terms) = 555 MUL and 1+2+...+9+9+...+9 (59 terms) = 495 ADD and for each of the following vector needs only 9 MUL and 9 ADD. Total need is 555+495+18*127=3336 operations.

Correlation and energy computation need the same operation: 60 MUL and 59 ADD for each vector, resulting (60+59)*128 = 15230 operations for correlation and 15230 operations for energy computing.

c) codebook search

After short and log term prediction that extract the spectrum information and pitch information from the speech signal, result the speech residual. Although the residual is noise like sequence, the CELP coders encode this residual using a vector coding and a noise like codebook. Most of CELP computational complexity is attributed to codebook search for residual coding represented in figure 2.

The coding goal is to find the optimum codeword that minimize the least square error between the current subframe speech signal s and the estimate $\hat{s}_k$ obtained from codebook c^k after gain adjust and perceptual weighting [3]:

$$LSE(k) = \| s - \hat{s}_k \| \qquad (6)$$

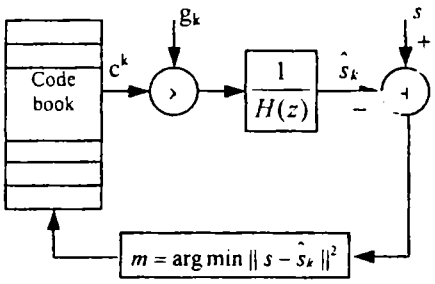

Fig. 2 Codebook search

or equivalent, maximizes the performance cost function C:

$$C(k) = \frac{\langle s, Hc^k \rangle^2}{\| Hc^k \|^2} \tag{7}$$

where notation $\langle a, b \rangle$ denote

$$\langle a, b \rangle = \sum_{n=0}^{N-1} a_n b_n \tag{8}$$

Convolution between codeword and perceptual weighting impulse response need:

For each vector: 1+2+3+...+10+10+...+10 (60 terms) = 555 MUL and 1+2+3+...+9+9+...+9 (59 terms) = 495 ADD total for 512 vector: (555+495)*512 = 537600

Correlation: For each vector: 60 MUL + 59 ADD

Total: (60+59)*512 = 60930

Energy: for each vector: 60 MUL + 59 ADD

Total: (60+59) *512 = 60930

These operations must be performed for each subframe of 7.5 ms, namely 1000/7.5 = 133.333 times per second, which result in a total complexity of (537600+60930+60930)*133.333 = 87.7MIPS, that is unacceptable for real time implementation.

III. DETERMINISTIC CODEBOOK

In order to reduce the computational complexity, a deterministic codebook is used. Considering the speech residual vector $\vec{r}$, the coder have to find a codebook vector $\vec{x}$, that after scaling with the gain value g will produce the minimum square error from the speech residual $\vec{r}$, as illustrated in figure 3.

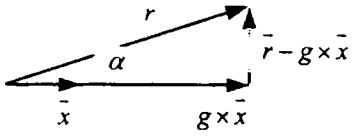

Fig. 3. Residual and codebook vector

Due to the gain scaling factor g, the minimization criterion is not the same as nearest neighbor in the Euclidian distance sense. The error minimization criterion is equivalent to maximizing:

$$\frac{|\vec{r} \cdot \vec{x}|^2}{|\vec{x}|^2} = \frac{|\vec{r}|^2 |\vec{x}|^2 \cos^2 \alpha}{|\vec{x}|^2} = |\vec{r}|^2 \cos^2 \alpha \tag{9}$$

or, equivalent, to maximizing $\cos^2 \alpha$, since $\vec{r}$ is fixed. To maximize $\cos \alpha$ means to find the codebook vector that is parallel to the speech residual vector. A good deterministic codebook in that sense is that codebook whose code vectors spans the n-dimensional hypersphere as uniformly as possible.

An algebraic ternary sparse codebook was used. The vector length is n = 60 and 48 component are zeros and the remaining are +1 or –1, according to figure 4.

Fig. 4 Sample of codebook vector

IV EXPERIMENTAL RESULTS

We have tested two different codebook. The first was derived from the above codebook, imposing some restriction to code vector, in order to reduce the codebook dimension at 2^9, as used in most CELP standard.

For the second codebook, we propose codebook of same dimensions (2^9), obtained by training the initial uniform distributed codebook of 2^{12} dimension. Both codebooks are the nonzero position uniformly distributed over the 60 possible positions. Only the elements with index 5n are nonzero, as in figure 4. This particular structure has an essential significance regarding the perceptual weighting and inner product computing. Because all code vector are different combination of signs +1 and -1, the inner product is simply to compute. The complexity result to be 0.16 MIPS. For trained codebook, the resulting complexity is about 1,2 MIPS, representing an great improvement compared with brute force search in an unstructured codebook.

To validate the proposed improvements of CELP coder by subjective comparative listening tests, a Matlab implementation was performed.

REFERENCES

[1] B. S. Atal, M. R. Schroeder, "Code Excited Linear Prediction (CELP): High quality speech at very low bit rates", *IEEE Int. Conf. Acoustics, Speech and Signal Proc.*, 1985.

[2] R. C. Rose, T. P. Barnwell, "Design and performance of an analysis-by-synthesis class of predictive speech coders", *IEEE Trans on ASSP, vol. ASSP-38, no.9*, sept. 1990.

[3] M. Mauc, G. Baudoin, M. Jelinek, "Complexity reduction for FS 1016 at 4800bps CELP coder", *Eurospeech, 93, Berlin*, sept. 1993.

[4] A. Buzo, A. H. Gray, R. M. Gray, J. D. Markel, "Speech coding based upon vector quantization", *IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28*, Oct. 1980.

[5] A. Gersho, R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1993.

[6] B. S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*, Kluwer Academic Publishers, 1991.