Proc. Biennial Symp. Commun. (Kingston, ON), pp. 25-28, May 1992

Wideband CELP speech coding at 12 kbits/sec[†]

K. Abboud[1] and P. Kabal[1,2]

[1]Electrical Engineering, McGill University, Montreal, Quebec, H3A 2A7
[2]INRS-Télécommunications, Université du Québec Verdun, Quebec, H3E 1H6

Abstract

This paper investigates the use of CELP (Code Excited Linear Prediction) as a coding scheme for wideband speech at an operating bit rate of 12 kbits/sec. With the help of different parameter coding techniques, the bit rate was lowered from 16 kbits/sec [2] to 12 kbits/sec while maintaining a similar speech quality. Three encoding schemes were used to improve the performance of the wideband CELP coder. The first approach used a combination of a three way split vector quantization and a new weighted distance measure for a set of line spectral frequencies (LSFs). The second approach used fractional pitch delays to improve the coder's performance for high pitched sounds. The third approach used perceptual noise weighting to improve coding in the high frequency region. The combination of all these three schemes resulted in a substantial increase in speech quality at a lower bit rate (12 kbits/sec).

1 Introduction

In recent years, the need for high quality speech coders (e.g. CELP) has been growing fast specially for applications such as teleconferencing, videophones, high quality voice-mail services and wideband telephone intended for the ISDN service. With a bandwidth of 50-7000 Hz corresponding to wideband speech, the bottleneck of a bandwidth limitation of 0.2-3.2 KHz is eliminated and a substantial increase in perceived quality is observed. The added low frequencies increase the voice naturalness and closeness while high frequencies make the speech sound sharper and more intelligible, specially in fricative sounds.

In recent years, different wideband CELP schemes were introduced as alternatives for the CCITT G.722 standard with bit rates almost similar to those found in high quality narrowband telephony systems: a 32 kbits/s low-delay CELP was introduced by Ordentlich and Shoham [1], as well as a 16 kbits/s CELP proposed by Roy and Kabal [2].

In this paper, the basic wideband CELP structure is first discussed. Then, the application of vector quantization on wideband LSFs is studied, followed by the impact of fractional pitch predictors and perceptual noise weighting on the coder's performance. Finally, the structure of the improved CELP with a complete parameter coding list is presented.

[†] This work was supported by FCAR.

2 Basic CELP coding

CELP coding is the result of combined features of both *waveform* and *analysis-by-synthesis* coding. The reconstruction of the input speech signal involves the use of a pitch synthesis filter, a formant filter and a residual codebook. An excitation waveform $\hat{r}_i(n)$ is first selected from the codebook as shown in Figure 1 and then goes through a cascade of two filters to give an initial reconstructed speech signal, the operation is repeated until the best match to the original signal is determined. This operation is divided into two stages.

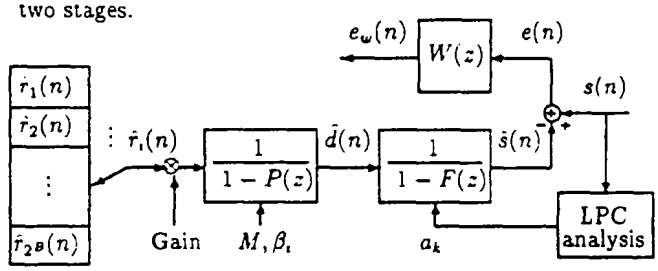

Fig. 1 Basic CELP coder.

During the analysis stage, the input speech is first divided into equal length blocks of samples or frames (e.g. 20ms). The input speech frame $s(n)$ is then passed through an *inverse formant* filter $A(z) = 1 - F(z) = 1 - \sum a_i^{-k}$ where the a_k's are the LPC coefficients, the resulting error signal $d(n)$ is:

$$d(n) = s(n) - \sum_{k=1}^{N_p} a_k s(n-k) \qquad (1)$$

where N_p is the number of LPC coefficients. These coefficients are determined by minimizing in the mean square sense (MS) the error signal $d(n)$ over the analysis frame.

During the synthesis stage, the reconstructed speech frame is generated using pre-determined synthesis parameters on speech sub-frames (i.e. excitation waveforms, gain values, lag values, pitch and LPC coefficients). Periodic components are added during voiced speech to the excitation waveform after its passage through the pitch synthesis

filter

$$G(z) = \frac{1}{1 - P(z)} = \frac{1}{1 - \beta z^{-M}} \quad (2)$$

where β is the pitch coefficient for 1-tap pitch predictor and M is the pitch lag.

The formant resonances are then added to the resulting signal $\hat{d}_n$ after the formant synthesis filter $H(z)$

$$H(z) = \frac{1}{1 - F(z)} = \frac{1}{1 - \sum_{k=1}^{N_p} a_k z^{-k}} \quad (3)$$

to obtain the initial synthesized frame of speech. This speech is then subtracted from the original speech and the result is weighted with $W(z)$ so as to cover the coding noise by the formant regions.

$$W(z) = \frac{H(\gamma z)}{H(z)} = \frac{1 - F(z)}{1 - F(\gamma z)} \text{ where } \gamma = 1/0.75 \quad (4)$$

Finally, this weighted difference $e_w(n)$ is minimized in a MS sense with the adequate combination of the gain G, the pitch lag M, the pitch coefficients β and the excitation waveform $\hat{r}_i(n)$.

3 Quantization of LPC parameters

In the past, two basic approaches for the quantization of LPC coefficients were taken into consideration. The first, *scalar quantization*, quantized each LPC coefficient individually, while the second, *vector quantization*, quantized all the LPC coefficient as a group. The first suffered from a high number of bits required for quantization while the second faced the misfortune of being highly complex in terms of memory and number of computations. Nevertheless, a new method for LPC quantization was introduced recently by Paliwal and Atal [5], it uses a split vector quantization technique and could substantially reduce the complexity of vector quantization.

In the present paper, a three way split vector quantization on wideband LSF parameters is used. The reference LPCs are first transformed into LSFs and then divided into three subgroups. A training data of LSF vectors is used to construct different codebook sets with varying levels of complexity (e.g. 30–33 bits used). This operation is performed with the use of the LBG algorithm [6].

A new weighted Euclidean LSF distance measure is also introduced. For a given reference LSF vector $\vec{v}_{ref}$, this measure determines the best matching spectral envelope $\vec{v}_{cod}$ from a vector quantization codebook:

$$d(\vec{v}_{ref}, \vec{v}_{cod}) = \sum_{k=1}^{N_{lsf}} \omega_k (f_k - \hat{f}_k)^2 \quad (5)$$

where f_k and $\hat{f}_k$ are the k-th LSFs in the reference and codebook vector, respectively, while ω_k is the k-th LSF weighting factor that considers both the frequency sensitivity and distance between LSFs:

$$\omega_k = \omega_k^{(i)} \omega_k^{(ii)} \quad (6)$$

The first weighting factor $\omega_k^{(i)}$ models the hearing sensitivity to frequency differences curve as shown in Figure 2. Specific weights are assigned to the LSFs according to their position in the frequency spectrum.

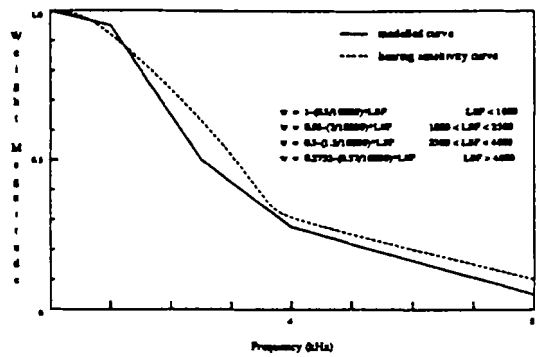

Fig. 2 Human and modeled hearing sensitivity to discriminating frequency differences.

The second weighting factor $\omega_k^{(ii)}$ refers to the distance between LSFs. The closer they are together, the more likely they are to fall near a formant.

$$\omega_k^{(ii)} = 0.05 + [1 - \frac{d_k}{d_{max}}]^2 \quad (7)$$

Table 1 shows the performance of this new weighting scheme in terms of spectral distortion (dB). The simulation uses 4800 LSF vectors for training and 1500 LSF vectors for testing. Both a 16-th and 20-th order LPC analysis are used with a frame update of 15.62 ms. The reference and training speech data have a sampling frequency of 16 kHz and a full spectrum information up to the Nyquist rate (8 kHz).

Order	Splits (number of LSFs and bits used)						SD
	Part 1	Bits	Part 2	Bits	Part 3	Bits	(dB.)
16-th	8	13	4	11	4	9	0.921
	8	13	4	9	4	8	0.996
	4	11	4	11	8	11	0.840
	4	10	4	10	8	10	0.934
	4	13	6	12	6	8	0.870
	4	13	6	10	6	7	0.960
20-th	10	13	5	11	5	9	1.052
	10	13	5	9	5	8	1.055
	5	11	5	11	10	11	0.933
	5	10	5	10	10	10	0.994

Table 1 Spectral distortion (SD) measures.

A combination of three and two LSF configurations with different bit assignment (total bits of 30) are used for a 16-th order LPC and a 20-th order LPC, respectively. The best candidates turned out to be the first (8-4-4) and fourth (4-4-8) entries.

4 Fractional pitch delays

The use of fractional pitch delays has proved to be a very efficient method to represent signal periodicity in CELP coders [3, 4], but so far studies were only made on narrowband speech. Considering the fact that the periodicity of a wideband signal is almost nonexistent in the 4-8 kHz band and that doubling the sampling frequency to 16 kHz meant improved resolution, the need for fractional delays in wideband speech is reduced.

In this paper, we investigate the actual impact of fractional pitch delays on wideband speech. In fact, the use of non-integer delays could be more beneficial in terms of lower bit rates ($\sim$ 10 bits/sub-frame) when compared to a multiple tap integer delay predictor ($\sim$ 11 bits/sub-frame for 3 pitch taps). High temporal resolution for pitch delays can be achieved by specifying the delay as an integer number of samples plus a fraction of a sample $\frac{l}{D}$ where $l = 0, 1, ..., D-1$, and l and D are integers.

The pitch delay in wideband speech ranges from $M = 40$ to $M = 320$ samples with some delays occurring more often than others, therefore it would be beneficial to assign finer resolution to these delays while leaving the others at a lower resolution level. With the use of interpolation and polyphase filters as described in [7, Section 6.3], fractional delays can be efficiently implemented for a first order pitch predictor.

The polyphase filters $p_l(n)$ can directly implement the operations of sampling rate increase and low-pass filtering. For each value of the delay l/D, a corresponding l-th polyphase filter branch is used. With a delay I for the low-pass filter, the expression for the new pitch predictor with a fractional delay of $M + l/D$ is:

$$P(z) = 1 - \beta \sum_{n=0}^{b-1} p_l(n) z^{-(M-I+n)} \qquad (8)$$

where b is the number of coefficients of the polyphase filter and β is the pitch predictor coefficient.

With the use of 38400 pitch sub-frames of 3.125 ms each, a pitch delay distribution is generated as shown in Figure 3.

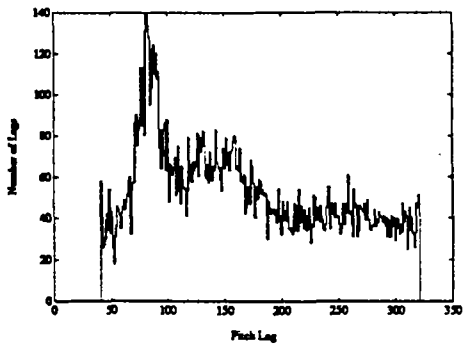

Fig. 3 Distribution of pitch delays.

A nonuniform distribution of non-integer delays can then be set up to construct the pitch delay codebook. Two configurations are set up accordingly with two levels of complexity giving higher resolution to frequently occurring pitch delays as shown in Table 2.

Coder	Pitch Range	Resolution
I	40-70	1/3
	71-100	1/4
	101-140	1/3
	141-320	1
II	40-70	1/4
	71-100	1/6
	101-197	1/4
	198-302	1/3
	302-320	1

Table 2 Configuration of the pitch delay codebook.

As shown in Table 2, the highest resolution is given to pitch lags in the range of 71-100 while the lowest resolution is given to the end of the lag range. In order to test these two configurations, eight speech files (4 female and 4 male speakers), a formant frame of 15.62 ms, a pitch sub-frame of 3.125 ms, a Gaussian codebook with 1024 codewords were used. The quantization of both the gain and LPC parameters was turned off.

Coder	Pitch Predictor		SegSNR (dB)	
	Order	Delays(bits)	female	male
I	1	non-integer (9)	14.48	13.38
II	1	non-integer (10)	14.66	13.48
III	1	integer (8)	14.34	13.32

Table 3 Effect of high resolution pitch prediction.

As shown in Table 3, non-integer delays improved the quality of the reconstructed speech by 0.1-0.3 dB in terms of segmental SNR and a substantial increase in perceived quality. These high resolution pitch predictor also played a very important role in the error matching procedure, making the size of the excitation codebook less significant.

5 Perceptual noise weighting

A further improvement that can be added to the CELP coder is the use of an enhanced noised weighting scheme [1]. This perceptual noise weighting technique solves the problem of unstructured high frequencies and a higher spectral dynamic range. This method is now extensively used because of the increased perceptual quality it adds to the reconstructed speech.

The major disadvantage of a normal noise weighting technique $W(z)$ is inadequate balancing of low and high frequency coding. The coding problem is mainly due to the

interdependency of both tilt and formant parameters. The enhanced noise weighting scheme introduces a decoupling factor that results in a independent control of the tilt with respect to the formants. An additional filter $P(z)$ is used and is responsible for the tilt only. The form of this new weighting filter is:

$$W'(z) = W(z)P(z) = \frac{H(\gamma z)}{H(z)} \frac{1}{1 + \sum_{k=1}^{3} p_k \delta^k z^{-k}} \quad (9)$$

where the coefficients p_k are determined by an LPC analysis on the first four correlation coefficients of the inverse filter $A(z)$ and δ is a spectral tilt controlling parameter and is set to 0.7.

The addition of perceptual noise weighting to the CELP coder did not improve the segmental SNR figures but the perceptual quality of the coded speech was definitely enhanced with no additional bit requirements.

6 Improved CELP

This section discusses the coding of parameters individually and gives a final coder configuration with all parameters quantized. Segmental SNR figures are also given.

- **Frame and sub-frame sizes:** a formant frame of 250 samples (64 Hz) and a pitch sub-frame of 50 samples (320 Hz) are used to control the update rate of all parameters.
- **LPC coefficients coding:** a 16-th order formant filter is used, the LPCs are first transformed into LSFs then coded with 30 bits/frame. Two configurations A and B corresponding to the first and the fourth entry in Table 1, respectively, are used.
- **Pitch coefficient coding:** one pitch tap is used and coded with a 4 bit non-uniform scalar quantizer.
- **Lag estimate and coding:** three configuration were used to determine the optimum lag value as seen in Tables 2 3.
- **Gain estimate and coding:** a 4 bit differential quantizer with a leaky predictor is used to code the differences in successive sub-frames magnitudes. An extra bit codes the sign.
- **Codeword design:** the codebook consists of normalized *iid* Gaussian sequences and the number of codewords is set to 1024.

The final operating rate of the improved CELP coder is listed in Table 4.

Parameter	Bits	Update rate (Hz)	Bits/sec
LPC coefficients	30	64	1920
β	4	320	1280
gain G	5	320	1600
lag M	10	320	3200
codebook	10	320	3200
		Total	11200

Table 4 Improved CELP coder configuration.

The resulting segmental SNR figures applied on eight speech files are shown in Table 5:

Coder	SegSNR (dB)	
	female	male
A	13.79	12.31
B	13.68	12.11

Table 5 Segmental SNR for two coders.

7 Conclusion

With a combination of three different schemes, an improved CELP coder was implemented. The new split vector quantization techniques helped reduce LPC parameter coding rate from 50 bits/frame to 30 bits/frame. The fractional pitch predictor improved both the perceived quality and the segmental SNR of the reconstructed speech. Further improvements that are still being investigated are the use of combined vector quantization on both the pitch coefficient and the gain as well as the implementation of a split-band version of this coder.

References

[1] Y. Ordentlich and Y. Shoham, "Low-delay code-excited linear-predictive coding of wideband speech at 32 kbps," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 9-12, Toronto, 1991.

[2] G. Roy and P. Kabal, "Wideband CELP speech coding at 16 kbits/sec," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 17-20, Toronto, 1991.

[3] J. S. Marques, I. M. Trancoso, J. M. Tribolet and L. B. Almeida "Improved pitch prediction with fractional delays," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 665-668, Albuquerque, 1990.

[4] P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 661-664, Albuquerque, 1990.

[5] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *Proc. Int. Conf. Acoust. Speech and Sign. Process.*, pp. 661-664, Toronto, 1991.

[6] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp.84-95, Jan. 1980.

[7] R. E. Crochiere and L. R. Rabiner, "Multirate Digital Signal Processing," Prentice Hall, Englewood Cliffs, NJ, 1983.