International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

AN IMPROVED 4 KBIT/S CELP SPEECH CODING ALGORITHM

Yanning Bai   Changchun Bao

Speech and Audio Signal Processing Lab, Beijing University of Technology, Beijing 100022
baiyanning@emails.bjut.edu.cn baochch@bjut.edu.cn

ABSTRACT

This paper• presents a 4 kbit/s CELP speech coder that utilizes the nonuniform and part-searching-area algebraic codebook technologies to overcome the insufficient number of signed pulses in fixed codebook (FCB). The nonuniform algebraic codebook is based on the non-uniform statistical properties of the FCB. The part-searching-area utilizes the periodicity of the FCB excitation signal at low bit rate. The latter is only employed when pitch delay is small enough. We also find that preserving the continuity of pitch is very important for voiced segment if these two technologies are used. So different pitch-detect methods are employed for voiced/unvoiced frame. Subjective and objective test results indicate that the qualities of reconstructed speech are improved, especially for the female speakers.

1. INTRODUCTION

Code excited linear prediction (CELP) has emerged in recent two decades as the most dominant technology for encoding telephone bandwidth speech signals. The CELP speech coding algorithm is the basis for many speech coding standards including ITU-T/G.728 at 16 kbps, ITU-T/G.729 [1] at 8 kbit/s and ITU-T/G.723.1 at 6.3 kbps and 5.3 kbit/s.

A key component of CELP is its excitation codebook. One of the most successful excitation codebooks is the algebraic codebook. Although ACELP (Algebraic CELP) algorithm has been proven to deliver toll quality or near toll quality at high to medium bit rates, it is still a challenge to achieve such quality at a bit rate as low as 4 kbit/s. The ITU-T has been in the process of standardizing a 4 kbit/s toll quality speech coding algorithm over the past number of years. But none of candidates has met all requirements for toll-quality speech as defined by the Terms of Reference until now.

One of the reasons for the difficulties is due to the insufficient number of signed pulses at this rate. In order to solve this problem, many methods have been proposed [2-4]. Some technologies have been successfully used in some low bit-rate coders.

In this paper, we proposed a new 4 kbit/s speech coder that has a nonuniform and part-searching-area FCB. This coder is based on a kind of DP-CELP speech coder [5], but adds some new features.

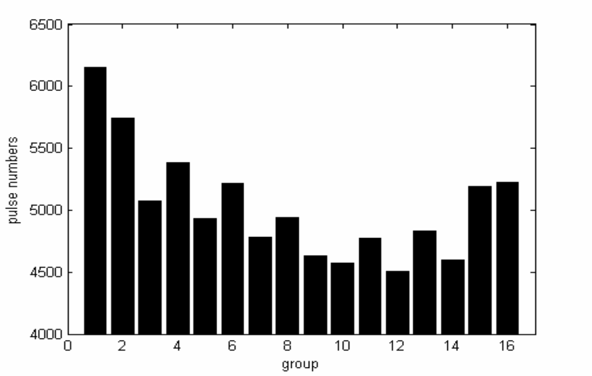

Fig.1. Non-uniform statistical properties of the FCB

Firstly, the algebraic codebook has a nonuniform structure. Our experiments show that FCB pulses have more chances of being placed in the front part of the subframe. Fig.1 gives an experiment result. In this case, the subframe size is 80 samples and one subframe has 4 signed pulses. The pulse locations are divided into 16 groups: locations from 0 to 4 belong to group 1, locations from 5 to 9 belong to group 2, and so on. Nearly 10, 000 frames are tested. Since one frame is divided into two subframes, so nearly 80, 000 pulses are counted. We can see that group 1 and 2 have more pulse numbers than other groups, so it is reasonable to make FCB denser in the front part of the subframe.

Secondly, when pitch delay is small enough, part-searching-area method is employed. Part-searching-area means that the candidate pulses area are part of the subframe area. Generally, female speakers' pitch delay is shorter than the male speakers', so this method is more effective for female speakers.

• This work was supported by National Natural Science Foundation of China under Grant 60372063, Natural Science Foundation of Beijing under Grant 4042009 and the Science and Technology Project of Beijing Municipal Education Commission (KM200310005024).

We also find that preserving the continuity of pitch is very important for voiced segment if these two technologies are employed. In the original DP-CELP coder [5], closed-loop method is employed. We find it is very difficult to preserve the continuity of pitch in voiced segment, especially for female speakers. So different pitch-detect methods are employed for voiced/unvoiced frame in the improved coder.

The rest of this paper is organized as follows: Section 2 explains the nonuniform and part-searching-area algebraic codebook in detail. The pitch-detect method also described here. In Section 3, the improved speech coder is proposed. Section 4 summarizes the results of experiments to measure the perceptual quality of the new speech coder, both in subjective and objective. Finally, in Section 5, we summarize our work.

2. NONUNIFORM AND PART-SEARCHING-AREA ALGEBRAIC CODEBOOK

In DP-CELP speech coder [5], the FCB that contains four non-zero pulses has two modes. Mode 1 and mode 2 correspond to even position and odd position of pulses, respectively. The sign of pulses for two modes is opposite. The positions and signs are given in Table 1. The ACELP only searches for the best-matched synthesis signal in minimum mean square error senses.

Table 1: 14 bits FCB structure

PULSE	SIGN AND POSITION	
	MODE1	MODE2
m_0	(+) 0,10,20,30,40,50,60,70	(-) 1,11,21,31,41,51,61,71
m_1	(-) 2,12,22,32,42,52,62,72	(+) 3,13,23,33,43,53,63,73
m_2	(+) 4,14,24,34,44,54,64,74	(-) 5,15,25,35,45,55,65,75
m_3	(-) 6,16,26,36,46,56,66,76 8,18,28,38,48,58,68,78	(+) 7,17,27,37,47,57,67,77 9,19,29,39,49,59,69,79

2.1. Nonuniform Algebraic Codebook

Since the nonuniform statistical properties of the FCB were found in Fig.1, we change the FCB structure of the original DP-CELP coder. The positions and signs are given in Table 2. Compared with the original one, the pulse density is different in one subframe and the front part of the subframe has larger pulse density. The method is dem-onstrated to be efficient for both male and female speakers.

Table 2: 14 bits non-uniform FCB structure (sub-codebook 1)

PULSE	SIGN AND POSITION	
	MODE1	MODE2
m_0	(+) 0, 6,16,26,36,46,56,66	(-) 0, 7,17,27,37,47,57,67
m_1	(-) 1, 8,18,28,38,48,58,69	(+) 1, 9,19,29,39,49,59,70
m_2	(+) 2,10,20,30,40,50,60,72	(-) 2,11,21,31,41,51,61,73
m_3	(-) 3,12,22,32,42,52,62,75 4,14,24,34,44,54,64,78	(+) 3,13,23,33,43,53,63,76 5,15,25,35,45,55,65,79

2.2. Part-Searching-Area Algebraic Codebook

Although non-uniform algebraic codebook can improve the reconstructed speech quality, the non-zero pulse numbers are the same as the original one. The part-searching-area algebraic codebook can solve this problem partly.

At medium bit-rate, the FCB excitation has a uniform pulse distribution over the subframe. But at lower bit rates, the FCB excitation begins to display strong periodicity [4]. This is particularly true in steady voiced and transient voiced segments of the speech. In theory, we expect the FCB reflects it in structure.

Just like G.729 [2] and DP-CELP [5] speech coders, when the pith delay is less than the subframe size, the true FCB *C(n)* is modified as

$$C(n)=\begin{cases} C(n) & n=0,...,T-1 \\ C(n)\ +\ \beta C(n-T) & n=T,...,79 \end{cases} \quad (1)$$

where β is pitch gain and bounded by 0.2 < β< 0.8. Equivalently, the above modification is performed in the FCB search by modifying the impulse response *h(n)* of the vocal model.

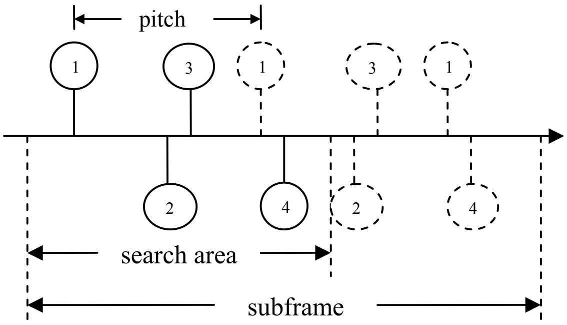

Fig.2. Response of Backward Pitch Enhancement

Since some nonzero pulses can be generated by equation (1) when pitch delay is small enough, another idea is proposed naturally. Why not decrease the search area to reduce the interval of nonzero pulses? In the original DP-CELP coder (the search area is the same as the subframe), the average interval between two locations is two samples, while if the search area size is half subframe size, the average interval is one sample. Fig.2 shows the idea of part-searching-area algebraic codebook. The pulses are constrained in the search area other than the subframe. The pulses outside the search area are generated by equation (1). It seems to be an adventure to decrease the size of search area, but this method certainly improves the reconstruction speech quality.

In our scheme, if pitch delay is less than 38 samples, the search area is 55 and β in equation (2) is equal to 0.96. The FCB structure is showed in Table.3. In this case, the average interval between two locations is 1.5 samples. Further more, if pitch delay is less than 28 samples, the search area is 40 and β is equal to 0.97. The FCB structure is showed in Table 4. In this case, the average interval between two locations is 1 sample.

Table 3: 14 bits non-uniform FCB structure (sub-codebook 2)

PULSE	SIGN AND POSITION	
	MODE1	MODE2
m_0	(+) 0,5,10,15,20,26,36,46	(-) 0,5,10,15,20,25,35,45
m_1	(-) 1,6,11,16,21,28,38,48	(+) 1,6,11,16,21,27,37,47
m_2	(+) 2,7,12,17,22,30,40,50	(-) 2,7,12,17,22,29,39,49
m_3	(-) 3,8,13,18,23,32,42,52	(+) 3,8,13,18,23,31,41,51
	4,9,14,19,24,34,44,54	4,9,14,19,24,33,43,53

Table 4: 14 bits non-uniform FCB structure (sub-codebook 3)

PULSE	SIGN AND POSITION	
	MODE1	MODE2
m_0	(+) 0,5,10,15,20,25,30,35	(-) 0,5,10,15,20,25,30,35
m_1	(-) 1,6,11,16,21,26,31,36	(+) 1,6,11,16,21,26,31,36
m_2	(+) 2,7,12,17,22,27,32,37	(-) 2,7,12,17,22,27,32,37
m_3	(-) 3,8,13,18,23,28,33,38	(+) 3,8,13,18,23,28,33,38
	4,9,14,19,24,29,34,39	4,9,14,19,24,29,34,39

So there are three sub-codebooks in the FCB. Once the best sub-codebook is chosen according to pitch delay, some fine searching for the best code-vector is done within this specific sub-codebook. Since which sub-codebook is selected is based on the pitch delay, no extra bit is required.

2.3. Adaptive Codebook Search

We find that preserving the continuity of pitch is very important for voiced segment if above technologies are used. In the original DP-CELP speech coder, the adaptive codebook parameters (or pitch parameters) are determined by closed-loop method. We find it is very difficult to preserve the continuity of pitch in voiced segment, especially for female speakers. The reason may be like this:

At the bit rate of 4 kbit/s, the frame and subframe sizes are relatively long to reduce the bit rates for coding the gains and the synthesis filter parameters. In the adaptive codebook approach for implementing the pitch filter, the excitation is repeated for delays less than the subframe length. When pitch delay is small, the excitation has to be repeated several times. So it is not always appropriate to do like this.

In order to preserve the continuity, different pitch-detect methods are employed for voiced/unvoiced frame in the improved coder. An open-loop pitch analysis is done per frame at first. According to the result and other parameters, such as peakiness [6], we classify this frame into voiced or unvoiced frame. Then different closed-loop search methods are used respectively. If this frame belongs to voiced frame, the final delay is found by searching a small range (eight samples) of delay values around the open-loop delay. If this frame belongs to unvoiced frame, the final delay is found by searching an entire range from 20 to 143 samples. It is noted that no extra bit is needed to denote this frame is voiced or unvoiced. Our experiment indicates that this method can preserve the continuity of pitch in voiced segment while pitch accuracy is also preserved.

3. IMPROVED 4 KBIT/S SPEECH CODING ALGORITHM

A system block diagram of improved speech coding algorithm is shown in Fig.3. The coder operates on speech frame of 20 ms at a sampling rate of 8 kHz. The 20 ms frame is divided into two sub-frames of 10 ms each. The LSF interpolation, the extraction of the adaptive and fixed-codebook indices and gains are operated in 10 ms sub-frame. The bit allocation of coder parameters is shown in Table 5. The issues of main focus are presented as follow

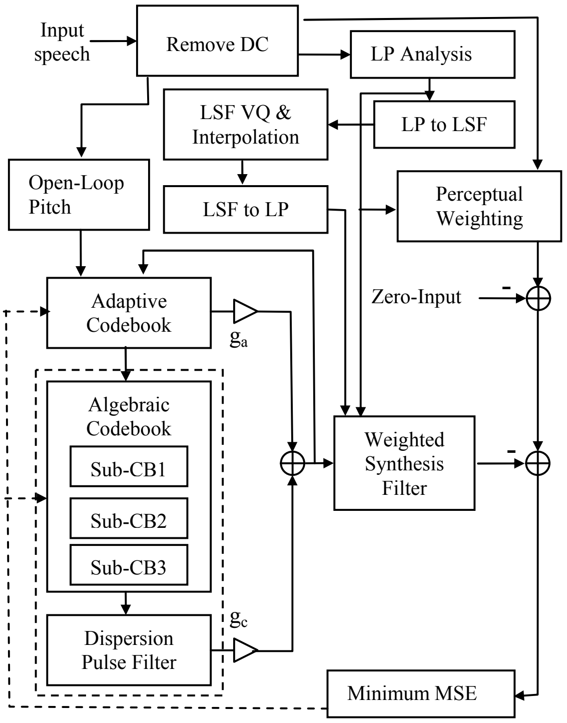

Fig.3. Block diagram of the speech encoder

Table 5: Bit allocation of the 4 kbit/s speech coder

Parameter	Bits per 20 ms
LSFs	20 bits
ACB	8 bits/subframe 16 bits
FCB mode	1 bits/subframe 2 bits
FCB	13 bits/subframe 26 bits
ACB gain	4 bits/subframe 8 bits
FCB gain	4 bits/subframe 8 bits
TOTAL	**80 bits**

3.1. LP Analysis and LSF Quantization

The LP analysis is done per 20 ms frame to compute the LP coefficients. The window for LP is the same as G.729, but there is no a look-ahead, that is to say that the algorithm delay is 20 ms. Then the LP coefficients are converted to Line Spectrum Frequencies (LSF) and vector quantized with 20 bits in one frame. Then the LP coefficients are converted to Line Spectrum Frequencies (LSF) and vector quantized with 20 bits in one frame. A one-step interpolation predictive vector quantization of LSP parameters [5, 7] is used in this coder.

3.2. Perceptual Weighting

The perceptual weighting is based on the unquantized LP filter, *A(z)*, and is given by

$$W(z)=\frac{A(z/\gamma_1)}{A(z/\gamma_2)} \qquad (2)$$

As in the ITU-T recommendation G.729 [1], the amount of perceptual weighting, controlled by γ_1 and γ_2 , is adaptive and depends on the spectral shape of the signal. Similar to G.729 the spectral shape is determined from the first two LAR coefficients, and it is mainly characterized as either tilted or flat.

3.3. Dispersion Pulse Filter

Just like the original DP-CELP speech coder, the dispersion pulse filter is applied to the FCB. The dispersion vector consists of the FIR filter impulse-response with the cut-off frequency of 3.4 kHz. The vector is used for dispersing the energies of signed pulses so as the ACELP coder performance to be improved.

4. EXPERIMENT RESULTS

To evaluate the performance of the 4 kbit/s CELP speech coding algorithm, a large number of subjective A/B tests were done. The test sentences are from 16 Chinese utterances including 8 female speakers (F1--F8) and 8 male speakers (M1--M8). The subjective listening test results show that the perceptible qualities of most sentences are improved.

In addition, we also gave the objective comparison in terms of predictive MOS obtained from ITU-T Recommendation P.862 [8] that is approved in 2001. This recommendation is an objective method named as Perceptual evaluation of speech quality (PESQ), which is for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. The PESQ test results are given in Table 6. It is seen that the improved 4 kbit/s CELP performs better than the original one in MOS value, especially for female speakers.

Table 6: Experiment results

Test Files	MOS (ori.)	MOS (new)	Test Files	MOS (ori.)	MOS (new)
F1	3.048	3.168	M1	2.774	2.812
F2	2.943	3.155	M2	2.867	3.013
F3	2.708	2.863	M3	3.141	3.161
F4	3.304	3.299	M4	3.073	3.112
F5	3.024	3.142	M5	2.926	2.969
F6	3.127	3.148	M6	3.128	3.279
F7	2.624	2.895	M7	2.965	3.084
F8	2.909	2.977	M8	3.183	3.217
Avg.	2.961	3.081	Avg.	3.007	3.081

5. CONCLUSION

In this work, an improved 4 kbit/s CELP speech coding algorithm was proposed. We demonstrate that to obtain good speech quality in low bit-rate CELP coders, it is important to utilize the nonuniform statistical properties and the periodicity of the FCB excitation signal. The method of this paper is able to achieve this goal by nonuniform and part-searching-area FCB technologies. Subjective and objective tests show that the reconstruction speech qualities are improved, especially for female speakers.

6. REFERENCES

[1] ITU-T Recommendation G.729. Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), 1996.

[2] K. Yasunaga et al., "Dispersed-pulse codebook and its application to a 4kb/s speech coder", *IEEE Proc.* ICASSP, 2000, pp. III-1503-1506.

[3] Y. Gao et al., "eX-CELP: A Speech Coding Paradigm", *IEEE Proc*, ICASSP, 2001, Vol. 2, pp. 689-692.

[4] Ajit V.Rao, Sassan Ahmadi et al, "Pitch Adaptive Windows for Improved Excitation Coding in Low-Rate CELP Coders," *IEEE Trans. Speech Audio Processing*, Nov, 2003, Vol. 11, pp. 648-659.

[5] Bao Changchun, "A High Quality Dispersed-Pulse CELP Speech Coding Algorithm at 4 Kb/s," *ACTA ELECTRONIC SINICA*, 2003, Vol. 31, No.2, pp. 309-313.

[6] Wai C. Chu, "Speech Coding Algorithms –Foundation and Evolution of Standardized Coders," *Wiley-Interscience*, 2003.

[7] Bao Changchun, "Harmonic Excited LPC (HE-LPC) Speech Coding at 2.3kb/s," Hongkong, *IEEE Proc.* ICASSP2003, 2003, pp. I-784~I-787.

[8] ITU-T Recommendation P.862. Perceptual evaluation of speech quality (PESQ), 2001.