1.4 Kbps 聲門激源線性預測(GELP)語音編解器之設計與實現

胡懷祖[1] **游竹**[2]

1. 國立宜蘭技術學院電子工程系 教授
2. 國立宜蘭技術學院電子工程系 助理教授

摘要

本文提出一個搭配 LPC 語音編碼器之聲門激源模型以使語音信號能在1400bps 的情況下有效編碼。其中頻譜參數的編碼工作是以轉轍式線性預測類神經網路伴隨多階段向量量化的方式來處理，而激源區分為兩類，屬於無聲的激源是搜尋自隨機代碼簿，至於有聲激源則是從聲門代碼簿加以挑選，所有涉及到激源信號的分析與合成以及代碼簿的構建程序皆有詳細交代。針對此一 1.4kbps 語音編解碼器所做之音質評鑑，其平均值為 2.993，而 2.4 Kbps 之 LPC 編碼器與4.8 Kbps 之 CELP 編碼器的相對分數則分別是 2.272 與 3.314。此外，我們亦嘗試推出簡化版使其能在 ADSP-2181 上執行，但由於演算法刪減與記憶體受限之故，致使最後的音質跟著下降，這似乎意謂著整套編解碼功能的完整實現都還是得借重擁有大記憶容量的浮點數 DSP 晶片。

關鍵詞： 聲門激源、LPC 語音編解碼器、轉轍式預測類神經網路

DESIGN AND IMPLEMENTATION OF A 1.4KBPS GLOTTAL EXCITATION LINEAR PREDICTION (GELP) VOCODER

Hwai-Tsu Hu [1]   Chu Yu [2]

1. Professor, Department of Electronic Engineering, National I-Lan Institute of Technology
2. Assistant Professor, Department of Electronic Engineering, National I-Lan Institute of Technology

Abstract

This paper presents a glottal excitation model to cope with the LPC vocoder for speech signals coded at 1400 bps. We encode the spectral parameters by using a switched-predictive neural network along with multi-stage vector quantization. While the unvoiced excitation is retrieved from a stochastic codebook, we use a glottal codebook to characterize the voiced excitation. Procedures are described for analysis and synthesis of the excitation signals in addition to codebook construction. The MOS test regarding the 1.4 Kbps GELP coder is 2.993, as compared with 2.272, 3.314 for the 2.4 Kbps LPC and 4.8 Kbps CELP coders. A simplified version is developed to work on the ADSP-2181 processor, but it suffers quality degradation due to the algorithm truncation and memory restriction. This suggests that fully implementation of the proposed coder may rely on a floating-point DSP chip integrated with large memory.

Key Words: glottal excitation, LPC vocoder, switched-predictive neural network

I. INTRODUCTION

Due to the booming demands of personal communication services, speech coders with low bit rates have been increasingly important for applications such as wireless telecommunications and internet telephony. One class of speech coders that has been extensively used in practice is the linear prediction coding (LPC) vocoder [1], which is developed based on a parametric model. The low bit rate is achieved by transmitting the involved parameters of the speech production model across the communication channel. At the receiving end the decoder is designed to regenerate speech signals according to these modeling parameters. Speech quality synthesized in this manner at low bit rates is often judged as unnatural due to incorrect voicing decisions, poor spectral resolution, and oversimplified excitation functions [2,3]. However, improvements are achievable with some sophisticated excitation models such as multipulse [4,5], regular pulse [6,7], or stochastic codeword [8,9]. Advancement in excitation modeling has successfully come out with several standards ranging from 4.8 to 16 Kbps in the past decade [10].

In fact, the ideal excitation for an LP coder is the residual signal obtained by inverse filtering the speech signal. Many investigators recognized that the glottal features residing in the residual signal are essential for synthesizing natural-sounding speech [11-13]. In this research we aim at developing an efficient excitation model to simulate the residual signal so that high-quality synthetic speech is attainable at a low-bit rate. Since the feasibility of the designed speech coder is also our concern, we intend to implement the designed coder subject to feature constraints of the ADSP-2181 processor.

II. CODING SCHEME

A. Speech production model

To incorporate the glottal excitation into the LP coder, we adopt a hybrid speech production model called “glottal excited linear prediction (GELP)” [14]. As presented in Fig. 1, this model inherits the basic structure of the LPC vocoder but exhibits the generic nature of the CELP coder. The analysis of speech signals is equivalent to the extraction of modeling parameters. Given that the speech signal is sampled at 8 KHz, we update the analysis frame at a rate of 240 samples. The analysis procedure begins with a voicing decision along with pitch estimation. Within each frame, a coarse estimate of the pitch period is obtained by adopting the average magnitude difference function (AMDF) [15] based on the lowpass filtered residual (LFR) [16]. The underlying frame is classified as voiced whenever the averaged magnitude of the speech signal is above 0.01 of the maximum allowable value and the resulting AMDF exhibits a distinct valley around the pitch period. In case the speech segment is categorized as voiced, we adjust the estimated pitch period using one frame of look-ahead to render a reliable pitch follower. A peak-picking procedure is then used to identify the glottal closure instant (GCI) subject to the constraint that the length deviation between any neighboring GCI's has to be less than 20% of the estimated pitch period.

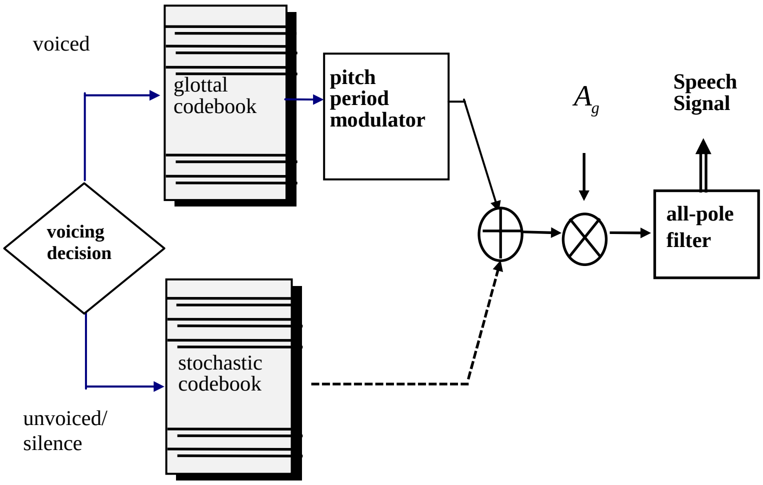

Fig. 1. Glottal excited linear prediction (GELP) speech production model

B. Spectral parameters

In order to characterize the spectral properties of speech signals, two sets of LP coefficients are derived respectively from the speech samples centered at the one-fourth and three-fourth places of the frame, each of which extending over a duration of 200 samples. Both of them are then converted to the line spectral frequency (LSF) parameters and jointly coded by using four-stage 22-bit (i.e., {6,6,5,5}) vector quantization [17]. Here we particularly apply a predictive neural network to render an initial estimate of the average of two sets of LSF parameters. As shown in Fig. 2, this network accommodates a single layer of neurons with purely linear functions, and each neuron collects LSF parameters of three past subframes as inputs. It is evident that the parameters for quantization are the residual LSF vector. Analogous to the switching method adopted by [18], two networks are prepared along with separate codebooks. The procedures for training the predictive networks and LSF codebooks are as follows. First, a primitive network is obtained by employing the quasi-Newton backpropagation method with all sampled LSF residual vectors involved. These residual samples are then categorized into two groups according the rule that samples with smaller errors are clustered together while samples with larger errors are attributed to the second group. Samples in each group are subsequently used to derive a new network individually. Finally, we perform the test with respect to all the samples based on two derived networks. The categorization is rearranged to associate each sample vector with a predictive network that attains a smaller sum-squared error. The above recategorization and network training are repeated iteratively until the reduction of the overall sum-squared error becomes negligible.

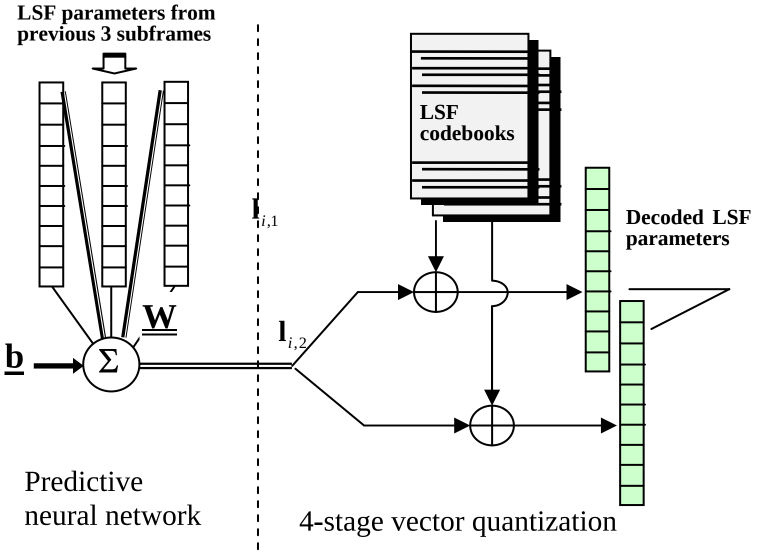
LSF parameters from previous 3 subframes
W
b
Σ
l_{i,1}
l_{i,2}
LSF codebooks
Decoded LSF parameters
Predictive neural network
4-stage vector quantization

Fig. 2. Predictive neural network and multi-stage vector quantization of LSF parameters

During the analysis phase, both network predictor/codebook pairs are tried, and the one that provides better quantization performance is selected for transmission along with one bit to represent the switch. The indices of the chosen codevector from all stages are transmitted to the decoder, and the quantized LSF vector is reconstructed by summing up the network output and all the codevectors from the multi-stage codebooks, i.e.,

$$\begin{bmatrix} \mathbf{l}_{i,1} \\ \mathbf{l}_{i,2} \end{bmatrix} = \mathbf{W}^{(n)} \begin{bmatrix} \mathbf{l}_{i-2,2} \\ \mathbf{l}_{i-1,1} \\ \mathbf{l}_{i-1,2} \end{bmatrix} + \mathbf{b}^{(n)} + \sum_{k=\mathbf{1}}^{m} \mathbf{c}_k{}^{(n)} \qquad (1)$$

where $\mathbf{l}_{i,j}$ represents the quantized LSF parameters at the jth subframe of the ith frame. $\mathbf{W}^{(n)}$ and $\mathbf{b}^{(n)}$ are the weighting matrix and bias vector of the nth predictive network, respectively. The codevector at the kth stage is denoted as $\mathbf{c}_{\mathbf{k}}{}^{(n)}$, which is determined by searching *M*-best paths that achieve the overall lowest distortion. In this study, the depth of codeword searching at each stage is 4 (i.e., *M*=4).

In the foregoing discussion, although there is only one bit to denote the switch, the codebook entries in all stages ought to be increased twice. The number of searched codevectors is therefore greater than that employed by the traditional method. We believe that it is the primary reason why the switched codebook generally yields better performance.

C. Source excitation

The analysis of the residual signal can be divided into three parts, namely, the main excitation, low-frequency component, and noise source. While the glottal features manifests itself as a low-frequency waveform, we quantize the waveform of the integrated residual signal on a pitch-period by pitch-period basis. Each pitch period is defined as the interval between two consecutive GCI’s.

The magnitude of the main excitation relates to the abruptness of the glottal closure and reflects on vocal quality directly. We describe the pulse magnitude, *m*, by

$$m = \alpha \left(\sum_{i=0}^{p-1} e^2(i) \right)^{\frac{1}{2}} \qquad (2)$$

where $e(i)$ denotes the low-frequency component of the residual obtained by using a 6th-order median filter, p represents the pitch period, and α is an adjustable factor that is included in the excitation model. Because the synthetic speech is not perceived differently unless α is changed considerably, we encode the value of α using one single bit per frame. The coded bit, termed m_b, is 1 whenever the averaged α is greater then 1.3 (which is approximately the mean value of all the measured α's), and it is 0 otherwise.

The analysis of the low-frequency component is also performed on the basis of pitch periods. We portray this component by using vector quantization. Steps of processing such a component comprise the integration, linear trend removal, and length adjustment. A total of 6353 normalized integrated residual templates obtained from twelve sentences uttered by four subjects are used in the training stage of vector quantization. All templates are unified to have a length of 64 samples and a power level of unity. The LBG algorithm [19] in conjunction with the maximum descent criterion [20] is employed to construct the codebook. We empirically find that 16 entries give adequate performance in describing the integrated residual. In particular, we use only one codeword to delineate the source characteristics for each voiced frame. The codeword index c_n is chosen to have the maximum correlation across the entire frame

$$c_n = \arg\max_i \left\{ \sum_{k=0}^{63} \left(\sum_{m=0}^{N_{gci}-1} e_m(k) \right) u_i(k) \right\} \tag{3}$$

where N_{gci} represents the number of GCI's, and $u_i(k)$ denotes the *k*th element of the *i*th glottal codeword.

Finally, as the glottal turbulent noise is important for producing natural voices, the proposed excitation model incorporates a noise source originally developed by McCree and Barnwell [21].

D. Excitation replication

The decoder at the receiving end is designed to reconstruct speech signals from coded bit string. For unvoiced frames, the synthetic speech is carried out by feeding the stochastic codewords to the synthesis filter obtained by converting the decoded LSF parameters. The synthesis of voiced speech is rather complicated since we have to replicate the glottal features based on quantized information. Since the synthesis of unvoiced speech already inherits the nature of waveform matching, we decide to keep track of GCI's for voiced frames.

As our coding scheme encodes the averaged pitch period and the last GCI position, we retrieve the rest GCI positions by means of interpolation and regulation to render a smoother transition between frames. The approach for pitch smoothing is arranged as follows. We first apply a first-order IIR lowpass filter, $0.3/(1-0.7z^{-1})$, to a data sequence of N_{gci} samples, each of which retaining an averaged pitch value p_{avg}. The initial filter memory holds the last pitch period derived in the previous frame. While the lowpass filter alters the pitch periods, we linearly scale these periods so that the last GCI is back to its original position. After determining each individual pitch period within the underlying frame, we generate the excitation waveform in a filterlike manner by

$$u_r(k) = 0.15^{1/N_{gci}} u_r(k) + (1-0.15^{1/N_{gci}})u_c(k), \qquad k = 0,1,2,\cdots,63. \tag{4}$$

where $u_r(k)$ represents the waveform of the running excitation, and $u_c(k)$ is the glottal codeword chosen in the present frame. The above recursive equation allows the forthcoming glottal features to merge into the intermediate excitation.

We remind that the excitation waveform, $u_r(k)$, is an integrated version of the residual. The differentiation with respect to such a waveform can only provide the low-frequency part of the glottal excitation. A complete glottal excitation model requires the involvement of the high-energy pulse and turbulent noise. Supposed that $v(k)$ is the derivative of an interpolated $u_r(k)$ with a length of p, a dispersed pulse sequence $i_p(k)$ given in Table I is introduced to $v(k)$ by

$$v(k) = \frac{i}{16}v(k) + \frac{16-k}{16}\beta i_p(k), \qquad k = 0,1,\cdots,14. \tag{5}$$

where β is computed recursively from its previous rendition and the encoded bit m_b by

$$\beta = 0.6\beta + 0.4 \times (1+0.6m_b). \tag{6}$$

In Eq. (6) the boundaries for β are 1.0 and 1.6, which correspond to the averages of the upper 50% and lower 50% of the measured α's, respectively.

Table I. Numerical values of the dispersed pulse

Index	Value				
0 → 4	1.000	0.039	-0.473	0.303	-0.155
5 → 9	0.112	-0.078	0.092	-0.044	0.040
10 → 14	0.039	0.020	-0.014	0.001	0.003

Our informal listening test confirms that the participation of such a dispersed pulse not only reduces buzzy quality but also produces expected vocal quality quite effectively. However, the spectrum of the resulting excitation is generally not flat. Such a consequence thus calls for spectral adjustment.

Note that the *n*th autocorrelation function, $R(n)$, with respect to the periodical excitation $v(n)$ is

$$R(n) = \sum_{k=0}^{p-1} v(k) v\big((k+n).\mathrm{mod}.p\big) \quad , \tag{7}$$

where the symbol .mod. denotes the modulus operation, and p is the length of excitation. The change to an arbitrary $v(i)$, termed $d(i)$, will yield a different autocorrelation function $\tilde{R}(n)$ such that

$$\tilde{R}(n) = d(i)\big[v(i+n) + v\big((p+i-n).\mathrm{mod}.p\big)\big] + R(n) \, . \tag{8}$$

While the desired excitation needs to possess a flat spectrum, we deliberately alter samples around the GCI (e.g., $i \in \{2,3,4,p-1\}$) to flatten the spectrum of the resulting excitation. Theoretically, the autocorrelation functions for an excitation with a flat spectrum should be zero except for the zero-lag term. Taking the first q nonzero-lag autocorrelation functions gives

$$\begin{cases} d(i)\big[v(i+1) + v\big((p+i-1).\mathrm{mod}.p\big)\big] + R(1) = 0 \\ d(i)\big[v(i+2) + v\big((p+i-2).\mathrm{mod}.p\big)\big] + R(2) = 0 \\ \qquad\qquad\qquad \vdots \\ d(i)\big[v(i+q) + v\big((p+i-q).\mathrm{mod}.p\big)\big] + R(q) = 0 \end{cases} \tag{9}$$

In case $q > 1$, the solution for the above over-determined system is

$$d(i) = -\frac{\sum_{n=1}^{q} R(n)\big[v(i+n) + v\big((p+i-n).\mathrm{mod}.p\big)\big]}{\sum_{n=1}^{q} \big[v(i+n) + v\big((p+i-n).\mathrm{mod}.p\big)\big]^2} \, . \tag{10}$$

Finally, we revise the new $\tilde{v}(i)$ as $\tilde{v}(i) = v(i) + d(i)$. In this research, q is chosen to be 5. We derive $\tilde{v}(2)$, $\tilde{v}(3)$, $\tilde{v}(4)$ and $\tilde{v}(p-1)$ one by the other to reach a near-flat spectrum.

Following the modification of $\tilde{v}(i)'s$, we have since added a noise source in series with the glottal codebook excitation to contribute additional high-frequency content.

Fig. 3 demonstrates typical examples of the derived glottal excitation.

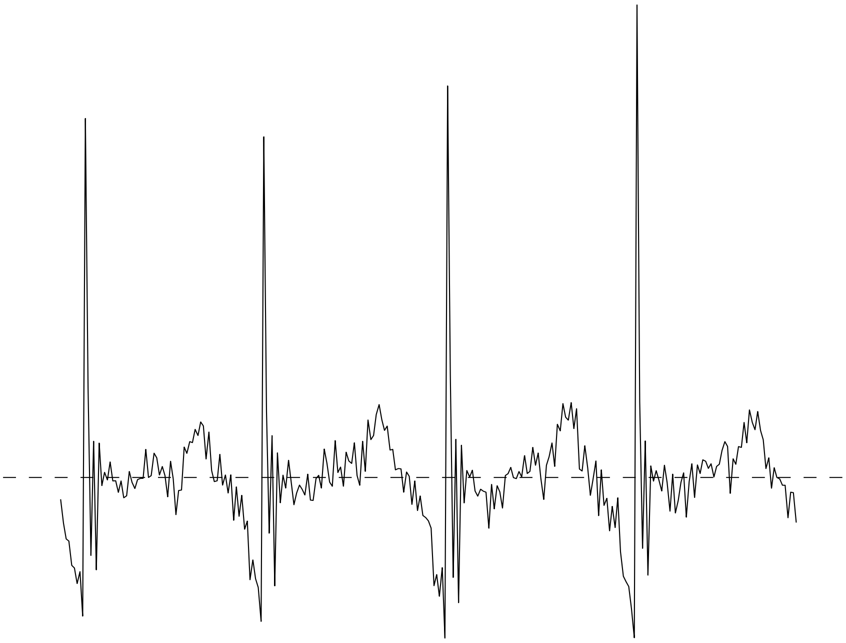

Fig. 3 Examples of glottal excitation

E. Gain control

The gain control is the remaining issue in the coding scheme. Following the derivation of the excitation, the synthetic speech is obtained by feeding the gain-adjusted excitation (either the glottal excitation or the innovation sequence) to a synthesis filter. Here we encode the logarithm of the segmental power instead of the excitation gain. For each voiced frame the pitch-synchronous segmental power on a logarithmic scale is obtained and interpolated into 4 representative values. These 4 values are vector quantized using 5 bits. The unvoiced frame is coded in a similar manner, but only the power levels derived from two subframes are involved in the quantization.

During the synthesis phase, we interpolate the LSP parameters to acquire the synthesis filter as the synthesis interval slides across frames. The excitation gain is then derived from the segmental power using a two-filter strategy. Namely, there are two filters employed to perform speech synthesis: the one holding the current LP coefficients is in charge of the excitation response $f(k)$, and the other retaining the previous LP coefficients takes care of memory contribution $h(k)$. The synthesized speech for each synthesis interval is the sum of the outputs of the two filters. Thus,

$$P_r = \frac{1}{M}\sum_{k=0}^{M-1}(h(k) + A_g f(k))^2 \quad , \tag{11}$$

where P_r is the segmental power and M is the length of the interval. The gain A_g can be obtained by solving the above quadratic equation. To avoid the obtainment of a negative or complex A_g, we damp the memory contribution $h(k)$ by multiplying each LP coefficient with 0.97 raised to the power of its index. This approach has been found to work well in our formal study [14]. Besides, it serves as the role of regulating energy deviation due to postfiltering [22], which is used in many pulse-excited vocoders to emphasize synthetic speech in formant regions.

F. Bit allocation

As manifested by the speech production model in Fig. 1, much emphasis is placed upon the LP filter and the source excitation. For each voiced frame we encode the spectral properties of the LP filter by applying the switched predictive-network with four-stage vector quantization to two sets of 10th-order LSF parameters. However, we have reserved the bits at the 4th stage to accommodate more stochastic codewords while processing unvoiced frames. The rationale behind the reallocation of bits is that the formant structure of unvoiced speech is generally not as distinctive as that of voiced speech. Therefore, the search among a larger stochastic codebook not only partially makes up for the deficiency in spectral modeling but also provides chances to produce plosive sounds.

The coding procedure of the unvoiced excitation is a direct modification of CELP. The analysis frame is merely partitioned into two subframes, each of which has a length of 120 samples. The design and search of the stochastic codebook follow the specifications of FS-1016 [23]. On the other hand, the coded source parameters for each voiced segment consist of the logarithmic segmental power, the glottal codeword, the pulse magnitude, and the number of GCI's associated with the last GCI position. The purpose of encoding the last GCI position rather than the pitch period is to achieve pitch synchronization and higher pitch resolution. The coding rate of our designed GELP coder is 1.4 Kbps. Table II presents the detailed coding scheme.

Table II. Bit assignment for the 1.4 Kbps GELP coder

	Voiced	Unvoiced
Voicing	1	1
Codebook switch	1	1
LSF parameters	22	17
Segmental power	5	5
Glottal codeword	4	
Pulse magnitude	1	
GCI number	3+(2)*	
Last GCI location	3+(2)*	
Stochastic codeword		8×2
Excitation polarity		1×2
Total	42	42

* Depending upon whether the number of GCI's exceeds 8, (2) bits are alternatively shifted from "last GCI location" to "GCI number" in the voiced frame.

III. LISTENING EVALUATION

We perform a listening test for a total of thirty files corresponding to speech uttered by six different speakers (3 male, 3 female), each speaker delivering five sentences. The mean opinion scores (MOS) judging from the 2.4Kbps LPC-10e v.55 (FS-1015) vocoder [24] and 4.8Kbps CELP (FS-1016) coder [23] are provided as reference. Twenty-four listeners participated in the test. During the testing period, each individual used a high-quality handset as a listening device in a quiet room. The sequence of the recordings in each presentation was played randomly.

Table III presents the results derived from the listening test. The scores associated with the LPC, GELP, CELP vocoders are 2.2722, 2.9931, and 3.3139, respectively. According to the opinions gathered from the listeners, the GELP coder suffers several weaknesses. For example, the utilization of a common frequency-shaping noise cannot fully characterize subtle differences of vocal quality. Faults among female synthetic speech are ascribable to inaccurate GCI identification, which leads to improper modeling of the glottal phase characteristics. The incorporation of a unique pulse sequence is probably another reason of quality degradation. All these defects lead to unfavorable responses from the listeners. Currently, a follow-up research project is undertaken to overcome these problems by using a frequency-domain approach.

Table III. Mean Opinion Scores (MOS) for speech sentences synthesized by CELP,

GELP, LPC coders

	4.8 Kbps CELP	**1.4 Kbps GELP**	**2.4 Kbps LPC**
Male speakers	3.4361	3.2667	2.2778
Female speakers	3.1917	2.7194	2.2667
AVERAGE	3.3139	2.9931	2.2722

IV. IMPLEMENTATION

The processor considered in our study is the ADSP-2181, which is a 16-bit fixed-point DSP chip operated at 33.3 MHz. In addition to three independent full-function computational units which support single-cycle instructions, the ADSP-2181 includes full-base on-chip memory: 16K×24-bit words of program memory and 16K×16-bit words of data memory. As the ADSP-2181 has demonstrated its capability of performing the well-known 4.8Kbps CELP coder without the need of external memory, it seems practicable to implement the proposed 1.4Kbps GELP coder on such a chip in consideration of the processing speed and memory requirement. In accordance to such a condition, the maximum sizes of both program and data memory for the GELP algorithm are set to 16K words, and the computation capability is counted as 10^6 ($=33.3\times10^6/8000\times240$) fixed-point operations for a frame of 240 speech data sampling at 8 KHz. Table IV lists the number of words required to account for global variables. It is observed that the most memory-consuming item is the switched LSF codebooks since each codevector comprises two sets of 10th-order LSF parameters. The number of bits for encoding spectral parameters appears to reach its limit since each additional bit will enlarge the codebook drastically, leading to the exhaustion of remaining memory space.

Table IV. Memory space required for the involved variables

Predictive Neural network	Weighting	2×30×10
	Bias	2×10
LSF codebooks	Stage. 1	2×64×20
	Stage. 2	2×64×20
	Stage. 3	2×32×20
	Stage. 4	2×32×20

Glottal codebook	16×64
Stochastic codebook	512+120
Voiced gain	32×4
unvoiced gain	32×2
Data buffer	240×2
Window	240
Total	**10868**

In contrast to the arrangement of data memory, the settlement of program memory is rather difficult because we have to confine the program size while restricting the number of instruction cycles down to 10^6. The manufacturer of the ADSP-2181 provides many useful libraries and illustrations in both C and assembly languages. However, for simplicity, except for few routines most of the program is written in C. This makes the resulting program less efficient not only in the memory utilization but in the computational speed. Furthermore, the necessity of floating-point operations in many situations aggrandizes the problem of insufficiency since the accomplishment of one floating-point operation often requires a bunch of fixed-point instructions. Eventually we learn that a complete version of the proposed coding scheme can not be achieved by simply transferring from a C program. Solutions to overcome such a predicament include the employment of a processor with a faster instruction rate clock and/or the substitution of optimized assembly codes plus the incorporation of external memory. However, our primary goal in this study is to evaluate the feasibility of the proposed algorithm. We therefore truncate this algorithm to make it executable on the ADSP-2181 anyway. For example, the M-L tree search procedure is not conducted during the search of the multi-stage codebooks. We also omit the pitch refinement subroutines that rectify the pitch contour and detect the GCI's position. From the experimental results, it is revealed that the above truncation helps us to save efforts on programming but inevitably result in the degradation of synthetic quality.

In fact, as faster floating-point DSP chips are available by nowadays technology, a full version of the algorithm could be implemented without sacrificing the performance. In the future we expect to make this algorithm fully executable on a floating-point DSP processor (such as ADSP-21061 or 21062) that retains a faster instruction rate and larger memory space.

V. CONCLUSIONS

This paper describes a 1.4 Kbps GELP coder that employs a glottal excitation model for voiced speech and innovation sequences for unvoiced speech. We formulate these two types of excitations as codevectors that are used to excite a synthesis filter. Only one of these two excitation functions is active at any time. The performance of the proposed coder is subjectively evaluated in comparison with the 2.4 Kbps LPC-10e (FS-1015) and 4.8 Kbps CELP (FS-1016) speech coders.

We have developed a real-time version based on the ADSP-2181, which is a 16-bit fixed-point DSP processor integrated with 80 Kbytes on-chip memory. Most of the algorithmic computation of this 1.4 Kbps GELP coder is implemented using the C language and then transferred into the assembly code. In order not to exceed the limit of on-chip memory space, we truncate the algorithm particularly in the search of LSF codevectors and the determination of GCI's. Such truncation not only reduces the required program memory but also accelerates the processing speed. However, it also leads to noticeable degradation of perceivable quality. We expect that a full version of the GELP coder could be made to run in real-time on a floating-point DSP chip with large memory. Furthermore, improvement of the proposed coder may be achievable by incorporating a more accurate pitch estimator and a frequency-shaping noise source as well as tractable pulse dispersion. Apparently, the increase of the computational burden can only be resolved by using a faster DSP chip.

ACKNOWLEDGMENT

This research was supported by the National Science Council, Taiwan, ROC, under Grant NSC89-2213-E-197-002.

REFERENCES

1. Atal, B. S. and S. L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am., Vol. 50, No. 2, pp.637-655.
2. Wong, D. Y. (1980), "On understanding the quality problems", in Proc. *ICASSP*, Vol. III, pp. 725-728.

3. Kahn, M. and P. Garst (1983), “The effects of five voice characteristics on LPC quality”, in Proc. *ICASSP*, Vol. II, pp. 539-543.
4. Atal, B. S. and J. R. Remde (1982), “A new model of LPC excitation for producing natural-sounding speech at low bit rates”, in proc. *ICASSP*, pp. 614-617.
5. Singhal, S. and B. S. Atal (1989), “Amplitude optimization and pitch prediction in multipulse coders”, *IEEE* Trans., ASSP, Vol. 37, No. 3, pp. 317-327.
6. Kroon, K., E. Deprettere, and R. Sluyter (1986), “Regular-pulse excitation: A novel approach to effective and efficient multipulse coding of speech”, *IEEE* Trans., ASSP, Vol. 34, No. 10, pp. 1054-1063.
7. Zhang, S. and G. Lockhart (1995), “An embedded scheme for regular pulse excited (RPE) linear predictive coding”, in Proc. *ICASSP*, pp. 37-40.
8. Schroeder, M. and B. Atal (1985), “Code excited linear prediction (CELP): High quality speech at low bit rate”, in Proc. *ICASSP*, pp. 937-940.
9. Rose, R. C. and T. P. Barnwell III (1990), “Design and performance of an analysis-and-synthesis class of predictive speech coders”, *IEEE* Trans., ASSP, Vol 38, No. 9, pp. 1489-1503.
10. Kondoz, A. M. (1994), *Digital Speech* (*Coding for Low Bit Rate Communication Systems*), Wiley, England.
11. Rosenberg, A. E. (1971), “Effect of glottal pulse shape on the quality of natural vowels”, J. Acoust. Soc. Am., Vol. 49, No. 2, pp. 583-590.
12. Kang, G. S. and S. Everett (1985), “Improvement of the excitation source in the narrow-band linear prediction vocoder”, *IEEE* Trans., ASSP, Vol. 33, No. 2, pp. 377-386.
13. Childers, D. G. and C. K. Lee (1991), “Vocal quality factors: Analysis, synthesis, and perception”, J. Acoust. Soc. Am., Vol. 90, No. 5, pp. 2394-2410.
14. Childers, D. G. and H. T. Hu (1994), “Speech synthesis by glottal excited linear prediction”, J. Acoust. Soc. Am., Vol. 96, No. 4, pp. 2026-2036.
15. Ross, M. J., H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley (1974), “Average magnitude difference function pitch extractor”, *IEEE* Trans, ASSP, Vol. 22, pp. 353-362.
16. Hu, H. T. (1995), “Method for extracting epochal information from noisy LP residual”, Electron. Lett., Vol. 31, No. 25, pp. 2145-1247.
17. LeBlanc, W. P., B. Bhattacharya, S. A. Mahmoud, and V. Cuperman (1993), “Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding”, *IEEE* Trans., SAP, Vol. 1, No. 4, pp. 373-385.
18. McCree, A. and J. C. De Martin (1998), “A 1.7 Kb/s MELP coder with improved analysis and quantization”, in Proc. *ICASSP*, pp. 2565-2568.
19. Linde, Y., A. Buzo, and R. M. Gray (1981), “An algorithm for vector quantization design”, *IEEE* Trans., COM, Vol. 28, No. 1, pp. 84-95.
20. Ma, C. K. and C. K. Chan (1991), “Maximum descent method for image vector quantisation”, Electron. Lett., Vol. 27, No. 19, pp. 1772-1773.
21. McCree, A. V. and T. P. Barnwell III (1991), “A new mixed excitation LPC vocoder”, in Proc. *ICASSP*, pp. 593-596.
22. Chen, J. H. and A. Gersho (1995), “Adaptive postfiltering for quality enhancement of coded speech”, *IEEE* Trans., SAP, Vol. 3, No. 1, pp. 59-71.

23. Campbell, J. P., T. E. Tremain, and V. C. Welch (1991), “The federal standard 1016 4800 bps CELP Voice Coder”, *Digital Signal Process*, Vol. 1, No. 3, pp. 145-155.
24. Tremain, T. E. (1982), “The government standard linear predictive coding algorithm: LPC-10”, Speech Tech. Mag., pp. 40-49.