針對硬體實現之高效率且低位元率的語音編碼演算法

胡懷祖

國立宜蘭技術學院電子工程系教授

摘要

本文提出一個低位元率之語音編解碼演算法，其編碼結構源自於傳統的激源激發之線性預測語音編解碼器，它一方面根據發聲情況決定激源究係屬於喉門脈衝與隨機雜訊二者當中之何者，另方面則以預定之高通濾波器結合高效率法求得之八階 LSF 參數的餘弦函數值來界定頻譜特性。為了讓此一演算法適用於特定應用晶片的設計以及信號處理器之程式撰寫，演算法的繁雜度以及特別的算數運算都會被加上不少限制，而所獲得之演算法在執行 1.6Kbps 語音編解碼時仍能展現令人滿意的品質。據非正式聽力測驗顯示，聽者對於本文提出之編解碼器的喜好程度要比 2.4 Kbps LPC-10e 高得許多，但整體的音質表現仍稍遜於 4.8 Kbps CELP 與 2.4 Kbps MELP 的兩種語音編解碼器。

關鍵詞：低位元率語音編解碼、硬體實現、編碼演算法

An Efficient Algorithm for Hardware Implementation of Low-Bit-Rate Speech Coding

Hwai-Tsu Hu

Professor, Department of Electronic Engineering, National Ilan Institute of Technology

Abstract

This paper presents an algorithm for hardware implementation of low-bit-rate speech coding. The coding scheme emerges from traditional pitch-excited linear prediction vocoders. While the excitation switches between glottal pulses and random noise according to the voicing condition, the spectral properties are characterized by a prefixed highpass filter in combination with 8th-order cosine functions of LSF parameters, which are attainable by an efficient root-solving method. To make this algorithm suitable for hardware implementation either by ASIC design or by programming on a signal processor, constraints are imposed on the complexity of the algorithm as well as the need of special-purpose arithmetic operations. Nonetheless, the resulting algorithm is still capable of carrying out 1.6 Kbps speech coding with acceptable quality. Informal listening tests reveal that listeners exhibit evident preference for the proposed speech coder over the 2.4 Kbps LPC-10e vocoder, though the overall synthetic quality is somewhat inferior to that of the 4.8 Kbps CELP and 2.4 Kbps MELP vocoders.

Key Words: low-bit-rate speech coding, hardware implementation, coding algorithm

I. Introduction

In recent years, low-bit-rate speech coding techniques have drawn much attention due to its wide application to telecommunications and intellectual appliances. The phrase of “low bit rate” is used to signify the reduction of transmission bandwidth and memory storage, which promotes the efficiency of speech-relevant devices and facilities. Prevailing products consist of the cellular handset, videophone, VoIP, dialogic system, digital recorder, and digital answering machine, etc. To economize the cost of such products, a sensible choice would be the chip implementation so that all the processing steps are accomplished by integrated circuits and control logics. However, the design and development of specific integrated circuits are generally time-consuming. This not only increases the crucial time to market but also reduces productivity. An alternative way to carry out the task is to resort to the microprocessors or digital signal processors. The focus of concern thus shifts to the algorithmic feasibility and programming efficiency. Designers are asked to implement the targeted algorithm based upon the inherent features of a selected processor.

The mapping of coding algorithms into hardware is a work-intensive process, since the designer must understand how to adjust the underlying algorithm subject to multiple constraints. Generally encountered constraints include the convergence of the algorithm and the latency, throughput, and timing characteristics of hardware implementation. For ASIC-based implementation a well-developed algorithm should be also parameterizable, which means that the algorithm should be coded using a hardware description language like VHDL and Verilog. In consequence, two simple rules are borne in mind when we develop the coding algorithm. First, the algorithm should avoid special operators and functions unless they belong to the built-in functions of the targeted hardware. Second, the statements not realizable by synthesis tools must be excluded from the programming construction of the algorithm.

This paper aims at the provision of an efficient algorithm for hardware implementation of low-bit-rate speech coding. To make the developed algorithm portable to both processor-based and ASIC-based implementations, it is essential for us to trim the algorithm as much as possible and to take away special-purpose arithmetic operations as many as possible without causing severe degradation of overall performance.

II. Speech analysis

In this paper, we adopt a model called glottal excited linear prediction (GELP) [1-3] in order to incorporate glottal excitation into the linear prediction (LP) coder. As shown in Fig. 1, this model inherits the fundamental structure of the LP coder, which is developed based on the source-filter theory. The source is denoted by an excitation model switching between the voiced and unvoiced types according to the voicing condition. On the other hand, the filter is employed to characterize the spectral properties of the glottal flow, vocal tract transfer function, and lip radiation. Consequently, the analysis of speech signals is equivalent to extracting the modeling parameters such as pitch, gain, and filter coefficients. Techniques that constitute the framework of speech analysis for LPC vocoders involve the pitch detection, gain determination, and estimation of filter coefficients.

As mentioned earlier, the complexity of the coding scheme must be deliberately reduced with caution in order to accelerate the processing speed. One of the important alterations occurs in the order selection for the autocorrelation method while deriving filter coefficients. In contrast with most prevailing vocoders that employed a 10th order LP analysis, our proposed coder adopts an 8th order predictor. As we expect to put more emphasis on the formant structure than on the spectral slope, a first-order highpass filter, $1-0.925z^{-1}$, is introduced to enhance the spectral components in high-frequency regions. This makes the filter order to be 9 altogether. Since the order directly relates to the computational requirements, such an arrangement makes the derivation of LP coefficients somewhat easier. However, the largest profit from the order reduction resides in the conversion between the LP coefficients and LSF parameters. The popularity of LSF parameters for spectral representation results from its superiority in stability check, excellent interpolation properties, and relative insensitivity to quantization errors [4,5]. Notice that the derivation of LSF parameters involves a numerical method for finding the roots of two polynomials whose order is half of that of the LP predictor. As indicated by [6,7], only the polynomials whose order not greater than 4 can be solved through closed form formulas. The choice of the LP order as 8 is therefore to render a 4th order polynomial. Here we pursuit the polynomial roots using Ferrari's solution [6]. The transfer function of an 8th order linear predictor is given as

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_8 z^{-8}. \tag{1}$$

A symmetric polynomial $P(z)$ and an antisymmetric polynomial $Q(z)$ are defined based on $A(z)$ as

$$\begin{cases} P(z) = A(z) + z^{-(m+1)} A(z^{-1}) \\ Q(z) = A(z) - z^{-(m+1)} A(z^{-1}) \end{cases}, \quad m = 8. \tag{2}$$

Either $P(z)$ or $Q(z)$ can be rewritten as

$$\begin{aligned} R(z) &= z^{-9}\left[\left(z^4 + z^{-4}\right) + r_3\left(z^3 + z^{-3}\right) + r_2\left(z^2 + z^{-2}\right) + r_1\left(z + z^{-1}\right) + r_0\right] \\ &= z^{-9}\left[x^4 + r_3 x^3 + \left(r_2 - 4r_4\right)x^2 + \left(r_1 - 3r_3\right)x + r_0 - 2\left(r_2 - r_4\right)\right] \end{aligned} \tag{3}$$

where $x = z + z^{-1} = 2\cos w$, and w denotes the corresponding line spectrum frequency. We rewrite the expression within the brackets on the right hand side of Eq. (3) as

$$x^4 + ax^3 + bx^2 + cx + d = 0 \tag{4}$$

The root-solving procedure of the above equation is given as follows. Let $E = \frac{a}{2}$, $A = \sqrt{a^2/4 - b + y_1}$, and $B = \frac{Ey_1 - c}{A}$, where y_1 is one of the roots for the resolvent cubic equation $y^3 - by^2 + (ac - 4d)y - a^2 d + 4bd - c^2 = 0$. We obtain the value of y_1 using the modified Newton-Raphson method [8]. The number of iterations is set to 3, and the initial value is chosen as the minimum of $\left\{1.9, a^2/4 - b + 1.5\right\}$. Then, the four real roots of Eq. (4) become

$$z_{1,2} = \frac{-(A+E) \pm \sqrt{(A+E)^2 - 2(y_1 + B)}}{2} \tag{5}$$

$$z_{3,4} = \frac{(A-E) \pm \sqrt{(A-E)^2 - 2(y_1 + B)}}{2} \tag{6}$$

While the above formulas necessitate square root operations, a fast algorithm developed by Tommiska [9] is employed to execute the process. The adopted square-root operation needs only n clock cycles for $2n$-bit wide numbers.

Because the cosine functions of LSF parameters are directly applicable to the synthesis filter, we encode them into bits using a scalar quantizer according to the sequence, {4,3,4,3,4,3,4,3}. The well-known generalized Lloyd algorithm is used to obtain the optimal nonuniform quantizer [10]. Our training data consist of 68314 speech frames extracted from Mandarin sentences uttered by 10 speakers (5 males and 5 females). Another 22112 samples extracted from different speakers and sentences are also prepared for verifying the competence of the obtained quantizer. For all utterances, the sampling rate is 8 KHz and the size of analysis frame is 20 ms. Within each frame, the LSF parameters are converted from the 8th-order LP coefficients, which are derived from pre-emphasized speech signals. Table 1 presents the results in the spectral distortion between the actual and predicted cosine function of LSF parameters. It is shown that the interlacing strategy for bit assignment comes up with an average spectral distortion (SD) of 0.93 dB. Among the distortion measures from all the training data, only 1.45% of them exceed 2 dB and 0.0029% are beyond 4 dB. Moreover, we observe no significant difference for the SD's measured either inside or outside the training set. We therefore reach a conclusion that the proposed 28-bit spectral quantizer achieves a transparent quantization of spectral information. However, it ought to be noted that the bit assignment plays an important role in reducing the average spectral distortion. Given that the number of bits is fixed as 28, we have attempted other kinds of bit allocation but ended up with worse results. Such consequences can be best realized by inspecting Table 2, which gives the average spectral distortion measures between the original parameters and the quantized parameters. In the case under study, only one parameter was quantized at a time using a specific number of bits and the other parameters remain intact. It is evident in Table 2 that not only the SD measures with odd-indices (as counted from one) are larger than that with even indices, but also the improvement is rather significant for the quantization of odd parameters. This suggests the assignment of more bits to quantize odd-indexed cosine function of LSF parameters.

Besides the spectral analysis and quantization, there are two parameters pertaining to the speech production model, namely, the gain and the pitch period. As the method of gain determination copes with the structure of the synthesis filter, we leave relevant discussion in the next section and place emphasis merely on the pitch detection. Here we employ the average magnitude difference function (AMDF) to determine the pitch period. The correlation measure between the waveforms of adjacent pitch periods is then adopted to perform the voicing dichotomization. In order to increase the processing speed, the speech signal of the analyzing frame is decimated by a factor of 2. Before the decimation, the speech signal is fed into a first-order zero-phase lowpass filter, $\left|1+0.5z^{-1}\right|^2$, to suppress the aliasing effect. The segment involved in the computation of AMDF is 9 ms (or equivalently 36 samples) and it is searched from the side that possesses a larger magnitude. We obtain a tentative pitch period by identifying the location of the minimum of the $AMDF(k)'s$. A supplementary scrutiny across the $AMDF(k)'s$ is then brought in to prevent from pitch doubling or tripling. More specifically, we examine the $AMDF(k)'s$ from one quarter to one half of the tentative pitch period to see whether there exists a valley with its magnitude satisfying

(i) $AMDF(k) < M_{\min} + 0.2(M_{\max} + M_{\min})$; (7.1)

(ii) $AMDF(k) < AMDF(k+1)$ & $AMDF(k) < AMDF(k-1)$, (7.2)

where $M_{\max}$ and $M_{\min}$ represent the maximum and minimum values of the $AMDF(k)'s$. A new pitch period is selected as the index of the first encountered $AMDF(k)$ if the above conditions are met.

Following the pitch estimation, we derive the correlation coefficient between the selected pattern of the speech signal and the one with one period away. This coefficient along with the first LP coefficient and the average magnitude of the analyzing frame are used to classify the speech signal into two different categories, i.e., voiced and unvoiced speech. The underlying frame is classified as voiced speech whenever either one of the following four conditions is satisfied.

(i) $(d < 0.015Q_{\max})$; (8.1)

(ii) $(d < 0.03Q_{\max})$ & $(h > 0.125)$; (8.2)

(iii) $(a_1 > 0)$ & $(h > 0.25)$ (8.3)

(iv) $(h > 0.375)$ (8.4)

where $Q_{\max}$ denotes the maximal quantized value of the signal. h, a_1, and d correspond to the correlation coefficient, the first LP coefficient, and the square root of the segmental power, respectively. The exploitation of above-mentioned criteria stems from the following considerations. The correlation coefficient describes the waveform similarity between adjacent periods. The first LP coefficient reflects the spectral tilt, while the average magnitude relates to the volume intensity. All of them are known to be strong indicators of voicing conditions [11,12]. Notice that both a_1 and d are just intermediate outcomes during the analysis phase. No extra computation is needed for these two measures.

III. Coding Scheme

Given that the speech signal is sampled at 8 KHz, we update the analysis frame at a rate of 200 samples. The designed coding rate is 1.6 Kbps. Table 3 presents the detailed coding scheme. For each individual frame, the speech signal is pre-emphasized by a highpass filter, $1-0.925z^{-1}$. An 8th order LP analysis based on the filtered signal is performed using the autocorrelation method, which is formulated as Levinson-Durbin recursion. The LP coefficients are then converted into LSF parameters and coded using a 28-bit scalar quantizer as illustrated in section II.

Depending on the method used for the gain retrieval at the synthesis stage, the gain factor of a speech frame is characterized by the square root of either the power of speech signal or the power of the prediction residual. Here we only reserve 5 bits to quantize the estimated power. On the other hand, 7 bits are used to describe the pitch period with the zero value denotes the unvoiced speech and nonzero values, 1~127, correspond to an acceptable pitch period in the range of 21~147 samples. The number of bit required for each frame adds up to 40 in total, which renders into a coding rate of 1.6 Kbps.

Although the modeling of glottal phase may help improve the quality of synthetic speech, we do not preserve any bits to code the glottal characteristics. This is because the extraction of glottal phase often requires extensive computation, which impedes its application in hardware implementation. However, as the glottal features are contributory to the naturalness of synthetic speech, we imitate the glottal phase characteristics by employing a prototype of glottal pulse during the speech synthesis.

V. Speech synthesis

Synthetic speech is the result attained by feeding the excitation (either the glottal pulse or random noise) to a synthesis filter. During the synthesis stage, we interpolate the involving parameters to render a smooth transition. The interpolation is performed as the synthesis interval slides across subframes, each of them extending one-fourth of the entire frame. That is,

$$q^{k}=\begin{cases}0.875f+0.125j\,, & k=0\\ 0.625f+0.375j\,, & k=1\\ 0.375f+0.625j\,, & k=2\\ 0.125f+0.875j\,, & k=3\end{cases} \tag{9}$$

where f and j denote the decoded parameters of the previous and current frames, respectively. q^{k} is the interpolated parameter for synthesis in the *k*th subframe. The synthesis of unvoiced speech is straightforward since the excitation is accessible from a random number generator. The synthesis of voiced speech is rather complicated because we have to modulate the pitch period apart from replicating the glottal features. In this study, the voiced excitation, denominated as glottal pulse in the sequel, is obtained by passing a phase-dispersed impulse into a lowpass filter followed by a 20-point Hamming window to truncate the length exactly down to 20, which is the minimum allowable duration of pitch period sampling at 8 KHz. Zeros are then padded to the resulting glottal pulse whenever the pitch period is larger than 20 samples.

Although the gain for excitation usually does not attract much attention, errors in gain determination can seriously corrupt the synthesis quality. For example, severe energy fluctuations in synthetic speech may result in warblelike or harsh artifacts. The gain can be determined in several ways. Makhoul computed the gain of excitation, termed G, by referring to the linear prediction relation [13].

$$G=\sqrt{R(0)-\sum_{k=1}^{p}a_{k}R(k)} \tag{10}$$

where $R(k)'s$ are the autocorrelation function of the speech signal. The above derivation is straightforward, and more importantly it is very efficient as the value within the square root operator is simultaneously available with the linear prediction analysis of speech signals. It is recalled that we have chosen twice the amount of the cosine function of line spectral frequency as the object to quantize. The quantized value can be readily inserted in a filter structure like Fig. 2 to produce synthetic speech. The input of the synthesis filter is, of course, the excitation signal multiplied by the gain constant.

Another existing approach besides Eq. (10) is proposed by Atal and Hanauer, who derived the gain constant by matching the power of the original speech signals with that of the synthetic samples [14], i.e.,

$$P_r = \frac{1}{N}\sum_{k=0}^{N-1}(q(k)+Gf(k))^2 \quad , \tag{11}$$

where $q(k)$ and $f(k)$ represent the memory contribution of the previous frame and the filter response of the present excitation, respectively. P_r is the segmental power of the speech signal, and N denotes the frame length. The gain is obtained by solving a quadratic equation. If G is negative or complex, it is set to zero to clear the filter memory. However, a zero setting for excitation may cause pitch doubling. Hence Tohkura et al. suggested to damp the filter response to make the memory contribution ignorable [15]. This gives

$$G = \left(NP_r \bigg/ \sum_{k=0}^{N-1} f^2(k) \right)^{1/2} . \tag{12}$$

We note that Eq. (12) is an indirect approach because the gain can only be obtained subsequent to the acquisition of $f(k)$, which is usually accessible at the synthesis stage rather than at the analysis stage. Therefore we quantize the square root of power segment P_r and compute the gain via the quantized version of P_r. It is particularly noted that the square root operation plays a role of reducing the dynamic range of the segmental power and letting the quantization more accurate in a perceptual sense. From the viewpoint of computational efficacy, the inferiority of Eq. (12) to Eq. (11) is for sure since the excitation response and filter memory should be dealt with separately, leading to the computational amount twice as many. However, one should never overlook its potential advantage. For example, one of the common techniques used to enhance formants is the use of a postfilter [16] defined by

$$H(z) = \frac{A(z/b)}{A(z/a)}, \quad 0 < b < a < 1 \tag{13}$$

Unfortunately, the incorporation of such a postfilter modifies the formant intensity and eventually alters the amplitude of filtered output. This makes the energy of synthetic speech vary considerably even within the same frame. Other types of manipulation regarding the excitation may encounter similar problems as well. Hence a wise strategy for regulating the amplitude of the synthetic waveform is to let the amplitude adapted to the expected energy, which is the exact goal of Eq. (12). Taking all the filtering operations into account, we depict in Fig. 3 a composite filtering structure which includes the pre-emphasis filter, the all-pole linear prediction filter, and the pole-zero type postfilter constructed in a Direct Form II manner. Unlike the filtering structure presented in Fig. 2, in which the cosine function is directly fed into the filter, we convert the cosine function of LSF parameters back to LP coefficients before starting the filtering process. The readers may soon find that this conversion is worthy as compared to the computation demanded by the postfiltering.

Given that the gain factor is carefully governed, we are now allowed to modify the glottal excitation further without worrying about energy fluctuation. In this study, a mixed excitation which comprises lowpass filtered glottal pulses and highpass filtered white noise has been attempted. Fig. 4 presents a speech production model, which involves the mixed excitation and formant enhancement. The mixture ratio of the power between the pulse and noise is set as 0.25%, which is a typical value suggested by [17]. In fact, the mixed excitation is the merit of the MELP vocoder, which was selected as the new 2.4 Kbps Federal Standard speech coder by the United States Department of Defense in 1996 [18-20]. Significant quality improvement was reported due to the adaptation of such a mixed excitation.

VI. Listening Evaluation

Usually, a formal listening assessment requires a series of auditory evaluations on intelligibility, naturalness, recognizability, noise robustness, etc. [21]. Due to the constraints on both finance and schedule, the scope of our study was restricted to informal evaluation.

An informal listening test is performed on a total of thirty files corresponding to speech uttered by six different speakers (3 male, 3 female), each speaker delivering five sentences. Two versions of synthetic speech signals, which are produced by the synthesis models presented in Figs 1 and 3, are evaluated. The measured scores from the 2400 bps LPC-10e vocoder (FS-1015) [22], 4800 bps CELP coder (FS-1016) [23], and 2400 bps MELP coder are provided as three baselines.

Our results clearly indicate a preference for the proposed vocoder over the LPC-10e vocoder. The participation of mixed excitation and formant enhancement further improves the subjective performance. However, the averaged score associated with the proposed coder is slightly inferior to that of the MELP and CELP coder. According to the opinions gathered from the listeners, the proposed coder suffers noticeable quality degradation from erroneous pitch detection. This urges the need of a reliable and robust pitch detector in the analysis phase along with an efficient pitch smoothing method in the synthesis phase.

VII. Conclusions

This paper presents an efficient algorithm for hardware implementation of low-bit-rate speech coding at 1.6 Kbps. The proposed coding scheme inherits the nature of the pulse-excited LP vocoders. The excitation, which is selected between glottal pulses and white noise, is fed into an all-pole filter to synthesize the speech signal. Only one of the two excitations is active at a time. Depending on the method adopted to retrieve the gain, we encode the power of either the original speech signal or the prediction residual. The quantization with respect to the signal power allows the exploitation of mixed excitation and postfiltering, which leads to quality improvement at the cost of extra computational load.

To make the developed algorithm suitable for chip design and/or programming using a fixed-point signal processor, most of the algorithm is carried out using the fundamental arithmetic and bit-wise operations. The only exception is the square root function, which is nonetheless transferred into a series of bit-wise shifts and comparisons. The performance of the proposed 1.6 Kbps vocoder has been subjectively evaluated in comparison with the LPC-10e (FS-1015), CELP (FS-1016), and MELP coders. The results show that our proposed coder compares favorably with the LPC-10e vocoder but is slightly worse than the CELP and MELP coders. Currently, a simple version of the speech synthesizer based on the proposed vocoder has been implemented on the programmable logic devices distributed by the Altera Corp. A fully functional speech vocoder, which includes the whole sequence of analysis/encoding/decoding/synthesis, is also under development. Algorithmic adjustments and simplification of the proposed coder are the investigative topics in the future as well.

Acknowledgment

This research was supported by the National Science Council, Taiwan, R.O.C., under Grant NSC 90-2213-E-197-001.

References

1. D. G. Childers and H. T. Hu, (1994), “Speech synthesis by glottal excited linear prediction,” J. Acoust. Soc. Am., vol. 96, no. 4, pp. 2026-2036.

2. H. T. Hu and H. T. Wu, (2000), "A glottal-excited linear prediction (GELP) Model for low-bit-rate speech coding," Proc. Natl. Sci. Counc. ROC(A)., vol. 24, no. 2, pp. 134-142.
3. H. T. Hu, F. J. Kuo, and H. J. Wang, (2000), "A pseudo glottal excitation model for the linear prediction vocoder with speech signals coded at 1.6 kbps," IEICE Trans. Inf. & Syst., vol. 83, no. 8, pp. 1654-1661.
4. F. K. Soong and B. H. Juang, (1984), "Line spectrum pair and speech data compression," Proceedings of ICASSP, pp. 1.10.1-1.10.4.
5. F. K. Soong and B. H. Juang, (1993), "Optimal quantisation of LSP parameters," IEEE Trans. Speech Audio Process., vol. 1, no. 1, pp. 15-24.
6. G. A. Korn and T. M. Korn, (1968), Mathematical Handbook for Scientists and Engineers, New York: McGraw-Hill.
7. C. H. Wu and J. H. Chen, (1997), "A novel two-level method for the computation of the LSP frequencies using a decimation-in-degree algorithm," IEEE Trans. Speech, Audio Processing, vol. 5, no. 2, pp. 106-115.
8. P. Henrici, (1964), Elements of Numerical Analysis, New York: Wiley.
9. M. T. Tommiska, (2000), "Area-efficient implementation of a fast square root algorithm," Proceedings of the Third IEEE International Conference on Devices, Circuits and Systems.
10. Y. A. Linde, Y., A. Buzo, and R. M. Gray, (1981), "An algorithm for vector quantization design," IEEE Trans. Commun., vol. 28, no. 1, pp. 84-95.
11. B. S. Atal and L. Rabiner, (1976), "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. 24, pp. 201-212.
12. L. J. Siegel and A. C. Bessey, (1982), "Voiced/unvoiced/mixed excitation classification of speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. 30, pp. 451-460.
13. J. Makhoul, (1975), "Linear prediction: A Tutorial Review," Proc. IEEE, vol. 63, pp.561-580.
14. B. S. Atal and S. L. Hanauer, (1971), "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, no. 2, 637-655.
15. Y. Tohkura, F. Itakur, and S. Hashimoto, (1978), "Spectral smoothing technique in PARCOR speech analysis-synthesis," IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, no. 6, pp. 587-596.
16. J. H. Chen, and A. Gersho, (1995), "Adaptive postfiltering for quality enhancement of coded speech," IEEE Trans. Speech, Audio Processing, vol. 3, no. 1, pp. 59-71.
17. D. G. Childers and C. K. Lee, (1991), "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am., vol. 90, no. 5, 2394-2410.
18. A. V. McCree and Barnwell, T. P. (1995), "A mixed excitation LPC vocoder model for low bit rate speech coding," IEEE Trans. Speech, Audio Processing, vol. 3, no. 4, pp. 242-250.
19. A. McCree, Truong, K., George, E. B., Barnwell, T. P., and Viswanathan, V. (1996), "A 2.4 Kbps/s MELP coder candidate for the new U.S. federal standard," in Proc. IEEE Conf. ASP, pp. 200-203.
20. L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, (1997), "MELP: the new federal standard at 2400 bps," Proceedings of ICASSP, 1591-1594.
21. M. A. Kohler, L. M. Supplee, and T. E. Tremain, (1995), "Progress towards a new government standard 2400 bps voice coder," Proceedings of ICASSP, pp. 488-491.
22. T. E. Tremain, (1982), "The government standard linear predictive coding algorithm: LPC-10," Speech Tech. Mag., pp. 40-49.
23. J. P. Campbell, 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.

91 年 08 月 26 日投稿
91 年 09 月 23 日接受

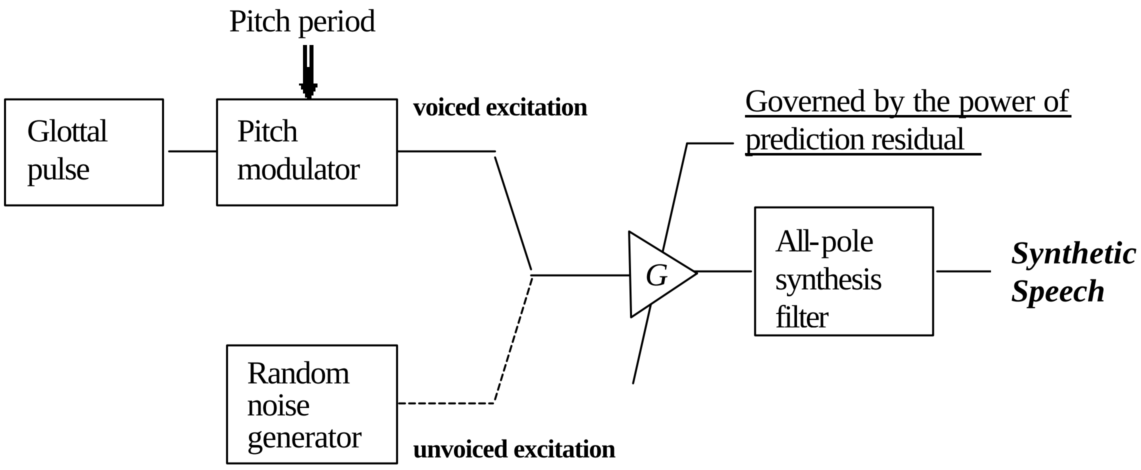

Fig 1 Block diagram of the traditional LPC vocoder.

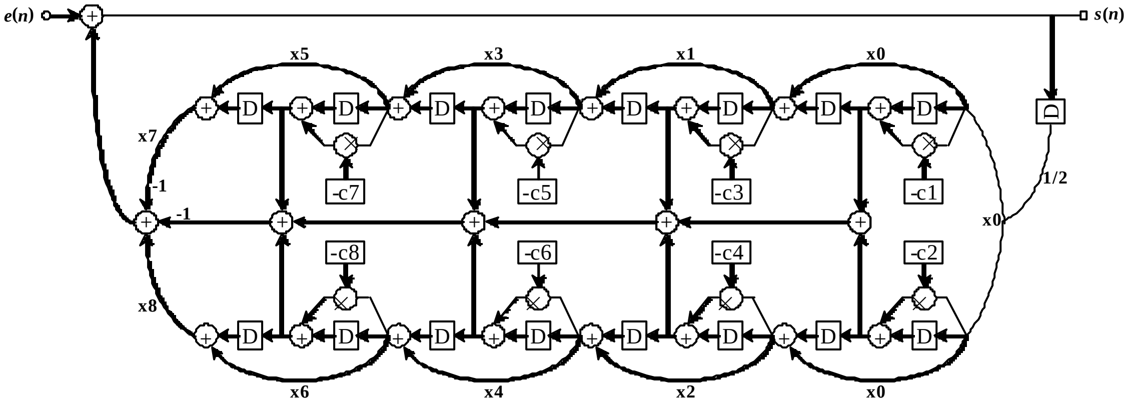

Fig 2 8th order LSP speech synthesis filter.

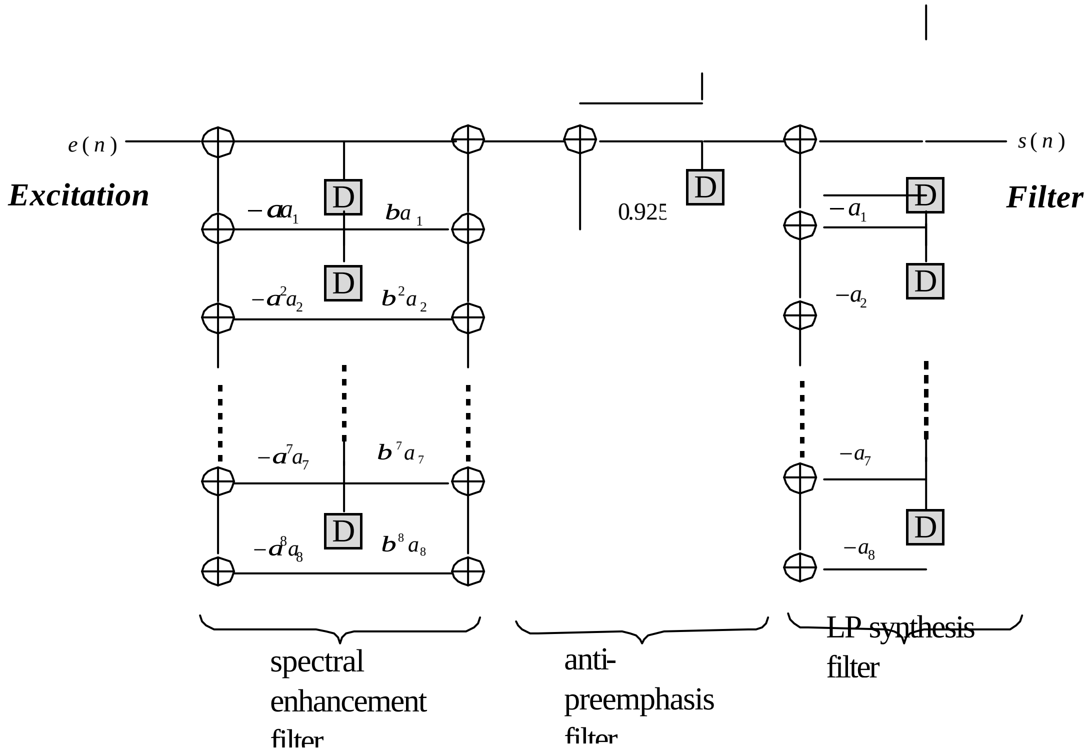

Fig 3 structure for the composite filter that includes pre-emphasis, all-pole filtering, spectral enhancement.

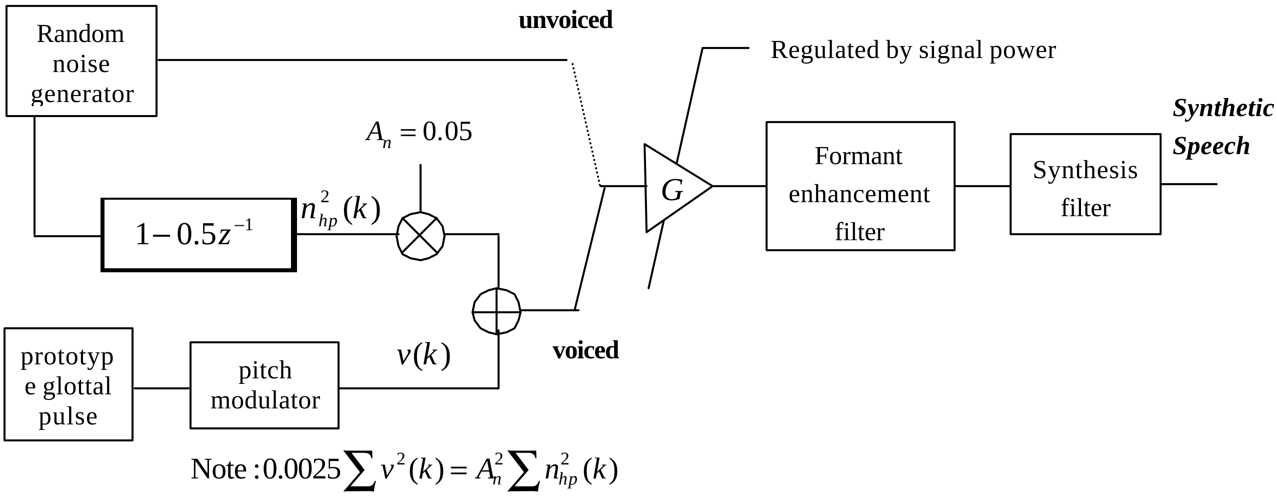

Fig 4 Speech production model with mixed excitation and formant enhancement included.

Table 1 Spectral distortion of the scalar quantizer with respect to various bit allocation.

Bit allocation	Data set	Spectral Distortion [dB]	Outliers (%)	
			2-4 dB	> 4 dB
{4,3,4,3,4,3,4,3}	within training	0.934	1.45	0.0029
	out-of-training	0.937	1.38	0.0000
{4,4,4,4,3,3,3,3}	within training	1.005	2.91	0.0498
	out-of-training	1.078	4.18	0.0814

Table 2 Influence due to quantization with respect to individual cosine function of line spectral frequency.

Parameter		1st	2nd	3rd	4th	5th	6th	7th	8th
Quantization error (spectral distortion, dB)	3 bits	0.3602	0.2792	0.3956	0.3775	0.4155	0.3572	0.3849	0.2643
	4 bits	0.1810	0.1488	0.2061	0.1931	0.2156	0.1835	0.1995	0.1362
Improvement due to an extra bit [dB]		0.1792	0.1305	0.1895	0.1845	0.1999	0.1736	0.1853	0.1281

Table 3 Bit assignment for the proposed 1.6 Kbps vocoder.

Sampling Rate: 8 KHz Frame Rate: 25 ms (200 samples/frame)	
Parameter	**bits/frame**
Voicing & Pitch	7
Gain	5
Spectrum	28
Total	40