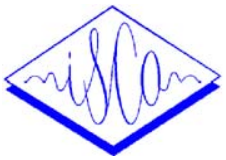

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

LOW BIT RATE SPEECH CODING USING AN IMPROVED HSX MODEL

Ridha Matmti, Milan Jelinek and Jean-Pierre Adoul

Department of Electrical Engineering, University of Sherbrooke
Sherbrooke, Québec, Canada, J1K 2R1
E-mail:ridha@gel.usherb.ca

ABSTRACT

This paper presents some improvements to the mixed Harmonic and Stochastic eXcitation (HSX) algorithm in the context of low bit rate speech coding (around 2.4 kbit/s). The dominant issue is the modeling of the excitation signal in order to improve the quality of the synthesized speech signal without increasing neither the bit rate nor the complexity. The pitch tracking algorithm is revised in order to increase the robustness and to reduce the complexity. The voicing analysis algorithm is also refined. Informal listening of the synthesized speech at 2.4 kbit/s shows a significant improvement.

1. INTRODUCTION

High quality speech coding at 4 kbit/s and below is of major interest in speech research. The demand for low bit rate speech coders for the telephone bandwidth (300-3400 Hz) is growing rapidly with the current and emerging applications like answering machines, paging, wireless communications, phone over Internet and with the interest in enhancing secure communications in government and military applications. The main motivations for low bit rates are the need to minimize storage and transmission costs, along with the demand to transmit over channels of limited capacity [1]. The new federal standard at 2.4 kbit/s selected by the United States Department of Defense (DoD) is based on the linear prediction coding (LPC) paradigm [2]. This proved that LPC is still an interesting method for low bit rate speech coding. Waveform coders such as CELP [3] are able to produce high quality speech at bit rates as low as 4.8 kbit/s. The strength of CELP coders resides in the use of the analysis-by-synthesis approach coupled with a time varying perceptual filter which is used to transform the speech signal into a space wherein the mean squared error measure is more significant to the human ear. However, a considerable amount of quantizing bits are assigned for modeling the phases of the speech waveform. As the bit rate is reduced to 4 kbit/s and below, the small number of available bits is not sufficient for matching the time domain waveform, and it is the reason why the CELP model fails at these bit rates. Recently, four candidate algotithms based on the CELP structure have been proposed for the ITU (International Telecommunication Union) 4 kbit/s standardization process. All of them failed to meet the requirements. One of the early submission was of a non-CELP type. The algorithm is based on Prototype Waveform Interpolation (PWI) [4] paradigm. It also failed to meet the requirements, but it demonstrated significant robustness against channel errors due to the fact that the coder parameters are highly independent of each other and from frame to frame unlike CELP coders, which are more sensitive to frame erasures, due to their inherent feedback structure. Generally, the recently developed non-CELP type algorithms, rely heavily on a correct pitch estimate which can require large delay to insure accuracy and to prevent halving or doubling the pitch. So, high quality low bit rate speech coding at 4 kbit/s and below is still a challenging problem for speech researchers. The recently proposed Harmonic Stochastic eXcitation (HSX) algorithm [5] demonstrates a good potential to produce comparatively high quality speech at 2.4 kbit/s. The basic idea consists of properly modeling the amplitude spectrum of the excitation signal at the input of the synthesis filter (all poles linear prediction filter). As the human perception is mainly related to the power spectrum of the input speech, a very good speech reproduction can be guaranteed provided the power spectrum is properly modeled.

All the parameters of the coder are estimated using an open loop analysis. Particular attention is taken to make sure that theses parameters represent, faithfully, the actual frame of the speech signal. This paper examines the effects of the introduced modifications on the excitation signal and insures that its general characteristics are closer to those of the prediction residual, which represents the ideal excitation signal for an LPC analysis/synthesis system.

The paper is organized as follows: section 2 presents an overview of the analysis part of the algorithm which includes linear prediction (analysis and quantization), pitch tracking and voicing estimation, spectral residual amplitudes (estimation and quantization) and energy estimation. The focus is on a 2.4 kbit/s implementation. Section 3 describes the synthesis part concentrating on the new low complexity and efficient synthesis methods for both the harmonic and stochastic components. In section 4, we give some insights of the performances of the algorithm and draw some conclusions.

2. OVERVIEW OF THE ANALYSIS

The block diagram of the analysis part is shown in figure 1. The frame size and the transmitted parameters

10.21437/Eurospeech.1997-194

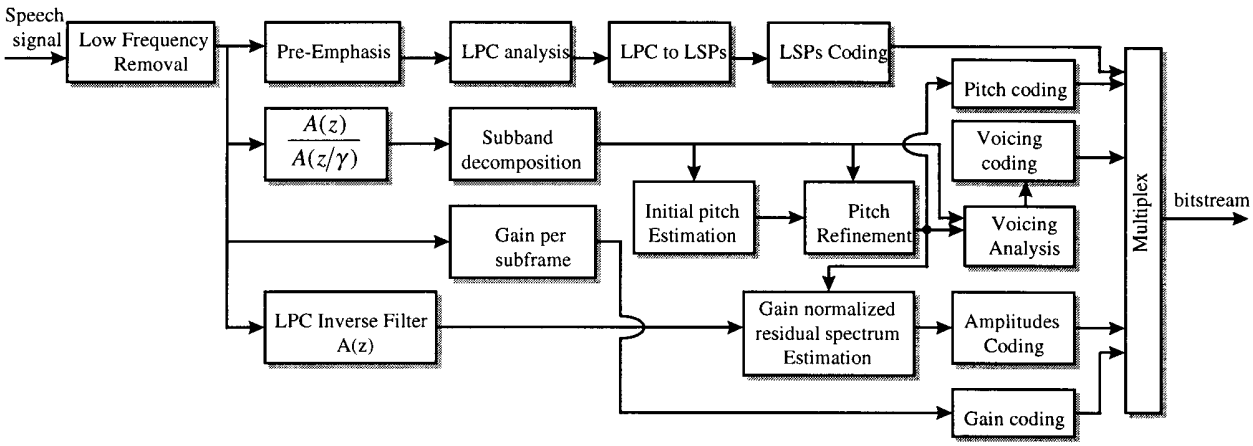

Figure1: Block diagram of the coder

depend on the bit rate and the delay requirement. In the following subsections, the different parts of the analysis algorithm are presented emphasizing the 2.4 kbit/s implementation and a frame size of 22.5 ms.

2.1 LP analysis and quantization

To estimate the LP parameters, the algorithm uses the autocorelation method of LPC analysis with a 30 ms Hamming window. Durbin's algorithm is used to evaluate the LP coefficients. In the first version of this algorithm, linear prediction is performed twice per speech frame. This results in an improved speech quality by insuring a better tracking of the varying speech spectrum. However, the quantization procedure of the LP parameters consumed the best part of the available bits (37 out of 53 in the case of the 2.4 kbit/s implementation). The new version uses a 14^{th} order LP filter for the entire speech frame in order to reduce the bit rate assigned to the LP quantization while retaining almost the same spectral accuracy. A 60 Hz bandwidth expansion is also performed in order to prevent sharp resonance in the spectral envelope. LP parameters are converted to Line Spectral Frequencies (LSFs). The LSF vector is split up into three sub-vectors of dimension, respectively, 4, 4 and 6. Each sub-vector is quantized with 9 bits using vector quantization with fixed 1^{st} order MA prediction and weighted LSF distance measure[6] .

2.2 Pitch tracking and voicing analysis

The residual signal is obtained by inverse filtering the input speech with the quantized LP filter interpolated four times per frame. By filtering the residual signal by the interpolated all-poles filter $1/A(z/\gamma)$, with γ typically equal to 0.85, the weighted speech signal is obtained. This latter signal was found more suitable than either the residual or the speech signal for tracking the pitch period due to the widened formants of a spectrum somehow between the residual and the speech spectra (depending on the value of γ). In this way we avoid some of interplay between pitch and formants. The pitch estimation procedure developed for this coding algorithm is based on an auto-correlation detector. Only the frequencies below 1000 Hz are used for the pitch search procedure. The pitch tracking algorithm is based on the normalized autocorrelation between the weighted speech frame and its delayed version in order to improve performance during energy transitions [8]. Normalization to the signal energy keeps the correlation values between -1 and 1 even during transition regions. In the initial pitch search procedure, the low pass weighted speech signal was downsampled by a factor of two in order to reduce the computational effort. In the refinement of the pitch period, a fine search is performed using an 1/8 sample resolution over a small range of pitch values around the initial pitch estimate. Also, an explicit check for pitch doubling and halving is included by using the past frame and one frame as look-ahead before deciding the final pitch value for the current frame. This results in a robust and reliable pitch follower. The pitch range of 19 to 143 is used and the lag is logarithmically quantized with 6 or 7 bits. At the receiver the pitch value is rounded to the nearest half-sample resolution.

In WI proposed by kleijn [9], there is no explicit checking of the voicing level, instead the prototypes are separated in Slowly Evolving Waveform (SEW) which represents the voicing part and Rapidly Evolving Waveform (REW) which represents the unvoiced part of speech signal. The major drawback of this method is the complexity and the extra delay needed to achieve this separation. In the proposed algorithm we don't use an explicit voiced/unvoiced classification, instead we use a voicing level in the form of is a continuous function of frequency. The voicing level is a monotonically decreasing function, that is, higher subbands are not permitted to be more voiced than lower subbands, which is in general in consistence with speech signal behavior. This allows for a limited number of bits needed for the voicing information.

The voicing level is determined using a subband decomposition of the weighted speech signal. In each subband, the degree of voicing is measured based on the normalized autocorrelation signal. The voicing measure is calculated over a range L which is a multiple of the

pitch period, $L = m\,T_0$. The parameter m depends on the maximum number N_{max} of signal samples available in the buffer (limited by the delay requirement). If T_{max} is the maximum delay, the constraint $mT_0 + T_{max} \le 2N_{max} + 1$ gives an appropriate estimate of the parameter m, $m = \mathrm{int}\{(2N_{max} + 1 - T_{max})\ /T_0\}$.

The correlation function is given by equation 1.

$$r_x(T) = \frac{\sum_{j=k_T}^{k_T+L-1} x_j x_{j-T}}{\sqrt{\sum_{j=k_T}^{k_T+L-1} x_j^2 \sqrt{\sum_{j=k_T+T}^{k_T+L+T-1} x_{j-T}^2}}} \qquad (1)$$

$$\text{where; } k_T = \mathrm{int}\{(T+L)/2\}$$

This method helps to maintain a significant correlation value even in the higher frequency bands where the peaks of the correlation function tends to be sharper and the true maximum peak could be missed due to the lack of time resolution. An upsampling operation is applied to the correlation function before choosing its maximum peak. The entire frequency band is divided in 7 frequency transition band each of 500 Hz. The voicing information is quantized using 3 bits. It indicates the frequency at which the excitation starts to be unvoiced.

2.3 Speech energy

The speech frame is divided into 4 subframes. The pitch-synchronous energy is computed in each subframe, and the energy per pitch period per sample is obtained. The four energies are vector quantized in the logarithmic domain using 7 bits.

2.4 Residual spectrum estimation and quantization

In early version of this work, the residual power spectrum was assumed completely flat and the power spectrum of the speech signal was described only by the LP model. A more accurate representation of the low frequency spectrum helps to improve significantly the quality of the synthesized speech, especially in the case of voiced speech. The number of points needed to represent the magnitude spectrum of the residual signal is a set of spectral magnitudes estimated by a pitch-synchronous DFT (or a multiple of the pitch period when a fine frequency resolution is necessary). As a result, the number of harmonics depends on the frequency resolution. In order to vector quantize the residual spectrum, a special procedure was developed to convert the estimated magnitudes into a fixed dimension vector. An interpolation/decimation procedure is applied in order to maintain one harmonic every 100 Hz. Only the first ten harmonics representing the 1000 Hz lower band are retained for transmission and are vector qunatized using 9 bits.

3. SYNTHESIS PROCEDURE

The aim of the synthesis procedure is to reproduce a speech signal as perceptually close as possible to the original speech signal. This means that from the received parameters we should be able to reproduce the main perceptual features of the synthesized signal. An accurate reproduction of the characteristics of the excitation signal is essential since the human ear is able to decide if a synthesized speech segment is either buzzy or noisy. In the HSX algorithm implementation, the received parameters that represent the signal in the middle of the presents frame are interpolated with the ones of the past frame throughout the synthesis procedure. In the frequency domain, the excitation can be seen as a combination of a harmonic component and a stochastic component ranging over all the frequency band. The normalized excitation spectrum can be represented by:

$$E_w(\omega) = \alpha(\omega) V_w(\omega) + [1 - \alpha(\omega)] U_w(\omega) \quad 0 \le \alpha(\omega) \le 1 \qquad (2)$$

The parameter $\alpha(\omega)$ represents the voicing level or the "degree" of voicing at frequency ω .

The harmonic (voiced) component can be generated in different ways: either by the sum of harmonic generators (oscillators) at the multiples of the pitch frequency [10, 11] or by the filtering of a train of pitch pulses through a bank of band pass filters [12]. The last approach has the advantage of simplicity since the excitation is given by the addition of the weighted impulse responses of the different filters. Although this method gives good results, it suffers from errors in pitch pulse locations and from a lack of consistence in timing alignment across different pitch periods. This difficulty occurs especially in the case of female or children speakers, where the shape of glottal impulse can be modified due to the short pitch period. In order to reduce the complexity problem, some researchers [13] proposed an approximation of the harmonic component by using an Inverse Fast Fourier Transform (IFFT). To maintain a good quality and avoid complexity, we propose to use a combination of the two methods, namely the harmonic oscillators and the bank of band pass filters, depending on the pitch value. As the pitch range is from 14 to 140 samples, the idea is to keep the same model of synthesis if there is not a significant change in the pitch value: the harmonic model for lower pitch period speakers and a bank of band pass filters for higher pitch period speakers. A reasonable value of the pitch transition is around 40 ± 4 samples. We have to insure a soft transition from one model to the other. The last pitch pulse location before the significant change of the pitch is identified and special measures are taken to insure that the next pitch pulse is in the right location.

The unvoiced contribution of the excitation signal is generated using inverse Short Time Fourier Transform (STFT) with the overlap and add method. The noise spectrum is conditioned based on the voicing information

and an interpolated version of the transmitted lower band gain normalized spectrum.

The gain of the excitation signal is estimated from the received gain vector on a subframe basis. Fisrt, the gain of the combination of synthesis filter and postfilter is calculated and the excitation gain is determined by dividing the subframe energy by the filters gain.

The gain-scaled excitation signal is then filtered through the synthesis filter and postfilter in order to find the synthesized speech signal. The postfilter is similar to that used at higher rate CELP coders [14]. At the output of the postfilter, an automatic gain control scaling is used to enforce the energy of the output speech signal to be close to the transmitted energy.

4. PERFORMANCE OF THE ALGORITHM

The bit allocation of the 2400 bit/s version of the improved algorithm is given in table 1. An informal comparison between the new USA Department of Defense (DoD) at 2400 bit/s and the proposed algorithm shows that the performance of the two algorithms is comparable. The proposed algorithm sounds slightly better when the input speech signal is corrupted by a background noise.

.

Parameters	Bits/frame
LSF parameters	27
residual spectrum	9
Gain (4 per frame)	8
pitch	6
voicing	3
Synchronization	1
54 bits per 22.5 ms =2.4 kbit/s	

Table1: Bit allocation of the algorithm at 2.4 kbit/s implementation

5. CONCLUSION

We have presented in this paper an improved version of the HSX low bit rate speech coding model. The proposed hybrid voiced excitation synthesis maintains a good quality with a reduced complexity.

6. ACKNOWLEDGMENT

The authors would like to thank Mr. Claude Laflamme for his many suggestions and fruitful discussions and Dr. Redwan Salami for his technical assitance.

REFERENCES

[1] Jayant, "Signal Compression: Technology Targets and Research Directions", IEEE Journal on Selected Areas in Communications, Vol. 10, n°5, pp.796-818, June 1992.

[2] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding", IEEE Trans. Speech and Audio Processing, vol. 3, no. 4, 242-250, July 1995.

[3] M. R. Schroder and B. S. Atal, "Coded-excited linear prediction (CELP): high-quality speech at low bit rates," Proc. ICASSP'85, pp. 937-940.

[4] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms", IEEE Transactions on Speech and Audio Processing, vol. 1, n° 4, pp.386-399, October 1993.

[5] C. Laflamme, R. Salami, R. Matmti and Adoul, "Harmonic-Stochastic eXitation (HSX) speech coding below 4 kbit/s", Proc. ICASSP-96, pp 204-207, 1996.

[6] K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC parameters at 24 Bits/frame", IEEE Tran. Acoust. Speech and Audio Processing, vol. 1, no , pp. 3-14, January 1993.

[7] V. Cuperman, P. Lupini and B. Bhattacharya "Spectral Excitation coding of speech at 2.4 Kb/s", Proc. ICASSP-95, pp 496-499, 1995.

[8] B.S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria ", IEEE Tran. Acoust. Speech and Signal Processing, vol. 27, pp. 247-254, June 1979.

[9] W. B. Kleijn and Jesper Haagen, "A Speech Coder Based on Decomposition of Characteristic Waveforms", Proc. ICASSP'95, pp. 508-511.

[10] M. Brandstein, J. Hardwick and J. Lim, "The Multibannd excitation speech coder", in advances in Speech Coding , B. S. Atal et al. Eds., Kluwer Academic Publishers, 1991, pp. 215-223.

[11] R. McAulay, T. Parks and M. Sabin "Sine-wave amplitude coding at low data rates ", in advances in Speech Coding , B. S. Atal et al. Eds., Kluwer Academic Publishers, pp. 203-213, 1991.

[12] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding", IEEE Trans. Speech and Audio Processing, vol. 3, no 4, pp. 242-250, July 1995.

[13] M. Nishiguchi and J. Matshumoto, "Harmonic and noise coding of LPC residuals with classified vector quantization", Proc. ICASSP-95, pp. 484-487, 1995.

[14] R. Salami, C. Laflamme, J-P Adoul and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications systems (PCS) " IEEE Trans. Veh. Technol., vol.43, no 3, pp. 808-816, Aug. 1994.