A Variable Rate Speech Compressor for Mobile Applications

S. Yeldener A. M. Kondoz B. G. Evans

Department of Electronics and Electrical Engineering
University of Surrey
Guildford, Surrey, GU2 5XH.

ABSTRACT

One of the most promising speech coder at the bit rate of 9.6 - 4.8 kbits/s is CELP [1]. CELP has been dominating 9.6 to 4.8 kbits/s region during the past 3 to 4 years, its set back however, is its expensive implementation. As an alternative to CELP, we have developed the Base-Band CELP (CELP-BB) [2] which produced good quality speech comparable to CELP and a single chip implementable complexity as reported in [3]. We have also improved its robustness to tolerate errors up to 1.0% and maintain intelligibility up to 5.0% and more [4]. Although, CELP-BB produces good quality speech at around 4.8 kbits/s, it has a fundamental problem when updating the pitch filter memory. We proposed a sub-optimal solution to this problem in this paper. Below 4.8 kbits/s, however, CELP-BB suffers from noticeable quantization noise as result of the large vector dimensions used. In this paper, therefore, we also report on efficient representation of speech below 4.8 kbits/s by introducing Sinusoidal Transform Coding (STC) [5] to represent the LPC excitation which is called Sine Wave Excited LPC (SWELP). In this case natural sounding good quality synthetic speech is obtained at around 2.4 kbits/s.

1. INTRODUCTION

The type of speech compression technique is very important for maritime and land mobile satellite communication systems. For these systems, the resources are very limited interms of the very small transceiver terminals requiring larger satellite power, and the very restricted bandwidth currently available. This especially applies to the land mobile satellite service which currently has only 4 MHz allocated on primary service transmission. For such services to be economic, they must employ very narrow bandwidth per channel. The competition is with analogue systems that employ Amplitude Companded Single Side Band (ACSSB) and achieve reasonable performance at C / N_o's of around 50 dB-Hz in 5 kHz transmitted bandwidth. Now in order to be competitive and to use modulation schemes that will not cause excessive distortion over the difficult land mobile propagation channel, digital speech coding of around 4.8 kbits/s or less is required. The performance must be better than the analogue contender in worst case degraded channels and the speech quality must be acceptable enough to be connected on to PSTN transmission.

With Land mobile satellite systems (INMARSAT, AVSAT, MSAT, etc.) proposing to operate voice services soon, the race is on to produce digital speech coders that can meet all the requirements. Until now Analysis By Synthesis (ABS) schemes such as CELP have only achieved these qualities down to 6 kbits/s. Its quality at 4.8 kbits/s can also be made adequate by post filtering the recovered speech. However, below 4.8 kbits/s, most of these schemes suffer from noticeable quantization noise as result of the large vector dimensions used. Another major disadvantage with these schemes is that they are not really practical for real time implementation owning to enormous computational demand. For the land mobile service we are looking for a speech coder whose cost is a fraction of the mobile terminal, thus we need a scheme which is simple to implement in a single DSP chip. The CELP-BB [2] satisfies these requirements, however below 4.8 kbits/s it suffers from noticeable quantization noise as mentioned earlier. In this paper we present results of a coder called Sine Wave Excited LPC (SWELP) which we believe is

capable of meeting all the requirements and thus is a serious contender for future land mobile satellite systems.

2. OVERVIEW OF CELP-BB SYSTEM

In CELP-BB system, a block of speech, $S(n)$ is first analized and 10 LPC coefficients are computed. These are transformed into Line Spectral Frequencies (LSF) and then quantized [4]. The quantized LSF are inverse transformed into LPC parameters which are then used to form an inverse filter to derive the LPC residual signal. The LPC residual signal is then divided into sub-blocks, each of which is filtered by the weighting filter (smoothing filter) separately. Filtered sub-blocks are split into a number of sequences equal to the decimation factor. These sequences are compared in terms of their energies and the one with the highest energy is selected for transmission. The position of the selected sequence in each sub-block is also transmitted to the decoder in order to place the sequence in the correct location in High Frequency Regeneration (HFR). The selected decimated signal is then treated as a continuous signal which now contains decimation factor times less samples than the original. This continuous signal is then quantized via a restricted ABS procedure operating around merely a pitch synthesis filter. First the pitch filter delay and gain of the continuous signal are computed by cross correlation with the past decoded samples in the pitch synthesis filter. Using these parameters in a pitch synthesis filter the memory response of the filter is computed and subtracted from the decimated signal to form the reference signal. Gaussian code-book sequences are then searched one by one to match the output response of the pitch synthesis filter with no memory, to the reference signal. The index of the optimum sequence, together with its scale value are transmitted to the decoder. At the decoder, selected code-book sequences are scaled up by their scale factors and passed through the pitch synthesis filter in order to recover the continuous decimated base-band signal. The recovered signal is then sub segmented and shifted to the correct positions with zeros inserted in between the samples, to form the LPC filter excitation sequence. Using this sequence the LPC synthesis filter is excited to recover the output speech.

2.1 The Performance Of CELP-BB

Although the encoder of CELP-BB appears very similar to CELP [1], it in fact offers one very significant advantage, namely simplicity. As the ABS procedure operates on the base-band, the vector dimensions are reduced by a factor of decimation. Also the ABS procedure is restricted around the pitch synthesis filter and noise shaping filter. This eliminates considerable amount of convolution computations required by the LPC filter as in standard CELP. These two differences contribute to the enormous savings in computations. Although, CELP-BB is much simpler coding scheme, its speech quality remains to be comparable to CELP [1].

Although, CELP-BB produces good quality speech at around 4.8 kbits/s with a single chip implementable complexity [3], it has a fundamental problem updating the pitch filter memory. The worst case situation arises, when the first and the last (third) decimated sequences are selected in two consequative sub-blocks as shown in figure 2.1.

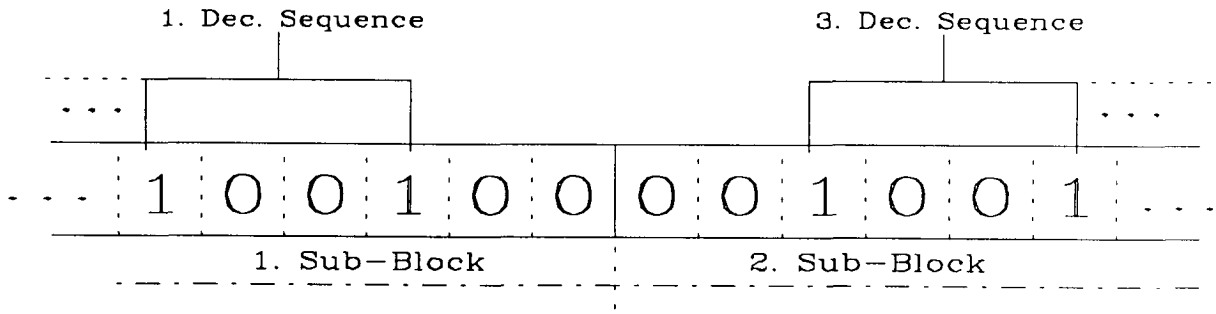

Figure 2.1 The representation of selected sequences in worst case.

In this figure, 1's represent the selected samples of each sub-block having the highest energy

comparing with the other sequences which are represented by 0's. In this case, The first and the third decimated sequences are selected for transmission for the first sub-block and second sub-block respectively. When the selected sequences are treated as a continuous signal which each consists of decimation factor times less samples than the original, in the worst case as shown in figure 2.1, an offset time occurs in the third sequence of the second sub-block. Since the pitch filter operates on the decimated sequence, this situation disturbes the pitch structure. Therefore, a method has to be found to update the pitch filter memory efficiently and accurately. One solution for this problem is to interpolate the decimated sequences in the pitch filter memory. With this strategy, pitch delay can be calculated by maximizing the equation 2.1 as,

$$E(p)=\sum_{i=1}^{N/D} h(iD+p)\,r(i) \qquad (2.1)$$

where N is the sub-block size, D is the decimation factor, p represents the pitch delay which ranges from 20 to 147 samples, $h(i)$ is the interpolated samples of pitch filter memory and $r(i)$ is the selected decimated sequence for the updated sub-block. In this case, although, the performance improvement is dependent on the accuracy of interpolation, we have noticed partial improvement with a simple interpolation schemes.

Below 4.8 kbits/s, however, CELP-BB suffers from noticeable quantization noise, back ground noise and hence roughness as result of the large vector dimensions used. The reason for this is that during vector quantization of excitation, all of the components are matched as a single vector. This results in poor matching performance for larger vector dimensions. Therefore, we have introduced the Sinusoidal Transform Coding (STC) [5] to represent the LPC excitation. Whit this it is possible to code speech below 4.8 kbits/s with a minimum quality loss. This system is called Sine Wave Excited LPC (SWELP) The results of this system is addressed in next sections.

3. SWELP CODING SYSTEM

In the speech production model, the speech waveform, $S(n)$ is assumed to be the output of passing a glottal excitation waveform, $r(n)$ through a linear time-varying system with impulse response $h(n)$ representing the characteristics of the vocal tract. Mathematically, this can be written as,

$$S(n)=\sum_{m=0}^{n} r(m)h(n-m) \qquad (3.1)$$

The excitation will be represented by a sum of sine waves of arbitrary amplitudes, frequencies and phases,

$$r_m(n)=\sum_{k=1}^{L_m} A_k^m\,Cos(n\,\omega_k^m+\Phi_k^m) \qquad (3.2)$$

where L_m is the number of sine-waves and A_k^m, ω_k^m and Φ_k^m are the amplitude, frequency and phase respectively for the k^{th} sine-wave components at the m^{th} frame.

A block diagram of the SWELP Analysis/Synthesis system is given in figure 3.1 which operates as follows:

3.1 Analysis

A block of speech, $S(n)$ is first LPC analized to obtain the LPC coefficients. These coefficients are then quantized. The quantized coefficients are used to form an inverse filter to derive the LPC excitation in sub-blocks, $r(n)$

$$r(n)=S(n)-\sum_{k=1}^{p} a_k S(n-k) \qquad (3.3)$$

where a_k's are the LPC filter coefficients and p is the order of the filter. The spectrum of LPC excitation is then analized using a 512 point FFT and a hamming window having a minimum width of 2.5 times the average pitch for accurate peak estimation. The spectrum, $R(\omega_k)$ can be computed as,

$$R(\omega_k)=\sum_{n=0}^{N-1} r(n)W(n)e^{-jn\omega_k} \qquad (3.4)$$

where $W(n)$ is a hamming window and ω_k's are the discrete frequencies ($\omega_k=\frac{2\pi k}{N}$).

The number of peaks L_m is typically about 40 - 50 over a 4 kHz range. The maximum number of peaks that can be specified is limited by a threshold that is also function of the measured average fundamental. In general, the performance can be affected by the choice of this

threshold only when too few peaks were allowed. The locations of the largest peaks are estimated by simply searching for a change of slope from positive to negative in the uniformly spaced samples of the short-time Fourier transform magnitude, ($|R(\omega_k)|$). The amplitude and phase components (modula 2π) of the sine waves are given by the appropriate samples of the high resolution FFT corresponding to $R(\omega_k)$ at the chosen frequency locations.

3.2 Synthesis

The LPC coefficients and the set of amplitudes, frequencies and phases which are estimated in the encoder, are transmitted to the decoder. Received set of amplitudes, frequencies and phases are used to generate the sine waves for each tone. The generated sine waves are then added together to form the LPC excitation, $\hat{r}(n)$ using equation 3.2. The final quantized speech, $\hat{S}(n)$ is then obtained by passing the recovered LPC excitation through the LPC filter,

$$\hat{S}(n)=\hat{r}(n)+\sum_{k=1}^{p} a_k \hat{S}(n-k) \qquad (3.5)$$

Due to the time-varying nature of the parameters, however, this straightforward approach leads to discontinuities at the frame boundaries, if this approach is directly applied to speech as in [5]. Therefore, a method was found which smoothly interpolates the parameters measured from one block to those that are obtained in the next. Although, recovered speech in [5] is interpolated, this interpolation procedure introduces some back ground noise and block edge effects in the recovered speech. In SWELP system, we used the LPC filter to interpolate the sine wave components. This way all the discontinuities were eliminated from the recovered speech with the cost of coding the LPC coefficients.

4. BIT RATE REDUCTION STRATEGIES

Since the parameters of the SWELP system are the LPC coefficients, amplitudes, frequencies and phases of the underlying sine waves, and since for a typical low-pitched speaker there can be as many as 80 sine waves in a 4 kHz speech bandwidth, it is not possible to code all of the parameters directly. Therefore, an important focus of this work has been on techniques for efficient coding of the model parameters. The first step in reducing the size of the parameter set to be coded, was to develop a pitch extraction algorithm, which leads to a harmonic set of sine waves. The computed harmonic set is perceptual best fit to the measured sine waves. With this strategy, coding of individual sine wave frequencies is avoided. A new set of sine wave amplitudes and phases is then obtained by sampling an amplitude and phase envelop at the pitch harmonics.

In addition to the development of pitch extraction algorithm which led to a harmonic set of sine waves, a predictive model for the phases of sine waves was developed. The model given in [6] is quite accurate during steady voiced speech, while during unvoiced speech, it is poor, resulting in phase excitations that appeared to be random values within $[-\pi,\pi]$. For this purpose, we have developed another phase prediction model which works in both voiced and unvoiced speech regions. Since the speech coder is independent of v/uv decision, another parameter which is called error compensation component, has to be coded and transmitted to the receiver. The recovered LPC excitation was then presented as,

$$\hat{r}(n)=\sum_{k=0}^{L_m} A_k^m \cos((n-n_o)k\omega_o+\Phi_k) \quad (4.1)$$

where $\Phi_k=(-\Phi_o)^{k+1}$ is the phase and frequency error compensation component, n_o is the onset time of the pitch pulse [6] and ω_o is the fundamental frequency. Experiments showed that during steady voiced speech, Φ_o was automatically selected as zero, on the other hand, during unvoiced speech ${\Phi_o}^{k+1}$ took a value between $[-\pi,\pi]$. Since the amplitudes of the LPC excitation sine waves are more or less flat, a good criterion to use is the minimum mean-squared error for seeking the optimum values of n_o and Φ_o.

A block diagram of the complete analysis/synthesis system is given in figure 3.1. A non-real time floating point simulation was developed in order to determine the effectiveness of the proposed approach in modeling real speech. In SWELP system, the LPC and Spectrum analysis took place on block by block and sub-block by sub-block basis respectively. In LPC analysis using 30 ms (240 sample) block intervals (each consists of 2 sub-blocks), 7 or 8 LPC coefficients was found to be sufficient for smoothly interpolating the sine wave

components. A 512 point DFT using a 20-22 ms Hamming window was found to be sufficient for accurate peak estimation for each sub-block of LPC excitation. The overall bit rate of SWELP system is chiefly determined by allocating a certain number of bits for the LPC coefficients and sine wave components for each block of speech. We therefore, feel that by allocating more sine wave components to the excitation representation, the overall bit rate can be varied from 2.4 to 4.8 kbits/s or higher. This also varies the quality of speech. Thus, this scheme can easily be operated in a variable rate environment if required. A 2.4 kbits/s SWELP was simulated using vector quantization for both LPC coefficients and amplitude components of sine waves. The bit allocation for the coder implementation is shown in table 4.1, and the waveforms of the original and decoded speech are shown in figure 4.1. In this case, 2.4 kbits/s SWELP system produces natural sounding good quality synthetic speech.

Parameter	Bits Per Frame	Bit Rate
LPC Coeff.	13	500.0
Fund. Freq.	14	400.0
Phases (n_0, Φ_0)	18	600.0
Amplitudes	27	900.0
Overall	72	2400.0

Table 4.1 Bit allocations for 2.4 kbits/s SWELP coder.

5. CONCLUSION

In this paper, the advantages of CELP-BB, its problems and solutions were examined. We saw that the interpolation of the decimated sequences in the pitch filter memory improved the subjective performance of CELP-BB system. Below 4.8 kbits/s, however, it seemed that efficient representation of the model parameters for good quality speech is very difficult. Therefore, we presented the results of SWELP system which is used to represent the LPC excitation at very low bit rates (around 2.4 kbits/s) and produced good quality speech. The strategies for bit rate reduction in transmission parameters were described. Depending on the detailed bit allocation rules, operation at rates from 2.4 to 9.6 kbits/s can be obtained with the variation of speech quality. Thus, this scheme can easily be operated in a variable rate environment if required. Finally, a 2.4 kbits/s SWELP system was simulated and bit allocation of its parameters has been given as an example. The speech produced by this system was found very intelligible and natural sounding. Currently, the complete analysis/synthesis blocks of the SWELP scheme is being modified to form an Analysis By Synthesis (ABS) select procedure for reduced set of amplitudes, frequencies and phases. Results of this modification will be published later.

6. REFERENCES

[1] M. R. Schroeder, B. S. Atal "Code Excited Linear Prediction (CELP): High quality speech at very low bit rates" Proc. of ICASSP-85 pp.937-940.

[2] A. M. Kondoz, B. G. Evans "CELP base-band coder for high quality speech coding at 9.6 to 2.4 kbits/s" Proc. of ICASSP-88 pp.159-162.

[3] G. H. Asjadi, A. M. Kondoz, B. G. Evans "A real-time implemented 8 kbits/s CELP base-band coder" 7. Fase Symposium on speech, Edinburgh, 1988, pp.1039-1042.

[4] S. A. Atungsiri, A. M. Kondoz B. G. Evans "Robust 4.8 kbits/s CELP-BB speech coder for satellite-land mobile communications ", First European Conf. on Satellite Communications, September, 1989, W. Germany.

[5] R. J. McAulay and T. F. Quatieri "Speech analysis/synthesis based on a sinusoidal representation" IEEE trans. ASSP-34, pp.744-754 (August 1986).

[6] R. J. McAulay and T. F. Quatieri "Phase modeling and its application to Sinusoidal Transform Coding" IEEE proc. Int. conf. on ASSP, Tokyo, Japan, April 1986.

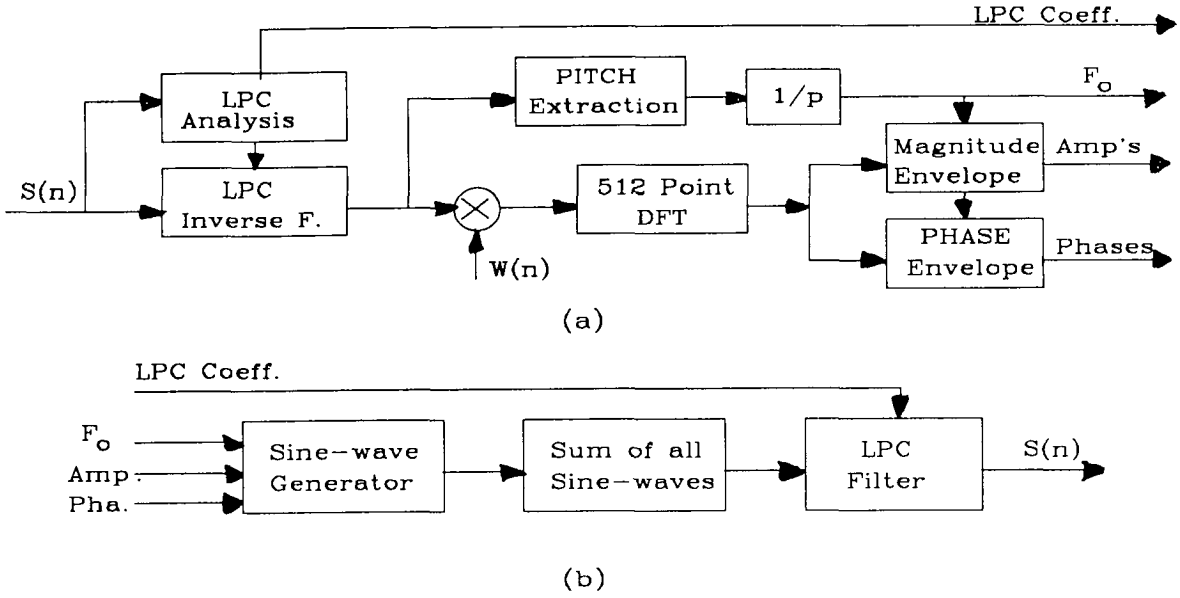

Figure 3.1 The block diagram of SWELP system a) Analysis b) Synthesis.

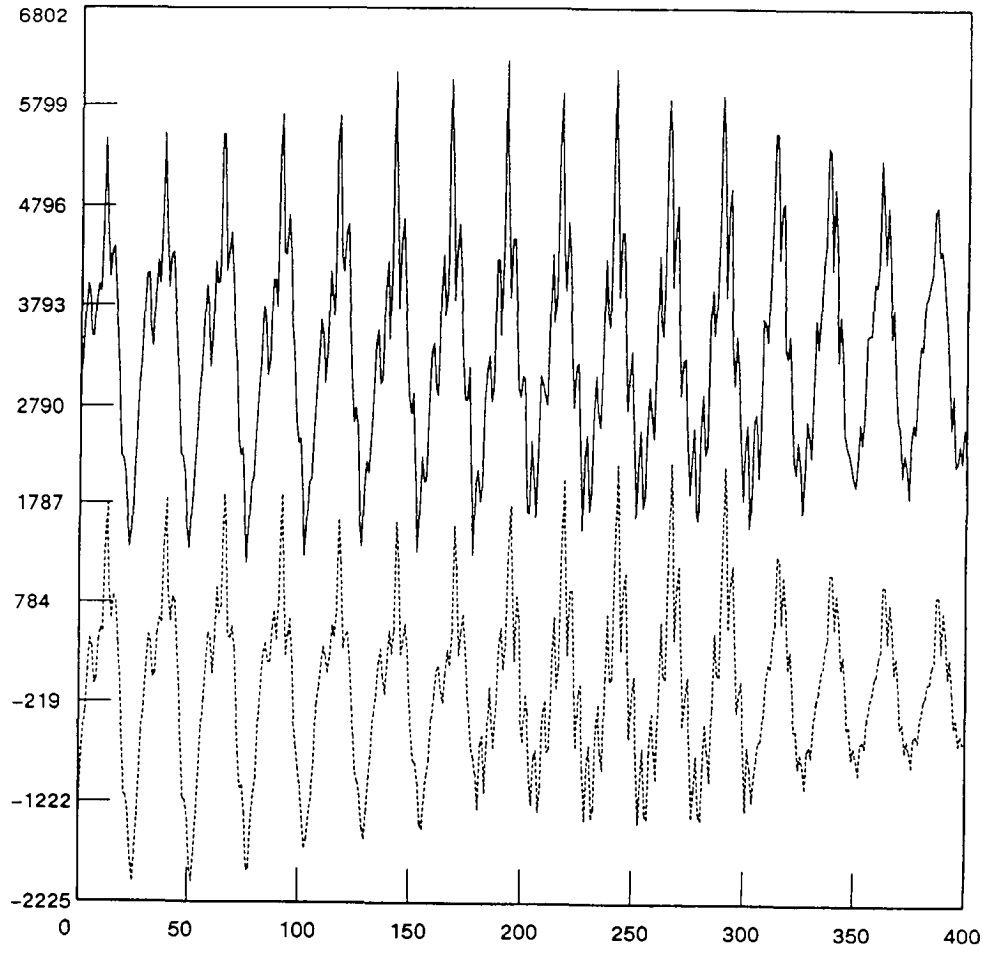

Figure 4.1 Typical waveforms of a) Original speech b) Recovered speech.