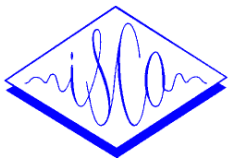

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

TIME ENVELOPE LP VOCODER: A NEW CODING TECHNIQUE AT VERY LOW BIT RATES.

I. A. Atkinson, A. M. Kondoz, B. G. Evans
Centre For Satellite Engineering Research
University of Surrey, Guildford, Surrey GU2 5XH
Tel. (+44) 1483 259803, Fax. (+44) 1483 259504

ABSTRACT

This paper presents a linear prediction (LP) based vocoder in which speech waveforms are considered as having a 'time envelope', the shape of which contains important perceptual information. By ensuring that the time envelope of the synthetic speech closely matches that of the original, natural sounding synthetic speech can be produced. The advantage over more traditional linear prediction vocoders is that the amplitude time envelope is preserved in addition to the spectral envelope, allowing the rapid amplitude transitions associated with onsets to be retained in the synthetic speech, resulting in a more intelligible output. This paper presents a complete vocoder scheme including details of techniques such as parameter interpolation, quantisation, spectrum shaping and pitch detection which have proven necessary to produce natural sounding synthetic speech.

1. INTRODUCTION

Speech compression algorithms with operating rates below 16 kbits/sec are now in widespread use with applications including second generation digital cellular radio, aeronautical communications, maritime communications, secure channels for military applications and very small aperture terminals. Code excited linear prediction (CELP), [1], techniques are widely used to code speech signals at 4.8 kbits/sec and above. Below this rate rapid speech quality degradation is observed owing to the reduced number of bits available to code the model parameters. LP vocoders, e.g. LPC10, [2], are naturally suited to very low bit rate applications because of the simplified model of the synthetic excitation. Unfortunately, the synthetic output from many LP vocoders can contain several types of distortion which is perceptually disturbing to an untrained user. Much of this distortion can be attributed to the simplified excitation model. This paper describes an improved excitation source and energy control scheme which when used in conjunction with reliable parameter determination algorithms results in natural sounding speech at very low bit rates.

2. SPEECH ANALYSIS

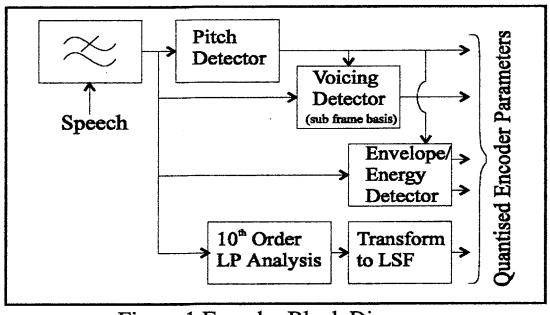

Figure 1 Encoder Block Diagram

Figure 1 shows a block diagram of the encoder. Speech analysis is performed on 20ms or 30ms frames depending on the bit rate, each frame is then further divided into three or four sub-frames. Pitch and LP analysis are performed half a frame ahead so that interpolation techniques may applied at the decoder. Envelope detection is also performed half a frame ahead, but this is to give the decoder knowledge of the early part of the next frame so that the effect of excitation at the end of the current frame can be considered during the energy envelope matching process. The first stage in the encoder is to high pass filter the input speech signal using a cut off frequency of 50Hz, this serves to remove any steady state bias superimposed upon the speech.

The first parameter to be calculated is the pitch since the voicing decision and time envelope detection algorithms require a pitch estimate. The pitch detector is autocorrelation based, however the autocorrelation is segmented into smaller units whose window length and lag are determined from the local maxima and minima of the speech signal. A brief description of the operation of the pitch detector follows. Figure 2 shows the block diagram of the pitch detector. A 480 sample speech segment centred on the end of the current analysis frame forms the input to the pitch detector. The pitch input frame is thus centred half a frame ahead. Peak detection is performed on the pitch window to determine separately both positive and negative peaks which are stored in separate lists. The peaks are determined using the following criteria:

1) Adjacent peaks must not be closer than 5 samples.
2) Each peak must be the largest value within ±5 samples.
3) Peaks falling below one twentieth of the largest peak amplitude are ignored..

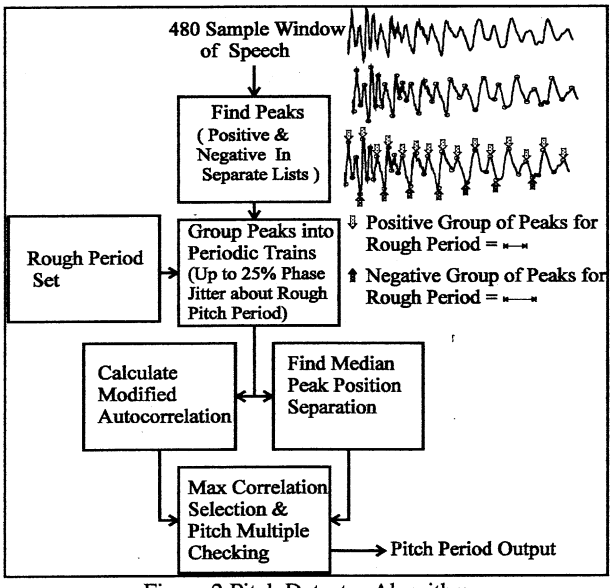

Figure 2 Pitch Detector Algorithm

For each rough pitch shown in Figure 2, a subset of the peaks is selected for both the positive and negative case. (The rough pitch values form an exponentially increasing set of numbers

rounded to the nearest integer.) Subset selection is commenced by assuming that the largest peak is included in the subset. The adjacent peaks are taken to be the largest peaks (same sign) in the region 75% to 125% of the rough pitch away from the last peak. A new search is performed using the previous peak as a starting point, this search procedure is repeated until both the start and end of the pitch window are reached, Figure 2 depicts the results of two subset selections.

For each subset (44 in all, positive and negative for each of 22 rough pitches), the segmented autocorrelation is calculated using Equation 1, where P_i is the i^{th} of N peaks in the relevant subset and d_{back} is the number of samples before the each peak at which point the autocorrelation begins. Figure 3 shows how the segmented autocorrelation is a summation of smaller autocorrelations whose position and lag are determined by the subset of peaks

$$R_{pk}=\frac{1}{P_N-P_0+N.d_{back}}\sum_{i=0}^{N-1}\sum_{-d_{back}}^{P_{i+1}-P_i}s(P_i+j).s(P_{i+1}+j) \quad (1)$$

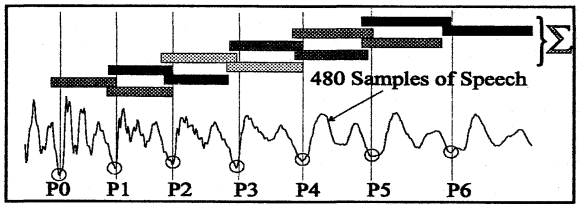

Figure 3 Segmented Autocorrelation

The shaded rectangles represent the speech segments over which each autocorrelation is performed. For each peak subset, the median peak separation is found and this is stored in a list together with the segmented autocorrelation value for that subset. The pitch period returned is the median pitch separation associated with the subset whose segmented autocorrelation is greatest, subject to pitch period multiple checking. The algorithm described operates effectively over the pitch period range of 16 to 180 samples. The advantages of this algorithm over a single autocorrelation are as follows:

1) During sections of speech whose pitch is changing the autocorrelation lag changes across the frame reducing the blurring effect which is observed using a single autocorrelation. This results in more reliable pitch detection.
2) The algorithm returns the median pitch period of the frame, thus rejecting outlier pitch values and when used in conjunction with pitch interpolation results in good pitch contour matching.
3) Only 44 correlations are performed resulting in lower computational load.
4) Formant interference is reduced, see Figure 4.

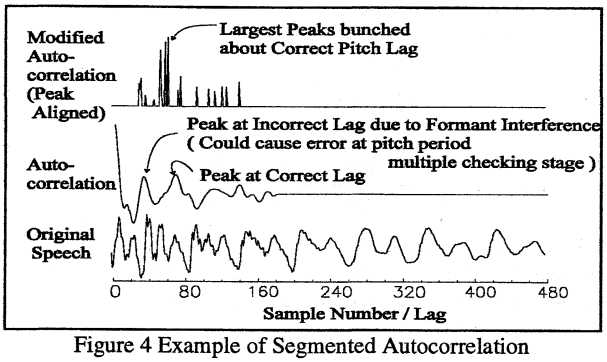

Figure 4 Example of Segmented Autocorrelation

A binary voicing decision is performed for each sub-frame based upon the values of two metrics, a mean square error (MSE) matching measure form one pitch cycle to the next and a high to low band energy measure. In both cases the window length (M) over which the metric is calculated is dynamic, set to be the smallest multiple of the pitch period greater than or equal to the sub-frame length. The centre of the M sample analysis window is set to be the centre of the current sub-frame. The first metric is given by equations 2 and 3 in which and taken to be the minimum value of $Met1(\tau)$ over the range $\tau=\pm$(0.75 to 1.25)Pitch Period. The first metric is the ratio of the matching error energy to the energy of the MSE matched speech

$$Met1(\tau)=\frac{\sum_{j=0}^{M-1}\left(s(j)-\alpha(\tau)s(j-\tau)\right)^2}{\sum_{j=0}^{M-1}\left(\alpha(\tau)s(j-\tau)\right)^2} \quad (2)$$

$$\alpha(\tau)=\sum_{j=0}^{M-1}s(j).s(j-\tau)\Big/\sum_{j=0}^{M-1}s(j-\tau).s(j-\tau) \quad (3)$$

Metric two is given by Equation 4 and is the ratio of the low band energy (<1800Hz) to the energy of the full band speech over the M length window. Note that a constant is added to the full band energy to bias the voicing decision of very low energy signals to that of unvoiced.

$$Met2=\frac{\sum_{j=0}^{M-1}s(j)^2+L_{noise\_floor}}{\sum_{j=0}^{M-1}s_{low\_pass}(j)^2} \quad (4)$$

A final voicing check is performed on sub-frames which have been declared unvoiced. The M length window of speech is filtered using the 2 point moving averager shown in Figure 5 and the maximum absolute output value, Y is determined. A third metric, $Met3$, is given by the ratio of Y to the RMS value of the speech over the window, shown in Equation 5. If the value of $Met3$ exceed a fixed threshold, then the sub-frame is redeclared voiced. This check was found necessary to ensure that certain mixed voiced sounds were declared voiced, the perceptually preferred choice.

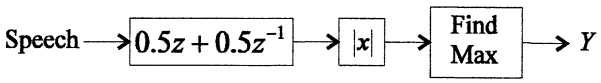

Figure 5 Averaging Filter

$$Met3=Y\Big/\sqrt{\frac{1}{M}\sum_{j=0}^{M-1}s(j)^2} \quad (5)$$

The next parameter to be ascertained is the time envelope, Equation 6, which is the RMS energy of the speech signal calculated over a window whose length is determined from the pitch. $P(n)$ is an interpolated pitch function over the frame, see decoder details.

$$Env(n)=\sqrt{\frac{1}{P(n)}.\sum_{i=0}^{P(n)-1}s^2\left(n-\frac{P(n)}{2}+i\right)} \quad (6)$$

The final parameters to be calculated are the linear prediction coefficients, using a tenth order Levinson-Durbin algorithm. 15Hz bandwidth expansion is applied prior to transformation into the line spectral frequency (LSF) domain, [3]. Table 1 shows the bit allocation schemes for rates of 1.7, 2.4 and 2.8 kbits/sec. A multi-stage split vector quantisation scheme [4] has been adopted to quantise the LSF parameters and a gain/shape approach has been used to quantise the time envelope. Logarithmic scalar quantisation is used for the pitch and time envelope energy.

		Bit Allocation					
Frame Length (ms)	**Sub-Frame Length (ms)**	**LSF**	**Pitch**	**TE Gain**	**TE Shape**	**Voicing**	**Bit Rate kbits/sec**
30	7.5	28	6	7	6	4	1.7
20	5	26	7	7	5	3	2.4
20	5	32	7	7	6	4	2.8

Table 1 Bit Allocation Scheme

	Total Bits	**1st Stage**	**2nd Stage Bit Allocation**				
LSF Order			**LSF1-3**	**LSF 4-7**	**LSF 8-10**	**LSF 8-9**	**LSF 10-12**
10	28	8	7	6	7	-	-
10	26	8	6	6	6	-	-
12	32	9	7	6	-	5	5

Table 2 LSF Quantiser Bit Allocation

Envelope coding is performed using a mean square error gain/shape codebook approach. The first step is to decimate the envelope to N samples per frame using an averager as the decimation filter, N is equal to eight in this case. The next step is to determine which shape codebook entry y_i yields the least value of the error criterion given by equations 7 and 8. In equations 7 and 8, x_k represents the decimated envelope which is to be coded, $y_{i,k}$ represents the i^{th} shape vector and g_i is the optimum gain matching the i^{th} vector to the input envelope. The envelope shape codebook contains 64 (6 bit) entries, the majority of which model rapid onsets.

$$Error_{env}(i) = \sum_{k=0}^{N-1} \left(x_k - g_i y_{i,k}\right)^2 \tag{7}$$

$$g_i = \sum_{k=0}^{N-1} x_k y_{i,k} \Big/ \sum_{k=0}^{N-1} y_{i,k} y_{i,k} \tag{8}$$

Once the optimum shape index has been determined, the associated gain is quantised to 7 bits using a logarithmic quantiser.

3. SPEECH SYNTHESIS

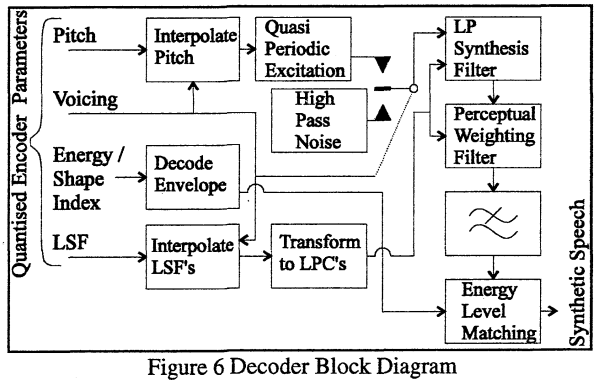

Figure 6 Decoder Block Diagram

Figure 6 shows a block diagram of the decoder. A pitch contour (P_n) for the current frame is calculated using the current and previously received pitch values together with voicing decision and energy information. Linear interpolation is only used when the frame is fully voiced and the energy and pitch parameters are not rapidly changing.

LSF parameters are interpolated every twenty samples using one of two methods depending on the sub-frame voicing:
During unvoiced sub-frames linear LSF interpolation is applied, however during voiced sub-frames the interpolation takes into account the energy of the original speech over which the LSF parameters were determined at the encoder. This information, i.e. E_{last} and E_{next} is obtainable from the decoded envelope. Equation 9 represents the function used to calculate the interpolation factor ($0 \le I(i) \le 1$), where x is the linear interpolation factor ($0 \le x \le 1$).

$$I(i) = \frac{E_{next} \cdot x}{E_{next} \cdot x + E_{last} \cdot (1 - x)} \tag{9}$$

Equation 9 biases the LSF interpolation towards the frame with the largest energy, this improves the performance of the coder at rapid unvoiced onsets which would otherwise introduce too much high frequency energy into the voiced speech immediately following the unvoiced portion.

During unvoiced speech frames the LP excitation takes the form of high pass filtered unit variance gaussian noise. Voiced excitation is the sum of all of the harmonics pitch fundamental within the 4kHz band, each harmonic having equal energy. Phase jitter is introduced to each harmonic to reduce the level of metallic sounds in the synthetic speech, resulting in more natural sounding speech. The expression for the voiced excitation is given by equation 10, where φ_n is the phase of the fundamental given by equation 11, and $\lambda_{n,i}$ is the phase jitter added to the i^{th} harmonic. The phase jitter added is obtained from equation 12 which is a discrete integration of unit gaussian random noise (R_{GEN}) weighted by a function which is dependant on the harmonic frequency, shown in Figure 7. This function has been obtained experimentally from repeated listening tests, the optimum value of β_{max} was found to be 0.75. Too high a value results in noisy speech, whilst too low a value produces metallic sounds in the synthetic speech. Equation 13, shows the relationship between $f_{n,i}$ and P_n, where F_s is the sampling frequency in Hertz.

$$ex_n = \frac{1}{N} \sum_{i=1}^{N} \cos\left(i.\varphi_n + \lambda_{n,i}\right) \tag{10}$$

$$\varphi_n = \varphi_{n-1} + 2\pi / P_n \tag{11}$$

$$\lambda_{n,i} = \lambda_{n-1,i} + R_{GEN} \cdot \beta\left(f_{n,i}\right) 2.\pi / P_n \tag{12}$$

$$f_{n,i} = i.F_s / P_n \quad (13)$$

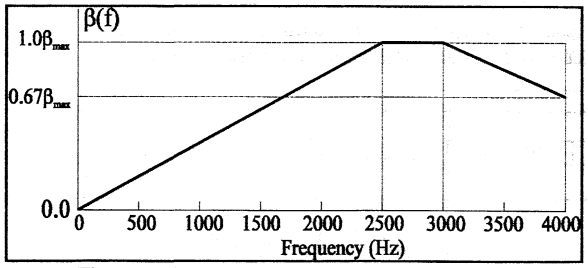

Figure 7 Phase Jitter, Variation with Frequency

The unit RMS energy excitation is passed through the filter whose structure shown in Figure 8. It consists of an LP synthesis filter followed by an adaptive post-filter whose coefficients ($\gamma,\lambda,\alpha,\beta$) are dependant upon the voicing decision. During unvoiced sub-frames high frequency emphasis is applied by the spectral tilt filter, but only moderate spectral emphasis is applied by the perceptual weighting filter, resulting in sharp unvoiced sounds. During voiced sounds, no spectral tilt is applied, but greater perceptual shaping is applied, producing more natural sounding voiced speech.

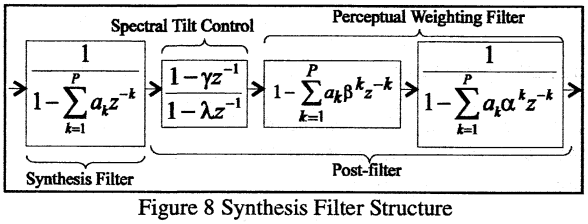

Figure 8 Synthesis Filter Structure

The output of the LP synthesis filter is high pass filtered (50Hz) to remove any steady state bias (c.f. encoder) and then scaled such that the energy envelope of the synthetic speech matches the decoded envelope. The scaling factor is the ratio of the decoded envelope to the envelope of the output of the high pass filter on a sample by sample basis

4. SIMULATION RESULTS

Figure 9 shows a 125ms time segment showing the original speech waveform together with the synthetic speech. Note that the coder does not attempt to match waveform shape nor maintain phase synchronicity. Figure 9 demonstrates the ability of the pitch tracking algorithm to rapidly acquire the correct pitch at onsets and also shows the operation of the voicing decision.

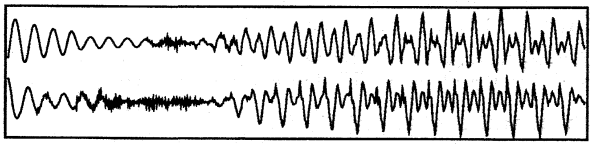

Figure 9 Synthetic Speech (Upper) and Original Speech (Lower)

Informal listening tests, Table 3, indicate that the speech quality of the TE vocoder is comparable to that produced by IMBE STD M [5] at 4.15kbits/sec and its improved version MB-LPC [6] at 2.4kbits/sec.

The listening tests were performed by both experienced and naive listeners using clean speech as source material. Table 3 also demonstrates the inherent robustness of this coder to bit errors, at 10^{-3} bit error rate (BER) little quality degradation was perceived. Additional listening tests were performed on source material containing multiple talkers, using worst case conditions, i.e. two talkers at equal signal level, the synthetic speech did not contain disturbing artefacts such as loud pops and clicks and pitch errors. A comparison with IMBE STD-M [5] showed that the TE vocoder requires half the computational complexity, making it very attractive for low cost and hand held (low power) applications.

	TE Vocoder 1.7kb/s	MBE-LPC 2.4kb/s	IMBE STD-M 4.15kb/s
Male - Error Free	2.8	2.8	2.9
Female - Error Free	2.7	2.5	2.7
Male - 10^{-3} BER	2.8	-	-
Female - 10^{-3} BER	2.7	-	-
Male - 10^{-2} BER	1.2	-	-
Female - 10^{-2} BER	2.0	-	-

Table 3 Results of Informal Listening Test for 1.7kbits/s Time Envelope Vocoder

5. CONCLUSION

In this paper a new form of LP vocoder has been presented. The principle innovation is the modelling of the time envelope which aids the reproduction of speech onsets which are perceptually important when attempting to discriminate between different consonants, especially plosives. This, together with reliable pitch detection and parameter interpolation techniques allow the frame to be extended to 30ms without introducing noticeable distortion, resulting a very low bit rate system. Such a low bit rate makes the Time Envelope Vocoder suitable for use in digital mobile radio systems, voice mail applications, multimedia etc..

7. REFERENCES

1 Schroeder M. R., Atal B. S., "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates.", IEEE Int. Conf. Acoust. Speech Sig. Proc. pp 937-940, 1985

2 Tremain T., "The Government Standard Linear Predictive Coding Algorithm (LPC10)", Speech Technology, 1(2), 1982

3 Soon F., B. Juang H.. "Line Spectral Pair and Speech Data Compression.", Proc IEEE Int. Conf. Acoust. Speech Sig. Proc. pp 1.10.1-1.10.4, 1984

4 Sirven F., "LSF Quantisation Applied to Low Bit Rate Speech Coders", Internal Document CSER University of Surrey, October 1994.

5 DVSI (Digital Voice Systems Inc.) "Inmarsat-M Voice Coding System Description: Draft version 1.3", February 1991

6 Yeldener S., Kondoz A. M., Evans B. G. "High Quality Multi-Band LPC Coding of Speech at 2.4 Kb/s", IEE Electronics Letters, Vol. 27 No. 14 pp1287-1289, July 1991