HIGH QUALITY SPLIT BAND LPC VOCODER OPERATING AT LOW BIT RATES

Ian Atkinson, Suat Yeldener and Ahmet Kondoz

Centre for Communication Systems Research
University of Surrey, Guildford, Surrey, UK.
i.atkinson@ee.surrey.ac.uk

ABSTRACT

LPC based speech coders operating at bit rates below 3.0 kbits/sec are usually associated with buzzy or metallic artefacts in the synthetic speech. These are mainly attributable to the simplifying assumptions made about the excitation source, which are usually required to maintain such low bit rates. In this paper a new LPC vocoder is presented which splits the LPC excitation into two frequency bands using a variable cut-off frequency. The lower band is responsible for representing the voiced parts of speech, whilst the upper band represents unvoiced speech. In doing so the coder's performance during both mixed voicing speech and speech containing acoustic noise is greatly improved, producing soft natural sounding speech. The paper also describes new parameter determination and quantisation techniques vital to the operation of this coder at such low bit rates.

1. INTRODUCTION

Many LPC based vocoders operating at bit rates of 3.0 kbits/sec and below employ simple excitation sources consisting of quasi-periodic impulse trains during voiced speech and gaussian noise during unvoiced speech. Whilst this may be sufficient for maintaining speech intelligibility, the synthetic speech often sounds robotic and the speaker identity is lost. If the speech quality could be improved without significantly increasing the bit rate then many new applications would open up, spawning new products such as voice pagers and internet telephony. Very low bit rate speech coders are also in demand as companions to video compression algorithms for use in future mobile videophone systems.

This Split-Band LPC Vocoder uses vector quantisation techniques to efficiently encode the LPC parameters in the LSF domain, freeing bits which may then be used to encode additional information about the excitation in order to improve the speech quality. The excitation information is extracted by applying an IMBE [1] type analysis to the LPC residual, which is then quantised using one of two schemes, resulting in an overall coded bit rate of 2.5 or 2.7 kbits/sec.

2. ENCODER

Figure 3 presents the schematic of the encoder. DC rejected, high frequency pre-emphasised speech is processed in 20ms frames. LPC parameters are determined using a 10th order Durbin's algorithm which are then quantised in the LSF domain. The quantised LPC parameters are then used to find the LPC residual required for determination of the excitation harmonic amplitudes. Both the speech signal and the LPC excitation are transformed into the frequency domain using a 512 point FFT, (note that these two real FFT's may be calculated using one complex FFT in order to reduce complexity).

Pitch analysis is performed in the spectral domain using a modified version of the algorithm described by McAulay [2] which determines the pitch period to half sample accuracy. The pitch frequency is the value of ω_0 which maximises $\rho(\omega_0)$ in Equation 1, where $E(\omega)$ is the exponentially decaying envelope of the speech spectrum shown in Figure 1, A_l and ω_l are the magnitudes and frequencies of the local peaks in the speech spectrum. $D(\omega\text{-}k\omega_0)$ is given by Equation 2 which is non-zero only for the main lobe of the 'sinc' function.

$$\rho(\omega_0) = \sum_{k=1}^{L(w_0)} E(k\omega_0)\left\{\max\left[A_l D(\omega_l - k\omega_0)\right] - \tfrac{1}{2}E(k\omega_0)\right\} \quad (1)$$

$$D(\omega - k\omega_0) = \mathrm{sinc}\left[2\pi\left(\tfrac{\omega - k\omega_0}{\omega_0}\right)\right] \quad (2)$$

This algorithm was further improved with regard to robustness to noise by incorporating a further energy based metric. All candidate pitch periods returned by Equation 1 were re-examined using Equation 3 and those whose value of $\varphi(l)$ exceeded treble the minimum value of $\varphi(l)$ for all candidate values of l (the candidate pitch period), were rejected. $\varphi(l)$ is a function which measures how much the RMS energy of the speech fluctuates as a function of the window length used in the RMS

calculation. If the window length is equal to the pitch period, then the variation is small and φ(*l*) is also small. An example of φ(*l*) is plotted in Figure 2 for four consecutive speech frames. This combined scheme was found to be highly reliable on all of the speech material tested, which is fundamental to the overall coder performance.

$$\varphi(l) = \left[\sum_{j=0}^{N} |d_l(j)|\right] / \left[\sum_{j=0}^{N} |e_l(j)|\right] \qquad (3)$$

$$e_l(j) = \sum_{k=0}^{l} \left[s\left(j - \tfrac{l}{2} + k\right)\right]^2 \qquad (4)$$

(where *s*(*n*) is the original speech)

$$d_l(j) = d_l(j-1) + 0.95\left[e_l(j) - e_l(j-1)\right] \qquad (5)$$

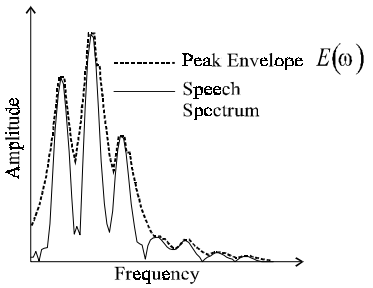

Figure 1: Example of Spectral Peak Envelope

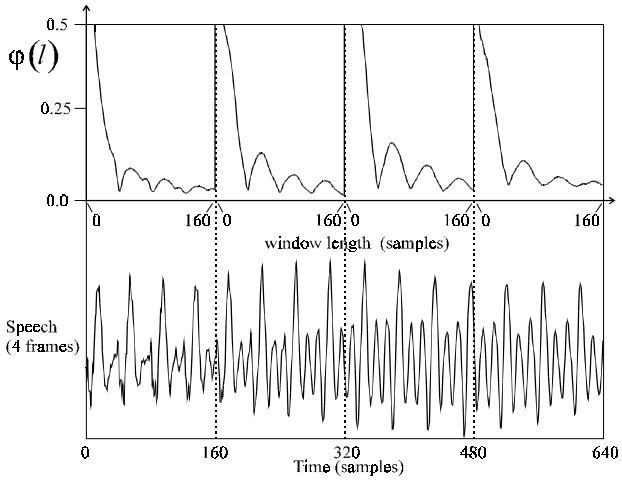

Figure 2: Energy Variance Vs Window Length

Pitch refinement is performed on the speech spectrum using the method similar to that described by Griffin [1] and a binary voicing decision is performed for each pair of harmonics using a technique similar to the APCO scheme [3]. Finally the harmonic amplitudes are determined from the excitation spectrum, using weighted spectral matching and the LSF, pitch, voicing and excitation parameters are finally quantised and transmitted to the decoder.

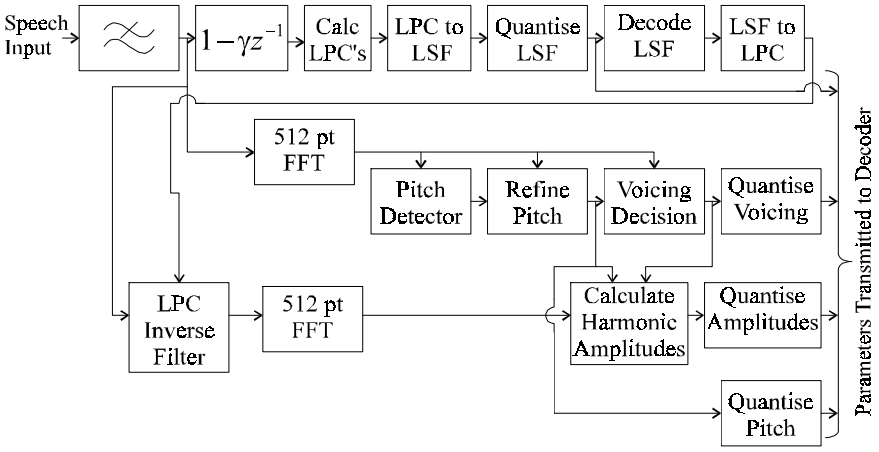

Figure 3: Encoder Schematic

3. DECODER

Figure 5 shows the decoder schematic. Once the parameters have been decoded, the harmonic excitation amplitudes are modified to reduce the noise in the LPC valleys, thereby perceptually improving the coder performance. This is an alternative to using a more traditional post-filtering technique as proposed by Chen [4], however this new method maintains a flat excitation spectrum. The excitation amplitudes are modified according to Equation 6, where $H(i\omega_0)$ is the LPC spectrum sampled at the harmonic frequencies, and $P(i\omega_0)$ is the peak interpolated LPC spectrum sampled at the harmonic frequencies, as shown in Figure 4. This also shows the valley suppressing effect on the LPC spectrum.

$$a'(i) = a(i)\left[\frac{H(i\omega_0)}{P(i\omega_0)}\right]^{0.3} \qquad (6)$$

The synthetic excitation is equal to the sum of the unvoiced and voiced generator outputs. The unvoiced generator is performed using FFT filtering, i.e. spectrally shaping a random noise source in the frequency domain according to the voicing and harmonic amplitude information, then transforming back to the time domain.

The voiced excitation is generated by summing up each sinusoidal harmonic, scaled by the decoded harmonic amplitude as per Equation 7. L is the number of harmonics in the 4kHz band, which is dependent upon the pitch. $a'(i)$ is the i^{th} perceptually modified amplitude. Note that Equation 7 does not indicate the use of amplitude interpolation, however this is performed but has been excluded from this discussion for simplicity. $\phi(i)$ is the phase of the fundamental given by Equation 8 which is an integration of the fundamental frequency $\omega(i)$ (in radians) which is linearly interpolated between frames if the deviation is less than 20%.

$$e_v(i) = \sum_{j=1}^{j \le L} a'(i)\cos\left[j\phi(i)\right] \quad (7)$$

$$\phi(i) = \phi(i-1) + \omega(i) \quad (8)$$

Finally the overall spectral shaping is added to the excitation using an LPC synthesis filter whose coefficients are linearly interpolated every 5ms.

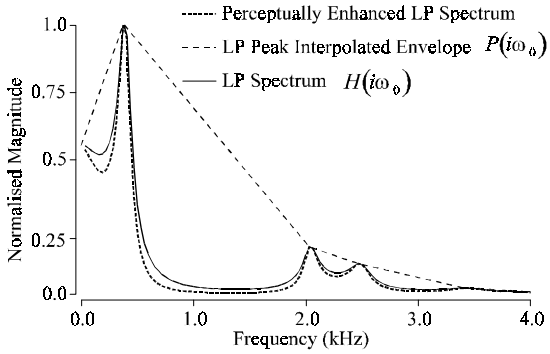

Figure 4: Effect of Perceptual Enhancing Process on LPC Spectrum

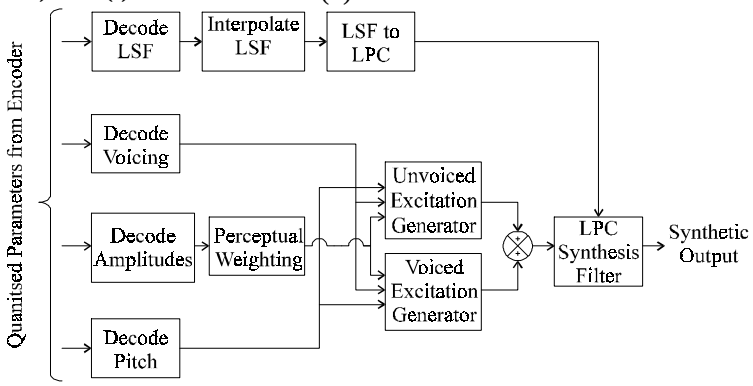

Figure 5: Decoder Schematic

5. QUANTISATION

Although a voicing decision is performed for each pair of harmonics the limited number of bits available for encoding forces us to restrict the voicing to adhere to the following rule: all harmonics up to a certain frequency are declared voiced, and those above are declared unvoiced. This means that the voicing information is now represented by a single frequency value which can easily be quantised using 3 bits. Previously, Yeldener [5] suggested the use of a voicing probability to determine the voicing frequency, given by the ratio of the number of voiced harmonics to the total number of harmonics. This was found to give good performance in general, however in certain circumstances strongly voiced frequencies in the mid-range of the spectrum were incorrectly declared unvoiced giving the synthetic speech a hoarse quality. A new single frequency voicing quantiser is defined by Equation 9 which takes into account the original speech harmonic amplitudes $\alpha(j)$ and determines the voicing frequency using a soft-decision process. $v(j)$ represents the unquantised harmonics voicing decisions, where values 1=voiced and -1=unvoiced. Additionally, another term was incorporated to give more weight to voiced harmonics, given by Equation 11. This scheme greatly reduced the "hoarseness" in the synthetic speech.

$$q(i) = \sum_{j=1}^{n \le L} \alpha(j)v(j)u_i(j)b\left(v(j)\right) \quad (9)$$

$$u_i(j) = \begin{cases} 1 & j \le \frac{iL}{7} \\ -1 & j > \frac{iL}{7} \end{cases} \quad (10)$$

$$b(x) = 1.5 \quad x \ge 0$$

$$b(x) = 1.0 \quad x < 0 \quad (11)$$

LSF quantisation was performed using the Linked Split-Vector approach described by Kim [6] which switches between quantiser tables depending upon already quantised LSF values. This gives good performance using 28 bits and requires only 448 words of storage. The LSF quantisation scheme is shown in Table 1.

LSF Group	Comment	Bits	Storage/words
1&2	Switched 2x32	5	64
3&4	Single 128	7	128
5&6	Switched 2x64	6	128
7&8	Switched 2x32	5	64
9&10	Switched 2x32	5	64
	Total:	**28**	**448**

Table 1: LSF Quantisation Scheme

The pitch and the RMS value of the excitation harmonic amplitudes is quantised by a logarithmic scalar quantiser, using 7 and 6 bits respectively. Finally, the harmonic amplitudes normalised to their RMS value are quantised. The first 8 values are vector quantised using 6 bits. The remainder are assumed to be equal to unity in the 2.5kbits/sec version. Alternatively, they are grouped into 8 bands and then vector quantised using 4 bits for the 2.7kbits/sec version. The overall bit allocation scheme is presented in Table 2.

Parameter	Version	
	2.5kbits/sec	2.7kbits/sec
LSF	28	28
Voicing Freq.	3	3
Pitch	7	7
Energy	6	6
1st 8 Harmonics	6	6
Excitation Shaping	-	4
Total:	**50**	**54**

Table 2: Bit Allocation Schemes

6. SIMULATION RESULTS

Listening tests were performed on the high-level simulation, comparing the Split-Band LPC Vocoder against DoD LPC10e [7], IMBE and DoD 1016 CELP [8]. Individual tests were performed on male and female speech using twenty subjects. The results are presented in Table 3. Although the scores are generally lower than expected due to the unfamiliarity of many of the subjects with low bit rate coders, the results clearly indicate the preference of the Split-Band vocoder over LPC10e, IMBE and 1016 CELP at both 2.5 and 2.7kbits/sec.

	Mean Opinion Score		
	Male	Female	Overall
LPC10 2.4kbs	1.3	1.3	1.3
IMBE 4.15kbs	3.4	3.0	3.2
1016 CELP 4.8kbs	3.2	3.0	3.1
Split Band Vocoder 2.5kbs	3.4	3.4	3.4
Split Band Vocoder 2.7kbs	3.5	3.7	3.6

Table 3: Split Band Vocoder Listening Test Results

The complexity of the coder was also investigated by considering every ADD, MULT, MOVE, COMPARE, etc. required by the program. These findings are presented in Table 4, expressed in terms of Millions of Operations per second (MOPS).

Function	Average MOPS	Maximum MOPS
Encoder and Decoder	14.0	15.3
Encoder Only	9.9	11.2
Decoder Only	4.1	4.4

Table 4: Split Band Vocoder Processor Cycle Estimation

Finally the coder was tested on speech containing acoustic noise (vehicle, gaussian, multiple talker). Although the background noise was modified by the coder, the speech retained its intelligibility and talker identity. Very little degradation was observed in the case of multiple talkers.

7. CONCLUSION

A new LPC based vocoder was presented which has been shown to produce higher quality speech at a bit rate of 2.5kbits/sec than both the IMBE coder operating at 4.15kbits/sec and the DoD 1016 CELP operating at 4.8kbits/sec. By splitting the speech into voiced and unvoiced bands, both mixed voicing speech and speech containing acoustic noise could be reliably coded without introducing excessive "buzziness" into the synthetic speech. The coder's high subjective performance may be directly attributed to improved parameter extraction techniques which have been described in this paper. The coder's quantisation schemes were designed with low storage in mind, this combined with a full duplex computational requirement of 14.0 MOPS should enable implementation of both encoder and decoder on a single mainstream fixed point DSP device.

REFERENCES

1. D. W Griffin., J. S. Lim, "Multi-Band Excitation Vocoder", IEEE Trans. ASSP, Vol. 36, No.8, pp 1223-1235, Aug. 1988.
2. R. J. McAulay, T. F. Quateri, "Pitch Estimation and Voicing Decision Based upon a Sinusoidal Speech Model", Proc. ICASSP, Vol. 1, pp 249-252, 1990.
3. APCO, "Vocoder Description", Document No. IS102BABA, Association of Public-Safety Communication Officials, 1993.
4. J. H. Chen, A. Gersho, "Real Time Vector APC Speech Coding at 4.8 kb/s with Adaptive Post-filtering", Proc. ICASSP, Vol. 3, pp 2185-2188, 1987.
5. S. Yeldener, "Sinusoidal Model Based Low Bit Rate Speech Coding for Communication Systems", PhD Thesis, University of Surrey, April 1993.
6. M. Y. Kim, N. K. Ha, S. R. Kim, "Linked Split-Vector Quantizer of LPC Parameters", Proc. IEEE ICASSP, Vol. 2, pp 741-744, 1996.
7. T. Tremain, "The Government Standard Linear Predictive Coding Algorithm (LPC-10)", Speech Technology, Vol. 1(2), pp 40-49, April 1982.
8. J. P. Campbell, T. E. Tremain, V. C. Welch, "The DOD 4.8KBPS Standard (Proposed Federal Standard 1016)", Speech Technology, Vol. 1(2), pp 58-60, April 1990.