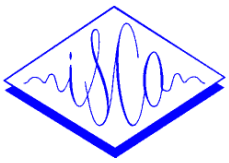

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A Robust 2.4kb/s LP-MBE With Iterative LP Modelling

Clifford I. Parris[1], Danny Wong[1] and François Chambon[2]

[1]Ensigma Ltd., Chepstow, UK (cliff@ensigma.com, danny@ensigma.com)
[2]ENST, Paris, France

1. ABSTRACT

In order to reduce the transmission bandwidth requirement for the spectral envelope parameters for Multiband Excitation (MBE) coders [1] Linear Prediction (LP) modelling has been widely investigated [2, 3, 4]. In this paper we compare the direct methods for evaluating the linear prediction filter [2, 3] with a novel iterative LP modelling technique. A novel switched Vector Quantizer (VQ) is employed to transmit 12 binary V/UV harmonic decisions [5] using only 4 bits which achieves only 5% errors. To enhance robustness joint source and channel coding codebook design and search has been incorporated into the coder.

2. INTRODUCTION

We have based our investigation on the IMBE algorithm [5] which operates with a 20ms frame rate. The unquantized spectral amplitudes are modelled by an LP envelope. The LP coefficients are transformed to Line Spectral Frequencies and quantized using a 26-bit 2 split VQ. The gain parameter related to the LP residual energy is quantized using 6-bits. A dimension 12 V/UV decision vector is vector quantized using a 4-bit switched VQ. The pitch parameter is quantized using 9 bits. Joint source and channel designed codebooks are used through out.

3. LP MODELLING

3.1 Introduction

Linear prediction is a well known approach in speech coding. It describes the average power spectrum as $|H(e^{jw})|^2$, with $H(e^{jw}) = \frac{G}{A(e^{jw})}$, where:

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \qquad (1)$$

where p is the order of the LP model. In the power spectrum domain, the LP coefficients a_k are calculated by minimising the integrated ratio of the signal and its LP approximation [6].

The gain factor G for minimising the error between original spectral amplitudes and estimated ones is given by

$$G = \frac{\sum_{m=0}^{M-1} A_m \hat{A}_m}{\sum_{m=0}^{M-1} \hat{A}_m \hat{A}_m} \qquad (2)$$

where A_m and $\hat{A}_m$ are the real and estimated spectral amplitudes for the M harmonics.

There are a number of ways to evaluate the LP coefficients. A common method is via the autocorrelation coefficients of the input signal. The autocorrelation coefficients can be derived in the time domain or in the frequency domain. In this work we use the frequency domain approach which is related to the time domain method via the Weiner-Khintchine theorem.

$$R_i = \frac{1}{N} \sum_{n=0}^{N-1} |F_n|^2 \cos(i\omega_n) \qquad (3)$$

where F_n is the signal spectrum at frequency ω_n.

Three different methods for calculating the autocorrelation coefficients have been investigated. They differ only in the number of spectral samples used and how the spectral samples are obtained.

3.2 Spectral Amplitudes Method (Method 1)

In this method the spectral envelope parameters A_m as determined by the MBE algorithm are used. The autocorrelation coefficients are computed as,

$$R_i = \frac{1}{M} \sum_{m=0}^{M-1} A_m^2 \cos(im\omega_0) \qquad (4)$$

where ω_0 is the pitch frequency, and M the number of harmonics. The gain energy term G can be conveniently calculated by equation (2). This technique was first reported in [2].

3.3 Full Spectrum Method (Method 2)

This method uses all of the frequency samples of the actual speech spectrum, F. This is the direct implementation of the Weiner-Khintchine theorem. The autocorrelation coefficients are computed as,

$$R_i = \frac{1}{N} \sum_{n=0}^{N-1} |F_n|^2 \cos(i\omega_n) \qquad (5)$$

where $\omega_n = \frac{2n\pi}{N}$. As the Fourier coefficients are readily available, the energy term can be directly computed by Parseval's Theorem. This technique was first reported in [3].

3.4 Spectrum Near Harmonics Method (Method 3)

This method is a mixture of the two preceding ones. It uses the frequency samples of the speech spectrum closest to the pitch harmonics. The autocorrelation coefficients are given by

$$R_i = \frac{1}{M} \sum_{m=0}^{M-1} |F_x|^2 \cos(im\omega_0) \qquad (6)$$

where x is the integer truncation of $\frac{256m\omega_0}{2\pi}$

Since the spectral amplitude at the harmonic is not used and no averaging of envelope in unvoiced regions is performed this technique should be expected to perform worse than the spectral amplitudes method.

3.5 Investigation of Direct LP Modelling Techniques

The number of spectral samples used in methods one and three is dependent on pitch and typically the harmonic sampling of the speech spectrum will not be lossless. In order to track the performance of the LP model, different orders of the LP analysis have been used. The metric we use to compare objectively the three methods for deriving the autocorrelation coefficients is given by

$$Q = -10 \log_{10} \frac{\sum_{m=0}^{M-1} (A_m - \hat{A}_m)^2}{\sum_{m=0}^{M-1} A_m^2} \qquad (7)$$

The overall performance score was averaged over a 30s speech segment of mixed male and female utterances. Note that the metric only considers the modelling of the harmonics not the entire spectrum, this was found to correlate well with informal listening tests and is similar to the performance metric used in [4].

The scores for the three methods are given by the dotted curves given in Figure 1. Method 1 performs better than method 3 as expected. Methods 1 and 3 which sub-sample the speech spectrum tend to saturate in performance at an LP order of 16.

According to [6] the error metric E minimised by the LP analysis is :

$$E = \frac{G^2}{N} \sum_{n=0}^{N-1} \frac{p(n)}{\hat{p}(n)} \qquad (8)$$

where $p(n)$ and $\hat{p}(n)$ are the input signal and modelled power spectra respectively.

But according to [6]

$$\sum_{n=0}^{N-1} \frac{p(n)}{\hat{p}(n)} = 1 \text{ for any } p \qquad (9)$$

The quality of the match is determined by how closely $\hat{p}(n)$ follows $p(n)$. Typically there are frequency regions for which $p(n) > \hat{p}(n)$ and regions for which $\hat{p}(n) > p(n)$. However, due to the nature of the metric, regions for which $p(n) > \hat{p}(n)$ contribute more to the error than regions for which $\hat{p}(n) > p(n)$.

Figure 4 illustrates spectral matches obtained for a typical strongly voiced frame for $p = 10$ and $p = 16$ when using the full spectrum method (method 2).

Notice that the harmonic peaks are rounded and reduced in magnitude and that the error is greatest in the spectral troughs. The error at the harmonics 'cancels' the error in the troughs since at the harmonics $p(n) > \hat{p}(n)$ where as at the troughs $\hat{p}(n) > p(n)$. As p increases the LPC analysis attempts to model the individual harmonics. Since the spectral troughs are more pronounced for the line spectrum assumed by method 1 the error may be expected to be larger. This forces a larger error in modelling at the harmonics to 'cancel' the error.

The synthetic speech quality obtained by these three direct methods was found to be acceptable but non-transparent for 16th order LP modelling and poor for 10th order modelling. These results contradict those given by [2, 3] where 10th order was claimed to give good speech quality. Our results do however agree with [4] who adopts 16th order model after first fitting a cubic-spline envelope to the spectral amplitudes to obtain a smooth target spectrum for LP modelling.

3.6 New Technique (Method 4)

The realisation that the quality of the spectral match at the harmonics is determined by the shape of $p(w)$ between the harmonics suggests that if we choose $p(w)$ so that $\hat{p}(w)$ matches well between the harmonics, i.e. $p(w)/\hat{p}(w) = 1$ then the match at the harmonics will improve. In [4] the cubic-spline envelope performs this function. We however suggest a simple iterative approach where $p(w)$ is modified so that the intra- harmonic spectrum is replaced by the model spectrum $\hat{p}(w)$. The iterative procedure is as follows:

1. Derive $\hat{p}(w)$ as described in method 1.

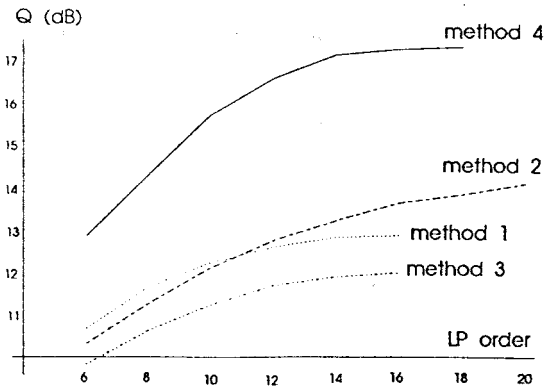

Figure 3: Performance versus LP analysis order

Figure 1. Performance versus LP analysis order

2. Modify $\hat{p}(w)$ to obtain $\tilde{p}(w)$ by replacing the amplitudes at the harmonics by the target values.

3. Derive $\hat{p}(w)$ as described in method 2 using $\tilde{p}(w)$ as the signal spectrum.

4. For another iteration goto step 2.

where $\tilde{p}(w)$ denotes the modified signal spectrum at each iterative stage.

To reduce complexity it is necessary to limit the maximum number of iterations. We observed that the majority of the improvement is obtained during the first three iterations. The performance of the new iterative method is given by the solid curve in Figure 1. The method clearly performs better than any other technique. Informal listening tests indicate that the novel iterative technique achieves transparent coding of the spectral amplitude for an LP order of 16, comparable results are quoted in [4]. The performance for a 10th order model was, however, found to be virtually transparent with only occasional formant shifting and pole merging accounting for non-transparent frames. These tests were carried out by listening to the synthetic speech obtained by the LP modelling of the spectral amplitudes (unquantised LP coefficients) and comparing it to the optimal synthetic speech obtained by using the unquantised spectral amplitudes directly at the synthesiser.

4. V/UV DECISION QUANTISATION

In the IMBE coder K bits are allocated to encode the V/UV decisions, where K is given previous.

To obtain a fixed vector length when $K < 12$ the IMBE V/UV vector is padded with unvoiced decisions until 12 decisions are obtained. This vector is

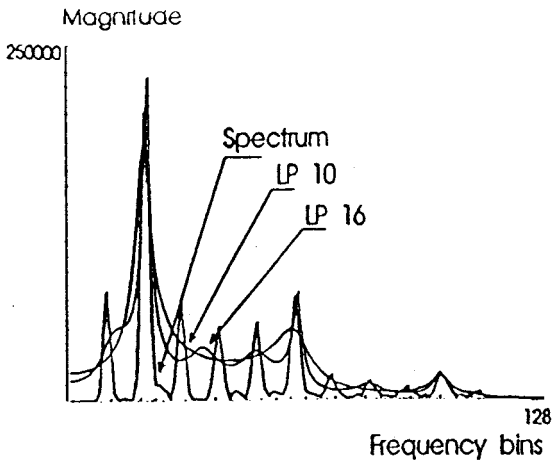

Figure 2. Spectral Modelling with 10th and 16th Order LP Models

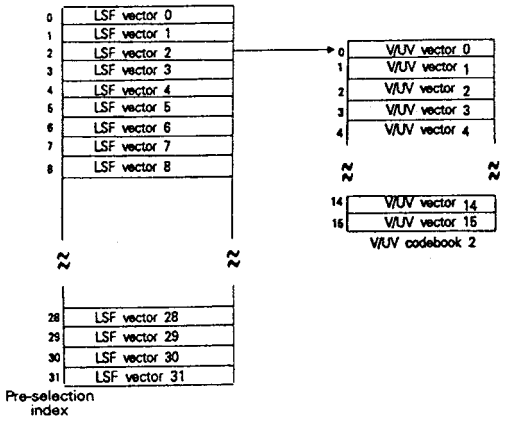

Figure 3. Voiced/Unvoiced Quantisation

then quantised using the switched VQ structure illustrated in Figure 3. A 4-bit codebook (16 entry) is searched to perform the vector quantisation. Thirty two such codebooks exist, the selected codebook is chosen based on a pre-selection search of a 32 entry LSF vector codebook. The switched VQ quantiser exploits the correlation between the spectral envelope (as represented by the LSFs) and the V/UV decisions. Figure 4 illustrates the improvement obtained as the size of the pre-selection codebook is increased. Notice that the proposed 2.4 kb/s coder has an error rate of about 5% on the V/UV decisions compared to the IMBE coding.

5. JOINT SOURCE AND CHANNEL CODING

5.1 Introduction

When bandwidth is limited Forward Error Correction (FEC) is often added at the expense of source coding information. Thus a tradeoff exists between

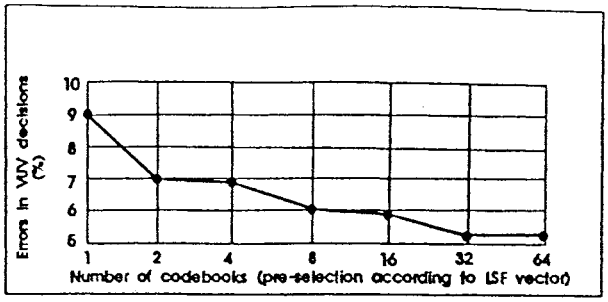

Figure 4. Improvement in V/UV Decisions Depending on Pre-Selection Codebook Size

source code and channel code bit allocation. By incorporating the expected channel transition probability and by reassigning or modifying the codewords, improved immunity to channel errors is obtained without the inclusion of explicit channel coding. Thus there is no increase in bandwidth due to redundant FEC codes and in many cases the overall system complexity is reduced.

Results given in [7] indicate that an LSF trellis quantiser can provide lower distortion than a conventional separate source and channel coder (operating at the same joint bit rate) for all random bit error rates below a design rate. In addition, the distortion degraded more gracefully for higher bit error rates.

5.2 Results

Table 1 illustrates the performance of the joint source and channel (JSC) coding technique. Notice that the JSC method does degrade clean channel perforance.

6. REFERENCES

	Quality Score	% of Errors on V/UV	Gain Distortion
(1) Uncorrupted without JSC quantisation	14.09	3.59	0.0098
(2) Uncorrupted with JSC quantisation	11.96	3.73	0.0124
(3) Corrupted without JSC training	7.33	11.74	2.42
(4) Corrupted without JSC quantisation	10.67	8.04	0.43
(5) Corrupted with JSC quantisation but no index optimisation	10.62	7.57	0.044
(6) Corrupted with JSC quantisation and index optimisation	10.75	7.3	0.0386

Table 1. Evolution of Scores and Robustness with Training and Quantisation Methods. Coruption is 1% Random Bit Errors

[1] D.W.Griffin and J.S.Lim.*Multiband Excitation Vocoder.*IEEE Transactions on ASSP, Vol.36, No.8, August 1988, pp 1223-1235.

[2] D.Rowe, W.Cowley and A.Perkis.*A Multiband Excitation Linear Predictive Speech Coder.*Proceedings of Eurospeech 91, 2nd European Conference on Speech Communication and Technology, Genova, Italy, 24-26 September 1991, pp 239-242.

[3] B. G. Evans, A. M. Kondoz, S. Yelder, M. R. Suddle, W. Ma. *A High Quality 2.4 kb/s Multi-Band LPC Vocoder and its Real Time Implementation.* Proc. ISSSE 1992.

[4] R.J.McAulay, T.Champion and T.F.Quatieri.*Sinewave Amplitude Coding Using Line Spectrum Frequencies.*Proceedings of IEEE Workshop on Speech Coding for Telecommunications, Canada, 13-15 October 1993, pp 53-54.

[5] DVSI.*Inmarsat-M Voice Coding System Description.*Draft Version 1.3, February 1991.

[6] J.Makhoul.*Linear Prediction: A Tutorial Review.*Proceedings of the IEEE, Vol. 63, No. 4, April 1975, pp 561-580.

[7] D. Rowe and P. Secker. *A Robust 2400bps MBE-LPC Speech Coder Incorporating Joint Source and Channel Coding.* Proc ICASSP 1992.