LP MODELLING OF SPECTRAL AMPLITUDES FOR SINE TRANSFORM CODERS

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. INTRODUCTION

Sine Transform Coders (STC) [1] compress speech signals in the frequency domain by parameterising separately the spectral envelope and the harmonic structure. In order to reduce the transmission bandwidth requirement of the spectral envelope parameters, we propose the use of line spectral frequencies (LSF) derived from a linear prediction front-end, as a substitute for the direct quantisation of the spectral amplitudes. We present in this paper three direct methods for evaluating the linear prediction filter from the spectral amplitudes, and compare their relative performance. A novel iterative technique is also described which performs significantly better than the direct methods.

2. MULTI BAND EXCITATION CODER

The Multi-Band Excitation (MBE) [2] coder is a class of the STC where the unvoiced harmonics are synthesised using bandpass filtered noise. The parametric representation consists of the magnitude of the short term spectral envelope at each harmonic, the pitch period and a voiced/unvoiced (V/UV) decision for each harmonic. Some STC/MBE coders also transmit phase information for each harmonic but this significantly increases the required bit rate. Typically contiguous harmonics are considered as bands and a binary V/UV decision is encoded for each band [3]. The MBE method gives good speech quality and intelligibility, even when the speech is in background noise.

The MBE analysis procedure determines the fundamental frequency and harmonic spectral envelope by matching a synthetic spectrum to the input speech spectrum. This synthetic spectrum is a windowed pulse train shaped optimally to fit the original speech. The matching process calculates the optimum spectral amplitude to be used for synthesis at each harmonic frequency, hence it determines the spectral envelope only at the harmonics. Subsequently the V/UV decisions are made by thresholding the error in this match within each frequency band. Voiced regions tend to have sharp harmonic peaks in the spectrum whereas noisy regions are much flatter. Unvoiced spectral amplitudes are then modified to reflect the mean level in the original speech spectrum over the harmonic band. Figure 1 illustrates the analysis process.

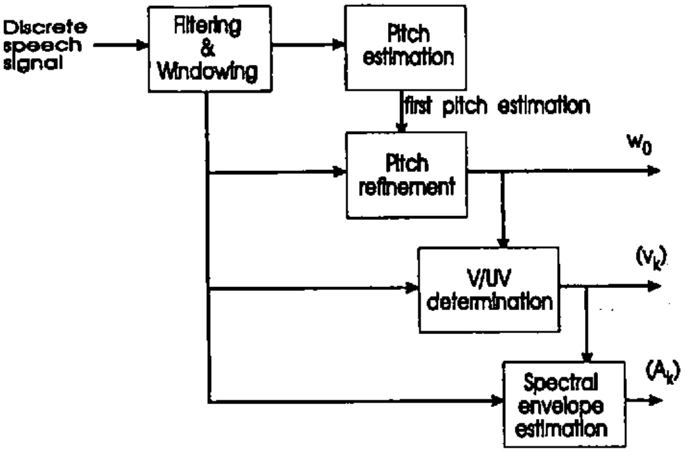

Figure 1: MBE encoder

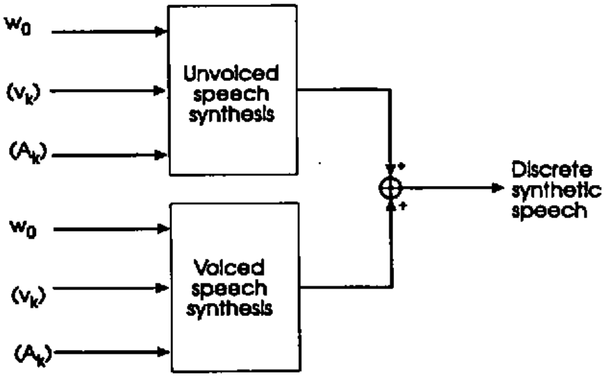

Figure 2: MBE decoder

The MBE synthesis proceeds separately for voiced and unvoiced frequency bands. The voiced parts are made up of the sum of sinusoidal oscillators with frequencies at the harmonics of the pitch frequency, and their amplitudes are given by the spectral envelope parameters. The unvoiced ones are bandpass filtered white noise. The relative phases of each frequency band are smoothed from one frame to another to avoid discontinuities. Figure 2 illustrates the synthesis process.

One of the disadvantages of MBE coding is the considerable number of parameters required. For 8kHz sampled speech, usually 15 to 50 V/UV decisions, as well as the same number of spectral envelope points, are necessary, depending on the pitch frequency. Recently all-pole modelling of spectral envelope parameters has been used to improve bandwidth efficiency [1, 4]. This is typically achieved by vector quantising the Line Spectral Frequencies (LSF) representation of the Linear Prediction (LP) description of the spectral envelope. In this paper different methods to compute the LP model within an MBE framework are presented.

3. LP MODELLING OF THE SPECTRAL AMPLITUDES

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:

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

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 [5].

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

$$G = \frac{\sum_{m=0}^{M-1} S_m \hat{S}_m}{\sum_{m=0}^{M-1} \hat{S}_m \hat{S}_m} \tag{2}$$

where $\hat{S}_m$ are the estimated spectral amplitudes. Equation 2 is based on minimisation of the mean squared error. Hence it requires the computation of the estimated spectral amplitudes. If computation time is an important factor, equation 3 is a good alternative [5].

$$G = R_0 + \sum_{k=1}^{p} a_k R_k \tag{3}$$

where R_k is the kth autocorrelation coefficient of the signal.

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.

$$R_i = \frac{1}{N} \sum_{n=0}^{N-1} |F_n|^2 \cos(i\omega_n) \tag{4}$$

where F_n is the signal spectrum at frequency ω_n.

Three different methods are possible to compute autocorrelation coefficients in our case. They differ only in the number of spectral samples used and how the spectral samples are obtained.

3.1 Spectral Amplitudes Method (Method one)

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

$$R_i = \frac{1}{M} \sum_{m=0}^{M-1} S_m^2 \cos(im\omega_0) \tag{5}$$

where ω_0 is the pitch frequency, and M the number of harmonics.

The energy term G can be conveniently calculated by equation 2. This technique was first reported in [4].

3.2 Full Spectrum Method (Method two)

This method uses all 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,

$$R_i = \frac{1}{N} \sum_{n=0}^{N-1} |F_n|^2 \cos(i\omega_n) \tag{6}$$

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.

3.3 Spectrum Near Harmonics Method (Method three)

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

$$R_i = \frac{1}{M} \sum_{m=0}^{M-1} |F_x|^2 \cos(im\omega_0) \tag{7}$$

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.

4. INVESTIGATION OF DIRECT LP MODELLING TECHNIQUES

We have based our investigation on the IMBE algorithm [3]. The direct evaluation and quantisation of spectral amplitudes have been replaced by a LP analysis. Each 20ms frame of speech is windowed (Kaiser-Bessel window) before a 256 point DFT is computed. Note that 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 covered. The metric we use to compare objectively the three methods for deriving the autocorrelation coefficients is given by

$$Q = -10\log_{10}\frac{\sum_{m=0}^{M-1}(S_m - \hat{S}_m)^2}{\sum_{m=0}^{M-1} S_m^2} \tag{8}$$

The overall performance score is then 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.

The scores for the three methods are given by the dotted curves given in Figure 3. 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. The saturation occurs since the harmonic line spectrum is exactly modelled by an all pole model of order 2M. Since the poles lie close to the unit circle LP analysis becomes unstable as LP order is increased towards 2M and the poles migrate to the unit circle. In our investigations instability occurred when $p > 16$. Methods 1 and 3 are therefore not suitable for higher order modelling.

The full spectrum method outperforms the spectral amplitude method for $p \geq 11$ this may be inferred from [5]. According to [5] the error metric E minimised by the LP analysis is :

$$E = \frac{G^2}{N}\sum_{n=0}^{N-1}\frac{p(n)}{\hat{p}(n)} \tag{9}$$

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

But according to [5]

$$\sum_{n=0}^{N-1}\frac{p(n)}{\hat{p}(n)} = 1 \text{ for any } p \tag{10}$$

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 $p = 16$.

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 the 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.

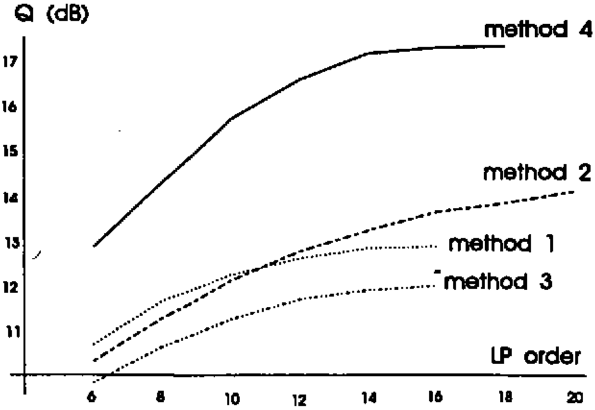

Figure 3: Performance versus LP analysis order

5. NEW ITERATIVE TECHNIQUE

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. This suggests an 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 one.
2. Modify $\hat{p}(w)$ to obtain $\bar{p}(w)$ by replacing the amplitudes at the harmonics by the target values.
3. Derive $\hat{p}(w)$ as described in method two using $\bar{p}(w)$ as the signal spectrum.
4. If improvement goto step 2.

Where $\bar{p}(w)$ denotes the modified signal spectrum at each iterative stage. The performance of the new iterative method is given by the solid curve in Figure 3. The method clearly performs better than any other technique.

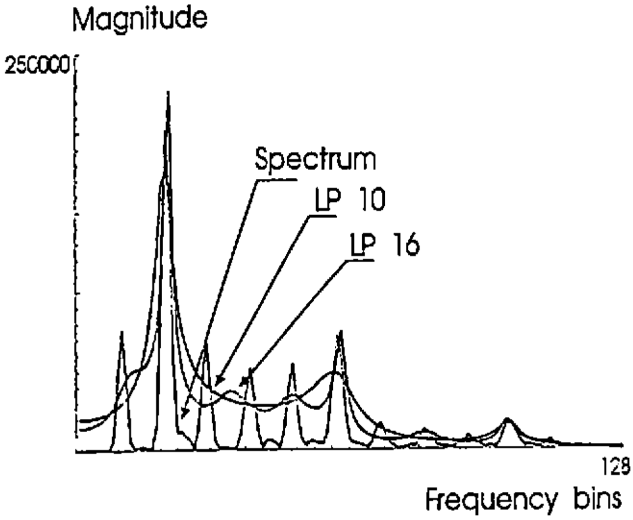

Figure 4: Spectral Modelling with 10th and 16th Order LP Models

6. CONCLUSION

In order to reduce the bit-rate requirement of STC, the direct quantisation of the spectral amplitudes can be substituted by transmitting the LSF representation of an LP model. Based on the structure of the IMBE coder, we have evaluated and compared the performance of four methods for calculating the LP coefficients. Informal listening tests indicate that the novel iterative technique achieves near transparent coding of the spectral amplitude for an LP order of 16. 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.

7. REFERENCES

[1] 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.

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

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

[4] 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.

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