COMPARISON OF SPECTRAL ESTIMATION TECHNIQUES FOR LOW BIT-RATE SPEECH CODING

D.J. Molyneux[*]*, C.I. Parris*[#]*, X.Q. Sun*[•]*, B.M.G. Cheetham*[*]

*School of Engineering, University of Manchester, Oxford Rd, M13 9PL, U.K.
#Ensigma Ltd., Turing House, Station Rd, Chepstow, Gwent, NP6 5PB, U.K.
• Voxware Inc., College Rd, Princeton, NJ08540, USA

ABSTRACT

Many low bit-rate speech coders represent the spectral envelope by an all-pole digital filter whose coefficients are calculated by a form of linear prediction (LP) analysis. The lower the bit-rate, the more critical will be the accuracy of the spectral analysis for achieving good quality speech. This paper compares four known techniques: a technique based on cubic spline interpolation, DAP, MVDR, and iterative all-pole modelling. First, the accuracy obtained for artificial and real speech spectra is assessed for each technique by calculating the degree of spectral distortion with reference to the spectral envelope sampled at the pitch-harmonics. Then, each technique is used to characterise the spectral amplitudes generated by a 2.4 kb/s multi-band excitation (MBE) coder. Results show that significantly better spectral accuracy is obtained using DAP. However listening tests on MBE encoded speech indicate that the advantage of DAP over the other techniques is not strongly perceptible.

1. INTRODUCTION

Many low bit-rate speech coders parametrise the short term spectral envelope of each speech segment as the gain response of an all-pole digital filter whose coefficients are calculated by a form of linear prediction (LP) analysis [1]. The accuracy of the spectral envelope analysis becomes a critical factor for high quality speech at bit-rates below 4kb/s, as the residual can carry less information than, for example, with higher bit-rate CELP coders. It is known that the accuracy of the LP analysis is affected, especially for higher pitch-frequencies, by the under-sampling of the spectral envelope in the frequency-domain, by pitch-period variation and by the influence of spectral fine structure due to long term correlation. Reduced accuracy can cause unnatural frame to frame variation of the envelope, especially around formants, as the frequencies of the pitch-harmonics vary. The accuracy can be improved by spectral estimation techniques which take these effects into account. Examples of such techniques are an approach based on cubic spline interpolation [3], discrete all-pole modelling (DAP) [4], minimum variance distortionless response (MVDR) [7] and iterative all-pole modelling (IAP) [9]. This work is concerned with the suitability of these four techniques for low bit-rate speech coding. Each of the techniques was tested initially with artificial and real speech, the spectra obtained being compared to the known magnitude spectra at the pitch-harmonics. The techniques were then applied to a version of multi-band excitation coding (MBE) [2] operating at 2.4 kb/s, as a means of efficiently representing the spectral amplitudes.

2. ALL-POLE ENVELOPE MODELLING

2.1. Cubic Spline Interpolation Approach

All-pole spectral estimation may be carried out by applying LP analysis to an autocorrelation function derived from the power spectral envelope of a speech segment. An advantage of this approach is that smoothing in the frequency domain may be applied to eliminate fine structure between pitch harmonics which may otherwise affect the accuracy of the LP analysis, especially at high orders or when forms of frequency warping [3] are used for perceptual weighting. For this technique, "cubic spline" polynomials are applied to successive pairs of pitch-harmonics as identified in the short term magnitude spectrum of a voiced speech segment. This fits a smoothed spectral envelope to the harmonic amplitudes and from the corresponding power spectral envelope an autocorrelation function is then obtained by means of an inverse DFT. A standard linear prediction analysis algorithm is now used to calculate the parameters of an all-pole digital filter whose gain response approximates the required spectral envelope.

2.2. Discrete All-Pole Modelling

This iterative technique [4] aims to produce an all-pole power spectrum $\hat{P}(\omega)$ which minimises a discrete version of the Itakura-Saito (IS) distance measure [4]:

$$E = \frac{1}{L}\sum_{m=1}^{L}\left(\frac{P_m}{\hat{P}(\omega_m)} - \ln\left(\frac{P_m}{\hat{P}(\omega_m)}\right) - 1\right) \qquad (1)$$

where ω_m for m = 1, 2, …, L are the pitch-harmonic frequencies in the range 0 to 2π radians/sample and P_m is the power spectral density of the original speech at frequency ω_m. Minimising this distance measure will produce a different all-pole model from that produced by conventional LP which minimises a continuous IS distortion measure [5], and is therefore affected by the spectral energy between the harmonics. The larger the number of harmonics the closer will be the envelope parameters produced by both methods. Let $\hat{P}(\omega)$ be the power spectrum of a p^th^ order all-pole transfer function, i.e.

$$\hat{P}(\omega)=\left|A(e^{j\omega})\right|^{-2} \quad \text{with} \quad A(e^{j\omega})=\sum_{i=0}^{p} a_i e^{-j\omega i} \qquad (2)$$

with a_0 not necessarily equal to 1. Substituting for $\hat{P}(\omega)$ in equation 1 and setting ∂E/∂a_k = 0 for k = 0, 1, …, p gives:

$$\sum_{i=0}^{p} a_i\left(R[k-i]-\hat{R}[k-i]\right)=0 \quad \text{for k = 0,1,…,p} \qquad (3)$$

$$\text{where} \quad R[k]=\frac{1}{L}\sum_{m=1}^{L} P_m e^{j\omega_m k} \qquad (4)$$

$$\text{and} \quad \hat{R}[k]=\frac{1}{L}\sum_{m=1}^{L} \hat{P}(\omega_m) e^{j\omega_m k} \qquad (5)$$

Note that a_i = 0 for i = 0, 1,…, p is not a solution to (3) as $\hat{P}(\omega)$ would be infinite. It is also unlikely that a set of a_i coefficients can be found to make $\hat{R}[k] = R[k]$ for k = 0, 1…, p. The proposed iterative method [4] addresses this set of non-linear equations by writing:

$$\hat{h}[-k]=\sum_{i=0}^{p} a_i \hat{R}[k-i] \ : \ \text{k = 0, 1,…, p} \qquad (6)$$

$$\sum_{i=0}^{p} a_i R[k-i]=\hat{h}[-k] \ : \ \text{k = 0, 1,…, p} \qquad (7)$$

Equation 6 is evaluated to obtain the sequence $\hat{h}[-k]$ for a set of coefficients $a_i = \hat{a}_i$ close to the required ones, initially derived by standard LP analysis. To do this, $\hat{P}(\omega)$ at ω = ω_1, …, ω_L and then $\hat{R}[k]$ for k = 0,1…p, may be derived from equations 2 and 5 respectively. Equation 7 is then solved to obtain a new set of coefficients a_i which, in general, will be closer to the required ones than the set $\hat{a}_i$. To begin a new iteration, set

$$\hat{a}_i(new)=(1-\alpha)\hat{a}_i + \alpha a_i \quad \text{for i = 0,1,…,p} \qquad (8)$$

where α is a “damping” factor between 0 and 1 (typically 0.5). Equation 6 is re-evaluated for $a_i = \hat{a}_i(new)$ and the process continues until E (equation 1) becomes sufficiently small. Alternative ways of solving equation 3 may be found

To explain the idea of DAP, it is easily shown that

$$\hat{h}[k] = \frac{1}{L}\sum_{m=1}^{L} \frac{1}{A\left(e^{j\omega_m}\right)} e^{j\omega_m k} \qquad (9)$$

thus revealing that $\hat{h}[-k]$ is a form of impulse response which corresponds to the complex conjugate of the all-pole spectrum $1/A(e^{j\omega})$ down-sampled at the pitch-harmonic frequencies. In comparison to the true impulse-response corresponding to $1/A(e^{j\omega})$ it will have suffered aliasing in the time-domain, since the number of pitch-harmonics is insufficient to accurately characterise the shape of the envelope. If the true vocal tract transfer function has the power spectrum P(ω), R[k] will be a similarly distorted version of the corresponding autocorrelation function. The ideal is to find coefficients a_i such that the aliassed autocorrelation function R[k] corresponds to the similarly aliassed impulse-response $\hat{h}[k]$ of 1/A(z) .

2.3. MVDR Approach

The minimum variance distortionless response (MVDR) method is an adaptation of Capon’s Maximum Likelihood theorem [8] commonly used in array processing. An application of this method [7] estimates the power spectrum at the pitch-harmonics by calculating the array of output powers that would be obtained if the input signal were applied to a parallel bank of Lth order FIR bandpass filters whose centre frequencies are the pitch harmonics. The impulse response $\underline{h}_m$ = {$h_m[0]$, $h_m[1]$, …, $h_m[L]$} of each of the L filters is calculated such that the pass-band gain is unity and the output power over the range 0 to 2π:

$$\hat{P}_m = \frac{1}{2\pi}\int_0^{2\pi} P(\omega)\left|\sum_{n=0}^{L} h_m[n]\, e^{-j\omega n}\right|^2 d\omega \qquad (10)$$

is the minimum possible. It may be shown [8] that the required impulse response for each m is

$$\underline{h}_m = R^{-1}\underline{e}_m/(\underline{e}_m^H R^{-1}\underline{e}_m) \qquad (11)$$

where R is the (L+1) by (L+1) autocorrelation matrix for the given speech segment, and $\underline{e}_m$ = {1, $e^{-j\omega m}$, $e^{-2j\omega m}$, …, $e^{-Lj\omega m}$}. It follows that

$$\hat{P}_m = \underline{h}_m^H R \underline{h}_m = 1/(\underline{e}_m^H R^{-1}\underline{e}_m) \qquad (12)$$

which is easily calculated for each m in the range 1 to L. This approach produces accurate power spectral estimates which are likely to be less affected by pitch variation and frequency domain sampling than direct DFT methods. To obtain an all-pole parametrisation, an order reduction procedure, involving the calculation of an autocorrelation matrix from the estimated power spectrum, and applying standard LP analysis to this is used to obtain the all-pole parameters.

2.4. Iterative All-Pole Modelling

This technique [9] first calculates the parameters of a standard LP synthesis filter of order M say. A gain factor is then computed which brings the gain response of this filter as close as possible to the amplitudes of the true harmonics. The differences between the true harmonics and the scaled gain response at the frequencies of the harmonics are then added to the scaled gain response, and the corresponding power spectrum is inverse DFT transformed to obtain a new auto-correlation function. Standard LP analysis is now applied to this new auto-correlation function, truncated to the order of the required LP filter. The new synthesis parameters thus obtained are again used to calculate an appropriate gain factor, and the scaled gain response is modified as before. This iterative procedure is continued until the discrete IS distortion (equation 1) between the LP spectrum and the true harmonics shows that significant improvements are no longer being made at the harmonic frequencies. In minimising the error function this technique is iterating towards an interpolation function with an all-pole shape rather than the flatter shape produced by the cubic spline method.

3. INVESTIGATIONS

3.1. Testing with Artificial Speech

To investigate each spectral estimation technique, segments of artificial voiced speech were generated by exciting a sixth order all-pole synthesis filter with an impulse train. Pitch periods in the range 2.5ms-12.5ms were used, and simulated formants were produced typical of a range of different vowel sounds. The pitch-periods were grouped into 'short' (2.5ms-5.6ms), 'medium' (5.6ms-8.1ms) and 'long' (8.1ms-12.5ms). The discrete IS and the log spectral distortion (LSD) measures (equations 1 and 13) were used to measure the differences between the estimated and true spectra at the pitch harmonics. The error measurements obtained for each set of pole positions were averaged for each of the three pitch period ranges. In all cases, the order of the all-pole estimation was six, i.e. the same as that of the synthesis filter. These tests demonstrated how spectral accuracy varies with pitch-period for each technique. The results of the tests are given in table 1.

$$LSD = \sqrt{\frac{1}{L}\sum_{m=1}^{L}(10\log_{10}(P_m) - 10\log_{10}(\hat{P}(\omega_m))^2} \quad (13)$$

PITCH	MEASURE	LP	DAP	MVDR	CS	IAP
SHORT	LSD (dB)	7.3	0.14	2.8	4.9	2.9
	IS	1	0.0006	0.38	0.46	0.19
MEDIUM	LSD (dB)	4.7	0.15	1.1	3.6	3
	IS	0.45	0.0006	0.3	0.26	0.19
LONG	LSD (dB)	3.4	0.16	2.3	3.4	3
	IS	0.29	0.0008	0.12	0.23	0.19

Table 1: Table of averaged spectral distortion results for short, medium and long pitch periods.

According to both measures, DAP provides by far the best spectral fit at all pitch periods, the accuracy being largely independent of pitch. MVDR outperforms IAP according to the LSD measure but not according to discrete IS except for long pitch-periods. CS is better than standard LP, and the accuracy of CS and LP techniques improve significantly as the pitch-period increases. A known problem with iterative techniques such as DAP and IAP is overestimation of the formant amplitudes when the order of analysis is higher than necessary. CS and MVDR do not have this problem.

3.2. Tests with Real Voiced Speech

Each technique was tested with spectra derived directly from segments of real speech, with a mostly voiced content. Tenth order analysis was now used. Again IS and LSD measures were used to examine the degree of spectral distortion with reference to a DFT spectrum for short, medium and long pitch-periods. The results (Table 2) show that again, according to both distance measures, DAP is significantly better than the other techniques. This is especially true for short and medium pitch-periods. For the other techniques IAP and MVDR show some improvements over CS and LP. A small amount of frame to frame averaging reduces the overestimation of formant amplitudes referred to earlier and improves DAP and IAP still further.

PITCH	MEASURE	LP	DAP	MVDR	CS	IAP
SHORT	LSD (dB)	6.5	2.26	4.32	6.06	5.38
	IS	0.68	0.12	0.29	0.58	0.55
MEDIUM	LSD (dB)	6	2.52	4.55	5.23	5.12
	IS	0.57	0.13	0.35	0.47	0.49
LONG	LSD (dB)	6.1	3.46	5.52	5.4	5.04
	IS	0.55	0.22	0.46	0.46	0.39

Table 2: Table of results of estimation technique using a 10th order all-pole model to create the speech envelope.

4. LOW BIT-RATE MBE

The modified LP-MBE algorithm [9] operates at 2.4 kb/s with a frame-rate of 20ms and the spectral amplitudes for each frame modelled by a tenth order all-pole synthesis filter parametrised by quantised LSP coefficients. IAP has previously been used [9] to derive the filter coefficients. A means of deriving a phase spectrum for LP-MBE from the encoded magnitude spectrum has also been investigated [10], and is expected to benefit from increased accuracy in the all-pole model. An investigation was carried out to compare the effectiveness of IAP for this purpose with standard LP analysis and each of the 3 alternative techniques referred to above. To test the performance of each of the techniques, objective measures were derived and informal listening tests were performed.

Overall performance scores, indicating how well each technique modelled the MBE spectral amplitudes, was obtained by averaging the LSD distances (equation 13) between the MBE amplitudes and corresponding samples of the all-pole envelope over approximately 5 second segments of mostly voiced speech. Separate scores were obtained for male and female speech segments.

The results in table 3 show that although DAP is still the most accurate technique, the advantage appears not to be as significant as it was for spectra derived directly from artificial or real speech. The results show that IAP also performs well in modelling the IMBE amplitudes for both male and female speech. The CS and MVDR techniques show little improvement over the LP technique.

Spectral estimation technique	LSD for male speech	LSD for female speech
LP	5.25dB	5.43dB
CS	5.2dB	5.46dB
DAP	4.25dB	4.58dB
IAP	4.52dB	4.86dB
MVDR	5.39dB	5.34dB

Table 3: Results for objective measures of MBE decoded speech.

Conclusions from preliminary informal listening tests carried out using for male and female speech segments correlated broadly with the objective scores. Compared with the standard LP algorithm, all techniques showed some improvement, but the differences between the 4 alternatives were difficult to discern.

CONCLUSIONS

The accuracy of the all-pole spectral model normally provided by classical LP analysis can be significantly improved by applying alternative spectral estimation methods. Of the 4 alternatives investigated, DAP was found to be the most accurate when applied to artificial speech or directly to real speech spectra. When applied to a low bit-rate LP-MBE coder the advantages of DAP over the other 3 alternatives appeared to be less significant.

ACKNOWLEDGEMENTS

This work was carried out at the Universities of Liverpool and Manchester with support from an EPSRC studentship and British Telecom.

REFERENCES

1. Makhoul, J. “Linear Prediction: A Tutorial Review,” Proceedings of the IEEE, Vol.63, No. 4, April1975, pp 561-580

2. Griffin, D.W., and Lim, JS. “Multiband Excitation Vocoder,” IEEE Transactions on ASSP, Vol.36, NO.8, August 1988, pp 1223-1235

3. McAulay, R.J., and Quatieri, T.F. “Low-Rate Speech Coding based on the Sinusoidal Model”, in “Advance in Speech Signal Processing” Edited by Sadaoki, Furui, M. Mohan Sondh, Marcel Dekker, Inc. 1992, (New York), pp.165-208.

4. El-Jaroudi, A., and Makhoul, J. “Discrete All-Pole Modeling,” IEEE Transactions on signal processing, Vol. 39 No. 2, February 1991 pp 411-423

5. Itakura, F., and Saito S. “A statistical method for estimation of speech spectral density and formant frequencies,” Electron. Commun.,Japan, vol. 53-A, pp.36-43, 1970

6. McAulay, R.J. “Maximum Likelihood Spectral Estimation and Its Application to Narrow-Band Speech Coding,” IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.32, NO. 2, April 1984

7. Murthi, M.N., and Rao, B.D. “Minimum Variance Distortionless Response (MVDR) Modelling of Voiced Speech,” ICASSP 97, International Conference on Acoustics, Speech and Signal Processing. Munich 1997

8. Zelniker, G., and Taylor, F.J. “Advanced Digital Signal Processing Theory and Applications,” Marcel Dekker Inc. 1994 pp 601-603

9. Parris, C.I., Wong, D., and Chambon, F. “A Robust 2.4kb/s LP-MBE With Iterative LP Modelling,” ESCA. Eurospeech95. 4th European Conference on Speech Communication and Technology. Madrid, September 1995

10. Cheetham, B.M.G., Choi, H.B., Sun, X.Q., Goodyear, C.C., Plante, F., and Wong, W.T.K., “All-pass excitation phase modelling for low bit-rate speech coding”, IEEE Symposium on circuits and Systems, ISCAS’ 97, Hong Kong, June 1997