A LOW-BIT-RATE SPEECH CODER USING ADAPTIVE LINE SPECTRAL FREQUENCY PREDICTION

C. W. Seymour and A. J. Robinson

Cambridge University Engineering Department
Cambridge, CB2 1PZ, UK
{cws, ajr}@eng.cam.ac.uk

ABSTRACT

This paper describes two aspects of a linear predictive coding (LPC) vocoder developed for operation on wide-band speech. The method for encoding the LPC parameters, based on the use of an adaptive predictor, is presented together with an extension to the vocoder model which enables it to operate on speech sampled at 16kHz rather than 8kHz. Good-quality operation on wide-band speech is achieved with an increase in bit rate of about 500 bits/s. Diagnostic rhyme test (DRT) results demonstrate the improvement in intelligibility gained through coding speech at the higher sample rate.

1. INTRODUCTION

The linear predictive coding (LPC) vocoder provides an efficient way to code speech at low bit rates. This paper focuses on two aspects of a coder developed for operation on wide-band speech. In the first part of the paper, the method used to code the LPC parameters is described. This is based on encoding the error signal for an adaptive predictor operating in the line-spectral frequency [1] (LSF) domain. The second part of the paper describes an improvement to the vocoder model which enables it to efficiently code speech sampled at 16kHz rather 8kHz.

Once the LPC parameters have been converted to the LSF domain, coding consists of 3 stages: prediction, quantisation and encoding. The use of scalar quantisation and encoding avoids the overhead of codebook storage and search encountered with VQ-based schemes [2], while the use of a predictor which includes previous elements from within the same frame allows intra-frame correlations to be modelled. The predictor is adaptively updated, ensuring that it is matched to the acoustic conditions and speaker.

Traditionally LPC vocoders have only been able to efficiently code speech sampled at little more than 8kHz (i.e. having a bandwidth of 4kHz). However, the higher-frequency components of the spectrum contribute to both speech intelligibility and quality. Therefore it would be desirable to encode speech at a higher sample rate if this could be achieved without substantially increasing the bit rate. This paper also presents a method for coding speech at higher rates which makes use of the fact that the spectral resolution required decreases with frequency. A low-order LPC model is employed to represent the region of the spectrum between 4 and 8kHz.

2. LSF ENCODING

2.1. Prediction

The prediction stage estimates the current LSF component from data currently available to the decoder. The entropy of the prediction error distribution is expected to be lower than that of the original values, hence it should be possible to encode this at lower bit rate for a given average error.

Let the LSF element i at time t be denoted $l_i(t)$ and the LSF element recovered by the decoder be denoted $\bar{l}_i(t)$. If the LSFs are encoded sequentially in time and in order of increasing index within a given time frame, then to predict $l_i(t)$, the following values are available: $\{\bar{l}_j(t)|1 \leq j < i\}$ and $\{\bar{l}_j(\tau)|\tau < t \text{ and } 1 \leq j \leq P\}$. Therefore a general linear LSF predictor can be written

$$\hat{l}_i(t) = c_i + \sum_{\tau=t-t_0}^{t-1} \sum_{j=1}^{P} a_{ij}(t-\tau)\bar{l}_j(\tau) + \sum_{j=1}^{i-1} a_{ij}(0)\bar{l}_j(t), \tag{1}$$

where $a_{ij}(\tau)$ is the weighting associated with the prediction of $\hat{l}_i(t)$ from $\bar{l}_j(t-\tau)$. A non-linear predictor is also considered, where $l_i(t)$ is estimated from a set of features augmented with the 2nd order cross terms such as $\bar{l}_i(t-1)^2, \bar{l}_i(t-1)\bar{l}_{i-1}(t)$ etc.

In general only a small set of values of $a_{ij}(\tau)$ will be used, as a high-order predictor is computationally less efficient both to apply and to estimate. Experiments were performed on unquantised LSF vectors (i.e. predicting from $l_j(\tau)$ rather than $\bar{l}_j(\tau)$) to examine the performance of various predictor configurations. For these experiments the predictor was trained and evaluated on separate data sets consisting of sentences from the WSJ-CAM0 database [3]. However, a scheme was also implemented where the predictor is adaptively modified to match each test sequence. The adaptive update is performed according to

$$\begin{aligned} \mathbf{C}_{xx}^{(k+1)} &= (1-\rho)\mathbf{C}_{xx}^{(k)} + \rho \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \\ \mathbf{C}_{xy}^{(k+1)} &= (1-\rho)\mathbf{C}_{xy}^{(k)} + \rho y_i \mathbf{x}_i, \end{aligned} \tag{2}$$

where ρ determines the rate of adaption[1]. The terms $\mathbf{C}_{xx}$ and $\mathbf{C}_{xy}$ are initialised from training data as $\mathbf{C}_{xx} = \frac{1}{N}\sum_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ and $\mathbf{C}_{xy} = \frac{1}{N}\sum_i y_i \mathbf{x}_i$. Here y_i is a value to be predicted ($l_i(t)$) and $\mathbf{x}_i$ is a vector of predictor inputs (containing $1, l_i(t-1)$ etc.). The updates defined in equation 2 are applied after each test frame, and periodically new MMSE predictor coefficients, $\mathbf{p}$, are calculated by solving $\mathbf{C}_{xx}\mathbf{p} = \mathbf{C}_{xy}$.

[1] A value of $\rho = 0.005$ was found suitable, giving a time constant of 4.5 seconds.

Table 1 shows the prediction error for a number of possible schemes. Each system is defined by the set of coefficients used to predict $l_i(t)$, listed in the column labelled "Elements"[2]. The suffixes -X and -A denote predictors containing cross terms and adaptive predictors respectively. Column "M" lists the maximum predictor order for each system. For these experiments, analysis was performed using a frame period of 22.5ms and an LPC order of 10 (hence giving 10 LSFs). Separate predictors were used for frames classified as voiced and unvoiced.

Sys	M	Elements	MS err./10^{-4}
A	0	-	4.50
B	1	$a_{ii}(1)$	2.42
C	2	$a_{ii}(1), a_{i\,i-1}(0)$	1.86
D	3	$a_{ii}(1), a_{i\,i-1}(0), a_{i\,i-1}(1)$	1.68
D-X	9		1.62
D-A	3		1.60
E	2	$a_{ii}(1), a_{ii}(2)$	2.38
F	19	$a_{ij}(1) \vert 1 \leq j \leq P$, $a_{ij}(0) \vert 1 \leq j \leq i-1$	1.60

Table 1: Prediction error for various LSF prediction configurations.

For system A, the predictors are simply the mean of each LSF. It can be seen that extending the predictor to estimate a given LSF from coefficients within the current and previous frames gives an improvement in prediction error. However, the small difference between the errors for systems B and E suggests that there is little to be gained by predicting from frames further back in time. The use of all the available LSFs from these two frames (system F) reduces the prediction error, but at the expense of a greatly increased predictor order. Therefore system D (shown in figure 1) was selected as giving the best compromise between efficiency and error. The use of cross terms (D-X) gives an improvement in performance, but again at the expense of increased predictor order. The improvement due to the adaptive predictor is quite small in this example; however, this was obtained using very similar training and test sets. In general the adaptive predictor can be expected to take into account differences between training and test conditions caused for example by speaker variations, channel differences or background noise. Histograms of prediction error over frames of speech classified as voiced for the 3rd LSF component for systems A and D are shown in figures 2 and 3 respectively.

2.2. Quantisation and Coding

Given a predictor output $\hat{l}_i(t)$, the prediction error is calculated as $e_i(t) = l_i(t) - \hat{l}_i(t)$. This is uniformly quantised to give an error $\bar{e}_i(t)$ which is then losslessly encoded. Coarser quantisation can be applied to frames classified as unvoiced. The coding scheme used is a Rice code [4], where the number of bits allocated increases with the magnitude of the error signal. Therefore this method is suitable for applications which do not require a fixed number of bits to be generated per frame. Figure 4 shows the overall coding scheme.

[2]In each case, c_i is included in the predictor.

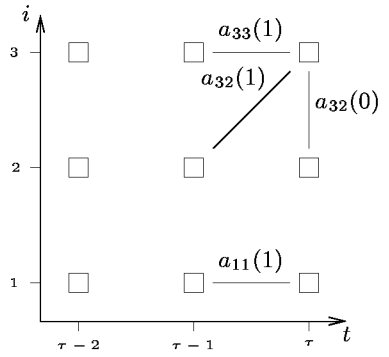

Figure 1: LSF linear predictors for system D.

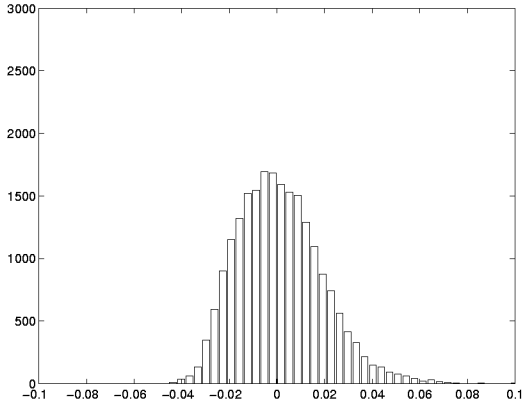

Figure 2: Prediction error for system A.

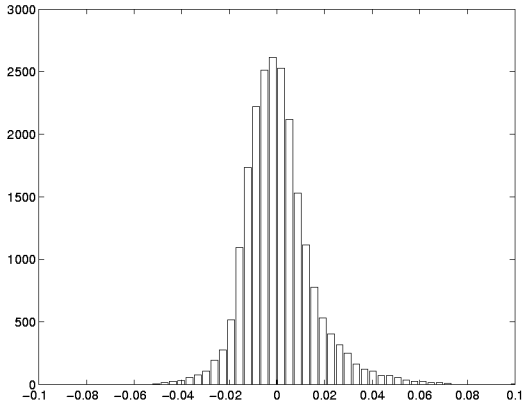

Figure 3: Prediction error for system D.

2.3. Results

Figure 5 shows average spectral distortion against LSF bit rate in terms of bits per frame (for frames of speech classified as voiced) achieved with a number of the systems listed in table 1. These experiments were performed using an LPC model of order 10 and a frame period of 22.5ms. The adaptive predictor was found to give an improvement in rate for a given spectral distortion in some instances, and in general ensures that the prediction error is minimised. However, for conditions reasonably matched to the training data, there was little difference

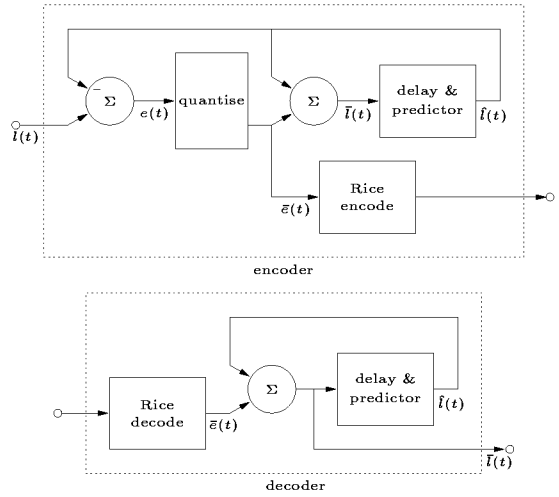

Figure 4: Overall coding scheme.

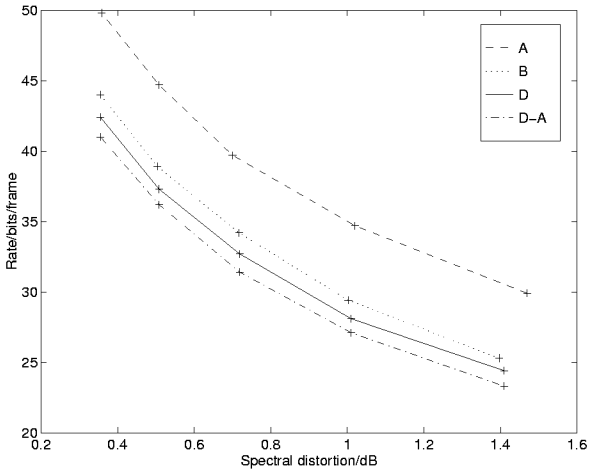

Figure 5: LSF bit rate against spectral distortion.

in performance. The results shown are for a segment of broadcast radio news, where adaption was found to be of benefit. An additional advantage of the adaptive scheme is that it allows low prediction error to be achieved without the use of initial predictors estimated from a period of training data.

3. WIDE-BAND OPERATION

For an LPC vocoder operating on speech sampled at 8kHz, an LPC model order of 10 is commonly used, giving a good compromise between quality and bit rate. With a sample rate of 16kHz, however, it is necessary to almost double the model order to achieve a good representation of the speech spectrum.

The approach used here is to first perform sub-band decomposition to split the speech signal into bands corresponding to the 0–4kHz and 4–8kHz regions of the spectrum. It was found that this could be achieved in a computationally efficient manner using 8th-order elliptic filters. High and low-pass filters are applied, and the resulting signals decimated to form the two sub-bands. The high sub-band contains a mirrored form of the 4–8kHz spectrum. As usual, 10th order LPC analysis is performed on the lower band. However, for the upper band, 2nd order analysis was found to be adequate. Therefore with this scheme, 3 additional parameters per frame of speech (2 LPCs and one gain) must be encoded. Figure 6 shows the short-time spectrum and two sub-band spectra for a frame of unvoiced speech, with the frequency axis for the upper sub-band adjusted to take into account the spectral shift and mirroring.

3.1. Synthesis from Sub-Band Models

Given LPC parameters corresponding to the two sub-bands, a method for resynthesising the speech waveform is required. The approach adopted is to combine the two sub-band models to form a higher-order LPC model representing the wide-band signal. A model order of 18 was found to be suitable here. Combination methods operating in the spectral and autocorrelation domains were considered. The spectral-domain technique involves computing the power spectral densities for the two sub-band models, which may be achieved using FFTs, and then combining the spectra and calculating the wide-band autocorrelation by an inverse FFT. The problem with this approach is that it is computationally inefficient due to need to use reasonably high-resolution FFTs. A more efficient autocorrelation-domain technique is presented below.

In the following work, subscripts L and H will be used to denote features of hypothesised low-pass and high-pass filtered[3] versions of the wide-band signal respectively, and subscripts l and h used to denote features of the lower and upper sub-band signals respectively.

The autocorrelations of low-pass and high-pass versions of the wide-band signal, $r_L(\tau)$ and $r_H(\tau)$, are generated. The low-pass filtered wide-band signal is equivalent to the lower sub-band up-sampled by a factor of 2. In the time-domain, this up-sampling consists of inserting alternate zeros (interpolating), followed by low-pass filtering. Therefore in the autocorrelation domain, up-sampling involves interpolation followed by filtering by the autocorrelation of the low-pass filter impulse response.

The autocorrelations of the two sub-band signals can be efficiently calculated from the sub-band LPC models [5]. If $r_l(m)$ denotes the autocorrelation of the lower sub-band, then the interpolated autocorrelation, $r'_l(m)$, is given by

$$r'_l(m) = \begin{cases} r_l(m/2) & \text{if m} = 0, \pm 2, \pm 4, \ldots \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The autocorrelation of the low-pass filtered signal, $r_L(m)$, is

$$r_L(m) = r'_l(m) * (h(m) * h(-m)), \quad (4)$$

where $h(m)$ is the low-pass filter impulse response. The autocorrelation of the high-pass filtered signal, $r_H(m)$, is found similarly, except that a high-pass filter is applied.

The autocorrelation of the wide-band signal, $r_W(m)$, can be expressed

$$r_W(m) = r_L(m) + r_H(m), \quad (5)$$

and hence the wide-band LPC model calculated. Figure 7 shows the resulting LPC spectrum for the frame of

[3]Assuming filters having cut-offs at 4kHz, with unity response inside the pass band and zero outside.

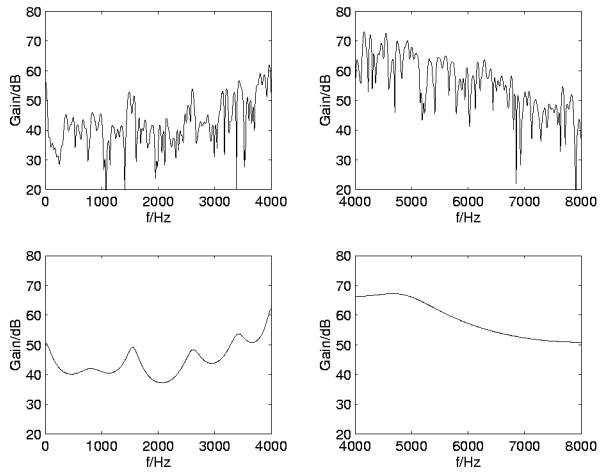

Figure 6: Split-band short-time and LPC spectra.

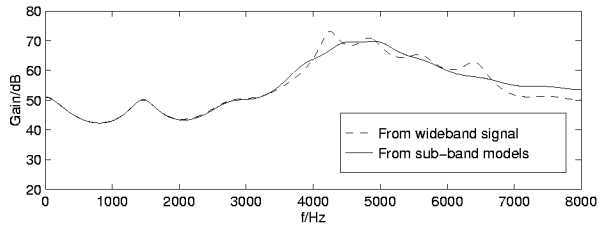

Figure 7: Wide-band LPC spectra from wide-band signal and sub-band LPC models.

speech considered above. The dotted line corresponds to an order 18 LPC model calculated from the wide-band signal, whereas the solid line is for an LPC model of the same order calculated using the combination technique described above.

FIR filters of order 30 were found to be sufficient to be used in the up-sampling. In this case, the poor frequency resolution implied by the low order filters is adequate because this simply results in spectral leakage at the crossover between the two sub-bands. This approach resulted in speech perceptually very similar to that obtained by using an high-order analysis model on the wide-band speech. The LPC parameters corresponding to the high sub-band are encoded using the prediction residual coding technique described in section 2. However, simple differences are used, rather than the output from a trained predictor. The two LPCs and gain for the high sub-band resulted in an additional average bit rate of about 500 bits/s.

4. RESULTS

Diagnostic rhyme tests [6] (DRTs) were performed to compare the intelligibility of the wide-band and narrow-band vocoders operating at a frame period of 16ms. The coder includes the adaptive spectral enhancement and pulse dispersion filters used by the MELP coder [7]. Mixed pulse and noise signals are used for voiced excitation, but the relative weights are kept constant in each frequency band. The pitch and gain are encoded by applying a Rice code to the time differences of these parameters. The degree of quantisation of the LSF prediction errors was adjusted so that the narrow-band and wide-band coders both operated at an average bit rate of approximately 2400bps.

Tests were also performed on the 4800 bps CELP coder (Federal Standard 1016) and the LPC-10e coder (compatible with Federal Standard 1015) operating on narrow-band speech. The results shown in table 2 indicate that wide-band operation gives an improvement in DRT score of about 2. Six listeners were used for the DRTs, with every person presented with 96 stimulus words from each coder. Informal listening tests indicated that in addition to improving intelligibility, wide-band operation resulted in a considerable improvement in overall speech quality.

Coder	DRT score
LPC-10e	68.0
CELP	83.8
Narrow-band coder	84.4
Wide-band coder	86.8

Table 2: DRT results for wide-band LPC vocoder and narrow-band CELP coder.

5. CONCLUSIONS

An LSF coding scheme using an adaptive predictor and a sub-band technique for wide-band vocoding have been presented. DRT results indicate that wide-band operation allows improved intelligibility to be achieved at a given bit rate.

6. ACKNOWLEDGEMENT

This research was funded by Hewlett Packard Laboratories, Bristol, UK.

7. REFERENCES

[1] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, S35(A), 1975.

[2] M. Yong, G. Davidson, and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988, pp. 402–405.

[3] A. J. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1995, pp. 81–84.

[4] R. F. Rice and J. R. Plaunt, "Adaptive variable-length coding for efficient compression of spacecraft television data," *IEEE Trans. Communication Technology*, vol. 19, no. 6, pp. 889–897, 1971.

[5] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*, p. 527, Addison-Wesley, 1987.

[6] W. D. Voiers, "Diagnostic evaluation of speech intelligibility," in *Speech Intelligibility and Speaker Recognition*, Mones E. Hawley, Ed., pp. 374–387. Dowden, Hutchinson & Ross, Inc., 1977.

[7] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 242–250, July 1995.