Proceedings of The Institute of Acoustics

IMPROVED VECTOR EXCITATION CODING USING WIENER FILTERING

M.Niranjan and F.Fallside

Cambridge University Engineering Department

ABSTRACT

This paper deals with a lov bit rate speech coding scheme based on a source filter model. We viev the excitation signal as a zero mean Gaussian stochastic process, which is vector quantised. The coefficients of an adaptive synthesis filter and an excitation vector are computed to minimise the expected squared error between the synthetic and natural versions of the vaveform. The synthesis filter models an ARMA process and is computed in two phases together vith a first order comb filter to synthesise the pseudo-periodicity of voiced speech. The AR part is first calculated by the autocorrelation method of Linear Prediction. Then, coefficients of the MA part are determined to minimise the error between the original and synthetic vaveforms.

INTRODUCTION

A number of speech production models suitable for data compression and synthesis have been reported in the past fev years. Most of these models are variations of the source filter model and are characterised by the assumption of pseudo-stationarity. Atal and Shroeder (1984) proposed one such model and called it the stochastic model. The encoding scheme vas called stochastic coding and vas essentially based on vector quantising the excitation signal for a source filter model. The filter used vas a pitch synthesis filter and an all-pole filter. The target vas to encode speech data at bit rates belov 9.6 kbits/sec vithout loss of quality. This coder is characterised by very high computational complexity.

In this paper ve will present our vork on improving the Stochastic Coder. The first part of the paper will introduce the problem, starting from a basic residual excitation scheme, and describe the operation of the stochastic coder. We will then describe the improved system and present some preliminary results.

VECTOR EXCITATION CODING

RESIDUAL EXCITATION

Fig.1 shows a speech analysis synthesis system. In the analysis part there are two filters. The filter $\sum_{k=1}^{p} a_k z^{-k}$, predicts a speech sample s(n) as a weighted sum of p samples immediately preceding it. Subtracting the predicted value from s(n) gives a difference signal d(n). This process attempts to remove the envelope of the short-time spectrum of the speech signal. The short-time spectrum of d(n) is therefore flat except for the comb-like fine structure introduced by the pseudo-periodicity of voiced speech. The second filter βz^{-M} has a long memory equal to one pitch period and predicts d(n) from d(n-M), a sample of the difference signal that occurred one pitch period before. The filter coefficients $a_1,\ldots,a_p$ and β are estimated to minimise the variance of the residual r(n) over a short period of time.

On the synthesis side, the same two filters are arranged in a feedback path shown in the right half of Fig.1. This will give overall transfer functions of $1/(1 + \beta z^{-M})$ and $1/(1 + \sum_{k=1}^{p} a_k z^{-k})$ for the two blocks respectively. The former gives the spectral fine structure of voiced speech and the latter is an all-pole synthesis filter. β ($0<\beta<1$) is a measure of the degree of voicing and will be very low for unvoiced sounds. It must be noted that with no quantisation present, $\hat{s}(n) = s(n)$, i.e. the synthetic signal is identical to the original signal.

Now, in order to achieve data compression, we need to introduce quantisation into the scheme. The quantisation properties of the all-pole filter have been studied in detail (Viswanathan and Makhoul, 1975). An optimum quantisation strategy is to transform the a_k into log area parameters and quantise these parameters.

Conventional vocoders based on Linear Prediction use only the all-pole synthesis filter, excited by a train of impulses, spaced one pitch period apart, for voiced sounds and white noise for unvoiced sounds. This can be viewed as a very crude quantisation of the signal d(n) that should excite the all-pole filter. This is reflected in the loss of quality in the synthetic signal produced by such vocoders. Further, the all-pole synthesis filter is not always capable of modelling the highly non-linear process that is going on in the vo-

cal system. This is especially true for nasal sounds where, due to the parallel nasal tract, the short-time spectrum is characterised by zeros as well.

A number of methods to quantise the excitation signal have been reported. Sluyter, Bosscha and Schmitz (1984) have developed a speech compression scheme for mobile radio. Their excitation is a down-sampled version of the residual. Atal (1982) describes a scheme with centre clipping, based on the observation that the high amplitude portions of the residual have to be coded more accurately. In such schemes, one explicitly computes the residual signal and then looks for a good quantisation scheme.

Multi-pulse excitation (Atal and Remde, 1982) and Stochastic coding (Atal and Schroeder, 1984) introduce a new approach to the problem of modelling the excitation. Both the schemes present the concept of computing the parameters of the excitation model in an analysis by synthesis framework; i.e. we compute these parameters such that some measure of error between the original and synthetic signals is minimised. The short-time Fourier transform (STFT) based model of Griffin and Lim (1985), presents this view as well. Such an analysis by synthesis framework offers the advantage that during the analysis we can make an allowance for the subsequent quantisation. Due to this, one part of the system will in fact compensate for quantisation errors introduced by the rest of the system.

VECTOR EXCITATION

The Stochastic Coder (Atal et al 1984) is based on vector quantising the excitation signal. Subsequently, this scheme has been referred to as Code Excited Linear Prediction or CELP (Schroeder and Atal, 1985) and Vector Excitation coding or VXC (Davidson and Gersho, 1986). A short block of residual is treated as a vector and is vector quantised. The encoder and decoder have an inventory of possible excitation vectors. During a particular interval of time, an optimum excitation vector is determined at the encoder and an index that points to this vector is transmitted to the decoder, together with information about the synthesis filter. Since the residual signal after short and long delay prediction is assumed to be white, the codebook contains a collection of random number sequences with Gaussian statistics. This approach has the benefit that the problem of designing a codebook by clustering, encountered in vector quan-

tisation, is in fact circumvented.

Since the optimum excitation vector is chosen from the codebook by synthesising with all the candidate vectors, vector excitation coding is computationally very complex. With a 10 bit codebook (1024 entries) and vectors of dimension 40, we need about 50000 multiply-add operations per sample. Trancoso and Atal, (1986) and Davidson et al, (1986) have suggested several techniques to improve the complexity aspects. The improvement proposed in this paper is on the synthesis quality, by improving on the synthesis filter. We describe this in the next section.

THE IMPROVED VECTOR EXCITATION CODER

In our improved vector excitation model, we not only choose an optimum excitation vector from a codebook, but also estimate part of the synthesis filter to give best possible synthesis, in the least squares sense. In general, the synthesis filter with a rational transfer function will have poles and zeros, and may be described as, $H(z) = \dfrac{\sum_{k=1}^{q} b_k Z^{-k}}{1 + \sum_{k=1}^{p} a_k Z^{-k}}$, $\quad q < p$.

Estimating a_k, $k=1,\ldots,p$ and b_k, $k=1,\ldots,q$ is an ARMA (Auto Regressive Moving Average) estimation problem and is in general non-linear. In our system these coefficients are estimated in two phases. The a_k are determined by the autocorrelation method of LP analysis. Only the MA part of the synthesis filter (i.e. b_k) are computed in the minimisation loop.

A 12^{th} order LP analysis is carried out on a 15 ms length of speech, sampled at 10kHz and weighted by a Hamming window. Pitch is determined by a time domain peak picking algorithm (Allerhand 1986). Every excitation vector v_i is first scaled by a gain factor γ_i and then filtered through a cascade of the pitch synthesis filter $1/(1 + \beta Z^{-M})$ and the all-pole synthesis filter $1/(1 + \sum_{k=1}^{p} a_k Z^{-k})$. The scale factor γ_i and the pitch synthesis filter gain β are computed to minimise the mean squared error at the output (Lin, 1986). This gives a number of potential synthesis waveforms. Then for each of these

waveforms, the coefficients b_k, k=1,...,q of a filter of the form $\sum_{k=1}^{q} b_k z^{-k}$ are calculated such that the filtered signal is closest to the original speech.

Let c(n) be a synthetic signal after pitch and all-pole synthesis and s(n), the original speech segment. The optimum b_k are the solution to the Wiener-Hopf equations,

$R_{cc}\, b = r_{sc}$,

where, R_{cc} is the qxq autocorrelation matrix of c(n) and r_{sc} is a q dimensional vector, the i^{th} element of which is given by,

$$r_i = \sum_{n=1}^{N} c(n)\, s(n-i)$$

RESULTS

We have simulated the coding scheme described above on a VAX-750. With a very large codebook and no quantisation of the filter parameters, the improvement due to the additional MA part is negligibly small. However, when the filter coefficients were quantised, the MA part does help in compensating for the quantisation noise. This is shown in the waveforms of Fig.3, where both the synthesis waveforms are obtained at equal bit rates of 10 kbit/s and the codebook used consisted of 64 vectors of dimension 50.

CONCLUDING REMARKS

We have described an improvement to a poverful speech compression technique. In this improvement we suggest that, in addition to searching a codebook for an optimal excitation sequence, one should also attempt to estimate the synthesis system to minimise the synthesis error. In our system these parameters are estimated sequentially and quantised. Therefore, the system is sub optimal.

Vector excitation coders are very complex because we have to search through a large inventory of innovation vectors for an optimum one. In our approach we attempt to make the best use of every candidate vector. This means that we can achieve low distortions with a smaller inventory leading to reduced complexity.

ACKNOWLEDGEMENT

Niranjan wishes to thank Ir. R.N.J. Veldhuis of Philips Research Laboratories, The Netherlands who created this interest in him.

REFERENCES

[1] Allerhand M.H., "A knowledge based approach to speech pattern recognition", Ph.D. dissertation, Cambridge University, 1986.

[2] Atal B.S, "Predictive coding of speech at lov bit rates", IEEE Trans. Commun. vol. COM-30, pp. 600-614, April 1982.

[3] Atal B.S. and Remde J.R., "A new model for LPC excitation for producing natural sounding speech at low bit rates", Proc. I.C.A.S.S.P., pp. 614-617, Lim1982.

[4] Atal B.S. and Schroeder M.R., "Stochastic coding of speech signals at very lov bit rates", Proc. Int. Conf. Commun. - ICC84, part2, pp. 1601-1613, May 1984.

[5] Davidson G. and Gersho A., "Complexity reduction methods for vector excitation coding", Proc. I.C.A.S.S.P., pp. 3055-3058, April 1986.

[6] Griffin D.W. and Lim J.E. "A new model-based speech analysis/synthesis system", Proc. I.C.A.S.S.P., pp. 513-516, March 1985.

[7] Lin D., "New approaches to stochastic coding of speech sources at very lov bit rates", Proc. EUSIPCO-86, pp 445-447, September 1986.

[8] Sluyter R.J., Bosscha G.J. and Schmitz H.M.P.T., "A 9.6 kbit/s speech coder for mobile radio applications", Proc. Int. Conf. Comm. ICC84, pp. 1159-1162, May 1984.

[9] Trancoso I.M. and Atal B.S., "Efficient procedures for finding the optimum innovation in stochastic coders", Proc. I.C.A.S.S.P., pp. 2375-2378, April 1986.

[10] Visvanathan R. and Makhoul J., "Quantisation properties of transmission parameters in linear predictive systems", IEEE Trans. A.S.S.P. ASSP-23, pp. 309-321, June 1975.

VECTOR EXCITATION CODING

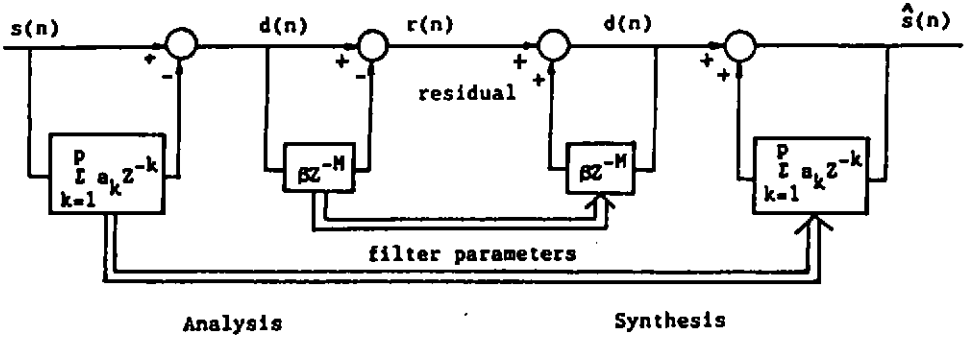

Figure 1. Speech analysis/synthesis scheme

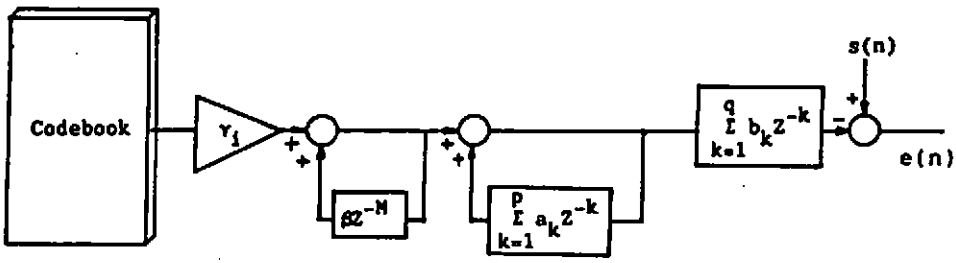

Figure 2. The Vector excitation coding system

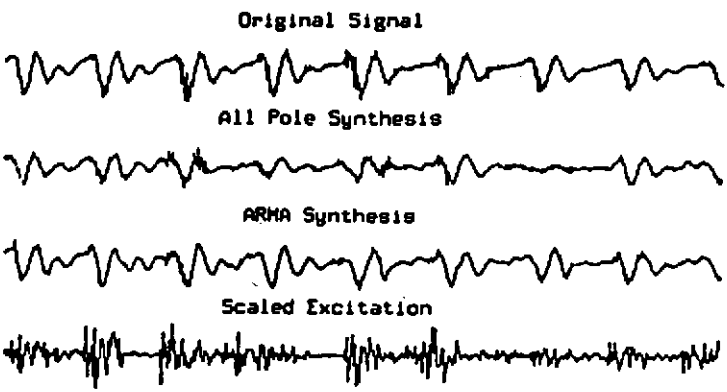

Figure 3. Original, synthetic and excitation waveforms