Proceedings of The Institute of Acoustics

Speech Coding at 32 k bit/s for use in the Switched Telephone Network

L.S. Moye and J.A.S. Angus

Standard Telecommunication Labs. Ltd.,
Harlow, Essex.

INTRODUCTION

If the bit-rate needed to give telephone quality speech is to be reduced below the 64 k bit/s currently used, it is necessary to have a coding system that makes use of the properties of speech production. This normally involves the introduction of delay as the signal is analysed and resynthesised. The use of 2-wire transmission in the switched telephone network causes echoes, and the conversational impairment that these cause can only be kept to a minimum by avoiding any unnecessary transmission delay. Thus, any 32 k bit/s speech coding system for the switched telephone network must introduce negligible delay, preferably none.

Only one class of speech coder is capable of making use of the properties of speech production without introducing delay - the backward adaptive predictive (BAP) coder (which most people misleadingly insist on calling ADPCM). There is only one disadvantage to this class of coder: in theory the idea appears very elegant, in practice it does not work well because of a phenomenon which we call capsizal [1,2]. Only by reducing the performance of the BAP coder can capsizal be overcome. This paper describes the nature of capsizal and discusses the merit of a new BAP coder structure to enable it to be overcome with minimum loss of system performance.

THE BACKWARD ADAPTIVE PREDICTIVE (BAP) CODER

The BAP coder is shown in fig. 1. It removes redundancy from the speech signal by predicting the current sample from the past output signal. This reduces the amplitude of the residual signal so that the residual coders (typically adaptive pcm coders) can code the smaller signal with less quantising noise. As shown in the equations accompanying fig. 1, the reconstruction of the signal at the receiver does not increase the quantising noise, so that the signal-to-noise ratio is enhanced over that of the residual coder on its own in proportion to the reduction of signal amplitude by the predictor (prediction gain).

The coder contains a local decoder to provide a replica of the output signal at the transmit end so that the prediction can be the same in both cases. No side information needs to be transmittted to control the filters, so there is no framing delay, and the straight through path from input to residual coder and residual decoder to output means that the system need be implemented to introduce no more delay than that of the residual coders, usually only one sample.

CAPSIZAL

The appropriate predictor for a speech signal is a transversal filter, which appears in a feed-forward configuration in the coder acting as an all-zero filter to remove the vocal tract poles, and in a recursive configuration in the

decoder to act as an all-pole filter and reinsert them. Because the excitation waveform is convolved with the vocal tract response, we call this transmit-end structure a deconvolver, and the recursive receive-end structure a reconvolver. This implies that the receive-end structure has the ability to be the exact inverse of the transmit end structure.

Unfortunately, this is not true, as can be seen from fig. 2 (which was used in the presentation of Ref. 1). This plot was obtained from a simulation in which the residual coders were removed and an "analogue" error introduced into the transmission path. Up to the point at which the error is introduced, the output is necessarily identical to the input because the signals in deconvolver and reconvolver are numerically identical and no quantization error has been introduced. After the error it can be seen to differ slightly, but eventually (in fig. 2(b)) the reconvolver starts replacing resonances different from those removed by the deconvolver. Careful study of the waveforms shows what has happened. The deconvolver has removed mainly the first formant and left mostly the second formant in the residual. The reconvolver has done the best it can and reinforced the resonances that it finds in the residual, giving an output dominated by second formant. We call this behaviour of the reconvolver "capsizal". Up to the point at which the error occurs,the reconvolver is in unstable equilibrium controlled precisely by the deconvolver which is connected in a feedback configuration. After the error, the reconvolver will never again have signals in it identical to those in the deconvolver (unless by chance). The deconvolver acts as a resonance attenuator and the reconvolver, not as its exact inverse, but as a resonance amplifier. Once something has occurred to destroy the perfect synchronism of the two - either a single error, or an asynchronous start - the reconvolver will for ever more amplify the largest peaks in the residual spectrum most. If the deconvolver attenuates the largest formant to below the level of a secondary one, then capsizal will occur.

Capsizal occurs because of the non-uniform removal of the spectral structure by the deconvolver. This is partly due to the limited order of the filter which means that it can only remove a few of the major features, possibly leaving the minor ones dominant; and partly due to the use of iteratively adapted filters. In these filters, the rate of adaptation at each frequency is proportional to the power in the input signal at that frequency, so that in a time-varying situation only the most prominent peaks of the spectrum are satisfactorily removed.

The potential for capsizal is implicit in the structure of the BAP coder. The only way to alleviate it is to ensure that the deconvolver does not modify the spectrum so that the wrong peaks become dominant. Various filter structures have been proposed for doing this, but all of them have the limitations mentioned above [3 - 11]. Partly because the phenomenon of capsizal does not seem to have been well understood, but confused with the problem of reconvolver filter instability (which can easily occur during capsizal), these structures have been devised to control stability by constraints on the coefficients and have reduced the tendency to capsize by poor adaptation algorithms which implicitly reduce the prediction gain. There is not room in this paper to describe these all in detail, and that will be done later elsewhere.

To reduce the tendency to capsize whilst maintaining prediction gain, it is

necessary to try to attenuate the input signal spectrum structure more uniformly. A new BAP coder filter structure for doing this is described in the next section.

THE PARTIAL DECONVOLUTION LATTICE STRUCTURE

The problem of predicting the current sample of a signal from past samples is the classic problem of linear prediction. Many methods exist for calculating the optimum coefficients for a linear predictor filter (as opposed to approximating them by iterative adaptation as described above), but most of these involve calculations on a block of data [12]. The most suitable structure for a filter that has to be updated sample-by-sample as in the BAP coder is the lattice filter [13,14] in which the coefficients are calculated in a feed-forward fashion in each stage as shown in fig. 3. Because such a filter has optimum coefficients for a given error weighting function, its use in a deconvolver results in complete deconvolution and ensures capsizal. This has already been observed, and methods involving coefficient decay and iterative coefficient adaptation used to overcome it [8], the attraction of the lattice filter being that by constraining its coefficients to lie between ±1 it can be prevented from becoming unstable. These techniques have all the disadvantages of using non-lattice filters as mentioned above.

If, instead of using the output A in fig. 3, the partially decorrelated output B is used, the correlation can be removed from the signal in a controlled but uniform way. When this structure is applied to the BAP coder, we get the structure shown in fig. 4. This shows the deconvolver, from which the reconvolver structure can be inferred.

PERFORMANCE OF THE LATTICE BAP CODER

Comparison of the performances of different forms of the BAP coder are rather difficult because of the difficulty of measuring the extent of capsizal and establishing comparability in adaptation rates etc. between systems which adapt in different ways. To illustrate the relative superiority of the configuration described in the previous section, we have used a system with no coding of the residual, but in which the reconvolver is started after the deconvolver. Fig. 5(a) shows the response of a lossy adaptive transversal filter system which, although it is not exhibiting the classic capsizal phenomenon, is behaving rather badly. Fig. 5(b) shows a similar plot for a partial deconvolution lattice structure. It can be seen that the coefficients adapt very quickly to a capsized condition, but then readapt to a completely correct state. It should be noticed that the residual for the lattice filter is much smaller than for the transversal filter, and that the transversal filter coefficients, apart from the first, seem to be doing very little useful and showing no tendency to stabilise in the reconvolver.

Examples using isolated speech sounds like this can be found to demonstrate almost any behaviour whatsoever. A proper comparison of the merits of the several BAP configurations requires tests on long representative samples of speech and adequate variation of their different parameters. This we have not yet done. As the performance is a compromise between capsizal and prediction gain, we propose to use an arrangement similar to that used to produce fig. 2 with random "analogue" errors inserted into a system with uncoded residual. A plot of some measure of mean segmental prediction gain against mean segmental

output signal-to-noise ratio (SNR) should then look something like fig. 6. The prediction gain should increase with little decrease in SNR until capsizal occurs, after which SNR should decrease dramatically with little further increase in prediction gain being attainable.

CONCLUSIONS

In the switched telephone network, backward adaptive predictive (BAP) coders have to be use for 32 k bit/s coding because they can introduce negligible delay. Unfortunately, they suffer from the little understood phenomenon of capsizal in the presence of transmission errors. Several different configurations of BAP coders have been proposed, but these only address the problem of filter instability associated with capsizal. The partial deconvolution lattice structure proposed in this paper appears to be capable of giving higher prediction gain with less danger of capsizal than the other configurations, and hence better overall coder performance. A complete evaluation of the system has yet to be carried out.

REFERENCES

(1) Moye, L.S., "Self-Adaptive Filter Predictive Coding System", Proc. Int. Zurich Seminar on Int. Sys. for Sp., Video and Data Comm., Zurich, Mar. 1972, paper F3.

(2) Moye, L.S., "Digital Transmission of Speech at Low Bit Rates", Elect. Comm., 47,4, 1972, pp. 212-223.

(3) Nishitani, T., Aikoh, S., Maruta, R., Areseki, T., and Ozawa, K., "A 32 k bit/s Toll Quality ADPCM Codec Using a Single Chip Processor", Proc. Int. Conf. on Acoust., Sp. and Sig. Proc., IEEE, ICASSP '82, Paris, May 3-5, 1982, pp. 960-963.

(4) Coinot, D., "A 32 k bit/s ADPCM Coder Robust to Channel Errors", ibid., pp. 964-967.

(5) Raulin, J.M. and Jeandot, J.L., "A 32 k bit/s PCM to ADPCM Converter", ibid., pp.968-971

(6) Mermelstein, P. and Millar, D., "Adaptive Predictive Coding of Speech and Voice-band Data Signals", ibid., pp. 972-975.

(7) Yatsuzuka, Y. and Suyderhoud, H.G., "A 32 k bit/s ADPCM Encoding with Variable Initially Large Leakage and Adaptive Dual Loop Predictors", ibid., pp. 976-979.

(8) Le Guyader, A. and Gilloire, A., "Comparison of Basic and Simplified Sequential Algorithms for the Computation of the Lattice Filter Predictor Coefficients in ADPCM Coding of Speech", ibid., pp. 1676-1679.

(9) Daumer, W.R. and Sullivan, J.L., "Subjective Quality of Several 9.6 - 32 k bit/s Speech Coders", ibid., pp. 1709-1712.

(10) Raulin, J.M., Bounerot, G., Jeandot, J.L. and Lacroix, R., "A 60 Channel PCM-ADPCM Converter", IEEE Trans. on Comm., Comm-30,4, April 1982, pp. 567-573.

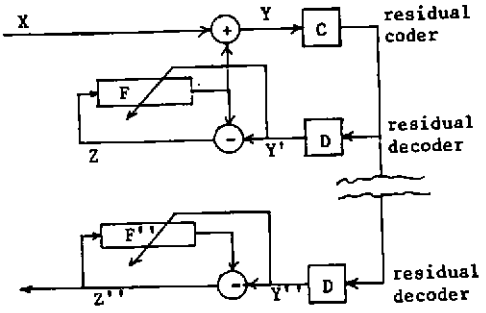
(11) CCITT, "32 k bit/s ADPCM-DLQ Coding", CCITT Study Group XVIII, Cont. No. EF, April 1982.

(12) Markel, J.D., and Gray, A.H.Jr.,, "Linear Prediction of Speech", Springer Verlag, 1976.

(13) Itakura, F. and Saito, S., "Digital Filtering Techniques for Speech Analysis and Synthesis", 7th. Int. Cong. on Acoust., Budapest, 1971.

(14) Makhoul, J.I., and Cosell, L.K., "Adaptive Lattice Analysis of Speech", IEEE Trans. on Circuits and Sys., CAS-28, 6, June 1981, pp.494-498.

Speech Coding at 32 k bit/s



In Z-transform terms

$Y = X - FZ$

$Y' = Y + Q$ (Q is quantising noise)

$Z = Y' + FZ$

$= Y + Q + FZ$

$= X - FZ + Q + FZ$

$= X + Q$

in the absence of errors,

$Y'' = Y'$

if, and only if

$F'' = F$

$Z'' = X - FZ + Q + F''Z''$

$= X + Q$

Fig. 1 Backward Adaptive Predictive Coder

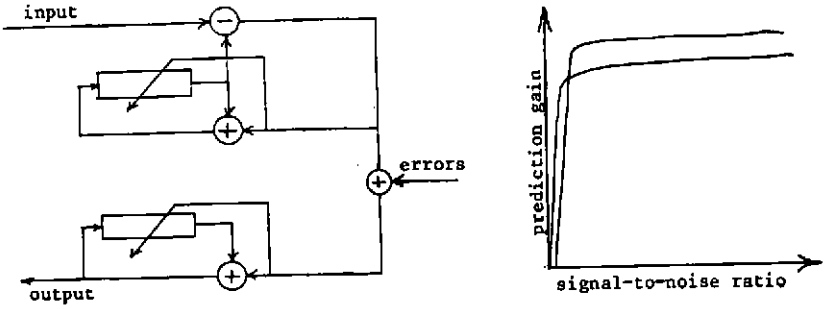
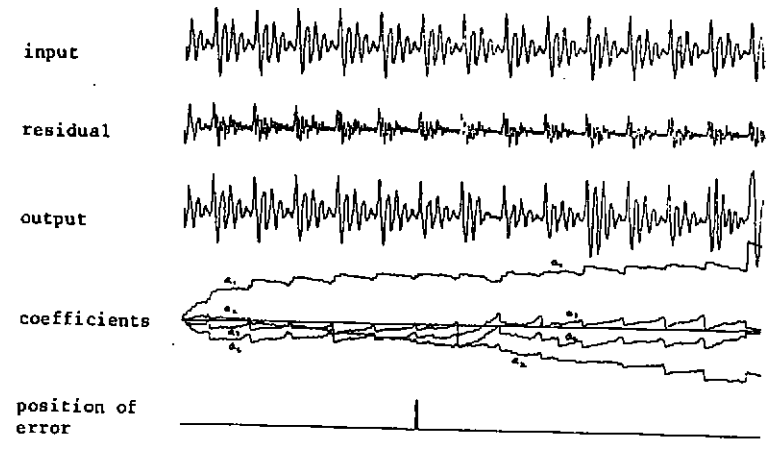
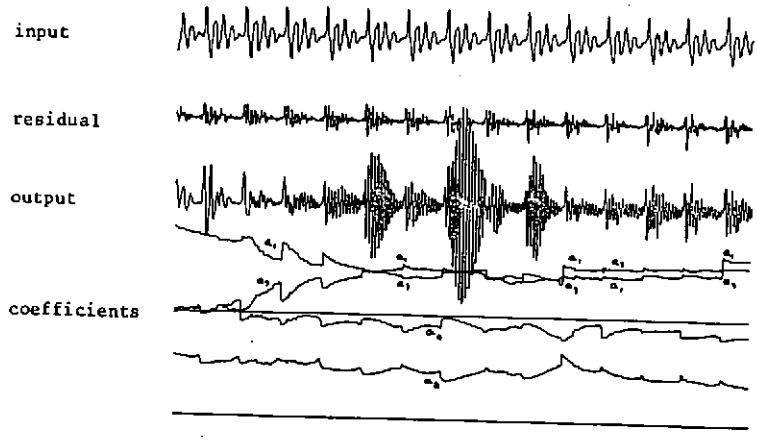


Fig. 6 Method for comparison of BAP coder filter structures.



(a) first part



(b) continuation

Fig. 2. Showing how the occurrence of a single error upsets the unstable equilibrium of a BAP coder.

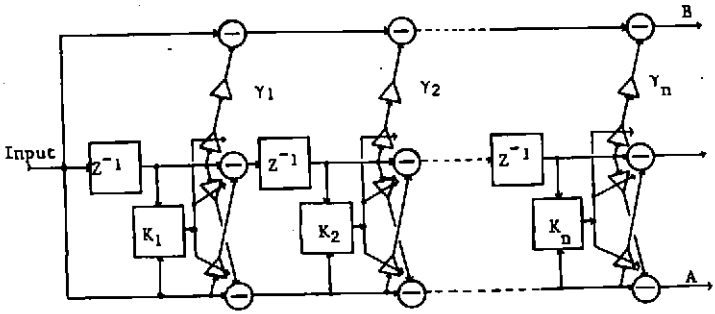


Fig. 3. Partial decorrelation lattice filter (Y_i <1)

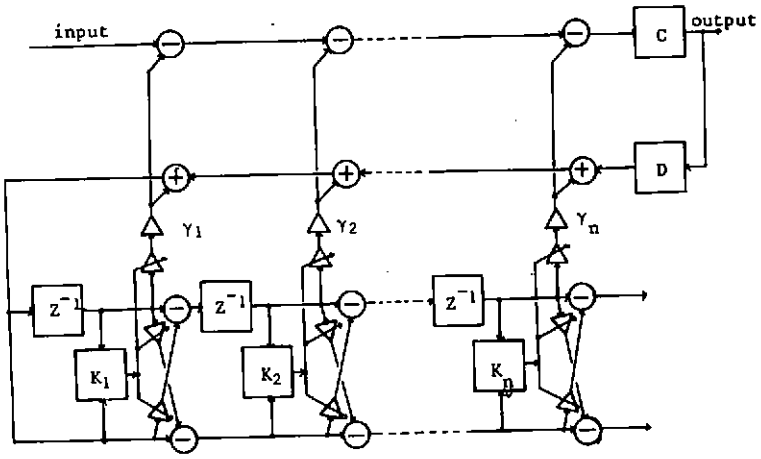


Fig. 4. BAP Coder using partial decorrelation lattice filter.

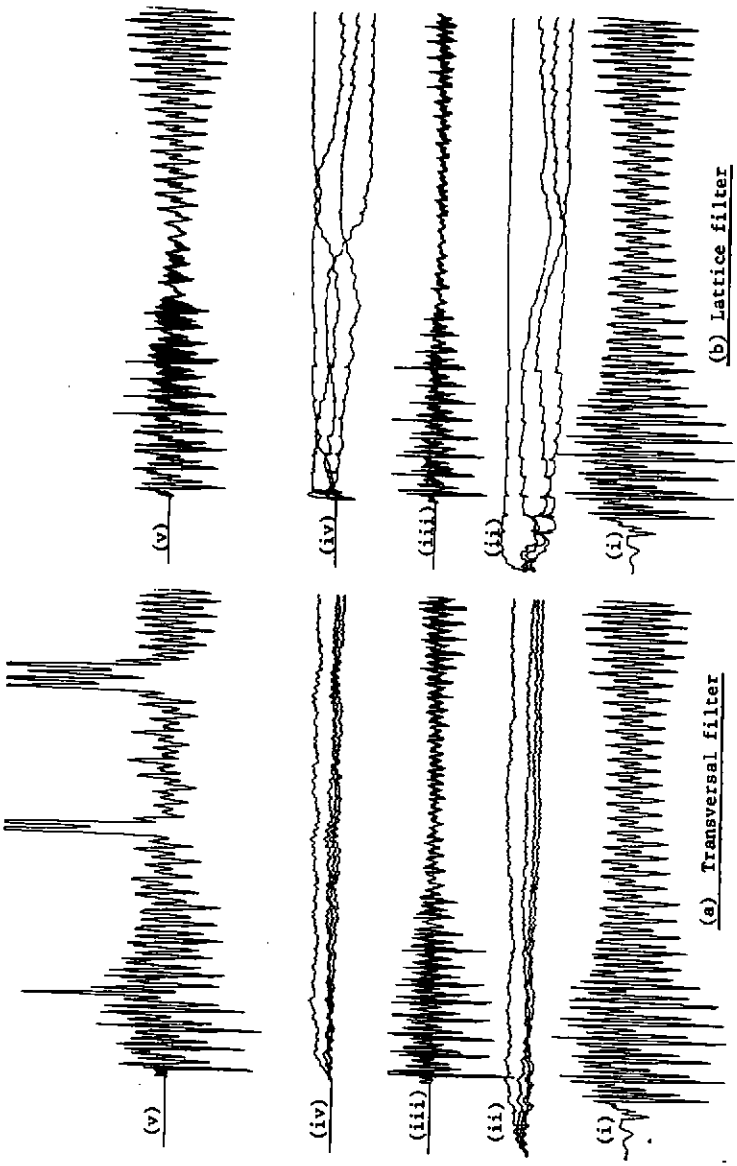


Fig. 5. Comparison of BAP system responses with no residual coding, but receive end started late to ensure lack of synchronisation. (i) input, (ii) send-end coefficients, (iii) residual, (iv) receive-end coefficients, (v) output.