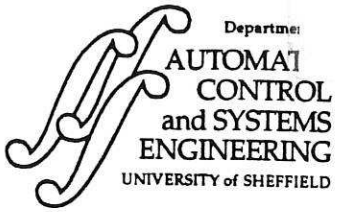

VARIABLE FRAME-RATE SPEECH CODING BY ADAPTIVE-FLUX INTERPOLATION

L Baghai-Ravary [*], S W Beet [**] and M O Tokhi [*]

[*] Department of Automatic Control and Systems Engineering,
[**] Department of Electronic and Electrical Engineering,
The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK.

Tel: + 44 (0)114 2825136.
Fax: + 44 (0)114 2731 729.
E-mail: O.Tokhi@sheffield.ac.uk.

Research Report No. 592

July 1995

Abstract

Variable frame-rate (VFR) speech coders have many desirable properties, but make implicit assumptions concerning the nature of the spectral evolution of speech (Peeling and Ponting, 1989). To date, these assumptions have been crude and unable to model speech parameters during extended periods of coarticulation. In particular they have been unable to cope with steadily changing formants. Thus existing VFR methods must transmit many more frames than are really necessary.

This paper presents a new technique; Adaptive-Flux Interpolation (AFI), which significantly extends the period over which accurate estimation can be performed, and is much more robust and accurate than other methods.

Key words: adaptive-flux interpolation, speech coding, variable frame-rate coding.

CONTENTS

Title		i
Abstract		ii
Contents		iii
List of figures		iv
1	Variable frame-rate coding	1
	1.1 Zero-order prediction	1
	1.2 Flow-based prediction	1
2	Adaptive-flux interpolation	1
3	Experiments	2
4	Conclusions	3
5	References	4

LIST OF FIGURES

Figure 1: Adaptive-flux Interpolation.

Figure 2: Variable frame-rate results.

Figure 3: Mean coding errors for single-frame estimation.

Figure 4: Original data and conventional, FBP and AFI error plots for the speech segment "in greasy".

1 Variable frame-rate coding

In variable frame-rate coding schemes, an underlying model of the data dynamics is assumed. This is used in the encoder (to assess when the reconstruction error will surpass a threshold) and in the decoder (to reconstruct the omitted frames of data). There are many possible models, most of which can be described by

$$\hat{\mathbf{o}}_n = \mathbf{C}(n)\mathbf{o}_{n-1} + \mathbf{v}_n \qquad (1)$$

where $\mathbf{o}_n$ is the observation vector at time n, $\mathbf{C}(n)$ is a (possibly time-dependent) coefficient matrix, $\mathbf{v}_n$ is the n^{th} vector of an innovation sequence, and $\hat{\mathbf{o}}_n$ is the estimated observation vector.

1.1 Zero-order prediction

In this conventional model, an unknown frame is assumed to be equal to its immediate predecessor. This process can be represented mathematically by setting $\mathbf{C}(n) = \mathbf{0}$ and $\mathbf{v}_n = \mathbf{o}_{n-1}$ in the above equation, yielding a zero-order prediction.

1.2 Flow-based prediction

This method allows for migration of features between appropriate elements within the data vectors (Beet, et.al, 1994). Two consecutive vectors are used to predict the direction of migration (the flow) and the change in value of the respective elements. From this, $\mathbf{C}(n)$ and $\mathbf{v}_n$ are estimated.

2 Adaptive-flux interpolation

Recently, more realistic flow-based models of speech evolution have been developed for accurate prediction of spectrogram-like vectors one or more steps ahead (Flow-Based Prediction, or FBP, (Baghai-Ravary et.al, 1994a,b; Beet, et.al, 1994). Adaptive-Flux Interpolation extends this concept, making the information content of each new

observation vector more nearly constant. This method seeks to make the evolution of every frame from, say, N_1 to N_2, explicit, given only the data in the two original frames.

The evolution of the data is represented by lines of flux. The likelihood of each pair of elements being linked, is assumed to be given by a zero-mean Gaussian distribution of the change in data value from one end of the line of flux to the other. As with magnetic flux, it is assumed that these links never cross, and the most likely set of links satisfying this constraint can be found by a simple form of dynamic programming.

Consider the link from element i of frame N-1 to element j of frame N, as shown in Figure 1. To find the links between frame N-1 and frame N, we require a local distance matrix, Γ, containing the likelihood of every potential link. The elements of Γ can be expressed in a normalised log-likelihood form:

$$\gamma_{i,j} = \left| R_{N_1,u} - R_{N_2,v} \right|^2 \qquad (2)$$

where $R_{N_1,u}$ is the value in element u of the known frame N_1 at the point of incidence of the link, linearly extrapolated to reach that frame. $R_{N_2,v}$ corresponds to the other (extrapolated) end of the link, where it intercepts frame N_2.

Dynamic programming is then used to find the set of links which give the minimum total log likelihood over all selected links. Various constraints can be imposed during the dynamic programming. In particular, links which would be extrapolated to pass beyond the ends of any frames, are disallowed, as are links which correspond to unrealistically rapid frequency transitions.

3 Experiments

AFI has been incorporated into a VFR coding scheme, and a spectrogram of the sentence "She had your dark suit in greasy wash water all year" encoded. The maximum allowable normalised RMS coding error was specified over a range from 5 to 80%.

For each block of data from frames N_1 to N_2, AFI was used to estimate frames N_1+1 and N_2-1. N_1 was then incremented and N_2 decremented, and the process repeated until

all frames had been estimated. No constraints were placed on the links of successive recursions to enforce any form of continuity on the lines of flux, but this does not appear to have any serious consequences. Figure 2 shows the percentage of the original frames needed to maintain all the errors in the interpolated frames below the specified level. Clearly, AFI is superior to both conventional VFR coding and FBP.

These methods have been applied to many other sentences and many other speakers. The results have always shown the same pattern, with close quantitative agreement. Only one instantiation is shown here for the sake of simplicity.

4 Conclusions

Adaptive-Flux Interpolation provides a powerful method for modelling non-stationary vector sequences, such as those produced by many frame-based speech analysis techniques. It achieves a high degree of accuracy and only uses local knowledge of the vector sequence characteristics.

Note that FBP is actually slightly worse than the conventional VFR in the example in Figure 2. This poor performance is due to the lack of robustness of FBP, when presented with abrupt spectral changes. To illustrate this, Figure 3 shows the average coding errors for the different algorithms, when performing single frame estimation on three segments of speech: the complete sentence mentioned previously, the segment "in greasy" taken from it, and the vowel segment taken from the word "greasy". The vowel segment is well estimated by FBP, but as the unpredictability of the data increases in the longer segments, its performance drops off rapidly. For the complete sentence, FBP is slightly worse than the conventional method, as expected. Figure 4 shows the reasons for this behaviour in more detail, showing the error magnitudes for "in greasy" as a function of time and frequency. In all cases, AFI is much better than either of the other two methods.

5 References

BAGHAI-RAVARY, L., BEET, S. W. and TOKHI, M. O. (1994a). Flow-based prediction: A method for improved speech recognition, *IEE Digest No 1994/138: Colloquium on Techniques for Speech Processing and their Application*, London, 01 June, 1994, pp. 5/1-5/5.

BAGHAI-RAVARY, L., BEET, S. W. and TOKHI, M. O. (1994b). Removing Redundancy from some common representations of speech, *Proceedings of the Institute of Acoustics*, **16**, (Part 5), pp. 467-474.

BEET, S. W., BAGHAI-RAVARY, L. and TOKHI, M. O. (1994). Non-stationary prediction of speech data. *Proceedings of EUSIPCO-94: Seventh European Signal Processing Conference*, Edinburgh, 13-16 September 1994, **III**, pp. 1653-1656.

PEELING, S. M and PONTING, K. M. (1989). Experiments in variable frame-rate analysis for speech recognition, *RSRE Memorandum 4330*, 1989.

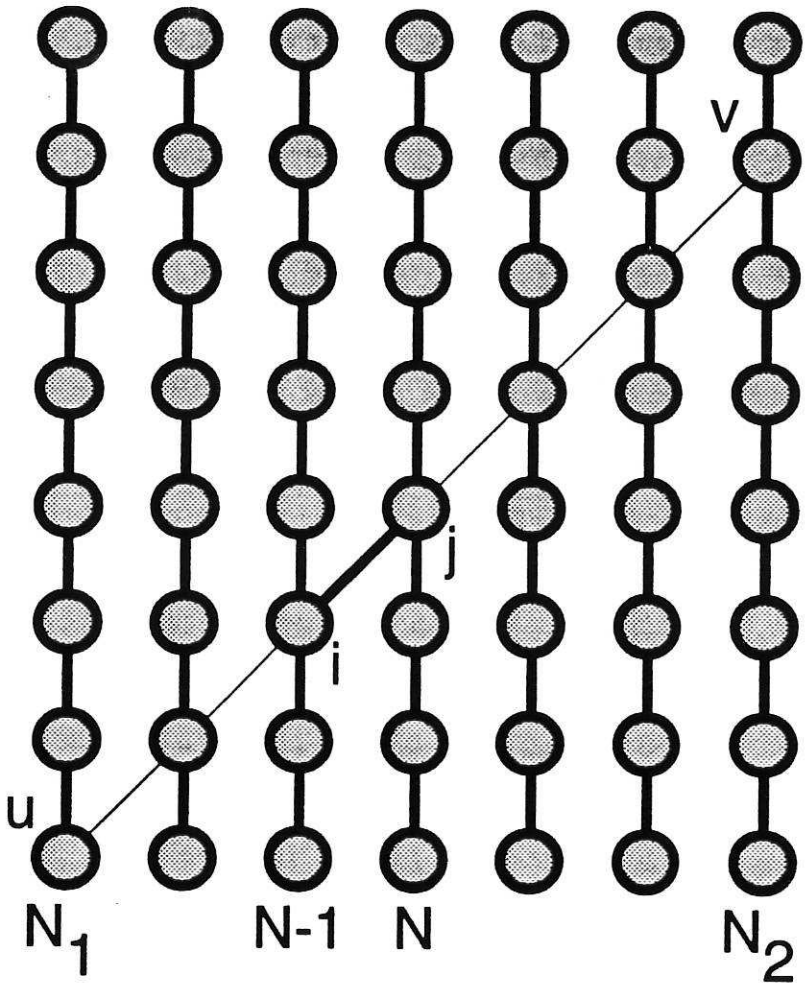

Figure 1: Adaptive-flux interpolation.

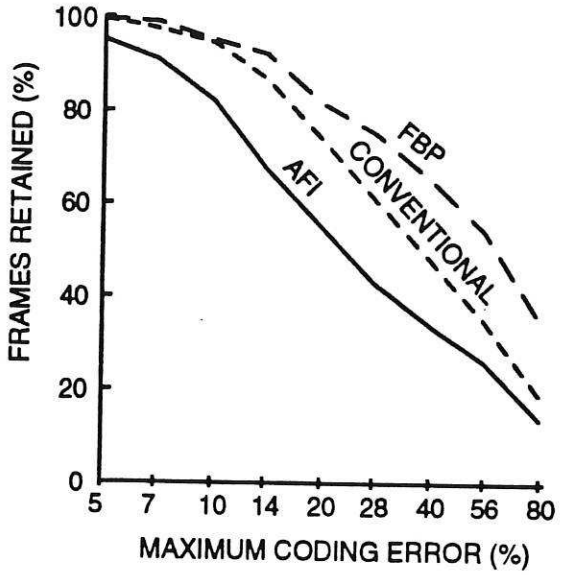

Figure 2: Variable-frame rate results.

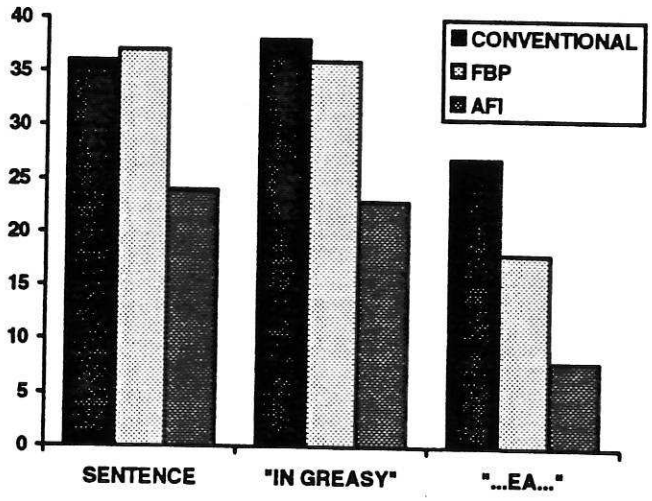

Figure 3: Mean coding errors for single-frame estimation.

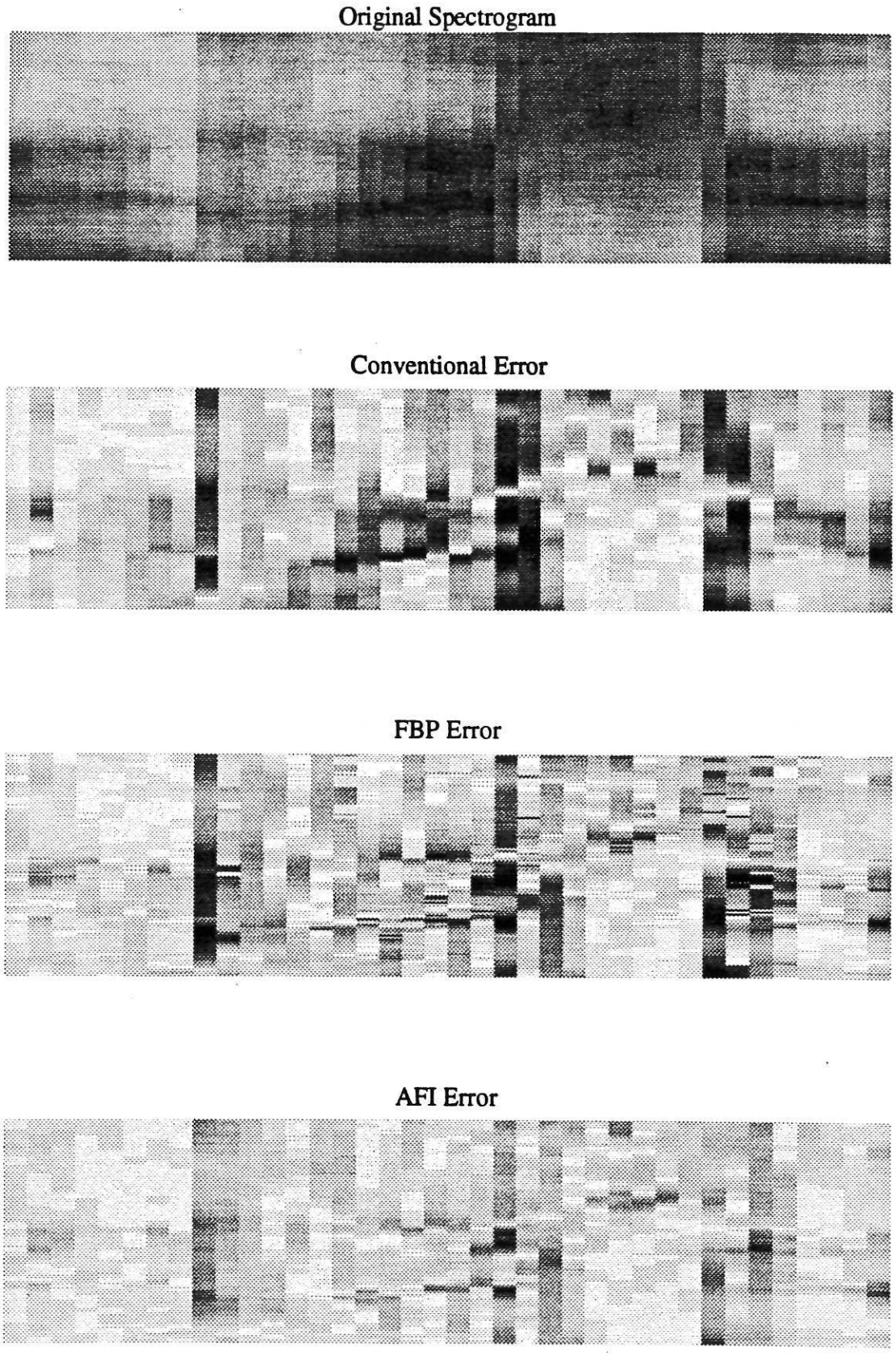

Figure 4: Original data and conventional, FBP and AFI error plots for the speech segment "in greasy".