ADAPTIVE FLUX INTERPOLATION, FLOW-BASED PREDICTION, DELTA OR DELTA-DELTA COEFFICIENTS: WHICH IS BEST?

Ladan Baghai-Ravary‡, Steve Beet† and Osman Tokhi‡

‡ Department of Automatic Control and Systems Engineering,
† Department of Electronic and Electrical Engineering,
Sheffield University, Mappin Street, Sheffield, S1 3JD, UK.

ABSTRACT

Four models of speech dynamics are compared for robustness and accuracy: Adaptive Flux Interpolation, Flow-Based Prediction, Delta and Delta-Delta Coefficients. These models are applied to a number of speech representations. The interpolative approach is shown to be superior to prediction in all cases.

1. INTRODUCTION

Speech is a nonstationary signal that is glibly treated as piecewise stationary in most systems. Although there are methods that aim to model speech dynamics, the most powerful of these are either computationally intense or require prohibitive quantities of training data [1, 2]. Conventional methods which avoid these problems include zero-order and first-order predictors, which yield the oft-used delta and delta-delta coefficients of automatic speech recognition systems.

A realistic model of speech evolution has recently been developed (Flow-Based Prediction, or FBP [3]), allowing accurate prediction of spectrogram data. This paper describes a method for *interpolation* between pairs of data vectors. Although both FBP and the Adaptive Flux Interpolation (AFI) technique, are based on the same assumptions regarding signal evolution, the addition of end-point constraints makes the interpolation approach more robust.

To conclude, AFI, FBP, delta and delta-delta coefficients are compared and the superiority of AFI is demonstrated.

2. BACKGROUND

Speech sounds are made up of periods of gradual spectral change (mostly voiced speech and fricatives), interspersed with other periods characterised by abrupt onsets or offsets (plosives, stops, etc.). Spectrograms and similar frame-based representations of speech, therefore have a piecewise-continuous form, but are rarely piecewise-constant.

There are several tractable methods for characterising the short-term dynamic behaviour of spectrograms and other related speech representations. The most important of these are listed below:

2.1. Zero-Order Predictor:

A zero-order predictor simply assumes that one data vector can be estimated to be the same as its predecessor. The error in the prediction is then the difference between successive vectors (i.e. the delta coefficients). This method relies on temporal over-sampling to ensure that changes between one vector and the next are not too abrupt.

2.2. First-Order Predictor:

A first-order predictor, which assumes the *change* between successive data vectors is constant, yields an error equal to the delta-delta coefficients. This method also requires over-sampling to operate correctly.

2.3. Flow-Based Predictor:

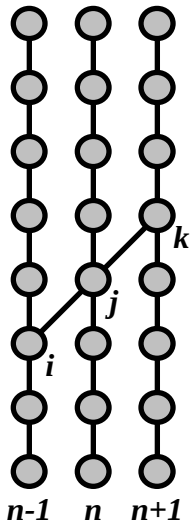
This method, described in [3], allows for migration of features between appropriate elements within the data vectors. Two consecutive vectors are used to predict the direction of migration (the flow) and the change in value of the respective elements.

3. ADAPTIVE FLUX INTERPOLATION

The fundamental aim of the Adaptive Flux Interpolation method described here, is to estimate data vector, n, from a sequence, given only the two adjacent vectors, *n*-1 and *n*+1. It is assumed that the data vectors contain related elements, which have similar values, and lie on non-crossing lines of flux. The direction of these lines of flux change with time. The AFI algorithm adapts to track these changes, and interpolates along the lines of flux.

The likelihood of each pair of elements lying on the same line of flux, can be assumed to be given by a zero-mean Gaussian distribution of the change in data value from one end of the line to the other. The most likely set of non-crossing lines can be found by a simple form of dynamic programming.

Figure 1: Link between consecutive frames



Consider the link from element *i* of vector *n*-1 to element *j* of vector *n*. To find the link, 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:

$$\gamma_{i,j} = \left| R_{n+1,u} - R_{n-1,v} \right|^2$$

where $R_{n+1,u}$ is the value in element *u* of vector *n*+1, at the point of incidence of the link. $R_{n-1,v}$ corresponds to the other (extrapolated) end of the link, where it intercepts vector *n*-1.

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

4. EXPERIMENTS

Figures 2 and 3 show the average normalised errors for typical speech representations of male and female speakers, when performing single-frame prediction and interpolation. The legends for these figures are explained in Table 1.

Most of the representations were produced by standard processing algorithms: the power spectrum was calculated by the Blackman-Tukey method [4], the ERB-scale cepstrum is a frequency-warped version of the standard cepstrum, based on the "Equivalent Rectangular Bandwidth" auditory frequency scale [5], the LP (Linear Prediction) spectrum is calculated by the Maximum Entropy Method (MEM) [4], and both the latter and the log vocal tract area functions were calculated from the reflection coefficients yielded by Burg's method [4].

5. CONCLUSIONS

The zero-order predictor, which yields the delta coefficients, is not particularly accurate but because it is robust, it rarely makes severe errors and the overall error score is not bad. By comparison, the first-order predictor (yielding

delta-delta coefficients) and FBP are more accurate during smoothly evolving regions, but are less robust. They overshoot following abrupt onsets or offsets, and their average error scores are therefore worse. Of these two, FBP is slightly better on average due to its ability to track migrating features, giving it a lower error score on most data types.

The performances of all these methods are virtually identical for different genders. This demonstrates that the normalisation applied during error score calculation has correctly compensated for differences in overall speech characteristics.

For all these speech representations, AFI gave a significantly more accurate estimate than any of the other methods. It is better at modelling speech dynamics and is the most robust of the methods presented here. Hence it has the lowest error score.

6. REFERENCES

[1] R. Schwartz, Y. L. Chow, et al.; "Context-Dependent Modelling for Acoustic-Phonetic Recognition of Continuous Speech"; in Proc. IEEE ICASSP-85, pp. 176-179; 1985.

[2] Y. Boram; "Construction of Linear Predictors for Stationary Vector Sequences"; IEEE Trans. on Automatic Control, vol. 35, no. 2, pp. 236-239; 1990.

[3] S. W. Beet, L. Baghai-Ravary and M. O. Tokhi; "Non-Stationary Prediction of Speech Data"; Signal Processing VII: Theories and Applications, pp. 1653-1656; 1994.

[4] J. G. Proakis and D. G. Manolakis; Digital Signal Processing: principles, algorithms and applications, second edition; Macmillan Publishing Company, 1992.

[5] B. C. J. Moore and B. R. Glasberg; "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns"; Hearing Research, vol. 28, pp. 209-225; 1987.

Figure 2: Comparative error scores for male speaker

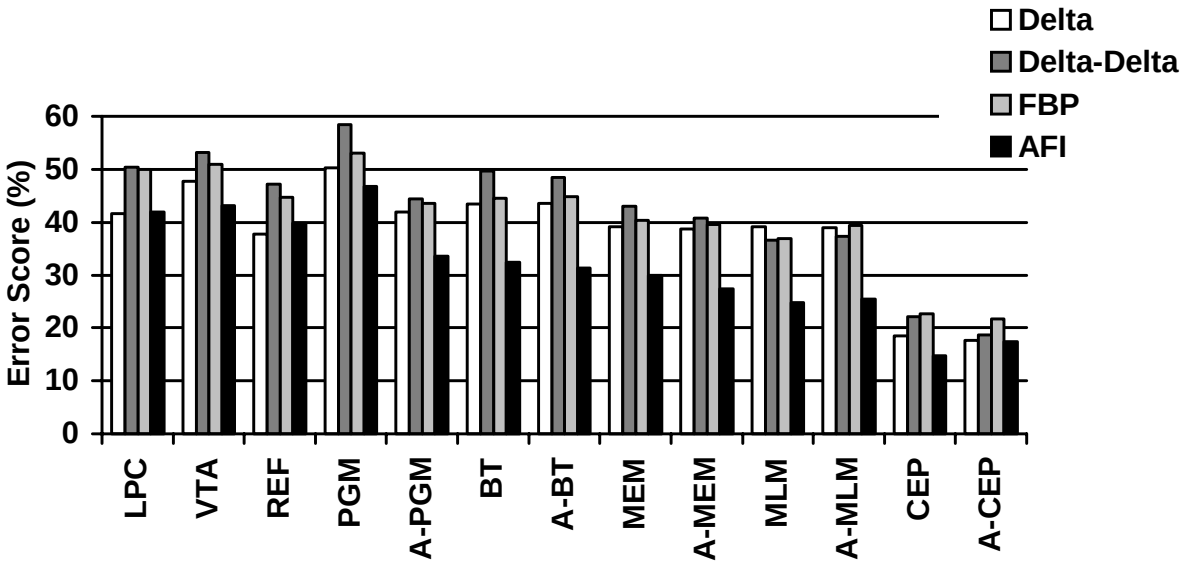


Figure 3: Comparative error scores for female speaker

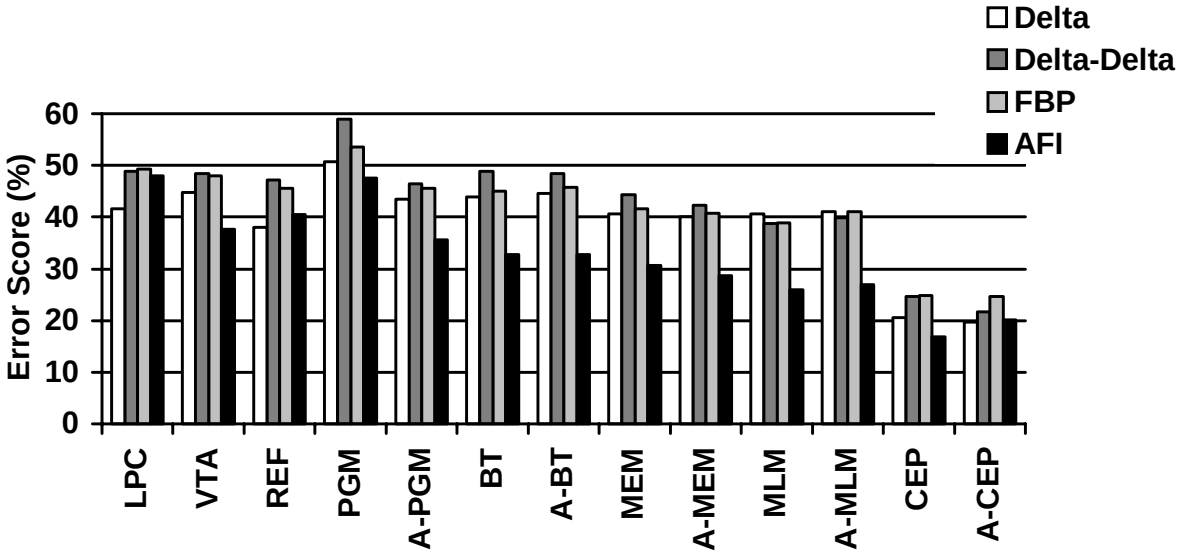


Table 1: Key to legends in figures 2 and 3

LPC	LP coefficients	MEM	LP (MEM) spectrum
VTA	Vocal tract areas	A-MEM	ERB-scale MEM
REF	Reflection coefficients	MLM	Maximum Likelihood PSD
PGM	Periodogram	A-MLM	ERB-scale MLM
A-PGM	ERB-scale PGM	CEP	Cepstrum
BT	Blackman-Tukey PSD	A-CEP	ERB-scale CEP
A-BT	ERB-scale BT		