Non-Stationary Prediction of Frame-Based Speech Data

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

†*Department of Electronic and Electrical Engineering, Sheffield University,*
P. O. Box 600, Mappin Street, Sheffield, S1 4DU, UK. E-Mail: s.beet@sheffield.ac.uk
‡*Department of Automatic Control and Systems Engineering,*
P. O. Box 600, Mappin Street, Sheffield, S1 4DU, UK.

Abstract. Most speech coding, synthesis and recognition algorithms are based on models which treat pre-processed speech data as a sequence of pseudo-stationary segments. This piecewise-constant assumption is significantly less accurate than one which allows for gradual changes in, for example, formant frequencies or pitch. This paper introduces a new technique for prediction and interpolation of speech representations, which merely assumes that the data is piecewise-continuous (i.e. continually evolving, but with occasional abrupt changes). This new approach ("Flow-Based Prediction", or FBP) imposes constraints on the form of a standard vector linear predictor which allow it to track rapidly changing data. It also provides a method for updating the predictor's coefficients from one frame to the next, so the predictor can be allowed to free-run for extended periods.

1. Background

Speech sounds are made up of periods of gradual spectral change (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 (as is assumed, for example, in hidden Markov models).

In general, it is possible for the human observer to predict the continuous segments of spectrograms with a high degree of accuracy, and to do so using only local information. Thus, these segments must be highly redundant. It is only when there is a sudden change in the evolutionary behaviour of the signal, that significant information is conveyed. Identifying the portions of speech representations which contain the most information, however, requires a realistic model of speech dynamics.

1.1 Previous Approaches

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

1.1.1 Delta Coefficients

Expanding the observation vector to include linear regression parameters [1] or "delta coefficients" [2], has been used to allow for the inter-dependency between successive frames of speech data. Such methods do offer noticeable performance improvements, but do not bear any particularly meaningful relationship to the true nature of speech dynamics. Consequently they have to rely on temporal over-sampling to ensure that changes between one frame and the next are not too abrupt.

1.1.2 Variable Frame-Rate Coding

Similarly, variable frame-rate coding schemes [3] are limited in usefulness because of their inadequate assumptions regarding the implied interpolation between transmitted frames (a high coding error can result from quite clearly predictable segments of speech). Their capabilities are further limited because they reduce redundancy by omitting complete observation vectors: they make no allowance for the different levels of redundancy in different elements within the vector.

1.1.3 Vector Linear Prediction

These more recent (and sophisticated) attempts to allow for inter-frame redundancy have achieved only limited success [4]. They have been based on linear, time-invariant (LTI) models of the signal's evolution over time, in general via some form of state-dependent vector linear prediction. Their biggest problem results from their assumption of time-invariance.

2. Flow-Based Prediction

The method proposed here employs a special form of vector linear predictor, where the coefficient matrix is sparse and is automatically recalculated at successive points in time. This allows the prediction to remain accurate for longer.

The prediction is based solely on the two most recent vectors in the observation sequence, so it is kept "up-to-date" over a very short time scale. This is a significant improvement over other linear prediction methods, which require accurate estimation of the data covariance matrices

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 1(a). An example prediction matrix.

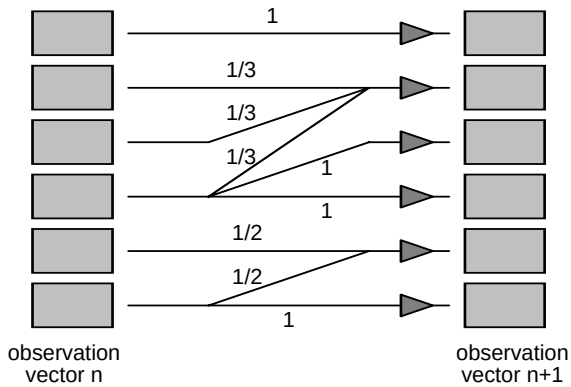

Figure 1(b). Mapping between observation vectors implied by the example prediction matrix.

[5]. To calculate the covariance matrices for these linear prediction methods, many example data sequences are required, and such data is not normally available for highly non-stationary data such as that derived from speech. Indeed, this form of calculation could only be used in [4] because each prediction matrix was tied to a particular state of the HMM[1].

2.1 The Prediction Model

Conventional vector linear prediction uses a model of the form:

$$\mathbf{o}_{n+1} = \mathbf{C}\mathbf{o}_n + \mathbf{v}_n$$

In this equation, $\mathbf{o}_n$ is the current observation vector, $\mathbf{v}_n$ is the innovation sequence (i.e. the unpredictable, information-bearing part of the data) and $\mathbf{C}$ is the prediction matrix. $\mathbf{o}_{n+1}$ is the next observation vector. The prediction matrix is usually assumed constant for long enough to allow reasonably accurate estimation of all its elements, but with most frame-based representations of speech [6] the data is rarely stationary for more than a few successive observations, if at all. The only way a prediction matrix can be constructed quickly enough to cope with such data, is by imposing some form of pre-defined structure on it.

2.2 Assumptions

The FBP algorithm assumes that the changes from one vector to the next can be modelled as an averaging and shift within the vector, together with a constant increment. This differs from the normal LTI model in that the signal statistics track the features in the data, rather than being tied to particular elements within the vectors.

To avoid having to estimate a full prediction matrix, each element of the nth observation vector, $\mathbf{o}_n$, is assumed to have a similar value to one or more consecutive elements of the following observation vector, $\mathbf{o}_{n+1}$. Each element in $\mathbf{o}_{n+1}$ is therefore approximated as the average of a small number (one or more) of consecutive elements in $\mathbf{o}_n$. This implies that the prediction matrix is sparse, with its rows summing to one, like the example shown in Figure 1(a). This example corresponds to a mapping between the elements in the observation vectors as shown in Figure 1(b).

It is further assumed that no "mapping links" cross, so that any spectral features appear in the same order in both vectors. Subject to these constraints, dynamic programming can be used to estimate $\mathbf{C}$, and hence $\mathbf{v}_n$, given $\mathbf{o}_n$ and $\mathbf{o}_{n+1}$. However, to predict $\mathbf{o}_{n+1}$ from *previous* values, we need to be able to estimate $\mathbf{C}$ and $\mathbf{v}_n$, given only $\mathbf{o}_n$ and $\mathbf{o}_{n-1}$. This requires a model of the evolution of $\mathbf{C}$ and $\mathbf{v}_n$ from one frame to the next. That is to say that $\mathbf{C}$ is assumed time-dependent, and will henceforth be referred to as $\mathbf{C}_n$, while the implied matrix for the previous transition will be $\mathbf{C}_{n-1}$. It is only constant where the data is totally stationary.

Consider a link between element j of vector $\mathbf{o}_n$ and element k of $\mathbf{o}_{n+1}$, and assume that the vectors have been evolving steadily. In that case, there would be expected to be a link between $o_{n-1,i}$ and $o_{n,j}$, where i = j - (k - j), as shown in Figure 2. Under the assumption of "steady evolution", it is also anticipated that the value of element k of $\mathbf{v}_{n+1}$ is equal to the value of element j of $\mathbf{v}_n$, i.e. $v_{n+1,k} = v_{n,j} = o_{n,j} - o_{n-1,i}$.

The inter-element links in Figure 1 are closely related to those produced by calculating the acoustic flow [7], and

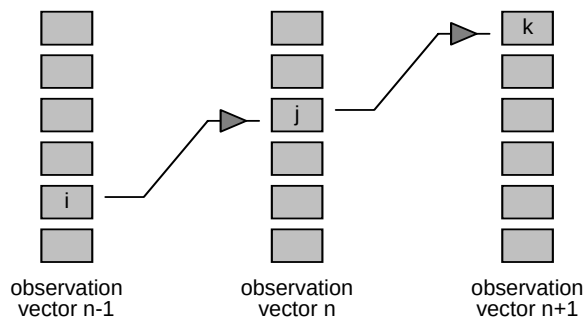

Figure 2. Relationship between links in successive observation vectors.

[1] The statistics from multiple occurrences of that state are then pooled to produce a composite estimate of the prediction matrix. Because of the variability between different occurrences of each state, however, the prediction is rarely precise enough to give significant improvement over delta coefficient methods.

these can be used as an approximation to the optimal set of links (i.e. that which would minimise $|\mathbf{v}_n|^2$).

2.3 Acoustic Flow

Acoustic flow uses dynamic programming to calculate the mapping of the elements between each pair of observation vectors, based on a local distance matrix. This contains the squared Euclidean distance between every possible "linked pair" of elements, $d_{i,j} = (o_{n,j} - o_{n-1,i})^2$. However, to allow for the assumed steady evolution between successive vectors, those values are replaced by $d_{j,k} = v_{n,j}^2 = (o_{n,j} - o_{n-1,2j-k})^2$, simultaneously yielding the links required to calculate the FBP parameters, $\mathbf{C}_n$ and $\mathbf{v}_n$.

3. Results

The Maximum Likelihood Method (MLM) [8] was used to produce a log power spectrum for successive 25ms frames of the words "...in greasy...", taken from the TIMIT database file `/DR5/MEWMO/SA1.WAV`. This speech segment has also been used previously as one of the standard data files in [6]. The analysis used was of 32nd order. There was no overlap between successive frames, so as to demonstrate the reduced need for temporal oversampling with Flow-Based Prediction.

During the central portion of this utterance, the formant frequency changes quite rapidly, as can be seen from the combined spectrogram and acoustic flow diagram in Figure 3. The spectrographic acoustic flow of the predicted version of this data is shown in Figure 4. Here the predictor was allowed to free-run until the error reached a threshold. At this point the prediction was assumed to be inaccurate and the predictor then used two consecutive vectors from the "real" data to restart the prediction. The threshold was chosen to give a 50% reduction in frame rate.

The similarity between the true data and the prediction can be seen more clearly in the spectrograms in Figure 5. The upper plot shows the original utterance and the lower, the predicted spectrogram. This shows smoothly-changing formant tracks, with occasional abrupt changes where the prediction error has become too large. Generally, these occur where the formants change direction or are absent (during frication, for example).

When the predictor "free-runs", the apparent formant tracks look quite natural, and can only be detected as artificial when they have been unable to follow unpredictable changes in the data, causing the predictor to be reset as a result of an increased error.

Careful examination of the speech segment in Figure 3, shows that the formant frequencies during the central portion of the data can be predicted easily over a short time-scale. The magnitude of the one-step-ahead FBP error is plotted in Figure 6. The lighter colours in the graph indicate high levels of redundancy, and the darker colours, low levels of redundancy. This plot shows that the FBP error is much smaller than the magnitude of the delta coefficients, especially within visibly predictable regions.

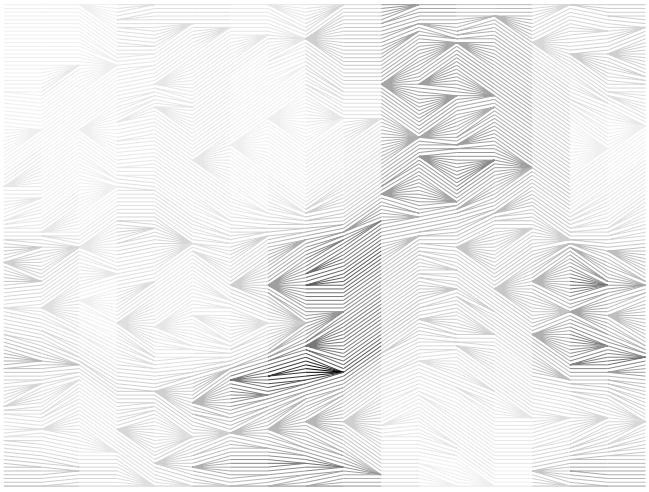

Figure 3. Spectrographic acoustic flow of natural "...in greasy..." utterance.

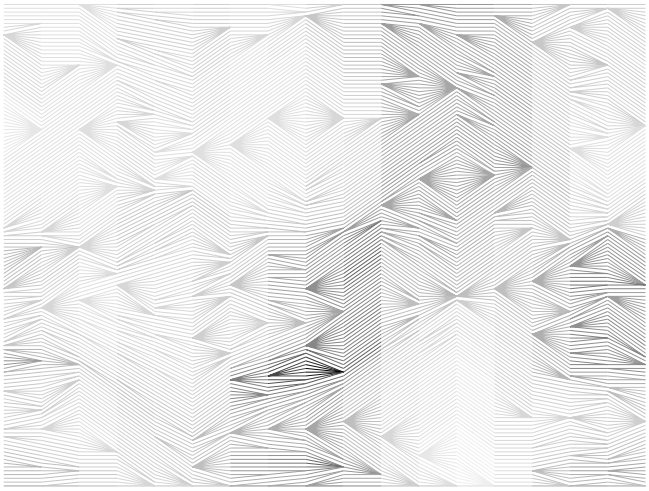

Figure 4. Spectrographic acoustic flow of FBP-encoded "...in greasy...", with 50% frame rate.

The dark regions in the FBP error are much more localised than those in the delta coefficient plot. In fact, the FBP error is only significant during the onset of frication and over a very narrow range of frequencies in the previous frame, where the second formant is rapidly fading out. Thus FBP has been more successful at detecting redundancy in this representation of speech.

4. Conclusions

Flow-Based Prediction provides a method for predicting 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.

FBP could be used not only for speech recognition, but also for speech coding, and to predict many other (non-speech) forms of vector sequence. However, it was origi-

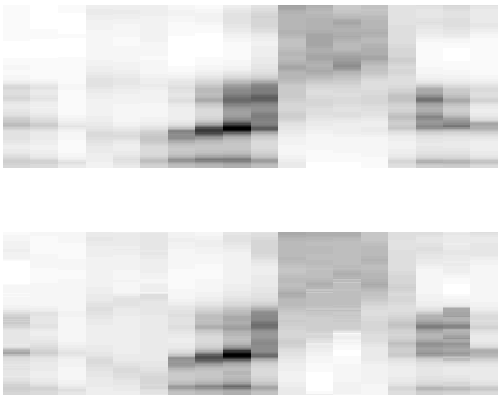

Figure 5. Original spectrogram (upper) and 50% frame rate FBP-encoded (lower) "...in greasy..." utterance.

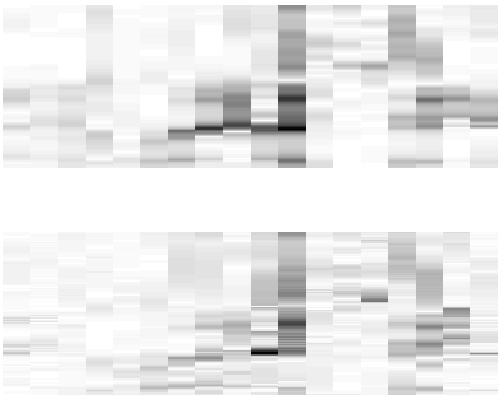

Figure 6. Delta coefficients (upper) and FBP prediction errors for "...in greasy..." utterance.

nally developed to estimate the information content of each new observation vector. This estimate is suitable for incorporation in a variable frame-rate coder (for use in HMM-based recognition [9]), in place of delta coefficients (to allow an HMM recogniser to identify slowly evolving speech segments [10]), or directly in a modified DTW recogniser, to concentrate attention on the most important (i.e. unexpected) vectors in the sequence.

In addition, the prediction matrix and the associated "dynamics" data could be incorporated directly into the recognition (or coding) process.

5. Potential Developments

FBP is most clearly applicable to spectrographic data, but can also be applied to other representations of speech [11]. The process can also be extended to the prediction of "composite" data made up of multiple simultaneous, but independent, processes (e.g. for the simultaneous prediction of pitch harmonics and vocal tract resonances) in a manner similar to that described in [12] for noise modelling.

References

[1] S. Furui; "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum"; IEEE Trans. ASSP, vol. 34, no. 1, pp. 52-59; 1986.

[2] K. Shirai and K. Mano; "A Clustering Experiment of the Spectra and Spectral Changes of Speech to Extract Phonemic Features"; Signal Processing, vol. 10, pp. 279-290; 1986.

[3] S. M. Peeling and K. M. Ponting; "Experiments in Variable Frame-Rate Analysis for Speech Recognition"; RSRE Memorandum 4330; 1989.

[4] P. Kenny, M. Lennig and P. Mermelstein; "A Linear Predictive HMM for Vector-Valued Observations with Applications to Speech Recognition"; IEEE Trans. ASSP, vol. 38, no. 2, pp. 220-225; 1990.

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

[6] Visual Representations of Speech Signals; Martin Cooke, Steve Beet and Malcolm Crawford (eds.); John Wiley and Sons, Ltd.; 1993.

[7] R. K. Moore, M. J. Tomlinson and S. W. Beet; "The Acoustic Flow of Speech"; Proc. IoA, vol. 6, pt. 4, pp. 241-248; 1984.

[8] B. R. Musicus; "Fast Power Spectrum Estimation from Uniformly Spaced Correlations", IEEE Trans. ASSP, vol. 33, no.4, pp. 1333-1335; 1985.

[9] A. P. Varga and R. K. Moore; "Hidden Markov Model Decomposition of Speech and Noise"; Proc. IEEE ICASSP-90, pp. 845-849; 1990.

[10] L. Baghai-Ravary, S. W. Beet and M. O. Tokhi; "Flow-Based Prediction: a Method for Improved Speech Recognition"; IEE Colloquium Digest, Techniques for Speech Processing and their Application, Savoy Place, London, 1st June, 1994.

[11] L. Baghai-Ravary, S. W. Beet and M. O. Tokhi; "Removing Redundancy from some Common Representations of Speech"; Proc. IoA, vol. 16; 1994.

[12] A. P. Varga and R. K. Moore; "Hidden Markov Model Decomposition of Speech and Noise"; Proc. IEEE ICASSP-90, pp. 845-849; 1990.