REMOVING REDUNDANCY FROM SOME COMMON REPRESENTATIONS OF SPEECH

L. Baghai-Ravary*, S. W. Beet† and M. O. Tokhi*

*Department of Automatic Control & Systems Engineering,
†Department of Electronic & Electrical Engineering,
University of Sheffield, Mappin Street, Sheffield, S1 3JD.

1. INTRODUCTION

This paper shows how a non-stationary vector predictor can be used to identify redundancy in various common forms of speech data. A number of different forms of data are used: some producing spectrogram-like representations, while others are rarely displayed graphically in the literature, and so many researchers are not familiar with the structure they exhibit (or the fact that they exhibit any significant structure at all).

The method used here is known as flow-based prediction (FBP) [1]. The prediction takes the form of a standard vector linear predictor [2], but with a sparse, time-varying, prediction matrix, which is updated over a very short time scale. This makes it eminently suitable for modelling speech dynamics, since large changes in, for example, formant trajectories, can occur over a very small number of analysis frames.

FBP, like the acoustic flow of Moore et al. [3], uses dynamic programming to estimate the most likely links between the elements of one observation vector and those of the next. However, FBP extends the acoustic flow concept to provide simultaneous estimates of the coefficient matrix and the innovation vector of a first-order vector linear predictor. These prediction parameters allow for positional shifts and merging of the features within the data vectors.

2. SPEECH DYNAMICS

The main articulators involved in speech production are not able to move abruptly. Speech signals can therefore be considered piecewise-continuous, except, for example, during plosives (where the signal statistics change rapidly). Plosive sounds have a short duration and the only other abrupt changes (from one continuous segment to the next) occur as a result of changes in voicing or nasalisation. Thus most of the speech signal evolves smoothly with respect to time.

This behaviour has previously been allowed for merely by temporal over-sampling, so that consecutive frames with smoothly-evolving characteristics can be identified as such by their small inter-frame Euclidean distances. In speech recognition, this approach is often implemented by the calculation of delta coefficients [4]. Thus, at present, speech recognition and coding systems do not fully account for speech dynamics, requiring significant temporal oversampling and even then, attributing undue importance to many insignificant parts of the signal.

The method described here, flow-based prediction (FBP), lowers the level of redundancy in speech data by tracking the features within the observation vectors and predicting their flow. FBP is computationally efficient and adapts very quickly to changes.

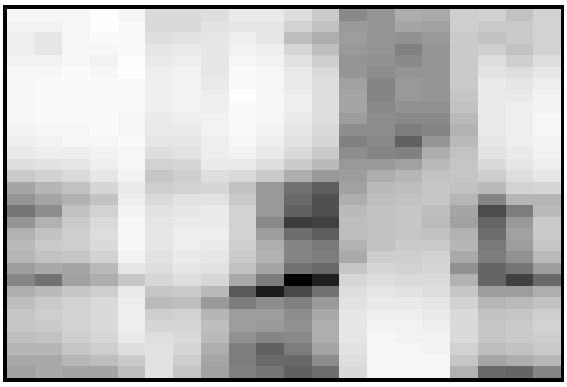


Figure 1: Spectrogram of the segment "...in greasy..."

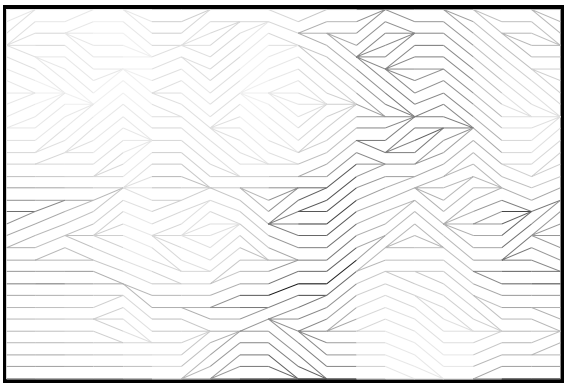


Figure 2: Spectrographic acoustic flow of the segment "...in greasy..."

3. FLOW-BASED PREDICTION

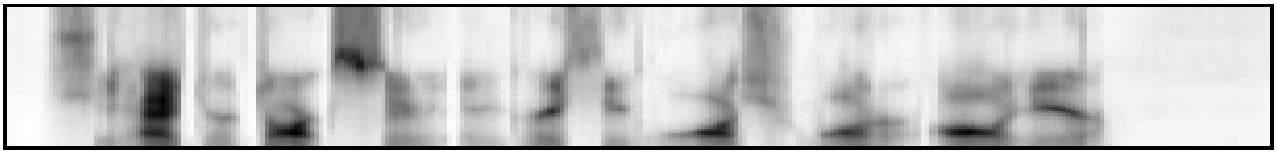
Acoustic flow uses dynamic programming to align consecutive observation vectors to make the evolution of the speech data manifest. The two plots in Figures 1 and 2 show a small segment of continuous speech. Figure 1 is the spectrogram of that utterance, which can be combined with the flow data, to give the spectrographic acoustic flow in Figure 2. Here, the darkness of each line illustrates the value of the observation vectors, and the lines themselves indicate the optimal links between one frame and the next. From these graphs it is apparent that, where the formants are changing smoothly, the flow has tracked that movement.

Flow-based prediction assumes that the change from one vector to the next can be modelled by averaging and shifting within a vector, together with a smoothly changing innovation. The process can be represented as non-stationary vector linear prediction:

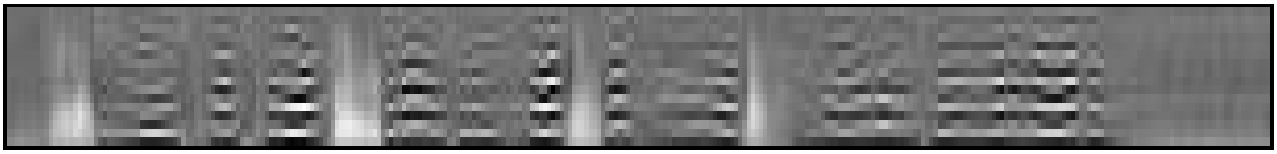
$$\mathbf{o}_{n+1} = \mathbf{C}_n \mathbf{o}_n + \mathbf{v}_n$$

However, there are two factors that differentiate it from conventional vector linear prediction. Firstly, the prediction matrix is automatically updated over a very short time scale, directly from the observation vectors, so as to track the features in the data more accurately. Secondly, the innovation vector is assumed to evolve steadily, following the lines of flow. The only exception to this is when an abrupt change occurs, when the innovation is assumed unpredictable, and estimated as zero.

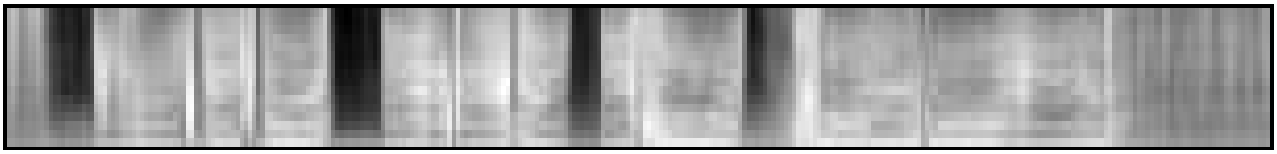
(a) Maximum entropy method power spectrum



(b) Maximum likelihood method power spectrum



(c) Linear prediction coefficients



(d) Log vocal tract area

Figure 3: Typical representations of the sentence "She had your dark suit in greasy wash water all year." spoken by an adult male.

4. SPEECH REPRESENTATIONS

There are many methods for analysing speech, each of which yields a different representation of the speech signal [5]. Even estimating the power spectrum of speech can give rise to a perplexing multitude of alternative algorithms, each with its own assumptions and peculiarities. For the purposes of this paper, the methods described below have been considered. Most of these are described in more detail in [6]. Wherever possible, the parameters of each analysis have been chosen to be comparable with each other. The details are given in the Appendix. Typical representations can be seen in Figure 3.

4.1 Periodogram

This is the most common method for visualising speech signals. It is formed by taking the discrete Fourier transform (DFT) of a windowed segment of speech, and finding the modulus squared of each complex output value. It provides an estimate of the power spectral density (PSD) which is degraded by the spectral effects of temporal windowing. The frequency resolution of the periodogram is inversely proportional to the length of the input frame (for a given window shape), and cannot be controlled independently, except by changing the window. The choice of window is

restricted by the expected dynamic range of the elements in each PSD estimate, and the required degree of temporal continuity. To give temporal continuity with adult male speech, this method can only give a narrow-band spectrogram, clearly resolving individual pitch harmonics, and making this representation unsuitable for simple HMM recognition.

4.2 Blackman-Tukey power spectrum

One method for controlling the resolution of a periodogram is to window an estimate of the autocorrelation function, rather than the data itself. This allows the frequency resolution to be reduced without losing temporal continuity. However, the window must have a non-negative Fourier transform for negative PSD estimates to be avoided. This method can give a broad-band spectrogram, characterising formant structure rather than pitch. Because of the limited frequency resolution, however, very closely-spaced formants are not always clearly resolved.

4.3 Maximum entropy power spectrum

The power spectrum of an autoregressive (AR) process can be obtained by calculating the parameters of the AR model from the autocorrelation function of the signal. This has been done here by Burg's method [6]. The maximum entropy method (MEM) PSD estimate is then obtained by multiplying the innovation power by the transfer function of the implied recursive filter. Since speech cannot always be approximated as an AR process (e.g. when corrupted by additive noise or reverberation, or during nasalised speech), the resulting PSD estimate can occasionally exhibit false peaks. Nonetheless, high quality speech recordings exhibit very clear formant tracks, and the resulting PSD estimate is visually very similar to that of a periodogram, but without any evidence of pitch harmonics.

4.4 Maximum likelihood power spectrum

This is variously referred to as the minimum variance PSD estimate, the maximum likelihood method (MLM) or Capon's method. It involves the design of an FIR filter for each frequency where an estimate of the PSD is required. These filters have unity gain at the design frequency, but with minimal overall output power. Thus the technique attempts to attenuate all but the frequency component of interest, and can be considered as a data-adaptive DFT. The power from each filter is calculated from the autocorrelation function of the signal, without explicitly implementing the filters, using the method described in [7].

The order of the filters determines the maximum number of frequency components which can be attenuated, and is chosen according to the application. To resolve formant structure while suppressing pitch information, the filter order should be chosen to be slightly more than twice the maximum number of formants, as in linear prediction analysis.

The frequency resolution is data-dependent, but generally intermediate between that of the maximum entropy and periodogram methods.

4.5 Cepstrum

Since speech can be considered as the product of a source spectrum and a vocal tract transfer function, pitch information can be separated from formant structure by homomorphic filtering. A

log-power periodogram is formed and then inverse-Fourier transformed to give a cepstrum containing formant data in its lower coefficients, with pitch being apparent at the higher end.

4.6 Linear prediction (LP) coefficients

Autoregressive modelling of speech signals can give a very concise description of the vocal tract transfer function. The results of this analysis are often presented as the coefficients of a ladder filter which can be used to predict one step ahead of the speech waveform. They generally exhibit a smooth, predictable structure during fricatives, but only their envelope consistently changes smoothly during voiced speech.

4.7 Reflection coefficients

Burg's method for calculating linear prediction coefficients is based on the calculation of reflection coefficients, which can be viewed as the parameters of an acoustic-pipe model of speech production [8]. These always have values between -1 and 1, so have lower dynamic range than standard linear prediction (ladder) coefficients, although many of their other properties are somewhat similar.

4.8 Vocal tract area functions

The shape of the acoustic pipe implied by a set of reflection coefficients can be calculated by adding successive log area ratios [8]. This gives a set of parameters which are loosely related to the cross-sectional area of the vocal tract, and therefore obey rules of motion similar to those of the real vocal tract. For example, as the tongue moves a constriction forward and backward, the vocal tract area function's values will move within the data vector, while the opening and closing of the mouth will affect the magnitude of the values at the respective end of that vector. However, there is an extremely abrupt change in their values between voiced and unvoiced speech. Such changes are difficult to predict and invariably give large prediction errors (at least for a short time).

5. RESULTS

Flow-based prediction was applied to one of the TIMIT files originally used as a standard test utterance in [5], taken from `/DR5/MEWMO/SA1.WAV`, "She had your dark suit in greasy wash water all year", spoken by a man from a Southern USA dialect region. The FBP algorithm was then compared with a zero-order predictor (which gives errors equal to the delta coefficients). The error magnitudes were calculated during steadily evolving segments of the utterance. Table 1 shows the FBP's degree of improved performance.

From Table 1 it is apparent that the speech representations which are most amenable to flow-based prediction (maximum likelihood, maximum entropy and Blackman-Tukey methods) are those PSD estimates which are tuned to resolve formant structure and suppress pitch (Figures 3(a) and 3(b)). The behaviour of the FBP algorithm, when predicting a diphthong, is shown in more detail in Figure 4. This demonstrates the FBP's ability to remove more of the redundancy from the data

than the zero-order predictor implicit in delta coefficient calculation, since the error magnitudes are smaller and exhibit less structure in the case of FBP.

Vocal tract area functions are also better modelled by FBP, because of their relationship to the positions of the articulators within the vocal tract. However, the advantage is only slight, because the longitudinal motion of those articulators only covers a limited range (see Figure 3(d)).

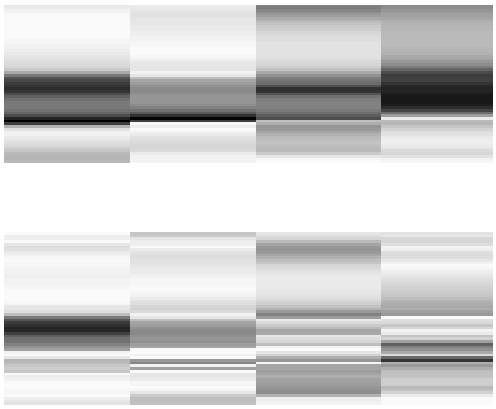


Figure 4: Delta coefficient (upper) and FBP error (lower) magnitude plots for the diphone "...rea..."

SPEECH REPRESENTATION	FBP IMPROVEMENT
Maximum likelihood PSD	57%
Maximum entropy PSD	49%
Blackman-Tukey PSD	36%
Vocal tract area	6%
Periodogram	None
Cepstrum	None
Linear prediction	None
Reflection coefficients	None

Table 1: Comparison of zero-order and flow-based prediction errors.

Those methods which yield parameters directly related to impulse response-like functions (linear prediction and reflection coefficients) only produce data with an appropriate structure during fricatives (see Figure 3(c), for example), which constitute only a small part of most utterances. Elsewhere, the most significant correlation between elements of consecutive vectors is between identical elements: there is little migration of features between elements, so the acoustic flow is rarely of any use. However, the information they contain must still be predictable, since they can be transformed into a suitable PSD form. All that these negative results show is that the evolution of the speech signal is more difficult to model in these domains.

In the case of the cepstrum, FBP gave no measurable advantage over a zero-order predictor, but in this case, the problem is attributable to the scoring method used, which took no account of the variance of individual elements within the vector. In the case of the cepstrum, the first coefficient has much greater variance than any of the others, so the error scores are heavily biassed towards that coefficient.

The narrow-band periodogram is inherently inappropriate to the model assumed in flow-based prediction, since the pitch harmonics move independently of the formants. This means that a more complicated model of combined pitch and formant evolution is required.

6. CONCLUSIONS

Flow-based prediction yields an accurate model of speech dynamics, provided the data changes smoothly. In this context, broad-band PSD estimates are therefore the most powerful representation for characterising smooth changes in speech. However, the current FBP model can only cope with one aspect of signal evolution at a time (pitch *or* formants, but not both), so data such as narrow-band periodograms is not appropriate.

When used on a suitable form of data, the FBP error has lower redundancy than delta coefficients (zero-order prediction error) and can be calculated at a reduced computational cost, and with fewer prior observation vectors, than full first-order vector linear predictor parameters.

7. FUTURE WORK

There is considerable scope for further work on FBP. One area would be the development of a combined model for simultaneous evolution of pitch and formant structure. Another would involve development of evolutionary models for non PSD-like speech representations (such as LP coefficients). One approach to this might involve calculating acoustic flow in a domain where FBP's assumption of "steady evolution" *is* valid, and then converting the prediction into a different domain (MEM, MLM, reflection coefficients, vocal tract areas and LP representations are all calculated from the same Burg algorithm and are inter-related).

8. APPENDIX: IMPLEMENTATION ISSUES

8.1 Input data

The data used here was taken from the TIMIT database, which was sampled at 16 kHz. The speech was pre-emphasised, giving roughly 6dB per octave gain above 500 Hz, prior to each analysis.

All the analysis methods used here are frame-based techniques, but the way the data is treated affects the temporal continuity of the resulting speech representations. For those analyses which analyse the data directly, each frame has been chosen to include at least two pitch pulses, and so the duration has been set to 25 milliseconds. However, those which initially window the data, have used a 50 millisecond minimum 4-sample Blackman-Harris window [9]. In either case a frame rate of 80 per second was chosen. This gives roughly 50% overlap correlation between successive data windows in both cases.

8.2 AR models

Much of the data presented in this paper was calculated by autoregressive (AR) modelling. In all cases, the order was set to 16, and Burg's method was used to estimate the AR parameters.

8.3 Logarithms

Power spectrum estimates are normally encoded on a log scale. In this paper, this scale is approximated by a function with similar, but more well-behaved, numerical properties. The same function is used to encode the vocal tract area functions, and in the intermediate calculations for the cepstrum.

In practice, log scales can cause problems when numbers become very small, and are totally impractical if numbers can become negative (due to rounding errors, etc.). To avoid this, it is usual to set a lower threshold on the data values, before the log is taken. This, however, assumes that the range of values is known a priori. To avoid having to estimate the respective ranges, a log scale can be approximated by taking the Nth root of the data value:

$$\ln(x) \approx N\left(\sqrt[N]{x} - 1\right) \quad ; \quad N >> 1, \quad x \approx 1$$

Here, N is a constant defining the range over which this formula is valid. The larger N is, the wider the range on either side of x = 1, for which the approximation holds. Furthermore, if N is chosen to be a positive, odd integer, this equation will be monotonic and calculable for any real value of x. In applications where scaling and offset on the resulting values is not important, it has useful properties related to amplitude-independence. In the data presented here, a value of N = 5 has been used, giving an effective dynamic range of 200:1 regardless of the mean level of the data. Interestingly, this value is similar to that used in many auditory models.

8.4 Autocorrelation functions

The autocorrelation function for the Blackman-Tukey PSD estimate was estimated from the inverse discrete Fourier transform of a periodogram, and windowed with the autocorrelation function of a minimum 4-sample Blackman-Harris window. This in itself is a finite-duration function, nonnegative for all time and at all frequencies. It therefore provides a valid PSD estimate.

9. REFERENCES

[1] S. W. Beet, L. Baghai-Ravary and M. O. Tokhi; "Non-Stationary Prediction of Speech Data"; Signal Processing VII: Theories and Applications; M. Holt, C. Cowan, P. Grant and W. Sandham (eds.), vol. 3, pp. 1653-1656; 1994.

[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] R. K. Moore, M. J. Tomlinson and S. W. Beet; "The Acoustic Flow of Speech"; Proc. IoA, vol. 6, pt. 4, pp. 241-8; 1984.

[4] 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.

[5] Visual Representations of Speech Signals; M. P. Cooke, S. W. Beet and M. D. Crawford (eds.); John Wiley and Sons, Ltd.; 1993.

[6] Digital Signal Processing: Principles, Algorithms and Applications, second edition; J. G. Proakis and D. G. Manolakis; Macmillan Publishing Company; 1992.
[7] B. R. Musicus; "Fast Power Spectrum Estimation from Uniformly Spaced Correlations", IEEE Trans. ASSP, vol. 33, no. 4, pp. 1333-1335; 1985.
[8] Digital Speech Processing, Synthesis and Recognition; S. Furui; Marcel Dekker, Inc.; 1989.
[9] Handbook of Digital Signal Processing Engineering Applications; D. F. Elliott (ed.); Academic Press; 1987.