RECOGNITION OF PHONEMES FROM ESTIMATION ERRORS

L Baghai-Ravary and S W Beet

Department of Electronic and Electrical Engineering,
The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK.
Tel: (+44) 114 282 5409; Fax: (+44) 114 272 6391
Email: l.baghai-ravary@shef.ac.uk, s.beet@shef.ac.uk

ABSTRACT

Speech recognition systems generally use delta and delta-delta (velocity and acceleration) coefficients to characterise the dynamics apparent in frame-based representations of speech. These coefficients can be thought of as the errors of simple predictors.

This paper describes the use of error coefficients derived from more advanced (and accurate) forms of prediction and interpolation. Both overall recognition accuracy and the detailed confusions observed are compared with those of the 'traditional' methods. The task used is speaker-independent phoneme recognition using a subset of the TIMIT database, and four different speech representations. The error coefficient performance on this task appears to be directly related to the robustness of the estimator used, with the best of the new methods out-performing delta-delta coefficients by around 10%.

1 EXPERIMENTAL PROCEDURE

For all the speech recognition experiments described in this paper, a phoneme recognition task was selected to show up differences in performance between the different analysis and modelling methods. This was chosen in preference to, for example, whole-word recognition because even the best current recognisers make sufficient errors on this type of task for comparative results to be meaningful. The identification of sub-word units also provides diagnostic information as to the forms of discrimination which are provided by each representation or recognition method.

The experiments are based on the phonetically-labelled speech database, TIMIT [1]. The recognition experiments were conducted using Cambridge University's HTK Hidden Markov Model Toolkit [2].

For the work here, a simple 3-state HMM, with no parameter tying, was used for each phoneme. The left-right structure of figure 1 was used. This is

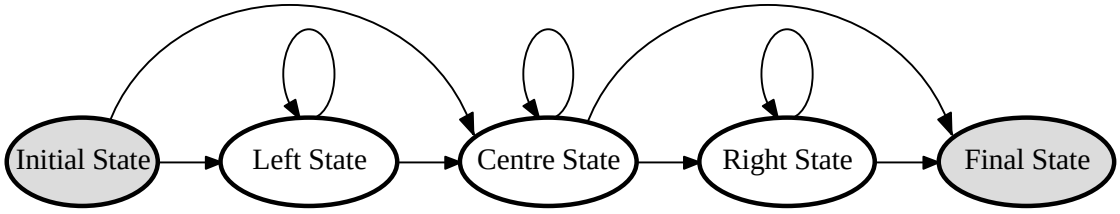

Figure 1: HMM used for each phoneme. Only those shown with white backgrounds emit observations.

only one of many possible structures [3], but it was chosen because it forces all instantiations of the respective phoneme to occupy the centre state at some point. This presumes that each phoneme has one segment which has a characteristic PDF, and which must be present for the phoneme to be identified. This segment may be preceded or followed by other less critical (and probably less well-defined) segments, so the other states may be skipped. Thus the essential segment which characterises each phoneme can be learnt as being at the beginning, middle or end of each phoneme, as appropriate.

Each state was allocated 3 mixtures in its PDF, as is commonly reported in the literature [4]. This represents a compromise between computational tractability and accuracy of modelling the true distributions, which frequently deviate from a simple Gaussian. Delta coefficients were also used in all experiments, being treated as a separate data stream of observations.

Because the available training data was somewhat restricted (due to the deliberate avoidance of dialect and gender-related variability, mentioned previously), the number of variables which must be learned by the HMM has been minimised wherever possible. In particular, a state-dependent diagonal covariance matrix has been used. In this case, the number of variables to be estimated is proportional to the dimensionality of the speech representation. Since the number of values present in that representation is also proportional to its dimensionality, there is no need to increase the size of the training data set even if a representation of higher dimensionality is used.

If a full covariance matrix had been used, the number of values to be estimated would have increased with the square of the dimensionality of the speech representation. Thus the size of the training data set would need to increase linearly with that dimensionality. In practice, the size of the training data set is fixed, so high-dimensional representations would be at a disadvantage with respect to those with lower dimensionality.

Finally, inter-phone-model transition probabilities were incorporated based on the original labelling of the TIMIT database' training set, subjected to a minimum value to cope with transitions which were under-represented in the training data.

2 VECTOR ESTIMATION

It is almost universal practice to augment the basic observation vector sequence with delta coefficients to improve recognition performance. A further improvement can be obtained by appending further dynamics information with delta-delta (or acceleration) coefficients. These can be viewed as a form of prediction error, but other forms of predictor can chracterise speech dynamics more accurately.

The errors of such models can be appended to the obeservation vectors in place of delta-delta coefficients. The estimators used here are described briefly below, but further details are to be found in [5].

2.1 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 requires over-sampling to model the sequence accurately.

2.2 Flow-Based Predictor

This method (FBP) 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. The resulting prediction is accurate during voiced sounds, but is susceptible to relatively large errors during abrupt onsets.

2.3 Adaptive Flux interpolation

The fundamental aim of the Adaptive Flux Interpolation (AFI) method 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. This method is at least as accurate as FBP, but is also more robust.

3 RESULTS

The results are shown in figure 2, where the improvement in recognition accuracy is shown for each of three auditory spectrum representations, together with the commonly used auditory cepstrum coefficients.

The improvement due to FBP and AFI is between 4% and 10%, depending on the representation. In general, FBP performed better than AFI (by about 1%), except in the case of the Blackman-Tukey power spectrum, where AFI was clearly superior.

4 CONCLUSIONS

The more complete removal of redundancy from the speech representations improves recognition accuracy. However, although AFI outperforms FBP in terms of mean square estimation error [5], it gives a better recognition result.

This may be because the statistical model at the heart of an HMM estimates the probability of an observation sequence using recursive application of the equation:

$$\begin{aligned} &\mathrm{p}(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_N \mid M) \\ &\quad = \mathrm{p}(\mathbf{o}_1, \mathbf{o}_2, \ldots \mid \mathbf{o}_N) \times \mathrm{p}(\mathbf{o}_N \mid M) \end{aligned}$$

where p is the probability operator, N is the most recent time index, M is the model and $\mathbf{o}_1$, $\mathbf{o}_2$, ..., $\mathbf{o}_N$ is the sequence of observation vectors.

When incorporating the estimation error statistics in this formalism, the probability of the observation sequence up to time N cannot be a function of future observations. Thus, interpolative approaches (such as AFI) will interfere with the assumed independence between $\mathbf{o}_N$ and its predecessors.

This effect is not apparent with the auditory Blackman-Tukey method since this was implemented with a longer time-domain window (to ensure comparable temporal continuity in its output data [6]). This increases the correlation between successive observation vectors and makes

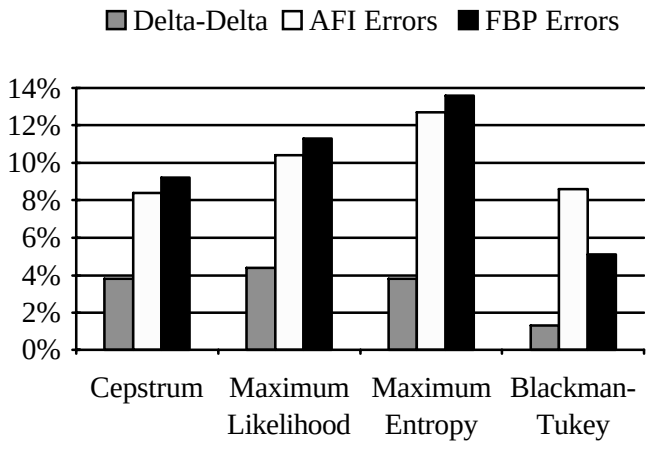

Figure 2: Relative increase in recognition accuracy.

the non-causality of the model less apparent. In this case, AFI's increased robustness outweighs FBP's causality.

REFERENCES

[1] Price P J, Fisher W, Bernstein J, et al. “A database for continuous speech recognition in a 1000-word domain”, Proc. ICASSP '88, New York, vol. 1, pp. 651-654, 1988.

[2] Young S J, “HTK version 1.4: Reference Manual and User Manual”, Cambridge University Engineering Department Speech Group, 1992.

[3] Deller J R, Proakis J G and Hansen J H, “Discrete-Time Processing of Speech Signals”, Macmillan Publishing Company, New York, 1993, pp. 724-728.

[4] Bahl L R, Jelinek F and Mercer L, “A maximum likelihood approach to continuous speech recognition”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 5, pp. 179-190, 1983.

[5] Baghai-Ravary L, Beet S W and Tokhi M O, “Adaptive flux interpolation, flow-based prediction, delta or delta-delta coefficients: which is best?”, Proc. Eurospeech '95, Madrid, vol. 2, pp. 1037-1040, 1995.

[6] Baghai-Ravary L, Tokhi M O and Beet S W, “Modelling the flow inherent in speech representations”, University of Sheffield Department of Automatic Control and Systems Engineering Research Report 551, 1994.