LINEAR DYNAMIC SEGMENTAL HMMS: VARIABILITY REPRESENTATION AND TRAINING PROCEDURE

Wendy J. Holmes and Martin J. Russell

e-mail: holmes@signal.dra.hmg.gb and russell@signal.dra.hmg.gb

Speech Research Unit, DRA Malvern,
St Andrews Road, Malvern, Worcs WR14 3PS, UK

ABSTRACT

This paper describes investigations into the use of linear dynamic segmental hidden Markov models (SHMMs) for modelling speech feature-vector trajectories and their associated variability. These models use linear trajectories to describe how features change over time, and distinguish between **extra-segmental** variability of different trajectories and **intra-segmental** variability of individual observations around any one trajectory. Analyses of mel cepstrum features have indicated that a linear trajectory is a reasonable approximation when using models with three states per phone. Good recognition performance has been demonstrated with linear SHMMs. This performance is, however, dependent on the model initialisation and training strategy, and on representing the distributions accurately according to the model assumptions.

1. INTRODUCTION

Two fundamental concepts in segmental HMMs [1,2,3] are modelling variability between different instantiations of a sub-phonemic speech segment separately from that within one example, and the notion of an underlying parametric trajectory describing how acoustic feature vectors change over time during a segment. The simplest case is a static SHMM [1] (or “target state segment model” [4]), where the trajectory is assumed to be constant over time and so is represented by a single “target” vector. A linear dynamic SHMM [2,3] is obtained by assuming that the underlying trajectory changes linearly, such that the trajectory is described by mid-point and slope vectors. A segment probability has two components: the **extra-segmental** probability of a trajectory given the model state, and the **intra-segmental** probability of the observations given the trajectory.

In the case of the linear model, we suppose that the distribution of trajectory parameters for a given state can be described by Gaussian distributions $N_{(\mu,\gamma)}$ and $N_{(\nu,\eta)}$ (with diagonal covariance matrices) for the slope m and mid-point c respectively. The intra-segmental distribution is assumed to be Gaussian with diagonal covariance τ. Ignoring any duration probability, the joint probability of a segment of observations $y = y_0, \ldots, y_T$ and a trajectory $f_{(m,c)}$ is specified as

$$P(y,m,c) = N_{(\mu,\gamma)}(m).N_{(\nu,\eta)}(c).\prod_{t=0}^{T} N_{(f_{(m,c)},\tau)}(y_t) .$$

We define the probability of a segment given a linear Gaussian segmental HMM (GSHMM) state as being the above quantity for the *optimal trajectory*, which is defined by a maximum *a posteriori* estimate of the slope $(\hat{m})$ and mid-point $(\hat{c})$. These values can be shown to be a weighted sum of the values which are optimal with respect to the data and the expected values as defined by the model, thus:

$$\hat{m} = \frac{\left(\sum_{t=0}^{T}(t-\frac{T}{2})y_t\right)\gamma + \mu\tau}{\left(\sum_{t=0}^{T}(t-\frac{T}{2})^2\right)\gamma + \tau} \quad \text{and} \quad \hat{c} = \frac{\left(\sum_{t=0}^{T} y_t\right)\eta + \nu\tau}{(T+1)\eta + \tau} \qquad (1).$$

A consequence of the two-stage model of variability is that different explanations of any one utterance use different numbers of intra- and extra-segmental probabilities. Hence, the models only perform appropriately for recognition if the two types of probability balance correctly. Experiments with the static model [3] demonstrated that a suitable balance can be achieved over a fairly wide range of segment durations, provided that the extra- and the intra-segmental distributions both fit the model assumptions. In particular, performance was greatly improved by using a two-component Gaussian mixture for the intra-segmental distribution. With appropriate model initialisation, these models outperformed conventional HMMs [5]. The current paper focuses on the linear GSHMM, with the aim of combining an appropriate trajectory description with accurate distribution modelling. The experiments use the same connected-digit recognition task with three-state phone models as in previous studies [2,5,6]. Speech data is analysed to investigate the validity of a linear trajectory assumption, and recognition experiments are described which demonstrate the importance of accurate distribution modelling and of adopting a suitable model initialisation and training strategy.

2. TRAJECTORY MODEL REPRESENTATIONS

2.1. Method

The aim of these studies was to analyse trajectories of acoustic features as described by a set of simple static and linear GSHMMs, independently from particular segmental models. The data was labelled at the segment level by using trained three-state-per-phone standard HMMs to perform a Viterbi alignment with the known transcription. A trajectory vector was estimated for each identified segment as the average of the observed feature vectors for the static model and as the best-fitting straight line parameters for the linear model.

2.2. Trajectory description

Figure 1 illustrates that, whereas the static model provides only a crude approximation to the observed features, the linear model generally follows the pattern of change very well. There

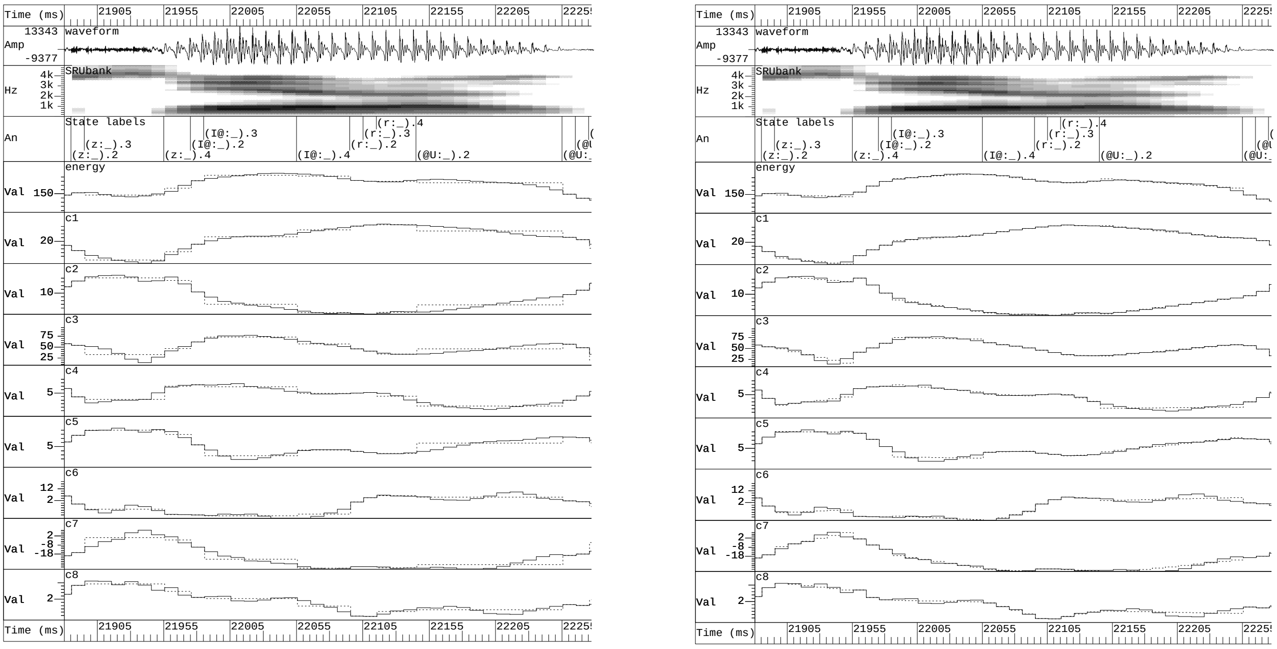

Figure 1 - Frame-by-frame values (solid lines) superimposed on calculated model values (dotted lines) for mel cepstrum features representing the digit "zero", as described by static (left plot) and linear modelling assumptions (right plot).

is some loss of detail in the linear approximation for higher-order cepstral features (from around the sixth upwards), which tend to change less smoothly than low-order features. A linear model should, however, be adequate to capture general time-evolving characteristics, with further variation around the linear trajectory modelled as random by the intra-segmental variance.

3. RECOGNITION EXPERIMENTS

3.1. Method

For the segmental HMMs, a strict left-right topology was used with no self-loops and the maximum segment duration was set to 10 frames. This model structure imposes a maximum phone duration of 300 ms, which was considered adequate for most speech sounds in connected speech. Self-loops were used for the non-speech models, to provide a simple way of allowing long periods of silence. All segment durations were assigned equal probability and duration distributions were not re-estimated. The other parameters were trained with five iterations of Baum-Welch re-estimation.

3.2. Model initialisation

It has been found [5] that a successful initialisation strategy is to (automatically) estimate the model parameters directly from the complete set of training data as segmented by trained standard HMMs. In the current experiments, different strategies of this type were investigated for the linear models. For each feature of every example of a segment, the best-fitting trajectory parameters were determined. The means and variances of the mid-points were initialised from the distributions of the individual mid-points. Different alternatives for using the slope parameters were investigated. One possibility is for the model to allow the slope of a feature to vary sufficiently to accommodate all observed trajectories for the segment, so the intra-segmental variance should be very small. An alternative would be for the slope to be more constrained and for the intra-segmental variance to be larger to allow for the greater variability of the observations around the optimal trajectory. The first approach should be able to represent all observed trajectories quite closely, while the second approach provides more model-dependent constraints, which may be better for short segments when it is difficult to compute a representative slope from the data alone. To investigate the properties of these alternative approaches, different initialisation strategies were compared, thus:

1. *fully-flexible slope:* the means and variances of the individual slopes were determined and the intra-segmental variance around the individual trajectories was estimated.
2. *constrained slope variance:* the slope means were set in the same way but their variances were set to a small fixed value. The intra-segmental variance was initialised by determining the variability of the observations around a line with segment-dependent mid-point but fixed mean slope.
3. *zero mean slope with constrained variance:* the slope means were all initialised to zero, and their variances were set to a small fixed value. The intra-segmental variances were initialised from the variability of the observations around a line with segment-dependent mid-point and a slope of zero.
4. *fixed zero mean slope with constrained variance:* models were initialised as for 3 above, but the slope means were fixed at zero during training. These models use the linear GSHMM structure but are in effect almost static (as the slope variance is small), and therefore allow for a direct comparison to evaluate the influence of modelling dynamics.

3.3. Recognition results

The recognition results for the different linear GSHMMS are summarised in Table 1. The recognition performance is very poor for the models (1) initialised with a fully-flexible slope parameter. The high proportion of word deletion and substitution errors reflects a problem with misrecognising sequences of short segments as smaller numbers of longer segments, frequently silence. When the slope variance was initialised to a small value (2,3), the word error rate was much lower. The slope variance remained small after training, and it thus appears that the models provide better discrimination if they do not attempt to describe variability in the dynamics. Some representation of dynamics is important however, as models initialised with zero-mean slope performed much better when allowed to deviate from the zero-mean condition during training (3) than if the slope mean was fixed (4). Recognition performance is generally quite disappointing for all model sets. The model of intra-segmental variability was therefore studied, as this was found to be an issue with static models [4,5].

Model set	% Corr	% Subs	% Del	% Ins	% Err
Standard HMM	93.2	5.6	1.2	1.0	7.8
LGSHMM 1	67.5	17.3	15.2	0.1	32.6
LGSHMM 2	91.7	4.2	4.1	0.1	8.4
LGSHMM 3	92.1	3.9	4.0	0.0	7.9
LGSHMM 4	74.2	16.1	9.7	0.1	25.9

Table 1: Connected-digit recognition results for standard HMMs and different sets of Linear GSHMMs.

4. MODELLING INTRA-SEGMENTAL VARIABILITY

4.1. Distributions describing segmental variability

4.1.1. Method

Based on the entire training corpus, intra-segmental distributions of the speech feature vectors were estimated for each model state, using the same procedure as in previous experiments [4,5]: a segmental Viterbi alignment procedure was performed to associate each speech frame with a single model state. For each segment identified, the optimal feature vector trajectory was computed and hence the distribution of differences between the trajectories and the observed feature values was derived. These distributions were compared with the distributions specified by the segmental models.

4.1.2. Results

As can be seen from Figure 2a showing typical example distributions, the intra-segmental variance varies according to the model set. Not surprisingly, this variance is smallest when the trajectory slope can vary to accommodate different examples. In addition, the importance of modelling dynamics is further supported by the observation that the intra-segmental variance is largest when the model mean slope is fixed at zero.

In all cases, a single-Gaussian model is not a very good representation for the intra-segmental distributions around optimal trajectories: the probability of very close matches to the mean is underestimated, while that of somewhat greater deviations is overestimated. There are two important influences determining the shapes of these distributions: the validity of the trajectory model, and the general problem of estimating a population mean and variance from a small sample of data. Thus, there will be a tendency to underestimate the variance, especially for very short segments, and this problem is greatest for the linear model with flexible slope. However, the trajectory assumptions are evidently more valid for the linear model, and so the true variance will be smaller. In general, the models with a constrained non-zero slope seem to provide the best compromise.

As with the static GSHMMs (although to a lesser extent), the distribution shapes should be improved by using a mixture of two Gaussians, each with the same mean but one with much smaller variance than the other. A theory of multiple-component intra-segmental mixture linear GSHMMs is therefore developed in the following section.

4.2. Theory of intra-segmental mixture linear GSHMMs

An intra-segmental mixture linear GSHMM is described by single-Gaussian distributions for the parameters defining the trajectory, and a mixture of I Gaussians to represent intra-segmental variance. Each component i has diagonal covariance τ_i and weight w_i. The probability of a segment of observations $y = y_0,\ldots,y_T$ for a given model state is defined as

$$\hat{P}(y) = N_{(\mu,\gamma)}(\hat{m}).N_{(\nu,\eta)}(\hat{c}).\prod_{t=0}^{T}\left(\sum_{i=1}^{I} w_i.N_{(f_{(\hat{m},\hat{c})},\tau_i)}(y_t)\right).$$

The values of $\hat{m}$ and $\hat{c}$ are given by

$$\hat{m} = \frac{\mu + \left(\sum_{t=0}^{T} p_t.(t-\frac{T}{2})(y_t-\hat{c})\right)\gamma}{1+\left(\sum_{t=0}^{T} p_t.(t-\frac{T}{2})^2\right)\gamma} \quad \text{and}$$

$$\hat{c} = \frac{\nu + \eta\sum_{t=0}^{T} p_t.(y_t-(t-\frac{T}{2})\hat{m})}{1+\eta\sum_{t=0}^{T} p_t} \text{ where } p_t = \frac{\sum_{i=1}^{I}\frac{w_i}{\tau_i}P_i(y_t|\hat{m},\hat{c})}{\sum_{i=1}^{I} w_i P_i(y_t|\hat{m},\hat{c})}.$$

As with the static model [4], the definition of the optimal trajectory is thus an iterative one, as it depends on existing values of trajectory parameters. In practice, provided the initial estimates are reasonable, the estimates converge within a very small number of iterations. For the case of a two-component mixture intra-segmental model, it is useful to consider the range of values which p_t can take, and the effect on the calculated optimal trajectory. If τ_2 is small relative to τ_1, then p_t will lie between $1/\tau_1$ and $1/\tau_2$, and τ_2 will only have any substantial influence for observations close to the trajectory. The result is that p_t will be highest for those observations nearest the trajectory and hence these observations will have the most influence on the optimal trajectory calculation. This effect of reducing the influence of occasional outliers seems an intuitively sensible one. In the special case where the intra-segmental variance is represented by a single Gaussian, p_t reduces to $1/\tau$ and hence the expressions for the optimal slope and mid-point simplify to those in (1).

4.3. Intra-segmental mixture experiments

4.3.1. Training procedure

The models were trained using the same approach adopted for the most recent experiments with static models [6]. The models were initialised from a standard-HMM segmentation in the same way as for single-Gaussian models, and the second mixture component then added with a small variance and low weight before training the models. Four different sets of models were trained (each with five iterations of Baum-Welch re-estimation), as for the single Gaussian models.

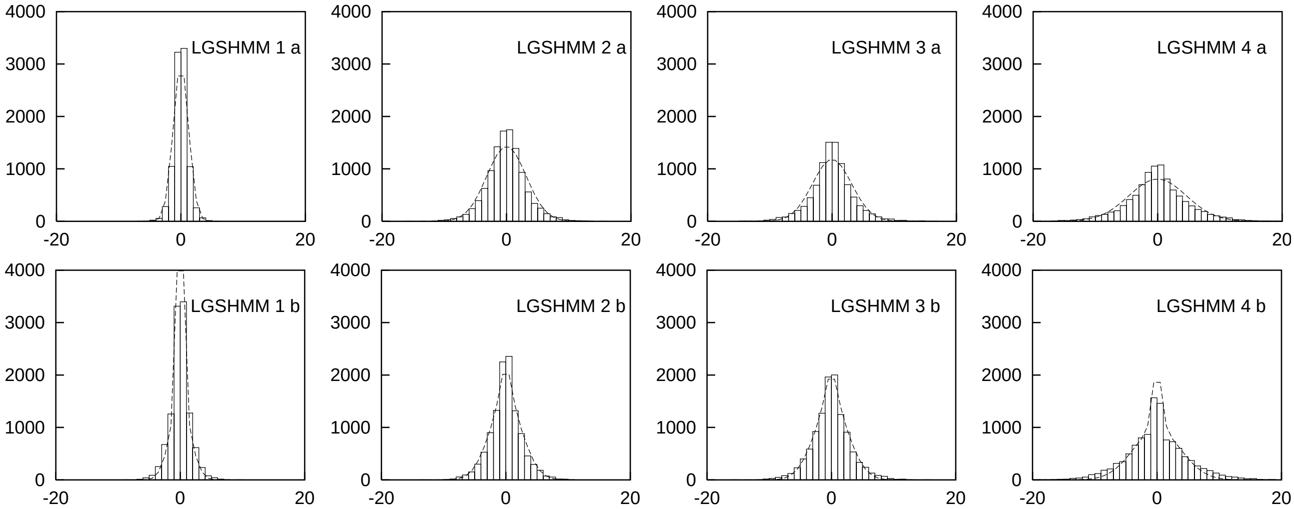

Figure 2 - Observed intra-segmental distributions plotted with calculated model distributions for the second cepstral coefficient representing the final state of /eI/ for **(a)** single-Gaussian models (top) and **(b)** two-component Gaussian mixture models (bottom).

4.3.2. Recognition results and discussion

With improved distribution modelling (Figure 2b), recognition performance has improved for all sets of models (Table 2). However, model initialisation strategy is still important. The models initialised with a constrained slope still outperform the fully-flexible models. It therefore appears that, even with quite accurate distribution modelling, attempting to model variability in the dynamics is detrimental to speaker-independent recognition performance. It is obviously important to model the general nature of temporal changes. However, variation in the detail of the dynamics, particularly across speakers, may not be consistent or important for distinguishing sounds. Another important factor is likely to be the difficulties in reliably estimating dynamics for short segments, which is probably why it is better to initialise the model slope means to zero (3) than to use the average slope over all segments (2), some of which will be unreliable.

The best set of linear models gives an error rate of only 3.3%, compared with 7.8% for conventional HMMs. However, if the conventional HMMs include derivative features computed using linear regression over five frames, the error rate reduces to 3.1%. Although the use of derivative features only provides implicit modelling of dynamics, some representation of change is provided for every frame. However, the segmental model only represents dynamics for the duration of a segment, and the representation is therefore only reliable for segments which are at least a few frames long. For this reason, further performance advantages may be obtained by using derivative features with the segmental models, as has been found by other researchers, for example Digalakis [7]. It would however be preferable to actually model dynamics across segments.

Model set	% Corr	% Subs	% Del	% Ins	% Err
LGSHMM 1	86.2	9.2	4.7	0.1	13.9
LGSHMM 2	93.6	3.4	3.0	0.1	6.5
LGSHMM 3	96.8	2.0	1.2	0.1	3.3
LGSHMM 4	91.5	6.1	2.4	0.1	8.6

Table 2: Recognition results for 2-component intra-segmental mixture Linear GSHMMs.

5. CONCLUSIONS

It has been demonstrated that linear GSHMMs can outperform conventional HMMs. As for static GSHMMs, recognition performance depends on describing the distributions accurately according to the model assumptions and on having an appropriate model initialisation strategy. It is also important that the parameters are used in the best way for reliable discrimination between segments. Current experimental results suggest that it may not be useful to attempt to represent variability in dynamics, at least for speaker-independent models. Further work on modelling dynamics is comparing speaker-independent with speaker-dependent modelling. Additional experiments are being carried out to investigate GSHMMs using a formant representation, which may be more suited to a linear trajectory model than the higher-order cepstra.

To obtain full benefit from a segmental approach, it is probably necessary to incorporate a model for dynamics across segments, as currently there are only advantages for fairly long segments. This is likely to be the main reason why linear GSHMMs have not so far outperformed HMMs with time derivative features.

6. REFERENCES

[1]M.J. Russell, "A segmental HMM for speech pattern modelling", *Proc. IEEE ICASSP*, Minneapolis, pp. 499-502, 1993.

[2]W.J. Holmes and M.J. Russell "Speech recognition using a linear dynamic segmental HMM", *Proc. EUROSPEECH'95*, Madrid, pp. 1611-1614, 1995.

[3]M.J. Russell and W.J. Holmes, "Linear Trajectory Segmental HMMs", to appear in *IEEE Signal Processing Letters*, 1997.

[4]M. Ostendorf, V. Digalakis and O. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition", *IEEE Trans. SAP*, Vol. 4, No. 5, pp. 360-378, 1996.

[5]W.J. Holmes and M.J. Russell, "Modeling speech variability with segmental HMMs", *Proc. IEEE ICASSP*, Atlanta, pp. 447-450, 1996.

[6]W.J. Holmes and M.J. Russell, "Modelling variability in speech patterns using dynamic segmental HMMs", *Proc. IOA*, Vol. 18, Part 9, 1996.

[7]V. Digalakis, 'Segment-based stochastic models of spectral dynamics for continuous speech recognition', PhD Thesis, Boston University, 1992.