Final Report on EPSRC (ROPA) Research Grant GR/M75204/01
Continuous-state dynamical system models for speech recognition

Simon King, Centre for Speech Technology Research, University of Edinburgh
www.cstr.ed.ac.uk

1st Oct 1999 - 30th Sept 2000 and 1st Apr 2001 - 31st Jan 2002 (22 months total)

1 Project aims

In our original proposal, we argued that discrete state Hidden Markov Models are inadequate for modelling observations produced by underlying continuous processes: for example, the movements of the articulators during speech production. We proposed to develop a continuous state model, known as a linear dynamical system model (LDM) or Kalman filter.

Our primary goal therefore was to come to a deeper understanding of these models and their advantages and disadvantages for use in speech recognition. Our strategy was to initially investigate these models using articulatory data [26, 27] because this data exhibits some very desirable properties (smoothness, continuity) which the LDM should be ideal for modelling, then move on to standard acoustic-only data, such as TIMIT [9].

1.1 Report structure

We will very briefly outline how linear dynamical system models (LDMs) operate, place our work in context, go on to report our experimental findings, then finish with a summary of our understanding of the power of these models and what directions our future work will be taking. Along the way we will refer to our publications where more details of our work are reported. We will show how our work fits in to the growing field of novel acoustic modelling that has been developing over that last 5 years or so by citing key papers from other groups.

2 Background

The following pair of equations define an LDM, where the initial state value $\mathbf{x}_0 \sim N(\boldsymbol{\pi}, \Lambda)$:

$$\mathbf{y}_t = H\mathbf{x}_t + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{v}, C) \tag{1}$$

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim N(\mathbf{w}, D) \tag{2}$$

2.1 How LDMs work

The basic premise of the model is that there is some underlying dynamic process which can be modelled by equation 2. This equation describes how $\mathbf{x_t}$, the state variable at time t, evolves from one time frame to the next. A linear transformation via the matrix F and the addition of some Gaussian noise, $\boldsymbol{\eta}_t$, provide this, the dynamic portion of the model.

The complexity of the motion that equation 2 can model is determined by the dimensionality of the state variable. For example, a 1 dimensional state space would allow exponential growth or decay with an overall drift ($\mathbf{v}$ can be non-zero) and 2 dimensions could describe damped oscillation with a drift. Increasing the dimensionality beyond 4 or 5 degrees of freedom allows fairly complex trajectories to be modelled. The observation vectors, given at time t by $\mathbf{y}_t$, represent realisations of this unseen dynamic process. A linear transformation with the matrix H and the addition of measurement noise, $\boldsymbol{\epsilon}_t$ (equation 1) relate the two. The observed trajectories could be modelled directly, however using a hidden state space in this way makes a distinction between the production mechanism at work and the parametrisation chosen to represent it.

H being a linear mapping means that trajectories which are continuous in state space are continuous in observation space. Points near each other in state-space are near each other in observation space.

For modelling speech, the model can be interpreted in several ways. If y is a vector of co-ordinates of selected points on the articulators, such as in data like [26], then x is the underlying articulator settings described in a more succinct form (x is typically of lower dimension than y). If on the other hand, y is acoustic observations (PLPs for example), then x can be interpreted as underlying abstract (or pseudo) articulator settings.

2.2 Previous work

Our investigations into LDMs grew out of earlier work in which we recovered phonological feature values from the acoustic signal [12, 13, 11]. We came to two conclusions from that work: 1) speech exhibits asynchronous properties (in our case, the fact that phonological features do not all change value synchronously, e.g. at phone boundaries) and 2) a discrete state model is not the best way to cope with such asynchrony. Our early experiments [14] led us to believe LDMs were worth further investigation. Richmond's thesis work on recovering articulation from acoustics [17, 28, 19, 18, 20] was also encouraging.

Work of others on factorial HMMs [15] *has* tackled problem 1) using discrete state models which factor out the underlying asynchronous processes using parallel Markov chains. This approach is promising but not without problems: the effective state space of the model is the product of the state spaces of the individual Markov chains, and can therefore very quickly explode – sophisticated parameter reduction schemes must be use to make the models tractable.

Our initial investigations built on the pioneering work on segment models by Mari Ostendorf and her group [3, 4, 5, 16]. Other groups have more recently become interested in LDMs, segmental HMMs and related models, e.g. Gales and Rosti in Cambridge [22, 23].

The work we carried out in this project took place in the context of a growing awareness in the speech recognition community that HMM performance has stopped improving and that alternative models must be investigated. The relationships between many of these models under investigation can be most easily seen by expressing them using the *graphical models* formalism – for background and surveys see [22] for example. The appearance of the graphical models toolkit (GMTK) from Bilmes & Zweig [1] typifies this emerging field.

3 Key advances and experimental results

3.1 Theoretical advances

We have performed the first full investigation using recognition, rather than classification or rescoring as in previous studies. Full results can be found in our publications (see References, particularly [7, 6]).

Our models are more sophisticated and general then in the previous or concurrent work of others: for example, Digalakis [3] didn't use any subspace modelling, and Rosti [22, 23] used zero mean noise distributions. We have experimented with more model variants that any other study and found that a model with subspace modelling, dynamic behaviour in the state space and non-zero mean noise distributions performs best. We therefore conclude that there is modelling power in the dynamic aspect of the model, and it is worth further development – details of our plans are given in section 5.

To perform the full search required for recognition, we have developed a number of strategies for reducing computation, most notably the caching of calculations and the use of best-first $A^\star$ time-asynchronous search (implemented with a stack decoder).

3.1.1 Subspace modelling

State-space models such as the LDM model apparently complex observation sequences using relatively simple processes which operate in the state space, along with a projection from the state space up to the observation space. In contrast, models such as HMMs attempt to model directly in the observation space, which can result in more complex and less parsimonious models.

3.1.2 Dynamic modelling

To confirm that the LDM is in fact operating in a dynamic fashion, we compared LDMs with factor analysers (which are like LDMs but with no dynamic process in the hidden space: equation 2 becomes simply $\mathbf{x}_t \sim N(1,0)$). LDMs were significantly more accurate in all experiments.

3.2 Preliminary results using MOCHA data

Our initial investigations used data which exhibits properties thought to be well matched to the modelling abilities of LDMs, namely parameters which are continuous (no discontinuities, such as those present in many acoustic representations) and smooth (low bandwidth).

Work was based on electromagnetic articulograph (EMA) data from a single-speaker, a southern English female (fsew0) from the MOCHA corpus [26]. We experimented with both the normalised EMA, automatically recovered EMA [18] (referred to as rEMA), standard Mel-scale cepstra (12MFCCs + energy, referred to as MELCEP) and a composite set of features similar to that used in [28] derived from EMA, electropalatograph and laryngograph data using linear discriminant analysis which we refer to as LDA. The rEMA data is the output of a single-hidden-layer MLP, trained using a scaled conjugate gradient method, with a skeletonisation algorithm to find the optimum network size, as reported in [18].

Our investigations started with the simplified task of classification (recognition with known segment boundaries, so search is much simpler) and results are shown in table 1. Full recognition results using a phone bigram language model are given in table 2.

Linear dynamical models	
EMA	58%
rEMA	56%
LDA	74%
EMA + MELCEP	76%
rEMA + MELCEP	69%
MELCEP	70%

Table 1: Classification results using MOCHA data (46 phone set).

Linear dynamical models	
LDA	61%
MELCEP	54%
HMM system from [28] using 5500 tied-state triphone models built with HTK	
LDA	63%

Table 2: Phone accuracy recognition results using MOCHA data (46 phone set).

3.3 Major results on TIMIT

We then progressed to the TIMIT corpus [9], performing the standard benchmark task of phone recognition using a simple bigram (phone) language model. We also used a perceptual linear prediction parameterisation (PLP), since this appears to be more smooth and continuous than MELCEPs.

With phone boundaries known, results are shown in table 3. Note that with PLPs, adding deltas makes almost no difference to classification accuracy - from this we infer that the LDM is indeed using dynamic behaviour in the hidden space, hence appending deltas to the observation vectors provides no extra information.

To our knowledge, we have obtained first full recognition results for LDMs on TIMIT. Digalakis [3] used a suboptimal 'split and merge' algorithm. Table 4 shows the best (and latest) result. This is of course still some way behind state of the art results for TIMIT using either HMMs or hybrid neural-net/HMM systems [21], but accuracy is steadily increasing as we further develop our models.

Linear dynamical models - 39 phone set	
MFCC + deltas	72%
PLP + no deltas	68%
Linear dynamical models - 61 phone set	
MFCC + deltas	61%
PLP + no deltas	59%

Table 3: Phone classification results using TIMIT data.

Linear dynamical models	
PLP	55%

Table 4: Phone accuracy recognition results using TIMIT data (39 phone set).

3.4 Analysis of results and models so far

A detailed investigation of the likelihoods accumulated by the LDM in the course of recognising a segment has revealed that it takes a few frames for the dynamic model the "lock on" to the correct trajectory – hence the first few frames of each segment are given rather low likelihoods. We think this is leading to deletion errors, and are investigating solutions. One approach would be to make the hidden state continuous across segment boundaries – this is on our longer-term agenda since it introduces a number of theoretical and practical difficulties. Another simpler approach would be to start the model running a few frames before the (hypothesised) segment start time, but not start accumulating likelihood until the segment actually begins. We are currently using a simple gamma distribution duration model, but with TIMIT there is enough data to estimate a histogram.

Perhaps the most significant limitation of our current system is that training uses the manually-assigned label times. We know this is suboptimal and are currently carrying out the first trials of an embedded training system which uses the previous iteration of models to perform a Viterbi re-alignment of the training data labels. Full EM embedded training may not be be necessary, although we need to perform experiments to confirm this.

3.5 Latest results: sub-phone modelling

One problem with our previous systems is that there is one setting for the LDM parameters per segment (phone). Our most recent experiments have divided each phone (deterministically for now, as in [3]) into a number of sub-phone regions, as shown in table 5.

phone type	regions per phone	phone type	regions per phone
affricates	2	fricatives	1
nasals	2	semivowels & glides	2
silence	1	stop closures	1
stops	3	vowels	3

Table 5: Number of sub-phone regions

This improves classification accuracy compared to the results given in table 3 for PLPs with no deltas to 62% (61 phones) and 70% (39 phones). This work is a precursor to some planned future research where we intend to both learn how many sub-phone (or other unit) regions to use and to control switching between regions with a Markov model. The topology of this Markov model may include branches, and not just linear chains.

3.6 Software

We have created a set of software tools, based on the same library of software as our FESTIVAL [2] speech synthesis toolkit. Along with our stack decoder, this gives us a very flexible framework for exploring future extensions of the models, particularly those outlined below. The tools allow rapid prototyping of new model types along with very straightforward use of those models in the decoder, since the stack decoder completely decouples the acoustic models from the language model and lexicon.

Once these tools are more stable and documented, we intend to freely distribute them under the same terms as FESTIVAL.

4 Collaboration

During the course of the project, we have been discussing ideas with various other groups, most notably Alan Wrench of Queen Margaret University College and members of the Institute for Adaptive and Neural Computation here at Edinburgh University. From these discussions we have come to a deeper understanding of the properties of LDMs and now feel that they have a great deal of potential for speech recognition. Our plans for future research are given in the following section.

4.1 Student projects

Joe Frankel (Ph.D) Joe's thesis work (submission expected Autumn 2002) has been investigating linear dynamic models for ASR and thus addresses the core issues in this project. Joe was co-author with the principal investigator (King) on the key publications [8, 7, 6].

Korin Richmond (Ph.D) Korin's thesis work (submitted late 2001 and being examined 29th April 2002) used neural networks and mixture density networks (whose output is a mixture-of-Gaussians PDF over articulator positions) to perform acoustic→articulatory mapping (the so-called inversion problem), and the recovered articulation mentioned earlier in this report was produced by his system. Korin has subsequently been funded by this grant, and his work included the pilot study preparing for using LDMs in speech synthesis (see below).

Fiona Couper (M.Sc) Starting 1st May 2002, Fiona will be carrying out her M.Sc. dissertation project using our new models. Her work will serve as the pilot study for two future projects. The first is Fiona's EPSRC-grant-funded Ph.D. which will investigate the learning of a (non-phonetic) unit **inventory** from data. The second future project is the subject of a grant proposal shortly to be submitted to EPSRC, where we propose to extend the sub-phone modelling work described above and explore methods for learning the **topology** and **parameters** of a finite state switching process for controlling the LDM parameters. This will ultimately link in with Fiona's Ph.D. work to make a general framework for learning both the switching process and the unit inventory from data.

5 Summary and future direction

This project was funded under the Realising Our Potential Awards (ROPA) scheme and carried out some novel work exploring the potential of a new type of acoustic model for ASR. We have demonstrated that linear dynamic models show promise for ASR and have identified a number of future directions for this research to take which we are already actively pursuing. The main thrust of our future work will be in adding a (Markov) switching process which sets the parameters of the LDMs. By having more than one state in series, this gives us sub-phone (or whatever unit we choose) modelling. By having more than one state in parallel this gives us the sort of modelling power that would be provided by Gaussian mixture distributions in equations 1 and 2. The models will then become switching linear dynamical systems [24, 25, 10]. We are also investigating ways of making the hidden state continuous across model boundaries without making the models intractable. Visit the CSTR website to get the latest developments: `www.cstr.ed.ac.uk`. *Items in the bibliography marked ★ are associated with this project.*

Use in speech synthesis One unexpected use we have found for our models is in concatenative speech synthesis. Two of the main problems facing state-of-the art unit-selection synthesisers are those of join cost and join smoothing. We have submitted an EPSRC proposal[1] to use LDMs to simultaneously compute join cost (a measure of perceptual discontinuity) and to smooth joins (currently done by spectral interpolation) by learning an underlying "articulatory-like" representation of the speech signal.

[1] Reviewers comments were recently received and are very favourable - this proposal is going to the panel on May 7th.

References

[1] J. Bilmes and G. Zweig. The graphical models toolkit: An open source software system for speech and time-series processing. In *Proc ICASSP 2001*, 2002.

[2] Alan Black, Paul A. Taylor, Richard Caley, and Robert Clark. The Festival speech synthesis system. Available from the Centre for Speech Technology Research `http://www.cstr.ed.ac.uk/projects/ festival`, 1997-2000.

[3] V Digalakis. *Segment-based stochastic models of spectral dynamics for continuous speech recognition*. PhD thesis, Boston University Graduate School, Boston, 1992.

[4] V Digalakis and M Ostendorf. Fast algorithms for phone classification and recognition using segment-based models. *IEEE Trans. on Speech and Audio Processing*, 40(12):2885–2896, 1992.

[5] V. Digilakis, J. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE Trans. Speech and Audio Processing*, 1(4):431–442, October 1993.

[6] ★ Joe Frankel and Simon King. ASR - articulatory speech recognition. In *Proc. Eurospeech*, September 2001.

[7] ★ Joe Frankel and Simon King. Speech recognition in the articulatory domain: investigating an alternative to acoustic hmms. In *Proc. Workshop for Innovations in Speech Processing (WISP)*, pages 37–46, Stratford-on-Avon, April 2001.

[8] ★ Joe Frankel, Korin Richmond, Simon King, and Paul Taylor. An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proc. ICSLP*, Beijing, 2000.

[9] J. S. Garofolo. *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

[10] Z. Ghahramani and G. Hinton. Switching state-space models, 1996.

[11] ★ Simon King, Joe Frankel, Paul Taylor, and Korin Richmond. Speech recognition via phonetically featured syllables. *Phonus*, 5:15–34, 2000.

[12] Simon King, Todd Stephenson, Stephen Isard, Paul Taylor, and Alex Strachan. Speech recognition via phonetically featured syllables. In *Proc. ICSLP '98*, pages 1031–1034, Sydney, Australia, December 1998.

[13] Simon King and Paul Taylor. Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language*, 14:333–353, 2000.

[14] ★ Simon King and Alan Wrench. Dynamical system modelling of articulator movement. In *Proc. ICPhS 99*, pages 2259–2262, San Francisco, August 1999.

[15] H. J. Nock and S. J. Young. Loosely coupled HMMs for ASR. In *Proc. ICSLP*, Beijing, 2000.

[16] M. Ostendorf, V. Digalakis, and O. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(5):360–378, Sept. 1996.

[17] ★ Korin Richmond. Estimating velum height from acoustics during continuous speech. In *Proc. Eurospeech*, volume 1, pages 149–152, Budapest, Hungary, 1999.

[18] ★ Korin Richmond. *Estimating Articulatory Parameters from the Acoustic Speech Signal*. PhD thesis, The Centre for Speech Technology Research, Edinburgh University, 2001. Submitted Oct. 2001.

[19] ★ Korin Richmond. Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech. In *Proc. Workshop for Innovations in Speech Processing (WISP)*, pages 259–277, Stratford-on-Avon, April 2001.

[20] ★ Korin Richmond, Simon King, and Paul Taylor. Modelling the uncertainty of recovring articulation from acoustics. *Computer Speech and Language*, forthcoming, 2002.

[21] Tony Robinson. The application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2), March 1994.

[22] A-V.I. Rosti and M.J.F Gales. Generalised linear Gaussian models. Technical Report CUED/F-INFENG/TR.420, Cambridge University Engineering, 2001.

[23] A-V.I. Rosti and M.J.F Gales. Factor analysed HMMs. In *ICASSP*, May 2002.

[24] R. H. Shumway and D. S. Stoffer. Dynamic linear models with switching. *Journal of the American Statistical Association*, 86(415):763, 1991.

[25] A. Stolcke and S. Omohundro. Best-first model merging for hidden markov model induction. Technical Report TR-94-003, ICSI, Berkeley, 1994.

[26] A. A. Wrench and W. J. Hardcastle. A multichannel articulatory speech database and its application for automatic speech recognition. In *Proc. 5th Seminar on Speech Production*, pages 305–308, Kloster Seeon, Bavaria, May 2000.

[27] Alan Wrench. A new resource for production modelling in speech technology. In *Proc. Workshop for Innovations in Speech Processing (WISP)*, Stratford-on-Avon, April 2001.

[28] ★ Alan Wrench and Korin Richmond. Continuous speech recognition using articulatory data. In *Proc. ICSLP*, Beijing, 2000.