DYNAMICAL SYSTEM MODELLING OF ARTICULATOR MOVEMENT

Simon King[†] Alan Wrench[‡]
† *Centre for Speech Technology Research, University of Edinburgh, UK*
‡ *Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, UK*
Simon.King@ed.ac.uk *www.cstr.ed.ac.uk*

ABSTRACT

We describe the modelling of articulatory movements using (hidden) dynamical system models trained on Electro-Magnetic Articulograph (EMA) data. These models can be used for automatic speech recognition and to give insights into articulatory behaviour. They belong to a class of continuous-state Markov models, which we believe can offer improved performance over conventional Hidden Markov Models (HMMs) by better accounting for the continuous nature of the underlying speech production process – that is, the movements of the articulators. To assess the performance of our models, a simple speech recognition task was used, on which the models show promising results.

1. INTRODUCTION

Our investigation of dynamical system models is motivated both by an interest in new models for speech recognition, and by the availability of new articulatory measurement data.

For speech recognition, we are investigating alternatives to Hidden Markov Models (HMMs) in which speech is generally seen as a sequence of phones, each of which is typically modelled by a three state model – three regions in which the observation is assumed to remain constant[1]. We propose that models which take more account of the continuous nature of the speech production process should provide better modelling (and therefore recognition) of the speech signal. One such model is the Continuous State Markov model (CSMM). Our models are a type of CSMM in which constraints are placed on the state behaviour in the form of a linear dynamical system. Our models belong to a larger class of models [4], which includes HMMs, and dynamic segment models [2], for example.

These models will find applications in articulatory modelling as well as speech recognition. Ultimately, we envisage models which will be able to (probabilistically) recover articulatory trajectories, or gestures [3, 7], from acoustic data. Dynamical system models of a similar form to the models here have been successfully used for intonation modelling [6], and have shown some potential for speech recognition [1] – although the task used in [1] was probably too difficult for testing a novel acoustic model.

New data, described in section 1.1 below, is becoming available which allows us to train models of articulator dynamics. We stress that these models are quite different from more literal mass-spring models or finite element simulations of the articulators. The models described in this paper are of articulatory *measurements*: two-dimensional coordinates of selected *points* on the articulators.

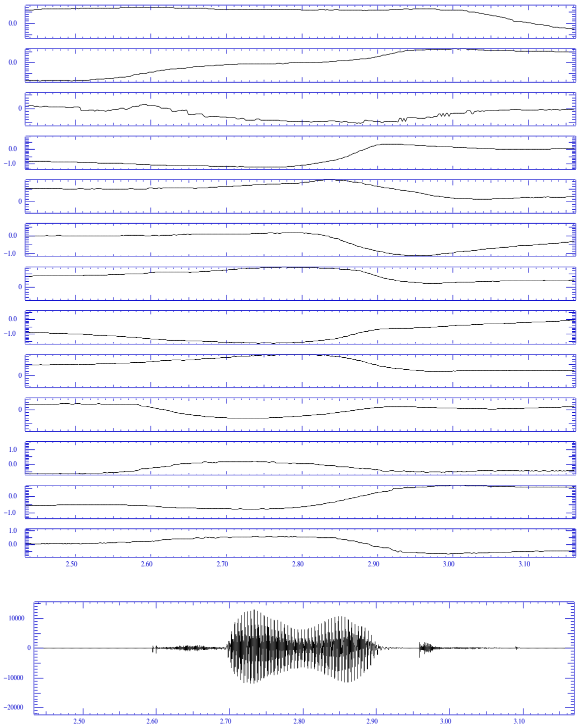

Figure 1: Example EMA data for the word "pod" . Tracks, reading from the top, are velum height (y), upper lip y, upper lip x, tongue tip y, tongue tip x, tongue dorsum y, tongue dorsum x, tongue blade y, tongue blade x, lower lip y, lower lip x, lower incisor y, lower incisor x. The y coordinate is vertical (increasing y means upward movement), and the x coordinate is horizontal (increasing x means forward movement).

1.1. Data

Physical measurements of articulatory behaviour have been recorded using a Carstens Electromagnetic Articulograph (EMA) system[2]. The data comprise two-dimensional co-ordinates in the midsaggital plane of 7 selected points on the articulators (3 tongue, 2 lips, jaw and velum), each recorded 500 times per second. For this study, a simple dataset was recorded, consisting of sixteen syllables, each repeated up to sixteen times by a single speaker. The syllables are all CVC patterns with C chosen from {d,p} and V chosen from {a,e,i,o}. A total of 248 tokens were recorded, of which a randomly chosen 64 were set aside for testing, leaving 184 training tokens. The velum position was represented by only its y coordinate (that is, the vertical coordinate), all other points have both x and y coordinates; thus the articulatory measurement data is 13 dimensional. An example from this data set is show in figure 1.

2. DYNAMICAL SYSTEM MODELS

2.1. Form of the model

The type of systems we are using to model the data are simply described by equations 1 and 2.

$$x_t = Fx_{t-1} + w_t \tag{1}$$

$$y_t = Hx_t + v_t \tag{2}$$

where x_t is the state of the system at time t, y_t is an observation, F describes the dynamical behaviour of the state, and H maps the state space to the observation space. w_t and v_t are uncorrelated noise with Gaussian distributions, with means and variances μ_w, Q_w and μ_v, Q_v respectively. In all our experiments, F and H are constant over time. See [4] for an excellent overview of this and related models as applied to speech recognition. The Markov property of these models comes from equation 2: the observation y_t depends only on the current state x_t (plus some uncorrelated noise). All our current experiments are with linear models (described by F) and linear transformations between hidden and observation spaces (described by H).

Equations 1 and 2 describe a single model. For speech recognition, we need a different model for each unit (syllable, for example) we wish to recognise. Thus, each model is segment-specific and has its own F, H, μ_w, Q_w, μ_v and Q_v. Some parameters can be shared between models – see section 2.4.1 below. Our system uses the syllable as the unit, rather than the more usual phone, as in [1].

2.2. Model configurations

We considered two forms for the models. In the first, the articulators themselves are treated as a linear dynamical system – only equation 1 is used, with y substituted for x. In the second, an "internal model" [5] is used – described by equation 1; the articulator positions are a linear transform of the state of this "internal" model – described by equation 2. In this system, x is the *hidden state* of the model; it cannot be observed. This second configuration has the advantage that the dimension of the hidden state space can be different from the dimension of the articulatory state space (which is determined by the number of articulators being modelled). In all our experiments, we used all 13 articulatory measurements (see figure 1); future experiments will explore ways of reducing the data dimensionality *a priori*, and of introducing Electro-palatograph (EPG) data.

2.3. Interpretation

Although we make no claims that the hidden space in the second of our model configurations relates to anything physical, it is interesting to examine how the dimension of the hidden space – that is, the number of free variables required to describe the state of the system - affects performance. At this time, our measure of the models' performance is accuracy on a simple classification task, as described in section 3 below. The best accuracies for systems with various hidden space dimensions are shown in figure 4.

2.4. Training

The models were trained by a simple Expectation-Maximisation algorithm (see, for example, [4] for mathematical details). This is a two step algorithm: in the first step (the E step), statistics are accumulated over the training tokens using the existing model parameters; in the second step (the M step) the model parameters are updated using those statistics.

2.4.1. Parameter tying In equations 1 and 2, the model parameters are different for each model. It is possible to share some of the parameters between models – we call this *parameter tying*. One interesting possibility is to have a single H matrix and observation noise process v for all models. This means that all models share the same hidden space: H, Q_v and μ_v are tied across all models. Implementation of parameter tying is trivial in the EM framework - parameters to be tied simply pool statistics during the E step.

2.4.2. Number of training iterations The EM algorithm only guarantees to increase the likelihood of the training data, given the models, each iteration. It is therefore possible to over fit the training data, and lose performance on the test data. Training can be stopped when the models' performance on a validation set is maximised. At this time, we have insufficient data to use a validation set, and use the test data itself to determine when training should be stopped (that is, how many iterations of the EM algorithm to perform).

3. EXPERIMENTS

3.1. A simple task

The measure we chose to assess the performance of our models is a simple syllable classification task. As described in section 1.1 above, the dataset comprised 16 different syllables (dap, pop, pid, and so on). The task was to classify (from endpointed data) unknown syllables as one of the 16 types. All types were equally likely, so the chance level of success would be 1 in 16, or 6%.

3.2. Acoustic-only observations

Clearly, articulatory measurements are not a practical proposition for automatic speech recognition; in the real world, we only have acoustic observations. One of our hypotheses is that using an articulatory state space for the dynamical system model will lead to improved performance even when the state space is hidden. In other words, if the state space is articulatory during training, we might expect improved recognition from acoustic-only data.

system	observations	state space
A	articulatory	articulatory
B	articulatory	hidden
C	LPC	LPC
D	LPC	hidden
E	LPC	articulatory*

* observable in training, but hidden during testing.

Table 1: The various systems

3.3. Experimental models

A number of systems were trained, as shown in table 1. Section 2.2 gave details of the configurations: systems A and C have observable state spaces, and systems B and D have a hidden space which can be optionally shared amongst all models, via parameter tying. System E uses a state space initialised from system A. Acoustic data was parameterised as Linear Prediction Coefficients (LPC) at the same frame rate (500Hz) as the EMA data[3].

system		accuracy
A		70%
B	non-tied H	**89%**
	tied H	81%
C		66%
D	non-tied H	59%
	tied H	41%
E	constant H	31%
	non-tied H	30%
	tied H	36%

Table 2: Classification test results. In systems B and D, the hidden state space dimension was varied to give the best result.

3.4. Results

3.4.1. Linear dynamical system model Models with observable state spaces were trained on both EMA and LPC data. The results in table 2 (systems A and C) indicate that the EMA data can be successfully modelled as a linear dynamical system, with a classification rate of 70% on the test set. This is an encouraging result and indicates that, firstly, our proposed system with a hidden articulatory state and only acoustic observations has some chance of success, and secondly, that even the very simple model used here fits the data remarkably well.

3.4.2. Hidden linear dynamical system model Models with hidden state spaces were trained on EMA and LPC data. The dimension of the hidden space was varied. Additionally, versions of both systems in which the hidden space was shared amongst models (by tying H, Q_v and μ_v across all models – indicated in table 2 simply as *tied H*) were trained. In table 2, the results are given for models using EMA observations (system B) and LPC observations (system D). The best system was the hidden state model with EMA observations, with a test set accuracy of 89% (hidden state order was 7, trained for 4 iterations of EM).

Our experiments with parameter tying did not generally show improvements in accuracy. Simply tying H, Q_v and μ_v across models may not be sufficient. This only ensures that the state space (that is, the space of x) has a fixed relation to the observation space. We could go further, and say that the *dynamical behaviour* of x should be the same in all models, since all speech is produced by the same "internal model" operating in the same hidden space – see section 4.1.

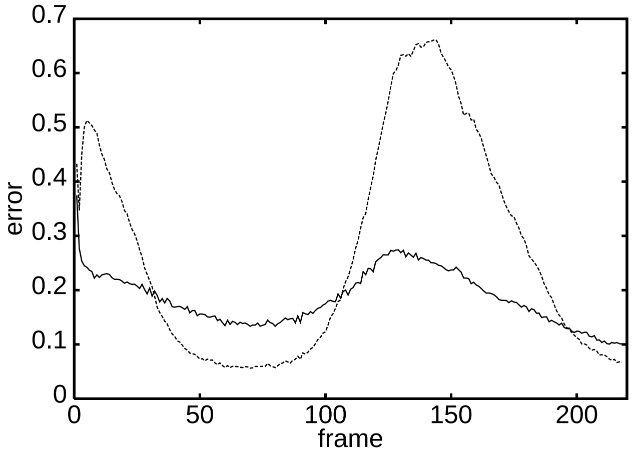

Figure 2: Error plot for test token "ded". Solid line is for model "ded"; dashed line is for model "pep". Frame rate is 500Hz.

Figure 2 shows the mean squared modelling error for one of the test tokens (the syllable "ded") using models for "ded" and for "pep". The model was type B, with non-tied H, the hidden state space dimension was 7 and the models were trained for 4 iterations of EM (each iteration uses each training token once). The error plotted is the squared difference between the observation and the model prediction, averaged across the 13 EMA channels. The correct model has a lower total error over the token: it matches the data better than the incorrect model. The regions of greatest error for the incorrect model are, unsurprisingly, in the onset and coda where the observed "d" does not match the model's "p". Figure 3 shows the effect of number of training iterations on test set accuracy.

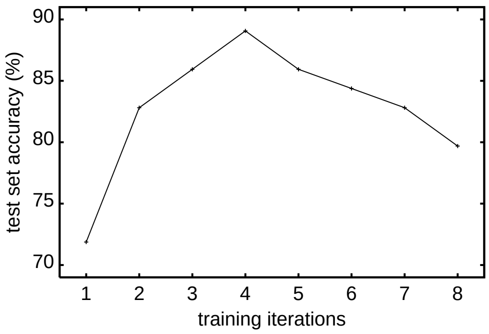

Figure 3: Test set accuracy vs. training iterations for hidden linear dynamical system model of order 7 with EMA data as observations.

System E was not successful. We tried three variations: keeping H, Q_v and u_v the same as system A (*constant H* in the table) and retraining H, Q_v and u_v in both tied and un-tied versions. The most likely reason for this failure is the non-linear relationship between articulatory and LPC spaces - see section 5.

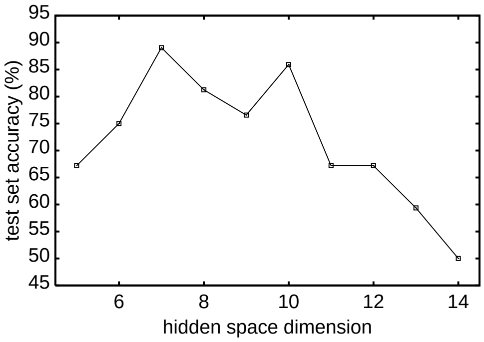

Figure 4: Effect of hidden space dimensionality on accuracy.

4. ANALYSIS

4.1. Interpretation of model parameters

The dynamical behaviour of the (hidden) state of the system is described by F, with the variability about this predicted behaviour represented by the Gaussian noise process w. Since we chose to make F segment-specific (and optionally tie H over all models) F can be seen as having two distinct components: one is the hidden dynamical system common to all models sharing the same hidden space – in other words, the underlying production mechanism; the other is a segment-specific component representing the state behaviour when a particular speech unit (syllable, in this case) is produced. The segment-specific part of F can be interpreted as the *gesture* required to produce the speech unit being modelled. The possible complexity of this gesture is governed by the form and size of F.

In future work we intend to explicitly separate these two components of F, which will allow us to tie the dynamical behaviour component across all models – based on the fact that the underlying mechanism of speech production is a constant across speech segments. This leads to two choices for the segment-specific component of the models: articulatory targets, or gestures. We are currently investigating this topic, with the intention of defining an inventory of articulatory units in terms of targets and/or gestures.

5. CONCLUSIONS

The most successful models were the hidden state models with articulatory observations. The models with an articulatory hidden state and acoustic observations were not successful. From this we draw two conclusions: the articulatory data is well modelled by a linear system – in other words, the underlying physical mechanism of speech production is sufficiently linear not to require non-linear models; however, the acoustic observations (LPC in this case) do not have a linear relationship to the articulatory parameters. This is not really surprising. The direction of our future work is therefore to keep the underlying (hidden state space) models linear, but explore non-linear mappings to the observation space. Equation 2 will be replaced by a non-linear relationship – a neural network, for example. Neural networks mapping between acoustic and articulatory spaces are a current research topic at CSTR.

ACKNOWLEDGEMENTS

Simon King is funded by EPSRC *Realising Our Potential Award* number GR/L59566 and Alan Wrench is funded by EPSRC grant number GR/L78680.

Notes

[1] The observation may include dynamic parameters known as delta coefficients

[2] Facility located at Queen Margaret University College, Edinburgh; see `http://sls.qmced.ac.uk/research/EMA/ema.htm`

[3] We avoided the most common parameterisation of the acoustic signal used in speech recognition – Mel-scale Cepstral coefficients – because this has a very non-linear relationship to vocal tract parameters

REFERENCES

[1] John S. Bridle, Li Deng, Joseph Picone, Hywel B. Richards, Jeff Ma, Terri Kamm, Micheal Shuster, Sandi Pike, and Roland Regan. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. In *CLSP/JHU Summer Workshop on Language Engineering*, Baltimore, 1998. Johns Hopkins University.

[2] Wendy J. Holmes and Martin J. Russell. Speech recognition using a linear dynamic segmental hmm. In *Proc. Eurospeech-95, Madrid*, pages 1611–1614, Sept. 1995.

[3] Bernd Kröger, Georg Shröder, and Claudia Opgen-Rhein. A gesture-based dynamic model describing articulatory movement data. *J. Acoust. Soc. Am.*, 98(4):1878–1889, October 1995.

[4] 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.

[5] Gordon Ramsay. Stochastic calculus, non-linear filtering, and the internal model principle: Implications for articulatory speech recognition. In *Proc. ICSLP '98*, pages 2987–2990, Sydney, Australia, December 1998.

[6] Ken Ross and Mari Ostendorf. A dynamical system model for recognising intonation patterns. In *EUROSPEECH 95*, pages 993–996, 1995.

[7] E. Saltzman and K. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.