AN AUTOMATIC SPEECH RECOGNITION SYSTEM USING NEURAL NETWORKS AND LINEAR DYNAMIC MODELS TO RECOVER AND MODEL ARTICULATORY TRACES.

Joe Frankel *Korin Richmond* *Simon King* *Paul Taylor*

Centre for Speech Technology Research,
University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK
www.cstr.ed.ac.uk joe@cstr.ed.ac.uk

ABSTRACT

We describe a speech recognition system which uses articulatory parameters as basic features and phone-dependent linear dynamic models. The system first estimates articulatory trajectories from the speech signal. Estimations of x and y coordinates of 7 actual articulator positions in the midsagittal plane are produced every 2 milliseconds by a recurrent neural network, trained on real articulatory data. The output of this network is then passed to a set of linear dynamic models, which perform phone recognition.

1. MOTIVATION

Hidden Markov Models (HMMs) have dominated automatic speech recognition for at least the last decade. The model's success lies in its mathematical simplicity; efficient and robust algorithms have been developed to facilitate its practical implementation. However, there is nothing uniquely speech-oriented about acoustic-based HMMs. Standard HMMs model speech as a series of stationary regions in some representation of the acoustic signal. Speech is a continuous process though, and ideally should be modelled as such. Furthermore, HMMs assume that state and phone boundaries are strictly synchronized with events in the parameter space, whereas in fact different acoustic and articulatory parameters do not necessarily change value simultaneously at boundaries.

We propose that modelling speech in the articulatory domain will inherently account for for the underlying processes of speech production, such as coarticulation, and will therefore offer improvements in recognition performance. Because the trajectories evolve smoothly over time, we chose to model them using a Linear Dynamic Model (LDM), extending the approach used in [2]. We have had access to real articulatory data, which has been used to train a neural network mapping from acoustic to articulatory domains. Both original and recovered data have been modelled.

1.1. Data

The data consisted of a corpus of 460 TIMIT sentences for which parallel acoustic-articulatory information was recorded using a Carstens Electromagnetic Articulograph (EMA) system (this facility is located at Queen Margaret University College, Edinburgh, see `sls.qmced.ac.uk`). Sensors were placed at three points on the tongue (tip, body and dorsum), upper and lower lip, jaw and also the velum. Their position in the midsagittal plane was recorded 500 times per second and the acoustic signal sampled with 16 bit precision at 16 kHz. 30% of the sentences were set aside for testing, and 70% used for training. The data was labelled using an HMM based system. Flat-start monophone models were forced-aligned to the acoustic data from a phone sequence generated from a keyword dictionary.

2. AUTOMATIC ESTIMATION OF ARTICULATORY PARAMETER VALUES

Many approaches have been tried during the long history of research into acoustic to articulatory inversion. The development of articulography technologies, such as electromagnetic articulography (EMA) have enabled the use of machine learning techniques in conjunction with real human data, for example [1], [3]. In the work we present, recurrent neural networks are trained to perform the inversion mapping. The test data was split equally into test and validation sets for training the networks.

2.1. Data processing

The raw acoustic and articulatory data is processed for use with the neural network. Silence is removed from the beginning and end of each recording. During silent stretches, the mouth may take any position and this would adversely affect network learning. Filterbank analysis of the waveform gives 16 filterbank coefficients for 16ms frames every 8ms. The EMA tracks are resampled to match this 8ms frame shift. The data is normalised so that network input (filterbank coefficients) lie in the range [0.0, 1.0] and network output (EMA data) lies in the range [0.1, 0.9].

Similar to [3], a large input "context" window of 25 acoustic frames (400 input units, as there are 16 filterbank coefficients for each frame), two hidden layers, and a single output unit for each articulator track was used. A key difference was the introduction of recurrence by adding context units for the second hidden layer. It has been found this generally decreases training time, and also gives smoother output trajectories from the trained network [4]. However we found that further smoothing using a 6 point moving average window gave an improvement in results.

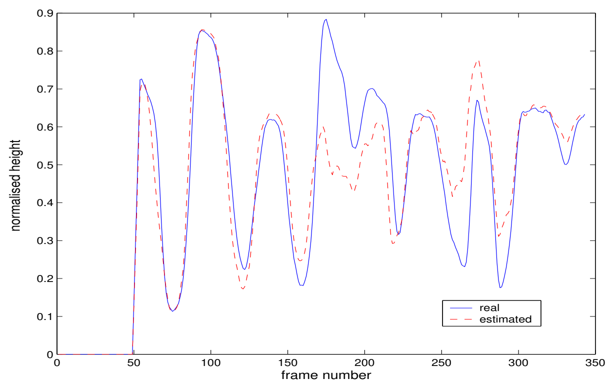

Figure 1: Actual and automatically estimated articulatory parameter (tongue tip height). "A large household needs lots of appliances"

Articulator	RMSE (mm)	Mean correlation
Upper lip X	1.1 (21%)	0.43
Upper lip Y	1.2 (17%)	0.58
Lower lip X	1.4 (22%)	0.42
Lower lip Y	2.6 (16%)	0.72
Lower incisor X	1.1 (22%)	0.44
Lower incisor Y	1.1 (14%)	0.76
Tongue tip X	2.5 (15%)	0.72
Tongue tip Y	2.5 (13%)	0.80
Tongue body X	2.0 (14%)	0.75
Tongue body Y	2.0 (13%)	0.80
Tongue dorsum X	2.1 (16%)	0.73
Tongue dorsum Y	2.7 (19%)	0.67
Velum X	0.8 (22%)	0.67
Velum Y	1.3 (27%)	0.68

Table 1: Automatic estimation of articulatory parameters on unseen data. RMSE is expressed in mm and as a percentage of the total range of movement for that articulator.

2.2. Results

Figure 1 shows an example from the test set for one articulatory parameter. Qualitatively, this shows that an accurate mapping is achieved. Two measures that have been used in the past are root mean square error, an indication of the distance between two trajectories, and the product moment correlation coefficient, an indication of similarity in "shape". Table 1 gives quantitative results: RMSE is given both in millimetres and as a percentage of the total range of movement for each articulator. However, these two measures are of limited use, as for the purposes of speech recognition, we are not necessarily interested in recovering the articulation as accurately as possible from the acoustics. There are many problems inherent in this, such as the critical / non-critical nature of the articulators (see section 4.1), and the fact that one sound can be produced by many articulatory configurations. A much more suitable performance measure for the network is that of phone classification score.

3. LINEAR DYNAMIC MODELS

The second stage of the system revolves around modelling the articulatory trajectories. For this task we have chosen a linear dynamic model described by the following pair of equations:

$$x_t = Fx_{t-1} + w_t \quad (1)$$

$$y_t = Hx_t + v_t \quad (2)$$

with y_t denoting the observation and x_t the hidden state variable of the system at time t. The basic premise of the model is that there is some underlying dynamic process which can be modelled by equation 1. Evolution from one time-frame to the next is described by a linear transformation F and the addition of some noise, $w_t \sim N(\mu_w, Q_w)$. The complexity of the motion is encapsulated in the dimensionality, for example a 1 dimensional state space would allow exponential growth or decay with an overall drift (μ_w 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 represent realisations of this unseen dynamical process; a linear transformation with the matrix H and the addition of more noise, $v_t \sim N(\mu_v, Q_v)$ (equation 2) relate the two. The 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 parameterisation chosen to represent it. This parameterisation is not necessarily optimal, in fact the system is best described using fewer degrees of freedom than originally present in the data. The models are segment-specific, with one set of parameters H, F, Q_v, Q_w, μ_v, and μ_w describing the articulatory motion for one unit of speech, although it is possible to share parameters between models. Segments used so far have been phones

The model can be thought of as a continuous state HMM [5]. Having a state which evolves in a continuous fashion, both within and between segments makes it an appropriate choice to describe speech. Attempts to directly model speech in the *acoustic* domain using LDMs have been made, however the defining feature of these models is that they are able to model smoothly varying (but noisy) trajectories. This makes them ideally suited to describing articulatory parameters. Furthermore, the asynchrony between the motion of different articulators is absorbed into the system, and the critical versus non-critical nature of articulators (see below) is captured in the state to observation mapping covariance Q_v. Lastly, parameter estimation is made much simpler through having a linear mapping between state and observation spaces, which is a reasonable assumption for observations in the articulatory domain.

3.1. Training

The Expectation Maximisation (EM) algorithm was used to train the models. [5]. The EM algorithm only guarantees to increase the likelihood of the *training* data. Over-training occurs quite rapidly as figure 2 shows. In this case, after 6 iterations classification performance starts to drop off as the models learn the specific behaviour of the training data, rather than the more gen-

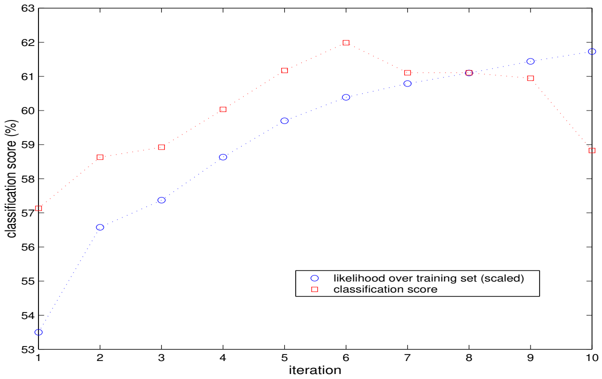

Figure 2: classification scores peak (here after 6 iterations) when the models start to become over-trained

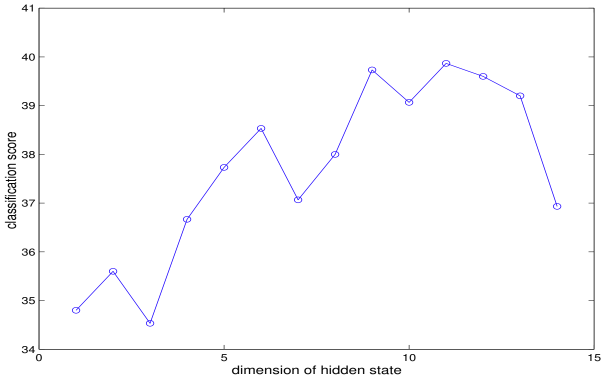

Figure 3: raw classification score against state dimension for a validation set consisting of 20 utterances. Models were trained and tested on real ema data only.

eral characteristics of each phone.

3.2. Classification

For each segment to be classified, the probability of the observations given the model parameters for each phone model is calculated. The hidden state prevents direct computation, so an idealised state sequence is generated using the posterior predictive distribution of the state variable given the observations, $\mathbf{x_t}|\mathbf{Y_t}$, where $\mathbf{Y_t} = (\mathbf{y_1}, \ldots, \mathbf{y_t})^T$. This is then used to arrive at the appropriate probability. The model likelihoods are then rescored using a bigram language model and finally ranked.

3.3. Feature set

Using a feature set consisting only of articulatory parameters lacks certain information. For instance making a voiced/voiceless classification, or indeed spotting silences is compromised by the lack of voicing information and energy. We have experimented with augmenting the feature set to use other parameters: Mel-scale cepstral coefficients, energy, (acoustic waveform) zero crossing rate, and a voiced/voiceless classification. High values for the zero crossing rate signify noise, ie frication and low values are found in periodic, ie voiced sections of speech. The voiced/voiceless decision was made from a laryngograph trace using a pitchmarking tool. The different experimental configurations are given in Table 2. Both real and simulated ema traces were used, and feature set E, just cepstral coefficients were included for comparison.

A	ema
B	ema + zero crossings + voicing
C	ema + 12 cepstra + energy
D	ema + 12 cepstra + energy + zero crossings + voicing
E	12 cepstra + energy

Table 2: summary of the different feature sets used for experimentation.

4. RESULTS

There was some degree of flexibility in the dimensionality chosen for the state space, in fact there did not seem to be a clear indicator of an 'optimal' dimensionality. Figure 3 shows how raw (no language model) classification scores are affecting by varying the state dimension. The feature set in use here is the real ema data. Anything between 4 and 13 degrees of freedom produces comparable results, with 11 producing the highest score. This suggests that only a few degrees of freedom are able to model the data, adequately and at present there is enough training data to learn some extra parameters for a minor improvement in results.

data	feature set	accuracy
real articulatory	A	49%
	B	59%
	C	68%
	D	69%
simulated articulatory	A	36%
	B	38%
	C	50%
	D	49%
acoustic	E	65%

Table 3: Classification results based on real and simulated articulatory data.

Table 3 summarises the results of experimentation with the system. The number of training iterations and state dimensions was optimised for each system, and the best result obtained quoted. Models trained on simulated articulatory parameters needed more iterations, generally 11-13, of the EM algorithm to converge than their real-data trained counterparts where 3-4 was sufficient. Training and testing models on the real articulatory data produced a classification score of 49%. Augmenting the feature set to also include zero crossing rate and the voiced/unvoiced decision gave a 10% improvement with a result of 59%. Adding then the 12 cepstra and energy to the real ema data, gave the result of 68%, and further adding zero crossing rate and the voicing decision yielded 69%. Performing the same tasks using the automatically esti-

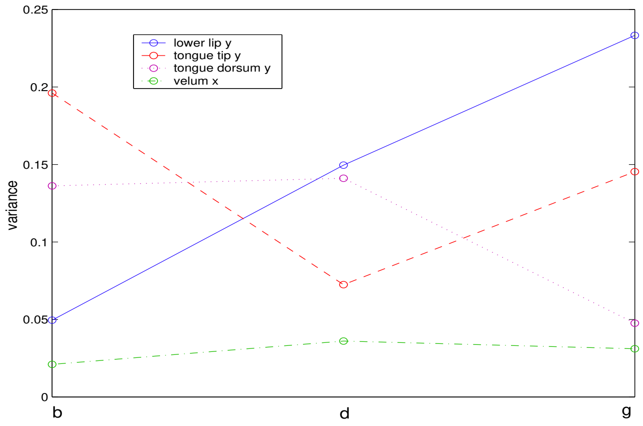

Figure 4: variances put on the projection from state to observation space for normalised, real ema data on segments /b/, /d/, and /g/

mated articulatory parameters gave a drop in performance. 36% and 38% were the scores based on using the articulatory data only and then including the zero crossing rate and voicing decision. The inclusion of the 12 cepstra and energy gave a result of 50% and then further adding zero crossing rate and voicing gave 49%. Training and testing models on just cepstra and energy gave a result of 65%.

4.1. Critical versus non-critical articulators

Papcun *et al* [3] reported that the movements of articulators critical to the production of a segment have a greater range and are less variable than the movements of non-critical articulators. For instance, the lips and velum have a fundamental role in producing a /p/, and would be termed *critical articulators*, while the movements of the tongue are far less important. Examination of the parameters of our trained models shows evidence of this effect. For example, figure 4 shows variance terms for selected articulatory parameters from trained models of the three voiced oral stops in English. These variances are the noise terms associated with the transformation from the hidden space to the observation space, and can be interpreted as an indication of the relative criticality of the articulators for a given phone model. If we first consider the velum, which is an articulator that all three phones have in common as critical, we see that the variance of its movement is uniformly relatively low for the three models. However, the picture is different for other articulators: for the /b/ model, the lower lip has the lowest variance; for the /d/ and /g/ models, it is the tongue tip and tongue dorsum respectively that show the lowest variance. In short, non-critical articulators exhibit higher variance, and lower variances are learned for more critical articulators. Both effects are useful in characterising and distinguishing segments.

Not only are these findings consistent with the notion of critical articulators, they potentially also offer clues to the nature of the acoustic-to-articulatory mapping necessary for the speech recognition system. Ultimately, recovering all articulation perfectly all the time need not be the goal for the inversion mapping.

5. DISCUSSION

Firstly, we would like to raise some points relevant to the system as used with just articulatory information. Training and classification have been performed on data forced-aligned using an HMM based system, which is clearly suboptimal. It is likely that a segmentation based on acoustic information is not the same as one based on a system using articulator positions; indeed there is asynchrony between changes in articulatory gestures and HMM-produced phone boundaries. The network-recovered traces display many of the same features as their real counterparts, but often slightly out of synchronisation. It is hoped therefore that moving away from acoustic-based segmentation will produce a marked improvement in the network-output based classification results. This problem will be solved by performing full recognition, rather than classification, and by embedded training (iterative forced alignment followed by model retraining).

In our opinion, the most important issue is the choice of unit the system should use. We will investigate alternative units which reflect the nature of the articulatory data. Phone-based systems typically use a large number of context-dependent models, which leads to elaborate parameter tying schemes to make training on limited data feasible. We envisage a model which works *with* coarticulation rather than treat it as a problem.

A practical speech recognition system clearly cannot use real articulatory data. We are using the data as a development tool, because of the useful properties it possesses: smoothly changing trajectories, explicit coarticulation, and so on. As the recogniser grows in scale, the articulatory aspect of the system will be reduced to that of a latent, or hidden variable, and the two parts of the system will be trained together.

6. ACKNOWLEDGEMENTS

JF and KR are funded by EPSRC studentships. SK is funded by EPSRC *Realising Our Potential Award* number GR/L59566.

7. REFERENCES

1. J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *J. Acoust. Soc. Am.*, 100(3):1819–1834, September 1996.
2. S. King, A. Wrench. Dynamical system modelling of articulator movement. *Proc. ICPhS '99* 2259–2263.
3. G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zachs, and S. Levy. Inferring articulation and recognising gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoust. Soc. Am.*, 92(2):688–700, August 1992.
4. K. Richmond. Estimating velum height from acoustics during continuous speech. *Eurospeech 1999*, (1):149–152.
5. 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), Sept. 1996.