SPEECH RECOGNITION VIA PHONETICALLY-FEATURED SYLLABLES

Simon King, Paul Taylor, Joe Frankel and Korin Richmond

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

Abstract

We describe recent work on two new automatic speech recognition systems. The first part of this paper describes the components of a system based on *phonological features* (which we call *Espresso-P*) in which the values of these features are estimated from the speech signal before being used as the basis for recognition. In the second part of the paper, another system (which we call *Espresso-A*) is described in which articulatory parameters are used instead of phonological features and a linear dynamical system model is used to perform recognition from automatically estimated values of these articulatory parameters.

1. Phonological feature-based system: *Espresso-P*

The first 5 sections of this paper report work on the components of a two stage recognition architecture based on *phonological features* rather than phones. While phonological features have been proposed before as the basis of a speech recognition system (see section 1.2 for a review), the use of features has been out of favour until recently because there had been little success in extracting them from speech waveforms, and a lack of suitable models with which to perform actual recognition. This paper reports a set of experiments which show that phonological features *can* be accurately and robustly extracted from speech; furthermore, we have shown that this is possible for speaker independent continuous speech.

1.1. *The theoretical basis of phonological features*

Most speech recognisers today are based on phones (or phonemes) which, in our opinion, are often given undue legitimacy in the speech community, particularly with respect to the assumption that a sequence of acoustic observations can be synchronised with a sequence of phones. Often phones are seen as being the "atoms" of speech in that they are the set of units from which all else (that is, word sequences) can be built. But just as with atoms in physics, it is now widely accepted in phonology that phones are decomposable into smaller, more fundamental units. There is no consensus as to what these units are, but the most popular view is that phones can be constructed from a set of *phonological distinctive features*. Phones are a useful representation because words can easily be re-written as phones using a lexicon. We argue here however that it is inappropriate to directly link acoustic observations to HMM states and phones: the HMM paradigm is not valid.

The principle of distinctive features was first proposed in the classic work of Jakobson, Fant and Halle (1952). Although this work gained much attention when published, many (e.g. (Jones, 1957)) regarded features as no-more than a useful classification scheme, whereby one could refer to the class of "nasal phones" or "voiced phones". The power of features became evident with the publication of *The Sound Pattern of English* by Chomsky and Halle (1968) (hereafter SPE), where the authors showed that what were otherwise complex phonological rules could be written concisely if features were used rather than phones. The goal of feature theory in phonology has been to discover the most basic set of fundamental underlying units (the features) from which surface forms (e.g. phones) can be derived; a small number of simple features can be combined to give rise to the larger number of phones, whose behaviour is more complex.

1.2. *Related work on Phonological Features*

The idea of using phonological features for speech recognition is not new, as many others have seen the basic theoretical advantages laid out above. Among others, the CMU Hearsay-II system (Goldberg & Reddy, 1976) made some use of features, as did the CSTR Alvey recogniser (Harrington, 1987). Often these these systems used knowledge based techniques to extract their features and in the end the performance of these systems was poor on speaker independent continuous speech. Some more recent work has continued in this vein. For example, Bitar and Espy-Wilson (Bitar & Espy-Wilson, 1995; Espy-Wilson & Bitar, 1995; Bitar & Espy-Wilson, 1996) used a knowledge-based approach to extract phonetic features from the speech signal. Lahiri and Reetz (Lahiri, 1999; Reetz, 1999) use a bottom-up rule based approach to extract phonological features from the speech signal which are subsequently decoded into lexical words. There is still no evidence that the techniques advocated have anywhere near the performance levels achieved by the statistical approaches of the techniques described in this paper or of those reviewed below.

Kirchhoff (1996) proposed a system which used HMMs to estimate feature values which are bundled into syllable units. In (Kirchhoff, 1998; Kirchhoff, 1999), Kirchhoff describes a different system, somewhat similar to that described here in which a neural network is used to predict manner and place features. She showed that the feature based recogniser performed comparatively better under noisy conditions and that a combination of a phone based recogniser and feature recogniser was better than either alone. Koreman *et al* (1999) use Kohonen networks to map between MFCCs and phonetic features, using these as observations in HMM monophone models.

A similar, but distinctly different, approach has been to use articulatory features in recognition. They share some interesting properties with phonological features, for example with respect to asynchronicity at phone boundaries. Deng and colleagues (Deng & Sun, 1994; Deng & Wu, 1996; Erler & Freeman, 1996) have modelled feature spreading explicitly in an HMM system via changes to the HMM topology. Harrington (1987) considers in detail a range of acoustic cues for automatic recognition of English consonants. Kirchhoff and Bilmes (1999) examined

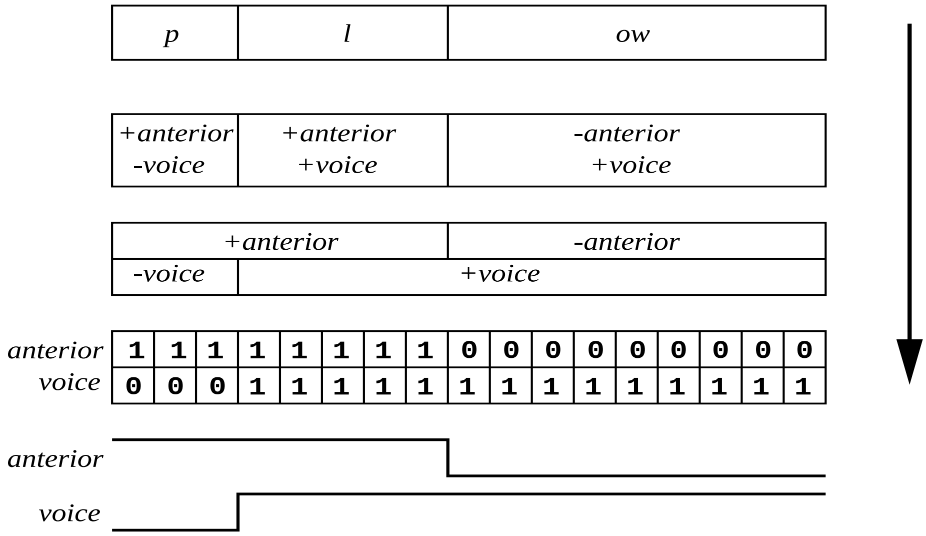

Figure 1: Deriving phonological feature values from phone labels.

conditional mutual information (CMI) between pairs of observations (MFCC, LPC, etc), conditioned on various co-articulatory conditions: speaking rate, stress type and vowel category. CMI is used as an indicator of co-articulatory effects in the speech signal. As expected, higher speaking rate, unstressed syllables and central/lax vowels all exhibit greater co-articulation. Papcun et al (1992) infer articulatory parameters from acoustics with a neural network trained on acoustic and X-ray microbeam data. Their articulatory parameters were very simple: vertical co-ordinates of the lower lip, tongue body and tongue dorsum. Zacks and Thomas (1994) use neural networks to learn acoustic–to–x-ray microbeam mapping, then do vowel classification on the output by simple template matching. Soquet *et al* (1999) report an increase in accuracy when appending articulatory and aerodynamic features to MFCCs in a speaker-dependent HMM recogniser.

2. Neural Networks for Feature Detection

This section describes the basic principles of our feature based approach. Perhaps the most useful way of describing the approach is by comparison with hybrid neural network/HMM recognisers such as Abbot (Robinson *et al.*, 1996). In these hybrid systems, the network performs an 1 from N classification over the set of phones. In our approach, the network has an output for each feature, and more than one feature can be “on” at any time. At run-time, the outputs of the trained network range continuously from 0 to 1 and this can be interpreted as a posterior probability. Another interpretation is that the network is performing a non-linear mapping problem from one space (acoustic) to another (phonological).

2.1. *Network Outputs*

Neural networks are typically trained by presenting successive pairs of known input and output patterns. The weights of the network are adjusted using the back propagation algorithm so as to minimise the mean squared error between network output and the target output. In our case each pair of patterns comprises an input of one frame of Mel cepstral coefficients and a phonological

Feature	Frames correct (%)	chance (%)
vocalic	88	71
consonantal	90	52
high	86	75
back	88	76
low	93	86
anterior	90	66
coronal	90	74
round	94	92
tense	91	78
voice	93	63
continuant	93	62
nasal	97	94
strident	97	85
silence	98	86
Average over all features	92	76
All correct together	52	14
Mapped to phone accuracy	59	14

Table 1: Results for the SPE feature system.

feature description for that frame. The cepstral coefficients can be directly calculated using signal processing on a frame by frame basis from the speech waveform, but the provision of the target output values is more tricky.

Our training corpus is fully labelled and segmented: we know the identity and boundaries of all phones. For each feature, the target is set to 1 if the feature is present in the canonical representation, and 0 otherwise. The outputs can therefore be interpreted as specifying a probability for each feature, which during training are either 0 or 1, but during run time, the outputs will take continuous values between 0 and 1. We interpret this as the probability of a feature being present. Figure 1 shows how we derive the target phonological descriptions from phone labels.

2.2. *Experimental setup*

Our experiments used the TIMIT database (Garofolo, 1988). The speech was parameterised as 12 Mel-frequency cepstral coefficients plus energy for 25ms frames, with a 10ms frame shift. All our experiments used networks with time-delaying recurrent connections, which give the network some "memory" from one pattern to the next. All networks had a single hidden layer. To allow optimisation of network size and training parameters, a validation set of 100 utterances was taken from the training set, leaving 3548 utterances for training network weights. None of the test speakers are in the training set, and hence all experiments are speaker independent.

3. Chomsky-Halle binary features

In experiment **I** we used the binary feature system from Chomsky and Halle's *"Sound Pattern of English"* (1968). There are 13 features in this system and each pronunciation unit is represented by a binary combination of these features. A single network was trained to recognise all features simultaneously, with one output for each feature and an additional network output for silence. A network with 250 hidden units and approximately 150 000 connections was found to give the best performance (measured on the validation set). The results for this network on the full test set are given in table 1. It is clear from the table that the general recognition accuracy is high, and in all cases substantially above chance levels. The performance on training and testing portions of the database did not differ greatly – this indicates that the network learned to generalise well. The chance level is the prior probability of the most likely value for a feature (given as a percentage)[1]. The "all correct together" figure gives the percentage that all features are correct for a given frame. This means that the network has found the right combination 52% of the time from a possible choice of $2^{14} = 16384$ feature combinations. The vast majority of these feature combinations don't give rise to valid phones. By forcing every frame to have a valid feature value combination (that is, a phone in the language), we can increase the phone accuracy from 52% to 59%. This is achieved by replacing invalid feature value combinations with the nearest valid combination (using a simple Euclidean distance measure). These two figures are only meant as a guide to overall network accuracy as they of course take no account of the asynchronous nature of the features: simple frame-wise phone classification is not our aim. Figure 2 shows the network output for an utterance from the test set, along with the canonical values (those that would have been used for targets had this utterance been in the training set).

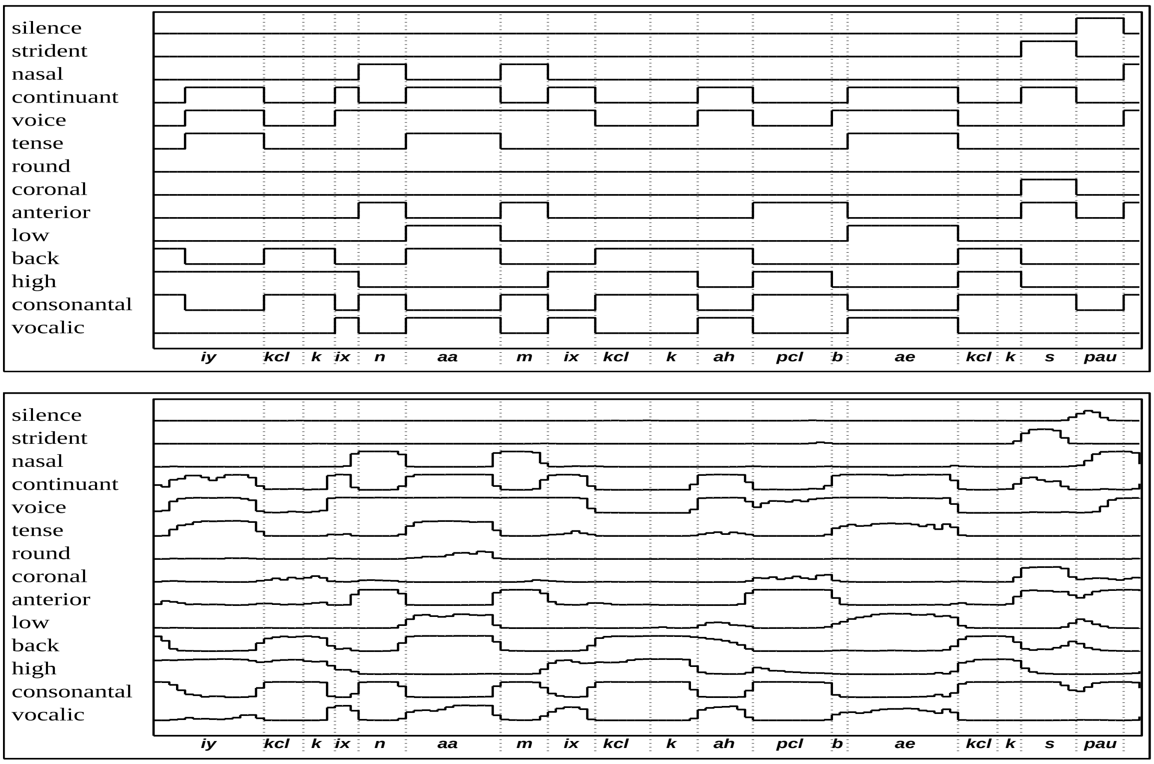

Figure 2: Example network output for the words "...economic cutbacks" for SPE feature system. The top plot shows the target values as derived from the canonical phone representation. The bottom plot shows the output of the neural net.

Feature	Possible Values		Frames correct (%)	Frames chance (%)
centrality	*central* *nil*	*full*	85	47
continuant	*continuant*	*noncontinuant*	86	45
frontback	*back*	*front*	84	59
manner	*vowel* *approximant* *nasal*	*fricative* *occlusive*	87	34
phonation	*voiced*	*unvoiced*	93	63
place	*low* *high* *coronal* *corono-dental* *velar*	*mid* *labial* *palatal* *labio-dental* *glottal*	72	25
roundness	*round*	*non-round*	92	78
tenseness	*lax*	*tense*	87	65
Average over all features			86	52
All correct together			53	14
Mapped to phone accuracy			60	14

Table 2: Results for the multi-valued feature system.

4. Multi-valued features

Experiment **II** investigated the use of a more traditional multi-valued feature system. In this system, there are fewer features, but each can take one of many values. In this experiment one network was trained for each feature in, so each network is performing a 1-of-N classification task. The size of each network was determined using the validation set, as for the previous experiment. The networks for **roundness** and **centrality** had 20 hidden units, for **phonation**, 40, and **place**, **frontback** and **manner** each had 80.

While the average per feature performance shown in table 2 is worse for these features than for the SPE features (86% as opposed to 92%), the average chance level is much lower also. The "all correct together" figures are about the same as for SPE, showing that performance of the networks on both feature systems is quite similar. Figure 3 shows the network output for an utterance from the test set, along with the canonical values (those that would have been used for targets had this utterance been in the training set).

	sil	*appr*	*fric*	*nasal*	*occ*	*vowel*
silence	89.0	1.3	2.3	1.3	3.1	3.0
approximant	0.9	68.6	1.8	1.8	1.3	25.7
fricative	1.9	0.9	88.2	1.1	4.6	3.1
nasal	1.8	1.9	2.1	84.4	2.6	7.3
occlusive	3.1	0.8	5.6	2.3	85.8	2.4
vowel	0.5	4.7	1.2	1.2	0.9	91.5

Table 3: Confusion matrix for the **manner** feature of the multi-valued system. Each row is for a correct feature value, and columns show the automatically determined values; for example, 4.7% of *vowel* frames were labelled *approximant*. All figures are percentage of frames correct.

[1] If we gave the most likely feature value to all frames, we would get the chance level of frames correct.

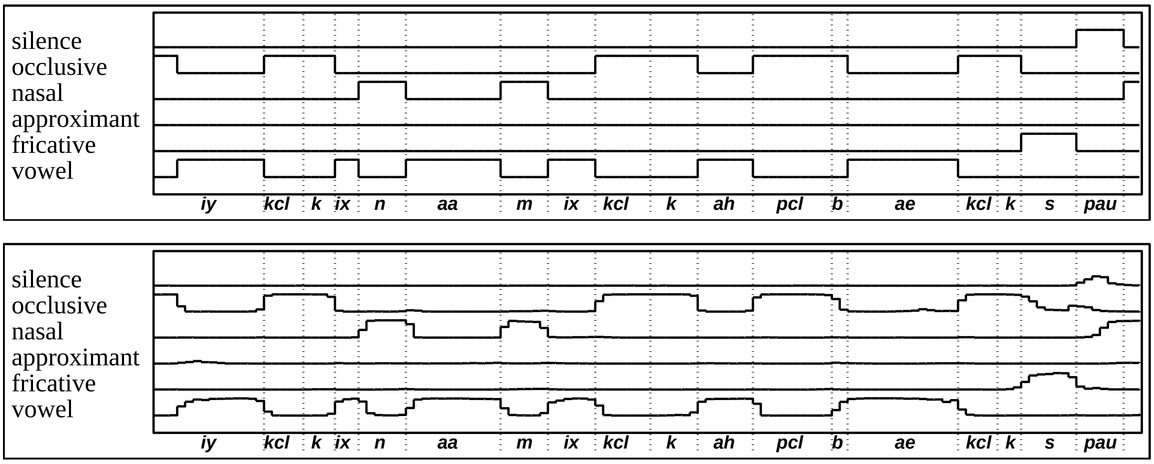

Figure 3: Example network output for the words "...economic cutbacks" for the manner feature of the multi-valued feature system. The top plot shows the target values as derived from the canonical phone representation. The bottom plot shows the output of the neural net. Compare with figures 2 and 4.

Feature		Frames correct (%)	Frames chance (%)
Primes	**A**	86	62
	I	91	79
	U	88	79
	@	88	75
	?	92	72
	h	95	79
	H	95	79
	N	98	94
Head	**a**	97	94
	i	96	90
	u	96	94
Average over all features		93	82
All correct together		59	14
Mapped to phone accuracy		61	14

Table 4: Results for Government Phonology primes.

5. Government Phonology primes

In *Government phonology* (Harris, 1994), or simply GP, sounds are described by combining *primes* in a structured way, and phonological phenomena are accounted for by the fusing and splitting of primes within a sound. GP also accounts for the combination of sounds into onset-rhyme groups; this allows elegant descriptions of phonological rules which operate on these structures. The primes **A**, **I**, **U** and **@** are known as the *resonance primes*, and capture consonant and vowel sounds. They are derived from examination of the spectral properties (formant structure) of vowels (Olive *et al.*, 1993). The **?** prime is present in sounds with a closure or any abrupt and sustained decrease in amplitude. Frication (acoustically evident as aperiodic energy) is indicated by the presence of the **h** prime, and the nasal prime **N** is present in sounds with an articulatory oral closure and acoustically with zeros in the spectrum. The **H** prime indicates unvoiced sounds, where the vocal folds are stiff and not vibrating periodically.

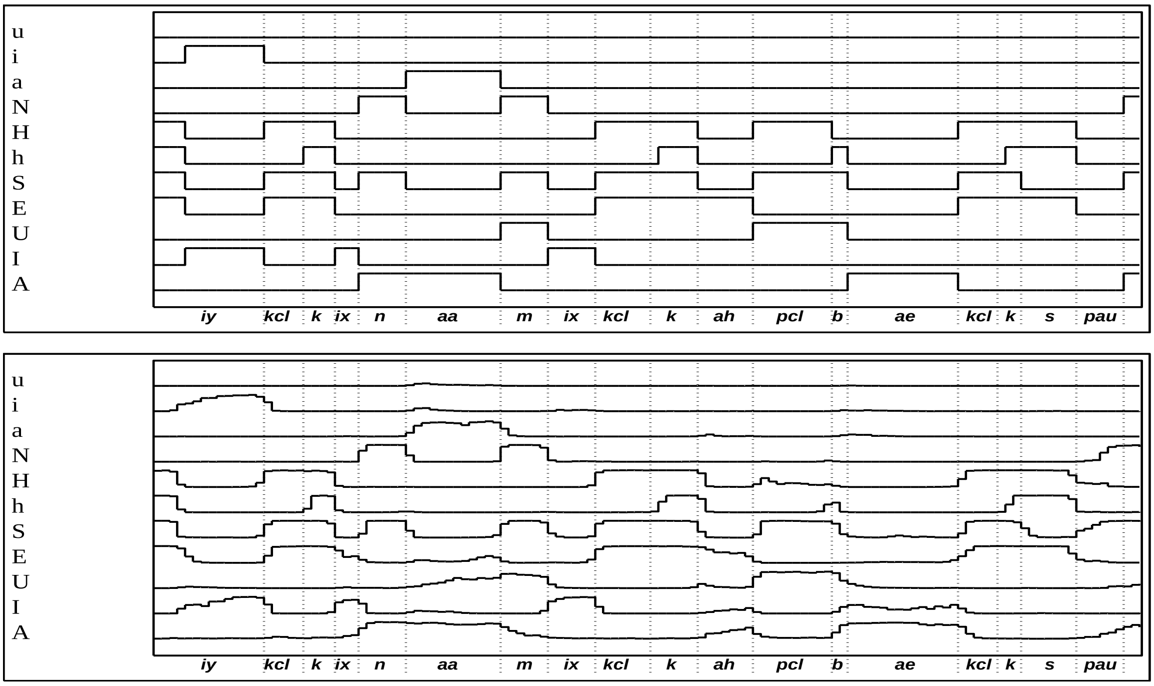

Figure 4: Example network output for the words "...economic cutbacks" for the government phonology system. The top plot shows the target values as derived from the canonical phone representation. The bottom plot shows the output of the neural net. Compare with figures 2 and 3.

The vowels */a/, /i/, /u/, /@/* are represented by just a single prime while all other sounds are made by fusing primes. For example, fusing **A** and **U** gives */o/* and fusing **A** and **I** produces */e/*. More complex sounds, like diphthongs, require the primes to be arranged in a structured way. As well as simply fusing two or more primes, one of the primes can optionally be made the *head* of the expression, denoting its greater significance both phonologically and in determining the phonetic realisation of the sound. As the GP representation is heavily structured, detecting the primes is not enough to distinguish all sounds. In experiment **III**, rather than attempt to recognise the structure directly, we have taken the approach of encoding the structure information as a set of pseudo-features. We allow three of the primes to be the head: **A**, **I** and **U**. Table 4 shows the results for the GP system and figure 4 shows the network output for an utterance from the test set. Again all features are recognised with high accuracy compared with the chance levels.

6. Articulatory parameter-based system: *Espresso-A*

Now we turn to the second system, in which articulatory parameters take the place of phonological features. We use recurrent neural networks to automatically estimate articulatory parameter values from speech; linear dynamical systems are employed to perform recognition.

6.1. *Data*

The data consists of TIMIT-like sentences (read text, continuous speech) recorded at Queen Margaret University College, Edinburgh. Articulatory measurements were recorded using a Carstens Electro-Magnetic Articulograph (EMA), along with high-quality audio. The raw acoustic and articulatory data is processed for use with the neural network by: endpoint de-

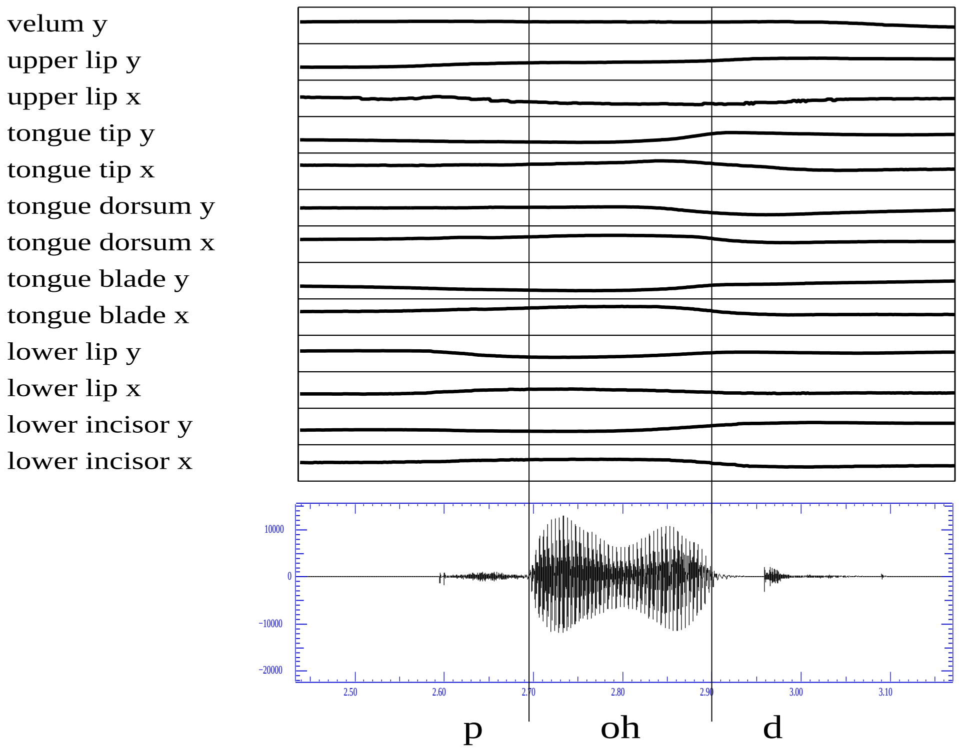

Figure 5: Example EMA data for the word "pod". Vertical lines show phone boundaries. The y coordinate is vertical (increasing y means upward movement), and the x coordinate is horizontal (increasing x means forward movement).

tection (during silent stretches, the mouth may take any position and this would adversely affect network learning); filterbank analysis (16 coefficients for 16ms frames every 8ms); resampling of EMA data 8ms frame rate; normalisation. The current system uses speech from a single speaker. 70% of the data is selected at random and used as training data. The remaining 30% is split into validation and testing sets of equal size.

7. Automatic estimation of articulatory parameter values

Researchers have sought to recover articulation from the acoustic signal for some time. Early work was typically based on analytical techniques, such as inverse filtering (e.g. (Wakita, 1973)). Recently, the development of X-Ray microbeam (XRMB) cinematography and electro-magnetic articulography (EMA) have enabled a few studies using machine learning techniques in conjunction with real human data, for example (Papcun *et al.*, 1992; Hogden *et al.*, 1996). Similar to (Papcun *et al.*, 1992), we use a large input "context" window of 25 acoustic frames and a network with two hidden layers, and a single output unit for each articulator track. A key difference was the introduction of Elman-style context units (recurrent in time) for the second hidden layer.

7.1. *Results*

Figure 6 shows an example from the test set for one articulatory parameter. Qualitatively, this shows that an accurate mapping is achieved. Table 5 gives quantitative results: the root mean squared error (RMSE) is given both in millimetres and as a percentage of the total range of

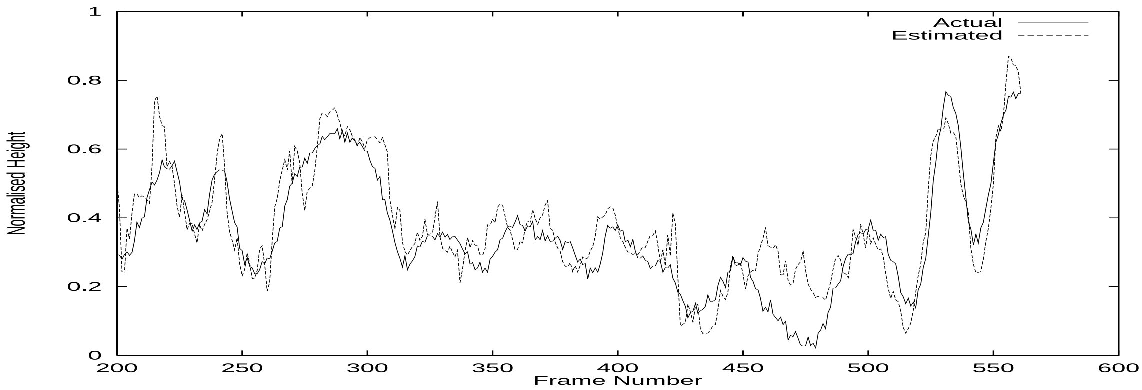

Figure 6: Actual and automatically estimated articulatory parameter (tongue tip height).

Articulator	av. RMSE mm	Correlation
Upper lip X	1.6 (25%)	0.84
Upper lip Y	1.6 (25%)	0.89
Lower lip X	3.6 (35%)	0.85
Lower lip Y	2.3 (22%)	0.86
Lower incisor X	2.9 (32%)	0.84
Lower incisor Y	1.5 (18%)	0.90
Tongue tip X	3.3 (18%)	0.88
Tongue tip Y	3.9 (18%)	0.88
Tongue body X	3.9 (23%)	0.88
Tongue body Y	2.5 (16%)	0.87
Tongue dorsum X	3.2 (18%)	0.89
Tongue dorsum Y	3.2 (19%)	0.84
Velum X	3.2 (26%)	0.91
Velum Y	1.7 (18%)	0.90

Table 5: Quantitative results for automatic estimation of articulatory parameter values.

movement for each articulator. The correlation figures indicate the similarity in the *shape* of the two trajectories.

8. Linear dynamical systems

The second stage of the process revolves around modelling these trajectories. We have chosen a linear dynamic model described by the following pair of equations:

$$\begin{aligned} y_t &= Hx_t + v_t \\ x_t &= Fx_{t-1} + w_t \end{aligned}$$

with x_t representing the hidden state and y_t the observation at time t. x's evolution from time $t-1$ to t is governed by the matrix F and some normally distributed error w_t, with non-zero mean μ_v and covariance C. This is projected onto the observation space via the matrix H and more normally distributed error v_t with non-zero mean μ_w and covariance D. One set of parameters H, F, C, D, μ_v, and μ_w describe the articulatory motion for one segment of speech; so far, the segments used have been phones; a different model is used for each phone. We chose this form of model for two reasons: the state space evolves in a continuous fashion (this is highly desirable given the nature of the physical system it describes); the observations y_t are in the articulatory domain, so a linear mapping from x to y is reasonable (and makes parameter estimation much simpler). Parameter estimation is performed using a Markov Chain Monte

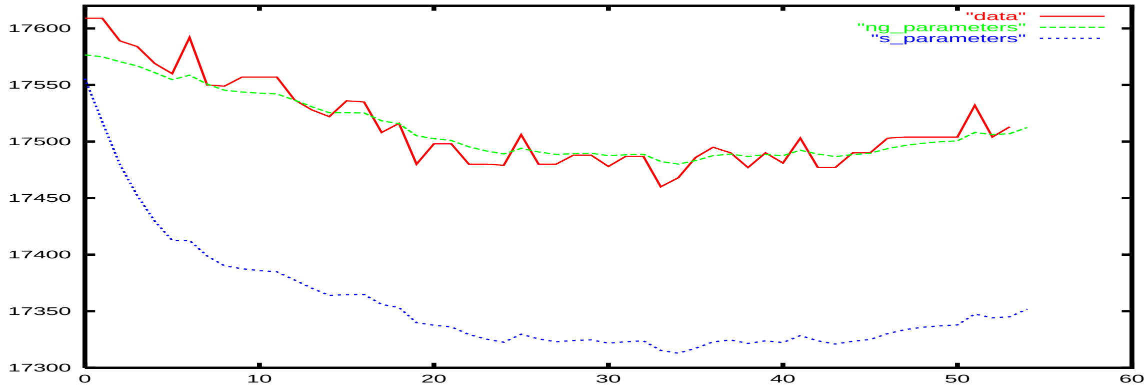

Figure 7: Linear dynamical models in action: solid line shows actual velum height for a token of /m/; predicted velum height from a model of /m/ is shown by the dotted line and for a model of /b/ by the dashed line.

Carlo technique (which is a Bayesian method): the Gibbs sampler. This is an alternative to the more obvious choice, Expectation-Maximisation (EM). It has some advantages over EM: given appropriate priors, a unique solution is found; it is less susceptible to local maxima; changing the form of the model, or the nature of the distributions on individual parameters is trivial. During recognition, we compute the probability of the observations, given the model parameters.

8.1. *Results for classification from real articulatory parameter values*

8.1.1. *Nasal vs. non-nasal*

A three way classification of segments into nasal, non-nasal and silence was performed using only the velum y-coordinate. The training set consisted of 8980 tokens from 259 utterances from a single female speaker, and the testing set had 2299 tokens from 66 utterances.Results are almost identical when testing is done on the training set, which suggests that the models have not been over-learning.

		classified as			
		nasal	silence	non-nasal	% correct
segment	nasal	134	43	8	72
	silence	41	222	1	84
	non-nasal	515	61	1274	69
Total					71

Table 6: Nasal classification from real articulatory parameter values.

8.1.2. *Phone classification*

In this experiment, the task was to classify tokens of /b/ and /m/. The training set consisted of 366 tokens from 259 utterances, and the testing set had 100 tokens from 66 utterances. Results are shown in table 7. Figure 7 shows the models performing classification on a token of /m/.

8.2. *Results for classification from automatically estimated articulatory parameter values*

In our most recent experiments, the automatically estimated articulatory parameter values were used for phone classification, in an experiment otherwise similar to that in section 8.1.2. The

		classified as b	m	% correct
segment	b	35	8	81
	m	2	55	96
Total				90

Table 7: Phone classification from real articulatory parameter values.

training set consisted of 146 tokens from 230 utterances, and the testing set had 69 tokens from 49 utterances. Results are shown in table 8.

		classified as b	m	% correct
segment	b	21	7	75
	m	4	37	90
Total				84

Table 8: Classification from automatically estimated articulatory parameter values.

9. Discussion

We now discuss some issues concerned with actual *recognition*, that is, the conversion of feature descriptions for an utterance into linguistic units such as words or phones. Our long term goal is to develop new statistical models designed to work with phonological features or articulatory parameters. These models will make explicit use of the benefits of features, for example by assuming conditional independence between the different feature values in a frame, and by modelling co-articulation with reference to the theory of critical articulators, etc. While this is the subject of current and future work, it certainly is reasonable to ask at this point what evidence we have that we are on the right track and that we haven't simply developed an interesting representation.

9.1. *Phone recognition*

A simple way of testing the information content of a feature representation is to treat it as a normal acoustic feature representation and train standard models. To this end, we performed an phone recognition experiment on TIMIT with a simple HMM speech recogniser. This used tied-state, cross word triphones, and a single Gaussian was used to model the observation density. A phone bigram language model was used. Our baseline system used Mel-scale cepstral features and using these as observations the phone accuracy was 63.3%. While this figure is lower than state of the art for TIMIT phone recognition, it should be noted that no particular optimisation of the recogniser was performed for the phone recognition task. An equivalent experiment was performed using exactly the same recognition architecture, but using multi-valued features rather than cepstra. That is, the trained neural network (as described in section 2) was used to produce multi-valued feature descriptions, and these were used as observations in the HMM system. This system gave a higher[2] phone recognition accuracy of 63.5%.

[2] not signifi cantly different

9.2. *Randomised features*

How do we know that the phonological feature-detecting neural networks are not simply doing phone classification in disguise? We repeated the experiment using SPE features from section 3 but with a randomised phone-to-feature-value table. Framewise accuracy drops from 52% to 37%. If the net was (internally) performing phone classification, then mapping to a binary representation, we would expect the two results to be the same.

9.3. *Conclusion*

While we do not actually advocate that phonological features should simply be used instead of acoustic features in a HMM recogniser, what this experiment shows is that they are at least as useful a representation, and the mapping from acoustics to features performed by the network hasn't been at the expense of information useful for recognition. Kirchhoff (1999) has also tried this approach and used features similar to ours in place of acoustic observations in Hybrid NN/HMM and HMM recognition systems. Her results show a similar pattern to ours, in that the systems using features have very close performance to systems using cepstra for the same recognition architecture. A number of interesting models have recently been proposed for use with acoustic features which we think would be suitable to serve as the basis of a phonological recognition model. A number of these approaches have been developed with the intention of modelling asynchrony. Multi-stream models (Bourlard & Dupont, 1996; Tibrewala & Hermansky, 1997) examine frequency bands separately and exploit the fact that listeners can perform partial recognition on individual bands and recombine the evidence relatively late in processing. In separate work, Sagayama *et al.* (1999) have proposed *asynchronous transition* HMMs (AT-HMMs) which model the temporal characteristics of each acoustic feature component separately. Their system uses a form of the successive state splitting algorithm (Takami & Sagayama, 1992; Ostendorf & Singer, 1997) to learn the temporal and contextual characteristics of each feature. Using mel-scale cepstra as observations, they report a significant reduction of errors compared to a standard HMM approach. These approaches are ideally suited to our task as they model asynchrony inherently. Our own work has been with linear dynamical system models, as described in section 8.

It is useful at this stage to say something about the nature of the features with regard to asynchrony. While the neural networks were trained on feature values which switched instantaneously at phoned boundaries, it is clear from their output that even when the networks are performing well, features often do not all change at phone boundaries, (for example the transition between */n/* and */aa/* in figure 2). To measure the size of this affect, we calculated the frame-wise classification accuracy if the features values were allowed some leeway near phone boundaries. We automatically corrected feature value transitions that were up to 20ms away from the phone boundary (but which had the correct value before and after the transition). Using this reclassification on the SPE features from section 3, the accuracy figure for "all frames correct" changes from 52% to 63%, and the figure for mapping to the nearest phone increases from 59% to 70%. These significant differences in performance show that asynchronous feature value changes are common, and indicate that recognition models which can model this properly should achieve significantly higher performance than the standard, frame synchronous HMM system reported above.

References

Bitar, N. N., & Espy-Wilson, C. Y. (1995). A signal representation of speech based on phonetic features. *Pages 310–315 of: Proc. 1995 IEEE Dual-Use Technologies and Applications Conference*.

Bitar, N. N., & Espy-Wilson, C. Y. (1996). A Knowledge-Based Signal Representation for Speech Recognition. *Pages 29–32 of: Proc. ICASSP '96*.

Bourlard, H., & Dupont, S. (1996). A new ASR approach based on independent processing and recombination of partial frequency bands. *Pages 426–429 of: Proc. ICSLP '96*.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. MIT Press.

Deng, L., & Sun, D. X. (1994). A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *J. Acoust. Soc. America*, **95**(May), 2702–2719.

Deng, L., & Wu, J. J.-X. (1996). Hierarchical Partition of the articulatory state space for overlapping-feature based speech recognition. *Pages 2266–2269 of: Proc. ICSLP '96*, vol. 4.

Erler, K., & Freeman, G. H. (1996). An HMM-based speech recognizer using overlapping articulatory features. *J. Acoust. Soc. Am.*, **100**(4), 2500–2513.

Espy-Wilson, C. Y., & Bitar, N. N. (1995). Speech Parameterization Based on Phonetic Features: application to speech recognition. *Pages 1411–1414 of: Proc. Eurospeech-95, Madrid*.

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

Goldberg, H. G., & Reddy, D. R. (1976). Feature extraction, segmentation and labelling in the Harpy and Hearsay-II systems. *J. Acoust. Soc. Am.*, **60**.

Harrington, J. (1987). Acoustic cues for automatic recognition of English consonants. *Pages 19–74 of:* **Jack, M. A., & Laver, J.** (eds), *Speech Technology: a survey*. Edinburgh University Press.

Harris, J. (1994). *English Sound Structure*. Blackwell.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., & Saltzman, E. (1996). Accurate Recovery of Articulator Positions from Acoustics: New Conclusions Based on Human Data. *J. Acoust. Soc. Am.*, **100**(3), 1819–1834.

Jakobson, R., Fant, G. M. C., & Halle, M. (1952). *Preliminaries to Speech Analysis: the Distinctive Features and their correlates*. MIT press.

Jones, D. (1957). *An Outline of English Phonetics*. 8 edn. Cambridge: Heffer & Sons.

Kirchhoff, K. (1996). Syllable-level desynchronisation of phonetic features for speech recognition. *Pages 2274–2276 of: Proc. ICSLP '96*, vol. 4.

Kirchhoff, K. (1998). Combining Articulatory And Acoustic Information For Speech Recognition In Noisy And Reverberant Environments. *In: Proc. ICSLP '98*.

Kirchhoff, K. (1999). *Robust Speech Recognition Using Articulatory Information*. Ph.D. thesis, Universität Bielefeld.

Kirchhoff, K., & Bilmes, J. A. (1999). Statistical Acoustic Indications Of Coarticulation. *In: Proc. ICPhS 99*.

Koreman, J., Andreeva, B., & Strik, H. (1999). Acoustic Parameters Versus Phonetic Features in ASR. *Pages 719–722 of: Proc. ICPhS 99*.

Lahiri, A. (1999). Speech Recognition With Phonological Features. *Pages 715–718 of: Proc. ICPhS 99*.

Olive, J. P., Greenwood, A., & Coleman, J. (1993). *The Acoustics of American English Speech, a Dynamic Approach*. New York: Springer.

Ostendorf, M., & Singer, H. (1997). HMM topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, **11**(1), 17–41.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., & Zacks, J. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoust. Soc. Am.*, **92**(2), 688–700.

Reetz, H. (1999). Converting Speech Signals To Phonological Features. *Pages 1733–1736 of: Proc. ICPhS 99*.

Robinson, T., Hochberg, M., & Renals, S. (1996). The use of recurrent neural networks in continuous speech recognition. *Chap. 7 of:* **Lee, C.-H., & Soong, F. K.** (eds), *Automatic Speech and Speaker Recognition - Advanced Topics*. Kluwer Academic Publishers.

Sagayama, S., Matsuda, S., Nakai, M., & Shimodaira, H. (1999). Asynchronous-transition HMM For Acoustic Modeling. *Page 99 of: International Workshop on Automatic Speech Recognition and Understanding*, vol. 1.

Soquet, A., Saerens, M., & Lecuit, V. (1999). Complementary Cues For Speech Recognition. *Pages 1645–1648 of: Proc. ICPhS 99*.

Takami, J., & Sagayama, S. (1992). A successive state splitting algorithm for efficient allophone modelling. *Pages 573–576 of: Proc. ICASSP 92*, vol. I.

Tibrewala, S., & Hermansky, H. (1997). Sub-band Based Recognition Of Noisy Speech. *Pages 1255–1258 of: Proc. ICASSP '97*.

Wakita, H. (1973). Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms. *IEEE Trans. Audio Electroacoust.*, **21**, 417–427.

Zacks, J., & Thomas, T. R. (1994). A new neural network for articulatory speech recognition and its application to vowel identification. *Computer Speech and Language*, **8**, 189–209.