Article Submitted to Computer Speech and Language

Detection of Phonological Features in Continuous Speech using Neural Networks

SIMON KING AND PAUL TAYLOR

Centre for Speech Technology Research
University of Edinburgh
80, South Bridge
Edinburgh
EH1 1HN
Contact: `Simon.King@ed.ac.uk`

Abstract

We report work on the first component of a two stage speech recognition architecture based on *phonological features* rather than phones. The paper reports experiments on three phonological feature systems: 1) the *Sound Pattern of English (SPE)* system which uses binary features, 2) a *multi valued (MV)* feature system which uses traditional phonetic categories such as manner, place etc, and 3) *Government Phonology (GP)* which uses a set of structured primes. All experiments used recurrent neural networks to perform feature detection. In these networks the input layer is a standard framewise cepstral representation, and the output layer represents the values of the features. The system effectively produces a representation of the most likely phonological features for each input frame.

All experiments were carried out on the TIMIT speaker independent database. The networks performed well in all cases, with the average accuracy for a single feature ranging from 86% and 93%. We describe these experiments in detail, and discuss the justification and potential advantages of using phonological features rather than phones for the basis of speech recognition.

1. Introduction

1.1. The theoretical basis of phonological features

This paper reports work on the first component 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 1.3 for a review), the use of features has been out of favour until recently because there had been little success in extracting them from the speech waveform. 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.

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*.

The principle of distinctive features was first proposed in the classic work of Jakobson, Fant and Halle [17]. Although this work gained much attention when published, many (e.g. [19]) 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 [7] (hereafter SPE), where the authors showed that what were otherwise complex phonological rules could be written concisely if features were used rather than phones. For example, consider the well-known phenomena of nasal assimilation in English. When a prefix such as "in" is added to a word, the nasal changes and takes on the place of articulation of the first phone of the word. E.g. "in" + "possible" $\rightarrow$ "impossible", "in"+ "balance" $\rightarrow$ "imbalance", "in" + "material" $\rightarrow$ "immaterial" etc. Rather than write a rule for each phone, this can be expressed neatly by saying the place of articulation is the same for the nasal as for the first phone of the word. Other uses of features include specification of phonotactics. In English it is very laborious to list the large set of possible syllable initial consonant sequences (e.g. "sp", "st" "sk", "spr", "spl" etc). But the same sequences can be described neatly by appealing to features (for example only unvoiced stops can appear after an */s/*, and only approximants can appear after a stop).

The SPE feature system used production-based binary features. In this system, each phone was composed of a vector of 13 binary components which represented production features such as *voicing*, *high*, *low* (representing tongue position during vowels), *round* (for lip rounding), *continuant* (to distinguish continuous sounds such as vowels and fricatives from stops), and so on. Features have also been proposed as the basis of spoken language universals, in the sense that while the phonemes of a language vary, the set of features does not and is the same for all languages.

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. Feature schemes should be minimal in the sense that the features should be independent and not contain redundancy. This follows on from the idea of finding the most basic set - if for example in a particular feature set, two features were found to regularly co-occur, then that could be taken as evidence that a different, as yet undiscovered, feature could replace them. The physics analogy is particularly useful here. If elements are described individually, they seem to exhibit idiosyncratic and somewhat arbitrary behaviour. However, by describing them in terms of their sub-atomic makeup, the picture becomes much clearer (cf. the periodic table). The important point is that a small number of relatively simple sub-atomic particles can be used to describe the complex behaviour of a much larger set of units from which they are made. Likewise, the principle behind phonological features is that 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. Advantages of Phonological Features over Phones

Phones are a useful representation because words can easily be re-written as phones using a lexicon. In conventional HMM systems, phones are then re-written as HMM states. HMMs are generative models, with each observation generated by a single state. During recognition, the state sequence is hidden, and the probability that a particular model has generated a given sequence of observations is often calculated approximately using Viterbi decoding. We argue here that it is inappropriate to align observations, phones and words in this strict fashion.

In HMM speech recognition, the acoustic model which relates states to observations effectively does two jobs: 1) it turns representations from the acoustic domain to a phonetic one; 2) it models the time dynamics of the acoustics so that a sequence of similar acoustic observations are modelled as single phone. Our feature based approach performs each of these operations separately. The feature extraction component maps from the acoustic domain to the phonetic domain, but the representation is still one of frame based time varying vectors. The phone model then operates in the feature domain. The phonological features could be regarded as latent variables: a simple, hidden process which gives rise to a (more complex) observable process.

The basic problem with performing the two steps at once is that the HMM has no inherent ability to model the dynamics of the acoustic observations*. Speech production from phones to acoustics is complex and non-linear and hence phenomena which can have relatively simple phonetic explanations can give rise to extremely complex acoustic patterns. While it is easy to model nasality spread phonetically, it is very difficult to do so in the acoustic domain as the effects of nasality can not be represented by a simple function operating on the acoustics.

Critical and non-critical articulators

Rather than view phonological features as an alternative frame-based vector representation to cepstra, it can be helpful to view feature representations as trajectories in time in which each component of the frame (each feature) exhibits different, but coupled, dynamic behaviour. This is clearer if we refer to actual articulatory movements. Figure 1 shows data recorded using an Electro-magnetic Articulograph (EMA). During production of the *d*, the critical articulator is the tongue tip – the place of articulation in the SPE system is *coronal*. Note the rapid movement of the tongue tip upwards (see *tongue tip y* track) to make the stop closure. Non-critical articulators, such as the lower lip (see *lower lip y* track) exhibit slower movement which is spread across phones: the lower lip is raised to produce the closure for the *[p]*, then lowers to make the vowel, and raises again for the *[d]*. Note how the lowering and raising movement takes place during the vowel – this articulator never reaches a stable position during the vowel as it anticipates the following stop. In other words, the degree of lowering of the lower lip to make the vowel is not required to be precise, but the closure for the *[p]* is. In addition, each feature is of varying importance, depending on the phone in question. The consequence of this is that when uttering a *[d]* in some context, only the tongue tip/alveolar ridge feature is distinctive - all the other features are governed by

*The combination of a discrete state and the Markov property restricts trajectories of parameter means to be piecewise constant. This situation is mitigated a little by appending first and second differences to the observation.

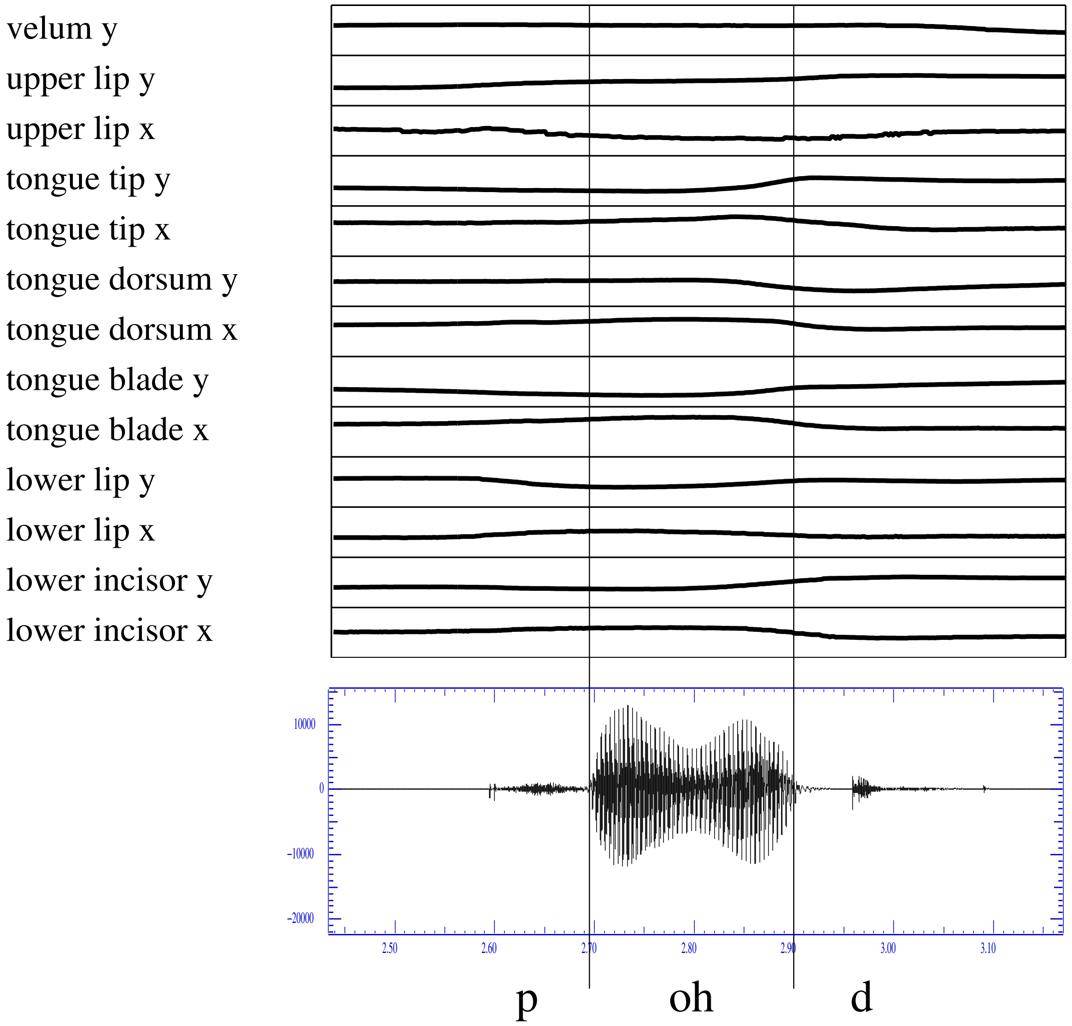

Figure 1: Example EMA data for the word "pod". Vertical lines show phone boundaries. The y coordinate is vertical (increasing y means upward movement), and the x coordinate is horizontal (increasing x means forward movement).

context, and during the *[d]* will probably just interpolate the positions required by the preceding and following phonological units. If this can be modelled in the phonetic domain, the crucial nature of a *[d]* can be represented quite simply by specifying high weighting for the alveolar feature while diminishing the importance of other evidence. Exactly the same *[d]* model can be used regardless of context. However, if acoustic observations are used, it will seem that *[d]* in different contexts exhibit a large amount of (unexplained) variance. So, the essence of the feature based model is that phenomena that seem extremely complex in the acoustic domain can have relatively simple explanations in the phonetic feature domain.

Contextual effects

HMMs make the following independence assumption:

$$P(O_1, ..., O_{N_p}|p_1, p_2, ...p_{N_p}) \quad = \quad \prod_{i=1}^{N_p} P(O_i|p_i)$$

where O_i is the sequence of observations associated with phone p_i. It is widely accepted that this assumption is not valid. That is, it is wrong to assume that $P(O_i|p_i)$ is independent of its neighbouring phones p_{i-1} and p_{i+1} (or indeed p_{i-2}, etc). In fact the observations associated with a phone are heavily influenced by phonetic context. If the parameters of a phone model are estimated from training examples from multiple contexts, the variances in the phone model will be excessively high, leading to poor modelling. However, rather than change the form of the model to make $P(O_i)$ explicitly depend on p_{i-1} (and so on), the technique of *context-dependent* phones is often used instead. A different model of each phone is required for each possible context; typically one phone to the left and right is used as context, giving *triphone* models. This divisive modelling strategy results in a large number of models, each with less training data, and means that techniques for reducing the number of model parameters are necessary; usually this means sharing (tying) parameters: state or mixture tying, for example. State tying is often performed using decision trees, in which a series of questions are asked, often about the phonological features of the phone and its context [39].

In search of a more principled system than this, we argue against the approach of assuming that all phonetic contexts are different, then grouping them together to reduce the number of parameters. Specifically, we propose that modelling context effects directly in terms of phonological feature trajectories should allow for better, more accurate models.

1.3. 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 [13] made some use of features, as did the CSTR Alvey recogniser [14]. Often 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 [2, 3, 11] used a knowledge-based approach to extract phonetic features from the speech signal. They measured properties of the signal, such as energy in certain frequency bands, and formant frequencies, and defined the phonetic features as functions of these acoustic measurements. Ali *et al* [1] also use rules, to segment and classify phones from a 30 speaker subset of the TIMIT database. Results are not reported in the conventional way, as a phone recognition accuracy, so are hard to compare with other work. Their results for phonetic feature detection, such as place of articulation, are only given for certain classes of segment. Lahiri and Reetz [25, 30] use a bottom-up rule based approach to extract phonological features from the speech signal which are subsequently decoded into lexical words. While these studies have shown some useful insights into the relationship between features and acoustics, there is still little evidence that such techniques have reached the performance levels required for speaker independent continuous speech recognition.

Kirchhoff [20] proposed a system which used HMMs to estimate feature values which are bundled into syllable units; complete temporal independence is allowed within units, with alignment at unit boundaries. In [21, 22], Kirchhoff describes a different system, somewhat similar to that described here in which a neural network is used to predict manner and place features. A particularly interesting point about this work was that 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* [24] use Kohonen networks to map between MFCCs and phonetic features, using these as observations in HMM monophone models. A great improvement was reported over HMMs using MFCCs as observations, albeit from a low baseline. Huckvale [16] proposes a tiered model based on non-linear phonology in which the "fairly independent tiers", such as *Excitation* and *Position* [of articulation] correspond roughly to phonological features. Neural networks are used to automatically label speech with this tiered representation. The database consists of a single repetition of 666 monosyllabic words for training and 359 for testing, and 51% of test words were correctly recognised.

A similar, but distinctly different, approach has been to use articulatory features (see figure 1) in recognition. Articulatory features differ from the features we are interested in here in that they are more closely linked to the physiology of the speech production mechanism rather than to phonological contrasts. Nevertheless, they share some interesting properties with phonological features, for example with respect to asynchronicity at phone boundaries. Deng and colleagues [8, 9, 10] have modelled feature spreading explicitly in an HMM system via changes to the HMM topology.

Kirchhoff [23] examined conditional mutual information (CMI) between pairs of observations (MFCC, LPC, etc) at differing times 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. Bridle *et al* [6] give evidence for the principle of critical articulators. For example, they explain that when uttering a *[d]*, the crucial thing is for the speaker is to make contact between the tongue and the alveolar ridge - the shape of the rest of the mouth and the tongue isn't important for phone identity.

Papcun et al [28] 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 [40] 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* [34] 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 [32, 5]. In these hybrid systems, the network performs a 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

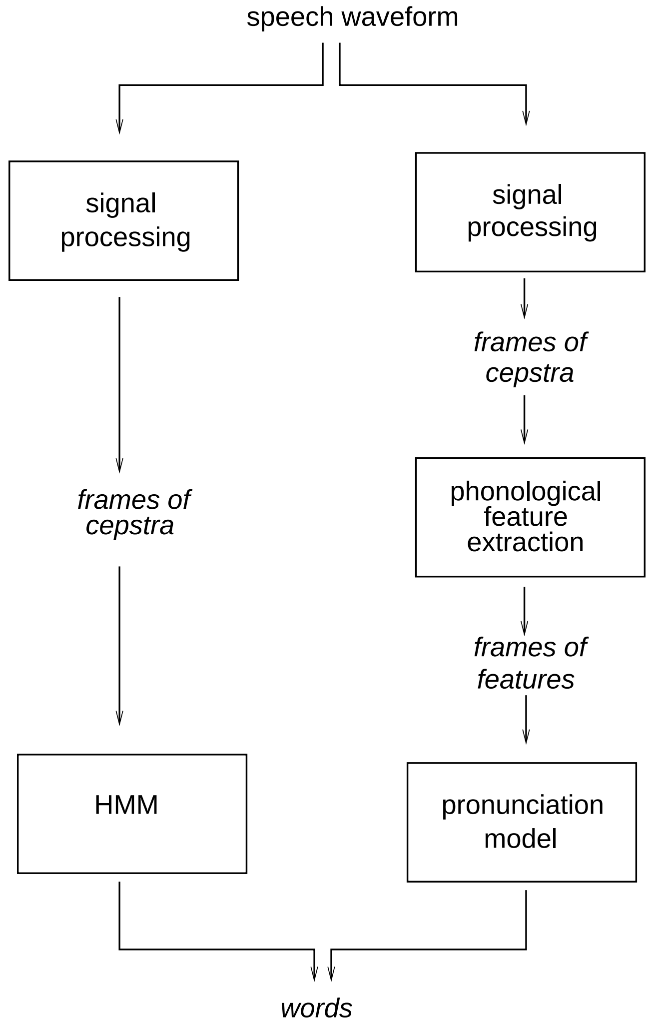

Figure 2: Speech recognition by HMMs, shown on the left, and by phonological features, shown on the right.

Mel cepstral coefficients and a phonological 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.

The problem arises because we can't actually determine what the phonological feature specification for a particular frame should be. Given a labelled corpus, it is of course possible to determine a *canonical* feature representation for each frame by assigning it the feature representation of the phone which is marked. However this doesn't take into account many of the points made above with respect to features changing asynchronously at phone boundaries. This problem raises the difficult issue of how concrete (close to the acoustics) or abstract (close to the phonology) we would

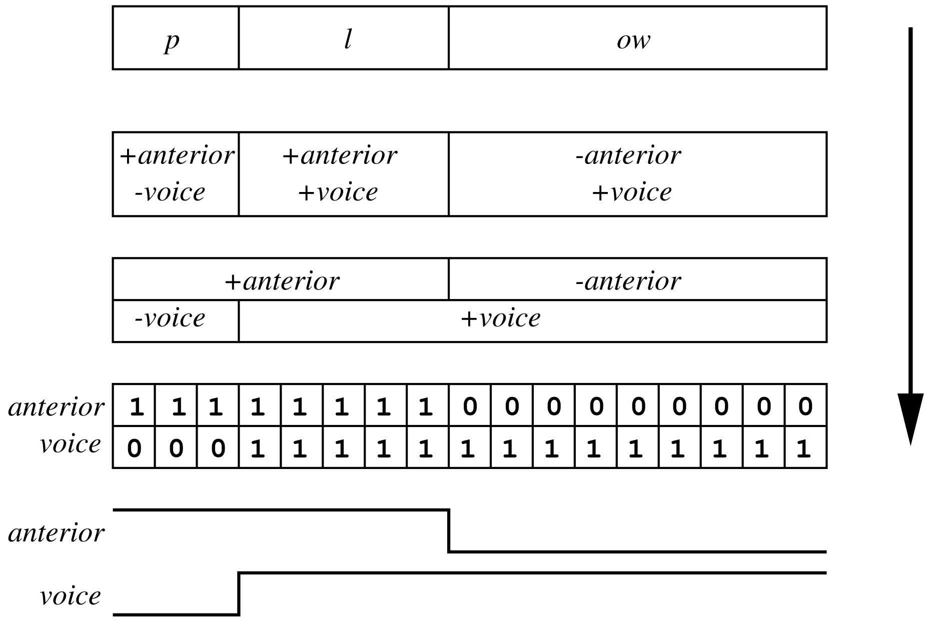

Figure 3: Deriving phonological feature values from phone labels.

like these feature representations to be. Given our framework, a concrete interpretation would make the features easier to recognise from the acoustics, implying that they would exhibit many of the properties mentioned above such as asynchronicity and critical articulations. It would however, be difficult to provide accurate training targets for concrete features as we have no mechanism at present to predict this from the phonetic transcription. On the other hand, it would be easy to specify the training targets for an abstract interpretation, but the recognition task is inherently harder and hence recognition errors may be more common.

Our solution to this problem is to use different interpretations for training and running the networks. Training is achieved by specifying the canonical targets for each labelled phone, but at run-time features frequently change at different frames near phone boundaries, which shows that the network naturally tends towards an asynchronous mode of behaviour. This can be seen more clearly in the following sections which describe our experiments.

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 the probability that each feature is present, which during training are either 1 or 0, but during run time, the outputs will take continuous values between

0 and 1. That is, for a 1-from-N classification task, the (normalised) network outputs can be interpreted as a discrete PDF; for an m-from-N task, individual outputs can be interpreted as posterior probabilities. Figure 3 shows how we derive the target phonological descriptions from phone labels.

2.2. Experimental setup

Our experiments used the TIMIT database [12]. This is a corpus of high-quality recordings of read continuous speech from North American speakers. The entire corpus is reliably transcribed at the word and surface phonetic levels. 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. The network thus can learn to use as much left context as required. Only a percentage of possible connections between units in successive layers are allowed; typically this parameter was varied from 25% to 100%. Higher values mean more weights to be trained. The Nico [35] toolkit was used throughout.

TIMIT is divided into 3648 training utterances and 1344 test utterances (only the si and sx sentences were used). 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.

During network training, performance on the validation set was observed. Training was terminated when this performance (measured either as a classification percentage, or as the error between target and actual network output) reached a plateau. To determine the optimum network size, networks with various numbers of hidden units were used. The training method employed prunes connections with low weights as training proceeds, allowing some automatic determination of the number of free parameters in the network. Variables such as learning rate (step size), momentum and degree of connection pruning were also varied to optimise training time and performance on the validation set.

3. Chomsky-Halle binary features

In experiment **I** we used the binary feature system from Chomsky and Halle's *"Sound Pattern of English"* [7]. There are 13 features in this system and each pronunciation unit is represented by a binary combination of these features. Table 1 gives examples of the feature specification of some phones.

	aa	ow	p	l	n	f
vocalic	+	+	-	-	-	-
consonantal	-	-	+	+	+	+
high	-	-	+	+	-	-
back	+	+	-	-	-	-
low	+	-	-	-	-	-
anterior	-	-	+	+	+	+
coronal	-	-	-	+	-	-
round	-	+	-	-	-	-
tense	+	+	-	-	-	-
voice	+	+	-	+	+	-
continuant	+	+	-	+	-	+
nasal	-	-	-	-	+	-
strident	-	-	-	-	-	+
silence	-	-	-	-	-	-

Table 1: SPE binary feature values for some phones from the TIMIT set.

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 2: Results for the SPE feature system.

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, 50% connectivity and approximately 150 000 connections was found to give the best performance (measured on the validation set). The results for this net-

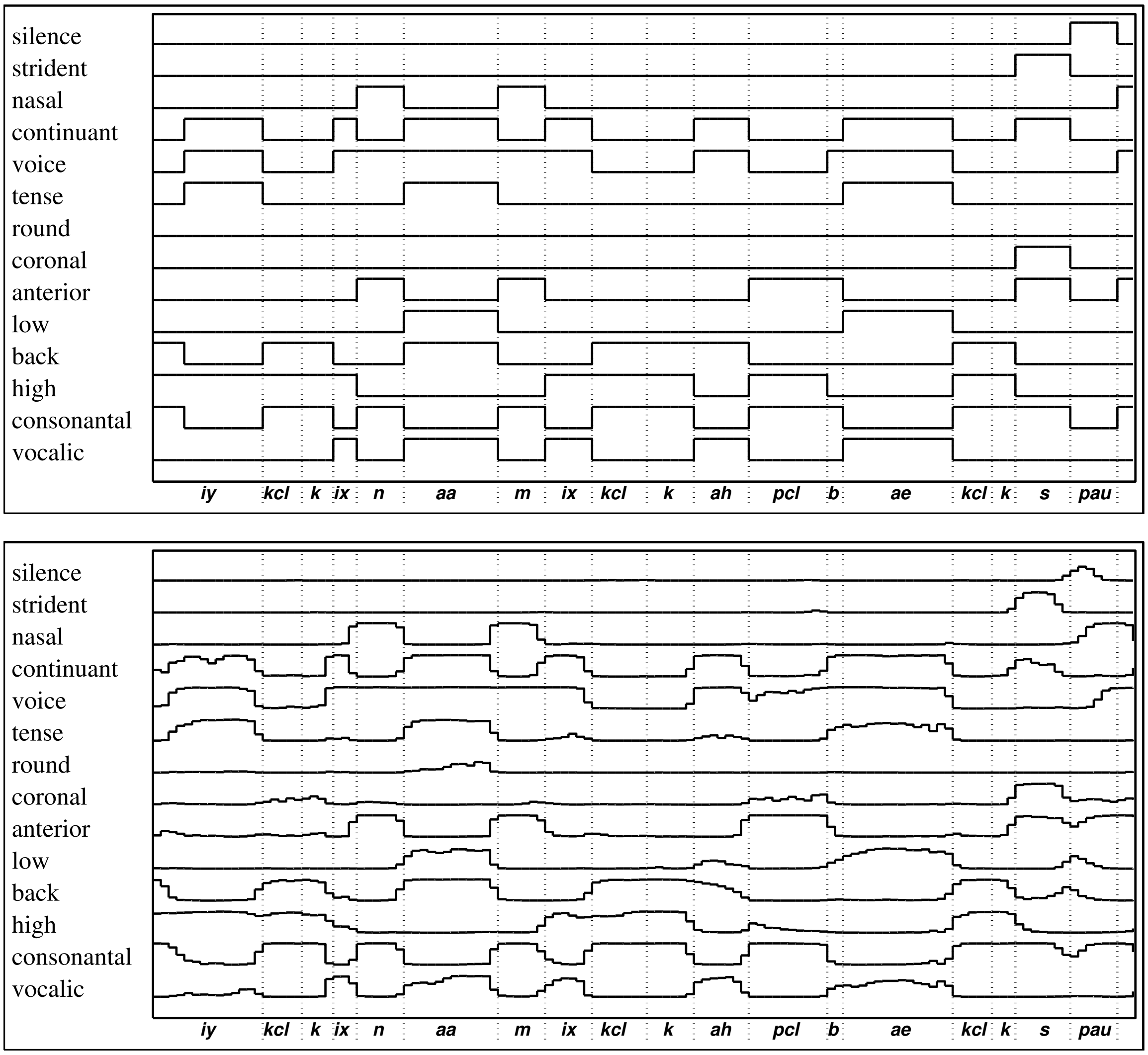

Figure 4: Example network output for the words "...economic cutbacks" for SPE feature system. The top plot shows the target values as derived from the canonical phone representation. The bottom plot shows the output of the neural net. It can be seen that in general the output of the network is very accurate and tracks the ideal values well. The major difference is that the target values switch between 0 and 1 at phone boundaries, whereas the network output values are often only at extremes in the middles of phones: values at boundaries tend to be intermediate.

work on the full test set are given in table 2. 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. When evaluating the results in the table, chance levels, giving the most common value for the feature, should be taken into account. The chance level is the prior probability of the most likely value for a feature given as a percentage. 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 framewise phone classification is not our aim. Figure 4 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).

4. Multi-valued features

	aa	ow	p	l	n	f
centrality	*central*	*full*	*nil*	*nil*	*nil*	*nil*
front-back	*back*	*back*	*front*	*front*	*front*	*front*
manner	*vowel*	*vowel*	*occlusive*	*approximant*	*nasal*	*fricative*
phonation	*voiced*	*voiced*	*unvoiced*	*voiced*	*voiced*	*unvoiced*
place	*low*	*high*	*labial*	*coronal*	*coronal*	*dental*
roundness	*unrounded*	*rounded*	*unrounded*	*unrounded*	*unrounded*	*unrounded*

Table 3: Multi-valued features for some phones from the TIMIT set.

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. Table 3 gives the feature specifications for some example phones. In this experiment one network was trained for each feature, 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 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 5 shows the network output for an utterance from the test set.

Feature	Possible Values		Frames correct (%)	Frames chance (%)
centrality	*central*	*full*	85	47
	nil			
continuant	*continuant*	*noncontinuant*	86	45
frontback	*back*	*front*	84	59
manner	*vowel*	*fricative*	87	34
	approximant			
	nasal	*occlusive*		
phonation	*voiced*	*unvoiced*	93	63
place	*low*	*mid*	72	25
	high	*labial*		
	coronal	*palatal*		
	corono-dental	*labio-dental*		
	velar	*glottal*		
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 4: Results for the multi-valued feature system. All features can additionally take the value 'silence'.

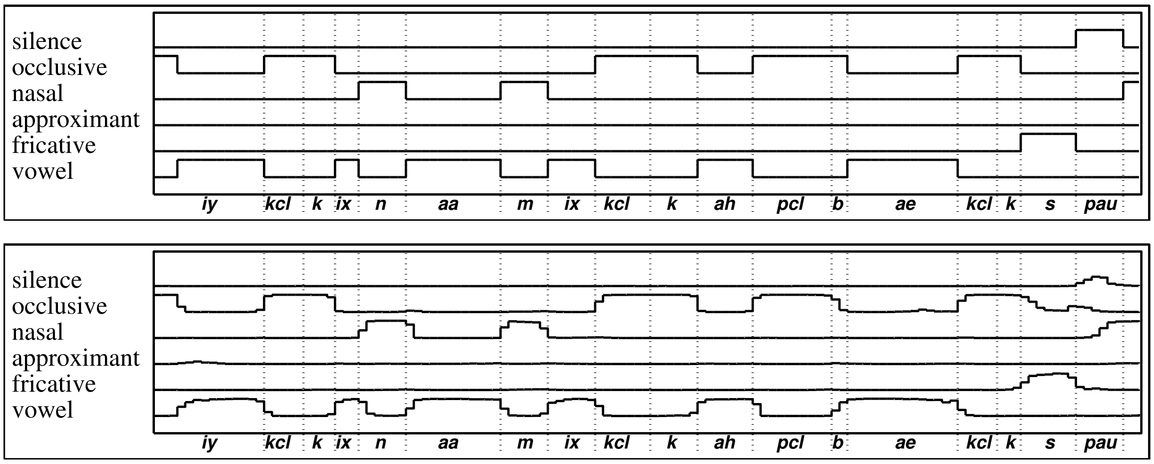

Figure 5: Example network output for the words "...economic cutbacks" for the manner feature of the multi-valued feature system. The top plot shows the target values as derived from the canonical phone representation. The bottom plot shows the output of the neural net. Compare with figures 4 and 6.

	silence	*approximant*	*fricative*	*nasal*	*occlusive*	*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 5: 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.

5. Government Phonology primes

In *Government phonology* [15], 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.

		aa	ow	p	l	n	f
Primes	**A**	✓	✓		✓	✓	
	I						
	U		✓	✓			✓
	@						
	?			✓	✓	✓	
	h			✓			✓
	H			✓			✓
	N					✓	
Head	**a**	✓					
	i						
	u		✓				

Table 6: Government Phonology primes for some phones from the TIMIT set.

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 [26]. 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.

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. In this way, a network can be trained to recognise a GP representation in the same way as for the other features. To represent the set of TIMIT segments, we allow three of the primes to be the head: **A**, **I** and **U**. Table 6 shows the GP primes for some example phones from the TIMIT set. Table 7 shows the results for the GP system. Again all features are recognised with high accuracy compared with the chance levels. Figure 6 shows the network output for an utterance from the test set.

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 7: Results for Government Phonology primes.

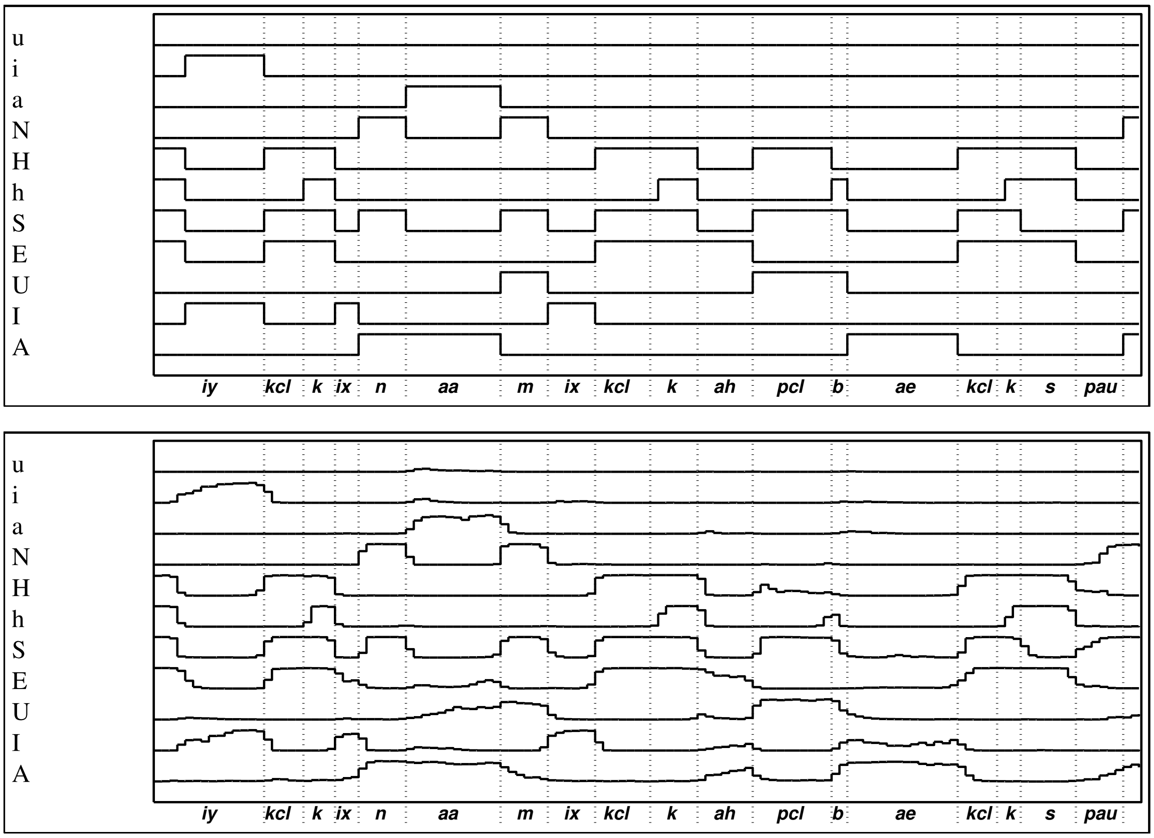

Figure 6: Example NN output for the words "...economic cutbacks" for the government phonology system. The top plot shows the target values as derived from the canonical phone representation. The bottom plot shows the output of the neural net. Compare with figures 4 and 5.

6. Discussion

The primary aim of this paper has been to describe techniques for the detection of phonological features from continuous speech. However, it is now worth discussing some issues concerned with actual *recognition*, that is the conversion of feature descriptions for an utterance into linguistic units such as phones or words. The long term goal is to develop or adapt statistical models which 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. 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 that will prove of no great benefit towards solving the larger problem.

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 a phone recognition experiment on TIMIT with a simple HMM speech recogniser. This used tied-state, cross word triphone models, and a mixture of Gaussians 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 (but not statistically significantly so) phone recognition accuracy of 63.5%.

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. We propose that phonological features are potentially a more useful modelling domain than MFCCs. Kirchhoff [22] 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 [4, 37] examine frequency bands separately and are motivated by 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.* [33] 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 [36, 27] 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. Bridle *et al* [6, 31] describe a hidden dynamic model in which deterministic segmental descriptions (targets) are filtered by a dynamical system model to simulate the effects of co-articulation and critical articulators, before being passed through a non-linear mapping (neural network) to generate acoustic vectors. These approaches are ideally suited to our task as they model asynchrony inherently.

It is useful at this stage to say something about the nature of the features with regard to asynchrony. In section 2.1 we discussed the issue of how concrete or abstract the feature representations should be, and explained that the more concrete (i.e. close to the acoustics) the representations are the more they could be expected to exhibit asynchrony. While the neural networks were trained on feature values which switched instantaneously at phone 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 4). To measure the size of this affect, we calculated the framewise classification accuracy if the features values were allowed some leeway near phone boundaries. Taking each feature individually, and examining two consecutive phones with differing values for that feature, if the value of the feature is correct for the first phone up to a point within two frames of the boundary (before or after), and is the correct value for the second phone after that point, then all the frames around the boundary of those two phones are also taken as correct. In other words, the feature must make the correct transition between the two values either side of the boundary ($+$ to $-$, or $-$ to $+$), but the timing of this transition may be up to two frames away from the reference phone boundary. This measure relaxes the requirement that feature values must change at exactly at phone boundaries. Using this reclassification on the SPE features, 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.

Given that it has been shown that neural networks can produce reasonable 1 from N phone classifications in hybrid systems, it is valid to ask at this point whether our networks are really performing feature detection, or whether they are just doing phone classification as with the hybrid systems, and then rewriting these phone classifications as features using a learned feature table. The proven asynchronous nature of the network behaviour shows this to some extent, but a further experiment was performed to show that more than phone boundary dynamics are involved. To test that the networks were really learning features, and not phones and converting them to features, we repeated the entire experiment (preparing the data, training a network and scoring the output) using SPE features from section 3 but with a randomised phone-to-feature-value table. The original table had one row per phone and one column per feature. The new table was made by randomly re-ordering the entries within each column. The re-ordering was different for each column. All phones *still had unique feature values* after the re-ordering. The original result was that all features were correct together in 52% of frames. For the randomised feature system, this figure drops to 37%. If the net was (internally) performing phone classification, then mapping to a binary feature representation, we would expect the two results to be the same.

A final point worth discussing concerns the nature and design of feature sets. The three sets we have tested here can be thought of as being representative of three different generations of research in phonological features, with the multi-valued set being the most traditional, the SPE set representing the original generative tradition and the GP set representing more current phonological theory. These three systems were chosen to be representative of different traditions in phonology, but there are many more feature sets and phonological theories that are equally valid candidates for the basis of recognition. While some have features in common with the three sets described here, many are completely different. It is therefore valid to ask whether it is sensible to base a recogniser on a particular phonological theory given that there is so much disagreement in the linguistics literature over what the best theory should be. But while there is much variation in the feature systems themselves, these differences are not arbitrary, and actually phonologists are pretty much agreed as to what an ideal feature system should look like: that is they agree on the *desiderata* of a feature system. In essence, the perfect feature system will be compact, have independent features, which combine naturally with pronunciation mechanisms to describe naturally occurring patterns as simply as possible.

It is interesting to note how closely this relates to the desiderata of the properties of observations in a probabilistic system, namely that fewer features are better, independent features are better (so that correlations don't have to be learned) and that features which combine naturally with the structure of the probabilistic model are also better (all other things being equal). Hence the goals of feature system design for phonologists and statisticians are the same. This leads us to our final conclusion, which is to say that perhaps self organising algorithms and other statistical mechanisms could be employed to learn more optimal feature representations, for use in both phonology and speech recognition.

Appendix

The following tables were used to map between TIMIT phone labels (the full set of 61 phones) and phonological feature values (refer to tables 4 and 7 for the full names of feature values in the MV and GP systems). Note that not all 61 phones have unique feature values, but when collapsed to the reduced set of 39, they do.

SPE feature system

phone	vocalic	consonantal	high	back	low	anterior	coronal	round	tense	voice	continuant	nasal	strident	silence
aa	+	-	-	+	+	-	-	-	+	+	+	-	-	-
ae	+	-	-	-	+	-	-	-	+	+	+	-	-	-
ah	+	-	-	+	-	-	-	-	-	+	+	-	-	-
ao	+	-	-	+	+	-	-	+	+	+	+	-	-	-
aw	+	-	-	+	+	-	-	+	+	+	+	-	-	-
ax	+	-	-	+	-	-	-	-	-	+	+	-	-	-
ax-h	+	-	-	+	-	-	-	-	-	+	+	-	-	-
axr	+	-	-	-	-	-	-	-	-	+	+	-	-	-
ay	-	-	-	+	+	-	-	-	+	+	+	-	-	-
b	-	+	-	-	-	+	-	-	-	+	-	-	-	-
bcl	-	+	-	-	-	+	-	-	-	-	-	-	-	-
ch	-	+	-	+	-	-	+	-	-	-	-	-	+	-
d	-	+	+	-	-	+	+	-	-	+	-	-	-	-
dcl	-	+	-	-	-	+	+	-	-	-	-	-	-	-
dh	-	+	-	-	-	+	+	-	-	+	+	-	-	-
dx	+	+	-	-	-	+	+	-	-	+	-	-	-	-
eh	-	-	-	-	-	-	-	-	-	+	+	-	-	-
el	-	+	-	-	-	+	+	-	-	+	+	-	-	-
em	-	+	-	-	-	+	-	-	-	+	-	+	-	-
en	-	+	-	-	-	+	+	-	-	+	-	+	-	-
eng	-	+	-	+	-	-	-	-	-	+	-	+	-	-
er	+	-	-	-	-	-	-	-	-	+	+	-	-	-
ey	-	-	-	-	-	-	-	-	+	+	+	-	-	-
f	-	+	-	-	-	+	-	-	-	-	+	-	+	-
g	-	+	-	+	-	-	-	-	-	+	+	-	-	-
gcl	-	+	-	+	-	-	-	-	-	-	-	-	-	-
hh	-	+	-	-	+	-	-	-	-	-	+	-	-	-
hv	+	+	-	-	+	-	-	-	-	+	+	-	-	-
ih	+	-	-	-	-	-	-	-	-	+	+	-	-	-
ix	+	-	+	-	-	-	-	-	-	+	+	-	-	-

(SPE continued)

phone	vocalic	consonantal	high	back	low	anterior	coronal	round	tense	voice	continuant	nasal	strident	silence
iy	-	-	+	-	-	-	-	-	+	+	+	-	-	-
jh	-	+	+	-	-	-	+	-	-	+	-	-	+	-
k	-	+	+	+	-	-	-	-	-	-	-	-	-	-
kcl	-	+	+	+	-	-	-	-	-	-	-	-	-	-
l	-	+	+	-	-	+	+	-	-	+	+	-	-	-
m	-	+	-	-	-	+	-	-	-	+	-	+	-	-
n	-	+	-	-	-	+	-	-	-	+	-	+	-	-
ng	-	+	-	+	-	-	-	-	-	+	-	+	-	-
nx	+	+	+	-	-	+	+	-	-	+	-	+	-	-
ow	+	-	-	+	-	-	-	+	+	+	+	-	-	-
oy	-	-	-	+	+	-	-	-	+	+	+	-	-	-
p	-	+	+	-	-	+	-	-	-	-	-	-	-	-
pcl	-	+	-	-	-	+	-	-	-	-	-	-	-	-
q	-	+	-	+	+	-	-	-	-	-	-	-	-	-
r	+	+	-	-	-	-	+	-	-	+	+	-	-	-
s	-	+	-	-	-	+	+	-	-	-	+	-	+	-
sh	-	+	+	-	-	-	+	-	-	-	+	-	+	-
t	-	+	-	-	-	+	+	-	-	-	-	-	-	-
tcl	-	+	-	-	-	+	+	-	-	-	-	-	-	-
th	-	+	-	-	-	+	+	-	-	+	+	-	-	-
uh	+	-	+	+	-	-	-	-	-	+	+	-	-	-
uw	+	-	+	+	-	-	-	+	+	+	+	-	-	-
ux	+	-	+	+	-	-	-	+	+	+	+	-	-	-
v	-	+	-	-	-	+	-	-	-	+	+	-	+	-
w	-	+	+	-	-	-	-	+	-	+	+	-	-	-
y	-	+	+	-	-	-	-	-	-	+	+	-	-	-
z	-	+	-	-	-	+	+	-	-	+	+	-	+	-
zh	-	+	+	-	-	-	+	-	-	+	+	-	+	-
sil	-	-	-	-	-	-	-	-	-	-	-	-	-	+

MV feature system

phone	phonation	manner	place	frontback	roundness	centrality	phone	phonation	manner	place	frontback	roundness	centrality
aa	v	v	lo	b	u	c	iy	v	v	h	f	u	f
ae	v	v	lo	f	u	f	jh	v	f	c	f	u	n
ah	v	v	lo	f	u	f	k	uv	o	v	b	u	n
ao	v	v	m	b	r	f	kcl	uv	o	v	b	u	n
aw	v	v	m	b	r	f	l	v	a	c	f	u	n
ax	v	v	m	n	u	c	m	v	n	l	f	u	n
ax-h	v	v	m	n	u	c	n	v	n	c	f	u	n
axr	v	v	m	n	u	c	ng	v	n	v	b	u	n
ay	v	v	m	f	u	f	nx	v	n	c	f	u	n
b	v	o	l	f	u	n	ow	v	v	h	b	r	f
bcl	uv	o	l	f	u	n	oy	v	v	h	b	r	f
ch	uv	f	c	f	u	n	p	uv	o	l	f	u	n
d	v	o	c	f	u	n	pcl	uv	o	l	f	u	n
dcl	uv	o	c	f	u	n	q	uv	o	g	b	u	n
dh	v	f	d	f	u	n	r	v	a	v	b	u	n
dx	v	o	c	f	u	n	s	uv	f	c	f	u	n
eh	v	v	m	f	u	f	sh	uv	f	c	f	u	n
el	v	a	c	f	u	f	t	uv	o	c	f	u	n
em	v	n	l	f	u	n	tcl	uv	o	c	f	u	n
en	v	n	c	f	u	n	th	v	f	d	f	u	n
eng	v	n	v	b	u	n	uh	v	v	h	b	r	f
er	v	a	v	b	u	f	uw	v	v	h	b	r	f
ey	v	v	h	f	u	f	ux	v	v	h	b	r	f
f	uv	f	d	f	u	n	v	v	f	d	f	u	n
g	v	o	v	b	u	n	w	v	a	l	f	r	n
gcl	uv	o	v	b	u	n	y	v	a	v	b	u	n
hh	uv	f	g	b	u	n	z	v	f	c	f	u	n
hv	v	f	g	b	u	n	zh	v	f	c	f	u	n
ih	v	v	h	f	u	f	sil	s	s	s	s	s	s
ix	v	v	h	f	u	f							

GP feature system

phone	A	I	U	E	S	h	H	N	a	i	u
aa	+	-	-	-	-	-	-	-	+	-	-
ae	+	-	-	-	-	-	-	-	-	-	-
ah	-	-	+	+	-	-	-	-	-	-	-
ao	+	-	+	+	-	-	-	-	-	-	+
aw	+	-	+	-	-	-	-	-	+	-	+
ax	-	-	-	+	-	-	-	-	-	-	-
ax-h	-	-	+	+	-	+	-	-	-	-	-
axr	+	-	-	+	-	-	-	-	-	-	-
ay	+	+	-	-	-	-	-	-	+	+	-
b	-	-	+	-	+	+	-	-	-	-	-
bcl	-	-	+	-	+	-	-	-	-	-	-
ch	-	+	-	-	+	-	+	-	-	-	-
d	+	-	-	-	+	+	-	-	-	-	-
dcl	+	-	-	-	+	-	-	-	-	-	-
dh	+	-	-	-	-	+	-	-	-	-	-
dx	+	-	-	+	+	-	-	-	-	-	-
eh	+	+	-	-	-	-	-	-	-	-	-
el	+	-	-	-	+	-	-	-	-	-	-
em	-	-	+	-	+	-	-	+	-	-	-
en	+	-	-	-	+	-	-	+	-	-	-
eng	-	-	-	+	+	-	-	+	-	-	-
epi	-	-	-	-	-	-	-	-	-	-	-
er	+	-	-	+	-	-	-	-	-	-	-
ey	+	+	-	-	-	-	-	-	-	+	-
f	-	-	+	-	-	+	+	-	-	-	-
g	-	-	-	-	+	+	-	-	-	-	-
gcl	-	-	-	-	+	-	-	-	-	-	-
hh	-	-	-	-	-	+	+	-	-	-	-
hv	-	-	-	-	-	+	-	-	-	-	-
ih	-	+	-	-	-	-	-	-	-	-	-
ix	-	+	-	+	-	-	-	-	-	-	-

phone	A	I	U	E	S	h	H	N	a	i	u
iy	-	+	-	-	-	-	-	-	-	+	-
jh	-	+	-	-	+	-	-	-	-	-	-
k	-	-	-	+	+	+	+	-	-	-	-
kcl	-	-	-	+	+	-	+	-	-	-	-
l	+	-	-	-	+	-	-	-	-	-	-
m	-	-	+	-	+	-	-	+	-	-	-
n	+	-	-	-	+	-	-	+	-	-	-
ng	-	-	-	+	+	-	-	+	-	-	-
nx	+	-	-	-	+	-	-	+	-	-	-
ow	+	-	+	-	-	-	-	-	-	-	+
oy	+	+	+	-	-	-	-	-	-	+	+
p	-	-	+	-	+	+	+	-	-	-	-
pau	-	-	-	-	-	-	-	-	-	-	-
pcl	-	-	+	-	+	-	+	-	-	-	-
q	-	-	-	-	+	-	-	-	-	-	-
r	+	-	+	+	-	-	-	-	-	-	-
s	-	-	-	+	-	+	+	-	-	-	-
sh	-	+	-	-	-	+	+	-	-	-	-
t	+	-	-	-	+	+	+	-	-	-	-
tcl	+	-	-	-	+	-	+	-	-	-	-
th	+	-	-	-	-	+	+	-	-	-	-
uh	-	-	+	+	-	-	-	-	-	-	-
uw	-	-	+	-	-	-	-	-	-	-	-
ux	-	-	+	-	-	-	-	-	-	-	-
v	-	-	+	-	-	+	-	-	-	-	-
w	-	-	+	-	-	-	-	-	-	-	-
y	-	+	-	-	-	-	-	-	-	-	-
z	-	-	-	+	-	+	-	-	-	-	-
zh	-	+	-	-	-	+	-	-	-	-	-
sil	-	-	-	-	-	-	-	-	-	-	-

References

[1] A. M. A. Ali, Jan Van der Spiegel, Paul Mueller, Gavin Haentjens, and Jeffrey Berman. An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. In *Proc. ISCAS*, volume 3, pages 118–121. IEEE, May 1999.

[2] Nabil N. Bitar and Carol Y. Espy-Wilson. A signal representation of speech based on phonetic features. In *Proc. 1995 IEEE Dual-Use Technologies and Applications Conference*, pages 310–315, May 1995.

[3] Nabil N. Bitar and Carol Y. Espy-Wilson. A knowledge-based signal representation for speech recognition. In *Proc. ICASSP '96*, pages 29–32, Atlanta, Georgia, 1996.

[4] Hervé Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. ICSLP '96*, pages 426–429, Philadelphia, 1996.

[5] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid approach*. Kluwer Academic Publishers, 1994.

[6] 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.

[7] Noam Chomsky and Morris Halle. *The Sound Pattern of English*. MIT Press, 1968.

[8] Li Deng and Don X. Sun. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *J. Acoust. Soc. America*, 95:2702–2719, May 1994.

[9] Li Deng and Jim Jian-Xiong Wu. Hierarchical partition of the articulatory state space for overlapping-feature based speech recognition. In *Proc. ICSLP '96*, volume 4, pages 2266–2269, Philadelphia, 1996.

[10] Kevin Erler and George H. Freeman. An HMM-based speech recognizer using overlapping articulatory features. *J. Acoust. Soc. Am.*, 100(4):2500–2513, 1996.

[11] Carol Y. Espy-Wilson and Nabil N. Bitar. Speech parameterization based on phonetic features: application to speech recognition. In *Proc. Eurospeech-95, Madrid*, pages 1411–1414, Sept. 1995.

[12] 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.

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

[14] J. Harrington. Acoustic cues for automatic recognition of English consonants. In M. A. Jack and J. Laver, editors, *Speech Technology: a survey*, pages 19–74. Edinburgh University Press, 1987.

[15] John Harris. *English Sound Structure*. Blackwell, 1994.

[16] M. A. Huckvale. Word recognition from tiered phological models. *Proc. Institute of Acoustics Conference on Speech and Hearing*, 16(5), 1994.

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

[18] Frederick Jelinek. *Statistical methods for speech recognition*. MIT Press, 1997.

[19] Daniel Jones. *An Outline of English Phonetics*. Cambridge: Heffer & Sons, 8 edition, 1957.

[20] Katrin Kirchhoff. Syllable-level desynchronisation of phonetic features for speech recognition. In *Proc. ICSLP ‘96*, volume 4, pages 2274–2276, Philadelphia, 1996.

[21] Katrin Kirchhoff. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proc. ICSLP ‘98*, Sydney, Australia, December 1998.

[22] Katrin Kirchhoff. *Robust Speech Recognition Using Articulatory Information*. PhD thesis, Universität Bielefeld, 1999.

[23] Katrin Kirchhoff and Jeff A. Bilmes. Statistical acoustic indications of coarticulation. In *Proc. ICPhS 99*, San Francisco, August 1999.

[24] Jacques Koreman, Bistra Andreeva, and Helmer Strik. Acoustic parameters versus phonetic features in asr. In *Proc. ICPhS 99*, pages 719–722, San Francisco, August 1999.

[25] Aditi Lahiri. Speech recognition with phonological features. In *Proc. ICPhS 99*, pages 715–718, San Francisco, August 1999.

[26] J. P. Olive, A. Greenwood, and J. Coleman. *The Acoustics of American English Speech, a Dynamic Approach*. Springer, New York, 1993.

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

[28] George Papcun, Judith Hochberg, Timothy R. Thomas, François Laroche, and Jeff Zacks. 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, 1992.

[29] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech signal processing*. Prentice Hall, 1993.

[30] Henning Reetz. Converting speech signals to phonological features. In *Proc. ICPhS 99*, pages 1733–1736, San Francisco, August 1999.

[31] Hywel B. Richards and John S. Bridle. Acoustic-phonetic modelling using the hidden dynamic model. In *Proc. ICPhS 99*, pages 691–694, San Francisco, August 1999.

[32] Tony Robinson, Mike Hochberg, and Steve Renals. The use of recurrent neural networks in continuous speech recognition. In C-H. Lee and F. K. Soong, editors, *Automatic Speech and Speaker Recognition - Advanced Topics*, chapter 11, pages 233–258. Kluwer Academic Publishers, 1996.

[33] Shigeki Sagayama, Shigeki Matsuda, Mitsuru Nakai, and Hiroshi Shimodaira. Asynchronous-transition hmm for acoustic modeling. In *International Workshop on Automatic Speech Recognition and Understanding*, volume 1, page 99, Keystone, Colorado, December 1999.

[34] Alain Soquet, Marco Saerens, and Véronique Lecuit. Complementary cues for speech recognition. In *Proc. ICPhS 99*, pages 1645–1648, San Francisco, August 1999.

[35] Nikko Ström. *The NICO Artificial Neural Network Toolkit*. `http://www.speech.kth.se/NICO`, 1996.

[36] J. Takami and S. Sagayama. A successive state splitting algorithm for efficient allophone modelling. In *Proc. ICASSP 92*, volume I, pages 573–576, Minneapolis, April 1992.

[37] Sangita Tibrewala and Hynek Hermansky. Sub-band based recognition of noisy speech. In *Proc. ICASSP '97*, pages 1255–1258, Munich, Germany, April 1997.

[38] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, and S.J. Young. The 1994 HTK large vocabulary speech recognition system. In *Proc. ICASSP '95*, Detroit, May 1995.

[39] Steve Young, Joop Jansen, Julian Odell, Dave Ollason, and Phil Woodland. *HTK manual*. Entropic, 1996.

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