SPEECH RECOGNITION VIA PHONETICALLY FEATURED SYLLABLES

Simon King *Todd Stephenson* *Stephen Isard* *Paul Taylor* *Alex Strachan*

Centre for Speech Technology Research,
University of Edinburgh, 80,South Bridge, Edinburgh EH1 1HN, UK
www.cstr.ed.ac.uk Simon.King@ed.ac.uk

ABSTRACT

We describe a speech recogniser which uses a speech production-motivated phonetic-feature description of speech. We argue that this is a natural way to describe the speech signal and offers an efficient intermediate parameterisation for use in speech recognition. We also propose to model this description at the syllable rather than phone level.

The ultimate goal of this work is to generate syllable models whose parameters *explicitly* describe the *trajectories* of the phonetic features of the syllable. We hope to move away from Hidden Markov Models (HMMs) of context-dependent phone units. As a step towards this, we present a preliminary system which consists of two parts: recognition of the phonetic features from the speech signal using a neural network; and decoding of the feature-based description into phonemes using HMMs.

1. INTRODUCTION

We will first discuss some of the shortcomings of the context-dependent phone HMM approach, then suggest the syllable as an alternative unit, with models of syllables specified in terms of phonetic features. HMM systems were trained to recognise phones from MFCCs (Mel Frequency Cepstral Coefficients) and from the feature representation.

1.1. The Problems with HMMs of Phones

Markov assumption The Markov assumption, namely that the observations generated by the states of a (Hidden) Markov Model are independent of one another, is not true for speech. Speech is produced by movements of articulators, and therefore, in some space, speech is constrained to follow a smooth trajectory with occasional abrupt accelerations. Smoothly moving articulators *can* produce sharply changing acoustics – when the lips open during the production of [p], for example. Therefore, we can say that there is some parametric description of speech, perhaps in terms of the articulators, in which the parameters follow piecewise smooth trajectories.

Trajectories The Markov property can be relaxed, so that states generate observation sequences, or *trajectories*. Such a model has been called a Segmental HMM [5], which is one of a family of stochastic models [6] which includes conventional HMMs. Constraints on the trajectory are applied in the model parameter space, which is generally also the observation (e.g. MFCCs) space since model parameters are generally means and variances of Gaussian distributions. The Markov assumption is still made for observations from different states, but the observations from any single state are no longer independent of each other. We wish to go one step further than this, and model trajectories at the syllable level, and, moreover, allow for temporal misalignment of those trajectories. We believe that the trajectories should be in a speech production-based feature space. The experimental work described here is a step towards such a model.

Context dependency HMM-based systems model co-articulation effects with context-sensitive models. In effect, a different variant of each phoneme model is required for each possible context. The large number of models required means that techniques for reducing the number of model parameters are used: state or mixture tying, for example. Typical HMM systems use decision trees to perform state tying. These trees use ad hoc questions which are typically about phonetic features. We want a more principled system than this, so we argue against the approach of assuming that all phonetic contexts are different, then grouping them together to reduce the number of parameters. Co-articulation is not simply a function of phonetic context. It depends rather on some *properties* of the context (articulatory targets or trajectories, perhaps), speech rate, position within a syllable, and so on. In other words, co-articulation should not be modelled as phonetic-context sensitivity, but more explicitly.

2. PHONETIC FEATURES

2.1. Feature system

The choice of feature system must consider both its descriptive power (its ability to distinguish all the required phoneme or syllable units) and the ease of automatically recognising the features from the speech signal. We will refer to the types of feature systems of interest as speech production-based phonetic features, or simply *phonetic features*, because they describe the way the sound is produced (place of articulation, for example).

A variety of feature systems exist. Binary systems, such as that of Chomsky and Halle [2], can be designed so that any combination

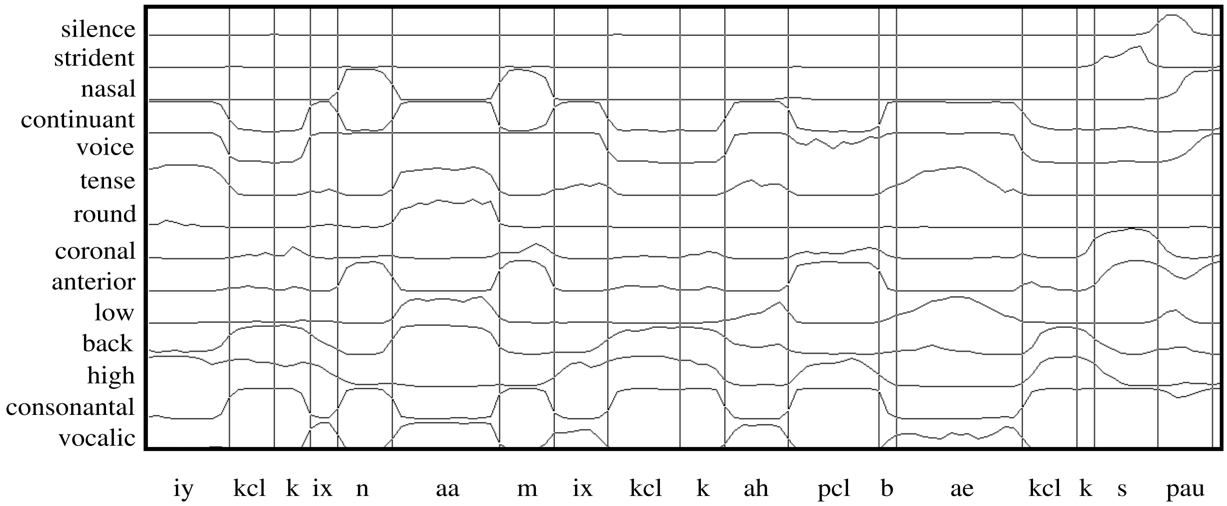

Figure 1: Example NN output for the SPE feature system. The utterance is part of TIMIT sentence dr4/mbns0/sx50 and is "...economic cutbacks..". Each feature takes values from 0 to 1. The phonetic labels are the correct transcription.

of values is allowed. In other words, setting the values of the features to any combination of + and - produces a description of a legitimate segment (although not necessarily one used in the particular language in question).

Multi-valued systems, on the other hand, tend to use fewer features, but allow these features to take more than two values. A typical feature in such a system might be place of articulation, which could take values *velar*, *dental*, *labial*, and so on. We considered one feature system of each type : Chomsky and Halle's binary system [2], which will will refer to as *SPE*; and a multi-valued system adapted (for English) from [7], which we will call *MV*.

2.2. Automatic feature detection

The output from the first component of the recogniser will be a set of parallel feature streams; an example is shown in figure 1. This representation allows temporal overlap, or misalignment, of features. In other words, features do not all change value simultaneously – at phone boundaries, for example – but tend to be staggered in time. Such overlap results from co-articulation and we believe that this representation allows the modelling of contextual effects more effectively than conventional triphone based recognisers.

Other work on automatic feature detection includes [1] in which the features are strictly acoustic – energy in certain frequency bands, for example. Such features, although intended to be phonetic, are more closely related to MFCCs than true phonetic features, and indeed, perform almost identically in speech recognition applications.

To train an automatic feature detector, we need data labelled with feature values. Since no such database exists, we must generate these feature labels from phonetic labels. This is far from ideal, since the labels generated in this way will not show all the effects of overlap that we intend to model. In future, embedded training techniques (repeated cycles of training and automatic segmentation using the trained models) may be able to improve the situation. The TIMIT [4] database was used for all experiments, because it is labelled throughout with phone and word boundaries. The training/testing division was the official one; the SA sentences were omitted, leaving approx. 3600 training sentences and 1300 test sentences.

Neural networks Two differing approaches were used. For the MV feature system, one neural network (NN) was trained for each feature; for the SPE feature set, a single network was trained to detect all features.

In both systems, the NNs were multilayer networks with one hidden layer, recurrent time-delaying connections and an input consisting of a total of 7 frames of context. Each frame is parameterised as 12 MFCCs plus energy. First and second derivatives were computed by special units. All nets were sparsely connected (only 1 in 4 possible connections between layers were made). The connectivity and time window arrangement were inspired by [10]. Frames are 25ms in duration and spaced 10ms apart. The Nico toolkit [10] was used for all experiments. The network sizes and training parameters were roughly optimised on a validation set – 100 files held out from the training set. No speaker appeared in both training and validation sets. The test set was only used for final evaluation. The total number of frames in the test set is $\sim 411\,000$.

Multi-valued features This system employed a number of smaller networks, each performing a 1-of-N classification task. The 8 features and their possible values are shown in table 1, along with the recognition accuracies for each net. The bottom line in the table indicates the percentage of frames in which all 8 features are assigned the correct value. This is effectively a frame-

Feature	Values		Frames correct (%)	
			NN	HMM
centrality	central nil	full	85	73
continuant	continuant	noncontinuant	86	n/a
frontback	back	front	84	64
manner	vowel approximant nasal	fricative occlusive	87	75
phonation	voiced	unvoiced	93	87
place	low high coronal corono-dental velar	mid labial palatal labio-dental glottal	72	61
roundness	round	non-round	92	83
tenseness	lax	tense	87	n/a
All correct together			53	32

Table 1: The multi-valued feature system. All features can additionally take the value 'silence'. Performance is measured on the full test set.

wise phone classification result, except the classification space contains nearly 6000 feature combinations (the product of the number of network outputs), and not just the 39 in the TIMIT phone set, for example. The confusion matrix for the manner feature is shown in table 2.

	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 2: Confusion matrix for the 'manner' neural network. All figures are percentage of frames correct.

SPE features The SPE feature system has 13 features. A single network was trained to recognise all features simultaneously, with an additional network output for 'silence'. All networks had 13 inputs and 14 outputs and the same context window and derivatives as the MV networks. Various numbers of hidden units were used, and a network with 250 hidden units 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 3 and typical network output is shown in figure 1. For these results, network outputs were thresholded (values over 0.5 become 1, the rest become 0). 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, it should be noted that on average, feature values are '0' 70% of the time and '1' 30% of the time; for some features, the value is '0' more than 90% of the time (94% for nasal). 14% of frames are silence.

Feature	Frames correct (%)		Feature	Frames correct (%)	
	NN	HMM		NN	HMM
vocalic	88	65	**consonantal**	90	n/a
high	86	83	**back**	88	58
low	93	87	**anterior**	90	71
coronal	90	70	**round**	94	63
tense	91	56	**voice**	93	76
continuant	93	65	**nasal**	97	68
strident	97	83			
All correct together				51	11

Table 3: The binary-valued feature system from [2]. Performance is measured on the full test set.

As for the MV feature results, the bottom line of table 3 shows the percentage of frames in which all 14 network outputs were correct; this figure includes the 'silence' output of the network. The space of feature combinations is effectively $2^{13}+1 = 8193$, since one output signifies silence. Fewer than 1% of these combinations are used in English.

Hidden Markov Models Mainly for comparison with the NN approach, the method in [7] was repeated for the TIMIT data. HMMs were used to recognise regions of the speech signal with feature values. Each feature was recognised independently: our approach was exactly as if we were doing phone recognition. Taking voicing (phonation) as an example, there are three models: voiced, unvoiced and silence. The "language model" was either a simple loop which allowed any sequence of these three values (but not two consecutive regions with the same value), or a bigram trained on data. The observation vectors were composed of 12 MFCCs and energy, with their 1st and 2nd derivatives (39 components). The training of the HMMs was not discriminative (in contrast with the NNs). This system produces feature value *labels* for speech, and not continuously valued features.

In order to compare the accuracy of the HMM systems to the NN systems, the results were converted into frame accuracies, as shown in tables 1 and 3. The HMM experiments were performed independently of the NN ones, and consequently there were are minor differences in the feature systems - those features not used in the HMM systems are indicated by *n/a* in the tables.

2.3. Analysis

The results in tables 1 and 3 show that the NN systems were more accurate for both feature systems. Furthermore, the nature of their output – continuously valued features – is preferable to the HMM symbolic output since it can be interpreted as feature value posterior probabilities. The superior performance of the NN systems may be because they use a longer context window and were discriminatively trained. Only the output from the NN system was used in the recognition experiments.

For the SPE feature system, the largest network (250 hidden units) gave the best performance. Training time for such large networks is considerable, even with sparse connectivity (12 hrs per epoch on a Sun Ultra 10); 14 epochs were required to achieve the quoted performance. We intend to experiment with larger numbers of hidden units and different degrees of connectivity.

The NN phoneme classification frame accuracies of around 52% are similar to results reported in [3], in which a NN using phonetic features was used for recognition of single-speaker data (Swedish and Hungarian).

3. RECOGNITION EXPERIMENTS

Only the output of the MV neural network system was used in the recognition experiments. This system gave marginally better frame accuracy than the SPE system. However, the SPE feature system is more compact, and future experiments will explore its use in syllable recognition.

3.1. Phones

As a test of the information content of the NN feature output, phone HMMs were trained. Models were trained on both MFCCs and the NN output for comparison, and the results are shown in table 4. In both cases the HMMs were tied state, cross-word triphones with single Gaussian observation densities, trained with HTK, and a simple phone bigram language model was used. The best reported performances on this task are summarised in [9] and are around 28% for HMM and NN-HMM hybrids, and around 36% for segment-type models. [8] contains a good review of segment models. Although our phone recogniser does not achieve state-of-the art performance, it compares well with segmental models.The feature representation carries as much information as MFCCs as far as phone recognition is concerned.

System	Phone error rate (%)
HMMs using MFCCs directly	36.7 %
NN feature detector + HMMs	36.5 %

Table 4: Phone recognition from feature-detecting NN output. The phone set is the standard 39 phone TIMIT set. The performance is quoted for the full test set. Error rate = 100 - accuracy.

4. CONCLUSION

We have introduced a new method for speech recognition, which shows promising results. The method allows explicit modelling of coarticulation effects by using a phonetic feature representation of the speech signal. We have achieved a high accuracy mapping from acoustics to this representation using neural networks, and have demonstrated the potential of the phonetic features for speech recognition.

Future work may use segment models of the types surveyed in [8], with explicit, parametric models of feature trajectories. Early work suggests that cubic polynomials are a reasonable fit to observed feature values within syllables.

Acknowledgements

SK is funded by EPSRC *Realising Our Potential Award* number GR/L59566. PT is funded by EPSRC grant number GR/L53250. TS and AS are students at the University of Edinburgh.

5. REFERENCES

1. N. Bitar and C. Espy-Wilson. A knowledge-based signal representation for speech recognition. In *Proc. ICASSP '96*, pages 29–32, Atlanta, Georgia, 1996.
2. Noam Chomsky and Morris Halle. *The Sound Pattern of English*. MIT Press, 1968.
3. Kjell Elenius and György Takás. Acoustic-phonetic recognition of continuous speech by artificial neural networks. Technical Report STL-QPSR 2-3/1990, Institutionen för tal, musik och hörsel, KTH, Stockholm, Sweden, 1990.
4. 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.
5. Wendy J. Holmes and Martin J. Russell. Speech recognition using a linear dynamic segmental hmm. In *Proc. Eurospeech-95, Madrid*, pages 1611–1614, Sept. 1995.
6. Ashvin Kannan and Mari Ostendorf. A comparison of constrained trajectory segment models for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6(3):303–306, May 1998.
7. Katrin Kirchhoff. Syllable-level desynchronisation of phonetic features for speech recognition. In *Proc. ICSLP '96*, volume 4, pages 2274–2276, Philadelphia, 1996.
8. M. Ostendorf, V. Digalakis, and O. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(5), Sept. 1996.
9. Tony Robinson. The application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2), March 1994.
10. Nikko Ström. Sparse connection and pruning in large dynamic artificial neural networks. In *Proc. Eurospeech 97*, volume 5, pages 2807–2810, Rhodes, Greece, 1997.