Final Report on Research Grant GR/L59566
ROPA: Phonetically-featured syllables for speech recognition

Stephen Isard and Simon King

October 1997 - September 1999

1 Project aims

We proposed to develop a new model for speech recognition based on phonetic features of syllables. The two aims of the project were therefore to automatically detect these features from the speech signal, and to develop a (syllable) model of these features.

We demonstrated the feasibility of our method by achieving almost identical syllable recognition rates on the TIMIT database to a conventional HMM system using MFCC observations directly. However, while there is a general impression in the speech recognition community that such conventional HMMs have been pretty thoroughly investigated and reached a performance plateau, there are a great many more avenues to be explored in connection to our new approach.

1.1 Report structure

This report is organised into three parts: first, we deal with the automatic detection of phonetic features from speech; second, we discuss syllable models for continuous speech recognition; finally, we summarise our findings and conclude with a plan for extending this work.

2 Phonetic feature detection

2.1 Feature systems

Three phonetic feature systems were examined:

SPE Chomsky and Halle's system from "The Sound Pattern of English" [5]

MV Multivalued place / manner type features

GP Government Phonology primes [8, 3, 4]

2.2 Data

All our experiments were conducted with the TIMIT corpus [7] of continuous, read speech. This corpus has the great advantage of being labelled throughout at the acoustic phonetic level.

Ideally, the database would be manually labelled according to each of the three chosen feature systems. This was impractical in terms of time and cost, so an alternative strategy was used: namely, mapping the existing phonetic labels into feature value labels. This is trivial, allowing us time to investigate all three feature systems, but has the serious disadvantage of placing feature value changes only at phone boundaries. This deficiency can be overcome by using training procedures that allow a re-alignment of labels with the speech signal – see section 2.3.2.

2.3 Models

Two different types of model were used to detect phonetic features from the speech signal: hidden Markov models (HMMs) and neural networks (NNs). Both these models have parameters which are trained from data.

The two models have differing capabilities. HMMs *segment* the signal into *classes* – for binary feature detection there will be two classes corresponding to presence and absence of the feature. NNs, on the other hand, map between input and output patterns, and can give continuously-valued output – so a binary-valued feature can be treated as continuously valued. This is advantageous for features like voicing, nasality, tongue height, etc.

2.3.1 Hidden Markov models

For a given feature system from the list in section 2.1, each feature is treated independently. We will take the *Place* feature from the *MV* system as an example. This feature can take one of 9 values : low, mid, high, labial, dental, coronal, velar, glottal and silence. An HMM is trained to recognise tokens of each of these values. These 9 models are then used exactly as phone models would be in speech recognition, and used to automatically label speech with a sequence of feature values.

On advantage of the HMM approach is that of "embedded training" in which the times of boundaries in the labelled training data are ignored, allowing the training algorithm to determine a more likely observation/model alignment. This alleviates the problem of phonetic feature labels derived only from phonetic labels mentioned in section 2.2.

A major disadvantage is that the output of these system is discrete (that is, the segmentation is "hard") rather then continuously-valued. For example, the system will always place a definite boundary between voiced and unvoiced regions, when in fact the transition between the two is gradual.

2.3.2 Neural networks

Neural networks with time-delaying recurrent connections, with a similar architecture to [18, 12], were trained on the data using the NICO toolkit [17].

In contrast to the HMM system, the NNs produce continuously-valued output, which we consider to be a major advantage. However, the training method for the NNs requires output activations (corresponding to feature-value labels for the HMMs) to be specified every frame. This means that "embedded training" is less straightforward than for HMMs. In systems like Abbot [13], so-called Viterbi training is performed. In short, this involves a re-alignment of labels and frames using the current, trained NN, and requires what amounts to constrained speech recognition. Forward-backward training of NNs is also possible [14], but has been shown to give only minor improvements over Viterbi training. A corresponding scheme for asynchronous features is possible in principle but we decided that developing one was too complicated for this project.

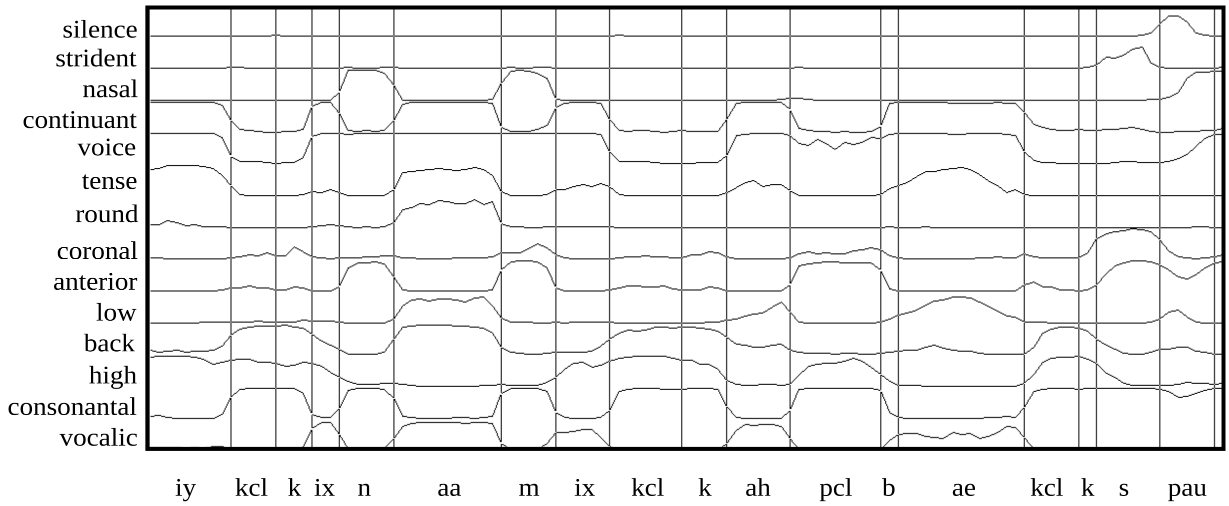

Figure 1: Example NN output for the SPE feature system. The utterance is part of TIMIT sentence dr4/mbns0/sx50 and is "...economic cutbacks..". Each feature takes values from 0 to 1. The phonetic labels are the correct transcription.

2.3.3 Results

Figure 1 shows an example of the NN output on a sentence unseen during training.

Objective comparison of results within and between feature systems is difficult for a number of reasons: the reference ("correct") transcription at the frame level was derived from phonetic labels and suffers from the defects mentioned in section 2.2; each feature is of varying importance in identifying the phone/syllable/word; the number of features in each system varies; the *GP* system is defined at the syllable rather than phoneme level.

Subjectively, results are impressive: a typical result is for the **manner** feature of the *MV* system, where 86% of frames were labelled by the NN with the correct value (out of 6 possible values) and in 98% of frames the correct value was amongst the top 2 most likely values estimated by the NN. Furthermore, when the most likely feature value output from the NN is the incorrect one, it tends to be similar to the correct value. For example, the NN detecting the **height** feature produces the following results on frames labelled as *low*: in 49% of frames the *low* output of the NN is highest; in 35% it is *mid*; 8% *high*; 8% *silence*. Full results are reported in [9] and [10].

We also looked at phone accuracy. For the *SPE* system, the average frames correct for individual features was 92%. All features had the correct value together in 52% of frames. By mapping each frame to the nearest valid feature combination (i.e. a phone), this increases to 59%. By correcting small timing errors near phone boundaries (we allowed features to change value up to 20ms early or late), the framewise phone accuracy is 70%.

3 Syllable modelling

Our proposal was that the phonetic features extracted from the speech signal should be modelled at the syllable level. The motivation for this was that coarticulation is stronger within syllables than across syllable boundaries.

The ultimate goal of this work was to find a model capable of expressing the asynchronous (and possibly continuous-valued) nature of the phonetic features. Conventional HMMs cannot express the asynchronous aspect. Alternative topologies for HMMs have been proposed [6], but we are seeking a more elegant and parsimonious solution.

3.1 HMMs

Initial experiments using conventional HMMs were undertaken to provide a baseline and to investigate the usability of the automatically extracted phonetic features.

3.1.1 Phone models

Two systems using cross-word tied-state triphone HMMs were trained and tested on the TIMIT database. The first system used conventional MFCC parameters (12 coefficients, plus energy, and the first and second derivatives of these); the second system used the output of the feature-detecting neural networks from section 2.3.2. The MV feature system was used, and the HMM observation vector comprised the concatenated outputs of the individual feature detecting networks (a total of 29 components).

3.1.2 Syllable models

For the TIMIT database, a set of 3413 word-internal syllable models was required. As for triphone HMMs systems, parameter tying was necessary to reduce the number of free parameters in the system and make training possible using limited data. A scheme was devised in which the onsets, nuclei and codas of syllables were tied across acoustically similar syllables. A decision tree procedure similar to that used for triphone clustering was used; some example questions used to build the tree for onsets are:

- Does **Phonation** = *voiced* occur anywhere in the onset ?
- Does **Phonation** = *voiced* occur in the final phone of the onset ?
- Does **Phonation** = *voiced* occur in the initial phone of the onset ?

There were also questions for combinations of feature values and for specific phones (which are simply complete feature value combination specifications). An advantage of the decision tree method is that any irrelevant questions in the list are simply not used to build the tree.

Two recognisers were trained: one using MFCCs, and one using the MV features automatically detected using NNs. On the full TIMIT test set, both systems achieved almost identical syllable error rates: 37% for the MFCC system and 38% for the MV system. This proves that the feature extraction process has not discarded essential information, but does not prove that the phonetic feature approach is superior to modelling MFCCs directly.

4 Summary and future direction

4.1 Major results

Automatic detection of phonetic features from the speech signal proved very successful – comparable to frame-by-frame estimation of posterior phone PDFs in systems like Abbot

[13].

The syllable modelling experiments showed that the MV phonetic feature system performs as well as MFCCs when using HMMs.

4.2 Student projects

This project provided a framework for several undergraduate and M.Sc. student projects whose reports are available from the project web pages at:

`http://www.cstr.ed.ac.uk/projects/espresso`

Undergraduate

- Alex Strachan used HMMs to label speech with *SPE* and MV features.

M.Sc. in Speech and Language Processing

- Stefanie Aalburg looked at pronunciation variation at the syllable level, with reference to phonetic features.
- Angela Michelfelder investigated a strictly bottom-up segmentation of articulatory data, looking for self-evident units in speech.
- Simon Ahern researched Government Phonology, and used neural networks to automatically extract GP primes from speech.

M.Sc. in Cognitive Science and Natural Language

- Todd Stephenson carried out experiments in MV feature detection using NNs, followed by syllable modelling using HMMs.

4.3 Extension of this work

One of the most interesting questions raised by this work is, can we find a better model (than the HMM) of the continuously-valued, asynchronous phonetic features? Clearly, we believe that we can, and this is the subject of our new project: "Continuous-state dynamical system models for speech recognition".

Items in the bibliography marked ★ are associated with this project.

References

[1] ★ Stefanie Aalburg. Syllable-based multiple-pronunciation recognition with discrete hidden markov models. Master's thesis, University of Edinburgh, 1997. `http://www.cstr.ed.ac.uk/projects/espresso`.

[2] ★ Simon Ahern. A government phonology approach to automatic speech recognition. Master's thesis, University of Edinburgh, 1999. `http://www.cstr.ed.ac.uk/projects/espresso`.

[3] W. Brockhaus. Skeletal and suprasegmental structure within government phonology. In Jacques Durand and Francis Katamba, editors, *Frontiers of Phonology: atoms, structures, derivations*, pages 180–221. Longman, 1995.

[4] W. Brockhaus and M. Ingleby. Unary primes and constituent structure. Unpublished research. University of Manchester, 1997.

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

[6] 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.

[7] 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.

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

[9] ★ Simon King, Todd Stephenson, Stephen Isard, Paul Taylor, and Alex Strachan. Speech recognition via phonetically featured syllables. In *Proc. ICSLP '98*, pages 1031–1034, Sydney, Australia, December 1998.

[10] ★ Simon King and Paul Taylor. Detection of phonological features in continuous speech using neural networks. *in preparation*, 2000.

[11] ★ Simon King and Alan Wrench. Dynamical system modelling of articulator movement. In *Proc. ICPhS 99*, pages 2259–2262, San Francisco, August 1999.

[12] Tony Robinson. The application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2), March 1994.

[13] 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 7. Kluwer Academic Publishers, 1996.

[14] Andrew Senior and Tony Robinson. Forward-backward retraining of recurrent neural networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. Morgan Kaufmann, 1996.

[15] ★ Todd Stephenson. Speech recognition using phonetically featured syllables. Master's thesis, University of Edinburgh, Sept. 1998. `http://www.cstr.ed.ac.uk/projects/espresso`.

[16] ★ Alexander Strachan. Final year project report on using HMMs to detect phonetic features from speech. `http://www.cstr.ed.ac.uk/projects/espresso`.

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

[18] Nikko Ström. Sparse connection and pruning in large dynamic artificial neural networks. In *Proc. Eurospeech 97*, volume 5, pages 2807–2810, Rhodes, Greece, 1997.