Summary report on EPSRC-funded research
ROPA grant number GR/L59566
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.

2 Experiments

Three phonetic feature systems were examined: *SPE* (Chomsky and Halle's system from "The Sound Pattern of English"); *MV* (Multivalued place / manner type features); *GP* (Government Phonology primes). We trained recurrent neural networks to recognise the feature value from the speech signal (the TIMIT database was used). For the TIMIT database, a set of 3413 word-internal syllable models was required. Two recognisers were trained: one using MFCCs, and one using the *MV* features automatically detected using NNs. Two HMM based syllable recognisers were trained; one using MFCCs, and one using features automatically detected using NNs. They achieved almost identical syllable error rates. This proves that the feature extraction process has not discarded essential information.

3 Summary and future direction

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

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".

For more information, contact Simon King (`Simon.King@ed.ac.uk`) or visit the project web pages at:

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