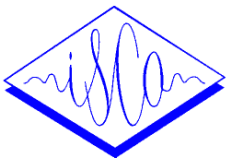

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

IMPROVING SPEECH RECOGNITION USING SPEAKER CLASSIFICATION

David O. Baldwin & Georg F. Meyer
e-mail: davidba@logica.com
Logica UK Limited
Betjeman House
104 Hills Road
Cambridge CB2 1LQ
UK

ABSTRACT

This work examines the use of speaker classification as a method of improving speech recognition. Basic speech recognisers based upon hidden Markov models and neural networks, are modified by the use of selective training. Speakers are clustered into speaker types and separate recognisers are trained for each type. Performance is shown not to improve significantly. A more individualistic system is proposed. The speaker space is mapped by the use of non-linear interpolation between speaker dependent recognisers. Performance using an abstract 'perfect' speaker classifier is shown to be significantly better than speaker independent recognition. A multi-layer perceptron based speaker classifier is introduced, but is shown to be unable to learn the mapping from speakers to recognisers.

10.21437/Eurospeech.1995-399

1. INTRODUCTION

1.1 Speech Recognition

Current automatic speech recognition systems are highly constrained. They typically only work with isolated words, limited vocabularies or only for single speakers. To improve speech recognition these constraints must be relaxed.

Psychophysical experiments show that human listeners 'adapt' to speakers so that single phonemes sound different, depending on the context ([1]). An ideal machine speech recognition system might adapt the pattern matching stage to the speaker currently using the system ([2],[3]). This work examines two strategies to achieve this goal.

1.2 Speaker Types

Two baseline recognition systems, based upon hidden Markov models (HMM) and multi-layer perceptrons (MLP), were used. Both were trained to recognise the four vowels /i/, /I/, /a/ and /o/, isolated from continuous speech, spoken by both male and female speakers (TIMIT database). Both systems were evaluated in a speaker independent mode and with two forms of speaker adaptation.

The basic hypothesis underlying this work is that it is possible to split speakers into groups of similar 'speaker type'. Speakers could be clustered using a variety of features that identify differences between speakers. However, speakers classified in this way may be similar with regard to the features used, but dissimilar with regard to the speech recognition process. To cluster speakers from the viewpoint of the recognisers, speaker dependent recognisers were trained on single speakers and clustering was performed on the recognisers themselves.

2. EXPERIMENTS

2.1 Training recognisers on speaker subsets

The MLPs and HMMs that perform the pattern matching in speech recognition are defined by parameters. Similar speakers produce similar recognisers with similar parameters. These parameters can be thought of as representing speakers as seen by the recognisers. Clustering on these parameters involved determining a suitable distance metric for the recognisers. For each cluster of speaker dependent recognisers, a *new* prototype recogniser was produced by training only on the speakers within that cluster. This new recogniser is 'speaker type' dependent. i.e. It is representative of only a subsection of all the speakers, and so will be tuned to the 'speaker type' of the cluster.

For an unknown speaker, if it is possible to determine what type of speaker he or she is, we can then use the 'speaker type' dependent recogniser to perform the speech recognition, resulting in improved performance.

2.2 Results

The upper bounds of system performance can be determined by assuming the best possible match of one of the available 'speaker type' dependent recognisers to an unknown speaker.

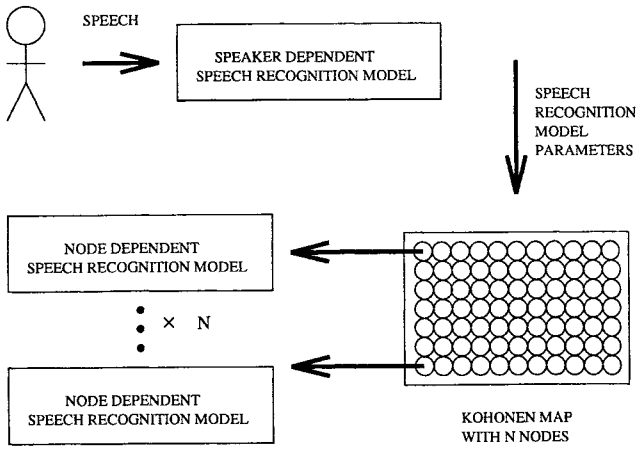

Fig.1. Producing a 2D grid of speech recognisers, mapping the speaker space

The system was tested using recognisers trained on one, two, four and eight speaker clusters. With increasing numbers of clusters average recognition performance increased slightly from 61.9% (standard deviation 9.8%) for a speaker independent HMM based recogniser to 64.48% (standard deviation 7.74%) for HMM based recognisers trained on 8 speaker clusters. Notably performance for test speakers on non-best-match prototypes reduces.

Using the MLPs for pattern matching improved average performance to around 75% but the changes with introduction of multiple speaker prototypes were still not statistically significant.

The reason performance does not improve significantly was shown to be the result of a trade-off between the amount of training data available for each speaker type and the optimisation to the speaker type. i.e. As performance improves, due to the recognisers being optimised to 'speaker types', it simultaneously worsens due to the amount of training data in each 'speaker type' being necessarily smaller than the total amount available. This was shown by the poor performance of the recognisers when trained on data from a similar number of random speakers as would be found in a typical 'speaker type' cluster.

2.3 Non-linear interpolation between recognition models

Self-organising maps (SOM) ([4]) can be trained to perform non-linear mappings of the input data. Here SOMs are used to map the space between speaker dependent speech recognisers by non-linear interpolation. Each training speaker produced a speaker dependent recogniser, which was represented by the node in the map with the most similar weights to the recogniser's parameters.

The Kohonen SOM produces a map with an ordered relationship between nodes. In this sense, nodes between 'training speaker nodes' are the result of non-linear interpolation between speaker dependent recognisers. It is now possible to turn the vectors underlying each node back into 'node dependent' recognisers which can be tested for unknown speakers.

A node between two 'training speaker nodes' will generate a recogniser for a speaker with characteristics 'between' the two training speakers. Thus a two-dimensional grid of recognisers was produced where adjacent recognisers have similar characteristics. The whole process is shown in Fig.1.

The performance of each recogniser in the grid was calculated for each test speaker. This could then be analysed to see how, for a given test speaker, the performance of the recognisers varied with the location of the recogniser on the grid. Fig.2. shows the scores for a grid of HMM recognisers on 12 utterances from a male and female test speakers. This shows how recognisers from a similar area have similar parameters and so produce similar scores.

A speaker identification stage could now determine the grid position of the best suited recogniser to deal with an unknown test speaker. Again, the upper bounds of performance of such a system were calculated assuming a perfect classifier.

Using the MLP based recognisers the system achieves recognition performance of 88.89% (standard deviation 8.27%) compared to a speaker independent performance of 76.19% (standard deviation 10.12%). The performance when using the 'in-between' interpolated MLPs was significantly better than the performance obtained from using the best fitting speaker dependent recogniser (85.71% (standard deviation 7.34%))

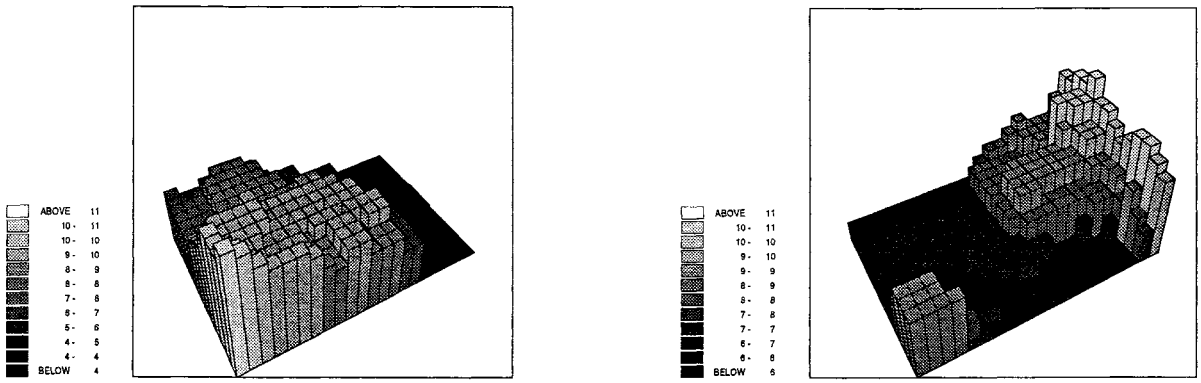

Fig 2. Recognition scores over a grid of HMMs for a female speaker (left) and a male speaker (right).

To use HMMs a concept of the distance between two models has to be developed that takes the distributions into account. Such a measure was developed and used as a distance measure in the SOM. Assuming perfect classification a recognition score of 77.18% (standard deviation 9.97%) was obtained. This is significantly better than the 62.12% (standard deviation 8.37%) score obtained for speaker independent recognition. The performance when using the interpolated HMMs was not significantly better than the performance from the best fitting speaker dependent HMMs.

3. CONCLUSIONS

The experiments discussed were vowel identification tasks carried out by two recognition models, MLPs and HMMs. The performance of the MLPs was higher than that of the HMMs which is probably due to the task of identifying stationary sounds.

- Crude speaker clustering does not improve performance over speaker independent systems. This is due to a trade-off between optimisation and the amount of available training data.

- Mapping the speaker space may improve recognition performance without need for increasing the amount of training data.

Throughout this work we assumed that a stage able to assign speakers to best suited recognition models exists. Such as stage is not yet available but under construction. The recognisers have been tested on vowels. This is a relatively easy task. The suitability of the approach proposed here will have to be tested for other phoneme classes and more complex recognition models.

4. REFERENCES

[1] Watkins. Journal of the Acoustical Society of America. 1990. pp 2942-2955.

[2] R. M. Stern & M. J. Lasry. IEEE Trans. Acoustics, Speech and Signal Processing. 1987. pp 751-758.

[3] K. F. Lee. *Automatic Speech Recognition. The SPHINX System.* Kluwer Academic Publishers. 1989.

[4] T. Kohonen. *Self-Organisation and Associative Memory.* Springer Verlag. 1989.