A SPEAKER ADAPTATION TECHNIQUE USING LINEAR REGRESSION

S.J.Cox

School of Information Systems, University of East Anglia, Norwich NR4 7TJ, UK
sjc@sys.uea.ac.uk

ABSTRACT

A technique for adapting speaker-independent speech recognition models to the voice of a new speaker is presented. The technique is capable of estimating adapted parameters for all the speech models when only a small subset of the recognition vocabulary is spoken by the new speaker. Whereas previous methods have often assumed a transformation between the speaker-independent models and the adapted models, this technique models the relationship between different speech units using linear regression. The regression models are built off-line using the training-set data. At recognition-time, the speech models are adapted using the regression models and the new speaker's data, a procedure which is computationally cheap. Experimental results show a halving of the recognition error-rate when only about 8% of the vocabulary is given as enrollment data, and when half the vocabulary is given, a reduction in the error-rate of 78%.

1. INTRODUCTION

Although advances in speech recognition techniques have enabled good performance to be obtained from speaker-independent (SI) speech recognition systems, a speaker-dependent (SD) system which has been adequately trained on the voice of the user will still give superior performance. However, in many situations it is impossible to obtain sufficient data (or perhaps any data) from the speaker before he uses the system and in such cases, it is necessary to use SI models.

In between these two extremes, perhaps as more data becomes available from a speaker in an on-line system, it is natural to attempt to adapt the SI models to work better on the new speaker's voice, a technique which has become to be known as "speaker adaptation". Most speaker adaptation schemes (e.g. [4]) require the new speaker to provide at least one example of each speech unit for full adaptation of the vocabulary. If examples of any of the units are not available, these units are not adapted. This may not be problematical if the vocabulary is fairly small but for large vocabulary systems, it means that the time taken for full adaptation will be long. Here, we present a method for adapting all sound classes in the vocabulary when only a subset of the vocabulary is available from a new speaker. The speech models are assumed to be continuous-density HMMs and the mean vectors of these models are adapted.

2. ADAPTATION PHILOSOPHY

In previous work ([1], [2]), we attempted to globally adapt SI models to a new speaker's voice by using a model of the form

$$\mathbf{s}_i = \boldsymbol{\mu}_i + \boldsymbol{\delta} \qquad (1)$$

where $\boldsymbol{\mu}_i$ is the i'th class mean, $\mathbf{s}_i$ is the speaker's realisation of this class and $\boldsymbol{\delta}$ is a "bias" vector characteristic of the speaker. A more sophisticated model was also studied in which the speaker's realisation of a sound-class was modelled as a linear transformation of the mean:

$$\mathbf{s}_i = A\boldsymbol{\mu}_i + \boldsymbol{\delta} \qquad (2)$$

where A is a $D \times D$ matrix (D = vector dimensionality). In both cases, the transformations were global i.e. $\boldsymbol{\delta}$ and A were used to transform all the models. However, the performance improvements obtained using these techniques were disappointing [2]. It was concluded that although the models of [1] and [2] have the advantage of simplicity and conciseness, their core assumption that a given speaker's speech can be modelled as a single invariant transformation applied to the 'prototype' speech models is not powerful enough. The techniques could be extended to use *sets* of neighbouring sound-units with a different bias term or transformation for each set [7]. As more sets are defined, the prediction accuracy within a given set increases but the model's predictive power lessens, since a given input sound can predict only the sounds within its own set. In the limit, the number of sets equals the number of sounds and an input sound is used only for predicting its own class i.e. the model ceases to be *predictive* of unheard classes.

0-7803-2431-5/95 $4.00 © 1995 IEEE

A shortcoming of the above approach is that it takes no account of the rich prior information (available in the training-data) about relationships between sounds. We have attempted to exploit this information as follows. Ignore for present purposes the temporal aspect of a speech sound and the desirability of representing it in a multi-dimensional space, and assume that a sound from a speaker can be represented by a scalar value on a certain axis. Other sounds made by the same speaker can then be represented by positions on orthogonal axes. Hence in this representation, if there are V sound-classes in the vocabulary, a speaker is a point in a V dimensional space. When several speakers are represented in this space, our assumption is that the resulting distribution of data points has some kind of structure and hence that we can assume transformations between sets of sounds. Once these transformation parameters are known, if we are given the values of a set of labelled sounds from a new speaker, values for any of the unheard sounds can be estimated.

There are several advantages to this approach over an approach in which a transformation between an SI model and an adapted model is assumed:

1. In principle, models can be built between any sets of sounds we choose and at recognition time, the ones which are most useful for exploiting the data presented by the new speaker may be used.

2. In general, there will be plenty of data to estimate the model parameters rather than the sparse amount of data provided by a new speaker. Hence we can assume more complex models and obtain good estimates of model parameters.

3. Model parameter estimation can be done off-line so that the computation required at recognition time is small.

3. MODEL DETAILS

For present purposes, we used the simplest assumption possible about the structure of the data in the 'speaker-space', which is that it is linear. Hence multiple linear regression can be used to model the data structure i.e we can use a model of the form:

$$\mathbf{Y}_k = \beta_0 + X\beta + \mathbf{e}_k, \qquad (3)$$

where $\mathbf{Y}_k$ is the set of S realisations of sound k from the speakers, X is an $S \times (V-1)$ matrix of realisations of all sounds except sound k, β is a set of $(V-1)$ regression coefficients and $\mathbf{e}_k$ is an error-vector. Given a subset of M of the vocabulary sounds, we can form a regression model of any sound k not in this subset using the $V \times V$ covariance matrix of the speaker-space [5]. However, to pre-compute each of these regressions would be prohibitive; in practice, they would have to be computed on-line once the set of M sounds available from the new speaker was known. To avoid this on-line computation, we use simple linear regression pairwise between all the sound-classes and construct a set of models of the form:

$$Y_k = \beta_0 + \beta_1 X_j + e_k \qquad j \neq k \qquad (4)$$

It will be seen later that the experimental paradigm we used enabled a fairly modest amount of storage for the pre-computed regression coefficients.

3.1. Model construction and data

Experiments were carried out on a database of isolated utterances provided by British Telecom [6]. Recognition of isolated utterances was used because it is the simplest practical speech recognition problem and the techniques developed for its application can be extended to more sophisticated systems.

In section 2, we assumed that examples of sound classes can be represented as scalars. To deal with sequences of vectors which represent isolated utterances, we first train speaker-independent HMMs for each class in the vocabulary and then use the HMM of class i and the Viterbi algorithm to segment the training-set utterances of class i. Hence each vector of the utterance is mapped to a single state in the HMM representing the utterance. For the purposes of the adaptation, the HMM states play the rôle of the sound-classes and the vectors associated with each state are the 'examples' of these pseudo-classes. We assume that the vector dimensions are independent and treat each dimension separately.

The data consisted of 3 utterances of the alphabet from 104 speakers and was divided equally into balanced sets of 13 enrollment classes and 13 testing classes. Isolated-word SI CDHMMs (10 states per model, single component Gaussian densities, diagonal covariance-matrix), were trained for each class using data from the 52 training-set speakers, and this data was then Viterbi-aligned to the appropriate model. For each speaker k, sample means for each enrollment-class (x) state and test-class (y) state were computed together with sample standard deviations. Linear regression models were then formed as follows:

```
For each vector dimension do
    For each state $j$ of each enrollment class $i$ do
        For each state $n$ of each test class $m$ do
            Form a scattergram using the 52 pairs of data points $\{\bar{x}^k_{i,j}, \bar{y}^k_{n,m}\}$ which have associated st.devs, $\{s^k_{i,j}, s^k_{n,m}\}$.
            Compute the best-fit line $y_{n,m} = a_{i,j,n,m}x_{i,j} + b_{i,j,n,m}$ through the scattergram and store the regression coefficients $a_{i,j,n,m}, b_{i,j,n,m}$ and correlation coefficient $r_{i,j,n,m}$
```

This algorithm was repeated for each vector dimension. Note that the above procedure is strictly linear fitting rather than linear regression, since both x and y data have associated uncertainties. Since there were 130 enrollment class states and 130 test class states in these experiments, the procedure involves the computation and storage of 16900 sets of regression coefficients.

3.2. Adaptation strategy

Given some utterances of a subset of the enrollment classes from a new (test-set) speaker, each utterance is Viterbi-aligned to the appropriate model (i.e. adaptation is supervised) and the mean computed for each state of each enrollment class. Hence a set of mean value estimates $u_i, i = 1, 2, \ldots, M$ (where $M =$ (no of enrollment classes supplied) x (no of states per model)) is available and each u_i is used with the appropriate regression model to produce an estimate $\hat{v}_i$ of each testing class state mean. If it is assumed that $\hat{v}_i$ has been drawn from a normal distribution, it can be shown that the ML estimate of a test class state mean (in a given vector dimension) is given by $\hat{v}_{ML}$ where

$$\hat{v}_{ML} = \left(\sum_{i=1}^{M} \frac{\hat{v}_i}{(1 - r_i^2)} \right) \Big/ \left(\sum_{i=1}^{M} \frac{1}{(1 - r_i^2)} \right) \qquad (5)$$

where $\hat{v}_i$ is the estimate from the i'th regression model which has sample correlation coefficient r_i. Given a prior distribution for the state mean vector (assumed normal with mean μ_{SI} and variance σ^2_{SI}), the MAP estimate of the mean is shown to be

$$\hat{v}_{MAP} = \left[\frac{\mu_{SI}}{\sigma^2_{SI}} + \sum_{i=1}^{M} \frac{\hat{v}_i}{(1 - r_i^2)} \right] \Big/ \left[\frac{1}{\sigma^2_{SI}} + \sum_{i=1}^{M} \frac{1}{(1 - r_i^2)} \right] \qquad (6)$$

Note that when $r_i = 0$, the regression line is a horizontal line through μ_{SI} and $\hat{v}_i = \mu_{SI}$ for any value of u_i. In the extreme case where all the available adaptation states are uncorrelated with v, both the ML and MAP estimates reduce to μ_{SI} i.e. the adaptation "fails safe".

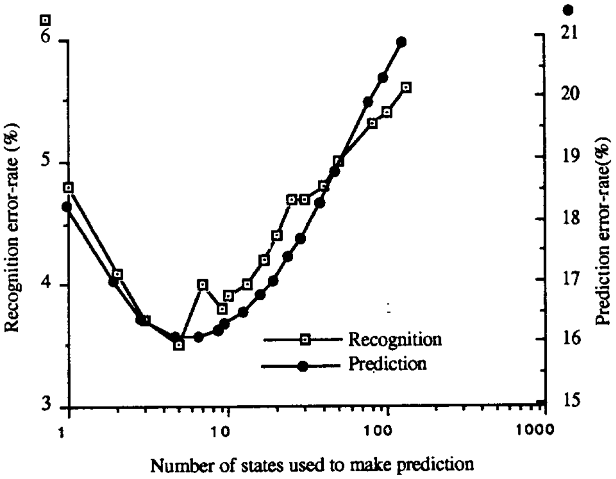

Fig 1: Effect of varying number of states used to make prediction (All 13 enrollment classes available, MAP estimate)

Adaptation of the test classes consisted of replacing the SI HMM test-class means with $\hat{v}_{ML}$ or $\hat{v}_{MAP}$. The means of the enrollment classes supplied by the new speaker were adapted by replacing the HMM means with the estimated mean for each enrollment class state, $\bar{x}^k_{i,j}$. The model variances were not adapted. To avoid testing with enrollment data, only the utterances of the test classes were tested, but all models were active during a test.

4. EXPERIMENTAL RESULTS

The 52 speakers in the test-set each gave three examples of the test classes, making a total test-set size of 1984 utterances (44 utterances were missing). The error-rate for the system with no adaptation was 17.0%. Firstly, the effect of using varying numbers of states to predict each test-class mean was studied. When examples of all 13 enrollment classes were given by the speaker, there are a total of 130 states available for prediction of each test class state mean value. The effect on (a) prediction error and (b) recognition accuracy of using only the 'best' P states to make the prediction was measured. By the 'best' P states, we mean that for each test class state predicted, the enrollment class states were ranked according to their average correlation with the test class state and the top ranking P states used to make the prediction. Fig 1 gives prediction and recognition error-rates vs. P when the MAP estimate was used. Prediction and recognition performance track quite closely (as might be expected) and peak when only 5–6 of the 130 available states are used. If more than about 5 states are used for prediction, the lower ranking (i.e. more poorly correlated) states contribute noise to the estimate and performance

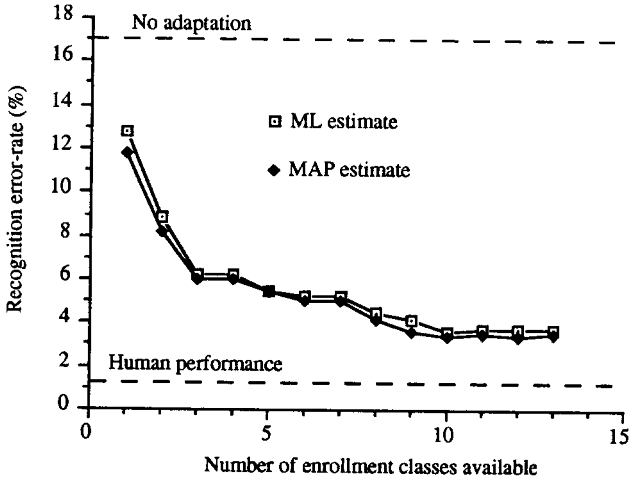

Fig 2: Recognition error vs. number of enrollment classes available for ML and MAP estimates

worsens slightly. Secondly, the number of available enrollment classes supplied by each test-set speaker was varied from 1 to 13 and recognition accuracy using the "best" 5 available states was measured. The result is shown in Fig 2 for both ML and MAP estimates of the adapted models. In both cases, the error-rate has dropped by over 50% when only the first two enrollment classes have been given. Full adaptation has been reached when the first 10 classes have been presented, when the error-rate is 3.4%. This figure should be compared with a best error-rate of 14.8 % produced by the "bias" technique (described in section 2) on this data. Also included on Fig 2 is an estimate of human performance on the same data. This figure (1.2%) was obtained from listening-tests on 25 subjects without any adaptation to the voice of each speaker in the database [3]. Although not definitive, it gives some idea of the performance obtainable by humans on this data.

5. SUMMARY

We propose a method of adapting SI speech models to the voice of a new speaker which is based on modelling the relationship between different speech units using linear regression and using the regression models to predict unheard sounds from the speaker. The technique offers several advantages:

- Efficiency—all speech models can be adapted given only a subset of the vocabulary from the new speaker.
- Flexibility—several models can be built off-line and the most useful ones used at recognition time.
- Parameter estimation—the models can be built from a training-set containing a large number of speakers and utterances, leading to good parameter estimates.
- Speed—computational requirements at recognition time are low because no parameter estimation is required, only model adaptation.

A technique for applying this scheme within the framework of continuous density hidden Markov models has been described and ML and MAP estimates for the predicted values derived. When applied to a database of 104 speakers each speaking utterances of the alphabet, the technique was shown to reduce the error-rate from 17.0% to 3.4%. Although the baseline recognition system was not a particularly sophisticated one and no attempt was made to optimise its performance, the adaptation performance is impressive and encourages us to experiment with the technique on more difficult speech recognition tasks.

6. REFERENCES

[1] S.J. Cox and J.S. Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 294–297, April 1989.

[2] S.J. Cox and J.S. Bridle. Simultaneous speaker normalisation and utterance labelling using Bayesian/neural-net techniques. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1990.

[3] S.J. Cox, R.D. Johnston, P.W. Linford, and K. Chikolowski. Human performance on an isolated alphabetic speech recognition task. In *Proc. The Institute of Acoustics*, 1994.

[4] C.H. Lee, C.H. Lin, and B.H. Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Transactions on Signal Processing*, 39(4):806–814, April 1991.

[5] D.F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill, 2nd edition, 1976.

[6] J.A.S. Salter. The RT5233 alphabetic database for the connex project. Technical Report RT52/G231, BT Technology Executive, April 1989.

[7] Y. Zhao. An acoustic phonetic speaker adaptation technique for improving speaker independent continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(3):380–394, July 1994.