PREDICTIVE SPEAKER ADAPTATION IN SPEECH RECOGNITION

Stephen Cox, School of Information Systems,
University of East Anglia, Norwich NR4 7TJ, UK.
Tel: +44 603 592582
e-mail:sjc@sys.uea.ac.uk

Address for correspondence from August 1st–December 15th 1994:

c/o Dr Richard Rose,
Speech Research Dept.,
AT+T Bell Labs.,
600 Mountain Hill Avenue
Murray Hill
N.J.
U.S.A.

ABSTRACT

A major problem with most speaker adaptation schemes is that they rely on the speaker providing at least one example of each acoustic unit (word, phone, triphone etc.) in the vocabulary in order to adapt the appropriate model. Rapid adaptation is difficult to achieve and some sounds may never be adapted because they are never heard. In this paper, a technique of adapting all the speech models to a new speaker's voice when he has given an incomplete set of the vocabulary is presented. The technique is based upon using the training-set to obtain estimates of correlations between sounds. Given some sounds from a new speaker at recognition time, these correlations are used to obtain estimates of unheard sounds which are used to adapt the speech models. The technique was applied to a database of 104 speakers speaking the English alphabet. When speakers spoke half of the vocabulary for enrollment prior to recognition, the technique gave a 78% decrease in error.

1 Introduction

One of the central problems in the design of automatic speech recognition (ASR) systems is how to model the variation in the acoustical signal representing a given speech unit. If the system is to be used by a single speaker, this variation is usually relatively small from utterance to utterance, and provided a sufficient amount of data can be collected from the speaker to 'train' the system, high accuracy can be achieved. Such a system is usually referred to as a 'speaker-dependent' system.

However, in many practical situations, a speech recognition system must be capable of functioning with good accuracy on any voice (a 'speaker-independent' system). These systems must contend with the variation in acoustical signals from speaker to speaker (caused by such effects as different vocal tract sizes and shapes, different accents, speaking styles etc.) which is much greater than the variation within a single speaker. For both speaker-dependent (SD) and speaker-independent (SI) systems, the approach most commonly used to model variation is to collect a large number of examples of the speech units comprising the recognition vocabulary and use these within a statistical framework to estimate parameters of models of the speech units. Because the speaker-independent models are derived from a large number of speakers, they have higher variance and hence higher overlap between different speech units than adequately trained speaker-dependent models. Hence given enough utterances from a speaker to generate adequately trained models, we would expect to obtain better performance on his voice using these models than using SI models. However, if there is no data available from the speaker, we are forced to use SI models (or SD models derived from a different speaker, a possibility we do no consider here).

In between these two extremes, perhaps as more data becomes available from a speaker in an on-line SI system, it seems natural to attempt to 'tune in' the SI models to work better on the new speaker's voice, i.e. to move the speaker-independent system towards a speaker-dependent system, a technique which has become known as 'speaker adaptation'. The way in which adaptation of the speech models is accomplished depends on the nature of the speech models and the recognition system. A popular and successful method for ASR is to use continuous density hidden Markov models (CDHMMs) and in such a system, adaptation can be formulated as a Bayesian learning procedure [11]. Using this method, the SI model parameters are regarded as the *a priori* information which is modified by the new speaker's data to obtain an *a posteriori* estimate of the model parameters. By adapting the means and variances of the HMM state probability density functions, it has been demonstrated [11] that a speaker adaptive system always obtained better performance

than a fully-trained speaker-dependent system.

2 Predictive adaptation

A disadvantage of the scheme proposed in [11] and similar schemes (e.g. [2], [10]) is that the new speaker must provide at least one example of each speech unit for full adaptation of the vocabulary. If examples of some 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. One system requires the new speaker to enrol by reading a specified text of about 100 sentences, a task which takes typically about 20 minutes [1]. Another system which does not specify the adaptation text requires about 10 000 words before complete adaptation is achieved [9].

By contrast, if a model for production of different sounds from a speaker is used, data supplied by the speaker can be used to drive the model to predict previously unheard sounds. The accuracy of such predictions will depend on the degree to which the model matches reality and on the number of free parameters required to be estimated in the model.

There are some grounds for believing that humans use some such procedure. Under normal environmental conditions, we usually understand effortlessly accents that we are accustomed to, regardless of the physiology and (to a large extent) speaking style of the speaker. This suggests that 'normalisation' to a speaker's acoustic signals is practically instantaneous if they follow the accent pattern that we expect. However, when we hear speech with an unfamilar accent, we are conscious of having to give the speaker extra attention to understand what he is saying and our ability to understand drops rapidly if the environment is noisy. If the accent is particularly severe, it may take a long period of exposure to it before we have complete and instantaneous understanding of the speech.

The fact that anyone can give some sort of impression of a well-known accent, or that we sometimes remark that someone speaks with a 'strange accent' implies that we have conscious 'models' of accents. It does not seem unreasonable to suggest that we have well-developed unconscious models of accents and speaker types and that hearing some speech from a new speaker triggers the appropriate model (i.e. set of expectations about the acoustical signal) to aid us in decoding their speech. The model may be refined as we receive more data from the speaker and 'tune-in' to his speech.

3 Linear predictive models

In this section, we consider two particularly simple predictive models which were reported in [5] and [6].

3.1 A 'bias' model

Suppose that an example of a sound-class c_i can be represented by a single D-dimensional vector and that an estimate of the mean vector of c_i over speakers is μ_i. The realisation of this sound class from a speaker is modelled as a vector $\mathbf{s}_i$, where

$$\mathbf{s}_i = \mu_i + \delta \tag{1}$$

The vector δ may be thought of as a 'bias' term and is characteristic of the speaker, or more correctly the speaker plus acoustic channel. For the purposes of predicting unheard sounds, *all* sound classes from the speaker are modelled as the mean value of the class plus the same offset vector δ i.e. we consider the summed effects of the speaker plus acoustic channel to be equivalent to a translation of the mean vector in the feature-space, and the translation is characteristic of the speaker and invariant. If the speaker gives N utterances, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ whose classifications are $c_{k(1)}, c_{k(2)}, \ldots, c_{k(N)}$, the maximum-likelihood value of $\boldsymbol{\delta}$ under this model is

$$\delta = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mu_{k(i)}) \tag{2}$$

3.2 A linear transformation model

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:

$$\mathbf{s}_i = \mathbf{A}\mu_i + \delta \tag{3}$$

where $\mathbf{A}$ is a $D \times D$ matrix. With no restrictions on the values of $\mathbf{A}$, it would be necessary to estimate $D^2 + D$ parameters to train this model. Since the technique is designed to work on small amounts of data, this is not viable for even modest values of D. Hence $\mathbf{A}$ was

restricted to be a tri-diagonal matrix with constant coefficients i.e.

$$\mathbf{A} = \begin{pmatrix} \beta & \gamma & & & & & \\ \alpha & \beta & \gamma & & & & \\ & \alpha & \beta & \gamma & & O & \\ & & . & . & . & & \\ & & & . & . & . & \\ & O & & & \alpha & \beta & \gamma \\ & & & & & \alpha & \beta \end{pmatrix}$$

The values of α, β and γ were estimated by least-squares fitting.

Both models had a phonetic motivation. The chosen feature representation in the work reported in [5] and [6] was a logged magnitude spectrum, so that if the speaker's 'transformation' of the mean is visualised as a linear filtering operation, this is implemented as an addition in this domain, as in equation 1. The effect of the tri-diagonal matrix transformation in equation 3 is to shift acoustic patterns up and down the spectrum, and it is known that realisations of a sound across speakers can be at least partially explained by upward or downward shifts of the spectrum [3].

However, recognition performance improvements obtained using these models were disappointing. Using a database of 104 speakers, each speaking three utterances of the alphabet, the greatest improvement in error-rate obtained was 32% (14.9% unadapted to 10.1% after adaptation). A study which used a very similar technique [14] also noted a small increase in performance. It was later discovered that a similar reduction in error-rate could be obtained by using a more sophisticated front-end which incorporated time-derivative information, and after this had been introduced, the adaptation techniques gave little improvement. It was concluded that although these models 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 to account for the complexity of the actual shifts in the feature-space.

4 A correlation-based model

The simplest and commonest type of variation between accents is in the way in which different vowels are pronounced. If there is a difference in two accents between the realisation of a certain vowel, there must be parallel and symmetrical realisational differences between neighbouring phonemes, or otherwise phonemes would not remain distinct. A similar reasoning would hold for differences due to other factors such as vocal-tract shape or acoustic channel.

This observation suggests that the techniques described in section 3 could be extended to use *sets* of neighbouring sound-units with a different bias term or transformation for each set. As more sets are defined, the model's predictive power lessens, since a given input sound can predict only the sounds within its own set, but the prediction accuracy within a given set increases. 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. Both conditions are present in the study by Zhao [15], in which the model of section 3.1 was extended by using both an invariant bias term δ and a sound-dependent term ϕ_i.

Rather than speculate about the number and membership of such sets, we have preferred to abandon the assumption that a sound from a speaker can be generated by a linear transformation of the appropriate prototype sound. Instead, we simplify the complexities of the speech signal by imagining that a sound from a speaker can be represented by a value on a certain axis, and that other sounds made by the same speaker can be represented by positions on orthogonal axes. In this representation, a speaker is a point in a V dimensional space, where V is the number of sounds in the vocabulary. Note that this space is quite different from the 'feature-space' used to initially represent the sounds from their waveforms. When several speakers are represented in this 'speaker-space', our assumption is that the resulting distribution of datapoints has some kind of structure and that knowledge of this structure is useful for prediction purposes. For instance, if the locations of the sounds whose classifications are $c_1, c_2, \ldots, c_N$ are known for a speaker, they can be used together with the knowledge of the overall structure to predict the positions of the remaining $(V - N)$ classes for this speaker.

The simplest model we can assume for this structure is a linear one. Suppose that the acoustic realisation of the i'th sound-class in the vocabulary by the j'th speaker in the training-set can be represented by a scalar $x_i^j, i = 1, 2, \ldots, V$. In practice, x_i^j may be the value for sound i in one dimension of the feature-space representation and we assume that the analysis is repeated for each feature-space component. We assume a multiple linear regression model for class n in terms of the other classes as follows:

$$x_j^n = \beta_0 + \sum_{k=1}^{V} \beta_k x_k^j + e_n^j \qquad k \neq n \tag{4}$$

where the β's are the regression coefficients and e_n^j is an error term.

The model of equation 4 would be useful only if we knew the values for all classes except class n for this speaker. In practice, we would like to be able to predict the value

for class n from a small set of classes—in the limit from a single class. In the latter case, it is appropriate to use simple linear regression between classes n and m:

$$x_n^j = \beta_0 + \beta_1 x_m^j + e_{n,m}^j \tag{5}$$

Equations 4 and 5 represent extremes—we can choose to model a sound-class as a linear regression on any set of different sound-classes. However, multiple linear regression is less useful than simple linear regression in this application because we do not know *a priori* what sounds from the new speaker will be available to us at recognition time. Hence we would need to build a set of regression models which modelled each class in terms of all the possible combinations of all the other classes, which is clearly out of the question. By restricting ourselves to simple linear regression, we need form only $V(V-1)/2$ regression models to enable us to predict any class in the vocabulary from any other class. However, these simple models will generally be less powerful than multiple linear regression models. This was shown in [4] where the technique underwent a preliminary investigation on a database of 11 isolated vowel sounds from each of 30 speakers, and the task was to predict 6 unheard vowels from 5 given vowels.

The procedure for speaker adaptation using the simple linear regression models is as follows: using the training-set speakers' data, estimate and store the values of β_0 and β_1 for all pairwise linear regression models. Then at recognition time, any unheard sound of class n can be predicted using any given labelled sound of class m by using the regression model which links these two classes. The prediction can then be used to adapt the SI model.

The success of this technique depends critically on how well the data fit a linear model, which can be measured by the correlation (or multiple correlation) coefficient between the data. The importance of this correlation is emphasized in section 7 where it is shown that it controls the amount of adaptation.

5 The data and speech models

The advantage of doing experiments using the isolated vowel data described in the previous section [4] was that each example of a vowel sound was represented as a single vector which dispensed with the need for time alignment procedures. Using this artificial data, the technique produced a 40% decrease in the error-rate and stimulated the experiments reported here. These were carried out on a database of isolated utterances provided by British Telecom [13]. 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.

The database consisted of 3 utterances of the alphabet from each of 104 speakers recorded in a soundproof room with a high-quality microphone at a bandwidth of approximately 8 kHz. Each utterance was manually endpointed and processed into frames of duration 16 ms, each frame consisting of a 17-d vector (frame) containing 8 MFCCs, 8 differential coefficients and a differential log-energy coefficient. The training-set (52 speakers) was used to construct a 10 state continuous density hidden Markov model (HMM) of each alphabetic class, the state PDFs being unimodal Gaussian with a diagonal covariance matrix. The topology of the HMM was a simple one in which state i was connected only to itself and state $(i+1)$.

Each test-speaker's data was divided into two sets. The 'enrollment classes' were the utterances of a set of classes which were to be used for adaptation purposes during the experiments and which were regarded as labelled; the 'test classes' were the utterances of the other classes in the vocabulary and were regarded as unlabelled. The identities of the 'enrollment' and 'test' classes were as follows:

Enrollment classes:	'A','D','E','F','G','I','K','M','O','P','Q','R','X'
Test classes:	'B','C','H','J','L','N','S','T','U','V','W','Y','Z'

The division between enrollment and test class sets was made in such a way as to keep the data as phonetically balanced between the two sets as possible. For instance, the 'E'-set ('B', 'C', 'D', 'E', 'G', 'P', 'T', 'V'), the 'A'-set ('A', 'J', 'K', 'H') and the set beginning with the vowel /ϵ/ ('F', 'L', 'M', 'N', 'S, 'X') were evenly divided between the two sets.

6 Building the correlation models

6.1 Dealing with sequences of vectors

Until now, we have assumed that examples of classes can be represented as single vectors and have not addressed the problem of how to apply the adaptation technique within a real ASR system in which utterances are represented by *sequences* of vectors.

Firstly, we train speaker-independent HMMs for each class in the vocabulary. We then use the HMM of class i and the Viterbi algorithm to 'segment' the training-set utterances of class i. By segmentation, we mean that each vector in the sequence of vectors describing an utterance is mapped to a single state in its associated HMM. For the purposes of the adaptation, we imagine the HMM states playing the rôle of the sound-classes. The vectors

associated with each state are then the 'examples' of these pseudo-classes. The simple topology of the HMMs ensures that some data is available for each state.

6.2 Model building

Once again, to simplify analysis, it was assumed that the vector dimensions were independent and could be considered separately. Denote:

$S_{i,j} = j$'th HMM state of the i'th class
$i = 1, 2, \ldots, V$ $\quad V =$ no of classes (26)
$j = 1, 2, \ldots, H$ $\quad H =$ no of states in an HMM (10)

After segmentation of the training-set speakers' data, denote:

$x^k_{i,j}(l) = l$'th value from speaker k associated with $S_{i,j}$
$k = 1, 2, \ldots, S_{tr}$ $\quad S_{tr} =$ no of speakers in training-set (52)
$l = 1, 2, \ldots, N^k_{i,j}$ $\quad N^k_{i,j} =$ no of vectors associated with state $S_{i,j}$

For each speaker k, a mean value $\bar{x}^k_{i,j}$ and associated standard deviation $s^k_{i,j}$ are estimated for each state of each HMM as follows:

$$\bar{x}^k_{i,j} = \frac{1}{N^k_{i,j}} \sum_{l=1}^{N^k_{i,j}} x^k_{i,j}(l) \qquad s^k_{i,j} = \sqrt{\frac{1}{N^k_{i,j} - 1} \sum_{l=1}^{N^k_{i,j}} (x^k_{i,j}(l) - \bar{x}^k_{i,j})^2}$$

Note that if we were building a speaker-dependent HMM from these utterances, $\bar{x}^k_{i,k}$ and $s^k_{i,j}$ would be estimates of the mean and standard deviation for state j of the HMM representing class i obtained using 'Viterbi' rather than 'Baum-Welch' style training.

Since we are to predict the test class states from the enrollment class states, denote the means of the m'th state of test class n as $\bar{y}_{n,m}$ (rather than $\bar{x}_{n,m}$). The linear regression models are constructed 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 S_{tr} pairs of datapoints $\{\bar{x}^k_{i,j}, \bar{y}^k_{n,m}\}$
      which have associated standard deviations, $\{s^k_{i,j}, s^k_{n,m}\}$.
      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 $r_{i,j,n,m}$

6.3 Straight line fitting

In 'standard' linear regression, it is assumed that the N values of the independent variable x_i are known exactly and that each value of the dependent variable y_i has an associated uncertainty σ_i. The values of a and b in the straight-line fit $y = ax + b$ are the values which minimise the chi-square value

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - b - ax_i}{\sigma_i} \right)^2 \tag{6}$$

and there are closed form solutions for a and b.

However, the data used to construct the correlation models described here has variance in both the x and y directions ($x_{i,j}$ and $y_{n,m}$ respectively), which makes the estimation of a and b more difficult. Suppose each x_i has a variance $\sigma_{x\,i}^2$ and each y_i a variance $\sigma_{y\,i}^2$. The chi-square value to be minimised is then

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - b - ax_i}{\sigma_{y\,i}^2 + a^2\sigma_{x\,i}^2} \right)^2 \tag{7}$$

A closed-form solution for b still exists but a must be found by a numerical method. In this work, we used a purpose-written routine to find a and b described in [12]. We also experimented with two other techniques in the straight-line fitting:

1. discarding the variance information by setting $\sigma_{x\,i}^2 = \sigma_{y\,i}^2 = 1.0$
2. minimising the *absolute* deviations of the data-points from the line

The differences in recognition performance obtained when using models prepared using these three methods were very small, but fitting using the variances in both directions was marginally better and was used for the results described in the later sections.

6.4 Analysis of correlations

Section 6.2 described how a set of data associated with a given test class state was correlated with data associated with each of the 130 enrollment class states. It is of interest to examine which states exhibit good correlation across 'test' and 'enrollment' classes. For each state n of each test class m, the enrollment state $S_{n,m}^*$ which had the highest absolute value of correlation when averaged over all vector dimensions was found i.e.:

$$S_{n,m}^* = \underset{i,\,j}{\text{argmax}}\, \bar{r}_{i,j,m,n}$$

$$\text{where:} \quad \bar{r}_{i,j,m,n} = \frac{1}{D} \sum_{k=1}^{D} |r_{i,j,m,n}(k)|$$

and $r_{i,j,m,n}(k)$ is the sample correlation coefficient in vector dimension k between enrollment class i, state j and test class m, state n. The identities of $S^*_{n,m}$ and the corresponding averaged absolute correlation values for two of the test classes ('H' and 'J') are given in Table 1.

(Table 1 here)

An analysis of the correlations between states of models of test class and enrollment class words reveals some interesting phonetic correspondences. Table 1 shows that the first two states of the model representing 'J' are most highly correlated with the first two states of the model representing 'G'. Both words begin with the phoneme /*dz*/ and these initial states would correspond to the position of that phoneme in the HMMs representing each word. The third state of 'J' is correlated best with state 2 of 'D', corresponding to the phoneme /*d*/ which is very close to /*dz*/. States 4–10 are correlated with states 4–10 of the classes 'A' and 'K' and all three words share the same final vowel, /*e*/. In the model for 'H', the first three states have best correlation to early states of the model representing 'A', which shares the same initial vowel-sound as 'H' (/*e*/). States 4 and 5 are best correlated to states 3 and 4 of 'X' which is probably the point in both models at which the transition from the vowel to the stop is being made. The final phoneme in 'H' (/$t\int$/) does not occur in the enrollment classes and the closest to it is probably the phoneme /*dz*/ which occurs at the beginning of 'G' and which may explain the correlation of states 6–9 of 'H' with states 1 and 2 of 'G'. The final state of 'H' correlates with the final state of 'X' which is again phonetically plausible as both words end with unvoiced affricates.

At first sight, the averaged correlation values seem too low to be very useful for prediction purposes. However, it was found that correlation between the first 8 components of the vectors (the MFCCs) was always higher than correlation between the last 9 components (the differential coefficients) and so the latter correlations significantly lower the average. The corollary is that it is more difficult to make good predictions of the differential coefficients than of the MFCCs. However, the differential coefficients contribute less to recognition than the MFCCs and hence their accurate estimation is less important.

7 Prediction using the correlation models

At recognition time, we assume that a new speaker gives labelled examples of a subset N of the enrollment classes. Suppose the classes given are $c(1), c(2), \ldots, c(N)$ (there may be more than one example of each class). To make this data available for prediction purposes we first:

1. Segment each example using the appropriate HMM
2. In each vector dimension, estimate the sample mean $\bar{x}_{i,j}$ of the values assigned to each state as described in section 6.2 , for $i = c(1), c(2), \ldots, c(N)$ and $j = 1, 2, \ldots, H$.

We now wish to make a prediction of the values of $\bar{y}_{n,m}$ and $s^2_{n,m}$ where n ranges over the test classes. For present purposes, we have confined ourselves to estimating $\bar{y}_{n,m}$. Estimation of $s^2_{n,m}$ is a much more difficult task and it seems an open question as to whether sensible estimates of $s^2_{n,m}$ could be obtained without some examples of each test class. Derivations of maximum-likelihood (ML) and maximum a posteriori (MAP) estimates of $\bar{y}_{n,m}$ are given in the next two sections.

7.1 Maximum likelihood estimate of the speaker-dependent mean

There are $M = N \times H$ values of $\bar{x}_{i,j}$ and using the stored coefficients $a_{i,j,m,n}$ and $b_{i,j,m,n}$, each can be used to give a prediction of all $V/2 \times H = 130$ values of $\bar{y}_{n,m}$. Clearly, the usefulness of any prediction of $\bar{y}_{n,m}$ depends on the correlation between $\bar{x}_{i,j}$ and $\bar{y}_{n,m}$. Let us assume that we are attempting to predict a particular $\bar{y}_{n,m}$ using M values of $\bar{x}_{i,j}$. To simplify notation, we drop the (class, state) subscripts and rename the target value $\bar{y}_{n,m}$ as v and $\bar{x}_{i,j}$ as u_i $(i = 1, 2, \ldots, M)$. Hence the i'th prediction of v is $\hat{v}_i$ where

$$\hat{v}_i = au_i + b \tag{8}$$

and where $a = a_{i,j,n,m}$, $b = b_{i,j,n,m}$.

It can be shown that the variance σ_i^2 of $\hat{v}_i$ is:

$$\sigma_i^2 = \sigma_y^2(1 - r^2)\left[1 + \frac{1}{n} + \frac{(u_i - \mu_{SI}^x)^2}{n\sigma_x^2}\right] \tag{9}$$

where σ_x^2 and σ_y^2 are respectively the overall variance of the $\bar{x}$'s and $\bar{y}$'s in the relevant scattergram, r is the sample correlation coefficient, n is the number of examples in the scattergram (in this case $n = S_{tr} = 52$) and μ_{SI}^x is the mean of the $\bar{x}$'s. In equation 9, the term $1/n$ increases the variance for small sample sizes and the term $(u_i - \mu_{SI}^x)^2/n\sigma_x^2$

increases the variance as the independent variable moves further from the overall mean μ_{SI}. However, for n sufficiently high, these two terms can be ignored and approximately:

$$\sigma_i^2 = \sigma_y^2(1 - r^2) \tag{10}$$

If it is assumed that each $\hat{v}_i$ has been drawn from a normal distribution with mean v and variance $\sigma_i^2 = \sigma_y^2(1 - r^2)$, we may write $\Pr(v|\hat{v}_i)$ as

$$\Pr(v|\hat{v}_i) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2}\frac{(\hat{v}_i - v)^2}{\sigma_i^2}\right] \tag{11}$$

Differentiating $\Pr(v|\hat{v}_i)$ w.r.t. v and setting to zero gives the maximum likelihood estimate of v as $\hat{v}_{ML}$ where

$$\hat{v}_{ML} = \left(\sum_{i=1}^{M} \frac{\hat{v}_i}{\sigma_i^2}\right) \Big/ \left(\sum_{i=1}^{M} \frac{1}{\sigma_i^2}\right) = \left(\sum_{i=1}^{M} \frac{\hat{v}_i}{(1 - r_i^2)}\right) \Big/ \left(\sum_{i=1}^{M} \frac{1}{(1 - r_i^2)}\right) \tag{12}$$

The $1/(1 - r_i^2)$ weighting of predictions means that predictions made from scattergrams where $|r|$ is high get higher weight than predictions from scattergrams where $|r|$ is low. At first sight, it would seem incorrect to give any weighting to a prediction of v when $r_i = 0$. However, when $r_i = 0$, the regression-line is a horizontal line through the mean of the y-points and so the prediction is μ_{SI}^y i.e. the speaker-independent mean. In fact if $r_i = 0\ \forall i$ is substituted into equation 12, $\hat{v}_{ML} = \mu_{SI}^y$. The fact that $\hat{v}_{ML} \to \mu_{SI}^y$ as $|r_i| \to 0$ means the adaptation 'fails-safe' in the sense that if the correlation is poor, the prediction reverts to the SI mean.

7.2 Maximum a posteriori estimate of the speaker-dependent mean

For a maximum a posteriori (MAP) estimate of v, we wish to take into account the prior distribution of the speaker-dependent means and to maximise $\Pr(v|\hat{v}_i, \lambda)$ where λ is the parameter set of the prior distribution of the speaker-dependent means. Using Bayes' theorem, we write

$$\Pr(v|\hat{v}_i, \lambda) = \frac{\Pr(\lambda)\Pr(v|\lambda)\Pr(\hat{v}_i|v, \lambda)}{\Pr(\lambda|\hat{v}_i)\Pr(\hat{v}_i)} \tag{13}$$

Terms which do not depend on v can be combined into a constant α. Also, we assume that we can write $\Pr(\hat{v}_i|v, \lambda) = \Pr(\hat{v}_i|v)\Pr(\hat{v}_i|\lambda)$ Hence equation 13 becomes:

$$\Pr(v|\hat{v}_i, \lambda) = \alpha \Pr(\hat{v}_i|v)\Pr(\hat{v}_i|\lambda) \tag{14}$$

(since $\Pr(\hat{v}_i|\lambda)$ does not depend on v). If we assume that the prior distribution of speaker-dependent means is normal with mean μ_{SI} and variance σ^2_{SI} we can write:

$$\Pr(v|\hat{v}_i,\lambda) = \alpha \frac{1}{\sqrt{2\pi}\,\sigma_{SI}} \exp\left[-\frac{1}{2}\frac{(v-\mu_{SI})^2}{\sigma^2_{SI}}\right] \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2}\frac{(\hat{v}_i - v)^2}{\sigma_i^2}\right] \tag{15}$$

Differentiating $\Pr(v|\hat{v}_i,\lambda)$ w.r.t v and setting to zero gives the MAP estimate of v, $\hat{v}_{MAP}$ as

$$\hat{v}_{MAP} = \left[\frac{\mu_{SI}}{\sigma^2_{SI}} + \sum_{i=1}^{M} \frac{\hat{v}_i}{(1-r_i^2)}\right] \Bigg/ \left[\frac{1}{\sigma^2_{SI}} + \sum_{i=1}^{M} \frac{1}{(1-r_i^2)}\right] \tag{16}$$

Note that if all the $\hat{v}_i$'s had the same variance σ^2 and $E\{\hat{v}_i\} = \bar{v}$, it can be shown that equation 16 becomes

$$\hat{v}_{MAP} = \frac{M\sigma^2_{SI}}{\sigma^2 + M\sigma^2_{SI}}\bar{v} + \frac{\sigma^2}{\sigma^2 + M\sigma^2_{SI}}\mu_{SI} \tag{17}$$

which is a well-known MAP estimate for a mean with prior density $\mathcal{N}(\mu_{SI}, \sigma^2_{SI})$ and a set of M examples with sample mean $\bar{v}$ and variance σ^2 (see, for instance [8]).

8 Experimental Procedure and Results

The data and models used in the experiments have already been described in section 5.

When testing a speaker, a subset of the enrollment classes were selected (see section 8.2) and all three of the speaker's utterances from each of these classes were used for adaptation. The HMMs of all test classes were then adapted, as were the HMMs for the enrollment classes for which data was available. However, only the utterances of the *test* classes from the new speaker were then tested. The reason for this is that testing on utterances which have already been used to adapt the models leads to biased results. Since only three utterances of each class were available from a speaker, the number of utterances available for enrollment would have been unacceptably small if some of them had been held back for testing purposes. Adaptation of a test class consisted of replacing the SI mean of the HMM state with either $\hat{v}_{ML}$ or $\hat{v}_{MAP}$; adaptation of an enrollment class consisted of replacing the SI mean of the HMM state j of class i with $\bar{x}_{i,j}$. It was found that adapting the enrollment class means made very little difference to the error-rate when testing only the test classes.

An error-rate of 17.0% was recorded for the unadapted system when testing on the test classes only. Note that the number of utterances in the test set is (52 speakers $\times$ 13 classes $\times$ 3 utterances) = 2028. In practice, some utterances were missing and the number actually used was 1984.

8.1 Effect of varying number of states used in prediction

With all 13 enrollment classes used, giving a total of 130 states available for prediction of each test class state value, the effect on (a) recognition accuracy and (b) prediction error 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 by absolute value of correlation coefficient averaged over the vector dimensions (see section 6.4), 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 error-rate' was measured as *(Actual-value – Predicted value)*^{2} / *Actual-value*^{2}. Prediction and recognition performance track quite closely (as might be expected) and peak when only 5–6 of the 130 available states are used. An examination of the identities of the best 5 enrollment class states contributing to prediction showed that a typical pattern is for two of them to be neighbouring states drawn from the same enrollment class, and the others to be states representing phonetically similar events from different classes. This means that the predicted value is synthesised from scattergrams of the test class state with several different enrollment class states, and the relative independence of each estimate leads to a good estimate of the predicted value. If more than about 5 states are used for prediction, the lower ranking states contribute noise to the estimate and performance worsens slightly.

Fig 2 compares prediction performance for the ML and MAP estimates. As expected, the MAP estimate is superior when only a few states are used for prediction but when more than about 6 states are used, the two estimates are very close and give very similar results.

8.2 Effect of varying number of enrollment classes available

The effect on recognition error-rate of increasing the number of enrollment classes available from a speaker is shown in Fig 3. When the number of classes available is $1, 2, \ldots, 13$ the number of states available for prediction is $10, 20, \ldots, 130$ and the result of section 8.1 showed that error-rate varied with P, where P was the number of 'best' states used for prediction. In Fig 3, the recognition error-rates were found by selecting an optimum value for P which was in the range 5–10. It was noticeable that P was larger when fewer classes were available. The implication is that when several enrollment classes are available, there is a good chance that one of them will be an excellent predictor for a particular test class and hence only a few states are required to make a good prediction. However, if there are only a limited number of enrollment classes available, the states of which all have mediocre correlation to the test class states, the best prediction is obtained by using a fairly large

number of such states.

For the result shown in Fig 3, the enrollment classes were made available in the following order:
'E','A','X','I','Q','D','K','M','P','F','G','O','R' e.g. when the abscissa in Fig 3 is 4, the classes used for enrollment were 'E','A','X' and 'I'. The rationale for the order given above was to make the classes most useful for adaptation purposes available early on. The greatest source of error in the test classes is the four 'E'-set words ('B','C','T' and 'V') and providing 'E' as the first enrollment class enables these to be rapidly adapted. 'A' is then given to aid adaptation of the classes 'J' and 'H', etc. A striking result from Fig 3 is that 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%.

Also included on Fig 3 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. For further details of this work, see [7].

The effect of presenting the enrollment classes in a different order is shown in Fig 4 where the result for the MAP estimate of Fig 3 is compared with the result when the order of presentation of the enrollment classes is reversed. Recognition performance for the reverse ordering is much worse until four enrollment classes have been presented, when it becomes almost indistinguishable from the original ordering. Experiments with other orderings confirmed that performance after any four enrollment classes have been provided is similar at an error-rate of roughly 6%.

Fig 5 plots the number of errors before adaptation against the number of errors after adaptation for each of the 52 speakers in the test set (using the MAP estimate and $P = 5$). In Fig 5a, the only enrollment class given was 'E', and although the average error-rate has dropped from 17% to 12.8%, the presence of several points above the equal error line shows that some speakers perform worse after adaptation. However, when the three enrollment classes 'EAX' are given (Fig 5b), only one speaker is worse after adaptation and several speakers are now error-free. When all 13 enrollment classes are available, no speaker makes more errors after adaptation than before and 19 speakers are error-free.

9 Summary

Model-based speaker adaptation systems are capable of adapting a speech recognition system very rapidly to a new speaker's voice by predicting sounds as yet unheard from the

speaker. A previously-investigated model-based speaker adaptation method based upon linear transformation has been reviewed and found to be insufficiently powerful. An alternative method proposed and investigated in the paper is to build linear regression models between sounds and use these models for predictive purposes at recognition time. A technique for applying this scheme within the framework of continuous density hidden Markov models (CDHMMs) is described and ML and MAP estimates derived for the predicted values. When applied to a database of 104 speakers each speaking utterances of the alphabet, the method reduced the error-rate from 17.0% to 3.4% after full adaptation. The method has very modest on-line computational requirements.

The results obtained in this study are very encouraging and further work on applying the technique to large vocabulary speech recognition is in progress.

References

[1] A. Averbuch et al. Experiments with the TANGORA 20 000 word speech recogniser. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 701–704, 1987.

[2] J.R. Bellegarda, P.V. de Souza, A. Nadas, D. Nahamoo, M.A. Picheny, and L.R. Bahl. The metamorphic algorithm: a speaker adaptation mapping approach to data augmentation. *IEEE Transactions on Speech and Audio Processing*, 2(3):413–420, July 1994.

[3] A Bladon. Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: a thread. In F Fallside and W.A. Woods, editors, *Computer Speech Processing*. Prentice Hall International, 1985.

[4] S.J. Cox. Speaker adaptation in speech recognition using linear regression techniques. *Electronics Letters*, 28(2):2093–2094, October 1992.

[5] 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.

[6] 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*, pages 161–165, April 1990.

[7] S.J. Cox, R.D. Johnston, P.W. Linford, and K. Chikolowski. Performance of human listeners on an isolated alphabetic speech recognition task. In *Proc. The Institute of Acoustics*, pages 23–30, 1994.

[8] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[9] Dragon Systems Inc. *DragonDictate Version 3.0 User's Guide.* Dragon Systems Inc., 1st edition, 1994.

[10] P. Kenny, M. Lennig, and P. Mermelstein. Speaker adaptation in a large-vocabulary Gaussian HMM recogniser. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):917–920, September 1990.

[11] 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.

[12] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vettering. *Numerical Recipes.* Cambridge University Press, 2nd edition, 1992.

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

[14] K. Shinoda, I. Ken-ichi, and T. Watanabe. Speaker adaptation for demi-syllable based continuous density hmms. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1991.

[15] 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.

Class 'J' state	S^*	—r—	Class 'H' state	S^*	—r—
1	'G', state 1	0.67	1	'A', state 1	0.52
2	'G', state 2	0.65	2	'A', state 4	0.65
3	'D', state 2	0.43	3	'A', state 6	0.56
4	'A', state 4	0.48	4	'X', state 3	0.43
5	'A', state 4	0.56	5	'X', state 4	0.29
6	'K', state 5	0.64	6	'G', state 1	0.34
7	'K', state 7	0.67	7	'G', state 1	0.48
8	'K', state 8	0.61	8	'G', state 2	0.37
9	'K', state 9	0.74	9	'G', state 2	0.34
10	'K', state 10	0.65	10	'X', state 10	0.36

Table 1: The enrollment class state S^* best correlated with the states of the test classes 'J' and 'H'.