UNSUPERVISED SPEAKER ADAPTATION BY PROBABILISTIC SPECTRUM FITTING.

Stephen Cox[†] *and John Bridle*[‡]

[†]British Telecom Research Laboratories, Martlesham Heath, Ipswich IP5 7RE, U.K.

[‡]Speech Research Unit, Royal Signals and Radar Establishment,
St. Andrews Road, Great Malvern, Worcs WR14., U.K.

ABSTRACT

A general approach to speaker adaptation in speech recognition is described, in which speaker differences are treated as arising from a paramterised transformation. Given some unlabelled data from a particular speaker, we describe a process that maximises the likelihood of this data by estimating the transformation parameters at the same time as refining estimation of the labels. The technique is illustrated using isolated vowel spectra and phonetically motivated linear spectrum transformations and is shown to give significantly better performance than non-adaptive classification.

1. INTRODUCTION

Speaker adaptation systems recently reported in the literature ([3], [6]) have concentrated on finding transformations which map (in some sense) a representation of the speech of a new speaker (spectral, cepstral, LPC etc.) to that of a reference speaker. The transformations are computed to optimise this mapping and take no account of the data being speech. We have been experimenting with models for speaker adaptation based on acoustic- phonetic principles following the methods originally described by Hunt [4]. In contrast to most current methods, which require known samples of speech from the new speaker, we describe a process that uses unlabelled utterances from an unknown speaker to update estimates of "speaker parameters" at the same time as deciding the labels of the utterances. Such unsupervised speaker adaptation is advantageous in situations in which it would be impractical (or at least undesirable) for the new speaker to recite a known text. So far we have dealt with the problem of labelling sets of vowel spectra from an unknown speaker;this task is related to the problem of identifying a single polysyllabic word from an unknown speaker, a task at which humans are highly competent.

2. THE MODEL

We use a model which supposes that speech from a given speaker can be explained as the application of a transformation $\mathbf{T}$ (whose parameters $\mathbf{q}$ are characteristic of the speaker), on the parameters $\mathbf{p_1}, \mathbf{p_2}, \ldots \mathbf{p_N}$ of 'prototype' sound classes, which might represent sub-word units, whole words or phrases etc. The likelihood of the observations given sound class c and speaker parameters $\mathbf{q}$ is:

$$P(\mathbf{x} \mid \mathbf{T}(\mathbf{q}) \odot \mathbf{p}_c) \tag{1}$$

where $\odot$ represents the action of the transformation $\mathbf{T}(\mathbf{q})$ upon the prototype.We shall also represent this likelihood as:

$$P(\mathbf{x} \mid \mathbf{p}_c, \mathbf{q}) \tag{2}$$

The complete stochastic model includes the processes that generate the speaker parameters $\mathbf{q}$, the sound classes $\mathbf{p}$ and the observations $\mathbf{x}$.

2.1. Model Estimation

Suppose we are given a set of labelled training-data $\{\mathbf{x}_i\}$. The likelihood of observing all this data, for given model and speaker parameters, is given by:

$$L = \prod_i P(\mathbf{x}_i \mid \mathbf{p}_{c_i}, \mathbf{q}_{s_i}) \tag{3}$$

where c_i is the sound-class and s_i the speaker number of the i'th example. Maximum likelihood estimates of the parameters of the models and the transformation are found by optimising L over the $\mathbf{p}$'s and $\mathbf{q}$'s simultaneously.

2.2. Transformation Estimation from Unlabelled Data

Now suppose that we are presented with a set of *un*labelled data from a single speaker, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_X$. Given a set of prototypes, we would like to label the data, but we should use the fact that the same (unknown) value of $\mathbf{q}$ applies to all the data. We find the speaker transformation parameters, $\hat{\mathbf{q}}$, which maximise the likelihood of the observations when we do not know the classes:

$$\begin{aligned} \hat{\mathbf{q}} = \operatorname*{argmax}_{\mathbf{q}} L' &= \arg max_{\mathbf{q}} \prod_{l=1}^{X} P(\mathbf{u}_l \mid \mathbf{q}) \\ &= \arg max_{\mathbf{q}} \prod_{l=1}^{X} \sum_{i=1}^{N} P(\mathbf{u}_l \mid \mathbf{p}_i, \mathbf{q}) \end{aligned} \tag{4}$$

where $P(\mathbf{u}_l \mid \mathbf{q})$ is the total posterior likelihood of $\mathbf{u}_l$, and we have assumed equal prior probabilities for the classes. The most likely labellings are then

$$\hat{c}_l = \arg max_{c_l} P(\mathbf{u}_l \mid \mathbf{p}_{c_l}, \hat{\mathbf{q}}), \qquad l = 1, 2, \ldots, X. \tag{5}$$

Proc. IEEE Conf. on Acoustics, Speech and Sig. Proc., Glasgow, 1989, pages 294–297

In practice we maximise the log of L'

$$\log L' = \sum_l \log \sum_i P(\mathbf{u}_l \mid \mathbf{p}_{c_l}, \mathbf{q}) \tag{6}$$

using the derivatives with respect to the **p**'s and **q**'s.

3. EXPERIMENTAL DATA

We have begun our experiments by using examples of individual vowel spectra from different speakers, because much of the inter- speaker acoustic variation is contained within the vowels and using isolated spectra dispenses with the need for time alignment procedures.The data consisted of a single example of each of 11 vowels from 30 speakers (15 male, 15 female). Each (RP) speaker spoke the following /hVd/ words:

heed	hoard	hid
hood	head	who'd
had	hudd	hard
heard	hod	

The vowel spectra were obtained by processing the digitised utterances using the SRUbank filterbank analysis facility [1] to produce a 27-channel log-amplitude spectrum vector every 10 ms. The frames corresponding to the steady-state portion of the vowel were excised and averaged to form a single 27-dimensional vector representing the spectral cross-section of the vowel. Each vector was normalised by subtracting the average value over its components from each component.The dataset was nominally divided into a training-set of speakers 1-16 (8 male, 8 female) and a test-set of speakers 17-30. The 11 sound-class models used here were 27-dimensional Gaussian densities with a common diagonal covariance matrix:

$$P(\mathbf{x} \mid \mathbf{p}) = \frac{1}{z} \exp\left[-\sum_{k=1}^{27} \frac{((x(k) - \mu(k))^2}{\sigma(k)^2} \right] \tag{7}$$

where z is a common normalising factor. In the baseline system (without transformations), the means of the prototypes μ_i were the sample means $\mathbf{m}_i$ of each vowel-class across the training-set speakers and the variances $\sigma_i(k)^2$ were the sample variances averaged across the classes.

4. THE SPECTRUM-BIAS TRANSFORMATION

The simplest speaker transformation we used (the "spectrum-bias model")supposes that a vowel spectrum of class i from a particular speaker $j$is generated by "filtering"the prototype spectrum of class i with a fixed spectrum shape δ^jcharacteristic of the speaker.Since log-amplitude spectra were used in these experiments, the transformation of the class means is purely additive.The best estimate of the prototype mean vector μ_i for class i is then the training-set mean vector $\mathbf{m}_i$ and the model takes the form:

$$\begin{aligned} P(\mathbf{x} \mid \mathbf{p}_i, \mathbf{q}^j) &= P(\mathbf{x} \mid \mathbf{m}_i, \sigma_i, \delta^j) \\ &= \frac{1}{z} \exp\left(-\sum_{k=1}^{27} \frac{(x(k) - m_i(k) - \delta^j(k))^2}{\sigma(k)^2} \right) \end{aligned}$$

Given X labelled examples from a particular speaker, the maximum likelihood estimate of δ is then:

$$\delta = \frac{1}{X} \sum_{i=1}^{X} (\mathbf{x}_{c_i} - \mathbf{m}_{c_i}) \tag{8}$$

where c_i is the class of the i'th example. Hence δ = the average distance between the speaker's vowel spectra and the corresponding means, a result which accords with intuition. The model was tested on the training-set data. Results are given in Table 1 below:

	No of errors (176 tests)
No adaptation	20
Spectrum-bias adaptation	13

Table 1: Comparison of no adaptation and spectrum-bias adaptation on training- set

4.1. The spectrum-bias model on unlabelled data

Assuming we are given a set of unlabelled data from a single speaker, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_X$, differentiating the log likelihood of the data given the model (equation 6) with respect to the unknown speaker transformation parameters $\delta$and setting to zero gives:

$$\hat{\delta} = \frac{1}{X} \sum_{l=1}^{X} \left(\mathbf{u}_l - \sum_{i=1}^{N} w_{il} \mathbf{m}_i \right) \tag{9}$$

where the w_{il}'s are weights($\sum_{l=1}^{N} w_{il} = 1$) derived from

$$w_{il} = \frac{P(\mathbf{u}_l \mid \mathbf{m}_i, \sigma_i, \delta)}{\sum_k P(\mathbf{u}_l \mid \mathbf{m}_k, \sigma_k, \delta)} = P(c_l = i \mid \mathbf{u}_l, \delta) \tag{10}$$

In practice we use a smoothed version of this posterior distribution over classes,by introducing a factor, T,which is analogous to the "temperature" in simulated annealing:

$$w_{il} = \frac{P(\mathbf{u}_l \mid \mathbf{m}_i, \sigma_i, \delta)^{\frac{1}{T}}}{\sum_k P(\mathbf{u}_l \mid \mathbf{m}_k, \sigma_k, \delta)^{\frac{1}{T}}}. \tag{11}$$

When $T = 1$, the weights are the probabilities derived from the normalised posterior likelihoods. However, $T \to 0$ has the effect of driving the weight of the most likely class to 1 and the others to 0, which is equivalent to assigning $\mathbf{u}_l$ to the most likely class. Conversely, as $T \to \infty$, the weights of each class become equal.To test this unlabelled adaptation/classification process, X vowel spectra were picked at random (without replacement) from the 11 available for a particular test-set speaker, δ was estimated using equation 9 and applied to the X spectra, the spectra were classified and the accuracy noted.This process was repeated for all ways of choosing X vowels from 11, for $X = 1, 2, \ldots, 11$ and for all speakers in the test-set. Fig 1 shows the recognition accuracy as a function of X (averaged over all ways

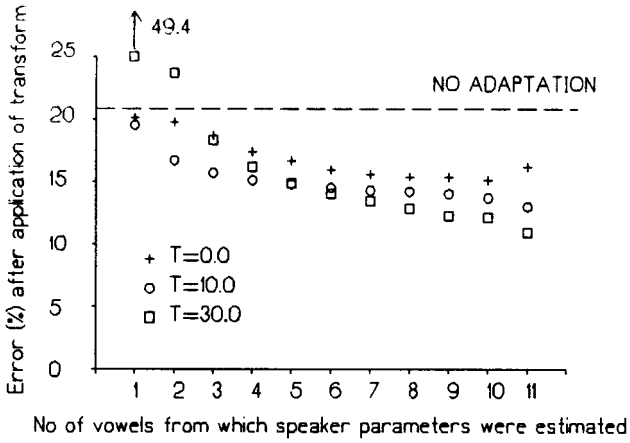

Figure 1: Unsupervised adaptation (using spectrum-bias model) at 3 different "temperatures"

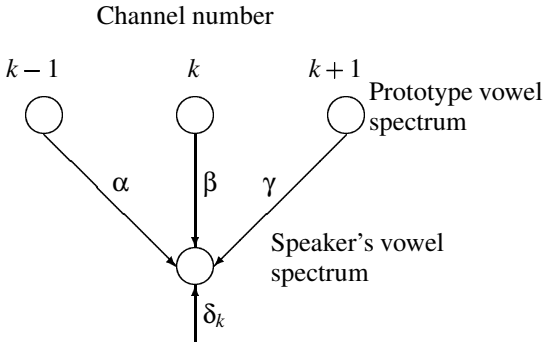

Figure 2: A model for vowel spectrum production incorporating shift and bias

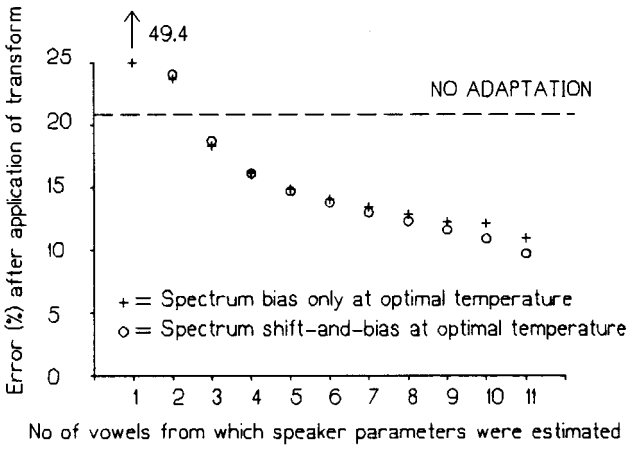

Figure 3: Comparison of unsupervised adaptation performance using spectrum-bias model and spectrum shift-and-bias model.

of choosing X from 11 and over all 14 test- set speakers) for 3 different 'temperatures': Fig 1 shows the beneficial effect of smoothing the likelihoods: as the number of vowels used for parameter estimation increases, the optimum value of T increases and the resulting percentage accuracy also increases. The curve marked '$T = 0$' was made by assigning each $\mathbf{u}_l$ to its most likely class.

5. THE SPECTRUM SHIFT-AND-BIAS TRANSFORMATION

A slightly more sophisticated model of inter-speaker spectrum differences takes into account the fact that, if the spectrum is measured on an appropriate scale, differences in realisations of the same vowel between speakers can be at least partially explained by sideways shifts of the spectrum [2]. An appropriate frequency-scale is the Bark scale, whose characteristics are approximated by the SRUbank filter-set. Given a filterbank representation of the vowel data, the following model represents a vowel produced by a particular speaker as a shifted and biased version of the corresponding vowel prototype:
For a given speaker (j),the prototype for a given sound-class (i) is modified as follows:

$$\mu_i^j = \mathbf{A}^j \mu_i + \delta^j \tag{12}$$

where $\mathbf{A}$ is the 27×27 tridiagonal matrix:

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

Notice that the α, β, γ do not depend on the channel number k, so that the shift is assumed to be the same throughout the spectrum and there are only 3 more parameters to estimate than in the spectrum-bias model.If $\alpha = \gamma = 0$, $\beta = 1$, the model reduces to the simpler spectrum-bias model.As sideways shifts of the spectrum are allowed in this model, it is clear that the best estimate of the prototype vowel spectrum means μ_i may no longer simply be the sample mean spectrum for each class. Prototype mean spectra were estimated from the training-set by applying a nonlinear optimisation technique (conjugate gradient method with an approximate line search [5]) to maximise the likelihood of the training-data. Given these prototype mean spectra, α, β, γ and δ for a test-set speaker were estimated from unlabelled data by least-squares fitting using the likelihood-weighting techniques described in section 4.1.A comparison of results using the spectrum shift- and-bias model and the simpler spectrum bias-model, both at their optimum 'temperatures' is given in Fig 3: When the number of unlabelled vowel spectra available for parameter estimation is fewer than 4, there is no advantage in using the shifting model, but the

shift-and-bias model is clearly superior when more than 4 are used.

6. CONCLUSIONS

We propose a model-based approach to speaker adaptation and have shown that two simple models (based on work by Hunt) are useful in explaining differences in vowel spectra between talkers. An exciting possibility demonstrated by this work is that, given some unlabelled utterances from a speaker, it is possible to estimate some parameters characteristic of the speaker whilst simultaneously labelling the utterances. Moreover, it has been shown that recognition accuracy is improved by this process. So far, we have confined ourselves to simple linear transformations of spectra,but the theory is quite general.The principles presented here have also been cast in terms of feed-forward non-linear networks and error back-propagation, which opens up the possibility of much more general, non-linear, discriminants and speaker transformations. To test the viability of the models, we also confined ourselves to isolated vowel spectra data. Extension of the techniques to whole word hidden Markov models should be straightforward and is the next immediate goal.

Acknowledgments

We are grateful for the help and advice of Dr K.Ponting, Dr A.Webb and Dr D.Lowe of the Speech Research Unit throughout the course of this work. We also thank Ros Temple of the Oxford University Phonetics Department and Neil McCulloch of the National Electronics Research Initiative in Pattern Recognition for providing the data used in the experiments.Acknowledgment is made to the Director of Research, British Telecom Research Laboratories for permission to publish this paper.

7. REFERENCES

[1] Details of the SRUbank specification may be obtained from the Head of the Speech Research Unit, RSRE, St Andrews Road, Gt Malvern, Worcs. WR14 3PS.

[2] 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.

[3] K. Choukri and G. Chollet. Adaptation of automatic speech recognisers to new speakers using canonical correlation techniques. *Computer Speech and Language*, 1:95–107, 1986.

[4] M.J. Hunt. Speaker adaptation for word based speech recognition systems (abstract only). *J. Acoust. Soc. Am.*, 69:S41–S42, 1981.

[5] J.C. Nash. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. Adam Hilger, 1979.

[6] R.M. Schwartz, Y.L Chow, and F Kubala. Rapid speaker adaptation using a probabilistic spectral mapping. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 633–636, April 1987.