# Simultaneous speaker normalisation and word recognition using neural networks/Bayesian techniques.

*Stephen Cox* † *and John Bridle* ‡

† British Telecom Research Laboratories, Martlesham Heath, Ipswich IP5 7RE, U.K.
‡ Speech Research Unit, Royal Signals and Radar Establishment, Malvern, Worcs., U.K.

## 1. Introduction

One of the most important problems in Automatic Speech Recognition (ASR) is methods of dealing with differences in the voices of different persons (and in some cases differences between the voice of a single person on different occasions). Most attempts to construct ASR systems which can be used by many people have used either multiple models for each word, or so called speaker-independent models. In either case, when decoding a short sequence of words there is no way of imposing our knowledge that all the speech is uttered by one person.Although there is probably a lot that can be done by using representations of the acoustic data that are less sensitive to speaker differences than the current techniques, it is generally accepted that high-performance ASR systems will need to "tune in" to characteristics of the speaker, probably at several levels such as acoustic and phonological. In this paper we are concerned with speaker characteristics which can be expressed as transformations of baseform models, with the transformation parameterised by continuous "speaker variables". We are especially interested in the possibility of estimating such parameters from quite small amounts of unlabelled speech, such as a few short words or one longer word, so adaptation and recognition are combined. Although the types of models and transformations we have used are very simple, we hope the general approach will be applicable to quite sophisticated models and transformations which will be necessary for future high-performance ASR systems. This is a continuation of work on simultaneous adaptation and recognition of vowel spectra reported in [4].

## 2. A Neural network approach to simultaneous recognition and normalisation

In this section we outline a rather general but intuitively accessible version of our approach, based on adaptive networks. Suppose we have a network with three (vector-valued) terminals, which encapsulates our knowledge of the relationship between acoustic patterns, **X**, class labels (e.g. word identities) **C**, and speaker parameters, **Q**. Imagine for now that the network works like a Boltzmann Machine [6]: we can clamp any pattern onto any of the three terminals, and the appropriate conditional distributions appear on the other terminals.Training such a network seems difficult, because although we can supply (**X**,**C**) pairs, we do not know the appropriate values of **Q**. However, we do have sets of (**X**,**C**) pairs from the same speaker, so we imagine a set of networks, one for each training-pair, with the **Q** terminals for networks for the same speaker strapped together (we do not mind what value of **Q** is used for each speaker, but it must be consistent.) Training the network proceeds in two phases. First, we keep the **Q** inputs at their default settings (e.g. all zero) and train a speaker-independent system. Then we use the same-speaker Q- strap trick: derivatives of the classification error measure are propagated back to the strapped **Q** terminals and used to adjust the speaker parameters. This is done separately for each speaker, and other parameters of the network are adjusted also. Depending on the network structure, the Q-strap can alter the parameters found by speaker-independent adaptation, and in particular we try to learn an appropriate mapping from speaker parameters **Q** to modifications to the $\mathbf{X} \rightarrow \mathbf{C}$ transformation. Once the network is trained we have two modes of use. If we have available one or more known utterances by our speaker, then we can "tune-in" to the speaker (as during training) except that only the **Q** inputs are adjusted. The case of most interest in this paper, however, is when we have a few *unknown* words from an *unknown* speaker. We set up a **Q**- strapped set of networks, one for each word, initialise the **Q** values to their defaults, propagate forwards to produce a set of distributions across word labels, and then we use a technique which tends to sharpen these distributions. In the simplest case, the sharpening process could be a matter of:

- for each utterance, pick the word label with the largest output;
- assuming this is correct, back-propagate derivatives to the common **Q**;
- repeat;

In practice, we can use a more gentle method in which large outputs get most 'encouragement'. For some networks it is possible to show that such a "phantom target" procedure can lead to hillclimbing on the likelihood of the data given an assumption about the form of the generator of the data [3]. As an example of a network with an interpretation in terms of a set of hidden Markov model (HMM) word recognisers, see [2]. In the following sections, we describe the system from a Bayesian point of view - we find it useful to have *both* persepectives.

## 3. The Bayesian perspective

In the experiments reported here, the **Q**-parameterised transformations were of a very simple kind (linear) and there was no learning of the control mapping. The $\mathbf{X} \rightarrow \mathbf{C}$ mapping is via a set of $N$ whole-word HMM word recognisers (one for each class) and we attempt to find modeland speaker parameters which together maximise the likelihood of the data. In practice, we estimate parameters in two stages as described in section 2.Firstly, all speaker-specific parameters are set to zero and the Baum-Welch algorithm is used to estimate parameters of speaker- independent HMMs. Standard optimisation techniques are then used to estimate parameters for a particular speaker. In our previous work on spectra of isolated vowels [4], we examined the effect on recognition accuracy of simple linear trans-

formations applied to a set of speaker's vowel spectra. If the adaptation is supervised, the likelihood to be maximised is:

$$L = \prod_i P(\mathbf{x}_i \mid \mathbf{p}_{c_i}, \mathbf{q}_{s_i}) \quad (1)$$

$\mathbf{x}_i$ is the $i$'th dataexample, $\mathbf{p}_{c_i}$ are the model parameters of the class of the $i$'th example and $\mathbf{q}_{s_i}$ are the transformation parameters appropriate to the speaker of the $i$'th example. For unsupervised adaptation, we must sum over classes (which we assume equiprobable). Hence for data from a single speaker, equation 1 becomes:

$$L' = \prod_i \sum_{j=1}^{N} L_j(i) = \prod_i \sum_{j=1}^{N} P(\mathbf{x}_i \mid \mathbf{p}_j, \mathbf{q}) \quad (2)$$

where $N$ is the number of different classes and $\mathbf{q}$ are the speaker-specific transformation parameters. We considered two parameterised transformations of the spectrum data, both originally proposed by Hunt [7].In each, the vowel classes were modelled as multivariate Gaussian distributions with common diagonal covariance matrices. In the first (the"spectrum-bias only" model), therealisation of a vowel class by a particular speaker was modelled as a bias spectrum shape $\delta$ added to the class mean spectrum shape:

$$\mathbf{x}_i = \mathbf{m}_{c_i} + \delta + \epsilon_i \quad (3)$$

where $c_i$ is the class of the $i$'th example and $\epsilon_i$ is the residual modelling error. A more sophisticated model (the "fixed shift and bias" model ) allowed for shifts along the frequency axis of the speakers's spectra relative to the mean spectra as well as the bias:

$$x_i(k) = \alpha m_{c_i}(k-1) + \beta m_{c_i}(k) + \gamma m_{c_i}(k+1) + \delta(k) + \epsilon_i(k) \quad (4)$$

Here, the values of $\alpha, \beta$ and $\gamma$ are constant with frequency-band so that the shift is the same throughout the spectrum. In this paper, we also report results using values of $\alpha, \beta$ and $\gamma$ which vary along the spectrum(the "variable shift and bias" model ).

# 4. The data and the speaker-independent models

The data used throughout these experiments was from the BT connectionism project database [9]. This consists of utterances from 104 speakers, divided into a training-set of 52 speakers and a test-set of 52 speakers, each set having equal numbers of males and females. The speakers recorded in random order three utterances of each of the letters of the alphabet. The recordings were made in an acoustic booth using a high- quality wideband microphone at a sampling frequency of 20kHz and using a 16 bit A/D converter. After checking the utterances,eliminating bad ones and manually endpointing good ones, a total of 3999 training utterances and 3978 test utterances remained. These were parameterised to 27 log magnitude filterbank channel outputs using the SRUbank filterbank analysis facility [1]. Each frame was normalised by subtracting the average value from each component and a 28th energy channel was appended to the frame.The speaker-independent models used in the experiments were continuous density HMMs with 15 states and a mixture of 3 Gaussian distributions per state.The topology of the models allowed transitions from state $j$ to state $j$ and $j+1$ only. The states and mixtures of a given model shared a common diagonal covariance matrix. Model parameters were estimated using the Baum-Welch training algorithm. The models also incorporated a "voicing factor" (see section 5.1).

# 5. Adaptation using whole word data

To extend the transformation techniques developed on single frame data to whole word data, it is neccessary to deal with *sequences* of frames representing a class. The class models are now whole-word HMMs and the adaptation is performed only on the state parameters of these models. The techniques described here could be extended to enable adaptation of model transition probabilities but for present purposes, these are not changed.

## 5.1. Voicing marking

The transformations discussed in section 3 are suitable for voiced speech only - our techniques do not yet attempt to deal with unvoiced sounds.Accordingly, every frame $j$ of the database was annotated with a real number $v_j^f$ ($0 \leq v_j^f \leq 1$), which gave the "voicing factor"; $v_j^f = 0$ for unvoiced sounds, $v_j^f = 1$ for fully voiced. This was done automatically using information derived from the amplitude of low order cepstral coefficents and low-frequency spectrum coefficents.The voicing information is used in two ways. Firstly, $v_j^f$ weights the 'significance'for adaptation of each data-frame: a frame with $v_j^f = 0$ is effectively ignored whenestimating a set of speaker parameters whereas a frame with $v_j^f = 1$ contributes fully. Frames with intermediate marking contribute in proportion to their voicing factor. Secondly, each HMM state $j$ of each word model $w$ was also marked with a voicing factor $v_{wj}^s$ at training time by applying Baum-Welch style re-estimation to $v_{wj}^s$. This voicing factor is used when the models are transformed to determine how much of the transformation each state should receive (see equation 9).

## 5.2. Likelihood-weighted adaptation

An important idea for unsupervised adaptation introduced in our previous work is that of *normalised likelihood weighting*. To illustrate this, consider the estimation of $\delta$ in the model given in equation 3. When theclass of each $\mathbf{x}_i$ is unknown, our estimate of $\delta$ is:

$$\hat{\delta} = \frac{1}{X} \sum_{i=1}^{X} \left( \mathbf{x}_i - \sum_{j=1}^{N} p_j(i) \mathbf{m}_j \right) \quad (5)$$

where $X$ is the number of examples and$\sum_j p_j(i) = 1$ i.e. a probability weighted sum of the prototypes is used. The weights, $p_j(i)$, are pseudo posterior probabilities derived from the likelihoods of the data given the models, $L_j(i)$, as follows:

$$p_j(i) = \frac{L_j(i)^{\frac{1}{T}}}{\sum_k L_k(i)^{\frac{1}{T}}} \quad (6)$$

$T$ is a factor which sharpens ($T \to 0$) or smooths ($T \to \infty$) these weighting functions. The same idea is used in the estimation of any speaker parameters from unknown data given models. In our work on isolated vowels, each unknown vowel spectrum had a likelihood associated with each vowel model. Using whole word data, each *frame* of data has $N * N_s * N_m$ associated likelihoods, where $N$ is the number of models, $N_s$ is the number of states in each model and $N_m$ the number of modes in the mixture in each state. The weighting factor for the mean of mode $k$ of state $j$ of model $i$ for a frame at time $t$ is

thus:

$$p_{wjk}(t) = \frac{\gamma_{wjk}(t)^{\frac{1}{T}}}{\sum_{i=1}^{N}\sum_{j=1}^{N_s}\sum_{k=1}^{N_m}\gamma_{wjk}(t)^{\frac{1}{T}}} \quad (7)$$

where $\gamma_{wjk}(t)$ is the probability of occupying mode $k$ of the mixture of state $j$ of model $w$ at time $t$. When these $p_{wjk}(t)$ have been calculated for each frame of each unlabelled utterance offered for adaptation, a weighted mean $\mathbf{m}^w(t)$ for each time $t$ is computed as:

$$\mathbf{m}^w(t) = \sum_{i=1}^{N}\sum_{j=1}^{N_s}\sum_{k=1}^{N_m} p_{wjk}(t)\mathbf{m}_{wjk} \quad (8)$$

where $\mathbf{m}_{wjk}$ is the mean of mode $k$ of the mixture of state $j$ of model $w$.This weighted mean is exactly analogous to the weighted mean in equation **??**. Once these weighted means have been calculated, the speaker parameters are calculated using the data-frame/weighted mean pair. In the caseof the spectrum bias transformation, this can be done directly using the equivalent of equation **??** where $X$ is the total number of frames. For the shifting transformations, we used a nonlinear optimisation technique(conjugate gradient method with an approximate line search [8]) to estimate parameters.In both cases, the mixture mode means are then updated according to:

$$\mathbf{m}'_{wjk} = (1 - v^s_{wj})\mathbf{m}_{wjk} + v^s_{wj}\mathbf{T}[\mathbf{m}_{wjk}] \quad (9)$$

where $\mathbf{T}[.]$ represents the effect of the transformation on themeans. The effect of the voicing factor is to adapt each mean according to the degree of voicing associated with it.In practice, we found this Baum Welch style procedure for calculating weighted means to be verycomputationally expensive and instead adopted the following Viterbi procedure which gavealmost no deterioration in performance:

1. Each utterance for adaptation was Viterbi aligned to each model. The alignment of frame $x(t)$ to a state in each model $M_w$ is noted as $s_w(t)$ and the mixture mode contributing the highest likelihood in this state was noted as $n_w(t)$.
2. The likelihoods of the whole utterances are normalised across the models i.e. for a given utterance:

$$p_w = \frac{V_w{}^{\frac{1}{T}}}{\sum_w V_w{}^{\frac{1}{T}}} \quad (10)$$

   where $V_w$ is the Viterbi likelihood of model $w$ for this utterance.
3. For each time $t$, a weighted mean is computed as:

$$\mathbf{m}^w(t) = \sum_{i=1}^{N} p_w \mathbf{m}_{w,s_w(t),n_w(t)} \quad (11)$$

4. Parameter estimation then proceeds as above.

## 6. Experimental procedure and results

In [4], our scenario was one in which a new speaker uttered a few vowel sounds of unknown class, from which we estimated the speaker transformation parameters. The new transformation was then applied to the models and the utterances were re-classified. We have used the same experimental paradigm in these experiments with whole word utterances. Sets of utterances were drawn randomly (without replacement) from the pool of 78 utterancesnominally available from each speaker, until the pool was exhausted. Because (no of utterances available) / (no of utterances in a set) was not alwaysinteger, the last set was sometimes smaller than the other sets, but each utterance was used only once in an adaptation set. The results shown are averages across all the sets used and across all 52 test-set speakers.Fig 1 indicates the best performance we can hope for using a given type of transformation. In these experiments, the speaker parameters were estimated using the true class of all the available utterances (supervised adaptation).The efect of using the "voicing factor" is also shown. In the bias only (BO) and fixed shift and bias (FSB) transformations, use of the voicing- factor produces a small but (using the test described in [5]) statistically significant improvement. The result for variable shift and bias (VSB) is interestingbecause the result is much better when the voicing information is ignored. This may be because the VSB transformation is able to produce appropriate shifts for unvoiced sounds as well as voiced sounds, but it works more effectively with more data because of the greater number of parameters associated with it (107 parameters, as opposed to 31 for FSB and 28 for BO). Fig 1 suggests that the power of these transformations is limited, although the very significant improvement in error rate when the VSB transformation was given more data for adaptation may indicate its power for supervised adaptation on larger amounts of data. Figs 2 and 3 show results using unsupervised adaptation. Fig 2 shows the effect of the "temperature" parameter used on the data likelihoods. All utterances were used for adaptation, as was voicing information. The adaptation is unstable for the two shifting transformations when$T \rightarrow \infty$ i.e. when the likelihoods of all classes are set equal. Although the modelling error ($\sum_{i,k}\epsilon_i^2(k)$ in equation 4) decreases during the process of parameter estimation, the sum of the Viterbi likelihoods of the data is lower after model adaptation than before adaptation. This points to the deficency of estimating first the model parameters and then the speaker parameters. These two sets of parameters should be jointly re-estimated starting from their defaults i.e. the null transformation for the speaker parameters and the speaker-independent values for the model parameters. Fig 3 simulates a real application with either 3,10,20 or 78 utterances for adaptation, using a hard decision on the class ($T \rightarrow 0$) and voicing information. This figure shows that the FSB transformation performed marginally better than the other two for a limited number of utterances, probably because ithas the best compromise of power and number of parameters to be estimated. For each transformation, the error-rate generally decreases as the number of adaptation utterances increases, but by only a small amount. However, the final error-rate is close (in all cases) to that obtained when supervised adaptation was used, which suggests that any limitation in power lies with the transformations rather than the unsupervised adaptation technique.In every case shown in these figures, the difference between the error-rate when using no adaptation and the error-rate after adaptation is significant at at least the 99.9% level.

## 7. Conclusions

The results reported here show that simultaneous word recognition and speaker normalisation can be made to work, that it improves performance over the corresponding speaker-independent version, and that given 3 to 10 unknown words, performance can be almost as good as when the adaptation is done using knowledge of the word identities. However, the rela-

tive improvements obtained by adaptation, although statistically significant, are possibly not of sufficient practical significance to warrant the extra complexity, and there are ways of improving speaker-independent performance.Nevertheless, we remain convinced that the general approach described here has potential. Among the improvements and additions we have in mind are:

- estimation and use of priors or other constraints on the speaker parameters
- integration of the model parameter estimation and the speaker transformation (even in the simplest (shift only) case this alters the variances of output distributions)
- training the set of word models (with speaker parameterisation) as a set of discriminators
- learning the form of a (non-linear) relationship between speaker parameters and word-model parameters
- application to connected word recognition or very large vocabulary isolated word recognition
- in an interactive application, the use of information from the dialogue to assist speaker adaptation

## 8. Acknowledgements

Acknowledgment is made to the Director of Communication Systems, Research and Technology, British Telecom for permission to publish this paper.

## 9. References

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

[2] J.S.Bridle. Alphanets: a recurrent neural network architecture with a hidden Markov model interpretation. *Speech Communication - special NeuroSpeech issue*, no 2, 1990.

[3] J.S.Bridle. Phantom target backpropogation networks for unsupervised data analysis - *forthcoming*

[4] 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.

[5] L.Gillick and S.J.Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. IEEE Conf. on Acoustics Speech and Signal processing*, pages 532- 535, April 1989.

[6] G.E.Hinton and T.J.Sejnowski. Optimal perceptual inference. In *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, pages 448-453, 1983.

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

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

[9] J.A.S. Salter. *The RT5233 Alphabetic Database for the Connex Project.* Technical Report RT52/G231/89, BT Technology Executive, 1989.