Proceedings of the Institute of Acoustics

# AN INVESTIGATION INTO DISCRIMINATIVE TRAINING OF INPUT TRANSFORMATIONS FOR CONTINUOUS SPEECH RECOGNITION

J.Bridle*, P.Nowell, L.Dodd

Speech Research Unit, Defence Research Agency, St. Andrews Rd, Malvern, WR14 3PS, U.K.

## 1 INTRODUCTION

The most successful continuous speech recognition systems use statistical techniques such as hidden Markov models (HMMs) to model and thereby recognise speech. The success of these techniques is based upon the availability of powerful and tractable techniques for automatic parameter reestimation and speech recognition. Most of these speech recognition systems operate in the space of log power spectra, although the data may have undergone some simple, mostly linear, transformations such as frequency scale transformations (Mel-scale warping), cosine transformations, and temporal differencing.

We assume that significant improvements could be made to speech recognition performance if we had data representations in which phonetically important properties of the acoustic pattern were made more explicit, so that a simple model such as a gaussian HMM might end up with good, useful information for speech recognition Some success has been reported using transformations based upon linear discriminant analysis (LDA) which produces transformed data that has average within-class unit covariance, and selects linear features which are most useful for discriminating between classes (in a certain sense).

One way to move beyond simple LDA is to introduce distortions into the training data, [1] so that the resulting transformation will be robust to the distortions. Another direction, which we explore here, is to optimise a measure of performance which is more relevant to speech recognition. There is also reason to believe that non-linear transformations might provide even better performance. Such transformations could, for example, enhance peaks in the spectrum which are known to be important for speech recognition.

A unification of the theory of back-propagation "neural networks" and HMMs leads to a method, which we call the *Alphanet* approach [2], which can be used to address the problem of designing better representations. The Alphanet approach views the forward (alpha) HMM likelihood calculations as the main part of a large recurrent network of a special kind which eventually delivers sets of numbers which are treated as posterior probabilities of classes (e.g. words or word sequences). Partial derivatives of some error criterion are back-propagated through time, and eventually form partial derivatives with respect to the adjustable parameters of the system (in our case parameters concerned with the data representation for the models). We apply this notion to a *continuous* speech recognition system, in which a class label corresponds to a string of words or phonemes.

*Now with Dragon Systems UK Ltd., Millbank, Stoke Road, Bishops Cleeve, Cheltenham GL52 4RW

## 2 BASIC THEORY

Most speech recognition systems attempt to choose the interpretation, $w$, which is most likely given the acoustic data $\boldsymbol{y}$ and the models.

The *posterior* probability, given $\boldsymbol{y}$, of an hypothesis $W=w$ (ie given the acoustic pattern, and assuming that one of the hypotheses is true) is

$$P_w \triangleq P(W=w \mid \boldsymbol{Y}=\boldsymbol{y}) = \frac{L_w}{L}$$

where

$$L_w \triangleq P(\boldsymbol{Y}=\boldsymbol{y},\ W=w) = P(\boldsymbol{Y}=\boldsymbol{y} \mid W=w)P(W=w).$$

The two terms correspond to the acoustic models and the language model respectively. We have defined the likelihood of the data given the general model as $L \triangleq \sum_v L_v$.

Suppose that for each acoustic training pattern we know the correct interpretation, $W=c$, then for a single example pattern the relative entropy based score [3] (which we need to minimise) is minus the log of the output for the correct class.

$$J = -\log P_c = -\log(L_c/L) = \log L - \log L_c.$$

So for any parameter $\theta$ of the system,

$$\frac{\partial J}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta} - \frac{1}{L_c}\frac{\partial L_c}{\partial \theta} \tag{1}$$

We obtain $L_c$ by a forward full-likelihood (alpha-) pass making use of the correct transcription. To obtain $L$ we perform another alpha-pass, over a syntax which allows any interpretation (eg a simple forward-pass word-loop or phone-loop) which we call the *free* syntax. The free syntax corresponds to that used by a recognition algorithm. The *clamped* syntax used for $L_c$ is a restriction of this free syntax.

$\frac{\partial L}{\partial \theta}$ and $\frac{\partial L_c}{\partial \theta}$ can then be computed via a backward pass (i.e. by back-propagation through time). In reference [4] it is shown that derivatives of $J$ wrt the HMM parameters, and wrt the pattern $\boldsymbol{y}$ which is treated as the output of the HMM, can be expressed in terms of differences between clamped and free state occupancy probabilities. For instance, for single unit-covariance gaussian output distributions

$$\frac{\partial J}{\partial \boldsymbol{y}_t} = \sum_j (\bar{\gamma}_{jt}^f - \bar{\gamma}_{jt}^c)(\boldsymbol{m}_j - \boldsymbol{y}_t)$$

where $\boldsymbol{m}_j$ is the mean of the gaussian associated with the $j^{th}$ state, and $\bar{\gamma}_{jt}^f$ is the posterior probability of state $j$ at time $t$.

When $\boldsymbol{y}_t$ is the output of a data transformation at time $t$ then we can use this result to compute derivatives of $J$ wrt the parameters of the transformation.

## 3 EXPERIMENTS

Our starting point was a subword-based continuous speech recognition system using single gaussians with diagonal covariance matrix [5]. To reduce the computation and storage for the free passes, we used context-insensitive phoneme models ('monophone').

The data was artificial airborne reconnaissance mission (ARM) reports, spoken by a single speaker (MR). The vocabulary for the ARM task is 519 words. Each report lasts about 30 seconds. and there are 37 reports for training and 10 for testing.

We trained only *linear* data transformations, and started with the results of linear discriminant analysis [6], which should be a good initial linear transformation from which to begin the adaptation process. We made the transformation slightly more general by including a constant bias vector (initially zero) which is added to the result of the matrix multiplication. 61.8% of the phonemes were correctly recognised when the LDA transformation was applied to the input data.

The models we started from were the result of running Baum-Welch re-estimation on the output of the LDA transformation. The variances were set to unity (diagonal covariance matrix). When we adapted the models as well as the transformation, we adjusted just the means of the output distributions, using the Baum-Welch method.

The following table shows the recognition results on test and training sets after 11 iterations of adaptation of the input transformation. Adapting the input transformation did not improve recognition performance.

| | | | Phonemes Correct | |
|---|---|---|---|---|
| Evaluation Set | Transform | Models | Before | After |
| Test Set | Adapted | Original | 61.8% | 61.1% |
| Training Set | Adapted | Original | 57.8% | 59.7% |

Table 1: Recognition results for the full training and test sets

We decided to investigate by using a single ARM report for training. This was much faster ( 5mins per iteration compared with 4hrs for the full training set) and we could look in detail at the types of errors. The LDA models were first adapted to the LDA data in the single report, followed by 200 iterations of transformation adaptation or of alternating transformation adaptation and model re-estimation. The result was similar to that for the large dataset– adaptation hardly changed the recognition performance on the training data (78.8%). (We expect that performance on unseen data would be much worse, because there is so little training data.)

## 4 INTERPRETATION

The following graphs (figs.1,2) give some insight into the adaptation process. They show the clamped and free log likelihoods and the $J$ criterion which is the difference between them, as a function of iteration number for the single file experiments.

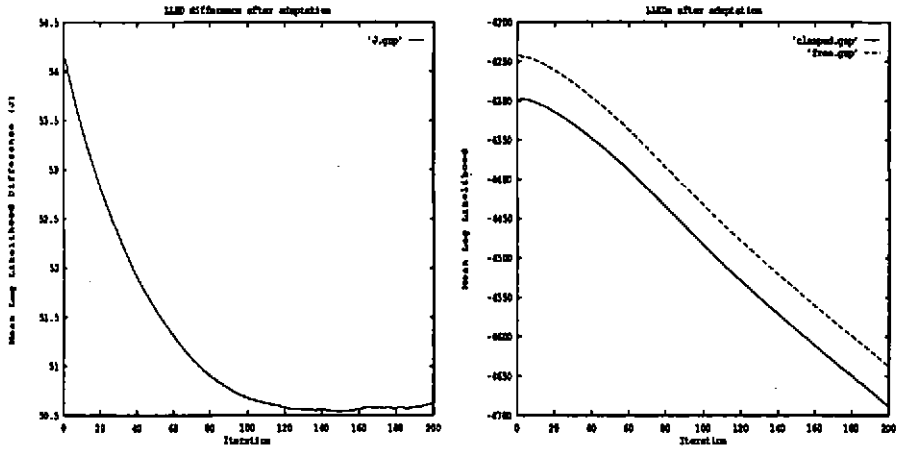

Figure 1: Error criterion 'J' (left) and clamped and free log likelihoods while adapting the transformation only

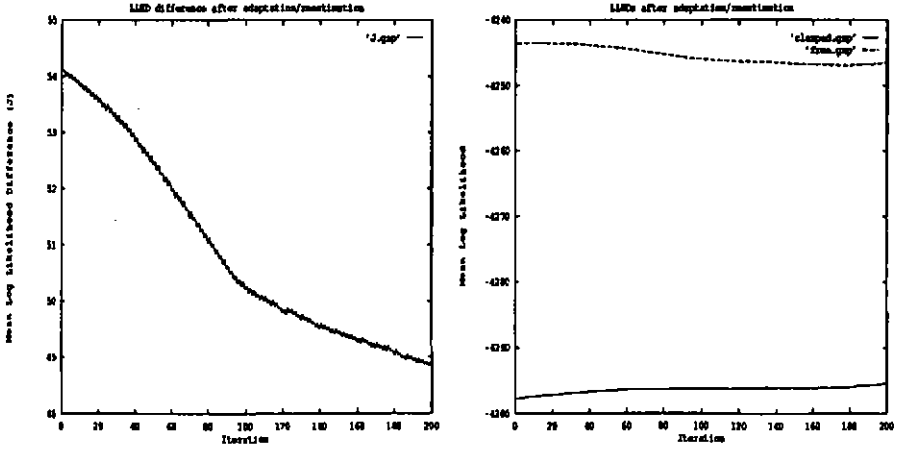

Figure 2: Error criterion 'J' (left) and clamped and free log likelihoods while adapting the transformation and model means

Adapting the transformation while leaving the models fixed causes both the clamped and free log likelihood to decrease (fig. 1 right), but they do get closer together (fig. 1 left).

When we alternately adapt the transformation and reestimate the models (fig. 2) the J plot looks somewhat similar, but has not settled into a minimum after 200 iterations. The clamped and free log likelihoods however behave very differently (note the change of scales between fig. 1 and fig. 2). The Baum-Welch re-estimation of the models increases the clamped log likelihood (on the even numbered iterations), which produces the jagged appearence of the J plot. This is not unreasonable since the Baum-Welch reestimation process aims only to minimise the clamped log likelihood distance and is not concerned with the difference between the clamped and free log likelihoods.

## 5 DISCUSSION

Further adaptive training of linear transformations starting with LDA transformations has produced no benefits in the performance of the speech recogniser over those obtained using the original LDA transformation. It is possible, though unlikely, that the LDA transformation is as good a linear transformation as we can get, and that whole-sentence criterion training cannot improve upon it. If this is the case we should look for ways of extending frame-by-frame style training, to include more relevant discrimination measures and non-linearities.

Alternatively, the gradient descent method which is used to adapt the transformation matrices may be too simple to cope with the task. More powerful techniques, such as conjugate gradients, could be needed to successfully adapt the transformation matrices.

In addition to these possibilities there are also some outstanding problems with the current experimental setup. The use of simple phoneme-type labels with context-insensitive models and single-mode state-conditioned distributions may be causing severe problems. We should repeat the experiments using either a much closer and more detailed phonetic transcription or a clamped syntax which allows for phonologically reasonable alternatives.

It also might help if adaptation and reestimation processes had the same goal. The adaptation process minimises the difference between the clamped and free log likelihoods whilst the Baum-Welch reestimations maximises the clamped log likelihood without regard to the free log likelihood. This results in a conflict of interests between the adaptation and reestimation processes which is highlighted by the jaggies in fig. 2. We should experiment with discriminitive training of the models as well as the transformation.

Further experiments could be performed which extend the current set of experiments on the ARM speaker dependent database using monophone models to the ARM speaker-independent database. Doing a free syntax forward-backward pass with triphone models would be very expensive, so we suggest using monophones or possibly a slightly more extensive set corresponding to acoustic phonetic elements, or a small set of context-sensitive models.

**References**

[1] M.J. Hunt and C Lefèbre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, 1989.

[2] S J Cox and J S Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, Glascow, 1989.

[3] J.S. Bridle. *Advances in Neural Information Processing Systems 1.* Morgan Kaufmann, 1989.

[4] J.S. Bridle and L. Dodd. An alphanet approach to optimising input transformations for continuous speech recognition. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, 1991.

[5] M.J.Russell, K.M.Ponting, S.M.Peeling, S.R.Browning, J.S.Bridle, R.K.Moore, I.Galiano, and P.Howell. The ARM continuous speech recognition system. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, 1990.

[6] S.M. Peeling and K.M. Ponting. The use of linear discriminant analysis in the ARM continuous speech recognition system. *RSRE Technical Memorandum 4512*, 1987.