INDICATOR VARIABLE DEPENDENT OUTPUT PROBABILITY MODELLING VIA CONTINUOUS POSTERIOR FUNCTIONS

A. Tuerk, S.J. Young

Cambridge University Engineering Department
Trumpington Street CB2 1PZ
e-mail:{at233,sjy}@eng.cam.ac.uk

ABSTRACT

This paper investigates the problem of inserting an additional hidden variable into a standard HMM. It is shown that this can be done by introducing a continuous feature which is used to calculate the probability of observing the different states of the hidden variable. The posteriors are modelled by softmax functions with polynomial exponents and an efficient method is developed for reestimating their parameters. After analysing a two dimensional reestimation example on artificial data, the proposed HMM is evaluated on the 1997 Broadcast News task with a particular focus on spontaneous speech. To derive a good indicator variable for this purpose, classification experiments are carried out on fast and slow classes of phones on the 1997 Broadcast News training data. Finally, recognition experiments on the test set of this task show that the proposed model gives an improvement over a standard HMM with a comparable number of parameters.

1. INTRODUCTION

Most HMM-based speech recognition systems use hidden states to model exclusively segments of phones. This ignores the fact that there are many other hidden states which influence the performance of a speech recogniser. Such states are, for instance, the current speaker, the acoustic background condition and the dynamics of speech like speaking rate. This paper describes a method of incorporating such hidden states into a standard HMM by introducing continuous posterior probability functions for a finite set of discrete hidden variables. The model discussed in this paper is a special case of the general framework developed in [1]. In contrast to [1], however, this paper makes use of softmax functions with polynomial exponents as posteriors and derives an efficient algorithm for the reestimation of the softmax parameters.

This work was in part funded by EPSRC grant reference GR/L49611

2. POSTERIOR PROBABILITIES OF HIDDEN STATES INSIDE A STANDARD HMM

In the following discussion d stands for a continuous feature which is used to calculate the posterior probability $p(v|d)$ of observing a discrete hidden variable v and o is a feature vector whose statistics are modelled by the output pdf's of the hidden states s of the HMM. One might think of o as a standard feature as used in most speech recognisers (PLP,MFCC, $\ldots$) and d as a feature which contains some additional information about the hidden variable v. For example, d should be a good measure of speaking rate if v is meant to be related to this hidden variable. Such measures have been discussed in [2] and [3]. In order to make use of the additional information contained in d and v the output probability for state s is altered as follows

$$b(o, d, v|s) = b(o|v, s)p(v|d, s)b(d|s) \tag{1}$$

Here $b(o|v, s)$ is the output pdf of feature o and states s and v, $p(v|d, s)$ is the posterior probability of observing v given feature d and state s and $b(d|s)$ is the output pdf of feature d and state s. Since v indicates which output pdf for feature o should be chosen for a given state s, v will be called an indicator variable. The posterior probabilities $p(v|d, s)$ are modelled by softmax functions $S_{v,s}$ which are defined as follows

$$S_{v,s}(d) = \begin{cases} \frac{1}{1+\sum_{k=2}^{K} e^{q_{k,s}(d)}} & : \quad v = 1 \\ \frac{e^{q_{v,s}(d)}}{1+\sum_{k=2}^{K} e^{q_{k,s}(d)}} & : \quad v \neq 1 \end{cases} \tag{2}$$

Here K is the number of hidden states of v and the $q_{v,s}$ are polynomials, i.e.

$$q_{v,s}(d) = \sum_{l=0}^{L} c_l^{v,s} d^l \tag{3}$$

where l can be a multi-index if d has dimension greater than one.

2.1. Reestimation

The parameters of the model introduced in (1) can be reestimated with the EM algorithm. The auxiliary function for this model is given by

$$Q(\lambda, \bar{\lambda}) = \sum_{\vec{s},\vec{v}} L_{\lambda}(O, D, \vec{s}, \vec{v}) \log L_{\bar{\lambda}}(O, D, \vec{s}, \vec{v}) \quad (4)$$

where

$$L(O, D, \vec{s}, \vec{v}) = \prod_{t=1}^{T} a_{s_{t-1},s_t} b(o_t, d_t, v_t|s_t) \quad (5)$$

To simplify the notation, the following abbreviations are introduced

$$\begin{aligned} \gamma_t(i,k) &= L(O, D, s_t = i, v_t = k) && (6)\\ \gamma_t(k|i) &= L(O, D, v_t = k|s_t = i) && (7)\\ \gamma_t(i) &= L(O, D, s_t = i) && (8) \end{aligned}$$

Note that these values can be efficiently calculated with the forward-backward algorithm. Now, omitting the term for the transition probabilities, (4) can be rewritten as

$$\begin{aligned} Q(\lambda, \bar{\lambda}) &= \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{t=1}^{T} \gamma_t(i,k) \log b(o_t|v_t = k, s_t = i) \\ &+ \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{t=1}^{T} \gamma_t(i,k) \log p(v_t = k|d_t, s_t = i)) \\ &+ \sum_{i=1}^{I}\sum_{t=1}^{T} \gamma_t(i) \log b(d_t|s_t = i) \quad (9) \end{aligned}$$

where I is the number of hidden states s. This shows that the parameters of the output pdf's $b(o|v,s)$ and $b(d|s)$ can be reestimated in the usual way. To reestimate the parameters of the polynomial $q_{k,i}$ the following equation has to be satisfied

$$\frac{\partial}{\partial c_l^{k,i}} Q(\lambda, \bar{\lambda}) = \sum_{t=1}^{T} \gamma_t(i)(\gamma_t(k|i) - S_{k,i}(d_t))d_t^l = 0 \quad (10)$$

This means that the moments up to order L of the error of the approximation to the true posterior $\gamma_t(k|i)$ by the softmax function $S_{k,i}$ weighted by the state likelihood $\gamma_t(i)$ have to vanish. Although this is not an explicit reestimation formula, the parameters of $S_{k,i}$ can be reestimated efficiently from (10) by using Newton's method with a line search and backtracking algorithm. Equation (10) is a necessary but not sufficient condition for the maximisation of the auxiliary function. In order to ensure that the estimated parameters are associated with a local maximum the second order derivative of the auxiliary function at this point has to be a negative definite matrix. This holds true because in the case of softmax functions with polynomial and more general exponents the error surface can even be shown to be concave [4]. Figures 1 and 2 give an example of a softmax parameter reestimation in two dimensions for a two-posterior problem. The feature d is located in a plane and v can have two different values. The state likelihood in this example was assumed to be uniform over the training region. Figure 1 shows one of the two true posterior probabilities (they both sum to one at each point in the d plane) and the initial guess for the approximating softmax function. The polynomial of degree 4 for the initial guess was chosen randomly to be $q(d_1, d_2) = d_1 + d_2$. The number of parameters that had to be reestimated in this case were therefore 15. Figure 2 shows different stages of the reestimation procedure. As can be seen the approximating softmax function converges to the proper solution. The fact that the softmax functions were initialised with a random polynomial illustrates the insensitivity of the reestimation procedure to bad initialisation. This robust behaviour of the reestimation algorithm was observed throughout the experiments.

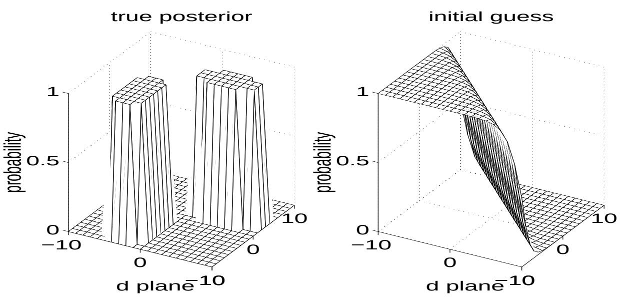

Fig. 1. Input posterior and initial guess in two dimensions.

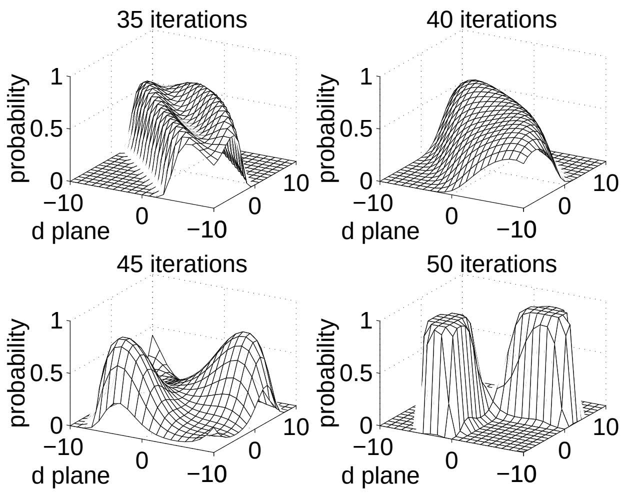

Fig. 2. Sequence of approximations to posterior probability in two dimensions.

2.2. Recognition

The models in (1) can be used for recognition in several different ways. Firstly, one can take the sum over the hidden variables v and therefore obtain

$$b(o, d|s) = \sum_{k=1}^{K} b(o, s, v = k|s) \tag{11}$$

In this case, the only states in the model are the s-states, and the v-states are factored out. Alternatively, one can expand the topology of the model to distinguish between different v-states as well. Figures 3 and 4 show two different ways of expanding a 5 state left-to-right HMM with two v-states. The S and E nodes in these figures are the non-emitting start and exit nodes. For the other nodes the first number refers to the s state and the second number refers to the v state. Therefore, figure 3 shows an HMM which has transitions from each (i, k) node to each other (j, l) node for which the transition probability $a_{i,j}$ is not zero. Figure 4, on the other hand, shows an HMM topology where transitions between different v-states have been removed. The reason why one would want to consider such a model is that the v-states might be expected to vary much slower than the s-states. This is, for instance, the case for hidden states associated with speaking rates which are usually not measured for each s-state separately and are defined on more macroscopic levels such as phones, words or utterances. Each of these three models was tested in the recognition experiments described in section 4.

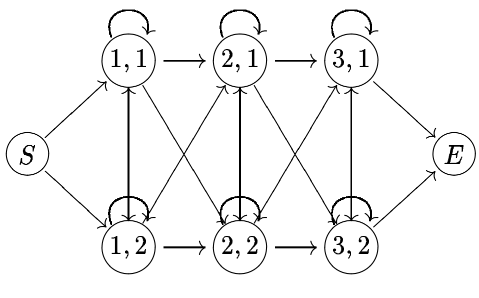

Fig. 3. Expanded 5 state left-to-right HMM with full topology

3. FEATURES AND INDICATOR VARIABLES

As mentioned in section 2 the feature d has to be a measure that distinguishes well between the hidden states of v. In order to obtain a model with a tractable number of parameters, however, the dimensionality of d has to be small. Since the feature d in this paper was directly derived from the feature vector o the aim was therefore to find a good dimensionality reduction of o that still contained most of the relevant information about the indicator variable v. As the main interest in the recognition experiments discussed in section 4

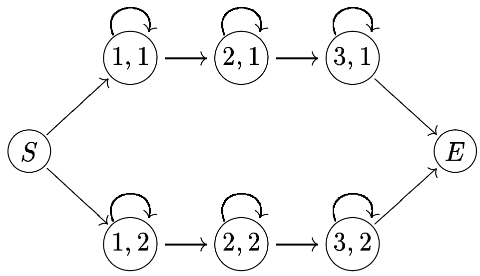

Fig. 4. Expanded 5 state left-to-right HMM with parallel topology

	full		static	
	slow	fast	slow	fast
slow	16915	4184	14962	6137
fast	5615	16388	8168	13835
	1st derivative		2nd derivative	
	slow	fast	slow	fast
slow	15588	5511	16362	4737
fast	8764	13239	5675	16328

Table 1. Classification results for phone “eh”.

was to develop a system that performed well on spontaneous speech, the quality of the feature d had to be evaluated on hidden states that were associated with different hidden dynamics. For this purpose each phone in the training data of the 1997 Broadcast News task was labelled as fast or slow depending on whether its duration was above or below a phone dependent threshold. The duration of the phones were determined by a forced alignment on the manual transcriptions of the training data. The statistics of feature d were determined on both the slow and fast instances of a phone and were subsequently used to classify the phones according to a MAP criterion. Table 1 shows the results of classification experiments that were carried out for phone “eh” on parts of the 39 dimensional feature vector. The feature vectors in these experiments consisted of 13 PLP features and first and second order derivatives. The columns in table 1 give the number of phones that were classified as fast or slow and the rows show which of the phones were truly fast or slow. As can be seen from table 1, using the full 39 dimensional feature vector gives a classifier with an error rate of 22.7 %, using only the static components results in an error rate of 33.2 %, and for the first and second order derivatives the error rates of the classifiers are 33.1 % and 24.2 %, respectively. This shows that the PLP features themselves are reasonable features for determining a speaking rate indicator variable and that the full feature vector can be reduced to the second order derivative with an increase in error rate by only 1.4 %. To decrease the dimensionality of d even further, the squared Euclidean norm of the second order derivatives was investigated. For this feature the corresponding classifier gave an error rate of 28.7 %. This was felt to be a reasonable performance and this feature was therefore used in the recognition experiments in section 4.

4. RECOGNITION EXPERIMENTS

The following recognition experiments were conducted on the 1997 Broadcast News task [5] by rescoring tri-gram lattices. This task consists of 72 hours of training data and 3 hours of test data. The feature vectors had 39 dimensions and consisted of 13 PLP coefficients and first and second order derivatives. The squared Euclidean length of the second order derivatives was used for feature d. Their output pdf's were modelled by Gamma densities as described in [3]. There are two different base line systems with 12 and 24 Gaussian mixtures, respectively. The 24 mixture system was created from the 12 mixture system by a conventional mixture splitting procedure that was applied iteratively to increase the number of mixtures by two at a time. The models introduced in this paper were initialised in two different ways. Firstly, each HMM in the 12 mixture model set was duplicated to give a slow and a fast instance of the model. These models were then trained on the training data which were relabelled with fast and slow tags for each triphone as described in section 3. The fast and slow instances of the HMM were then combined into a single HMM and the output pdf's of d and the posterior probabilities were added. Six reestimation iterations were performed on all the parameters of this model. For the second method of initialisation the output pdf's of d were added to the original HMM. The parameters of the resulting model were trained and subsequently the posterior probabilities were prescribed to intersect at the median of the trained pdf's of feature d. This method ensured that the training sets for the two states of the indicator variable had the same size. Both initialisation methods gave a model set with a number of parameters that is comparable to the 24 mixture baseline model. Table 2 shows the results of the recognition experiments on average and for the different F-conditions of the 1997 Broadcast News task [5]. As can be seen the models that were trained initially on the relabelled training data performed worse than the models that were initialised by splitting the indicator variable at the median of the pdf of feature d. This is due to the fact that by splitting phones into fast and slow instances by a duration criterion the size of the two classes can sometimes be rather different. As a result one of the posteriors that were trained for these HMM's was sometimes almost identically one while the other was almost identically zero. Consequently, some of the output pdf's for the PLP features were unreliable. The results in table 2 show that the full HMM topology as described in figure 3 and the standard HMM topology with the v states factored out gave the same performance. The parallel topology described in figure 4 had an error rate that was worse than the 12 mixture baseline. The reason for the bad performance of this model might be the result of the small number of v-states which prevents smooth transitions between them. Finally, table 2 shows that the first two models that were trained with the median split method give a small improvement over the 24 mixture baseline.

	avg.	F0	F1	F2	F3	F4	F5	FX
12mix base	21.8	12.4	20.5	31.5	31.5	24.3	27.5	43.9
24mix base	21.3	12.2	20.3	29.9	32.0	24.4	25.7	41.7
relabelled training data								
sum	21.4	12.5	20.1	30.2	31.2	24.7	25.1	42.2
full	21.4	12.5	20.1	30.6	31.0	24.6	24.8	42.4
par	22.7	13.1	21.7	32.4	32.7	25.2	27.4	45.1
median split								
sum	21.0	12.2	19.9	29.5	31.4	24.1	25.9	41.1
full	21.0	12.2	19.9	29.6	31.5	23.8	25.9	41.2
par	22.6	13.2	21.0	32.4	32.5	26.1	28.1	44.1

Table 2. Recognition experiments on the 1997 Broadcast News task

5. CONCLUSIONS AND FURTHER WORK

This paper has described a method of including information about hidden states into a standard HMM using continuous posteriors that are modelled by softmax functions. As an example of the use of this technique, speaking rate was introduced in a large vocabulary speech recognition system using a binary hidden variable. Different topologies of the model were evaluated on the 1997 Broadcast News task and some small improvements were obtained. Future work will include the use of indicator variables other than the one discussed here and application to the modelling of hidden states other than speaking rate.

6. REFERENCES

[1] M. Ostendorf et al., "Modelling systematic variations in pronunciation via a language-dependent hidden speaking mode," Tech. Rep., CLSP, 1996.

[2] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," *Proc.Eurospeech*, vol. 4, pp. 2079 – 2082, 1997.

[3] A. Tuerk and S. J. Young, "Modelling speaking rate using a between frame distance metric," *Proc. Eurospeech*, vol. 1, pp. 419–422, 1999.

[4] A. Tuerk and S. J. Young, "Polynomial softmax functions for pattern classification," Tech. Rep. CUED/F-INFENG/TR.402, Cambridge University Engineering Department, 2001.

[5] D. S. Pallett, J. G. Fiscus, J. S. Garofolo, A. Martin, and M. Przybocki, "1998 Broadcast News Benchmark Test Results: English and non-English Word Error Rate Performance Measures," *Proc. DARPA Broadcast News Workshop*, pp. 5–11, 1999.