MAXIMUM MUTUAL INFORMATION TRAINING OF HIDDEN MARKOV MODELS WITH VECTOR LINEAR PREDICTORS

K.K. Chin and P.C. Woodland

Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

HMM makes a piece-wise constant assumption about the temporal evolution of the speech signal. This is not true for speech signals which are known to be highly temporally correlated. Many approaches had been proposed to overcome the limitation of HMMs in modelling temporal context. One of these approaches uses a Vector Linear Predictor (VLP) to model the relationship between a nearby frame and the current frame. In this paper, Maximum Mutual Information Estimation (MMIE) of VLP-HMMs is explored. The MMIE training of VLP-HMMs are evaluated on WSJ data.

1. INTRODUCTION

In using HMMs for acoustic modelling, it is assumed that successive frames from a given output state are independent and identically distributed (IID), i.e. the state-conditional stationary assumption. It is known that this is not a valid assumption for speech signals which are by nature highly temporally correlated. Many approaches had been proposed to overcome this limitation of HMMs. They can be roughly classified into two broad types. Firstly, there are approaches that employ front end processing that try to capture the temporal information in speech data. The most successful example of these approaches is the use of dynamic parameters where the “static” features are augmented with the first and/or second differentials. The second class of approach involves extending the standard HMM to allow the explicit modelling of temporal information. Some examples include the Segment Model, Linear Dynamical System and Vector Linear Predictor (VLP) HMM [1]. Even though using dynamic parameters can give a significant gain in recognition accuracy, it is a direct violation of the state-conditional stationary assumption. It also increases the dimension of the feature vector hence causing difficulties when using full covariance matrices. Furthermore, an explicit model of speech dynamics will allow better understanding of the nature of speech data and may allow a more intuitive adaptation of the model parameters. VLP-HMMs have one significant advantage compared to other extension to the standard HMM.

While extending the model to capture the temporal dynamics, it maintains the efficient training and decoding algorithms of standard HMMs.

The concept of using nearby frames to predict the current frame via linear predictors was proposed in [2]. It was shown in [3] that VLP-HMMs show promising results but the performance of these VLP-HMMs are still poorer than standard HMMs with dynamic parameters. In [4] it was shown that applying VLP to a system with first differential parameters can yield good results.

Most studies of VLP-HMMs have concluded that these extended models are able to model the speech data more accurately. This implies that in the Maximum Likelihood Estimation (MLE) framework, the variance of the models is reduced. However this doesn’t necessarily give better discrimination since the model is not the true speech production model. Furthermore, the selection of the predictor offsets are arbitrary (this is true for all the published work in VLP-HMMS except for the extension in [5] where the dependence structure is optimised through an EM iterative algorithm). This paper intends to address both problems. Training the parameters of VLP-HMMs using the MMIE criterion will give better discrimination. In addition, it also provides a measure that can be use to select the optimum dependence structure.

2. MODEL FORMULATION

The model formulation is similar to that described in [4]. This will be briefly outlined here for completeness. All aspects of VLP-HMMs are the same as the standard HMMs except the definition of the output probability density. For VLP-HMMs, the output probability density for state $j$ of a particular single Gaussian model is defined as:

$$b_j(O_t) = \frac{1}{\sqrt{(2\pi)^d|C_j|}} \exp - \frac{1}{2} (e_j(t)e_j(t)')$$

where $C_j$ is the covariance matrix of the prediction error $e_j(t)$ for state $j$ and the prediction error of state $j$ at time $t$.
is defined as:
\[ e_j(t) = O_t - \left( \mu_j + \sum_p A_{jp} (O_{t+q_p} - \mu_{j,p}) \right) \]
where \( O_t \) is the observation vector at time \( t \), \( A_{jp} \) is the \( p \)th predictor for state \( j \), \( q_p \) is the "offset" associated with the \( p \)th predictor and \( \mu_{j,p} \) is the mean value for vectors at offset \( q_p \). Note that the predictors can have arbitrary offsets. We can use either a full matrix or a diagonal matrix as the predictor. If the predictor is diagonal, the correlations between the elements in the observation vector across the prediction time lag are ignored. Eqn. 1 can be revised to support systems with mixture distributions.

3. MLE TRAINING

The well established MLE HMM training algorithm was extended to support VLP-HMM[3][4]. This will be reviewed briefly to facilitate comparison with the MMIE training of VLP-HMMs. The parameter update equations can be obtained by differentiating the standard auxiliary function with respect to these unknown parameters and equating to zero.

For those parameters that are present in both standard HMMs and VLP-HMMs such as, \( \mu_j \), transition probabilities and mixture weights, the update equations are similar to those in standard HMMs. For the mean at offset \( q_p \), \( \mu_{j,p} \), the update equation is

\[ \bar{\mu}_p = \frac{\sum \psi(t)^j O_{t+q_p}}{\sum \psi(t)^j} \]

where \( \psi(t)^j \) is the posterior probability of being in state \( j \) at time \( t \) (the occupation probability).

The inter-frame covariance matrix between observations from offset \( q_x \) and \( q_y \) is defined as (the state index \( j \) is dropped for clarity):

\[ \gamma_{xy} = \frac{\sum \psi(t) (O_{t-q_x} - \bar{O}_x) (O_{t-q_y} - \bar{O}_y)}{\sum \psi(t)} \]

Note that \( \gamma_{00} \) is just the covariance matrix for the standard HMMs (assuming \( q_0 = 0 \) and \( \mu_{j,0} = \mu_j \).

Solving the following matrix equation will then yield the new estimates for a set of \( P \) predictors:

\[ Z' L' = R' \]

where

\[ Z = \begin{bmatrix} C_{11} & \cdots & C_{1P} \\ \vdots & \ddots & \vdots \\ C_{P1} & \cdots & C_{PP} \end{bmatrix} \quad L = \begin{bmatrix} \bar{A}_1 \cdots \bar{A}_P \end{bmatrix} \quad R = \begin{bmatrix} \bar{C}_{10} \cdots \bar{C}_{P0} \end{bmatrix} \]

For the correlation matrix of the prediction error, \( C_{ij} \), the update equation is (without the state index \( j \)):

\[ \bar{C} = \gamma_{00} - L \bar{R} - L' R + L' Z' L' \]

4. MMIE TRAINING

MMIE training for standard HMMs using the extened Baum-Welch algorithms was proposed in [6] to achieve better discrimination and has been improved and applied to LVCSR successfully [7]. The VLP-HMMs parameter update equation can be derived by following similar steps to those described in [8]. For a corpus with \( R \) sentences, the MMIE criterion to maximise is

\[ F(\theta) = \sum_{r=1}^{R} \log P(O_r | \theta) \]

where \( \theta \) represents the set of parameters for all the models, while \( O_r, M_r, P(w_r) \) and \( M'_{gen} \) are the sequence of observation vectors, the correct model sequence, the language model probability and models in the confusion lattice respectively for sentence \( r \). Using the same approach described in [8], an auxiliary function is defined by assuming that there exists an equivalent discrete distribution for each continuous distribution in the system.

The update equations for the models are again derived by differentiating the appropriate auxiliary function w.r.t each of the model parameters and equating to zero. For most parameters, the update equation is the same as those for the standard HMMs. For the extra LP parameters, the equations are revised accordingly. For example, \( \pi_p \) is given by (the sentence index \( r \) and state index \( j \) are dropped for clarity):

\[ \bar{\pi}_p = \frac{\sum \hat{\psi}(t) O_{t+q_p} + D \mu_p}{\sum \hat{\psi}(t) + D} \]

where \( \hat{\psi}(t)^r = \psi(t)^r - \psi(t)_{gen}^r \) and \( \psi(t)_{gen}^r \) is the occupation probability for frame \( t \) of sentence \( r \) from the confusion lattice. \( D \) is a constant selected to ensure convergence of the training iteration. Note that this is just an extension of the MMIE update equation of the mean for standard HMMs. For \( \bar{A}_1, \cdots, \bar{A}_P \) and \( \bar{C} \) the update equations are the same as Eqn. 2 and 3 respectively but with a revised estimation of the correlation matrix:

\[ \bar{\gamma}_{xy} = \frac{\sum \hat{\psi}(t) (O_{t-q_x} - \bar{O}_x) (O_{t-q_y} - \bar{O}_y) + D \gamma_{xy}}{\sum \hat{\psi}(t) + D} \]

Note that the MMIE update equation for the correlation matrix of the standard HMMs is consistent with the new MMIE definition of \( \gamma_{00} \).

Eqn. 4 can be rewritten to illustrate more explicitly the role and implication of the value \( D \) to MMIE:

\[ \bar{\pi}_p = \lambda \left( \frac{\sum \hat{\psi}(t) O_{t+q_p}}{\sum \hat{\psi}(t)} \right) + (1 - \lambda) \mu_p \]

where \( \lambda = \frac{\sum \hat{\psi}(t)}{\sum \hat{\psi}(t) + D} \). The choice of \( D \) will determine the step size in MMIE training. In standard HMMs \( D \) is selected to ensure convergence and that the diagonal elements
of the matrix $\Sigma_{ii}$ are positive. The criteria for selecting the value of $D$ described in [7] is applied to all the MMIE training in this paper.

For the VLP-HMMs, due to insufficient training data, the matrix $Z$ is often uninvertable. Especially when the number of predictors $P$ is more than 2 in a system with multiple mixture components. This will limit the number of predictors that can be employed practically. $D$ can be selected to ensure that the matrix $Z$ is positive definite. This ensures that the training of a system with larger number of predictors will not be hampered by limited data for a particular Gaussian.

5. EXPERIMENTS AND RESULTS

For evaluation, the VLP-HMMs is trained using the WSJ corpus SL-284 (66 hours). All VLP-HMMs in the experiments are initialise using tree-based state clustered standard HMMs trained using conventional MLE, by just adding one or more null predictors to the Gaussians in the standard HMMs. In all the experiments, we use only diagonal predictors. The observation vectors consist of 13 PLP coefficients derived from a mel-scale filter bank, including $\alpha_0$, with per-sentence cepstral mean normalisation. MMIE training lattices were generated by a similar 12 mixture components standard HMM system using a bigram language model, but unigram probabilities are used during MMIE training [7]. All results shown here are taken after either 5 MLE or 7 MMIE iterations. The acoustic scaling described in [7] is applied to all MMIE training.

Recognition results presented are from re-scoring of trigram lattices with 65k word vocabulary which were generated by a standard HMM system with 12 mixture components. The 1994 DARPA Hub1 development and evaluation test sets, denoted csrnab1, with per-sentence cepstral mean normalisation. MMIE training lattices were generated by a similar 12 mixture components standard HMM system using a bigram language model, but unigram probabilities are used during MMIE training [7]. All results shown here are taken after either 5 MLE or 7 MMIE iterations. The acoustic scaling described in [7] is applied to all MMIE training.

Fig. 1 shows the changes in MMI criteria for both standard HMMs and VLP-HMMs.

Table 1. %WER for HMMs and VLP-HMMs with single diagonal predictors that can be uninvertable. By selecting a $D$ value (and the corresponding MMIE step size $\lambda$) that will ensure that the matrix $Z$ is positive definite, i.e. $\lambda$ will reflect the amount of available data.

<table>
<thead>
<tr>
<th>Training Scheme</th>
<th>#mix comp.</th>
<th>csrnab1</th>
<th>MLE</th>
<th>HMM</th>
<th>VLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>1</td>
<td>19.06</td>
<td>16.80</td>
<td>22.25</td>
<td>20.01</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>17.95</td>
<td>15.96</td>
<td>20.00</td>
<td>18.66</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>17.20</td>
<td>15.09</td>
<td>20.02</td>
<td>17.80</td>
</tr>
<tr>
<td>MMIE</td>
<td>1</td>
<td>17.49</td>
<td>15.06</td>
<td>20.49</td>
<td>17.51</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>15.73</td>
<td>14.05</td>
<td>18.40</td>
<td>15.92</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>15.35</td>
<td>13.49</td>
<td>18.03</td>
<td>15.52</td>
</tr>
</tbody>
</table>

Fig. 1. Changes to MMI Criteria during training for both standard HMMs and VLP-HMMs

table1

3The results for MLE trained 6 mixture components system in Table 2 are obtained by controlling the training step size using the constant $D$ as in the MMIE case.

---

\[ E = 1 \text{ was used} \]

\[ ^2 \text{The MLE scheme could be adapted to use the same "trick".} \]
trained three mixture component system with the same predictors does better than a similar system with only one predictor (compare with Table 1). WER reduces, from 14.05 to 13.55 and 15.92 to 15.67 for test set csrnab1dt_h1 and csrnab1et_h1 respectively. It might be possible to use more then three predictors with MMIE, but it is well known that MMIE give the largest gain with small number of parameters. It will be interesting to find the maximum number of predictors this current framework can reliably train.

<table>
<thead>
<tr>
<th>Training Scheme</th>
<th>#mix</th>
<th>csrnab1dt_h1</th>
<th>csrnab1et_h1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>3</td>
<td>16.13</td>
<td>18.71</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>15.36</td>
<td>17.86</td>
</tr>
<tr>
<td>MMI</td>
<td>3</td>
<td>13.55</td>
<td>15.67</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>13.14</td>
<td>15.07</td>
</tr>
</tbody>
</table>

Table 2. %WER for VLP-HMMs with 3 diagonal predictors at offset -7, -5 and -3. Feature vectors contain only static parameters.

Applying VLP to a system with dynamic parameters also gives small improvements in most cases (see Table 3). However VLP-HMMs using static features (even with 3 predictors) did not perform as well as standard HMMs with dynamic parameters. The dynamic parameters are computed across 5 frames of speech data, so the dependence structure is not as localised and rigid compared to the current framework where the temporal structure is fixed (by using fix offset for the predictors). Due to variable speaking rate and other factors, this is a poor approximation for speech data. To overcome this limitation, one can extend the model to dynamically adjust the predictor offsets [5] or employ a front-end such as variable frame rate technique that tries to compensate for temporal variation in speech.

<table>
<thead>
<tr>
<th>Training Scheme</th>
<th>#mix comp.</th>
<th>csrnab1</th>
<th>csrnab1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HMM</td>
<td>VLP</td>
</tr>
<tr>
<td>MLE</td>
<td>1</td>
<td>13.71</td>
<td>13.64</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10.72</td>
<td>10.45</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>9.66</td>
<td>9.61</td>
</tr>
<tr>
<td>MMI</td>
<td>1</td>
<td>11.23</td>
<td>11.67</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.52</td>
<td>9.72</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>9.24</td>
<td>9.08</td>
</tr>
</tbody>
</table>

Table 3. %WER for HMMs and VLP-HMMs with single diagonal predictors at offset -5. Feature vectors contain static parameters, first and second differential.

6. CONCLUSION

This paper has extended the MMIE framework to include VLP-HMMs and experiments on WSJ data show that the training of such models under this new framework converges and gives improved MMI criterion compared to standard HMMs. Experimental results demonstrate that MMIE does gives significant improvements in term of recognition accuracy to both standard HMMs and VLP-HMMs. Some of the temporal dynamics in speech is captured with just one diagonal predictor. It is possible to use more predictors to model more of the temporal dynamics. MMIE is able to scale the contribution from each of the predictors to give the optimum dependence structure from the selected predictors.

7. REFERENCES