This work was funded by the European Commission under the Language project Le-5 Coretex. Extensive use was made of equipment donated by IBM under a SUR award.
need to determine the appropriate learning rate, or use second-order statistics. The weak-sense auxiliary function may be selected so that it has a simple closed-form for the parameter estimation. Normally there will need to be smoothed in some form to try to ensure that the value of the original function increases.

There are thus two functional forms to select when using weak-sense auxiliary functions. First the auxiliary function of the function to be optimised is required. Second an appropriate form of smoothing function is required; it must be some function with its maximum at \( \lambda \).

2.2. Weak-sense auxiliary functions for MMIE

This section describes how a weak-sense auxiliary function may be used to optimise the MMI criterion for training HMMs and how, given the appropriate smoothing function, it yields the standard extended Baum-Welch (EBW) update rules. Considering only a single training utterance, \( O = \{ o_1, \ldots, o_T \} \) and using a fixed language model\(^1\), the MMI criterion may be expressed as

\[
F(\lambda) = \log p(O|\mathcal{M}^{num}) - \log p(O|\mathcal{M}^{den})
\]

where \( \mathcal{M}^{num} \) and \( \mathcal{M}^{den} \) are HMMs corresponding to the correct transcription (numerator term) and all possible transcriptions (denominator term) respectively. It is not possible to define a strong-sense auxiliary function for this expression, since the second term is negative. Therefore the inequality of equation (1) will no longer hold. However, it is possible to linearly combine individual weak-sense auxiliary functions to form an overall weak-sense auxiliary function, even when there is negative feedback.

As a strong-sense auxiliary function is by definition also a weak-sense auxiliary function, it is natural to use the standard strong-sense auxiliary function associated with ML estimation as an appropriate form for the weak-sense auxiliary function. Thus a possible weak-sense auxiliary function for the numerator term (considering a single Gaussian per state with a single dimension) is

\[
G^{num}(\lambda, \hat{\lambda}) = \sum_{t=1}^{T} \sum_{j=1}^{J} \gamma_j^{num}(t) \log p(x_t|\mu_j)
\]

where \( \lambda_j = \{ \mu_j, \sigma_j^2 \} \),

\[
Q(\gamma_j, \theta_j(O), \theta_j(O^2), \lambda_j) = -\frac{1}{2} \left( \gamma_j \log(2\pi\sigma_j^2) + \theta_j(O^2) - 2\theta_j(O)\mu_j + \gamma_j\mu_j^2 \right)
\]

\( s_j \) indicates state \( j \) of the system, \( \gamma_j(t) \) is the posterior probability of being in state \( s_j \) at time \( t \) given \( \hat{\lambda} \), and the sufficient statistics to evaluate the function for the numerator are given by \( \theta_j^{num}(O) = \sum_{t=1}^{T} \gamma_j^{num}(t) \), \( \theta_j^{num}(O^2) = \sum_{t=1}^{T} \gamma_j^{num}(t) \mu_j^2 \) and \( \gamma_j^{num} = \sum_{t=1}^{T} \gamma_j(t) \) the occupancy of the state. Similarly the auxiliary function for the denominator term alone can be defined. These two may then be combined to yield a candidate weak-sense auxiliary function for the MMI criterion.

As previously mentioned, in order to improve stability of the training process, a smoothing function, \( G^{num}(\lambda, \hat{\lambda}) \), can be added. This may be any function with a zero differential w.r.t. \( \lambda \) around the current estimate \( \lambda = \hat{\lambda} \). As such combining this with any weak-sense auxiliary will still be a valid weak-sense auxiliary function. Hence, for MMIE the complete weak sense auxiliary function will have the form

\[
G^{mmi}(\lambda, \hat{\lambda}) = G^{num}(\lambda, \hat{\lambda}) - G^{den}(\lambda, \hat{\lambda}) + G^{sm}(\lambda, \hat{\lambda}).
\]

One possible form for \( G^{sm}(\lambda, \hat{\lambda}) \) is to use \( D_j \) “effective” observations which yield the current state parameters, \( \hat{\lambda} \), as the ML estimate, thus automatically satisfying the requirements for the smoothing function. This may be written in the same form as equation (4)

\[
G^{sm}(\lambda, \hat{\lambda}) = \sum_{j=1}^{J} Q(D_j, \beta_j, \sigma_j^2, \lambda_j),
\]

where \( D_j \) are positive smoothing constants for each state \( j \). The above analysis can be simply extended for multiple Gaussian components per state.

Optimising the weak-sense auxiliary function simply requires combining the sufficient statistics for each of the individual auxiliary functions. The global maximum of \( G^{mmi}(\lambda, \hat{\lambda}) \) for the mean and variance of component \( m \) of state \( j \) are given by

\[
\mu_{jm} = \frac{\theta_{jm}^{num}(O) - \theta_{jm}^{den}(O)}{\gamma_{jm}^{num} - \gamma_{jm}^{den}} + D_{jm} \beta_{jm}
\]

\[
\sigma_{jm}^2 = \frac{\theta_{jm}^{num}(O^2) - \theta_{jm}^{den}(O^2)}{\gamma_{jm}^{num} - \gamma_{jm}^{den}} + D_{jm} (\beta_{jm}^2 + \mu_{jm}^2) - \mu_{jm}^2
\]

where \( D_{jm} \) is set on a per-Gaussian level as described in [8] and determines the convergence-rate and stability of the update rule. These are the standard update rules obtained from the extended Baum-Welch (EBW) algorithm [3], though derived using weak-sense auxiliary functions. Similarly, update equations may also be derived for the component priors and transition probabilities.

3. Incorporating Prior Information

In this section the incorporation of a prior into the weak-sense auxiliary function framework is discussed. The derivation of I-smoothing and discriminative MAP based on MMI (MMI-MAP) and MPE (MPE-MAP) is described.

By definition, any function is both a weak and strong-sense auxiliary function. Therefore the inequality of equation (1) will no longer hold. However, it is possible to linearly combine individual weak-sense auxiliary functions to form an overall weak-sense auxiliary function, even when there is negative feedback.

\[
F(\lambda) = \log p(O|\mathcal{M}^{num}) - \log p(O|\mathcal{M}^{den}) + \log p(\lambda)
\]

The exact form of the log-prior distribution affects the nature of the MAP update. One of the major issues, and choices, in MAP estimation is how to obtain this prior distribution.
3.1. I-smoothing

I-smoothing for discriminative training [5] may be regarded as the use of a prior over the parameters of each Gaussian, with the prior being based on the ML statistics. The log prior likelihood is defined as

$$\log p(\lambda jm) = Q \left( \tau , \tau' \frac{\theta \nu m(O)}{\gamma jm} , \tau \frac{\theta \nu m(O^2)}{\gamma jm} , \lambda jm \right)$$  

This log-prior is the log-likelihood of $\tau$ points of data with mean and variance equal to the numerator (correct model) mean and variance. The MMIE update formula for the mean is then

$$\mu jm = \frac{\{\theta \nu m(O) - \theta \nu m(O')\} + D jm \hat{\mu jm} + \tau' \mu jm}{\gamma jm + \tau'}$$

where $\mu jm = \frac{\theta \nu m(O)}{\gamma jm}$.

I-smoothing can also be directly implemented by altering the numerator statistics [6]. A similar form of prior with ML-MAP training yields I-smoothing for MPE.

3.2. MMI-MAP

In the context of adapting a HMM set, the use of ML statistics accumulated from the adaptation data as the center of the prior may not be robust since there may not be enough data to estimate the ML Gaussian parameters. In this case it is preferable to estimate the center of the prior in a fashion similar to standard ML-MAP. The technique denoted MMI-MAP is the use of a prior over the parameters of each Gaussian, with the prior distribution used. The value of $\tau$ is typically in the same range used for I-smoothing (e.g. 100) and $\tau'$ is normally in the range used for ML-MAP (e.g. 10).

3.3. MPE-MAP

In MPE [5], as for MMI, the auxiliary function to be optimised is represented in the form given in equation (11); but the statistics $\gamma jm^m$, $\gamma jm^m$ etc. are accumulated from the training data in a different way as described in [5]. The combination of the auxiliary function with the prior distribution used in I-smoothing follows the same pattern, with one difference: in MPE the numerator ("num") statistics are defined differently and do not correspond to the correct transcription. Therefore, where the correct-model statistics are needed (e.g., in equation 15) a separate set of statistics with the superscript "mle" are used in place of the "num" statistics; the "mle" statistics are the same statistics used in normal ML training.

4. Experiments

The performance of discriminative MAP was evaluated on two tasks. The first is to port a well-trained Switchboard system to the Voicemail task using limited training data. These results have previously been published in [6] and are summarised in this paper to allow an overview of the scheme. The second application examined is to build gender-specific HMMs using Broadcast News data by discriminative adaptation from gender independent models.

4.1. Porting Switchboard to Voicemail

Initial Switchboard HMMs were trained used 265 hours of data. Cross-word state-clustered triphones were generated. The system had 6684 distinct states and 16 Gaussians per state. For further details of the acoustic training see [1]. Two “initial” models were trained: an ML-trained system and one discriminatively trained using MMIIE. The Voicemail database consists of voicemail messages left by IBM employees. This data was partitions into a 94 minute test set and 28.1 hours of training data. The training data was further partitioned into nested subsets of approximately 1h, 4h, 15h and 20h. See [1] for more details of the database set-up.

All test set WERs reported here are from testing with a Switchboard language model (LMs). The baseline acoustic-model porting used a single iteration of ML-MAP. It was found that additional iterations yielded no further gains in performance. MMI-MAP task adaptation used four iterations of model parameter updates. The various forms of $\tau$ were approximately tuned, but there was little sensitivity to the precise values used.

Figure 1: WERs for MMI-MAP and ML-MAP from MMI and ML baselines against amount of Voicemail adaptation data.
4.2. Gender Dependent Broadcast News Models

The Broadcast News acoustic model training data consists of two sub-sets referred to as BNtrain97 and BNtrain98, reflecting the years of their release. The combined set gives a total of 142 hours of training data [9]. A cross-word state-clustered triphone system was built using MLE with 6,976 speech states and 16 Gaussian components per state using MF-PLP parameterised speech with static, first and second order differences. MMIE and MPE trained models were also built. In addition a gender dependent system was generated using the training data speaker gender labels and only updating the Gaussian mixture weights and mean values. All experiments reported below used single pass decoding without adaptation. The decoder used a 65k word trigram language model which was taken from the 1998 Cambridge University broadcast news evaluation system [9]. The pronunciation dictionary was based on the 1993 LIMSI WSJ lexicon with many additions.

<table>
<thead>
<tr>
<th>System</th>
<th>WER (%)</th>
<th>Std</th>
<th>HLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE-GI</td>
<td>19.6</td>
<td>17.9</td>
<td></td>
</tr>
<tr>
<td>MLE-GD</td>
<td>18.8</td>
<td>17.1</td>
<td></td>
</tr>
<tr>
<td>MMI-GI</td>
<td>17.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPE-GI</td>
<td>16.2</td>
<td>15.0</td>
<td></td>
</tr>
<tr>
<td>MPE-MAP</td>
<td>15.7</td>
<td>14.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: WER on BNeval98 using gender independent (GI) and gender dependent (GD) models with ML, MMI and MPE training and also MPE-MAP adaptation to GD models.

The error rates of the gender independent (GI) and gender dependent (GD) systems on the 1998 NIST Broadcast News evaluation data (BNeval98) is shown in table 1. Initially the system was tested using the standard front-end. The ML-GD system reduced the error rate by about 4% relative, 0.8% absolute, over the ML-GI system. Table 1 also shows the performance of MMI training and MPE training. Both discriminative training schemes show significant gains over ML training. MPE training gave a lower WER than MMI training yielding a 17% relative reduction in error rate over the MLE-GI system and 14% over the MLE-GD performance. As GD systems significantly reduced the error rate for the MLE system, it would be useful to generate gender dependent systems for the discriminative models. As the MPE-GI system outperformed the MMI-GI system, the MPE system was used as the original models for adaptation and MPE-MAP was applied. Table 1 lists the error rate for the MPE-GI system adapted with MPE-MAP to form GD models. These gender-dependent discriminative models gave an additional 3% relative reduction in WER over the MPE-GI system.

Table 1 also shows the performance of using the various training schemes with an HLDA frontend. Here third order differences were added to the feature vector and then projected down to 39 dimensions. The use of HLDA significantly reduced the WER for all systems. Using MPE-MAP yielded a 0.5% absolute reduction in error rate over the gender-independent system. An alternative approach to generating the GD model would rely on the I-smoothing to perform the regularisation and to simply do MPE training on the male and female training data separately. This gave an error rate of 14.8%, 0.3% higher than using MPE-MAP.

5. Conclusions

This paper has described techniques for incorporating prior information into discriminative training schemes. Versions based on both MPE, MPE-MAP, and MMI, MMI-MAP, have been described. It was shown that by using the appropriate form of the prior, these discriminative MAP schemes may be implemented by count smoothing. Depending on the exact form of the prior distribution used, this yields either versions of MAP estimation or I-smoothing. The discriminative adaptation schemes were investigated for both task porting, in this case from Switchboard to Voicemail, and for generating gender dependent models on the Broadcast News task. In both cases the methods were effective and allowed the performance advantage of discriminatively trained HMMs to be retained.

6. References