Confidence-informed unsupervised minimum Bayes risk acoustic model adaptation

Matthew Gibson & Thomas Hain
Technical report CS-11-01
January 2011

Department of Computer Science
Regent Court, 211 Portobello
Sheffield, S1 4DP
U. K.
E-mail: Matt.Gibson@sheffield.ac.uk

Abstract

In a supervised speaker adaptation scenario, discriminative techniques have yielded performance gains over standard approaches to adaptation e.g. maximum likelihood linear regression (MLLR). However the discriminative approaches have failed to yield such gains when the task is unsupervised. This paper addresses this issue by applying a novel confidence-informed minimum Bayes risk (MBR) criterion to the task of unsupervised MBR linear regression (MBRLR) speaker adaptation. Experimental evaluations on a large vocabulary recognition task demonstrate that confidence-informed unsupervised MBRLR adaptation delivers significant performance improvements over standard unsupervised MBRLR when using posterior-based confidence measures. Several aspects of confidence-informed unsupervised MBRLR adaptation are analysed and evaluated, including the use of sub-word confidence measures and MBR criteria, and use of the I-smoothing technique. It is demonstrated that, given improved confidence measures, confidence-informed MBRLR yields performance superior to state-of-the-art confidence-informed MLLR adaptation.
1 Introduction

Speaker dependent (SD) automatic speech recognition (ASR) systems are designed to recognise the speech of a single speaker. Speaker independent (SI) systems are designed to recognise the speech of any speaker. Given an identical amount of training data, an SD system typically displays significantly better performance than an SI system [Woodland, 2001]. This is due to a greater mismatch between the test data and the speech models for the case of the SI system.

SI recognition systems use techniques known collectively as speaker adaptation to reduce this mismatch as the system encounters new speakers. A relatively small (in comparison to the volume of data used to train the SI system) quantity of speaker data, known as the adaptation data, is used to adapt the SI system. When the correct transcription of the adaptation data is available the task is referred to as supervised adaptation. An unsupervised adaptation scenario exists if the correct transcription of the adaptation data is not available.

One successful approach to speaker adaptation is to adjust the acoustic model to better match the adaptation data, known as acoustic model adaptation [Gauvain and Lee, 1994, Leggetter and Woodland, 1995, Kuhn et al., 2000]. The parameters governing these model adaptation methods have typically been determined by maximisation of model likelihood (ML) or a-posteriori probability (MAP) with respect to the adaptation data. However, recent research has focussed on the use of discriminative criteria to determine these parameters [Woodland and Povey, 2002, Povey and Woodland, 2002, Juang and Katagiri, 1992]. Discriminative acoustic model adaptation methods alter the acoustic model such that a discriminative criterion is optimised. Discriminative versions of MAP [Povey et al., 2003a,b], linear regression [Gunawardana and Byrne, 2001, Wu and Huo, 2002, Wang and Woodland, 2004], speaker adaptive training [Tsakalidis et al., 2002, Wang and Woodland, 2002] and cluster adaptive training [Yu and Gales, 2006] have been previously proposed. These discriminative approaches have yielded recognition performance improvements over MAP [Povey et al., 2003b] and ML [Gunawardana and Byrne, 2001] parameter estimation in the case of supervised speaker adaptation. However the discriminative approaches have failed to yield such impressive gains when the task is unsupervised [Gunawardana and Byrne, 2001, Wang and Woodland, 2004].

This paper focuses upon unsupervised linear regression speaker adaptation using the discriminative Bayes risk criterion. The resulting framework is termed minimum Bayes risk linear regression (MBRLR). Previous work [Wang and Woodland, 2004] is extended by incorporating confidence information into the Bayes risk criterion. The impact of this confidence information upon the performance of the resulting acoustic models is then quantified.

The paper is structured as follows. Sections 2 and 3 respectively introduce the theory and implementation of MBRLR. Section 4 explains how confidence information may be integrated into the MBR criterion to yield a confidence-informed MBR criterion suitable for unsupervised speaker adaptation. The confidence measures used in this work are introduced in Section 5. Section 6 describes the large vocabulary recognition system used to evaluate the technique. Evaluations of several aspects of confidence-informed MBRLR are presented in Section 8. Analysis of the different confidence measures used is essential to the interpretation of the results of these evaluations and is therefore presented beforehand in Section 7. A concluding discussion and propositions for future research are found in Section 9.

2 Discriminative linear regression

The linear regression speaker adaptation paradigm [Leggetter and Woodland, 1995] uses adaptation data from a speaker to estimate one or more affine transforms of the SI acoustic model parameters. In the case of continuous density hidden Markov models (HMMs) with Gaussian mixture state output distributions of dimension $n$, the mean of a SI Gaussian mixture component $\hat{\mu}_s$ is transformed according to Equation 1.

$$\hat{\mu}_s = A\mu_s + b = W\xi_s$$  (1)
The symbol \( \hat{\mu}_s \) represents the adapted mean, \( W = [ \beta A ] \), \( \xi_s \) is the extended mean vector \([ 1 \quad \mu_s^T ]^T\), \( A \) is an \( n \times n \) matrix and \( \beta \) is an \( n \)-dimensional vector called the bias.

The affine transform described above is typically shared across many mixture components. This parameter-tying is implemented via a regression class tree [Leggetter and Woodland, 1995, Young et al., 2003]. This structure groups the mixture components of the ASR system into hierarchically-defined clusters called regression classes. In addition to defining how parameters are tied, the regression class tree is used to ensure that the generated transforms correspond to the volume and nature of the available adaptation data.

Discriminative linear regression adaptation deploys the same adaptation framework as maximum likelihood linear regression (MLLR, [Leggetter and Woodland, 1995]). The difference between MLLR and discriminative linear regression is that the affine transforms are chosen to optimise a discriminative criterion instead of the model likelihood function. Discriminative linear regression corresponds to the conditional maximum likelihood criterion [Gunawardana and Byrne, 2001], the minimum classification error criterion [Wu and Huo, 2002] and the minimum phone error criterion [Wang and Woodland, 2004] have been previously proposed. The minimum phone error (MPE) criterion may be viewed as a specific instance of the minimum Bayes risk (MBR) criterion, which is now introduced.

### 2.1 Minimum Bayes risk linear regression

The MBR criterion \( R_{\text{MBR}}(\theta) \), also referred to as the overall risk criterion [Na et al., 1995, Kaiser et al., 2002], is defined by Equation 2.

\[
R_{\text{MBR}}(\theta) = \frac{1}{R} \sum_{r=1}^{R} \sum_{w_1^f} p(w_1^N(o_r, \theta) | L(w_1^N, \hat{\mu})^{M(r)}) \tag{2}
\]

The set \( W \) comprises all possible transcriptions of the acoustic data \( o_r \), \( w_1^{M(r)} \) is the correct transcription of \( o_r \) and \( L(w_1^N, \hat{\mu})^{M(r)} \) is the Levenshtein distance between the correct transcription and hypothesis \( w_1^N \). The set \( W \) is called the hypothesis space, \( \theta \) represents the model parameters and \( R \) is the number of training set examples.

Choosing the parameters of the affine transform \( W \) to minimise the MBR criterion results in a technique which will be referred to as minimum Bayes risk linear regression (MBRLR). Re-estimation equations for the transform parameters are derived in [Wang and Woodland, 2004]. Assuming that HMM state output distributions are Gaussian mixtures with diagonal covariances, the \( i \)-th row of the mean transform \( W \) of mixture component \( s \) for a particular speaker is given by Equation 3.

\[
W^{(i)} = G_{\text{MBR}}^{(i)} k_{\text{MBR}}^{(i)} \tag{3}
\]

The matrix \( G_{\text{MBR}}^{(i)} \) and the vector \( k_{\text{MBR}}^{(i)} \) are defined by Equations 4 and 5 respectively.

\[
G_{\text{MBR}}^{(i)} = \sum_{m \in R(s)} \frac{1}{\sigma_m^2} (\gamma_m + D_m) \xi_m \xi_m^T \tag{4}
\]

\[
k_{\text{MBR}}^{(i)} = \sum_{m \in R(s)} \frac{1}{\sigma_m^2} (\theta_m^{(i)} + D_m \theta_m^{(i)}) \xi_m \tag{5}
\]

In Equations 4 and 5, \( R(s) \) is the regression class containing component \( s \), \( D_m \) is a learning rate discussed in Section 3.3, \( \mu_m^{(i)} \) is the \( i \)-th dimension of the mean of component \( m \), \( \sigma_m^{(i)} \) is the variance of the \( i \)-th dimension of component \( m \), \( \gamma_m \) is described by Equation 6 and \( \theta_m^{(i)} \) is given by Equation 7.

\[
\gamma_m = \sum_{r=1}^{R} \sum_{w_1^f} K(r, w_1^N, \theta) \sum_{t=1}^{T(r)} \gamma_m(t|w_1^N, o_r, \theta) \tag{6}
\]
\[ \theta^{(i)}_m = \sum_{r=1}^{R} \sum_{w_1 \in \mathcal{W}} K(r, w_1^N, \theta) \sum_{t=1}^{T(r)} \gamma_m(t|w_1^N, o_r, \theta) \varphi^{(i)}(r) \] (7)

In the above equations, the \(i\)-th dimension of the \(t\)-th frame of the \(r\)-th training utterance is denoted by \(\varphi_i^{(i)}(r)\) and \(\gamma_m(t|w_1^N, o_r, \theta)\) is the posterior probability that state \(m\) is the \(t\)-th element of the hidden state sequence. The quantity \(K(r, w_1^N, \theta)\) is described by Equation 8.

\[ K(r, w_1^N, \theta) = p(w_1^N|o_r, \theta)[L_{av}(r) - L(\hat{w}_1^M(r), w_1^N)] \] (8)

In Equation 8, \(L_{av}(r)\) is the average error of all hypotheses, given by Equation 9.

\[ L_{av}(r) = \sum_{w_1^1 \in \mathcal{W}} p(w_1^N|o_r, \theta)L(\hat{w}_1^M(r), w_1^N) \] (9)

The statistics necessary for MBRLR transform estimation, namely the quantities \(\theta^{M_{MBR}(i)}_m\) and \(\gamma^{M_{MBR}}_m\), are calculated via a lattice-based implementation discussed in Section 3. The quantity \(D_m\) is a learning rate discussed in Section 3.3.

3 Lattice-based MBR

A lattice-based implementation of MBR estimation is introduced in [Povey and Woodland, 2002, Povey, 2003]. Lattices which include temporal alignment information, i.e. label start and end times, are used, and the lattice encodes the alignments of the acoustic data of highest posterior [Young et al., 2003]. A lattice is generated via a recognition pass of a speech utterance. Additionally, the alignments of the correct label sequence of highest posterior, generated using a constrained recognition pass, are added to the lattice produced by recognition. The resulting lattice represents a set of alternative alignments of the acoustic data associated with an utterance.

The idea behind lattice-based MBR is not only to use the lattice as an approximation to the hypothesis space, but also to use the alignment information which is present in the lattice to save computation. Note that Equation 6 can be re-phrased as a sum over all possible alignments of the acoustic data as in Equation 10. A similar re-formulation for Equation 7 is given by Equation 11.

\[ \gamma_m = \sum_{r=1}^{R} \sum_{z \in \mathcal{Z}_r} K(r, z, \theta) \sum_{t=1}^{T(r)} \gamma_m(t|z, o_r, \theta) \] (10)

\[ \theta^{(i)}_m = \sum_{r=1}^{R} \sum_{z \in \mathcal{Z}_r} K(r, z, \theta) \sum_{t=1}^{T(r)} \gamma_m(t|z, o_r, \theta) \varphi^{(i)}(r) \] (11)

The set \(\mathcal{Z}_r\) comprises all possible alignments of the utterance \(o_r\) and \(K(r, z, \theta)\) is given by Equation 12.

\[ K(r, z, \theta) = P(z|o_r, \theta)[L_{av}(r) - L(\hat{w}_1^M(r), w_z)] \] (12)

In the above equation, \(w_z\) is the hypothesis associated with alignment \(z\). Notice also that the average error \(L_{av}(r)\) may also be expressed as a sum over alignments as in Equation 13.

\[ L_{av}(r) = \sum_{z \in \mathcal{Z}_r} P(z|o_r, \theta)L(\hat{w}_1^M(r), w_z). \] (13)

Substituting the set of all possible alignments \(\mathcal{Z}_r\) with the set of alignments specified by the lattice, Equations 10 and 11 yield practical approximations to the statistics required for MBRLR mean transform estimation. Further, since an alignment corresponds to a sequence of lattice arcs,
Figure 1: Symmetrically normalised frame error approximation to the Levenshtein distance.

Equation 10 can be expressed in terms of lattice arcs as in Equation 14. A similar rearrangement
of Equation 11 in terms of lattice arcs may be performed.

$$\gamma_m = \sum_{r=1}^{R} \sum_{a \in A_r} K(r, a, \theta) \sum_{t=a_{\text{start}}}^{a_{\text{end}}} \gamma_m(t|a, o_r, \theta) \quad (14)$$

The symbol \(a\) represents a lattice arc which in turn represents a label, its start time \(a_{\text{start}}\) and end
time \(a_{\text{end}}\). The set \(A_r\) contains all arcs in the lattice and \(K(r, a, \theta)\) is expressed by Equation 15.

$$K(r, a, \theta) = p(a|o_r, \theta) \left[ L_{av}(r) - L(\hat{w}_M^{(r)}, a) \right] \quad (15)$$

In the above equation, \(p(a|o_r, \theta)\) is the posterior probability that arc \(a\) is included in any path, i.e.
any contiguous sequence of arcs from the lattice start node to the lattice end node. The quantity
\(L(\hat{w}_M^{(r)}, a)\) is the posterior-weighted sum of the Levenshtein error of all the lattice paths which
include arc \(a\), while \(L_{av}(r)\) is the posterior-weighted sum of the Levenshtein error of all the lattice
paths.

Calculation of the Levenshtein distance between a path in the lattice and the reference label se-
quence \(\hat{w}_M^{(r)}\) is non-trivial. This involves a dynamic programming alignment of the label sequence
associated with the path and the reference word sequence. Since a lattice encodes many such paths,
calculation of the quantities \(L(\hat{w}_M^{(r)}, a)\) and \(L_{av}(r)\) becomes computationally expensive. Section
3.1 explains how such costly computation is avoided by approximating the Levenshtein distance
between a lattice path and the reference label sequence. This approximation assigns an error \(l(a)\)
to each lattice arc \(a\) such that the overall error of each path is the sum of the errors associated
with its composite arcs.

### 3.1 Error approximation

The error approximation technique used in this work uses alignment information from the reference
and hypothesis transcriptions as shown in Figure 1. A path in a lattice defines what will be referred
to as a hypothesis alignment. The most likely alignment of the reference transcription is called
the reference alignment. Using the hypothesis and reference alignments, the Levenshtein error of
the hypothesis label sequence is approximated as follows.

Firstly, the temporal region of each label of the hypothesis alignment is divided into segments
corresponding to regions of overlap with different labels of the reference alignment. For example,
in Figure 1 the label C is split into two segments, the first being associated with the label A of the
reference alignment and the second associated with label B. The segment boundaries are illustrated
by vertical dashed lines in Figure 1. Each segment has a corresponding frame error; the number
of frames within the segment at which the hypothesis label differs from the label specified by the
reference alignment. Then for each segment a normalisation factor is defined. This is the length
of the shorter of the overlapping hypothesis and reference labels. For example, the normalisation
factor for the segment overlapping reference label A (length 100 frames) and hypothesis label C

<table>
<thead>
<tr>
<th>Reference</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Length (frames)</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>Frame error</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Normalisation factor</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Symmetrically normalised frame error</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>
The frame error for each segment is then divided by the normalisation factor to yield the symmetrically normalised frame error (SNFE) for each segment. The overall SNFE for the hypothesis is the sum of the SNFE over all segments. The SNFE approximation has been shown to yield improved estimation of errors when compared to alternative error approximation methods [Gibson, 2008].

The SNFE approximation to the Levenshtein error is expressed by Equation 16. The approximate error between hypothesis label sequence $w^n_1$ and reference label sequence $\hat{w}_1^M$ is denoted by $L_{\text{approx}}(w^n_1, \hat{w}_1^M)$. $A$ represents the set of aligned reference labels corresponding to sequence $\hat{w}_1^M$, $\hat{A}$ is the set of aligned hypothesis labels corresponding to sequence $w^n_1$, and $l(a, \hat{a})$ is the SNFE between the aligned labels $a$ and $\hat{a}$.

$$L_{\text{approx}}(w^n_1, \hat{w}_1^M) = \sum_{\hat{a} \in \hat{A}} \sum_{a \in A} l(a, \hat{a})$$ (16)

The SNFE between two aligned labels $l(a, \hat{a})$ is defined by Equation 17.

$$l(a, \hat{a}) = \frac{e(a, \hat{a})}{n(a, \hat{a})}$$ (17)

The quantity $e(a, \hat{a})$ is the number of frames at which the aligned labels $a$ and $\hat{a}$ differ (defined as zero if no temporal overlap exists between the aligned labels). The divisor $n(a, \hat{a})$ is the normalisation term, defined as the length of the shorter of $a$ and $\hat{a}$.

### 3.2 Forward-backward algorithms

This section explains how the quantities $\gamma_m(t|a, o_r, \theta)$ and $K(r, a, \theta)$ of Equation 14 are calculated for each arc $a$. The mixture component occupancies $\gamma_m(t|a, o_r, \theta)$ of Equation 14 are calculated via a forward-backward pass over the models defined by each lattice arc $a$ using the segment of acoustic data aligned with arc $a$. This is a standard forward-backward procedure, as implemented in the Baum-Welch algorithm. In order to calculate the quantities $K(r, a|\theta)$ for each arc $a$, a single lattice-level forward-backward pass suffices when an error $l(a)$ is assigned to each lattice arc, as described in Section 3.1. This forward-backward algorithm is introduced in Povey [2003].

### 3.3 Setting the learning rate

The choice of an appropriate learning rate $D_m$ in Equations 4 and 5 has mainly been studied in the context of the conditional maximum likelihood (CML) discriminative criterion [Valtchev et al., 1997, Schluter et al., 1997, Povey, 2003]. An occupancy-dependent scheme for determining the learning rate in the case of CML training is adopted for MBR training in Povey [2003]. The procedure is as follows. For each mixture component $m$:

1. Calculate $D_{m\text{min}}$, the minimum $D$ required to ensure all variance updates are positive for component $m$.

2. Define $\gamma_{m\text{den}}$ as in Equation 18

$$\gamma_{m\text{den}} = \sum_{r=1}^{R} \sum_{a \in A_{\text{den}r}} K(r, a, \theta) \sum_{t=a_{\text{start}}}^{a_{\text{end}}} \gamma_s(t|a, o_r, \theta)$$ (18)

where $A_{\text{den}r}$ denotes the subset of lattice arcs in the set of all lattice arcs $A_r$ for which $K(r, a, \theta)$ is negative. The symbols $a_{\text{start}}$ and $a_{\text{end}}$ denote the start and end time of arc $a$, respectively.

3. Set the learning rate $D_m$ to $\max\{2D_{m\text{min}}, E\gamma_{m\text{den}}\}$ where $E$ is a configurable constant which typically assumes a value in the interval $[1, 2]$. 

Since setting the learning rate in this way also leads to successful MB RLR transform estimation [Wang and Woodland, 2004], it is the approach used in the experimental work described in this paper. Since this experimental work also uses I-smoothing and complexity control, these methods and now explained in the context of MBRLR adaptation.

3.4 I-smoothing for MBR linear regression

To address the issue of overfitting the adaptation data, a version of the I-smoothing regularisation technique, previously introduced in the context of MBR acoustic model estimation [Povey and Woodland, 2002, Povey, 2003], has also been applied to MBR-based acoustic model adaptation in [Wang and Woodland, 2008]. A prior distribution upon the transform $W$ is proposed, as described by Equation 19.

$$
\log p(W) = k + \frac{T}{2} \sum_{m \in R(s)} \sum_{r=1}^{R} T(r) \sum_{t=1}^{T(r)} \gamma_m(t) |w_1^M(r), o_r, \theta)(o_r(r) - W \xi_m)^T C_m^{-1} (o_r(r) - W \xi_m)
$$

The quantity $k$ is a normalisation term used to ensure the prior probability distribution sums to one and $C_m$ is the covariance of component $m$. Note that this prior is formulated for each regression class $R(s)$ and is globally maximal at the maximum likelihood estimate of the transform $W$. The prior probability described by Equation 19 is subtracted from the MBR criterion to yield an I-smoothed MBR criterion. The subsequently estimated transforms optimise the I-smoothed MBR criterion, a combination of likelihood and MBR criteria. This technique is used in the experiments involving I-smoothed MBRLR adaptation in Section 8.

3.5 Complexity control and MBRLR adaptation

Complexity control using regression class trees has been introduced in the case of MLLR transform estimation [Leggetter and Woodland, 1995]. In the experimental work of Section 8, the MBRLR transform estimation procedure adopts the same complexity control mechanism. To achieve this, the overall occupancies of each mixture component are accumulated in an initial pass, prior to accumulation of the statistics required for MBRLR transform estimation. These statistics are subsequently input to the MBRLR adaptation procedure to control the amount and type of MBRLR transforms estimated. Note that this ensures a fair comparison between MLLR and MBRLR since the number of transforms are identical in each case.

4 Confidence-informed MBRLR

The idea behind confidence-informed MBRLR is to either disregard or de-emphasise errors assigned with respect to low-confidence labels of the reference transcription. A general form of confidence-informed MBR criterion is defined by Equation 20.

$$
R_{MBR}^C(\theta) = \frac{1}{R} \sum_{r=1}^{R} \sum_{w_1^N \in W} p(w_1^N | o_r, \theta) L_C^C(w_1^N, \hat{w}_1^M(r))
$$

Here $L_C^C(w_1^N, \hat{w}_1^M(r))$ is a confidence-sensitive error between the label sequences $w_1^N$ and $\hat{w}_1^M(r)$. This error function is a modification to the Levenshtein error approximation, designed to incorporate knowledge of confidence associated with the reference transcription labels. Comparing Equation 20 with the standard MBR criterion given by Equation 2, it is clear that the only difference is the use of this confidence-sensitive error function $L_C^C(w_1^N, \hat{w}_1^M(r))$. Since confidence-informed MBR amounts to a refinement of the error function, confidence-informed MBRLR transforms are estimated using the same theory as standard MBRLR, presented in Section 2.1. The only
difference is that the confidence-sensitive error replaces the standard error approximation in the MBRLR transform estimation formulae.

Two confidence-informed MBRLR techniques are proposed. The first technique is referred to as confidence-thresholded MBRLR while the second technique is referred to as confidence-weighted MBRLR. An experimental evaluation of both instances of confidence-informed unsupervised MBRLR adaptation is presented in Section 8. Note that confidence-thresholded MBRLR was first introduced in [Gibson and Hain, 2007], where the results of some initial experiments are reported.

4.1 Confidence-thresholded MBRLR

Confidence-thresholded MBRLR deploys a modification to the Levenshtein error approximation as illustrated in the example of Figure 2. Section A of Figure 2 shows an alignment of an estimated reference transcription and a hypothesis alignment. The symmetrically normalised frame error (Section 3.1) of the hypothesis is 2, the sum of the SNFE for each aligned hypothesis label. Section B of Figure 2 shows the confidence associated with each label of the estimated reference transcription.

Figure 2: (A) Standard and (B) confidence-thresholded error approximations. A confidence threshold of 0.5 is used in this example.

The confidence-thresholded frame error is a modified version of the frame error which assigns errors only with respect to high-confidence (i.e. above a specified confidence threshold) labels of the reference alignment. More precisely, for each segment of the hypothesis alignment, the confidence-thresholded frame error is zero if the segment overlaps with a low-confidence label (i.e. below a specified threshold) of the reference alignment and equal to the standard frame error otherwise. The confidence-thresholded error for each hypothesis segment is then the normalised confidence-thresholded frame error, where the normalisation factor is the length of the shorter of the overlapping labels, as described in Section 3.1. The overall confidence-thresholded error for the hypothesis is then the sum of the confidence-thresholded error over each segment.

In the example of Figure 2, the confidence threshold is defined to be 0.5. The reference label B of the estimated reference is therefore deemed unreliable while the other labels are deemed reliable. Therefore the confidence-thresholded frame error is zero for segments overlapping with the second reference label. The overall confidence-thresholded error is 1 and the error incurred with respect to the second reference label is disregarded. Modifying the error in this way thus reduces the impact of errors associated with low-confidence labels of the estimated reference transcription.
Recall that the symmetrically normalised frame error approximation to the Levenshtein error is expressed by Equation 16. Reusing the notation of Equation 16 ($\hat{A}_r$ represents the set of aligned labels in the alignment of reference sequence $\hat{w}_1^M(r)$ and $l(a, \hat{a})$ is the symmetrically normalised frame error between the aligned labels $a$ and $\hat{a}$) the confidence-thresholded error $L_C(w_1^N, \hat{w}_1^M(r))$ is expressed by Equation 21, where $A_C$ is the set of aligned hypothesis labels which overlap only with reference labels of confidence greater than some threshold $C$.

$$L_C(w_1^N, \hat{w}_1^M(r)) = \sum_{\hat{a} \in \hat{A}_r} \sum_{a \in A_C} l(a, \hat{a})$$ (21)

An evaluation of confidence-thresholded MBRLR is found in Section 8.4.

### 4.2 Confidence-weighted MBRLR

Confidence-weighted MBRLR is an alternative to the use of a threshold to decide which labels of the reference alignment to disregard. Instead, the error function uses confidence to weight the error derived from each aligned reference label. This is expressed by Equation 22, where the notation of Equation 21 has been reused and $C(\hat{a})$ is the confidence assigned to label $\hat{a}$.

$$L_C(w_1^N, \hat{w}_1^M(r)) = \sum_{\hat{a} \in \hat{A}_r} \sum_{a \in A_C} l(a, \hat{a}) C(\hat{a})$$ (22)

The need to specify an additional threshold parameter is avoided when using confidence-weighted MBRLR. It may be also argued that use of the confidence-weighted technique does not wastefully disregard low-confidence labels of the reference transcription, opting instead to de-emphasise their impact in proportion to their associated confidence. The performance yielded by confidence-weighted MBRLR-adapted models is compared to that yielded by confidence-thresholded MBRLR-adapted models in Section 8.5.

### 4.3 Sub-word confidence-informed MBRLR

Word and sub-word MBR criteria have been compared with regard to acoustic model estimation [Povey, 2003, Gibson, 2008]. These word and sub-word MBR criteria may also be used for the purposes of acoustic model adaptation. Further, confidence-informed MBR criteria can be formulated at the word and sub-word levels. In the experimental work of this paper, the confidence measure and hypothesis space correspond when using confidence-informed MBRLR, so, for example, a word-level confidence measure is used in conjunction with the word-level MBR criterion. Note that it is possible to use, for example, a state-level confidence measure in conjunction with a phoneme-level MBR criterion. However additional motivation is required to argue in favour of such a configuration.

The following argument may be presented in favour of use of sub-word confidence measures for confidence-informed MBRLR. Use of a word-level confidence measure lacks the ability to identify high-confidence sub-word units within low-confidence words. It is therefore conceivable that word-level confidence-informed MBRLR is a suboptimal configuration. Constraint of the adaptation procedure to errors corresponding to high-confidence sub-word labels may more efficiently exploit the available adaptation data.

The confidence-informed MBRLR techniques described above naturally accommodate use of confidence measures at either the word or sub-word levels. In the case of confidence-thresholded MBRLR, a comparison between word and sub-word criteria is found in Section 8.4. A similar comparison is found in Section 8.5 in the case of confidence-weighted MBRLR.

### 4.4 Confidence-informed MBRLR and generalisation

Consider the issue of generalisation with regard to the unsupervised adaptation scenario. Since the adaptation data and the test data coincide, one might reasonably argue that the question of
generalisation does not apply to this situation. However, when using confidence-based refinements to unsupervised adaptation, the issue of generalisation becomes relevant for the following reason. When using confidence-informed adaptation, the adaptation data is effectively subdivided into high-confidence and low-confidence subsets. The confidence-informed adaptation procedure is designed to respect, or emphasise, information learnt from the high-confidence adaptation data while the low-confidence adaptation data is disregarded or de-emphasised. The evaluation tests how well the adapted acoustic models perform upon both the low-confidence and high-confidence subsets of the adaptation data. This evaluation is therefore a measure of how well the models adapted using the high-confidence data generalise to the low-confidence data.

Experimental results have shown that acoustic models estimated using sub-word MBR criteria display significantly superior generalisation to that yielded by word-level MBR [Povey, 2003, Gibson, 2008]. Given these results, and the relevance of generalisation to the unsupervised confidence-informed MBRLR scenario, it is hypothesised that sub-word level confidence-informed MBRLR will display superior generalisation to word-level confidence-informed MBRLR. This hypothesis is experimentally tested in Section 8.3.

4.5 Confidence-informed I-smoothing for MBRLR

It has been shown that I-smoothing can improve the generalisation of MBR-estimated acoustic models [Povey, 2003, Gibson, 2008]. Since the issue of generalisation is relevant to confidence-informed MBRLR adaptation, as explained in Section 4.4, it is hypothesised that unsupervised confidence-informed MBRLR adaptation will also benefit from the use of I-smoothing.

The question of how to formulate the I-smoothing prior distribution in the case of confidence-informed MBRLR arises. It is inconsistent to use confidence information to influence the MBR criterion but to ignore this confidence information when specifying a prior for the transform $W$. In this work, a confidence-adjusted prior $p(W | C)$, of the form described by Equation 23, is used. This prior is similar to the standard MBRLR prior (Equation 19) with the exception that confidence-adjusted occupancies replace the standard occupancies.

$$\log p(W | C) = k + \frac{\tau}{2} \sum_{m \in R(s)} \sum_{r=1}^{R} \sum_{t=1}^{T(r)} \gamma_m^C(t | \hat{w}_1^M(r), o_r) (o_t^r - W \xi_m)$$

These confidence-adjusted occupancies are given by Equation 24, where $C(t)$ is the confidence associated with the label overlapping frame $t$ in the most likely alignment of the reference transcription $\hat{w}_1^M(r)$.

$$\gamma_m^C(t | \hat{w}_1^M(r), o_r, \theta) = \begin{cases} 0 & \text{if } C(t) < C \\ \gamma_m(t | \hat{w}_1^M(r), o_r, \theta) & \text{otherwise} \end{cases}$$

The I-smoothed confidence-thresholded MBR criterion is formulated by subtracting the prior defined above (Equation 23) from the confidence-thresholded MBR criterion (Equation 20), as specified by Equation 25. The scalar $\tau$ determines the influence of the prior term.

$$R^C_{MBR}(\theta) = \frac{1}{R} \sum_{r=1}^{R} \sum_{w_1^N \in \mathcal{W}} p(w_1^N | o_r, \theta) L^C(w_1^N, \hat{w}_1^M(r)) - \log p(W | C)$$

Optimisation of this criterion requires that the reference labels which are deemed as low-confidence (with respect to the threshold $C$) are ignored both for the purpose of error computation (as explained in Section 4) and for the purpose of accumulation of the statistics used to smooth the transform estimation. The latter accumulation process is implemented by using confidence-adjusted occupancies of Equation 24. With infinitely large $\tau$, the I-smoothed confidence-thresholded MBR criterion becomes the confidence-informed MLLR criterion examined in [Pitz et al., 2000]. The experiments described in Section 8.6 evaluate the performance of I-smoothed confidence-informed MBRLR.
4.6 Confidence-informed complexity control for MBRLR

The complexity control mechanism for MBRLR transform generation is explained in Section 3.5. Essentially, the same complexity control formalism used for MLLR is used for the purposes of MBRLR transform generation. In the case of confidence-informed MBRLR, errors corresponding to much of the data may be de-emphasised or disregarded. To introduce sensitivity to the volume and type of de-emphasised data, the confidence-adjusted occupancies are used to inform the complexity control framework.

5 Confidence measures

Much research has been done to identify suitable measures of word-level confidence in large vocabulary ASR. In this paper, confidence measures related to word posterior probability are used since there is empirical evidence of their superiority (with regard to classification of words as correct or incorrect) over several other confidence measures [Wessel et al., 2001].

Lattices are used as follows to derive posterior-based confidence measures. Consider the lattice shown in Figure 3. Each lattice arc is marked with its associated label and its posterior probability. For example B/0.3 means that the arc has label B and posterior probability of 0.3. The lattice path of maximum posterior probability is shown as the MAP alignment. For each label of the MAP alignment, an associated confidence measure is sought. For each label $X$ in the MAP alignment, with start time $X_s$ and end time $X_e$, the posterior probability $p(X, X_s, X_e | o^T, \theta)$, given the acoustic data $o^T$ and model parameters $\theta$, may be calculated by summing the posterior probabilities of all the lattice arcs with matching label and identical start and end times. This quantity is referred to as the alignment posterior. In Figure 3, the alignment posterior of the label E is the sum of the arc posteriors for the arcs labelled E since these arcs have matching start and end times.

The alignment posterior yields a somewhat poor measure of confidence [Wessel et al., 2001], again with respect to the classification of words as correct or incorrect. This is because, typically,
a lattice comprises several alignments containing the same label with slightly different start and end times. In such a situation the alignment posterior underestimates the confidence in the label since posterior probabilities are distributed across each different alignment. Notice that this is the case for label A in Figure 3.

Alternative confidence measures involve the use of the frame posterior, defined as follows. Each frame defines a set of lattice arcs overlapping the frame and an overlapping label in the MAP alignment, X say. The frame posterior is defined as the sum of the posteriors of each of the arcs overlapping frame \( t \) which have label X. In Figure 3, the label A has frame posterior of 1.0 between the frames of 0 and 50 due to the contributions of each of three lattice arcs overlapping this time period.

The confidence associated with a label of the MAP alignment may be derived from the frame posteriors in several ways. The most reliable confidence measure reported in [Wessel et al., 2001] is defined by choosing the maximal value of the frame posterior between the start and end frames of the label. This measure is referred to as the maximal frame posterior and displayed in Figure 3. To illustrate, note that label A in the MAP alignment has frame posterior of 1.0 in the time period between 0 and 50 frames and frame posterior of 0.7 in the time period between 50 and 60 frames. So maximisation of the frame posterior in the region between the start and end frames of the label A gives this label a maximal frame posterior of 1.0. The maximal frame posterior does not necessarily obey the sum-to-one constraints of a probability distribution. However it yields a more successful confidence measure than the alignment posterior because alignments which differ slightly from the MAP alignment also contribute to the confidence measure.

There are several other techniques which may be used to derive a confidence measure from frame posteriors. In [Wessel et al., 2001], a confidence measure for the label X is defined as the value of the frame posterior of label X at the mid-point of the label between the start and end times. This yields a confidence measure with similar classification performance (with regard to the classification of words as correct or incorrect) as the maximal frame posterior. The geometric mean of the frame posterior over each of the frames corresponding to the label has also been effectively used as a confidence measure [Evermann and Woodland, 2000]. In the experimental work of this paper, the maximal frame posterior is used as a confidence measure since it yields consistently superior classification performance to the alternative measures studied in [Wessel et al., 2001].

The maximal frame posterior has mainly been used for the computation of word-level confidence measures from lattices. Note however that by using a lattice and MAP alignment marked with sub-word alignment information, the technique naturally yields a sub-word confidence measure.

6 Evaluation system

The evaluation system is based upon the individual headset microphone (IHM) 2005 AMI meeting speech transcription system [Hain et al., 2005a,b]. This section gives further information on this system including details of acoustic features, acoustic and language models, system operation and datasets used in training and evaluation.

6.1 Acoustic features

The IHM system uses 39-dimensional features to represent the speech signal. In the first pass this feature vector comprises 13 Mel-frequency-based perceptual linear prediction coefficients (MF-PLP, [Woodland et al., 1997]) as well as the first and second time derivatives of these features.

In subsequent passes the feature vector is extracted via smoothed heteroscedastic linear discriminant analysis (SHLDA, [Burget, 2004]) from a 52-dimensional vector comprising 13 MF-PLP coefficients and the first, second and third time derivatives of these features. Additionally, in the second pass, these features are normalised using speaker-specific vocal tract length normalisation [Lee and Rose, 1996] as well as speaker-specific cepstral mean (CMN) and variance (CVN) normalisation [Atal, 1974].
6.2 Acoustic models

The acoustic models are triphone HMMs with three emitting states and left-to-right topology. The model states are clustered using a phonetic decision tree [Young et al., 1994] and trained using 104 hours of speech (see Section 6.4) and the maximum likelihood criterion. Approximately 4000 tied states are used and state output distributions are modelled with 16-component Gaussian mixture models.

6.3 Dictionary and language models

The recognition dictionary contains the 50000 most frequently used words as specified by a procedure outlined in [Hain et al., 2005b]. The pronunciations are based on the UNISYN pronunciation lexicon [Fitt, 2000].

Language models derived from several text corpora are estimated using ML-based techniques. A trigram language model suitable for meeting speech is then interpolated from these language models. More details of this procedure and the text corpora are found in [Hain et al., 2005a,b].

6.4 Training and evaluation datasets

The training and evaluation datasets are recordings of spontaneous speech in meetings. Each meeting participant wore a head-mounted microphone during recording. While the language of all meetings is English, not all speakers are native English speakers.

The training dataset used to estimate acoustic models comprised 104 hours of transcribed speech in meetings from a selection of corpora including the ICSI meeting corpus [Janin et al., 2003], the NIST meeting room pilot corpus [Garofolo et al., 2004], the ISL meeting corpus [Burger et al., 2002], the NIST RT04s development and evaluation sets (rt04sdev and rt04seval) and the AMI meeting corpus [Carletta et al., 2005].

The National Institute of Standards and Technology (NIST) conference meeting evaluation datasets are used as test data. These datasets are labelled rt06seval and rt07seval. The number of hours of speech, speakers and words (including hesitations, partial words and fillers) associated with each dataset is shown in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>rt06seval</th>
<th>rt07seval</th>
</tr>
</thead>
<tbody>
<tr>
<td># hours</td>
<td>2.4</td>
<td>3.1</td>
</tr>
<tr>
<td># speakers</td>
<td>43</td>
<td>31</td>
</tr>
<tr>
<td># words</td>
<td>33321</td>
<td>37314</td>
</tr>
</tbody>
</table>

Table 1: NIST conference meeting speech evaluation datasets.

6.5 System operation

As illustrated in Figure 4, the evaluation system uses three passes for recognition. All passes use the same trigram language model, details of which are found in [Hain et al., 2005a], and a language model scaling factor of 14. ML-estimated acoustic models are used in the first recognition pass, the output of which is used for unsupervised estimation of speaker-specific VTLN, CMN and CVN normalisation parameters used in the second pass. The second recognition pass uses ML-estimated acoustic models which incorporate a smoothed HLDA transform of the acoustic features.

The experiments described in this paper evaluate the third pass of the recognition procedure, illustrated in Figure 5. The acoustic models used in the second pass are adapted using MBRLR and the second pass transcription is used as the reference transcription. The adaptation process alters only the means of the Gaussian mixtures of the acoustic model. Two regression classes

---

1 More details of these datasets are found at http://www.nist.gov/speech/tests/rt/.
are used for adaptation, one corresponding to speech models and one for non-speech models. An occupancy threshold of 1000 and full transform matrices are used, preliminary experiments having indicated the suitability of such a configuration. Phoneme, model and state-marked lattices are generated prior to MBRLR adaptation using a bigram language model (derived from the trigram used in recognition) and the unadapted second pass acoustic models. The lattices are subsequently pruned to a maximum density of 500 arcs per second to reduce the computational cost of the MBR criterion optimisation process, retaining only the lattice paths of highest posterior probability.

When using confidence-informed MBRLR, these pruned lattices and the second pass transcription are used as input to a confidence estimation procedure to calculate the maximal frame posterior confidence measures described in Section 5. These measures are subsequently used as input to confidence-informed adaptation procedures.

The adapted acoustic models are evaluated via a recognition pass identical in configuration to the second pass recognition process. The \texttt{rt06seval} and \texttt{rt07seval} test datasets (see Section 6.4) are used in the evaluations reported in this paper.

Before reporting the results of the experimental evaluation of confidence-informed MBRLR in Section 8, an evaluation and analysis of the confidence measures used in these experiments is presented. This analysis is essential to the interpretation of the results of the evaluation of the confidence-informed adaptation techniques.

7 Evaluation: confidence measures

The \texttt{rt06seval} and \texttt{rt07seval} datasets are processed as shown in Figure 5. The initial two recognition passes produce a transcription of the data. This transcription is then aligned at the word, phoneme, model and state levels to give the MAP alignment at each level. A lattice generation process, as illustrated in Figure 5, generates phoneme, model and state-marked lattices. The maximal frame posterior confidence measures at the word, phoneme, model and state level are then calculated from these lattices and the MAP alignment at the appropriate level (as detailed in Section 5). Note that acoustic probability scaling is used when calculating lattice arc posterior probabilities to achieve a reasonable distribution of arc posterior probabilities [Wessel et al., 2001]. The acoustic scale factor is $\frac{1}{14}$, the inverse of the language model scale factor used in the second
Figure 5: Confidence-informed MBRLR adaptation process.

7.1 Performance evaluation

The performance of the threshold-driven classifier defined above is measured by comparing the decision made by a classifier at each frame (i.e. to classify the frame as correctly or incorrectly labelled) with the correct decision. The correct decision is defined via the word, phoneme, model or state-level alignment of the correct word sequence. This alignment is referred to as the correct or reference alignment. The correct decision is to classify a frame as correct if the label of the MAP alignment agrees with the label at this frame in the corresponding correct alignment, and to classify the label as incorrect otherwise. So, e.g. the labels of the word-level MAP alignment and the word-level correct alignment define the correct decision at the word-level, while the labels of the state-level MAP alignment and the state-level correct alignment define the correct decision at the state-level. The performance of the classifier associated with a particular level is then measured in terms of how many frames are correctly classified, with respect to the correct decision at that level.

The performance of the classifiers induced by the different maximal frame posterior confidence measures is illustrated by the detection error tradeoff (DET, [Martin et al., 1997]) curves of Figure 6. A false alarm is defined as a frame which is misclassified as correct while a miss is defined as a frame which is misclassified as incorrect. Each datapoint corresponds to the performance

recognition pass. This factor is used to scale the acoustic likelihoods associated with each lattice arc.

Categorisation of a particular frame of an alignment as correctly or incorrectly labelled defines a binary classification task. Given a confidence threshold and a frame, a confidence measure defines a binary classifier which either classifies this frame as correctly labelled (if the confidence of the label spanning this frame exceeds the threshold) or incorrectly labelled (otherwise). This section evaluates the performance and analyses the behaviour of these simple threshold-driven classifiers with respect to four confidence measures of interest; the maximal frame posterior at the word, phoneme, model and state levels. The model and state levels correspond to the labels assigned to state-clustered clustered triphone models and states respectively.
of a classifier at a particular confidence threshold. The confidence thresholds range from 0.99, corresponding to high miss rates, to 0.1, corresponding to high false alarm rates. The horizontal axis measures the number of frames for which a false alarm occurs, normalised by the total number of incorrect frames (frames at which the label in the MAP alignment differs from the reference alignment). The vertical axis measures the number of frames for which a miss occurs, normalised by the total number of correct frames (frames at which the label in the MAP alignment agrees with the reference alignment label). Figure 7 displays these miss and false alarm rates as a function of the confidence threshold. The rt07seval dataset is used in this evaluation.

Figure 6: Performance of classifiers corresponding to word, phoneme, model and state-level maximal frame posterior confidence measures (rt07seval dataset).

7.1.1 Discussion

Only small classification performance differences are yielded by the classifiers corresponding to the four confidence measures: the maximal frame posterior at the word, phoneme, model and state-level. Each classifier yields an equal error rate (the point where the miss and false alarm rates are equal) in the range of 24.5% to 25.8%. Analysis is now presented to lend insight into the behaviour of these classifiers.

7.2 Analysis

Figure 8 plots the fraction of frames which are retained (i.e. deemed as correctly labelled) as a function of the confidence threshold, using the rt06seval and rt07seval datasets. Each curve corresponds to the threshold-driven classifier derived from the maximal frame posterior confidence measure indicated by the legend. Figure 9 gives an example of the maximal frame posterior confidence measures associated with a small segment of speech in the rt07seval dataset.
Figure 7: Miss and false alarm (FA) rates of classifiers corresponding to word, phoneme, model and state-level maximal frame posterior confidence measures (rt07seval dataset).

Figure 8: Data retained as a function of confidence threshold using the maximal frame posterior at the word, phoneme, model and state levels (rt06seval and rt07seval datasets).

It is clear from Figure 8 that use of the classifier induced by the phoneme-level confidence measure retains more adaptation data than the classifiers induced by the other confidence measures. This is true at all non-zero confidence thresholds.

Table 2 is a contingency table recording the behaviour of the classifiers corresponding to the word and phoneme-level maximal frame posterior at the confidence threshold of 0.7. A reasonably high level of agreement is observed. The classifiers agree on whether to retain or dismiss (i.e.
Figure 9: Maximal frame posterior confidence measures at the word, phoneme, model and state levels for a sample of speech in the rt07seval dataset.

decide as incorrectly labelled) a frame of data for 85.07% of the frames at this threshold (the sum of the diagonal entries of Table 2). The main disagreement (13.85% of all frames) is due to frames which are retained by the phoneme-level classifier but dismissed by the word-level classifier at a threshold of 0.7. This is because the phoneme-level confidence of a phoneme is often greater than the word-level confidence of its containing word, an effect is observed in Figure 9. A smaller percentage (1.08%) of frames are retained by the word-level classifier but dismissed by the phoneme-level classifier. On inspection, these frames are dismissed by the phoneme-level classifier as a consequence of differences between the within-word phoneme-level alignments of the MAP alignment and the phoneme-marked lattice.

Table 2: Contingency table corresponding to classifiers induced by the word and phoneme-level maximal frame posterior confidence measures at a threshold of 0.7 (rt06seval and rt07seval datasets).

<table>
<thead>
<tr>
<th>Frame</th>
<th>206</th>
<th>211</th>
<th>213</th>
<th>217</th>
<th>220</th>
<th>222</th>
<th>224</th>
<th>225</th>
<th>227</th>
<th>228</th>
<th>230</th>
<th>231</th>
<th>232</th>
<th>234</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retained at phoneme-level (%)</td>
<td>72.6</td>
<td>13.85</td>
<td>1.08</td>
<td>12.81</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8 shows that the classifiers corresponding to the word, model and state-level maximal frame posterior retain a similar volume of data at all confidence thresholds. Despite similar volumes of data being retained in the case of the model and word-level confidence measures, the frames of retained and dismissed data are somewhat complementary, as shown in Table 3.

The classifier corresponding to the word-level confidence measure retains 6.75% of the frames whose model-level confidence is below 0.7. This effect is due to the exclusion of data corresponding to models which display a lower model-level confidence than the word-level confidence of their containing word. This effect occurs because a word appears in several different contexts within a
lattice. These different contexts correspond with different models due to the presence of different context-sensitive triphone models at the start and end of the word. So while a word may have the dominant portion of posterior probability mass at a frame near the start or end of the word, the corresponding model at this frame may be less dominant. The result is that start or end models occasionally have lower model-level confidence than the word-level confidence of the containing word. This effect is observed in Figure 9, the first model of the word ‘NOT’ having a model-level confidence of 0.889, while the containing word has word-level confidence of 0.998.

Notice also that 6.74% of frames in the datasets considered have model-level confidence above 0.7, but word-level confidence below this threshold. This is due to the model-level confidence of a model in the MAP alignment occasionally being greater than the word-level confidence of its containing word. This effect is again observed in Figure 9, where the confidence of the model m-iy+n (0.921) exceeds that of the containing word MEAN (0.888).

A comparatively high level of agreement (95.71% of all frames) is found between the classifiers derived from the state and model-level confidence measures at the threshold of 0.7, as shown in Table 4. The classifier corresponding to the state-level confidence measure retains slightly more data (2.98% at this threshold) than the model-level classifier. This is due to the inclusion of data corresponding to states which display a higher state-level confidence than the model-level confidence of their containing model. Again this phenomenon is observed in Figure 9. Due to alignment differences between the state-level MAP and lattice alignments there are a small number (1.32% at this threshold) of frames which are retained by the model-level classifier but dismissed by the state-level classifier.

<table>
<thead>
<tr>
<th>frames (%)</th>
<th>retained at word-level</th>
<th>dismissed at word-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>retained at model-level</td>
<td>66.59</td>
<td>6.74</td>
</tr>
<tr>
<td>dismissed at model-level</td>
<td>6.75</td>
<td>19.91</td>
</tr>
</tbody>
</table>

Table 3: Contingency table corresponding to classifiers induced by the word and model-level maximal frame posterior confidence measures at a threshold of 0.7 (rt06seval and rt07seval datasets).

While the analysis of this section has compared the classifiers induced by the different confidence measures at a threshold of 0.7, this analysis characterises the different behaviour of these classifiers in general, i.e. at all non-zero confidence thresholds.

8 Evaluation: confidence-informed MBRLR

This section presents a series of experiments designed to evaluate aspects of confidence-informed unsupervised MBRLR adaptation. Standard MBRLR performance is measured in Section 8.1 to provide a performance baseline for confidence-informed MBRLR. Section 8.2 reports the per-
formance of confidence-informed MLLR and is provided as a further benchmark for comparison with confidence-informed MBRLR. The optimal performance of the confidence-informed MBRLR technique is measured in Section 8.3 by using ideal confidence measures. Section 8.4 evaluates the performance of word and sub-word confidence-thresholded MBRLR using imperfect confidence measures. These results are compared with the performance of confidence-weighted MBRLR in Section 8.5. Lastly, the impact of I-smoothing upon the performance of confidence-thresholded MBRLR is established in Section 8.6.

The multi-pass recognition system described in Section 6 is used, where the third-pass adaptation step is standard MBRLR/MLLR or confidence-informed MBRLR/MLLR. The MBRLR mean transforms are calculated as described in Sections 2.1 and 3. Acoustic likelihood scaling is used within the MBRLR lattice forward-backward process. The acoustic scale factor is $1/14$, the inverse of the language model scale factor used in recognition. The learning rate constant $E$ is set to 2 in all experiments. No I-smoothing is used unless specified otherwise. Preliminary experiments indicated that, with this configuration, twenty to thirty adaptation iterations are sufficient for convergence of the test set WER.

### 8.1 Experiment 1: Standard MBRLR

Before evaluating the confidence-informed MBRLR technique, the performance of standard unsupervised MBRLR adaptation is established. Table 5 compares the performance of unsupervised MLLR (four iterations) and MBRLR (twenty iterations) using the $rt06seval$ and $rt07seval$ test datasets. Note that three different MBRLR-adapted systems are evaluated, corresponding to the word, phoneme and state-level MBR criteria. The model-level MBR criterion is omitted due to its close similarity with the state-level MBR criterion [Gibson, 2008]. It is clear from Table 5 that the best performance is provided by MLLR adaptation, which yields an average WER of 33.9%. This is a significant improvement over both the average unadapted system performance of 35.8% and the performance of all MBRLR-adapted systems. A smaller, but significant, performance improvement (35.6% average WER for all criteria) over the unadapted performance is provided by the MBRLR-adapted systems. No significant difference is found between the different MBRLR-adapted systems. Note that throughout this paper a significant improvement is defined as significant at the 95% confidence level using the matched pairs sentence segment word error test (MPSSWE, [Gillick and Cox, 1989, Pallett et al., 1990]).

The superior performance of MLLR over MBRLR in the unsupervised scenario reveals an important issue, namely that the discriminative MBR adaptation technique is more sensitive to the quality of the unsupervised transcription than ML-based adaptation. This issue is directly addressed by this work by using confidence measures to alleviate the negative impact of errors in the estimated transcription.

### 8.2 Experiment 2: confidence-informed MLLR

This section presents the results of an experimental evaluation of unsupervised confidence-informed MLLR adaptation. Section 8.2.1 evaluates the performance of word and sub-word confidence-

<table>
<thead>
<tr>
<th>System</th>
<th>WER (%)</th>
<th>rt06seval</th>
<th>rt07seval</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadapted</td>
<td>33.3</td>
<td>38.0</td>
<td>35.8</td>
<td></td>
</tr>
<tr>
<td>MLLR</td>
<td>31.4</td>
<td>36.2</td>
<td>33.9</td>
<td></td>
</tr>
<tr>
<td>MBRLR (word)</td>
<td>33.2</td>
<td>37.8</td>
<td>35.6</td>
<td></td>
</tr>
<tr>
<td>MBRLR (phoneme)</td>
<td>33.2</td>
<td>37.7</td>
<td>35.6</td>
<td></td>
</tr>
<tr>
<td>MBRLR (state)</td>
<td>33.2</td>
<td>37.7</td>
<td>35.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Performance of standard MLLR and MBRLR ($rt06seval$ and $rt07seval$ datasets).
thresholded MLLR. The results are compared with the performance of confidence-weighted MLLR in Section 8.2.2.

The multi-pass recognition system described in Section 6 is used, where the third-pass adaptation step is standard MLLR or confidence-informed MLLR. Four adaptation iterations of adaptation are deployed in all experiments, no significant changes in WER being observed with a larger number of iterations.

8.2.1 Confidence-thresholded MLLR

Figure 10 displays the WER of the confidence-thresholded MLLR-adapted models as a function of the confidence threshold. The rt06seval and rt07seval datasets are used as test data. Each curve corresponds to a maximal frame posterior confidence measure as indicated by the legend. The unadapted models yield a WER of 35.8% and standard unsupervised MLLR corresponds to a confidence threshold of 0.0, a WER of 33.9%.

![Figure 10: Performance of confidence-thresholded MLLR-adapted models as a function of confidence threshold (rt06seval and rt07seval datasets).](image)

The curves of Figure 10 describe a tradeoff between the amount of adaptation data retained and the accuracy of the transcription of the retained data. At thresholds above 0.95 the volume of retained adaptation data quickly decreases and consequently the performance of the confidence-based adaptation scheme is compromised. At thresholds nearer 0.0 the confidence-based adaptation scheme uses almost all the adaptation data and the adapted system yields performance similar to that of standard MLLR. Each curve achieves a minimum within the range of 0.7 to 0.95, corresponding to the optimal confidence threshold for the associated confidence measure.

While the best performance is achieved using the phoneme-based confidence measure at a
threshold of 0.9, no significant difference is found between this system and the word, model and state-level confidence-thresholded systems at their optimal thresholds of 0.9, 0.7 and 0.7 respectively. The optimal confidence threshold corresponds with a WER of 33.7% in the case of all four confidence measures. This constitutes a significant reduction over the WER of standard MLLR in all cases.

### 8.2.2 Confidence-weighted MLLR

Table 6 displays the performance of the confidence-weighted MLLR-adapted models for the *rt06seval* and *rt07seval* datasets. The initial row details the performance of standard MLLR adaptation and the remaining rows correspond to the maximal frame posterior confidence measures indicated in the second column. Small, consistent improvements over standard MLLR are yielded by confidence-weighted MLLR when using the maximal frame posterior confidence measure at the word, phoneme, model and state levels. Significance testing reveals that these are significant improvements in the case of the word, phoneme and state-level confidence measures. The model-level confidence-weighted MLLR system gives a significant improvement over standard MLLR with 93.7% confidence, which is below the 95% confidence limit deemed as significant. No significant difference is found between any pair of confidence-weighted MLLR systems.

The improvements yielded by confidence-weighted MLLR over standard MLLR are smaller than those yielded by the optimally-thresholded systems of Section 8.2.1. However, when applying the MPSSWE significance test, no significant difference is found between any confidence-weighted MLLR system and, for example, the word-level confidence-thresholded MLLR system at a threshold of 0.9.

<table>
<thead>
<tr>
<th>System</th>
<th>WER (%)</th>
<th>rt06seval</th>
<th>rt07seval</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard MLLR</td>
<td></td>
<td>31.4</td>
<td>36.2</td>
<td>33.9</td>
</tr>
<tr>
<td>Confidence measure</td>
<td></td>
<td>Word</td>
<td>31.3</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>Phoneme</td>
<td>31.3</td>
<td>36.0</td>
<td>33.8</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>31.3</td>
<td>36.1</td>
<td>33.8</td>
</tr>
<tr>
<td></td>
<td>State</td>
<td>31.3</td>
<td>36.1</td>
<td>33.8</td>
</tr>
</tbody>
</table>

Table 6: Performance of confidence-weighted MLLR-adapted models (*rt06seval* and *rt07seval* datasets).

8.3 Experiment 3: Ideal confidence-informed MBRLR

To establish the optimal performance of confidence-informed MBRLR adaptation techniques, an experiment using ideal confidence measures is conducted. The ideal confidence measures are those which dismiss the frames of incorrectly transcribed data (with respect to the correct alignment) and retain the correctly-transcribed frames. So, for example, the ideal phoneme-level confidence measure dismisses all frames for which the phoneme label of the MAP alignment disagrees with the phoneme label of the correct alignment. The most likely alignment of the correct transcription is used in this way to identify the ideal confidence measures at the word, phoneme and state levels. Figure 11 plots the WER of the adapted system when using ideal confidence-thresholded MBRLR and the *rt06seval* and *rt07seval* test datasets. The horizontal axis represents the adaptation iteration. Each solid curve corresponds with a different MBR criterion formulation and the associated ideal confidence measure.

---

2 A similar result is found when comparing confidence-thresholded and confidence-weighted MLLR using the word-level maximal frame posterior confidence measure in Chapter 9 of Pitz (2005).

3 An ideal confidence measure can be defined as 1 if the label is correct and 0 otherwise. So confidence-weighted MBRLR behaves identically to confidence-thresholded MBRLR (with a threshold strictly between 0 and 1) under these ideal conditions.
Recall from Section 4.5 that confidence-informed MLLR may be implemented by effectively ignoring the discriminative term of the I-smoothed MBRLR criterion. For the purposes of comparison, the ideal model-level confidence-informed MLLR performance is measured and marked on Figure 11 (33.1%).

After thirty adaptation iterations, all ideal confidence-thresholded MBRLR formulations achieve a significantly lower WER than ideal confidence-informed MLLR. This is because the negative impact of the erroneous transcription has been constrained. Without this negative impact, the benefit of the discriminative MBRLR adaptation technique over MLLR adaptation method becomes observable. Thus, given perfect knowledge of the correct and incorrectly-transcribed reference labels, confidence-informed MBRLR delivers performance significantly superior to that of confidence-informed MLLR. This observation provides motivation for the use of confidence-informed MBRLR.

8.4 Experiment 4: Confidence-thresholded MBRLR

Using imperfect confidence measures, confidence-thresholded MBRLR is now compared with standard MBRLR. This experiment is designed to gain insight into the effect of the confidence threshold upon the adaptation process, and to compare word and sub-word confidence-informed MBR criteria. Note that the confidence measure and MBR criterion correspond, so, for example, a phoneme-level confidence measure is used in conjunction with the phoneme-level MBR criterion.

Figure 12 displays the WER yielded by the confidence-thresholded MBRLR-adapted models after twenty adaptation iterations at a range of confidence thresholds between 0.1 and 0.99. Each curve represents a different confidence-thresholded MBR criterion, corresponding to the criterion and confidence measures at the word, phoneme and state levels. Note that the model-level criterion and confidence measure does not feature in these experiments due to the close similarity of both the state and model-level criteria [Gibson, 2008] and the state and model-level confidence measures (see Section 7.2).

The threshold of 0.0 in Figure 12 corresponds to standard MBRLR adaptation (Table 5).
Figure 12: Confidence-thresholded MBRLR performance as a function of confidence threshold (rt06seval and rt07seval datasets).

Each of the three curves displays a gradual decrease in WER to a minimum at a confidence threshold in the range of 0.8 to 0.9 inclusive. This demonstrates the benefit of ignoring errors associated with low-confidence labels of the reference transcription. At confidence thresholds above 0.9 an increase in WER is witnessed in all cases. This illustrates the tradeoff between ignoring errors derived from low-confidence labels and ignoring all error information. Notice that, in the case of the phoneme-level formulation, the increase in WER at thresholds above 0.9 is less pronounced than the increase observed in the case of the word and state-level formulations. This is because a comparatively large volume of data has confidence above 0.9 in the case of the phoneme-level confidence measure, as displayed in Figure 8.

In the case of all three criteria considered, confidence-thresholded MBRLR at the optimal confidence threshold yields significantly improved performance over the performance of standard word-level MBRLR (35.6% WER); a WER of 35.4% in the case of phoneme-level MBRLR at a confidence threshold of 0.9 and a WER of 35.2% at the threshold of 0.8 in the case of both word and state-level MBRLR. Note further that word and state-level confidence-thresholded MBRLR provides a larger absolute gain (0.4%) than the 0.2% gain provided by word and state-level confidence-thresholded MLLR (Figure 10).

8.5 Experiment 5: Confidence-weighted MBRLR

Use of confidence-weighted MBRLR in the third-pass adaptation step of the evaluation system described in Section 6 yields the results presented in Table 7. Again, the MBR criterion and confidence measure correspond. So, for example, the system labelled ‘state-level’ deploys a confidence-weighted state-level MBR criterion, where the confidence measure used is the state-level maximal frame posterior. Twenty transform estimation iterations are performed in all cases.

Let us compare the performance of confidence-weighted MBRLR with the performance of standard unsupervised MBRLR, shown in Table 5. A small performance improvement is yielded over standard MBRLR (which yields 35.6% WER for all criterion formulations) in the case of
all criteria and corresponding confidence measures. However confidence-weighted MBRLR yields more modest WER gains compared to the optimal confidence-thresholded systems of Section 8.4.

Significance testing reveals significant differences between the performance of all three confidence-weighted MBRLR-adapted systems and the standard word-level MBRLR-adapted system. No significant difference is found between the performance of the three confidence-weighted MBRLR-adapted systems.

Significance testing also shows that, for example, the confidence-thresholded state-level MBRLR-adapted models, using a confidence threshold of 0.8, delivers significantly better performance than any of the confidence-weighted MBRLR-adapted models. Given the optimal confidence threshold, it is more beneficial to completely disregard, rather than to merely de-emphasise, errors derived from low-confidence reference labels. A similar phenomenon has been witnessed in the case of confidence-informed MLLR adaptation (Chapter 9 of [Pitz, 2005]) when using the word-level maximal frame posterior confidence measure.

### 8.6 Experiment 6: Confidence-thresholded MBRLR and I-smoothing

The experiments described previously in this paper have not used the I-smoothing technique for MBRLR, introduced in Section 3.4. An experimental evaluation of the impact of the I-smoothing method is now presented. As discussed in Section 8.3, the issue of generalisation is relevant to confidence-thresholded MBRLR. Since I-smoothing can improve the generalisation of MBR-estimated acoustic models, it is feasible that unsupervised confidence-thresholded MBRLR adaptation will also benefit from the use of I-smoothing.

#### 8.6.1 Ideal confidence case

Figure 13 plots the performance of the first 140 iterations of I-smoothed confidence-thresholded phoneme-level MBRLR adaptation in the case of the ideal phoneme-level confidence threshold. Each curve corresponds to a different value of the constant $\tau$ (see Equation 19), as indicated in the legend.

The use of a zero-valued I-smoothing constant $\tau$ corresponds to no use of smoothing and the ideal phoneme-level confidence-thresholded performance of Figure 11. This curve converges at a WER of 31.1% after 140 iterations. As $\tau$ increases, the criterion approximates more closely the phoneme-level confidence-thresholded ML criterion and the transform re-estimation equations approximate more closely the phoneme-level confidence-thresholded MLLR transform re-estimation. So as $\tau$ increases, the associated curve approximates more closely the curve corresponding to ideal phoneme-level confidence-informed MLLR. With sufficiently large $\tau$, the discriminative part of the smoothed MBR criterion is effectively ignored and the transform re-estimation equations effectively implement MLLR transform re-estimation.

One side-effect of increasing $\tau$ is quicker convergence of the criterion. This is consistent with the relatively quick convergence of ML re-estimation procedures over MBR re-estimation procedures. However no beneficial effect is observed in terms of performance for non-zero values of $\tau$. At a $\tau$ value of 0.05 the WER converges to 31.4%, and at a $\tau$ value of 0.1 the WER converges to 31.7%. So, in the case of the ideal confidence measures, I-smoothed confidence-thresholded MBRLR fails to deliver improved performance over unsmoothed MBRLR.
8.6.2 Imperfect confidence case

Table 8 presents the performance yielded after twenty iterations of I-smoothed phoneme-level confidence-thresholded MBRLR for a range of smoothing factors $\tau$. In this case the imperfect phoneme-level maximal frame posterior confidence measure at a threshold of 0.9 is used. Again, a $\tau$ value of 0.0 corresponds to no use of I-smoothing. An arbitrarily large value of $\tau$ corresponds to phoneme-level confidence-thresholded MLLR. As in the case of ideal confidence-thresholded MBRLR, as $\tau$ increases, the performance more closely approximates the performance of confidence-thresholded MLLR. In this case, increasing $\tau$ yields significant performance improvements. However, no value of $\tau$ delivers improvements over confidence-thresholded MLLR performance.

To summarise the results of this section, in the case of unsupervised MBRLR adaptation using the dataset considered in this experiment, and with an ideal confidence measure, I-smoothing de-
livers no improved generalisation over unsmoothed confidence-thresholded MBRLR adaptation. In the case of unsupervised MBRLR adaptation using the dataset considered in this experiment, and with an imperfect confidence measure, I-smoothed MBRLR delivers no improved generalisation over standard confidence-informed MLLR adaptation.

9 Summary and future work

This paper has introduced novel confidence-informed MBR criteria by refining the approximate error function of the standard MBR criterion. Experimental evaluations on a large vocabulary recognition task have demonstrated that both confidence-thresholded and confidence-weighted unsupervised MBRLR adaptation deliver significant performance improvements over standard unsupervised MBRLR adaptation when using confidence measures based on frame posteriors.

While standard unsupervised MBRLR adaptation gives significantly inferior performance to standard unsupervised MLLR adaptation, it has been shown that, given an ideal confidence measure (as defined in Section 8.3), confidence-thresholded unsupervised MBRLR adaptation yields significantly superior performance to confidence-informed unsupervised MLLR.

Experimentation with confidence-thresholded MBRLR formulations corresponding to word and sub-word hypothesis spaces reveals that, in the case of ideal confidence measures, sub-word formulations yield significantly superior performance to ideal word-level confidence-thresholded MBRLR.

Although the use of the I-smoothing technique is motivated with regard to unsupervised confidence-informed MBRLR adaptation, no performance improvement over unsmoothed MBRLR has been observed when using this technique in the case of ideal confidence measures. Nor has any improvement over confidence-thresholded MLLR adaptation been witnessed when using I-smoothed confidence-thresholded MBRLR adaptation with maximal frame posterior confidence measures.

9.1 Future work

There is much scope for future research in confidence-informed MBRLR adaptation. This work could focus on several different aspects of the adaptation technique. For example, investigative work could determine if techniques used to enhance the generalisation of standard MBR parameter estimation, e.g. acoustic scaling and choosing a less specific language model, are applicable also to supervised and unsupervised MBRLR. Additionally, the impact of different learning rates (the constant $E$ in the implementation used in this work) upon MBRLR adaptation should be measured and understood.

There is certainly a need to develop an understanding of the relationship between errors made by the confidence-based classifier (when classifying frames as correctly or incorrectly labelled) and the performance of the confidence-thresholded MBRLR adaptation process.

The field of confidence estimation is an area of research which directly impacts confidence-informed MBRLR. As evidenced by the experimental work of this paper, with improved confidence measures, confidence-thresholded MBRLR adaptation has the capacity to yield performance which is superior to state-of-the-art confidence-informed MLLR adaptation. Future improvements on confidence measures are the key to access this superior performance.

References


