Hypothesis Spaces For Minimum Bayes Risk Training In Large Vocabulary Speech Recognition

Matthew Gibson, Thomas Hain

Department of Computer Science, University of Sheffield,
211 Portobello Street, Sheffield, S1 4DP, UK
{mgibson,th}@dcs.shef.ac.uk

Abstract

The Minimum Bayes Risk (MBR) framework has been a successful strategy for the training of hidden Markov models for large vocabulary speech recognition. Practical implementations of MBR must select an appropriate hypothesis space and loss function. The set of word sequences and a word-based Levenshtein distance may be assumed to be the optimal choice but use of phoneme-based criteria appears to be more successful. This paper compares the use of different hypothesis spaces and loss functions defined using the system constituents of word, phone, physical triphone, physical state and physical mixture component. For practical reasons the competing hypotheses are constrained by sampling. The impact of the sampling technique on the performance of MBR training is also examined.

Index Terms: discriminative training, Minimum Bayes Risk.

1. Introduction

Discriminative training of acoustic models has yielded test set performance improvement over maximum likelihood (ML) training in large vocabulary continuous speech recognition (LVCSR). Recently acoustic models optimised using the maximum mutual information (MMI) criterion have outperformed those trained using the ML criterion for the task of conversational telephone speech (CTS) transcription [1].

The MMI criterion aims to increase the posterior probability of the correct transcription of the acoustic training data. Hence it is not directly linked to the standard performance measurement, word error rate (WER). Contrastingly the Minimum Bayes Risk (MBR) estimation framework [2] incorporates a performance measurement, known as the *loss function*, into the training criterion.

The expected value of the loss of a MAP decoder with parameters θ is given by

$$\mathcal{R}_T(\theta) = \sum_{\bar{W}\in\mathcal{W}} \sum_{W'\in\mathcal{W}} \int P(W'|\mathcal{O},\theta) l(\bar{W},W') P(\bar{W},\mathcal{O}) d\mathcal{O} \tag{1}$$

where $\mathcal{O}$ is an acoustic observation sequence and $\bar{W}$ is the corresponding correct hypothesis. $\mathcal{W}$ is the *hypothesis space* and $l(\bar{W},W')$ is the loss function. If the hypothesis space is the set of all possible word sequences and the loss function is the Levenshtein (string edit) distance between word sequences $\bar{W}$ and W' then the expected loss is identical to the expected WER.

Since reduction of the expected WER is the aim of most parameter estimation techniques $\mathcal{R}_T(\theta)$ is the ideal training objective function. However it is not useful since the joint distribution of acoustics and word sequences $P(\bar{W},\mathcal{O})$ is unknown. The normalised MBR objective function $\mathcal{R}(\theta)$ given by

$$\mathcal{R}(\theta) = \frac{1}{N} \sum_{r} \sum_{W'\in\mathcal{W}} P(W'|\mathcal{O}^r,\theta) l(\bar{W}^r,W') \tag{2}$$

approximates the expected loss using a finite training dataset. Here $\mathcal{O}^r$ represents the acoustic feature sequence associated with the r^{th} training utterance, $\bar{W}^r$ is the corresponding correct hypothesis and N is the number of training examples. Note that as the training set size increases the normalised MBR objective function converges to the expected loss given by Equation 1. Using a hypothesis space equal to the set of all possible word sequences and the Levenshtein word error loss function the MBR technique has been applied to LVCSR with reported improvements over MMI-trained systems [3].

Minimum phone error (MPE) training [4] can be interpreted as an instance of MBR training where the set of all possible phone sequences forms the hypothesis space. The Levenshtein word error loss function is replaced by a phoneme error loss function. Use of this alternative criterion has been shown to both outperform MMI and yield test set performance gains over an equivalent word-level MBR criterion on a CTS task [5].

This paper builds on the ideas of MPE by further exploring the use of alternative hypothesis spaces within the MBR parameter re-estimation framework. Frame error rate is used as a loss function, allowing the definition of a range of alternative criteria covered by the MBR framework. Techniques for predicting the effectiveness of these criteria are presented and their actual effectiveness is evaluated.

The rest of the paper is organised as follows. Section 2 overviews the theory and technical implementation of MBR parameter re-estimation. Section 3 introduces a method for predicting the effectiveness of the MBR training criterion. Experimental evaluations of the criteria are reported in Section 4 while Section 5 discusses possible future research.

2. MBR Theory and Implementation

The MBR parameter updates for hidden Markov models with Gaussian output distributions are derived in [2]. We repeat the mean update here for convenience.

$$\hat{\mu}_s = \frac{\sum_{r=1}^{N} \sum_{W'\in\mathcal{W}} K^r(W'|\theta) \sum_{t=1}^{T(r)} \gamma_s(t|W',\mathcal{O}^r)\mathbf{o}_t^r + D\mu_s}{\sum_{r=1}^{N} \sum_{W'\in\mathcal{W}} K^r(W'|\theta) \sum_{t=1}^{T(r)} \gamma_s^r(t|W',\mathcal{O}^r) + D} \tag{3}$$

where $\hat{\mu}_s$ is the updated mean of mixture component s, μ_s is the current mean, $\mathbf{o}_t^r$ is the acoustic feature vector of the t^{th} frame of the r^{th} training example. D is a learning rate discussed in Section 2.4, $\mathcal{W}$ is the hypothesis space, $\gamma_s^r(t|W', \mathcal{O}^r)$ is the posterior probability of occupancy of component s at time t given hypothesis W' and observation sequence $\mathcal{O}^r$ and

$$K^r(W'|\theta) = P(W'|\mathcal{O}^r, \theta)\big[l_{av}^r - l(\bar{W}^r, W')\big] \tag{4}$$

where l_{av}^r is the average loss given by

$$l_{av}^r = \sum_{W \in \mathcal{W}} P(W|\mathcal{O}^r, \theta) l(\bar{W}^r, W) \tag{5}$$

Here $\bar{W}^r$ is the reference hypothesis.

2.1. Lattice-based MBR

In the context of large vocabulary systems, when using the set of all possible word sequences as the hypothesis space $\mathcal{W}$, a prohibitively large amount of computation is required to calculate the statistics required for MBR estimation. Practical solutions to this problem are to approximate this space either using an N-best list of the most likely hypotheses [2] or to use lattices as a more compact representation [1]. Word lattices which encode temporal alignment information (i.e. word start and end times) [6] are used in this work. These lattices are generated via an unconstrained recognition pass. The most likely alignments of the correct word sequence, generated using a constrained recognition pass, are merged into the recognition lattice to form a *consolidated recognition lattice*. This lattice can be viewed as a representation of alternative word-level alignments of the acoustic data.

Constraining the search space to only those alignments specified by the lattice, Equation 3 may be written

$$\hat{\mu}_s = \frac{\sum_{r=1}^{N} \sum_{a \in \mathcal{A}^r} K^r(a|\theta) \sum_{t=a_{\text{start}}}^{a_{\text{end}}} \gamma_s^r(t|a, \mathcal{O}^r) \mathbf{o}_t^r + D\mu_s}{\sum_{r=1}^{N} \sum_{a \in \mathcal{A}^r} K^r(a|\theta) \sum_{t=a_{\text{start}}}^{a_{\text{end}}} \gamma_s^r(t|a, \mathcal{O}^r) + D} \tag{6}$$

where a is a lattice arc representing a word, its start time a_{start} and end time a_{end}, $\mathcal{A}^r$ is the set of all arcs in the lattice and

$$K^r(a|\theta) = p(a|\mathcal{O}^r, \theta)\big[l_{av}^r - l(\bar{W}^r, a)\big] \tag{7}$$

where $p(a|\mathcal{O}^r, \theta)$ is the posterior probability that arc a is included in a *path* i.e. a contiguous sequence of arcs from the lattice start node to the lattice end node. $l(\bar{W}^r, a)$ is the posterior-weighted sum of the loss of all the lattice paths which include arc a, while l_{av}^r is the posterior-weighted sum of the loss of all the lattice paths.

Calculation of $l(\bar{W}^r, a)$ can be problematic when using a Levenshtein loss function. These difficulties arise because in this case an arc may make different contributions to the loss of two different containing paths. A possible solution to this issue is presented in [3] and involves assigning to each lattice arc a sequence of words within the reference transcription, referred to as lattice-to-string alignment. This is both conceptually and practically problematic and so in this work, likewise in the approximate MPE technique [5], an alternative loss function is defined which avoids the difficulties imposed by the Levenshtein loss function.

2.2. Hypothesis Spaces and Loss Function

To fully specify the MBR criterion one must define the hypothesis space $\mathcal{W}$ and the loss function $l(\bar{W}^r, W)$. In this work the theoretical hypothesis spaces investigated are defined by the set of all possible temporal alignments of the following five system constituents: words, phones, physical triphone models, physical states and physical mixture components. Words and phones are those labels used in the recognition dictionary. A physical state represents a state cluster and a physical triphone model represents each HMM comprising a unique set of physical states. A physical mixture component is a Gaussian mixture component of the output distribution of a physical state.

A sample of the hypothesis space is represented by the word lattice described above in the following way. Associated with each arc of the consolidated recognition lattice is a *constituent alignment*. This is a temporal alignment of one of the following: words, phones, physical triphone models, physical states or physical mixture components. Thus any lattice path has an associated constituent alignment; the concatenation of the constituent alignments of its arcs. The lattice may therefore be viewed as a representation of competing constituent alignments i.e. a sample of the hypothesis space $\mathcal{W}$.

2.2.1. Loss Function Definition

The loss of an alignment W is defined as the number of frames at which the constituent specified within W differs from the constituent specified by the reference hypothesis $\bar{W}^r$. This loss function is referred to as the *frame error rate* (FER).

The loss associated with each lattice path is defined as the loss of its associated alignment. Defining the FER of an arc as the number of frames for which the associated alignment differs from the reference hypothesis one can see that the FER of a path is simply the sum of the FER of its arcs. Each arc therefore contributes an equal amount to the loss of its containing paths and the difficulties encountered when using a Levenshtein distance are avoided.

Note that the reference hypothesis $\bar{W}^r$ is the most likely constituent alignment of the correct (word-level) transcription, generated using a constrained recognition pass.

2.3. Forward-Backward Algorithms

Calculation of the posterior probabilities $\gamma_s^r(t|a, \mathcal{O}^r)$ necessary to perform the update of Equation 6 requires a standard forward-backward pass over the models defined by each lattice arc a using the segment of acoustic data assigned to arc a. In order to calculate $K^r(a|\theta)$, a lattice-level forward-backward pass is conducted as detailed in [5].

2.4. I-Smoothing and the Learning Rate D

MBR parameter updates can be unstable and require regularisation. The I-smoothing technique [4] defines a prior distribution over the acoustic model parameters, the sharpness of which is determined by a parameter τ^I. This prior distribution is then integrated into the MBR objective function to smooth the parameter updates. This technique is used in the experiments described in Section 4.

The learning rate D of Equation 6 is specific to each Gaussian mixture component. To determine its value the occupancy-dependent scheme described in [5] is used, i.e. for each mixture component s

1. Calculate D_s^{min}, the minimum D required to ensure all variance updates are positive for component s.
2. Set $\gamma_s^{den} = \sum_r \sum_{a \in \mathcal{A}_{den}^r} K^r(a|\theta) \sum_{t=a_{start}}^{a_{end}} \gamma_s^r(t|a, \mathcal{O}^r)$

where $\mathcal{A}^r_{den}$ denotes the subset of lattice arcs in $\mathcal{A}^r$ for which $K^r(a|\theta)$ is negative.

3. Set the learning rate D_s to $\max(2D_s^{min}, E\gamma_s^{den})$ where E is a configurable parameter.

3. Performance Prediction

To predict the effectiveness of a particular MBR criterion one can measure the strength of the correlation between the criterion function and the performance measurement of interest, in this case the test set WER. The higher this correlation the more effective (in terms of test set WER improvement) the criterion should be.

To measure such correlations several datapoints are required. A datapoint is a 2-tuple $(\mathcal{R}(\theta), \mathcal{E}(\theta))$ where $\mathcal{R}(\theta)$ is the criterion function and $\mathcal{E}(\theta)$ is the test set WER. Multiple datapoints are generated by sampling the parameter space then measuring the values of $\mathcal{R}(\theta_i)$ and $\mathcal{E}(\theta_i)$ for each point θ_i in the parameter space. A point in parameter space is generated by adapting the baseline system.

3.1. Baseline System

The baseline system is trained using maximum likelihood training and the `WSJ0` corpus `SI84` Sennheiser microphone dataset. This comprises 12.67 hours of speech data, 83 different speakers and approximately 7000 utterances.

The acoustic models used are tied-state triphone models. Maximum likelihood clustering techniques [6] are used to cluster the triphone states. 8 Gaussian mixtures model the state output distributions and 3877 tied states are used. A 39-dimensional feature vector is used to represent the acoustic data. This comprises 12 perceptual linear prediction (PLP) coefficients, log energy and the first and second time derivatives of these variables. The features are normalised using cepstral mean normalisation to reduce the effects of the input channel. All training and test utterances are pre-processed to contain a maximum of 0.1 seconds of silence at the start and end to ensure the effectiveness of the cepstral mean normalisation step.

3.2. Sampling the Parameter Space

Parameter space samples θ_i are generated via speaker adaptation of the baseline system described above. For each speaker in the `WSJ0 SI84` Sennheiser dataset the speaker-specific subset of utterances are used to adapt the baseline system using maximum likelihood linear regression [7]. Thus 83 speaker-dependent (SD) systems are generated, a sample of the parameter space.

3.3. Measuring the Test Set WER

The `WSJ0` speaker-independent 5k Sennheiser evaluation dataset is used as test data. The closed vocabulary 5k bigram language model provided with the `WSJ0` corpus is used in decoding with a language model scale factor of 16. The test set WER $\mathcal{E}(\theta_i)$ is measured for each SD system θ_i.

3.4. MBR Criterion Correlations

Good parameter optimisation criteria should perfectly correlate with the word error rate on an independent test set. In practice this correlation is not only dependent on the optimal choice of loss function but also on the implementation detail (e.g. hypothesis space construction) and the amount of training data available. Hence the logical choice of criterion function does not necessarily yield optimal performance. In this section the criterion function $\mathcal{R}(\theta_i)$ is calculated for each SD system θ_i using a unigram language model. The values $\mathcal{R}(\theta_i)$ are then correlated with the test set word error rates $\mathcal{E}(\theta_i)$ described in Section 3.3.

To measure the MBR criterion the hypothesis space is sampled since exploration of all hypotheses is computationally infeasible for the spaces considered here. Sampling the space $\mathcal{W}$ means choosing an appropriate subset of $\mathcal{W}$ for each utterance.

3.4.1. 1-Best Sampling

A first approximation uses the first-best hypothesis as a simplified representation of confusability. The most likely constituent alignment of the most likely word sequence is used as the sole sample of the hypothesis space. The MBR criterion of each SD system θ_i is then measured for each hypothesis space using both Levenshtein distance and FER. Note that it is possible to use the Levenshtein loss in this case because a single hypothesis does not present the difficulties discussed in Section 2.1. Table 1 shows the correlation coefficient of $\mathcal{R}(\theta_i)$ with $\mathcal{E}(\theta_i)$ for each hypothesis space/error metric combination. Using the significance test for the difference

Table 1: *Correlation of* 1*-Best MBR Criterion with Test Set WER*

	Error Metric	
Hypothesis Space	Levenshtein	FER
Word	0.11	0.13
Phone	0.14	0.27
Physical Triphone	0.21	0.30
Physical State	0.21	0.31
Physical Mixture	0.39	0.39

between dependent correlations [8] it is observed that use of the FER metric yields a significantly greater (at the 95% confidence level) correlation coefficient than the Levenshtein metric when using phone, physical triphone and physical state hypothesis spaces. Defining the *space resolution* as the average number of system constituent labels per utterance, note that this metric increases on descending the rows of Table 1. The correlation coefficient increases with space resolution in the case of both the FER and the Levenshtein metric. In the case of the FER metric almost all coefficient pairs are significantly different at the 95% confidence level.

The use of temporal information within the loss function definition and deployment of higher-resolution hypothesis spaces both result in a 1-best MBR criterion more closely correlated with the test set WER. One therefore predicts that usage of such information will result in a criterion which is more robust to sparse sampling of the hypothesis space. Smaller sample sizes reduce the amount of computation required for MBR training so this observation is also of pragmatic importance.

3.4.2. Sampling Multiple Hypotheses

Two different methods are used to sample multiple hypotheses. The first technique is to sample the most likely constituent alignments. This sample set is referred to as $\mathcal{W}^L$.

The second sampling method firstly identifies the most likely word-level alignments. Then, for each of these word-level alignments, the most likely constituent alignment is identified and added to the sample set. The resulting sample set is referred to

as $\mathcal{W}^W$. Note that when word-level constituent alignments are used $\mathcal{W}^L$ and $\mathcal{W}^W$ coincide.

In all cases the sample set is represented in lattice format as described in Section 2.2. A threshold is applied to limit the density of this lattice. This threshold is identical for both of the above sampling techniques.

Table 2 displays the correlation coefficient of $\mathcal{R}(\theta_i)$ with $\mathcal{E}(\theta_i)$ for each hypothesis space/sampling method combination. The benefit of incorporation of multiple hypotheses is evident since the correlation coefficients of Table 2 are, in general, significantly higher than those of Table 1. However very few significant differences are observed between the coefficients of Table 2. One therefore predicts similar test set performance after MBR parameter re-estimation using each hypothesis space/sampling method combination.

Table 2: *Correlation of MBR Criterion with Test Set WER*

	Sample	
Hypothesis Space	$\mathcal{W}^W$	$\mathcal{W}^L$
Word	0.51	0.51
Phone	0.42	0.44
Physical Triphone	0.47	0.47
Physical State	0.39	0.40
Physical Mixture	0.39	0.40

4. Evaluation

The test of the effectiveness of an MBR training criterion is to measure its influence on test set WER. This section reports the evaluation of each of the MBR configurations.

The baseline system described in Section 3.1 is re-estimated using the `WSJ0 SI84` Sennheiser dataset and 7 iterations of MBR training. An I-smoothing τ^I of 100 and an E value of 4 are used in parameter re-estimation. A unigram language model [9] and acoustic probability scaling [1] are deployed to improve the generalisation of the procedure. A language model scale factor of $\frac{1}{2}$ and an acoustic model scale factor of $\frac{1}{16}$ are used.

The decoding procedure is as described in Section 3.3. Table 3 displays the test set WER when using each of the MBR configurations. The ML baseline WER for this task is 6.63%.

Table 3: WSJ0 5k *WER*

	Sample	
Hypothesis Space	$\mathcal{W}^W$	$\mathcal{W}^L$
Word	6.55	6.55
Phone	6.52	6.59
Physical Triphone	6.52	6.59
Physical State	6.53	6.67
Physical Mixture	6.53	6.61

When using the hypothesis sample $\mathcal{W}^W$ test set WER improvements over the ML baseline are observed in all cases and no significant difference is observed between the different hypothesis spaces. This concurs with the predictions of Section 3.4.2.

Using the hypothesis sample $\mathcal{W}^L$ much smaller test set WER improvements are observed in general. This discrepancy between the sampling techniques is not predicted by the results of Section 3.4.2. This is because much larger samples are used in parameter re-estimation and the benefits of using sample $\mathcal{W}^W$ over $\mathcal{W}^L$ are not evident when using the smaller samples deployed for the purpose of calculating the correlation coefficients of Section 3.4.2.

5. Conclusions and Future Research

This paper has motivated the exploration of alternative hypothesis spaces and loss functions within the MBR formulation. An empirical technique for predicting the effectiveness of the MBR configuration has been described. The impact of hypothesis space definition and sampling technique has been examined. Evidence has been presented to support the utilisation of temporal information in the loss function definition and the use of a high-resolution hypothesis space for robust MBR in cases of sparse sampling.

One deficiency of the MBR formulation described in this paper derives from the fact that the reference hypotheses are generated using constrained recognition and an imperfect recognition model. This reference alignment is then used when applying the FER loss function. Possible future work could incorporate not just one but several reference alignments into the FER loss function definition to minimise punishment of favourable hypotheses which disagree with the most likely reference alignment.

Another potential line of future research is to track the relative performance of the MBR training configurations whilst varying the quantity of training data. This is particularly interesting in the case of small amounts of training data, e.g. in the case of speaker adaptation.

6. References

[1] Woodland, P.C. and Povey, D., "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition", Computer Speech and Language, Vol. 16, 2002, pp. 25-47.

[2] Kaiser, J., Horvat B. and Kacic, Z., "Overall Risk Criterion Estimation of Hidden Markov Model Parameters", Speech Communication, Vol. 38, 2002, pp. 383-398.

[3] Doumpiotis, V. and Byrne, W., "Pinched lattice minimum Bayes risk discriminative training for large vocabulary continuous speech recognition", Proceedings Interspeech, pp. 1717-1720, 2004.

[4] Povey, D. and Woodland, P.C., "Minimum Phone Error and I-Smoothing for Improved Discriminative Training", Proceedings ICASSP, Vol. 1, pp. 105-108, 2002.

[5] Povey, D., "Discriminative Training for Large Vocabulary Speech Recognition", PhD Thesis, Department of Engineering, University of Cambridge, UK, 2003.

[6] S.J. Young et al., "The HTK Book for HTK version 3.2.1", 2003.

[7] Leggetter, C.J. and Woodland, P.C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech & Language, Vol. 9, 1995, pp. 171-185.

[8] Chen, P. Y. and Popovich, P. M., Correlation: Parametric and nnparametric measures, Sage Publications, Thousand Oaks, CA, 2002.

[9] Schlueter, R., Macherey, W., Muller B., and Ney, H., "Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition" Speech Communication, Vol. 34, 2000, pp. 287-310.