interspeech 2008 incorporating SST 08

Penalty Function Maximization for Large Margin HMM Training

George Saon and Daniel Povey

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
e-mail: {gsaon,dpovey}@us.ibm.com

ABSTRACT

We perform large margin training of HMM acoustic parameters by maximizing a penalty function which combines two terms. The first term is a scale which gets multiplied with the Hamming distance between HMM state sequences to form a multi-label (or sequence) margin. The second term arises from constraints on the training data that the joint log-likelihoods of acoustic and correct word sequences exceed the joint log-likelihoods of acoustic and incorrect word sequences by at least the multi-label margin between the corresponding Viterbi state sequences. Using the soft-max trick, we collapse these constraints into a boosted MMI-like term. The resulting objective function can be efficiently maximized using extended Baum-Welch updates. Experimental results on multiple LVCSR tasks show a good correlation between the objective function and the word error rate.

1. INTRODUCTION

Recently there has been a lot of interest in large margin approaches for training HMMs in speech recognition [1, 2, 3]. We previously introduced the technique of Boosted MMI [4] which uses the traditional framework of MMI training involving lattices and Extended Baum-Welch updates, but incorporates ideas from large margin classification. In this paper we make the connection to large margin more explicit and propose modifications that draw directly from large margin techniques, and serve to optimize an arbitrary factor that we introduced in our previous work.

Chronologically, the first application of large margin training for ASR appears to have been done in [1] and related references. Here, the authors use a generalized probabilistic descent algorithm to maximize a quantity termed relative margin which is one minus the ratio between the likelihood of the closest competitor and the likelihood of the correct sentence. One potential shortcoming of this Maximum Relative Margin Estimation (MRME) technique is that it doesn't handle well variable length utterances. This observation has been exploited in [2], where the authors propose a Soft-Margin Estimation (SME) technique which has the advantage of incorporating utterance length normalization. In addition, the SME objective function balances margin maximization and constraints violation which we are also advocating in this paper. However, both MRME and SME only deal with the closest competitor sentence and we feel that this can be a limiting factor especially for LVCSR.

Concomitantly, in [3] the authors propose a large margin training technique for ASR which has some appealing properties. First, they deal efficiently with an exponential number of constraints by using a "soft-max" trick. Second, they incorporate the margin definition proposed in [5] for sequence (or multi-label) classification which, in this case, becomes the scaled Hamming distance between HMM state sequences. Interestingly, the authors consider the margin scale to be a constant (i.e. 1) and only optimize the amount by which the margin constraints are violated given this fixed scale. Another characteristic of their approach is a particular parameterization of the Gaussian means and covariances which allows them, under some assumptions, to formulate and solve a convex optimization problem.

We differ with [3] in two important aspects: in our case, the margin scale becomes an integral part of the objective function (as in SME) and, more importantly, we use Extended Baum-Welch for the optimization by exploiting the connection with MMI as opposed to resorting to gradient descent procedures. Additional minor differences have to do with the consideration of language model scores and with the removal of the hinge function which leads to a smooth objective function.

The remainder of this paper is organized as follows. Section 2 introduces the large margin framework and shows how we can naturally adapt this to a MMI-like update for HMMs; Section 3 provides some experimental results on two LVCSR tasks and Section 4 summarizes our findings.

2. LARGE MARGIN TRAINING

2.1. General setting

We are given a set of training vector sequences with corresponding label sequences $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_r, Y_r), \ldots, (\mathbf{X}_R, Y_R)\}$, $\mathbf{X}_r = \mathbf{x}_{r1}, \ldots, \mathbf{x}_{rT_r}$, $Y_r = y_{r1}, \ldots, y_{rT_r}$, $\mathbf{x}_{rt} \in \mathbb{R}^n$, $y_{rt} \in \mathcal{Y}$ where $T_r = |\mathbf{X}_r| = |Y_r|$ represents the length of the sequences $\mathbf{X}_r$ and Y_r. The idea is to form a discriminant function $D(\mathbf{X}, Y)$ which has as arguments vector sequences and label sequences such that

$$D(\mathbf{X}_r, Y_r) \geq D(\mathbf{X}_r, Y), \qquad \forall\, Y \neq Y_r,\ |Y| = T_r \tag{1}$$

i.e. we want the discriminant function for the correct label sequence to be higher than for competitor label sequences of the same length. Furthermore, $D(\mathbf{X}_r, Y_r)$ has to exceed $D(\mathbf{X}_r, Y)$ by some positive quantity termed the margin. Maximizing this margin will increase the difference between the scores of the true label sequence and the closest competitor which, in turn, will increase the confidence of the classification. Since we are predicting multiple labels, we want to generalize the notion of margin to take into account the number of labels that are misclassified. In particular, we would like the margin between Y_r and Y to scale linearly with the number of different labels in Y as in [5]. One possibility is to define the margin between Y_r and Y as the scaled Hamming distance $\rho H(Y_r, Y)$ where:

$$H(Y_r, Y) := \sum_{t=1}^{T_r} I(y_{rt} \neq y_t) \tag{2}$$

Accepted after peer review of full paper

and $I(\cdot)$ is the $0-1$ loss (or indicator) function. $\rho > 0$ represents the margin scale. Armed with these simple definitions, we can formulate the margin constraint between Y_r and Y as:

$$D(\mathbf{X}_r, Y_r) - D(\mathbf{X}_r, Y) \geq \rho H(Y_r, Y) \tag{3}$$

Note that this inequality is trivially satisfied for $Y = Y_r$. We can therefore include the case $Y = Y_r$ in the subsequent derivations. Assuming the previous inequalities hold for multiple ρ's, it is natural to search for the maximum ρ subject to the constraints of (3). We then arrive at the following fairly general setup for large margin sequence classfication problems:

$$\begin{aligned} &\max \rho \\ &\text{s.t. } D(\mathbf{X}_r, Y_r) - D(\mathbf{X}_r, Y) \geq \rho H(Y_r, Y),\ \forall\, Y,\ 1 \leq r \leq R \end{aligned} \tag{4}$$

This problem formulation differs from the work of [3], where the authors assume $\rho = 1$ throughout their derivation and only minimize the constraints violation part. We will adopt however some steps from [3] which have to do with how to deal efficiently with exponentially many constraints. One such step is to replace (3) with the maximum constraint and reformulate (4) as:

$$\begin{aligned} &\max \rho \\ &\text{s.t. } D(\mathbf{X}_r, Y_r) - \max_Y \{D(\mathbf{X}_r, Y) + \rho H(Y_r, Y)\} \geq 0 \end{aligned} \tag{5}$$

A standard technique in optimization theory is to create a penalty (or merit) function which combines the original objective function with the constraints in order to form an unconstrained optimization problem. We opt for an L_1 exact penalty function[1] which can be written in the following manner cf. [6]:

$$\max \left\{ \rho - \right.$$
$$\left. -\frac{1}{\lambda} \sum_{r=1}^{R} \left[D(\mathbf{X}_r, Y_r) - \max_Y \{D(\mathbf{X}_r, Y) + \rho H(Y_r, Y)\} \right]^- \right\} \tag{6}$$

where $[\cdot]^-$ denotes the hinge function:

$$[x]^- = \max\{0, -x\}$$

and $\lambda > 0$ is the penalty parameter. By driving λ to zero, we penalize the constraint violations with increasing severity. It is the case in many practical applications that not all constraints can or should be satisfied. A more reasonable approach is to treat these constraints as soft and to have λ control the trade-off between margin maximization and constraint violation. The idea of a penalty function which balances margin and constraints has also been proposed in [2]. We differ however significantly with [2] in that our final objective function is differentiable and considers multiple competing sequences which can be encoded in a lattice.

The next task at hand is to obtain differentiable expressions for the constraints. First, we can replace the maximum in (6) by a soft-max upper bound leading to:

$$\max \left\{ \rho - \frac{1}{\lambda} \sum_{r=1}^{R} \left[D(\mathbf{X}_r, Y_r) - \log \sum_Y e^{D(\mathbf{X}_r, Y) + \rho H(Y_r, Y)} \right]^- \right\} \tag{7}$$

[1] The term "exact" means that there exists $\lambda^* > 0$ such that for any $\lambda \in (0, \lambda^*]$, any local solution of (5) is a local solution of (6).

Next, we notice that the resulting constraint terms are always strictly negative. Indeed,

$$\log \sum_Y e^{D(\mathbf{X}_r, Y) + \rho H(Y_r, Y)} > \log \sum_Y e^{D(\mathbf{X}_r, Y)} > D(\mathbf{X}_r, Y_r)$$

since the summation includes Y_r. It follows that we can rewrite (7) without the hinge function:

$$\max \left\{ \rho + \frac{1}{\lambda} \sum_{r=1}^{R} \left(D(\mathbf{X}_r, Y_r) - \log \sum_Y e^{D(\mathbf{X}_r, Y) + \rho H(Y_r, Y)} \right) \right\} \tag{8}$$

2.2. HMM parameter estimation

Let θ be a shorthand notation for all the HMM parameters: transition probabilies, Gaussian mixture component priors, means and covariances. We aim at finding θ^* which maximizes an objective function similar to (8) suitably formulated for HMM's. In the context of LVCSR, it makes sense to reason in terms of observation sequences and word sequences and to define discriminant functions of the form:

$$D_\theta(\mathbf{X}, W) := \log[p_\theta(\mathbf{X}|W)^\kappa P(W)], \tag{9}$$

with $P(W)$ being the language model probability of W which we assume to be constant for the purpose of this discussion, and κ being an acoustic scaling factor which will normally be the inverse of the language model power e.g. $\frac{1}{15}$. $p_\theta(\mathbf{X}|W)$ represents the likelihood of the acoustic sequence given the word sequence and depends on the HMM parameters θ. We define the margin between two word sequences W and W' as $\rho H(W, W')$ where:

$$H(W, W') := H(Y_W, Y_{W'}) \tag{10}$$

$Y_W, Y_{W'}$ are the Viterbi state sequences corresponding to W, W' and $H(Y_W, Y_{W'})$ is given by (2). By rewriting (8) in terms of word sequences and by plugging in (9) we get, after some manipulations:

$$(\theta^*, \rho^*) = \operatorname*{argmax}_{\theta,\rho} \left\{ \rho + \frac{1}{\lambda} \sum_{r=1}^{R} \log \frac{p_\theta(\mathbf{X}_r|W_r)^\kappa P(W_r)}{\sum_W p_\theta(\mathbf{X}_r|W)^\kappa P(W) e^{\rho H(W_r, W)}} \right\} \tag{11}$$

Lastly, we would like our objective function to be normalized by the number of frames. This can be achieved by setting

$$\lambda = \lambda' \sum_{r=1}^{R} T_r$$

where λ' is a constant which can be reused across tasks (in practice $\lambda' = 0.5$). Our constant λ' represents the proportion of the denominator in Equation 11 which we expect to consist of wrongly labeled frames. By fixing λ' in this way and maximizing Equation (11) over ρ, we believe we can choose an appropriate ρ in a way that is less dependent on the task.

2.3. Connection with boosted MMI

In [4], we introduced an HMM parameter estimation technique called boosted MMI (BMMI) which can be viewed as a variant of MMI where we increase (or boost) the likelihood of sentences which have more errors, thereby generating more confusable data. It was mentioned that BMMI can be construed as imposing a soft margin which is proportional to the number of errors in a hypothesized sentence. Using the notations introduced so far, the boosted MMI objective function is:

$$\theta^{BMMI} = \underset{\theta}{\operatorname{argmax}} \sum_{r=1}^{R} \log \frac{p_\theta(\mathbf{X}_r|W_r)^\kappa P(W_r)}{\sum_W p_\theta(\mathbf{X}_r|W)^\kappa P(W) e^{-\rho A(W_r,W)}} \tag{12}$$

with $A(W_r, W)$ denoting the accuracy of W with respect to W_r. The accuracy is expressed in terms of the number of correct phones in W as in MPE [7]. Comparing (11) and (12), we notice that the former includes the margin explicitly in the objective function whereas, for BMMI, ρ has to be tuned manually. The second difference is more pedantic and has to do with using a frame-based, state-level Hamming distance versus a negative phone-level accuracy. Indeed, phone-based and frame-based metrics have been found to produce similar results cf. [8] and negative accuracy versus (positive) distance leads to identical objective functions in the model parameters modulo a constant term.

If we ignore the margin term ρ, any form of optimization that works for (12) is obviously applicable to (11). To deal with the margin term, we follow the suggestion in [2], namely, we try multiple values of ρ and optimize the constraints term assuming a fixed ρ. In the end, we pick the pair (ρ^*, θ^*) which achieves the maximum. The hope is that the maximum is fairly broad in ρ so that only a small number of scale values will have to be tested.

The constraints term is optimized using the Extended Baum-Welch equations which can be found in many papers (see for instance [7, 4]). The only modification has to do with the forward-backward algorithm on the denominator lattice: for each word arc, we add to the acoustic log-likelihood ρ times the number of incorrectly labeled frames during the time span of that arc. This constitutes the contribution of the arc to the overall Hamming distance of the hypothesis which contains that arc.

3. EXPERIMENTS AND RESULTS

We report some experimental results on two large vocabulary broadcast news transcription tasks which differ in language (English versus Arabic), amount of training data (50 hours versus 1400 hours) and amount of speaker adaptation performed (speaker-independent versus VTLN, FMLLR and MLLR). Both systems have pentaphone acoustic cross-word context and cepstral mean (and variance) normalization. In this work, neither of the systems uses feature-space discriminative transformations.

The acoustic features for the English system are 40-dimensional vectors obtained via an LDA+MLLT projection of 9 consecutive spliced frames of 19-dimensional PLP features which are mean normalized on a per utterance basis. The baseline system has 2200 context-dependent HMM states and 50K Gaussians and is referred to as the EBN50 setup in [4] meaning that the numbers in Figure 2 are directly comparable with those from our previous paper.

The acoustic features for the Arabic system are 40-dimensional vectors obtained via an HDA+MLLT projection[9] of 9 consecutive spliced frames of 13-dimensional VTLN-warped PLP features which are mean and variance normalized on a per speaker basis. Additionally, the features are transformed through feature-space MLLR at both training and test time. The baseline system uses unvowelized (or graphemic) acoustic models with 5000 states and 400K Gaussians and was trained on 1400 hours of data as opposed to the ABN2300 setup from [4], where the models were trained on 2300 hours of data. Also, we report results on a more recent test-set (DEV'07 versus EVAL'06). More details about the Arabic system can be found in [10].

For both scenarios, the experimental setup is as follows. First, we decode the training data and generate denominator lattices with a unigram language model using the decoder and lattice generation procedure described in [11] (with a lattice n-best degree of 8). Next, we accumulate MMI-like statistics for the objective function (11) for various margin scale parameters ρ with per-frame canceled statistics. Finally, we perform an EBW update with I-smoothing to the previous iteration models. The statistics canceling method and the particular form of I-smoothing are described in [4]. We used four iterations of EBW in both scenarios for best results.

In Figure 1, we plot the objective function (11) for the two tasks. More precisely, we plot (11) multiplied by $\lambda' = 0.5$ so that for $\rho = 0$ we get the per-frame MMI objective function. Observe that, without the margin scale term (as in boosted MMI), the objective function would be monotonic decreasing in ρ reaching the maximum for $\rho = 0$ (which is the MMI case). This validates the use of the margin term in (11) to counter-balance the decrease of the constraints term as a function of ρ.

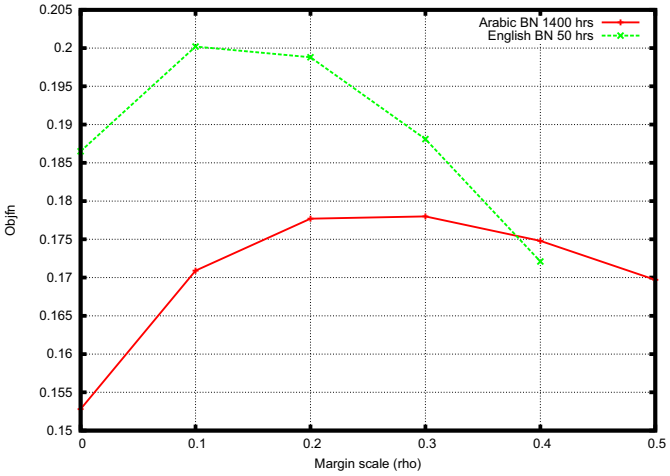

Figure 1: Objective functions for the English and Arabic BN systems.

In Figure 2, we present the results for the English BN system on the RT'04 testset which comprises 4 hours of speech. The best results were obtained for $\rho = 0.2$ with a broad maximum range for $\rho \in [0.1, 0.3]$. This corresponds roughly to the region of the maximum of the large margin objective function depicted in Figure 1. The lowest word error rate achieved is 21.2% and the corresponding ML-trained baseline has a WER of 25.3%.

Additionally, in Table 1, we compare the performances of various discriminative training algorithms on two different testsets (DEV'04f and RT'04). As can be seen, MPE outperforms MMI and is outperformed by the proposed large margin technique which is in line with our previous findings [4].

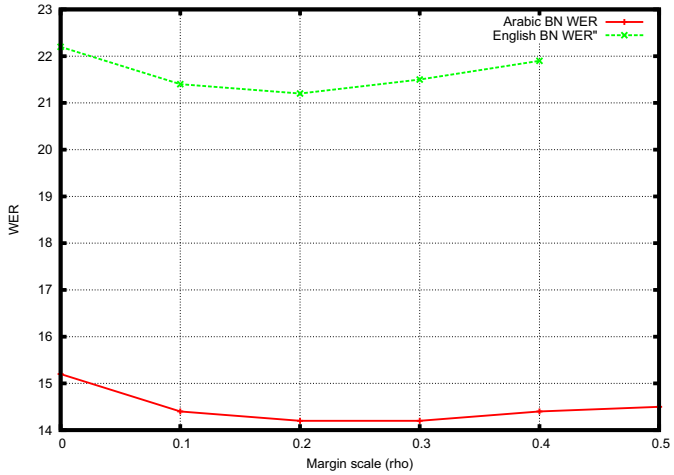

Figure 2: Word error rates for the English and Arabic BN systems (on RT'04 and DEV'07 respectively).

Training criterion	DEV04f	RT04
Maximum Likelihood	28.7%	25.3%
Maximum Mutual Information	25.3%	22.2%
Minimum Phone Error	24.7%	21.9%
Large Margin (ρ=0.2)	24.2%	21.2%

Table 1: Word error rates for different discriminative training criteria on English BN.

A similar picture can be encountered on the Arabic setup, where again, the best results are obtained for $\rho = 0.2$ with a broad optimum range for $\rho \in [0.1, 0.3]$ which corresponds to the optimum region of the objective function. The results are presented on the DEV'07 testset which has 3 hours of speech. The lowest word error rate obtained is 14.2% and the corresponding ML-trained baseline has a WER of 17.1%.

4. CONCLUSION

The main contribution of this work is to show the connection between boosted MMI and large margin training in the sense of [3]. As a side-effect, we have constructed an objective function which attains its maximum for a margin parameter which also achieves the lowest word error rate. The objective function arises from turning a constrained optimization problem into a penalty function maximization problem. This penalty function is a weighted combination of the margin scale and the constraints violation part and can be efficiently optimized using the traditional framework of MMI training involving lattices and Extended Baum-Welch updates. While the experimental results have focused here only on model parameter estimation, it is straightforward to extend these ideas to feature-space discriminative training.

5. REFERENCES

[1] C. Liu, H. Jiang, and L. Rigazio, "Recent improvements on maximum relative margin estimation of HMMs for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2006.

[2] J. Li, M. Yuan, and C.-H. Lee, "Soft margin estimation of Hidden Markov Model parameters," in *Interspeech-2006*, 2006.

[3] F. Sha and L. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by Hidden Markov Models," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2007.

[4] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2008.

[5] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Neural Information Processing Systems - NIPS*, 2003.

[6] J. Nocedal and S. J. Wright, "Numerical optimization," in *Springer Series in Operations Research*, 1999.

[7] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2002.

[8] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2007.

[9] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2000.

[10] H. Soltau, G. Saon, B. Kingsbury, J. Kuo, L. Mangu, D. Povey, and G. Zweig, "The IBM 2006 GALE Arabic ASR system," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2007.

[11] G. Saon, D. Povey, and G. Zweig, "Anatomy of an extremely fast LVCSR decoder," in *Interspeech-05*, 2005.