# LATTICE-BASED OPTIMIZATION OF SEQUENCE CLASSIFICATION CRITERIA FOR NEURAL-NETWORK ACOUSTIC MODELING

*Brian Kingsbury*

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
bedk@us.ibm.com

**ABSTRACT**

Acoustic models used in hidden Markov model/neural-network (HMM/NN) speech recognition systems are usually trained with a frame-based cross-entropy error criterion. In contrast, Gaussian mixture HMM systems are discriminatively trained using sequence-based criteria, such as minimum phone error or maximum mutual information, that are more directly related to speech recognition accuracy. This paper demonstrates that neural-network acoustic models can be trained with sequence classification criteria using exactly the same lattice-based methods that have been developed for Gaussian mixture HMMs, and that using a sequence classification criterion in training leads to considerably better performance. A neural network acoustic model with 153K weights trained on 50 hours of broadcast news has a word error rate of 34.0% on the `rt04` English broadcast news test set. When this model is trained with the state-level minimum Bayes risk criterion, the `rt04` word error rate is 27.7%.

***Index Terms***— speech recognition, neural networks, discriminative training

## 1. INTRODUCTION

There are a number of arguments why neural networks are a useful alternative to Gaussian mixture models (GMMs) for acoustic modeling in speech recognition. First, neural networks make minimal assumptions about the distribution of the input features, allowing for significant flexibility in front-end feature extraction. Second, evidence from multiple feature streams can be easily combined in a single HMM/NN recognition system because neural networks can estimate posterior probabilities. Third, neural network training criteria are discriminative, while the maximum likelihood criterion commonly used for GMM acoustic models is not. With the recent development of discriminative training algorithms for GMM acoustic models [1, 2, 3], this third argument is less compelling than it used to be. In fact, GMMs may enjoy an advantage because the criteria used for discriminatively training them are based on sequence classification, while the most common criterion for training neural network acoustic models is based on frame classification. Criteria based on sequence classification are more closely related to word error rate than criteria based on frame classification, and should provide better speech recognition performance.

Algorithms for training HMM/NN systems to discriminate between sequences have been proposed before. Alphanets [4] view the HMM as a recurrent neural network in which the maximum mutual information (MMI) criterion is optimized through gradient descent, and backpropagation in the HMM takes exactly the same form as the backward pass in EM training. The tasks to which Alphanets were applied were small enough that the contribution of the competing hypotheses to the gradient could be computed through a forward-backward pass using a phone-loop grammar. The REMAP algorithm [5] maximizes the a-posteriori probability of the reference word sequence, and relies upon sum-to-one constraints to penalize competing, incorrect hypotheses. Other studies [6, 7] demonstrate the effectiveness of global normalization and conditional maximum likelihood (CML) training (equivalent to MMI for a fixed language model) on TIMIT phone recognition and broad class recognition.

The work presented in this paper extends the prior work in two ways. First, it uses word lattices to compactly represent the reference and the competing hypotheses, making it possible to train on large-vocabulary tasks with large training sets. Second, it shows that criteria other than MMI, such as state-level minimum Bayes risk, can be optimized. The rest of the paper is organized as follows. Section 2 reviews the use of the frame-based cross-entropy criterion in neural network training. Section 3 shows how sequence classification criteria may be optimized using the lattice-based framework developed for discriminatively training GMM acoustic models. Section 4 describes the experimental conditions, baseline results, and discriminative training results. Section 5 summarizes the findings of this study and discusses future work.

## 2. THE CROSS-ENTROPY CRITERION

Let the training set be a collection of acoustic feature sequences, $\mathbf{X}_r$, and corresponding word-level transcripts, $W_r$. For each training sample $(\mathbf{X}_r, W_r)$, there is also a label sequence, $\hat{\mathbf{Y}}_r$, of the same length as $\mathbf{X}_r$, which specifies for each time $t$ a multinomial distribution, $\hat{\mathbf{y}}_{rt}$ over $N$ physical states. Typically, the labels are "hard," $\hat{y}_{rt}(i) \in \{0, 1\}, i = 1, \cdots, N$, and are defined either by manual labeling (e.g., TIMIT) or through forced alignment. In some cases the labels may be "soft," $\hat{y}_{rt}(i) \in [0, 1], i = 1, \cdots, N$, having been derived from a forward-backward pass over a reference transcript or lattice. $T_r$ denotes the length of the $r$-th sequence in the training set.

The cross-entropy criterion is

$$\mathcal{L}_{XENT}(\theta) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{i=1}^{N} \hat{y}_{rt}(i) \log \frac{\hat{y}_{rt}(i)}{y_{rt}(i)}, \quad (1)$$

where $\theta$ denotes the parameters of the neural network (weights and biases for all layers) and $y_{rt}(i)$ is the network output for physical state $i$ at time $t$ in sample $r$.

In training, an error backpropagation procedure adjusts $\theta$ to minimize $\mathcal{L}_{XENT}(\theta)$. When the cross-entropy criterion is used with a network having a softmax output nonlinearity,

$$y_{rt}(i) = \frac{e^{a_{rt}(i)}}{\sum_{j=1}^{N} e^{a_{rt}(j)}}, \quad (2)$$

where the inputs to the softmax (activations) are denoted $\mathbf{a}_{rt}$, gradient-descent training can be based on a convenient expression for the derivative of the loss with respect to the activations:

$$\frac{\partial \mathcal{L}_{XENT}(\theta)}{\partial a_{rt}(i)} = y_{rt}(i) - \hat{y}_{rt}(i). \quad (3)$$

## 3. SEQUENCE CLASSIFICATION CRITERIA

There has been considerable interest in discriminative training methods for Gaussian mixture HMMs in recent years. One approach to discriminative training, which has been quite successful for large-vocabulary tasks, relies on lattices to compactly represent the space of competing hypotheses [1] and uses extended Baum-Welch (EBW) updates with appropriate smoothing to train model parameters. While this framework was originally developed for the MMI criterion [8], other criteria have been developed around it, including minimum phone error (MPE) [3], minimum Bayes risk (MBR) [9, 10], and a maximum-margin criterion [11].

Let $\mathcal{L}_{SEQ}(\theta)$ be any sequence classification criterion (e.g., MMI, MPE or MBR). Note that in some cases these criteria are formulated as objective functions to be maximized instead of loss functions to be minimized. In such cases, a loss function is derived by multiplying the original objective function by $-1$. The expected occupancies $\gamma_{rt}^{NUM}(i)$ and $\gamma_{rt}^{DEN}(i)$ for each physical state required by the EBW updates are computed with forward-backward passes over the numerator and denominator lattices, respectively. Recall that the numerator lattices represent the reference transcriptions and the denominator lattices represent competing hypotheses.

These expected occupancies are also related to the gradient of the loss with respect to state log-likelihoods [3]:

$$\frac{\partial \mathcal{L}_{SEQ}}{\partial l_{rt}(i)} = \kappa\big(\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)\big), \quad (4)$$

where $l_{rt}(i)$ is the log-likelihood of physical state $i$ at time $t$ in sample $r$ and $\kappa$ is the acoustic scaling used in the lattice generation and forward-backward passes to improve generalization. In an HMM/neural network hybrid, $l_{rt}(i) = \log y_{rt}(i) - \log p(i)$, where $p(i)$ is the prior probability of state $i$, computed from the training set. By the chain rule,

$$\frac{\partial \mathcal{L}_{SEQ}}{\partial y_{rt}(i)} = \kappa \frac{\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)}{y_{rt}(i)}. \quad (5)$$

Applying the chain rule and some algebra, derivatives with respect to the softmax activations, like Equation (3), are

$$\frac{\partial \mathcal{L}_{SEQ}}{\partial a_{rt}(i)} = \kappa\big(\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)\big). \quad (6)$$

To a factor of $\kappa$, Equations (3) and (6) are nearly identical, with the lattice-based denominator and numerator counts, $\gamma_{rt}^{DEN}(i)$ and $\gamma_{rt}^{NUM}(i)$, in Equation (6) taking the place of the estimated and reference posteriors, $y_{rt}(i)$ and $\hat{y}_{rt}(i)$, in Equation (3). This provides a simple recipe for training neural-network acoustic models using any of the sequence classification criteria developed for Gaussian mixture HMMs in the lattice-based EBW framework: the gradient with respect to the cross-entropy criterion is replaced with the gradient with respect to the sequence-classification criterion (Equation (5) or Equation (6)), and backpropagation is run as usual.

Equation (6) is not new [4, 7]. What is new is the adoption of lattices to represent the reference and competing hypotheses, and the use of sequence classification criteria other than MMI and CML for neural-network training.

## 4. EXPERIMENTS

Sequence classification training for neural network acoustic models is evaluated on an English broadcast news transcription task by comparing the performance of models trained with the frame-based, cross-entropy error criterion to that of models trained with the state-level minimum Bayes risk (sMBR) criterion [9, 10, 12],

$$\mathcal{L}_{sMBR}(\theta) = \sum_{r=1}^{R} \frac{\sum_{W \in \mathcal{W}_r} P(\mathbf{X}_r|W,\theta)^{\kappa} P(W) d(\mathbf{Y}, \hat{\mathbf{Y}}_r)}{\sum_{W \in \mathcal{W}_r} P(\mathbf{X}_r|W,\theta)^{\kappa} P(W)}, \quad (7)$$

where $\mathcal{W}_r$ is the set of word hypotheses represented by the denominator lattices for sample $r$, $\mathbf{Y}$ is the label sequence for word hypothesis $W$, and $\hat{\mathbf{Y}}_r$ is the label sequence for the reference. $d(\mathbf{Y}, \hat{\mathbf{Y}}_r)$ is the Hamming distance between the label sequences if the reference labels are hard [10]; otherwise, it is defined as in [12]. The sMBR criterion was chosen over MMI and related criteria [11] because it had the best performance in pilot experiments. Results are also provided for Gaussian mixture acoustic models as a point of reference. The acoustic model training set comprises 50 hours of data from the 1996 and 1997 English Broadcast News Speech corpora (LDC97S44 and LDC98S71), and was created by selecting entire shows at random. The EARS Dev-04f set (`dev04f`), a collection of 3 hours of audio from 6 shows collected in November 2003, is used for system development. The EARS RT-04 test set (`rt04`), a collection of 6 hours of audio from 12 shows collected in December 2003, is used for system evaluation.

The acoustic features are 19-dimensional PLP features with speaker-based mean and variance normalization. Aside from the normalization, the features are speaker-independent. For the training data, speaker labels are provided in the reference transcripts, while for test data the "speakers" are actually clusters of segments produced by an automatic diarization system [13]. Phones are modeled as three-state, left-to-right HMMs with no skip states. States are quinphone context-dependent, except for silence states, which are context-independent. The decision trees that cluster contexts into states are trained to maximize likelihood gain with single-Gaussian, diagonal covariance models, modeling 40-dimensional features computed from a linear discriminant analysis (LDA) projection of vectors computed by splicing 9 frames of normalized PLP features. Recognition is done using a dynamic decoder similar to the one described in [14], but which uses a statically compiled and minimized word network, allowing for multiple pronunciations and contexts spanning more than one word. The language model used for decoding is a 54M n-gram, interpolated backoff model trained on a collection of 335M words from the following sources: 1996 CSR Hub4 Language Model data (LDC98T31), EARS BN03 closed captions, GALE Phase 2 Distillation GNG Evaluation Supplemental Multilingual data (LDC2007E02), Hub4 acoustic model training transcripts (LDC97T22 and LDC98T28), TDT4 closed captions (LDC2005T16), TDT4 newswire (LDC2005T16), GALE Broadcast Conversations (LDC2005E82, LDC2006E33, LDC2006E84, LDC2006E91, LDC2007E05, and LDC2007E45), and GALE Broadcast News (same catalog numbers as GALE Broadcast Conversations). The source-specific LMs in the interpolation are 4-gram models with modified Kneser-Ney smoothing. The recognition lexicon contains 84K word tokens, with an average of 1.08 pronunciation variants per word. Where possible, pronunciations were based on PRONLEX (LDC97L20).

The neural network acoustic models are multilayer perceptrons with a single hidden layer and full connectivity between layers. The

| model | # states | WER |
|---|---|---|
| NN | 126 | 39.9 |
| | 192 | 36.8 |
| | 256 | 36.5 |
| | 384 | 35.7 |
| | 512 | 35.7 |
| GMM | 126 | 44.9 |
| | 192 | 41.9 |
| | 256 | 41.1 |
| | 384 | 40.4 |
| | 512 | 39.2 |
| | 640 | 39.0 |
| | 768 | 38.7 |
| | 896 | 38.8 |

**Table 1**. dev04f word error rate (WER, stated as a percentage) as a function of the number of context-dependent states used in the model, for both neural networks and GMMs. Note that 126 states corresponds to context-independent modeling.

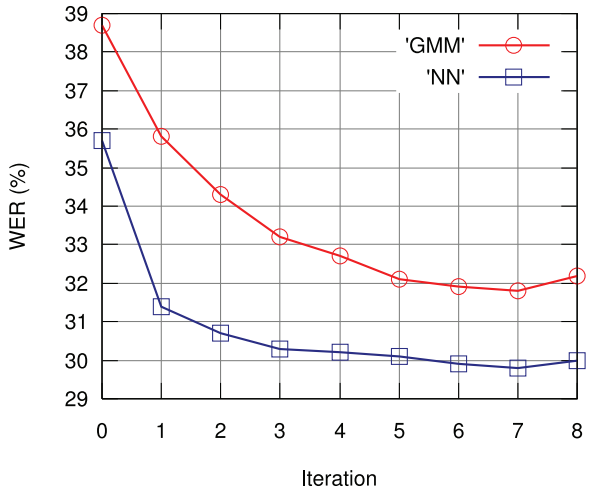


**Fig. 1**. dev04f performance as a function of discriminative training iteration for neural networks and GMMs.

input is a sliding window of 9 frames on the normalized PLP features. The hidden units use logistic nonlinearities, while the output layer is a softmax nonlinearity with the output units corresponding to context-dependent HMM states. This network structure was chosen because it is fairly standard in neural-network acoustic modeling [15, 16]; any other network for which error backpropagation training is possible, including recurrent neural networks, could also be used. Training with both criteria, cross-entropy and sMBR, is done using on-line stochastic gradient descent, with a weight update after the presentation of each utterance. The order in which utterances are visited is randomized. After each pass over the training data, the loss is measured on a held-out set (a sample of 10% of the full training set, with selection of entire shows). If a pass over the data increases the held-out loss, the weights revert to their previous values and, if additional training is to be done, the step size is multiplied by 0.5 [15]. Training halts after the step size is reduced five times. The initial step size is chosen to optimize word error rate on dev04f. The networks contain roughly 153K weights: a small size that allows for fast-turnaround experiments.

The GMMs use 40-dimensional features computed from an LDA+MLLT (LDA followed by a maximum-likelihood linear transform) projection of 9 spliced frames of normalized PLP features. The LDA discriminates between context-dependent states, and semi-tied covariance updates using a single, global class are interleaved with standard HMM updates to diagonalize the original LDA projection. Baseline HMM training is maximum-likelihood (ML), using a fixed, state-level alignment of the training data. The GMMs use 2048 mixture components, for a total of 165K trainable parameters.

The setup for discriminative training is quite similar to that in [12]. For each utterance, a lattice with fixed state alignments is used to represent the set of competing hypotheses used in the loss function optimization. Accumulation of denominator counts is done through forward-backward passes over the phone-marked training lattices [2]. The reference for each utterance is a Viterbi alignment of the reference word transcript. Training of the neural network models uses error backpropagation as described above. The GMM models are trained using EBW updates with $E = 2.0$ and cancellation of the numerator and denominator statistics [11]. I-smoothing is performed with $\tau = 500$, using the models from the previous training iteration as a prior [11]. For both the neural network and GMM models, the number of iterations of discriminative training is chosen to optimize word error rate on dev04f. Note that the neural network and GMM discriminative training procedures use exactly the same routines for the collection of denominator counts.

### 4.1. Baseline results

The first experiments determined the best number of context-dependent states to use in the neural network and GMM acoustic models. Results are summarized in Table 1. The minimum number of states, 126, is determined by the use of 42 phones and 3-state HMMs. As the number of context-dependent states increases, the size of the hidden layer in the neural networks decreases and the number of Gaussian mixtures per state in the GMMs decreases, to keep the number of trainable parameters constant. Both neural networks and GMMs show a similar pattern: performance improves as the number of states increases, until a plateau is reached. Neural network models reach this plateau earlier than GMMs, possibly because the discriminative training criterion used with the neural networks compensates for incorrect modeling assumptions.

### 4.2. Discriminative training results

In the second set of experiments, a single neural network model and a single GMM model are trained to optimize the state-level minimum Bayes risk criterion. The neural network with 384 states and the GMM with 768 states were selected as models that had the best tradeoff between dev04f word error rate and model complexity. The results of these experiments are summarized in Figure 1. The recognition performance for both models improves with discriminative training, with the neural network improving from 35.7% WER to 29.8% and the GMM improving from 38.7% WER to 31.8% on dev04f. It is not surprising that the GMM improves more than the neural network because the GMM baseline is trained with the maximum-likelihood criterion, while the neural network baseline is trained with a frame-based, discriminative criterion. Both models appear to be overtrained by the eighth iteration of sMBR training. Because the sMBR training entails a realignment of the training data, further cross-entropy training of the neural network using the new

| model | criterion | WER |
|---|---|---|
| NN | XENT | 34.0 |
| | sMBR | 27.7 |
| GMM | ML | 36.7 |
| | sMBR | 28.9 |

**Table 2**. rt04 word error rate (WER) for neural networks and GMMs with baseline and sequence classification (sMBR) training.

alignments was also tried. The best dev04f improvement from this additional training was only 0.2%, from 35.7% WER to 35.5%. The results in Figure 1 are obtained with acoustic scaling factors that were optimized on the baseline systems. After sMBR training, the optimal scaling for both models was reduced. The best scaling for the neural network dropped from 0.14 to 0.11, improving dev04f performance from 29.8% WER to 29.1%, and the best scaling for the GMM dropped from 0.07 to 0.06, improving dev04f performance from 31.8% WER to 31.4%.

Finally, the baseline and sMBR-trained neural network and GMM acoustic models were tested on rt04, using the best acoustic weights from the dev04f experiments. The results, which are provided in Table 2, are consistent with the results obtained on dev04f. Training with a sequence classification criterion (sMBR) greatly improves performance compared to either the frame-based, cross-entropy criterion used with neural networks or the maximum likelihood criterion used with GMMs. The GMM enjoys a larger absolute improvement in performance with sMBR training than the neural network because the GMM baseline is not discriminatively trained, while the neural network baseline is. The sMBR-trained neural network outperforms the sMBR-trained GMM, although this comes at the cost of significantly more training time.

The small model size used in these experiments leads to higher word error rates than are achieved by the best systems for this task. A standard HMM system with 50K mixture components and the same features has an rt04 word error rate of 25.3% [11], which is better than any of the results reported here, but uses 24 times as many parameters. The neural network is also underparameterized, using roughly 120 frames of training data per weight when 10–40 frames/weight is preferable [16]. It is not clear if the NN systems would continue to outperform the GMM systems as the number of parameters is increased.

## 5. CONCLUSIONS

Neural network acoustic models can be trained with sequence classification criteria instead of frame classification criteria using exactly the same lattice-based framework developed for discriminatively training GMM acoustic models, and the use of a sequence-based criterion (sMBR) in training leads to a substantial improvement in word error rate on a large vocabulary, continuous speech recognition task. Directions for future work include evaluation of other sequence classification criteria [11] for neural network training, comparison of frame-based training and sequence-based training in tandem systems [17], and development of methods for neural network training that scale to larger networks and data sets.

## 6. ACKNOWLEDGMENTS

Thanks to George Saon for the lattice generation and count accumulation code used in this work. This work was partially supported by the Defense Advanced Research Projects Agency under contract No. HR0011-06-2-0001.

## 7. REFERENCES

[1] Y. Normandin, R. Lacouture, and R. Cardin, "MMIE training for large vocabulary continuous speech recognition," in *Proc. ICSLP*, 1994.

[2] D. Povey and P. C. Woodland, "Improved discriminative training techniques for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 2001.

[3] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002.

[4] J. S. Bridle and L. Dodd, "An Alphanet approach to optimising input transformations for continuous speech recognition," in *Proc. ICASSP*, 1991.

[5] Y. Konig, *REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities in Transition-based Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1996.

[6] F. T. Johansen, "A comparison of hybrid HMM architectures using global discriminative training," in *Proc. ICSLP*, 1996.

[7] A. Krogh and S. K. Riis, "Hidden neural networks," *Neural Computation*, vol. 11, no. 2, pp. 541–563, 1999.

[8] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP*, 1986.

[9] J. Kaiser, B. Horvat, and Z. Kačič, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proc. ICSLP*, 2000.

[10] M. Gibson and T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in *Proc. Interspeech*, 2006.

[11] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008.

[12] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *Proc. ICASSP*, 2007.

[13] H. Soltau, G. Saon, B. Kingsbury, J. Kuo, L. Mangu, D. Povey, and G. Zweig, "The IBM 2006 GALE Arabic ASR system," in *Proc. ICASSP*, 2006.

[14] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Proc. ASRU*, 2001.

[15] N. Morgan and H. Bourlard, "Neural networks for statistical recognition of speech," *Proc. IEEE*, vol. 83, no. 5, pp. 742–772, May 1995.

[16] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1999.

[17] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000.