RECENT IMPROVEMENT ON MAXIMUM RELATIVE MARGIN ESTIMATION OF HMMS FOR SPEECH RECOGNITION

Chaojun Liu[†]*, Hui Jiang*[‡]*, Luca Rigazio*[†]

[†] Panasonic Digital Networking Laboratory, Panasonic R&D Company of America
550 S. Winchester Blvd., Suite 300, San Jose, CA 95128, USA
[‡] Department of Computer Science and Engineering, York University
4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA
Email: {chaojunl,rigazio}@research.panasonic.com hj@cs.yorku.ca

ABSTRACT

Our previous study on maximum relative margin estimation (MRME) of HMM ([7, 8]) demonstrated its advantage over the standard minimum classification error (MCE) training. In this paper, we report our recent improvement on MRME. Specifically, two novel approaches are proposed to handle recognition errors in training sets for the MRME. One is a new training criterion based on a combination of MRME and MCE objective functions. The other approach proposes to remove a strong constraint in the original MRME algorithm, so that MRME algorithm can be applied to all training data as opposed to only correctly recognized data in the original MRME approach. Both new approaches can take advantage of more training data during the large margin training and can bootstrap directly from MLE models without a separate MCE training step. Improvement on recognition accuracy has been achieved on a speaker independent connected digit strings recognition task using the TIDIGITS database.

1. INTRODUCTION

In automatic speech recognition (ASR), discriminative training has been extensively studied over the past decade and it has been proved quite effective to improve performance over the traditional maximum likelihood (ML) method for HMM-based speech recognition systems. Two popular discriminative training methods are minimum classification error (MCE) training and maximum mutual information (MMI) training. But as reported by many researchers, all these discriminative training methods have poor generalization capability. In other words, they can significantly improve HMMs and leads to a dramatic error reduction on training data but such a significant performance gain can hardly be maintained or generalized in any unseen test set. Usually only a marginal gain can be achieved over the ML method in a new data set, especially for large-scale tasks.

Motivated by some recent advances in machine learning about large margin classifier, recently we proposed two novel training methods, namely large margin estimation (LME) [5, 6] and maximum relative margin estimation (MRME) [7, 8] for speech recognition. In LME or MRME, HMM parameters are estimated to maximize the minimum margin among all training utterances. Significant accuracy improvement over MCE has been achieved on both isolated-word and continuous digits recognition task. Both LME and MRME only use correctly recognized training data to estimate HMM models based on the principle of large margin and they usually ignore all mis-recognized data in the training set. They typically rely on a prior MCE training step to reduce the total number of mis-recognized utterances in the training set. As a result, they usually bootstrap from the MCE-trained model [5, 6, 7, 8] and therefore the overall training time is longer than the MCE-only method. Moreover, as we extend the string-level MRME [8] to large vocabulary continuous speech recognition (LVCSR) tasks, there is usually only a very small percentage of the training utterances that can be correctly recognized in string level even after the separate MCE training stage. The benefit of the original LME or MRME method may be greatly limited due to lack of applicable training utterances.

In this paper, we propose two different approaches to handle mis-recognized utterances in training sets in MRME to further improve the MRME method. The first approach is to minimize a new objective function, which is a weighted linear combination of the original MRME objective function and an MCE objective function. This new training approach basically applies MRME training and MCE training at the same time. The second approach proposes to remove a strong constraint in the original LME and MRME methods. It changes the definition of support token set in LME and MRME so as to include misrecognized training utterances as well. Now the MRME objective function is optimized over all training utterances as opposed to only correctly recognized utterances in the original MRME approach. Both new approaches can bootstrap directly from MLE models and does not require a seperate MCE training any more. They have achieved improvement on recognition accuracy over the original MRME approach on TIDIGITS corpus. The second approach has achieved a string error rate as low as 0.76% on the standard TIDIGITS test set, which is, to our knowledge, the best result that has ever been reported in this task.

The remainder of this paper is organized as follows. First of all, we briefly summarize the MRME formulation for HMMs in speech recognition in section 2. Next, in section 3, we will present the first new approach based on the new objective function as well as a gradient descent method using the new training criterion. Then, in section 4 we will propose our second new approach, which is based on a new definition of support token set. Experimental results on the TIDIGITS task are reported and discussed in section 5. Finally, the paper is concluded with our findings and future works in Section 6.

2. MAXIMUM RELATIVE MARGIN ESTIMATION

In ASR, given any continuous speech utterance X, a speech recognizer will choose the string $\hat{S}$ as output based on the MAP decision rule as follows (without loss of generalization, here we only give the

formulation of string model based MRME for continuous speech, where S represent the concatenated model for a string. But it can be applied to other cases as well, for example isolated word case as in [7]):

$$\begin{aligned}\hat{S} &= \arg\max_S \ p(S|X) = \arg\max_S \ p(S)\cdot p(X|S) \\ &= \arg\max_S \ p(S)\cdot p(X|\lambda_S) = \arg\max_S F(X|\lambda_S) \end{aligned} \quad (1)$$

where λ_S denotes the HMM representing the string S and $F(X|\lambda_S)$ is called discriminant function. As we are only interested in HMM λ_S, assume $p(S)$ is fixed.

In MRME, the HMM parameters are estimated in such a way that the decision boundary of the HMM-based classifier achieves the maximum classification margin as in other large margin classifiers, such as support vector machine. According to the statistical learning theory [10], a large margin classifier generally yields good generalization capability when classifying new unseen data. In [3, 6], the margin is defined as the minimum difference of log likelihood values between the true model and all of its competing models. However, the margin defined in this way has been shown to be unbounded. In MRME, we instead define the margin as a relative margin [7, 8], which is bounded by definition.

For a speech utterance X_i in the training data set $\mathcal{D} = \{X_1, X_2, \cdots, X_N\}$, assume its true transcript as $\{S_1^T, S_2^T, \cdots, S_N^T\}$. The relative separation margin for X_i is defined as:

$$\begin{aligned} d_1(X_i) &= \min_{S_j\in\Omega \ S_j\neq S_i^T} \left[\frac{F(X_i|\lambda_{S_i^T}) - F(X_i|\lambda_{S_j})}{F(X_i|\lambda_{S_i^T})}\right] \\ &= \min_{S_j\in\Omega \ S_j\neq S_i^T} \left[1 - \frac{F(X_i|\lambda_{S_j})}{F(X_i|\lambda_{S_i^T})}\right] \end{aligned} \quad (2)$$

where Ω denotes the set of all possible strings. As in continuous speech recognition, there is enormous or even infinite number of possible strings, which makes it impossible to enumerate all of them. In [8] we proposed to use N-best strings from Viterbi decoding, which represent the most confusable strings with the true string, to approximate the admissible set Ω for each training utterance. As Ω now becomes different for each training utterance, we denote it as Ω_i for X_i.

As observed in most ASR systems, the values of discriminative functions, $F(\cdot)$, determined by the likelihood (NOT log-likelihood) values of HMMs and language models in eq.(1), are within the range of $[0, 1]$. For any training utterance, X_i, which is correctly recognized by the current models based on the rule in eq.(1), $1 > F(X_i|\lambda_{S_i^T}) > F(X_i|\lambda_{S_j}) > 0$ holds for any $S_j \in \Omega$ and $S_j \neq S_i^T$. As a result, we have $0 < \left[1 - \frac{F(X_i|\lambda_{S_j})}{F(X_i|\lambda_{S_i^T})}\right] < 1$. Therefore $d_1(X_i)$ is bounded between $[0, 1]$ for any correctly recognized training utterances.

In practice, usually log-likelihood functions rather than the original likelihood functions are used in most HMM-based ASR systems. Let's denote discriminant function in the logarithm scale as $\mathcal{F}(X_i)$ for each $F(X_i)$, i.e., $\mathcal{F}(X_i) = \log\ F(X_i)$. Based on logarithm discriminant functions, we define the relative margin for any training utterance, X_i, as follows:

$$\begin{aligned} d_2(X_i) &= \min_{S_j\in\Omega_i \ S_j\neq S_i^T} \left[\frac{\mathcal{F}(X_i|\lambda_{S_j}) - \mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})}\right] \\ &= \min_{S_j\in\Omega_i \ S_j\neq S_i^T} \left[1 - \frac{\mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})}\right] \end{aligned} \quad (3)$$

As discussed above, $F(\cdot) < 1$, so $\mathcal{F}(\cdot) < 0$. So for any correctly recognized utterance X_i, $\mathcal{F}(X_i|\lambda_{S_j}) < \mathcal{F}(X_i|\lambda_{S_i^T}) < 0$ for any $S_j \in \Omega_i$ and $S_j \neq S_i^T$. As a result, $0 < \left[1 - \frac{\mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})}\right] < 1$ Therefore the relative margin $d_2(X_i)$ is also bounded between $[0, 1]$ for any correctly recognized training utterance X_i.

Analogous to support vector machine, we define a subset of training utterances from $\mathcal{D}$ for model estimation, denoted as the support token set, $\mathcal{SV}$.

$$\mathcal{SV} = \{X_i \mid X_i \in \mathcal{D} \text{ and } 0 \le d_2(X_i) \le \gamma\} \quad (4)$$

where $\gamma > 0$ is a pre-set positive number. Each X_i in $\mathcal{SV}$ is called a support token.

To achieve better generalization power, it is desirable to adjust decision boundaries, which are implicitly determined by all models, through optimizing all HMM parameters to make all support tokens in $\mathcal{SV}$ as far from the decision boundaries as possible, which will result in a robust classifier with better generalization capability. This idea leads to estimating HMM models based on the criterion of maximizing the minimum relative margin, either $d_1(X_i)$ or $d_2(X_i)$, of all support tokens in $\mathcal{SV}$. This estimation method is called maximum relative margin estimation (MRME). It can be formulated as a minimax optimization problem:

$$\begin{aligned} \tilde{\Lambda} &= \arg\max_{\Lambda} \min_{X_i\in\mathcal{SV}} \ d_2(X_i) \\ &= \arg\min_{\Lambda} \max_{X_i\in\mathcal{SV},\, S_j\in\Omega_i, S_j\neq S_i^T} \left[\frac{\mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})} - 1\right] \end{aligned} \quad (5)$$

where Λ denotes the parameter set of all CDHMMs in the classifier.

To solve the minimax optimization problem, a gradient descent method is used. First, we define a differentiable objective function, $Q_1(\Lambda)$, to approximate the maximization in eq.(5) as follows.

$$\begin{aligned} &\max_{X_i\in\mathcal{SV},\, S_j\in\Omega_i,\, S_j\neq S_i^T} \left[\frac{\mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})} - 1\right] \\ &\approx \log\left\{\sum_{X_i\in\mathcal{SV},\, S_j\in\Omega_i,\, S_j\neq S_i^T} \exp\left[\eta\cdot e_2(X_i, \lambda_{S_j}, \lambda_{S_i^T})\right]\right\}^{1/\eta} \\ &= Q_1(\Lambda) \end{aligned} \quad (6)$$

where $\eta > 1$ and

$$e_2(X_i, \lambda_{S_j}, \lambda_{S_i^T}) = \frac{\mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})} - 1.$$

As $\eta \to \infty$, $Q_1(\Lambda)$ will approach the maximization in eq.(5).

Next, a gradient descent method is used to minimize $Q_1(\Lambda)$ to solve eq.(5) in an approximate way.

3. NEW TRAINING CRITERION BASED ON COMBINATION OF MRME AND MCE

As we can see in eq.(4), the MRME algorithm uses only support tokens for training, which are correctly recognized utterances. It relies on a prior MCE step to handle those negative (or mis-recognized) tokens. Thus, the original MRME algorithm usually uses MCE models to bootstrap the training. In other word, a separate step of MCE

training is performed first to reduce the total number of negative tokens as much as possible and then the generated model is used as the seed model for the MRME training. In this way, the overall time for training an MRME model is longer than training an MCE model. On the other hand, separation of MRME and MCE training may not be optimal in model estimation.

In this paper, we propose an alternative way to handle recognition errors in the training data. Instead of requiring a seperate MCE training before MRME training, we combine the objetive function for MCE and the objective function for MRME, and minimize the new combined objective function. In that way, both positive tokens and negative tokens are taken care at the same time, that is, the relative margins for positive tokens are maximized and the number of recognition errors is minimized. This idea is similar to the combined objective function of support vector machine (SVM) in non-separable cases, where some slack variables are introduced to represent classification errors.

Based on the the current model set Λ, we first identify all misrecognized utterannces, which have negative margins (or equivalently, negative relative margins), as the error set $\mathcal{E}$. Let's define the margin $d(X_i)$ as

$$d(X_i) = \min_{S_j \in \Omega_i \ S_j \neq S_i^T} \left[\mathcal{F}(X_i|\lambda_{S_i^T}) - \mathcal{F}(X_i|\lambda_{S_j}) \right]$$

so the error set is given as follows.

$$\mathcal{E} = \{X_i \mid X_i \in \mathcal{D} \text{ and } d(X_i) \leq 0\} \tag{7}$$

For utterances in $\mathcal{E}$, following the MCE training, we optimize HMM model parameters, Λ, to minimize the total number of utterances in $\mathcal{E}$. In practice, the total count of utterances in $\mathcal{E}$ must be smoothed by plugging the margin into the following sigmoid function:

$$l(d(X_i)) = \frac{1}{1 + exp[\gamma d(X_i)]} \tag{8}$$

where $\gamma > 1$ is a constant to control the slope of the sigmid function. As in the MCE formulation, the *max* in the definition of margin need to be approximated by summation of exponential functions. Finally, the smoothed count of total mis-recognized utterances in $\mathcal{E}$ can be expressed as:

$$Q_2(\Lambda) = \sum_{X_i \in \mathcal{E}} l(d(X_i)) = \sum_{X_i \in \mathcal{E}} \frac{1}{1 + exp[\gamma d(X_i)]} \tag{9}$$

Given a training set D, we can estimate the whole CDHMM set, Λ, to minimize the following new objective function, $Q(\Lambda)$, which is a weighted linear combination of the objective function for MRME, $Q_1(\Lambda)$ in eq.(6) and $Q_2(\Lambda)$. Thus,

$$Q(\Lambda) = \alpha\, Q_1(\Lambda) + \beta\, Q_2(\Lambda) \tag{10}$$

where $\alpha \geq 0, \beta \geq 0$ are parameters to make a good balance between $Q_1(\Lambda)$ and $Q_2(\Lambda)$. The optimal value for α and β can be selected experimentally. The new training criterion is given as

$$\tilde{\Lambda} = \arg\min_{\Lambda} \ Q(\Lambda) \tag{11}$$

Similar to MRME, this minimization problem can be solved by any gradient descent algorithm, such as the generalized probabilistic descent (GPD) algorithm. This new training criterion can be regarded as a generalized MRME criterion. When $\beta = 0$, it is equivalent to the original MRME criterion. When $\alpha = 0$, it becomes a variant MCE criterion, which use only misrecognized training data as opposed to using all training data in the normal MCE.

4. MRME WITH NEGATIVE TOKENS

As discussed in section 3, the MRME algorithm uses only correctly recognized utterances to estimate the models. This constraint greatly limits the potential of MRME algorithm as when the tasks become large, the string error rate (even after MCE training) will be so high that only a very small percentage or even none of the training utterances can be used for MRME training.

The MRME algorithm was originally motivated by large margin classifiers in machine learning such as support vector machine (SVM). Therefore the definition of support token set in MRME takes similar form as that in SVM. By closely looking at MRME criterion, we find that we can include the negative tokens in the support token set as well (maybe it should not be called support token set any more). Considering the relative margin $d_2(X_i)$ defined in eq.(3), for any (including misrecognized) utterance X_i, $\mathcal{F}(X_i|\lambda_{S_i^T}) < 0$ and $\mathcal{F}(X_i|\lambda_{S_j}) < 0$, so $\left[1 - \frac{\mathcal{F}(X_i|\lambda_{S_i^T})}{\mathcal{F}(X_i|\lambda_{S_j})}\right] < 1$ always holds. Therefore the relative margin $d_2(X_i)$ has an upper bound of 1 for any training utterance X_i, regardless correctly or incorrectly recognized. This at least guarantees that the minimax optimization in eq.(5) is still solvable when we include all negative tokens.

Let's define a new support token set as

$$\mathcal{SV} = \{X_i \mid X_i \in \mathcal{D} \text{ and } d_2(X_i) \leq \gamma\} \tag{12}$$

where $\gamma > 0$ is a pre-set positive number. Obviously, the new support token set includes all mis-recognized utterances as well as all original support tokens given in eq.(4). The training criterion is the same as eq.(5). Given the new definition of support token set, the minimization in the criterion eq.(5) will choose the most negative token, which is the farthest from the decision boundary and locates in the wrong decision region. This is very different from the original MRME training where the minimization will always choose the token that is the nearest to the decision boundary but locates in the correct decision region. According to the criterion, the maximization will push the negative tokens to cross the decision boundaries (so they will have positive margins), similar to the way that MCE does. In this way, MRME no longer needs to bootstrap from MCE-trained model.

The new algorithm has a few potential advantanges over the orginal MRME. First, it can take full benefit of MRME because more training utterances participate in the training and therefore may be able to achieve better model. Especially in large vocabulary continuous speech recognition (LVCSR) tasks, where only a very small percentage of the training utterances is correctly recognized by the baseline models (MLE- or MCE- trained models), the benefit of the original MRME will be greatly limited due to lack of applicable training utterances. But the new algorithm has no such problem and can be directly applied to LVCSR tasks. Second, unlike the original MRME, the new algorithm does not need to use MCE models to bootstrap the training. As MRME itself is faster than MCE, training an MRME model takes even less time than training an MCE model. This is very important especially for LVCSR tasks where hundreds of hours training data may be used.

On the other hand, a drawback of this new algorithm is that the training may be vulnerable to outliers in training data.

5. EXPERIMENTAL RESULTS

The new approaches have been evaluated in the TIDIGITS corpus. The experimental setup is the same as in [8]. The database in our

experiments contains utterances from 225 speakers (112 for training, 113 for testing). The vocabulary has only digits ('1' to '9', plus 'oh' and 'zero'). The lengths of the digit strings are from 1 to 7. The training set has 8623 digit strings and the test set has 8700 strings. Our model set has 11 whole-word CDHMMs representing all digits. Each HMM has 12 states and use a simple left-to-right topology without state-skip. The data sampling rate is 16KHz. Standard MFCC feature is used (39 dimensions including 12 MFCC's and normalized energy, plus their first and second order time derivatives). The size of N-Best list is five (N=5). As opposed to the MCE training, only Gaussian means are updated during all MRME trainings.

Table 1 gives a performance comparison of the best results obtained by different training criteria on TIDIGITS test set. *mrme1* is the original MRME algorithm that uses only positive tokens in the support token set and bootstraps from MLE model. *mrme2* uses the same algorithm but bootstraps from MCE model. For references, results for best MLE models (*mle*) and MCE models (*mce*) are also listed. **cmb** is the new algorithm that combines MCE and MRME as discussed in section 3. **nsv** is the new algorithm that includes negative tokens in the support token set as discussed in section 4. Both **cmb** and **nsv** bootstrap from MLE model.

Table 1. Results (sentence accuracy in %) of different training criteria on TIDIGIT test data. The first row lists the number of Gaussions per HMM state.

mix/state	1m	2m	4m	8m	16m	32m
mle	85.10	93.66	95.74	97.63	97.84	98.34
mce	92.24	95.29	97.24	98.30	98.64	98.89
mrme1	95.10	97.90	98.55	98.90	99.03	99.16
mrme2	95.21	97.97	98.59	98.90	99.03	99.16
cmb	96.10	98.10	98.82	99.00	99.03	99.16
nsv	96.16	98.25	98.89	99.10	99.13	99.24

We can see that bootstrapping from MLE model (*mrme1*) as opposed to from MCE model (*mrme2*) affects MRME's performance only when the model is simple. When more complicated models (8 or more mixtures) are used, the recognition rates become the same. This is mainly because the accuracy difference between the MLE model and MCE model is so small when complicated models are used, so the support token sets in both cases are very similar. **cmb** achieves higher accuracy than the original MRME algorithm (*mrme1* or *mrme2*) when the model is simple and therefore there are more mis-recognized utterances in the training set. This confirms our hypothesis that joint optimization of MRME and MCE may work better than a separate MCE followed by an MRME. But when complexity of the models and therefore recognition accuracy increase, the improvement becomes smaller or even disappear. One reason is that when the number of recognition errors in the training set becomes smaller, **cmb** converges to MRME (when there is no mis-recognized utterance in the training set, **cmb** is equivalent to the original MRME).

The new MRME algorithm (**nsv**) that includes negative tokens in support token set achieves small but consistent improvement over the original MRME algorithm. The best result achieved by **nsv** is 0.76% string error rate, which is 10% relative lower than the best result (0.84%) for the original MRME algorithm. This is, to our knowledge, the best result that has ever been reported on this task. When the tasks become more difficult, for example, in WSJ 20K task, a very good MLE model can only achieve 40% sentence accuracy although the word accuracy is as high as 90%. Even after MCE training, the error rate will be still very high. In this case, the benefit of the original MRME algorithm will be greatly limited and the advantages of the new MRME algorithms may become much larger.

Although we use word as the basic modeling unit here, it is straightforward to extend to sub-word model unit (e.g. phone), where the string model will become a concatenation of phone models as opposed to word models used in this experiment. So far, our preliminary results on a WSJ 5K task using a simple state-clustered within-word triphone model and the new MRME algorithm (**nsv**) cut the word error rate of MLE model from 19.4% to 16.6%.

6. SUMMARY

Two new approaches to improve MRME are proposed and discussed. One is a new training criterion based on a combination of MRME and MCE objective function. The other is to include negative tokens into the support token set, so that MRME algorithm can be applied to all training data as opposed to only correctly recognized data in the original MRME approach. Both proposed new approaches can bootstrap directly from MLE models. Improvement on recognition rate has been achieved on TIDIGITS database. More research and experiments on sub-word (phone) based systems for large vocabulary continuous speech tasks (e.g. Wall Street Journal tasks) are in progress.

7. REFERENCES

[1] 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", *Proc. of ICASSP86*, pp.49-52, Tokyo, Japan, 1986.

[2] W. Chou, C.-H. Lee and B.-H. Juang, "Minimum error rate training based on N-best string models", *Proceedings of ICASSP93*, 1993.

[3] H. Jiang, "Discriminative training for large margin HMMs", *Technical Report CS-2004-01, CSE Department, York University*, March 2004.

[4] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Processing (SAP)*, pp.257-265, Vol.5, No.3, May 1997.

[5] X. Li, H. Jiang and C. Liu, "Large margin HMM for speech recognition", *Proceedings of ICASSP05*, May 2005.

[6] X. Li and H. Jiang, "A Constrained Joint Optimization Method for Large Margin HMM Estimation, " *Proceedings of IEEE ASRU Workshop*, November 2005.

[7] C. Liu, H. Jiang and X. Li, "Discriminative training of CDHMM for maximum relative separation margin", *Proceedings of ICASSP05*, May 2005.

[8] C. Liu, H. Jiang and L. Rigazio, "Maximum Relative Margin Estimation of HMMs based on N-Best String Models for Continuous Speech Recognition", *Proceedings of IEEE ASRU Workshop*, November 2005.

[9] Y. Normandin, R. Cardin and R. Demori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. on SAP*, Vol. 2, No. 2, Apr. 1994.

[10] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.