Optimization on Decoding Graphs Using Soft Margin Estimation

Abdelaziz A.Abdelhamid and Waleed H.Abdulla
Department of Electrical and Computer Engineering, Auckland University, New Zealand
E-mail: aabd127@aucklanduni.ac.nz, w.abdulla@auckland.ac.nz

***Abstract*—This paper proposes a discriminative learning algorithm for improving the accuracy of continuous speech recognition systems through optimizing the language model parameters on decoding graphs. The proposed algorithm employs soft margin estimation (SME) to build an objective function for maximizing the margin between the correct transcriptions and the corresponding competing hypotheses. To this end, we adapted a discriminative training procedure based on SME, which is originally devised for optimizing acoustic models, to a different case of optimizing the parameters of language models on a decoding graph constructed using weighted finite-state transducers. Experimental results show that the proposed algorithm outperforms a baseline system based on the maximum likelihood estimation and achieves a reduction of 15.11% relative word error rate when tested on the Resource Management (RM1) database.**

I. Introduction

Weighted finite state transducer (WFST) is an appropriate and flexible method for integrating various speech knowledge sources together into an elegant recognition network [1]. The strength of WFST comes from the simple but powerful operations, such as composition, determinization, and weight pushing [2]. The process of building an integrated recognition network usually starts with representing each knowledge source as a WFST. Then, a series of WFST operations are applied to produce the final recognition network (also called decoding graph). The resulting network can be used to decode the speech signal efficiently through the application of a search algorithm, such as the Viterbi search with beam pruning [3].

Most of current state-of-the-art speech research efforts are directed towards optimizing the parameters of speech knowledge sources separately without taking into consideration the interdependency between them. However, this direction is susceptible to achieve a sub-optimal performance of the overall speech decoding process [4]. One key for enhancing the accuracy of speech decoders is to find out a reliable estimation procedure for optimizing the parameters of the various knowledge sources jointly. This joint optimization can be achieved through optimizing the parameters of these knowledge sources while being integrated together into a single decoding graph using a discriminative training technique.

Discriminative training techniques are considered as an interesting approach for optimizing the parameters of pattern classifiers [5]. The basic idea of discriminative training is to penalize the parameters that are liable to confuse the correct and competing hypotheses. Various discriminative training criteria have been employed to optimize the acoustic and language models parameters, such as minimum phone error (MPE) [6], minimum word error (MWE) [7], maximum mutual information estimation (MMIE) [8], minimum sample risk (MSR) [9], minimum classification error (MCE) [10], and reranking techniques based on the perceptron algorithm [11].

Discriminative training of decoding graph parameters has received less attention compared to the discriminative training of acoustic and language models [12][13]. The discriminative training of decoding graph parameters is first introduced by Lin and Yvon in [4]. In that research, the authors applied discriminative training based on the MCE criterion to optimize transition weights of a WFST-based decoding graph composed of lexical, n-gram and acoustic models. This approach is asserted to achieve better performance when compared with the standard maximum likelihood estimation (MLE) approach. An extension to that research is presented by Kuo et al. in [14] in which the authors extended the work presented in [4] by using context dependent acoustic models instead of the context independent acoustic models. The benefit from optimizing decoding graph parameters lies in the ability to include both the language and acoustic scores in the optimization process thus may yield better parameter adjustment.

The approach of margin-based discriminative training became the current fashion in designing pattern classifiers due to its robustness towards the miss-matched conditions between training and testing data sets. However, it has been widely applied for optimizing the parameters of acoustic models [15], [16], but limitedly used for optimizing the parameters of language models [17]. To best of our knowledge, margin-based discriminative training has not been addressed yet for optimizing the parameters of WFST-based decoding graphs. This makes our findings are valuable for further improvements in discriminative training of language models on WFST-based decoding graphs. In this paper we propose a discriminative training algorithm for optimizing the parameters of speech decoding graphs using SME. The key advantage of soft margin classifiers is that they does not require a well trained ground models to achieve better recognition performance which makes these classifiers more advantageous over large margin classifiers [18].

This paper is organized as follows. In Section II the mathematical formulas of the SME are derived and the parameters optimization procedure is also discussed. Experimental results comparing the proposed method with both MLE and MMIE

are presented in Section III. Finally, Section IV presents the conclusion and the perspectives for future work.

II. Discriminative training using SME

Assume the speech utterance is represented as a sequence of observation vectors, denoted by X, and the corresponding word sequence is denoted by $W = w_1, w_2, ..., w_n$. The score of this observation sequence given the acoustic model parameters, denoted by Λ, and the language model parameters, denoted by Γ, is defined as [4]:

$$g(X, W, \Lambda, \Gamma) = logP(X|W, \Lambda) + \alpha.logP(W|\Gamma) \quad (1)$$

where α is the language model scaling factor, $P(X|W, \Lambda)$ and $P(W|\Gamma)$ are the acoustic and language models scores respectively. The task of the speech decoder is to select the best word sequence W_{best} that maximizes the score of X as follows:

$$g(X, W_{best}, \Lambda, \Gamma) = \underset{W}{arg\,max}\ g(X, W, \Lambda, \Gamma) \quad (2)$$

During the parameter optimization process, we need to compare the score of the reference word sequence, W_{ref}, with that of the 1-best competing hypothesis, W_{best}. For this purpose, an anti-discriminant function is defined as [19]:

$$d(X, \Lambda, \Gamma) = -g(X, W_{ref}, \Lambda, \Gamma) + g(X, W_{best}, \Lambda, \Gamma) \quad (3)$$

For simplicity, we incorporated only 1-best decoding hypothesis in the preliminary experiments discussed in this paper.

A. Expected Risk

The purpose of classification and recognition is to minimize the expected risk, which is calculated in terms of the classification errors of a representative test set. However, we don't know exactly the property of the test set to be considered, but we can only assume that the training and the test sets are independently and identically distributed from the same expected density. Since there is no explicit knowledge of the expected density, it can be approximated by an empirical density. In this case, the expected risk can be expressed in terms of the empirical density as follows [18]:

$$R(\Lambda, \Gamma) \leq R_{emp}(\Lambda, \Gamma) + R_{gen}(\Lambda, \Gamma) \quad (4)$$

where the expected risk, $R(\Lambda, \Gamma)$, and empirical risk, $R_{emp}(\Lambda, \Gamma)$, are the system's recognition error on testing and training data, respectively. The generalization risk, $R_{gen}(\Lambda)$, is a regularization term proportional to model complexity. Most current discriminative training methods focus only on how to minimize the empirical risk, $R_{emp}(\Lambda, \Gamma)$, with the hope to achieve a significant minimization of the expected risk. However, an optimal performance on the training set does not guarantee an optimal performance on the test set. The minimum expected risk can be obtained when a good balance between the empirical risk and the generalization risk is achieved. The generalization risk usually depends on the margin of the model and thus, this risk can be reduced if the margin shown in Fig. 1 is increased. The margin serves as a desired minimum distance between the training samples and the decision boundary. During training the model parameters, the objective is to pull those samples that fall within the margin away from the decision boundary. Those samples already far from the decision boundary do not contribute to model parameters estimation. After training the model parameters, all or most training samples will be outside the margin. Consequently, if a test sample deviates from the training samples of its correct class but the distance between the test sample and its nearest training sample is less than the margin, a correct decision can still be made.

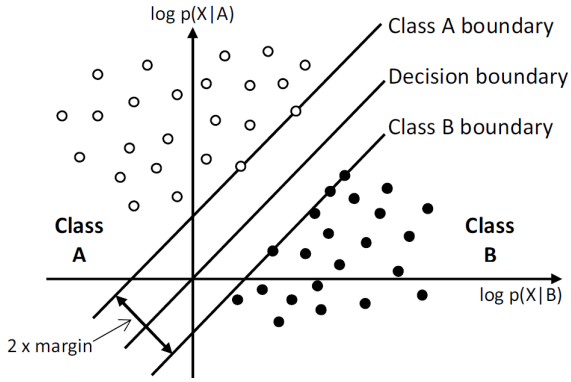

Fig. 1. Maximizing the margin between the two classes A and B to improve the model generalization capability.

B. Maximizing the margin

The key to improve the capability of model's generalization is to use a large margin. In our experiments, we use SME [18] to maximize the margin due to its good approximation of the expected risk. A brief description of SME is presented in this section. For more detailed implementation and discussion about SME, please refer to [18]. In SME, the language model parameters on a decoding graph are estimated by minimizing an approximated expected risk as follows:

$$L^{SME}(\rho, \Lambda, \Gamma) = \frac{\lambda}{\rho} + R_{emp}(\rho, \Lambda, \Gamma) \quad (5)$$

where Γ is the set of language model parameters, Λ is the acoustic model parameters, ρ is the soft margin, and $\frac{\lambda}{\rho}$ is the generalization risk. The variable λ is used to control the relative weight of the two terms of Eq. (5). With large λ, the training process will focus on reducing the generalization term and the margin will be large and vice versa. To obtain a good performance, it is important to obtain a good balance of these two terms. In this paper, the empirical risk is defined as:

$$R_{emp}(\rho, \Lambda, \Gamma) = \begin{cases} \rho - d(X, \Lambda, \Gamma) & \text{if } \rho > d(X, \Lambda, \Gamma) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

using this empirical risk, the SME objective function can be re-written as:

$$\begin{aligned} L^{SME}(\rho,\Lambda,\Gamma) &= \frac{\lambda}{\rho} + R_{emp}(\rho,\Lambda,\Gamma) \\ &= \frac{\lambda}{\rho} + (\rho - d(X,\Lambda,\Gamma))I(X \in U) \end{aligned} \quad (7)$$

where I is an indicator function, and U is the set of utterances that have the separation measures less than the soft margin. The separation measure usually represents how well the correct model is separated from the competing models corresponding to X, or how far X is from the decision boundary. If the separation measure is not large enough (i.e., it is less than the margin), a loss is generated that equals to $(\rho - d(X,\Lambda,\Gamma))$.

C. Solution to SME

There are two solutions to SME similar to those presented in [18]. One solution is to optimize the soft margin and the decoding graph parameters jointly. The other is to set the soft margin in advance, then find the optimal decoding graph parameters. In this paper, we adopted the second solution and the first one is left for future work. For both of these solutions, the indicator function $I(X \in U)$ is approximated with a sigmoid function [18]. Therefore, Eq. (7) is reformulated as:

$$\begin{aligned} L^{SME}(\rho,\Lambda,\Gamma) &= \frac{\lambda}{\rho} + (\rho - d(X,\Lambda,\Gamma)) \\ &\times \frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))} \end{aligned} \quad (8)$$

where γ is a smoothing parameter of the sigmoid function. Equation (8) can be viewed as a smoothing function of the soft margin ρ and the decoding graph parameters. Therefore, these parameters can be optimized using the generalized probabilistic descent (GPD) algorithm [20] on the training set as follows:

$$\Gamma_{t+1} = \Gamma_t - \epsilon \bigtriangledown L^{SME}(\rho,\Lambda,\Gamma_t) \quad (9)$$

For simplicity, we keep the parameters of the margin ρ and acoustic model Λ unchanged and calculate $\frac{\partial L^{SME}(\rho,\Lambda,\Gamma)}{\partial\Gamma}$, then the gradient of (8) becomes:

$$\frac{\partial L^{SME}(\rho,\Lambda,\Gamma)}{\partial\Gamma} = A + B \quad (10)$$

where

$$A = \frac{\partial(\rho - d(X,\Lambda,\Gamma))}{\partial\Gamma}\frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))} \quad (11)$$

and

$$B = (\rho - d(X,\Lambda,\Gamma))\frac{\partial}{\partial\Gamma}\frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))} \quad (12)$$

The two derivatives in Eq. (11) and Eq. (12) can be further written as:

$$\frac{\partial(\rho - d(X,\Lambda,\Gamma))}{\partial\Gamma} = \frac{\partial(-d(X,\Lambda,\Gamma))}{\partial\Gamma} \quad (13)$$

where

$$\frac{\partial d(X,\Lambda,\Gamma)}{\partial\Gamma} = -\digamma\left(W_{ref},s\right) + \digamma\left(W_{best},s\right) \quad (14)$$

where $\digamma(W,s)$ represents the number of occurrences of the transition weight, sm in the decoding hypothesis, W. For the derivative in Eq. (12), it can be written as:

$$\begin{aligned} &\frac{\partial}{\partial\Gamma}\frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))} \\ =\ & -\left\{\frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))}\right\}^2 \\ \times\ & exp(-\gamma(\rho - d(X,\Lambda,\Gamma)))(-\gamma)\frac{\partial(\rho - d(X,\Lambda,\Gamma))}{\partial\Gamma} \\ =\ & \gamma\left\{1 - \frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))}\right\} \\ \times\ & \frac{1}{1+exp(\gamma(\rho - d(X,\Lambda,\Gamma)))}\frac{\partial(\rho - d(X,\Lambda,\Gamma))}{\partial\Gamma} \end{aligned} \quad (15)$$

where $\frac{\partial(\rho - d(X,\Lambda,\Gamma))}{\partial\Gamma}$ is defined in Eq. (14). Then, by substituting Eq. (14) and Eq. (15) into Eq. (11) and Eq. (12) respectively, we can get the derivative required for the GPD update rule defined in Eq. (9).

D. Optimization procedure

The training procedure used to optimize the transition weights (carrying the language model parameters) of WFST-based decoding graph consists of the following steps:

1) For each training utterance, we extract a reference sub-graph, S_{ref}, by constructing an acceptor-type WFST, Y_{ref}, which has an arc sequence that inputs and outputs the same word labels and composing it with the large decoding graph, R, as follows: $S_{ref} = R \circ Y_{ref}$.
2) Decode the training utterance using the large decoding graph, R, and store the corresponding transitions of the competing hypothesis along with the associated decoding score.
3) Decode the training utterance using the extracted reference sub-graph, S_{Ref}, and store the corresponding reference path along with the associated decoding score.
4) Count the transitions in reference and competing hypotheses based on the transition weights.
5) Calculate the score difference using Eq. (3), then calculate the gradient of the loss function Eq. (10).
6) Update the transition weights of the large decoding graph using the update rule Eq. (9).
7) Repeat from step 2 as long as the performance converges or reaching a certain number of iterations.

Only the first transition in the set of candidate transitions with different weight counts is updated [4].

III. Experiments

A. Experimental setup

The experiments conducted in this paper are performed in terms of the RM1 speech database. The utterances containing out-of-vocabulary (OOV) words were removed from both

training and testing sets. In all experiments, the speech signal is sampled at 16kHz, 16bits/sample and framed at a rate of 30msec with 75% overlap between successive frames. Each frame is represented using 39 dimensional feature vector consisting of 12 static Mel Frequency Cepstral Coefficients (MFCC), energy, 26 dynamic coefficients (13 Δ, 13 $\Delta\Delta$).

The HMM set contains physical acoustic models for 41 phones, 882 diphones and $26,412$ triphones. These physical models are trained using Wall Street Journal (WSJ) speech corpora. Additionally, $38,229$ logical models are synthesized using state tying based on decision trees [21]. Each acoustic model consists of 3 states with left to right transitions without skip. There is a total of $8,000$ distinct states, each of which is associated with 39-dimensional probability density function taking the form of 32 mixtures per state with diagonal covariance matrix. The language model consists of $5,000$ uni-grams, $258,669$ bi-grams and $171,064$ tri-grams. These n-grams are trained using Gigawords text corpus and used to construct the large decoding graphs with a vocabulary containing 5k words. The acoustic and language models were freely available at the location referred to in [22] at the time of writing this paper.

The decoder presented in [23] is used in our experiments. This decoder runs at $1.5\times$RT and $0.02\times$RT on the large decoding graph R and the reference sub-graph S_{Ref}, respectively, when tested on 2.3 Ghz Intel Core i5 processor and after applying some pruning thresholds. In the literature, there are many faster decoders (eg. [24]), but these decoders only keep track of the word history of hypotheses, thus the complete sequence of state transitions which play a crucial role in discriminative training cannot be recovered.

B. Results and discussion

Before experimenting with the GPD procedure, we performed a number of experiments to set the slope of the sigmoid function, γ, which was chosen as 0.01. Also, the training step size for both SME and MMIE was chosen as 0.1. One way to select these values is to cut and try. In all experiments, five iterations of the GPD procedure were conducted. The optimized models resulting from the five experiments were incorporated in the evaluation of the RM1 test set.

The baseline system consists of various knowledge sources trained using the standard MLE approach. While performing the parameter optimization using the proposed SME approach, and after each iteration, the optimized graph is saved on disk and used for testing. The detailed testing results using the trained graphs from each training iteration are listed in Tables I and II for the MMIE and SME based training, respectively. The first row of these tables is the testing results using the MLE trained acoustic and language models. It is shown from these results that the word error rate (WER) achieved by SME approach outperforms the results achieved by both MLE and MMIE approaches. The best WER achieved by the SME approach was 25.00%, which is better than the WER achieved by MLE (29.45%) and MMIE (25.65%).

An important factor affects the generalization capability of the trained graphs is the margin size ρ defined in Eq. (7). Since we set the margin value as a constant, several experiments have been conducted for finding the best value for this margin. The margin value of 15 is found to achieve the best recognition performance. However, it is expected to achieve much better results if the margin value is adjusted adaptively during the training process and depending on the training samples whose distance from the decision boundary is less than the margin.

TABLE I
RECOGNITION PERFORMANCE USING THE MMIE TRAINED GRAPH.

Iteration	Sub (%)	Del (%)	Ins (%)	WER (%)
Baseline [22]	*23.56*	*3.27*	*2.62*	*29.45*
1	23.95	2.88	2.88	29.71
2	21.86	2.09	2.75	26.70
3	21.73	1.44	2.88	26.05
4	21.07	1.31	3.27	**25.65**
5	20.03	1.31	4.45	25.79

TABLE II
RECOGNITION PERFORMANCE USING THE SME TRAINED GRAPH.

Iteration	Sub (%)	Del (%)	Ins (%)	WER (%)
Baseline [22]	*23.56*	*3.27*	*2.62*	*29.45*
1	23.95	2.88	2.88	29.71
2	21.86	2.09	2.88	26.83
3	21.07	1.70	3.01	25.79
4	20.55	1.31	3.14	**25.00**
5	20.03	1.18	4.32	25.52

IV. CONCLUSION

In this paper we presented a discriminative training method for learning the parameters of speech decoding graphs. Experimental results emphasized the effectiveness of the proposed method when compared with one of the well-known discriminative training criteria, namely MMIE. The research presented in this paper can be extended from several perspectives. Firstly, the use of N-best decoding hypotheses is proved to give better performance when SME is applied for acoustic modelling. Therefore, the proposed approach is expected to achieve better results if N-best hypotheses are incorporated. Secondly, changing the margin size adaptively also can give better performance. In this case, the margin size can be learnt as well as the transition weights using the GPD algorithm. Thirdly, the proposed method is applied using the GPD, whereas other learning algorithms, such as Rprob or Quickprop can be used to improve the quality of the learnt parameters. Fourthly, The proposed approach is tested on a small task, so it may be further tested on a larger task.

V. ACKNOWLEDGEMENT

This work is supported by the R&D program of the Korea Ministry of Knowledge and Economy (MKE) and Korea Evaluation Institute of Industrial Technology (KEIT) [KI001836]. We thank ETRI for their contributions and help with the work. The authors would like to acknowledge the HealthBots Project Leader A/P Bruce A.MacDonald for the great support in developing this research.

REFERENCES

[1] M. Mohri, F. Pereira, and M. Riley, "Weighted finite state transducers in speech recognition," *Transactions on Computer Speech and Language*, vol. 16, pp. 69–88, 2002.

[2] C. Allauzen, M. Mohri, M. Riley, and B. Roark, "A generalized construction of integrated speech recognition transducers," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[3] S. Young, N. Russel, and J. Thornton, "Token passing: A simple conceptual model for connected speech recognition systems," Tech. Rep., 1989.

[4] S. Lin and F. Yvon, "Discriminative training of finite state decoding graphs," in *Proceedings of International Speech Communication Association (InterSpeech)*, 2005, pp. 733–736.

[5] H. Jiang, "Discriminative training for automatic speech recognition: A survey," *Transactions on Computer Speech and Language*, vol. 24, pp. 589–608, 2010.

[6] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 105–108.

[7] J. Kuo and B. Chen, "Minimum word error based on discriminative training of language models," in *Proceedings of International Speech Communication Association (InterSpeech)*, 2005, pp. 1–4.

[8] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 1986, pp. 49–52.

[9] J. Gao, H. Yu, W. Yuan, and P. Xu, "Minimum sample risk methods for language modeling," in *Proceedings of HLT/EMNLP*, October 2005, pp. 209–216.

[10] Z. Chen, M. Li, and K. Lee, "Discriminative training on language model," in *Proceedings of International Conference on Spoken Language Processing*, 2000.

[11] B. Roark, M. Saraclar, and M. Collins, "Corrective language modeling for large vocabulary ASR with the perceptron algorithm," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[12] A. Abdelhamid and W. Abdulla, "Discriminative training of context-dependent phones on WFST-based decoding graphs," in *Proceedings of International Conference on Communication, Signal Processing and their Application*, 2013.

[13] ——, "Optimizing the parameters of decoding graphs using new log-based MCE," in *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2012.

[14] H. Kuo, B. Kingsbury, and G. Zweig, "Discriminative training of decoding graphs for large vocabulary continuous speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, April 2007, pp. 45–48.

[15] H. Jiang, X. Li, and C. Liu, "Large margin hidden Markov models for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1584–1595, September 2006.

[16] D. Yu and L. Deng, "Large-margin discriminative training of hidden Markov models for speech recognition," in *International Conference on Semantic Computing*, September 2007, pp. 429–438.

[17] V. Magdin and H. Jiang, "Large margin estimation of n-gram language models for speech recognition via linear programming," in *IEEE International Conference on Acoustics Speech and Signal Processing*, March 2010, pp. 5398–5401.

[18] J. Li, M. Yuan, and C.-H. Lee, "Approximate test risk bound minimization through soft margin estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2393–2404, November 2007.

[19] H. Kuo, E. Lussier, H. Jiang, and C.-H. Lee, "Discriminative training of language models for speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 325–328.

[20] S. Katagiri, C.-H. Lee, , and B.-H. Juang, "A generalized probabilistic descent method," in *Proceedings of Acoustical Society of Japan*, 1990, pp. 141–142.

[21] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of ARPA Human Language Technology Workshop*.

[22] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cambridge University, Tech. Rep., 2006.

[23] A. Abdelhamid, W. Abdulla, and B. MacDonald, "WFST-based large vocabulary continuous speech decoder for service robots," in *Proceedings of International Conference on Imaging and Signal Processing for Healthcare and Technology*, 2012, pp. 150–154.

[24] G. Saon, D. Povey, and G. Zweig, "Anatomy of an extremely fast LVCSR decoder," in *Proceedings of International Speech Communication Association (InterSpeech)*, 2005, pp. 549–552.