Optimizing the Parameters of Decoding Graphs Using New Log-based MCE

Abdelaziz A.Abdelhamid
Electrical and Computer Engineering
The University of Auckland
Email: abdelaziz.ece@hotmail.com

Waleed H.Abdulla
Electrical and Computer Engineering
The University of Auckland
Email: w.abdulla@auckland.ac.nz

***Abstract*—This paper proposes a new class loss function as an alternative to the standard sigmoid class loss function for optimizing the parameters of decoding graphs using discriminative training based on minimum classification error (MCE) criterion. The standard sigmoid based approach tends to ignore a significant number of training samples that have a large difference between the scores of the reference and their corresponding competing hypotheses and this affects the parameters optimization. The proposed function overcomes this limitation through considering almost all the training samples and thus improved the parameter optimization when tested on large decoding graphs. The decoding graph used in this research is an integrated network of weighted finite state transducers. The primary task examined is 64K words, continuous speech recognition task. The experimental results show that the proposed method outperformed the baseline system based on both the maximum likelihood estimation (MLE) and sigmoid-based MCE and achieved a reduction in the word error rate (WER) of 28.9% when tested on the TIMIT speech database.**

I. Introduction

Weighted finite state transducer (WFST) is considered as an appropriate and flexible approach for combining various speech knowledge sources together in an integrated recognition network (also called decoding graph)[1]. The strength of WFST comes from the simple but powerful operations i.e. composition, determinization, and weight pushing [2].

The process of building speech decoding graphs usually starts with representing each kind of speech knowledge sources (namely, lexical, acoustic and language models) as a WFST. Then a series of WFST operations are applied to produce the final decoding graph. The resulting graph can be used to decode speech utterances efficiently using a search algorithm i.e. Viterbi search with beam pruning and token passing techniques [3].

Most of current state-of-the-art speech research efforts are biased towards optimizing the parameters of speech knowledge sources separately without taking into consideration the interdependency among these knowledge sources. The fact that yields a sub-optimal performance of the overall speech decoding process [4]. The key for enhancing the accuracy of the speech decoders is to find a reliable estimation procedure for optimizing the parameters of the various knowledge sources jointly. One way to do that is to optimize the parameters of speech knowledge sources while being integrated together into a unified decoding graph and using discriminative training.

In the literature, MLE is the commonly used technique for optimizing the language and acoustic models parameters separately [5]. However, MLE is based on some assumptions i.e. speech observations are taken from a family of distributions (typically Gaussian), training data is unlimited, and the true language model is known. In real, none of these assumptions holds for speech so, MLE is not guaranteed to give optimal decoding results [6], [7].

Currently, discriminative training techniques are considered as a complementary approach to MLE and overcome some of its limitations [8]. The basic idea of discriminative training is to penalize the knowledge sources parameters that are liable to confuse the correct and incorrect utterances. Various discriminative training criteria have been used to optimize the acoustic model parameters i.e. Minimum phone error (MPE) [9], Minimum word error (MWE) [10], and Maximum Mutual Information (MMI) [11]. It has been also shown that discriminative training can be used to optimize the language model parameters using various criteria such as Minimum Sample Risk (MSR) [12], Minimum Classification Error (MCE) [13], and the reranking techniques based on the perceptron algorithm [14].

Although the variations of discriminative training criterion achieved better performance than MLE, but still deal with the knowledge sources as separate and independent optimization tasks. There are some research efforts have been done on the joint optimization of the acoustic and language models parameters i.e. [4], [6]. These researches are based on optimizing the language model parameters in decoding graphs with taking into consideration the acoustic score while training the language model parameters using only the first best competing hypothesis.

In this paper we improved the previous work on discriminative training presented in [4], [6] through the utilization of the N-best hypotheses in the training process. Additionally, a new class loss function is presented as an alternative to the standard sigmoid class loss function to improve the parameter optimization of decoding graphs.

This paper is organized as follows: in section II the new class loss function is defined and derived. The parameters optimization procedure is discussed in section III. Experimental results comparing the proposed method with both sigmoid based MCE and MLE are presented in section IV. Finally,

section V presents the conclusion and future perspectives.

II. Proposed Approach

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

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

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:

$$W_{best} = \underset{W}{argmax} \;\; g\left(X, W; \Lambda, \Gamma\right) \quad (2)$$

where g is the total decoding score of the speech utterance X. The N-best word sequences can be retrieved through keeping the top N decoding hypotheses as follows:

$$W_r = \underset{W \neq W_1, ..., W_{r-1}}{argmax} \;\; g\left(X, W; \Lambda, \Gamma\right) \quad (3)$$

where W_r is the r^{th} best decoding hypothesis. Let W_{ref} denotes the correct reference word sequence, we need to compare the score of W_{ref} with that of its corresponding N-best competing hypotheses $W_1, ..., W_N$. For this purpose, an anti-discriminant function can be defined as:

$$\begin{aligned} d(X) &= d\left(X; \Lambda, \Gamma\right) \\ &= G\left(X, W_1, ..., W_N; \Lambda, \Gamma\right) \\ &- g\left(X, W_{ref}; \Lambda, \Gamma\right) \end{aligned} \quad (4)$$

where the contribution of N-best hypotheses is represented in terms of P_{norm} anti-discriminant function G defined as [8]:

$$G\left(X_i, W_1, ..., W_N; \Lambda, \Gamma\right) = log\left(\frac{1}{N}\sum_{r=1}^{N} exp\left[g\left(X_i, W_r; \Lambda, \Gamma\right)\eta\right]\right)^{\frac{1}{\eta}} \quad (5)$$

where η is a positive parameter that controls the weighting of N-best hypotheses. The anti-discriminant function is used to differentiate between the correct and incorrect word hypotheses; since it gives $+ve$ value if the best hypothesis is correct and gives $-ve$ value vice versa.

In order to use the anti-discriminant function in a gradient descent optimization, one way is to formulate it into a smoothed and differentiable 0-1 function. The standard and common choice is the sigmoid class loss function defined as:

$$l\left(d\left(X\right)\right) = \frac{1}{1 + exp\left(-\gamma d\left(X\right) + \theta\right)} \quad (6)$$

where γ and θ are constants used to control slope and shift of the sigmoid function respectively. Based on the Generalized Probabilistic Descent (GPD) algorithm [15], the decoding graph parameters can be iteratively adjusted using the following update rule [4], [6]:

$$\Gamma_{t+1} = \Gamma_t - \epsilon \bigtriangledown l\left(X; \Lambda_t, \Gamma_t\right) \quad (7)$$

To simplify the problem of joint optimization of both the acoustic and language model parameters, we keep the parameters of the acoustic models unchanged and calculate $\frac{\partial d(X;\Lambda,\Gamma)}{\partial \Gamma}$, then the gradient of (6) becomes:

$$\bigtriangledown l\left(d(X)\right) = \frac{\partial l}{\partial d}\frac{\partial d\left(X; \Lambda, \Gamma\right)}{\partial \Gamma} \quad (8)$$

where the first term is the gradient of the sigmoid class loss function and is given by:

$$\frac{\partial l}{\partial d} = \gamma l(d)(1 - l(d)) \quad (9)$$

However, $l(d)$ approaches the value One for the utterances with a large value of $d(X)$, yielding the slope of the sigmoid function to approach the value $Zero$, and resulting in ignoring the contribution that we may obtain from these utterances to the gradient in the optimization process. In order to overcome this limitation, we propose the following log class loss function:

$$\hat{l}\left(d\left(X\right)\right) = log(l\left(d\left(X\right)\right)) = -log\left(1 + e^{-\gamma\, d(X)}\right) \quad (10)$$

The gradient of this new loss function is defined as:

$$\frac{\partial \hat{l}}{\partial d} = \frac{\gamma e^{-\gamma d(X)}}{1 + e^{-\gamma d(X)}} \quad (11)$$

The advantage of this gradient is that almost all the training speech utterances are included in the training process even if the speech utterance is highly misclassified. In this case, it is obvious that for speech utterances with score difference between the reference and competing hypotheses is large $+ve$ value, the gradient of the proposed function approaches $\gamma/2$ and thus these speech utterances will be considered in the training process.

Back to the derivation in (8), to compute $\frac{\partial d(X;\Lambda,\Gamma)}{\partial \Gamma}$, we need to differentiate $d\left(X; \Lambda, \Gamma\right)$ with respect to all the parameters of the weight vector Γ. However, we can simplify this by taking the partial derivative with respect to the transition weight s, then the derivation becomes [8]:

$$\frac{\partial d(X;\Lambda,\Gamma)}{\partial \Gamma} = \left[-I\left(W_{Ref}, s\right) + \sum_{n=1}^{N} C_n I(W_n, s)\right] \quad (12)$$

where

$$C_n = \frac{exp\left[g\left(X, W_n; \Lambda, \Gamma\right)\eta\right]}{\sum_{j=1}^{N} exp\left[g\left(X, W_j; \Lambda, \Gamma\right)\eta\right]} \quad (13)$$

where $I(W, s)$ represents the number of occurrences of the transition weight s in the decoding hypothesis W.

III. Optimization Procedure

The speech knowledge sources are compiled and integrated together into a unified static and large decoding graph through the application of a series of WFST operations [16]. An optional silence is added at the boundaries of each word in the dictionary and a scaling factor $\alpha = 13$ is used to scale the language model parameters. Table I shows the size of WFST representing each knowledge source along with the operations applied to get the integrated decoding graph $(C \circ det(L)) \cdot (G \circ T)$.

TABLE I
Sequence of operations applied to build large WFST.

Graph/Operation	Num. of states	Num. of trans.
C	1,681	84,080
L	523,083	592,837
G	595,765	1,327,969
T	63,999	191,997
$det(L)$	209,919	279,673
$C \circ det(L)$	346,452	550,709
$G \circ T$	886,099	1,932,311
$(C \circ det(L)) \cdot (G \circ T)$	5,579,208	8,082,205

C : Context dependency WFST, L : Lexicon WFST, G : Tri-gram WFST, T : Silence WFST, $\circ$: Composition operation, det : Determinization operation, $\cdot$: Lookahead composition.

The training procedure followed to optimize the language model parameters consists of the following steps:

1) For each training sentence we extract a reference subgraph S_{ref} by constructing an acceptor-type WFST Y_{ref} which has an arc sequence that inputs and outputs the same word sequence and composing it with the large decoding graph R as follows: $S_{ref} = R \circ Y_{ref}$.
2) Decode the training speech utterance using the large decoding graph R. For each sentence, store the corresponding transitions of the N-best decoded sequences along with the associated decoding score for each sequence.
3) Decode the training speech utterance using the extracted reference subgraph S_{Ref} and store the corresponding reference path along with the associated decoding scores.
4) Count the transitions in the reference and N-best hypotheses based on the transition weights.
5) Calculate the score difference using (Eq. 4), then calculate the gradient of the proposed loss function (Eq. 11) and the gradient of the language model (Eq. 12).
6) Update the transition weights of the large decoding graph using the update rule (Eq. 7).
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].

IV. Experiments

A. Experimental setup

The experiments are performed on the TIMIT speech database. Approximately 2 hours of continuous speech represented in 1,019 utterances of manually transcribed data was used for training the language model parameters of a large decoding graph. The test set contains around 3.5 hours of continuous speech represented in 2,819 utterances is used for intensive testing. The transcriptions that contain 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 with frame rate of 30msec with 75% overlap between successive frames. Each frame is represented using 39 dimensional feature vectors with 12 static Mel Frequency Cepstral Coefficients (MFCC), 24 dynamic coefficients (12 Δ, 12 $\Delta\Delta$) and 3 log energy values.

The HMM set contains physical acoustic models for 38 phonemes, 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. Each acoustic model consists of 3 states with left to right transitions without skip. There are 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 and the covariance matrix is diagonal.

The language model consists of 64,000 uni-grams, 522,530 bi-grams and 173,445 tri-grams. These n-grams are trained from Gigawords text corpus and used to construct the large decoding graphs with a vocabulary containing 64k words. The resulting decoding graph and the operation applied are shown in table I.

The decoder proposed in [17] is used in our experiments. This decoder runs at 1.5xRT and 0.02xRT on the large decoding graph R and the reference subgraph 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. [18]), but these decoders only keep track of the word history of hypotheses and thus, the complete sequence of state transitions which play a crucial role in discriminative training cannot be recovered.

B. Parameters selection

Before experimenting with discriminative training procedure, we performed a number of experiments aiming at setting three parameters of the procedure for the proposed log-based training. This set of parameters includes: γ, which controls the slope of the sigmoid function, ϵ, which is the increment parameter of the gradient descent, and η which controls the contribution of the N-best hypotheses in the parameters optimization. To simplify the training procedure, we assumed that α, the language model weight, and the word insertion penalty, δ, are fixed and take the values $\alpha = 13, \delta = 0.1$ we also set $\theta = 0$ for both the sigmoid based and log based training. Since the proposed method can deal with the utterances with large difference between the reference and competing hypotheses scores, we can easily set the values of the parameters γ and ϵ as 0.05 and 0.9 respectively depending on the convergence speed of the training algorithm. The rationale for finding a reasonable

value of the third parameter η is based on the following remark: when the difference d_f, between the reference and competing hypotheses scores is large, the value of C_n which represents the contribution of the N-best hypotheses tends to $Zero$ for large values of η. Choosing a small value for $\eta = 0.001$ would have the effect of increasing the contribution of the N-best hypotheses. While more experimental work is required to fine tune these parameters seemed to yield a reasonable convergence rate. For discriminative training using the standard sigmoid class loss function, we used the values 0.01, 10 and 0.0001 for both γ, ϵ and η respectively [4]. The number of best decoding hypotheses included in the anti-discriminant function of Eq. (5) is $N = 2$.

C. Results and discussion

The baseline system consists of various knowledge sources trained using the standard MLE approach. While performing the parameters optimization using the proposed MCE approach, and after each iteration, the optimized graph is saved on disk and used for testing. For both sigmoid based and log based training, four training iterations were performed. The reason for choosing this number of iterations is based on the convergence of optimization process, since we observed that the optimization diverges after the fourth iteration for both the sigmoid and log based training.

The detailed testing results after each training iteration are listed in tables II and III for the sigmoid and log based training respectively. It is shown from these results that the recognition accuracy and WER obtained using log-based approach outperforms the results obtained when from the sigmoid-based approach. The reason behind this improved performance is that the sigmoid function ignores from the training procedure all the utterances with high difference between the reference and competing hypotheses especially in the first few iterations. However, the log-based approach takes into consideration most of the training utterances even when the score difference between the reference and competing hypotheses is high.

Figure 1 shows the number of ignored utterances that are skipped from the training process for both the sigmoid-based and log-based approaches. It is shown that the number of missed utterances in case of log-based approach is much lower than that of the sigmoid based approach which proves our claim. The number of ignored utterances in the first iteration is 160 utterances in the case of sigmoid-based approach while the number of ignored utterances is only 15 utterances in the case of log-based approach, this explains why the achieved word accuracy of the log-based method is much higher than that obtained using the sigmoid-based method. It's also noted that some of the ignored utterances are taken into consideration in the next iterations, but still the number of ignored utterances from the sigmoid-based approach is much higher than that from the log-based one.

Table II shows the detailed results of each training iteration of the sigmoid based approach. The first row shows the details of the output from the decoding process for the set of test utterances using MLE trained language model. The baseline

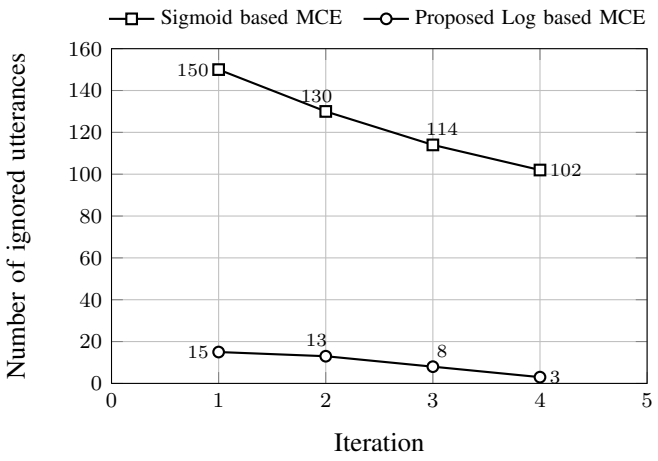

Fig. 1. Number of ignored utterances while training.

word recognition accuracy is 77.01% and the WER is 25.19%. The next rows in this table show the details of the output from the decoder for the set of test utterances using the optimized decoding graph after each training iteration. The best word accuracy achieved is 83.98% and the best WER is 18.09% when the sigmoid function is used as a class loss function for the MCE training.

TABLE II
DETAILED PERFORMANCE OF TRAINED GRAPHS USING THE SIGMOID BASED MCE.

Iteration	Corr%	Sub%	Del%	Ins%	WER%
Baseline	*77.01*	*18.51*	*4.47*	*2.20*	*25.19*
1	80.39	15.95	3.66	2.06	21.67
2	81.93	14.72	3.35	1.90	19.97
3	83.24	13.59	3.18	1.93	18.69
4	**83.98**	13.15	2.87	2.07	**18.09**

The results obtained from discriminative training using the proposed log-based approach is shown in table III. The best word recognition accuracy obtained from the log-based approach is 84.19% which is higher than the obtained word accuracy from the sigmoid-based method. Besides, the WER obtained from the log-based method is 17.89% which is also better than that of the sigmoid-based method.

TABLE III
DETAILED PERFORMANCE OF TRAINED GRAPHS USING THE PROPOSED LOG BASED MCE.

Iteration	Corr%	Sub%	Del%	Ins%	WER%
Baseline	*77.01*	*18.51*	*4.47*	*2.20*	*25.19*
1	81.21	15.30	3.49	1.86	20.65
2	83.24	13.69	3.07	1.82	18.59
3	83.98	13.25	2.76	1.96	17.98
4	**84.19**	13.05	2.76	2.08	**17.89**

V. CONCLUSION

In this paper we presented an improvement to the previous work presented in [4], [6] through the inclusion of N-best

hypotheses in the training process. Additionally, we presented a new class loss function that overcomes the limitation of the standard sigmoid class loss function for discriminative training based on MCE criterion. The experimental results show that the proposed log-based approach achieves better performance than both MLE and the standard sigmoid-based MCE approaches when tested on the TIMIT speech database. A future direction could be to evaluate the performance of the proposed method with different number of N-best hypotheses included in the calculation of the anti-discriminant function. Another future direction is to test the proposed approach with different strategies for updating the weights of the candidate transitions like updating the final transition or randomly selecting the transition used in the weight adjustment.

VI. 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, Development of Mediated Interface Technology for HRI]. The authors would like to register their acknowledgement to the HealthBot Project Leader A/P Bruce A.MacDonald and team 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 the International Conference on Acoustics, Speech, and Signal Processing*, 2004.

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

[4] S. S. LIN and F. YVON, "Optimization on decoding graphs by discriminative training," in *Proceedings of the International Speech Communication Association*, 2007.

[5] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 2, pp. 179–190, 1983.

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

[7] K. Vertanen, "An overview of discriminative training for speech recognition," in *Cambridge University*, 2004.

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

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

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

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

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

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

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

[15] W. Chou, C. H. Lee, and B. H. Juang, "Segmental GPD training of hidden Markov model based speech recognizer," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1992, pp. 473–476.

[16] J. R.Novak, N. Minemaysu, and K. Hirose, "Painless WFST cascade construction for LVCSR-Transducersaurus," in *Proceedings of the International Speech Communication Association*, 2011.

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

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