Joint Discriminative Learning of Acoustic and Language Models on Decoding Graphs

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*—In traditional models of speech recognition, acoustic and language models are treated in independence and usually estimated separately, which may yield a suboptimal recognition performance. In this paper, we propose a joint optimization framework for learning the parameters of acoustic and language models using minimum classification error criterion. The joint optimization is performed in terms of a decoding graph constructed using weighted finite-state transducers based on context-dependent hidden Markov models and tri-gram language models. To emphasize the effectiveness of the proposed framework, two speech corpora, TIMIT and Resource Management (RM1), are incorporated in the conducted experiments. The preliminary experiments show that the proposed approach can achieve significant reduction in phone, word and sentence error rates on both TIMIT and RM1 when compared with conventional parameter estimation approaches.**

I. Introduction

Discriminative learning of generative models has been emerged recently as a promising approach for boosting performance of automatic speech recognition (ASR) [1], [2]. However, most of current, research based on this approach, treats acoustic and language models as separate and independent components, and the parameters of these models are usually optimized individually using various criteria [3], [4], [5], [6], [7]. The drawback of the models independence assumption can be obvious if we take into consideration the hierarchical matching from phonetic to linguistic levels. In this case, this assumption becomes unrealistic for achieving better optimization of the parameters of acoustic and language models. Therefore, we propose in this paper a generic framework that optimizes the parameters of both acoustic and language models jointly to benefit from the inherent correlation between these models. To verify the effectiveness of the proposed framework, we conducted as set of experiments using the minimum classification error rate (MCE) criterion for the sake of improving the phone, word and sentence error rates of TIMIT and RM1 speech corpora.

This research is organized as follows. A review on the related work is presented in Section II, followed by a description of the proposed framework in Section III. The experimental results are then presented and discussed in Section IV. Finally, the conclusions come in Section V.

II. Related work

To the best of our knowledge, there is a little research in the literature *explicitly* addressing the joint optimization of the parameters of acoustic and language models. In [8], the authors optimized the parameters of a decoding graph in the form of log-linear distributions, but no significant improvement was achieved when compared with traditional MCE-based learning especially for acoustic models with large number of mixtures per state (i.e., 16 and 32) in benchmark testing of the TIMIT corpus. Also, the baseline models trained using maximum likelihood estimation (MLE) were quite poor or mismatched with the task, so that the method presented in that research showed some improvements. In [9], the authors optimized the parameters of acoustic and language models jointly using maximum entropy criterion, but string recognition error has not been explicitly targeted. In [7], the authors optimized the acoustic models on weighted finite-state transducers (WFSTs) with taking the language model score into consideration, but the language model parameters are assumed fixed. Similarly, the authors in [6] and [10] optimized the language model parameters with considering the acoustic model score, but assuming the acoustic model parameters as constants.

Despite the improvements achieved by that research, there is still more improvements can be achieved if we could optimize the acoustic and language models' parameters simultaneously. In this research, we engaged the optimization approaches presented in [6] and [7] into a single framework based on the MCE criterion to optimize the parameters of acoustic and language models jointly on a WFST-based decoding graph.

The gain from choosing the MCE as a criterion in the proposed framework is the direct minimization of word string errors which may yield a significant improvement in speech recognition performance. Another interesting approach in discriminative learning is maximum mutual information estimation (MMIE) [4], but this approach is derived from information theory rather than decision theory [7]. The standard learning approach used in MMIE models is usually based on MLE. However, MLE depends on several inaccurate assumptions i.e., modelling assumption, and thus, both MLE and MMIE cannot achieve an optimal classifier design [7]. There are other effective learning criteria, such as minimum phone/word error (MPE/MWE) criteria that share features with MCE criterion but with different formulations [5]. The advantage of MPE/MWE criteria is that they model phone/word accuracy, whereas MCE criterion models the string accuracy. Practically, we can view MPE/MWE as model-based estimates of the recognition accuracy in which the phoneme or word accuracies are explicitly weighted by the model-based posterior probabilities. This makes MPE/MWE criteria are also limited

by the modelling assumption. Whereas model-based estimate using MCE criterion does not require an accurate posterior probability, which makes MCE criterion advantageous over the other discriminative learning criteria [7].

III. MCE-BASED JOINT LEARNING

In this research, the joint models refer to both acoustic and language models. The acoustic models are modelled using context-dependent hidden Markov models (HMMs) consisting of N states, and each state contains Gaussian mixtures of K components. The parameter set of an HMM model, denoted by Λ, is: $\Lambda = \{A, c_{jk}, U_{jk}, R_{jk}\}$ where $A = [a_{sj}]$ is the state transition matrix with state indices denoted by s and j, c_{jk} is the weight for the k^{th} mixture component in the j^{th} state, $U_{jk} = [\mu_{jkl}]_{l=1}^{D}$ is the mean vector and R_{jk} is the corresponding covariance matrix which, for simplicity, is assumed diagonal, i.e. $R_k = [(\sigma_{jkl})^2]_{l=1}^{D}$. We also assumed that $X = (X_1, ..., X_t, ..., X_T)$ is D-dimensional feature vectors of length T.

In the proposed framework, we optimize only the mean and variance vectors from the parameter set of acoustic models. To maintain the constraints imposed on these mean and variance vectors during the parameter optimization, the following parameter transformations, $(\Lambda \rightarrow \tilde{\Lambda})$, are applied [11]:

$$\mu_{jkl} \rightarrow \tilde{\mu}_{jkl}, \quad where \;\; \tilde{\mu}_{jkl} = \frac{\mu_{jkl}}{\sigma_{jkl}} \tag{1}$$

$$\sigma_{jkl} \rightarrow \tilde{\sigma}_{jkl}, \quad where \;\; \tilde{\sigma}_{jkl} = log\sigma_{jkl} \tag{2}$$

On the other hand, the parameter set of the n-gram language model, denoted by Γ, is: $\Gamma = \{\psi(w_i), p(w_i|w_{i-1}), p(w_i|w_{i-1}, w_{i-2})\}$, where $\psi(w_i)$ is the back-off probability of the word w_i, $p(w_i|w_{i-1})$ is the bi-gram probability of the word sequence $w_{i-1}\ w_i$, and $p(w_i|w_{i-2}w_{i-1})$ is the tri-gram probability of the word sequence $w_{i-2}\ w_{i-1}\ w_i$.

These n-grams are integrated with other speech knowledge sources, such as pronunciation lexicon and context-dependent phonemes, into a decoding graph using a sequence of WFST operations same as those presented in [12]. The resulting decoding graph carries the n-gram probabilities as weights distributed on the transition arcs. Consequently, optimizing the transition weights of the decoding graph corresponds to optimizing the n-gram models.

Now, the MCE-based parameter optimization formula for the joint models, denoted by $\theta = \{\tilde{\Lambda}, \Gamma\} = \{\tilde{U}_{jk}, \tilde{R}_{jk}, \Gamma\}$, using the gradient probabilistic descent (GPD) [13] is:

$$\theta(n+1) = \theta(n) - \epsilon \frac{\partial l_n(X, \Lambda, \Gamma)}{\partial \theta(n)} \tag{3}$$

where n and ϵ are the learning iteration and step size, respectively, and $l(X, \Lambda, \Gamma)$ is the sigmoid class loss function which is defined as [14]:

$$l(X, \Lambda, \Gamma) = \frac{1}{1 + e^{-\alpha d(X,\Lambda,\Gamma) + \beta}} \tag{4}$$

where α and β are parameters used to control the slope and shift of the sigmoid function, respectively, and d is the score difference between the reference and competing hypotheses:

$$d(X, \Lambda, \Gamma) = -g(X, W_{ref}, \Lambda, \Gamma) + g(X, W_{best}, \Lambda, \Gamma) \tag{5}$$

where W_{ref} and W_{best} are the reference and best decoding hypotheses. g is the sum of the acoustic and language models' scores and is defined as $g(X, W, \Lambda, \Gamma) = log\ P(X|\Lambda) + \gamma\ log\ P(W|\Gamma)$ where γ is the language model scaling factor. The gradient part of Eq. (3) can be further written as:

$$\begin{aligned} \frac{\partial l_n(X, \Lambda, \Gamma)}{\partial \theta(n)} &= \frac{\partial l_n(X, \Lambda, \Gamma)}{\partial d_n(X, \Lambda, \Gamma)} . \frac{\partial d_n(X, \Lambda, \Gamma)}{\partial \theta(n)} \\ &= \alpha l_n(1 - l_n) . \frac{\partial d_n(X, \Lambda, \Gamma)}{\partial \theta(n)} \end{aligned} \tag{6}$$

A. Acoustic model optimization

As the parameters of the joint models are defined as $\theta = \{\tilde{\Lambda}, \Gamma\}$, the gradient of the update equation defined in Eq. (3) can be written in terms of the acoustic model parameters, Λ, as:

$$\begin{aligned} \frac{\partial l_n(X, \Lambda, \Gamma)}{\partial \tilde{\Lambda}(n)} &= \frac{\partial l_n(X, \Lambda, \Gamma)}{\partial d_n(X, \Lambda, \Gamma)} . \frac{\partial d_n(X, \Lambda, \Gamma)}{\partial \tilde{\Lambda}(n)} \\ &= \alpha l_n(1 - l_n) . \frac{\partial d_n(X, \Lambda, \Gamma)}{\partial \tilde{\Lambda}(n)} \end{aligned} \tag{7}$$

Using the above equation, the update formula for optimizing the Gaussian mean vectors of Λ can be obtained by partial derivative of $\frac{\partial d_n(X,\Lambda,\Gamma)}{\partial \tilde{\Lambda}(n)}$ with respect to $\{\tilde{\mu}_{jkl}\}$ as follows [11]:

$$\frac{\partial d_n(X, \Lambda, \Gamma)}{\partial \tilde{\mu}_{jkl}(n)} = \sum_{t=1}^{T} \delta(q_t - j) \frac{c_{jk} . b_{jk}(X)}{b_j(X)} \left(\frac{x_{tl}}{\sigma_{jkl}} - \tilde{\mu}_{jkl} \right) \tag{8}$$

where $\delta(.)$ is the Kronecker delta function, q_t is the state number at time t and

$$b_j(X) = \sum_{k=1}^{K} c_{j\,k} . b_{j\,k}(X) \tag{9}$$

is the observation distribution probability of the speech utterance X with respect to Gaussian mixtures at state j and $b_{jk}(X)$ is the probability of mixture component k which is:

$$b_{jk}(X) = \frac{1}{\sqrt{|(2\pi)^D \Sigma_{jk}|}} e^{\left(-\frac{1}{2}(x-\mu_{jk})^T \Sigma_{jk}^{-1} (x-\mu_{jk})\right)} \tag{10}$$

Similarly, the update formula for optimizing the Gaussian variance vectors of Λ is expressed in terms of the partial derivative of $\frac{\partial d_n(X,\Lambda,\Gamma)}{\partial \Lambda(n)}$ with respect to $\{\tilde{\sigma}_{jkl}\}$ as follows [11]:

$$\frac{\partial d_n(X, \Lambda, \Gamma)}{\partial \tilde{\sigma}_{jkl}(n)} = \sum_{t=1}^{T} \delta(q_t - j) \frac{c_{jk} . b_{jk}(X)}{b_j(X)} \left(\left(\frac{x_{tl} - \mu_{jkl}}{exp(\tilde{\sigma}_{jkl})} \right)^2 - 1 \right) \tag{11}$$

Then, using the GPD algorithm, the mean and variance vectors of Λ can be iteratively adjusted using the following update rule:

$$(\tilde{\Lambda})_{n+1} = (\tilde{\Lambda})_n - \epsilon_\Lambda \frac{\partial l_n(X, \Lambda, \Gamma)}{\partial \tilde{\Lambda}(n)} \tag{12}$$

where $\epsilon_\Lambda > 0$ is a preselected constant used to control the step size of the parameter update. From Eq. (12), we can see

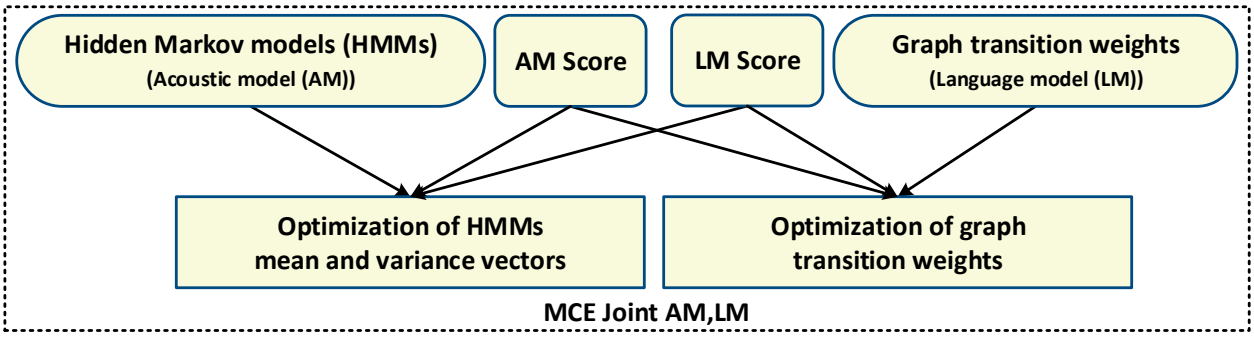

Fig. 1. *Block diagram of the proposed joint optimization process at each training iteration.*

that the acoustic model parameters $(\Lambda)_{n+1}$, at iteration $n+1$, are optimized in terms of both the acoustic parameters, $(\Lambda)_n$, and the language model score (embedded in $l_n(X,\Lambda,\Gamma)$), at iteration n. This process is depicted in Fig. 1. Finally, the inverse transformation is applied to restore the actual updated parameters as follows [11]:

$$(\mu_{jkl})_{n+1} = (\sigma_{jkl})_n.(\tilde{\mu}_{jkl})_{n+1} \tag{13}$$

$$(\sigma_{jkl})_{n+1} = exp(\tilde{\sigma}_{jkl})_{n+1} \tag{14}$$

B. Language model optimization

Back to Eq. (6), since $\theta = \{\tilde{\Lambda}, \Gamma\}$, the gradient of the update rule presented in Eq. (3) can be written in terms of the language model parameters, Γ, as:

$$\begin{aligned}\frac{\partial l_n(X,\Lambda,\Gamma)}{\partial \Gamma(n)} &= \frac{\partial l_n(X,\Lambda,\Gamma)}{\partial d_n(X,\Lambda,\Gamma)}.\frac{\partial d_n(X,\Lambda,\Gamma)}{\partial \Gamma(n)} \\ &= \alpha l_n(1-l_n).\frac{\partial d_n(X,\Lambda,\Gamma)}{\partial \Gamma(n)}\end{aligned} \tag{15}$$

Viewing the parameters of language model, Γ, as a vector of transition weights, then the partial derivative of Eq. (15) is [6]:

$$\frac{\partial d_n(X,\Lambda,\Gamma)}{\partial \Gamma(n)} = [-I(W_{ref},s) + I(W_{best},s)] \tag{16}$$

where $I(W_{ref},s)$ and $I(W_{best},s)$ refer to the number of occurrences of the transition weight, s, in the reference and best decoding hypotheses, respectively [6]. Then, using the GPD algorithm, the transition weights can be iteratively adjusted using the following update rule:

$$(s)_{n+1} = (s)_n - \epsilon_\Gamma \frac{\partial l_n(X,\Lambda,\Gamma)}{\partial \Gamma(n)} \tag{17}$$

where $\epsilon_\Gamma > 0$ is another preselected constant used to control the update value of the transition weights. Similarly, from Eq. (17), we can see that transition weights $(s)_{n+1}$, at iteration $n+1$, are optimized in terms of both the transition weights, s_n, and the score of acoustic model, Λ (embedded in $l_n(X,\Lambda,\Gamma)$), at iteration n, as shown in Fig. 1.

IV. Experiments

In this research, we conducted two sets of experiments aiming at evaluating the effectiveness of the proposed framework. In each set, five experiments were conducted namely "MLE", "MCE AM", "MCE LM", "MCE AM & MCE LM" and "MCE Joint AM/LM" for the evaluating MLE-based acoustic and language models (baseline [15]), MCE-trained acoustic models while fixing language models [7], MCE-trained language models while fixing acoustic models [6], combined MCE-trained acoustic and language model, and the proposed jointly optimized models, respectively. The first set of experiments is performed in terms of the TIMIT corpus [16], whereas the second set is performed on another corpus, the RM1 [17]. The following subsections discuss the experimental setup followed by a discussion on the recorded results.

A. Experimental setup

It has been asserted in [7], [18], [19] that training utterances containing out-of-vocabulary (OOV) words negatively affect the MCE-based parameter optimization, for this reason, we removed the training utterances containing OOV words before conducting the experiments. The training and test sets of TIMIT (after removing utterances containing OOV words) consist of $2,888$ and $1,015$ utterances, respectively. For the RM1, the numbers of training and testing utterances (after removing utterances containing OOV words) are, 865 and 371, respectively. The speech utterances are sampled at 16 kHz sampling rate with 16 bit quantization and framed using a Hamming window of length 30 ms and frame shift of 10 ms. The length of the feature vector extracted from each frame is 39 including static, energy and dynamic ($\triangle$ and $\triangle\triangle$) components. A set of baseline context-dependent acoustic models (based on $8,000$ states and 32 mixtures/state) along with tri-gram language models (containing $64,000$ uni-grams, $594,160$ bi-grams and $237,579$ tri-grams, after Kneser-Ney smoothing and pruning using SRILM toolkit [20]) are used in building the WFST-based decoding graph. These baseline models are freely available for research purposes and can at the time of writing be found at the location specified in [15]. The constructed WFST-based decoding graph has $6,223,933$ states and $9,092,597$ transitions. It is worth noting that, the following transducers were incorporated in the construction of

TABLE I
Recognition performance on the TIMIT complete test set. PER, WER and SER refer to phone, word and sentence error rates.

Approach	PER	PER Reduction	WER	WER Reduction	SER	SER Reduction
MLE (Baseline) [15]	18.67	-	22.27	-	67.23	-
MCE LM [6]	17.70	5.20	15.26	26.99	52.22	22.33
MCE AM [7]	14.73	21.10	16.01	28.11	54.79	18.50
MCE AM [6] & MCE LM [7]	16.51	11.57	13.70	38.48	47.29	29.66
Proposed MCE Joint AM/LM	16.28	12.80	13.37	39.96	45.11	32.90

TABLE II
Recognition performance on the RM1 test set. PER, WER and SER refer to phone, word and sentence error rates.

Approach	PER	PER Reduction	WER	WER Reduction	SER	SER Reduction
MLE (Baseline) [15]	27.04	-	33.57	-	81.62	-
MCE LM [6]	24.34	9.99	29.08	13.38	75.68	7.28
MCE AM [7]	24.15	10.69	29.98	10.69	78.11	4.30
MCE AM [7] & MCE LM [6]	25.70	4.96	26.51	21.03	74.32	8.94
Proposed MCE Joint AM/LM	21.30	21.23	25.73	23.35	73.24	10.27

the decoding graph: context-dependent phonemes (C), lexicon (L), tri-grams (G) and silence (T) transducers, using the following sequence of operations, $((C \circ det(L)).(G \circ T))$, where $\circ$, det and . refer to composition, determinization and lookahead composition operations, respectively [12]. Whereas, the acoustic models transducer (H) is not integrated with the decoding graph, but accessed on demand while decoding.

B. Results and discussion

Before experimenting with the GPD procedure, we performed a number of experiments to set the parameters of the sigmoid function; α and β, which were chosen as 0.01 and 0, respectively. The training set is used in tuning these parameters. Also, the step sizes of learning acoustic and language models, ϵ_Γ and ϵ_Λ, were chosen as 2.5 and 100, respectively, for the TIMIT training, and 10 and 100, respectively, for the RM1 training. The language model scaling factor, γ is set as 13. One way to select these values is to cut and try. In all experiments, the best decoding hypothesis (1-best) is considered in calculating the score difference of Eq. (5). In the both sets of experiments on the TIMIT and RM1, five iterations of the GPD procedure were conducted. The optimized models resulting from the five experiments were incorporated in the evaluation of the TIMIT complete test set and RM1 test set.

Table I presents the performance of the speech recognition using the trained models from each experiment. This performance is expressed in terms of phone error rate (PER), word error rate (WER) and sentence error rate (SER), along with the percentage of reduction with respect to the performance of the "MLE" baseline. Although PER using the jointly optimized models is better than that of "MCE AM & MCE LM", it is better than that of "MCE AM" and "MCE LM". This can be interpreted as due to the multiple pronunciation of the dictionary [15]. However, on the word and sentence levels, the

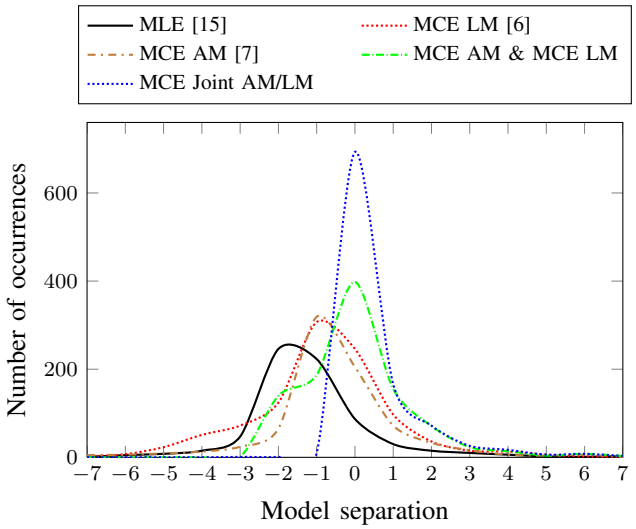

Fig. 2. Histogram of model separation calculated by MLE, MCE LM, MCE AM, MCE AM & MCE LM and MCE Joint AM/LM models on the TIMIT complete test set.

performance of the jointly optimized models outperforms that of all the other approaches. The experimental results from the second test set on the RM1 are shown in Table II. In this table, the performance of the jointly optimized models outperforms the other approaches and achieved significant improvements. Also, it is worth noting that the jointly optimized models achieved better than the combination of separately optimized models, "MCE AM & MCE LM", which emphasizes our expectation that the joint optimization benefits from the inherent correlation between acoustic and language models. Additionally, to emphasize the effectiveness of the proposed approach, the following logarithm of sentence posterior probability of training data (X, W) was examined:

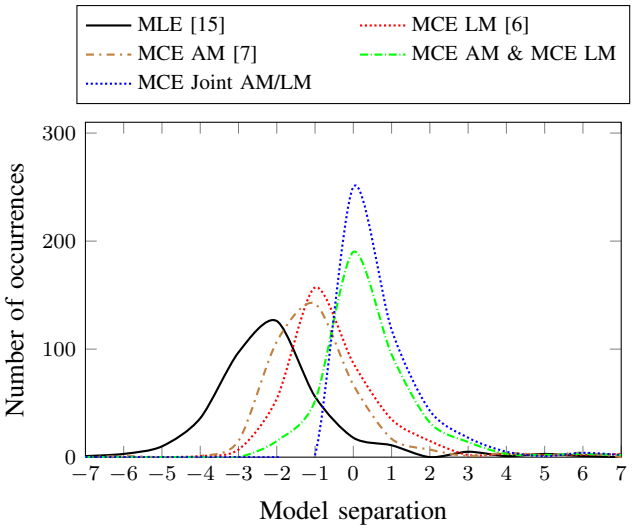

Fig. 3. Histogram of model separation calculated by "MLE", "MCE LM", "MCE AM", "MCE AM & MCE LM" and "MCE Joint AM/LM" models on the RM1 test set.

$$log\ p(W|X) \approx log\ [p(X|W_{ref}).p(W_{ref})] - log\ [p(X|W_{best}).p(W_{best})] \quad (18)$$

It was asserted in [9] that, the larger $log\ p(W|X)$ is measured, the bigger model separation between the reference and the competing strings is obtained. Figures 2 and 3 illustrate the histograms of the model separation from the four experiments on TIMIT and RM1, respectively. From these figures, we can note that, the distribution of "MCE Joint AM/LM" models is shifted rightward making a peak when compared to that of other approaches. Both of histogram right-shift and high peak refer to better optimization of the acoustic and language models [9].

V. Conclusion

In this research, a joint optimization framework is proposed for learning the parameters of acoustic and language models on WFST-based decoding graphs using the GPD procedure. The experimental results on TIMIT and RM1 emphasized the effectiveness of the proposed framework in reducing PER, WER and SER when compared with the conventional separate optimization approaches. Also, the model separation histogram of the jointly optimized models gave more emphasize on the validity of the proposed framework as it gives better model separation and thus better recognition performance. The proposed framework is generic, so we can use it in future work with various discriminative learning criteria.

VI. Acknowledgements

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] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition- A unifying review for optimization-oriented speech recognition," *IEEE Signal Processing Magazine*, vol. 5, pp. 14–36, 2008.

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

[3] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.

[4] Y. Normandin, R. Lacouture, and R. Cardin, "MMIE training for large vocabulary continuous speech recognition," in *Proceedings of International Conference on Spoken Language Processing*, vol. 3, 1994, pp. 1367–1370.

[5] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2004.

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

[7] E. McDermott, T. Hazen, J. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 203–223, 2007.

[8] S. Watanabe, T. Hori, E. McDermott, and A. Nakamura, "A discriminative model for continuous speech recognition based on weighted finite state transducers," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4922–4925.

[9] J. Chien and C. Chueh, "Joint acoustic and language modeling for speech recognition," *Transactions on Speech Communication*, pp. 223–235, 2010.

[10] 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.

[11] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[12] J. Novak, N. Minemaysu, and K. Hirose, "Painless WFST cascade construction for LVCSR-Transducersaurus," in *Proceedings of International Speech Communication Association (InterSpeech)*, 2011, pp. 1537–1540.

[13] 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.

[14] 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.

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

[16] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Transactions on Speech Communication*, vol. 9, pp. 351–356, 1990.

[17] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, pp. 651–654.

[18] 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.

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

[20] A. Stolcke, "SRILM: An extensible language modeling toolkit," in *Proceedings of International Conference on Spoken Language Processing*, 2002, pp. 901–904.