Discriminative Maximum Entropy Language Model for Speech Recognition

Chuang-Hua Chueh, To-Chang Chien and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{sgxxx,dosun,chien}@chien.csie.ncku.edu.tw

10.21437/Interspeech.2005-10

Abstract

This paper presents a new discriminative language model based on the whole-sentence maximum entropy (ME) framework. In the proposed discriminative ME (DME) model, we exploit an integrated linguistic and acoustic model, which properly incorporates the features from *n*-gram model and acoustic log likelihoods of target and competing models. Through the constrained optimization of integrated model, we estimate DME language model for speech recognition. Attractively, we illustrate the relation between DME estimation and the maximum mutual information (MMI) estimation for language modeling. It is interesting to find that using the sentence-level log likelihood ratios of competing and target sentences as the acoustic features for ME language modeling is equivalent to performing MMI discriminative language modeling. In the experiments on speech recognition, we show that DME model achieved lower word error rate compared to conventional ME model.

1. Introduction

Automatic speech recognition (ASR) has been increasingly important in many human computer interaction systems. The goal of speech recognition aims to find the optimal word sequence $\hat{s}$ from an observed speech signal X. This is a pattern classification problem which can be solved according to Bayesian decision theory

$$\hat{s} = \arg\max_{s} P(s|X) = \arg\max_{s} P(X|s)P(s), \qquad (1)$$

where $P(X|s)$ and $P(s)$ represent the probabilities using acoustic model and language model, respectively. In general, language model is used to characterize the linguistic regularities in natural language. In addition to the application of speech recognition, it is also popular to apply language models for machine translation, information retrieval and document classification.

As we know, the statistical *n*-gram model is effective for extracting the lexical regularities. The latent semantic language model [2] was presented to explore long distance word relations. In [4], the structured language model was developed to exploit the relevant syntactic regularities. Typically, different language model approaches characterized varying linguistic features in natural language. Based on the maximum entropy (ME) principle [3][11], we can integrate different features to establish optimal probability distribution model satisfying all available knowledge sources. ME principle was first applied to language modeling in [3]. Using ME principle, the information of trigger pairs representing semantically related two words was incorporated into *n*-gram model so that long distance dependencies with local lexical characteristics can be properly combined [11]. This principle was also adopted to build hybrid language models, which combine *n*-grams and topic information [6]. All of these approaches have been successfully applied to improve model perplexities and speech recognition rates.

Based on ME paradigm, an alternative ME language model called whole-sentence ME language model was proposed [12]. This model was suitable to characterize the linguistic regularities in sentence level. Because the whole-sentence ME model treated a sentence as a "bag of feature", the language model of a whole sentence was estimated directly. This approach was effective to extract the lexical relation between sentences or paragraphs. Different from conventional ME model, the whole-sentence ME had no need of computing normalization term. The computational cost was potentially economic. Either ME or whole-sentence ME models can be viewed as maximum likelihood (ML) models of log-linear or Gibbs distribution. However, ML training only concerned on maximizing the likelihood of training data, which could not guarantee the decrease of word error rates. Due to this reason, some discriminative training criteria, for example, minimum classification error (MCE) and maximum mutual information (MMI) [1][9] have been proposed for acoustic modeling as well as language modeling [8][10]. In [10], MCE criterion was applied to estimate parameters of Gibbs distribution for language modeling. The resulting ME model was discriminative naturally. In this paper, we propose a new discriminative ME language model for speech recognition. The features of sentence-level ratio of log likelihoods from competing and target models are merged in whole-sentence ME framework. Using the integrated linguistic and acoustic ME model, we can estimate discriminative parameters for ME language modeling. We also show the equivalence relation between DME and MMI language models.

2. Whole-Sentence ME Language Model

In whole-sentence maximum entropy (ME) language modeling, all available knowledge sources serve as the constraints to be integrated in sentence level. Under the ME framework, we maximize the entropy to find the optimal language model. Correspondingly, we restrict the estimated model to be consistent with all the information we have and simultaneously make the distribution in the model as uniform as possible. Let $f_1^{\mathrm{L}}, \cdots, f_F^{\mathrm{L}}$ denote a set of linguistic features, e.g. bigram or trigram, specifying the properties that we want to integrate in the desired model. Feature functions can be expressed by

$$f_i^{\mathrm{L}}(s) = \begin{cases} 1, & \text{if } s \text{ matches feature } i \\ 0, & \text{otherwise} \end{cases}. \qquad (2)$$

In *n*-gram modeling, bigram and trigram are viewed as the

features for ME modeling. The expectation of feature functions can be calculated with respect to empirical distribution $\tilde{p}(s)$ and actual distribution $p(s)$ in sentence level as follows

$$\tilde{p}(f_i^{\mathrm{L}}) = \sum_s \tilde{p}(s) f_i^{\mathrm{L}}(s) = \frac{1}{R}\sum_{r=1}^{R} f_i^{\mathrm{L}}(s_r) \tag{3}$$

$$p(f_i^{\mathrm{L}}) = \sum_s p(s) f_i^{\mathrm{L}}(s)\,, \tag{4}$$

where R is the number of training sentences. Because the desired model satisfies all constraints expressed by these features, the expectation functions should follow the equality

$$p(f_i^{\mathrm{L}}) = \tilde{p}(f_i^{\mathrm{L}}), \quad \text{for} \quad i = 1, \cdots, F\ . \tag{5}$$

To solve this constrained optimization problem, the Lagrange optimization procedure is adopted. The Lagrangian objective function $\Lambda(p, \lambda)$ is formed as

$$\Lambda_{\mathrm{ME}}(p, \lambda) = H(p) + \sum_{i=1}^{F} \lambda_i^{\mathrm{L}} \left[p(f_i^{\mathrm{L}}) - \tilde{p}(f_i^{\mathrm{L}}) \right], \tag{6}$$

where $\lambda = \{\lambda_i^{\mathrm{L}}\}$ is Lagrange multiplier and $H(p)$ is the entropy due to a model $p(s)$

$$H(p) = -\sum_s p(s) \log p(s)\,. \tag{7}$$

When maximizing the objective function $\Lambda(p, \lambda)$ with respect to $p(s)$, the whole-sentence ME language model is derived by [11][12]

$$p(s) = \frac{\exp\left(\sum_{i=1}^{F} \lambda_i f_i^{\mathrm{L}}(s) \right)}{\sum_{s'} \exp\left(\sum_{i=1}^{F} \lambda_i f_i^{\mathrm{L}}(s') \right)}\,. \tag{8}$$

Using this approach, we need to compute expectation values of feature functions. However, considering all possible sentences is infeasible in real implementation. We may use Gibbs sampling to deal with the problem. When calculating the expectation values via sampling, the generalized iterative scaling (GIS) algorithm [7] can be realized to estimate Lagrange multipliers $\lambda = \{\lambda_i^{\mathrm{L}}\}$. GIS algorithm is briefly described below.

Input: Feature functions $f_1^{\mathrm{L}}, \cdots, f_F^{\mathrm{L}}$, and empirical distribution $\tilde{p}(s)$.

Output: Optimal Lagrange multipliers $\hat{\lambda}$.

1. Initialization with $\lambda_i^{\mathrm{L}} = 0$ for all $i = 1, \cdots, F$.
2. For each $i = 1, \cdots, F$, update λ_i^{L} based on

$$\lambda_i^{\mathrm{L}} \leftarrow \lambda_i^{\mathrm{L}} + \frac{1}{F_i} \log \frac{\tilde{p}(f_i^{\mathrm{L}})}{p(f_i^{\mathrm{L}})}\,, \tag{9}$$

$$F_i = \frac{1}{\sum_{s'} p(s') f_i^{\mathrm{L}}(s')} \sum_s p(s) f_i^{\mathrm{L}}(s) \sum_{i'} f_{i'}^{\mathrm{L}}(s)\,. \tag{10}$$

3. Go to step 2 if λ_i^{L} has not converged.

After finding the optimal parameters $\hat{\lambda}$, we can calculate the ME language model using (8).

3. Discriminative ME Language Model

In general, ME can be considered as a maximum likelihood (ML) model using log-linear distribution. Namely, we can estimate the parameters by maximizing the likelihood function of training data. In general pattern recognition, using ML for model training can not guarantee good recognition performance. To tackle this issue, we propose a discriminative language model based on whole-sentence ME and illustrate its relation with MMI language model.

3.1. Acoustic features for ME estimation

To attain discriminability for language modeling, we build a new discriminative ME (DME) language model $p_{\mathrm{DME}}(s)$, which incorporates the acoustic information from the target models as well as the competing models. In additional to original linguistic features $f_1^{\mathrm{L}}, \cdots, f_F^{\mathrm{L}}$, for each training sentence, we further merge a new acoustic feature $f_X^{\mathrm{A}}(s)$ associated with the current speech signal X. Importantly, we extract the discriminative acoustic feature using the *sentence-level log likelihood ratio of competing and target sentences* as defined by

$$f_X^{\mathrm{A}}(s) = \begin{cases} \log \dfrac{p(X|s)}{p(X|s_X)} & \text{if } s \neq s_X \\ 0 & \text{if } s = s_X \end{cases} \tag{11}$$

where X is the input speech signal and $p(X|s_X)$ and $p(X|s)$ are the likelihood scores given the acoustic models of target sentence s_X and a competing sentence s, respectively. In particular, we don't calculate the expectation values $p(f_X^{\mathrm{A}})$ and $\tilde{p}(f_X^{\mathrm{A}})$ to determine Lagrange multiplier λ_X^{A} corresponding to acoustic feature $f_X^{\mathrm{A}}(s)$. In a meaningful way, the feature weight parameter λ_X^{A} is assigned by zero-one function

$$\lambda_X^{\mathrm{A}} = \begin{cases} 1 & \text{if } X \in \Omega \\ 0 & \text{if } X \notin \Omega \end{cases}\,. \tag{12}$$

Namely, we activate feature parameter to be one for those speech signals observed in training database $X \in \Omega$. Feature parameters are zeros for those unseen speech signals $X \notin \Omega$. Using this augmented set of feature functions $f^{\mathrm{LA}} = \{f_1^{\mathrm{L}}, \cdots, f_F^{\mathrm{L}}, f_X^{\mathrm{A}}\}$, the linguistic parameters in parameter set $\lambda^{\mathrm{LA}} = \{\lambda_1^{\mathrm{L}}, \cdots, \lambda_F^{\mathrm{L}}, \lambda_X^{\mathrm{A}}\}$ can be upgraded through GIS algorithm by substituting an integrated linguistic and acoustic model

$$p_{\mathrm{LA}}(s) = \frac{\exp\left(\sum_{i=1}^{F} \lambda_i^{\mathrm{L}} f_i^{\mathrm{L}}(s) + \lambda_X^{\mathrm{A}} f_X^{\mathrm{A}}(s) \right)}{\sum_{s'} \exp\left(\sum_{i=1}^{F} \lambda_i^{\mathrm{L}} f_i^{\mathrm{L}}(s) + \lambda_X^{\mathrm{A}} f_X^{\mathrm{A}}(s') \right)}\,, \tag{13}$$

into whole-sentence ME model of (9)(10). Notably, the sentence-level acoustic feature f_X^{A} measures the discriminability between target and competing sentences. Using the integrated model $p_{\mathrm{LA}}(s)$ in GIS algorithm, we accordingly upgrade $\lambda_1^{\mathrm{L}},\cdots,\lambda_F^{\mathrm{L}}$ to *discriminative linguistic* parameters $\lambda_1^{\mathrm{DL}},\cdots,\lambda_F^{\mathrm{DL}}$ and finally obtain the DME language model

$$p_{\mathrm{DME}}(s)=\frac{\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{DL}}f_i^{\mathrm{L}}(s)\right)}{\sum_{s'}\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{DL}}f_i^{\mathrm{L}}(s')\right)}, \tag{14}$$

for speech recognition. In (14), the acoustic feature $f_X^{\mathrm{A}}(s)$ is zero because each test speech signal serves as target sentence. The procedure of DME language model is shown in Figure 1. In general, we integrate the features of conventional *n*-gram and new discriminative acoustic features to estimate ME language model.

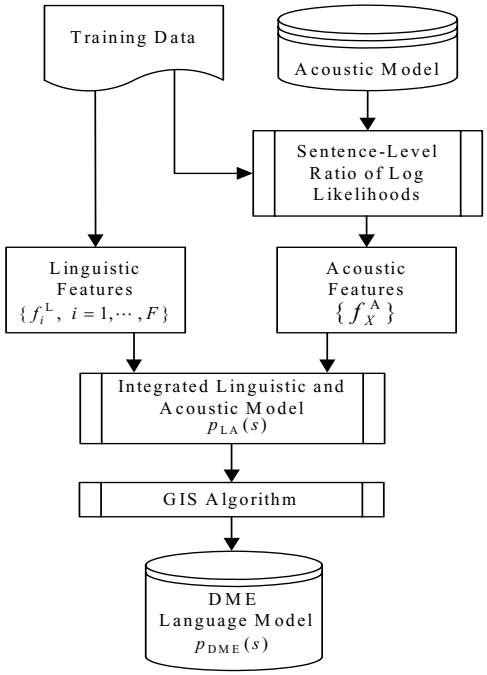

Figure 1: Implementation procedure of DME language model.

3.2. Relation between DME and MMI language models

As we know, the maximum mutual information (MMI) is a popular objective function for discriminative acoustic model training [1][9]. Using MMI, we maximize the mutual information between speech signal $X=\{X_r\}$ and its target word sequence $S=\{s_r\}$. MMI criterion is written by

$$\Lambda_{\mathrm{MMI}}=\log\frac{p(S,X)}{p(S)p(X)}=\log\frac{p(X|S)}{\sum_{S'}p(X|S')p(S')}. \tag{15}$$

In (15), the possible competing sentences S' appear in the denominator. Here, we modify MMI criterion by multiplying the numerator with a language model of target sentences $p(S)$ [9]. The resulting criterion becomes

$$\tilde{\Lambda}_{\mathrm{MMI}}=\log\frac{p(X|S)p(S)}{\sum_{S'}p(X|S')p(S')}=\sum_{r=1}^{R}\log\frac{p(X_r|s_r)p(s_r)}{\sum_{s'}p(X_r|s')p(s')} \tag{16}$$

which also measures the log posterior distribution. To investigate the relation of MMI and DME language models, here, we express ME model as a ML model using log-linear distribution with parameter λ. The optimal parameter $\hat{\lambda}$ is estimated by maximizing the integrated likelihood function of all sentences $S=\{s_r\}$

$$\begin{aligned}\hat{\lambda}&=\arg\max_{\lambda}\{\Lambda_{\mathrm{DME}}=\log p_{\mathrm{LA}}(S)=\sum_{r=1}^{R}\log p_{\mathrm{LA}}(s_r)\}\\&=\arg\max_{\lambda}\sum_{r=1}^{R}\log\frac{\exp\left(\sum_{i=1}^{F+1}\lambda_i^{\mathrm{LA}}f_i^{\mathrm{LA}}(s_r)\right)}{\sum_{s_r'}\exp\left(\sum_{i=1}^{F+1}\lambda_i^{\mathrm{LA}}f_i^{\mathrm{LA}}(s_r')\right)}\\&=\arg\max_{\lambda}\sum_{r=1}^{R}\log\frac{\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{L}}f_i^{\mathrm{L}}(s_r)+\lambda_{X_r}^{\mathrm{A}}f_{X_r}^{\mathrm{A}}(s_r)\right)}{\sum_{s'}\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{L}}f_i^{\mathrm{L}}(s')+\lambda_{X_r}^{\mathrm{A}}f_{X_r}^{\mathrm{A}}(s')\right)}\end{aligned}. \tag{17}$$

Using the assignment of (11) and (12), we obtain

$$\begin{aligned}\Lambda_{\mathrm{DME}}&=\sum_{r=1}^{R}\log\frac{\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{L}}f_i^{\mathrm{L}}(s_r)+\log p(X_r|s_r)\right)}{\sum_{s'}\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{L}}f_i^{\mathrm{L}}(s')+\log p(X_r|s')\right)}\\&=\sum_{r=1}^{R}\log\frac{p(X_r|s_r)\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{L}}f_i^{\mathrm{L}}(s_r)\right)}{\sum_{s'}p(X_r|s')\exp\left(\sum_{i=1}^{F}\lambda_i^{\mathrm{L}}f_i^{\mathrm{L}}(s')\right)}\\&=\sum_{r=1}^{R}\log\frac{p(X_r|s_r)p(s_r)}{\sum_{s'}p(X_r|s')p(s')}=\tilde{\Lambda}_{\mathrm{MMI}}\end{aligned}. \tag{18}$$

It is attractive to see that the discriminative ME criterion is equivalent to the MMI criterion for language modeling. Through the integrated linguistic and acoustic ME modeling, we can find that the discriminative parameters $\lambda_1^{\mathrm{DL}},\cdots,\lambda_F^{\mathrm{DL}}$ maximizing $p_{\mathrm{LA}}(S)$ are in accordance with the MMI criterion. In Table 1, we compare some properties of using MMI and DME criteria. Under the assumption of log-linear distribution with parameter λ, the optimal $\hat{\lambda}_{\mathrm{MMI}}$ can be estimated according to MMI criterion. The corresponding MMI language model $p_{\mathrm{MMI}}(s)$ is determined following the unconstrained optimization procedure. On the other hand, DME modeling performs the constrained optimization. Under the extended constraints, DME criterion allows us to obtain the integrated linguistic and acoustic model $p_{\mathrm{LA}}(s)$ with the highest entropy. The whole-sentence based DME model $p_{\mathrm{DME}}(s)$ can be obtained. Interestingly, these two estimation methods achieve the same objective function.

Table 1: Relation between MMI and DME criteria

	MMI	**DME**
Criterion	Maximize $\sum_{r=1}^{R} \log \frac{p(X_r \mid s_r) p(s_r)}{\sum_{s'_r} p(X_r \mid s'_r) p(s'_r)}$	Maximize $-\sum_{r=1}^{R} \sum_{s_r} p(s_r) \log p(s_r)$
Type of search	Unconstrained optimization	Constrained optimization with $p(f_i^{\mathrm{LA}}) = \tilde{p}(f_i^{\mathrm{LA}})$
Solution	$p_{\mathrm{MMI}}(s)$	$p_{\mathrm{LA}}(s)$ composed by $\{\lambda_i^{\mathrm{L}}, \lambda_X^{\mathrm{A}}\}$
$p_{\mathrm{MMI}}(s) = p_{\mathrm{DME}}(s)$ composed by $\{\lambda_i^{\mathrm{DL}}\}$		

4. Experiments

In the experiments, the benchmark speech corpus TCC300 [5] was used to establish speaker-independent hidden Markov models for Mandarin speech. Referring to [5], each Mandarin syllable was modeled by right context dependent states. Each state had 32 mixture components at most. Feature vectors consisted of twelve Mel-frequency cepstral coefficients, one log energy and their first derivation. We selected 4,200 sentences (about 4 hours) uttered by 100 speakers for model training and 450 sentences were used for testing. Language models were estimated using training sentences of TCC300. The estimated models were interpolated with the *n*-gram models trained from Academia Sinica CKIP balanced corpus, which was composed of about five million words with a vocabulary of 32,909 words. We carried out baseline *n*-gram, whole-sentence ME [12] and proposed whole-sentence DME for comparison. In the implementation of ME modeling, we used a large set of training sentences to compute the normalization (denominator) term in ME model.

The baseline *n*-gram, ME and DME language models were evaluated through the supervised N-best rescoring. For simplification, N-best lists were constructed using the decoded hypotheses and corresponding transcriptions. The final recognition result was obtained by rescoring the lists using language model. In Table 2, we report word error rate (%) and corresponding error rate reduction (%) using different models. Baseline system obtains WER of 14.85 %. Using ME and DME language models, WER's are reduced to 13.05% and 12.69%, respectively. Especially, the error rate reduction 14.54% can be achieved when using DME language model. This shows that the proposed DME language model is able to alleviate the confusion between target and competing sentences in speech recognition compared to baseline and ME language models. We have obtained promising preliminary results. We are working on different implementation procedures to investigate varying properties of using DME. Also, we are evaluating the experimental and theoretical difference between DME and MCE language models.

Table 2: Word error rates (%) of different language models

	Baseline	ME	DME
WER	14.85	13.05	12.69
Error rate reduction	-	12.12	14.54

5. Conclusions

We have presented a new ME language model integrating linguistic and acoustic features for speech recognition. New approach was developed under the whole-sentence ME framework where two streams of features were extracted in sentence level that the model computation could be efficient. Interestingly, the acoustic features had the form containing log likelihoods from target and competing models. The derived ME language model was inherent with discriminability power. Such approach was different from ML based modeling, which did not guarantee the decrease of model confusion. In this paper, the relation between DME model and MMI model was also illustrated and compared. Specially, DME model involved a constrained optimization procedure and was powerful for knowledge integration. In preliminary speech recognition results, we obtained desirable performance compared to baseline *n*-gram and conventional ME language model.

6. References

[1] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", In *Proc. of ICASSP*, vol. 1, pp. 49-52, 1986.

[2] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279-1296, 2000.

[3] A. Berger, S. Della Pietra and V. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, vol. 22, no. 1, pp. 39-71, 1996.

[4] C. Chelba and F. Jelinek, "Structured language modeling", *Computer Speech and Language*, vol. 14, no. 4, pp. 283-332, October 2000.

[5] J.-T Chien and C.-H. Huang, "Bayesian learning of speech duration models", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 558-567, 2003.

[6] C.-H. Chueh, J.-T. Chien and H..Wang, "A maximum entropy approach for integrating semantic information in statistical language models", In *Proc. of ISCSLP*, pp. 309-312, Hong Kong, 2004.

[7] J. Darroch and D. Ratcliff. "Generalized iterative scaling for log-linear models", *The Annals of Mathematical Statistics*, vol. 43, pp. 1470-1480, 1972.

[8] H. J. Kuo, E. Fosler-Lussier, H. Jiang and C.-H. Lee, "Discriminative training of language models for speech recognition", In *Proc. of ICASSP*, pp. 325-328, 2002.

[9] Y. Normandin, R. Cardin and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation", *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 299-311, 1994.

[10] C. Paciorek and R. Rosenfeld, "Minimum classification error training in exponential language models", In *Proc. of NIST/DARPA Speech Transcription Workshop*, 2002.

[11] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling", *Computer Speech and Language*, vol. 10, pp. 187-228, 1996.

[12] R. Rosenfeld, S. F. Chen and X. Zhu, "Whole-sentence exponential language models: a vehicle for linguistic-statistical integration", *Computer Speech and Language*, vol. 15, pp. 55-73, 2001.