INTERDEPENDENCE OF LANGUAGE MODELS AND DISCRIMINATIVE TRAINING

R. Schlüter, B. Müller, F. Wessel, H. Ney

Lehrstuhl für Informatik VI, RWTH Aachen – University of Technology
Ahornstraße 55, 52056 Aachen, Germany
schlueter@informatik.rwth-aachen.de

ABSTRACT

In this paper, the interdependence of language models and discriminative training for large vocabulary speech recognition is investigated. In addition, a constrained recognition approach using word graphs is presented for the efficient determination of alternative word sequences for discriminative training. Experiments have been carried out on the ARPA *Wall Street Journal* corpus. The recognition results for MMI training show a significant dependence on the context length of the language model used for training. Best results were obtained using a unigram language model for MMI training. No significant correlation has been observed between the language model choice for training and recognition.

1. INTRODUCTION

Since the first publication on discriminative training for speech recognition [1], many authors have shown the improvements that are obtainable using discriminative training in comparison to *Maximum Likelihood* (ML) training. Considerable effort has been dedicated to discriminative training of isolated word and small vocabulary continuous speech recognizers (e.g. [1, 4, 8]). There also exist a number of publications investigating discriminative training for large vocabulary continuous speech recognition (LVCSR) [2, 3, 7, 11].

In contrast to isolated word and small vocabulary continuous speech recognition, LVCSR introduces additional problems to discriminative training. Firstly, there is the problem of determining representative sets of alternative word sequences for each training iteration. For small vocabulary applications, this is usually done by full recognition on the training data. On the other hand, for LVCSR it would be very time consuming to perform full recognition in every training iteration. Therefore it has been proposed to perform full recognition only once to produce either N-best lists [3] or word graphs [7, 11], which subsequently are used to restrict the search space in every discriminative training iteration. In this work, an efficient constrained search algorithm is proposed, that restricts recognition to word graphs initially obtained on the training data, while retaining the advantages of a tree structured pronounciation lexicon.

As a further aspect, discriminative training introduces language models to training in several views. Firstly, the language model for the - at least initial - recognition of alternative word sequences for training has to be chosen. Secondly, the choice of language models for discriminative training itself will have impact on the resulting acoustic models. Finally, the question arises to what extent recognition results using a particular language model depend on the language models chosen for training. In a *Minimum Classification Error* (MCE) training approach with a vocabulary of 1000 words, using no language model for training at all gave better results than using a word pair grammar, where in both cases a word pair grammar was used for evaluation [2]. In [11] a bigram language model was used for MMI training of a speech recognizer with 65k vocabulary. Clearly, improvements in comparison to the baseline ML results diminished with increasing context length of the language model for recognition. In this work, systematic investigations on the interdependence between language model choice for MMI training and recognition are presented. It is shown that the recognition performance of the MMI trained models significantly depend on the choice of the language model context length used for *training*. Moreover, results are presented that do not indicate considerable correlation between the choice of language models for training and recognition.

The rest of the paper is organized as follows. In Section 2 our basic approach to MMI training is summarized. In Section 3 the determination of alternative word sequences is discussed and an efficient restricted recognition approach using word graphs is presented. The interdependence of language models and discriminative training is discussed in Section 4 and the corresponding experiments on the ARPA *Wall Street Journal* (WSJ) corpus are presented in Section 5. The paper is closed by the conclusions in Section 6.

2. DISCRIMINATIVE TRAINING

In this section we will introduce the discriminative methods applied here to large vocabulary discriminative training, which are based on the MMI criterion [1, 11]. Let the training data be given by training utterances r with $r = 1...R$, each consisting of a sequence X_r of acoustic observation vectors $x_{r1}, x_{r2}, ..., x_{rT_r}$ and the corresponding sequence W_r of spoken words. The *a posteriori* probability for the word sequence W_r given by the acoustic observation vectors X_r shall be denoted by $p_\theta(W_r|X_r)$. Similarly, $p_\theta(X_r|W_r)$ and $p(W_r)$ represent the corresponding emission and language model probabilities respectively. In the following, the language model probabilities are supposed to be given. Hence the parameter θ represents the set of all parameters of the emission probabilities $p_\theta(X_r|W_r)$. Finally, let $\mathcal{M}_r$ denote the set of alternative word sequences. The MMI criterion $\mathcal{F}_{\text{MMI}}$ could then be defined by:

$$\begin{aligned}\mathcal{F}_{\text{MMI}}(\theta) &= \sum_{r=1}^{R} p_\theta(W_r|X_r) \qquad (1)\\ &= \sum_{r=1}^{R} \log p_\theta(X_r|W_r) - \log \sum_{W \in \mathcal{M}_r} \frac{p(W)}{p(W_r)} p_\theta(X_r|W).\end{aligned}$$

Ideally, the set $\mathcal{M}_r$ has to contain all possible word sequences. In practice, $\mathcal{M}_r$ is approximated by those word sequences, which significantly compete with the spoken word sequences, including the spoken word sequence itself. In the *corrective training* (CT) approximation to the MMI criterion, the set of alternative word sequences is reduced to the best recognized word sequence only. For this case, the language model dependence cancels out, i.e. the language model dependent term could be separated from the emission probabilities and thus has no effect on MMI training. Furthermore, CT only considers those utterances for training, which are misrecognized. In our experiments on small vocabulary speech recognition [10] we found that MMI training outperforms CT. In addition, for both MMI and CT the main computational complexity is produced by the recognition step. Therefore, we chose the MMI criterion for our investigations on discriminative training for LVCSR.

An optimization of the MMI criterion tries to simultaneously maximize the emission probabilities of the spoken spoken word sequences of the training corpus and to minimize the corresponding sum over the emission probabilities of each alternative word sequence weighted by its language model probability relative to the spoken word sequence. Thus, MMI training optimizes class separability according to the words under consideration of the language model. Since the language model is supposed to be given, it has to be chosen according to optimal training performance. For ML as well as MMI training the *Viterbi* approximation was applied for state alignment. Methods for parameter optimization and convergence control as well as efficient estimation of discriminative statistics using word graphs were taken over from previous work on discriminative training for small vocabulary speech recognition [10].

3. CONSTRAINED RECOGNITION USING WORD GRAPHS

Discriminative training always involves the definition of a competing model. For MMI training this model is defined by a sum over the set of alternative word sequences, $\mathcal{M}_r$ (cf. Eq. (1)). Because of the combinatorial complexity, it is certainly unrealistic to include all possible word sequences with all possible word boundaries, especially for large vocabulary applications. Therefore, alternative word sequences usually are determined by a recognition pass on the training data.

For discriminative training with small vocabulary we usually perform unconstrained recognition every iteration step. For LVCSR applications, unconstrained recognition for whole training corpora in every iteration of discriminative training would clearly be unrealistic by means of computation time. In [11], discriminative training using the WSJ SI-284 training corpus is reported, where unconstrained recognition was performed only once in order to produce an initial word lattice, which was then used for constrained recognition in each iteration step of discriminative training. Preliminary experiments for discriminative training applying acoustic and language model rescoring on word graphs with fixed boundary times showed only little effect or even degradations in performance as shown in Table 2. In consequence, we developed a method of constrained recognition, where the boundary times are relaxed to intervals around the boundary times given by the word graph. At each time frame τ where new word hypotheses are to be started, not only the word hypotheses starting at exactly this time frame in the word graph are allowed in this approach, but also those words starting at time frames in the vicinity of time frame τ defined by the interval $[\tau - \Delta\tau, \tau + \Delta\tau]$, as shown by a section of a word graph in Fig.1. The successor word candidates thus obtained from the word graph are then used to reduce the possible search space of the recognizer by constraining the lexical tree, as illustrated in Fig. 2. This method [illegible]ognition even allows for

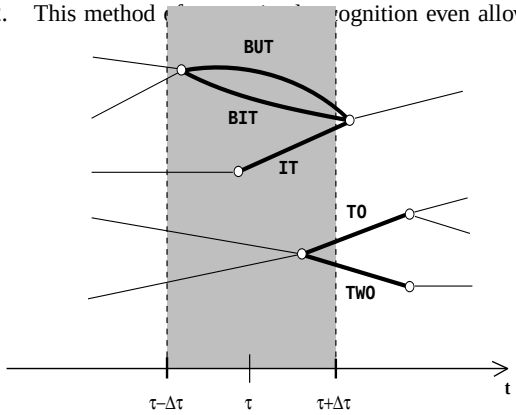

Figure 1: Constrained recognition: words of the word graph, which are allowed to be started at time τ.

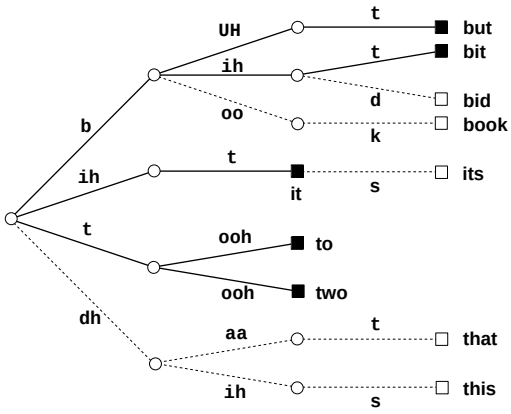

Figure 2: Lexical prefix tree for constrained recognition. Allowed words are marked with filled squares. Inactive arcs are drawn with dashed lines.

recognition of new word sequences not originally represented by the according word graph, which would not be produced by simple acoustic or language model rescoring on the word graph, because boundary times of subsequent word hypotheses might not match. In addition the approach still takes advantage of the efficiency of a tree lexicon. In our experiments a time interval of 11 frames was used, i.e. $\Delta\tau = 5$. In order to further reduce computation time, the *Viterbi* state alignment paths from constrained recognition were saved on disk, such that it was not necessary to estimate them again word-wise for accumulation of statistics.

4. CHOICE OF LANGUAGE MODELS

From the definition of the MMI criterion it is not at all clear, what the best choice of language models for MMI training would be. Firstly, there are three levels, at which the choice of language models might be important:

1. the determination of alternative word sequences;
2. the MMI criterion itself; and
3. the correlation between training and recognition.

The first aspect should not have any considerable effect. In the worst case, a non-matching language model for the recognition of alternative word sequences would lead to missing word sequences in the word graphs, which should not cause any problem, if the word graph densities are high enough. The second point should certainly be significant, since the acoustic parameters obtained by MMI training directly depend on the language model. It is not clear, what effect different language models will have on MMI training; and if there are any correlations between the language models used for training and those used for recognition on unseen test data. For MMI training, it could easily be shown that the contributions of parts of training utterances decrease with increasing probability difference to corresponding competing parts. This applies for whole sentences, as well as words or even single HMM-states. Therefore, two diametrical hypotheses are conceivable:

Correlation hypothesis: With respect to the recognition situation, one would expect that only those acoustic models need optimization, which do not sufficiently discriminate between correct and incorrect word sequences. If this argument holds, a strong correlation between the language models chosen for training and evaluation has to be concluded.

Covering hypothesis: With respect to the quality of the acoustic model, the language model usually largely improves the recognition accuracy and might cover or lead away from deficiencies of the acoustic models. Such an effect would call for suboptimal language models for training. Moreover, the choice of language models for training should not considerably correlate with those chosen for evaluation.

5. EXPERIMENTS

In order to investigate the interdependence of language models and discriminative training, experiments have been performed on the ARPA *Wall Street Journal* (WSJ) corpus. The main properties of the corresponding recognition system are summarized as follows. Training was done on the WSJ0 84-speaker corpus (15h speech) and testing on the WSJ0 Nov. '92 development and evaluation test sets. The recognition lexicon contained 4986 words plus 668 pronunciation variants plus silence. For state tying the number of 23509 triphone states was reduced to 2001 (including silence) by a decision tree based method. Acoustic features were given by mel-frequency cepstral coefficients, which were transformed by linear discriminant analysis (LDA). Acoustic emission probabilities were given by Gaussian mixture densities with approx. 96k densities and one pooled diagonal covariance. Further details on the baseline RWTH large vocabulary continuous speech recognition system could be found in [5].

The number of different words observed in the training corpus is more than twice the number of words contained in the recognition lexicon. Therefore these words had to be added to the recognition lexicon *for discriminative training*, which contains 10108 words plus 668 pronounciation variants. This presented an additional problem: About half of the words of the training recognition lexicon are unknown to the language models for recognition. Preliminary tests with special language models for discriminative training did not produce improvements using the original language models on the test corpora. Therefore, all words, which were unknown to the language model for recognition, were mapped to the unknown word class, which was renormalized according to the number of words included into it. As a consequence, the language model perplexities on the training corpus were significantly higher, than those on the test corpora. The perplexities of all language models used for the corresponding corpora are summarized in Table 1. All discriminative training experiments presented here

Table 1: Language model perplexities: ARPA WSJ0 training and testing corpora. The notations "bi-phr" and "tri-phr" refer to language models containing phrases/multiwords.

corpus	perplexity					
	zero	uni	bi	bi-phr	tri	tri-phr
Training	10110	1372	398	–	289	–
Nov. '92 Dev.	–	–	107	94	58	54
Nov. '92 Eval.	–	–	107	91	53	48

were initialized with the parameters obtained by a standard ML training. These initial parameters were also used to perform an unconstrained recognition pass on the WSJ0 training data. Word graphs [9] with a word graph density of 44 were produced, which were used for word graph rescoring or constrained recognition in every training iteration. Table 2 shows recognition results for MMI training with rescoring and constrained recognition in comparison to the initial ML results. Clearly, the determination of alternative word sequences using constrained recognition performs better than word graph rescoring, since the word boundaries from the initial word graphs are left unchanged by rescoring. Therefore, constrained recognition was chosen in all subsequent experiments on MMI training presented here. As shown in Table 3, the constrained recognition algorithm reduced the corresponding recognition time by a factor of more than 5, resulting in an RTF of 1.9 on an ALPHA 5000 PC. Including the calculation of word probabilities and

Table 2: Comparison of rescoring and constrained recognition using word graphs for the determination of alternative word sequences during discriminative training. Results on ARPA WSJ0 Nov. '92 corpus, training and recognition with bigram language model.

training criterion	determination of alternative word sequences	word error rates [%]		
		dev	eval	dev& eval
ML	–	6.91	6.78	6.86
MMI	rescoring	6.96	6.41	6.72
	constrained recogn.	6.71	6.20	6.48

Table 3: Comparison of full (unrestricted) recognition and constrained recognition using word graphs with $\Delta\tau = 5$. Recognition with bigram language model. The search space is indicated by the numbers of state, arc, tree, and word hypotheses. The real time factors (RTF) correspond to an ALPHA 5000 PC. Results on the ARPA WSJ0 Nov. '92 corpus.

recognition method	search space: number of				WER [%]	RTF
	states	arcs	trees	words		
full	6472	1835	36	106	6.86	10.5
constrained	989	239	17	67	6.86	1.9

the reestimation process, a single iteration step of MMI training on the ARPA WSJ0 training corpus took about 1.5 days resulting in an RTF of about 2.3 on an ALPHA 5000 PC.

In order to check the hypotheses on the interdependence of language models and discriminative training stated in Section 4, experiments using language models of varying context length for training and recognition were performed on the WSJ0 corpus, as shown in Table 4. The initial recognition and the constrained recognition for the trigram training has been performed using the trigram language model, and the constrained recognitions for the zerogram, unigram and bigram training were performed using the bigram. In order to distinguish the recognition for MMI training from the test set recognition, the latter will be referred to as 'test' in the following. For testing with either bigram or trigram language

Table 4: Comparison of several language models for MMI training and recognition. Results on ARPA WSJ0 Nov. '92 corpus.

language models		criterion	word error rates[%]		
test	training		dev	eval	dev& eval
bi	–	ML	6.91	6.78	6.86
	zero	MMI	6.71	6.03	6.41
	uni		**6.59**	**6.00**	**6.33**
	bi		6.71	6.20	6.48
	tri		6.87	6.54	6.72
tri	–	ML	4.82	4.11	4.51
	zero	MMI	4.63	4.05	4.38
	uni		**4.30**	**3.64**	**4.01**
	bi		4.48	3.94	4.24
	tri		4.58	4.00	4.33
bi-phrase	–	ML	6.40	5.79	6.13
	bi	MMI	5.91	5.60	5.78
tri-phrase	–	ML	4.76	4.26	4.54
	bi	MMI	4.48	4.07	4.30

models, clearly the best results are obtained using a unigram language model for MMI training resulting in relative improvements of up to 11% in word error rate. Moreover, for testing with the bigram, the results for training with the trigram language model are even worse than those for training with the zerogram. Even for testing with the trigram, the results for training with the trigram language model are only slightly better than those for training with the zerogram. Best results were obtained using a unigram language model for MMI training, which resulted in a word error rate of 4.01% using a trigram language model for testing.

In another experiment, the correlation between the language models chosen for training and testing was examined. As shown in Table 4, in comparison to ML training the improvements obtained by MMI training using a bigram language model for training remained approximately the same for testing with a bigram, trigram, phrase-bigram and phrase-trigram language model. For these cases, the relative improvements in word error rate in comparison to ML training ranged between 5 and 6%.

It should be noted that both sets of experiments clearly support the covering hypothesis as stated in Section 4. It suggests that language models, which are too accurate are in fact able to cover deficiencies of acoustic models by weigthing down their contributions from MMI training. Moreover, the experiments presented here indicate that the improvements obtained by discriminative training using a particular language model are fairly independent of the choice of language model for evaluation.

6. CONCLUSION

In this work, discriminative training for large vocabulary speech recognition has been investigated with special reference to its interdependence with the choice of language models for training and recognition. In addition, an efficient approach for constrained recognition using word graphs has been presented, which reduces the time for the determination of alternative word sequences for MMI training by a factor of more than 5 in comparison to unconstrained recognition. Experiments were performed on the ARPA WSJ0 corpus. Best results were obtained using a unigram language model for MMI training. Using a trigram language model for recognition, a relative improvement of 11% was obtained in comparison to ML training leading to a word error rate of 4% on the test data. No significant correlation between the choice of language models for training and recognition has been observed.

Acknowledgement. This work was partly supported by Siemens AG, Munich.

7. REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer. "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," Proc. *1986 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 49-52, Tokyo, Japan, May 1986.

[2] W. Chou, C.-H. Lee, B.-H. Juang. "Minimum Error Rate Training based on N-Best String Models," Proc. *1993 Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 652-655, Minneapolis, MN, April 1993.

[3] Y.-L. Chow. "Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition using the N-best Algorithm," Proc. *1990 Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 701-704, Albuquerque, NM, April 1990.

[4] B.-H. Juang, W. Chou, C.-H. Lee. "Statistical and Discriminative Methods for Speech Recognition," *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, K. K. Paliwal (eds.), pp. 109-132, Kluwer Academic Publishers, Norwell, MA, 1996.

[5] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel. "The RWTH Large Vocabulary Continuous Speech Recognition System," Proc. *1998 Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 853-856, Seattle, WA, April 1998.

[6] Y. Normandin, R. Cardin, R. De Mori. "High–Performance Connected Digit Recognition Using Maximum Mutual Information Estimation," IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 2, pp. 299-311, April 1994.

[7] Y. Normandin, R. Lacouture, R. Cardin. "MMIE Training for Large Vocabulary Continuous Speech Recognition," Proc. *1994 Int. Conf. on Spoken Language Processing*, Vol. 3, pp. 1367-1370, Yokohama, September 1994.

[8] Y. Normandin. "Maximum Mutual Information Estimation of Hidden Markov Models," *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, K. K. Paliwal (eds.), pp. 57-81, Kluwer Academic Publishers, Norwell, MA, 1996.

[9] S. Ortmanns, H. Ney, X. Aubert. "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition," Computer, Speech and Language, Vol. 11, No. 1, pp. 43-72, January 1997.

[10] R. Schlüter, W. Macherey, B. Müller, H. Ney. "A Combined Maximum Mutual Information and Maximum Likelihood Approach for Mixture Density Splitting," to appear in Proc. *1999 Europ. Conf. on Speech Communication and Technology*, Budapest, Hungary, September 1999.

[11] V. Valtchev, J. J. Odell, P. C. Woodland, S. J. Young. "MMIE Training of Large Vocabulary Recognition Systems," Speech Communication, Vol. 22, No. 4, pp. 303-314, September 1997.