Log-linear Model Combination with Word-dependent Scaling Factors

Björn Hoffmeister, Ruoying Liang, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University, Germany,
hoffmeister@cs.rwth-aachen.de

Abstract

Log-linear model combination is the standard approach in LVCSR to combine several knowledge sources, usually an acoustic and a language model. Instead of using a single scaling factor per knowledge source, we make the scaling factor word- and pronunciation-dependent. In this work, we combine three acoustic models, a pronunciation model, and a language model for a Mandarin BN/BC task. The achieved error rate reduction of 2% relative is small but consistent for two test sets. An analysis of the results shows that the major contribution comes from the improved interdependency of language and acoustic model.
Index Terms: speech recognition, model combination, system combination, log-linear modeling, minimum risk training

1. Introduction

Nowadays, the standard approach to large vocabulary continuous speech recognition (LVCSR) is to combine several knowledge sources in a log-linear model. In this approach each knowledge source gets a scaling factor (the exponent in the log-linear model) which is optimized in a discriminative manner: either by direct error minimization on a tuning set or by optimizing some objective function on a training set [1, 2]. The models of the knowledge sources are usually trained independently and the task of the log-linear model combination is to capture the dependencies between the individual models.

In this work we aim at improving the log-linear model combination by introducing word- and pronunciation-dependent scaling factors, instead of using a single scale per model. The idea is that the additional parameters can better describe the interdependency of the individual models. For example, a word-dependent scale can adjust the impact of the language model for a given word depending on how well the acoustic model can discriminate the pronunciation of the word. A detailed description of the log-linear model combination and the dependencies we try to capture as well as a short discussion of related work is given in the next section.

We apply our approach to a Mandarin broadcast news and conversations (BN/BC) task. The log-linear model combination consists of several acoustic models, a pronunciation model, a language model, and word penalties. We train the scaling factors on a separate 120h set using minimum risk training. The models are applied in a lattice rescoring followed by a confusion network (CN) decoding on character level. We also present experiments with character- and syllable-dependent scaling factors, which reduce the problem of data sparseness and are consistent with the character level CN decoding. The models and the experimental setup are described in Section 3.

Experimental results and an analysis of the results are presented in Section 4 and 5. The last section draws conclusions and gives an outlook.

2. Log-linear Model Combination with Word-dependent Scaling Factors

The general form of the log-linear model we use is shown in Equation (1). It consists of a set of word-level feature functions $f_i(w_n;\cdot)$ and a corresponding set of word-dependent scaling factors $\lambda_i(w_n)$.

$$p_\Lambda(w_1^N|x_1^T) := \frac{\exp(\sum_n \sum_i \lambda_i(w_n) f_i(w_n; w_1^N, x_1^T))}{\sum_{v_1^M} \exp(\sum_m \sum_i \lambda_i(v_m) f_i(v_m; v_1^M, x_1^T))} \quad (1)$$

In the following we assume that for each word its acoustic spelling is known, that is, we consider a word w_n to be a tuple of the orthography $orth(w_n)$ of the word and the pronunciation $pron(w_n)$. Furthermore, by $x_{t_{n-1}+1}^{t_n}$ we denote the consecutive sequence of acoustic features assigned to word w_n. The feature functions used are the logarithms of several acoustic models $p(pron(w_n)|x_{t_{n-1}+1}^{t_n})$, of the pronunciation model $p(pron(w_n)|orth(w_n))$, of the language model $p(orth(w_n)|orth(w_{n-1}^{n-L}))$, where L is the context length of the language model, and a word penalty.

In many experiments it was shown that model-dependent scaling factors are crucial for highly accurate speech recognizers, e.g. [1, 2, 3]. Going from a single scaling factor per model to word-dependent scaling factors is motivated by the following observations, which give reason to assume a word- and pronunciation-dependent interaction between the models.

- varying discriminative power of the acoustic model: the discriminative power of an acoustic model is usually unsteady across phones and thus across pronunciations
- varying discriminative power among different acoustic models: different acoustic front-ends differ in their ability to discriminate among phones
- several modeling and training issues of the acoustic model, e.g. the severe independence assumptions and the presumably underestimated variances of the GMMs

Furthermore, due to the word-dependent scaling factors the training of the model in Equation (1) estimates word-dependent pronunciation scores and penalties in a discriminative manner.

Word- or word class-dependent scaling factors were used in [4, 5]. In the first paper a joint training of the acoustic model, the language model, and the scaling factors is performed. In the latter work word class-dependent scaling factors are used among other techniques in an adaptation step. Neither paper investigated the improvement coming solely from the word-dependent scaling factors. Another approach was applied in [6], where an improvement of around 3% relative is reported by using scaling factors that depend on classes derived from several acoustic features.

Table 1: *The Mandarin BN/BC system: Training, development, and test sets. The word-dependent scaling factors are trained on the 120h "Λ-training" set. For the first test set no word-segmented transcripts are available.*

		running		vocabulary	
corpus	duration	words	char.s	words	char.s
AM-training	∼230h	2.4M	4.0M	42.1K	5.3K
Λ-training	∼120h	1.3M	2.2M	33.7K	4.4K
held-out	1.5h	12.7K	21.5K	4.4K	1.8K
development	2.5h	27.5K	46.8K	5.3K	1.9K
test1	1.6h	-	28.1K	-	1.7K
test2	1h	10.5K	18.2K	2.9K	1.4K

An alternative approach is to consider the interdependency of the models directly in model training or model adaptation. Discriminative language model training was done e.g. in [7]. The major problem for LVCSR tasks is that the vast amount of language model training data comes without a spoken form. In the recent years some work was done on considering the dependencies between different acoustic front-ends already in acoustic model training yielding small improvements, e.g. [8].

3. Experimental Setup

The word-dependent scaling factors are tested in a Mandarin BN/BC system [9]. We use two different training sets for acoustic model training (230h) and scaling factor training (120h). On the 230h set we train models for three different acoustic front-ends: MFCCs, PLPs, and gammatone filter bank based features (GTs). The acoustic models are maximum likelihood estimates and use an LDA, VTLN, and constrained-MLLR in training and in addition MLLR in recognition.

The log-linear model combination of the three acoustic models with word-dependent scaling factors is applied in a lattice rescoring step. Lattices are produced with the MFCC system and are subsequently arc-wise rescored with fixed word boundaries. For experiments on character or syllable level the word arcs are first split into character arcs using the time information from an arc-wise forced alignment with the MFCC model. The lattices for the 120h training and the development and test sets are produced with exactly the same setup.

The 60K vocabulary and the four-gram language model are kindly provided by SRI. Unfortunately, the language model training data includes both training sets which results in a much lower perplexity on the 120h set than on the development and test sets. In order to get an idea of how much we loose due to the discrepancy we created an additional held-out set by removing each hundredth segment from the 120h set. Table 1 summarizes the corpora statistics.

3.1. Scaling Factor Classes

If a word in the 120h training set occurs less often than a cut-off N_{min}, then the corresponding scaling factor is replaced by a backing-off scale. The backing-off scaling factor depends on the number of phonemes in the pronunciation of the word:

$$\lambda_i(w) = \begin{cases} \lambda_{i,w}, & \text{if } \#w > N_{min} \\ \lambda_{i,|pron(w)|}, & \text{else} \end{cases} \quad (2)$$

For experiments on character level we use only a single backing-off class. In order to get an idea of how important the lexical information is we build an alternative set of scaling factors where we tie character-dependent scaling factors among equal pronunciations, i.e. we build syllable classes.

Table 2 contains the number of word-, character-, and syllable-dependent scales and the corresponding cut-offs.

We measured the coverage of running words in the development set which have a word-dependent scaling factor: for the cut-offs of 200, 50, 20, 10, and 5 the coverage is 67%, 83%, 90%, 93%, and 96%. For character- and syllable-dependent scaling factors the coverage is almost complete.

3.2. Scaling Factor Training

For most experiments we combine five models: the three acoustic models, the pronunciation model, and the language model. The interdependency between these models is sufficiently described by putting the word-dependent scaling factors on four of the five models. Following the considerations from Section 2 we put the word-dependent scaling factors on the acoustic models and the pronunciation model (and on the word penalty, if used).

For parameter estimation we apply minimum risk training with either the smoothed phoneme error (MPE) or word error (MWE) as objective function, where only the scaling factors are optmized. The estimation is done iteratively using Rprop, a gradient-descent algorithm. Time boundaries (and thus acoustic model scores and costs) are kept fixed among the training iterations. The implementation of the MPE objective function follows directly [10]. For character level experiments we try in addition MWE training, where we derive the costs from a confusion network (CN). The CNs are build from the training set lattices using our standard CN-decoder. In preliminary experiments we also tried MMI but it was inferior to minimum risk training.

Regularization turns out to be important, similar to the I-smoothing used in [10] for GHMM training. Equation (3) shows the objective function for minimum risk training, where L denotes the loss function.

$$\mathcal{F}(\Lambda) = \sum_{w_1^N} p_\Lambda(w_1^N|x_1^T) L(w_1^N, \tilde{w}_1^{\tilde{N}}) + \frac{R}{2}||\Lambda - \Lambda_0||_2^2 \quad (3)$$

Λ_0 is the set of model-dependent scaling factors derived from a direct error rate minimization on the development set using our CN-decoder. Thus, the initial LM scaling factors are around one and the acoustic model scaling factors are close to the inverse language model scale (as used in Viterbi decoding) divided by the number of acoustic models. We optimize the scaling factors until convergence in the objective function occurs and take the scaling factors from the last training iteration for decoding. The regularization constant R is optimized on the development set, which is costly and therefore is not done in fine-grained steps.

For discriminative acoustic model training it is known that using a weak language model improves the training result. We did some preliminary experiments using uni- and bigram language models, but all results were clearly worse.

For lattice decoding we use the same character-based CN-decoder that we employ in MWE training. For all log-linear model combinations (using one or several acoustic models) we observe a slight 2-3% relative improvement over the Viterbi results. The improvements persist for experiments with word-dependent scaling factors.

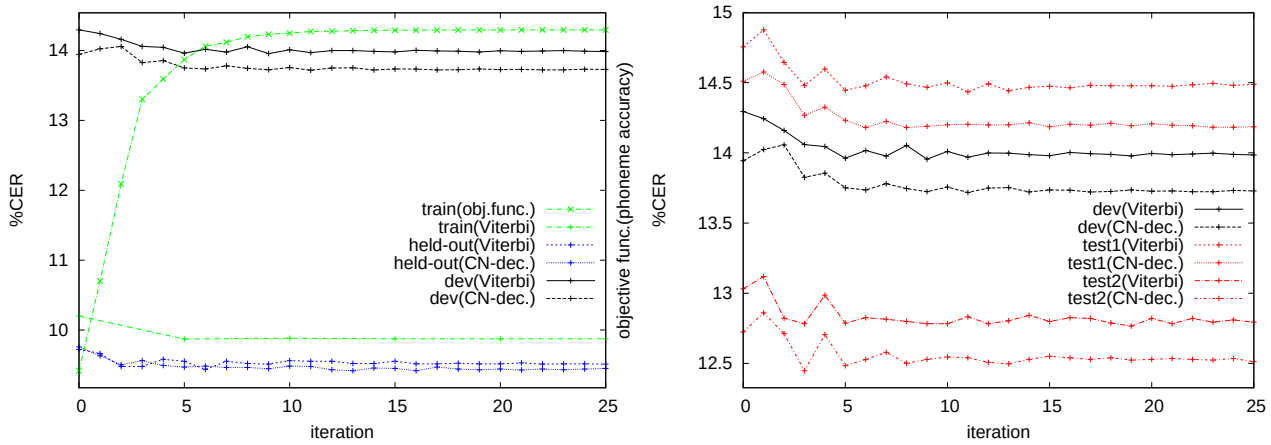

Figure 1: *Model-combination results for 25 training iterations and 6,904 word-dependent scaling factors. The word-dependent scaling factors are trained on 120h. The left plot shows the objective function and character error rates for the training set, the held-out set, and the development set. The right plot shows the progression of the error rates for the development set and the two test sets.*

4. Experimental Results

Figure 1 shows detailed results for the training and evaluation of the best performing setup which consists of 6K word-dependent scaling factors. In the left plot we see that the objective function (phoneme accuracy) improves smoothly and the character error rates (CER) on training, held-out, and development set smoothly decrease. The right plot shows again the development set together with the two test sets. Both, the Viterbi and the CN results are plotted.

The plots for the other setups look rather similar. Table 2 assembles the results for word-, character-, and syllable-dependent scaling factors for different cut-off values. The number of classes refers to the number of scaling factors per model; throughout the language model gets only a single scaling factor. The baseline is the setup using a single scale per model.

4.1. Word level

The best improvement is achieved with 6K scaling factors, but the differences among the cut-off values is tiny and especially for 3K and more scaling factors it might even disappear with a more fine-grained optimization of the regularization constant. The relative improvement in CER is around 2%, a little better for the held-out set where we observe a relative improvement of 3%. On the training set we measured the error rate of the Viterbi decoding and even here we observe at most a gain of 4% relative. Additional word penalties do not help.

4.2. Character level

The results with character-dependent scaling factors are similar to the word-dependent results. The differences in the word- and character-level baselines are due to fixing the boundaries of the character arcs with the MFCC model. When rescoring with the PLP and GT model the results are suboptimal compared to a word arc-wise rescoring.

The MWE results are a little worse than for MPE. We hoped that the CN-decoder benefits from MWE trained, character-dependent scaling factors, but the gap to the corresponding Viterbi results do not widen.

Going from character- to syllable-dependent scaling factors changes only the results for one test set, for which already moving from words to characters yields slightly worse results.

5. Analysis

The only small improvements we get from word-dependent scaling factors are sobering. Even on the training set the improvement is rather small.

A further analysis is to show which interdependencies are eventually captured by our approach. The first question is: how much of the improvement comes from the pronunciation weights? To answer this question we train a log-linear model combination with a single pronunciation model scale. The results remain almost equal, which is not surprising as in Mandarin only few pronunciation variants are known.

Next, we investigate the influence of the word length. In preliminary experiments we used scaling factors that depend only on the number of phonemes in the pronunciation of a word. The improvements, if at all, were much smaller than for word-dependent scaling factors. Looking at the results in Table 2 we observe that character-dependent scaling factors perform almost as good as the word-dependent ones. All the differences are rather tiny which does not allow us to draw reliable conclusions, but the results indicate that the interdependency between the models does not strongly depend on the word length.

The remaining question is: how important are the word-dependent scaling factors for modeling the interdependency of the three acoustic models? We performed the following experiment: we build a MFCC, a PLP, and a GT system using the best word-dependent scaling factors trained on the log-linear model combination containing the three acoustic models. Obviously, in the resulting systems the impact of the acoustic and the language model are not balanced anymore. To compensate for that we introduce an additional scaling factor for each model, which we optimize directly on the development set. For a fair comparison we performed the same optimization for the log-linear model combination of all three acoustic models. The results are given in Table 3. We see that the average relative improvement for the three systems using only a single acoustic model is almost equal to the relative improvement for the log-linear combination containing all three acoustic models. The conclusion is that the word-dependent scaling factors presumably do not capture the dependencies between the acoustic models, but solely model the interdependency of acoustic and language model. Which in turn explains why we don't benefit from applying a weaker language model in training.

Table 2: *CN-decoding results for the log-linear model combination using word-, character-, and syllable-dependent scaling factors. The scaling factors are trained on 120h using either minimum phone error (MPE) or minimum character error (MWE) training.*

opt. criterion	#classes (cut-off)	[%CER] dev	test1	test2	held-out
word-dependent scaling factors					
MPE	1	13.94	14.51	12.73	9.76
	997(200)	13.80	14.26	12.56	9.59
	3,596(50)	13.76	14.23	12.57	9.54
	6,904(20)	13.73	14.19	12.51	9.43
	10,911(10)	13.73	14.28	12.53	9.57
	16,665(5)	13.74	14.25	12.46	9.52
character-dependent scaling factors					
MPE	1	13.95	14.59	12.63	9.77
	2,708(20)	13.79	14.37	12.40	9.54
	3,707(5)	13.80	14.37	12.40	9.55
MWE	1	13.94	14.60	12.65	9.77
	2,708(20)	13.84	14.42	12.55	9.77
	3,707(5)	13.83	14.42	12.50	9.78
syllable-dependent scaling factors					
MPE	1	13.95	14.59	12.63	9.77
	1,064(20)	13.79	14.51	12.40	9.60
MWE	1	13.94	14.60	12.65	9.77
	1,064(20)	13.91	14.61	12.46	9.84

A last experiment that points in the same direction is to combine and decode the three systems in a confusion network combination (CNC). We compute separate lattice sets with the MFCC, PLP, and GT system and rescore them using the best word-dependent scaling factors from the log-linear model combination. The results of the CNC are shown in Table 3, the relative improvements are even slightly higher than for the log-linear model combination.

6. Conclusions and Outlook

In this work we investigated the influence of word- and pronunciation-dependent scaling factors in a log-linear combination of three acoustic models, a pronunciation and a language model. The scaling factors are trained on a large, separate 120h training set. But only an improvement of 2% relative is observed for a Mandarin BN/BC task. We see the maximum improvement already for around 3K scaling factors.

A further analysis of the results shows that the improvement comes presumably almost solely from the interdependency of the language and the acoustic models. The word-dependent scaling factors cannot boost the gain from assembling three acoustic models in the log-linear model combination.

7. Acknowledgements

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

Table 3: *CN-decoding results for log-linear model combinations using one or several acoustic models, and for a system combination. The word-dependent scaling factors are optimized for the model combination containing the three acoustic models. Each model has an additional scale; for each setup these scales are optimized directly on the development set.*

acoustic front-end(s)	#classes	[%CER] dev	test1	test2
model combination with one acoustic model				
MFCC	1	14.79	15.18	13.32
	6,904	14.44	15.07	12.98
PLP	1	15.07	15.20	13.67
	6,904	14.71	15.05	13.54
GT	1	15.47	15.97	14.05
	6,904	15.20	15.87	13.68
model combination with all acoustic models				
MFCC+PLP+GT	1	14.01	14.51	12.71
	6,904	13.69	14.18	12.52
system combination (CNC)				
MFCC+PLP+GT	1	13.72	14.13	12.44
	6,904	13.35	13.89	12.13

8. References

[1] P. Beyerlein, "Discriminative model combination," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, California, USA, Dec. 1997, pp. 238–245.

[2] D. Vergyri, "Integration of multiple knowledge sources in speech recognition using minimum error training," Ph.D. dissertation, Johns Hopkins University, 2000.

[3] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Philadelphia, PA, USA, Mar. 2005, pp. 457–460.

[4] X. Huang, M. Belin, F. Alleva, and M. Hwang, "Unified stochastic engine (USE) for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, Minneapolis, MN, USA, Apr. 1993, pp. 636–639.

[5] R. R. Sarukkai and D. H. Ballard, "Improved spontaneous dialogue recognition using dialogue and utterance triggers by adaptive probability boosting," in *Proc. Int. Conf. on Speech Communication and Technology*, vol. 1, Philadelphia, PA, USA, Oct. 1996, pp. 208–211.

[6] D. Vergyri, S. Tsakalidis, and W. Byrne, "Minimum risk acoustic clustering for multilingual acoustic model combination," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 873–876.

[7] H. J. Kuo, E. Fosler-Lussier, H. Jiang, and C.-H. Lee, "Discriminative training of language models for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Orlando, FL, USA, May 2002, pp. 325–328.

[8] C. Breslin and M. J. F. Gales, "Generating complementary systems for speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sep. 2006.

[9] C. Plahl, B. Hoffmeister, M.-Y. Hwang, D. Lu, G. Heigold, J. Lööf, R. Schlüter, and H. Ney, "Recent improvements of the RWTH GALE mandarin LVCSR system," in *Proc. Int. Conf. on Spoken Language Processing*, Brisbane, Australia, Sep. 2008, pp. 2426–2429.

[10] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Orlando, FL, USA, May 2002, pp. 105–108.