COMPARING CONFIDENCE-GUIDED AND ADAPTIVE DYNAMIC PRUNING TECHNIQUES FOR SPEECH RECOGNITION

Tibor Fabian, Günther Ruske

Institute for Human-Machine-Communication, Technische Universität München
Arcisstr. 21, 80333, Munich, Germany
{fab,rus}@mmk.ei.tum.de

ABSTRACT

Improvement in pruning algorithms for automatic speech recognition leads directly to a more efficient recognition process. Efficiency is a very important issue in particular for embedded speech recognizers with limited memory capacity and CPU power. In this paper we compare two pruning algorithms, the *confidence-guided* pruning and the *adaptive control* pruning technique. Both methods set the pruning threshold for the Viterbi beam search process dynamically for each time frame depending on search space properties. We show that both dynamic pruning techniques are applicable in reducing the time consumption of the recognizer whereas our novel confidence-guided pruning approach outperforms the adaptive control technique clearly.

1. INTRODUCTION

The computation time efficiency of automatic speech recognition systems (ASR) is still an important and topical issue. More and more speech recognizers will be deployed in embedded systems e.g. portable mobile devices which often have limitations in computation and memory capacity. Nevertheless speech applications running on such resource constrained devices also have to meet users expectations e.g. reasonable system response times. Therefore the main issue is to minimize ASR response delays respectively to speed up the recognition process.

The most of time consumption during the recognition process will happen in the search process. Managing alternative hypotheses for each time frame could be very time costly and memory loaded depending on the complexity of the search network. The size of the Viterbi search space of HMM-based ASR systems increases usually non-linearly with the vocabulary size. That's why different pruning strategies have been already proposed to reduce the time consumption of the recognition process. The main purpose of pruning is the ability to limit the size of the search space which has direct impact on the CPU and memory requirements of the recognizer.

Before we introduce the dynamic pruning techniques which we compare in this work let us recapitulate how pruning generally works in HMM-based ASR on the example of the classical pruning methods:

Probability-based pruning controls the pruning threshold B_{set} of the Viterbi beam search process at each time frame and keeps only those hypotheses whose score is no less than a threshold from the score of the best hypothesis. The threshold is generally a constant value for the whole recognition process. However the number of hypotheses which can be cut-off depends on the distribution of the hypotheses scores. If they are close to each other only few of them can be pruned.

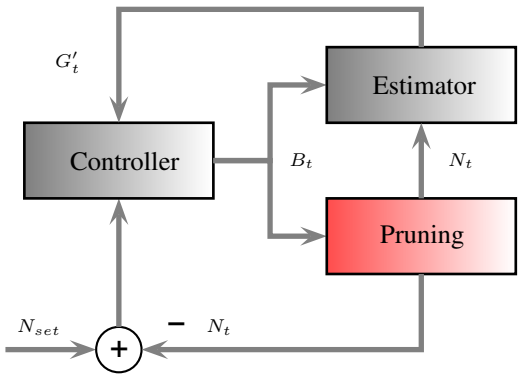

Figure 1: Schematic view of the adaptive controller for pruning.

Rank-based pruning avoids this problem by limiting the maximum number of alternatives to a fixed preset value N_{set}. In contrast to the probability-based technique rank pruning controls the number of hypotheses allowed for each time step independently of their distribution. For this reason all alternative hypotheses have to be ranked by their log probabilities keeping only the best N_{set} hypotheses. To improve the efficiency of the ranking procedure, usually a histogram of the hypotheses scores is computed - *histogram rank pruning*. It is a common practice to combine both, probability-based and rank pruning to increase pruning efficiency.

These classical pruning techniques generally use constant pruning thresholds during the whole search procedure. Both, B_{set} of the probability-based pruning and N_{set} of the rank-based approach are predefined values. However these thresholds could be adjusted dynamically to take the time-variant behavior of the search process into consideration.

In this work we compare two dynamic pruning algorithms the *confidence-guided dynamic* technique (CGD pruning) and the *adaptive control dynamic* pruning method (ACD pruning). Both algorithms set the pruning threshold for the Viterbi beam search process dynamically for each time frame depending on search space properties. In our earlier work [1] we presented the novel *confidence-guided dynamic* pruning method which uses confidence measurement to minimize the

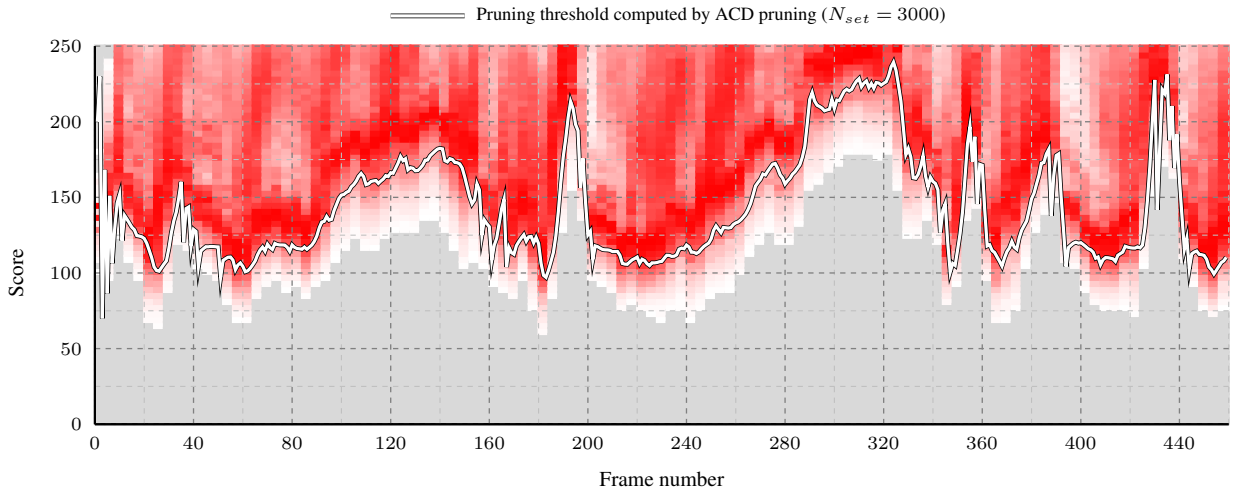

Figure 2: Example for dynamic pruning threshold during the appointment negotiation utterance 'Ja genau, lassen wir gleich die letzte Woche im März, prima!' (English: That's correct, let's keep right the last week in March, great!).

computation time effort of the Viterbi search process by reducing the search space to an acceptable level (i.e. the number of active hypotheses). The decision about the appropriate threshold at each time frame is based on the utilization of confidence measurement. The other dynamic pruning approach ACD was presented in [2] and [3]. This pruning method uses adaptive control techniques to steer the pruning threshold dynamically. We show in this work that both dynamic pruning techniques are applicable in reducing the time consumption of the recognizer whereas our novel confidence-guided pruning approach outperforms the adaptive control technique clearly.

In the next Sections 2 and 3 we describe these dynamic pruning approaches which compute the beam pruning threshold B_t of the HMM-based Viterbi search process frame by frame. A comparison of both pruning techniques is given in Section 4. Section 5 describes the evaluation material and the ASR system we used for our evaluations. In Section 6 we present the results of our experiments.

2. CONFIDENCE-GUIDED DYNAMIC PRUNING

CGD pruning is a combination of the widely used classical probability based pruning and runtime confidence measurement. As we described detailed in [1] this approach allows to take the time-variant behavior of the search process into account. As confidence measurement we used accumulated normalized log likelihood C'_{acc} which is computed frame by frame as follows:

$$C'_{acc,t} = \log\left(\frac{\prod_{t=1}^{T} p(\vec{x}_t|c)}{\max(\prod_{t=1}^{T} p(\vec{x}_t)|W_{t,best})}\right). \qquad (1)$$

For normalization in Equation 1 we use the combined maximum of accumulated observation probability $p(\vec{x}_t)$ and best word end likelihood $W_{t,best}$. The observation probability $p(\vec{x}_t)$ is estimated using a so called catch-all model (s. [4] for details). The schematic view of this approach is shown in Fig. 1 in [1]. C'_{acc} is expressed in the logarithmic space and can be viewed as a zero-centered confidence score where positive scores indicate good and negative scores bad confidence.

To optimize the computation of the dynamic pruning threshold we also use the so called dynamic lift B_{dyn} which is calculated framewise with following formula:

$$\Delta B_{dyn,t} = T_{upp} - \frac{T_{low}}{1 + e^{(\alpha - C'_{acc,t})/\beta}}. \qquad (2)$$

The thresholds (T_{upp}, T_{low}) and the parameters α and β for the modified sigmoid function in Equation 2 can be determined using a cross evaluation corpus. Reasonable setting for our experiments was $\alpha = \beta = 20$.

Let us recapitulate the computation steps of the CGD pruning technique:

1. compute C'_{acc} using Equation 1,
2. calculate dynamic lift using Equation 2,
3. set the dynamic pruning threshold as follows:

$$B_t = \Delta B_{dyn,t} + C'_{acc,t}. \qquad (3)$$

The dynamic pruning threshold B_t computed in Equation 3 can be directly used as the pruning threshold for the Viterbi search process for each time frame.

3. ADAPTIVE CONTROL DYNAMIC PRUNING

Another possibility to steer the dynamic beam width frame by frame is taking the advantages of adaptive control algorithms into account (s. [2]). ACD pruning is a technique which changes the pruning threshold for the Viterbi search process at runtime to compensate the variations in the search environment and to achieve the preset threshold of maximum

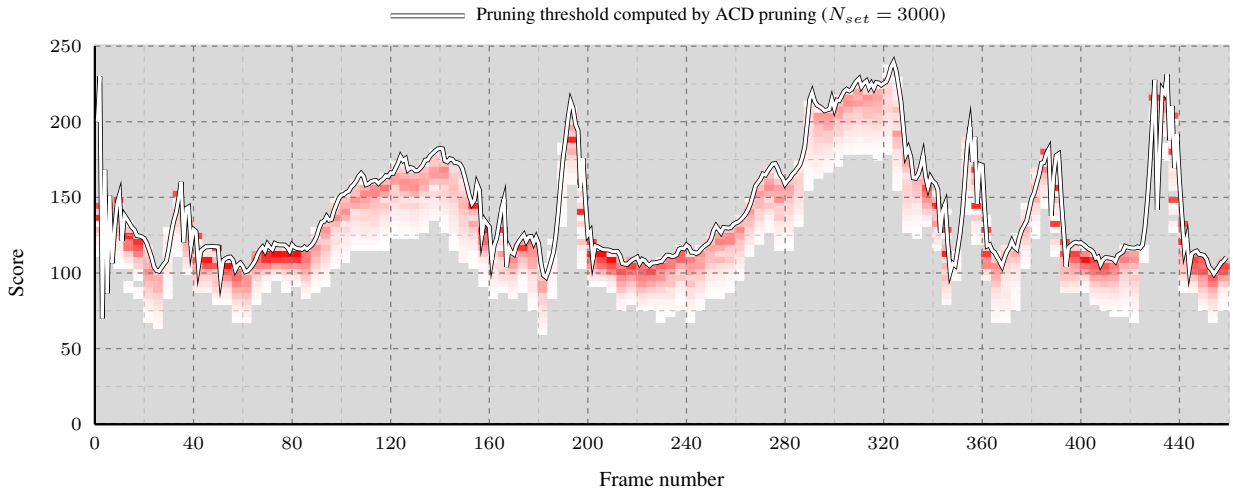

Figure 3: Histogram of the number of active hypotheses for the same example utterance as Fig. 2 but using ACD pruning technique.

number of hypotheses N_{set}. Fig. 1 shows the topology of the ACD pruning approach.

This method uses a feedback-control mechanism that contains adjustable coefficients. The ACD pruning system consists of an inner and an outer loop. The inner loop contains an ordinary feedback loop and the plant. These are in Fig. 1 the feedback *controller* and the *pruning* process. The parameters of the controller are adjusted by the outer feedback loop which is composed of a recursive parameter *estimator*.

For simplicity the pruning process is modeled as a 0th order linear system with slowly varying gain G_t:

$$N_t = G_t B_t.$$

The controller is an integrator

$$B_{t+1} = b_t + \alpha(N_{set} - N_t)/G'_t\ , \tag{4}$$

where α, the parameter of the controller, can adjust the response speed of the feedback loop. The time variant gain G'_t in Equation 4 can be estimated using least squares estimation with the following formula:

$$G'_t = \frac{\sum_{i=1}^{L} N_{t-i} B_{t-i}}{\sum_{i=1}^{L} B_{t-i}^2}, \tag{5}$$

based on the past L observations of the pruning threshold B_t. For the dynamic pruning approach reasonable values are for the parameter $L = 5$ and for $\alpha = 0.2$ as proposed in [2].

The computation steps of the pruning process based on this adaptive controller are as follows:

1. estimate the gain of the pruning process, Equation 5,
2. compute the pruning threshold with Equation 4.

Equations 5 and 4 must be calculated only once per frame. This is acceptable and their computation costs should not influence pruning efficiency negatively. To catch side effects of the controller especially at the beginning of an utterance the computed pruning threshold should be limited by maximum and minimum values.

4. COMPARING BOTH PRUNING TECHNIQUES

The main advantage of both dynamic pruning techniques CGD and ACD prunings is the framewise computation of the pruning threshold for the search process. That way both of them are able to take time-variant characteristics of the search process into account. We illustrate this clearly in Fig. 2 with an example sentence of the evaluation set. The diagram shows the dynamic pruning threshold of the ACD pruning depending on the frames of the example utterance. In this diagram the horizontal line at $y = 0$ (i.e. x-axis) represents the best hypothesis scores for each frame. The pruning threshold is plotted relative to it as y-value frame by frame. The pruning threshold curve for the CGD pruning looks similar to the plotted curve of the ACD pruning so we omitted it to keep clarity.

One main difference between CGD and ACD pruning is: at the beginning of the B_t curve of the ACD pruning is a kind of transient oscillation to observe as Fig. 2 shows this clearly. This is caused by the integrator because of the insufficient number of observation values for the computation of the plant gain in Equation 5. At this point the CGD pruning has a clear benefit because its computation is based on confidence measurement and therefore it does not have this negative effect at the beginning of the utterance.

Additional to the pruning threshold curve the histogram of the number of the active hypotheses is also plotted in Fig. 2 as color coded z-axis at the background. This histogram was computed on equidistant score intervals (width = 1) in the range of [0-250] from the x-axis. For the histogram computation we used the classical probability based pruning technique with a pruning threshold of 250. The color

transition from white to red color (in gray scale from white to black) illustrates the number of active hypotheses in the range from 1 to ∞ for each time frame and score interval. Light gray color beneath the color gradient represents no active hypothesis for the specified intervals.

In the histogram plot of Fig. 2 we can see that the number of active hypotheses generally increases with increasing distance to the best hypothesis (x-axis). If we used a preset constant pruning threshold (i.e. horizontal line parallel to the x-axis e.g. by score = 210) there were a lot of hypotheses of poor quality which could not be pruned because they fell below the constant threshold. In contrast to the constant threshold the time dependent dynamic pruning threshold of CGD or ACD pruning methods is able to cut off clearly more hypotheses. This is possible because the dynamic threshold demarcates the edge of the histogram transition to the increasing number of active hypotheses framewise at different score distances from the x-axis. As a result the number of the active hypotheses can be reduced dramatically as we show this in Fig. 3 which means the ASR can be speeded up and on the other hand the memory usage can be saved enormously.

5. EXPERIMENTAL SETUP

The experiments described in this paper were performed on the commonly used speech recognition system HTK (Release 3.2) [5]. As test material we used the German Verbmobil '96 corpus [6], which contains 343 sentences, i.e. 6428 words. The computation of the scaling factors in Equation 2 and controller parameters in Equations 4 and 5 was performed on a distinct cross-validation set, which contains 599 sentences (11577 words). For the recognition process, we used a bigram language model, a dictionary with 5343 entries, and triphone acoustic models with about 25000 mixtures trained on the Verbmobil '96 training corpus.

6. EXPERIMENTS AND RESULTS

The goal of the experiments presented in this section was to compare both dynamic pruning techniques CGD and ACD. For this purpose we ran several tests on the Verbmobil '96 test data (s. previous Section 5) using different parameters. Our results are presented in Fig. 4 which shows *word error rates* (WER) depending on the *time factor*. The time factor is defined as the ratio of the time consumption of the ASR with particular pruning parameter settings and the time consumption of the ASR without any pruning.

We focused our investigation on the comparison of our CGD pruning technique and the adaptive control approach. Additionally Fig. 4 also allows to compare their results with the classical probability-based fixed pruning (PB) and its combination with rank-based pruning (PBR).

As far as the results of the classical pruning techniques are concerned the curve of constant pruning threshold in Fig. 4 was determined by computing the WER for the evaluation corpus using different pruning values B_{set} in a range of [80-250]. To the greater pruning threshold value belongs lower WER but higher time factor. The combination of probability-based and rank pruning was evaluated by keeping B_{set} at 210 and varying N_{set} in the range of [500-9000].

Regarding to the dynamic approaches the resulting WER curve of the CGD pruning method was found using $T_{upp} =$ 110 and different T_{low} in a range of [20-70] (please refer to Equation 2 for details). To get results for the ACD pruning method we varied N_{set} in the range of [300-10000].

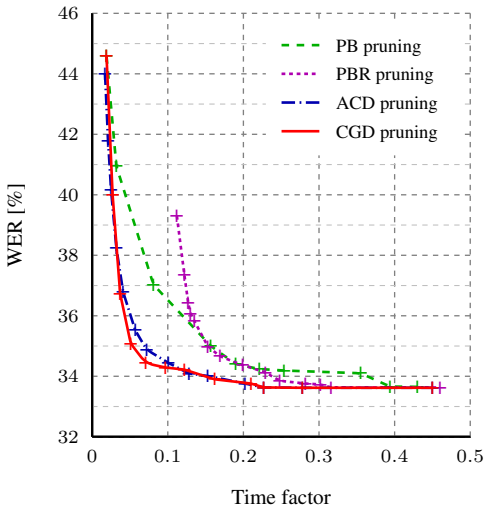

Figure 4: Word error rates (WER) of different pruning techniques depending on time factor: probability-based beam width (PB), combined probability-based and rank (PBR), adaptive control (ACD), and confidence-guided dynamic pruning (CGD).

Fig. 4 shows that both dynamic pruning techniques outperform the static methods significantly. The time factor of the ASR could be decreased to 0.23 without increasing WER using CGD or ACD pruning. Furthermore if we accept an increase of WER less than 1 %, ACD pruning achieves a time factor of 0.1 which corresponds to the acceleration of the ASR 10 times (reciprocal time factor). Respectively compared with the best PB pruning result ACD pruning makes the ASR 1.9 times faster. The best result was achieved by our CGD beam pruning approach namely a time factor of 0.07 which corresponds to the acceleration of the ASR about 14 times or compared with the best PB pruning result 2.7 times. Further details of our evaluation tests are shown in Table 1.

How is it possible to get better results using ACD pruning technique with the preset value of $N_{set} = 3000$ than with the classical rank-based approach with $N_{set} = 2000$? The explanation is that ACD pruning controls the beam width of the search process to avoid to exceed the maximum number of active hypotheses. In contrast the rank-based pruning needs two passes: First the Viterbi search step is performed and only afterward the number of the active hypotheses is reduced to the preset value for the next search step. As a result ACD pruning achieves indeed an average N_{set} of 3000 but the ASR using the classical rank-based pruning approach has to handle often 3 or 4 times more hypotheses which leads to increased computation time effort.

7. CONCLUSION

The comparison of two dynamic pruning methods CGD and ACD pruning has shown that both of them are applicable to reduce the computation time of the speech recognizer. As well CGD as ACD pruning approaches perform signifi-

Pruning method; parameters	WER [%]	Time factor
PB; $B_{set} = 250$	33.63	0.43
PB; $B_{set} = 150$	34.4	0.19
PBR; $B_{set} = 210, N_{set} = 9000$	33.63	0.32
PBR; $B_{set} = 210, N_{set} = 2000$	34.37	0.2
ACD; $N_{set} = 8000$	33.63	0.23
ACD; $N_{set} = 3000$	34.44	0.1
CGD; $T_{upp} = 110, T_{low} = 40$	33.63	0.23
CGD; $T_{upp} = 110, T_{low} = 70$	34.43	0.07

Table 1: Word error rates and the corresponding time factors of different pruning methods.

cantly better than classical pruning techniques. As a result a significant improvement in decoding speed of the ASR system could be achieved. We found the best results using our confidence-guided dynamic pruning approach which clearly outperforms not only the classical pruning techniques but also the ACD pruning.

The next step of our work will be to investigate the combination of these dynamic pruning strategies. Further improvements in pruning could be achieved if instead of a preset threshold N_{set} of the ACD pruning the maximum number of the active hypotheses was varied by utilization of an appropriate confidence measurement. That way the pruning threshold could be decreased in case of high confidence and increased in case of low confidence for each time frame.

REFERENCES

[1] Fabian, T. and Lieb, R. and Ruske, G. and Thomae, M., "A Confidence-Guided Dynamic Pruning Approach -Utilization of Confidence Measurement in Speech Recognition-" *Proceedings of INTERSPEECH*, Lisbon, Portugal, 2005, pp. 585-588.

[2] Hamme H.V. and Aellen F.V., "An Adaptive-Beam Pruning Technique for Continuous Speech Recognition" *In Proceedings of ICSLP*, Philadelphia, Pennsylvania, 1996, pp. 2083-2086.

[3] Zhang, D. and Du, L., "Dynamic Beam Pruning Strategy Using Adaptive Control" *Proceedings of INTERSPEECH*, Jeju Island, Korea, 2004, pp. 285-288.

[4] Kamppari S.O. and Hansen T.J., "Word and Phone Level Acoustic Confidence Scoring" *In Proceedings of IEEE ICASSP*, Istanbul, Turkey, 2000, pp. 1894-1897.

[5] Young S.J., "The HTK Hidden Markov Model Toolkit: Design and Philosophy" *Technical Report, Department of Engineering*, Cambridge University (UK), 1994.

[6] Bub T. and Schwinn J., "VERBMOBIL The Evolution of a Complex Speech-to-Speech Translation System" *In Proceedings of ICSLP*, Philadelphia, Pennsylvania, 1996, pp. 2371-2374.