A Confidence-Guided Dynamic Pruning Approach - Utilization of Confidence Measurement in Speech Recognition -

Tibor Fabian, Robert Lieb, Günther Ruske, Matthias Thomae

Institute for Human-Machine-Communication
Technische Universität München, Germany
{fab,lie,rus,tho}@mmk.ei.tum.de

Abstract

Improved efficiency of pruning accelerates the search process and leads to a more time efficient speech recognition system. The goal of this work was to develop a new pruning technique which optimizes the well known probability-based pruning (beam width) by utilization of confidence measurement. We use normalized hypotheses scores to guide the beam width of the pruning process dynamically frame per frame during the whole utterance. Compared with classical pruning techniques like fixed beam pruning and histogram rank pruning we achieved significantly better results concerning the time consumption of the recognizer. The speed of the recognition process could be accelerated up to 14 times with a slight degradation in recognition accuracy.

1. Introduction

Making speech recognition more efficient in computation time is still an important and topical issue. More and more speech applications will be deployed in embedded systems which often have only a limited computation capacity. In order to meet users expectations we need acceptable runtime behavior by minimizing system response delays.

This situation motivated us to analyze commonly used speech recognition algorithms in order to optimize their efficiency in computation time. 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 Viterbi search space of HMM-based automatic speech recognition systems (ASR) 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.

Probability-based pruning controls the beam width B_{set} of the Viterbi 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 set 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.

Rank-based pruning avoids this problem by limiting the absolute number of alternatives to a fixed value. In contrast to the beam width 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_{max} hypotheses. The main disadvantage of this method is that two passes through all hypotheses are required and the ranking can be very time costly. 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. The combination allows achieving better results by memory saving and reduction of computational time effort by keeping recognition accuracy on an acceptable level.

Proven confidence measures like maximum a posteriori probability (P_{MAP}) or normalized log likelihood score (C_{NLL}) allow an assessment of the classification correctness at phone or word level during the search process as described in [1], [2], and [3]. Especially in the last years some pruning algorithms were introduced concerning confidence measurement as a guide for pruning techniques (among others Ortmanns [4], Liu [5], Renals [6], and Abdou [7]). In [7] a complex look-ahead technique was presented which has to manage HMM-specific thresholds of posterior confidence scores to support the pruning procedure. This could lead to an enormous management effort in case of thousands of triphones which are often used in current ASR systems. The a posteriori based look-ahead approach proposed in [4] is based on a different framework than HMM, namely a neural network.

All of the mentioned pruning techniques use generally constant pruning thresholds during the whole search procedure. Both B_{set} of the probability-based pruning and N_{max} of the rank-based approach are predefined thresholds which have to be justified during cross validation tests. However, these thresholds could be adjusted dynamically to fit time-variant requirements by taking variable search quality into consideration.

In the next section we introduce a novel dynamic pruning approach which controls the beam width B_{set} of HMM-based Viterbi search process framewise. The decision about the appropriate threshold at each time frame is based on the utilization of confidence measures. Section 3 describes the evaluation material and the ASR system we used for our evaluations. In section 4 the results of our experiments will be shown.

2. Confidence-guided pruning method

Our novel approach is a combination of the widely used classical probability-based beam pruning technique and runtime confidence measurement. As already mentioned above probability-based pruning uses a constant threshold B_{set} to set beam width of the Viterbi search process at each time frame of the whole utterance. In contrast to this an appropriate confidence measure of the best hypothesis allows to take the time-variant behavior of the search process into account. As a result beam width $B(t)$ can be set dynamically at each time frame t according to the

From: Interspeech 2005, Lisboa, ISCA

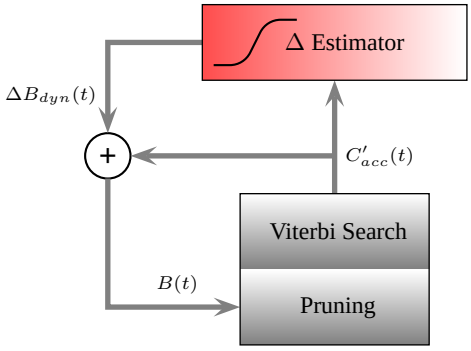

Figure 1: *Schematic view of the confidence-guided dynamic pruning method (CGD).*

confidence estimation.

2.1. Computation of the confidence measure

In this work we use a confidence measurement which is a variation of the maximum a posteriori probability P_{MAP}, the mathematical formulation is:

$$P_{MAP} = P(c_i|x) = \frac{p(\vec{x}|c_i)P(c_i)}{p(\vec{x})} \; .$$

P_{MAP} can be thought of as the ratio of a proposed score $p(\vec{x}|c_i)P(c_i)$ of a class c_i and the so called catch-all score $p(\vec{x})$. Where c_i corresponds to a HMM state and $\vec{x}$ to the acoustic observation vector. The catch-all score reflects how well the acoustic models estimate the observation probabilities.

As slight variation of P_{MAP}, the normalized log likelihood C_{NLL}, will be computed as the logarithm of P_{MAP} normalized by the prior probability $P(c_i)$, s. Equation 1. C_{NLL} 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.

$$C_{NLL}(c_i|\vec{x}) = \log\left(\frac{p(\vec{x}|c_i)}{p(\vec{x})}\right), \qquad (1)$$

where the observation probability will be computed as follows:

$$p(\vec{x}) = \sum_{j=1}^{N_c} p(\vec{x}|c_j)P(c_j) \; .$$

Earlier works (e.g. [1]) provide adequate results about the good quality of C_{NLL} as a reliable confidence measurement. However the confidence of the hypotheses at a specific time frame of the utterance cannot be directly used to control the pruning of the Viterbi beam search process. That is because the confidence of the hypotheses can change in time in a large manner. A particular hypothesis could be pruned at a specific time frame because of its low confidence, even if this hypothesis would be the best one at the end of the utterance.

This is the reason why pruning techniques generally work on accumulated quantities. Therefore the main task is to find a way which allows to use confidence measurement for accumulated values of hypotheses. They have to be computed step by step during the search process. We define the accumulated normalized log likelihood C_{acc}, the ratio of the accumulated hypothesis likelihood probability and the accumulated observation's probability for each time frame from 1 to T of the utterance:

$$C_{acc} = \log\left(\frac{\prod_{t=1}^{T} p(\vec{x}|c)}{\prod_{t=1}^{T} p(\vec{x})}\right) .$$

Unfortunately this score increases continuously because of the steady increasing difference between $p(\vec{x}|c)$ and $p(\vec{x})$ from frame to frame. Therefore we use a modified normalization instead of $p(\vec{x})$, the combined maximum of accumulated $p(\vec{x})$ and best word end likelihood W_{best}:

$$C'_{acc} = \log\left(\frac{\prod_{t=1}^{T} p(\vec{x}|c)}{\max(\prod_{t=1}^{T} p(\vec{x})|W_{best})}\right) . \qquad (2)$$

Equation 2 allows to generate a normalization quantity which can be used for each time step to compute the confidence value of the hypothesis score. Fig. 2 shows an example of normalized hypothesis score C'_{acc} of the best hypothesis. As we can see in the diagram the curve of the normalized score (dashed line) depends on the time frame. Especially high local maximum values appears in correlation to pauses in the utterance.

In the classical pruning case we have to set an appropriate constant beam width which allows to keep the best hypothesis from the first frame until the last. Such a constant beam width would correspond to a horizontal line in Fig. 2 at a specific score level of e.g. 200. In contrast to this our dynamic approach allows to use a constant threshold B_{set} relative to the normalized score (dashed line). At each time frame only those hypotheses will be kept which score is no less than a threshold $B(t) = B_{set} + C'_{acc}(t)$ from the score of the best hypothesis.

Further optimization of this dynamic approach can be achieved if the threshold which is set relative to the normalized score (B_{set}) will also be computed dynamically. For this purpose we vary the value of B_{set} depending on the value of the normalized score C'_{acc} itself which indicates the observation quality of acoustic models. Low score indicates poor certainty of the best hypothesis therefore the beam width should be increased. This kind of dynamic lift (dotted line in Fig. 2) gives some compression to the dynamic beam width. On the one hand greater C'_{acc} score indicates good confidence of the best local hypothesis therefore the dynamic beam width ΔB_{dyn} could be decreased. On the other hand ΔB_{dyn} should be increased in case of low local confidence to avoid the pruning of the global best hypothesis. To implement this kind of dynamic lift we use a modified sigmoid function to control the beam width between appropriate upper and lower thresholds:

$$\Delta B_{dyn} = \frac{T_{upp}}{1 + e^{(\alpha - C'_{acc})/\beta}} + T_{low} \; . \qquad (3)$$

The parameters α and β in Equation 3 can be easy determined using a cross evaluation corpus. Reasonable setting for our experiments was $\alpha = \beta = 20$. The threshold for the pruning decision is computed in this case as follows:

$$B(t) = \Delta B_{dyn}(t) + C'_{acc}(t) \; . \qquad (4)$$

Fig. 1 shows the schematic block diagram of the confidence-guided dynamic (CGD) pruning approach. CGD pruning computes beam width of the probability-based pruning dynamically in accordance with the confidence assessment of the best hypothesis (s. Equation 4). The Δ estimator is responsible to compute $\Delta B_{dyn}(t)$ for each time

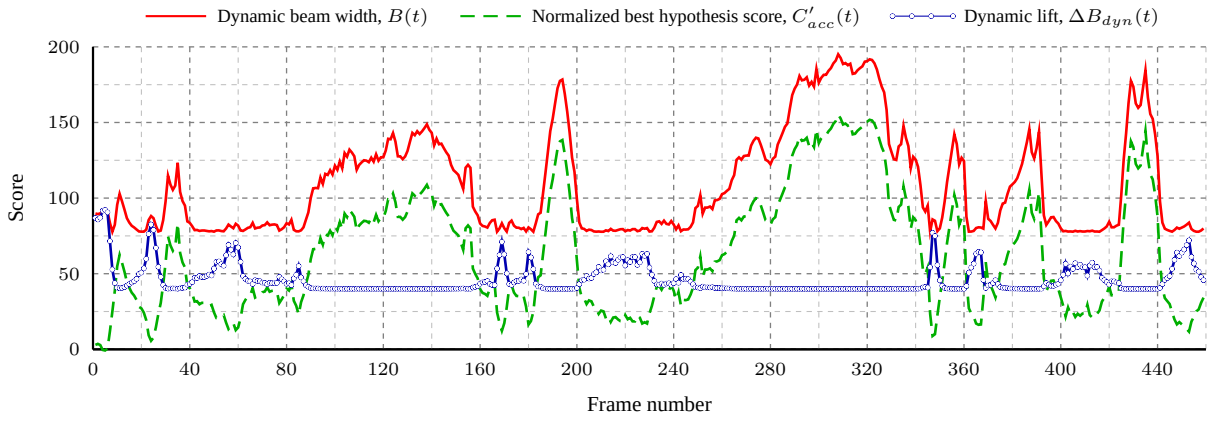

Figure 2: *Example for dynamic beam width during the appointment negotiation utterance 'Ja genau, lassen wir gleich die letzte Woche im März, prima!' (English: That's correct, let's keep right the last week in March, great!)*

frame based on the confidence score of the best hypothesis $C'_{acc}(t)$ using Equation 3. The results we achieved with this dynamic approach for different T_{upp} and T_{low} can be seen in section 4.

The main challenge in computing C'_{acc} in HMM based systems is to get a correct assessment of $p(\vec{x})$. That is because they generally do not have dedicated models for this purpose. The computation of $p(\vec{x})$ requires the calculation of the emission of all HMMs which could be very time expensive. However Kamppari in [1] proposed a way to get managed this problem by the reduction of the catch-all model's size in term of the number of Gaussian components.

2.2. Catch all model generation

The process of reduction of catch-all model size is an iterative bottom-up clustering process. In each iteration step two Gaussians which are most similar to each other are found and then combined into a new one. As the measure for the similarity of two Gaussians the weighted Battacharyya distance measure will be used:

$$D_{Batt} = -\log \int \sqrt{P_1(x) \cdot P_2(x)}\, dx\,.$$

Battacharyya distance is a measure of overlap between two Gaussians with ranges between 0 and ∞ corresponding to full and no overlap. The specific implementation of D_{Batt} for Gaussians is

$$D_{Batt} = \frac{1}{8}(\vec{\mu}_1 - \vec{\mu}_2)^T \cdot \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \cdot (\vec{\mu}_1 - \vec{\mu}_2) + \frac{1}{2}\ln\left(\left|\frac{\Sigma_1 + \Sigma_2}{2}\right| \cdot |\Sigma_1|^{-1/2} \cdot |\Sigma_2|^{-1/2}\right),$$

where μ_1 and μ_2 are the means of the Gaussians and Σ_1 and Σ_2 the covariance matrices. D_{Batt} is scaled to compress the acoustic space so that the entire acoustic space is covered with acceptable resolution using weights of the Gaussians w_1 and w_2:

$$D_{scale} = \sqrt{\frac{w_1^2 + w_2^2}{2 w_1 w_2}}\,.$$

In HMM systems these weights can be computed based on the weights of the mixture distribution functions (s. [7] for details). The combination of the most similar Gaussians will be processed based on Equation 5, 6 and 7 for each dimension d

$$w_{new} = w_1 + w_2 \tag{5}$$

$$\mu_{new,d} = \frac{w_1}{w_1 + w_2} \cdot \mu_{1d} + \frac{w_2}{w_1 + w_2} \cdot \mu_{2d} \tag{6}$$

$$\Sigma_{new,d} = \frac{w_1}{w_1 + w_2} \cdot \Sigma_{1d} + \frac{w_2}{w_1 + w_2} \cdot \Sigma_{2d} + \frac{w_1}{w_1 + w_2} \cdot \frac{w_2}{w_1 + w_2} \cdot (\mu_{1d} - \mu_{2d})^2\,. \tag{7}$$

After a new Gaussian was computed it is added to the pool of Gaussians of the catch-all model and the Gaussians from which the new one was created, are removed. This iteration is repeated as long as required to achieve the desired compression ratio of the acoustic space. As presented in [1] and [7] catch-all model allows acceptable estimation of the observation probability $p(\vec{x})$ even with a compression ratio of 95 %. Based on these findings the evalutaions for this work were also made with a catch-all model of the same compression ratio of 95 %. In that way we were able to reduce the complexity of the acoustic models, trained on the Verbmobil '96 training material, from about twenty-five thousand mixture components to about thousand.

3. Experimental setup

The evaluations described in this paper were performed on the commonly used speech recognition system HTK (Release 3.1) [8]. As test material we used the German Verbmobil '96 corpus [9], which contains 343 sentences, i.e. 6428 words. The computation of the scaling factors 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.

4. Experiments and results

The goal of the experiments presented in this section was to evaluate the capability of the confidence-guided dynamic prun-

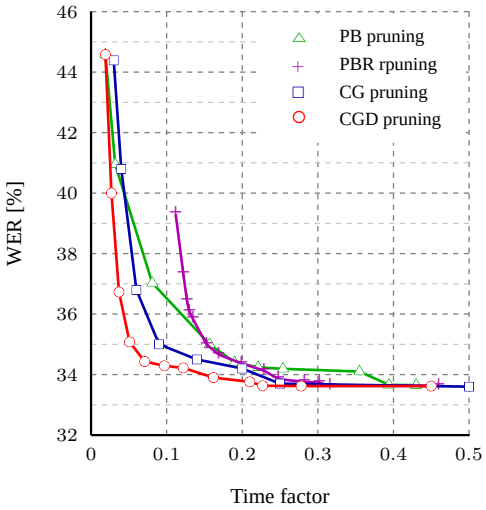

Figure 3: *Word error rates (WER) of different pruning techniques depending on time factor: probability-based beam width (PB), combined probability-based and rank (PBR), confidence-guided (CG), and confidence-guided dynamic pruning (CGD).*

ing approach to accelerate the search process of an ASR system. For this purpose we ran several tests on the Verbmobil '96 evaluation data (s. section 3) using different pruning techniques and parameters. Our results are presented in Fig. 3 which shows *word error rates* (WER) depending on the *time factor*. The time factor was 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. The tests were performed always on all utterances of the evaluation corpus.

We compared our dynamic pruning techniques with the classical probability-based fixed beam width pruning and with the combination of beam width and rank pruning. The curve of constant beam width in Fig. 3 was determined by computing the WER for the evaluation corpus using different beam width values B_{set} in a range of [80-250]. To the greater beam width value belongs lower WER but higher time factor. The combination of beam width and rank pruning was evaluated by keeping B_{set} at 210 and varying N_{max} in the range of [500-9000]. The curve of confidence-guided (CG) beam width was found using static beam width relative to the normalized score of the best hypothesis in a range of [55-200]. The curve of confidence-guided dynamic (CGD) beam width was plotted using $T_{upp} = 110$ and different T_{low} in a range of [20-70] (s. Equation 3 for details).

Fig. 3 shows that our confidence-guided dynamic beam pruning approach outperforms the static methods significantly. The time factor of the ASR using CGD pruning could be decreased to 0.23 without increasing WER. Furthermore if we accept an increase of WER of about 1 % CGD achieves a time factor of 0.07 which corresponds to the acceleration of the ASR about 14 times (reciprocal time factor). In comparison the classical constant beam width pruning achieves with the same WER increase a time factor of 0.19 (in accelaration 5 times). Further results of our evaluation tests can be shown in Table 1.

5. Conclusion

This paper has presented a novel dynamic beam width pruning method using confidence measurement for accumulated hypothesis score normalization. Confidence-guided pruning performs significantly better than classical pruning techniques. As a result a significant improvement in decoding speed of the ASR system could be achieved. This technique suggests a kind of dynamic beam control behavior, therefore the next step of our work will be to investigate the combination of confidence measurement and adaptive control pruning strategies [10].

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; $B_{set} = 210, N_{set} = 2000$	34.66	0.17
CG; $B_{set} = 200$	33.63	0.5
CG; $B_{set} = 90$	34.5	0.14
CGD; $T_{upp} = 110, T_{low} = 70$	33.63	0.23
CGD; $T_{upp} = 110, T_{low} = 40$	34.66	0.07

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

6. References

[1] 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.

[2] Fabian T., Lieb R., Ruske G., and Thomae M., "Impact of Word Graph Density on the Quality of Posterior Probability Based Confidence Measures", In Proceedings of EUROSPEECH, Geneve, Switzerland, 2003, pp. 917-920.

[3] Williams G. and S. Renals, "Confidence Measure for Hybrid HMM/ANN Speech Recognition, In Proceedings of EUROSPEECH, Rhodes, Greece, 1997, pp. 1955-1958.

[4] Ortmanns S., Eiden A., Ney H., and Coenen N., "Look-Ahead Techniques for Fast Beam Search", In Proceedings of IEEE ICASSP, Munich, Germany, 1997, Vol. 3. pp. 1783-1786.

[5] Liu F., Afify M., Jiang H., and Siohan O., "A New Verification-Based Fast Match Approach to Large Vocabulary Continuous Speech Recognition", In Proceedings of EUROSPEECH, Aalborg, Denmark, 2001, pp. 851-854.

[6] Renals S. and Hochberg M., "Start-Synchronous Search for Large Vocabulary Continuous Speech Recognition", IEEE Trans. Speech and Audio Processing, 1999, pp. 542-553.

[7] Abdou S. and Scordilis M.S, "An Efficient, Fast Matching Approach Using Posterior Probability Estimates in Speech Recognition", In Proceedings of EUROSPEECH, Geneve, Switzerland, 2003, pp. 1161-1164.

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

[9] 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.

[10] 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.