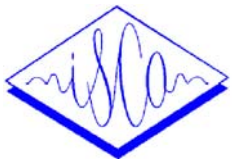
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Dynamic Beam Pruning Strategy Using Adaptive Control

Dongbin Zhang, Limin Du

Labs for Speech Interaction Technology Research
Institute of Acoustics, Chinese Academy of Sciences
{zhangdb,dulm}@iis.ac.cn

Abstract

In large vocabulary continuous speech recognition, huge search space results in vast computational cost. While most pruning search strategies can reduce the computation, but the recognition rate often decreases. This paper aims to reduce the computation time without any sacrifice of the recognition rate. By means of the adaptive control theory, a novel pruning method is presented. It can automatically steer beam to make search space attain an expected size. In order to make the decoder more efficient, the average number of active model instances is used as the dynamic reference of the adaptive system. Compared with the baseline system with fixed beam and histogram pruning, the proposed method leads to a significant reduction in the computation time and a slight improvement in word accuracy.

10.21437/Interspeech.2004-139

1. Introduction

Automatic speech recognition (ASR) is an essential component for human-computer interactive systems. Recent advances have brought the state-of-the-art ASR systems close to reality. When the vocabulary size becomes large, the search space increases to a prohibitively high level. Beam search enables the system to avoid an exhaustive search and focuses the search on those hypotheses that are most likely to result in the best state sequence. Unfortunately, a path that is unlikely up to now may become more likely in the future. Therefore, the challenge of designing a good pruning mechanism is to reduce the computation time maximally without removing useful hypotheses.

Recently, a lot of different pruning mechanisms have been proposed in literatures. Among them, there are several major heuristics, rank-based (histogram pruning), score-based (acoustic pruning/language model pruning), or the combination of these two. Some new data-driven techniques also have been brought into use for optimizing these methods. Hamme and Aellen [1] applied adaptive control theory and used a standard Self-Tuning Regulator (STR) to control pruning process. But as they said in the paper, the method was more suitable for reducing computation than for limitation of the memory resource.

This paper goes one step further and presents a self-tuning pruning strategy with dynamic target tracing. Experimental results demonstrate that the proposed method can reduce the computation time greatly with a slight improvement in accuracy.

2. Conventional pruning techniques

Viterbi search and all its variant forms belong to time-synchronous and breadth-first searching techniques, which often make the searching space huge. Pruning is to reduce the searching space.

Score-based beam search is widely used in speech recognition. At each time frame, only hypotheses with a probability, which is no less than a preset threshold T from the best hypothesis, are kept for further consideration. Usually, the preset threshold is fixed for the whole recognition process, which is called Fixed-Beam pruning (FB Pruning). However, if all scores are close to each other, only few paths can be cut-off. It may force the system to a prohibitive complete search ins such worst case.

Rank-based beam search avoids this problem, by retaining top N candidates for further consideration. Similar to score-based beam search, N is normally a preset fixed value. This method can be called as Fixed-Height Histogram pruning, such as Maximum Active Phone Model Instance (MAPMI) pruning [5][6]. In large vocabulary continuous speech recognition, histogram pruning usually performs better than fixed beam pruning [4]. A larger N results in a lot of unlikely hypothesis, and takes computational cost greatly, while a smaller N may decrease the recognition rate.

However, MAPMI pruning is not economical. Masses of overshoot waste computation time and contribute little to the recognition rate. Hence, the combination of MAPMI pruning and Beam pruning is used instead of the MAPMI pruning alone. In the following experiments, it's verified that using the combined pruning can gain the same word accuracy rate (WAR) with much less computation time. Our baseline system is a paradigm of such combination (MAPMI + FB Pruning). The proposed method is used to optimize the combined pruning further.

All the conventional pruning methods cannot adjust beam width dynamically to fit the time-variant requirement precisely. In order to make the decoding most economically, adaptive control method is applied to develop a self-tuning beam pruning system.

3. Pruning control mechanism

The proposed pruning system includes the following two steps:

- At every time frame, it can automatically regulate beam-width to make the number of active phone model instances (PMI) close to a preset value.
- Based on the one-step system, the preset value is chosen dynamically to guide the number of active PMI to make the decoder system more economic.

3.1. Adaptive pruning system

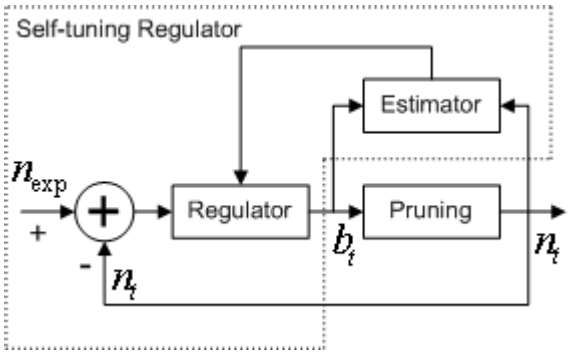


Figure 1: Dynamic beam pruning procedure using a standard Self-Tuning Regulator.

Our pruning system is based on a standard STR of adaptive control theory; a block diagram of this system is shown in Figure 1. The adaptive controller can be considered as being composed of two loops. The inner loop consists of the pruning process and an ordinary feedback regulator. The parameters of the regulator are adjusted by the outer loop, which is composed of a recursive parameter estimator. This regulator can automatically tune its parameters to obtain the desired properties of the closed-loop system [5].

In Figure 1, b_t represents beam width at time t, it's regarded as the input value of the pruning process, and the output n_t represents current number of active PMI. The reference signal of the STR $n_{\exp}$ is the expected number of active PMI in MAPMI pruning. The goal of the system is to drive n_t close to $n_{\exp}$.

The pruning process can be regard as a non-linear time-variant dynamic system. For simplicity, a simple zero order difference equation is used to describe the pruning process,

$$n_t = \theta_t b_t \tag{1}$$

where θ_t, the parameter of the pruning process, is a time-variant gain, which reflects the relation of beam width and the number of active PMI at t frame. The regulator is an integrator,

$$b_{t+1} = b_t + \alpha(n_{\exp} - n_t)/\hat{\theta}_t \tag{2}$$

where α, the parameter of the regulator, can affect the response speed of the feedback loop. Here, a reasonable setting for α is 0.2. Based on observations of beam width and number of active PMI in past L frames, $\hat{\theta}_t$ is an estimation of θ_t. A good setting for L is 5, as reported in [1].

Using least squares estimation, $\hat{\theta}_t$ can be got from the following equation:

$$\hat{\theta}_t = \frac{\sum_{i=1}^{L} n_{t-i} b_{t-i}}{\sum_{i=1}^{L} b_{t-i}^2} \tag{3}$$

By means of the STR, the one-step system can steer n_t close to fixed $n_{\exp}$ quickly by the feedback loop. The experimental results show the one-step system can gain a slightly better recognition rate with fairly faster speed than the baseline system.

But in some cases, especially at the beginning and the tail of an utterance, it's impossible to reach the preset $n_{\exp}$. The regulator however increases the beam width greatly to attempt impossibility. Obviously, it's expensive in computation.

3.2. Expected number-of-active-PMI curve

In order to avoid the above problem and decrease the computation time further more, $n_{\exp}$ should be heuristic to guide n_t precisely.

For this purpose, the mechanism of the two different pruning systems should be considered. Figure 2 shows the different topology of the two pruning methods at t time frame.

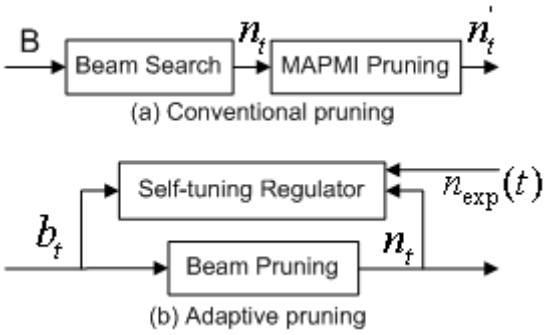


Figure 2: Comparison of the two pruning mechanisms

To make the two-step system have the same recognition rate as the baseline system, $n_{\exp}(t)$ should reflect the change of n_t'.

In baseline system, the number of active PMI after pruning (n_t') often obey such a "law": increasing rapidly at the beginning, and keeping close to the height of MAPMI pruning, then falling a little smoother near the end than at the beginning. The mean-value curve of n_t' is shown as the solid line in Figure 3, while the dash-dot line represents the variance bound. The mean value of n_t' is chosen as $n_{\exp}(t)$, which is estimated at every time interval. Considering of the unequal length in each utterance, normalized frame indices and the corresponding $n_{\exp}(t)$ are used to form a curve, which is called beam search ranking curve (BSRC) [6].

$$n_{\exp}(t) = \frac{\sum_{i=1}^{n_{total}} n_t^i}{n_{total}} \quad (4)$$

In equation (4), t is the percent of frame indices, n_t^i represents the number of active PMI after combination pruning ($n_t^{'}$) at time t in i_{th} utterance, and n_{total} represents the total number of utterance in the train-set. Since only 100 sample hits are chosen in whole frame indices, n_t^i must be obtained by linear interpolation.

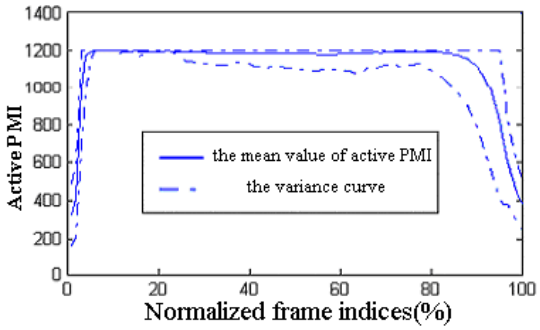


Figure 3: Expected number of active PMI curve estimated on baseline system with combination of FB (where T=300) and MAPMI pruning (where N=1200).

As can be seen from Figure 3, the curve of $n_{\exp}(t)$ "covers" a little less area than fixed $n_{\exp}$ of the one-step system (where $n_{\exp}$=1200), which demonstrates a little less search space used in the two-step system. After using the time-variant expected number of active PMI, the self-tuning pruning will be more flexible.

4. Evaluation

The decoding algorithm is one-pass Viterbi beam search. The speaker-independent recognition task is performed on a loop-back network of all the 408 Chinese syllables.

In this study, all utterance for the test-set and the $n_{\exp}(t)$ estimate-set are randomly selected from Chinese 863 Speech Corpus. The $n_{\exp}(t)$ estimate-set consists of 5146 utterances by 8 people (4 male and 4 female), while the test-set is an entirely separate set of speech data that consists of 800 utterances by other 8 people (4 male and 4 female). The average utterance length is 3.4 seconds in the test-set, while 6.7 seconds in the estimate-set.

Analysis condition of the feature extraction is summarized in Table 1. All the experimental evaluations are performed on an 866MHz PIII processor with 512MB RAM, and every experimental setting described in the following was tested three times to obtain averaged results.

The real-time factor (RTF) is defined as the ratio of total computation time to total length of processed speech, which is used to judge speed performance of the decoding system with different pruning methods.

Sampling Rate	16kHz
Pre-emphasis	1-0.97 z^{-1}
Windowing	Hamming (25msec)
Frame size	10msec
Feature Vectors	39

Table 1: Analysis condition of the feature extraction.

4.1. Baseline system

Our baseline system is the combination of MAPMI Pruning and FB Pruning. RTF are shown in Table 2, it shows the speed performance of the baseline system. The WAR is converged at T=300. While at T→∞, the baseline system changes to a true MAPMI pruning system. As can be seen from Table 2, compared with true MAPMI pruning, the combined pruning (at T=300) reduces 28.4% of search space and 21.1% of computation time.

Pruning Beam Width (T)	*WAR (%)*	*Average active PMI/frame*	*RTF (x RT)*
100	70.25	798.9	1.22
150	71.44	1713.9	2.23
200	72.35	2671.9	3.12
250	72.52	3481.4	3.80
300	72.67	4038.8	4.26
350	72.67	4440.0	4.57
∞ (MAPMI alone)	72.67	5634.2	5.40

Table 2: Experimental evaluations of baseline system (where N=1200).

4.2. Self-tuning beam pruning

n_t is a good measure for the search space, which has direct impact on the computation time. The behavior of n_t, at t time frame in the baseline system and two adaptive systems (the one-step and the two-step STR pruning) are shown in Figure 4. The area under each curve can represent search space. The two adaptive systems "covers" quite less search space than the baseline system. Compared with the one-step pruning, the two-step system is cheaper. Near the end of utterance, n_t still attempts to withhold 1200 active PMI especially in the one-step system,, while descends reasonably in the two-step pruning system.

As can be seen from Table 3, compared with the baseline system (MAPMI+FB pruning, where T=300), the two-step STR system can save 71% of search space and 55% of computation time, with an improvement in WAR. And most important, compared with the one-step system, the two-step STR pruning can reduce 4.0% in computation time and 5.3% in search space at a slightly better WAR.

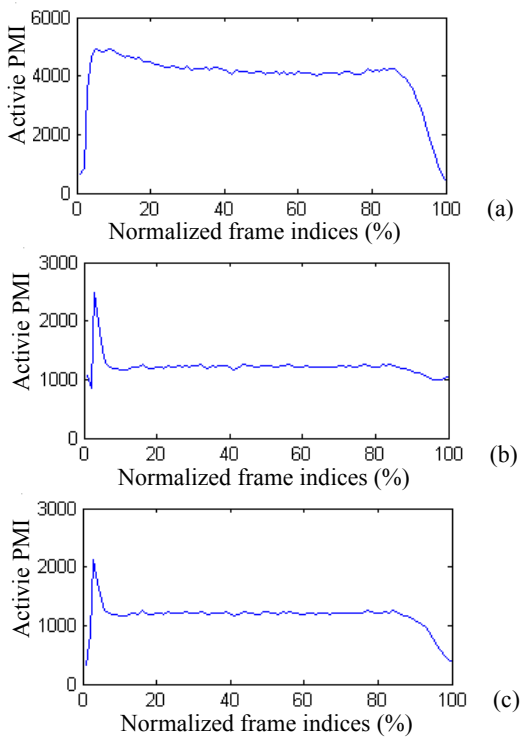


Figure 4: The behavior of n_t in two pruning systems: (a) combined pruning (T=300, N=1200); (b) for the one-step STR pruning (where $n_{\exp}$ =1200); (c) for the two-step STR pruning with the time-variant $n_{\exp}(t)$.

Pruning Mechanism	*WAR (%)*	*Average active PMI/frame*	*RTF (x RT)*
MAPMI Pruning (N=1200)	*72.67*	*5634.2*	*5.40*
Baseline system (N=1200, T=300)	*72.67*	*4038.8*	*4.26*
One-Step STR ($n_{\exp}$ =1200)	*72.68*	*1237.1*	*2.00*
Two-Step STR	*72.72*	*1172.1*	*1.92*

Table 3: Comparison of four pruning mechanisms at the same MAPMI (where N or $n_{\exp}$ =1200). Time-variant $n_{\exp}(t)$ in the two-step STR is trained under the baseline system, where N=1200, T=300.

5. Conclusion

In this paper, a data-driven adaptive-beam pruning method is presented. By using the dynamic expected number of active PMI, beam-width is self-tuned to make the number of active PMI follow the expected direction. The system is apt to decrease the computation much more greatly with no loss in word accuracy.

6. References

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

[2] Odell, J. J., Valtchev, V., Woodland, P. C. and Young, S. J., "A One-Pass Decoder Design for Large Vocabulary Recognition," in Proceedings of the DARPA Human Language Technology Workshop, pp. 405-410, March 1995.

[3] Deshmukh, N., et al., "An Efficient Public Domain LVCSR decoder," in Proceedings of Hub-5 Conversational Speech Recognition Workshop, Washington D.C., September 1998.

[4] Tran, B. H., Steinbiss, V. and Ney, H., "Improvement in beam search," In Proceedings ICSLP, pp. 2143-2146, 1994.

[5] Astrom, K. J., Wittenmark, B., *Adaptive Control (2nd Ed.)*, Addison-Wesley, 1995.

[6] Jang, J.-S. and Lin, S.-S., "Optimization of Viterbi Beam Search in Speech Recognition," In Proceedings ISCSLP, pp. 177-180, 2002.