Reducing Computation on Parallel Decoding using Frame-wise Confidence Scores

Tomohiro Hakamata, Akinobu Lee, Yoshihiko Nankaku, Keiichi Tokuda

Department of Computer Science and Engineering,
Nagoya Institute of Technology
Nagoya, 466-8555 Japan
{cab5,ri,nankaku,tokuda}@ics.nitech.ac.jp

Abstract

Parallel decoding based on multiple models has been studied to cover various conditions and speakers at a time on a speech recognition system. However, running many recognizers in parallel applying all models causes the total computational cost to grow in proportion to the number of models. In this paper, an efficient way of finding and pruning unpromising decoding processes during search is proposed. By comparing temporal search statistics at each frame among all decoders, decoders with relatively unmatched model can be pruned in the middle of recognition process to save computational cost. This method allows the model structures to be mutually independent. Two frame-wise pruning measures based on maximum hypothesis likelihoods and top confidence scores respectively, and their combinations are investigated. Experimental results on parallel recognition of seven acoustic models showed that by using the both criteria, the total computational cost was reduced to 36.53% compared to full computation without degrading the recognition accuracy.

Index Terms: parallel decoding, robust speech recognition, multi model recognition, confidence measure, search.

1. Introduction

Speech recognition using multiple models has been studied recently to overcome diverse acoustic conditions and variety of target speakers on real-world speech applications. A multi-mixture model [1], combining mixture components of several acoustic models trained from different training sets independently, can achieve better accuracy than a single model with multi-condition training. However, multi-mixture method requires the acoustic models to have strictly the same model structure.

Another multi-model recognition scheme is the parallel decoding [2], where input speech is recognized by several decoders for each model / feature setup, and then the transcribed outputs are integrated to produce the final result [3]. The results can be further combined to produce better result on the basis of classifier combining, typically called as ROVER [4]. This approach is promising in that it allows different feature sets, acoustic model structures, and even different language models to be applied simultaneously, which can lead to robust recognition against real-world utterances. This approach can be further applied to multi-environment, multi-domain, multi-speaking-style, and multi-language recognition where the possible models are processed concurrently.

However, the increasing computational amount may prevent the parallel decoding from its practical use. The computational cost will grow proportional to the number of combinations. Since the combination of conditions (i.e., target speakers, noise conditions, utterance topics) becomes exponentially larger for rich application, it is essential to reduce the computational cost in the framework of multi-model speech recognition.

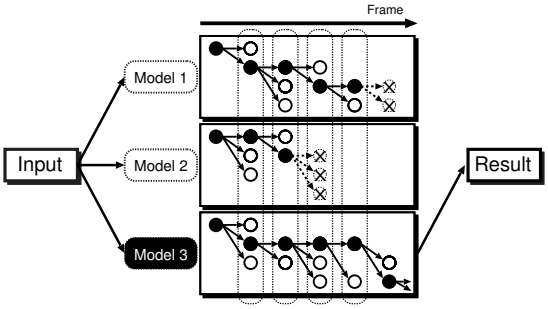

Figure 1: Inter-decoder frame-wise pruning.

In this paper, an efficient parallel decoding scheme pruning the decoders in the middle of recognition is investigated. Frame-wise search statistics extracted from the running decoders at each input frame are compared concurrently during recognition to judge which recognition processes should be terminated as unpromising ones. Several search statistics based on maximum hypothesis likelihoods and confidence scores are investigated to estimate which recognizer would have the best discriminative ability.

2. Inter-decoder pruning

2.1. Basic idea

In the parallel decoding scheme, the search is performed on all the possible model sets for each input, and then the best result will be chosen or the results are combined to form lattice. In this scheme, the computation for unpromising models that will not contribute to the final result can be cut off. By terminating decoders at the time of being determined as unpromising, the total computational cost can be saved.

The proposed pruning process is illustrated in Figure 1. Several search statistics are gathered among all decoders on each frame of beam-synchronous search step, and the ones whose scores get relatively low will be terminated at the frame and will not be computed further. We call this method "inter-decoder pruning" in this paper. Though there is a mathod for selecting the reserve of the model before decoding by supervised model selection,

evaluating models during decoding is examinated for the realtime quality.

Another multi-model recognition approach is to handle models on a single decoder and search with independent hypotheses for every model sets to find the best result among the model variations. In this framework, the hypothesis for the most matched models will be survived on the basis of normal recognition procedure. However, using various kinds of language model and acoustic model simultaneously at single decoder can be a loss of computational efficiency.

Previous works on parallel recognition have focused on consulting partial hypotheses among decoders during search to strengthen most common word hypothesis [5]. However, this method assumes that the vocabularies are common, and also assumes hypothesis equally appeared among decoders to be promising. The later assumption will not always be true under various diverse conditions.

2.2. Pruning measures

The pruning performance depends on how early and how stably it can find unpromising decoders as early as possible in the middle of search process, while keeping accuracy. We investigate two criteria based on maximum hypothesis likelihood and posterior probability based confidence scores, which can be obtained on frame-wise recognition process.

2.2.1. Maximum hypothesis likelihood

The hypothesises on an mismatched model tend to get lower likelihoods than on matched models. Thus, the maximum hypothesis likelihood at each frame can be used as a pruning criteria to determine which model is mismatched toward the input, if the acoustic models have similar resolutions (i.e., model parameter size).

Let $[w, t]$ denote a specific hypothesis of word w that ends at frame t, and $W_{\rm best}(k)$ indicates the best hypothesis path from start that ends at hypothesis k. The word likelihood $g_m([w, t])$ is computed on model set m,

$$g_m([w,t]) = \log P(x_t|W_{\rm best}([w,t]))P(W_{\rm best}([w,t])) \quad (1)$$

where x_t indicates input sequence from frame 0 to t. Here let $W_m(t)$ denote the set of survived words at frame t on decoder m. The maximum hypothesis likelihood of each decoder $g_m(t)$ is then defined as

$$g_m(t) \quad = \quad \max_{w \in W_m(t)} g_m([w,t]). \quad (2)$$

When recognition, at frame t, a decoder m will be terminated if it meets the criterion

$$g_m(t) + g_{\rm off} \quad < \quad \max_m g_m(t) \quad (3)$$

where the offset threshold $g_{\rm off}$ is a fixed value.

Additionally, a parameter $start$ should be defined to skip the first $start$ frames. It is needed to avoid pruning only by the scores of noisy input, which often exists at the beginning part of a speech input. Thus the parameter $start$ should be specified to start pruning after that frame.

When two or more decoders are survived till the end of input T, the one which gets the maximum value of $g_m(T)$ will be selected as the final result.

2.2.2. Top confidence score

Confidence scoring based on posterior probabilities is one of the popular methods to assign confidences to the speech recognition results. We propose using the confidence score computed from likelihoods of survived hypotheses at each frame as a pruning criteria to be compared between decoders. Since the posterior probability reflects the relative distribution of word likelihoods among competing alternatives [6] and their sum is always 1, the top confidence scores can be counted as a measure of discriminative ability of the model.

At each frame, confidence scores of the survived hypotheses on each decoder can be calculated based on their posterior probabilities derived from their accumulated likelihoods at that time. Let τ denote the starting time and t the ending time of word w. $W_{[w;\tau,t]}$ denotes all paths that contain the hypothesis $[w; \tau, t]$. The posterior probability $P([w; \tau, t]|x)$ of a specific word hypothesis $[w; \tau, t]$ over the whole acoustic observations x_T can be computed by summing up the posterior probabilities of all paths which contain the hypothesis $[w; \tau, t]$,

$$P([w;\tau,t]|x_T) \quad = \quad \sum_{W \in W_{[w;\tau,t]}} \frac{P(x_T|W)P(W)}{P(x_T)}. \quad (4)$$

While recognition process, by approximating the sum of all paths by the current best path [7], the posterior probability of the hypothesis at t ($0 < t \leq T$) can be computed as

$$P([w;\tau,t]|x_t) \quad = \quad \sum_{W \in W_{[w;\tau,t]}} \frac{P(x_t|W)P(W)}{P(x_t)} \quad (5)$$

$$\approx \quad \frac{P(x_t|W_{\rm best}([w,t]))P(W_{\rm best}([w,t]))}{P(x_t)} \quad (6)$$

Since $P(x_t)$ is approximated by the sum over all existing paths at x_t, the posterior probability of a word w on time t at decoder m can be expressed as follows:

$$P_m([w,t]|x_t) \quad \approx \quad \frac{e^{g_m([w,t])}}{P(x_t)} \quad (7)$$

$$\approx \quad \frac{e^{g_m([w,t])}}{\sum_{w' \in W_m(t)} e^{g_m([w',t])}}. \quad (8)$$

Thus, its confidence score is defined with a scaling factor α,

$$C_m([w,t]) \quad = \quad \frac{e^{\alpha \cdot g_m([w,t])}}{\sum_{w' \in W_m(t)} e^{\alpha \cdot g_m([w',t])}}. \quad (9)$$

We assume that the discriminative ability of the model set on a decoder can be estimated by the top word confidence score. When a word in a decoder has high confidence, it can be expected that the models in the decoder discriminates the word well from other competing alternatives. Actually, the sum of top N confidence scores is used to get the overall decoder confidence $C_m(t)$,

$$C_m(t) \quad = \quad \sum_{\text{best } N \text{ words at } t} C_m([w,t]). \quad (10)$$

When recognition, a decoder m will be terminated if it meets the criterion

$$C_m(t) + c_{\rm off} \quad < \quad \max_m C_m(t). \quad (11)$$

Since the confidence score is normalized, it is possible to compare the scores of competing hypotheses on the same scale among models. This is the distinctive feature compared with the word likelihood.

3. Experiment

The proposed method was evaluated through an experiment below. The proposed methods were implemented on an open-source recognition engine Julius [8] rev.3.5. Since Julius is a 2-pass decoder, the proposed pruning methods are implemented at the first pass of frame-synchronous beam search. The second pass will be executed for only survived decoders to get the final result.

3.1. Set up

All acoustic and language models available from the Continuous Speech Recognition Consortium [9] were used.

The training databases are ATR-BLA and JNAS+ASJ. The ATR-BLA database consists of 3,769 speakers, total 162 hours of spontaneous speech. The JNAS+ASJ database consists of 361 speakers, total 98 hours of read speech.

Seven speaker-independent acoustic models were chosen to this experiment as listed in Table 1. "JNAS-PTM" and "JNAS-tri" are trained using JNAS database, and "ATR-PTM" and "ATR-tri" are using ATR database. "Senior-PTM" is trained from senior people's utterances with the same amount and sentences of JNAS. "Child-PTM" is trained using 100 reading utterances of 400 children. "JNAS-Tel-tri" is trained using JNAS database, with bandwidth limited for telephone-based recognition, so this is unmatched model. The suffixes "-PTM" and "-tri" denote the model structures: the former is phonetic tied-mixture model and the latter is shared-state triphone model. Speech waves were analyzed with 25-ms Hamming window, and the sampling rate is 16kHz. The feature vectors had 25 elements comprising of 12 MFCC, their delta, and delta energy. The Language model is a word 3-gram of 20k words.

The test set were gathered from the two database. 50 utterances by 23 adult male speakers and 23 adult female speakers are extracted from JNAS database, and 50 utterances by four adult male speakers are recorded for ATR database. They are not included in the training set.

The beam width was set to 2000, and the scaling parameter of confidence scoring α was set to 0.05 for all experiments.

3.2. Evaluation measure

In addition to word error rate (WER), an approximate computational amount measure $Cost$ is defined as the rate of processed frames over all the decoders to assess the efficiency of proposed decoder pruning. For instance, $Cost = 100.00$, if all the decoders are fully computed, and $Cost = 14.29$, if only a single model is computed while other six decoders had been terminated at the first frame.

3.3. Preliminary Experiments

First, the performance of each model was examined. Word error rates of all acoustic models when each model was used individually are listed in Table 1. In this experiment, the best word error rate of 9.40% was obtained by applying ATR-PTM model. In this case, $Cost$ is calculated as 14.29 as stated in section 3.2.

Next, the performance of conventional parallel decoding was examined. All models are fully computed as conventional parallel decoding, and the best result was selected after all the models are computed. The results are shown in Table 2. By selecting the best model at each utterance by hand, a word error rate of 5.36 % was obtained. This is the upper bound when applying automatic selection. It was confirmed that multi-model speech recognition with parallel decoding gives slightly lower word error rates than single-model one.

Table 1: WER by single model.

model	WER		
	JNAS	ATR	Total
JNAS-PTM	7.28%	21.37%	10.11%
ATR-PTM	7.70%	16.16%	9.40%
Senior-PTM	11.81%	27.22%	14.90%
Child-PTM	39.81%	36.97%	51.16%
JNAS-Tel-tri	96.67%	92.12%	95.77%
JNAS-tri	6.56%	21.49%	9.55%
ATR-tri	7.27%	20.16%	9.85%

Table 2: WER and cost by static model selection.

method	WER	$Cost$
best of single model (ATR-PTM)	9.40%	14.29
oracle model selection	5.36%	100.00
selection by likelihood of final result	8.94%	100.00
selection by likelihood of first pass	9.33%	100.00

3.4. Results

Examples of transitions of the two criteria based on maximum hypothesis likelihood and top confidence score are plotted on Figure 2 and Figure 3, respectively. A reliminary experiment has shown that the performance using confidence score is optimal at $N = 4$ on this test set, accodingly we use that value at Figure 3. In Figure 2, both the child model and telephony model got much lower values, while other models are comparable. Thus it is considered that the likelihood should reflects the acoustic space of the training set. In Figure 3, not only the child model and telephony model, but also senior model gets lower value in average. It is considered that decoders with relatively much lower confidence than the maximum can be terminated.

Finally, the total relationship between WER and $Cost$ when using the proposed two criteria separately and together are plotted in Figure 4. The performance was measured while changing the parameter values $g_{\rm off}$, $start$, $C_{\rm off}$ and N. Table 3 summarizes

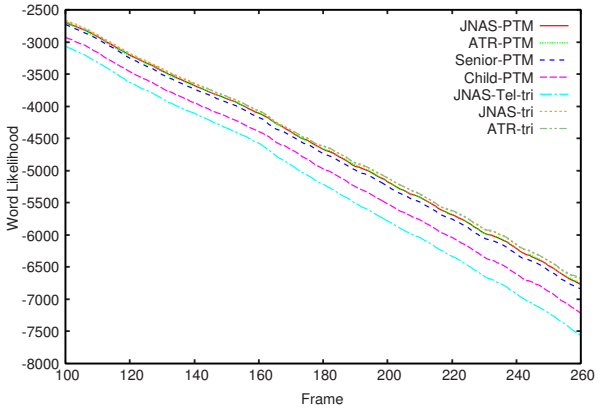

Figure 2: Likelihoods of maximum hypothesis per frame.

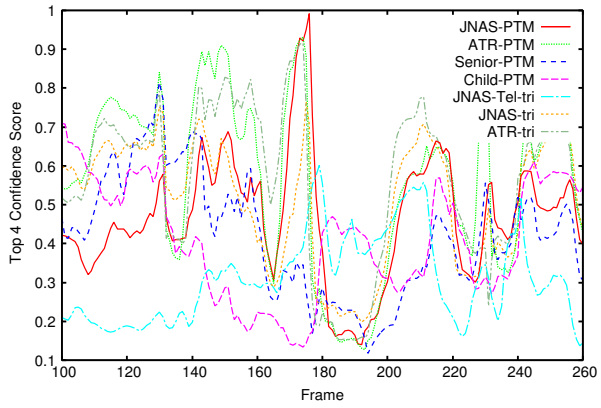

Figure 3: Sum of top 4 confidence scores per frame.

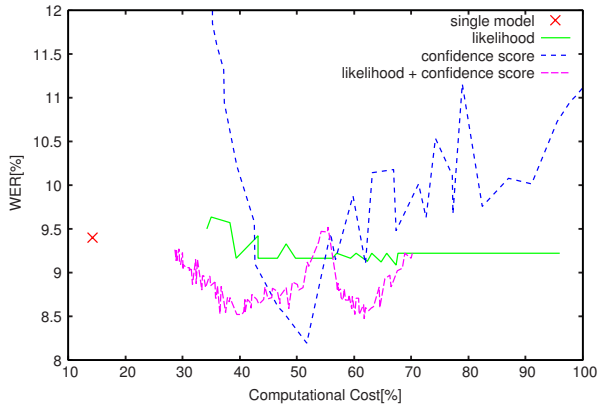

Figure 4: WER and cost by proposed method.

the best result, where $Cost$ becomes minimal while best keeping WER equal to the static model selection.

The likelihood-based pruning reduced the computational cost to 39.42, with a slight loss of accuracy. On the contrary, by confidence-based pruning, lower WER was obtained though some computational amount increased. Conclusively, $Cost = 36.53$ was achieved by using combination of both criteria.

By using only word likelihood as criterion, the correlation was not seen in the change in computational cost and WER. The range of word likelihoods among each models changes place rarely therefore, the selected model did not depend on the setting of offset threshold g_{off} strictly.

Via the combination of word likelihood and confidence score as pruning criteria, the result was settled to a roughly better value. In the beginning of input frame, the model which the domain is quite different was rejected by word likelihood, and afterward, comparison among remaining models was advanced by confidence score, therefore high accurate model selection was conducted by low computational cost.

Table 3: WER and cost by proposed method.

criteria	WER	cost
word likelihood	9.17%	39.42
confidence score	8.61%	46.75
word likelihood + confidence score	8.53%	36.53

4. Conclusions

An efficient parallel decoding scheme which contributes to the reduction of computational cost has been introduced. Our method judges model reliability toward the input in the middle of recognition process by comparing maximum hypothesis likelihoods and top N confidence scores among surviving decoders at each input frame. Experiments demonstrated that the proposed method has an ability to achieve much computational reduction in the framework of parallel speech recognition without impairing recognition accuracy. Future work will be dedicated to examination on broad acoustic condition, language model combination and more stable criteria, and compare this method with other approach, such as supervised model selection.

5. Acknowledgements

A part of this work is supported by the e-Society project provided by MEXT, Japan.

6. References

[1] Motoyuki Suzuki, Yusuke Kato, Akinori Ito, and Shozo Makino, "Construction method of acoustic models with various backgroundnoises based on combination of HMMs," in *Proc. INTERSPEECH2005*, 2005, pp. 973–976.

[2] Takahiro Shinozaki and Sadaoki Furui, "Spontaneous speech recognition using a massively parallel decoder," in *Proc. ICSLP*, 2004, pp. 1705–1708.

[3] Masahiko Matsushita, Hiromitsu Nishizaki, Yasuhiro Kodama, Takehito Utsuro, and Seiichi Nakagawa, "Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven web retrieval task," in *Proc. Eurospeech*, 2003, pp. 1205–1208.

[4] Jonathan G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *Proc. IEEE ASRU*, 1997, pp. 347–352.

[5] Yonghong Yan, Chengyi Zheng, Jianping Zhang, Jielin Pan, Jiang Han, and Jian Liu, "A dynamic cross-reference pruning strategy for multiple feature fusion at decoder run time," in *Proc. Eurospeech*, 2003, pp. 1177–1180.

[6] Simo O. Kamppari and Timothy J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. IEEE ICASSP*, 2000, pp. 1894–1897.

[7] Akinobu Lee, Kiyohiro Shikano, and Tatsuya Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," in *Proc. IEEE ICASSP*, 2004, pp. 793–796.

[8] Tatsuya Kawahara, Akinobu Lee, Kazuya Takeda, Katsunobu Itou, and Kiyohiro Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," in *Proc. ICSLP*, 2004, pp. 688–691.

[9] Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda, Masato Mimura, Atsushi Yamada, Akinori Ito, Katsunobu Itou, and Kiyohiro Shikano, "Continuous speech recognition consortium — an open repository for CSR tools and models —," in *Proc. IEEE LREC*, 2002, pp. 1438–1441.