On Improvements to CI-based GMM Selection

Arthur Chan, Mosur Ravishankar, Alexander I. Rudnicky

Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, 15213.
`{archan, rkm, air}@cs.cmu.edu`

Abstract

Gaussian Mixture Model (GMM) computation is known to be one of the most computation-intensive components of speech recognition. In our previous work, context-independent model based GMM selection (CIGMMS) was found to be an effective way to reduce the cost of GMM computation without significant loss in recognition accuracy. In this work, we propose three methods to further improve the performance of CIGMMS. Each method brings an additional 5-10% relative speed improvement, with a cumulative improvement up to 37% on some tasks. Detailed analysis and experimental results on three corpora are presented.

1. Introduction

Most modern large vocabulary continuous speech recognition (LVCSR) systems rely on *continuous density* hidden Markov models (HMMs) for acoustic modeling. They consist of thousands of HMM states, each state modeled by a separate *Gaussian mixture model* (GMM), consisting of tens of multi-dimensional Gaussian densities. The function of the *GMM computation component* in an LVCSR system is to provide HMM state scores for the search routine. Given the large number of GMMs in the acoustic model, this computation is expensive unless done intelligently. In the past, several attempts have been made to speed up GMM computation [3-6]. The key to most of these speed-up techniques is being able to intelligently ignore some parts of the computation without significant loss of accuracy. As no single technique is sufficient to provide enough speed gain, it becomes necessary to apply several techniques simultaneously. A major concern in applying these methods in a practical system is that the individual techniques are not usually orthogonal to each other; gains from each do not necessarily accumulate. Researchers can have a hard time combining them together effectively.

In our previous work [1], a four-level categorization scheme of GMM computation was proposed. The basic idea is that fast GMM computation techniques can be categorized into four orthogonal levels (Figure 1):
1. **Frame-level**, determining which frames of speech should be considered in detail.
2. **GMM-level**, identifying which GMMs are to be evaluated in any given frame.
3. **Gaussian-level,** determining which densities are relevant within any given GMM.
4. **Component-level,** reducing the dimensionality of an individual Gaussian distribution.

This conceptual framework provides an effective way to easily incorporate multiple speed-up techniques into an LVCSR system. In principle, any arbitrary technique can be simply "plugged in" at the appropriate level, without disturbing techniques at other levels.

In our previous work, this framework was implemented and incorporated in the CMU Sphinx-3 recognizer. We evaluated five representative GMM speedup techniques and showed that the framework provides a significant speedup by a factor of 4-5, with only about 5% relative degradation in recognition word error rate (WER).

Among the techniques we evaluated, we observed that **C**ontext-**I**ndependent model based **GMM S**election (CIGMMS), first introduced in the Julius system [4], was a major contributor to the speed gains. The idea behind CIGMMS is simple: instead of computing the entire set of *context dependent* (CD) GMMs in each frame, the much smaller set of *context independent* (CI) GMMs is first computed. A beam is applied to the resulting scores, identifying the *active CI set* for that frame. Only those CD GMMs whose parent CI model is active are computed in detail. The remaining CD GMMs back-off to (or inherit) the parent CI GMM scores.
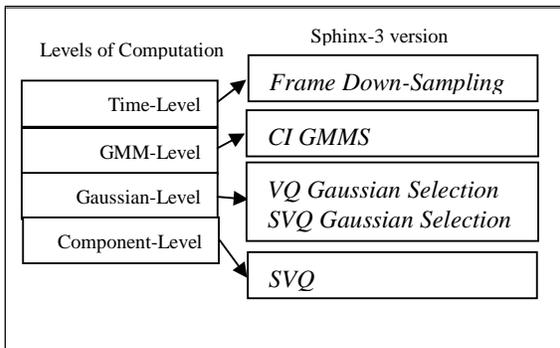
Figure 1. *Four Level Categorization Scheme of Fast GMM Computation proposed in [1].*

Despite its advantages, we observed that there are several issues with CIGMMS. In particular:

Issue 1. Unpredictable per-frame performance. When a beam is used to determine the active CI GMMs, the number of CD GMMs computed varies widely from frame to frame. If most CI GMM scores fall within the threshold, most CD GMMs have to be computed. In particular, noisy utterances can take much longer than average to decode.
Issue 2. Poor pruning characteristics. Since a large number of CD GMMs fall back (share) the same set of CI GMM

scores, a subtle consequence is that the search routine can no longer distinguish between competing hypotheses. This makes the pruning performance of search become less effective. The increased size of the search space increases the search cost, negating at least some of the speedup in GMM computation.

In this paper, we suggest three improvements to CIGMMS that address the above issues. The organization of the paper is as follows. In Section 2 we provide a brief description of our corpora and testing conditions. In Section 3 we describe **improvement I** for addressing Issue 1 above, by bounding the number of CD GMMs computed at every frame. In Section 4 we address Issue 2 and discuss other approximations to CD GMM scores, ending up with **improvement II** to CIGMMS. In Section 5 we discuss **improvement III**, a novel technique we call adaptive CIGMMS (A-CIGMMS), to further improve CIGMMS by dynamically varying its pruning threshold. In Section 6 we present evaluation results for each improvement, using the Communicator, WSJ 5k and the ICSI meeting tasks as benchmarks. Finally, we conclude the paper with Section 7. We note that each improvement provide significant speed-gains (5-10%) on multiple tasks.

2. Description of Corpora for Evaluation

All the experimental results presented in this paper are based on three different corpora: the Communicator task for planning air travel [7], the Wall Street Journal task [8], and the ICSI meeting transcription task [9][1]. The acoustic model configurations we used, optimized for each task to minimize WER, are summarized in Table 1. However, models with fewer component densities/GMM were also used for the analysis presented in Section 4.

Task	Vocabulary	#GMM	Comp/GMM
Communicator	2-3k	2000	64
WSJ-5k	~5k	5000	8
ICSI	12k	5000	32

Table 1. Corpora used in the experiments

3. Improvement I: CIGMMS with a bound on GMMs computed

As we described in the introduction, when a beam in used to prune the set of CD GMMs evaluated based on CI GMM scores, the number of CD GMMs computed at every frame tends to vary widely. As a result, this computation can take much longer than average for some utterances, especially if they are noisy. Clearly, this is undesirable for real-time applications. One simple way to deal with this problem is to modify the original CIGMMS scheme to limit the number of CD GMMs computed to some upper bound. In other words, incorporate absolute pruning into the scheme. The absolute pruning threshold is chosen such that it gets activated only when there are a large number of competing candidates to be evaluated. This reason this heuristic works is the same reason

---

[1] The test set we were using in the ICSI task were bdb001 and bro002 in the ICSI meetings. The model was trained by the rest of the 71 meetings.

that absolute pruning works in limiting the number of active models during search. Specifically, when a large number of models lie within the beam pruning threshold, they all have relatively good scores. In such a situation, the chances of picking out the correct candidate are not much better without absolute pruning anyway.

The following procedure summarizes the implementation:
1. Compute all CI GMM scores and sort them in descending score order.
2. Run down this ordered list, computing the corresponding CD GMM scores, until either the CI GMM scores become worse than the beam pruning threshold, or the absolute pruning limit on the number of CD GMMs evaluated is exceeded.

4. Improvement II: Using Best Gaussian Index

The other problem with the CIGMMS scheme is that with many CD GMMs backing off to the same CI GMM score, the search algorithm is unable to distinguish between competing candidates. The key to improving the pruning of search with CIGMMS is to introduce a better approximation scheme than backing off to a common score. One possibility is to use a score derived from the individual CD GMM models, but without incurring the cost of computing it entirely. In this section, we perform a detailed analysis of the behavior of a commonly known such scheme. We dub this the ***best Gaussian index*** (BGI) scheme.

The BGI scheme can be described as follows. When a particular GMM is fully computed in any frame, we also remember the index of the component density with the highest score. In the next frame, if this GMM is not fully computed, we use the stored best Gaussian index and compute only that component of the GMM. We note that when the chosen component is truly the best within the GMM in a given frame, its score is an excellent approximation to the full GMM score.

Despite the widespread knowledge of the BGI scheme, its behavior has not been fully documented in detail. The assumptions behind the above approximation can be broken down in two parts:
1. Given an input frame of speech and a GMM, the score of the single best Gaussian within the GMM is an excellent approximation to the score of the entire GMM.
2. Given any GMM, the best-scoring Gaussian in one frame is *usually* the same as in a neighboring frame (the principle of **locality** of the best Gaussian index).

Literature provides ample evidence showing that assumption 1 is accurate. For example, in [6] it was shown that the single best Gaussian contributes 95% of the likelihood or score of the entire GMM. This is the major assumption for all Gaussian selection ideas.
Assumption 2, however, has rarely been studied in detail in the past. A notable exception can be found in [5], in which the authors studied the behavior of GMM computation and measured the locality rate. They found that this number is roughly 70%. Unfortunately, there was no further analysis beyond that measurement.

There are two questions we need to ask about assumption 2:

1. **How does locality rate relate to the number of densities in a mixture?** The number of Gaussians can vary from system to system. Clearly, locality rate should decrease when the number of Gaussians per GMM increases.
2. **Can we assume that the high scoring GMMs in any frame also have a high locality rate?** It is intuitive that the highest-ranking GMMs play a more important role in guiding the search process. They are usually the "correct" models for a given frame of speech and therefore one should avoid underestimating their scores. For the BGI scheme to be successful, it is crucial that the high scoring GMMs have a high locality rate.

In Sections 4.1 and 4.2, we describe experimental results investigating the above two issues. Based on this analysis, in Section 4.3 we propose a modified scheme for CIGMMS that improve its pruning characteristics.

4.1. Effect of mixture size on BGI locality rate

In this experiment, we measured the locality rate of best Gaussian indices *vs.* the number of component densities/GMM in three different tasks. The results are shown in Table 2. Note that the models with higher densities/GMM were essentially derived by splitting the densities in a lower density/GMM model, and retraining. This is a common and representative method for obtaining several acoustic models with varying densities/GMM.

Task	1	2	4	8	16	32	64
Comm.	100	93.1	88.2	84.5	80.7	76.2	70.7
WSJ5k	100	90.7	84.7	80.3	NA	NA	NA
ICSI	100	89.5	82.6	77.3	71.7	64.6	NA

Table 2. *Variation of locality rate with GMM size for various tasks. (NA stands for "not available".)*

A couple of points are worth mentioning here. First, the locality rate decreases as the number of densities/GMM increases, as is to be expected. Second, the rate of decline of the locality rate depends on the corpora used, and perhaps conditions such as noisiness of speech.

4.2. Locality rate for high scoring GMMs

In the next experiment, we investigated the locality rate specifically for the high-ranking GMMs in different frames. In Table 3, we compare the locality rate for the 50-best scoring GMMs to that for all GMMs on average. The result is interesting: *the best scoring GMMs have a much lower locality rate.* This result is consistent across the different domains tested. The implication is that the BGI method should only be used selectively. In particular, it is not advisable to use the method directly for the high scoring GMMs.

Task	Best 50 GMMs	All
Comm. (64 den/GMM)	42.4	70.7
WSJ5k (8 mix)	67.5	84.7
ICSI (32 mix)	49.2	64.6

Table 3. *Locality rates for best 50 GMMs vs all GMMs.*

4.3. Summary of insights from the two experiments and a modified CIGMMS scheme

The previous two sections provide several insights about GMM approximation using the best Gaussian index. The first is that the number of densities per GMM dictates how aggressively one can pursue the BGI scheme for speeding up GMM computation. The second insight is that the BGI approximation is quite inappropriate for the high-ranking GMMs in general, owing to their poor locality rate. A blind application of this scheme would result in their scores being underestimated too frequently, and recognition accuracy to degrade significantly.

The above analysis suggests that the CIGMMS and BGI schemes might be complementary to each other, and the two may be used in conjunction very effectively. Assuming that the active CI set of the CIGMMS scheme accurately predicts which CD GMMs are high-scoring ones, the scheme ensures that these CD GMMs are evaluated fully. (The assumption is usually valid because CD GMMs are usually trained by bootstrapping from CI GMMs.) Thus, the problem of their poor BGI locality is avoided. For the remaining CD GMMs, instead of simply using the corresponding CI GMM scores, we use the best Gaussian index from the previous frame whenever possible. We know from the above analysis that these GMMs do enjoy a higher locality rate and hence the approximation is mostly valid. Thus, the use of approximate scores more specific to individual CD GMMs addresses the problem of increased search space in plain CIGMMS.

Of course, this scheme is slightly more expensive than the original version of CIGMMS evaluated in [1] and [4], since a CD GMM cannot simply back off to CI GMM score. However, their complementary nature is expected to improve recognition accuracy, while also providing better pruning characteristics.

5. Improvement III: Adaptive CI GMMS (A-CIGMMS)

In this section we explore the possibility of limiting the CD GMM computation even further, by linking CIGMMS to *frame down-sampling* [1]. We hypothesize that, even with CIGMMS and the absolute limit scheme of Section 3, in many frames too many CD GMMs are computed unnecessarily. At the other extreme, frame down-sampling (briefly mentioned in Section 1) ignores some frames of speech entirely, as far as GMM computation is concerned. Instead, one simply re-uses GMM scores from a previous frame. This technique is certainly effective in speeding up GMM computation, but is also very damaging to recognition accuracy. (Experimental results can be found in our evaluation in [1].)

One can explain the poor performance of frame down-sampling based on the analysis of section 4. Namely, it is not advisable to apply any approximation scheme to the high-scoring GMMs in any frame. What is needed is a version of CIGMMS that *approaches* frame down-sampling in those frames where most GMMs need not be fully computed.

We propose a new method, called Adaptive CIGMMS (A-CIGMMS) that effectively achieves this. First, we identify

selected frames to be "dropped" (as in frame down-sampling). Then, instead of ignoring the frame completely, we resort to CIGMMS, but with a much tighter beam for selecting the active CI set. The tighter beam is determined by multiplying the regular beam (in log-space) by a *tightening factor* (TF; $0<=TF<=1$). TF=1 is equivalent to normal CIGMMS, and TF=0 is equivalent to frame down-sampling. In between, we have a tradeoff between the two. Thus, far fewer CD GMMs are computed fully than with a normal CIGMMS beam. However, the high-scoring CD GMMs (predicted by the active CI set) will still be computed correctly.

6. Experimental Results

We evaluated the improvements described in this paper on the three corpora outlined in Section 2. To summarize, the methods evaluated are:

1. CIGMMS with an absolute pruning threshold on the number of CD GMMs computed (Sec. 3).
2. CIGMMS with non-computed GMMs using the BGI scheme (best Gaussian in the previous frame, Sec. 4.3).
3. A-CIGMMS (Sec. 5).

The experiments were carried out on a 2.2GHz P4 machine using Sphinx 3.5 [1]. Table 4 summarizes the experimental results. BL represents the baseline condition in which the following techniques were already applied:

1. The tightest (well-tuned) state, phone and word-level beam as well as histogram pruning.
2. The search structure optimized using a tree lexicon, with unigram look-ahead.
3. CIGMMS in its original form, without the improvements of this paper.
4. Sub-vector quantization-based Gaussian selection [10].

The improvements were applied to the baseline in succession, and cumulatively. The overall speed performance with the addition of each new method was measured. The results summarized in Table 4 are discussed below:

*   Improvement I provides significant speed gain on both Communicator and ICSI tasks (19% and 16% relative, respectively) with only a 3% relative increase in WER. The advantage of the scheme is twofold. First, the average number of CD GMMs computed per frame is reduced from 916 to 536 in the Communicator task, and from 1210 to 807 in the ICSI task. Second, the worst case speed improves from 2.08 xRT to 1.20 xRT for the ICSI task and 1.56 xRT to 0.96 xRT for the Communicator task. The WSJ5k task, however, is already highly tuned, and it is very hard to reduce the average number of CD GMMs computed. Still, its worst case behavior improves from 1.12 to 1 xRT.
*   Improvement II also provides significant gains on the Communicator and ICSI tasks (11 and 5%, respectively).
*   Improvement III provides consistent improvement on all three tasks. In our tests, we only experimented with tightening the CI beam every other frame, blindly. We expect that the gains could be higher if a more intelligent strategy of "ignoring frames" is used (e.g., based on similarity of successive frame feature vectors).

Task	Comm.	WSJ5k	ICSI Meeting
BL	12.85	6.73	34.45
	0.89xRT	0.64xRT	1.10xRT
+Method 1	12.84	6.73	35.35
	0.73xRT	0.64xRT	0.93xRT
+Method 2	12.84	6.73	35.35
	0.64xRT	0.63xRT	0.88xRT
+Method 3	13.11(TF=0.7)	6.90(TF=0.4)	36.43(TF=0.5)
	0.56xRT	0.59xRT	0.73xRT

Table 4. *Performance of proposed improvements to CIGMMS on different tasks.(xRT stands for "times real-time".)*

7. Conclusion

In this work, we analyzed in detail the behavior of CIGMMS for improving GMM computation performance. One of our main observations is that it is necessary to preserve the integrity of the high-scoring GMMs in each frame, and avoid approximating them as far as possible. With this in mind, we proposed several variants of CIGMMS, by combining it with absolute pruning, the previous best Gaussian index scheme, and with frame down-sampling, respectively. We found that it is possible to improve over the basic CIGMMS by incorporating each one of these, cumulatively. The improvements are consistent across multiple corpora.

8. Acknowledgment

9. References

[1] Chan, A. et al, "Four-Layer Categorization of Fast GMM Computation Techniques in Large Vocabulary Continuous Speech Recognition Systems", *ICSLP-2004*.

[2] Ravishankar, M. et al, "The 1999 CMU 10x real time broadcast news transcription system", *Proc. DARPA workshop on Automatic Transcription of Broadcast News 2000.*

[3] Hwang, M., "Sub-phonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition", *Ph.D. Thesis, Computer Science, Carnegie Mellon University,* Dec 1993.

[4] Lee, A., "Gaussian Mixture Selection using Context independent HMM", IEEE ICASSP 2001.

[5] Pellom, B. et al, "Fast Likelihood Computation in Nearest-Neighbor Based Search for Continuous Speech Recognition" *IEEE Signal Processing Letters*, vol. 8, no. 8, pp. 221-224, August, 2001

[6] Gales, M. J. F. et al, "Use of Gaussian Selection in Large Vocabulary Speech Recognition Using HMMs" *ICSLP 1996.*

[7] Bennett, C. and Rudnicky, A., "The Carnegie Mellon Communicator Corpus", *ICSLP 2002.*

[8] Douglas, P. and Baker, J. M., "The Design for the Wall Street Journal-based CSR corpus", *Workshop of Speech and Natural Language 92.*

[9] Janin, A. et al. "The ICSI Meeting Project: Resources and Research", *NIST ICASSP 2004 Meeting Recognition Workshop.*

[10] Mosur, R. et al, "Sub-Vector Clustering to Improve Memory and Speed Performance of Acoustic Likelihood Computation**.",** *Eurospeech 1997.*