Selectively Computing Dynamic Features in the Likelihood Computation of ASR Systems

Xiao Li, Jeff Bilmes
{lixiao, bilmes}@ee.washington.edu

Dept of EE, University of Washington
Seattle WA, 98195-2500

UWEE Technical Report
Number UWEETR-2003-0009
May 2003

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Selectively Computing Dynamic Features in the Likelihood Computation of ASR Systems

Xiao Li, Jeff Bilmes
{lixiao, bilmes}@ee.washington.edu

Dept of EE, University of Washington
Seattle WA, 98195-2500

University of Washington, Dept. of EE, UWEETR-2003-0009
May 2003

Abstract

This paper proposes a novel technique to reduce the likelihood computation in ASR systems that use continuous density HMMs. Based on the nature of dynamic features and the numerical properties of Gaussian mixture distributions, we approximate the observation likelihood computation to achieve a speedup. Although the technique does not show appreciable benefit in an isolated word task, it yields significant improvements in continuous speech recognition. For example, 50% of the computation can be saved on the TIMIT database with only a negligible degradation in system performanc

1 Introduction

As speech technology becomes widely used in mobile devices, decoding speed and power consumption are becoming new metrics of ASR performance. Many fast decoding techniques for Viterbi search have been developed for this purpose, such as beam search [5] and tree-structured lexicons [6].

However, for speech recognition tasks using continuous HMMs, the computation of the state likelihood remains one of the most expensive parts. Since the observation distribution is typically represented by a Gaussian mixture, the computational complexity of the likelihood evaluation is proportional to the total number of Gaussians in the system and the dimensionality of these Gaussians. This gives two potential research directions to improve efficiency: reducing the number of Gaussians and reducing their dimensionality. There has been much of research in both areas. Bocchieri [2] developed the idea of Gaussian selection by vector quantization, where only a fraction of Gaussians with means close to the observation are precisely computed for each state, and all remaining ones are assigned approximated values. As for the second direction, various feature selection methods [8, 3, 9] tackle the dimensionality problem by selecting only the features with the highest discriminative power.

Dynamic features [10, 4, 7, 11] in conjunction with static features have long been used in most modern ASR systems. They are generally a linear combination of static-feature frames over a fixed time span. [1] gives a theoretical explanation why dynamic features can help in speech recognition in spite of the fact they are generated merely from static features. However, such a concatenated feature vector inevitably contains redundant information and may introduce a certain degree of correlation between the likelihoods of the static and dynamic features. Furthermore, dynamic features improve the recognition rate at the cost of doubling or even tripling the computational effort in likelihood evaluation. Therefore, it should at least be questioned whether the improvement in performance is worth the increase in computational cost.

In this paper, we propose a novel and simple technique to reduce the cost that dynamic features have brought to likelihood evaluation. In section 2, we observe the correlation between the log probability of the static-feature part of an observation vector and that of the dynamic-feature part, and analyze the potential likelihood computation approximation. We describe our method in section 3. By selectively computing the dynamic features, we can reduce the computation substantially while maintaining the baseline performance for continuous speech task, as is shown in section 4.

2 Motivation

In continuous density HMMs, the observation distribution for a certain state is typically parameterized by a Gaussian mixture. We let N denote the feature dimension, M the number of components in the mixture, and we assume diagonal covariance matrices. The observation probability of a state q is given by

$$\begin{aligned} P_q(x) &= \sum_{i=1}^{M} w_i \mathcal{N}(x|\mu_i, \Sigma_i) \\ &= \sum_{i=1}^{M} w_i \exp\{-\frac{a_{0i}}{2}\} \exp\{-\sum_{j=1}^{N} \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\} \end{aligned} \tag{1}$$

where a_{0i} is a constant independent of the observation. Without loss of generality, we analyze the case with only static features and their deltas. Equation 1 can be further decomposed as follows,

$$\begin{aligned} P_q(x) &= \sum_{i=1}^{M} w_i exp\{-\frac{1}{2}(a_{0i} + a_{1i}(x) + a_{2i}(x)\} \\ &= \sum_{i=1}^{M} w_i exp\{-\frac{1}{2} A_i(x)\} \end{aligned} \tag{2}$$

where $a_{1i}(x) = \sum_{j=1}^{L} \frac{(x_j - \mu_{ij})^2}{\sigma_{ij}^2}$ is the square Mahalanobis distance of the static-feature part, between the observation and the mean of the i^{th} Gaussian component; similarly, $a_{2i}(x) = \sum_{j=L+1}^{2L} \frac{(x_j - \mu_{ij})^2}{\sigma_{ij}^2}$ is that of the delta part and L is the dimensionality of static features only. $A_i(x) = a_{0i} + a_{1i}(x) + a_{2i}(x)$ governs the observation probability in the i^{th} Gaussian component. The greater $A_i(x)$ is, the more unlikely the observation in that component. Furthermore, since such a component probability is exponential in $-A_i(x)$, the component with *relatively* large $A_i(x)$ is negligible in the likelihood computation, and hence could potentially be approximated or even removed. Based on these facts, we will argue that the computation of $a_{2i}(x)$ is both theoretically and numerically approximable when $a_{1i}(x)$ is high.

Let's first consider the joint distribution of $a_{2i}(x)$ and $a_{1i}(x)$, $\forall i$, for a certain phoneme. Assuming $a_{1i}(x)$ and $a_{2i}(x)$ were independent random variables, if x itself was a Gaussian, $a_{1i}(x)$ and $a_{2i}(x)$ would each have a χ-square distribution. As a simple demonstration, we draw from two identical and independent χ-square random variables, as depicted in the upper-left plot (Pattern A) of Figure 1, where we are unable to obtain any information about $a_{2i}(x)$ from $a_{1i}(x)$. On the other hand, if $a_{1i}(x)$ and $a_{2i}(x)$ are fully correlated as in the upper-right plot (Pattern B), we can estimate $a_{2i}(x)$ completely from $a_{1i}(x)$, and the dynamic features are absolutely redundant in this sense.

Static and dynamic features are generated from the same source, so there is likely to exist a certain degree of correlation between them. Consequently, $a_{1i}(x)$ and $a_{2i}(x)$ are unlikely to be independent. Their real joint distribution, therefore, lies somewhere between the two above extreme cases. The bottom plots of Figure 1 show the true $a_{2i}(x)$ against $a_{1i}(x)$, $\forall i$, with observation x chosen randomly from various utterances from the NYNEX Phonebook corpus. The bottom-left figure plots the data distribution for all states of a vowel and the bottom-right one for a consonant. As shown in the figure, the conditional variance of $a_{2i}(x)$ tends to decrease as $a_{1i}(x)$ increases. In other words, for a large $a_{1i}(x)$, $a_{2i}(x)$ has a relatively narrow dynamic range, which allows us to more accurately approximate $a_{2i}(x)$ based on the value of $a_{1i}(x)$. The plot in the other case follows the same trend, although the dynamic range of $a_{2i}(x)$ shrinks to a different extent.

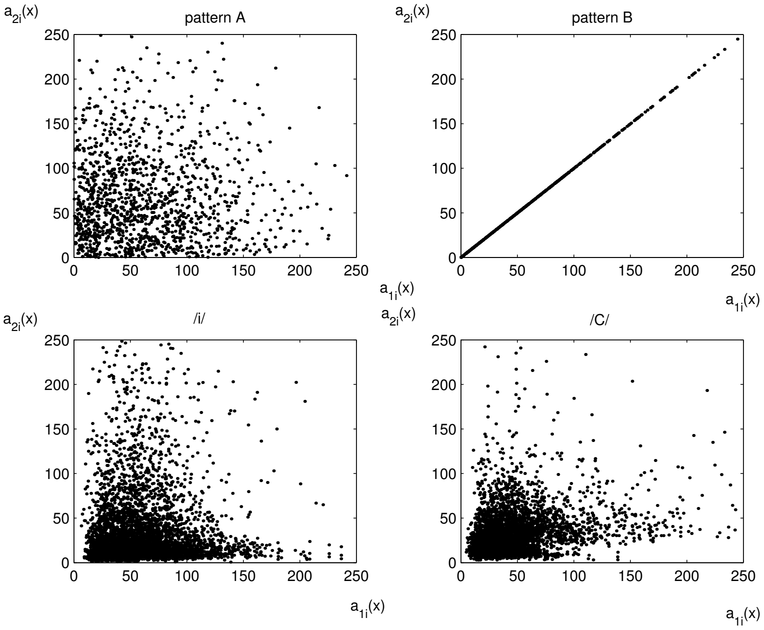

Figure 1: $a_1(x)$ against $a2(x)$. Upper: two extreme distribution patterns; Bottom: two real distributions for vowel and consonant phonemes respectively

Moreover, from purely a numerical perspective, if $a_{1i}(x)$ is small, $a_{2i}(x)$ becomes a dominate factor in determining the probability $\mathcal{N}(x|\mu_i, \Sigma_i)$. Any approximation to $a_{2i}(x)$ might incur an error in likelihood evaluation. However, when $a_{1i}(x)$ is high, the i^{th} component has already become negligible in the overall likelihood computation of a mixture model. Therefore we don't need to further compute $a_{2i}(x)$ precisely.

The above analysis also works for the static features with both 1^{st} and 2^{nd} order dynamic features. In that case, by adding the 2^{nd} delta, $A_i(x)$ can be written as

$$A_i(x) = a_{0i} + a_{1i}(x) + a_{2i}(x) + a_{3i}(x) \tag{3}$$

With both deltas and 2^{nd} deltas, the issue can become even more inefficient since computation can triple relative to just the static features.

3 Proposed Technique

Our goal is to approximate $A(x)$ to reduce computational cost. For simplicity, we consider only a certain component and drop the index i from our notation. We let $e_k(x)$, $k = 0, 1, 2$, denote the intermediate sums in the accumulation process of the $A(x)$ computation, with $e_k(x) = \sum_{l=0}^{k} a_l(x)$, $k = 1, 2$ and $e_0(x) = a_0$. Also we define (below) a set of approximation functions $f_k(t)$, $k = 0, 1, 2$ which we assume to be linear.

We let $\hat{A}_i(x)$ denote the approximated $A_i(x)$. For the case of static and delta features only:

$$\hat{A}(x) = \begin{cases} f_0(e_0(x)) & \text{if } e_0(x) > t_0; \\ f_1(e_1(x)) & \text{if } e_0(x) \leq t_0 \text{ and } e_1(x) > t_1; \\ A(x) & \text{otherwise.} \end{cases} \tag{4}$$

In this approach, we precisely compute the Mahalanobis distance of the delta part only when that of the static-feature part is low enough (presumably when the delta-feature part is less predictable). Otherwise, we approximate $A_i(x)$ as a linear function of $e_1(x)$, a quantity already been computed. A similar definition holds in the case of static features + deltas + 2^{nd} deltas:

$$\hat{A}(x) = \begin{cases} f_0(e_0(x)) & \text{if } e_0(x) > t_0; \\ f_1(e_1(x)) & \text{if } e_0(x) \leq t_0 \text{ and } e_1(x) > t_1; \\ f_2(e_2(x)) & \text{if } e_1(x) \leq t_1 \text{ and } e_2(x) > t_2; \\ A(x) & \text{otherwise.} \end{cases} \tag{5}$$

To produce the linear functions, we applied a simple Least Mean Square estimation method on the speech training data $\{x$: $e_k(x) > t_k\}$, and obtain the optimal coefficients for $f_k(t) = r_k t + s_k$, $k = 0, 1, 2$. Note we assign the same set of approximation functions to all phoneme models, though they would have different linear functions if trained separately. Experiments demonstrated that in fact $r_k \approx 1$. We therefore just let $r_k = 1$ for simplicity. It is consistent with the bottom plots in Figure 1, where $a_2(x)$ becomes approximately a constant when $a_1(x)$ is high. The exact algorithm for the case of static features + deltas + 2^{nd} delta can be described as follows:

Input: observation x; $\{\mu, \Sigma\ \}$ of a Gaussian

Output: observation probability $A(x)$ in that Gaussian

```
1:  $sum \leftarrow 0$
2:  compute $a_0$; $sum \leftarrow sum + a_0$;
3:  **if** $sum > t_0$ **then**
4:      $sum \leftarrow sum + s_0$; goto 15;
5:  **end if**
6:  compute $a_1(x)$; $sum \leftarrow sum + a_1(x)$;
7:  **if** $sum > t_1$ **then**
8:      $sum \leftarrow sum + s_1$; goto 15;
9:  **end if**
10: compute $a_2(x)$; $sum \leftarrow sum + a_2(x)$;
11: **if** $sum > t_2$ **then**
12:     $sum \leftarrow sum + s_2$; goto 15;
13: **end if**
14: compute $a_3(x)$; $sum \leftarrow sum + a_3(x)$;
15: **return** $sum$**;**
```

In this way, only a fraction of observation vectors are precisely computed to full dimension. Some of the dynamic features are ignored according to the likelihoods of their static-feature. If we set the thresholds t_k, $k = 0, 1, 2$ very low, substantial computational savings can result, but at the cost of more noise and consequently performance degradation. In the following section, we will experimentally demonstrate this tradeoff and show the degree to which dynamic features can be ignored without losing performance.

4 Experiments and Results

We ran two sets of experiments, one on an isolated word corpus and the other on a continuous speech recognition task. We used 26-dimensional features on the first task and 39-dimensions for the second. Although our proposed technique does not show significant benefit in the former case, it works remarkably well on the latter.

The isolated word experiment was done on NYNEX Phonebook, a telephone-speech database. We use 42 continuous HMMs as acoustic model with 165 states altogether. Each state is represented by a Gaussian mixture comprised of 12 Gaussian components. We extract a 26-dimensional feature, MFCC + delta with mean subtraction and variance normalization.

The test was carried out on 8 different test sets with 75 words each. And the final WER is an average over them all. The baseline WER is 2.07%. We applied the approximation algorithm in the previous section to the likelihood computation. Here we made $t_1 = t_0 = t'$ and $s_1 = s_0 = s'$ and obtained WER for different t''s. Note that for a reasonably high t_0, it is rarely the case that the constant a_0 can be high enough so that $e_0(x) > t_0$. Therefore, the static-feature part is always computed even though t_0 is set as a threshold. We counted the delta chunks ignored according to different t's. Since the overhead introduced is only trivial, the percentage computation saved just from the observation likelihood evaluation is approximately:

$$\eta = \frac{\text{\# of delta chunks ignored}}{\text{\# of MFCC chunks computed} \times 2} \times 100\%$$

Figure 2 shows the plot of WER against computation saved. The saving is obviously not significant. If we would like to keep baseline performance, only 3.5% computational savings can be achieved. And if we allow an acceptable loss of performance by $+0.5\%$ absolute increase in WER, 8% saving is the best we can get. We also tested on 600-word set and found a similar trade-off between WER and η.

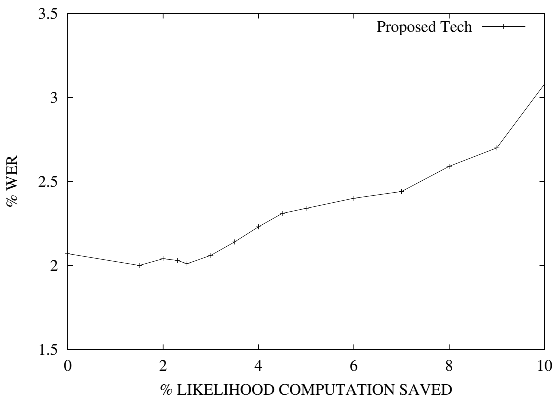

Figure 2: WER vs. η for 75-word PBK testing set

The result of our second experiment, however, is remarkably good. It is based on TIMIT, a corpus designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for speech recognition evaluation. We use simple bi-gram language model and 42 uni-phone HMMs for the acoustic model. We assign 3 emitting states to each HMMs which amounts to 124 states altogether. Again each state distribution is represented by a Gaussian mixture with 12 components. The features we use are MFCC + delta + 2^{nd} delta, so 39 dimensions in total. The baseline computes each observation vector to its full dimension for all $124 \times 12 = 1488$ Gaussian components with an accuracy 88.07%. We applied the same technique, calculating η as follows:

$$\eta = \frac{\text{\# of delta or } 2^{nd} \text{ delta chunks ignored}}{\text{\# of MFCC chunks computed} \times 3} \times 100\%$$

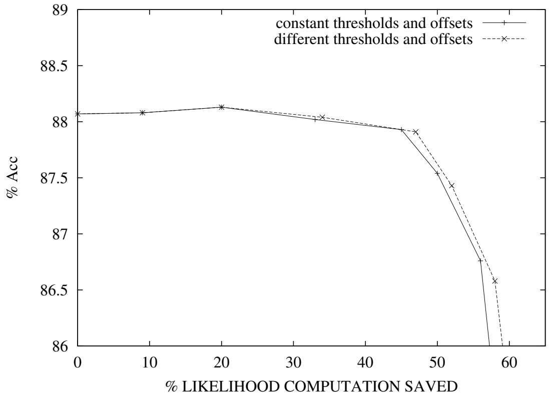

Figure 3: Acc vs. η for TIMIT core testing set

As in the first experiment, we used a constant threshold $t_k = t\prime$ and offset $s_k = s\prime$, $k = 0, 1, 2$. Also we tried a set of thresholds $\{t_k\}$ and $\{s_k\}$ varying with k, which obeys $t_0 < t_1 < t_2$ and $s_0 > s_1 > s_2$. The specific values of the thresholds and offsets were chosen using heuristics. As shown in Figure 3, both of them work well. We can halve the likelihood computation with only a slight degradation in system performance, results which are hugely significant. The varying-threshold experiment worked slightly better, but not significantly.

5 Discussion and Conclusion

In comparison with conventional fast decoding techniques, our approach focuses on achieving computational saving by partially computing the observation probability in a Gaussian component. It ignores computing the dynamic-feature part of an observation vector when its static-feature part already falls in the tail of a Gaussian.

This technique doesn't require a complicated training procedure and brings almost no overhead to the decoding process. It is effective on both isolated word and connected word speech tasks, but works especially well on connected word recognition with high-dimensional dynamic features. This is probably because on the isolated-word task, the recognition becomes a “hard decision” without a language model available. Any slight approximation in acoustic model could affect the results. Another reason is the 26-dimensional feature we used is relatively compact compared to the 39-dimensional feature, and therefore benefits less from our technique. Also, taking the database conditions into consideration, speech in Phonebook has lower quality than that in TIMIT, which makes the system more sensitive to approximations.

Lastly, it is worth noting that the phonemes do not share the same $a_1(x), a_2(x)$ joint distribution. Using different thresholds and offsets for each phoneme (or HMM states) might improve our technique.

This work was funded by NSF grant ITR/RC-0086032.

References

[1] J. A. Bilmes. Graphical models and automatic speech recognition. In R. Rosenfeld, M. Ostendorf, S. Khudanpur, and M. Johnson, editors, *Mathematical Foundations of Speech and Language Processing*. Springer-Verlag, New York, 2003.

[2] E. Bocchieri. Vector quantization for the efficient computation of continuous density likelihoods. In *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, volume 2, pages 692–695, 1993.

[3] E. Bocchieri and J. Wilpon. Discriminative feature selection for speech recognition. *Computer, Speech and Language*, pages 229–246, 1993.

[4] K. Elenius and M. Blomberg. Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system. In *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, pages 535–538, 1982.

[5] F. Jelinek. *Statistical methods for speech recognition.* MITPress, Cambridge, Massachusetts, 1998.

[6] J.Suontausta, J.Hakkinen, and O.Viikki. Fast decoding in large vocabulary name dialing. In *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, volume 3, pages 1535–1538, 2000.

[7] C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini, and A.E. Rosenberg. Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition. *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, 1991.

[8] E. Lleida and C. Nadeu. Principal and discriminant component analysis for feature selection in isolated word recognition. In *Singal Processing V: Theories and Applications*, pages 1251–1254. Elsevier Sc. Publ. B.V., 1990.

[9] J. Nouza. On the speech feature selection problem: Are dynamice features more important than the static ones? In *Eurospeech*, pages 919–922, Sep 1995.

[10] S.Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34:52–59, 1986.

[11] J.G. Wilpon, C.-H. Lee, and L.R. Rabiner. Improvements in connected digit recognition using higher order spectral and energy features. *Proc. Intl. Conf. on Acoustic, Speech and Signal Processing*, 1991.