Fast Likelihood Computation Techniques in Nearest-Neighbor Based Search for Continuous Speech Recognition

Bryan L. Pellom, Ruhi Sarikaya, and John H. L. Hansen

指導教授：陳嘉平

報告者：陳柏含

Introduction

- One of the computationally most expensive steps in speech recognition based on CDHMM is the state likelihood computation.

- Typically, there computations take up a major proportion (30%-70%) of the overall recognition time. This is due to multiple number of Gaussian mixtures used to model each state (4-64).

State Likelihood Computation

- The likelihood of an HMM state λ_s for a given feature vector x_t can be expressed as a weighted sum of likelihoods from individual Gaussian densities (with diagonal covariance matrices):

$$p(x_t|\lambda_s)=\sum_{m=1}^{M_s}\frac{w_m}{(2\pi)^{J/2}\left(\prod_{j=1}^{J}\sigma_m^2(j)\right)^{1/2}}\times\exp\left(-\frac{1}{2}\sum_{j=1}^{J}\frac{(x_t(j)-\mu_m(j))^2}{\sigma_m^2(j)}\right)$$

$$=\sum_{m=1}^{M_s}C_m\exp\left(-\frac{1}{2}\sum_{j=1}^{J}\frac{(x_t(j)-\mu_m(j))^2}{\sigma_m^2(j)}\right)$$

C_m: constant for each density

M_s: the number of Gaussian mixture

w_m: mixture weight for m-th density in state λ_s

μ_m: mean for m-th density in state λ_s

σ_m: variance for m-th density in state λ_s

Nearest-Neighbor Approximation

- Computation of $p(x_t / \lambda_s)$ is expensive due to the J multiplications, J divisions, and M_s exponential operations.

- In the log-domain, using nearest-neighbor approximation

$$\log(p(x_t | \lambda_s)) \approx \max_{1 \le m \le M} \left\{ \log(C_m) - \frac{1}{2} \sum_{j=1}^{J} \frac{(x_t(j) - \mu_m(j))^2}{\sigma_m^2(j)} \right\}$$

Partial Distance Elimination (PDE)

- Denote the likelihood for mixture m given x_t as $D(x_t/y_m)$:

$$D(x_t|y_m) = C'_m - \sum_{j=1}^{J} (x_t(j) - \mu_m(j))^2 \frac{1}{2\sigma_m^2(j)}$$

- Note that the weighted (with variance) squared error is separable measure, and $D(x_t/y_m)$ can be evaluated component-wise.

Partial Distance Elimination (PDE)

- Partial Distance Elimination:
  1. computing the likelihood of the first mixture over all J components to get the initial $D_{\max}$

  Elimination :

  Before finishing the computation of a complete likelihood, for any $j < J$, if the negative accumulated weighted squared error for the first j components of the input vector plus C'_m is smaller than highest $\hat{D}_{\max}$ yet in the search, the likelihood of this mixture is not possible to be the final maximum value.

Best Mixture Prediction (BMP)

- The efficiency of the PDE technique heavily depends on how quickly a high estimate of $D_{\max}$ is obtained.

$$m^{t-1} = \underset{1 \le i \le M_s}{\operatorname{argmax}} D\left(x_{t-1} \middle| y_m \right)$$

- Choosing the previous best match, Gaussian as the current best match and computing its first result in a high D_{max} speeds up the elimination process.

- Because of overlapping frames during feature extraction x_t and x_{t-1} is usually similar, we expect $D(x_{t-1}/y_m{}^{t-1}) = D^{t-1}{}_{max}$ to be close to $D(x_t/y_m{}^t)$

Feature Component Reordering (FCR)

- By analyzing the components of the feature vectors and the densities, the contribution of some of the components are heavier than others.

- Reorder: Let $j \rightarrow o[k]$ be a mapping of the location of component j in the vectors into a new location, $o[k]$

$$D(x_t|y_m) = C_m^{'} - \sum_{j=1}^{J} \left(x_t(o[j]) - \mu_m(o[j])\right)^2 \frac{1}{2\sigma_m^2(o[j])}$$

- The mapping function can be learned from a portion of the development test set offline.

Analysis of feature component reordering

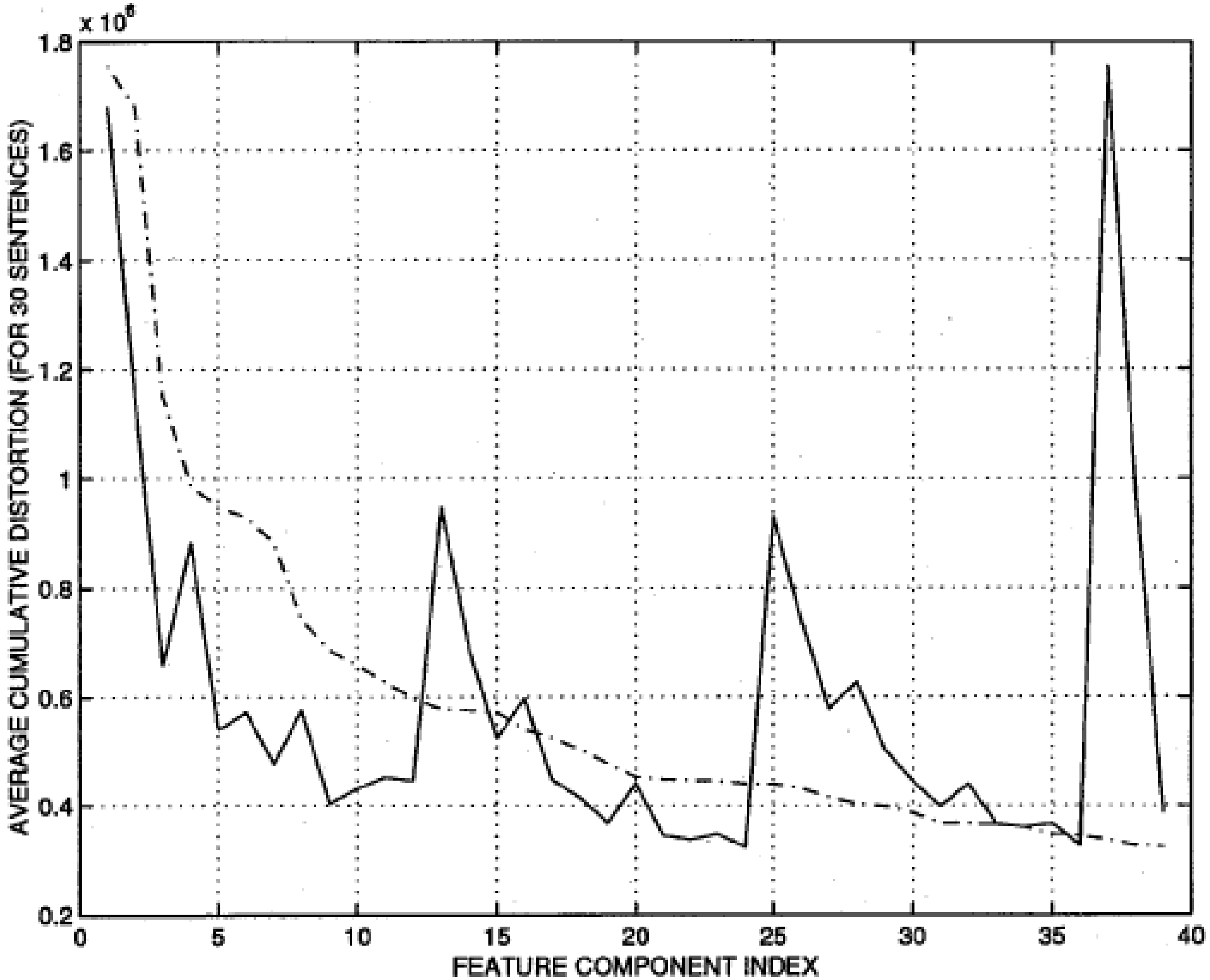

Fig. 1. The solid curve is the average distortion before reordering, and the dashed curve is the distortion of the reordered elements in descending order: $\{37, 1, 2, 38, 13, 25, 4, 26, 14, 3, 28, 16, 27, 8, 6, 5, 15, 29, 7, 11, 17, 12, 30, 20, 32, 10, 18, 9, 31, 39, 35, 19, 33, 34, 23, 21, 22, 36, 24\}$.

Recognition System

Corpus: 1992 ARPA WSJ 5k vocabulary continuous speech recognition.

Training set: SI-284 WSJ training set.

System: cross-word gender-dependent system.

Acoustic model:

- triphone acoustic model
- 3 state left-to-right topology per HMM
- 6-16 Gaussian mixtures per state.

Language model: trigram language model.

- Word error rate for the baseline as well as proposed techniques is 11.8%.

TABLE I

COMPARISON OF THE EXPERIMENTAL EVALUATION OF THE SPEEDS OF NEAREST-NEIGHBOR BASEDBASELINE AND PROPOSED TECHNIQUES ON NOVEMBER 1992 DARPA WSJ EVALUATION USING WSJ0-DEV SET

Likelihood Computation Time (%)			
Baseline	PDE	PDE + FCR	PDE + FCR + BMP
100.0	96.0	74.0	70.2
Average # of multiplications for likelihood computation per sentence (million)			
687.75	405.52	324.36	293.09