

Gaussian Mixture Modeling with the EMLLT Model with Applications to Speech Recognition

Ramesh Gopinath

`rameshg@us.ibm.com`

IBM T. J. Watson Research Center

Yorktown Heights, NY 10598

<http://www.research.ibm.com/people/r/rameshg>

joint work with Peder Olsen and Scott Axelrod

Outline

- Statistical Speech Recognition
 - Acoustic Modeling
- Gaussian Mixture Models
 - MLLT model
 - EMLLT model
- Experimental results on two speech recognition tasks.

Speech Recognition - How does it work?

- Training: Learn the parameters θ of a statistical model that maps the speech waveform, \mathbf{a} , to a word sequence, \mathbf{w} .

$$p_{\theta}(\mathbf{w}|\mathbf{a})$$

- Recognition: Given a waveform \mathbf{a} hypothesise the most likely word sequence \mathbf{w} .

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} p_{\theta}(\mathbf{w}|\mathbf{a}).$$

Speech Recognition Building Blocks

- Signal Processing:

$$\mathbf{a} \mapsto \mathbf{x} = (x_1, x_2, \dots, x_T), x_i \in \mathbb{R}^d.$$

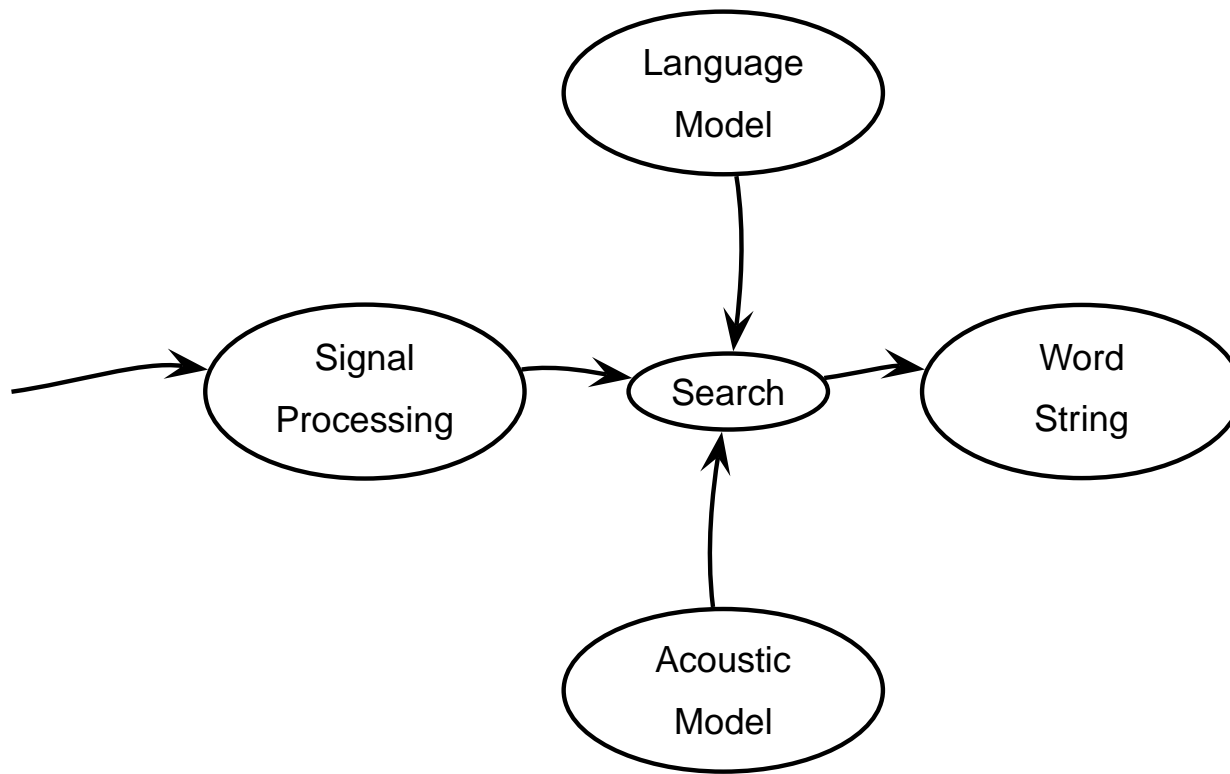
- Language Model: $p_{\theta_L}(\mathbf{w})$

- Acoustic Model: $p_{\theta_A}(\mathbf{x}|\mathbf{w})$

- Search:

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} p_{\theta_L}(\mathbf{w}) p_{\theta_A}(\mathbf{x}|\mathbf{w})$$

How a Speech Recognizer Works



Acoustic Modeling - Hidden Markov Model

- Inherent variation in the **duration** of speech
- Inherent variation in the **pronunciation** of words
- Inventory of primitive sounds or states \mathcal{S}
- Word sequences map to state sequences
- Each state in the sequence can produce a variable number of feature vectors

$$p(\mathbf{x}|\mathbf{w}) = \sum_{\mathbf{s}} \underbrace{p(x_1|s_1) \dots p(x_T|s_T)}_{p(\mathbf{x}|\mathbf{s})} p(\mathbf{s}|\mathbf{w}).$$

Acoustic Modeling - Example

Given $\mathbf{x} = (x_1, \dots, x_{100})$ and hypothesis $\mathbf{w} = \text{'cat'}$ want
 $p(\mathbf{x}|\mathbf{w}) = \sum_s p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|\mathbf{w})$.

cat

CAT.1

K

AE

T

K_1

K_2

K_3

AE_1

AE_2

AE_3

T_1

T_2

T_3

K_1.2

K_2.3

K_3.1

AE_1.2

AE_2.1

AE_3.1

T_1.3

T_2.2

T_3.2

$x_{1..8}$

$x_{9..21}$

$x_{22..35}$

$x_{36..40}$

$x_{41..52}$

$x_{52..59}$

$x_{60..72}$

$x_{72..84}$

$x_{84..100}$

$$p(x_1 \dots x_{100} | s_1 \dots s_{100}) = p(x_1 | s_1) p(x_2 | s_2) \dots p(x_{100} | s_{100})$$

Acoustic Modeling

- Each distribution $p(x|s)$, $s \in \mathcal{S}$ is modeled as a Gaussian Mixture

$$p(x) = \sum_{i=1}^m \pi_i G(x; \mu_i, \Sigma_i)$$

where

$$G(x; \mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}{(2\pi)^{d/2} |\det(\Sigma)|^{1/2}}$$

and $\sum_{i=1}^m \pi_i = 1$.

Gaussian Mixture Model

- Parameters: Priors, Means and Covariances

$$x \sim \{\pi_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, \dots, m.$$

- Practical considerations (computational, storage etc. when m and d are large) lead to parameter constraints, e.g., Σ_i is diagonal, say, D_i .
- Constrained GMM may not be Invariant to Linear Transformations (ILT) of the data; in which case one hopes to better satisfy constraints by linearly transforming the data.

Maximum Likelihood Linear Transform (MLLT)

- MLLT: Diagonal Covariance GMM is not ILT. If $y = Ax, A \in \mathbb{R}^{d \times d}$ is modeled by a Diagonal Covariance GMM,

$$Ax = y \sim \{\pi_i, \mu_i, D_i\},$$

then, x is also a GMM with parameters of the form

$$A^{-1}y = x \sim \{\pi_i, A^{-1}\mu_i, A^{-1}D_iA^{-T}\}.$$

- ML Estimation of MLLT Model Parameters - Generalized EM.

MLLT - Another Viewpoint

- MLLT: $\Sigma_i = A^{-1} D_i A^{-T}$
- MLLT in terms of Precision Matrices:

$$P_i = \Sigma_i^{-1} = A^T \Lambda_i A = \sum_{k=1}^d \lambda_k^i \mathbf{a}_k \mathbf{a}_k^T.$$

- Precision matrices are restricted to be in a linear subspace
 - Basis: rank-one symmetric matrices $\{\mathbf{a}_k \mathbf{a}_k^T\}_{k=1}^d$,
 - Expansion Coefficients: $\{\lambda_k^i\}$

The EMLLT Model

- Extended ML Linear Transform (EMLLT)

$$P_i = A^T \Lambda_i A = \sum_{k=1}^D \lambda_k^i \mathbf{a}_k \mathbf{a}_k^T, \quad d \leq D \leq \frac{d(d+1)}{2}$$

- EMLLT is very flexible

- $D = d, \mathbf{a}_k = \mathbf{e}_k \Leftrightarrow$ Diagonal Covariance GMM

- $D = d \Leftrightarrow$ MLLT

- $D = d(d+1)/2 \Leftrightarrow$ Full Covariance (Precision)

- Generalized EM algorithm for ML estimates from training data of of the basis and expansion coefficients.

The EMLLT Model - What does it model?

- In a diagonal covariance GMM the ML variance estimate is equal to the the sample variance along the co-ordinate directions.

$$\mathbf{e}_k^T \bar{\Sigma}_i \mathbf{e}_k = \mathbf{e}_k^T D_i \mathbf{e}_k, 1 \leq k \leq d.$$

- In an EMLLT based GMM the ML variance estimate is equal to the sample variance along the directions \mathbf{a}_k .

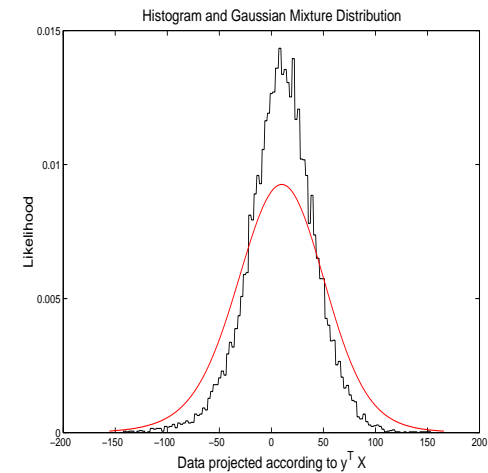
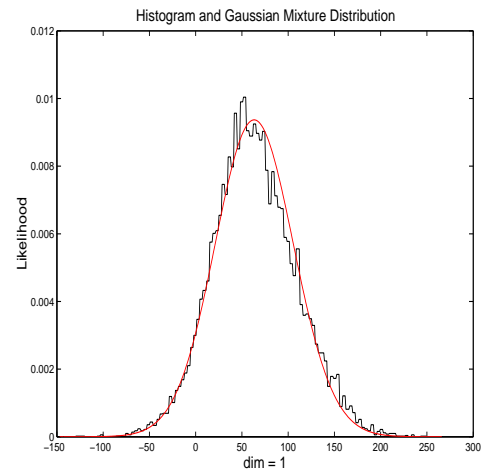
$$\mathbf{a}_k^T \bar{\Sigma}_i \mathbf{a}_k = \mathbf{a}_k^T P_i^{-1} \mathbf{a}_k = \mathbf{a}_k^T (A^T \Lambda_i A)^{-1} \mathbf{a}_k, 1 \leq k \leq D.$$

Histogram vs. Projections of GMMs

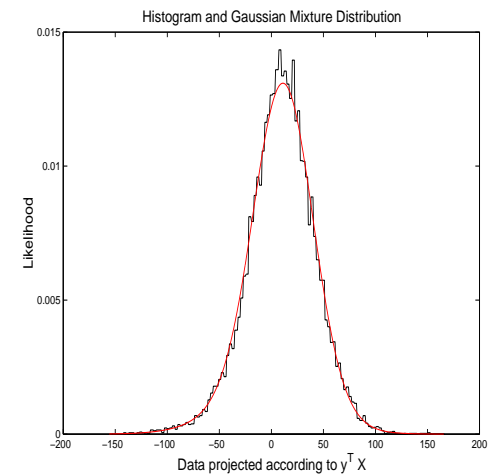
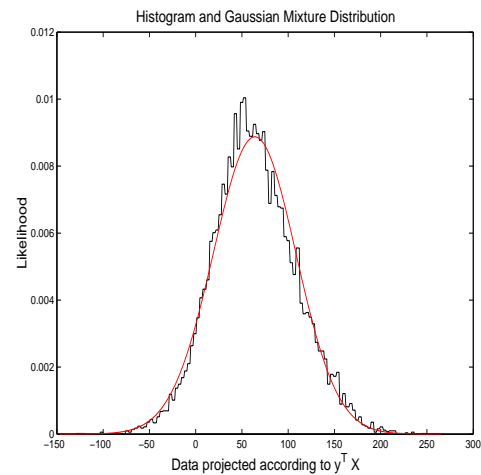
Diagonal

Coordinate

Random



EMLLT



EMLLT - ASR Results on IBM Internal Car Database

- Full Covariance: 2.11%, EMLLT ($D = 14d$): 2.04%

| Diagonal | | MLLT | | EMLLT | | |
|----------|-------|------|-------|-------|------|-------|
| m | WER | m | WER | m | D | WER |
| 10K | 3.14% | 10K | 2.84% | 10K | d | 2.84% |
| 17K | 3.08% | 17K | 2.74% | 10K | $2d$ | 2.54% |
| 26K | 3.01% | 26K | 2.58% | 10K | $4d$ | 2.34% |
| 46K | 2.84% | 46K | 2.50% | 10K | $8d$ | 2.10% |

EMLLT - ASR Results on IBM Internal Telephony Database

| D | # gaussians | WER |
|-------|-------------|-------|
| d | 45K | 3.56% |
| $4d$ | 45K | 2.75% |
| $8d$ | 45K | 2.67% |
| $14d$ | 45K | 2.49% |

Summary

- Covariance modeling for GMMs still largely unexplored field with significant research potential.
- Numerical optimization is central to the estimation of complex GMMs.
- EMLLT model is a step in the right direction.
- EMLLT expands precision matrices in a basis of rank-one symmetric matrices and hence is extremely flexible.