JHU CLSP Workshop 2009

Motivating Sub-Space Modeling for Automatic Speech Recognition

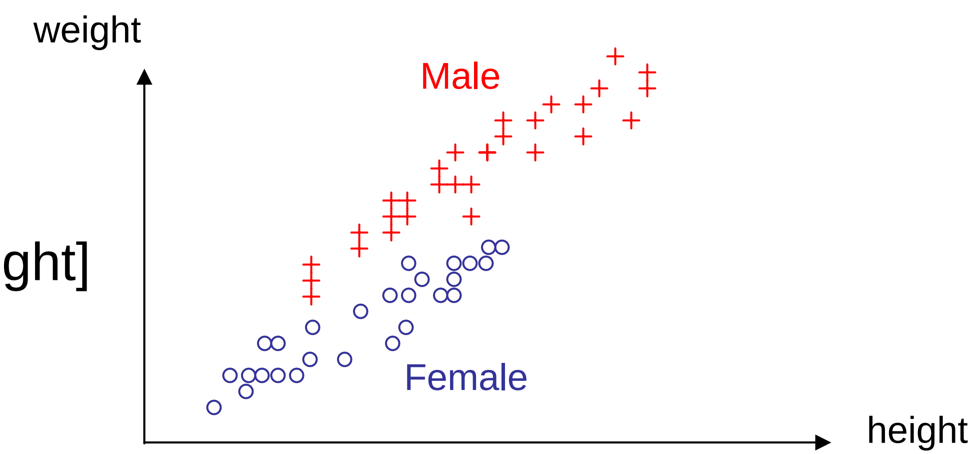
OUTLINE

- Introduce the Notion of Identifying Low Dimensional Subspaces over Model Parameters
- Subspace Based Speaker Adaptation
- Generalization to a Generalized Joint Subspace Model of Acoustic Variability in ASR

Identifying Low Dimensional Feature Space – Dimensionality Reduction

- Problems with high dimensional ***feature spaces***:
  - High computational complexity
  - Poor generalization to unseen data
  - "Curse of dimensionality"
- Starting with high dimensional ***data***, identify low dimensional feature space
  - Principal components analysis (PCA) – Capture maximum variance
  - Linear Discriminant Analysis (LDA) – Maximum class separability

Example:
2-D Feature Space: [height,weight]
2 Classes: male,female



Identifying Low Dimensional Feature Space

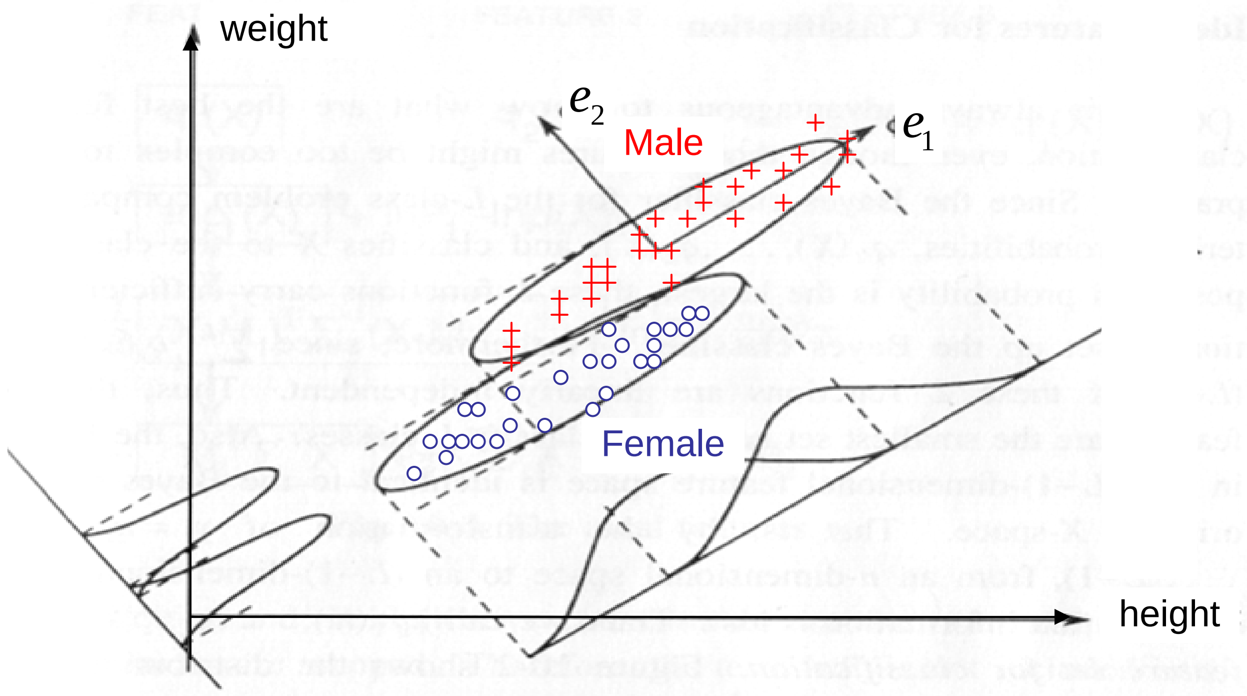
- Low dimensional feature, y, obtained from high dimensional feature, x, by a linear transformation:

Maximum Separability: $y = e_2^T x$

(Linear Discriminant Analysis)

Maximum Variance: $y = e_1^T x$

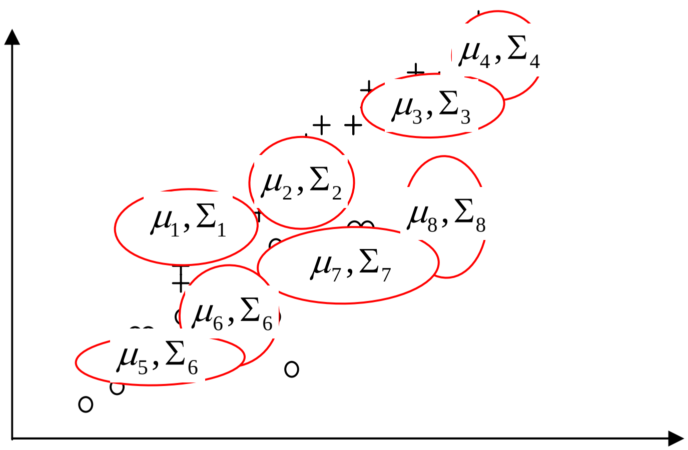
(Principal Components Analysis)



Identifying Low Dimensional Model Space

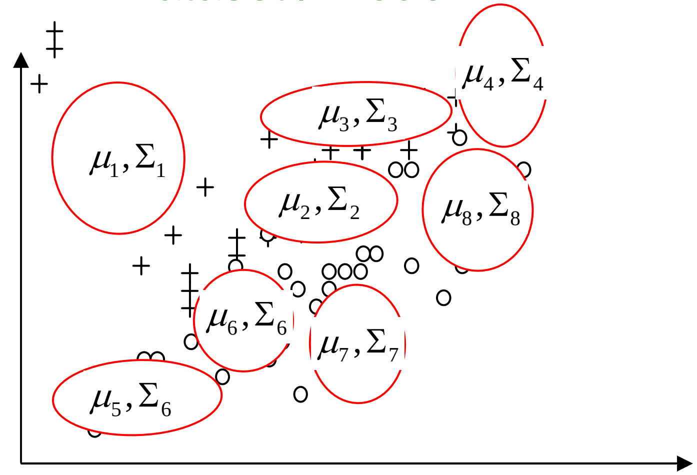
- Suppose there are multiple data sets describing similar populations and models are to fit each data set:

Dataset / Model 1



• • •

Dataset / Model N



- A low dimensional subspace can be identified that describes variation of the parameters

Form Super Vectors from Models:

$$\boldsymbol{\mu}^1 = \begin{bmatrix} \mu_1^1 \\ \vdots \\ \mu_8^1 \end{bmatrix}, \cdots, \boldsymbol{\mu}^N = \begin{bmatrix} \mu_1^N \\ \vdots \\ \mu_8^N \end{bmatrix}$$

Estimate Sub-space Projection:

$$\boldsymbol{\mu} = \mathbf{E}\boldsymbol{\nu}$$

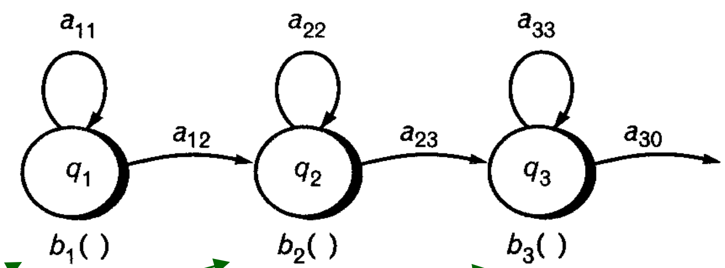
… or:

$$\boldsymbol{\mu} = \mathbf{m}_0 + \mathbf{E}\boldsymbol{\nu}$$

Speaker Space Adaptation – Super-vector

- Continuous Gaussian Mixture Observation Density HMMs

State dependent Gaussian Mixtures



$p(\vec{x} \mid s_j) = \sum_{m=1}^{C_j} w_{m,j} f_{m,j}(\vec{x})$, with $C = \sum_m C_m$ mixture components, and $f_{m,j}(\vec{x}) : N\left[\vec{x}; \vec{\mu}_{m,j}, \Sigma_{m,j}\right]$

- A speaker, *S*, is generally defined over a "super-vector" of the concatenated means of component Gaussians:

$$\vec{\boldsymbol{\mu}}^s = \left.\begin{bmatrix} \mu_1^s \\ \mu_2^s \\ \vdots \\ \mu_C^s \end{bmatrix}\right\} \text{Dimension: } M = CF$$

- Example: Wall Street Journal HMM
  - Component Gaussians $C \approx 100{,}000$
  - Feature Vector Dimension $F \approx 40$
  - Super Vector Dimension $CF \approx 4{,}000{,}000$
- Super-vector dimension can be very large

Speaker Space Based Adaptation

- Adapt Super-vector in Low Dimensional Subspace

- Training (Off-Line): Identify basis vectors of low dimensional speaker subspace from speaker dependent super-vectors:

$$\vec{\boldsymbol{\mu}}^1, \ldots, \vec{\boldsymbol{\mu}}^S \quad \longrightarrow \quad \mathbf{E} = \vec{\mathbf{e}}^1, \ldots, \vec{\mathbf{e}}^K$$

where $\vec{\boldsymbol{\mu}}^s$ is dimension, *M*, and $\mathbf{E}$ is dimension $M \times K$ where $K << M$

- Adaptation: Estimate weights $w_k^s, k = 1, \ldots, K$ from adaptation data to obtain adapted super-vector:

$$\hat{\vec{\boldsymbol{\mu}}}^s = \vec{\boldsymbol{\mu}}^{SI} + \sum_{k=1}^{K} w_k^s \vec{e}(k)$$

- Requires only a few seconds of adaptation data
- Speaker subspace dimension $K \approx 10 \rightarrow 100$

Electrical & Computer Engineering

Subspace Identification – Training

- Principal Components Analysis (EigenVoices)
  - Starting from M dimensional super-vectors for each of S speakers to a K dimensional subspace

$$\left[\begin{array}{c|c|c} \vec{\mu}_1^1 & \cdots & \vec{\mu}_1^S \\ \vdots & \ddots & \vdots \\ \vec{\mu}_C^1 & \cdots & \vec{\mu}_C^S \end{array}\right] \xrightarrow[K<<CF]{\text{PCA}} \left[\begin{array}{c|c|c} \vec{e}(1,1) & \cdots & \vec{e}(1,K) \\ \vdots & \ddots & \vdots \\ \vec{e}(M,1) & \cdots & \vec{e}(M,K) \end{array}\right]$$

Speaker Super-vectors for Speaker 1 through Speaker S

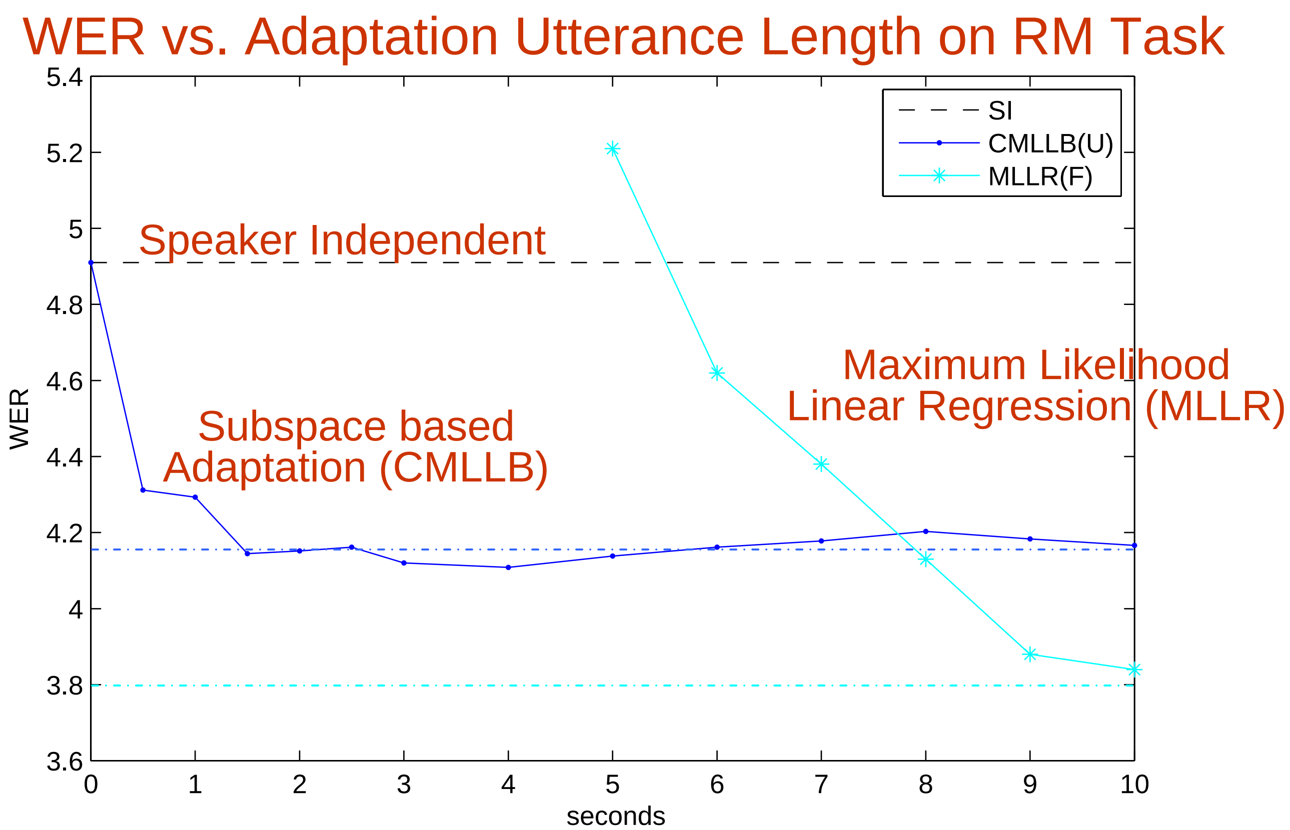
Speaker Subspace basis vectors

- Maximum Likelihood Clustering (Cluster Adapt. Training)
  - Given SD training observation vectors, $X^s = x_1^s, \ldots, x_T^s$, SI HMM model, λ^{SI}, and initial estimate of $\mathbf{E}^{(0)} = \vec{\mathbf{e}}^1, \ldots, \vec{\mathbf{e}}^K$,
  - Use EM algorithm to iteratively estimate weights and basis vectors

$$\hat{\Lambda}: \quad \hat{\vec{\boldsymbol{\mu}}}^s = \vec{\boldsymbol{\mu}}^{SI} + \sum_{k=1}^{K} w_k^s \vec{e}(k)$$

Electrical & Computer Engineering

Limited Effect of "Global" Subspace Adaptation



[From Tang and Rose, 2008]

- Substantial improvement with only 1 or 2 seconds of adaptation data
- Does not exhibit desirable asymptotic behavior

Generalization to Subspace Models of Phonetic Variability

- Speaker space model is limited in the form of the variability it can represent
  - Single vector in speaker space describes speaker specific variability

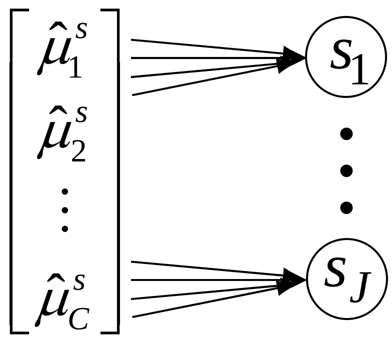
- Generalize this model in three ways:
  1. Define ***multiple model subspaces*** over different regions of the feature space
  2. Define ***state-specific***,(rather than speaker specific) ***weight vectors*** to describe phonetic variation within these subspaces
  3. Define ***joint model / speaker subspaces***

- This is conceptually a straightforward generalization of the speaker subspace approach

Generalization of Sub-Space HMM

- ***Review:*** Subspace Based Speaker Adaptation
  - Single global subspace, $\mathbf{N}$, defined over all Gaussians in HMM
  - Single weight vector, $v^{(s)}$, describes variation in subspace

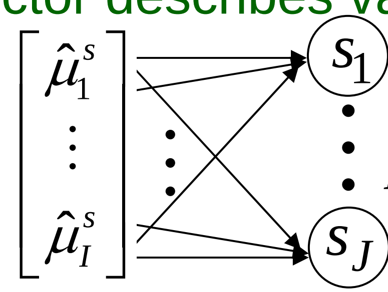
$$\hat{\mu}_{j,m}^{(s)} = \mu_{j,m}^{SI} + \mathbf{N}v^{(s)}$$



$$p(x \mid s_j) = \sum_{m=1}^{M} w_{j,m} p(x; \hat{\mu}_{j,m}^{(s)}, \Sigma_{j,m})$$

- ***Generalization***: Multiple Region-Specific Subspaces
  - Separate Subspaces, $\mathbf{N}_i$, defined for each Gaussian in a GMM
  - Single weight vector describes variation in subspaces

$$\hat{\mu}_i^{(s)} = \mu_i^{UBM} + \mathbf{N}_i v^{(s)}$$



$$p(x \mid s_j) = \sum_{i=1}^{I} w_{j,i} p(x; \hat{\mu}_i^{(s)}, \Sigma_i)$$

Electrical & Computer Engineering

Generalization to Joint Subspace HMM

- ***Generalization***: State-Specific Weight Vectors
  - Subspaces, $\mathbf{M}_i$, defined over shared pool of Gaussians
  - State-specific weight vectors, $\mathbf{v}_j$, describe phonetic var. in subspace

$$\hat{\mu}_{j,i} = \mathbf{M}_i \mathbf{v}_j \qquad p(x \mid s_j) = \sum_{i=1}^{I} w_{j.i}\, p(x; \hat{\mu}_{j,i}, \Sigma_i)$$

- ***Generalization***: Joint Model / Speaker Subspaces
  - Model and Speaker subspaces, $\mathbf{M}_i$ and $\mathbf{N}_i$
  - State-specific and speaker specific weight vectors $\mathbf{v}_j$ and $v^{(s)}$

$$\hat{\mu}_{j,i}^{(s)} = \mathbf{M}_i \mathbf{v}_j + \mathbf{N}_i v^{(s)}$$

Subspace HMM

Observation Prob. for State j:

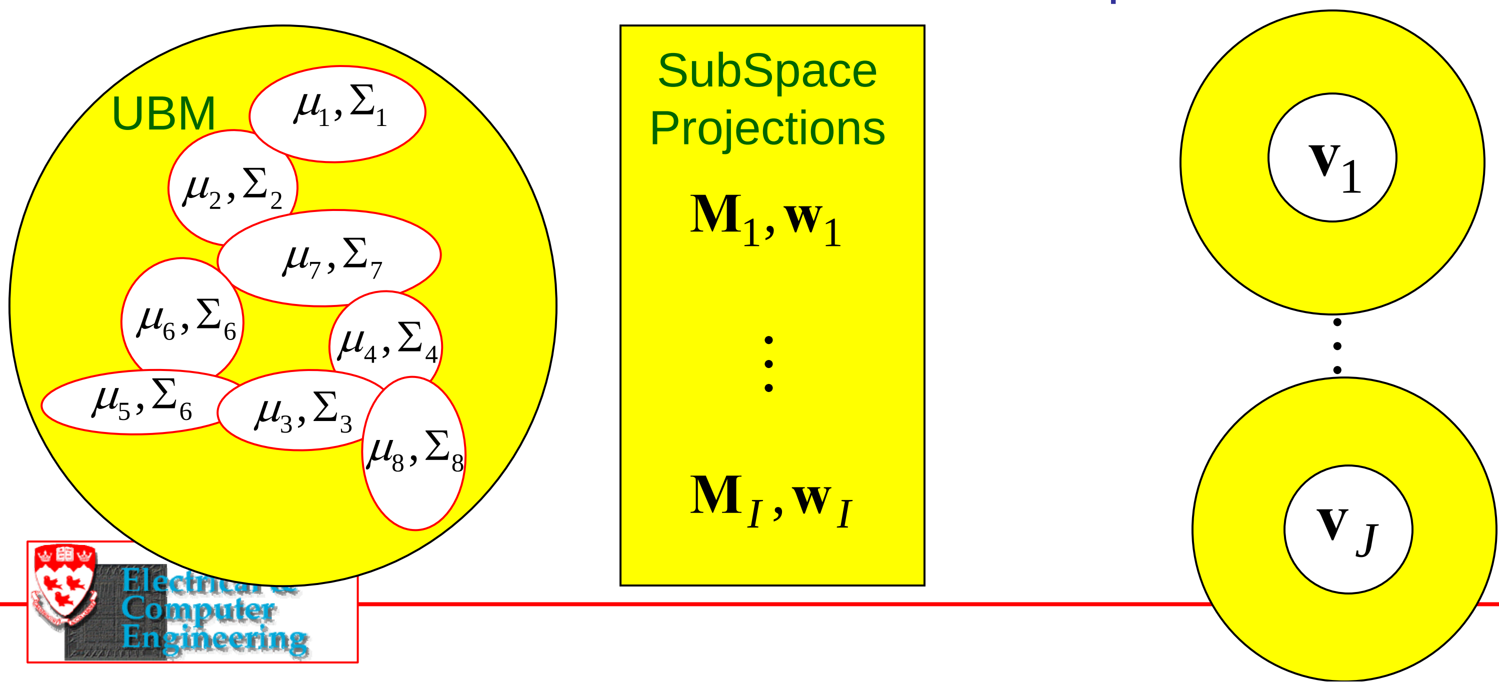
$$p(x \mid s_j) = \sum_{i=1}^{I} w_{j,i}\, p(x; \mu_{j,i}, \Sigma_i)$$

Projection of State-SpecificVectors:

$$\mu_{j,i} = \mathbf{M}_i \mathbf{v}_j \qquad w_{j,i} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{l=1}^{I} \exp \mathbf{w}_l^T \mathbf{v}_j}$$

Shared Parameters

State Specific Parameters



Subspace HMM – Multiple Substates

Observation Prob. for State j:

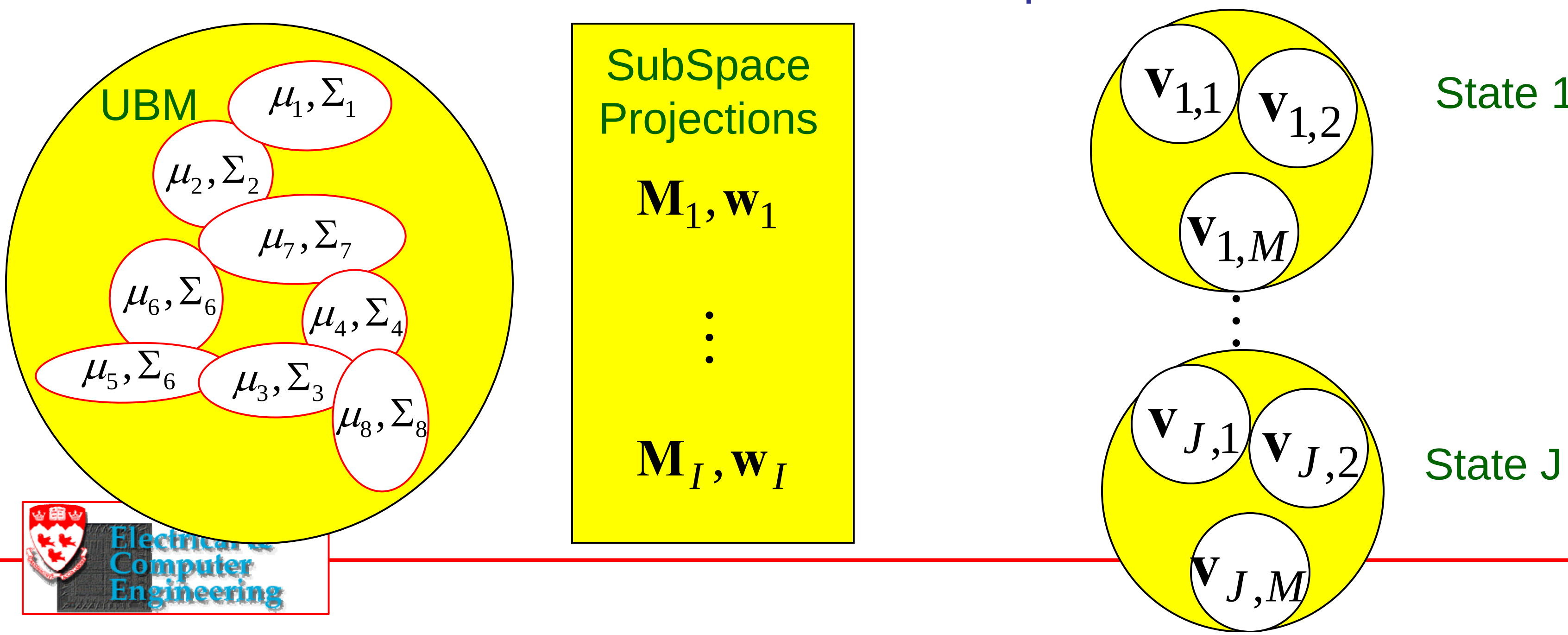
$$p(x \mid s_j) = \sum_{m=1}^{M} c_{j,m} \sum_{i=1}^{I} w_{j,m,i}\, p(x; \mu_{j,m,i}, \Sigma_i)$$

Projection of State-SpecificVectors:

$$\mu_{j,m,i} = \mathbf{M}_i \mathbf{v}_{j,m} \qquad w_{j,m,i} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{j,m}}{\sum_{l=1}^{I} \exp \mathbf{w}_l^T \mathbf{v}_{j,m}}$$

Shared Parameters

State / Sub-state Specific Parameters



Empirical Trade-Off: Shared and State-Specific Parameters

- Example: Wall Street Journal HMM model:
- 6000 states, 100,000 Gaussians ⇨ ***~8 Million parameters***
- Possible Substate HMM Parameterizations:

Parameter Allocation			Number of parameters		
UBM Gaussians	Sub-space Dim.	Sub-States	Shared	State-Specific	Total
256	39	1	600K	235K	835K
1024	39	1	2.4M	235K	2.65M
1024	100	1	4.8M	600K	5.4M
256	39	16	600K	3.7M	4.3M

Summary

- The workshop is investigating a subspace based alternative to HMM models that includes:
  1. ***multiple model subspaces*** defined over different regions of the feature space
  2. ***state-specific weight vectors*** for describing phonetic variation within these subspaces
  3. ***joint model / speaker subspaces***

- Workshop goals are to investigate:
  - Potential for sharing training data across languages and task domains
  - Empirical trade-off between number of Gaussians, states, sub-states, and sub-state dimension
  - Effects of joint subspaces: modeling both phonetic and speaker variation

Electrical & Computer Engineering