# Improved Linear Discriminant Analysis Considering Empirical Pairwise Classification Error Rates

Hung-Shin Lee and Berlin Chen

*Spoken Language Processing Laboratory*
*Taiwan Normal University, Taipei*

# Outline

- Linear Discriminant Analysis
- Limitation of LDA
- Approximate Pairwise Empirical Accuracy Criterion (aPEAC)
  - Weighted LDA
  - Approximate Accuracy Function
- Experiments and Results
- Conclusions

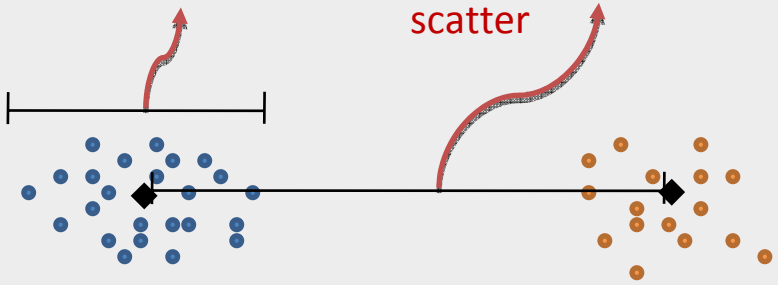
# Linear Discriminant Analysis (1/4)

- The problem formulation of LDA
  - To seek a linear transformation $\Theta \in R^{n\times p}$ that reduces the dimensionality of *n*-dimensional feature vectors to *p* (*p* < *n*) by maximizing the following discrimination criterion in the transformed space

$$J_{LDA}(\Theta) = \text{trace}\left(\left(\Theta^T \mathbf{S}_W \Theta\right)^{-1}\left(\Theta^T \mathbf{S}_B \Theta\right)\right)$$

within-class scatter

between-class scatter



$$\mathbf{S}_W = \sum_{i=1}^{C} \frac{n_i}{N}\mathbf{S}_i = \sum_{i=1}^{C} p_i \mathbf{S}_i$$

$$\mathbf{S}_B = \sum_{i=1}^{C} p_i (\mathbf{m}_i - \overline{\mathbf{m}})(\mathbf{m}_i - \overline{\mathbf{m}})^T$$

$$= \frac{1}{2}\sum_{i,j=1}^{C} p_i p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$$

# Linear Discriminant Analysis (2/4)

- The physical meaning of LDA:
  - If all classes are assumed to share the same covariance $\mathbf{S}_W$, $J(\Theta)$ can be regarded as **the average square of the Mahalanobis distance between each class mean pair in the transformed space**

$$\Delta_{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j)}$$

$$\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \tilde{\Delta}_{ij}^2 \quad (\text{"}\sim\text{" denotes the transformed subspace})$$

$$= \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{S}}_W^{-1} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)$$

$$= \frac{1}{2}\text{trace}\left(\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T \tilde{\mathbf{S}}_W^{-1} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)\right)$$

$$= \frac{1}{2}\text{trace}\left(\tilde{\mathbf{S}}_W^{-1}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)^T\right) = \text{trace}(\tilde{\mathbf{S}}_W^{-1}\tilde{\mathbf{S}}_B) = J_{LDA}(\Theta)$$

# Linear Discriminant Analysis (3/4)

- The derivation of LDA has an attractive property: Lightweight solvability
  - With no need of an iterative optimization technique
  - Solved as a generalized eigen-analysis problem

$$\mathbf{S}_B\boldsymbol{\theta}_i = \lambda_i \mathbf{S}_W \boldsymbol{\theta}_i$$

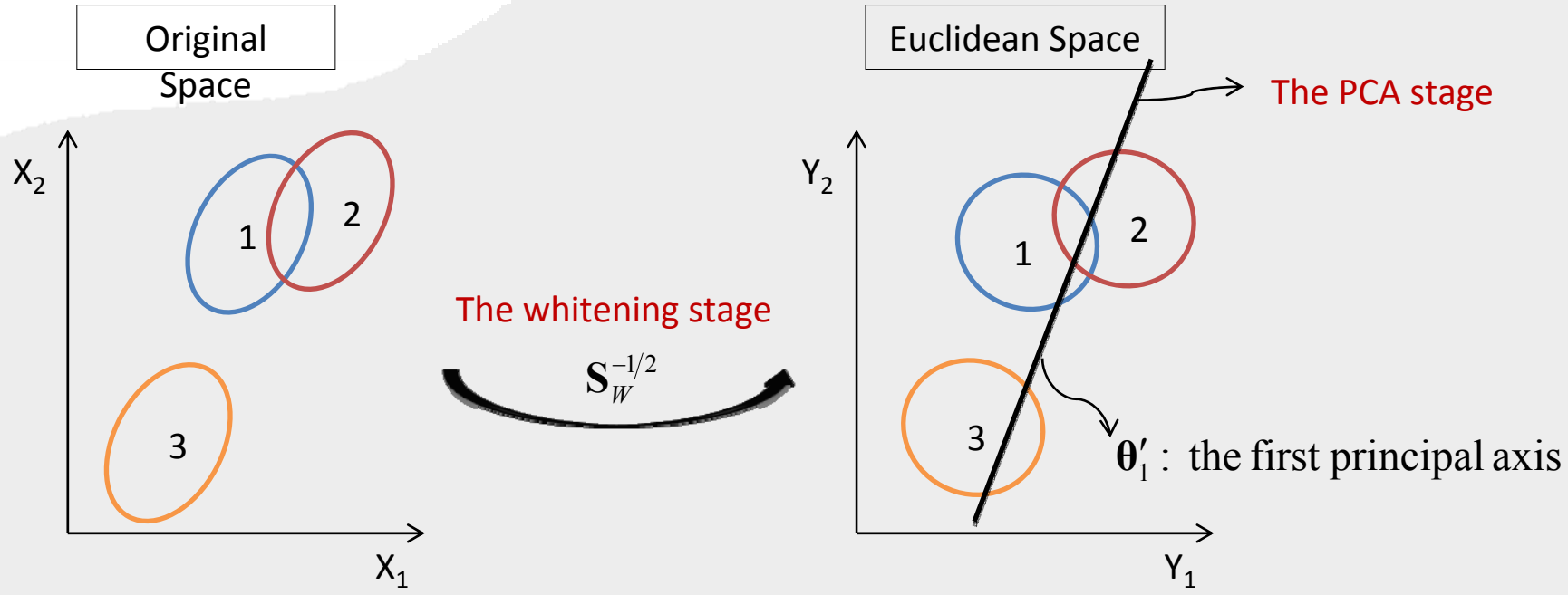
$\lambda_i$ : the $i$th largest eigenvalue of $\mathbf{S}_W^{-1}\mathbf{S}_B$

$\boldsymbol{\theta}_i$ : the corresponding eigenvector

$$\boldsymbol{\Theta} = \left[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p\right]$$

# Linear Discriminant Analysis (4/4)

- The derivation of the LDA matrix can be also geometrically viewed as a two-stage procedure




– It can be shown that $\boldsymbol{\Theta} = \mathbf{S}_W^{-1/2}\left[\boldsymbol{\theta}_1' \ldots \boldsymbol{\theta}_p'\right]$ also maximizes $J_{LDA}(\boldsymbol{\Theta})$

# LDA's Limitation – Geometrical Separation vs. Error Rate

- From the formulation of LDA, it appears that LDA does not **directly** relate itself to classification error rates

- Actually, LDA does not guarantee to find the optimal subspace in the Bayesian sense
  - However, what we really care about is the classification accuracy for a given classifier (like HMMs), but not the geometrical separation between each class pair

# Weighted LDA (1/2)

- To modify the LDA formulation such that it is more closely related to the figure of merit for classification while retaining its lightweight solvability
  - The new weighted criterion becomes

$$J_w(\boldsymbol{\Theta}') = \text{trace}\left( \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} p_i p_j w(i,j) \boldsymbol{\Theta}'^T (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T \boldsymbol{\Theta}' \right)$$

("$\cap$" denotes the whitened subspace)

weighting function

mean difference in the whitened space

where $\boldsymbol{\Theta}'$ can be found by solving the eigenvectors of the matrix

$$\frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} p_i p_j w(i,j) (\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T$$

# Weighted LDA (2/2)

- Then, the final transformation matrix $\boldsymbol{\Theta}$ can be derived by pre-multiplying $\boldsymbol{\Theta}'$ with a whitening transform $\mathbf{S}_W^{-1/2}$

$$\boldsymbol{\Theta} = \mathbf{S}_W^{-1/2}\boldsymbol{\Theta}'$$

- Here, we focus on how to choose a suitable $w(i,j)$, which can convert the class separation measurement from the pairwise distance to the pairwise empirical classification accuracy

# Empirical Error Function (1/3)

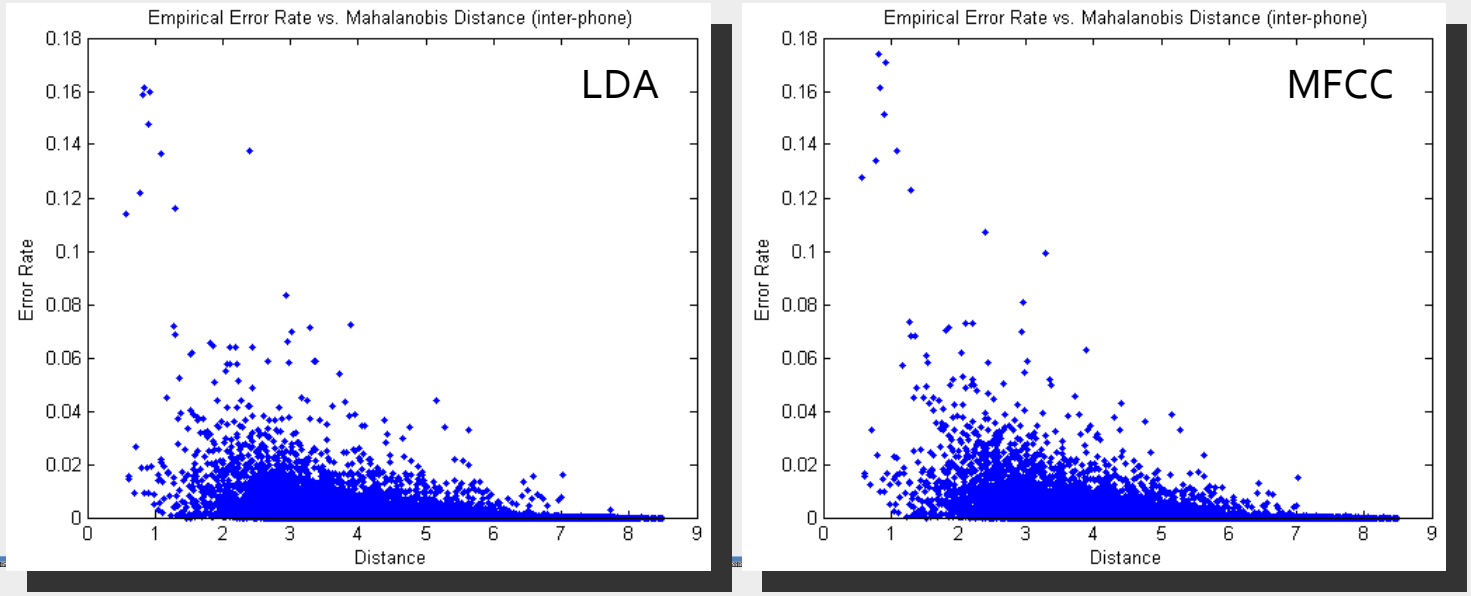
- First we define the empirical pairwise classification error rate between classes *i* and *j* as

$$ER_{ij} = \frac{e_{ij} + e_{ji}}{n_i + n_j}$$

the number of samples that originally belong to class *i* but are misallocated to class *j* by the classifier

the size of class *i*

- Dot Plots:



# Empirical Error Function (2/3)

- To appropriately model the phenomenon, we use the data-fitting (or regression) scheme to find out a function of the Mahalanobis distance $\mathrm{E}(\Delta_{ij})$, which can approximate the relationship between the empirical pairwise classification error rate and the corresponding Mahalanobis distance

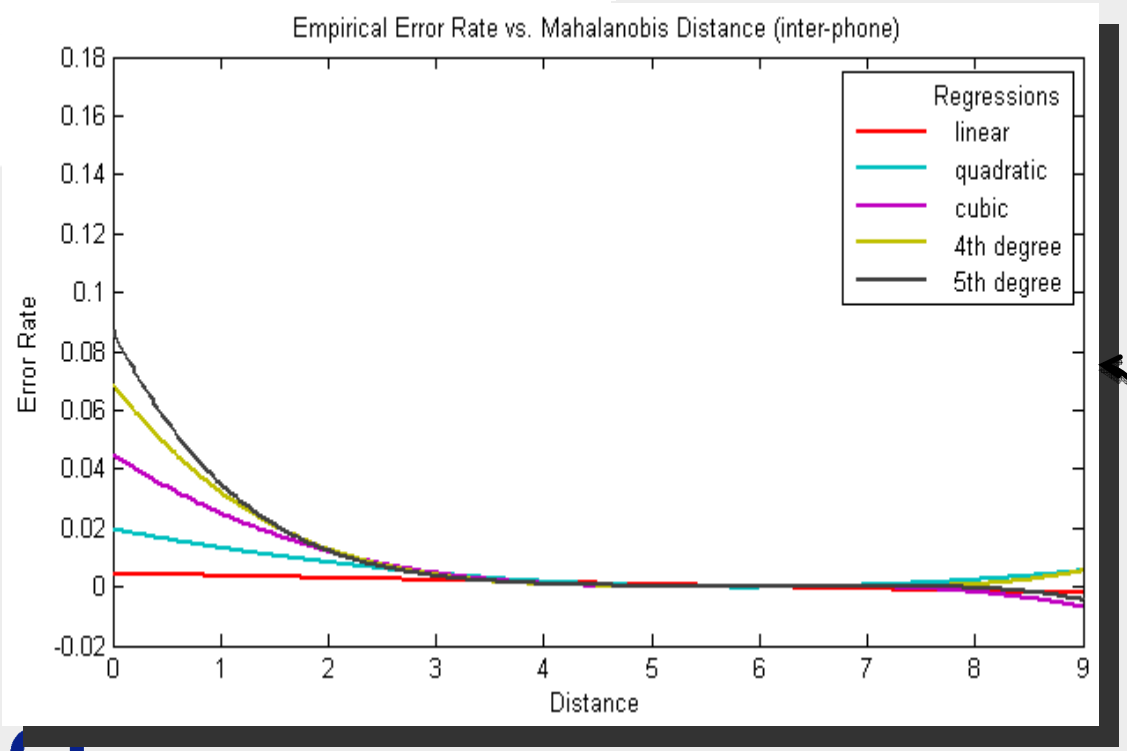
- Data Fitting
  - if $\mathrm{E}(\Delta_{ij})$ is a **quadratic polynomial**

$$\mathrm{E}(\Delta_{ij}) = a\Delta_{ij}^2 + b\Delta_{ij} + c$$

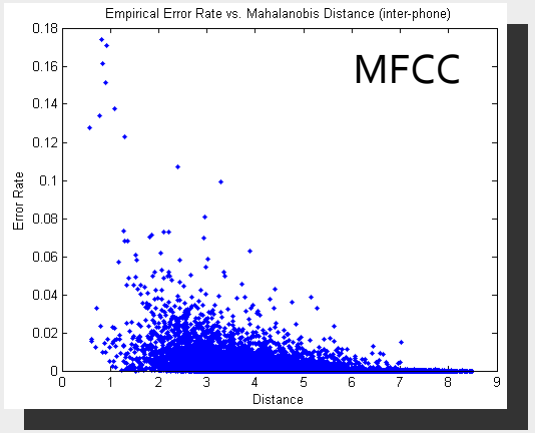
$$\{\hat{a},\hat{b},\hat{c}\} = \arg\min_{a,b,c} \sum_{i=1}^{C-1} \sum_{j=i}^{C} \left[ \left( a\Delta_{ij}^2 + b\Delta_{ij} + c \right) - ER_{ij} \right]^2$$

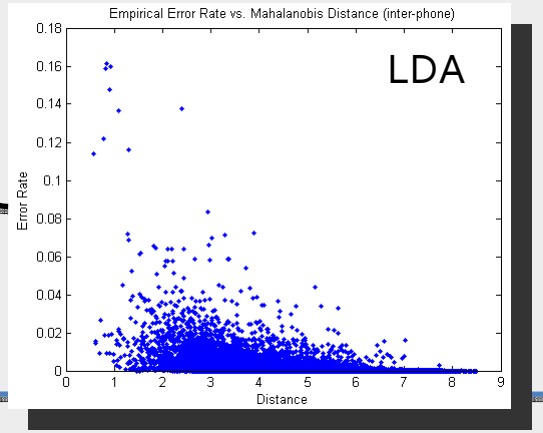
# Empirical Error Function (3/3)

- The curves of the error functions, derived on the basis of data-fitting for polynomials of degrees 1 up to 5




**Dot plots**



# Approximate Pairwise Empirical Accuracy Criterion (aPEAC) (1/5)

- The derived error function $\hat{\mathrm{E}}(\Delta_{ij})$ can be used to predict the pairwise classification error rate, and to approximate the pairwise **empirical accuracy** for any class pair *i* and *j*

$$\hat{\mathrm{A}}(\Delta_{ij}) = 1 - \hat{\mathrm{E}}(\Delta_{ij})$$

- Based on the weighted LDA, we define a new accuracy-based between-class scatter matrix in the whitened space

$$\mathbf{S}'_B = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} p_i p_j \frac{\left(1-\hat{\mathrm{E}}(\Delta_{ij})\right)}{\Delta_{ij}^2}(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)(\hat{\mathbf{m}}_i - \hat{\mathbf{m}}_j)^T$$

$$\text{weighting function} : w(i,j)$$

# Approximate Pairwise Empirical Accuracy Criterion (aPEAC) (2/5)

- Our proposed new criterion, named approximate pairwise empirical accuracy criterion (aPEAC), can be defined as

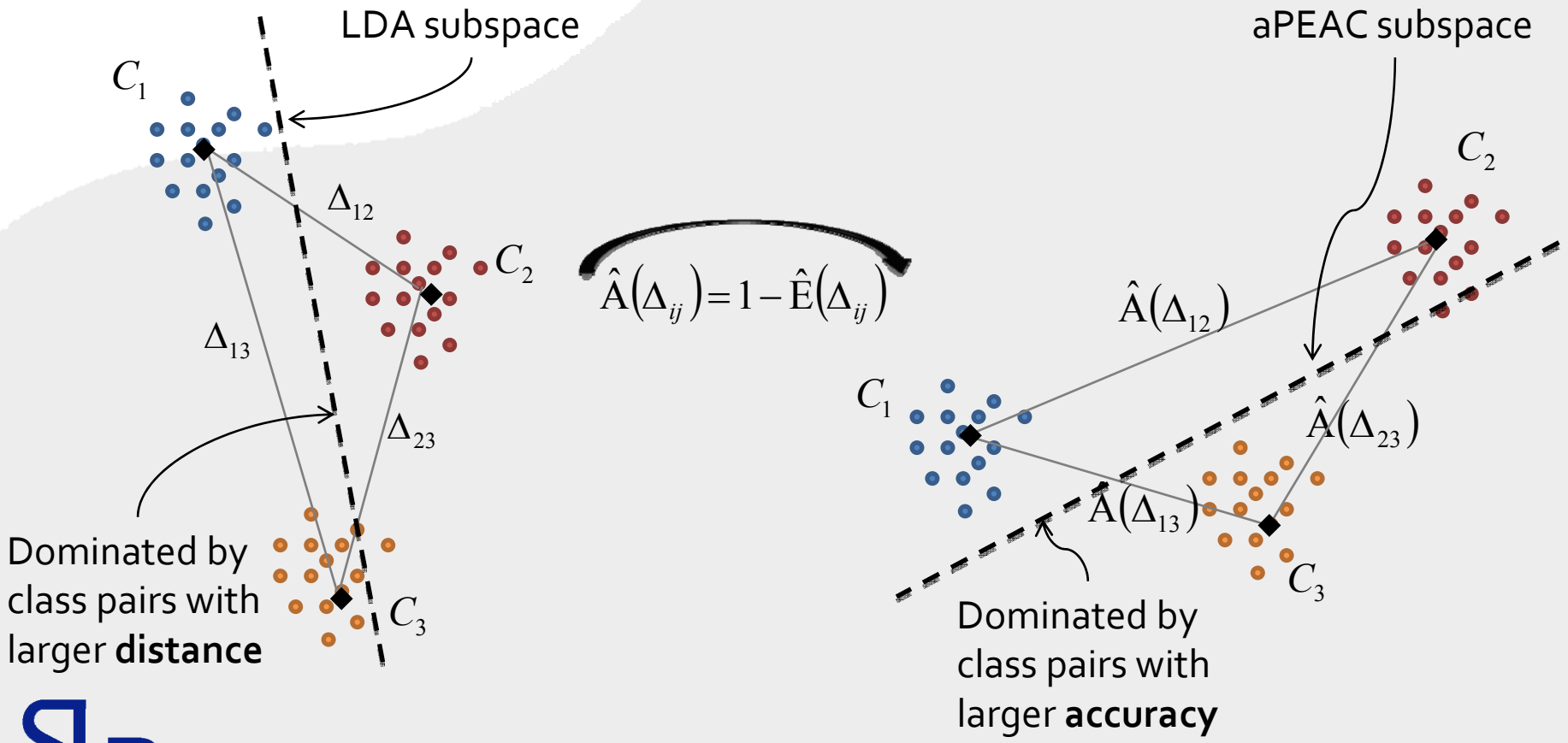
("∩" denotes the whitened subspace)

$$J_{aPEAC}\left(\widehat{\boldsymbol{\Theta}}\right)=\operatorname{trace}\left(\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}p_i p_j \frac{1-\hat{\mathrm{E}}(i,j)}{\Delta_{ij}^2}\widehat{\boldsymbol{\Theta}}^T\left(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j\right)\left(\hat{\mathbf{m}}_i-\hat{\mathbf{m}}_j\right)^T\widehat{\boldsymbol{\Theta}}\right)$$

$$\text{weighting function}: w(i,j)$$

# Approximate Pairwise Empirical Accuracy Criterion (aPEAC) (3/5)

- Geometrical interpretation of aPEAC:



# Approximate Pairwise Empirical Accuracy Criterion (aPEAC) (4/5)

- Mathematical meaning of aPEAC:
  - Assuming all classes share the same covariance $\mathbf{S}_W$ , $J_{aPEAC}\left(\widehat{\mathbf{\Theta}}\right)$ can be regarded as the average pairwise empirical classification accuracy between each class pair in the transformed space

$$J_{aPEAC}\left(\widehat{\mathbf{\Theta}}\right)=\operatorname{trace}\left(\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}p_i p_j\frac{1-\hat{\mathrm{E}}\left(\Delta_{ij}\right)}{\Delta_{ij}^2}\widehat{\mathbf{\Theta}}^T\left(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j\right)\left(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j\right)^T\widehat{\mathbf{\Theta}}\right)$$

$$=\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}p_i p_j\left(1-\hat{\mathrm{E}}\left(\Delta_{ij}\right)\right)\operatorname{trace}\left(\frac{\widehat{\mathbf{\Theta}}^T\left(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j\right)}{\Delta_{ij}}\frac{\left(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j\right)^T\widehat{\mathbf{\Theta}}}{\Delta_{ij}}\right)$$

unit vector

$$=\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}p_i p_j\left(1-\hat{\mathrm{E}}\left(\Delta_{ij}\right)\right)\operatorname{trace}\left(\frac{\left(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j\right)^T\widehat{\mathbf{\Theta}}}{\Delta_{ij}}\frac{\widehat{\mathbf{\Theta}}^T\left(\widehat{\mathbf{m}}_i-\widehat{\mathbf{m}}_j\right)}{\Delta_{ij}}\right)$$

idnetity matrix

$$=\frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C}p_i p_j\left(1-\hat{\mathrm{E}}\left(\Delta_{ij}\right)\right)$$

generated in the transformed space

# Approximate Pairwise Empirical Accuracy Criterion (aPEAC) (5/5)

- Comparisons among aPEAC, LDA, HLDA, and another weighting-based approach – aPTAC (Loog, 2001) :
  - aPTAC: approximate pairwise **theoretical** accuracy criterion

$$w(i,j)=\frac{1}{2\Delta_{ij}^2}\operatorname{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right)$$

| | LDA | aPEAC | aPTAC | HLDA |
|---|---|---|---|---|
| Distribution Assumption | No | **No** | Gaussian | Gaussian |
| Separation Criterion | Geometrical Distance | **Empirical Accuracy** | Theoretical Accuracy | No (Likelihood) |
| Error Rate Related | No | **Yes** | Yes | No |
| Classifier Related | No | **Yes** | No | No |
| Solvability | Lightweight | **Lightweight** | Lightweight | Complicated |

# Experiments and Results (1/2)

- Recognition task: Mandarin LVCSR
  - Speech corpus: about 200 hours of MATBN Mandarin television news (a corpus widely used in Taiwan)
  - Acoustic models (diagonal HMMs): 112 right-context-dependent INITIAL's and 38 context-independent FINAL's
  - Language models: consisting of unigram, bigram and trigram models

- LDA-specific setting
  - The unit for class assignment: the states of each HMM
  - The spliced feature vectors were 162-dimensional , which will be reduced to 39 dimensions

# Experiments and Results (2/2)

- Preliminary Results:
  - The CER results (%) with respect to various degrees of polynomials:

| Regressions | Linear | Quadratic | Cubic | 4th Degree | 5th Degree |
|---|---|---|---|---|---|
| without MLLT | 30.62 | 30.40 | 30.59 | 30.69 | 30.47 |
| with MLLT | 28.37 | 28.15 | **27.80** | 28.57 | 28.03 |

    - MLLT: a widely-used transformation for feature de-correlation
  - Comparison among the CER results (%) of various approaches:

| | LDA | aPEAC (3rd degree) | aPTAC | HLDA |
|---|---|---|---|---|
| without MLLT | 31.44 | 30.59 | 30.39 | 44.56 |
| with MLLT | 28.95 | **27.80** | 28.51 | 28.38 |

    - The MFCC baseline: 32.16 %

# Conclusions

- *a*PEAC:
  - A weighting-based LDA retaining lightweight solvability
  - Successfully converted the LDA derivation from class-pair distance maximization to empirical accuracy maximization

- Future work:
  - Compared with other well-known methods, such as MCE-based LDA and fMPE