Speaker adaptation

JHU Summer Workshop, 2009

29 July 2009

Outline

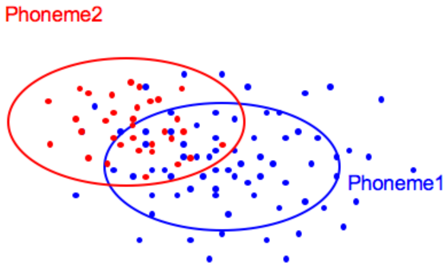
- Speaker-dependent characteristics present in acoustic data
- Modeling speaker characteristics vastly improve recognition performance

Outline

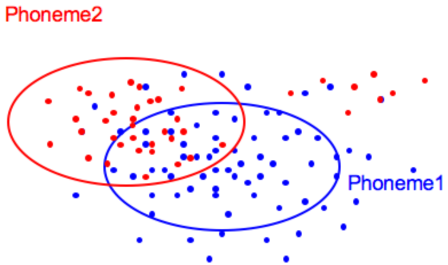
- Speaker-dependent characteristics present in acoustic data
- Modeling speaker characteristics vastly improve recognition performance

1. Speaker vectors
2. Constrained Maximum Likelihood Linear Regression (CMLLR)
3. Subspaces for CMLLR

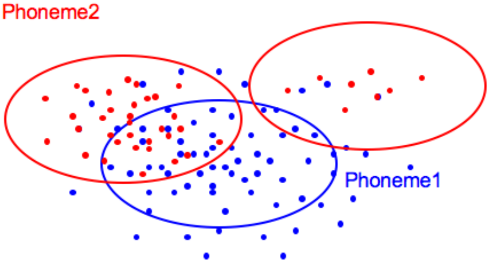
Introduction to speaker subspace

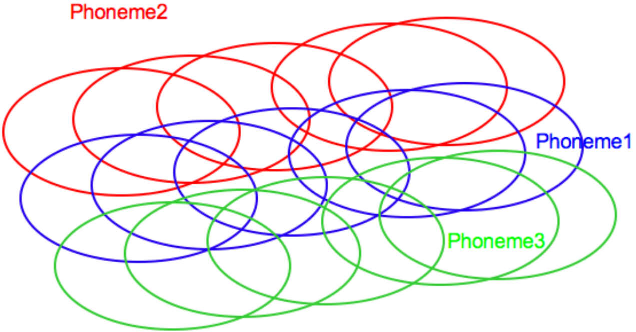


Introduction to speaker subspace

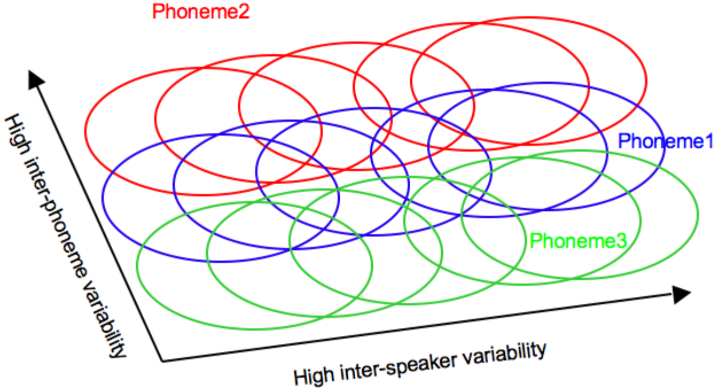


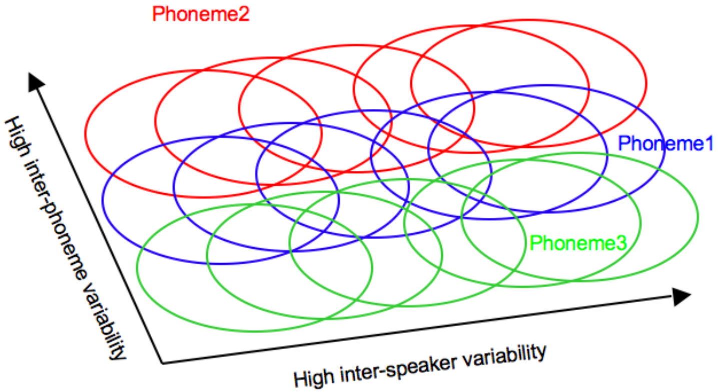
Introduction to speaker subspace




Phoneme2
Phoneme1
Phoneme3

Introduction to speaker subspace

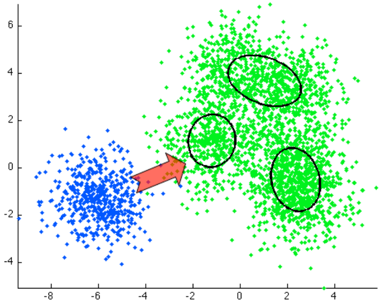




- Speaker-factor in Gaussian mean: $\mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}\,\mathbf{v}^{(s)}$
- Widely used in speaker identification systems

Constrained Maximum Likelihood Linear Regression

- Transform of the observation space to maximize likelihood under current model: $\mathbf{x}^{(s)} = \mathbf{A}^{(s)}\mathbf{x} + \mathbf{b}^{(s)}$



- Speaker-specific mean: $\mu^{(s)} = \mathbf{A}^{(s)}\mu + \mathbf{b}^{(s)}$
- Speaker-specific variance: $\Sigma^{(s)} = \mathbf{A}^{(s)}\,\Sigma\,\mathbf{A}^{(s)\top}$

CMLLR (contd.)

- The auxiliary function is quadratic in $\mathbf{W} = [\,\mathbf{A}\,,\,\mathbf{b}\,]$
- Estimating $\mathbf{W}$ requires solving $(d^2 + d)$ simultaneous equations in $(d^2 + d)$ variables
  - Invert $(d^2 + d) \times (d^2 + d)$ matrix. $O(d^6)$ complexity.

CMLLR (contd.)

- The auxiliary function is quadratic in $\mathbf{W} = [\,\mathbf{A}\,,\,\mathbf{b}\,]$
- Estimating $\mathbf{W}$ requires solving $(d^2 + d)$ simultaneous equations in $(d^2 + d)$ variables
  - Invert $(d^2 + d) \times (d^2 + d)$ matrix. $O(d^6)$ complexity.
- Can be simplified for diagonal covariance case. Row-by-row update [Gales & Woodland, 1996].

CMLLR (contd.)

- The auxiliary function is quadratic in $\mathbf{W} = [\,\mathbf{A}\,,\,\mathbf{b}\,]$
- Estimating $\mathbf{W}$ requires solving $(d^2 + d)$ simultaneous equations in $(d^2 + d)$ variables
  - Invert $(d^2 + d) \times (d^2 + d)$ matrix. $O(d^6)$ complexity.
- Can be simplified for diagonal covariance case. Row-by-row update [Gales & Woodland, 1996].
- For full covariance case, row-by-row update still requires computing $O(d^4)$ statistics [Sim & Gales, 2005].

CMLLR (contd.)

- The auxiliary function is quadratic in $\mathbf{W} = [\,\mathbf{A}\,,\,\mathbf{b}\,]$
- Estimating $\mathbf{W}$ requires solving $(d^2 + d)$ simultaneous equations in $(d^2 + d)$ variables
  - Invert $(d^2 + d) \times (d^2 + d)$ matrix. $O(d^6)$ complexity.
- Can be simplified for diagonal covariance case. Row-by-row update [Gales & Woodland, 1996].
- For full covariance case, row-by-row update still requires computing $O(d^4)$ statistics [Sim & Gales, 2005].

- Optimal $\mathbf{W}$ can be computed using Newton's method.
- Computing inverse Hessian (matrix of second derivatives) will still require $O(d^4)$ storage, $O(d^6)$ computation.

Our approach

- Transform the *data* and *model*, such that:
  - average within-class variance is unit
  - covariance of the mean vectors is diagonal
  - average mean is zero.
- Transformation with *expected Hessian* simplifies to $O(d^2)$ computation.

Our approach

- Transform the *data* and *model*, such that:
  - average within-class variance is unit
  - covariance of the mean vectors is diagonal
  - average mean is zero.
- Transformation with *expected Hessian* simplifies to $O(d^2)$ computation.

Optimation steps

1. Compute the local gradient: $\mathbf{P}$
2. Compute gradient in transformed space: $\tilde{\mathbf{P}}$
3. Proposed change in $\mathbf{W}$ in this space: $\tilde{\Delta} = \frac{1}{\beta}\tilde{\mathbf{P}}$
4. Transform $\tilde{\Delta}$ back to the original space, and update: $\mathbf{W} \leftarrow \mathbf{W} + k\Delta$
   - Optimal value of k can be computed iteratively

CMLLR Subspaces

- Extending the idea of parameter subspaces to CMLLR transforms
- Express $\mathbf{W}^{(s)}$ as a linear combination of orthonormal basis matrices

$$\mathbf{W}^{(s)} = \mathbf{W}_0 + \sum_{b=1}^{B} \lambda_b^{(s)} \mathbf{W}_b$$

CMLLR Subspaces

- Extending the idea of parameter subspaces to CMLLR transforms
- Express $\mathbf{W}^{(s)}$ as a linear combination of orthonormal basis matrices

$$\mathbf{W}^{(s)} = \mathbf{W}_0 + \sum_{b=1}^{B} \lambda_b^{(s)} \mathbf{W}_b$$

- Parameter update in the transformed space modified as:

$$\tilde{\Delta} = \frac{1}{\beta} \sum_{b=1}^{B} \tilde{\mathbf{W}}_b \, \mathrm{Tr}(\tilde{\mathbf{W}}_b \tilde{P}^\top)$$

CMLLR Subspaces

- Extending the idea of parameter subspaces to CMLLR transforms
- Express $\mathbf{W}^{(s)}$ as a linear combination of orthonormal basis matrices

$$\mathbf{W}^{(s)} = \mathbf{W}_0 + \sum_{b=1}^{B} \lambda_b^{(s)} \mathbf{W}_b$$

- Parameter update in the transformed space modified as:

$$\tilde{\Delta} = \frac{1}{\beta} \sum_{b=1}^{B} \tilde{\mathbf{W}}_b \, \mathrm{Tr}(\tilde{\mathbf{W}}_b \tilde{P}^\top)$$

- Helps us to perform speaker-adaptation with relatively little adaptation data

Results

- Adaptation per speaker

System	% Accuracy
Baseline	50.3
+ speaker vectors	51.0
+ CMLLR	51.7
+ speaker vectors + CMLLR	52.0

- Adaptation per utterance

System	% Accuracy
Baseline	50.3
+ CMLLR	50.3
+ CMLLR subspaces	50.8