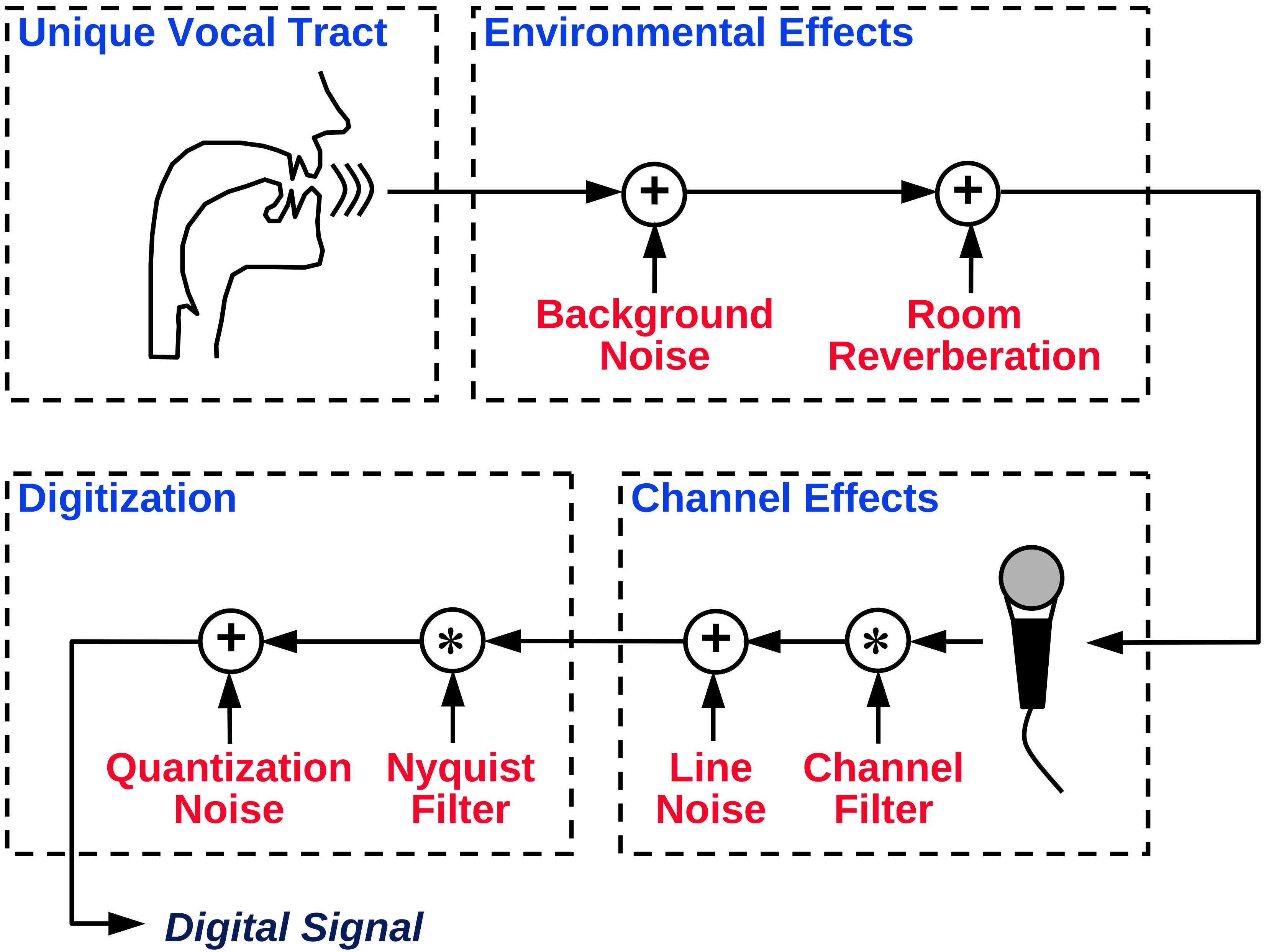
Speaker Adaptation

Lecturer: T. J. Hazen

- Overview
- Adaptation Methods
  - Vocal Tract Length Normalization
  - Bayesian Adaptation
  - Transformational Adaptation
  - Reference Speaker Weighting
  - Eigenvoices
  - Structural Adaptation
  - Hierarchical Speaker Clustering
  - Speaker Cluster Weighting
- Summary

Typical Digital Speech Recording



Accounting for Variability

- **Recognizers must account for variability in speakers**
- **Standard approach: Speaker Independent (SI) training**
  - **Training data pooled over many different speakers**
- **Problems with primary modeling approaches:**
  - **Models are heterogeneous and high in variance**
  - **Many parameters are required to build accurate models**
  - **Models do not provide any speaker constraint**
  - **New data may still not be similar to training data**

Providing Constraint

- **Recognizers should also provide constraint:**
  - **Sources of variation typically remain fixed during utterance**
  - **Same speaker, microphone, channel, environment**
- **Possible Solutions:**
  - **Normalize input data to match models (i.e., Normalization)**
  - **Adapt models to match input data (i.e., Adaptation)**
- **Key ideas:**
  - **Sources of variability are often systematic and consistent**
  - **A few parameters can describe large systematic variation**
  - **Within-speaker correlations exist between different sounds**

Probabilistic Framework

- **Acoustic model predicts likelihood of acoustic observations given phonetic units:**

$$P(A \mid U) = P(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_N \mid u_1, u_2, \ldots, u_n)$$

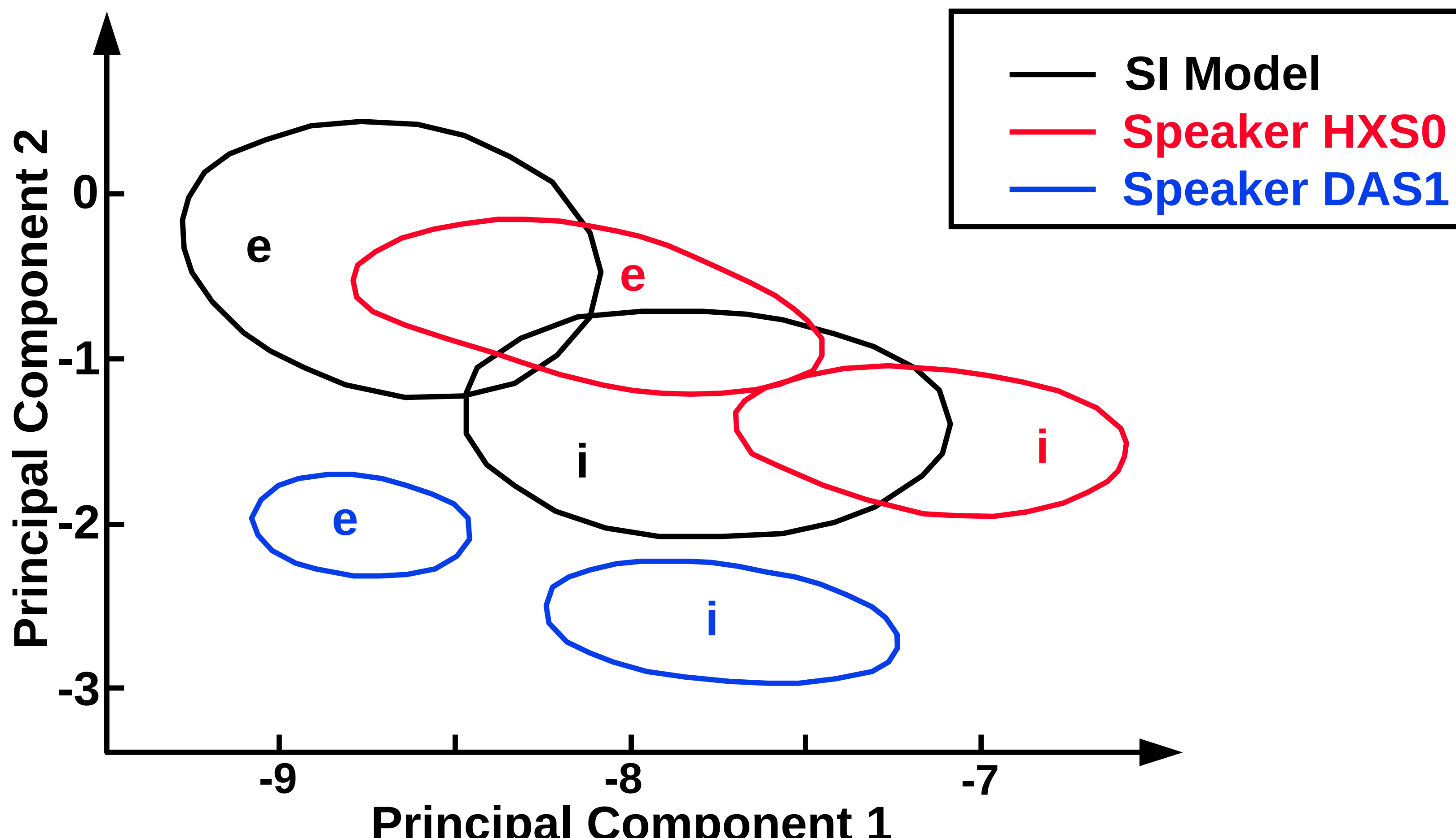
- **An independence assumption is typically required in order to make the modeling feasible:**

$$P(A \mid U) = \sum_{i=1}^{N} P(\vec{a}_i \mid U)$$

- **This independence assumption can be harmful!**
  - **Acoustic correlations between phonetic events are ignored**
  - **No constraint provided from previous observations**

Variability and Correlation

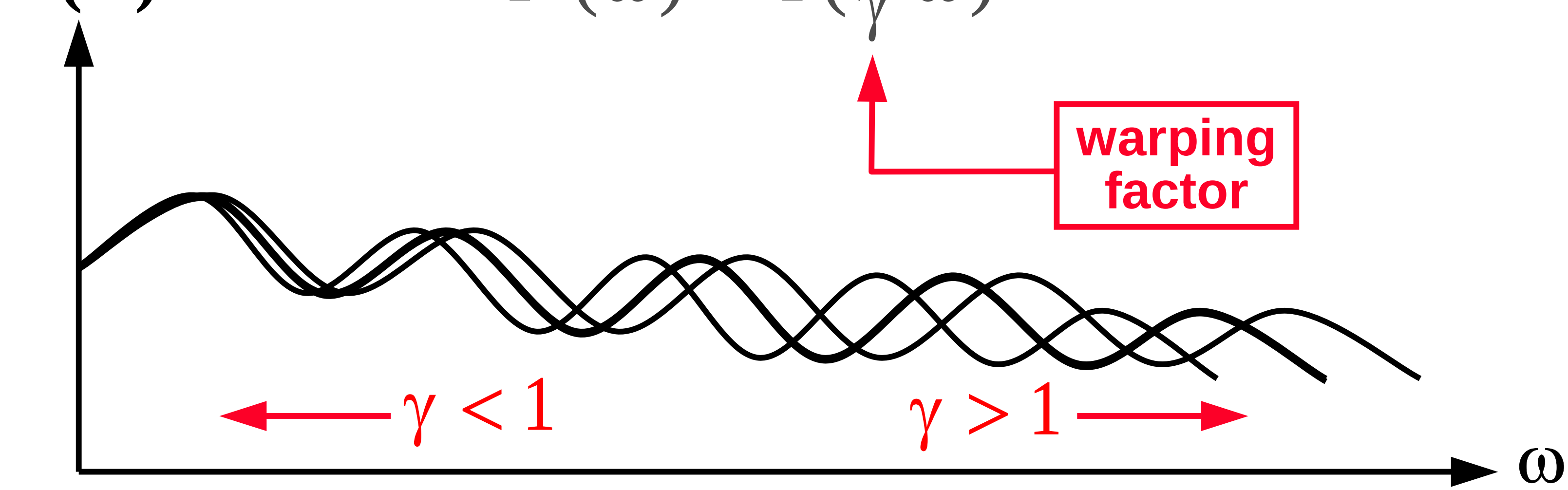
- Plot of isometric likelihood contours for phones [i] and [e]
- One SI model and two speaker dependent (SD) models
- SD contours are tighter than SI and correlated w/ each other



Vocal Tract Length Normalization

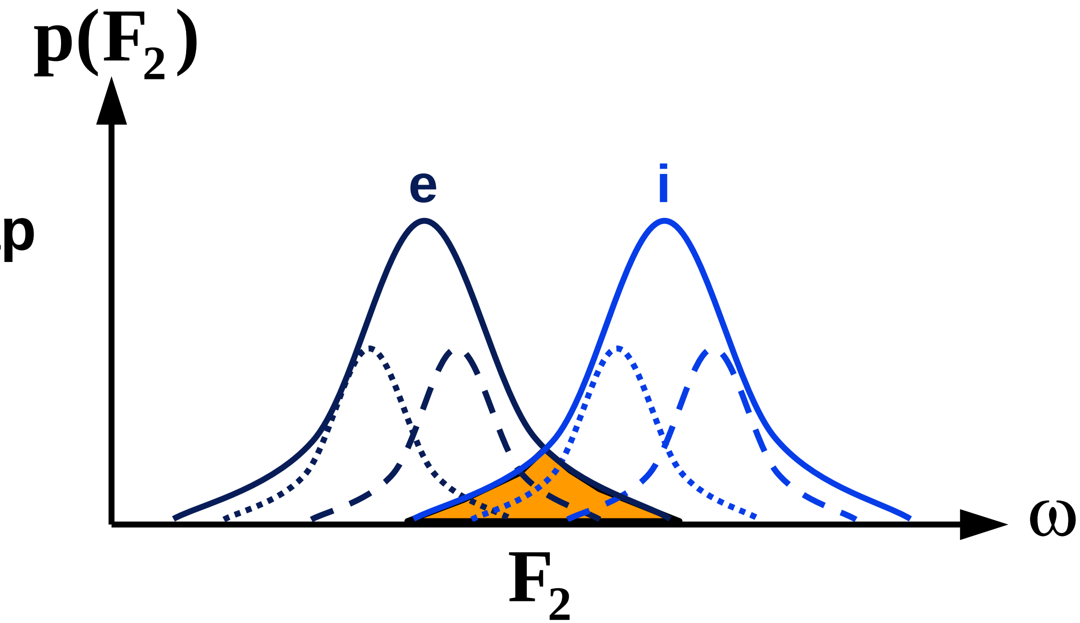
- Vocal tract length affects formant frequencies:
  - shorter vocal tracts $\Rightarrow$ higher formant frequencies
  - longer vocal tracts $\Rightarrow$ lower formant frequencies
- Vocal tract length normalization (VTLN) tries to adjust input speech to have an "average" vocal tract length
- Method: Warp the frequency scale!

$$Y'(\omega) = Y(\frac{1}{\gamma}\omega)$$

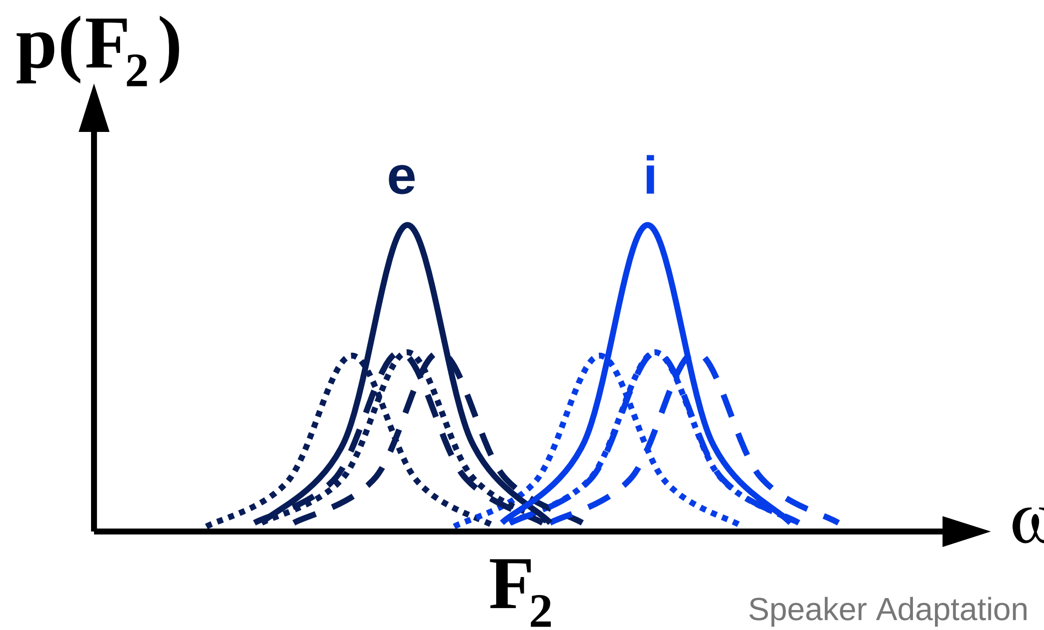


Vocal Tract Length Normalization (cont)

- **Illustration: second formant for [e] and [i]**
- **SI models have large overlap (error region)**
- **SD models have smaller variances & error region**



- **Warp spectrums of all training speakers to best fit SI model**
- **Train VTLN-SI model**
- **Warp test speakers to fit VTLN-SI model**



Vocal Tract Length Normalization

- **During testing ML approach is used to find warp factor:**

$$\gamma = \arg\max_{\gamma} p(X^{\gamma} \mid \Theta_{\text{VTLN}})$$

- **Warp factor is found using brute force search**
  - **Discrete set of warp factors tested over possible range**
- **References:**
  - **Andreou, Kamm, and Cohen, 1994**
  - **Lee and Rose, 1998**

Speaker Dependent Recognition

- **Conditions of experiment:**
  - **DARPA Resource Management task (1000 word vocabulary)**
  - **SUMMIT segment-based recognizer using word pair grammar**
  - **Mixture Gaussian models for 60 context-independent units:**
  - **Speaker dependent training set:**
    - 12 speakers w/ 600 training utts and 100 test utts per speaker
    - ~80,000 parameters in each SD acoustic model set
  - **Speaker independent training set:**
    - 149 speakers w/ 40 training utts per speaker (5960 total utts)
    - ~400,000 parameters in SI acoustic model set
- **Word error rate (WER) results on SD test set:**
  - **SI recognizer had 7.4% WER**
  - **Average SD recognizer had 3.4% WER**
  - **SD recognizer had 50% fewer errors using 80% fewer parameters!**

Adaptation Definitions

- **Speaker dependent models don't exist for new users**
- **System must learn characteristics of new users**
- **Types of adaptation:**
  - **Enrolled vs. instantaneous**
    - Is a prerecorded set of adaptation data utilized or is test data used as adaptation data?
  - **Supervised vs. unsupervised**
    - Is orthography of adaptation data known or unknown?
  - **Batch vs. on-line**
    - Is adaptation data presented all at once or one at a time?

Adaptation Definitions (cont)

- **Goal: Adjust model parameters to match input data**
- **Definitions:**
  - X **is a set of adaptation data**
  - Λ **is a set of adaptation parameters, such as:**
    * Gender and speaker rate
    * Mean vectors of phonetic units
    * Global transformation matrix
  - Θ **is a set of acoustic model parameters used by recognizer**
- **Method:**
  - Λ **is estimated from** X
  - Θ **is adjusted based on** Λ

Adaptation Definitions (cont)

- **Obtaining Λ is an estimation problem:**
  - **Few adaptation data points $\Rightarrow$ small # of parameters in Λ**
  - **Many adaptation data points $\Rightarrow$ larger # of parameters in Λ**
- **Example:**
  - **Suppose Λ contains only a single parameter λ**
  - **Suppose λ represents the probability of speaker being male**
  - **λ is estimated from the adaptation data X**
  - **The speaker adapted model could be represented as:**

$$P(\vec{a} \mid \Theta_{sa}) = \lambda P(\vec{a} \mid \Theta_{male}) + (1 - \lambda) P(\vec{a} \mid \Theta_{female})$$

Bayesian Adaptation

- **A method for direct adaptation of models parameters**
- **Most useful with large amounts of adaptation data**
- **A.k.a. maximum *a posteriori* probability (MAP) adaptation**
- **General expression for MAP adaptation of mean vector of a single Gaussian density function:**

$$\vec{\mu} = \arg\max_{\vec{\mu}} p(\vec{\mu}|X) = \arg\max_{\vec{\mu}} p(\vec{\mu}|\vec{x}_1,\ldots,\vec{x}_N)$$

- **Apply Bayes rule:**

$$\vec{\mu} = \arg\max_{\vec{\mu}} \underbrace{\boxed{p(X\,|\,\vec{\mu})}}_{\text{observation likelihood}} \underbrace{\boxed{p(\vec{\mu})}}_{\textit{a priori} \text{ model}}$$

Bayesian Adaptation (cont)

- **Assume observations are independent:**

$$p(X \mid \vec{\mu}) = p(\vec{x}_1, \ldots, \vec{x}_N \mid \vec{\mu}) = \prod_{n=1}^{N} p(\vec{x}_n \mid \vec{\mu})$$

- **Likelihood functions modeled with Gaussians:**

$$p(\vec{x} \mid \vec{\mu}) = N(\vec{\mu}; S) \qquad p(\vec{\mu}) = N(\vec{\mu}_{ap}; S_{ap})$$

- **Adaptation parameters found from** X**:**

$$\Lambda = \{\vec{\mu}_{ml}, N\} \qquad \vec{\mu}_{ml} = \frac{1}{N} \sum_{n=1}^{N} \vec{x}_n$$

maximum likelihood (ML) estimate

Bayesian Adaptation (cont)

- **The MAP estimate for a mean vector is found to be:**

$$\vec{\mu}_{map} = S(NS_{ap} + S)^{-1}\vec{\mu}_{ap} + NS_{ap}(NS_{ap} + S)^{-1}\vec{\mu}_{ml}$$

- **The MAP estimate is an interpolation of the ML estimates mean and the *a priori* mean:**
  - **If N is small:** $\vec{\mu}_{map} \approx \vec{\mu}_{ap}$
  - **If N is large:** $\vec{\mu}_{map} \approx \vec{\mu}_{ml}$

- **MAP adaptation can be expanded to handle all mixture Gaussian parameters**
  - **Reference: Gauvain and Lee, 1994**

Bayesian Adaptation (cont)

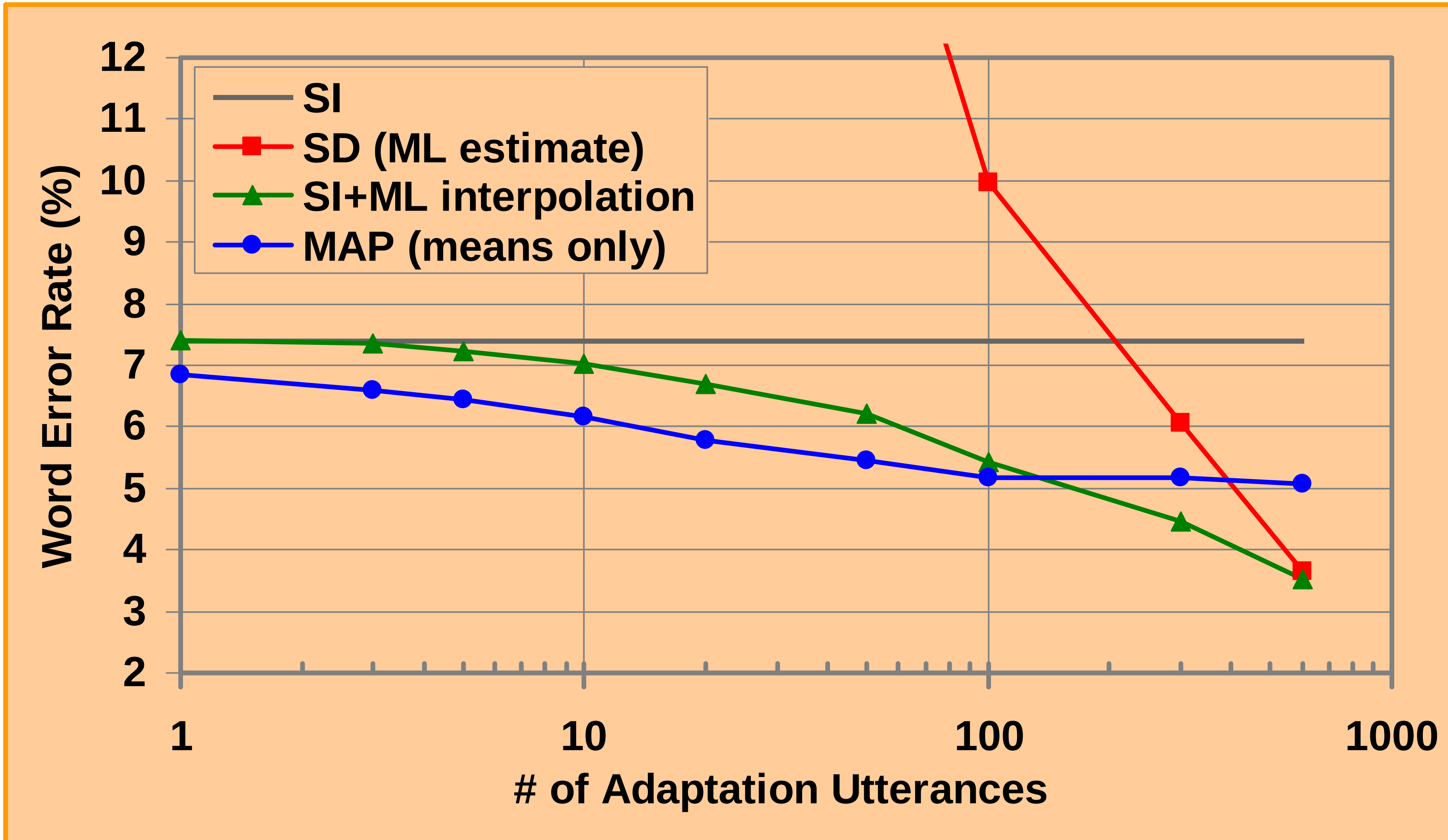
- **Advantages to MAP:**
  - **Based on solid mathematical framework**
  - **Converges to speaker dependent model in limit**
- **Disadvantages to MAP:**
  - **Adaptation is very slow due to independence assumption**
  - **Is sensitive to errors during unsupervised adaptation**
- **Model interpolation adaptation approximates MAP**
  - **Requires no a priori model**
  - **Also converges to speaker dependent model in limit**
  - **Expressed as:**

$$p_{sa}(\vec{x}_n \mid u) = \frac{N}{N+K} p_{ml}(\vec{x}_n \mid u) + \frac{K}{N+K} p_{si}(\vec{x}_n \mid u)$$

K determined empirically

Bayesian Adaptation (cont)

- **Supervised adaptation Resource Management SD test set:**



Transformational Adaptation

- **Transformation techniques are most common form of adaptation being used today!**
- **Idea: Adjust models parameters using a transformation shared globally or across different units within a class**
- **Global mean vector translation:**

$$\boxed{\forall \mathbf{p}} \quad \vec{\mu}_{\mathbf{p}}^{\mathbf{sa}} = \vec{\mu}_{\mathbf{p}}^{\mathbf{si}} + \boxed{\vec{\mathbf{v}}}$$

adapt mean vectors of all phonetic models

shared translation vector

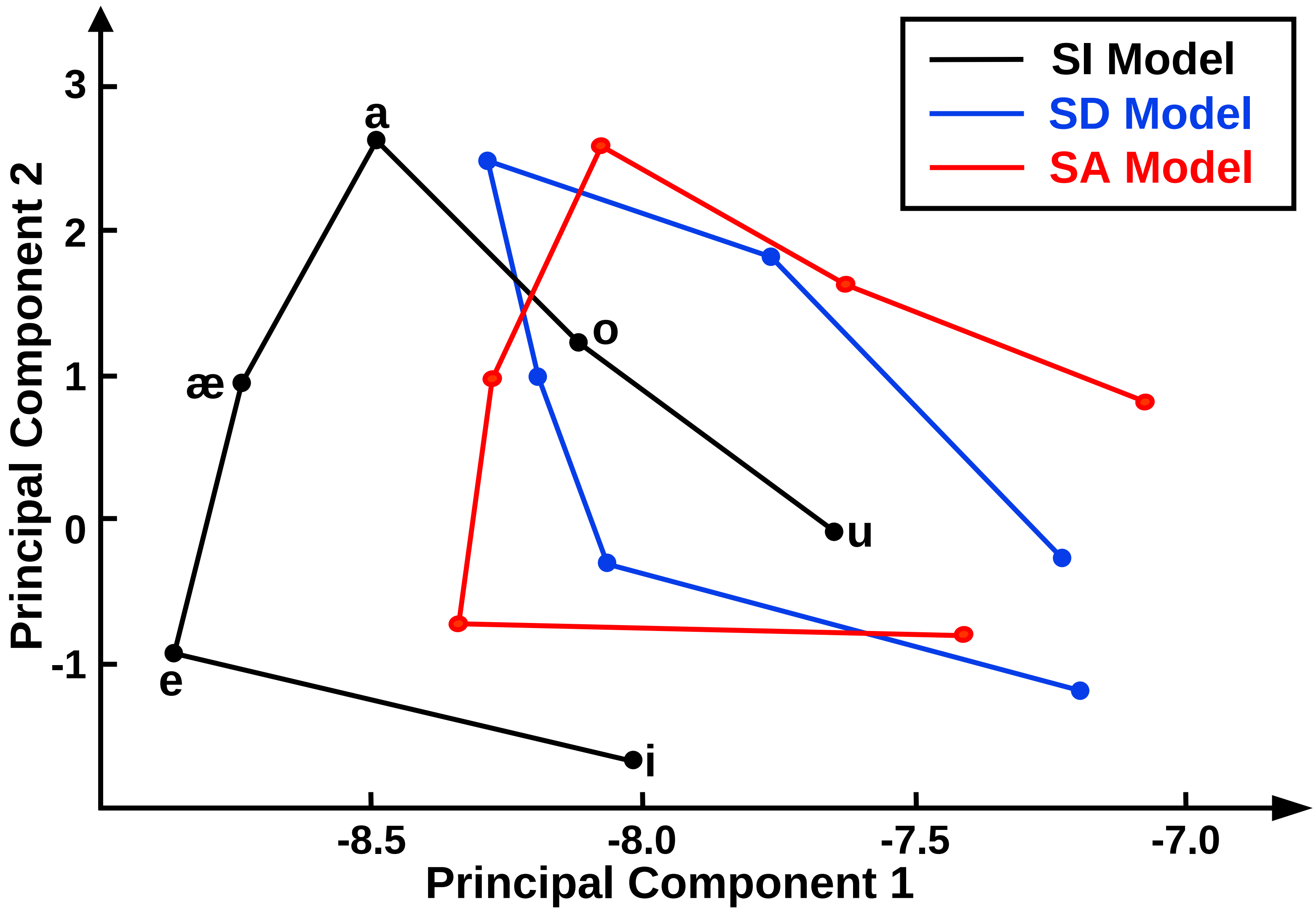
- **Global mean vector scaling, rotation and translation:**

$$\forall \mathbf{p} \quad \vec{\mu}_{\mathbf{p}}^{\mathbf{sa}} = \boxed{\mathbf{R}}\vec{\mu}_{\mathbf{p}}^{\mathbf{si}} + \vec{\mathbf{v}}$$

shared scaling and rotation matrix

Transformational Adaptation (cont)

- **SI model rotated, scaled and translated to match SD model:**



Transformational Adaptation (cont)

- **Transformation parameters found using ML estimation:**

$$[R,\vec{v}] = \arg\max_{R,\vec{v}} p(X|R,\vec{v})$$

- **Advantages:**
  - **Models of units with no adaptation data are adapted based on observations from other units**
  - **Requires no *a priori* model (This may also be a weakness!)**
- **Disadvantages:**
  - **Performs poorly (worse than MAP) for small amounts of data**
  - **Assumes all units should be adapted in the same fashion**
- **Technique is commonly referred to as maximum likelihood linear regression (MLLR)**
  - **Reference: Leggetter & Woodland, 1995**

Reference Speaker Weighting

- Interpolation of models from "reference speakers"
  – Takes advantage of within-speaker phonetic relationships
- Example using mean vectors from training speakers:
  – Training data contains $\mathbf{R}$ reference speakers
  – Recognizer contains $\mathbf{P}$ phonetic models
  – A mean is trained for each model $\mathbf{p}$ and each speaker $\mathbf{r}$: $\vec{\mu}_{\mathbf{p,r}}$
  – A matrix of *speaker vectors* is created from trained means:

$$\vec{\mathbf{m}}_{\mathbf{r}} = \begin{bmatrix} \vec{\mu}_{\mathbf{1,r}} \\ \vdots \\ \vec{\mu}_{\mathbf{P,r}} \end{bmatrix} \qquad \mathbf{M} = \begin{bmatrix} \vec{\mu}_{\mathbf{1,1}} & \cdots & \vec{\mu}_{\mathbf{1,R}} \\ \vdots & \ddots & \vdots \\ \vec{\mu}_{\mathbf{P,1}} & \cdots & \vec{\mu}_{\mathbf{P,R}} \end{bmatrix}$$

speaker vector

speaker matrix

each column is
a speaker vector

Reference Speaker Weighting (cont)

- **Goal is to find most likely speaker vector for new speaker**
- **Find weighted combination of reference speaker vectors:**

$$\vec{\mathbf{m}}_{\mathbf{sa}} = \mathbf{M}\vec{\mathbf{w}}$$

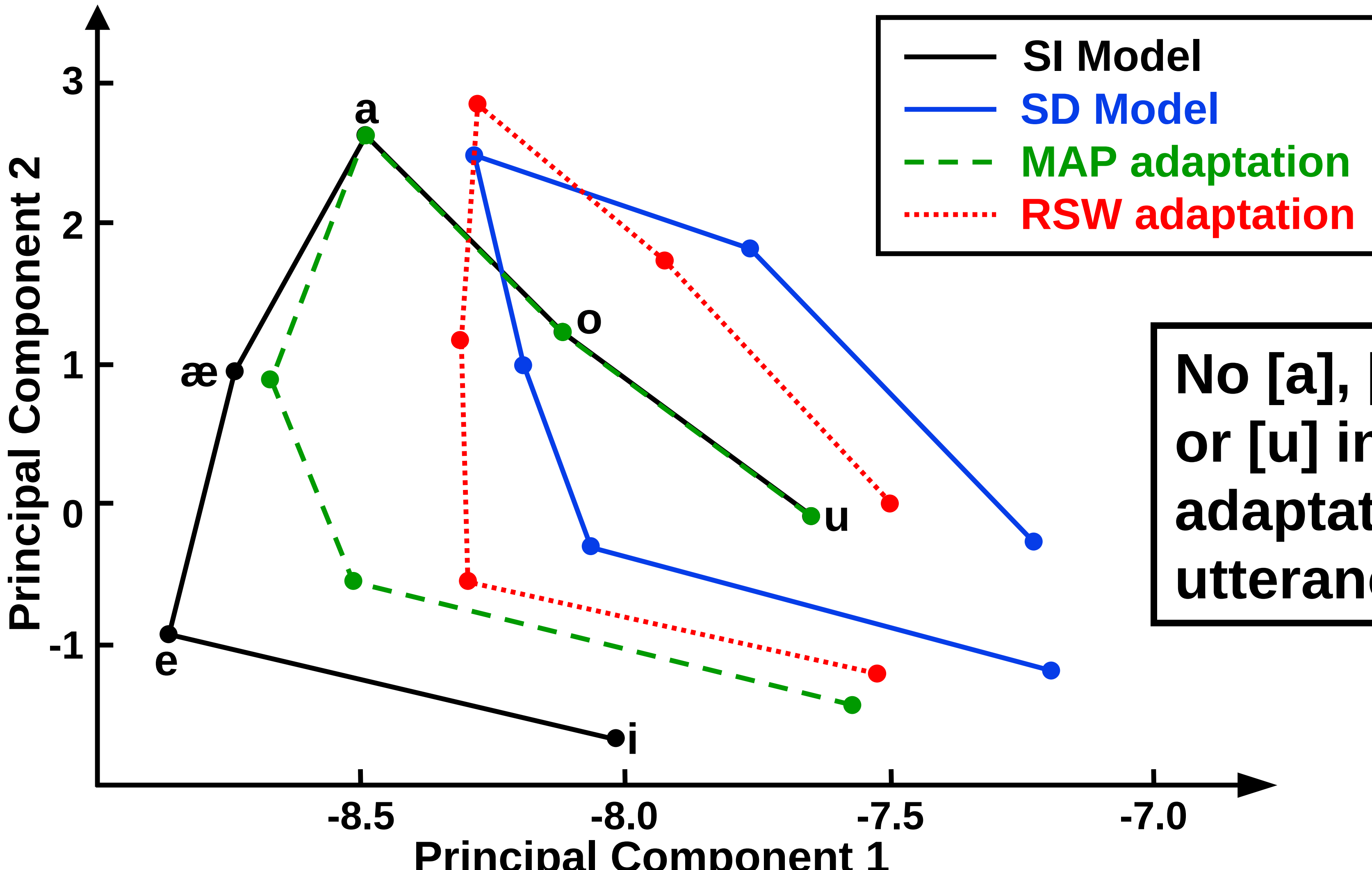
- **Maximum likelihood estimation of weighting vector:**

$$\vec{\mathbf{w}} = \arg\max_{\vec{\mathbf{w}}} \mathbf{p}(\mathbf{X} \mid \mathbf{M}, \vec{\mathbf{w}})$$

- **Global weighting vector is robust to errors introduced during unsupervised adaptation**
- **Iterative methods can be used to find the weighting vector**
  - **Reference: Hazen, 1998**

Reference Speaker Weighting (cont)

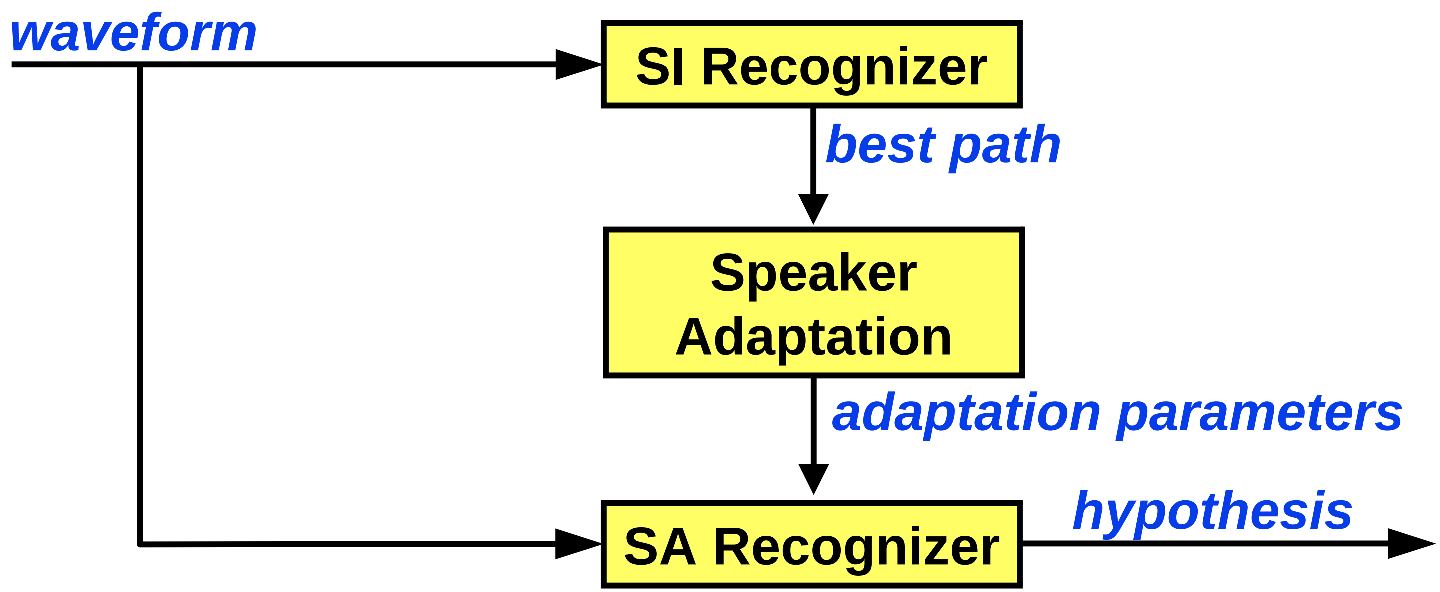
- **Mean vector adaptation w/ one adaptation utterance:**



No [a], [o], or [u] in adaptation utterance

Unsupervised Adaptation Architecture

- **Architecture of unsupervised adaptation system:**



- **In off-line mode, adapted models used to re-recognize original waveform**
  - **Sometimes called instantaneous adaptation**
- **In on-line mode, SA models used on next waveform**

Unsupervised Adaptation Experiment

- **Unsupervised, instantaneous adaptation**
  - Adapt and test on same utterance
  - Unsupervised $\Rightarrow$ recognition errors affect adaptation
  - Instantaneous $\Rightarrow$ recognition errors are reinforced

Adaptation Method	WER	Reduction
SI	8.6%	---
MAP Adaptation	8.5%	0.8%
RSW Adaptation	8.0%	6.5%

- **RSW is more robust to errors than MAP**
  - RSW estimation is "global" $\Rightarrow$ uses whole utterance
  - MAP estimation is "local" $\Rightarrow$ uses one phonetic class only

Eigenvoices

- Eigenvoices extends ideas of Reference Speaker Weighting
  - Reference: Kuhn, 2000
- Goal is to learn uncorrelated features of the speaker space
- Begin by creating speaker matrix:

$$\mathbf{M} = \begin{bmatrix} \vec{\mu}_{1,1} & \cdots & \vec{\mu}_{1,R} \\ \vdots & \ddots & \vdots \\ \vec{\mu}_{P,1} & \cdots & \vec{\mu}_{P,R} \end{bmatrix}$$

- Perform Eigen (principal components) analysis on $\mathbf{M}$
  - Each Eigenvector represents an independent (orthogonal) dimension in the speaker space
  - Example dimensions this method typically learns are gender, loudness, monotonicity, etc.

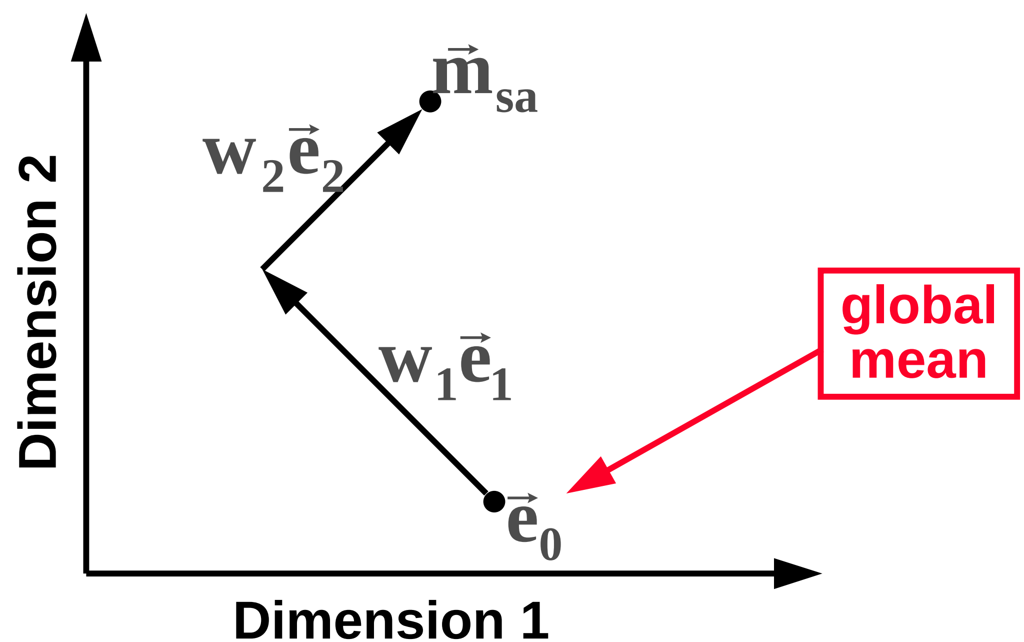
Eigenvoices (cont)

- **Find R eigenvectors:**

$$E = \left\{ \vec{e}_0 ; \vec{e}_1 ; \cdots ; \vec{e}_R \right\}$$

- **New speaker vector is combination of top N eigenvectors:**

$$\vec{m}_{sa} = \vec{e}_0 + w_1 \vec{e}_1 + \cdots + w_N \vec{e}_N$$



Eigenvoices (cont)

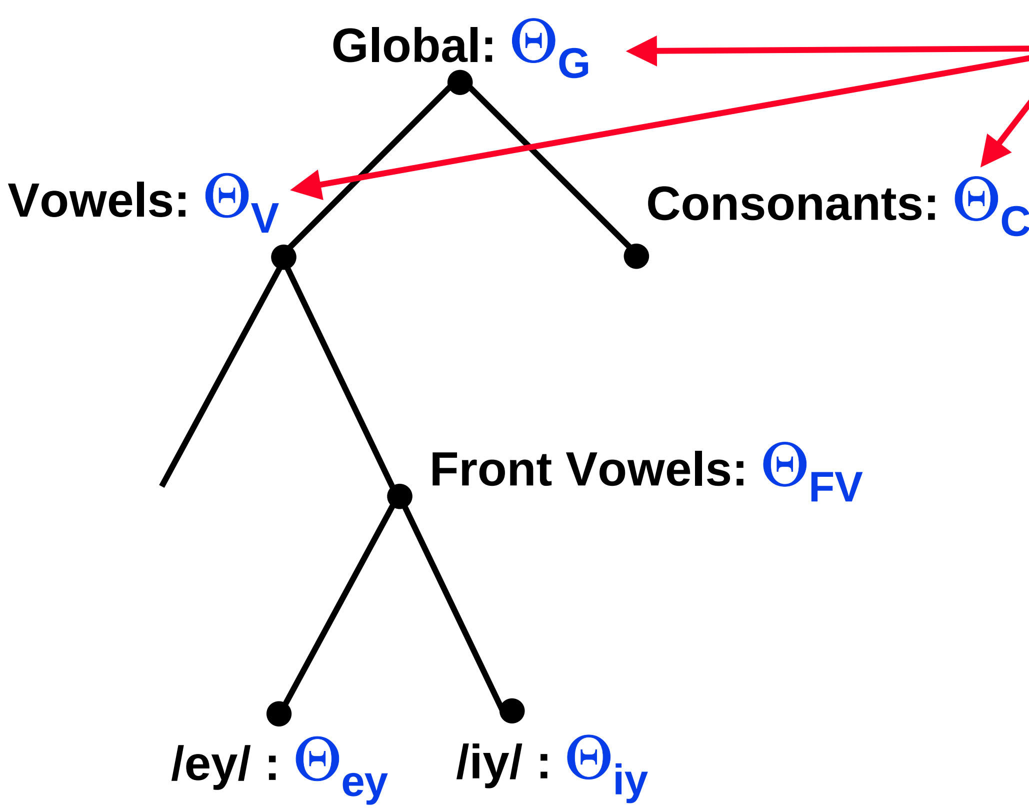
- **Adaptation procedure is very similar to RSW:**

$$\vec{w} = \arg\max_{\vec{w}} p(X \mid E, \vec{w})$$

- **Eigenvoices adaptation can be very fast**
  - **A few eigenvectors can generalize to many speaker types**
  - **Only a small number of phonetic observations required to achieve significant gains**

Structural Adaptation

- Adaptation parameters organized in tree structure
  - Root node is global adaptation
  - Branch nodes perform adaptation on shared classes of models
  - Leaf nodes perform model specific adaptation



- Adaptation parameters learned for each node in tree
- Each node has a weight: w_{node}
- Weights based on availability of adaptation data
- Each path from root to leaf follows this constraint:

$$\sum_{\forall node \in path} w_{node} = 1$$

Structural Adaptation

- **Structural adaptation based on weighted combination of adaptation performed at each node in tree:**

$$p_{sa}(\vec{x} \mid u, tree) = \sum_{\forall nodes \in path(u)} w_{node}\, p(\vec{x} \mid u, \Theta_{node})$$

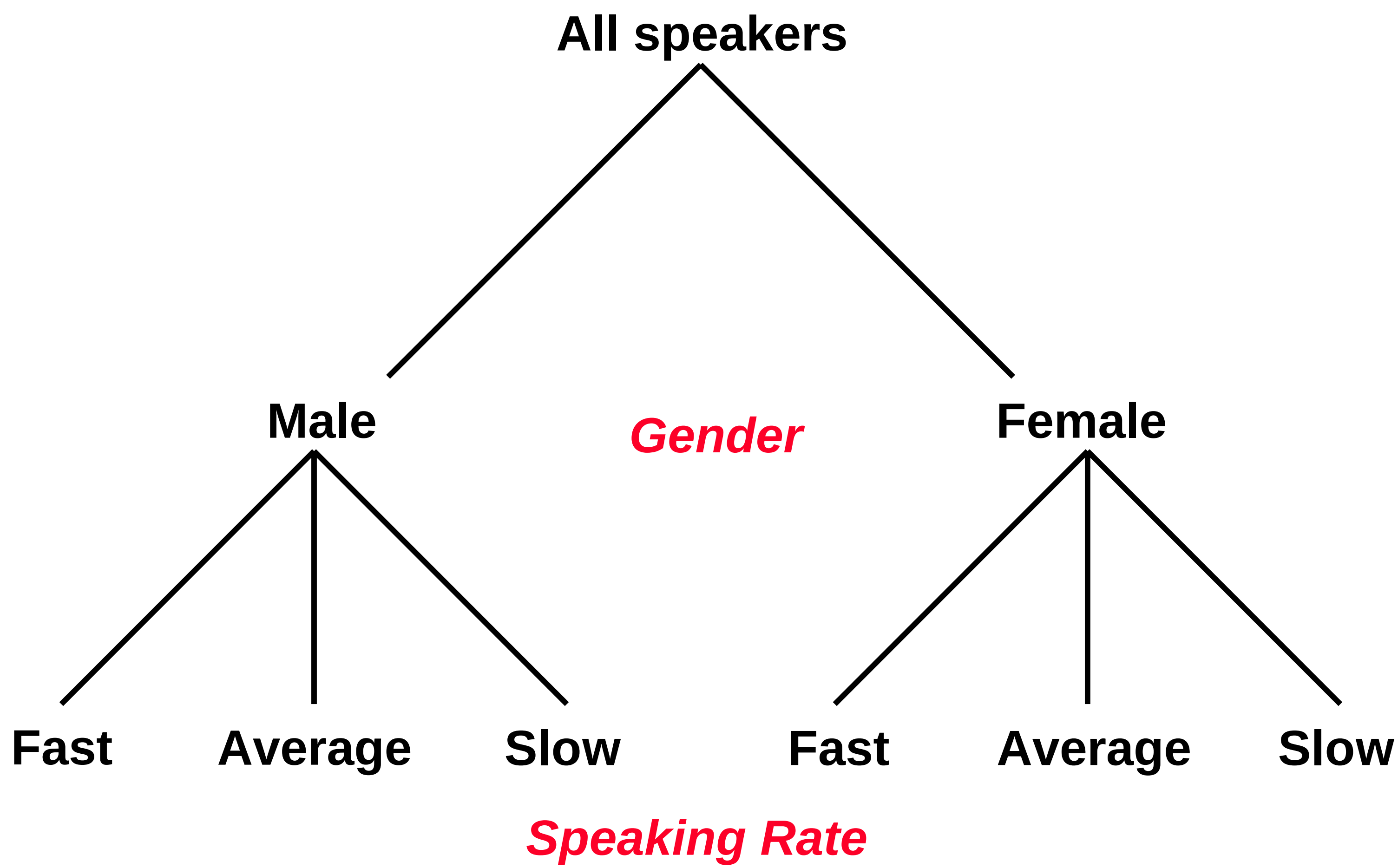
- **Structural adaptation has been applied to a variety of speaker adaptation techniques:**
  – **MAP** (Reference: Shinoda & Lee,1998)
  – **RSW** (Reference: Hazen, 1998)
  – **Eigenvoices** (Reference: Zhou & Hanson, 2001)
  – **MLLR** (Reference: Siohan, Myrvoll & Lee, 2002)

Hierarchical Speaker Clustering

- **Idea: Use model trained from cluster of speakers most similar to the current speaker**
- **Approach:**
  - **A hierarchical tree is created using speakers in training set**
  - **The tree separates speakers into similar classes**
  - **Different models build for each node in the tree**
  - **A test speaker is compared to all nodes in tree**
  - **The model of the best matching node is used during recognition**
- **Speakers can be clustered…**
  - **…manually based on predefined speaker properties**
  - **…automatically based on acoustic similarity**
- **References:**
  - **Furui, 1989**
  - **Kosaka and Sagayama, 1994**

Hierarchical Speaker Clustering

- **Example of manually created speaker hierarchy:**



Hierarchical Speaker Clustering (cont)

- **Problem: More specific model $\Rightarrow$ less training data**
- **Tradeoff between robustness and specificity**
- **One solution: interpolate general and specific models**
- **Example combining ML trained gender dependent model with SI model to get interpolated gender dependent model:**

$$\mathbf{p}_{\mathbf{igd}}(\vec{\mathbf{x}}_{\mathbf{n}} \mid \mathbf{u} = \mathbf{p}) = \lambda \mathbf{p}_{\mathbf{mlgd}}(\vec{\mathbf{x}}_{\mathbf{n}} \mid \mathbf{u} = \mathbf{p}) + (1 - \lambda)\mathbf{p}_{\mathbf{si}}(\vec{\mathbf{x}}_{\mathbf{n}} \mid \mathbf{u} = \mathbf{p})$$

- **λ values found using the deleted interpolation**
  - **Reference: X.D. Huang, *et al*, 1996**

Speaker Cluster Weighting

- **Hierarchical speaker clustering chooses one model**
- **Speaker cluster weighting combines models:**

$$p_{sa}(\vec{x}_n \mid u = p) = \sum_{m=1}^{M} w_m p_m(\vec{x}_n \mid u = p)$$

- **Weights determined using EM algorithm**
- **Weights can be global or class-based**
- **Advantage: *Soft* decisions less rigid than *hard* decisions**
  - **Reference: Hazen, 2000**
- **Disadvantage:**
  - **Model size could get too large w/ many clusters**
  - **Need approximation methods for real-time**
  - **Reference: Huo, 2000**

Speaker Clustering Experiment

- **Unsupervised instantaneous adaptation experiment**
  - **Resource Management SI test set**
- **Speaker cluster models used for adaptation:**
  - **1 SI model**
  - **2 gender dependent models**
  - **6 gender and speaking rate dependent models**

Models	WER	Reduction
SI	8.6%	---
Gender Dependent	7.7%	10.5%
Gender & Rate Dependent	7.2%	16.4%
Speaker Cluster Interpolation	6.9%	18.9%

Final Words

- **Adaptation improves recognition by constraining models to characteristics of current speaker**
- **Good properties of adaptation algorithms:**
  - **account for a priori knowledge about speakers**
  - **be able to adapt models of units which are not observed**
  - **adjust number of adaptation parameters to amount of data**
  - **be robust to errors during unsupervised adaptation**
- **Adaptation is important for “real world” applications**

References

- A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," *CAIP Workshop: Frontiers in Speech Recognition II*, 1994.
- S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering," ICASSP, 1989.
- J. Gauvain and C. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Trans. On Speech and Audio Processing,* April 1994.
- T. Hazen, *The use of speaker correlation information for automatic speech recognition*, PhD Thesis, MIT, January 1998.
- T. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communication*, May 2000.
- X.D. Huang, *et al*, "Deleted interpolation and density sharing for continuous hidden Markov models," ICASSP 1996.
- Q. Huo and B. Ma, "Robust speech recognition based on off-line elicitation of multiple priors and on-line adaptive prior fusion," ICSLP, 2000.

References

- **T. Kosaka and S. Sagayama, "Tree structured speaker clustering for speaker-independent continuous speech recognition," ICASSP, 1994.**
- **R. Kuhn, *et al*, "Rapid speaker adaptation in Eigenvoice Space," *IEEE Trans. on Speech and Audio Processing*, November 2000.**
- **L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. On Speech and Audio Proc.*, January 1998.**
- **C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, April 1995.**
- **K. Shinoda and C. Lee, "Unsupervised adaptation using a structural Bayes approach,", ICASSP, 1998.**
- **O. Siohan, T. Myrvoll and C. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language*, January 2002.**
- **B. Zhou and J. Hanson, "A novel algorithm for rapid speaker adaptation based on structural maximum likelihood Eigenspace mapping," Eurospeech, 2001.**