

Probabilistic Feature Transformation for Channel Robust Speaker Verification

Man Wai MAK and Kwok Kwong YIU
Dept. of Electronic and Information
Engineering,
The Hong Kong Polytechnic University

Outline

- What is Speaker Verification
- Challenge in Speaker Verification and How to Deal With it
- Feature Transformation
 - Blind Methods
 - Non-Blind Methods
- Evaluations on NIST2000 Dataset
- Results and Conclusion

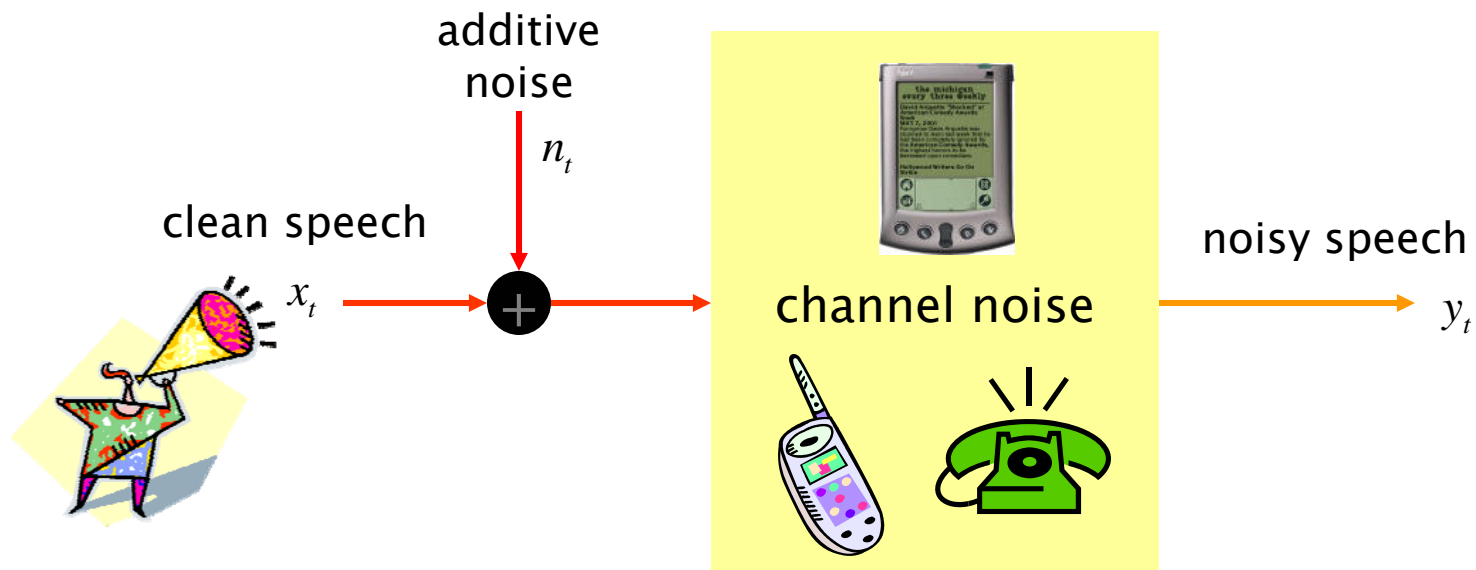
What is Speaker Verification?

- To verify the identify of a claimant based on his/her own voices

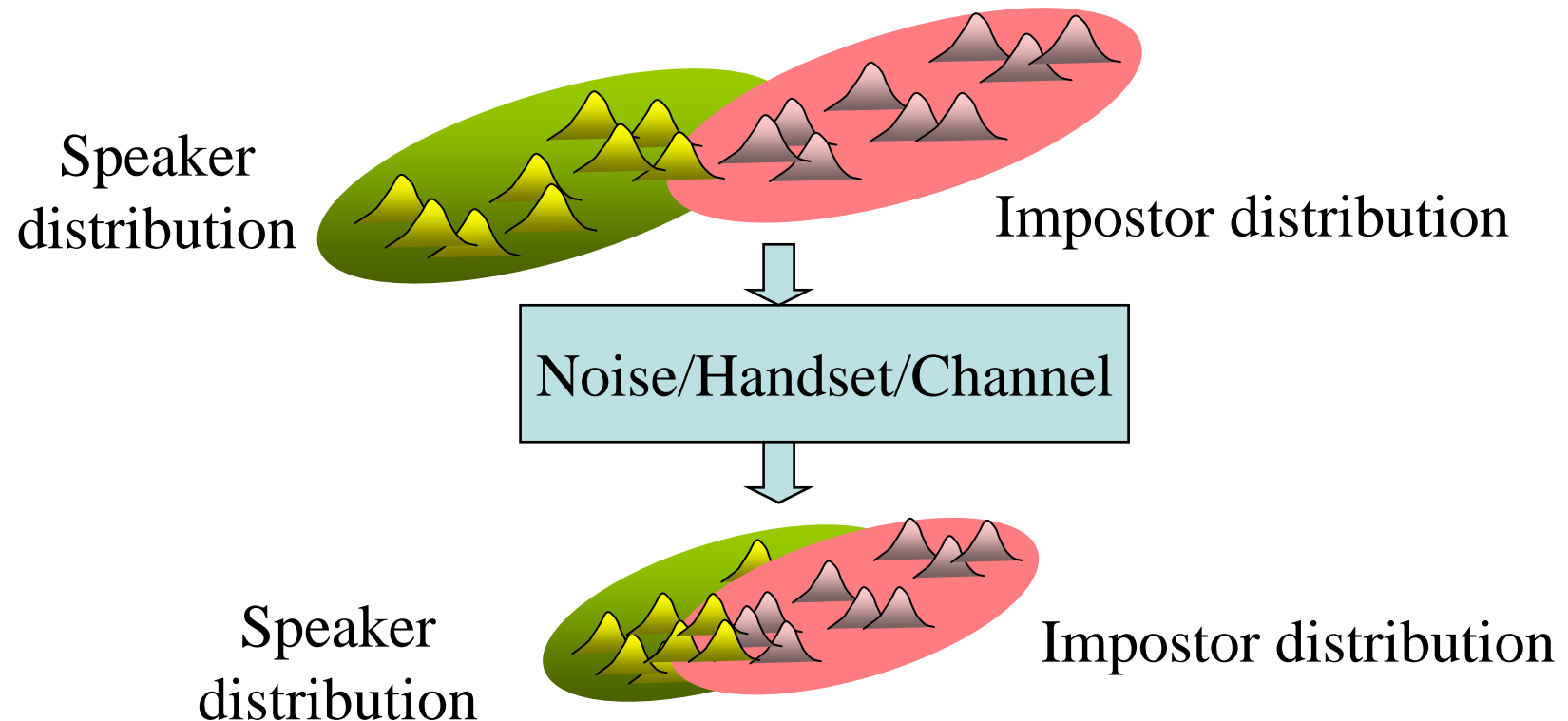


Challenges in Speaker Verification

- Handle distortion of speech signal due to background noise
- Handle distortion of speech signal due to handsets / transducers



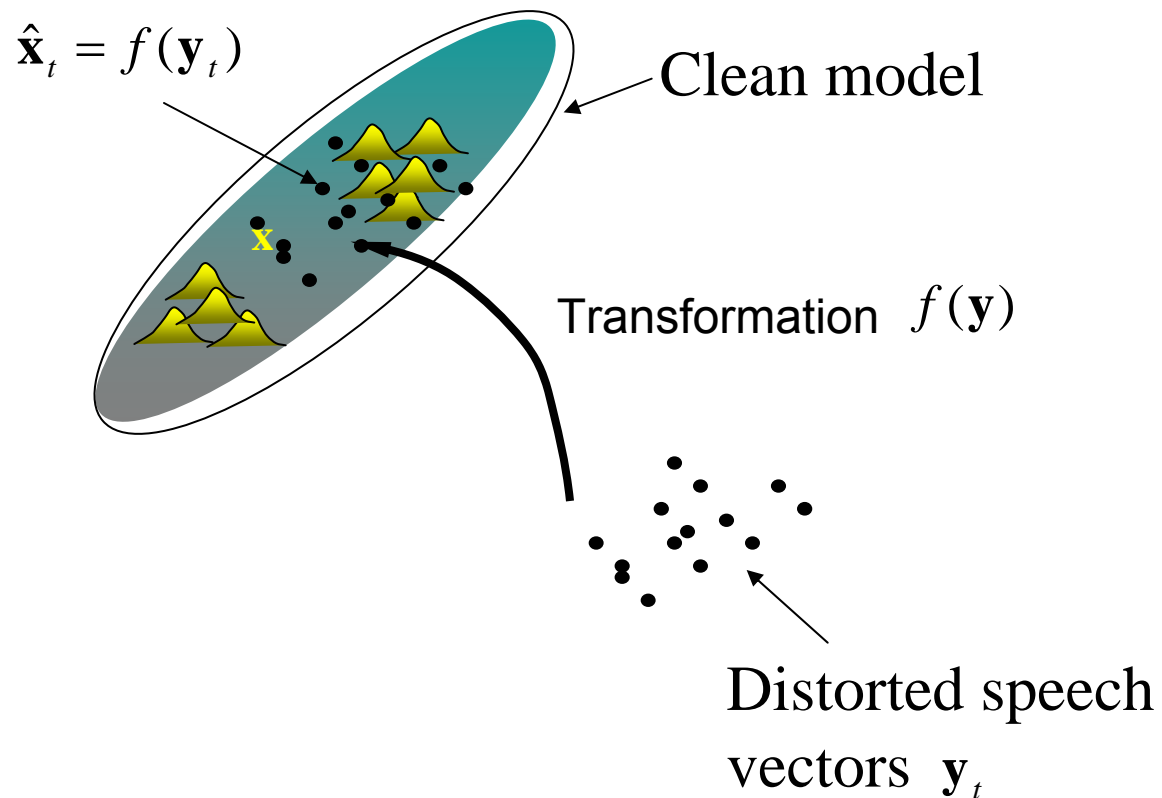
Effect of Noise/Channel on Speech Features



Problem: Noise and channel effect will make the speech of the true speaker and impostors less discriminative.

How do We Deal with the Challenge?

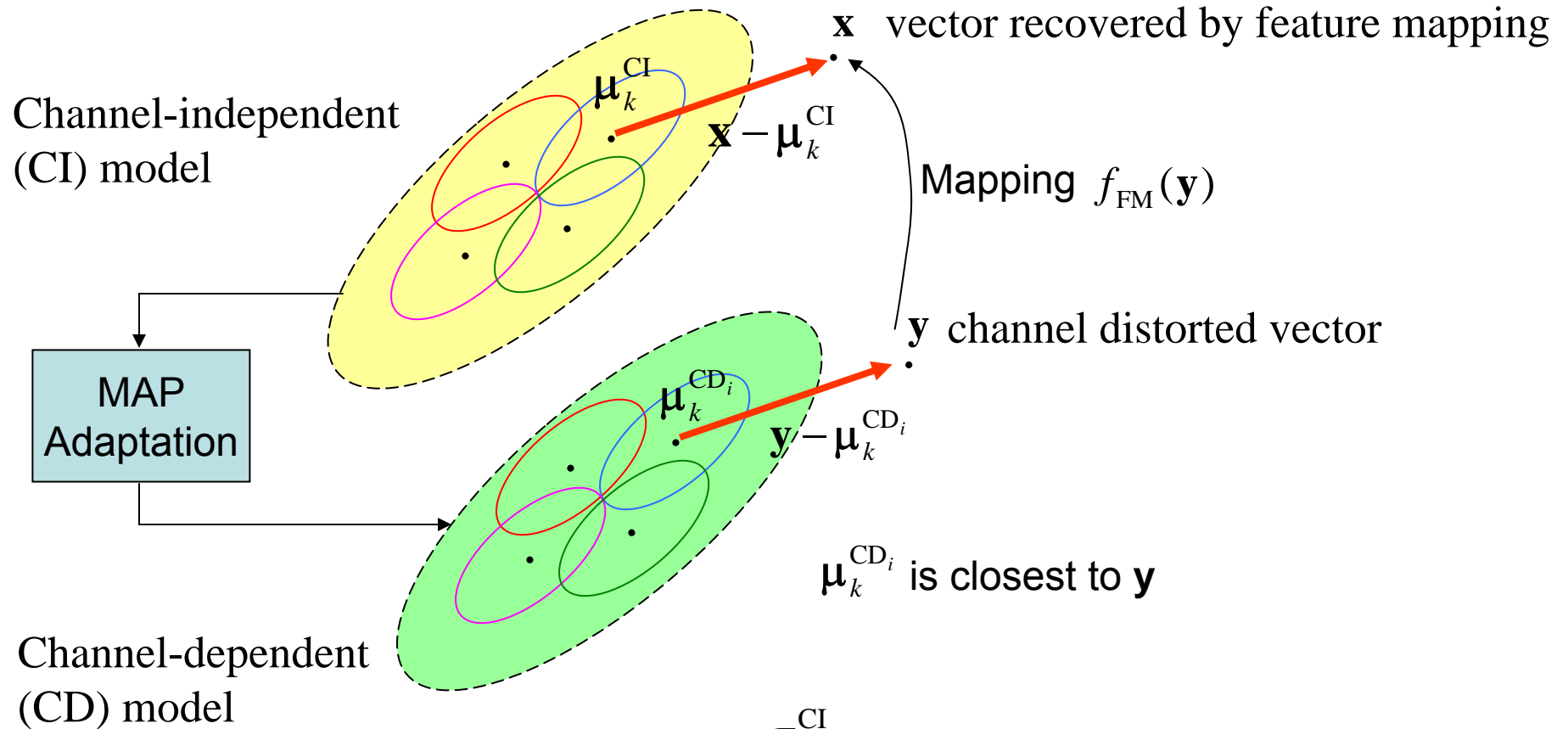
- **Feature transformation** aims to reduce the effects of channel- and handset-distortion by transforming distorted features to bring them closer to the clean speech models.



Type of Feature Transformation

- Non-blind compensation – based on a priori knowledge of the all possible channels
 - Feature Mapping (FM), by Reynolds (2003)
 - Probabilistic feature mapping (PFM), proposed in this paper
 - Fast probabilistic feature mapping (fPFM), proposed in this paper
 - Stochastic feature transformation (SFT), by Mak & Kung (2002)
- Blind compensation – without a priori knowledge of the channel characteristics
 - Cepstral Mean Subtraction (CMS), by Atal (1974)
 - Blind stochastic feature transformation (BSFT), by Yiu, Mak and Kung (2004, 2005)
 - Fast BSFT (fBSFT), proposed in this paper

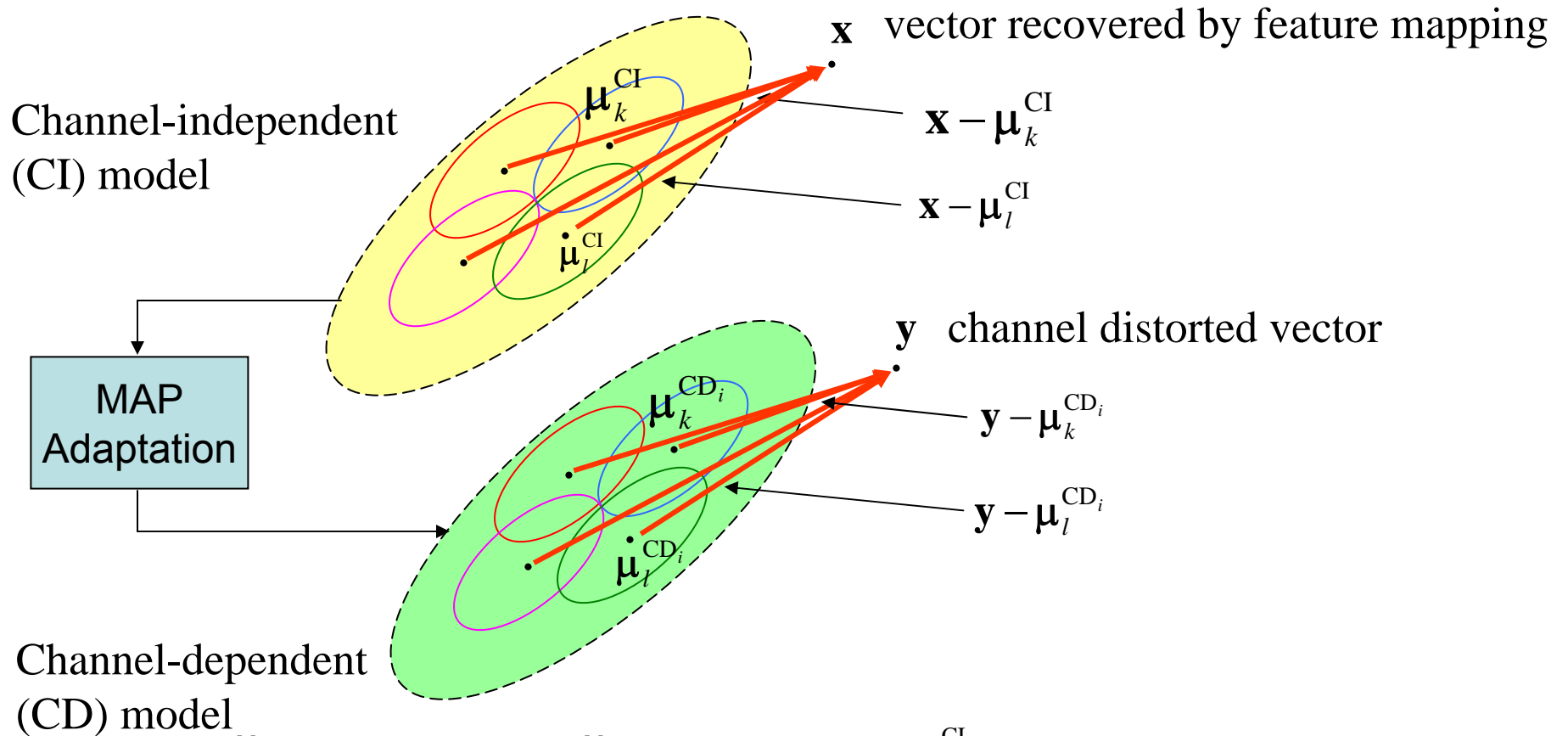
Feature Mapping (Non-Blind)



$$\mathbf{x} - \mu_k^{CI} = (\mathbf{y} - \mu_k^{CD_i}) \frac{\sigma_k^{CI}}{\sigma_k^{CD_i}} \leftarrow \text{account for the difference in variance}$$

$$\Rightarrow \mathbf{x} = f_{FM}(\mathbf{y}) = (\mathbf{y} - \mu_k^{CD_i}) \frac{\sigma_k^{CI}}{\sigma_k^{CD_i}} + \mu_k^{CI}$$

Probabilistic Feature Mapping (Non-Blind)

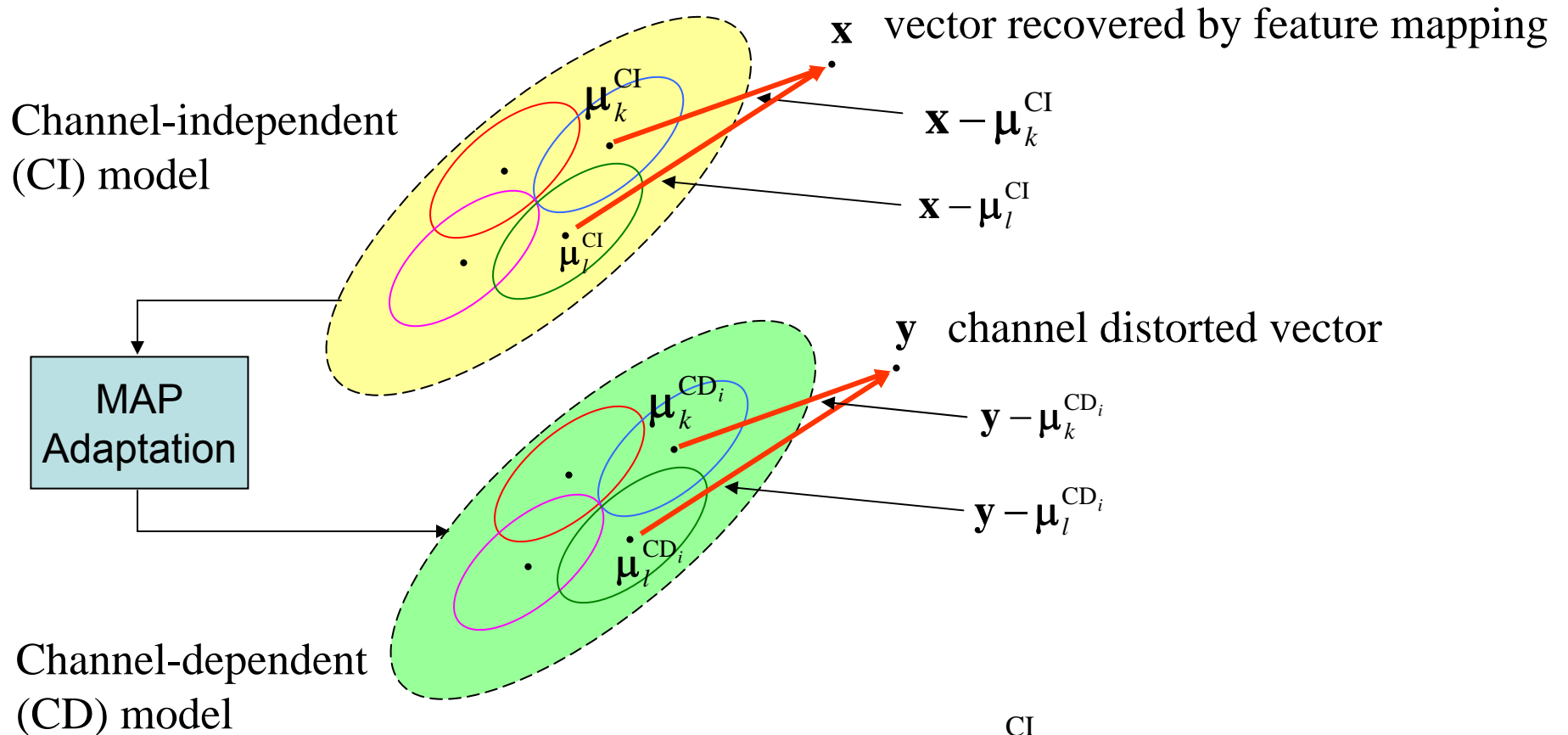


$$\sum_{j=1}^M P(j | \mathbf{y}) (\mathbf{x} - \mu_j^{\text{CI}}) = \sum_{j=1}^M P(j | \mathbf{y}) (\mathbf{y} - \mu_j^{\text{CD}_i}) \frac{\sigma_j^{\text{CI}}}{\sigma_j^{\text{CD}_i}}$$

$$\Rightarrow \mathbf{x} = \sum_{j=1}^M P(j | \mathbf{y}) \left[(\mathbf{y} - \mu_j^{\text{CD}_i}) \frac{\sigma_j^{\text{CI}}}{\sigma_j^{\text{CD}_i}} + \mu_j^{\text{CI}} \right]$$

$$P(j | \mathbf{y}) = \frac{\pi_j^{\text{CD}_i} p(\mathbf{y} | \mu_j^{\text{CD}_i}, \Sigma_j^{\text{CD}_i})}{\sum_{l=1}^M \pi_l^{\text{CD}_i} p(\mathbf{y} | \mu_l^{\text{CD}_i}, \Sigma_l^{\text{CD}_i})}$$

Fast Probabilistic FM (Non-Blind)



$$\sum_{j \in C} P(j | \mathbf{y}) (\mathbf{x} - \mu_j^{\text{CI}}) = \sum_{j \in C} P(j | \mathbf{y}) (\mathbf{y} - \mu_j^{\text{CD}_i}) \frac{\sigma_j^{\text{CI}}}{\sigma_j^{\text{CD}_i}}$$

$$\Rightarrow \mathbf{x} = \sum_{j \in C} P(j | \mathbf{y}) \left[(\mathbf{y} - \mu_j^{\text{CD}_i}) \frac{\sigma_j^{\text{CI}}}{\sigma_j^{\text{CD}_i}} + \mu_j^{\text{CI}} \right] \quad P(j | \mathbf{y}) = \frac{\pi_j^{\text{CD}_i} p(\mathbf{y} | \mu_j^{\text{CD}_i}, \Sigma_j^{\text{CD}_i})}{\sum_{l=1}^M \pi_l^{\text{CD}_i} p(\mathbf{y} | \mu_l^{\text{CD}_i}, \Sigma_l^{\text{CD}_i})}$$

Different Type of Feature Mapping

Feature Mapping

$$\mathbf{x} = \left(\mathbf{y} - \boldsymbol{\mu}_k^{\text{CD}_i} \right) \frac{\boldsymbol{\sigma}_k^{\text{CI}}}{\boldsymbol{\sigma}_k^{\text{CD}_i}} + \boldsymbol{\mu}_k^{\text{CI}}$$

$$k = \arg \max_{j=1}^M \pi_j^{\text{CD}_i} p(\mathbf{y} | \boldsymbol{\mu}_j^{\text{CD}_i}, \boldsymbol{\Sigma}_j^{\text{CD}_i})$$

Probabilistic Feature Mapping

$$\Rightarrow \mathbf{x} = \sum_{j=1}^M P(j | \mathbf{y}) \left[\left(\mathbf{y} - \boldsymbol{\mu}_j^{\text{CD}_i} \right) \frac{\boldsymbol{\sigma}_j^{\text{CI}}}{\boldsymbol{\sigma}_j^{\text{CD}_i}} + \boldsymbol{\mu}_j^{\text{CI}} \right] \quad P(j | \mathbf{y}) = \frac{\pi_j^{\text{CD}_i} p(\mathbf{y} | \boldsymbol{\mu}_j^{\text{CD}_i}, \boldsymbol{\Sigma}_j^{\text{CD}_i})}{\sum_{l=1}^M \pi_l^{\text{CD}_i} p(\mathbf{y} | \boldsymbol{\mu}_l^{\text{CD}_i}, \boldsymbol{\Sigma}_l^{\text{CD}_i})}$$

Fast Probabilistic Feature Mapping

$$\Rightarrow \mathbf{x} = \sum_{j \in C} P(j | \mathbf{y}) \left[\left(\mathbf{y} - \boldsymbol{\mu}_j^{\text{CD}_i} \right) \frac{\boldsymbol{\sigma}_j^{\text{CI}}}{\boldsymbol{\sigma}_j^{\text{CD}_i}} + \boldsymbol{\mu}_j^{\text{CI}} \right]$$

where C contains the indexes of top- C Gaussians

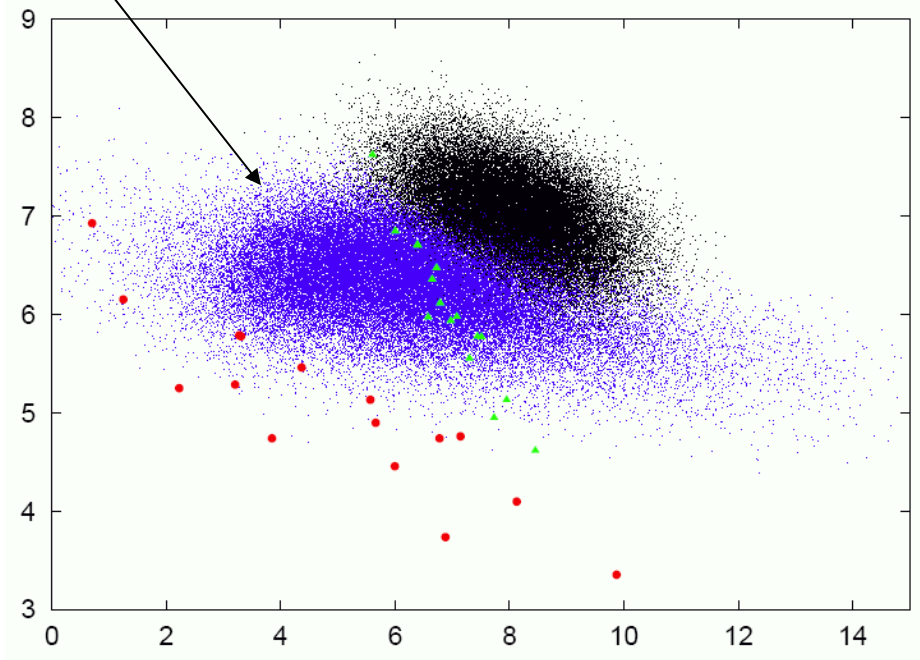
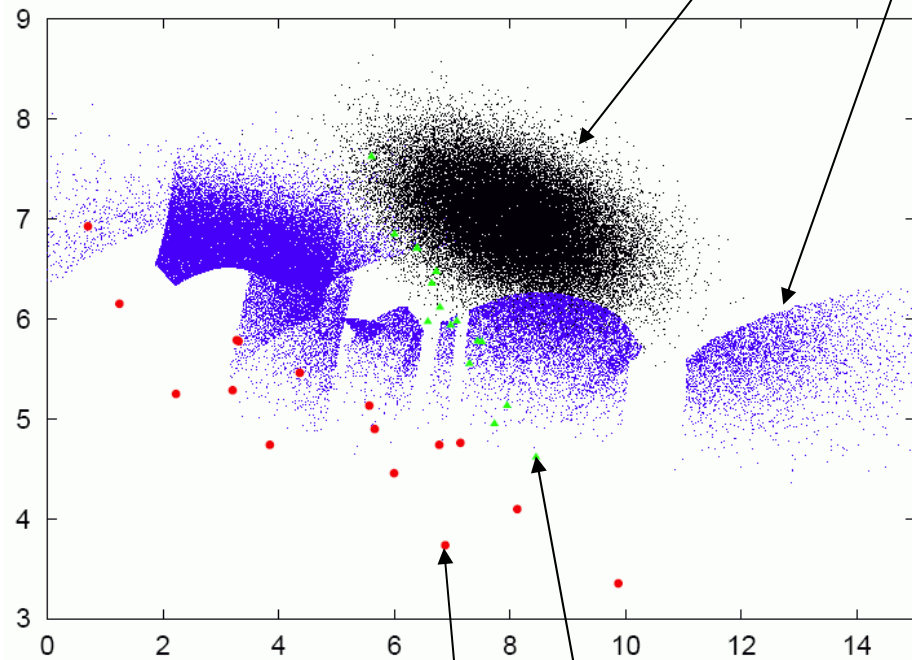
Clustering Effect of Feature Mapping

y : data from channel-dependent source

x : transformed features

Feature Mapping

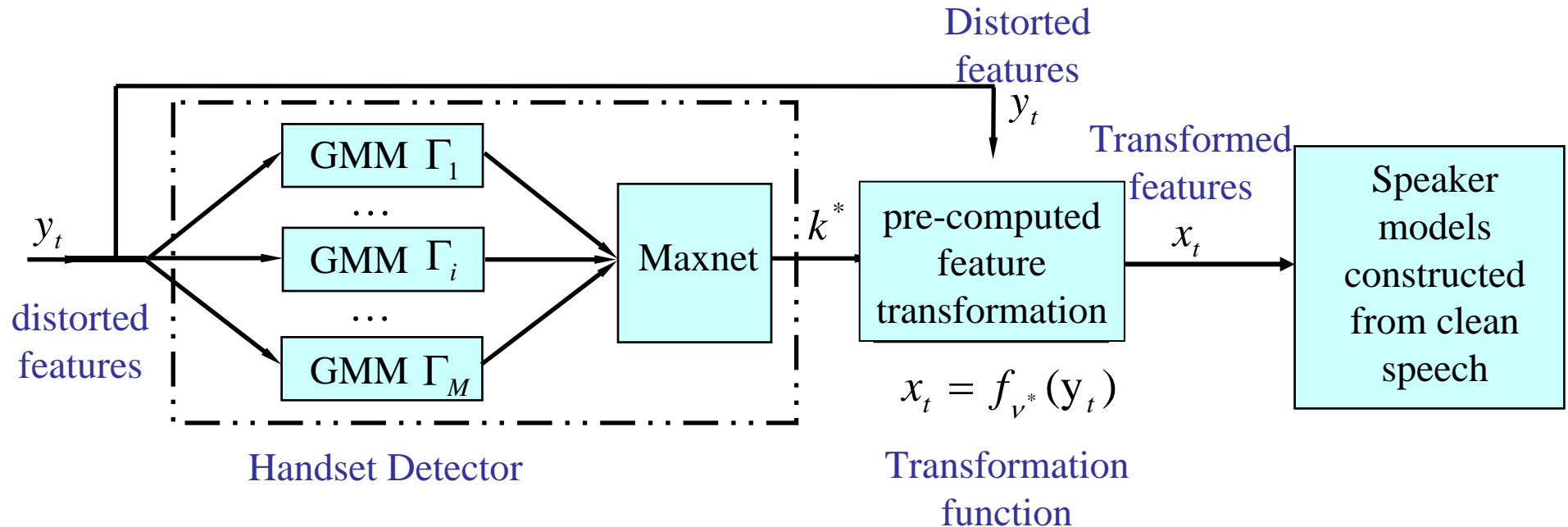
Probabilistic Feature Mapping



$\mu_k^{CD_i}$: Centers of channel-dependent model
 μ_k^{CI} : Centers of channel-independent root model

Stochastic Feature Transformation (Non-Blind)

Aim: To transform distorted features to fit the clean speech models using handset detection

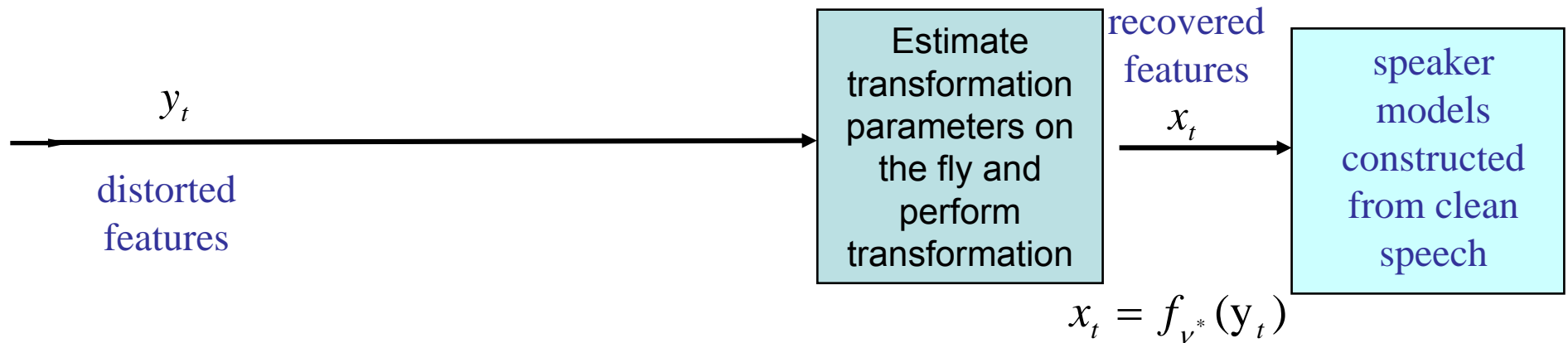


$$k^* = \arg \max_{k=1}^H \sum_{t=1}^T \log p(y_t | \Gamma_k)$$

Problem: The handset detector may not work for “unseen” handsets

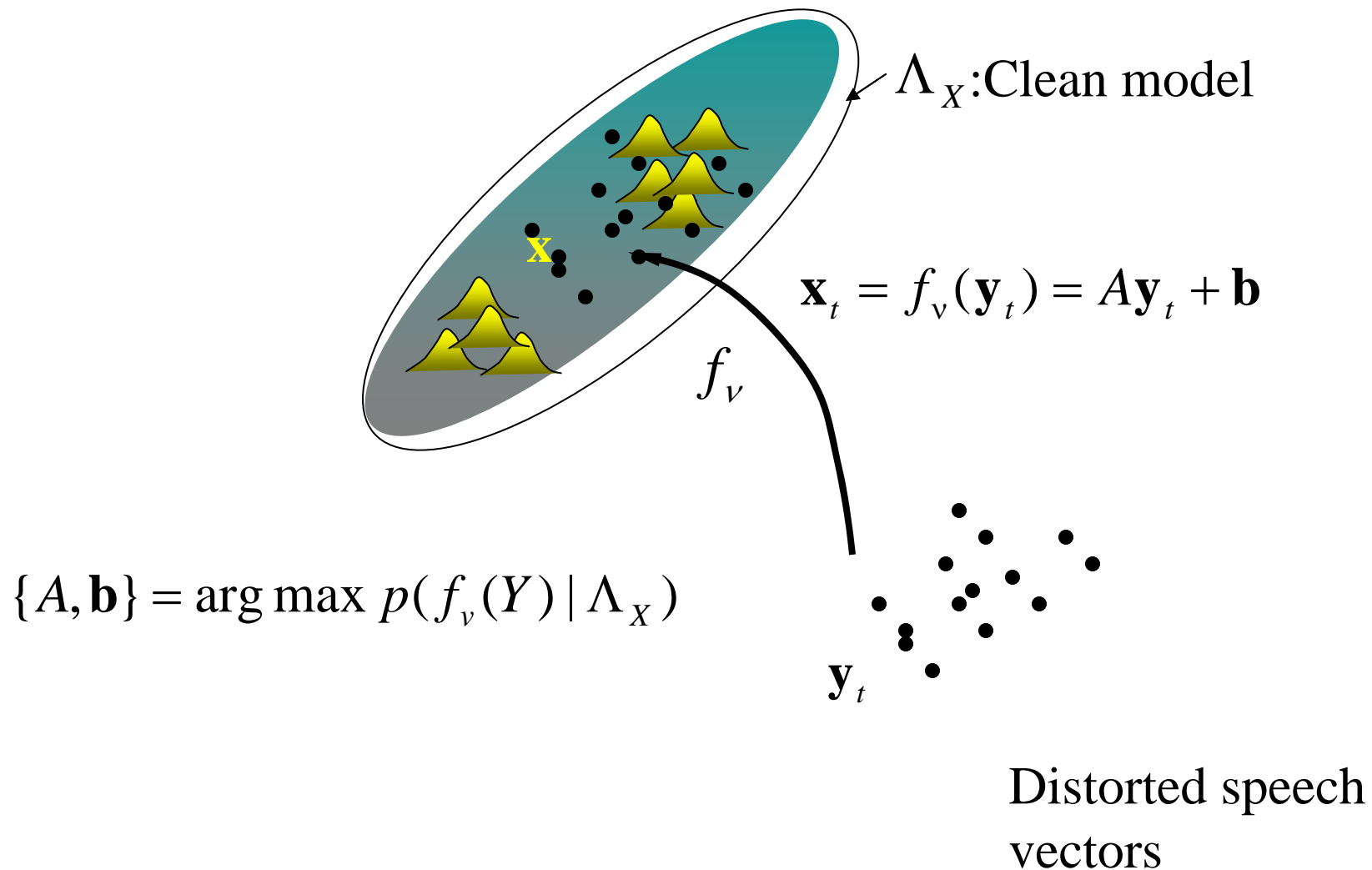
Blind Stochastic Feature Transformation

Aim: To transform distorted features to fit the clean speech models **without** handset detection



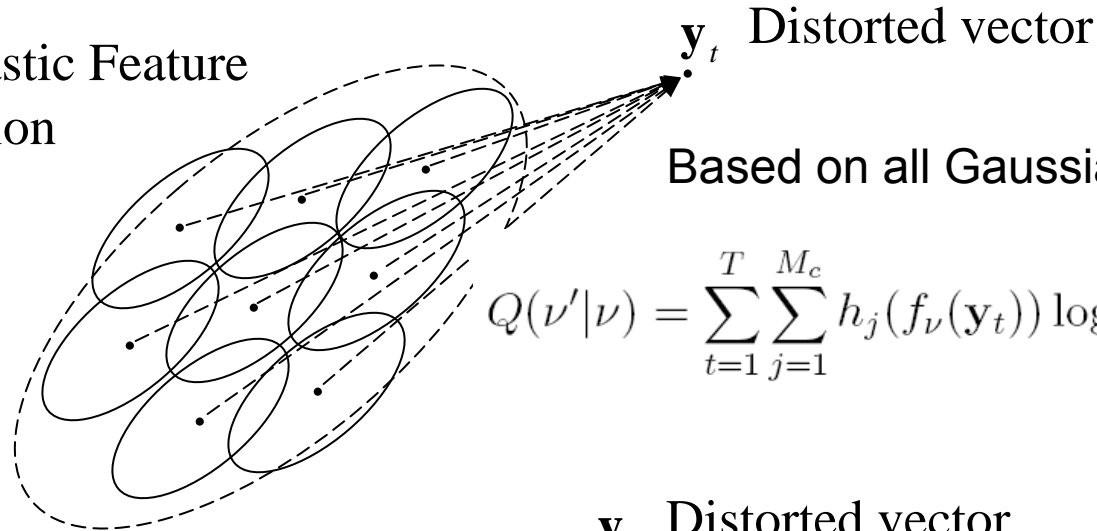
Blind Stochastic Feature Transformation

- Distorted speech vectors can be transformed to fit the clean speech model.



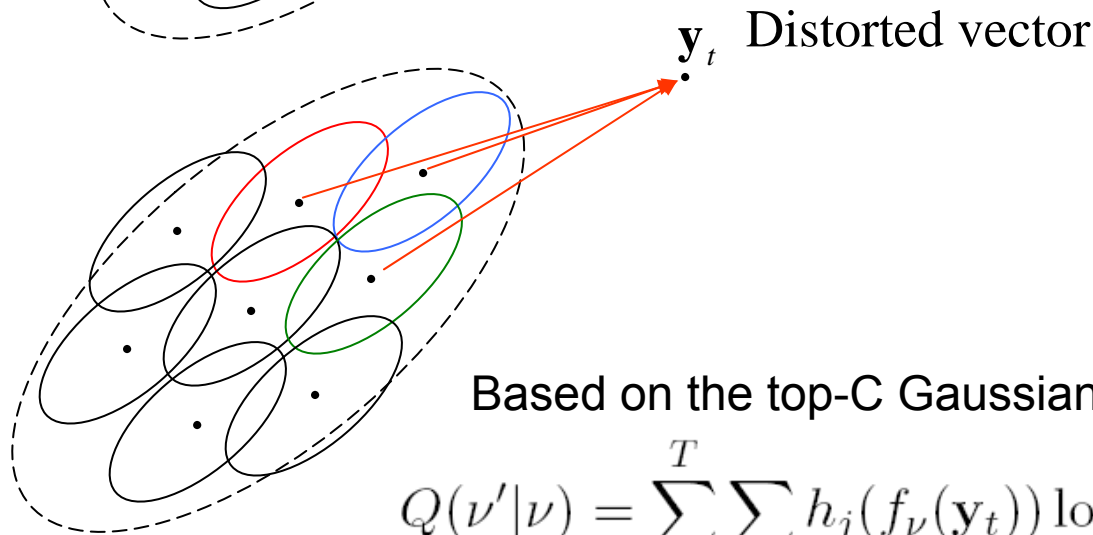
Fast Blind Stochastic Feature Transformation

Blind Stochastic Feature Transformation



Based on all Gaussians:

$$Q(\nu'|\nu) = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_\nu(\mathbf{y}_t)) \log \left\{ \frac{p(f_{\nu'}(\mathbf{y}_t)|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{|J_{\nu'}(\mathbf{y}_t)|} \right\}$$



Based on the top-C Gaussians only:

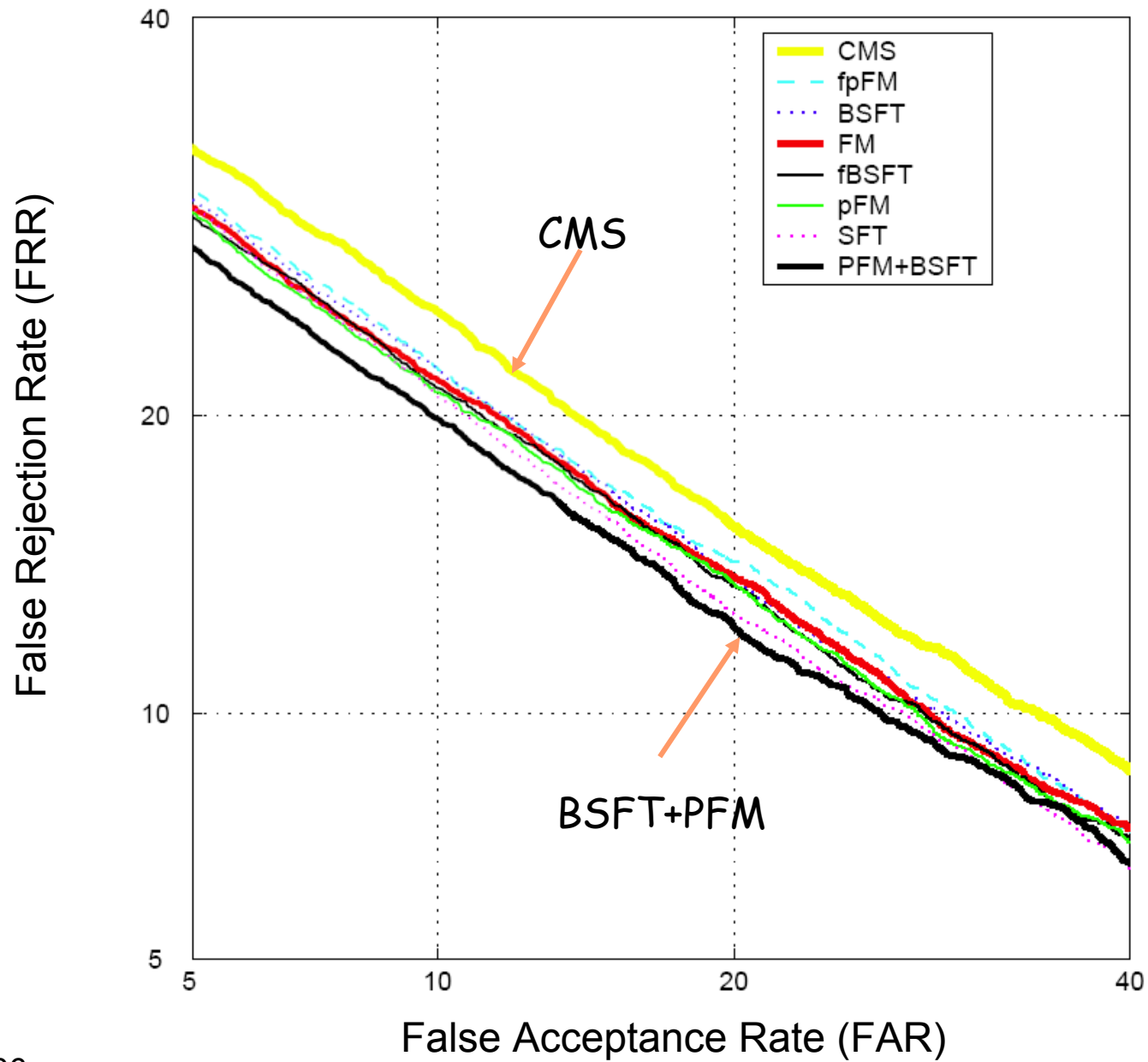
$$Q(\nu'|\nu) = \sum_{t=1}^T \sum_{j \in \mathcal{C}} h_j(f_\nu(\mathbf{y}_t)) \log \left\{ \frac{p(f_{\nu'}(\mathbf{y}_t)|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{|J_{\nu'}(\mathbf{y}_t)|} \right\}$$

Fast Blind Stochastic Feature Transformation

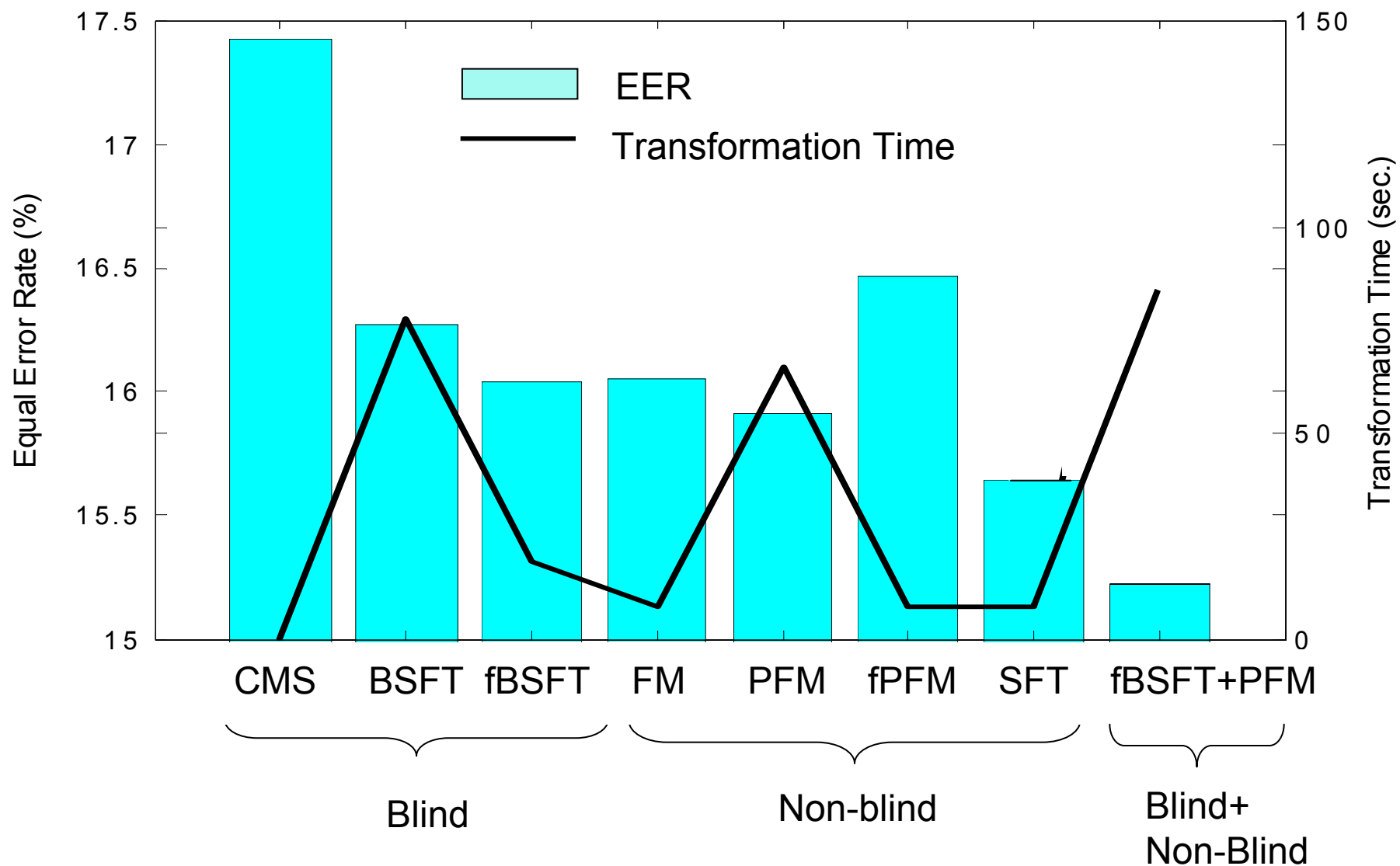
Experiments

- 2000 NIST speaker recognition evaluation set.
- 1003 target speakers (457 male and 546 female).
- Enrollment: approximately 1 minutes of speech.
- Verification: 6052 utterances (3026 male and 3026 female).
- Each verification utterance is evaluated against 11 hypothesized gender-matched speakers.
- For each gender, gender-dependent evaluation utterances from NIST99 were used to train a 1024-component gender-dependent universal background models (UBMs).
- Speaker and background models:
 - MFCC + Δ MFCC
 - Speaker models were adapted from the gender-dependent background model using MAP adaptation

Results: DET plots



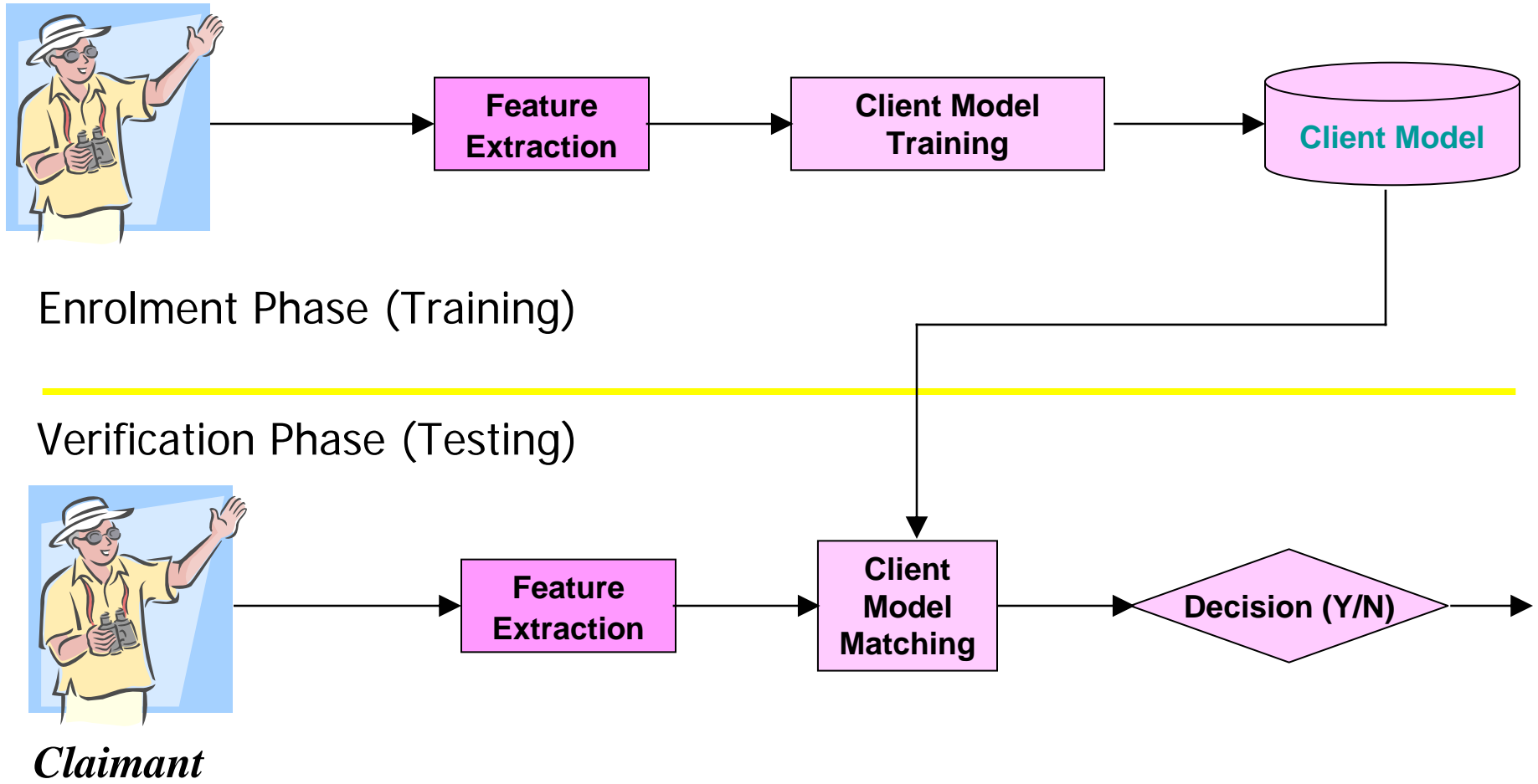
Results: EER & Transformation Time



Conclusions

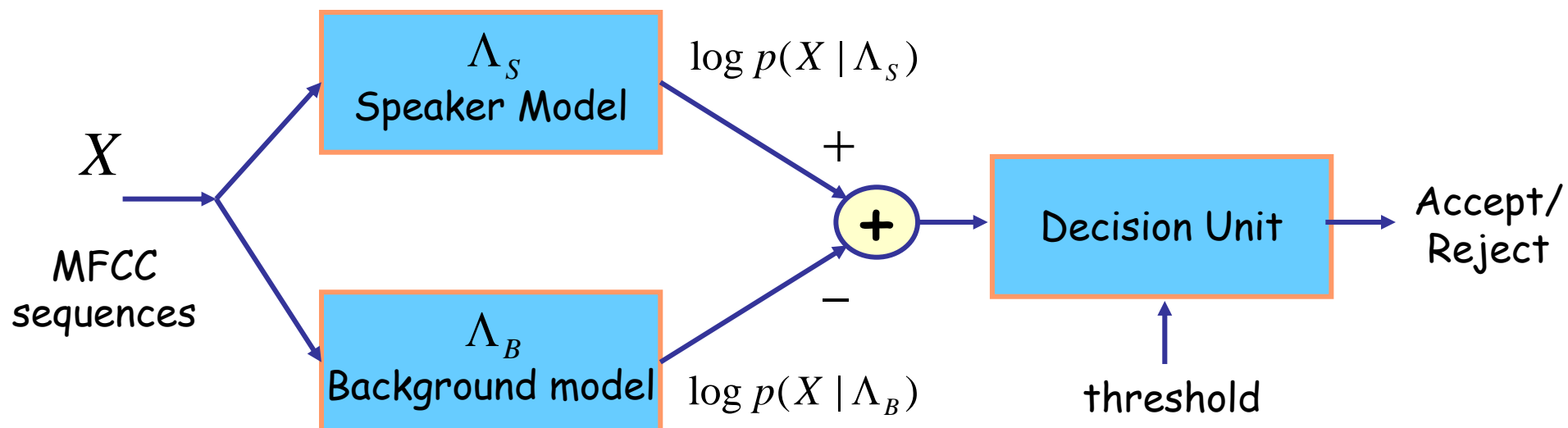
- Probabilistic feature map performs slightly better than feature map and that the fast approach can reduce computation time substantially.
- Among the approaches investigated, the fast BSFT (fBSFT) strikes a good balance between computational complexity and error rates.
- We advocate Fast BSFT for robust speaker verification because it achieves good performance without any a priori knowledge of the communication channel.

Biometric Verification



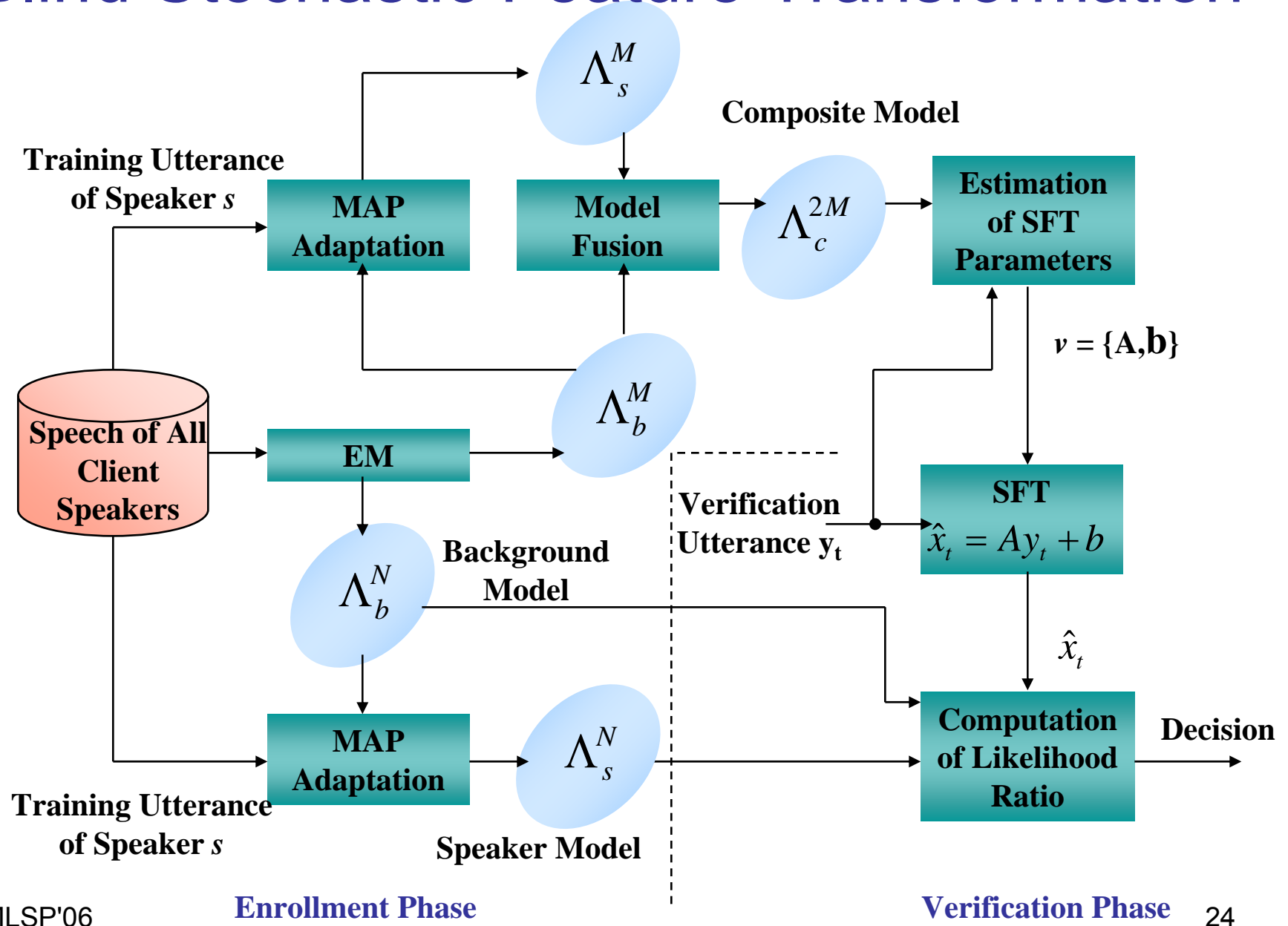
State-of-the-art Speaker Verification Systems

- Speaker verification is a biometric technology that aims to authenticate users via their voice patterns.



- The lack of robustness to channel variability and the acoustic mismatch between enrollment and verification conditions remain a major practical challenge.
- Currently, this problem is addressed by a technique called **channel mismatch compensation**.

Blind Stochastic Feature Transformation



Blind Stochastic Feature Transformation

- In BSFT, the transformed feature vector is

$$\mathbf{x} = f_{\mathbf{v}}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{b}$$

- The auxiliary function is

$$Q(\mathbf{A}', \mathbf{b}' | \mathbf{A}, \mathbf{b}) = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_{\mathbf{v}}(\mathbf{y}_t)) \log \left\{ \frac{p(f_{\mathbf{v}'}(\mathbf{y}_t) | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{|J_{\mathbf{v}'}(\mathbf{y}_t)|} \right\}$$

$$\frac{\partial Q(\mathbf{A}', \mathbf{b}' | \mathbf{A}, \mathbf{b})}{\partial a'_i} = 0 \quad \frac{\partial Q(\mathbf{A}', \mathbf{b}' | \mathbf{A}, \mathbf{b})}{\partial b'_i} = 0$$

$$\Rightarrow \begin{cases} b'_i = \frac{p_i - q_i a'_i}{r_i} \\ \left(s_i - \frac{q_i^2}{r_i} \right) a_i'^2 + \left(\frac{q_i p_i}{r_i} - u_i \right) a'_i - T = 0 \end{cases}$$

Blind Stochastic Feature Transformation

$$p_i = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_v(\mathbf{y}_t)) \boldsymbol{\mu}_{ji} \boldsymbol{\sigma}_{ji}^{-2}$$

$$q_i = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_v(\mathbf{y}_t)) \mathbf{y}_{ji} \boldsymbol{\sigma}_{ji}^{-2}$$

$$r_i = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_v(\mathbf{y}_t)) \boldsymbol{\sigma}_{ji}^{-2}$$

$$s_i = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_v(\mathbf{y}_t)) \mathbf{y}_{ji}^2 \boldsymbol{\sigma}_{ji}^{-2}$$

$$u_i = \sum_{t=1}^T \sum_{j=1}^{M_c} h_j(f_v(\mathbf{y}_t)) \boldsymbol{\mu}_{ji} \mathbf{y}_{ti} \boldsymbol{\sigma}_{ji}^{-2}$$

- In the fast BSFT, the auxiliary function is

$$Q(A', \mathbf{b}' | A, \mathbf{b}) = \sum_{t=1}^T \sum_{j \in C} h_j(f_v(\mathbf{y}_t)) \log \left\{ \frac{p(f_{v'}(\mathbf{y}_t) | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{|J_{v'}(\mathbf{y}_t)|} \right\}$$

where C contains the indexes of top- C Gaussians

Procedures for creating target speaker models for FM, PFM, and BSFT

