



**A New Adaptation Method for Speaker-Model
Creation in High-Level Speaker Verification**

Shi-Xiong Zhang and Man-Wai MAK

Dept. of Electronic and
Information Engineering

 **The Hong Kong
Polytechnic University**



Outline

Introduction of Speaker Verification

GMM system and MAP Adaptation

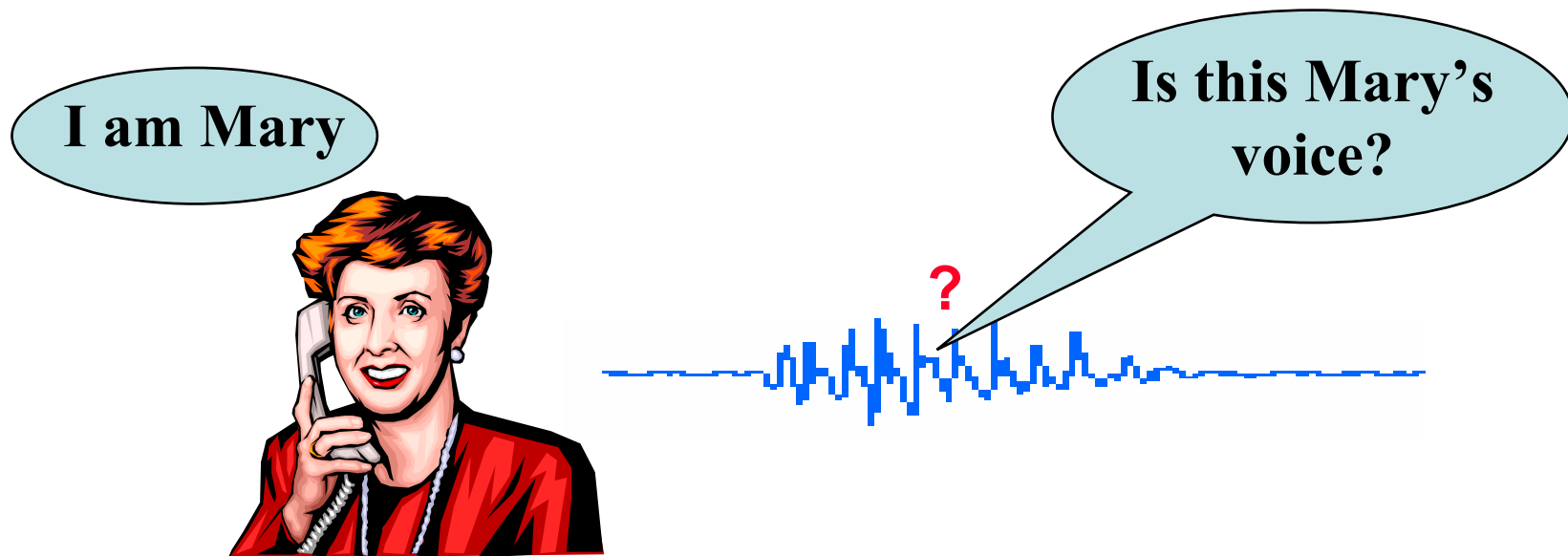
New Adaptation for Speaker Modeling

Experiments and Results



What is Speaker Verification?

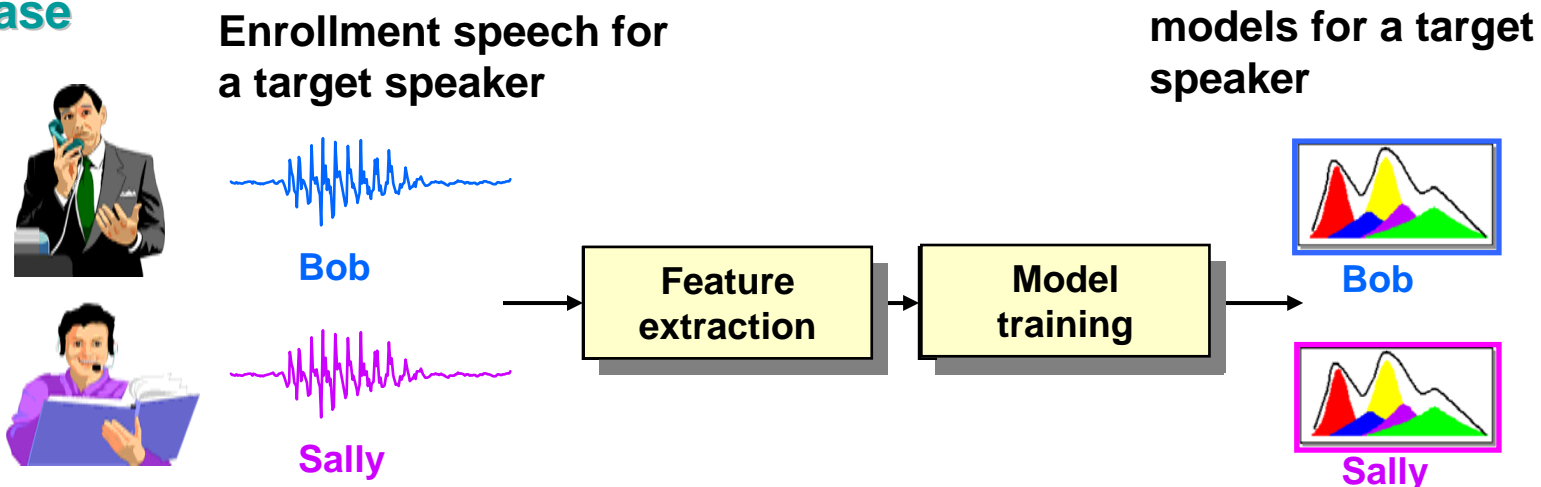
- To verify the identity of a claimant based on his/her own voices
(Determine whether a person is who he/she claims to be)



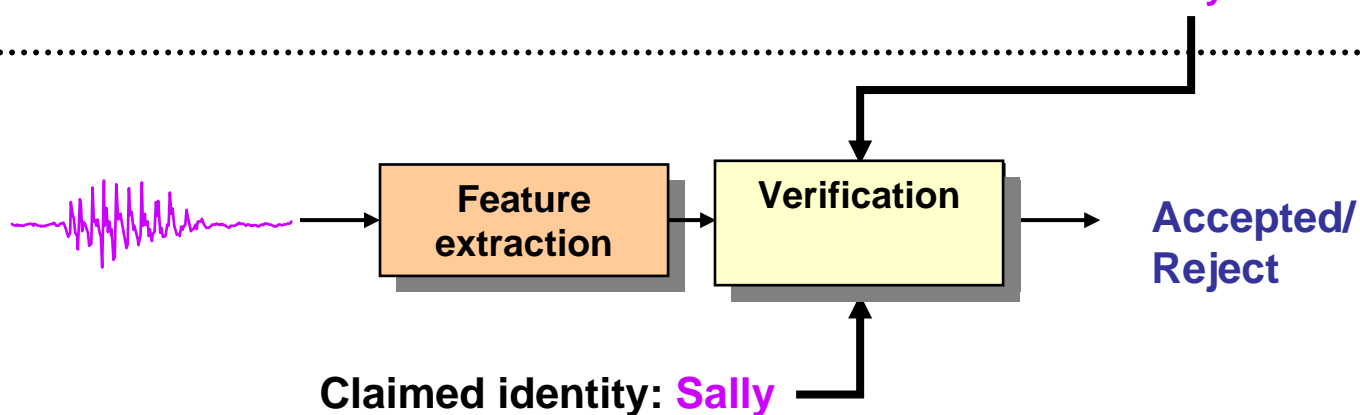


Two Phases of Speaker Verification

Enrollment Phase



Verification Phase

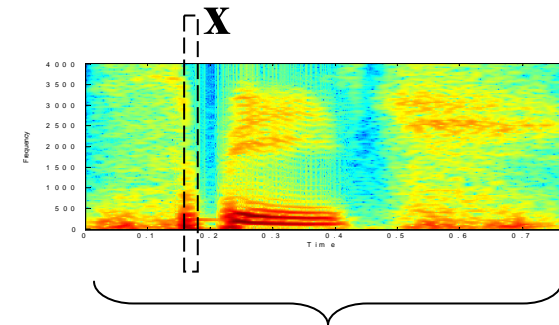




Traditional Speaker Modeling

- The mixture density function is a linear combination of several Gaussian densities
- Gaussian mixture model (GMM):

$$p(\mathbf{x} | \Lambda_s) = \sum_{i=1}^M w_i^s p_i^s(\mathbf{x})$$



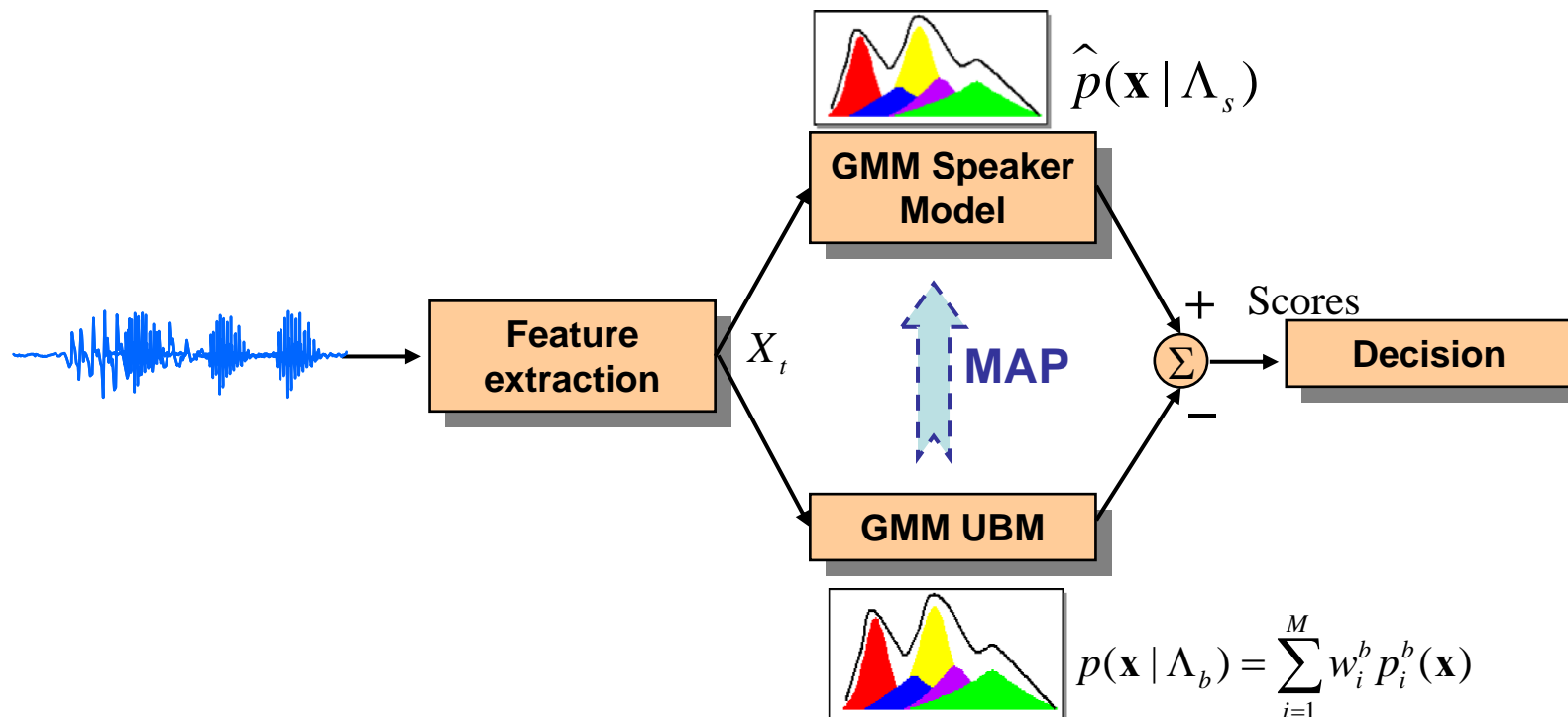
$$\Lambda_s = \{w_i^s, \boldsymbol{\mu}_i^s, \boldsymbol{\Sigma}_i^s\}$$

$$p_i^s(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_i^s|^{1/2}} \exp\left\{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i^s)' (\boldsymbol{\Sigma}_i^s)^{-1} (\mathbf{x} - \boldsymbol{\mu}_i^s)\right\}$$



Verification based on Speaker and Background GMM Model

- **Universal Background Model (UBM)**
The UBM is a large GMM trained to represent the distribution of speaker-independent features.
- **Speaker GMM is used to represent a specific user**





High-Level Features

- Humans use several levels of perceptual cues for speaker recognition

High-level cues
(learned traits)



Low-level cues
(physical traits)

Perceptual Cues	Depends on
<ul style="list-style-type: none">• Pronunciations• Idiolect (word usage)	Socio-economic status, education, place of birth
<ul style="list-style-type: none">• Prosodic (Rhythm)• Speed of speech• Intonation	Personality type, parental influence
<ul style="list-style-type: none">• Acoustic aspect of speech	Physical structure of vocal apparatus

Difficult to
automatically
extract

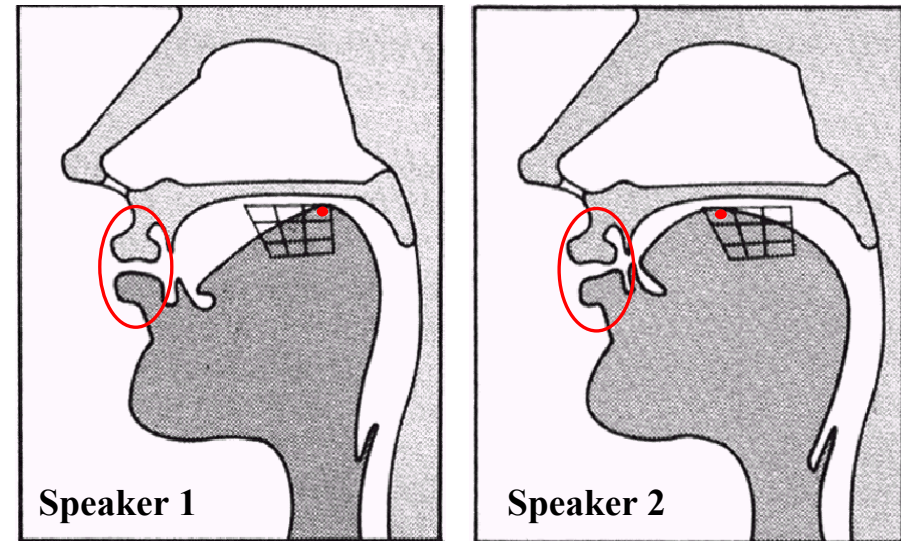


Easy to
automatically
extract



What's the **Articulatory Feature**?

Articulatory features (AFs) are abstract classes that describe the **movements** and **positions** of different articulators during speech production.



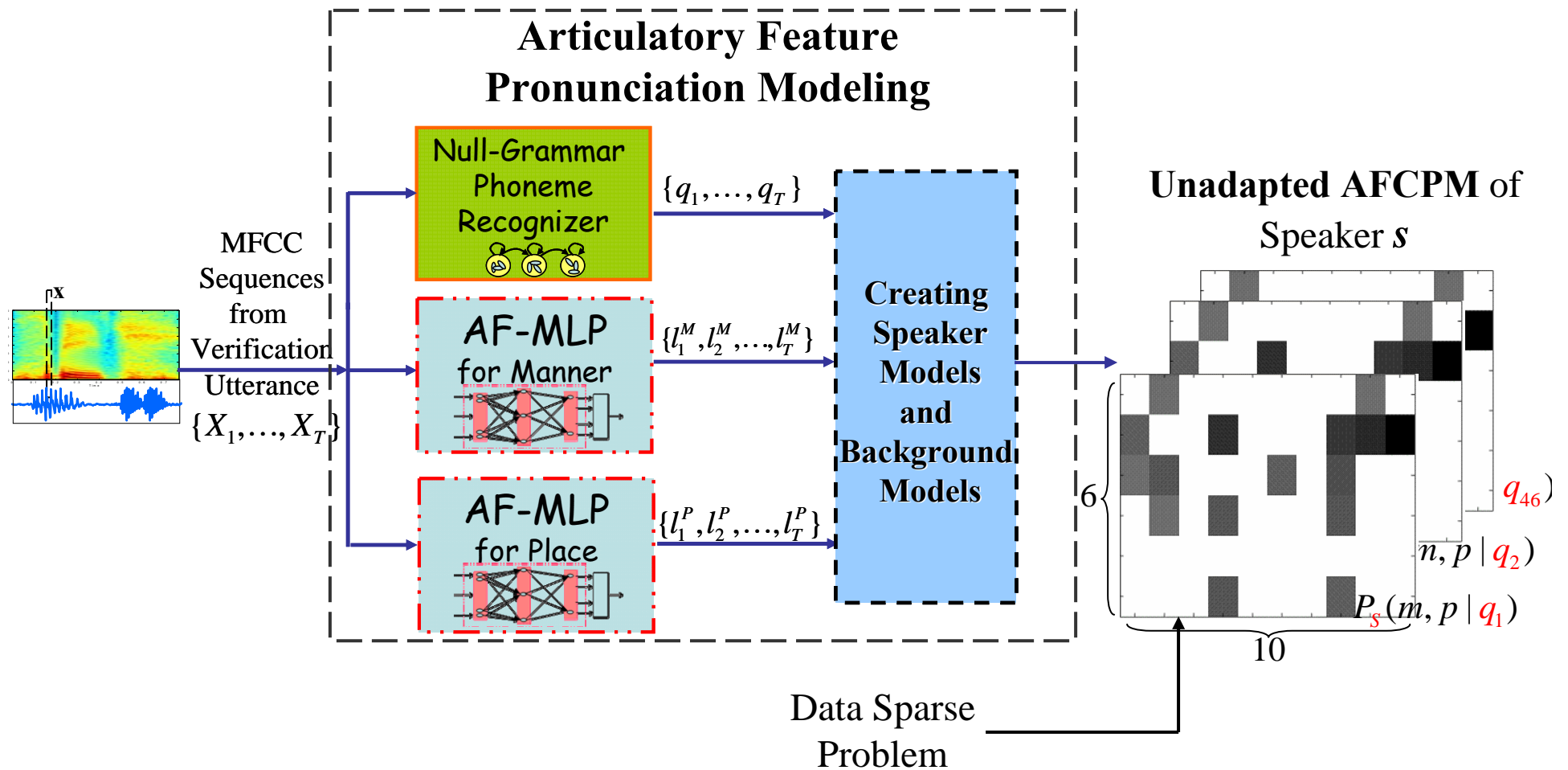
/u/

Two AFs were adopted for Pronunciation Modelling (AFCPM):

Articulatory Properties	Classes	Number of Classes
Manner(\mathcal{M})	Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral	6
Place(\mathcal{P})	Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal	10



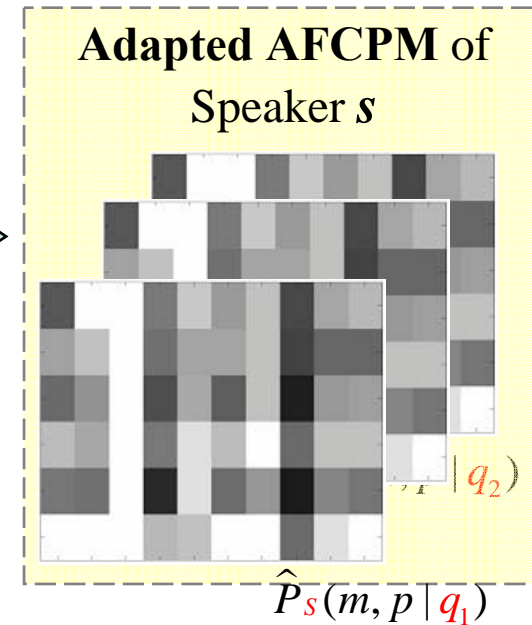
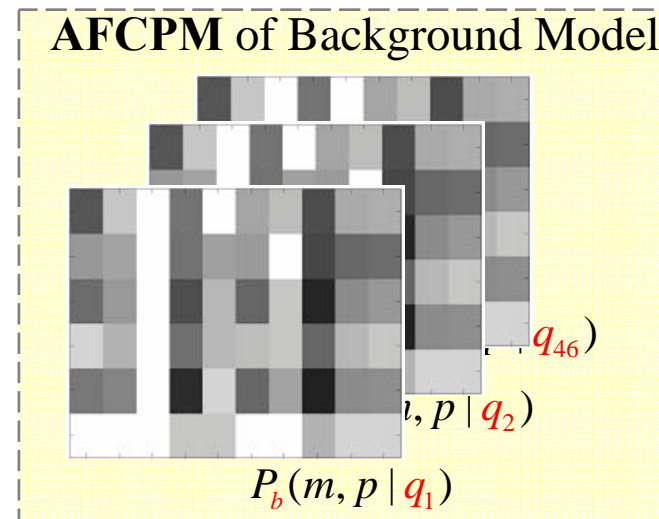
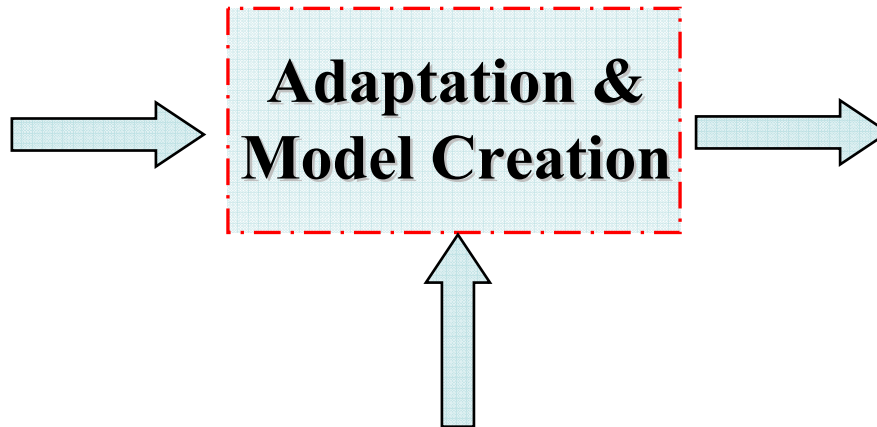
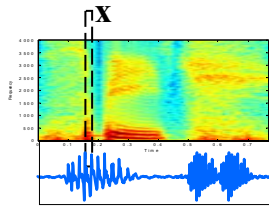
AFCPM Training





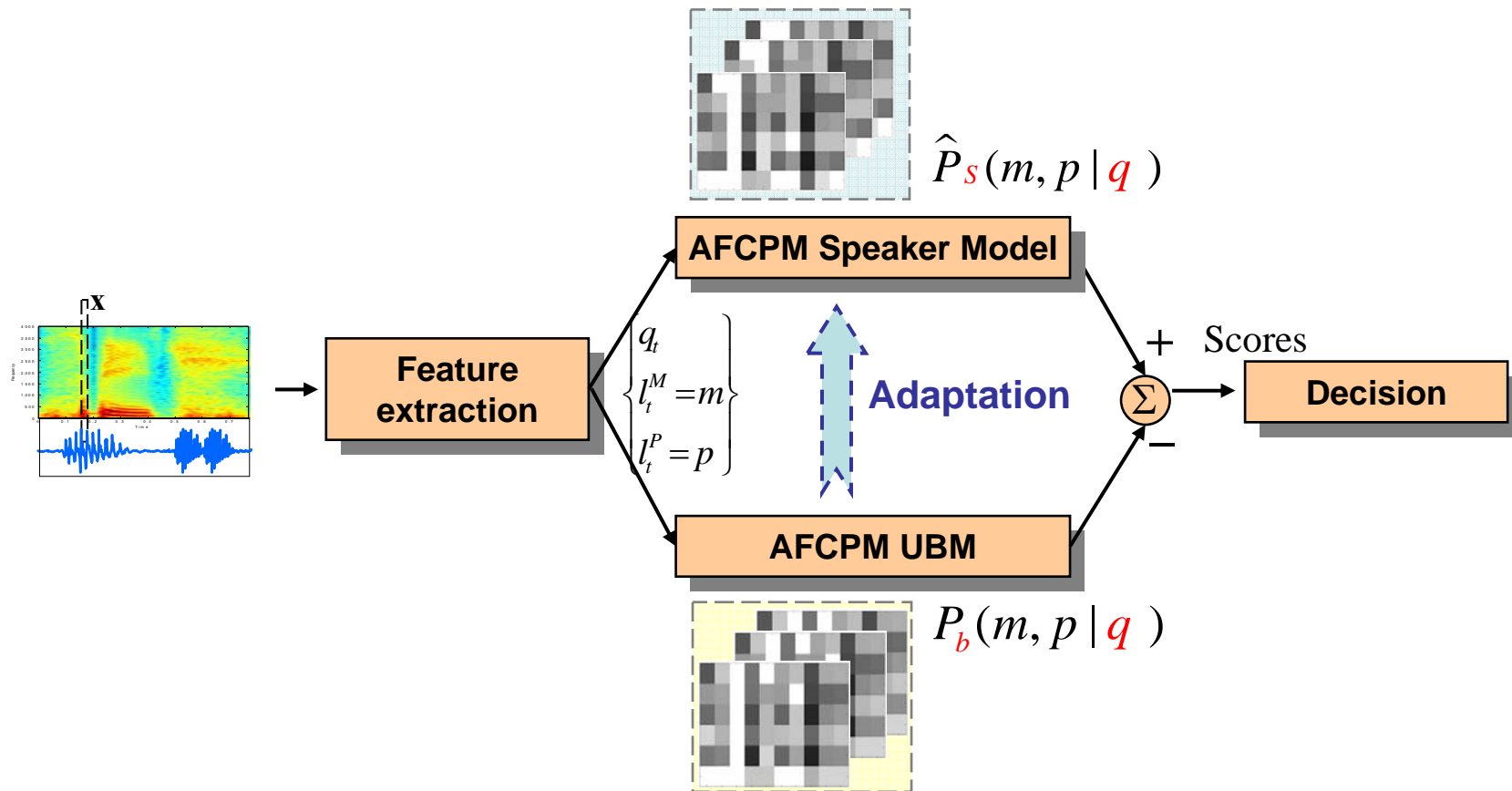
Contribution of our paper

Enrollment Data for
a target speaker s





Verification based on Speaker and Background AFCPM Model

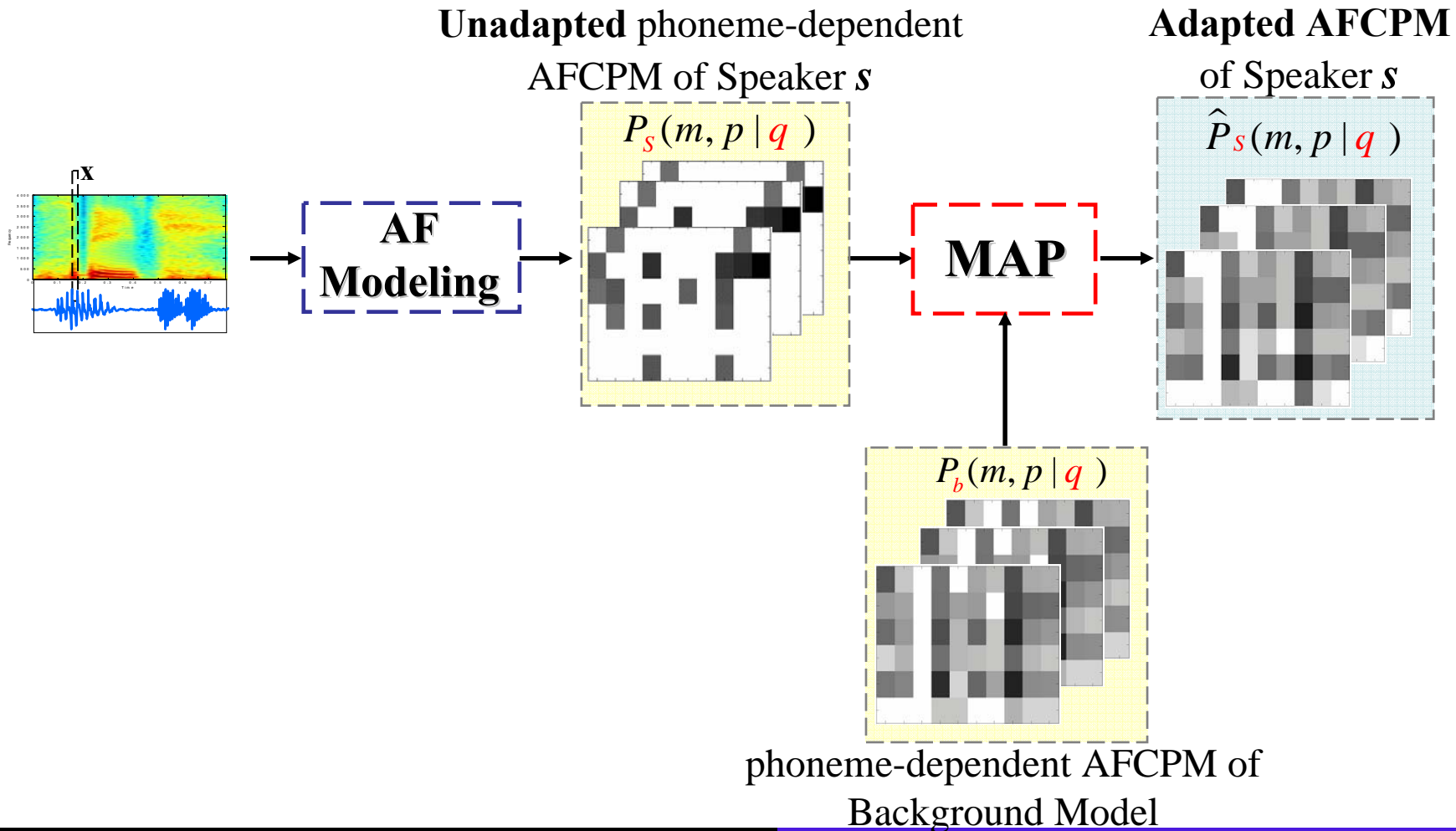




Traditional MAP Adaptation

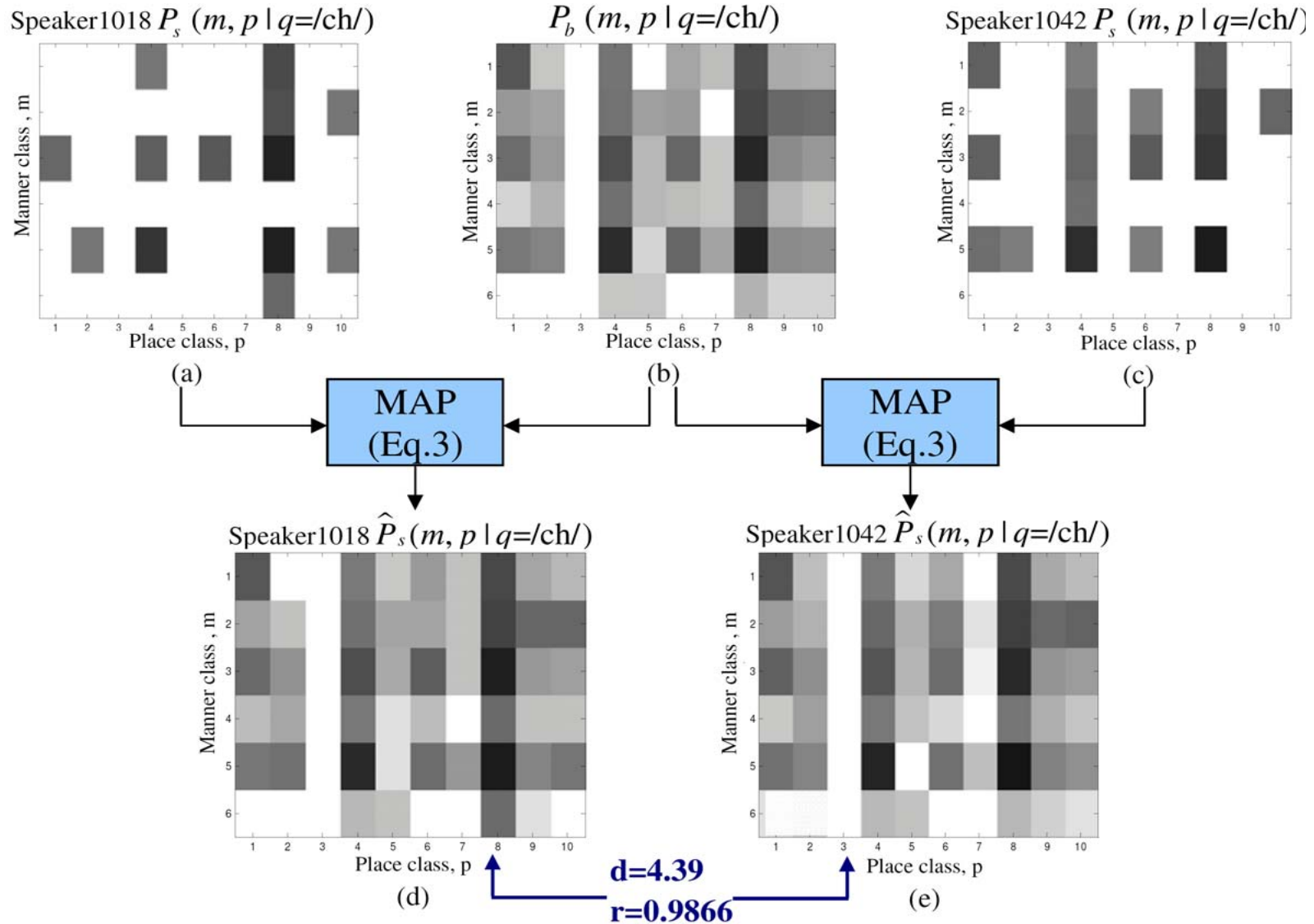
MAP
Adaptation:

$$\hat{P}_s(m, p | q) = \beta P_s(m, p | q) + (1 - \beta) P_b(m, p | q)$$
$$\beta = \frac{\#((*, *, q) \text{ in the utterances of speaker } s)}{\#((*, *, q) \text{ in the utterances of speaker } s) + r}$$



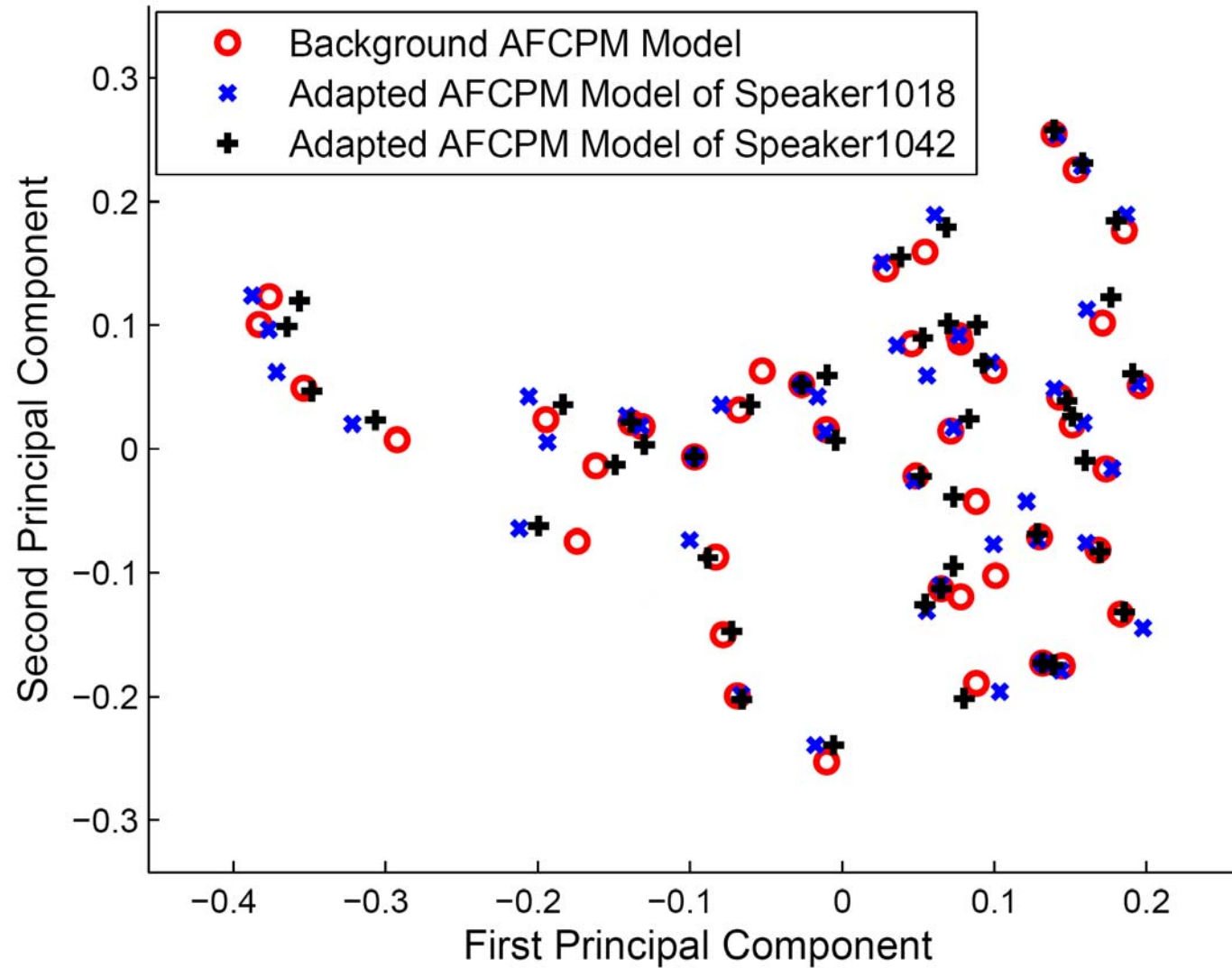


Limitation of Traditional MAP Adaptation





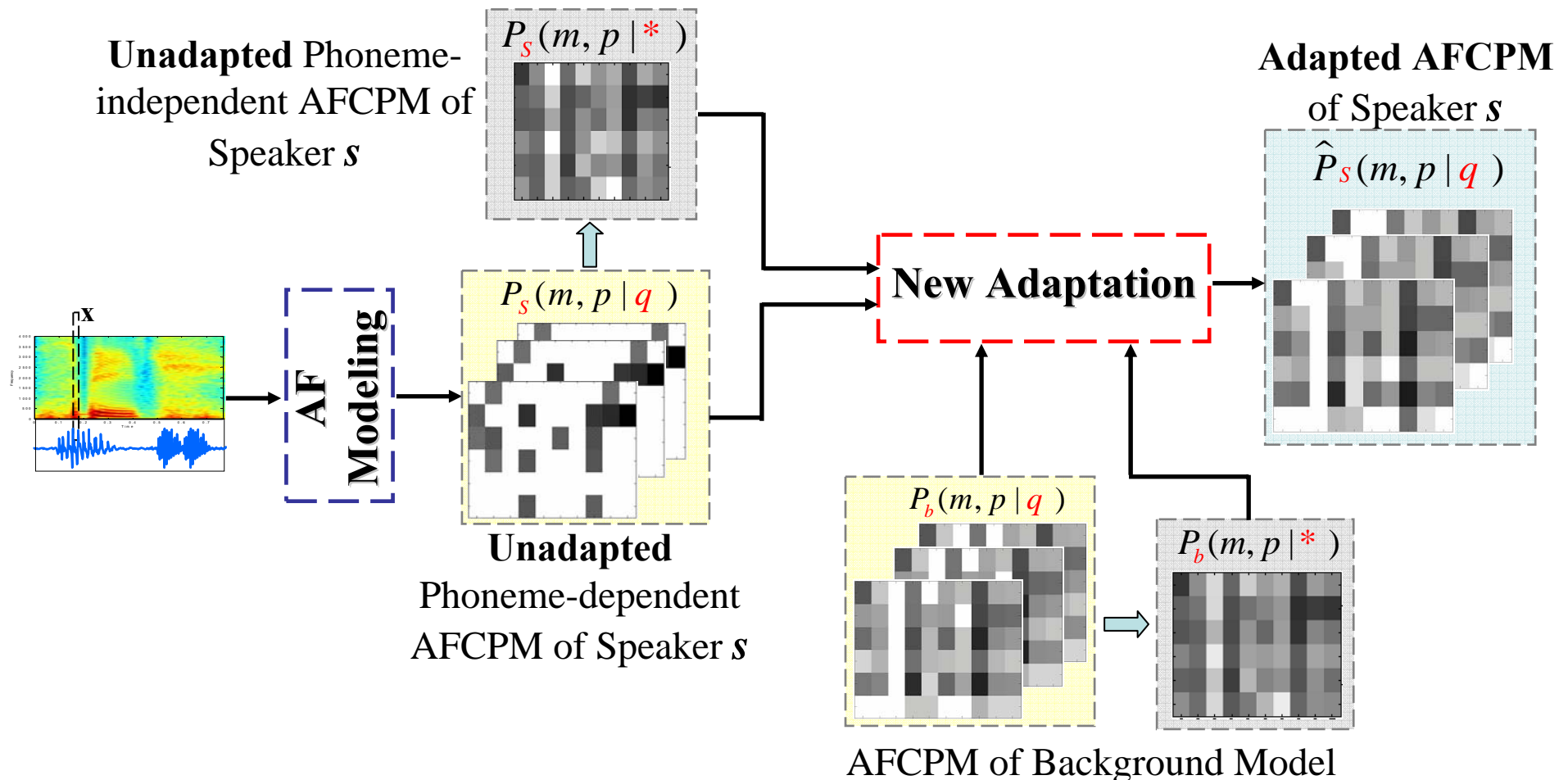
Limitation of Traditional MAP Adaptation





Proposed New Adaptation Method

$$\hat{P}_s(m, p | q) = \beta_s^q P_s(m, p | q) + (1 - \beta_s^q) \left[\alpha_b^q P_b(m, p | q) + (1 - \alpha_b^q) \frac{P_b(m, p | q)}{P_b(m, p | *)} P_s(m, p | *) \right]$$
$$\alpha_b^q = \frac{\#((*, *, q) \text{ in the utterances of all background speakers})}{\#((*, *, q) \text{ in the utterances of all background speakers}) + r}$$

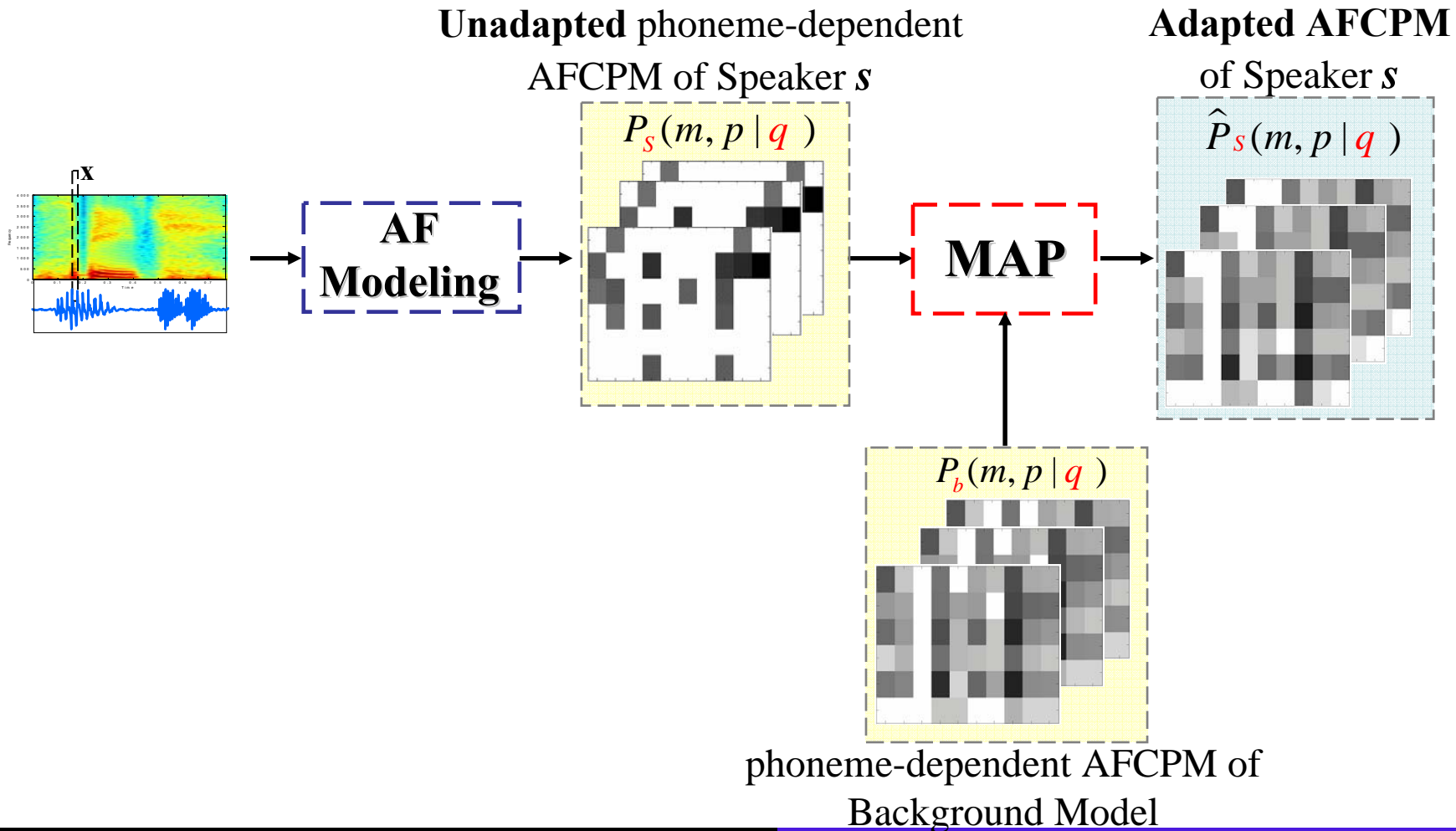




Traditional MAP Adaptation

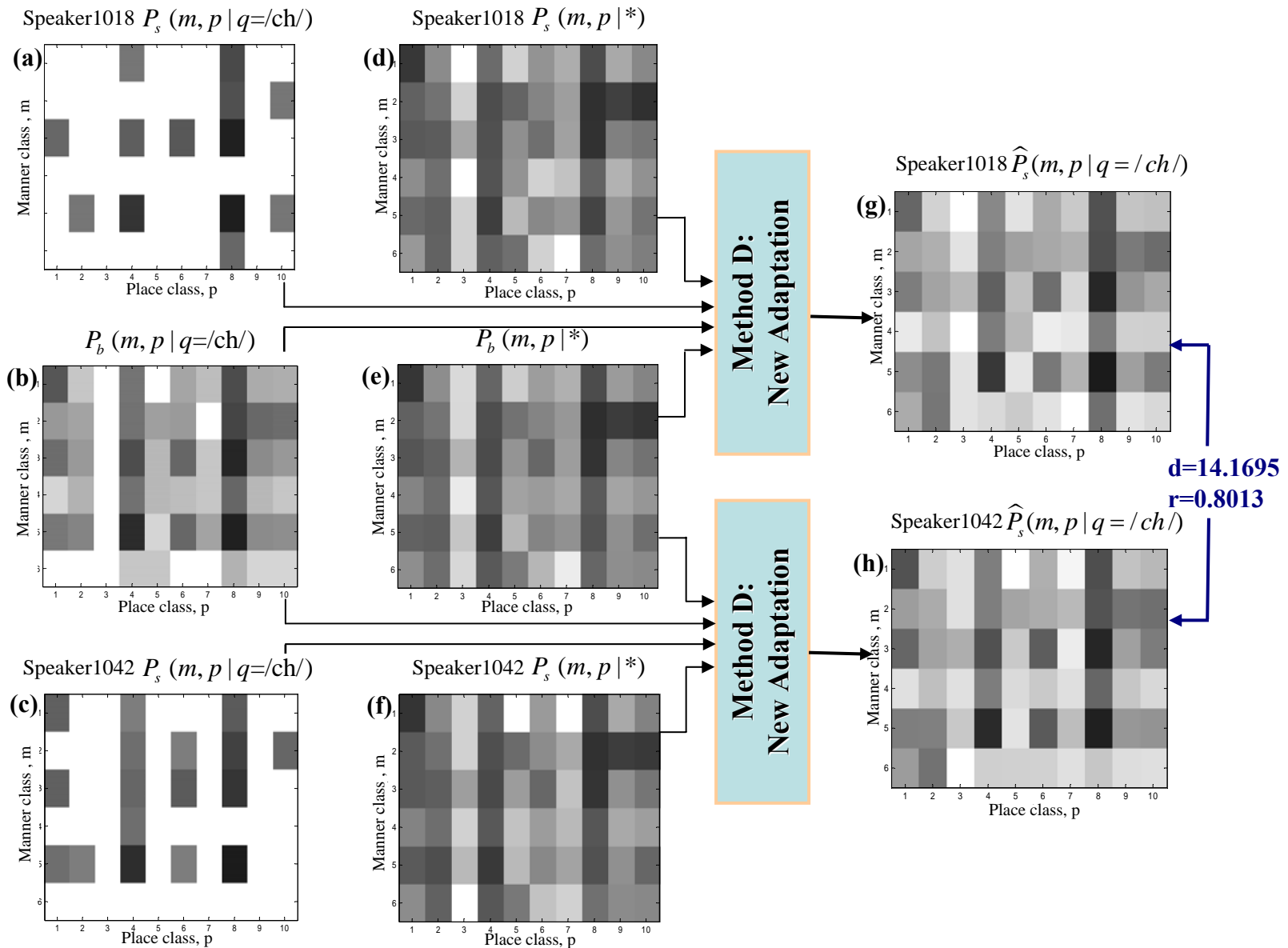
MAP
Adaptation:

$$\hat{P}_s(m, p | q) = \beta P_s(m, p | q) + (1 - \beta) P_b(m, p | q)$$
$$\beta = \frac{\#((*, *, q) \text{ in the utterances of speaker } s)}{\#((*, *, q) \text{ in the utterances of speaker } s) + r}$$



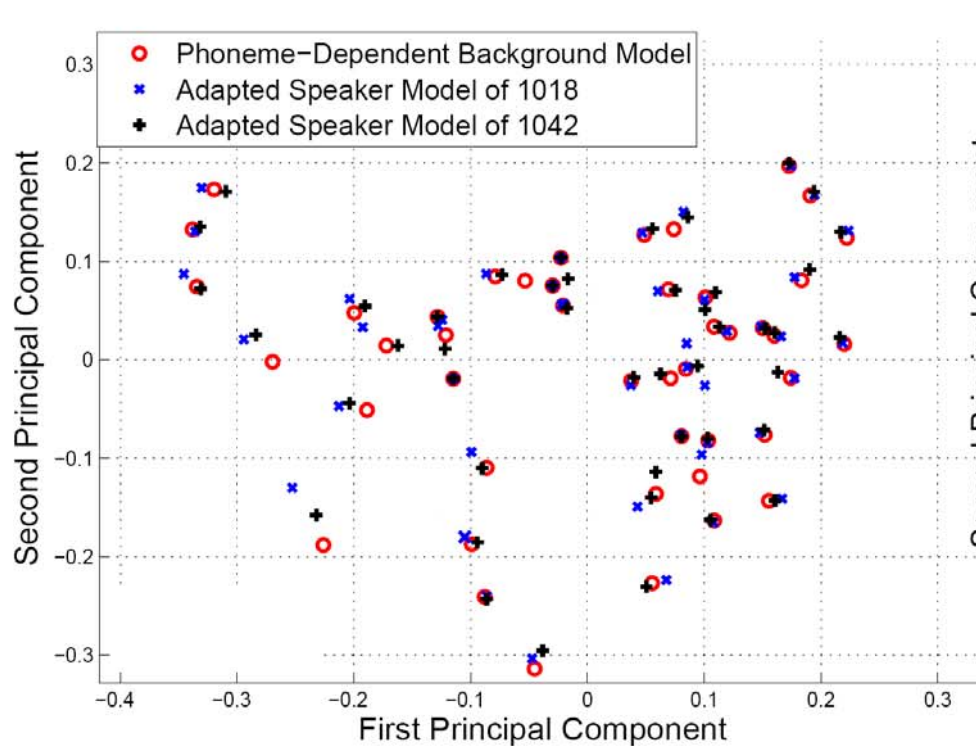


Proposed Adaptation Method for High-Level Speaker Verification

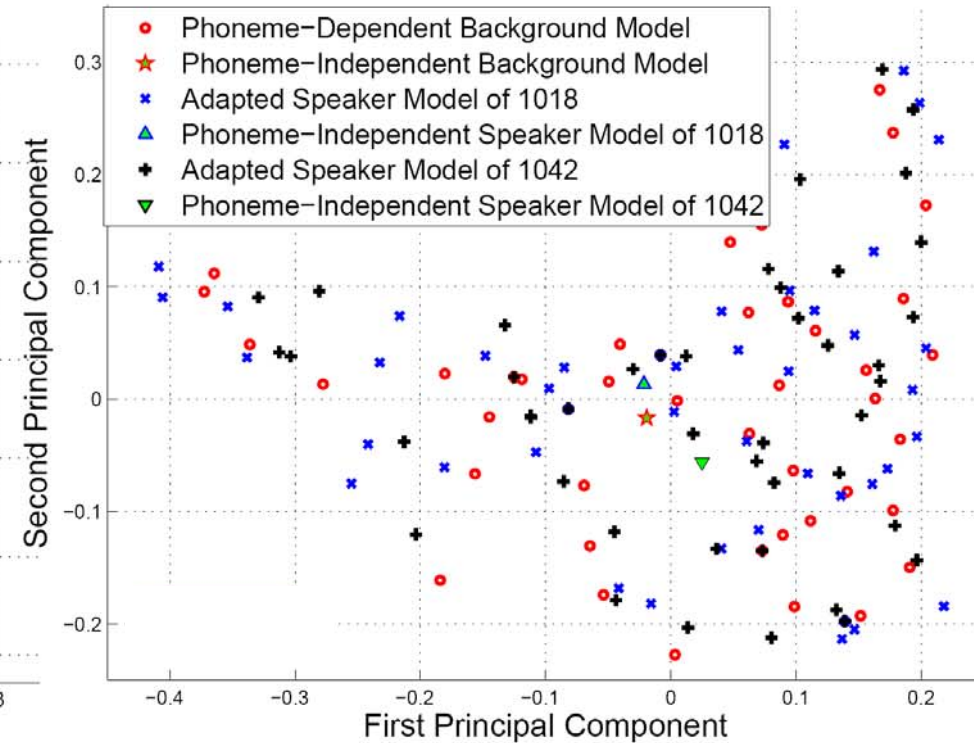




Comparing the created speaker models based on Method A and D



Conventional MAP Adaptation

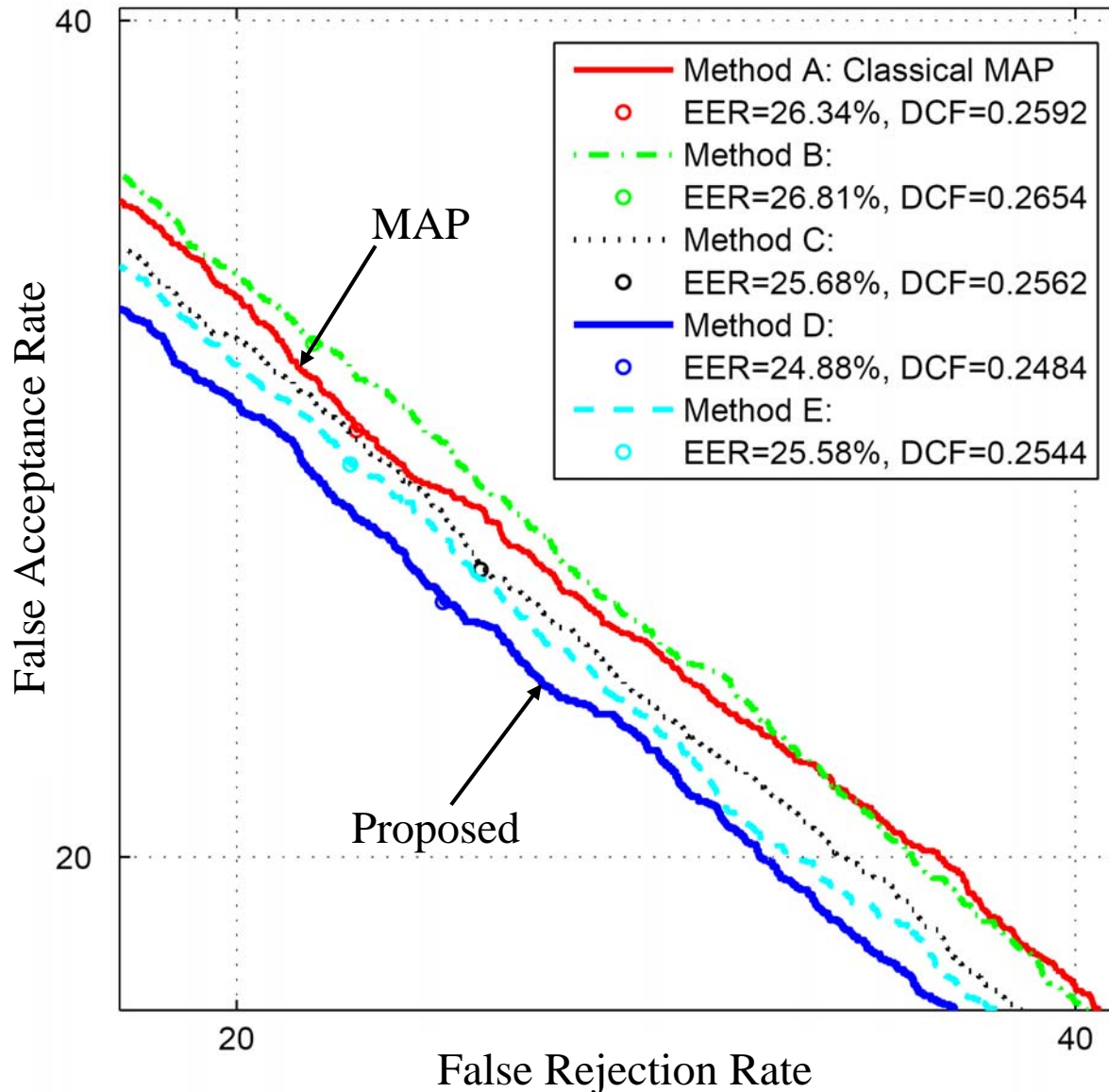


Proposed New Adaptation



Comparison and Results

Performance comparison of different adaptation methods



Database	Purpose
SPIDRE	To train null-grammar phone recognizer
HTIMIT	To train the AF-MLPs
NIST99	To create the background models and mapping functions
NIST00	To create speaker models and evaluate the performance

NO. of target speakers: **547**
NO. of speaker trails: **3,135**
NO. of impostor attempts: **3,0151**



THANKS !

QA



Complementary



How to extract AFs?

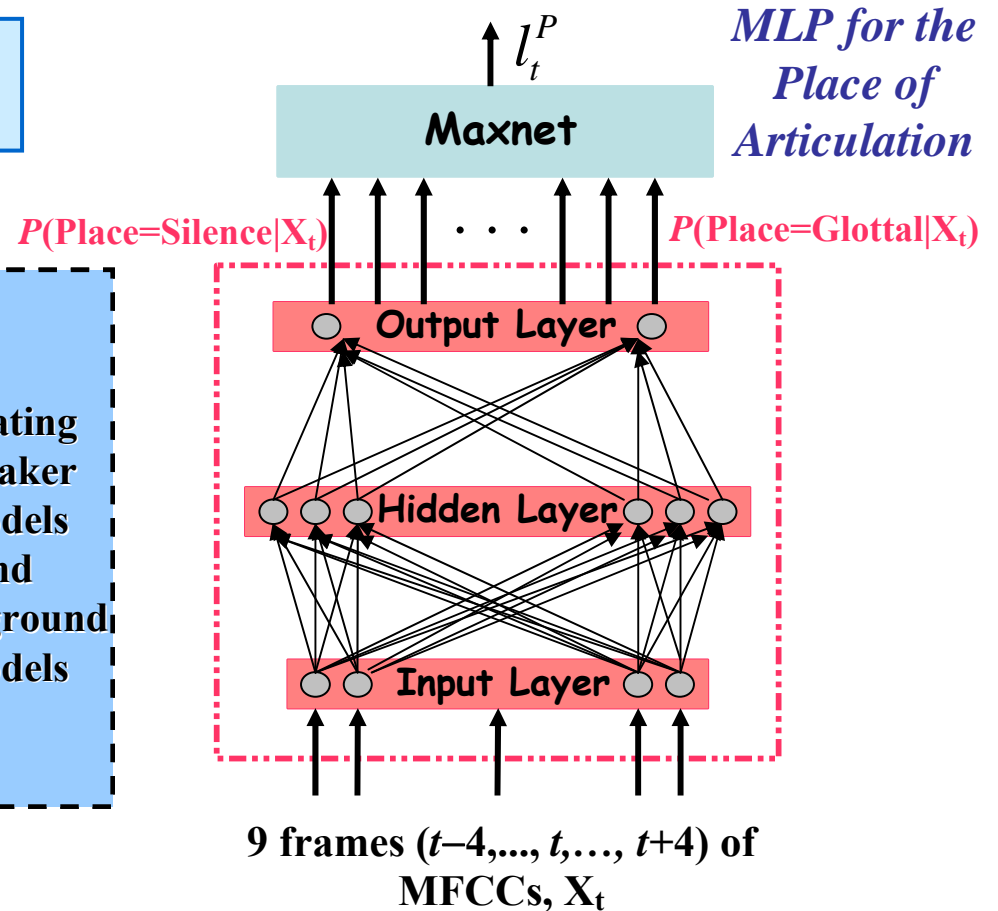
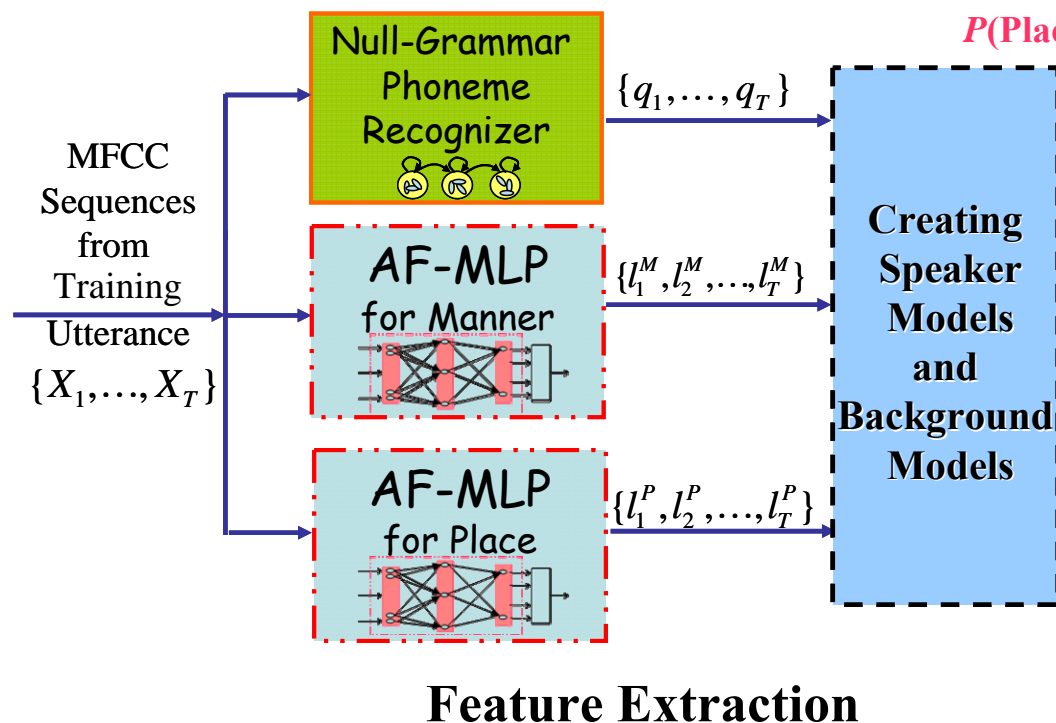
Manner labels

$$l_t^M = \arg \max_{m \in M} P(\text{Manner} = m | X_t)$$

Place labels

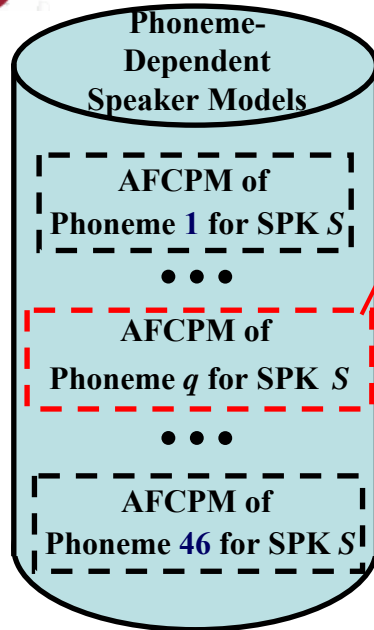
$$l_t^P = \arg \max_{p \in P} P(\text{Place} = p | X_t)$$

At frame position t , 9 consecutive frames of MFCCs (X_t) centred at frame t were input to an MLP.





Phoneme-Dependent AFCPM Training

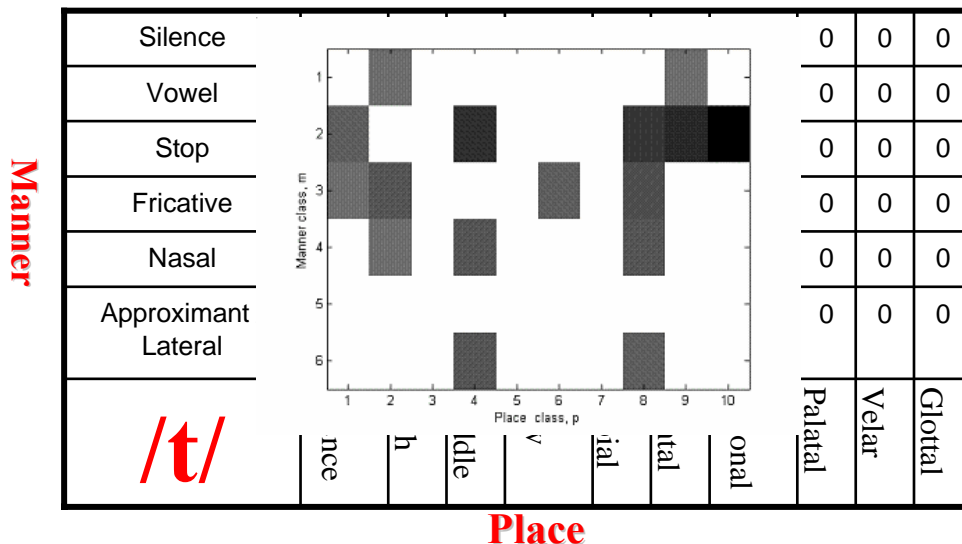


$$P_s(m, p | q) = P_s(\text{Manner} = m, \text{Place} = p | \text{Phoneme} = q)$$

$$= \frac{\text{No. of } \{m, p, q\} \text{ from the data of speaker } s}{\text{No. of } \{q\} \text{ from the data of speaker } s}$$

$P(\text{Manner}=\text{Vowel}, \text{Place}=\text{Low} \quad | /t/) = 1/6,$
 $P(\text{Manner}=\text{Silence}, \text{Place}=\text{Silence} \quad | /t/) = 4/6,$
 $P(\text{Manner}=\text{Stop}, \text{Place}=\text{Coronal} \quad | /t/) = 1/6,$
 and all other entries are equal to 0.

$$P_s(m, p | q = /t/)$$



Frame, t	Phoneme, q_t	l^{m_t}	l^{p_t}
1	/t/	Vowel	Low
2	/t/	Silence	Silence
3	/t/	Silence	Silence
4	/t/	Silence	Silence
5	/t/	Silence	Silence
6	/t/	Stop	Coronal
7	/aa/	Vowel	Low
...

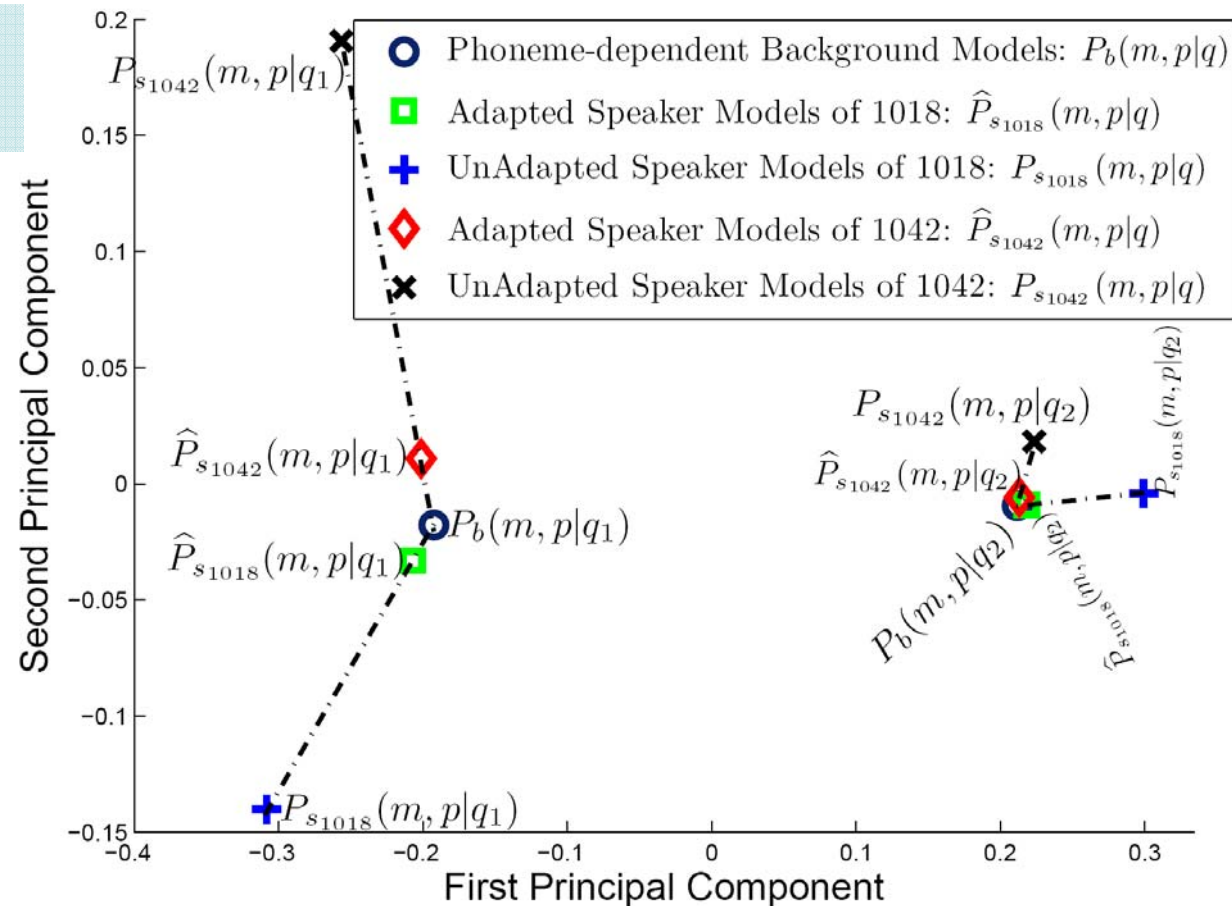


Traditional MAP Adaptation

$$\hat{P}_s(m, p | q) = \beta P_s(m, p | q) + (1 - \beta) P_b(m, p | q)$$

$$\beta = \frac{\#((*, *, q) \text{ in the utterances of speaker } s)}{\#((*, *, q) \text{ in the utterances of speaker } s) + r}$$

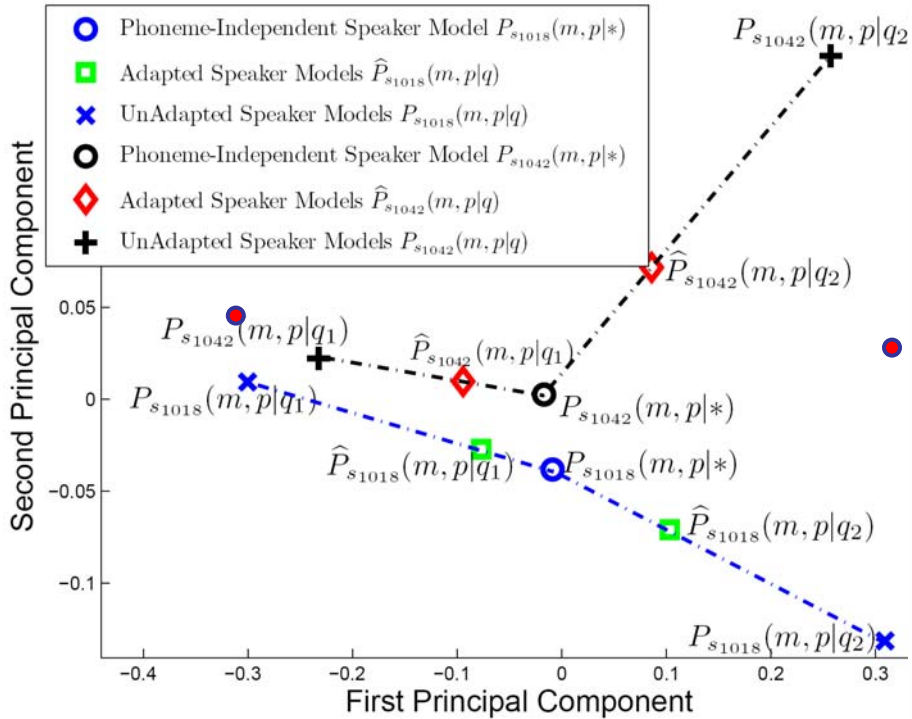
Method A MAP Adaptation:





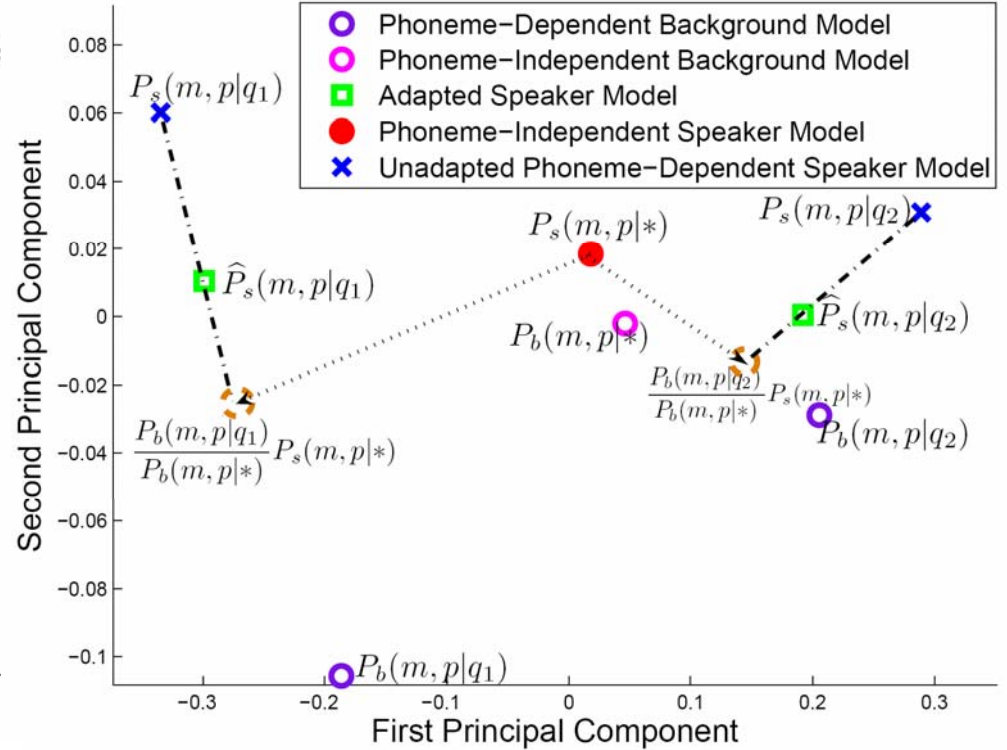
Proposed Adaptation Method B and C

Adaptation Method B:



$$\hat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) P_s(m, p|*)$$

Adaptation Method C:



$$\hat{P}_s(m, p|q) = \beta_s^q P_s(m, p|q) + (1 - \beta_s^q) \cdot \left[\frac{P_b(m, p|q)}{P_b(m, p|*)} \cdot P_s(m, p|*) \right]$$



Proposed New Adaptation Method

$$\hat{P}_s(m, p | q) = \beta_s^q P_s(m, p | q) + (1 - \beta_s^q) \left[\alpha_b^q P_b(m, p | q) + (1 - \alpha_b^q) \frac{P_b(m, p | q)}{P_b(m, p | *)} P_b(m, p | *) \right]$$

$$\alpha_b^q = \frac{\#((*, *, q) \text{ in the utterances of all background speakers})}{\#((*, *, q) \text{ in the utterances of all background speakers}) + r}$$

Adaptation Method D:

