

**Phoneme-Based Speaker Verification
using
Adapted Phoneme
Gaussian Mixture Models**

**Yuval Bistriz
EE Dept. Tel Aviv University**

Based on M.Sc. Thesis of
D. Gutman

Speaker Recognition vs. Speaker Verification

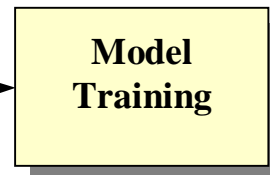
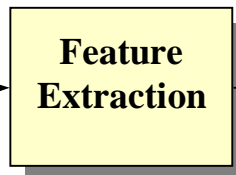
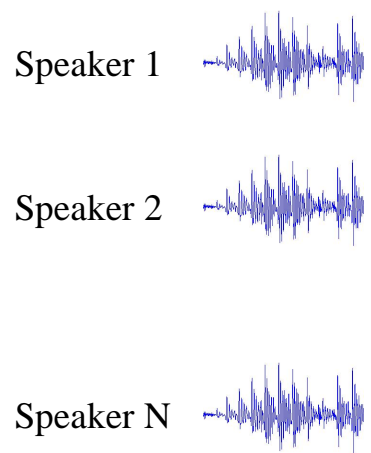
- **Speaker recognition is the process of automatically recognizing a speaker from information included in the sound waves of his speech.**
- **Speaker Verification (SV) is the binary decision of acceptance or rejection of an identity claim of a speaker.**

Other classifications: text dependent or text independent, etc.

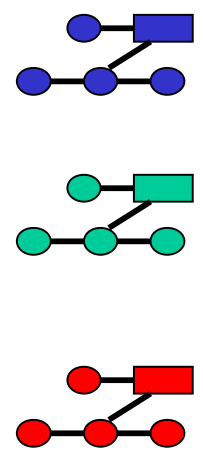
Speaker Verification

Enrollment Phase (Training)

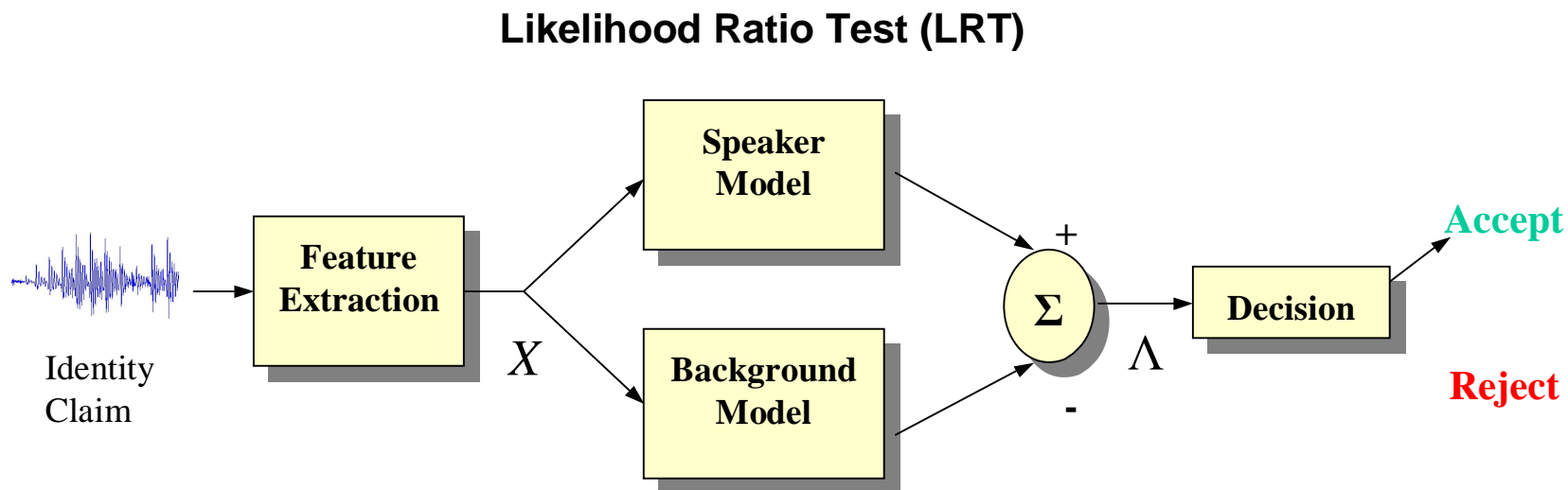
Enrollment speech for each speaker



Trained Model for each speaker



Verification Phase (Testing)



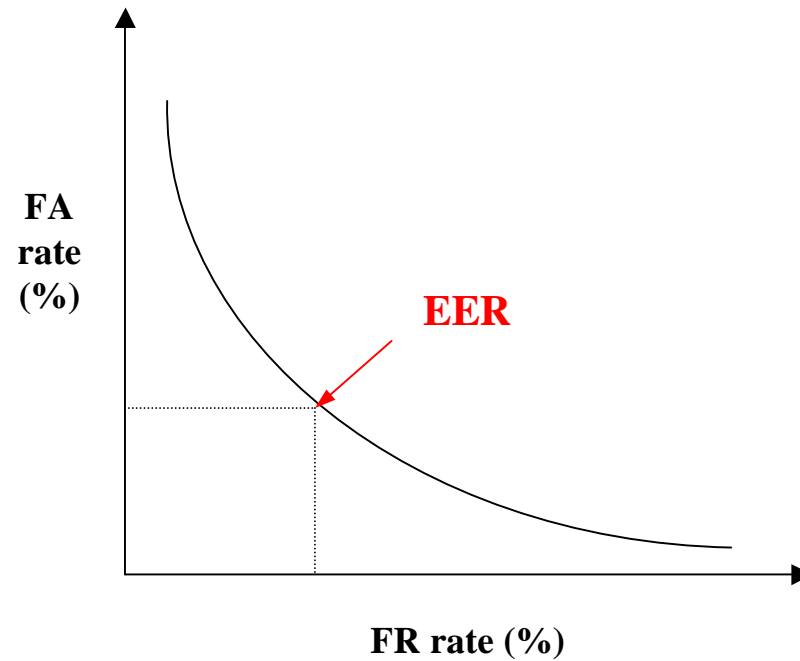
X is from the claimed speaker: probability $p(\lambda_c | X)$

X is *not* from the claimed speaker: probability $p(\lambda_{\bar{c}} | X)$

Use Bayes, can measure (log) likelihood by:

$$\Lambda(X) = \log p(X | \lambda_c) - \log p(X | \lambda_{\bar{c}})$$

Receiver Operating Curve (ROC)

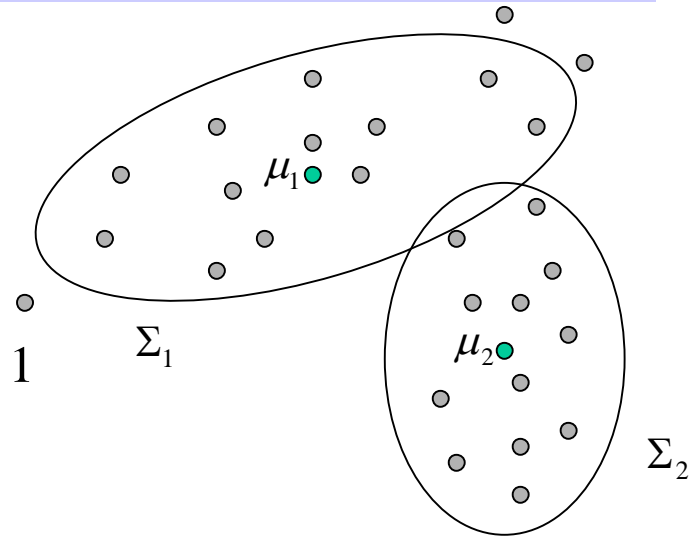


Equal Error Rate (EER) : False Acceptance rate = False Rejection rate

Gaussian Mixture Models (GMM)

GMM is a model that consists of a weighted sum of M Gaussian densities used to measure probability for a feature vector, say $x_0 \in \mathbb{R}^{D \times 1}$

$$p(x_0 | \lambda) = \sum_{i=1}^M w_i g_i(x_0) \quad ; \quad \sum_{i=1}^M w_i = 1 \quad ; \quad 0 \leq w_i \leq 1$$



$$g_i(x_0) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x_0 - \mu_i)' (\Sigma_i)^{-1} (x_0 - \mu_i)\right\}, \quad \mu_i \in \mathbb{R}^{D \times 1}, \Sigma_i \in \mathbb{R}^{D \times D}$$

- A GMM is denoted as: $\lambda = \{w_i, \mu_i, \Sigma_i\}_1^M$
- The log-likelihood of a sequence of T feature vectors, $X = \{x_1, \dots, x_T\}$

$$\log p(X | \lambda) = \sum_{t=1}^T \log p(x_t | \lambda)$$

GMM Training

Training data is used to refine a model by the algorithm depicted below
(called EM or BW)

$$\lambda = \{w_i, \mu_i, \Sigma_i\}_1^M \Rightarrow \hat{\lambda} = \{\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i\}_1^M$$
$$p(X | \hat{\lambda}) \geq p(X | \lambda)$$

Compute

$$p(i | x_t, \lambda) = \frac{w_i g_i(x_t)}{\sum_{j=1}^M w_j g_j(x_t)}$$
$$n_i = \sum_{t=1}^T p(i | x_t, \lambda)$$
$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^T p(i | x_t, \lambda) x_t$$
$$R_i(x) = \frac{1}{n_i} \sum_{t=1}^T p(i | x_t, \lambda) x_t x_t'$$

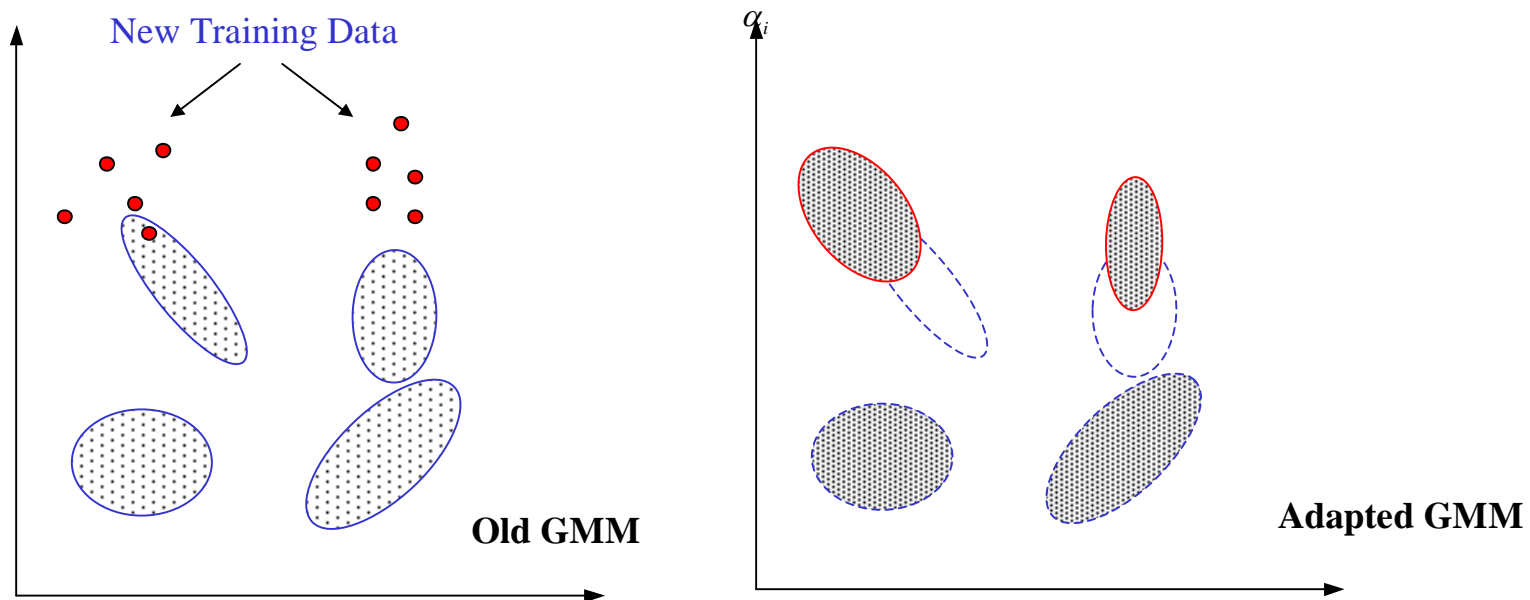
Update

$$\hat{w}_i = \frac{1}{T} n_i$$
$$\hat{\mu}_i = E_i(x)$$
$$\hat{\Sigma}_i = R_i(x) - \hat{\mu}_i \hat{\mu}_i'$$

Adapted GMM

- Adaptation to an extent dependent on data determined by coefficients α_i
 $\alpha_i \rightarrow 1$ full exposure to new data; $\alpha_i = 0$ stay with old parameters

$$\alpha_i = \frac{n_i}{n_i + r} \quad r \text{ is a "relevance factor"}$$



Adaptation of a GMM

Use training data to gradually adjust a model

$$\{w_i, \mu_i, \Sigma_i\}_1^M \Rightarrow \{\tilde{w}_i, \tilde{\mu}_i, \tilde{\Sigma}_i\}_1^M$$

Use training data to compute:

$$p(i | x_t, \lambda) = \frac{w_i g_i(x_t)}{\sum_{j=1}^M w_j g_j(x_t)}$$

$$n_i = \sum_{t=1}^T p(i | x_t, \lambda)$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^T p(i | x_t, \lambda) x_t$$

$$R_i(x) = \frac{1}{n_i} \sum_{t=1}^T p(i | x_t, \lambda) x_t x_t'$$

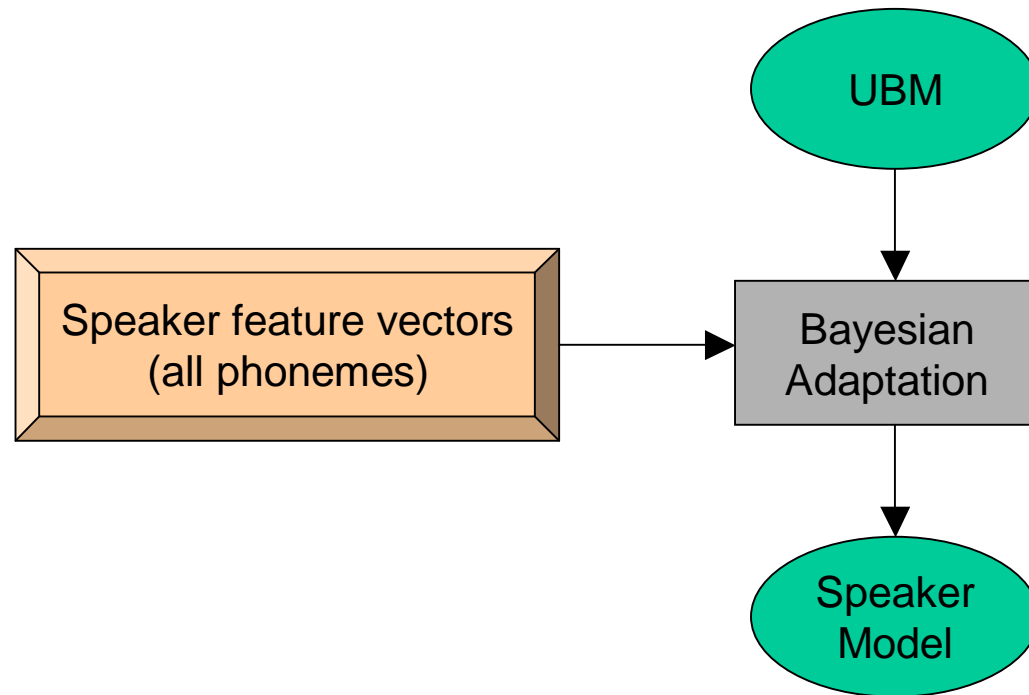
Update by Combination:

$$\tilde{w}_i = [\alpha_i^w \frac{1}{T} n_i + (1 - \alpha_i^w) w_i] \gamma$$

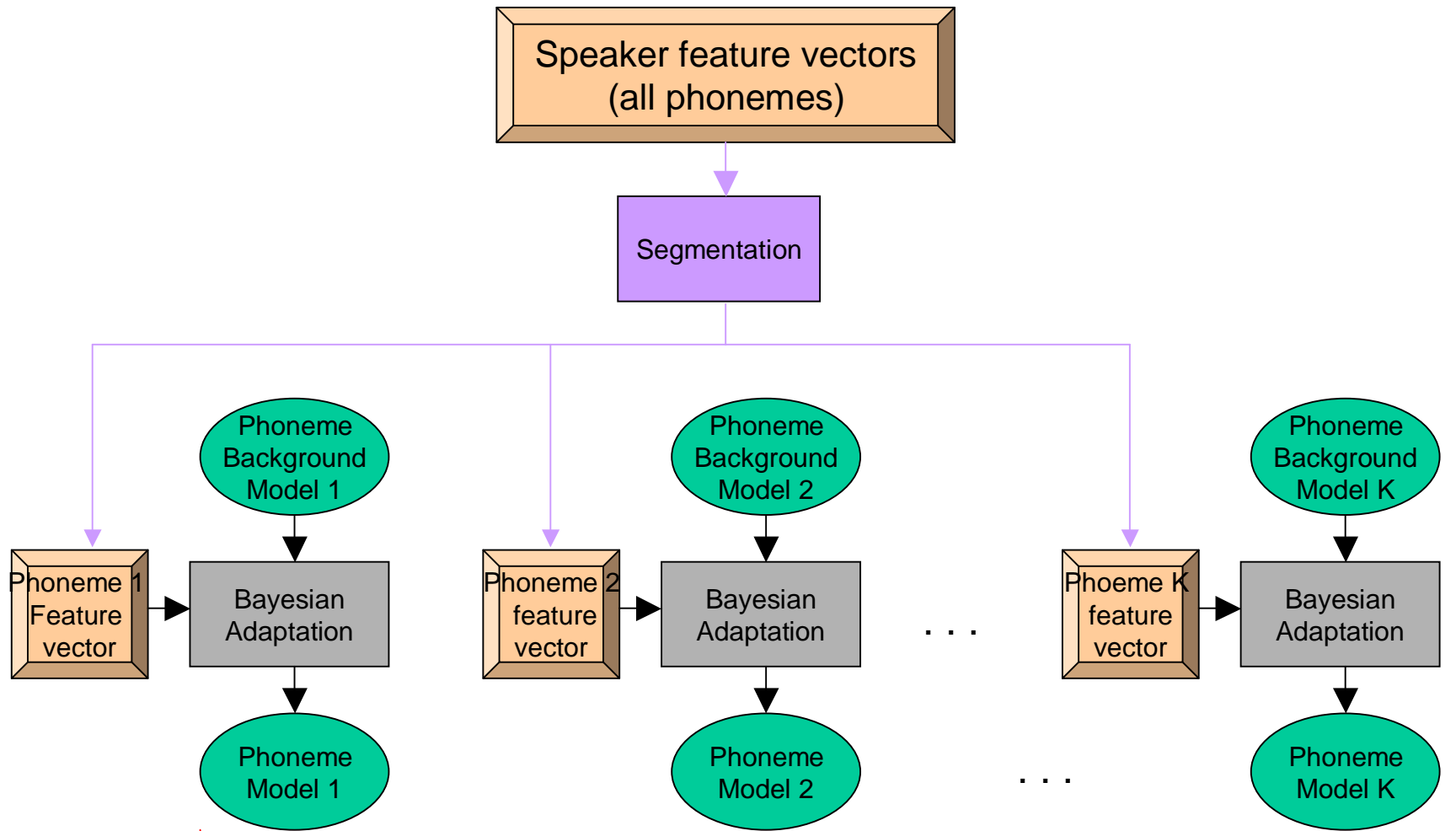
$$\tilde{\mu}_i = \alpha_i^m E_i(x) + (1 - \alpha_i^m) \mu_i$$

$$\tilde{\Sigma}_i = \alpha_i^v [R_i(x) - E_i(x) E_i(x)'] + (1 - \alpha_i^v) \Sigma_i$$

Phoneme Independent (PI) system



Phoneme Dependent Direct (PD) SV System



Experimental Setting

- **Feature Extraction – 12 MFCC concatenated w. 12 delta MFCC**
- **Training duration: 5, 10 and 20 seconds**
- **Test duration: 2-3 seconds**

- **Database – TIMIT + NTIMIT (only male speakers)**
 - 88 speakers for UBM training**
 - 350 speakers for true speakers**
 - 350 imposters**

- **Phoneme segmentation – using SI phoneme HMM with 3 states and phoneme sequence files “.phn”**

(Experiments 1a)
Phoneme Dependent Direct (PD) system
vs.
Phoneme Independent (PI) system
Clean Speech

TIMIT – clean speech

Train Duration	PI GMM Size	PI (EER in %)	DP (EER in %)
5 secs	16	4.28	8.57
10 secs	16	2.28	4.42
20 secs	32	1.71	1.71

- PD system: 2-size GMMs for vowels
uni-modal Gaussians for consonants

(Experiments 1b)
Phoneme Dependent Direct (PD) system
vs.
Phoneme Independent (PI) system
Telephone Speech

NTIMIT – telephone lines

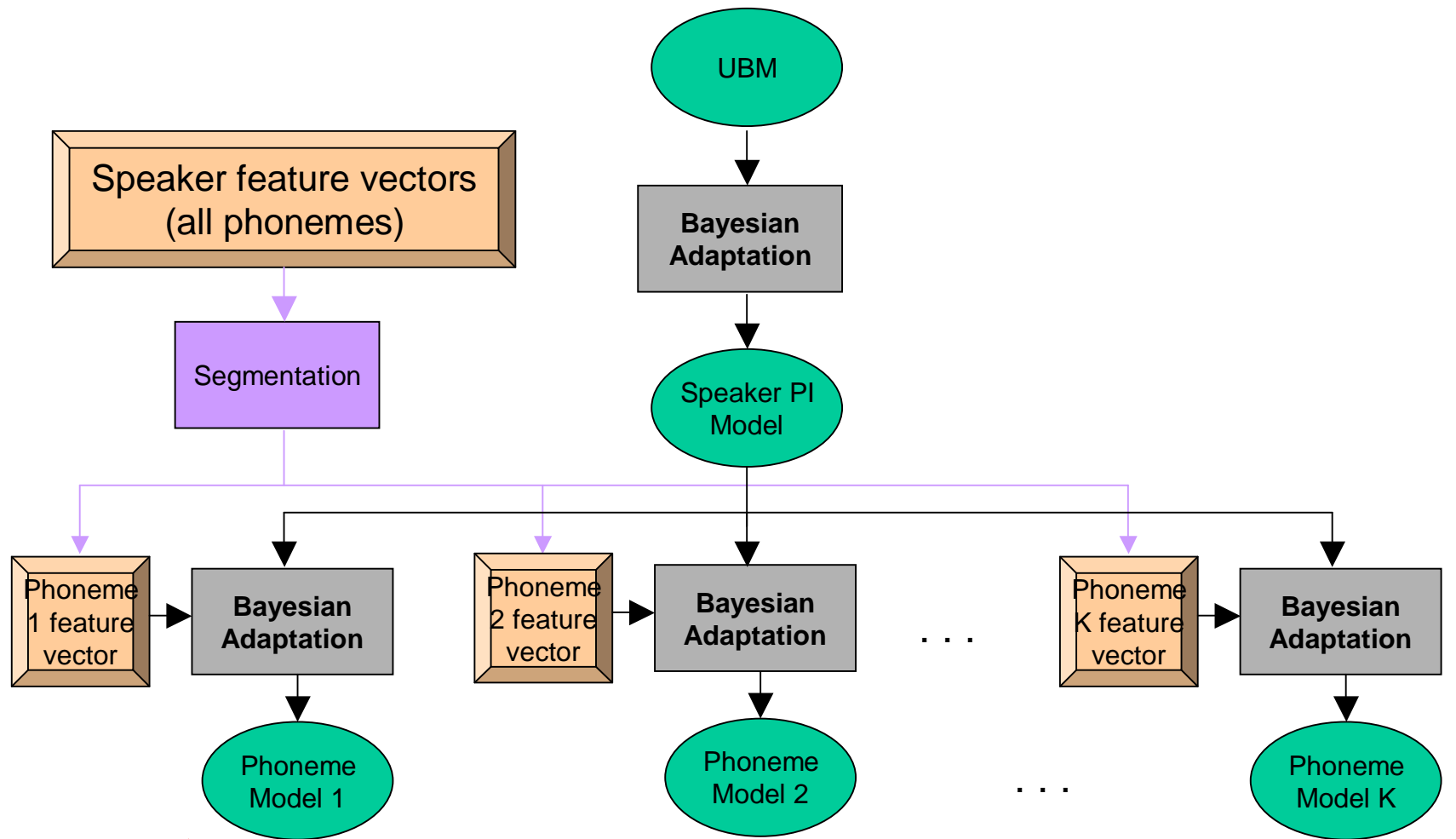
Train Duration	PI GMM Size	PI (EER in %)	PD (EER in %)
5 secs	16	25.28	30.42
10 secs	16	19	25.71
20 secs	32	13	20.14

- PD system: 2-size GMMs for vowels
uni-modal Gaussians for consonants

Why direct phoneme SV systems perform less well than phoneme independent SV systems?
(A counter-intuition observation found also in other works so far)

- **The phonetic segmentation layer introduces errors?**
- **Each phoneme model is built using a relatively small amount of data?**
- **Phoneme not included in the training data are ignored in the testing phase?**
- **Discards events connected with correlation and transition between different phonemes?**

New approach – Adapted Phoneme GMM (AP-GMM)



AP-GMM Training Procedure 1/3

Construct PI GMM for a specific speaker using the whole training data of the speaker (including all phonemes):

$$\lambda_s = \{w_i^s, \mu_i^s, \Sigma_i^s\}_1^M$$

Cluster the training feature vectors into K phoneme groups:

$$X_k = \{x_1, \dots, x_{T_k}\} \quad k = 1, \dots, K$$

AP-GMM Training Procedure 2/3

- **“Expectation” step:**
Use data for each phoneme k , to compute new parameters:

$$P(i | x_t, \lambda_s) = \frac{w_i g_i(x_t)}{\sum_{j=1}^M w_j g_j(x_t)}$$

$$n_{i,k} = \sum_{t=1}^{T_k} P(i | x_t, \lambda_s)$$

$$E_{i,k}(x) = \frac{1}{n_{i,k}} \sum_{t=1}^{T_k} P(i | x_t, \lambda_s) x_t$$

$$R_{i,k}(x) = \frac{1}{n_{i,k}} \sum_{t=1}^{T_k} P(i | x_t, \lambda_s) x_t x_t'$$

AP-GMM Training Procedure 3/3

- **“Combination” step:**
Combine for each phoneme k , the parameters estimated from its data with the parameters of the GMM model subject to adaptation:

$$\alpha_{i,k} = \frac{n_{i,k}}{n_{i,k} + r_k}$$

$$\tilde{w}_{i,k} = [\alpha_{i,k} n_{i,k} / T_k + (1 - \alpha_{i,k}) w_i] \gamma_k$$

$$\tilde{\mu}_{i,k} = \alpha_{i,k} E_{i,k}(x) + (1 - \alpha_{i,k}) \mu_i^s$$

$$\tilde{\Sigma}_{i,k} = \alpha_{i,k} [R_{i,k}(x) - E_{i,k}(x)E_{i,k}(x)'] + (1 - \alpha_{i,k}) \Sigma_i^s$$

(Experiment 2a)
Adapted Phoneme (AP) system
vs.
Phoneme Independent (PI) system
clean speech

TIMIT – clean speech

Train Duration	PI GMM Size	AP (EER in %)	PI (EER in %)
5 seconds	16	4.14	4.28
10 seconds	16	2.57	2.28
20 seconds	32	1.71	1.71

- AP system: Fixed relevance factor: $r_k=12$.

(Experiment 2b)
Adapted Phoneme (AP) system
vs.
Phoneme Independent (PI) system

NTIMIT – telephone lines

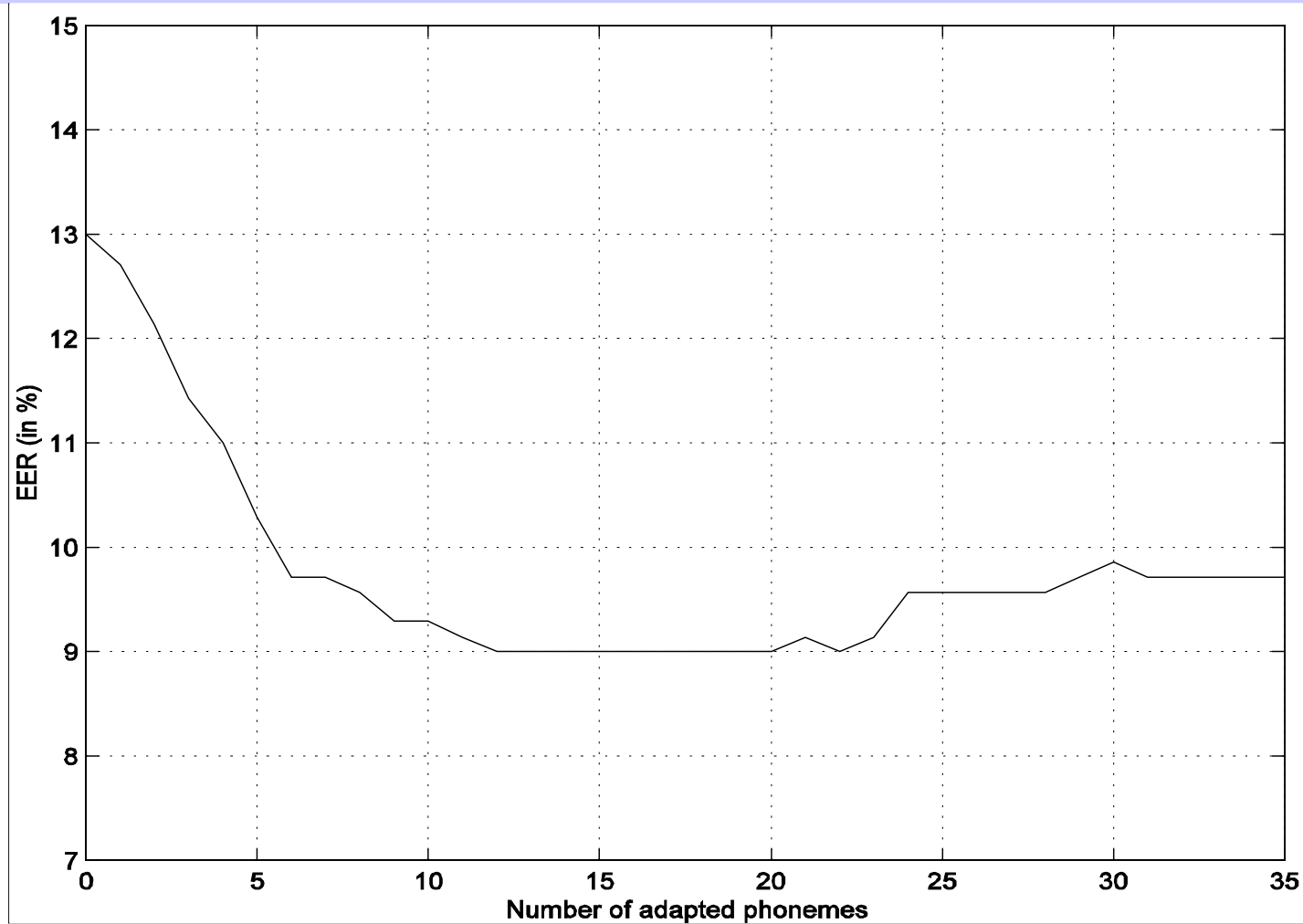
Train Duration	PI GMM Size	AP (EER in %)	PI (EER in %)
5 seconds	16	20.71	25.28
10 seconds	16	14.85	19
20 seconds	32	9.71	13

- AP system: Fixed relevance factor: $r_k=12$.

Experiment 3

Adapting Subsets of Phonemes

NTIMIT - 20 seconds experiment



Experiment 4

Adapting subsets of parameters

NTIMIT - 20 seconds experiment

	w	m	v	w,m	w,v	m,v	w,m,v
EER (in %)	11.28	9.71	8.71	8.85	10.28	9.85	9.71

w- weights

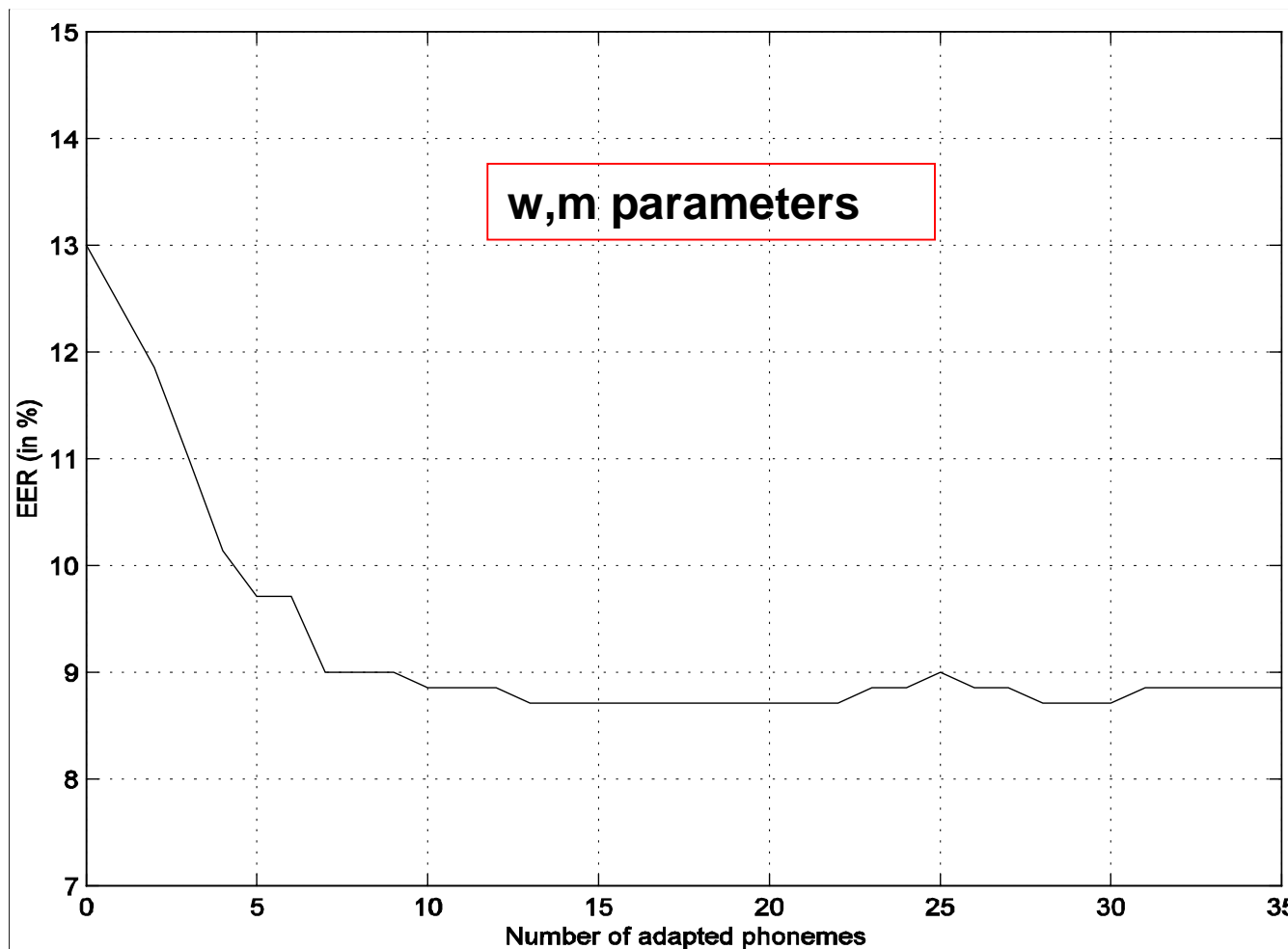
m- mean values

v- variance values

Experiment 5a

Adapting Subsets of Phonemes and Parameters

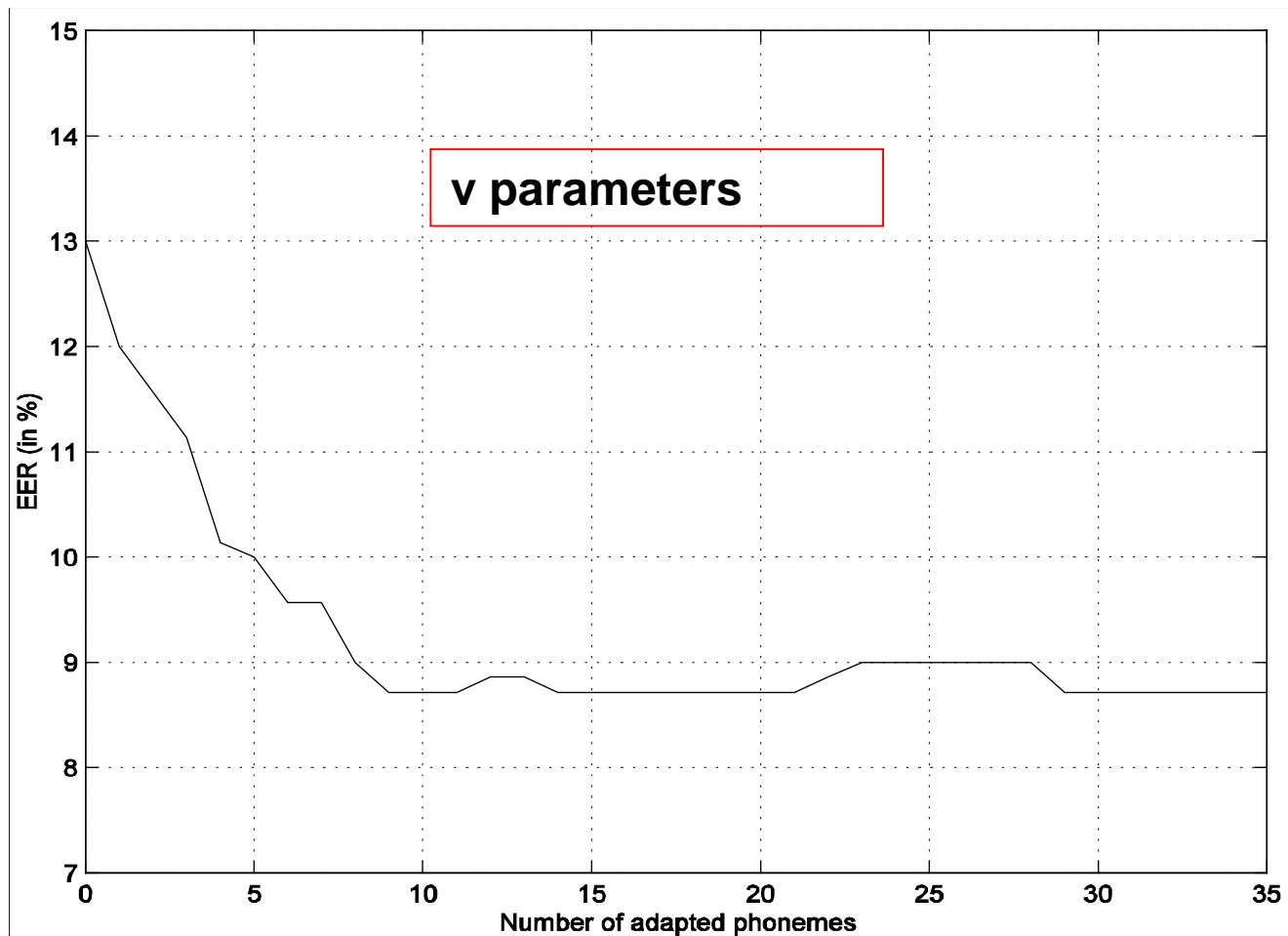
NTIMIT - 20 seconds experiment



Experiment 5b

Adapting Subsets of Phonemes and Parameters

NTIMIT - 20 seconds experiment



Conclusions

- **A new approach introduced - AP GMM.
The AP-GMM perform better than PI or DP GMMs.**
- **The extent of the adaptation is controlled by an adaptation factor α which is a function of the amount of data available on the phoneme.**
- **Adaptation of subsets of phonemes and parameters may reduce models size without degradation and with even improvement of verification scores.**

References

- D. A. Reynolds [Speech Communication 1995]
- I. Margin-Chagnoleau, J. F. Bonastre, F. Bimpot [Eurospeech 1995]
- M. Newman et al. [ICSLP 1997]
- J. Ø. Olsen [Pattern Recognition Letters 1997]
- R. Auckenthaler, E. S. Parris, M. J. Carey [ICASSP 1999]
- D. A. Reynolds, T. F. Quatieri, R. B. Dunn [DSP Review J. 2000]

This presentation

- D. Gutman and Y. Bistriz [Eusipco 2002]
- D. Gutman and Y. Bistriz “Speaker verification by adapted phoneme Gaussian mixture models” submitted 2003.