Separate MAP Adaptation of GMM Parameters for Forensic Voice Comparison on Limited Data

Chee Cheun Huang, Julien Epps and Ewald Enzinger

School of Electrical Engineering & Telecommunications, The University of New South Wales, Sydney, Australia

National ICT Australia (NICTA), Sydney, Australia

Paradigm for evaluation of FVC evidence

- Likelihood-ratio framework:
  - Statement of strength of the evidence as an answer to a specific question

$$\mathrm{LR} = \frac{p(E \mid H_p)}{p(E \mid H_d)}$$

- Quantitative measurements, statistical models, databases representative of the relevant population

- Testing of validity and reliability under conditions reflecting those of the case

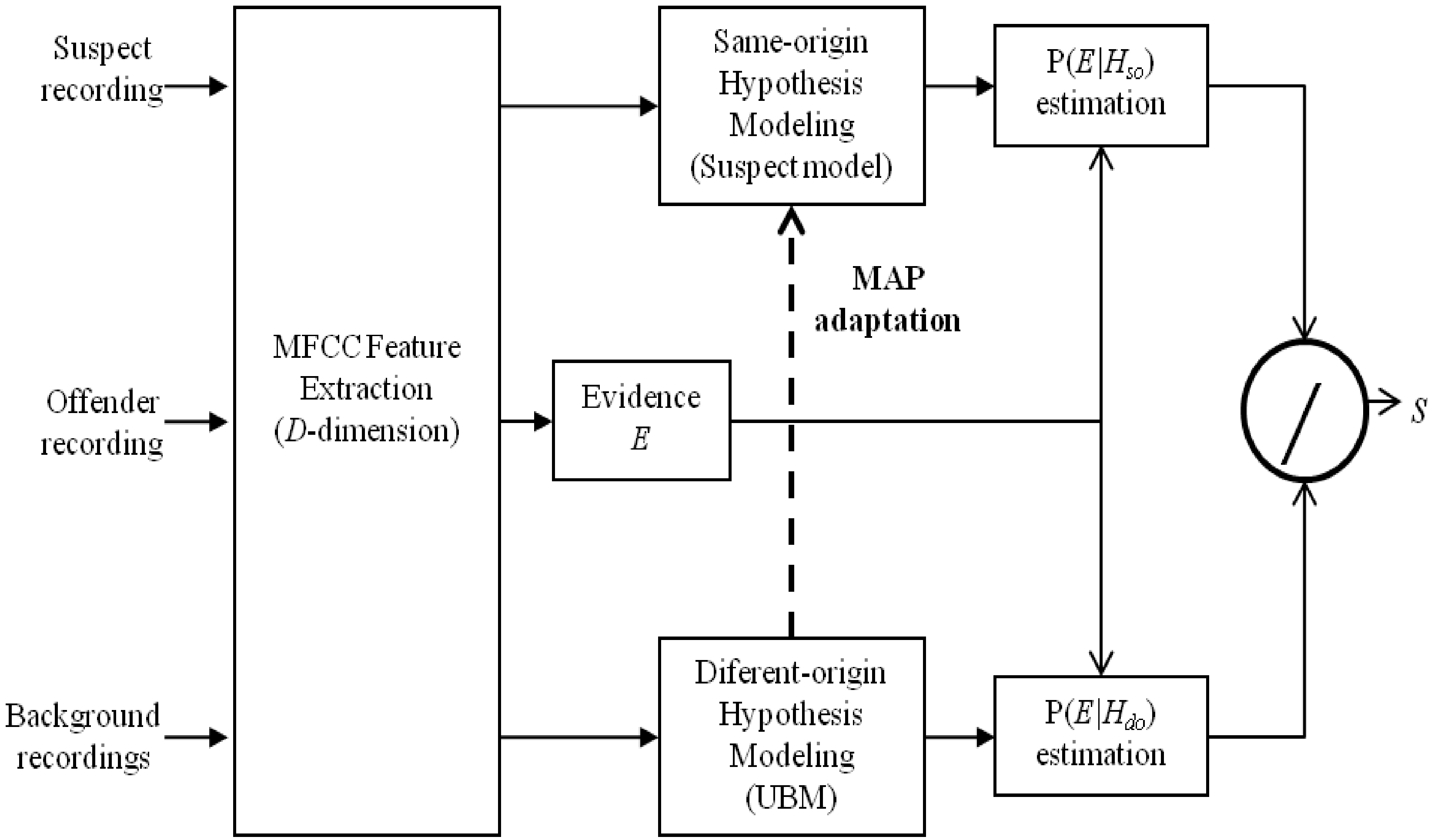
GMM-UBM statistical modeling (1)

- Gaussian mixture model-Universal background model (GMM-UBM) often used in automatic forensic-voice-comparison (FVC) systems

1. Feature extraction
2. Train GMM λ_{UBM} from sample of relevant population
   - Model of the defence hypothesis H_d
3. Adapt suspect speaker GMM λ_{sp} from UBM using maximum a-posteriori (MAP) adaptation
   - Model of the prosecution hypothesis H_p
4. Calculate score
5. Transform score to likelihood ratio using calibration

GMM-UBM statistical modeling (2)

Gaussian mixture model-Universal background model system



Maximum a-posteriori (MAP) adaptation

- Initialize suspect GMM parameters $\lambda_{\mathrm{sp}} = (w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)_{i=1,\ldots,M}$ from universal background model GMM λ_{UBM}
- Maximum a-posteriori (MAP) adaptation
  - Calculate occupancy and sufficient statistics:

$$E_i(\mathbf{x}_t) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t) \mathbf{x}_t \qquad \Pr(i \mid \mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j=1}^{M} w_j p_j(\mathbf{x}_t)}$$

$$E_i(\mathbf{x}_t^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t) \mathbf{x}_t^2 \qquad n_i = \sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t)$$

  - Update parameters:

$$\hat{w}_i = \left[\alpha_i n_i / T + (1-\alpha_i) w_i\right] \gamma$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i E_i(\mathbf{x}_t) + (1-\alpha_i) \boldsymbol{\mu}_i$$

$$\hat{\boldsymbol{\sigma}}_i = \alpha_i E_i(\mathbf{x}_t^2) + (1-\alpha_i)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2$$

$$\alpha_i = \frac{n_i}{n_i + r}$$

r ... relevance factor

Motivation

- Conventionally, only mean parameters adapted
  - ➢ Comparison of mean / variance / weight / full MAP adaptation

- Modification: Separate MAP adaptation
  - – Often short suspect and/or offender samples
  - – Problem of overfitting to suspect data

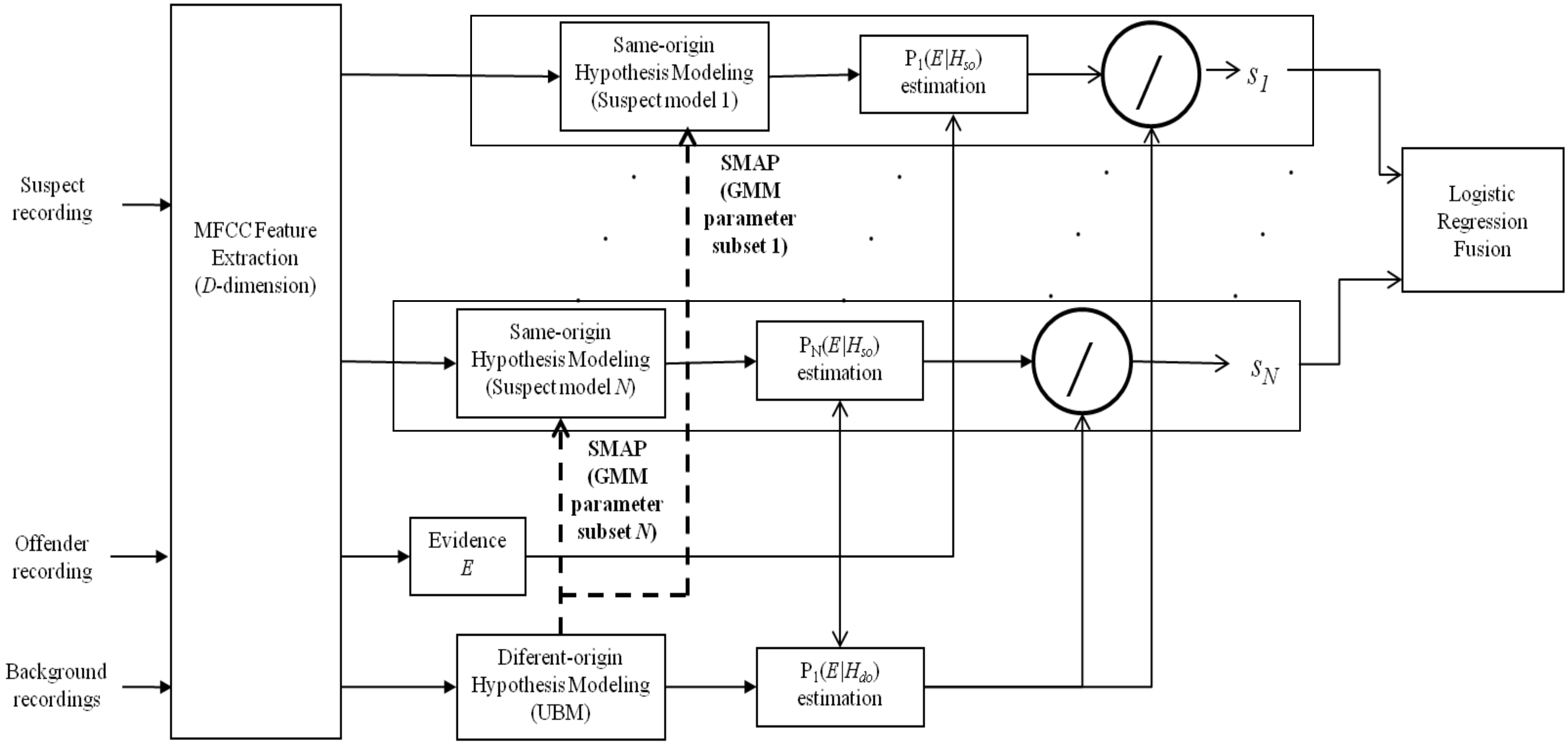
  - ➢ Adaptation that operates on fewer parameters than mean-only MAP adaptation?

Separate MAP Parameter Adaptation (1)

- Define N non-overlapping subsets of GMM mean parameters: $S_n \subset \{1,2,\ldots,D\},\ \bigcup_{n=1}^{N} S_n = \{1,2,\ldots,D\},\ \bigcap_{n=1}^{N} S_n = \varnothing$
- Each subset forms separate MAP system:
  - Perform mean-only MAP adaptation
    - Calculate occupancy and sufficient statistics
    - Update mean parameters
  - "Reset" parameters j not in S_n

    $$\hat{\mu}_i(j) = \mu_i(j), \forall i$$
- Logistic regression fusion of all N separate MAP systems

Separate MAP Parameter Adaptation (2)



Data

- 60 female Standard Chinese speakers
- Split into 3 groups of 20 speakers
  - background set
  - development set
  - test set
- Information-exchange task over telephone
- High quality studio recordings
- Two recording sessions separated by 2–3 weeks

`http://databases.forensic-voice-comparison.net/`

Experimental setup

- GMM-UBM FVC system
  - Entire speech-active portion of recording
  - 16 MFCC + 16 delta (Δ) coefficients (D=32)
  - 512 Gaussian mixture components (UBM)
  - 3 MAP iterations

- Logistic regression calibration and fusion

- Metric of validity / accuracy:
  - log-likelihood ratio cost (C_{llr}) metric:

$$C_{llr} = \frac{1}{2}\left[\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}}\log_2\left(1+\frac{1}{\mathrm{LR}_{ss,i}}\right)+\frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}}\log_2\left(1+\mathrm{LR}_{\mathrm{ds,j}}\right)\right]$$

Results: Comparisons of MAP variants

Individual systems	C_{llr}
Mean-only adaptation	0.196
Variance-only adaptation	0.221
Weight-only adaptation	0.848
Full adaptation	0.302

Fusion	C_{llr}
Fusion mean-only + variance-only adaptation	0.183
Fusion mean-only + weight-only adaptation	0.187
Fusion variance-only + weight-only adaptation	>1
Fusion mean-only + variance-only + weight-only adaptation	**0.182**

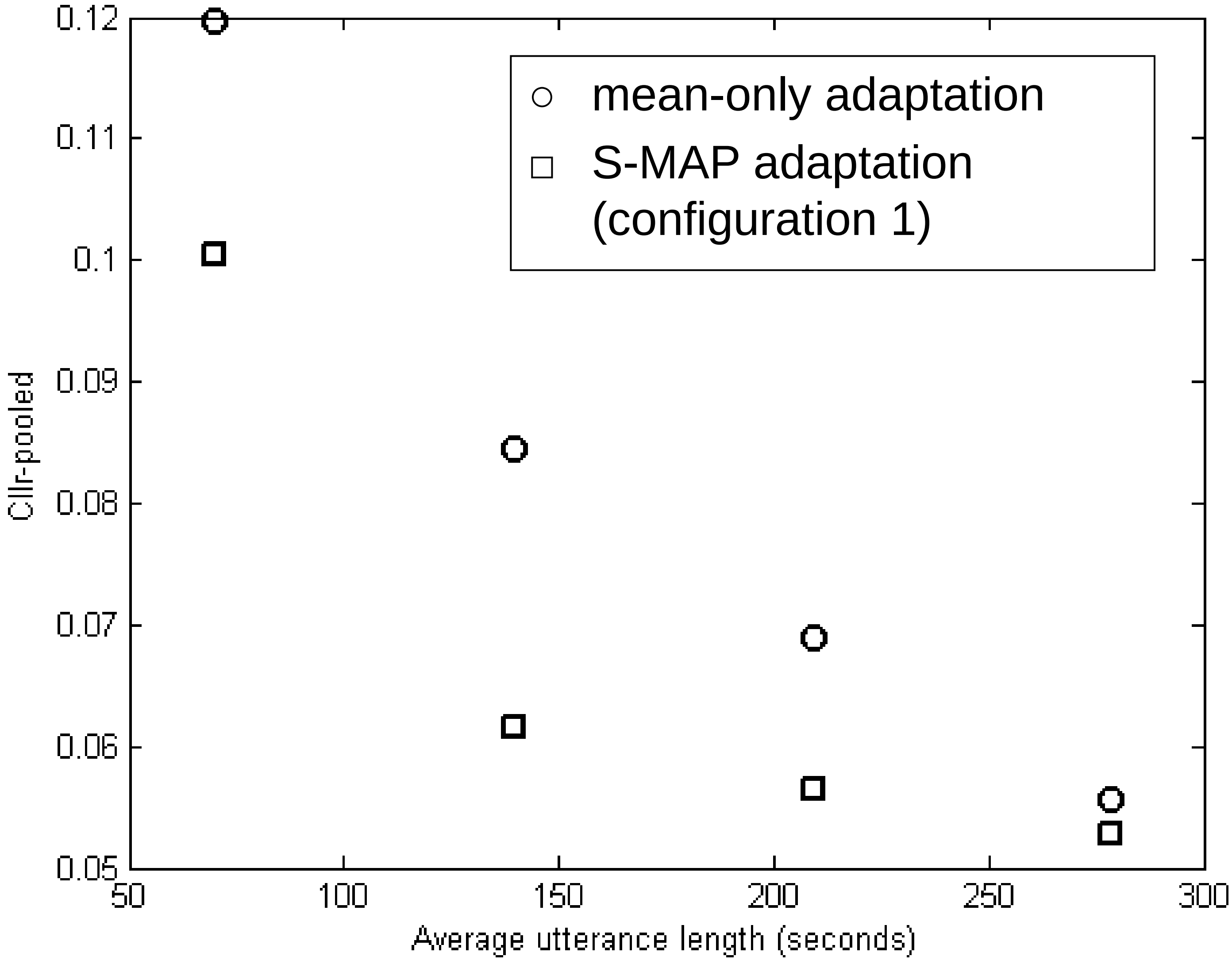
Fused system: **6.8%** improvement over mean-only

Results: Separate MAP

- 2 Separate MAP (S-MAP) configurations:
  - Configuration 1: N=2
    S_1={MFCC$_1$,…, MFCC$_{16}$}, S_2={Δ_1,…, Δ_{16}}
  - Configuration 2: N=32
    S_1={MFCC$_1$},…, S_{16}={MFCC$_{16}$},
    S_{17}={Δ_1},…, S_{32}={Δ_{16}}

	C_{llr}
Mean-only adaptation	0.056
S-MAP configuration 1	0.053
S-MAP configuration 2	**0.042**

Results: S-MAP v mean-only in **limited data**



Conclusion

- Mean / variance / weights / full MAP adaptation:
  - Mean-only adaptation: best individual performance
  - Fusion with other variants can improve performance
- Separate MAP adaptation can achieve substantial improvements compared with the traditional mean-only MAP adaptation
- For increasingly small amounts of suspect speaker data, there seems to be an increasingly large advantage of S-MAP

Thank You!!