Bayesian Method in Speaker Verification

Chengyuan Ma and Jinyu Li

School of Electrical and Computer Engineering

Georgia Institute of Technology

Outline

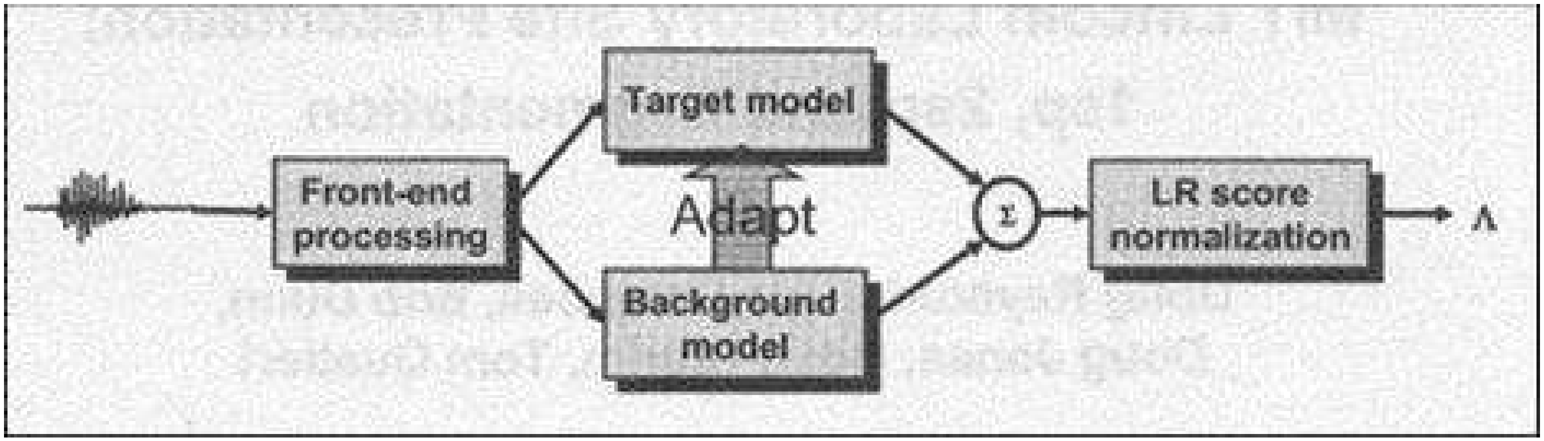
- Speaker Verification System
- MAP adaptation of Speaker Model
- Experiment Setup, Conclusion and Future Work

Speaker Verification System

Speaker Verification System Framework

- Speaker Verification
  - Determines whether person is who they claim to be
  - User makes identity claim (one to one mapping)
- Three Main Components
  - Front-end processing (Feature Extraction)
  - Speaker and background modeling (Speaker Modeling)
  - Log-likelihood-ratio score normalization (Decision Making)

Speaker Verification System Framework (Cont.)



Speaker and Background Modeling

- GMM *(Gaussian Mixtures Model)* represent static acoustic event distribution for each speaker

$$p(o_t|\Lambda) = \sum_{m=1}^{M} \frac{w_m}{(2\pi)^{\frac{d}{2}}|\Sigma_m|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(o_t-\mu_m)^\tau \Sigma_m^{-1}(o_t-\mu_m)}$$

$$\sum_{m=1}^{M} w_m = 1 \quad \text{and} \quad w_m > 0$$

Λ is used to represent speaker model parameters. w_m is the m^{th} weight of Gaussian mixture $\mathcal{N}(o_t; \mu_m, \Sigma_m)$ with mean μ_m and covariance matrix Σ_m.

Georgia Institute of Technology

Speaker Recognition Scoring

$$p(O|\Lambda) = p(o_1, o_2, \cdots, o_T|\Lambda) = \prod_{t=1}^{T} p(o_t|\Lambda)$$

The score (average *Log-Likelihood Ratio* (LLR)) of a given test segment O is computed as follow,

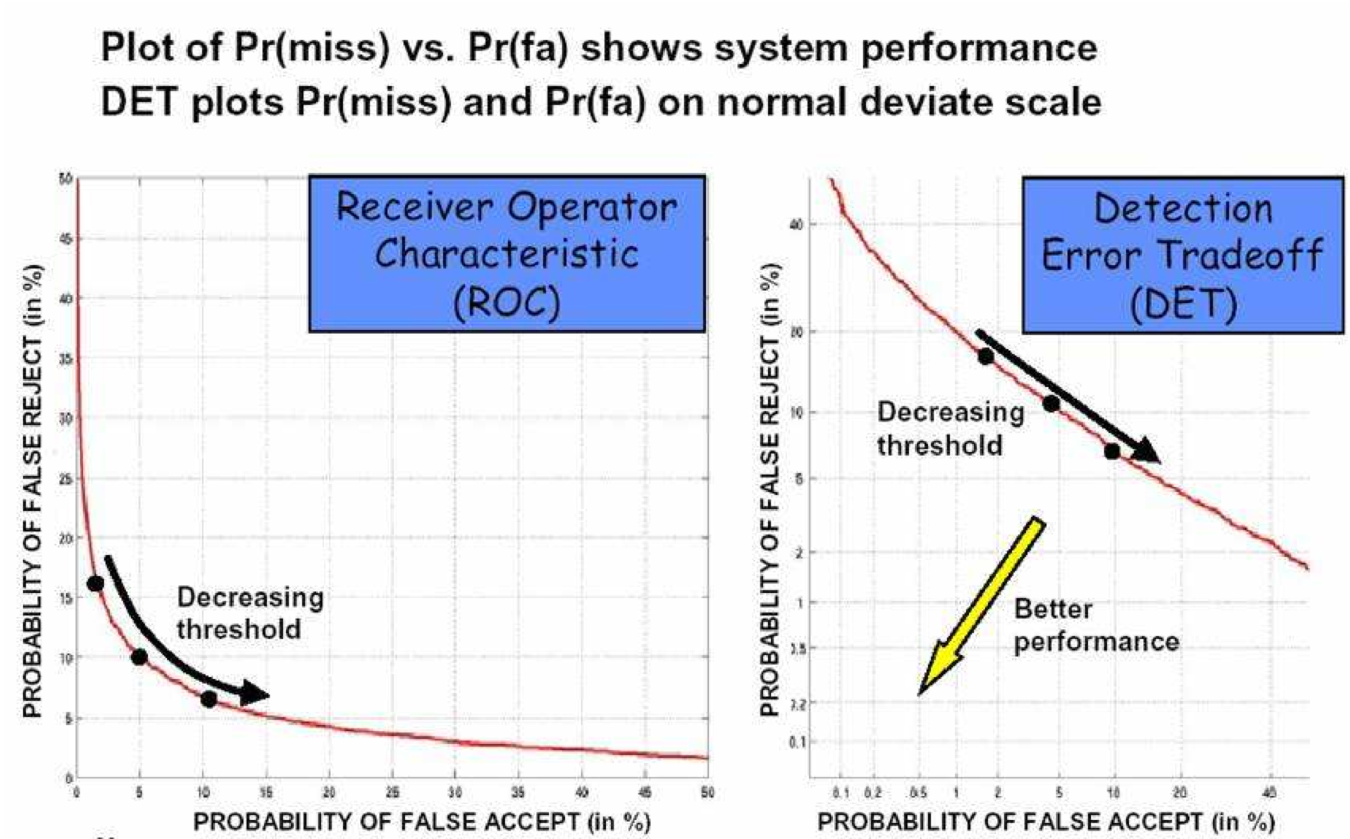
$$score = \frac{1}{T}\sum_{t=1}^{T}(\log p(o_t|\Lambda_{\mathrm{tar}}) - \log p(o_t|\Lambda_{\mathrm{UBM}}))$$

T is the length of the test segment, o_t is the feature vector at time t and Λ_{tar} and Λ_{UBM} are parameters of the target model and UBM (Universal Background Model) respectively.

Georgia Institute of Technology

Performance Measurement of Speaker Verification System

- ROC and DET plots are used for performance measurement



MAP adaptation of Speaker Model

Speaker Model Adaptation Using MAP Method

- EM training is reliable only when sufficient training data are available
- If only a small amount of training data, adaptation method will be used for model training.
- Background model is regarded as prior information for each speaker, speaker-specified training data are the observation samples, these two can be combined in the Bayesian framework.

Georgia Institute of Technology

Speaker Model Adaptation Using MAP Method (Cont.)

- Speaker model parameters to be estimated

$$\Lambda = (w_1, w_2, \cdots, w_M, \mu_1, \mu_2, \cdots, \mu_M, \Sigma_1, \Sigma_2, \cdots, \Sigma_M)$$

- MAP estimation of speaker model parameters

$$\theta_{MAP} = \arg\max_{\theta} p(O|\Lambda)g(\Lambda)$$

Georgia Institute of Technology

Speaker Model Adaptation Using MAP Method (Cont.)

- Prior distribution of mixture weights is a conjugete density such as Dirichlet density

$$g(w_1, w_2, \cdots, w_M) \quad \propto \quad \prod_{m=1}^{M} w_m^{v_k - 1}$$

where $v_k > 0$ are the parameters for the Dirichlet density. [a]

[a] J.-L. Gauvain and C.-H. Lee, "Maximum a Posterior Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, April 1994.

Speaker Model Adaptation Using MAP Method (Cont.)

- Prior distribution of (μ_m, Σ_m) is a normal_Wishart density

$$
\begin{aligned}
g(\mu_m, \Sigma_m) \quad &\propto \quad |\gamma_m|^{(\alpha_m - p)/2} \\
&\exp\left[-\frac{\tau_m}{2}(\mu_m - m_m)^t \Sigma_m (\mu_m - m_m)\right] \\
&\exp\left[-\frac{1}{2} tr(u_m \Sigma_m)\right]
\end{aligned}
$$

where $(\tau_m, m_m, \alpha_m, u_m)$ are the prior density parameters such that $\alpha_m > p - 1, \tau_m > 0$, m_m is a vector of dimension p, and u_m is a $p * p$ positive definite matrix.

Speaker Model Adaptation Using MAP Method (Cont.)

- Practical method for parameter adaptation only the mean vectors will be updated

$$\hat{\mu}_m = \alpha_m E_m(O) + (1 - \alpha_m)\mu_m$$

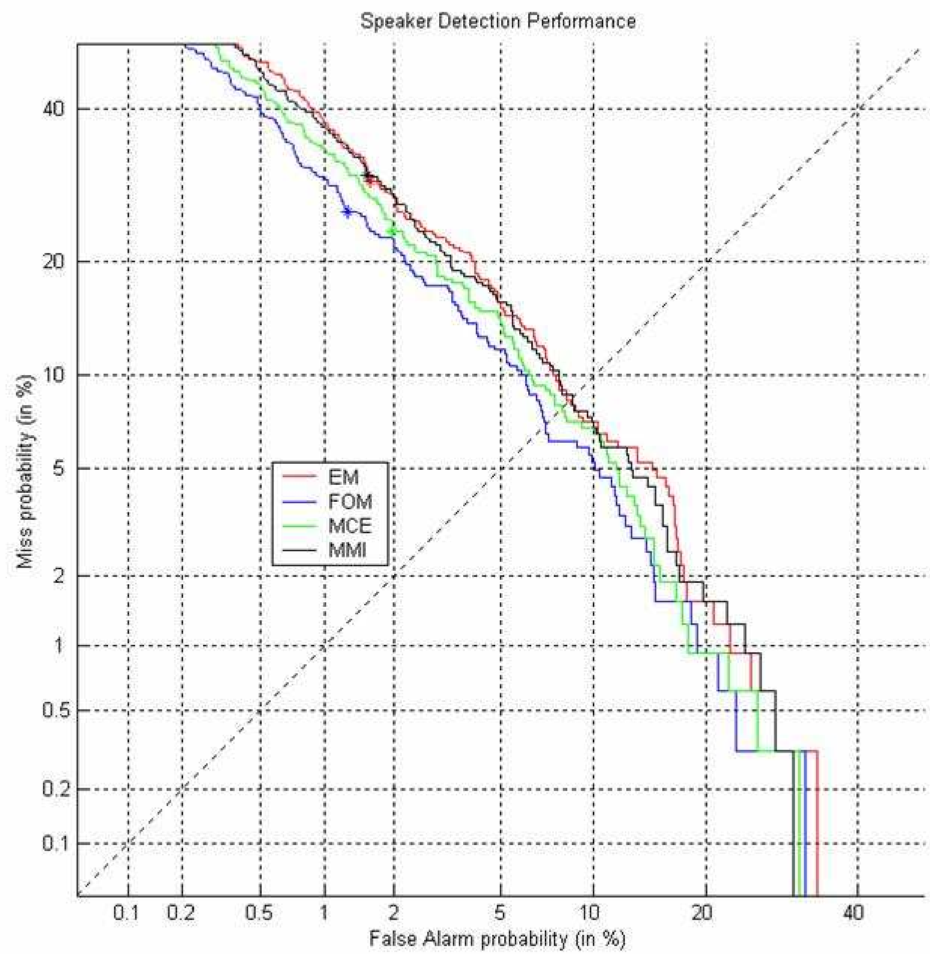
Georgia Institute of Technology

Experiment Setup, Conclusion and Future Work

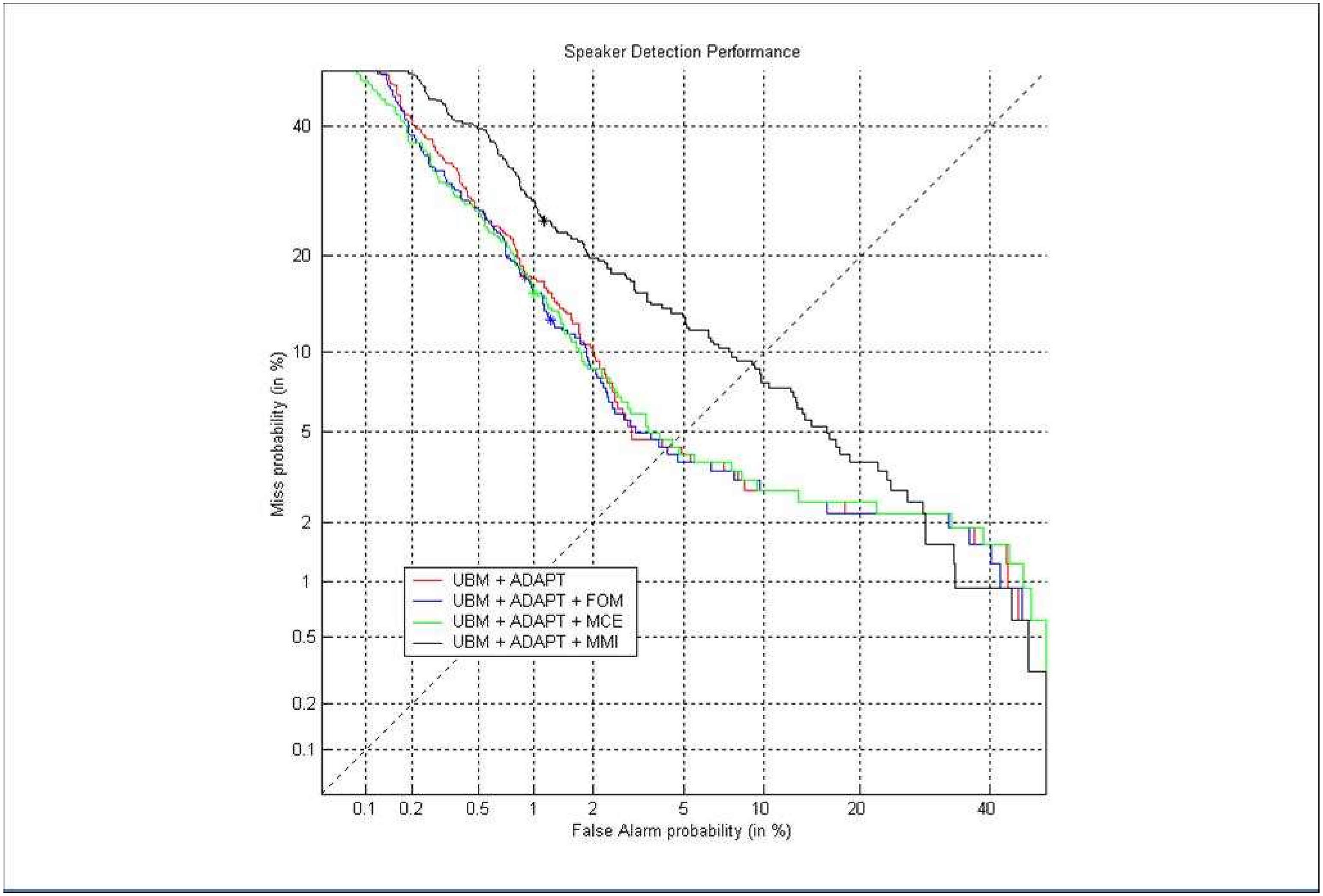
Experiment Setup

- 1996 NIST(National Institute of Standards and Technology) dataset was used for experiment,
- 225 Speakers in the Corpus
- 1 minutes for training and 15 second for testing for each speaker

System Performance using EM Training



System Performance using MAP Adaptation



Future Work

- Acoustic features robust against noisy environment and different channels
- Incorporation dynamical acoustic features and prosodic features
- Higher level information
- novel probabilistic models for speaker modeling

Q & A