Computationally Efficient Speaker Identification for Large Population Tasks using MLLR and Sufficient Statistics

A. K. Sarkar • S. Umesh • S. P. Rath

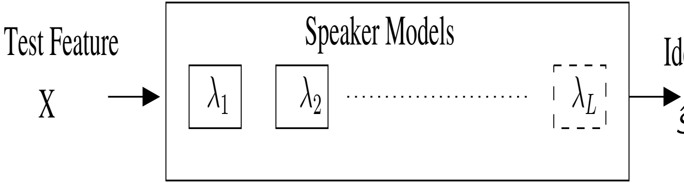
Department of Electrical Engineering
Indian Institute of Technology Madras

June 28^{th}, 2010

Outline

- Overview of Speaker Identification (Closed-Set)
- MAP adaptation and Top-C mixtures based Likelihood Estimation
- Speaker Identification using MLLR matrices
- Efficient Likelihood Calculation using MLLR matrices
- Comparison of GMM-UBM with Fast MLLR system
- Cascade Identification System to improve Performance
- Summary

Overview of Speaker Identification (Closed-Set)



$$\hat{S} = \max_{1 \leq i \leq L} Pr(X|\lambda_i)$$

- Evaluate likelihood for all speaker models
  - Computationally expensive for large databases.

MAP adaptation and Top-C mixtures based Likelihood Estimation

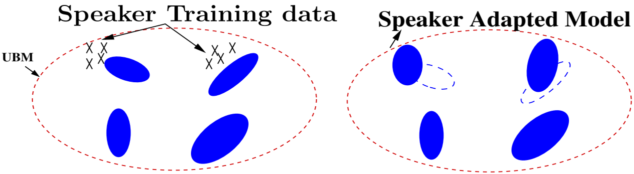


Figure: *Adapted Speaker dependent model with MAP.*

- Top-C scoring steps
  1. Align test data w.r.t UBM and find Top-C mixtures/feature vector
  2. Evaluate Top-C mixtures for all speaker models
     i.e. $2048 + L \times C$ mixtures for L speaker models
- As L becomes large computation grows.

Speaker Model Training using MLLR adaptation

Propose: Use of MLLR to adapt "UBM mean" to "speaker-model mean"

- MLLR matrix is estimated using speaker's training data w.r.t UBM.

$$\hat{\mu}_{spkr} \;=\; W_{spkr}\,\mu_{ubm} \;; \qquad spkr = 1, 2, \ldots, L$$

- Speaker is characterized by MLLR matrix, W_{spkr}.
  - No model is formed.

Speaker Identification using MLLR matrices

Spkr 1	Spkr 2	Spkr 3	- - - - - - - - - - -	Spkr L
W_1	W_2	W_3		W_L

- For a given unknown Test utterance and MLLR matrices of Speakers
- We identify speaker as:

$$\hat{S} = \max_{1 \le i \le L} Pr(X | \lambda_{UBM}, W_i)$$

- It looks like we still need to calculate likelihood for all speakers!
  - but this can be efficiently done.

Speaker Identification using MLLR and EM

$$Q(W_s, I) = \sum_{j=1}^{M} Pr(j|X, \lambda_{UBM}, I)\ log Pr(X, j|\lambda_{UBM}, W_s)$$

$$\hat{S} = \arg\max_{W_s} Q(W_s, I)$$

where,

$W_s \Rightarrow$ MLLR matrix for speaker, s

$I \Rightarrow$ identity matrix

- W_s can be represented as

$$W_s = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

Efficient Likelihood Calculation using MLLR matrices

- Do one alignment of test data w.r.t UBM (same as MAP+Top-C)
- Compute two statistics over all Gaussian components in the GMM-UBM using the test utterance, X, only once

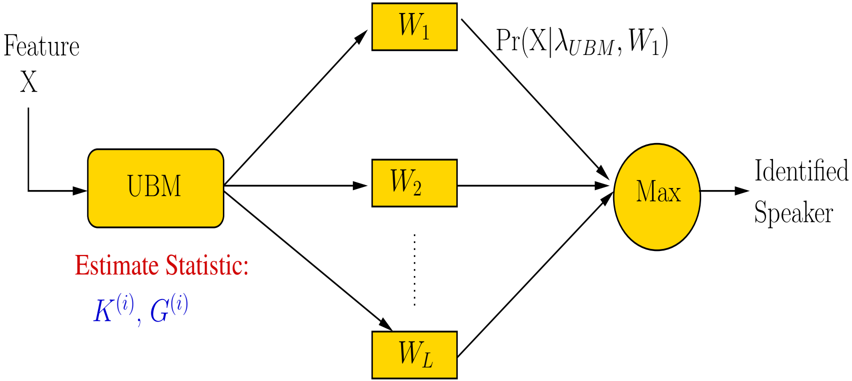
$$K^{(i)} = \sum_{j=1}^{M} \frac{\mu_j^{(i)}}{\sigma_j^{(i)^2}} \sum_{t=1}^{T} \gamma_j(t) x^{'}(t); \qquad G^{(i)} = \sum_{j=1}^{M} \frac{1}{\sigma_j^{(i)^2}} \mu_j \mu_j^{'} \sum_{t=1}^{T} \gamma_j(t)$$

- **Using** $K^{(i)},\ G^{(i)}$

  ▷ only matrix multiplication to get speaker model likelihood

$$\hat{S} = \arg\max_s \underbrace{\left\{ -\frac{1}{2} \left\{ \sum_{i=1}^{D} (w_{s,i} G^{(i)} w_{s,i}^{'} - 2K^{(i)} w_{s,i}^{'}) \right\} \right\}}_{Pr(X|\lambda_{UBM}, W_s)}$$

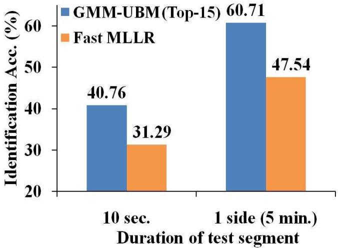
Illustration of Fast MLLR Speaker Identification System



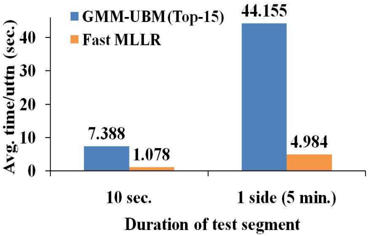
$$\hat{S} = \arg\max_{s} \left\{ -\frac{1}{2} \left\{ \sum_{i=1}^{D} \left(w_{s,i} G^{(i)} w_{s,i}^{'} - 2K^{(i)} w_{s,i}^{'} \right) \right\} \right\}$$

Comparison of GMM-UBM with Fast MLLR system

Performance

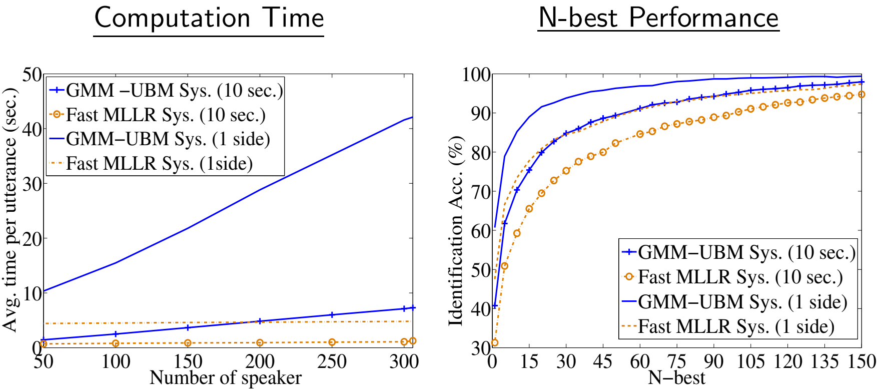


Computation Time



- 306 (122 Male, 184 female) speakers are used for evaluation (Closed-Set identification) from NIST 2004 SRE.
- Fast MLLR system performs poorer than GMM-UBM.
- But Fast MLLR system faster than GMM-UBM system.
- Longer utterance ▷ more gain in computation time.

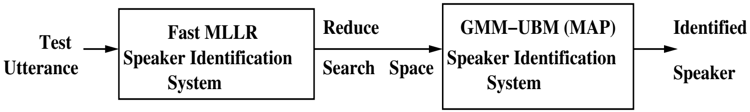
Analysis of Computation Complexity & Performance

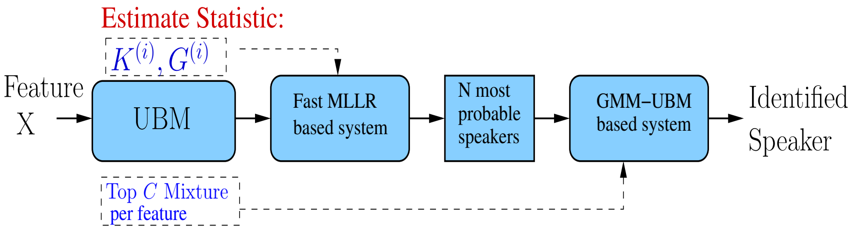


(a) (b)

- Fast MLLR system: Time taken to identify speaker does not increase significantly as database size increases (Fig. a).
- N-best performance of both systems converge as N increases (Fig. b).

Cascade Identification System to improve Performance

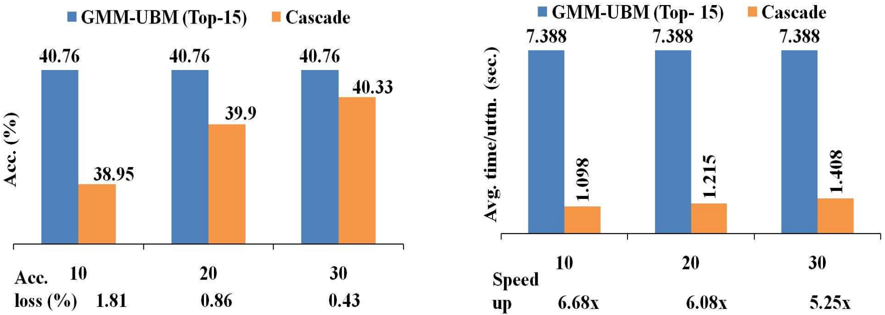




Requires only one alignment of test data w.r.t UBM

Trade-off between Computation Complexity & Performance

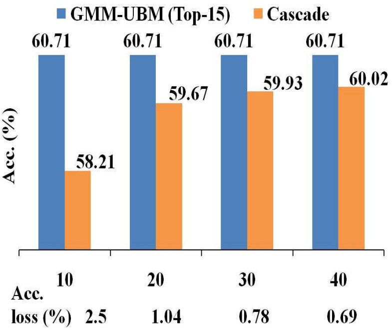
- Experiment Result for 10sec. test segment

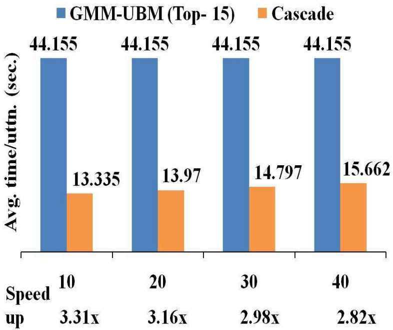


- 306 (122 Male, 184 female) speakers (1163 test uttn.) are used for evaluation (Closed-Set identification) from NIST 2004 SRE

Trade-off between Computation Complexity & Performance

- Experiment Result for 1-side test segment





- For 1-side test segment cascade system becomes comparatively less faster than 10 sec. utterance due to the slower backend GMM-UBM identification system.

Summary

- As we increase value of N-best the performance of cascade system comes closer to GMM-UBM system.
- Tuning the value of N: a compromise between accuracy loss and system speed that can be achieved.
- For N=20, cascade system with 306 speakers.
  - For 10 sec. $\Rightarrow$ 6.08$\times$ faster than GMM-UBM and 0.86% loss in Acc.
  - For 1-side $\Rightarrow$ 3.16$\times$ faster than GMM-UBM and 1.04% loss in Acc.

Experimental setup

- Front End
  - 20 ms frames for every 10 ms
  - 21 mel filters over 300 – 3400 Hz
  - MFCC with (C_1 to C_{13} with Δ and $\Delta\Delta$ excluding C_0)
  - Frame Selection: Gaussian modeling of the energy component of frames
  - 0-mean and 1-Variance utterance level
- Background Modeling
  - Speaker Independent UBM (2048 mixt.) with diagonal covariance matrix
  - Training Data: NIST 2002 SRE and Switchboard-1 Release-2
- Evaluation: 1 side trn.: 10s & 1 side test condition of NIST-04 SRE
  - Speaker model & MLLR matrix using 1-iteration of MAP and MLLR w.r.t UBM (only mean adaptation) respectively.
  - Relevance factor, 16 is used during MAP.
  - There are 306 (122 male, 184 female) speakers for evaluation.
- Computer used for the experiment
  - Intel Quad Core (Q9550) processor @ 2.83Hz
  - 8 GB RAM

Thank You!