Investigation of Speaker-Clustered Background Models based on Vocal Tract Lengths and MLLR matrices for Speaker Verification

A. K. Sarkar • S. Umesh

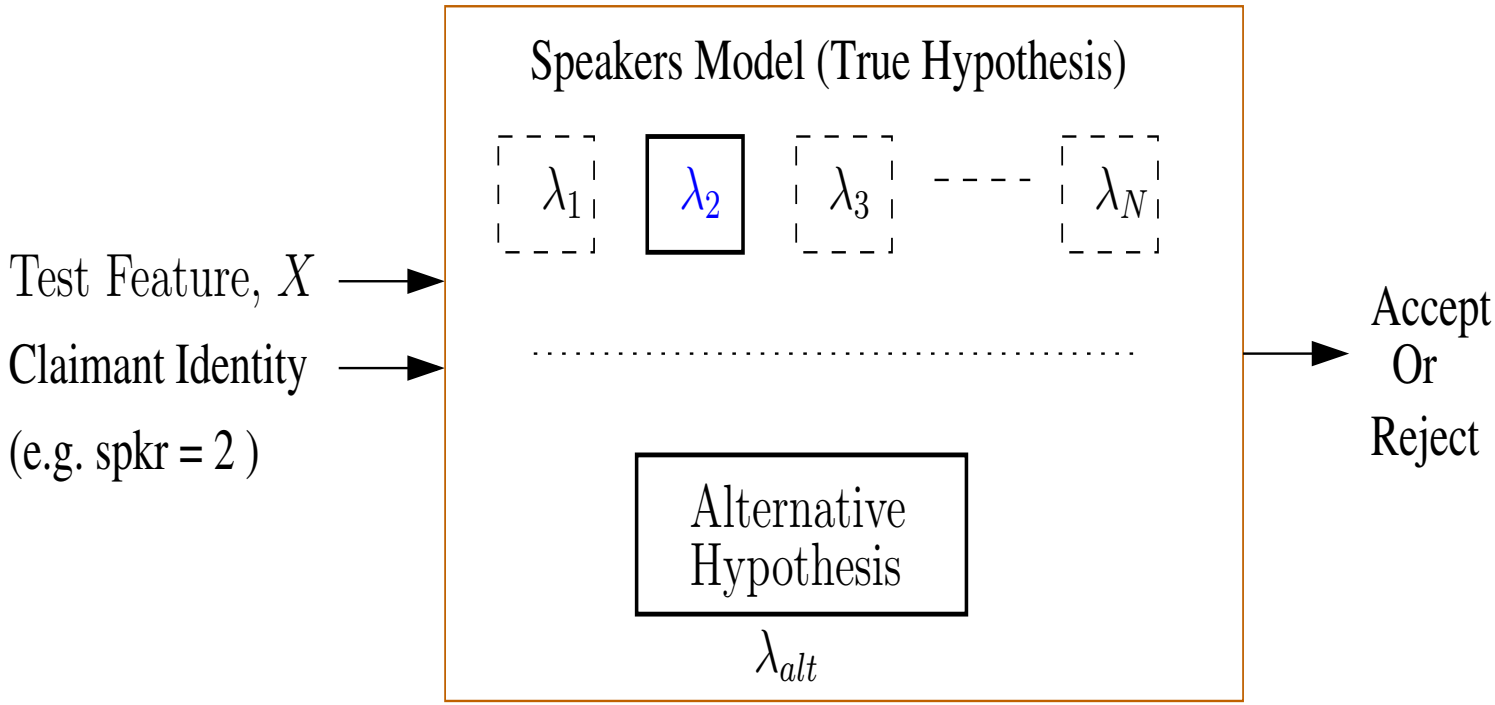
Department of Electrical Engineering
Indian Institute of Technology Madras
June 28^{th}, 2010

Outline

- Overview of Speaker Verification
- Propose use of Speaker Cluster-Wise Background Model (SC-BM)
  - using VTLN and MLLR super-vector
- Building Speaker Cluster-Wise Background Model (SC-BM)
- Comparison of performance using a single gender independent UBM
- Verification using Gender-wise SC-BM Vs gender dependent UBM
- Summary

Overview of Speaker Verification

- Accept or reject the claimant based on their claim.
- It is a binary decision problem.



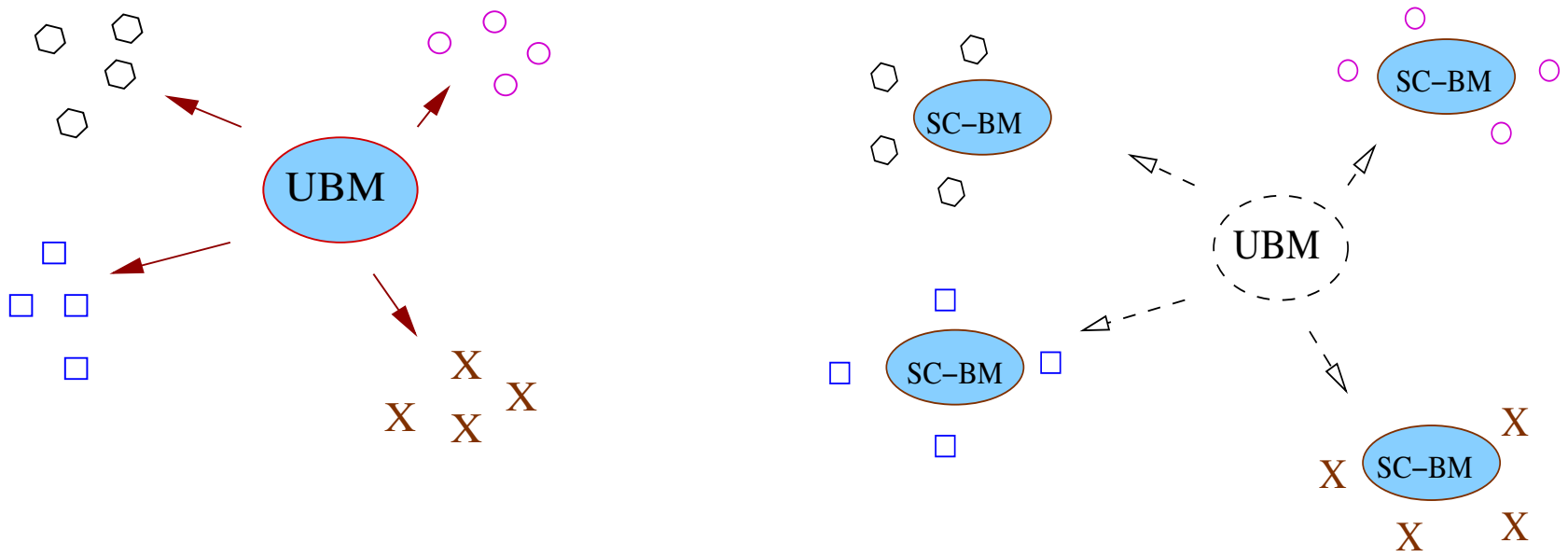
Log Likelihood Ratio: $\Lambda(X) = \log Pr(X|\lambda_2) - \log Pr(X|\lambda_{alt})$

Selecting Alternative Hypothesis

- Speaker independent Universal Background Modeling (UBM) [1]
  - a single model for all speakers in the database.
- Cohort based
  - Maintain a set of closest speaker models (cohorts) **per speaker** [2-5]
  - Separate Background Model (BM) obtained from cohorts called Individual Background Model (IBM) [6] for **each speaker**.
- Propose:
  - use of a separate Background Model (BM) for a **group of speakers**.

1. D. A. Reynolds and et. al, "Verification using Adapted Gaussian Mixture Models," DSP, vol. 10, pp. 19-41, Jan2000.
2. A. E. Rosenberg and et. al, "The Use of Cohort Normalized Scores for Speaker Verification," ICSLP, 1992.
3. D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," Speech Communication, vol. 17, pp. 91-108, 1995.
4. A. E. Rosenberg and et. al, "Speaker Background Models for Connected Digit Password Speaker Verification," ICASSP, 1996.
5. A.M. Ariyaeeinia and et. al, "Analysis and Comparison of Score Normalization Methods for Text Dependent Speaker Verification", Eurospeech, 1997.
6. Yossi Bar-Yosef and et. al, "Adaptive Individual Background Model for Speaker Verification," Interspeech, 2009.

Speaker Cluster-Wise Background Model



- Separate Background Model (BM) for each group of speakers
- Speakers are clustered/grouped based on
  - Vocal Tract Length Normalization (VTLN) factor
  - Maximum Likelihood Linear Regression (MLLR) super-vector

Use of VTLN parameter, α for Speaker Clustering

- Cluster/group speakers who have the same VTLN parameter
- Speakers with similar α have similar spectra for same sound

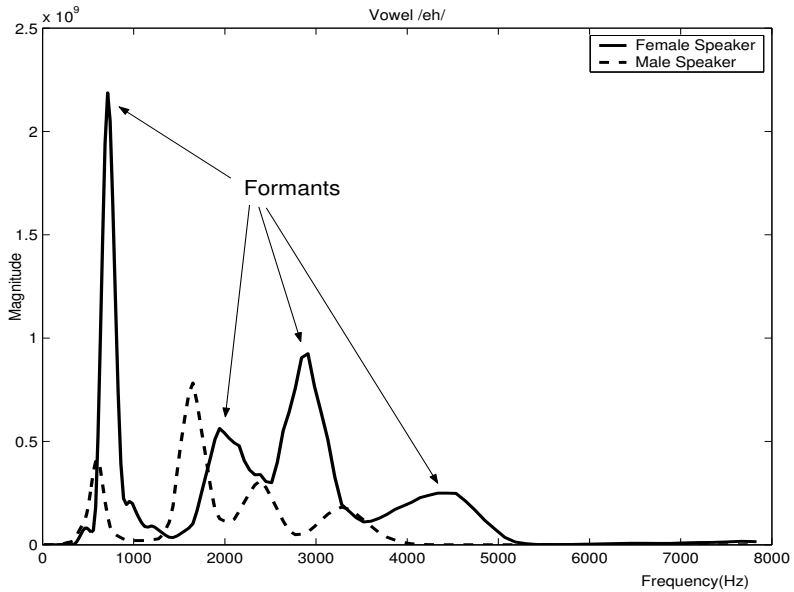


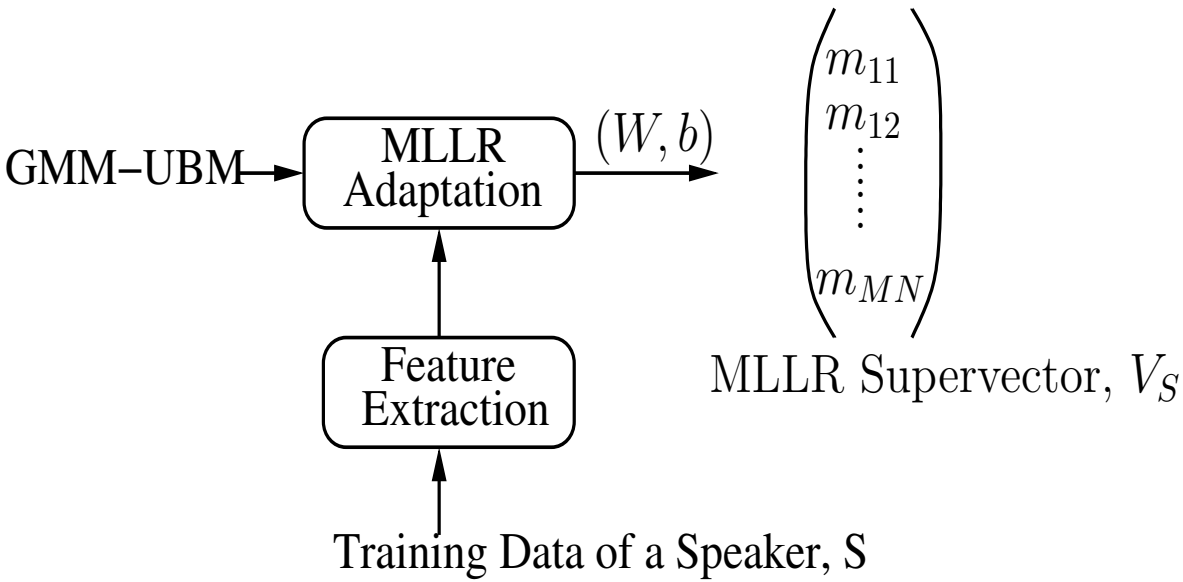
Figure: The spectra of vowel /eh/ for male and female speaker.

- Estimate optimal α, using warped features and UBM

$$\hat{\alpha}_r = \arg \max_{\alpha} Pr(X_r^{\alpha}; \lambda_{UBM}) \; ; \quad \alpha \, \epsilon \, [0.80, 1.20]$$

Use of MLLR super-vector for Speaker Clustering

- Characterize speaker by MLLR super-vector using his/her training data
- Cluster/group the speakers using MLLR super-vector with K-means algorithm and Euclidean distance measure.

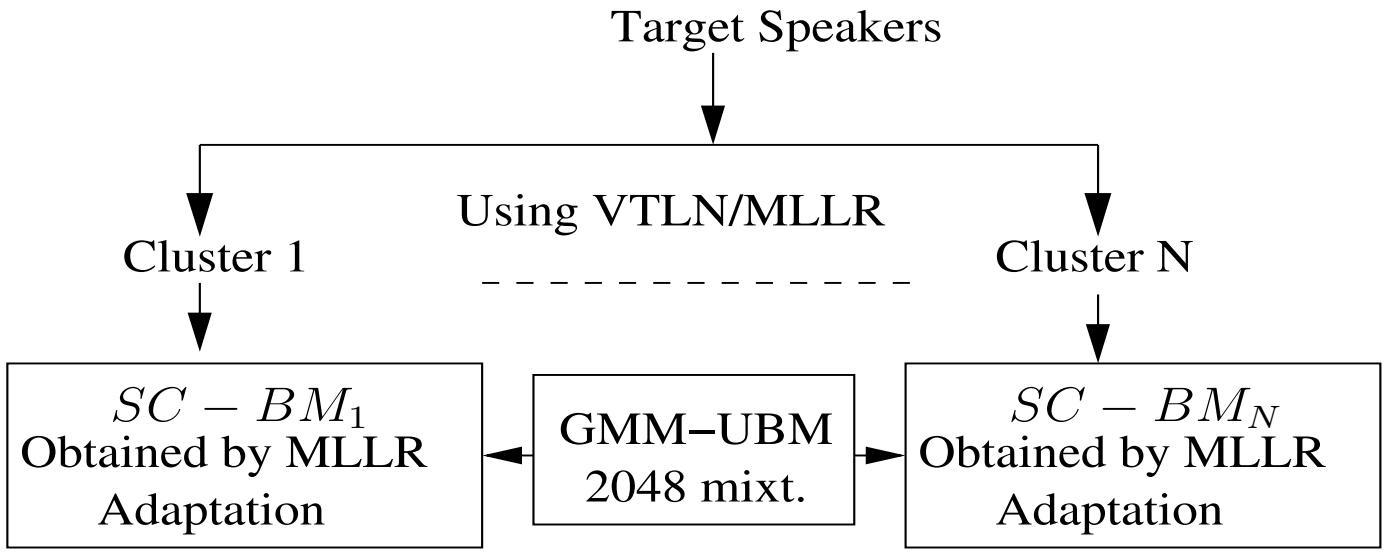


Building Speaker Cluster Wise BM (SC-BM)

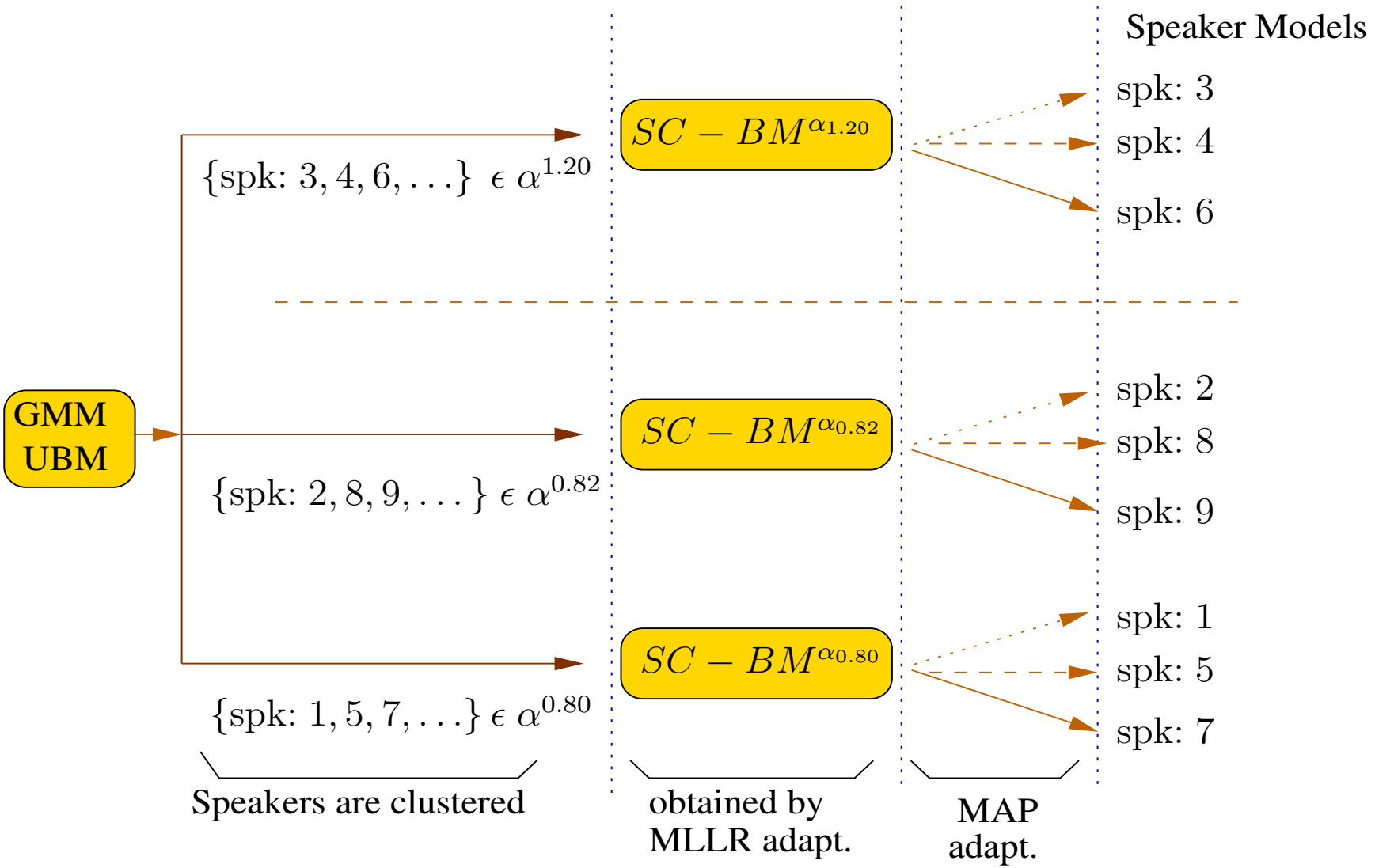
- Speakers are clustered either using VTLN, α or MLLR super vector.
- Each SC-BM is adapted from UBM using MLLR adaptation:

$$\hat{\mu}_{sc-bm_\alpha} = W\mu_{ubm} + b$$

$\hat{\mu}_{sc-bm_\alpha}$ is estimated using data from particular speaker cluster



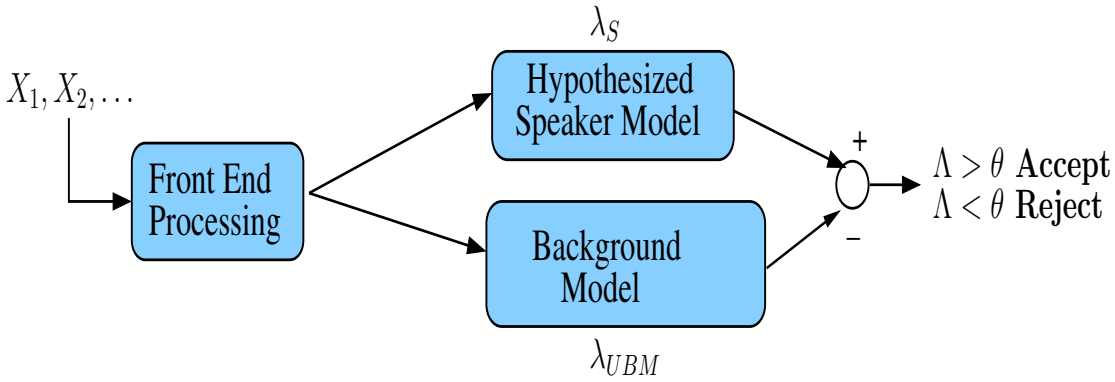
Training Speaker models for proposed multiple BMs



- Use similar steps for speakers clustered using MLLR super-vector.

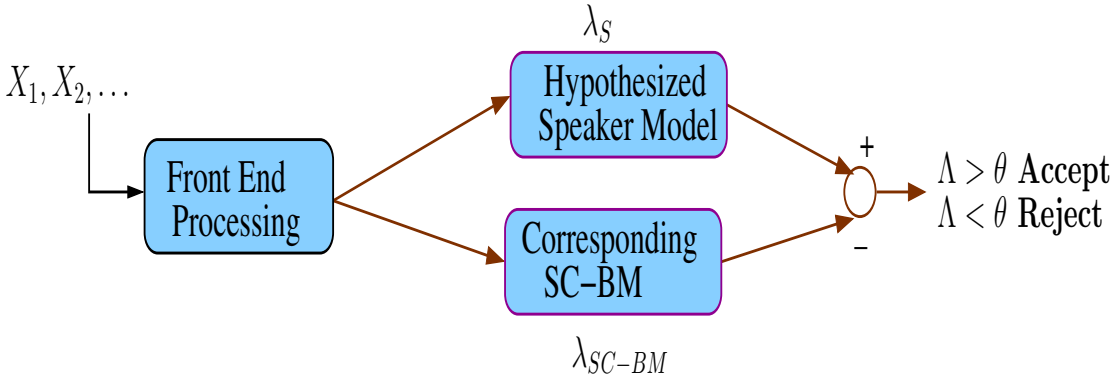
Test Phase: Single BM Vs Multiple BM

- Conventional:



$$\Lambda(X) = \log Pr(X|\lambda_S) - \log Pr(X|\lambda_{UBM})$$

- Proposed:



$$\Lambda(X) = \log Pr(X|\lambda_S) - \log Pr(X|\lambda_{SC-BM})$$

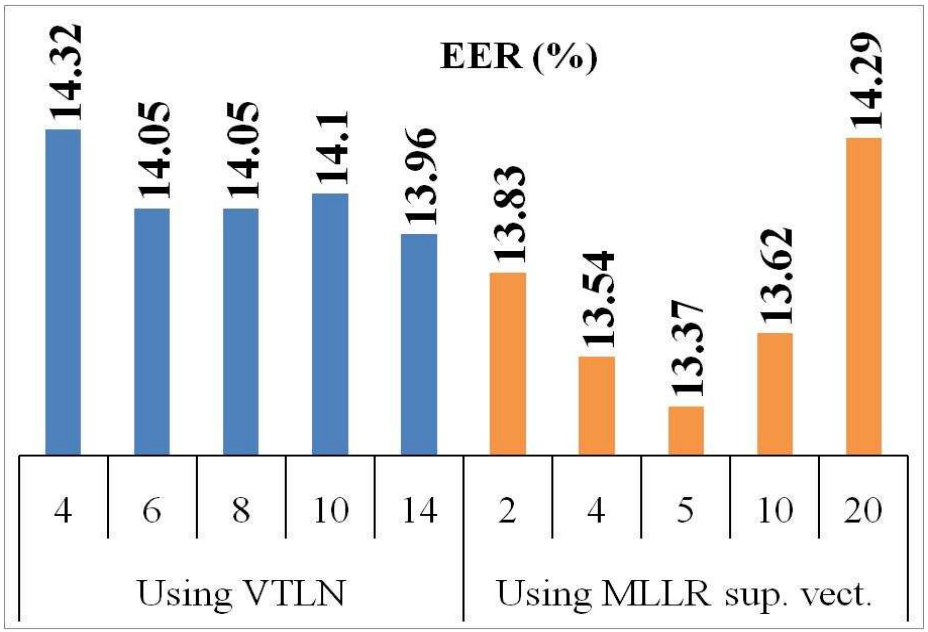
Both systems require same computation cost during test phase.

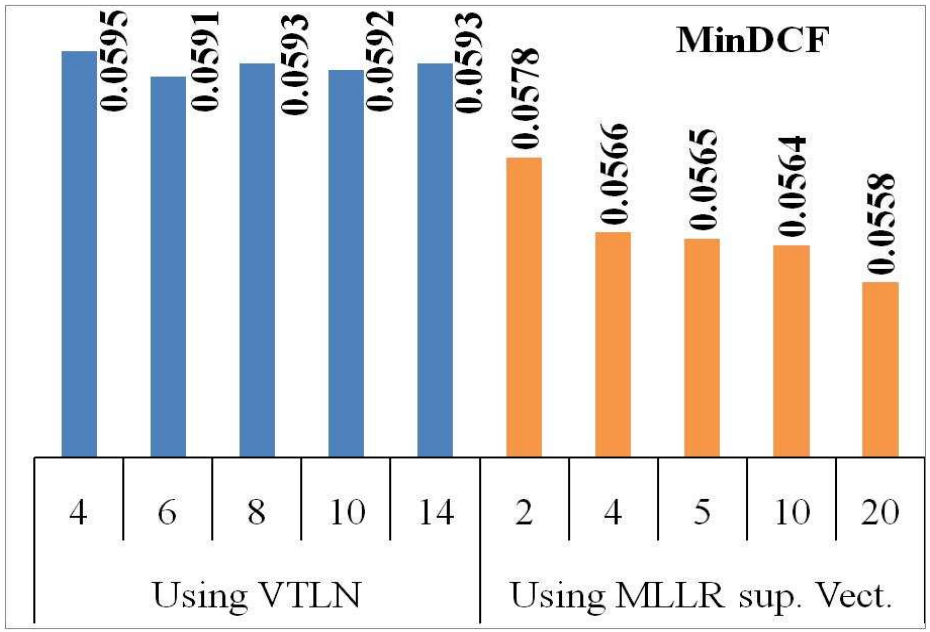
Experiment setup

- **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 energy component of frames
  - 0-mean and 1-Variance utterance level
- **Background Modeling**
  - Speaker Independent UBM model
  - 2048 mixtures with diagonal covariance matrix
  - Training Data: NIST 2002 SRE and Switchboard-1 Release-2
- **Evaluation**: 1 side trn. & 1 side test (core) condition of NIST 2004 SRE
  - Target Model: 2 iteration of MAP wrt UBM/SC-BMs
  - SC-BMs: 1 iteration of MLLR wrt Speaker Independent UBM

Comparison of performance using single UBM

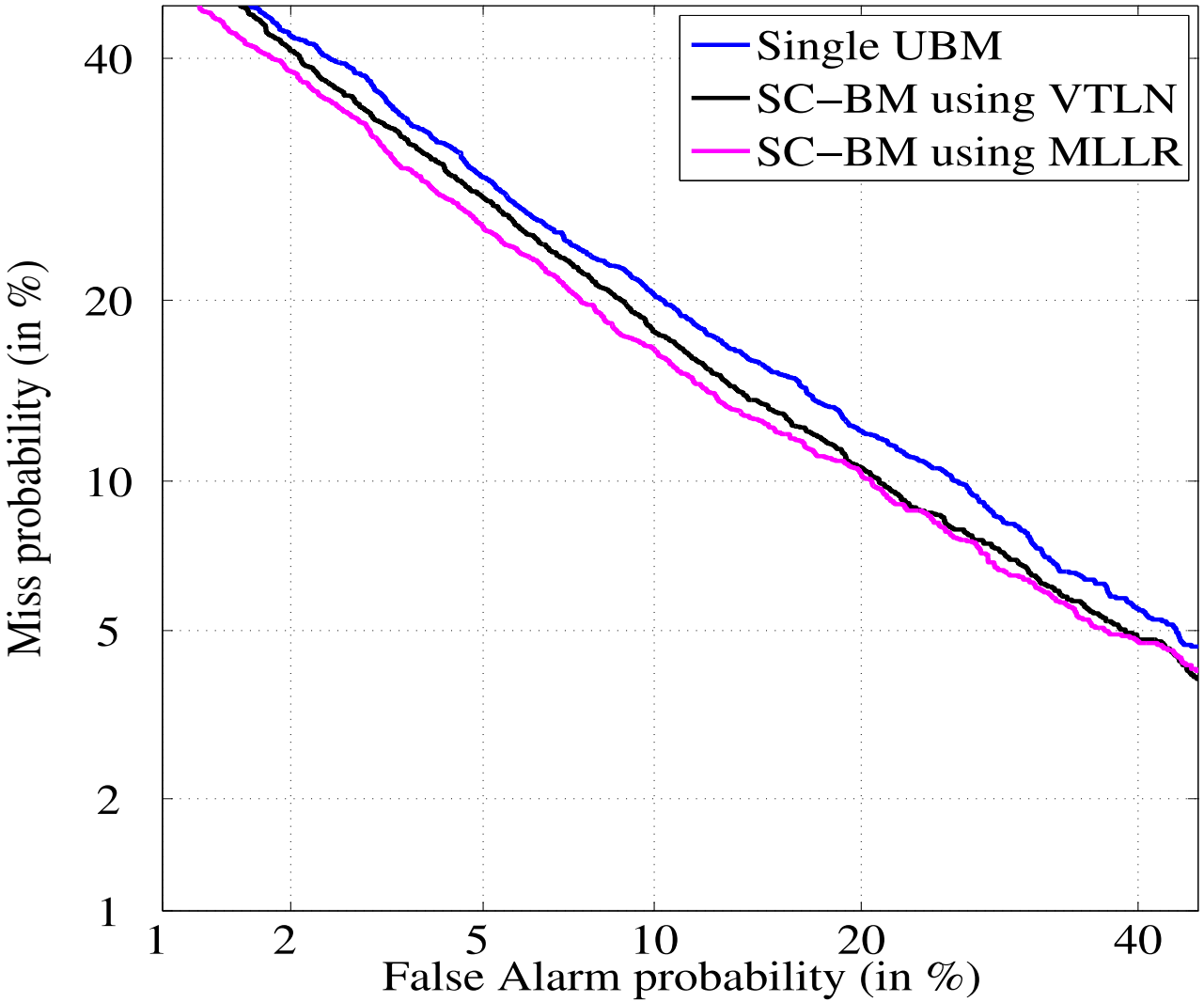
- Single UBM: EER=15.42% , MinDCF=0.0597





- 10% relative improvement in EER for both techniques
- MLLR super-vector provides better performance than VTLN.
- VTLN and MLLR super-vector technique show best result for 14 and 5 clusters respectively (in terms of EER).

Results for gender independent SC-BM

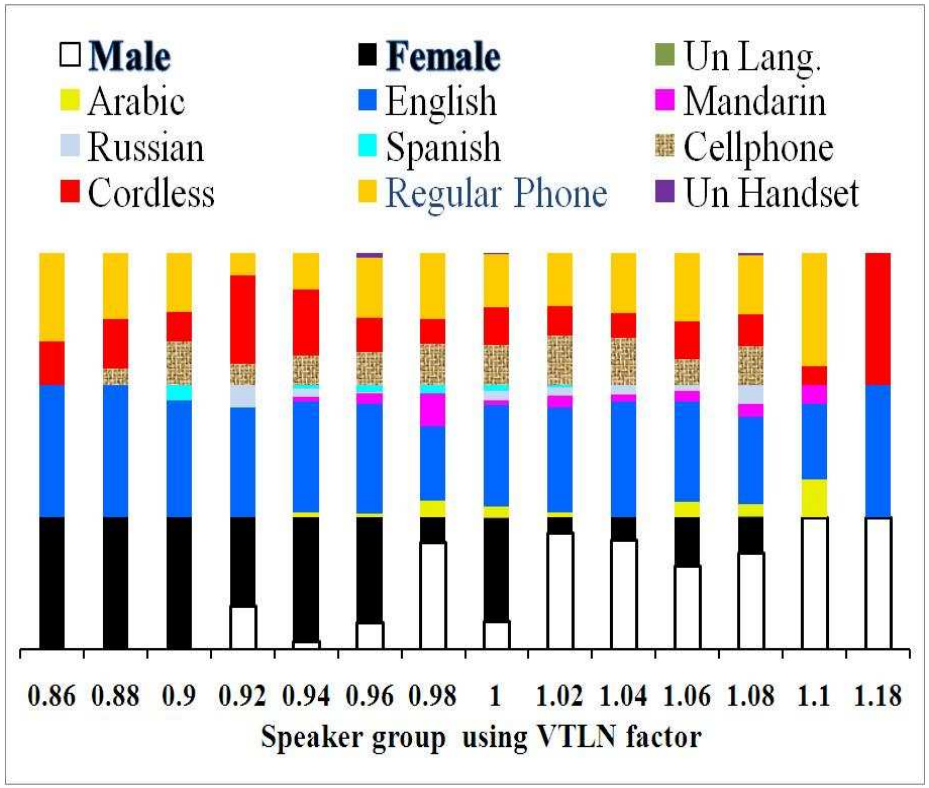


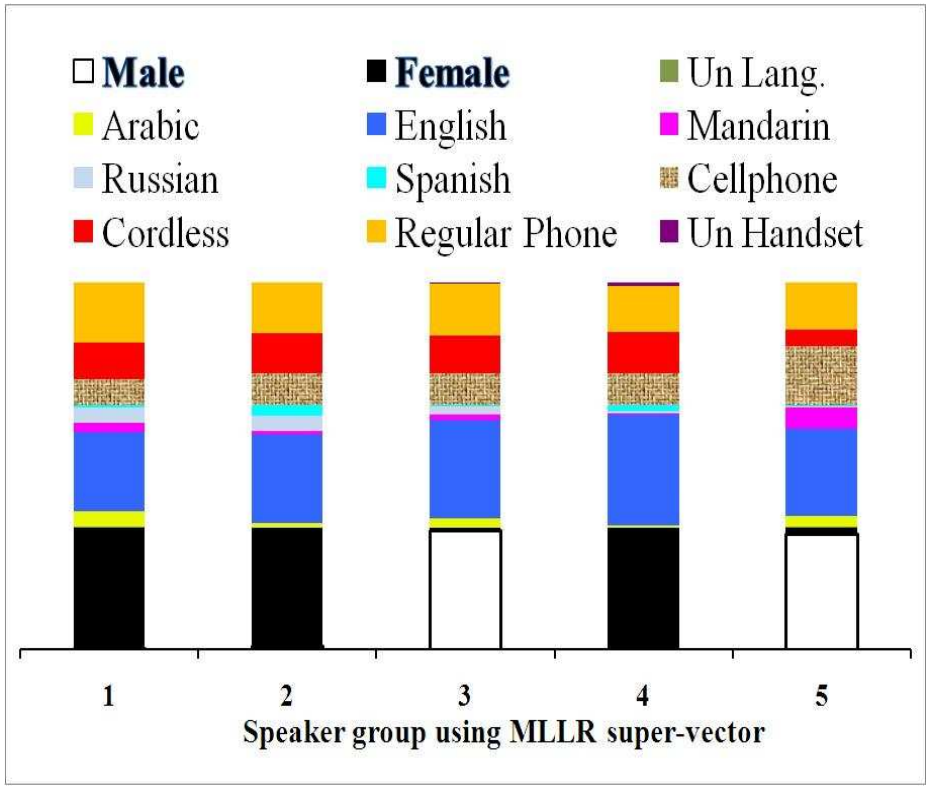
- Single UBM
  EER=15.42%
  MinDCF=0.0597

- SC-BM using VTLN
  EER=13.96%
  MinDCF=0.0593

- SC-BM using MLLR
  EER=13.37%
  MinDCF=0.0565

Distribution of gender in each Speaker Cluster

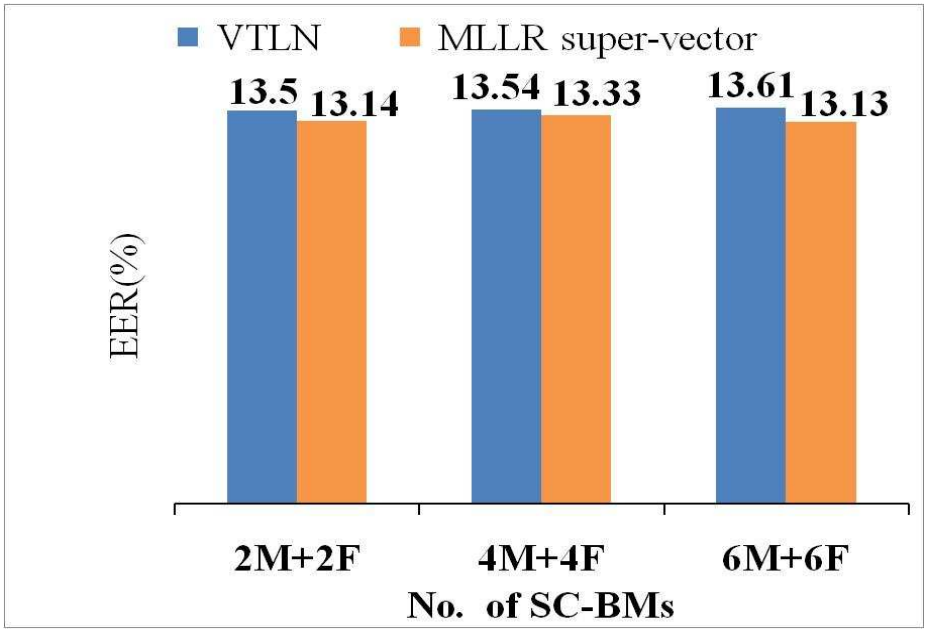


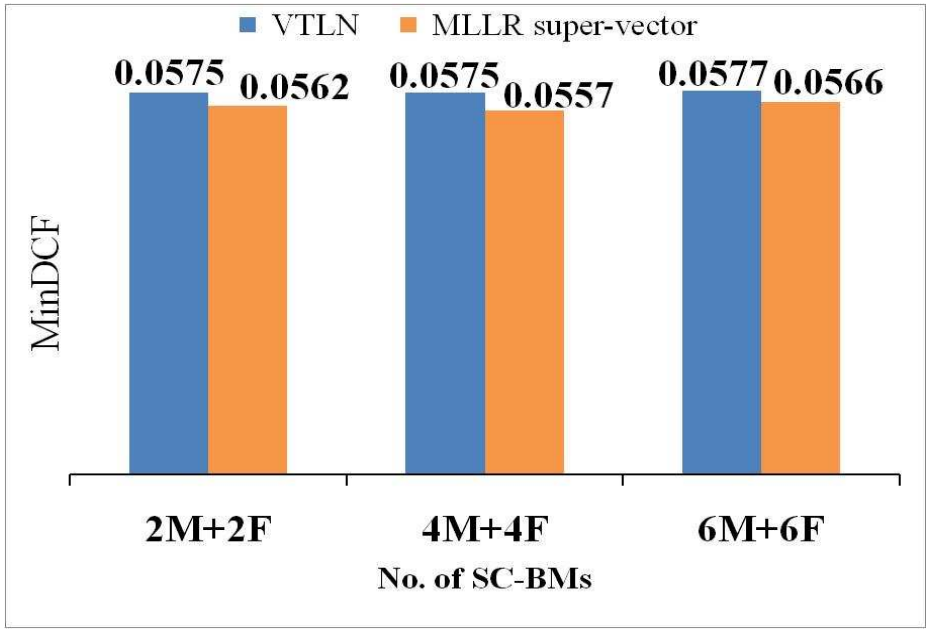


- MLLR clusters are "pure" in terms of gender
  - may be the reason for better performance
- Use of gender wise cluster may improve in VTLN

Verification using Gender-wise SC-BM

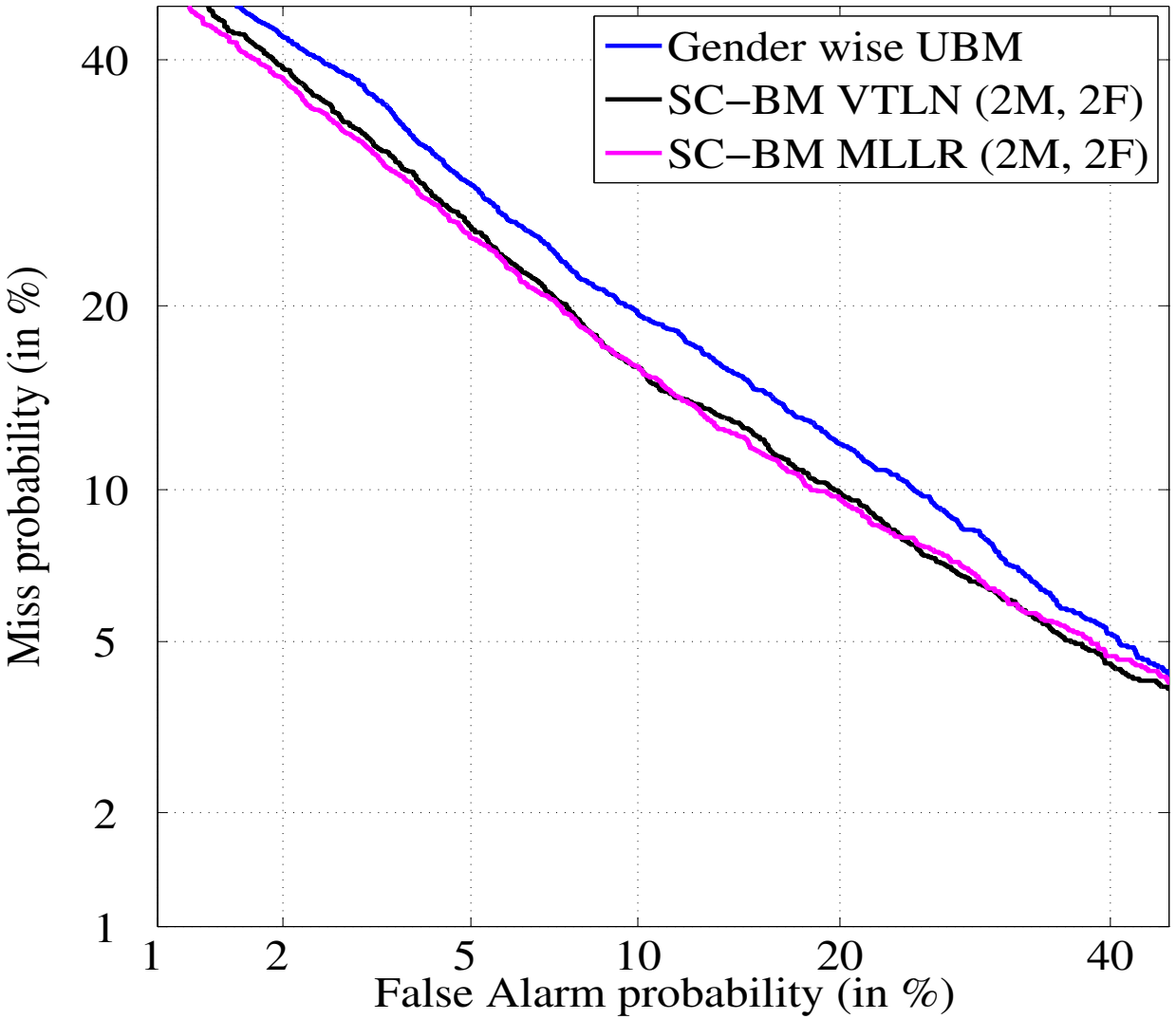
- Gender wise UBM: EER=15.07% , MinDCF=0.0597





- Gender wise cluster makes MLLR/VTLN performance comparable.
- SC-BM outperform gender-wise UBM again by about 10% relative

Result for gender-wise SC-BM



- Gender-wise UBM
  EER=15.07%
  MinDCF=0.0597
- SC-BM using VTLN
  EER=13.50%
  MinDCF=0.0575
- SC-BM using MLLR
  EER=13.14%
  MinDCF=0.0562

Summary

- Better performance with small increase in the number of BMs.
  - both in gender independent and dependent case.
- Computational cost during test same as that of a single UBM case.
- MLLR super-vector clustering performs better than VTLN.
- Using gender-wise SC-BM narrows the gap between
  - using VTLN and MLLR super-vector for speaker clustering.

Thank You!