The 2006 France Telecom Research and Development Center (Beijing) Speaker Recognition System

June 26, 2006

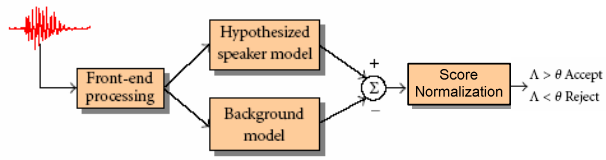
Xianyu Zhao, Yuan Dong, Hao Yang, Jian Zhao

Outline

- Overview
- Front End
- Speaker Modeling
- Score Normalization
- Speaker Segmentation and Clustering
- Conclusion

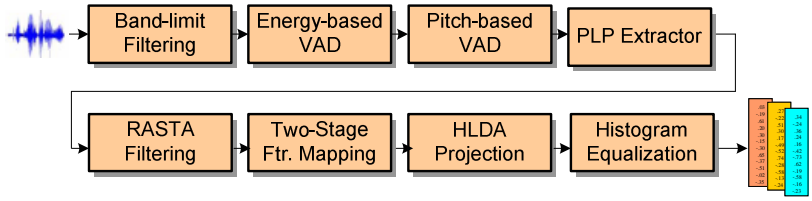
GMM-UBM based Text Independent Speaker Verification System

- Likelihood-ratio-based speaker verification system using GMM-UBM
  - Front End Feature extraction
  - Speaker (and alternative hypothesis) Modeling
  - Score normalization



Front-End Processing: Aim and Structure

- The aim of robust front-end processing:
  - To reduce the influences of transmission channel, microphone type and environmental noises
  - To facilitate the following speaker modeling with GMM
- The FTRD SRE'06 Front-End

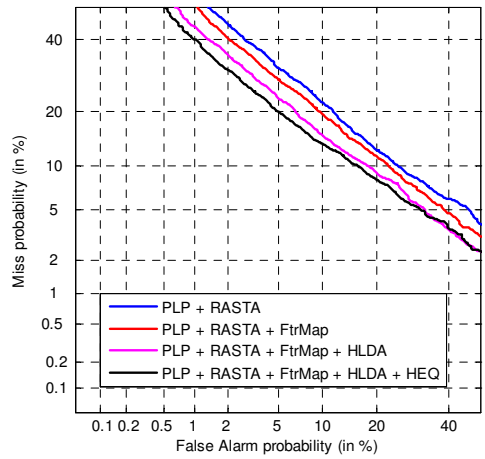


Front-End Processing: Some Details

- Two-Stage Feature Mapping
  - Stage-1 (Channel Mapping): Land, GSM, CDMA, Cellular, TDMA, Cordless
  - Stage-2 (Microphone Mapping): Speaker phone, Head phone, Handheld, Ear-bud
- HLDA Projection
  - Original Feature: PLP_0 + Delta + Delta-Delta + Delta-Delta-Delta (52-dim)
  - Projected Feature: 51-dim
- Histogram Equalization
  - Reference Cumulative Distribution Function (CDF) was estimated using all the development data

Front-End Processing: Performance

- NIST SRE04 1side-1side

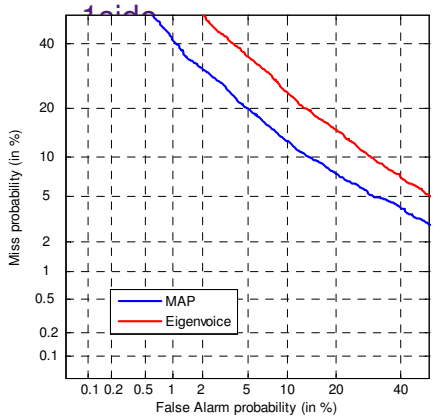


Speaker Modeling

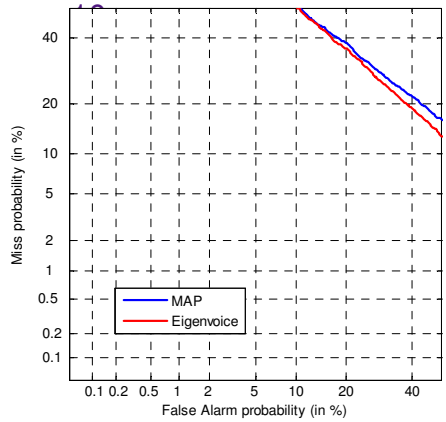
- The UBM model is a gender independent GMM with 2048 mixture components trained using about 40 hours of data from the Switchboard and SRE'2004 evaluation database.
- Speaker models are obtained by adapting from the UBM with their individual training data through
  - Bayesian learning approach (Maximum a Posteriori adaptation): good asymptotic behavior
  - Speaker clustering based approach (Eigenvoice adaptation): rapid adaptation

Speaker Modeling: Performance

- NIST SRE04 1side-1side



- NIST SRE04 10sec-10sec



Score Normalization

- Each verification score is normalized by subtracting the mean and then dividing by the standard deviation of imposter score distribution.
- Estimation of imposter score distribution:
  - TNorm: Each utterance is compared against to a set of imposter models to estimate the imposter score distribution.
  - ATNorm: Similar to TNorm, but the set of imposter models is dependent on the hypothesized speaker of a test segment.

Score Normalization: some details

- Multi-language imposter pool

	English	Mandarin	Arabic	Russian	Spanish	Total
Male	268	31	32	14	9	354
Female	349	24	27	31	37	468
Total	617	55	59	45	46	822

- TNorm: for male segment speaker, all 354 male imposters are used; for female segment speaker, all 468 female imposters are used.

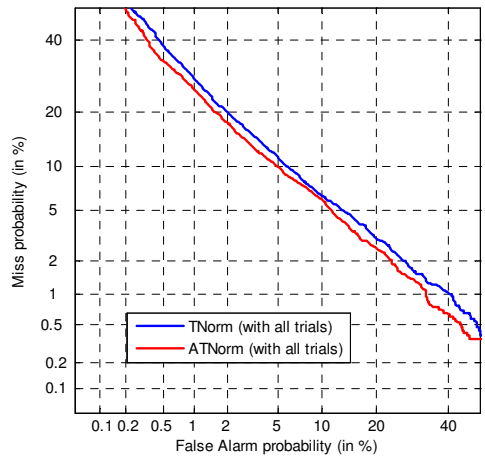
Score Normalization: some details

- ATNorm: for each male target speaker, 55 nearest imposters are selected from the male imposter pool; for each female target speaker, 55 nearest female imposters are selected from the pool.
- The following cross-model log likelihood ratio distance is used to define the neighborhood of two models

$$d(\lambda_i, \lambda_j) = -\frac{1}{N_i}\log\left(\frac{p(x_i|\lambda_j)}{p(x_i|\lambda_{UBM})}\right) - \frac{1}{N_j}\log\left(\frac{p(x_j|\lambda_i)}{p(x_j|\lambda_{UBM})}\right)$$

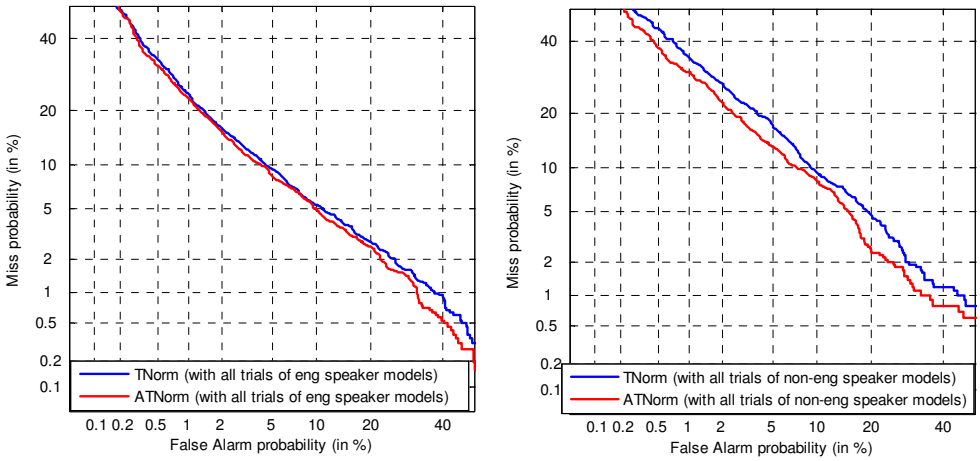
Score Normalization: Performance

- NIST SRE06 1conv4w-1conv4w test



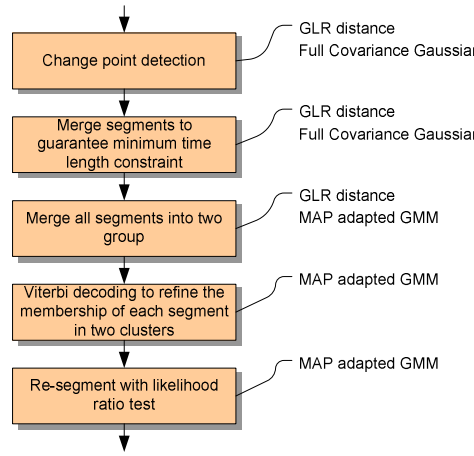
Score Normalization: Performance

- NIST SRE06 1conv4w-1conv4w test



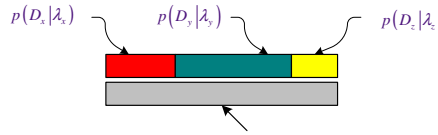
Speaker Segmentation

- MFCC + energy, no channel compensation
- GLR distance is used for change point detection and agglomerative clustering
- Two-Stage agglomerative clustering: as more data are in cluster, more complicated models are used



Speaker Clustering

- MFCC + Delta + RASTA: reduce channel mismatch across different conversations
- A form of GLR distance is employed to cluster multiple segments across several conversations



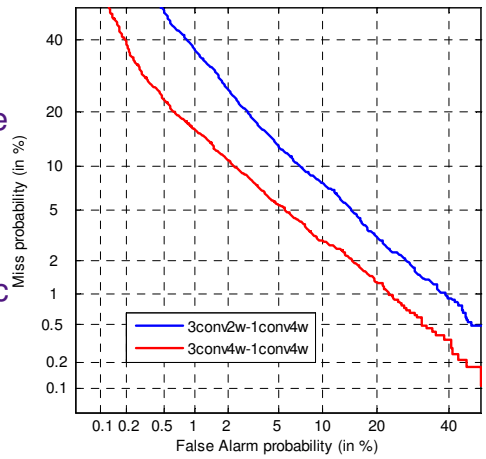
Hyp: $\{D_x, D_y, D_z\}$ are from the same distribution $p(\{D_x, D_y, D_z\}|\lambda_0)$

$$GLR(D_x, D_y, D_z) = \log\left(\frac{p(D_x|\lambda_x)\cdot p(D_y|\lambda_y)\cdot p(D_z|\lambda_z)}{p(\{D_x, D_y, D_z\}|\lambda_0)}\right)$$

- Gender of target speaker is taken into account to make decision

Summed Channel SRE Performance

- NIST SRE06
- There is still large gap between 2-wire and 4-wire training; our speaker segmentation/clustering process need improvements



Conclusion

- It is the first time that FTRD (Beijing) participates NIST SRE and we present a simple GMM-UBM likelihood ratio based system.
- Some improvements have been made on front-end, model adaptation and score normalization processes.
- To catch up with the state-of-art, there are still a lot of things to do, e.g. prosodic/idiolectal features, session variability modeling, cooperation with ASR, etc.

Thanks

- To NIST for providing the wonderful evaluation platform!
- To all pioneers in speaker recognition/verification research area from who/whose publications we have learned a lot!