NIST SRE 2010: TokyoTech Speaker Recognition

Marc Ferràs, Sangeeta Biswas, Koichi Shinoda and Sadaoki Furui

Tokyo Institute of Technology, Japan

1. Introduction

- TokyoTech participated in the core condition
  - Focusing on telephone speech
- Two SVM-based acoustic systems:

  Primary System : GLDS-SVM

  Alternate System : Fusion of GLDS-SVM and GMM-SVM
- System fusion was performed by a weighted average of the system scores
- Decision thresholds were optimized using the new cost and priors used in the core condition of NIST SRE 2010

$$C_{Det} = 1 \times P_{Miss|Target} \times 0.001 + 1 \times P_{FalseAlarm|NonTarget} \times 0.999$$

- Three different thresholds for English phn-phn, int-int and int-phn conditions were estimated on NIST SRE 2008 scores

2. Front-End

- Speech Enhancement
  - ICSI-OGI-Qualcomm Wiener filter for interview segments
  - FIR echo canceller for phonecall segments
- Feature Extraction
  - 15 Perceptual Linear Prediction (PLP) coefficients + 15 Δ + 15 $\Delta\Delta$ + log-E + ΔE + $\Delta\Delta$E (48 dimensions)
  - Feature warping with 3s sliding window
  - Energy-based speech/non-speech segmentation
    - Threshold set to select 30% of the frames

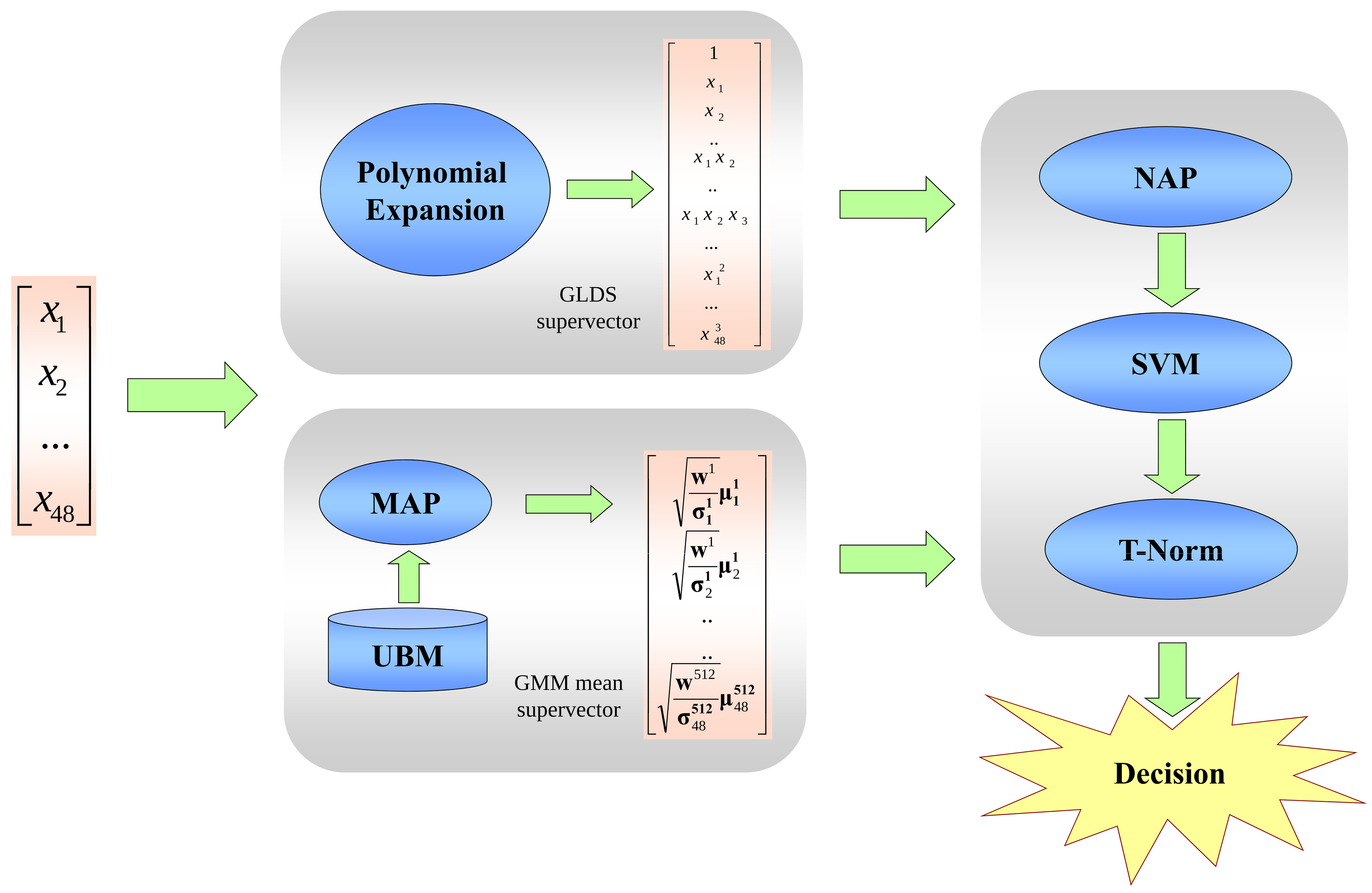
3. GLDS-SVM

- SVM system using Generalized Linear Discriminant Sequence (GLDS) kernel by explicit polynomial expansion
  - Polynomial features up to the 3rd order (20824 dimensions)
- Nuisance Attribute Projection (NAP) session compensation
  - 50 dimensions for the session subspace
  - Projection matrix trained using NIST SRE 2004 training data
- Feature scaling to normalize dot products
- Soft margin C-SVM classifier (LIBSVM)
  - Linear kernel
  - 4000 impostor speakers from NIST SRE 2004 data
- Gender-dependent T-norm score normalization
  - 250 cohort speakers per gender from NIST SRE 2005 data
  - Minimum segment length was 2 minutes

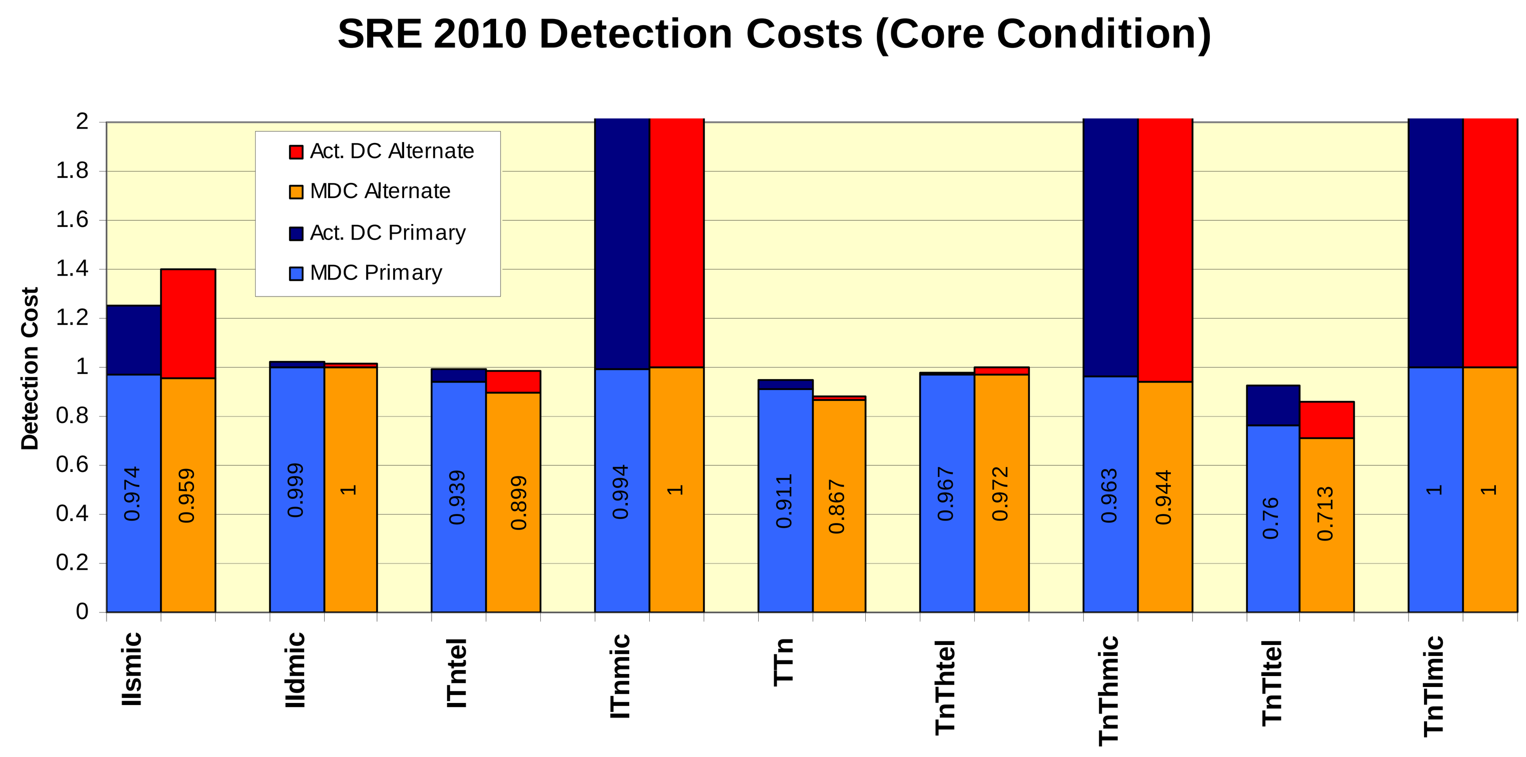
4. GMM-SVM

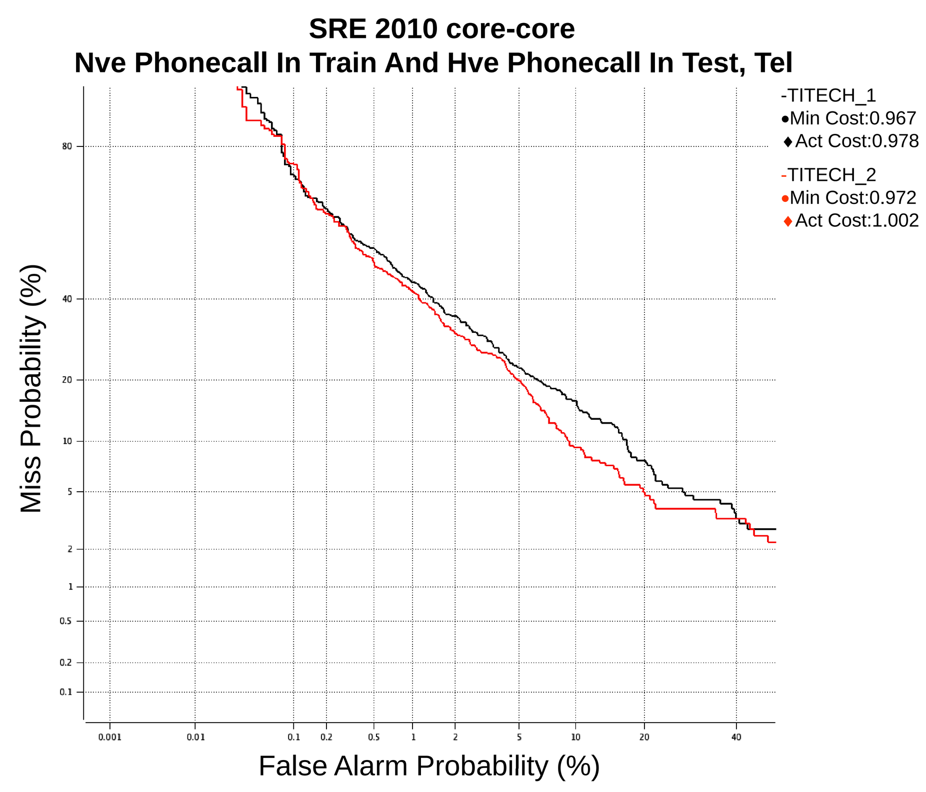
- SVM system using mean vectors of speaker models as features
- Universal Background Model (UBM)
  - Training data of 40 hours from the NIST SRE 2004
  - 512-Gaussian components
  - Diagonal covariance matrices
  - 3 iterations of maximum likelihood estimation
- Speaker models obtained by standard MAP adaptation of UBM
- GMM linear kernel based on K-L divergence
- NAP, SVM and score normalization set-ups were the same as in GLDS-SVM

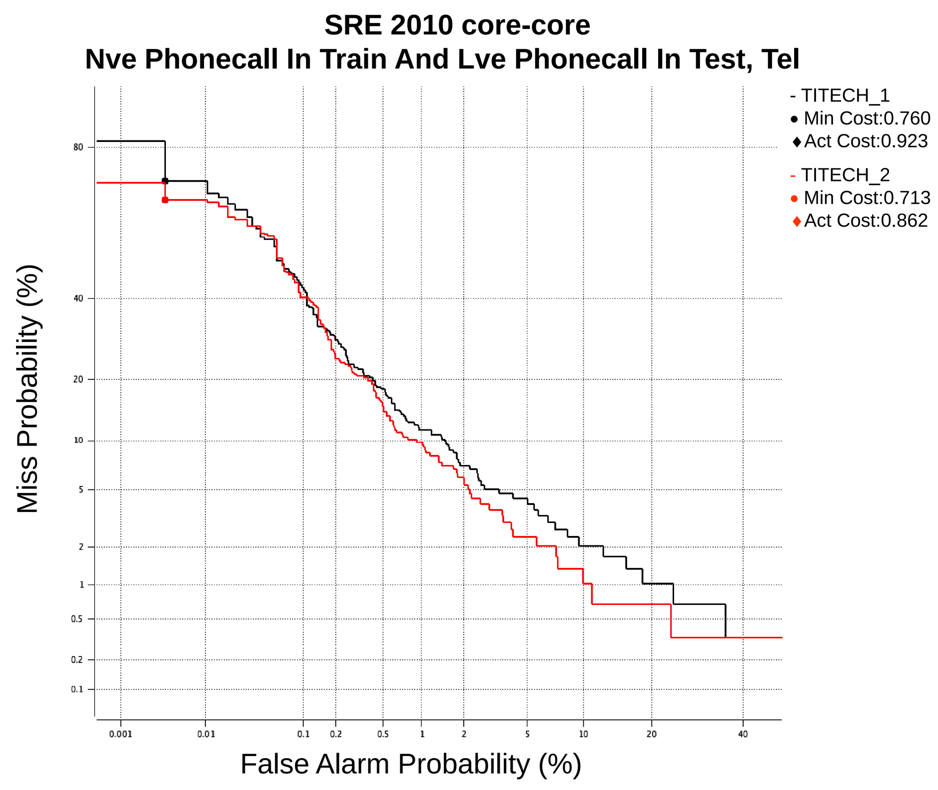
5. System Diagram

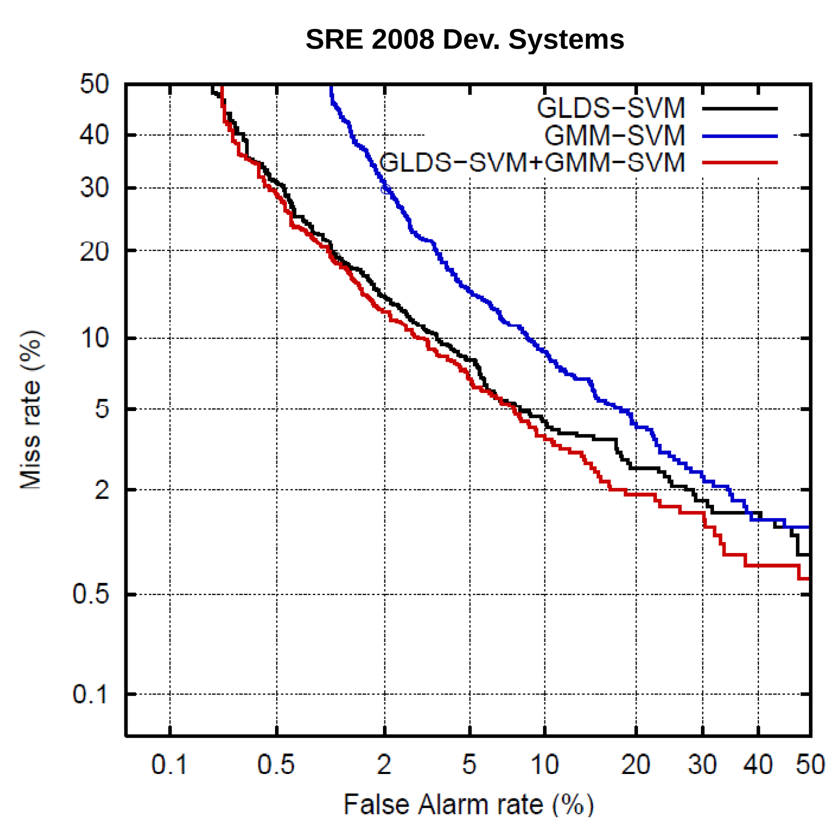


6. Results









- GLDS-SVM outperforms GMM-SVM system
  - GMM-SVM system is not mature
  - Fusion improvement relies on GLDS-SVM
    - 0.9 vs.0.1 weights
- Similar performance for low vocal effort speech
- Big performance degradation for high vocal effort speech
- Good overall calibration
  - Different thresholds for different conditions
  - Long segments in T-norm might improve stability
- Bad calibration for conditions involving telephone speech recorded with room microhone