論文 / 著書情報
Article / Book Information

Title	NIST SRE 2010:Tokyo Tech Speaker Recognition
Authors	Marc Ferras, Sangeeta Biswas, Koichi Shinoda, Sadaoki Furui
Citation	Proc. NIST 2010 Speaker Recognition Evaluation Workshop, Vol. , No. , pp.
Pub. date	2010, 6

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

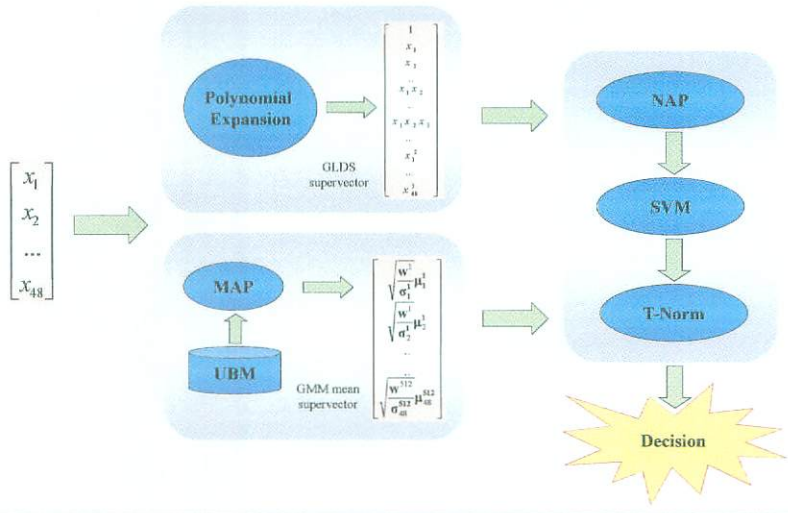
3. GLDS-SVM

- 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

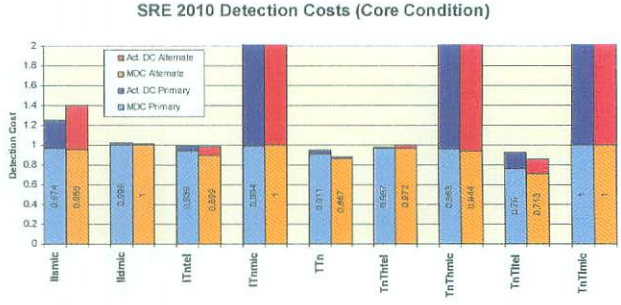
4. GMM-SVM

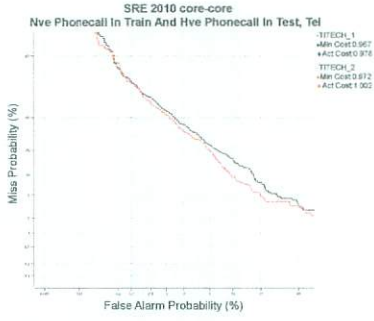
- SVM system using mean vectors of speaker models as features
- Universal Background Model (UBM)
  - Training data of 40 hours from the NIST SRE 2004
  - 512-Gaussian components
  - Diagonal covariance matrices
  - 3 iterations of maximum likelihood estimation
- Speaker models obtained by standard MAP 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

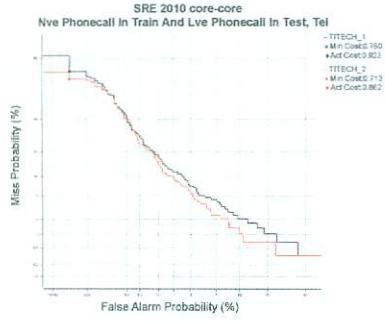
5. System Diagram

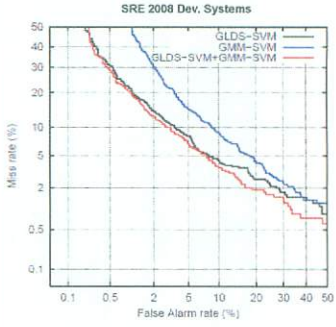


6. Results









- 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

System Description

Our submission for the 2010 NIST SRE used two SVM-based acoustic systems, GLDS-SVM and GSV-SVM, combined by a weighted average of their respective scores.

Front-end
For both systems, we used 15 Perceptual Linear Prediction (PLP) coefficients along with log-energy plus all first and second order derivatives, making up a total of 48 features per vector. Vectors were computed from the speech signal, previously pre-emphasized and bandpass-filtered from 30Hz to 3400Hz every 10ms and within a 30ms sliding window. Except for the energy coefficients, feature warping was applied on all other features using a single Gaussian and a window length of 3s. Speech/non-speech segmentation was performed by thresholding the energy coefficient. The threshold was set so that the 30% of the frames were classified as speech.

GLDS-SVM System
The GLDS-SVM system uses a simplified Generalized Linear Discriminant Sequence (GLDS) kernel. Polynomial expansions of orders one, two and three of the PLP feature vectors were computed and concatenated to yield an expanded vector. Each of the expanded vector components was normalized to have unity variance within each speaker segment. We used these vectors as base features for the GLDS-SVM system. The base features were compensated for session variation using Nuisance Attribute Projection (NAP), with the transform being trained on the SRE04 training data and a subspace dimension of 50. We normalized each of the resulting feature components using min-max scale-and-shift across all speaker in the impostor set to normalize any dot product performed later in the SVM.

About 4000 speaker segments were taken from the SRE04 training data and were used as impostors for SVM training. We used a linear kernel C-SVM classifier (LIBSVM library). In the test phase, every trial score was normalized using T-norm with 500 cohort speakers from the NIST SRE05 training data.

The real-time runtime factors are x0.25RT and x0.12RT for the training and test phases respectively on an Intel Core2 2.4GHz CPU. Memory usage was below 1Gb.

GSV-SVM System
The GSV-SVM system uses Gaussian Mean Supervectors (GSV) as base features. A Universal Background Model (UBM) with 512 Gaussian components was trained using 3 iterations of Maximum Likelihood per Gaussian-split and 40 hours of speech training data. The UBM was adapted to the speech data of every speaker of interest using standard MAP adaptation. The mean vectors of all Gaussian components were concatenated into a supervector and each mean coefficient normalized by its standard deviation and the square root of its weight, which is equivalent to classification using the GMM Linear Kernel. These supervectors were post-processed using NAP with the same setup as in the GLDS-SVM system. SVM training and test were performed in the same way as in the GLDS-SVM system, including T-norm score normalization.

The real-time runtime factors are x0.33RT and x0.19RT for the training and test phases respectively on an Intel Core2 2.4GHz CPU. Memory usage was below 1Gb.

Submission
We submitted two systems. The primary submission consisted of the GLDS-SVM system alone. A contrastive system was submitted as the score average of the GLDS-SVM and GSV-SVM scores with 0.75 and 0.25 weights respectively.