TÜBİTAK – UEKAE + Sabancı University System for NIST 2010 Speaker Recognition Evaluation

Yusuf Ziya Işık[1,2], Mehmet Uğur Doğan[1], Hakan Erdoğan[2]
(1) TÜBİTAK – UEKAE, Kocaeli, TURKEY
(2) Sabancı University, İstanbul, TURKEY

System Description

Overview

- MFCC features with energy based VAD
- GMM supervector + SVM with linear kernel
- Znorm score normalization

UBM Training

- 2048 mixture gender-dependent UBM's are trained from 323 hours of speech for male and 463 hours of speech for female.
- We started from 1 mixture and increment the mixture number by a factor of 2 after 25 EM iterations.
- At each EM iteration only randomly selected 2% of each utterance is used in training.

Database Organization

- SRE06, SRE08 and SRE08 follow-up databases are used for system training and development.
- SRE06 database is used solely for system training.
- SRE08 and SRE08 follow-up databases are splitted into two parts according to speaker ids. The first part is used in all parts of system development: UBM training, impostor model generation and znorm score normalization. The second part is used for system testing.

SVM Training and Test

- Speaker models are obtained by mean-only r-MAP adaptation with a relevance factor of 8.
- To create supervectors:
  - UBM means are subtracted from speaker model means
  - Resulting mean vectors are normalized by standard deviation and scaled by the weight of that mixture
  - Supervector is obtained by stacking these vectors and scaling the final vector to be unit-norm.
- SVM models are trained using 1095 male and 1695 female impostor utterances.
- Linear kernel is used for testing. Znorm score normalization is applied using 1258 male and 1832 female utterances.

Feature Extraction

- Acoustic vector used: 19 dimensional static MFCC features + Δ + ΔE: Total dimension 39
- 300-3400 Hz bandwidth
- Energy based VAD using a bi-Gaussian model.
- Feature Warping with a 3s sliding window

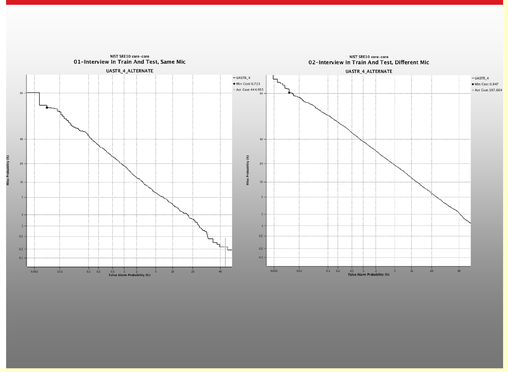
Software Specifications

- All the software used for the evaluation has been written in TUBITAK – UEKAE as an extension to Torch3 Machine Learning Library. SVMTorch present in Torch3 library is used for SVM training.
- We have run our system on 4 server machines: One server with 24 cores (2.4 GHz Intel Xeon) and 64 GB of memory, and 3 servers with 8 cores (3 GHz Intel Xeon) and 32 GB of memory each.
- Execution times for each task is given in the below table for a 2.4 GHz Intel Xeon CPU. In this table training task consists of relevance MAP adaptation, supervector creation and SVM model training. Testing task consists of relevance MAP adaptation, supervector creation and score generation.

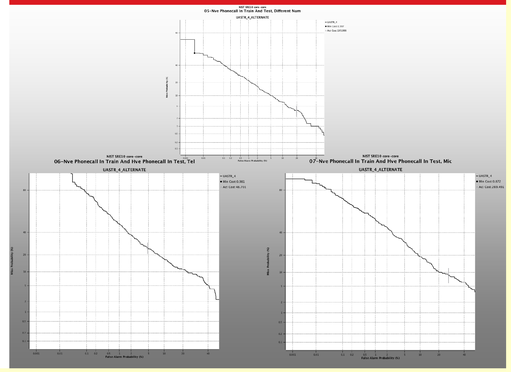
Task	Execution Time
Feature Extraction	0.01RT
Training	0.25RT
Testing	0.05RT

Results

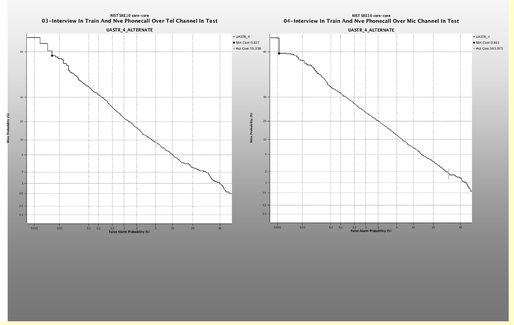
SRE10 Interview



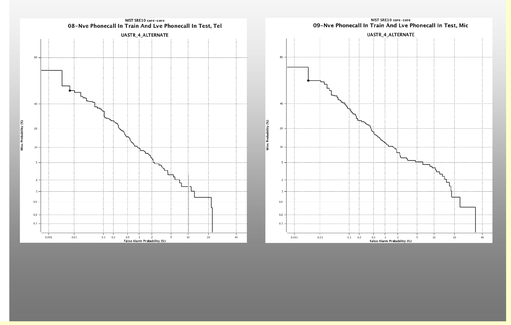
SRE 10 – NVE vs HVE



SRE 10 - Interview



SRE 10 – NVE vs LVE



CONCLUSION

- First participation to NIST Speaker Recognition Evaluations.
- A speaker verification system using GMM supervector + SVM with linear kernel has been implemented.
- The system has not been satisfactory especially for the interview trials with different microphones and for NVE vs HVE telephone trials.
- Algorithms to achieve robustness to channel and session variabilities (NAP, JFA) seems to be a necessary component.
- Great expertise has been achieved on working speaker verification problem with large datasets.