The IESK-Magdeburg Speaker Detection System for the NIST 2006 Speaker Recognition Evaluation

Marcel Katz

IESK, Cognitive Systems
University of Magdeburg, Germany
marcel.katz@e-technik.uni-magdeburg.de

Abstract

In this paper we describe the speaker detection system developed by the *Institute of Electronics, Signal processing and Communications* (IESK) of the University of Magdeburg for the NIST 2006 Speaker Verification Evaluation. The system is based on a *Gaussian Mixture Model* (GMM) system which is used to model the low-level acoustic features and a *Support Vector Machine* (SVM) subsystem which is used to model the high-level prosodic and durational features.

Datasets

- Participated task: 1conv4w-1conv4w

Dataset	UBM Training	T-Norm Training	Development	Evaluation
NIST SRE 2000	100/100	-	-	-
SWITCHBOARD Cell.	50/50	-	-	-
NIST SRE 2004	75/75	50/50	121/245	-
NIST SRE 2006	-	-	-	354/462

Feature Extraction

- Speech Detection was realized by a *Voiced Speech Detection* approach
- Bandwidth limitation: $300Hz - 3400Hz$
- Window size: $25ms$, window shift: $10ms$
- 12 dim MFCCs + energy + first and second time differences
- Cepstral Mean Subtraction, no feature warping

UBM System

- Two gender dependent UBMs with 1024 mixture components
- Initialization: K-means clustering,
  Estimation: Expectation Maximization (EM) Algorithm
- Speaker Models: *Maximum A-Posteriori* (MAP) adaptation,
  only means were adapted (relevance factor $\tau = 16$)
- only the N-best mixture components with respect to the UBM model were used for scoring
- Score Normalization with T-Norm

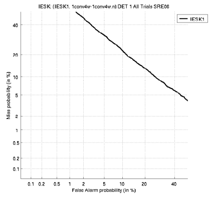
SVM System

- Classification: speaker-versus-background approach
- Features:
  - characteristics of the pitch frequency (mean and standard deviation)
  - energy (logarithm of the segment energy)
  - duration of the voiced speech segments
- non-probabilistic SVM output was transformed to class probability by the algorithm of Platt
- Score Normalization with T-Norm

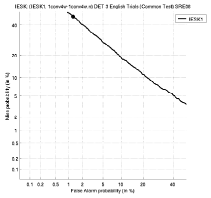
Results

- Experiments were performed on a Linux cluster, consisting of 24 dual Xeon 2.0 GHz processors and 4 GB RAM
- Feature mapping (using 4 models) on all feature vectors was used
- T-Norm was applied to the log-likelihood scores of the subsystems
- Score Fusion by a simple weighted sum approach
- Unique gender independent decision threshold was used
- Decision threshold was set on the optimal DCF of the DevSet

All Trials:



English Trials:



Conclusion

- First evaluation of the IESK-Cognitive System
- More investigation on the standard UBM is needed
- The used datasets (SWBD Cell) are not optimal
- Only small gains in performance using the SVM system
- Investigation of different speech detection approaches
- Alternative characteristic features (phonemes, words)
- Investigation of a lexical subsystem
- Better score fusion: LNKNet, SVM