ETI NIST Speaker Recognition Evaluation 2006 system

Ciano Frost & Thomas Byskov Østergaard

The ETI system is based on a UBM-GMM. In the results reported to the NIST evaluation CMS (Cepstrum Mean Subtraction) was used and the scores were normalized with Tnorm. For the workshop we also present results, where the CMS has been replaced by feature warping, which improved the verification performance significantly.

UBM-GMM description.

The ETI system is based on a Universal Background Model (UBM). The UBM is a gender independent representation of the world. All individual speaker models are created by adaptation of the UBM to the speaker. In our system the speaker models and the UBM have 1024 mixture components.

Parameterization

Speech that is given to the system has been preprocessed by a Voice Activity Detector (VAD). The VAD is based on a Generalized Log likelihood Ratio. The features extracted are:

- 12 MFCC (C1-C12, C0 is discarded) with Cepstrum Mean Subtraction (CMS)
- 12 Δ-MFCC

The UBM

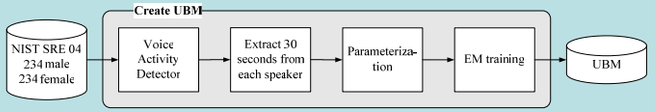
The UBM is trained by the EM algorithm. 468 sessions from NIST 2004 data have been used for training the UBM. The following applies for the training data.

- Data balanced between genders.
- The first 30 seconds of each session are skipped.
- 30 seconds from each session are used in the training.

The division between language in the UBM training material is given in the table.

Language	English	Arabic	Mandarin	Russian	Spanish
In percent	81%	8%	6%	4%	1%

The flowchart below shows the training of the UBM.

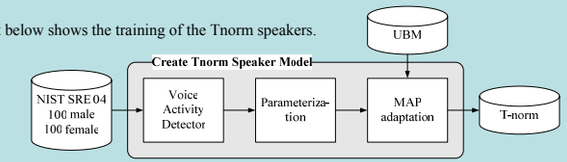


The Tnorm speakers

A set of 100 male and 100 female Tnorm speakers is trained by MAP adaptation of the UBM [1]. The data for the models is from NIST 2004.
The division of the Tnorm speakers between language is given in the table.

Language	English	Arabic	Mandarin	Russian	Spanish
In percent	66%	2%	10%	12%	10%

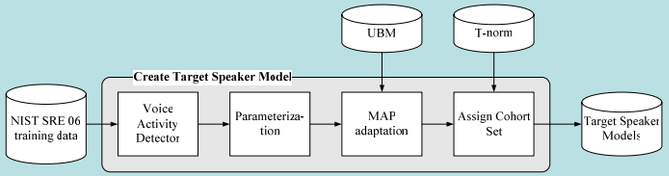
The flowchart below shows the training of the Tnorm speakers.



The Target speakers

A target speaker model is created for each model in the NIST trials. The target speaker model is trained by MAP adaptation of the UBM. A cohort set of 30 Tnorm speakers is assigned to the target speaker model. The cohort set is chosen as the Tnorm speaker models, which have the shortest weighted euclidean distance to the target model.

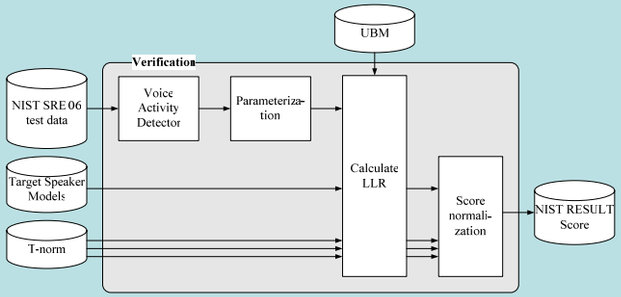
The flowchart below shows the training of the target speakers.



The Verification

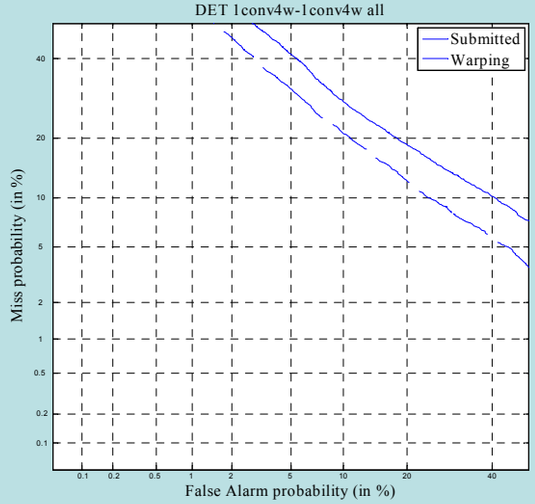
The LLR is calculated for the target speaker model. A test score normalization is performed as follows. The LLR for the test session is calculated for the 30 Tnorm speakers in the target speaker's cohort set. The mean and and standard deviation of the 30 Tnorm LLR are calculated, and used for normalizing the LLR of the target speaker model.

The flowchart below shows the verification process.



Result

The DET plot below shows the results from the mandatory NIST SRE 06 test. After the submission of the results to NIST we have incorporated feature warping [2] instead of CMS. The DET plot below shows the results for both the submitted system and the new system. It can be seen that feature warping gives a significant improvement of the verification performance.



Discussion

For this year of NIST we have implemented a GMM-UBM based speaker verification system. The result for the submitted system is shown in the result section. We assume that the performance is degraded because of channel mismatch between training and test data. To get a more channel robust system we incorporated feature warping after submission. The idea is that the features are mapped from a channel dependent distribution to a channel independent distribution. For our implementation of feature warping we have a separate distribution that consists of a mixture of 1024 gaussian distributions for each feature coefficient. The result for the feature warping is shown in the result section. As it can be seen the feature warping gives a significant performance increase.

Comparing our GMM-UBM system with systems using other types of feature warping we achieve comparable performance.

However, we still think that a more channel robust system can be developed to further increase performance. Recent research shows that channel compensation schemes based on factor analysis give promising results [3][4].

Refences
[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn "Speaker Verification using Adapted Gaussian Mixture Models", Digital Signal Processing 10
[2] Jason Pelecanos, and Sridha Sridharan, "Feature Warping for Robust Speaker Verification".
[3] Patric Kenny and Pierre Dumouchel, "Experiments in Speaker Verification using Factor Analysis Likelihood Ratios", The Speaker and Language Recognition Workshop 2004.
[4] Robbie Vogt, Brendan Baker, Sridha Sridharan, "Modelling Session Variability in Text-Independent Speaker Verification", Interspeech 2005