UNIFR-INT
University of Fribourg
Institut National des Télécommunications

Asmaa El Hannani

Dr Dijana Petrovska-Delacrétaz

NIST Speaker Recognition Workshop

26-27 June 2006

UNIVERSITAS FRIBURGENSIS

Overview

1. Motivation
2. Systems Description
   - GMM system
   - ALISP-Ngram system
   - ALISP-LM system
   - ALISP-Duration system
3. Results
4. Conclusions

1. Motivation

- Use the ALISP data-driven units instead of phonemes in speaker verification
- Exploit High-level features automatically derived using language and task independent technology
  - Data-driven speech segmentation ***ALISP*** (Automatic Language Independent Speech Processing) tools
  - No annotated databases needed
  - Language and task independent

2.1 GMM system

- Front-end
  - 16 Frequency Cepstral Coefficients + First order Deltas + Delta-energy
  - 20ms frames every 10ms
  - Only bands in 300-3400 Hz frequency range are used
  - The parameter vectors are normalized to fit a zero mean and a unit variance distribution

- Description
  - Based on ALIZE-LIA-SpkDet tools
  - The feature vectors are modeled by a 2048 GMMs
  - The background models are trained using Fisher and 2003 NIST SRE data
  - Speakers' models are trained via a MAP adaptation.
  - The verification is performed using the 10-best Gaussian components

ALISP recognizer

- Front-end
  - 15 Mel Frequency Cepstral Coefficients + energy + First order Deltas
  - 20ms frames every 10ms
  - Only bands in 300-3400 Hz frequency range are used
  - Cepstral Mean Subtraction is applied
- Description:
  - Gender dependent ALISP HMMs
  - 65 ALISP classes
  - Left-to-right HMMs having three emitting states and containing up to 8 Gaussians each
  - Trained on the (1999, 2001 and 2003) NIST SRE data

2.2 ALISP-Ngram system

- Exploiting Speakers-specific ALISP-sequences
  - Only ALISP sequences are used to model speakers
  - ALISP-sequences models are generated using a n-gram (1-2-3 gram) frequency count
  - For the scoring phase each ALISP-sequence is tested against a speaker specific model and a background model using a traditional likelihood ratio
  - The gender dependent background models are trained using Fisher and 2003 NIST SRE data.

2.3 ALISP-LM system

- Exploiting Speakers-specific ALISP-sequences
  - Label sequences produced by the ALISP recognizer are used to train ALISP trigrams using the HTK LM tools
  - The trigram language models is used to predict each symbol in the sequence given its tow predecessors.
  - Speaker models created by interpolation:
    - 8c-1c: Speaker-models = 0.3 BM + 0.8 Speaker-data-model
    - 1c-1c: Speaker-models = 0.9 BM + 0.1 Speaker-data-model

  - The gender dependent background models are trained using Fisher and 2003 NIST data.

2.4 ALISP-Duration system

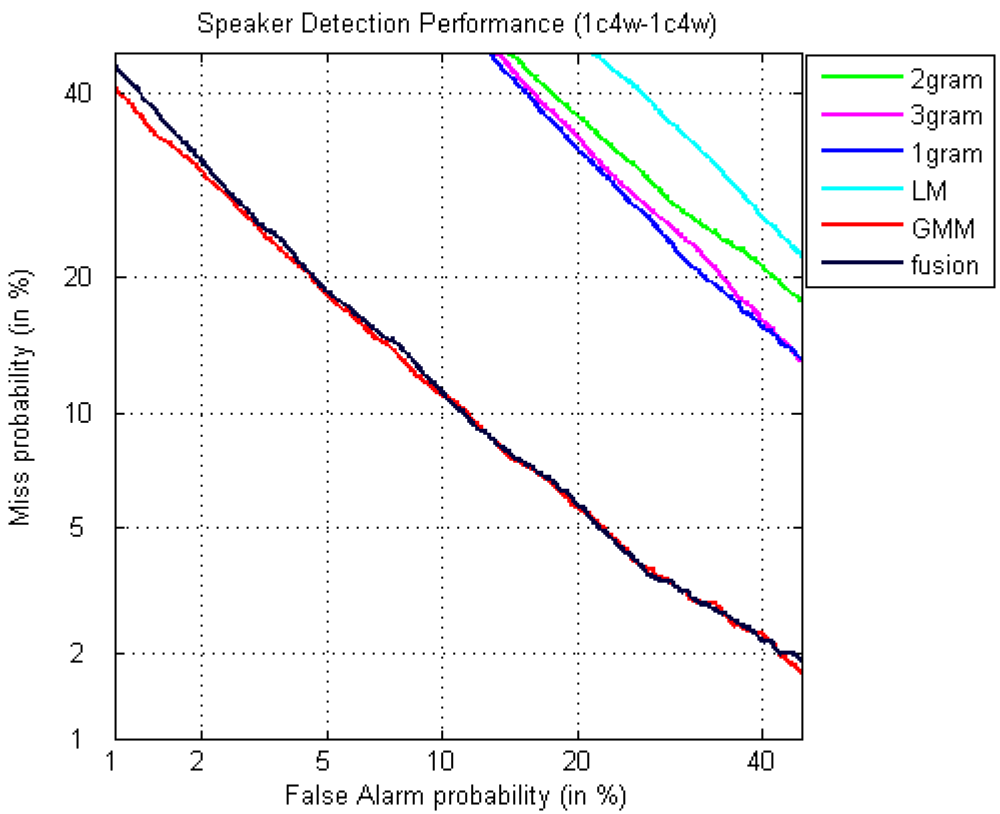
- Exploiting Speakers-specific ALISP-duration
  - Each ALISP unit is represented by a feature vector comprised of its duration
  - A background GMM of 8 mixtures is trained for each ALISP class (65 models)
  - Target models (65 models per speaker) are trained via MAP adaptation.
  - The gender dependent background models are trained using Fisher and 2003 NIST data.

2.5 FUSION

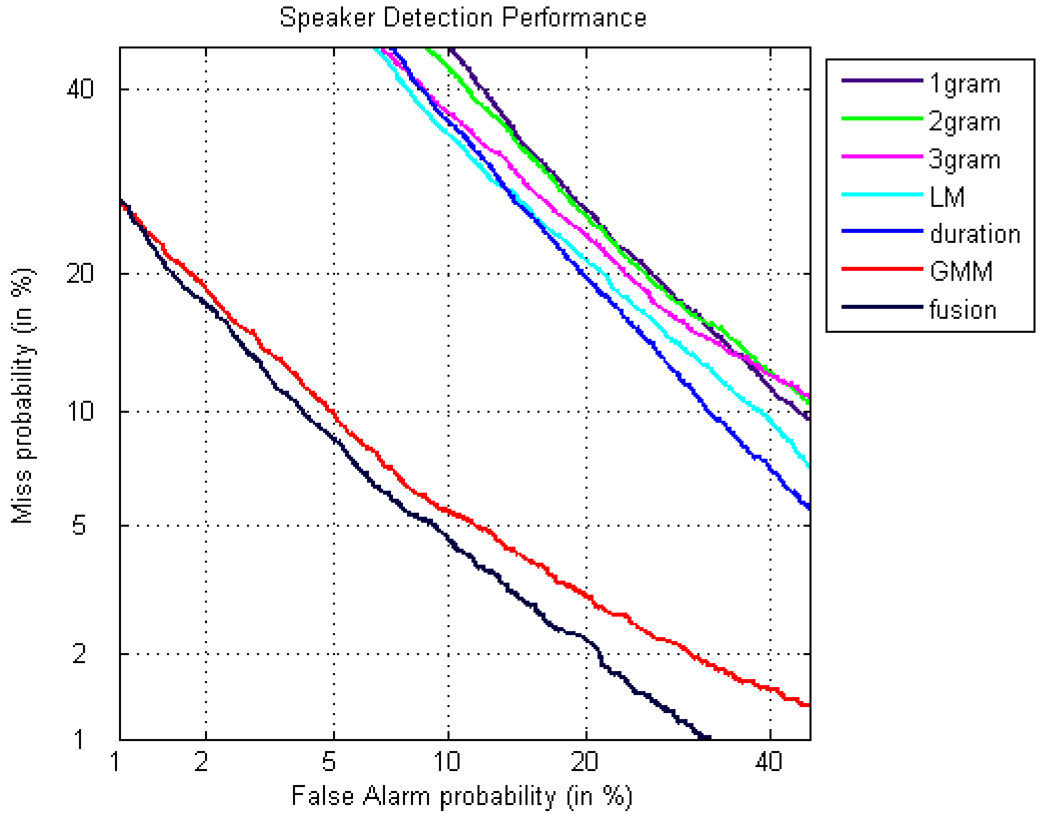
- Scores from the systems:
  - GMM
  - ALISP-Ngram
  - ALISP-LM
  - ALISP-Duration (post-eval)

are fused with an SVM

3. Results (1c4w-1c4w)



3. Results (8c4w-1c4w)



4. Conclusions

- Using data-driven (ALISP) segmentation, instead of the phonetic segmentation, for speaker verification

- Fusing the ALISP and GMM systems improves speaker recognition results for the 8c4w-1c4w task but not for the 1c4w-1c4w task