UNIFR-INT

University of Fribourg

Institut National des Télécommunications-Evry

PhD. Asmaa El Hannani

Dr. Dijana Petrovska-Delacrétaz

NIST Speaker Recognition Workshop

7-8 June 2005

Overview

1. Motivation
2. Description of
   - ø different systems and
   - ø their fusion
3. Experimental setup
4. Results
5. Conclusions

1. Motivation

- Exploit data-driven speech segmentation for the speaker verification task
  - Using ***ALISP***

    Automatic Language Independent Speech Processing
    - No annotated databases needed
    - Language and task independent

- Use the ALISP data-driven segmentation in different ways in order to extract complementary types of information

2. Submitted systems

- Description of different systems
  - GMM system
  - GMM-ALISP system
    - "frame-based" ALISP scoring
  - HMM-ALISP system
    - modeling of the ALISP segments with HMM
  - ALISP-Ngram systems

- Fusion of different systems with Multilayer Perceptrons (MLPs)

2.1 GMM system

- GMM system (Based on BECARS free-software)
  1. Background models
     - 512 Gaussians
  2. Target models
     - MAP adaptation
  3. Scoring

2.2 GMM-ALISP system

- ALISP-GMM system
  1. Background models
  2. Target models
  3. Segmental scoring
     - Frame score:

$$s_t = \log \hat{p}(y_t \mid X) - \log \hat{p}(y_t \mid \overline{X})$$

     - ALISP segment score:

$$S_i = \frac{1}{N} \sum^{N} s_t$$

N : number of frames in the ALISP segment i

  4. Segmental score fusion

2.3 HMM-ALISP system

- ALISP-HMM system (Based on HTK tools)
  1. 64 Background models (1 model per ALISP class)
  2. Target models (64 models per speaker)
  3. Segmental scoring
  4. Segmental score fusion

2.4 ALISP-Ngram system

- Exploiting Speakers-specific ALISP-sequences
  - Only ALISP sequences are used to model speakers
  - ALISP-sequences models are generated using an n-gram frequency count:

1. Background model : $L_{Bm}(k) = \frac{C_{Bm}(k)}{\sum_{n=1}^{N} C_{Bm}(n)}$
2. Speaker model : $L_i(k) = \frac{C_i(k)}{\sum_{n=1}^{N} C_i(n)}$
3. Scoring : $S_{ti} = \frac{\sum_{n=1}^{M} (C_t(n).log\,[L_i(n) - L_{Bm}(n)])}{\sum_{n=1}^{M} C_t(n)}$

2.5 FUSION

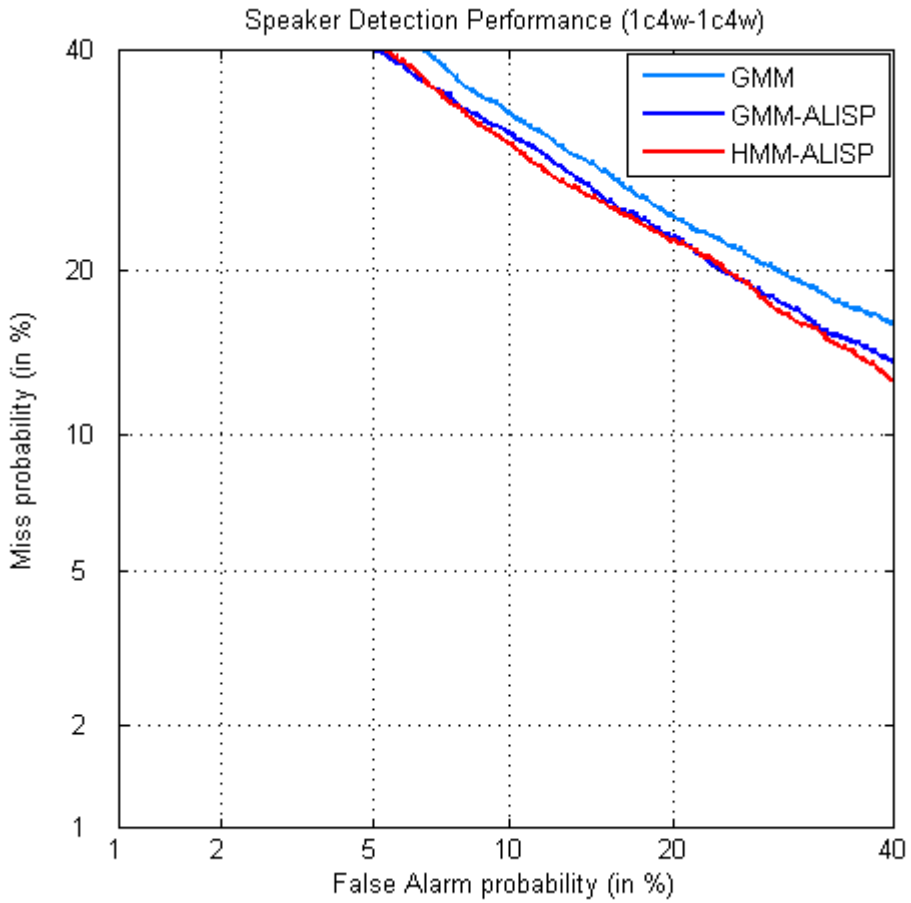
- Scores from systems:
  - GMM-ALISP
  - HMM-ALISP
  - ALISP-NGRAM

  were fused with a Multi-Layer Perceptron (MLP)

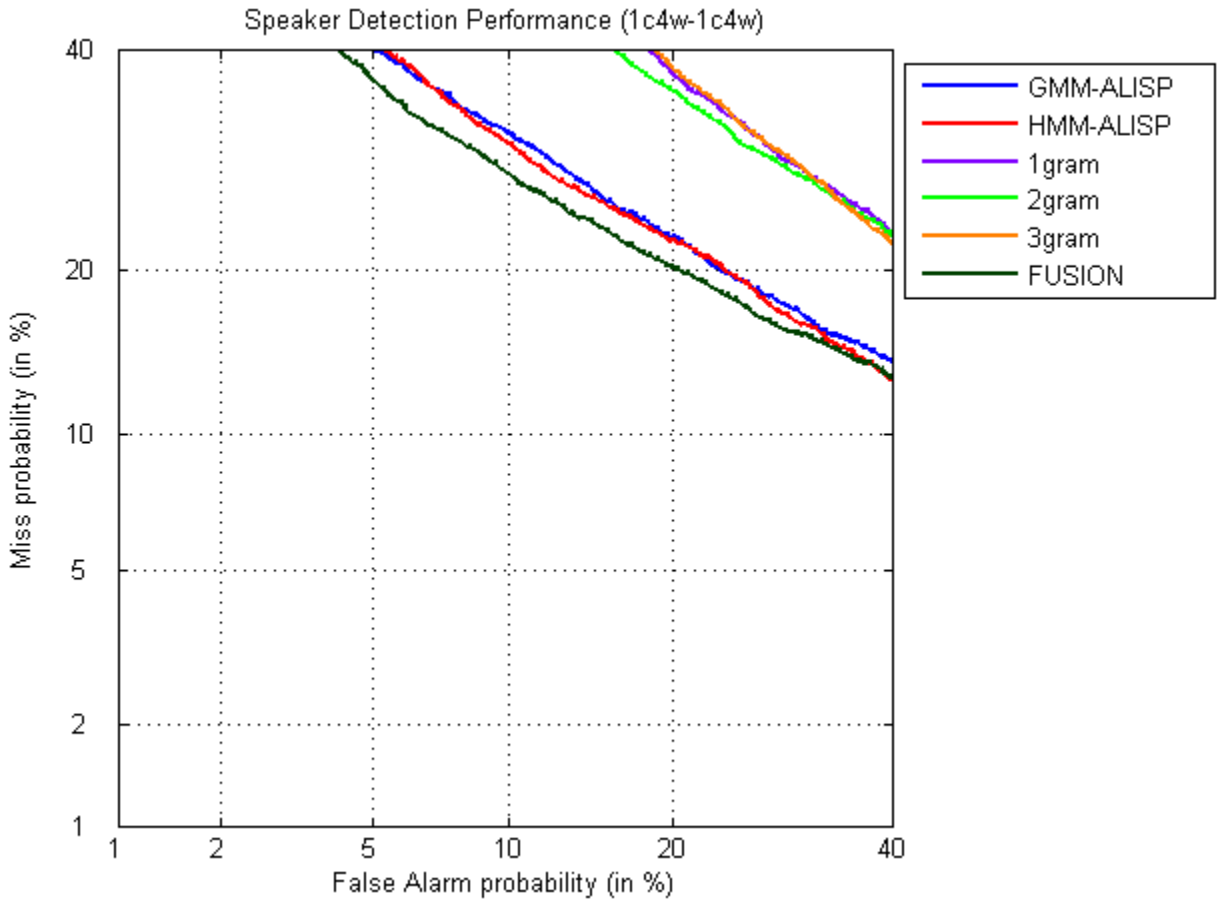
3. Experimental Setup

- 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
- Two-step silence removal
- GMM background models and ALISP recognizer are:
  - Gender dependent
  - Trained on 1999 and 2001 NIST data (with approx. 6 hours of speech data)
- Speakers' models with
  - MAP adaptation
- ALISP
  - 64 ALISP classes

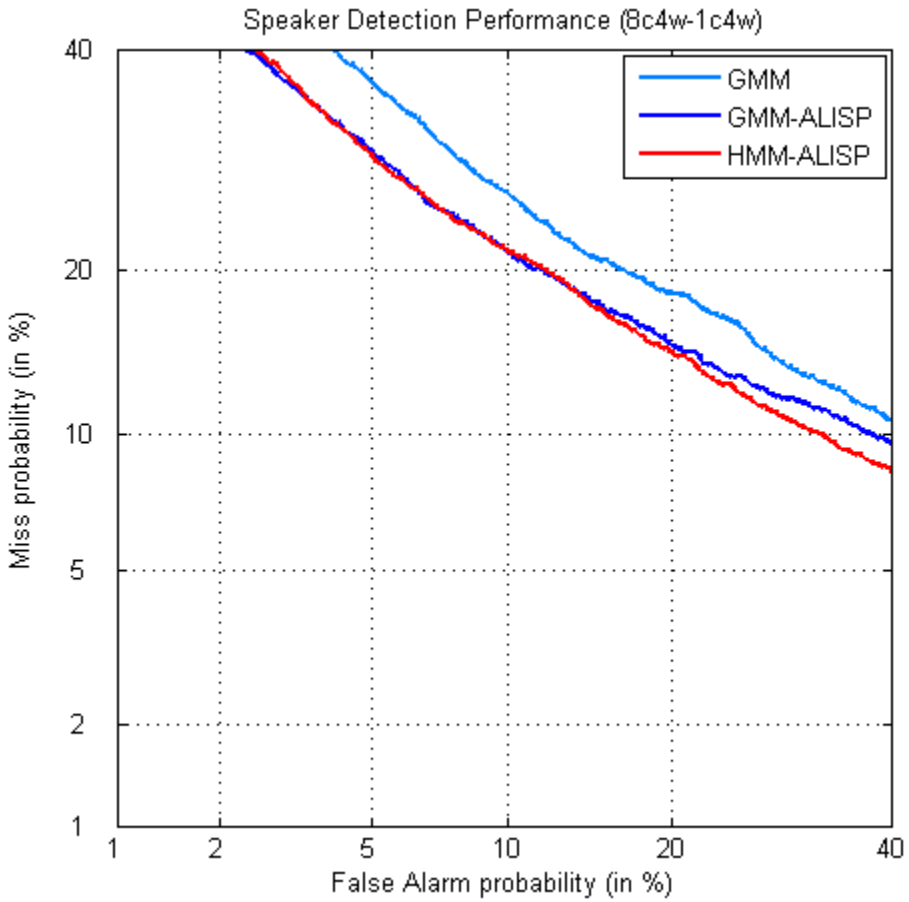
4.1 Individual results (1c4w-1c4w)



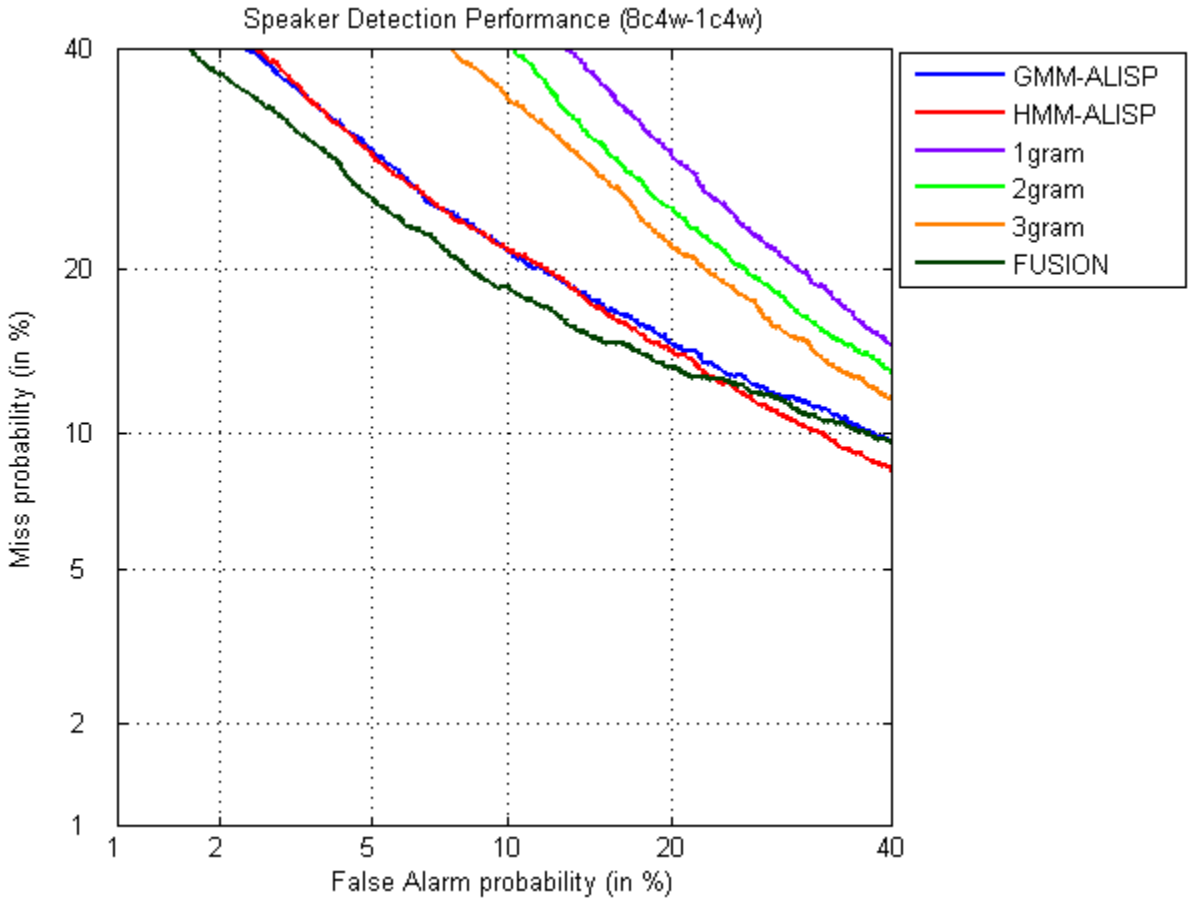
4.2 Fusion results (1c4w-1c4w)



4.3 Individual Results for GMM and HMM based systems (8c4w-1c4w)



4.4 Fusion Results for (8c4w-1c4w)



4.5 Confidence Intervals

	EER	95% Conf. Interval
1c4w-1c4w		
GMM system	22.8 %	± 0.5
GMM-ALISP system	21.5 %	± 0.5
HMM-ALISP system	21.6 %	± 0.5
Fusion	20.2 %	± 0.5
8c4w-1c4w		
GMM system	18.8 %	± 0.5
GMM-ALISP system	16.7 %	± 0.5
HMM-ALISP system	16.3 %	± 0.5
Fusion	15.1 %	± 0.5

5. Conclusions

- Some “bells and whistles” missing to the baseline GMM’s !
- Improvement of our baseline GMM based systems using “simple” speaker specific frequency counts of ALISP-sequences
  - 6% for 1c4w-1c4w
  - 7% for 8c4w-1c4w