IM2.MPR

Themes at core to Multimodal Processing and Recognition

Multimodality means:

- Take inspiration from methods for one modality to analyze another modality
  - Features from vision used for speech analysis

- Combine modalities to enhance recognition
  - Speaker Localization
  - Speech recognition
  - Mutual information versus redundancy

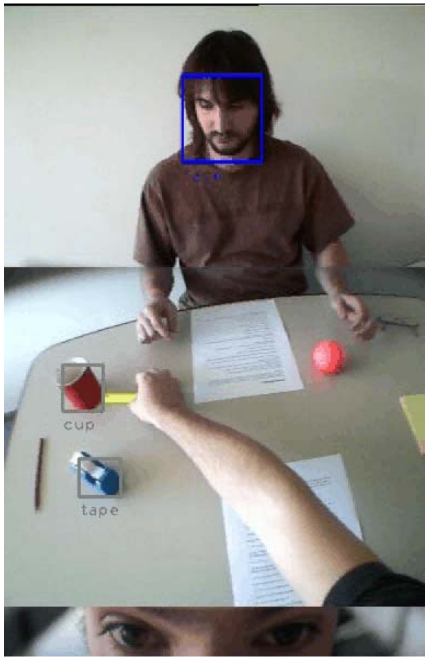
- Combine modalities to allow higher-level analysis of scenes
  - Focus of attention and its role in speaker-listener interaction
  - Turn-taking and leadership
  - Biometric classification

IM2.MPR

Multimodal Processing and Recognition

Structure of the IP: **8 Teams**

- UNIGE (Voloshynovskiy)
- IDIAP (Fleuret, Gatica, Marcel, Dines/Friedland)
- EPFL (Billard, Drygajlo, Thiran)

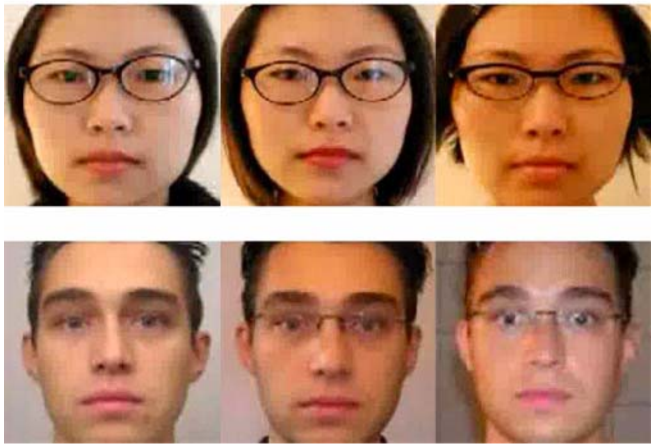


Determining the object focus of attention

Audio-Visual Analysis of Scenes

Determining the dominant person in a meeting

User Authentification



Biometric Identity Verification

IM2.MPR

Themes at core to multimodal data analysis

Multimodality means:

- Take inspiration from methods for one modality to analyze another modality
  - Determine the best features for further analysis

- Combine modalities to enhance recognition
  - Speaker Localization
  - Speech recognition
  - Mutual information versus redundancy

- Combine modalities to allow higher-level analysis of scenes
  - Focus of attention and its role in speaker-listener interaction
  - Turn-taking and leadership
  - Biometric classification

IDIAP – A. Roy, S. Marcel

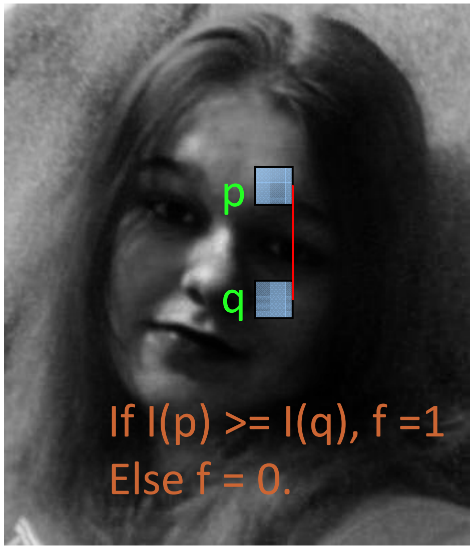
Speaker Authentication - *Fern-Audio Features*

- Investigate novel audio features for speaker authentication in the presence of additive white Gaussian noise

- ~ analogous to visual object detection under varying illumination conditions

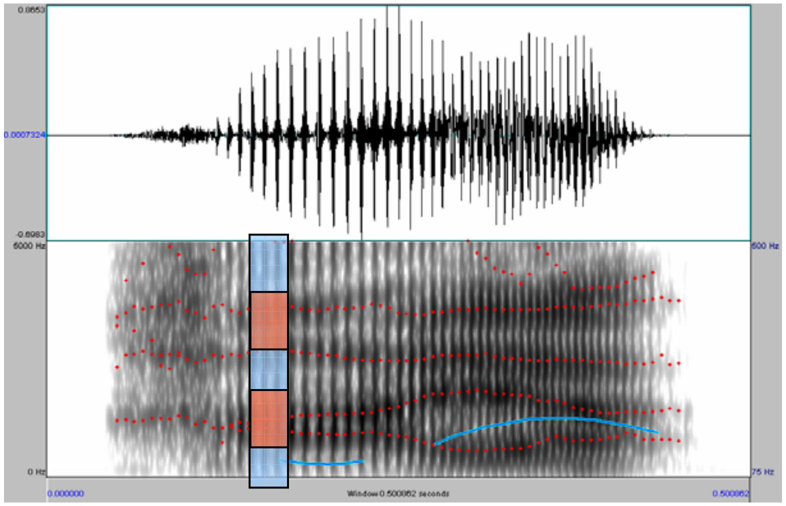
  – Inspired by binary features (Ferns, LBP) in computer vision.

  – Multiple pairs are combined to achieve robust to noise (illumination) object Recognition.



Speaker Authentication - *Fern-Audio Features*

- Considering a spectral vector as an 1D-image
- Boosting used to find frequency pairs in time



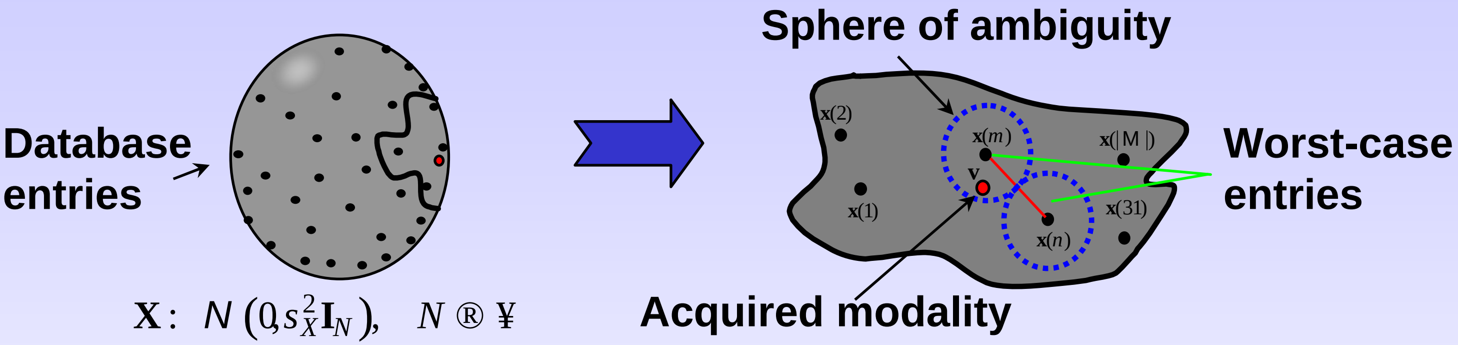
1-D image

- It outperforms a MFCC-GMM baseline in noisy conditions

UNIGE - O. Koval and S. Voloshynovskiy

Performance analysis of one-vs-one multimodal worst-case classification with independent modalities in projected domain:

- Random projections can be considered as an efficient solution to performance/complexity/storage trade-of;



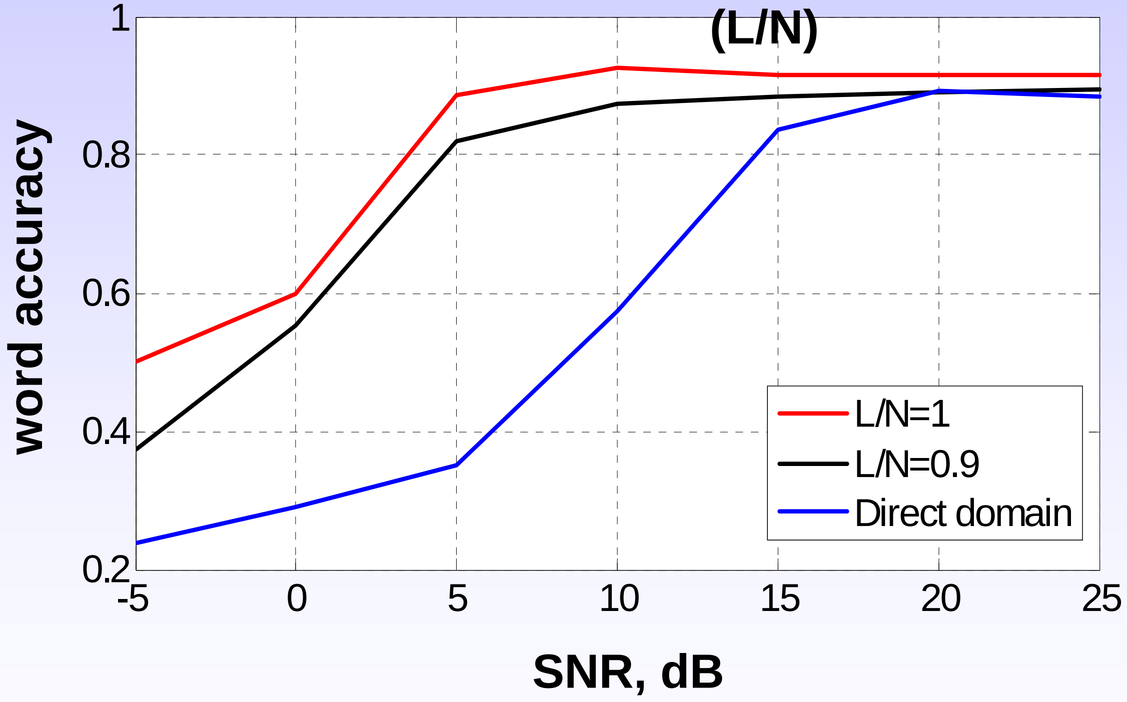
$\mathbf{X} \sim \mathcal{N}(0,\sigma_X^2\mathbf{I}_N),\quad N \to \infty$

- Authentication distortions models analysis: the use of Gaussian distribution to model authentication distortions is justified for the projected domain when the projector is generated from a certain statistical distribution. Furthermore, it is demonstrated that in the direct domain one-vs-one classification some distortion models exist that are discrete and can be used to lower bound the performance of authentication more accurately than Gaussian distribution;

Performance analysis of one-vs-one multimodal worst-case classification with independent modalities in projected domain:

- Practical impact of random projections: random projections satisfy assumption of Gaussian pdf for the output → useful in HMM-based recognition scenarios

Audio HMM-based speech recognition with random projections for two ratios (L/N)



N - dimensionality of the projection input;

L – dimensionality of the projection output.

IM2.MPR

Themes at core to multimodal data analysis

Multimodality means:

- Take inspiration from methods for one modality to analyze another modality
  - Ferns feature for speech analysis

- Combine modalities to enhance recognition
  - Speaker Localization
  - Speech recognition
  - Mutual information versus redundancy

- Combine modalities to allow higher-level analysis of scenes
  - Focus of attention and its role in speaker-listener interaction
  - Turn-taking and leadership
  - Biometric classification

IM2.MPR

Themes at core to multimodal data analysis

Multimodality means:

- Take inspiration from methods for one modality to analyze another modality
  - Ferns feature for speech analysis
- Combine modalities to enhance recognition
  - Speaker Localization
  - Speech recognition
  - Mutual information versus redundancy
- Combine modalities to allow higher-level analysis of scenes
  - Focus of attention and its role in speaker-listener interaction
  - Turn-taking and leadership
  - Biometric classification

EPFL – M. Gurban & J-Ph. Thiran

Speaker Localization & Speech Recognition

Speaker localization

A speaker localization method based using the joint probability density of optical flow differences and audio energy.

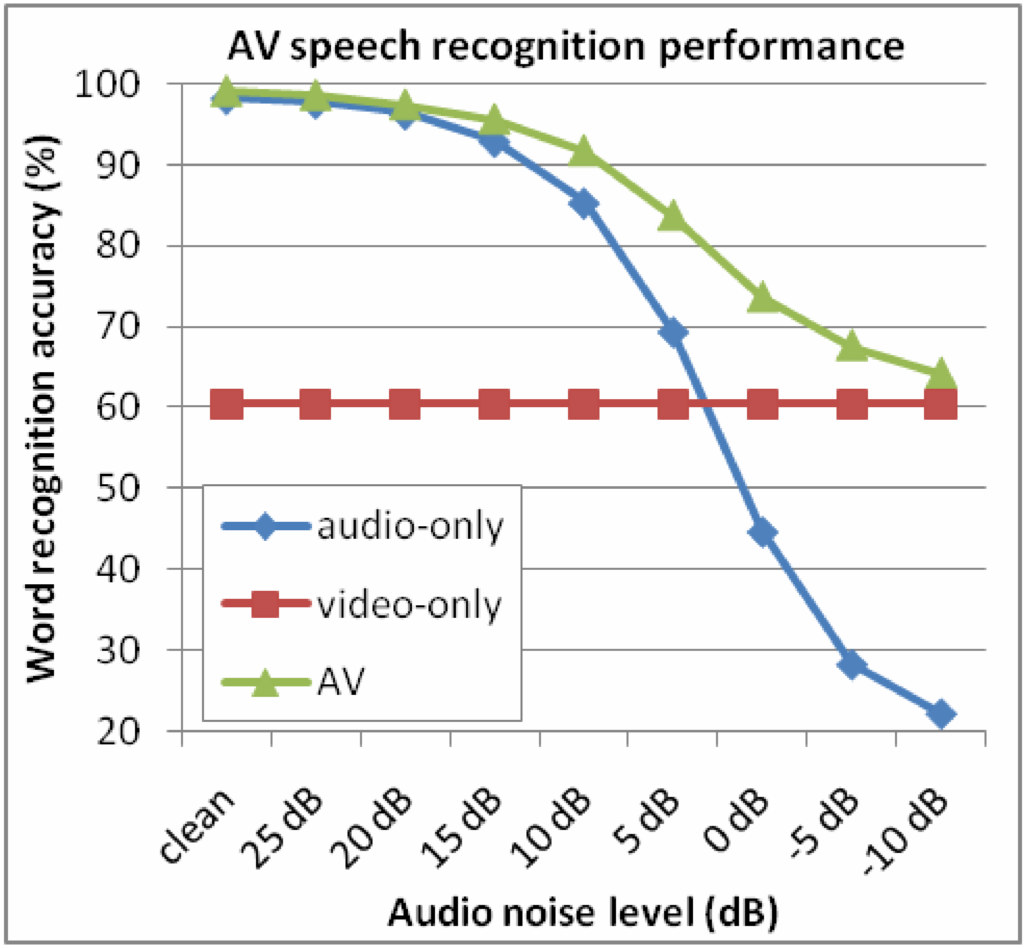
Feature extraction for audio-visual speech recognition

Novel low-dimensional visual features based on optical flow

Feature selection methods using mutual information for maximum relevance and also including a penalty for redundancy.

EPFL – M. Gurban & J-Ph. Thiran

Speaker Localization & Speech Recognition



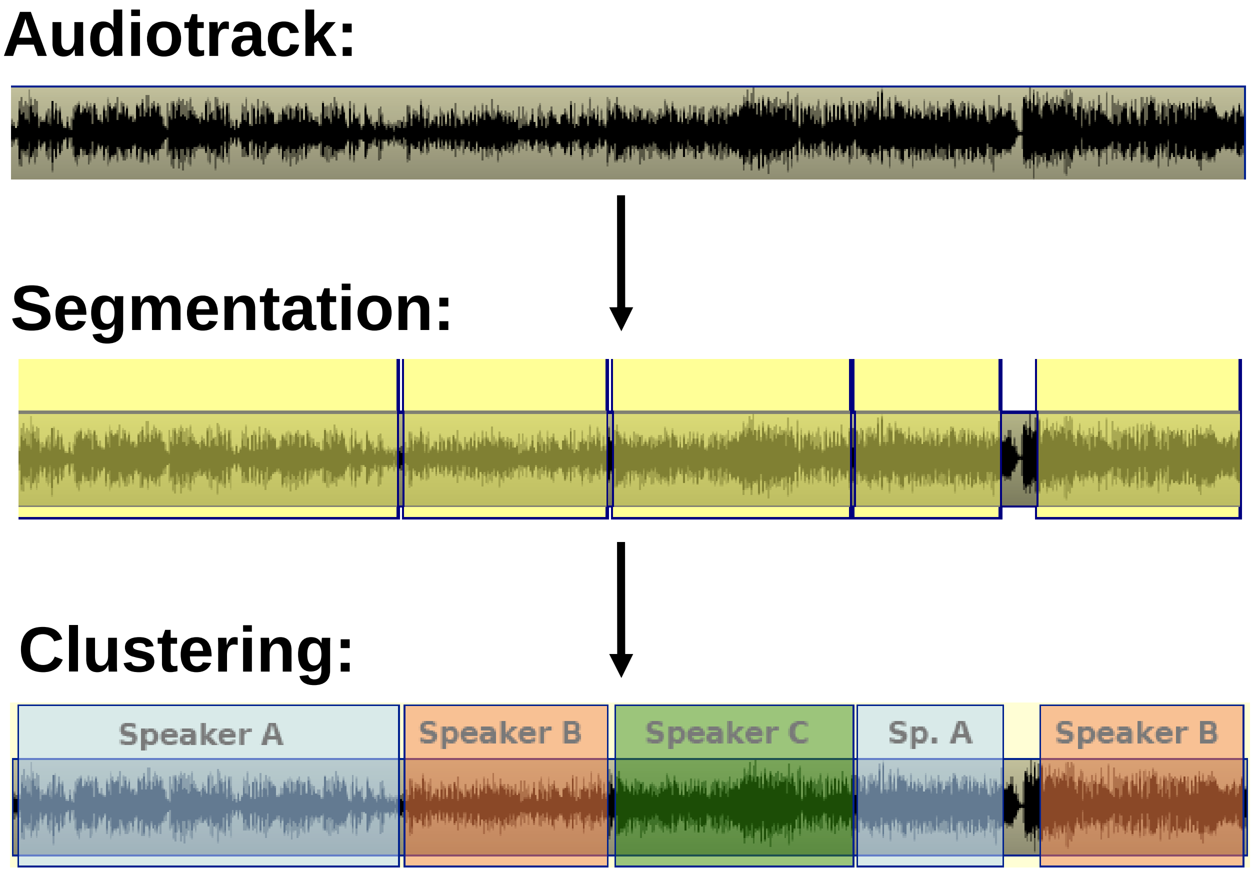
Multimodal integration for audio-visual speech recognition

- An adaptive stream weighting method based on the entropies of instantaneous stream posterior distributions
- Asynchronous models for audio-visual speech classifiers
- Addition of a processing step aligning the audio and visual

Dealing with asynchrony in audio-visual speech recognition, Virginia Estellers (EPFL)
Schedule 17:30 – 18:00

IDIAP – G. Friedland

Classical Diarization



Audiotrack:

Segmentation:

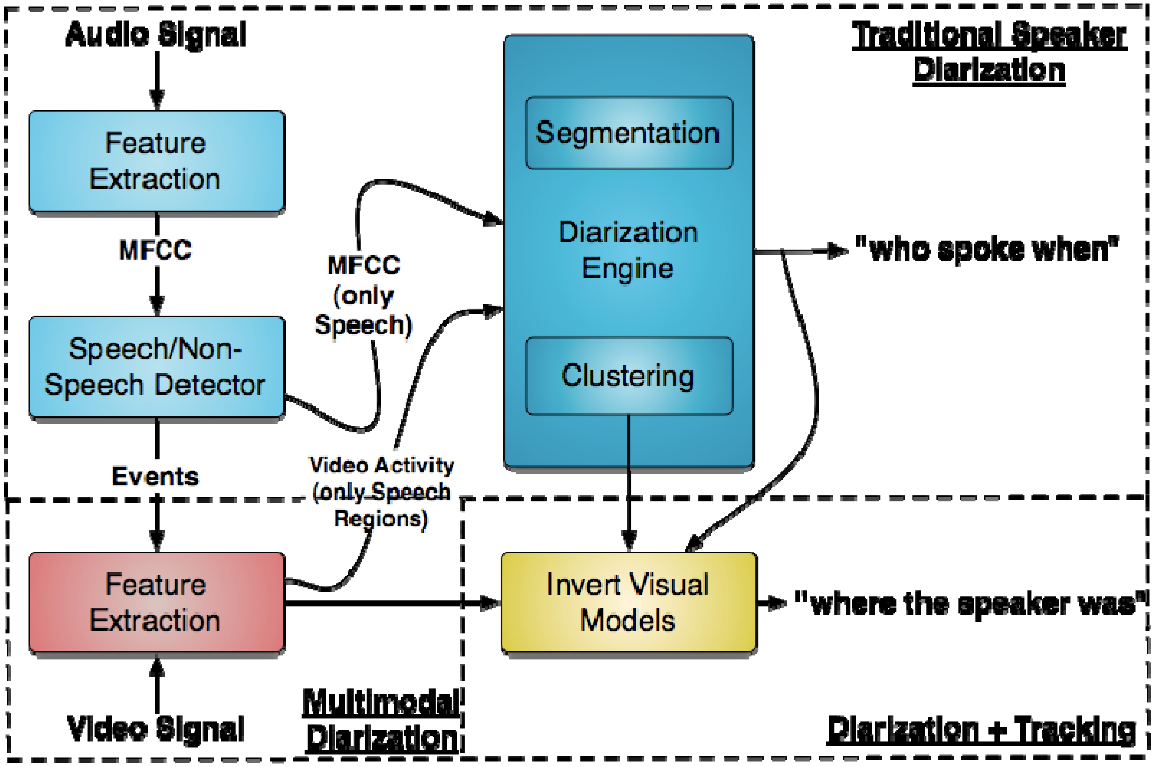
Clustering:

"Who spoke when?"

IDIAP – G. Friedland

Joint Speaker Diarization and Tracking

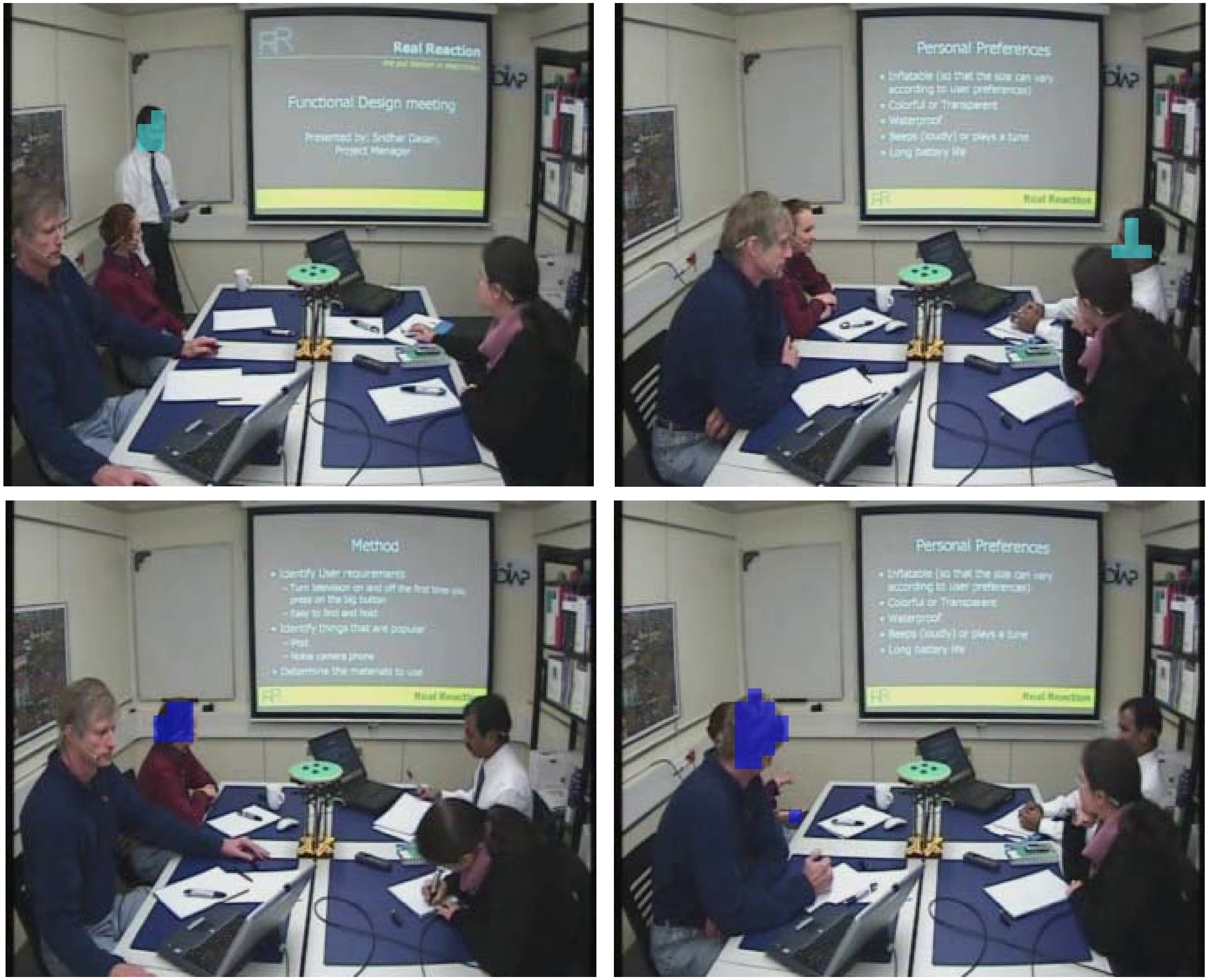
Treat Speaker Diarization and Speaker Localization as a joint unsupervised optimization problem.



Single-camera, single-mic case:

higher accuracy (30%) of diarization at low computational overhead

Example: Obfuscated Speaker tracked



Speaker localisation as a by-product: Robust against visual changes such as different cloth, occlusions, etc...

IM2.MPR

Themes at core to multimodal data analysis

Multimodality means:

- Take inspiration from methods for one modality to analyze another modality
  - Ferns feature for speech analysis
- Combine modalities to enhance recognition
  - Speaker Localization
  - Speech recognition
  - Mutual information versus redundancy
- Combine modalities to allow higher-level analysis of scenes
  - Focus of attention and its role in speaker-listener interaction
  - Turn-taking and leadership
  - Biometric classification

IDIAP – G. Garau

Audio Visual Speaker Diarisation

- Estimating “who spoke when” using audio and visual cues

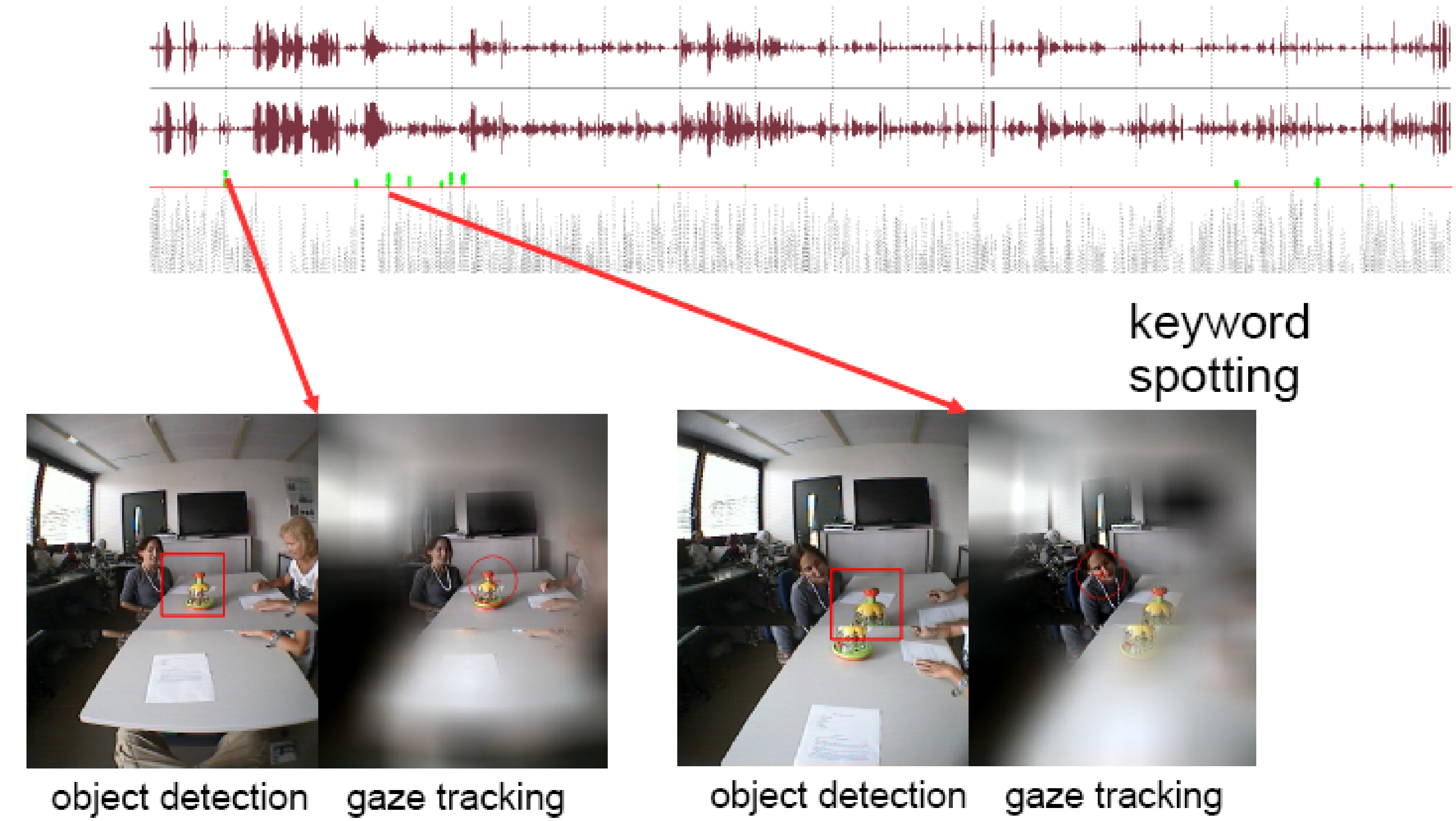
- Using psychology inspired visual features:
  - Visual Focus of Attention: *role of gaze in a conversation*:
    - **Listeners mostly look at the person who’s talking**
    - **Speakers look at the person they are addressing**
    - VFoA features were defined as a **measure of the number of persons looking at each meeting participant**

      (Experiments both on manually annotated and automatic VFoA)
    - **Head pose likelihoods (i.e. probabilities that each meeting participant is looking at a given target)** were also investigated

  - Motion features: *speaker’s movement for speech production and use of gestures for conversation floor management*
    - For each close-up camera the average pixel-by-pixel difference between adjacent frames was computed

EPFL: Basilio Noris, Martin Duvanel, Weifeng Li, Aude Billard
IDIAP: Johnny Marietthoz,Francois Fleuret

Combining Keyword Spotting, Gaze Tracking and Object Detection



Concurrent Speech and Gaze **Disjointed Speech and Gaze**

EPFL: Basilio Noris, Martin Duvanel, Aude Billard
IDIAP: Johnny Marietthoz,Francois Fleuret

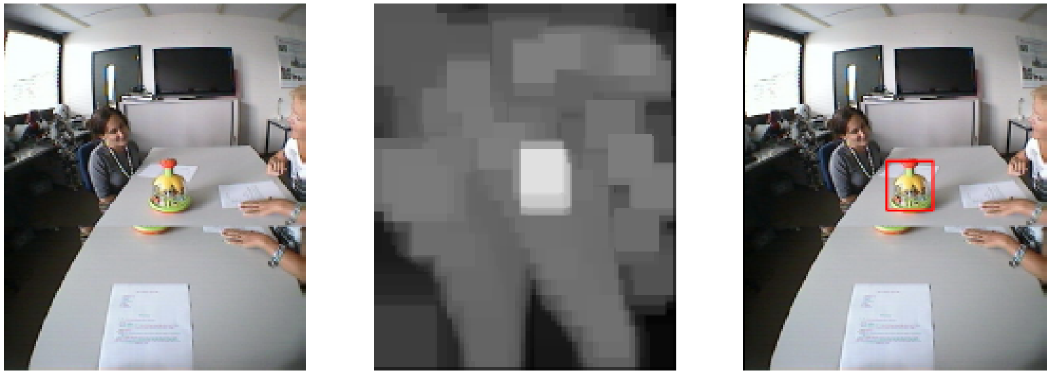
Combining Keyword Spotting, Gaze Tracking and Object Detection

Keyword Spotting:

- A left-right model Word Hypothesis Phoneme sequence
- A fully connected model for garbage

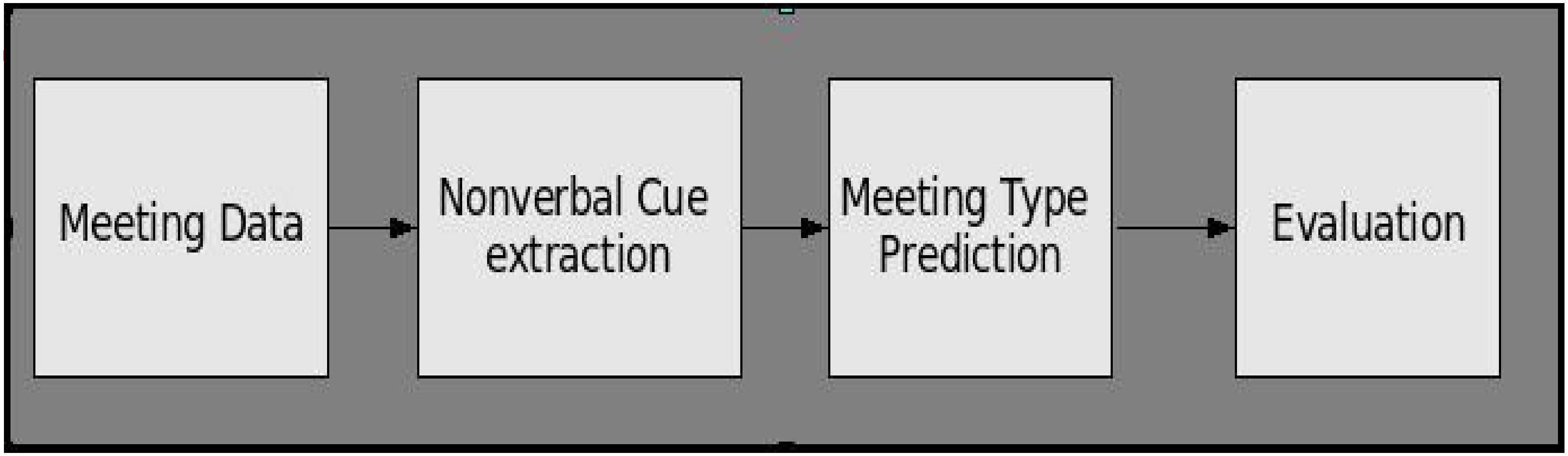
Object Detection with limited amount of samples

- user-based positive/negative examples for training set
- Adaboost on mixture of SVM classifiers



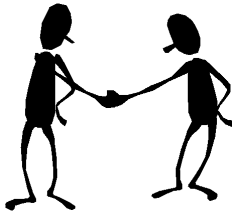
IDIAP – D. Jayagopi, J. Biel, D. Gatica-Perez

Classifying group dynamics



Meeting Data → Nonverbal Cue extraction → Meeting Type Prediction → Evaluation

AMI Dataset



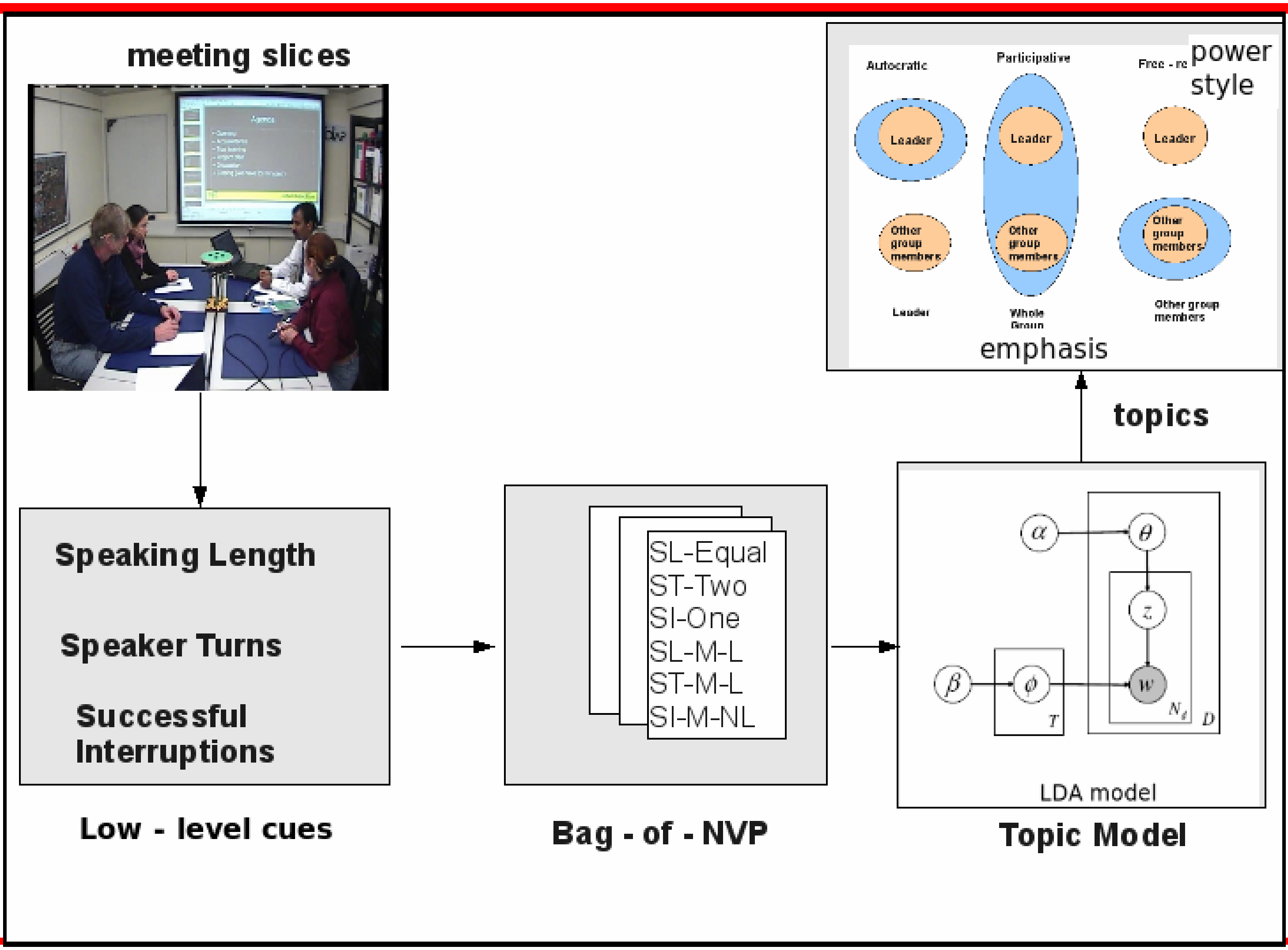
Apprentice Dataset

Characterization of group by the aggregation (both temporal and person-wise) of their nonverbal behaviour

Cue fusion - Naive Bayes Classifier and SVM with quadriatic kernel

Best cues – up to 100 % accuracy

Mining group dynamics



meeting slices

Speaking Length

Speaker Turns

Successful Interruptions

Low - level cues

SL-Equal
ST-Two
SI-One
SL-M-L
ST-M-L
SI-M-NL

Bag - of - NVP

LDA model

Topic Model

topics

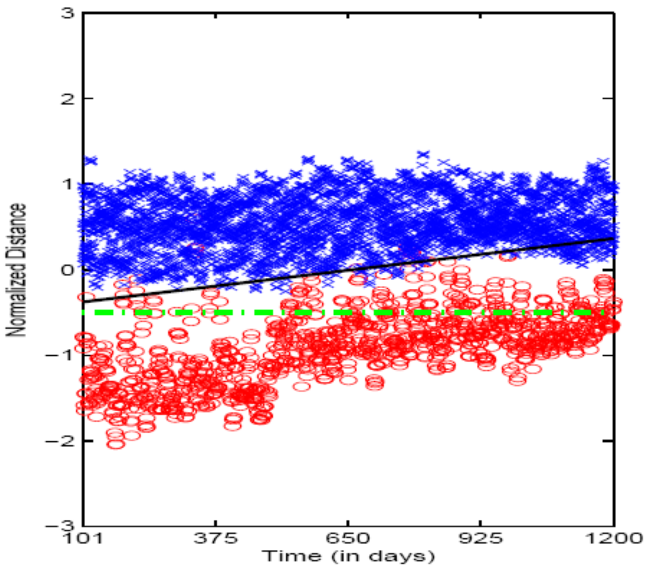
power style

emphasis

EPFL: Andrzej Drygajlo, Weifeng Li, Kewei Zhu

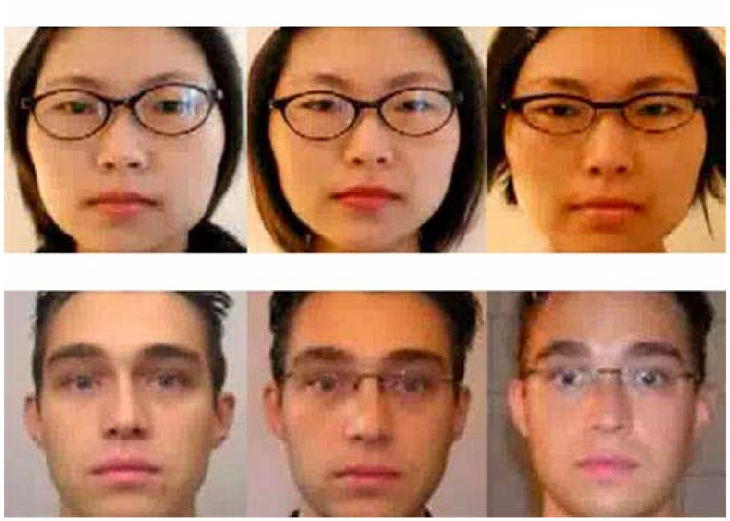
Reliable biometric classification in adverse environmental conditions

- Incorporating age into the biometric recognition process. Age as metadata quality measure.
- Experiments using real-world data recorded every day during more than 3 years and MORPH database + TV series “friends”
- Reduces the error rates below those of baseline classifier created at the time of enrolment.



threshold of baseline classifier

Q-stack decision boundary



Quality Measures and Stacking Classifiers in Multimodal Biometric Recognition

<u>Speaker</u> Andrzej Drygajlo and Weifeng Li (EPFL), <u>Schedule</u> 16:30 – 17:00

IM2.MPR

Themes at core to multimodal data analysis

Multimodality means:

- Take inspiration from methods for one modality to analyze another modality
  - Ferns feature for speech analysis

- Combine modalities to enhance recognition
  - Speaker Localization
  - Speech recognition
  - Mutual information versus redundancy

- Combine modalities to allow higher-level analysis of scenes
  - Focus of attention and its role in speaker-listener interaction
  - Turn-taking and leadership
  - Biometric classification