NIST Speaker Recognition Evaluation 2006

Speech@FIT, BRNO UNIVERSITY OF TECHNOLOGY
(STBU consortium)

Lukas Burget, Pavel Matejka, Petr Schwarz, Ondrej Glembek, Martin Karafiat, Jan Cernocky and Frantisek Grezl

Outline

- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks

Submitted systems

- BUT01 - primary (6 systems)
  - GMM with and without T-norm
  - SVM GMM with and without T-norm
  - SVM MLLR with and without T-norm
- BUT02 - (3 systems)
  - GMM with T-norm
  - SVM GMM with T-norm
  - SVM MLLR with T-norm
- BUT03 - (1 system)
  - Only GMM without T-norm

Outline

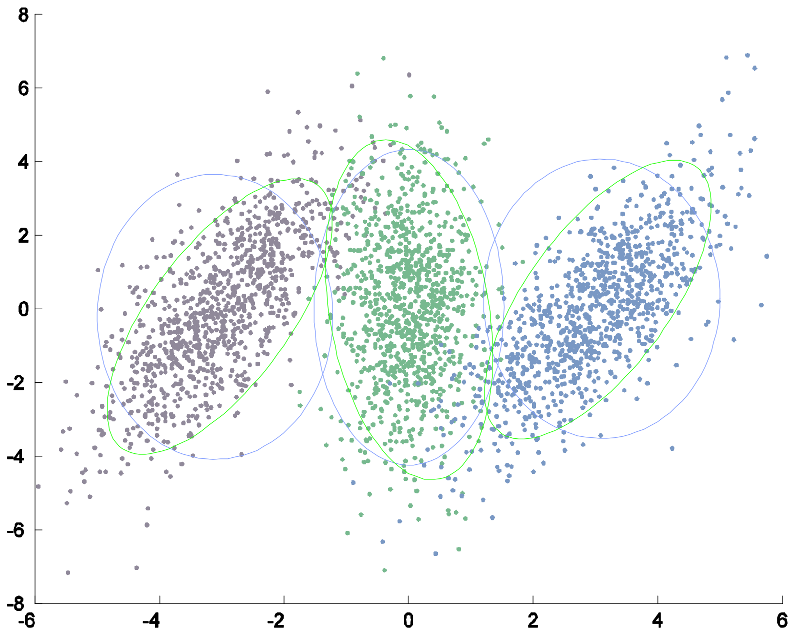
- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks

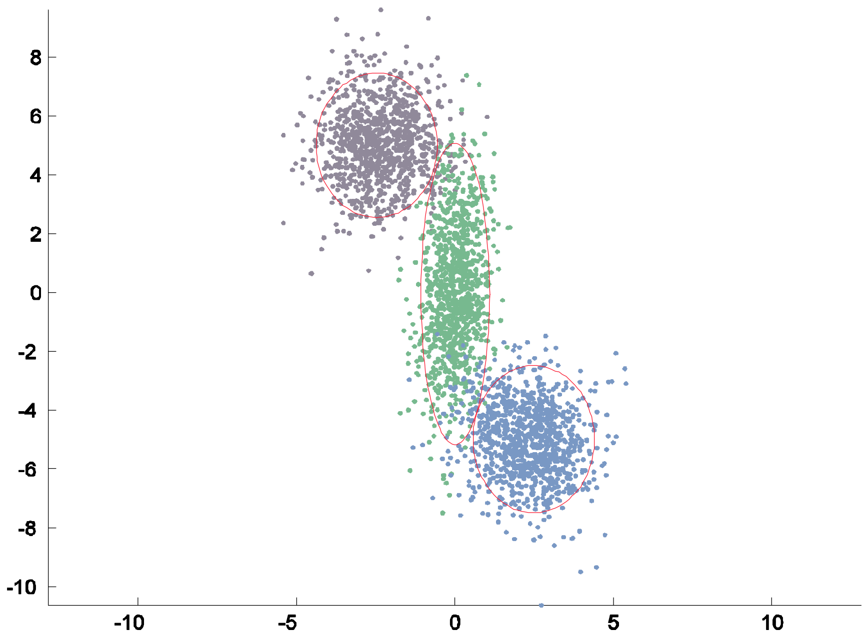
GMM System

- MAP adapted UBM with 2048 Gaussian components
  - Single UBM trained on NIST 2004 test data
- 12 MFCC + C0 (20ms window, 10ms shift)
- Cepstral mean normalization (over whole conversation)
- Short time Gaussianization
  - Rank of current frame coefficient in 3sec window transformed by inverse Gaussian cumulative distribution function.
- RASTA filtering
- Delta + double delta + triple delta coefficients
  - Together 52 coefficients, 12 frames context
- HLDA (dimensionality reduction from 52 to 39)
- Feature Mapping (7 channels, 2 gender)
- Eigen-channel adaptation
  - 30 eigen-channels derived on 310 speakers from NIST 2004
- T-norm: 230 speakers from NIST 2002

HLDA

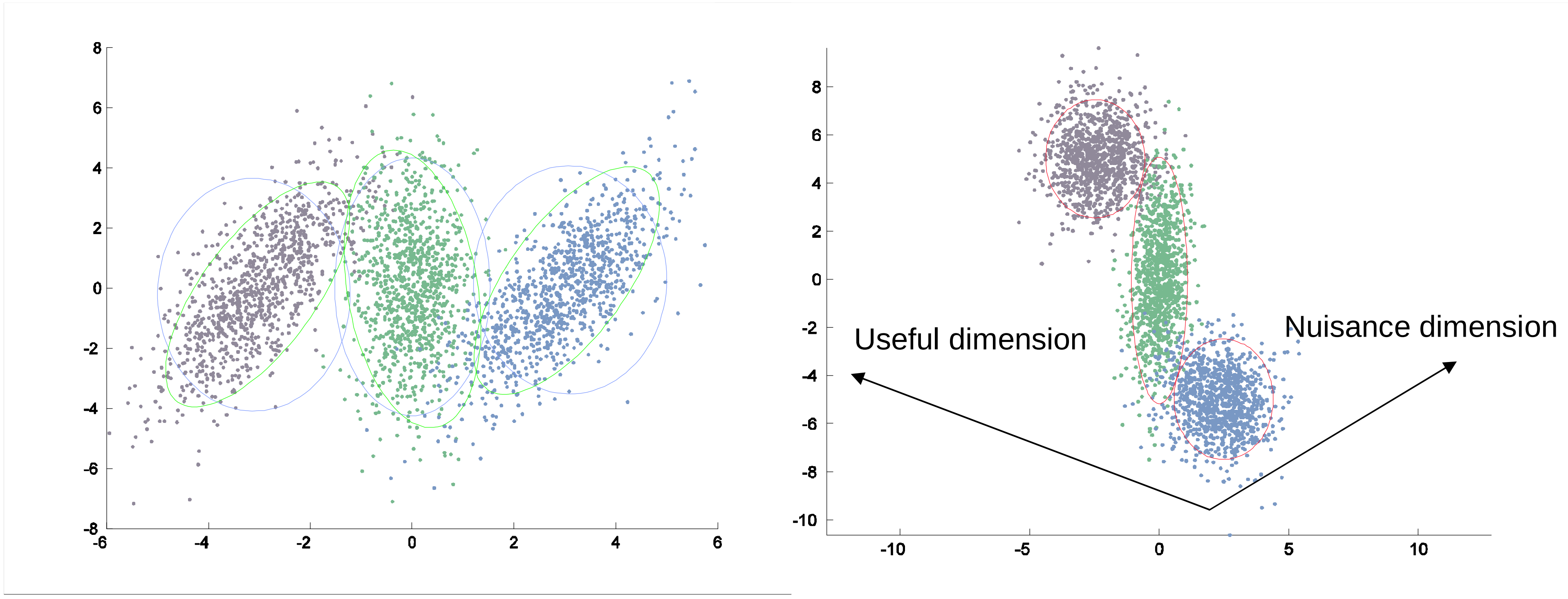
Heteroscedastic Linear Discriminant Analysis provides a linear transformation that de-correlates classes.





HLDA

HLDA allows for dimensionality reduction while preserving the discriminability between classes (HLDA without dim. Reduction is also called MLLT)



Feature Mapping

- 2004 data used for training
- Supervised adapted channel models
  - 3 channels per gender (cell,cord,stnd) derived from 2004 data
- Unsupervised adapted channel models [Mason2005]
  - Initial clustering given by recognition FM output from TNO SRE 2005 (4 channels (elec, cord, gsm, cdma) - per gender)
  - Iteration on NIST 2004 data
  - In each iteration:
    - One model is adapted for each cluster of conversations
    - Conversations are re-clustered by new models
  - Converges in about 20 iterations
- All 14 models from both supervised and unsupervised adaptation used for feature mapping
- Feature mapping is **not important when applied together with eigen-channel adaptation!**

Eigen-channel adaptation I.

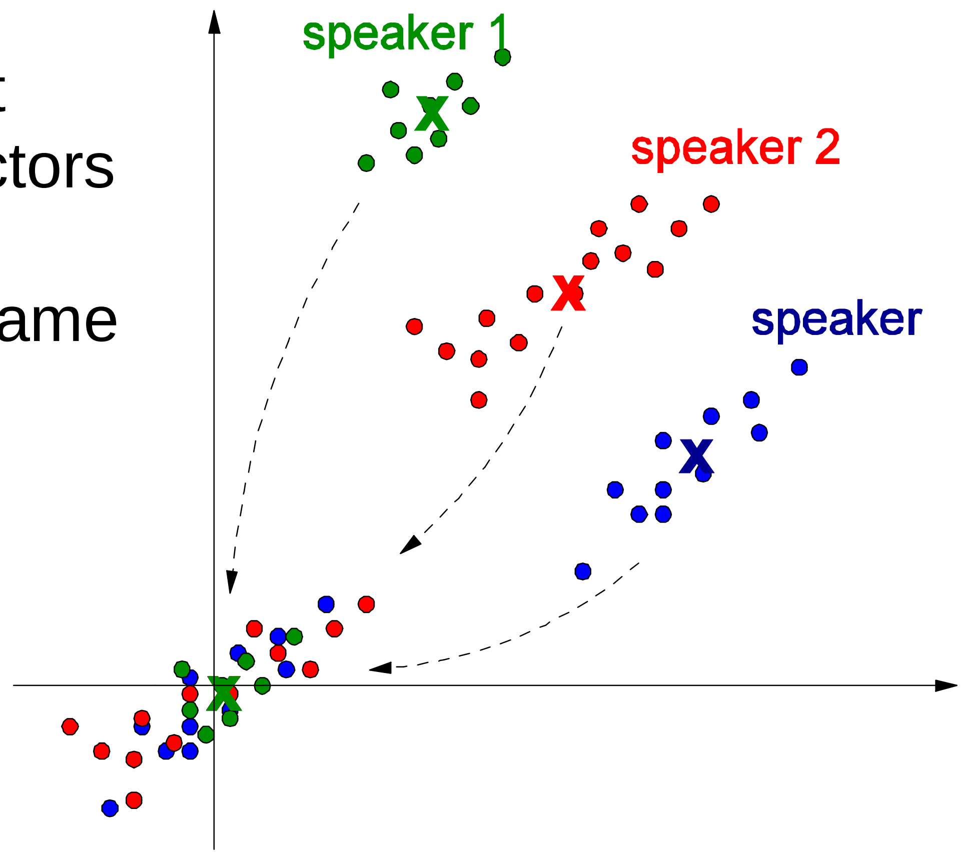
- We used the simplest version of eigen-channel adaptation [Brummer2004]
  - adaptation is applied only in test (speaker model is obtained using normal UBM MAP adaptation from enrolment data)
  - as the score, we use LLR computed using channel (MAP or ML) adapted speaker model and UBM model (or T-norm model)

Likelihood of data: $$\sum_t \log p(x_t \mid s)$$

- speaker model is defined by supervector s = concatenated mean vectors of UBM adapted to enrolment data normalized by standard deviations

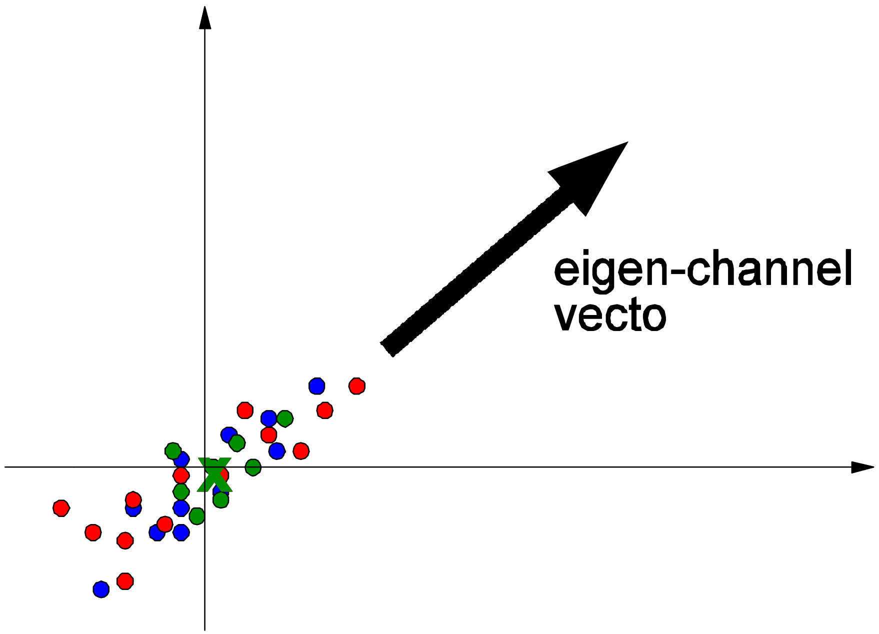
Eigen-channel adaptation II.

- We want to find the direction(s) of highest variability of supervectors obtained for different utterances from the same speaker – eigen-channel(s).



Eigen-channel adaptation III.

- The direction is obtained by PCA of average within-class covariance matrix, where classes are supervectors corresponding to the same speaker.

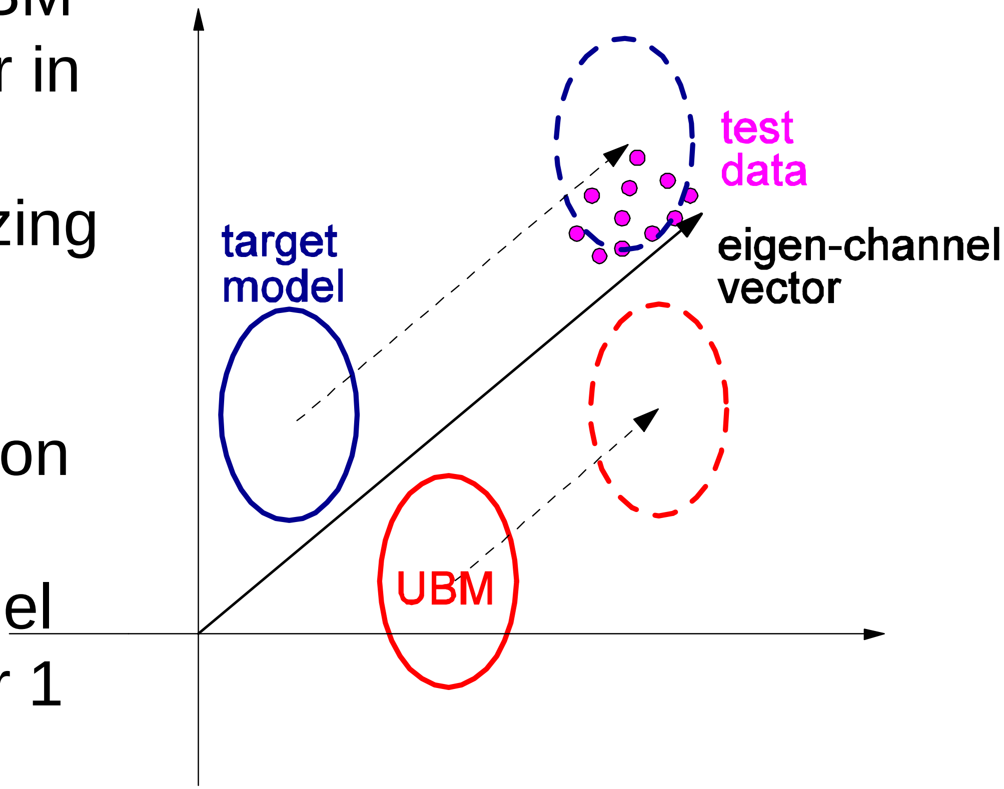


Eigen-channel adaptation IV.

- During the test, we adapt speaker model and UBM by moving supervector in the direction of eigen-channel(s) => Maximizing

$$\sum_t \log p(x_t \mid s + Vx)$$

- p(x) - models distribution of speaker variability along the eigen-channel direction; negligible for 1 conversation



Outline

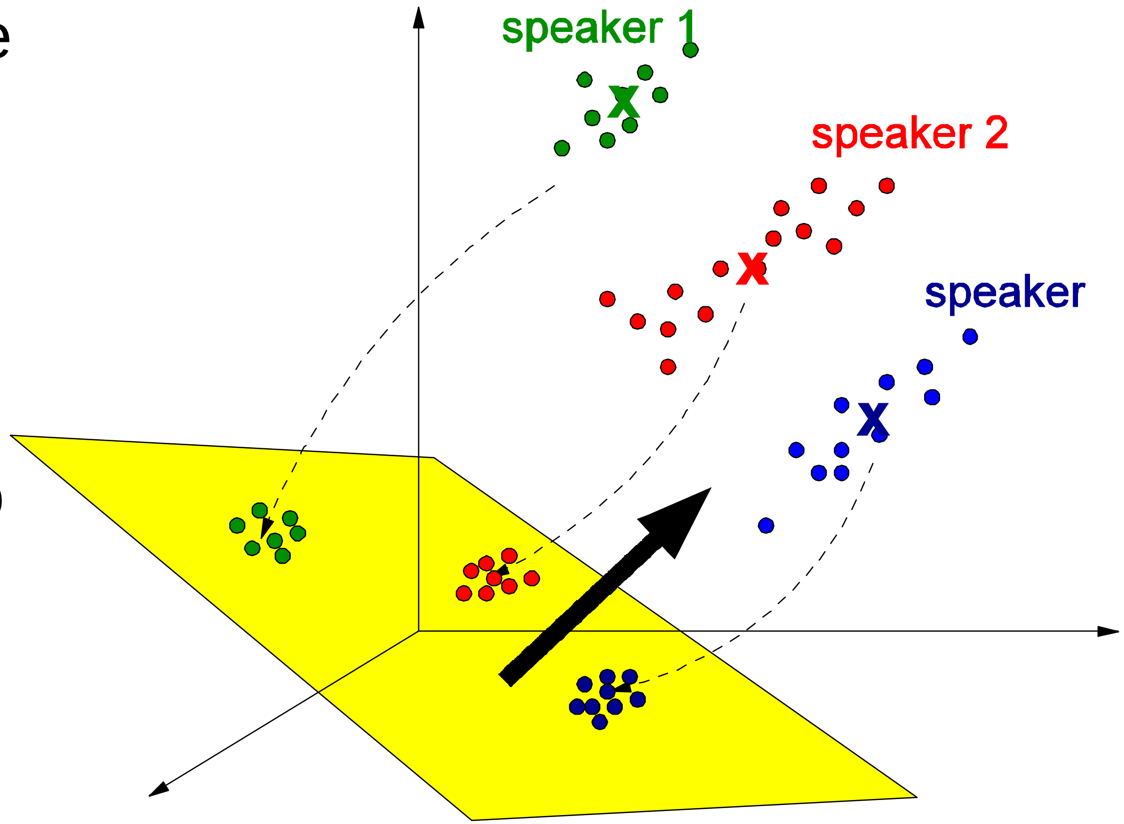
- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks

SVM systems

- Linear kernels
- Rank normalization
- LibSVM C++ library [Chang2001]
- Pre-computed Gram matrices
- Nuisance attribute projection (NAP) [Campbell2006]

NAP

- Nuisance attribute projection
- Removes the unwanted variability from features by projecting them to useful space.



SVM - GMM

- Feature extraction and UBM adaptation is the same as for GMM system
- Only 512 Gaussian components
- Supervector 512*39=19968
- NAP with 30 eigen-vectors derived on 310 speakers from NIST 2004
- Impostors: 230 speakers from NIST 2002 and 2606 speakers from Fisher
- T-norm: 230 speakers from NIST 2002 and 800 speakers from Fisher

SVM CMLLR/MLLR [Stolcke2005/6]

- LVCSR system is adapted to speaker (VTLN factor and (C)MLLR transformations are estimated) using ASR transcriptions provided by NIST
- AMI 2005(6) LVCSR system incorporates [Hain2005]:
  - 50k word dictionary (pronunciations of OOVs were generated by grapheme to phoneme conversion based on rules trained from data)
  - PLP, HLDA
  - CD-HMM with 7500 tied-states each modeled by 18 Gaussians
  - Discriminatively trained using MPE
  - Adapted to speaker: VTLN, SAT based on CMLLR, MLLR

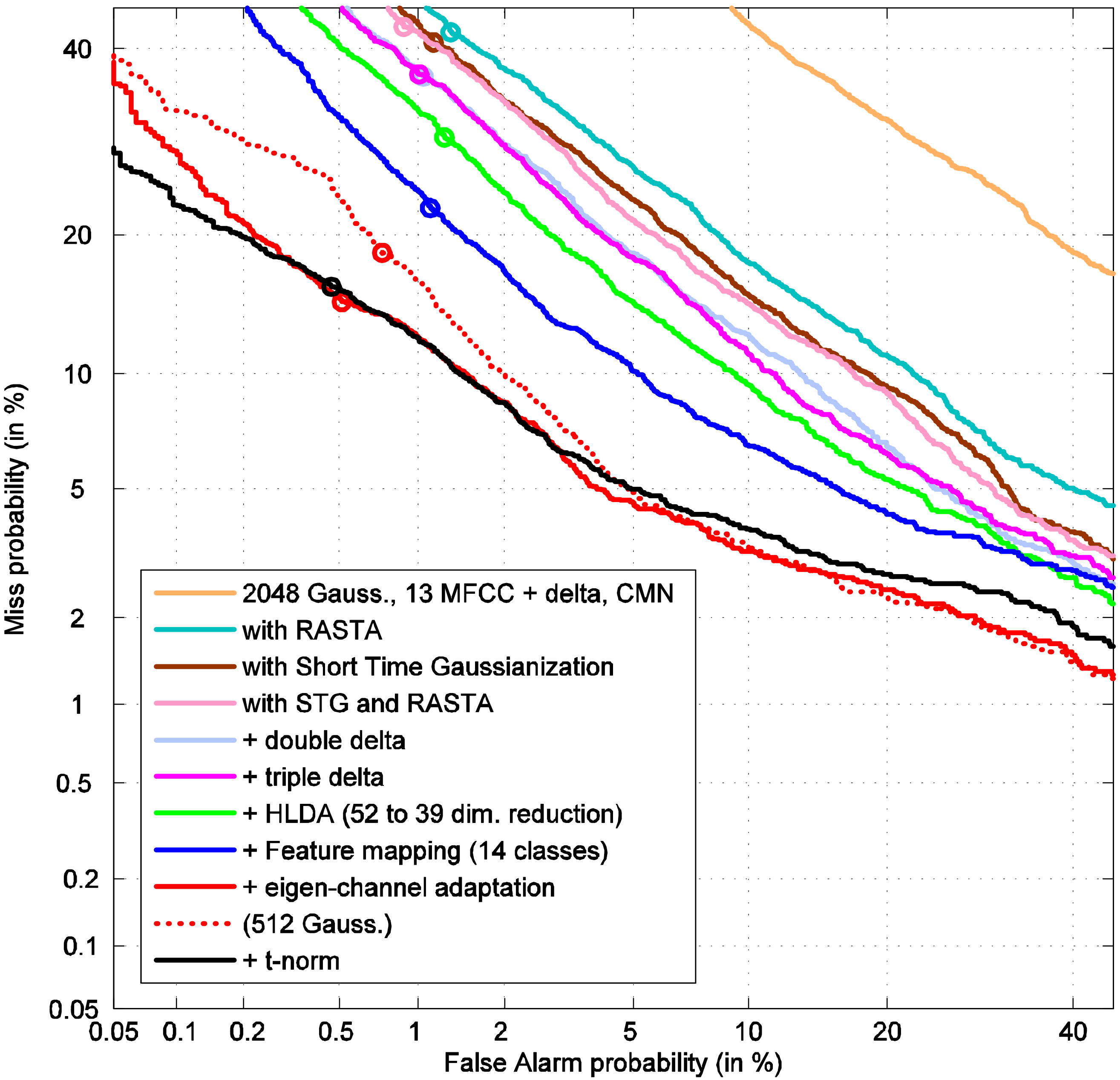
SVM - CMLLR/MLLR

- Cascade of CMLLR and MLLR
  - CMLLR: 2 classes – silence and speech
  - MLLR: 3 classes – silence and 2 speech classes derived from data
- Silence class discarded for SRE
- Supervector = 1 CMLLR + 2 MLLR =

  $= 3*3*13^2+3*39=1638$
- NAP with 20 eigen-vectors derived on NIST 2004
- Impostors: 310 speakers from NIST 2004
- T-norm: 310 speakers from NIST 2004

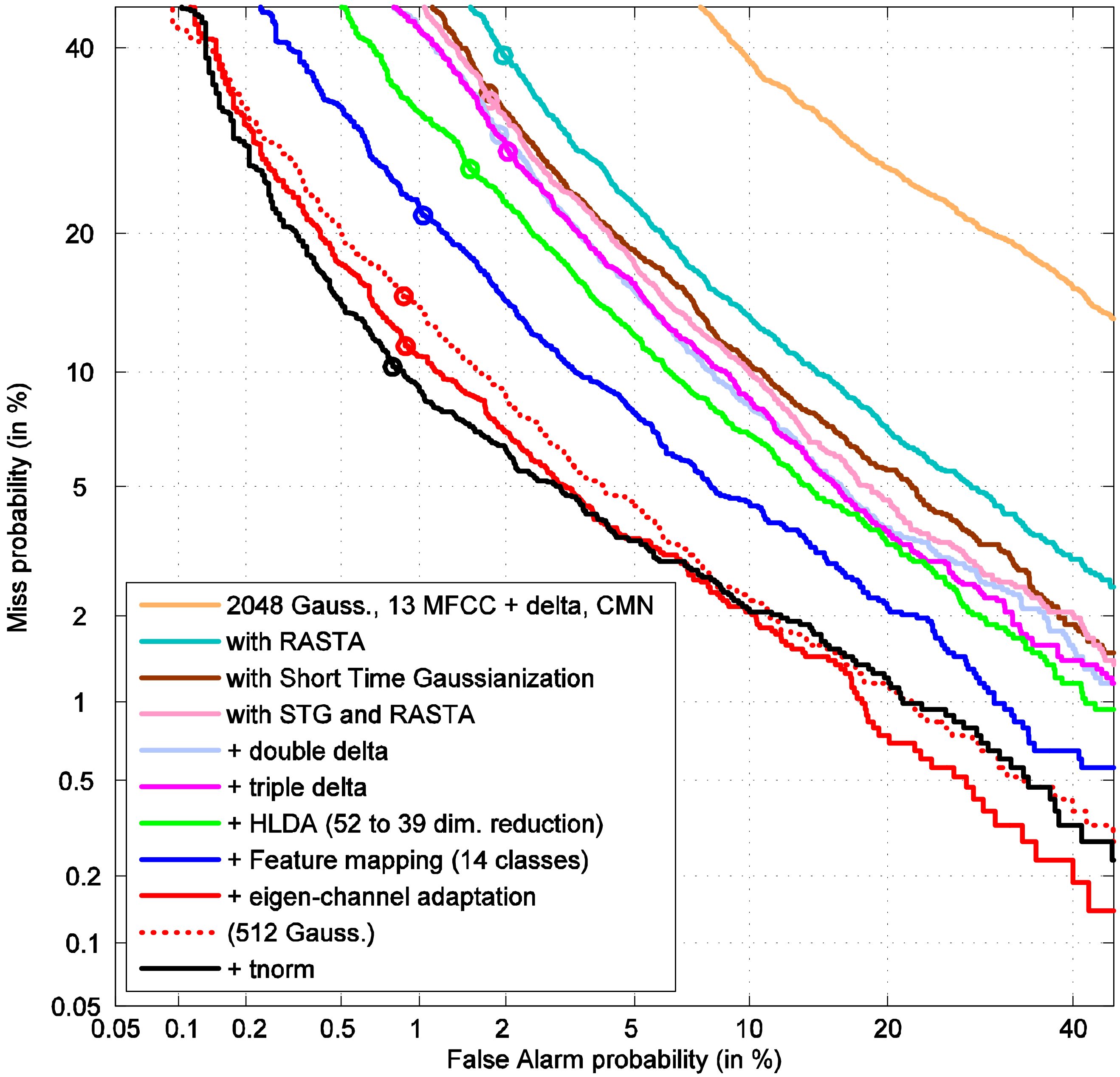
Outline

- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks



NIST 2005

all trials

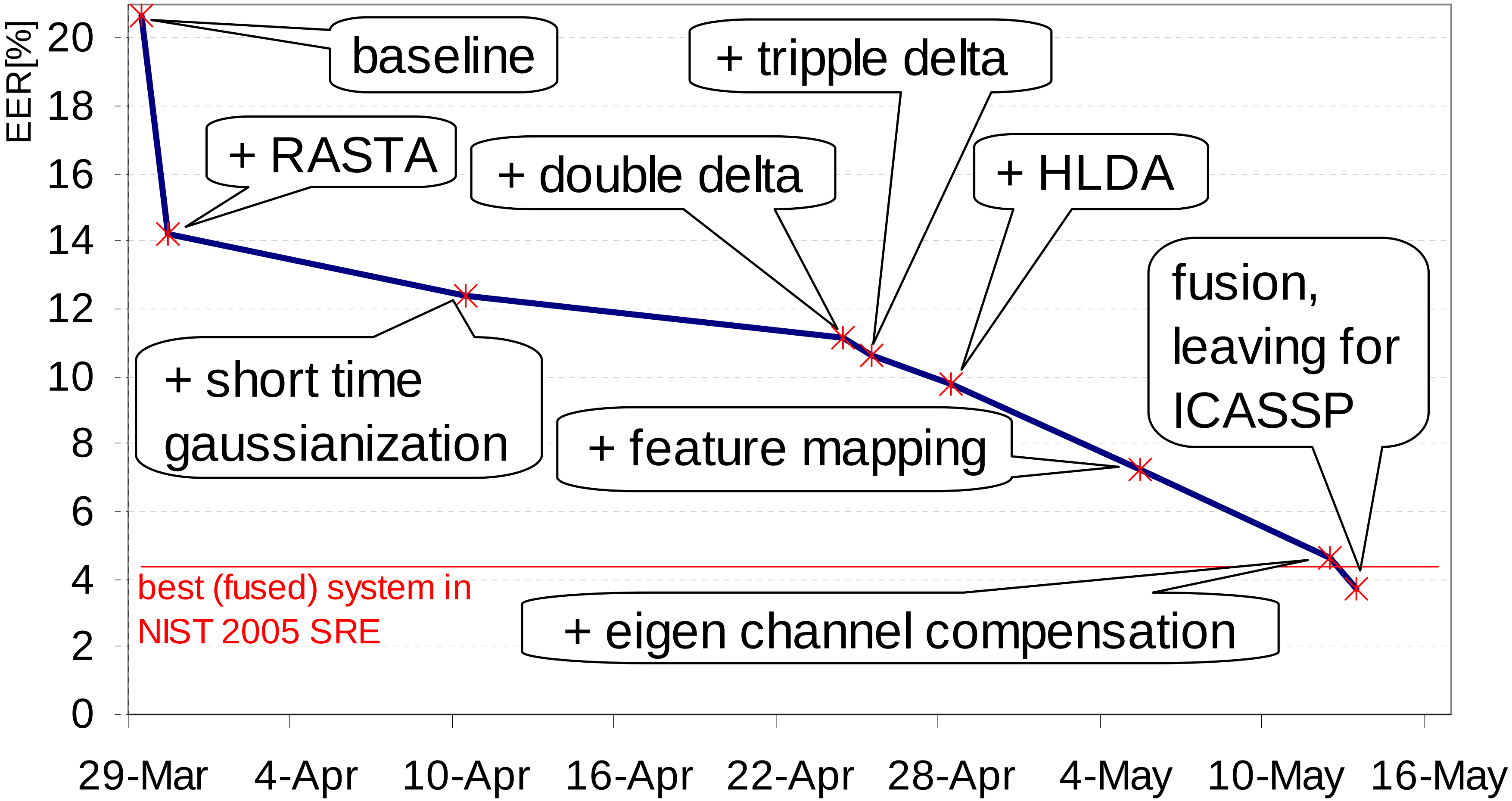


NIST 2006 English only trials

GMM System Analysis in numbers

system	2005 all trials		2006 all trials		2006 Engslish only	
	EER [%]	DCF	EER [%]	DCF	EER [%]	DCF
Baseline GMM – MFCC + C0, zero mean normalization, deltas, 2048 Gaussian	26,6	0,089	24,1	0,089	23,8	0,088
+ RASTA channel compensation	14,3	0,055	12,9	0,063	11,8	0,059
+ short-time Gaussianization (3 sec window)	12,4	0,052	10,9	0,054	10,0	0,051
+ acceleration coefficients	11,2	0,047	10,1	0,053	9,1	0,049
+ tripple deltas (bad for 2006)	10,6	0,047	10,3	0,053	9,3	0,048
+ HLDA 52->39 dimensions	9,7	0,042	9,5	0,047	8,2	0,041
+ Feature Mapping (7channel 2gender)	7,3	0,033	7,8	0,040	6,2	0,032
+ eigen-channel adaptation (30 dimensions)	**4,6**	0,020	5,4	0,028	4,0	0,020

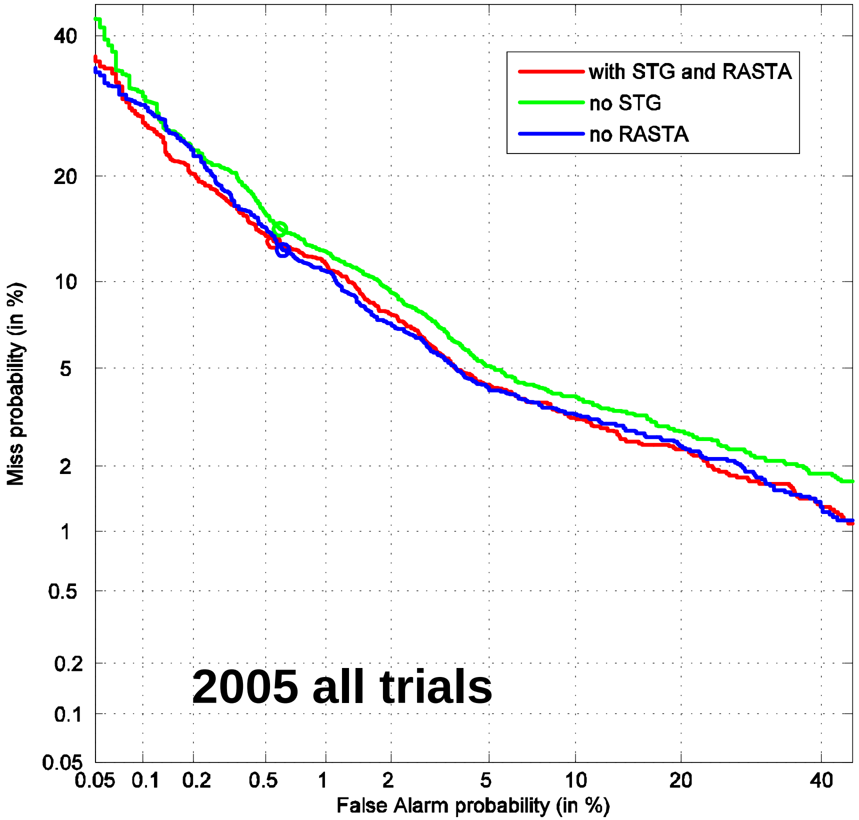
Things to improve GMM

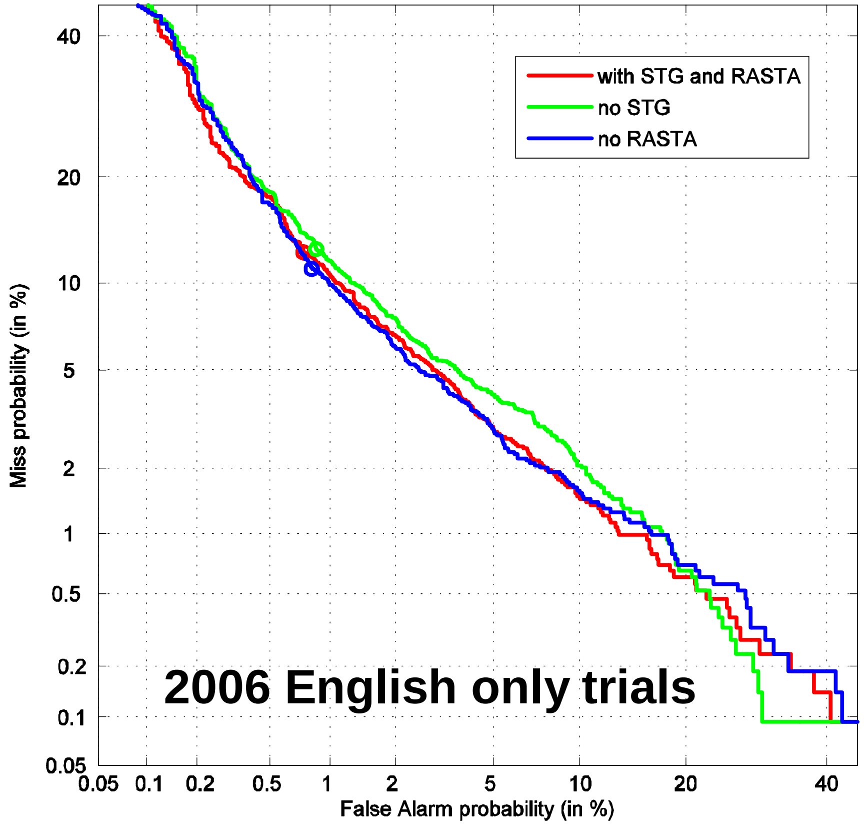


Outline

- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks

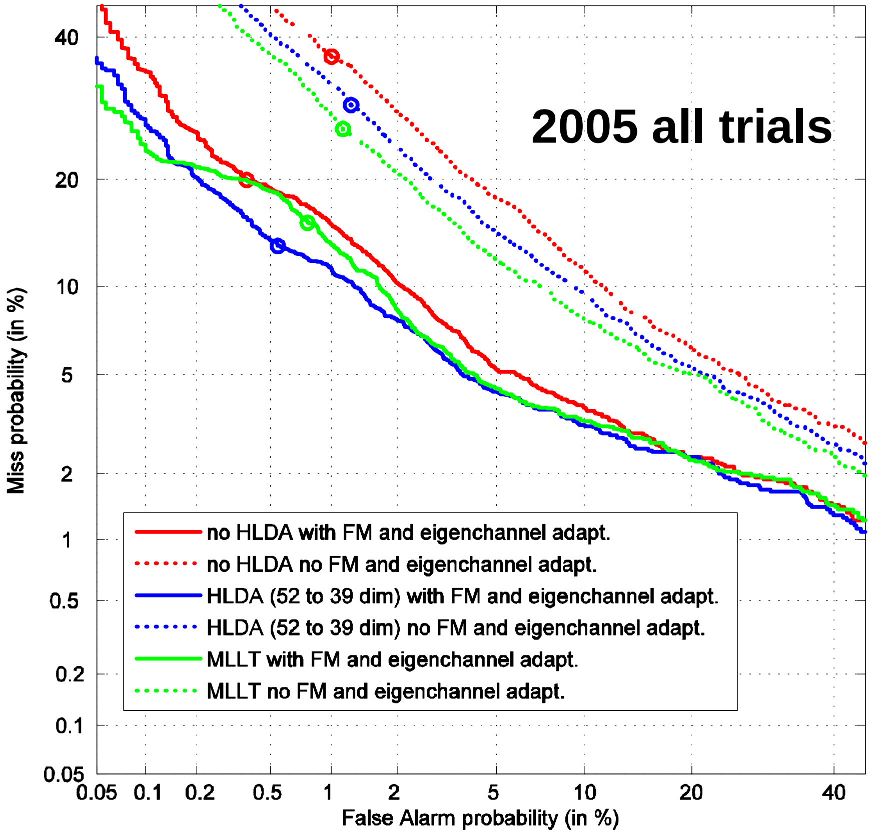
Importance of RASTA and STG

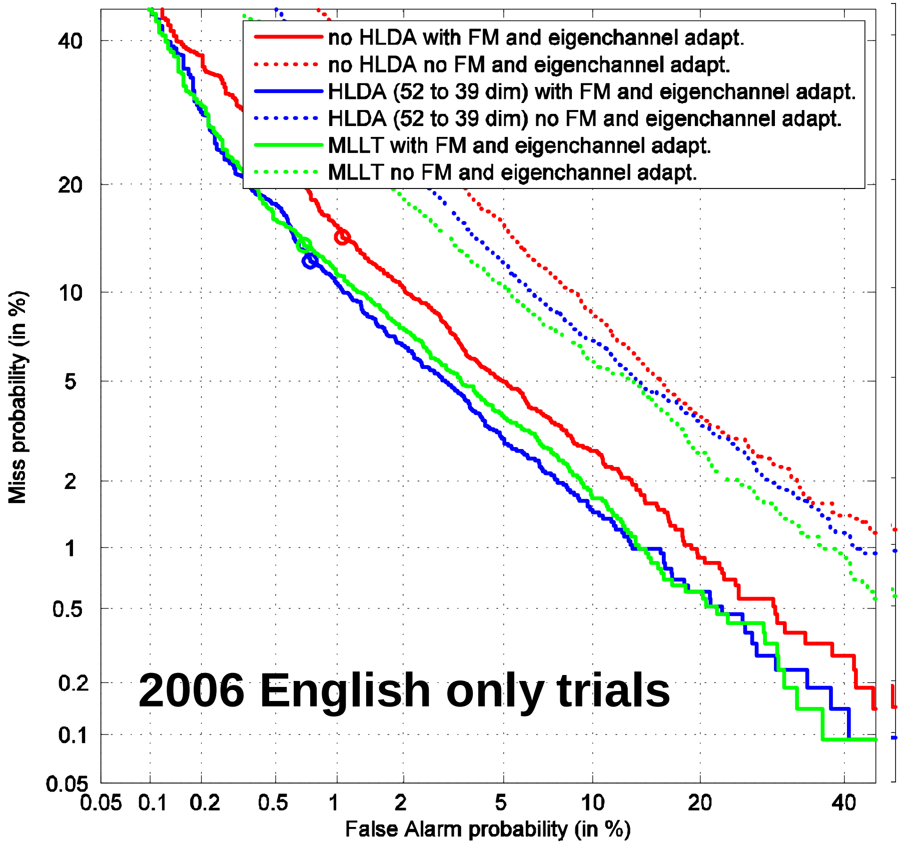




=> RASTA does not help in the final system

Is HLDA worthy to implement?

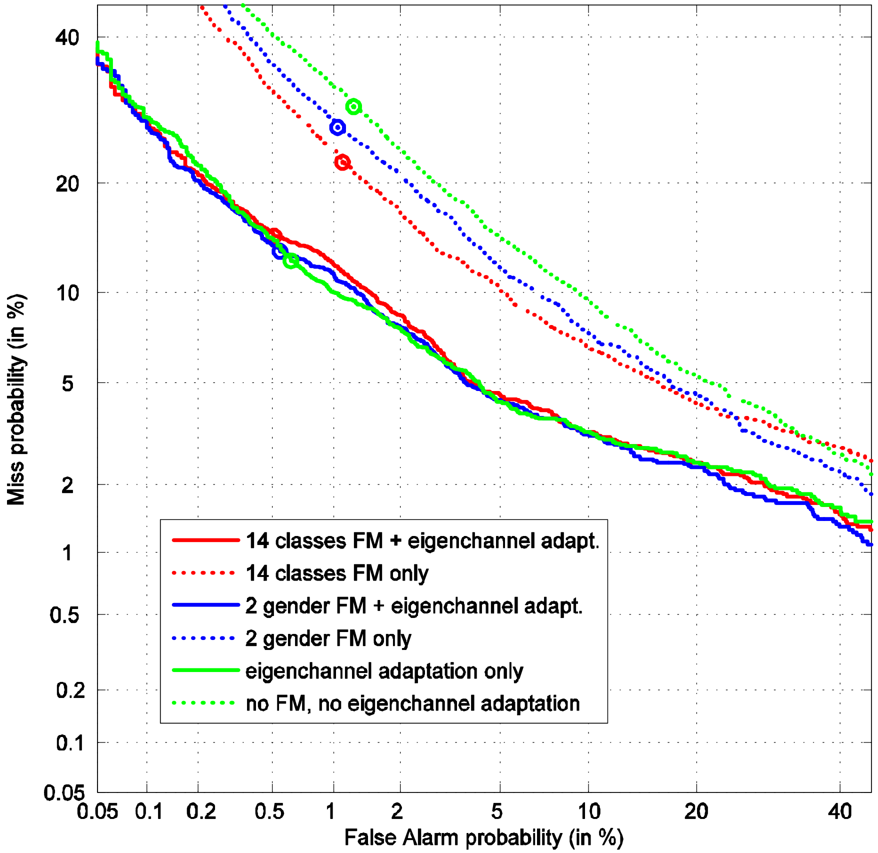




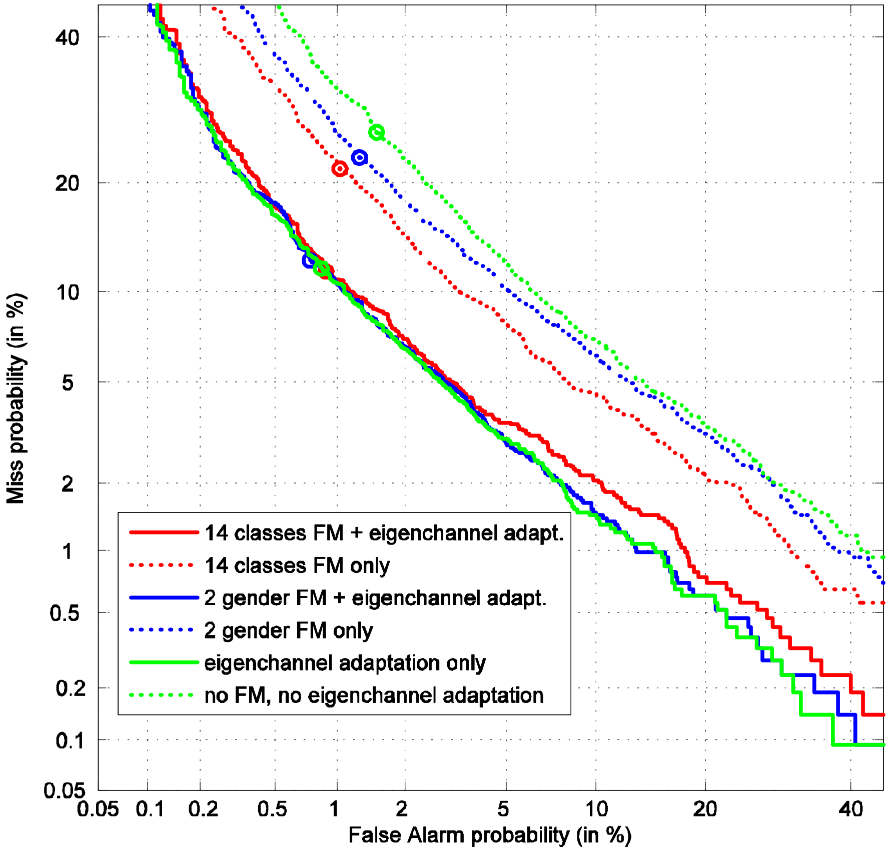
=> Dimensionality reduction is probably advantageous for correct estimation of eigen-channels

Eigen-channel adaptation vs. Feature mapping

2005 all trials

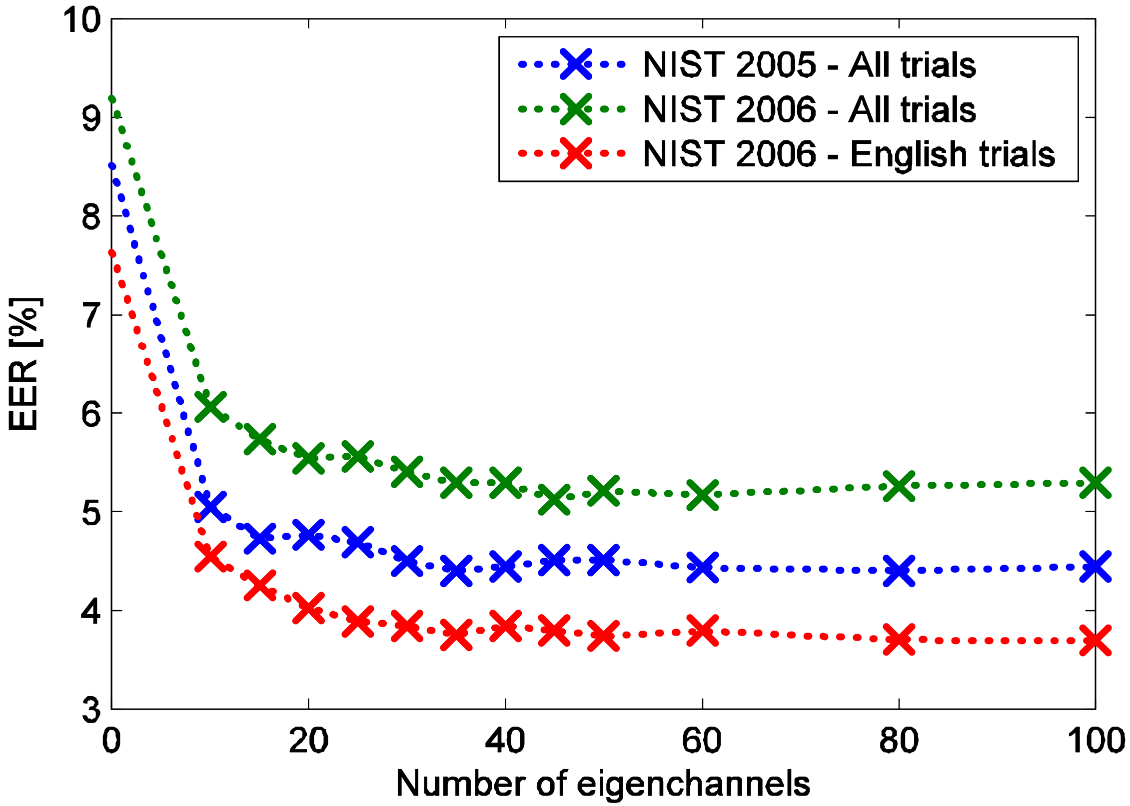


2006 English only trials



=> Feature mapping is not important when applied together with eigen-channel adaptation

How many eigen-channels to use?



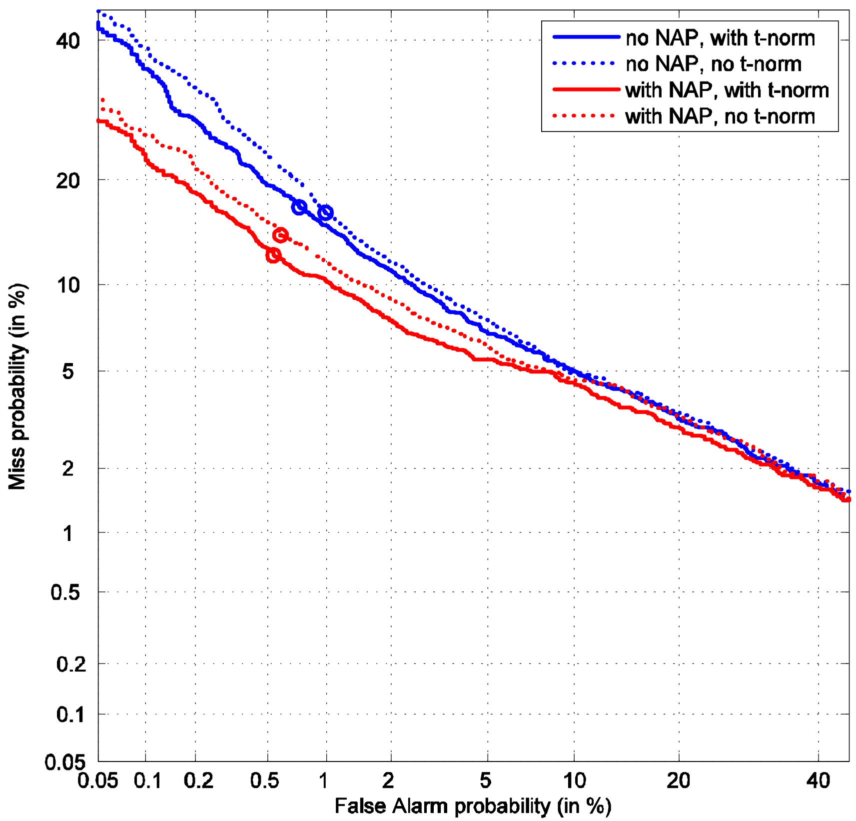
=> Channel adaptation is not very sensitive to the number of eigen-channels used

Outline

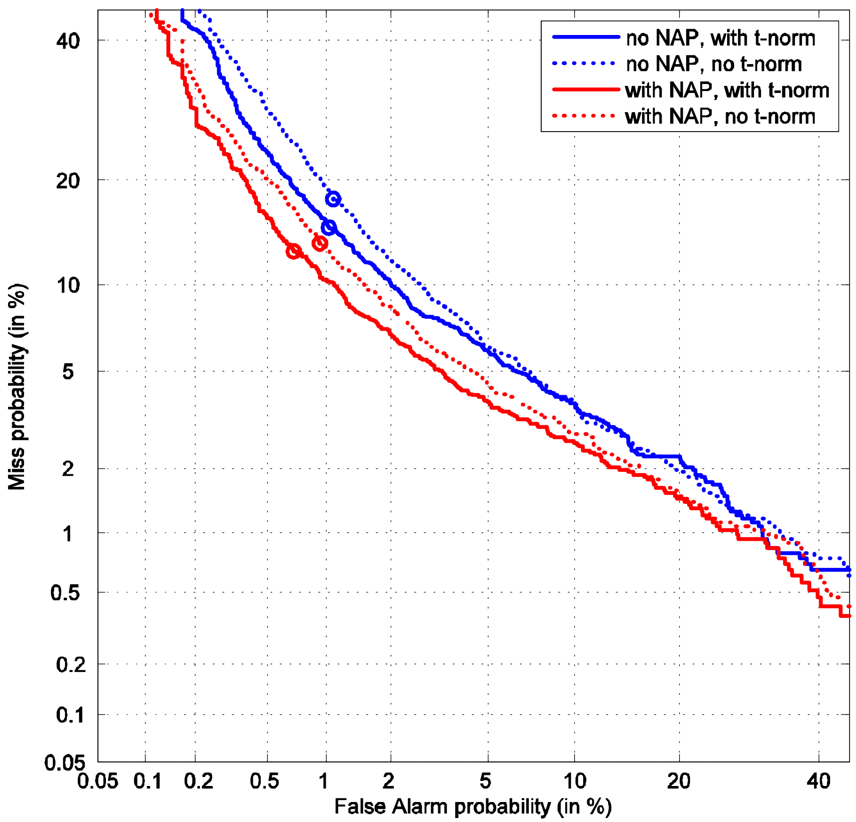
- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks

SVM-GMM system analysis

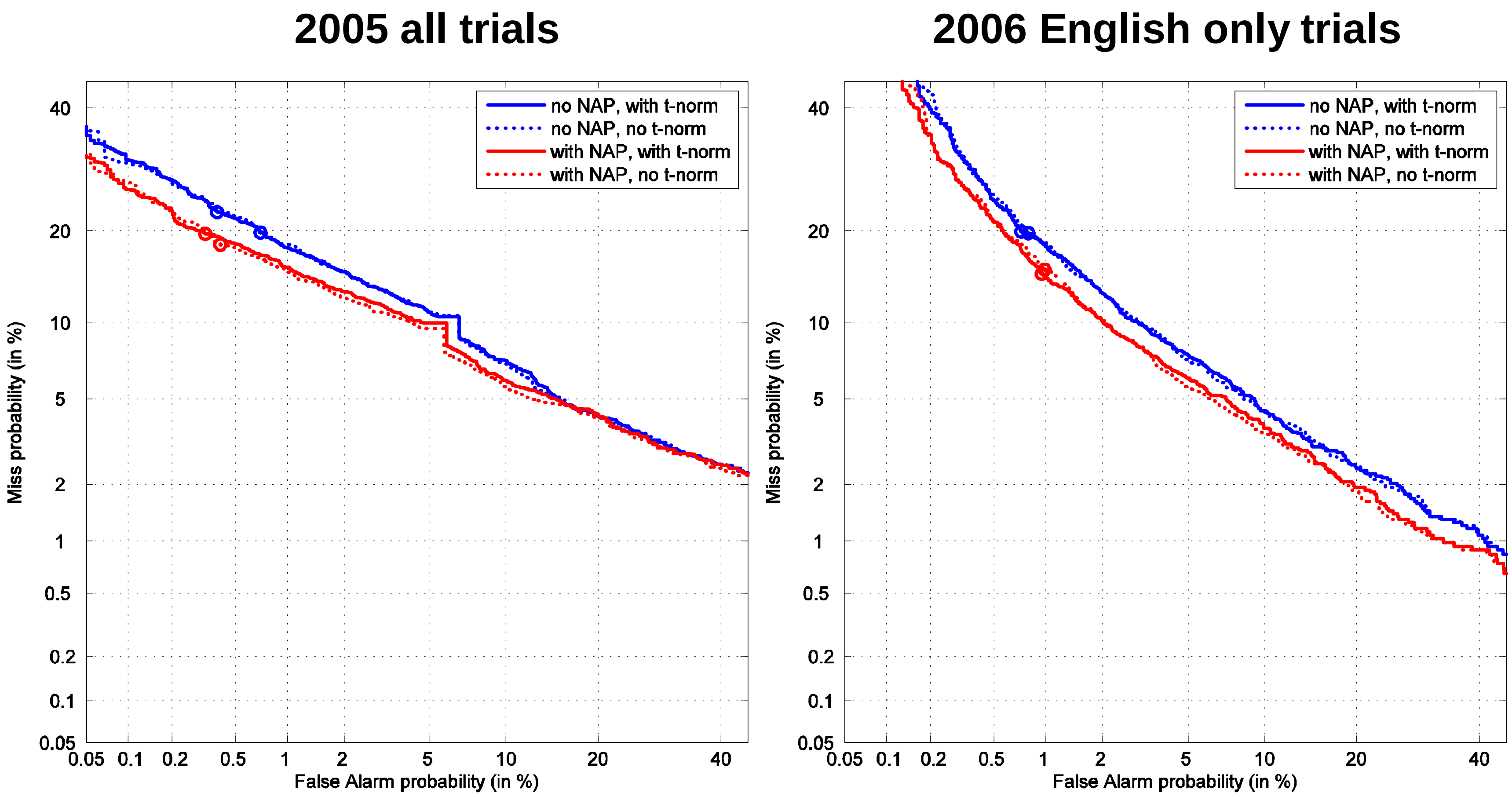
2005 all trials



2006 English only trials



SVM-MLLR system analysis

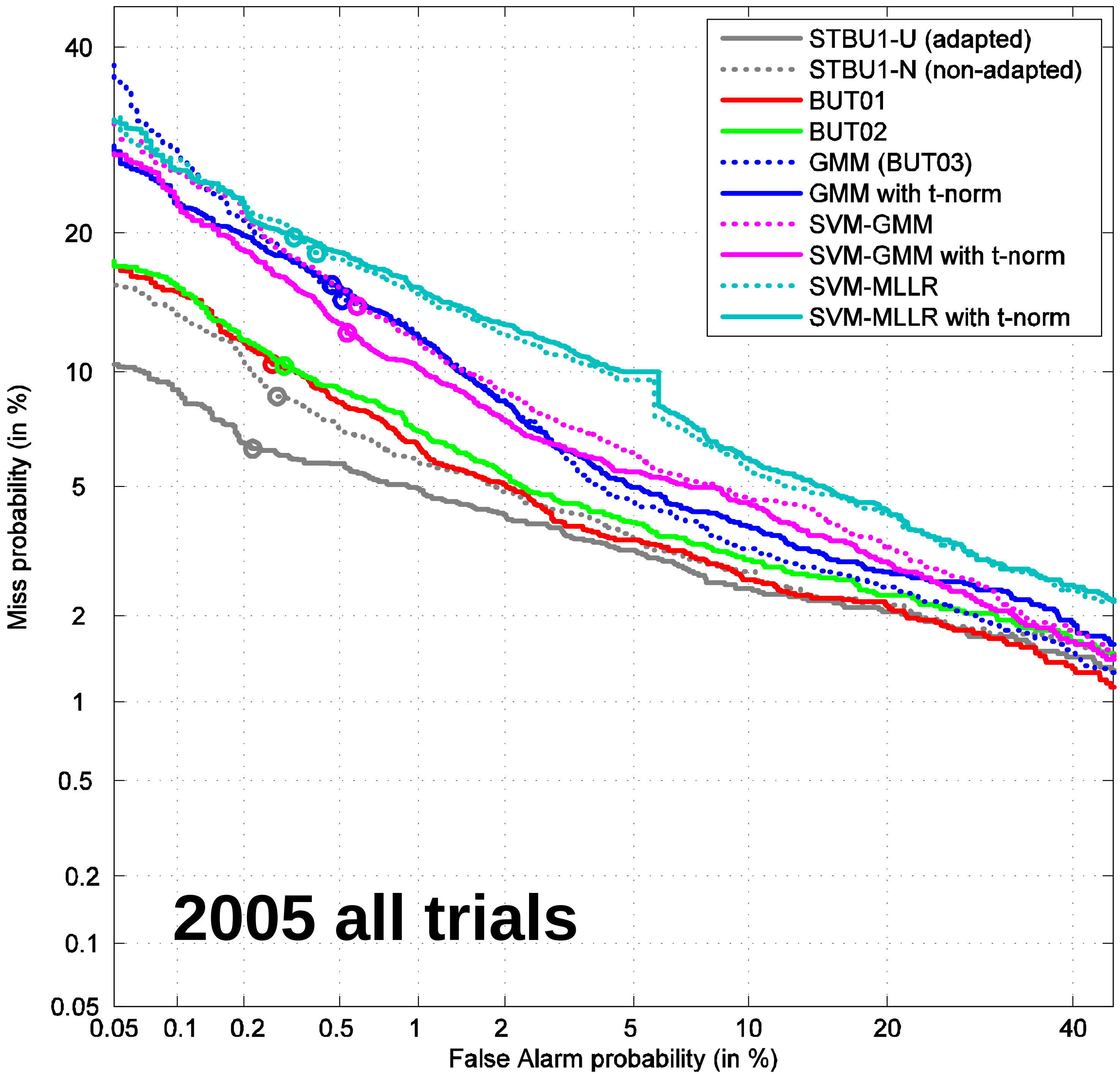


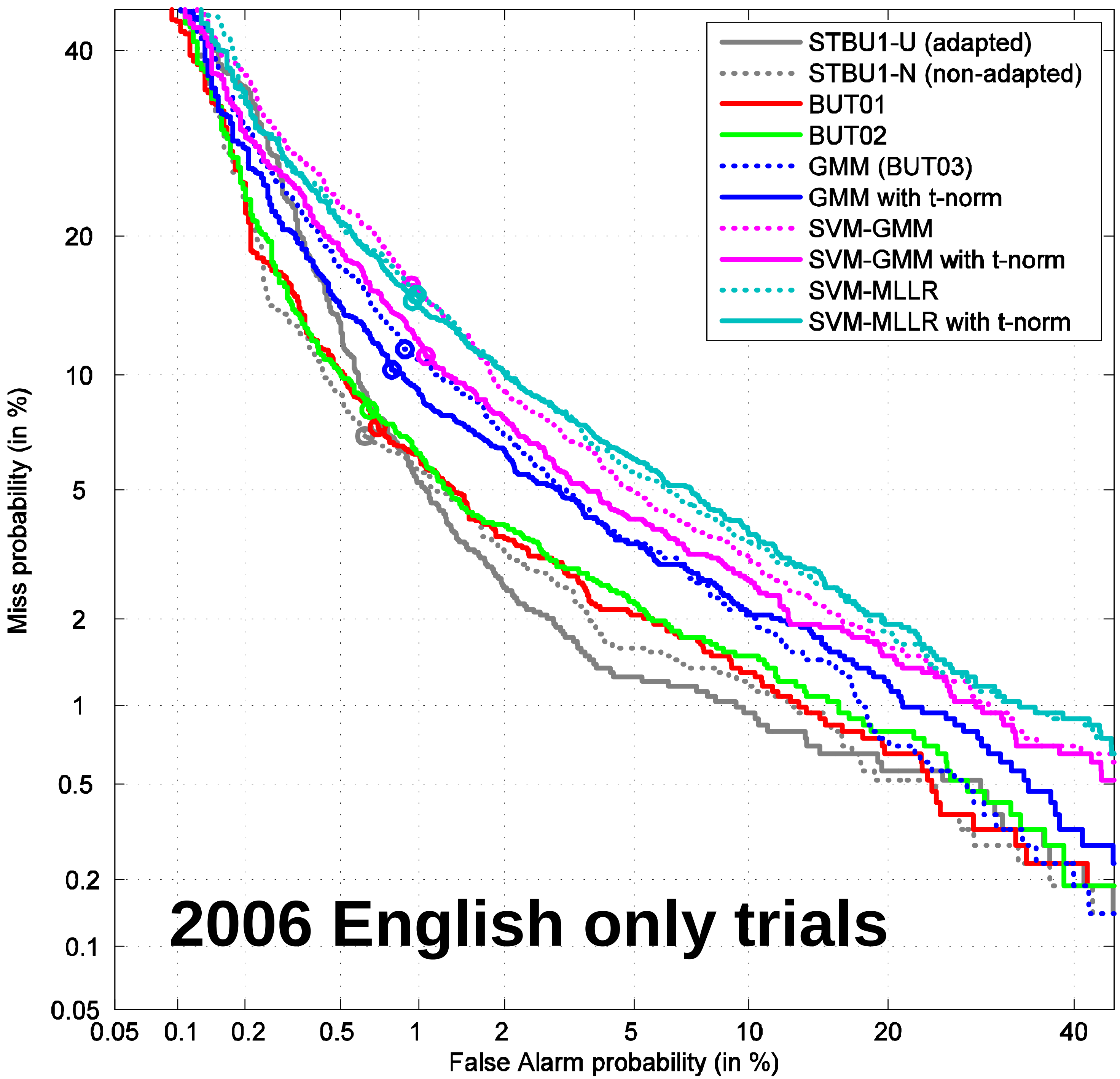
Outline

- Submitted systems
- Description of individual systems
  - GMM
  - SVM-GMM
  - SVM-MLLR
- System analysis
  - Building the GMM system
  - Importance of individual components in the final GMM system
  - Importance of NAP in SVM systems
- Fusion
- Conclusions and thanks

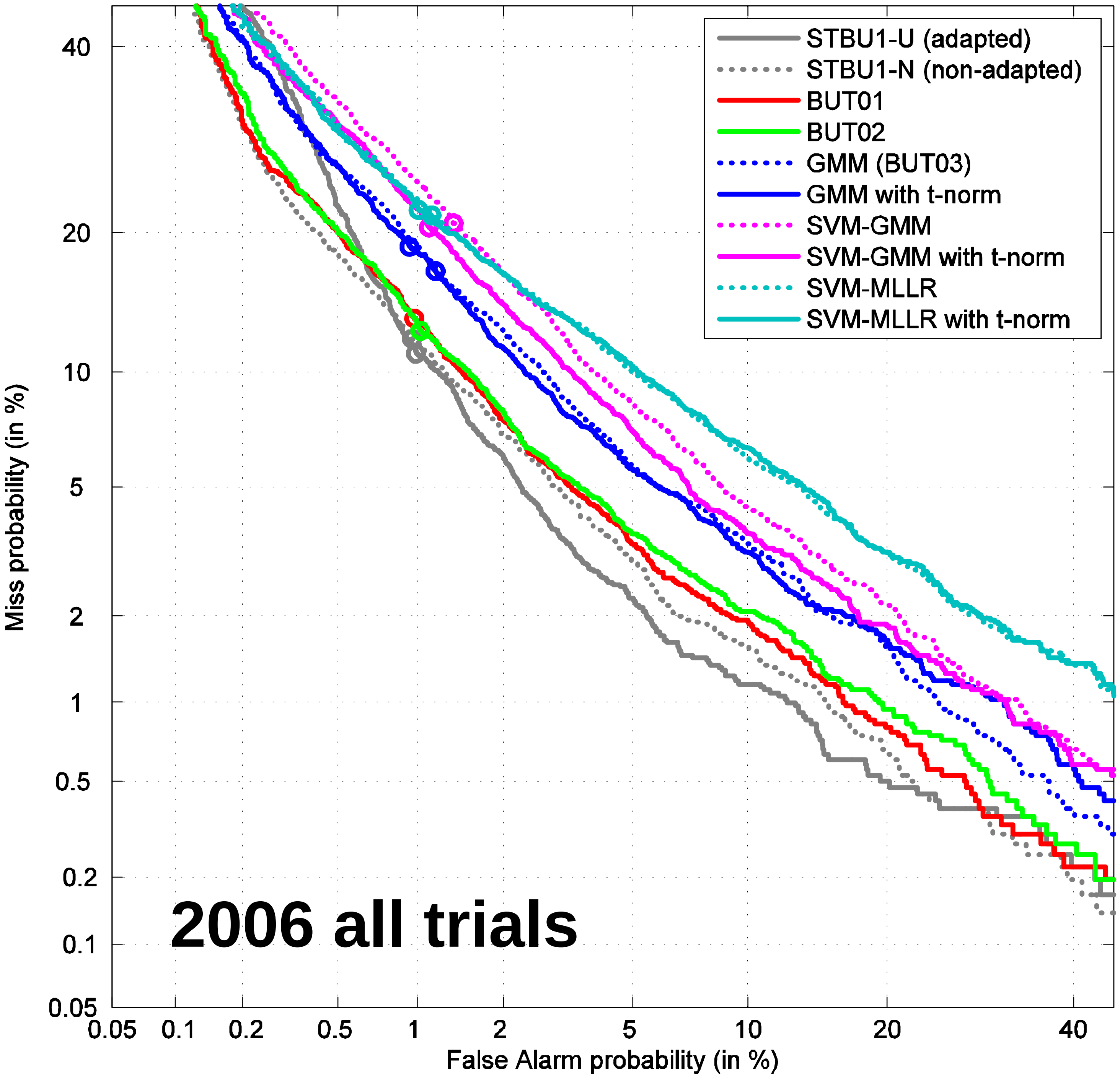
Fusion

- Linear logistic regression used to fuse:
  - all 6 systems with and without t-norm - **BUT01**
  - 3 T-normed systems - **BUT02**
- Niko's FoCal toolkit was used for this purpose [BrummerFoCal]


STBU1-U (adapted)
STBU1-N (non-adapted)
BUT01
BUT02
GMM (BUT03)
GMM with t-norm
SVM-GMM
SVM-GMM with t-norm
SVM-MLLR
SVM-MLLR with t-norm
2005 all trials
Miss probability (in %)
False Alarm probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.05
0.05
0.1
0.2
0.5
1
2
5
10
20
40



STBU1-U (adapted)
STBU1-N (non-adapted)
BUT01
BUT02
GMM (BUT03)
GMM with t-norm
SVM-GMM
SVM-GMM with t-norm
SVM-MLLR
SVM-MLLR with t-norm
Miss probability (in %)
False Alarm probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.05
0.05
0.1
0.2
0.5
1
2
5
10
20
40
2006 English only trials



STBU1-U (adapted)
STBU1-N (non-adapted)
BUT01
BUT02
GMM (BUT03)
GMM with t-norm
SVM-GMM
SVM-GMM with t-norm
SVM-MLLR
SVM-MLLR with t-norm
2006 all trials
Miss probability (in %)
False Alarm probability (in %)
0.05
0.1
0.2
0.5
1
2
5
10
20
40

Summary of results

system	2005 all trials		2006 all trials		2006 English only	
	EER [%]	DCF	EER [%]	DCF	EER [%]	DCF
GMM	4,62	0,0196	5,40	0,0283	4,02	0,0203
GMM with t-norm	4,98	0,0203	5,35	0,0280	4,03	0,0182
SVM-GMM	5,42	0,0176	6,04	0,0314	4,40	0,0314
SVM-MLLR	7,05	0,0222	7,58	0,0327	5,42	0,0327
Fusion	3,71	0,0131	4,15	0,0229	3,04	0,0143

Conclusions

- *We considered NIST 2006 evals as a good occasion to build BUT's "baseline"…*
- Looks like we have a good one ;-)

Thanks

- Thanks **a lot: NIKO, DAVID** and **ALBERT** for great cooperation, many advices, support and enormous help.

- Everything we have in our system was already published by others. Thanks all the authors.

References

[Mason2005] M. Mason et al: Data-Driven Clustering for Blind Feature Mapping in SpkID, Eurospeech 2005.

[Chang2001] C. Chang et al.: LIBSVM: a library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm

[Hain2005] T. Hain et al.: The 2005 AMI system for RTS, Meeting Recognition Evaluation Workshop, Edinburgh, July 2005.

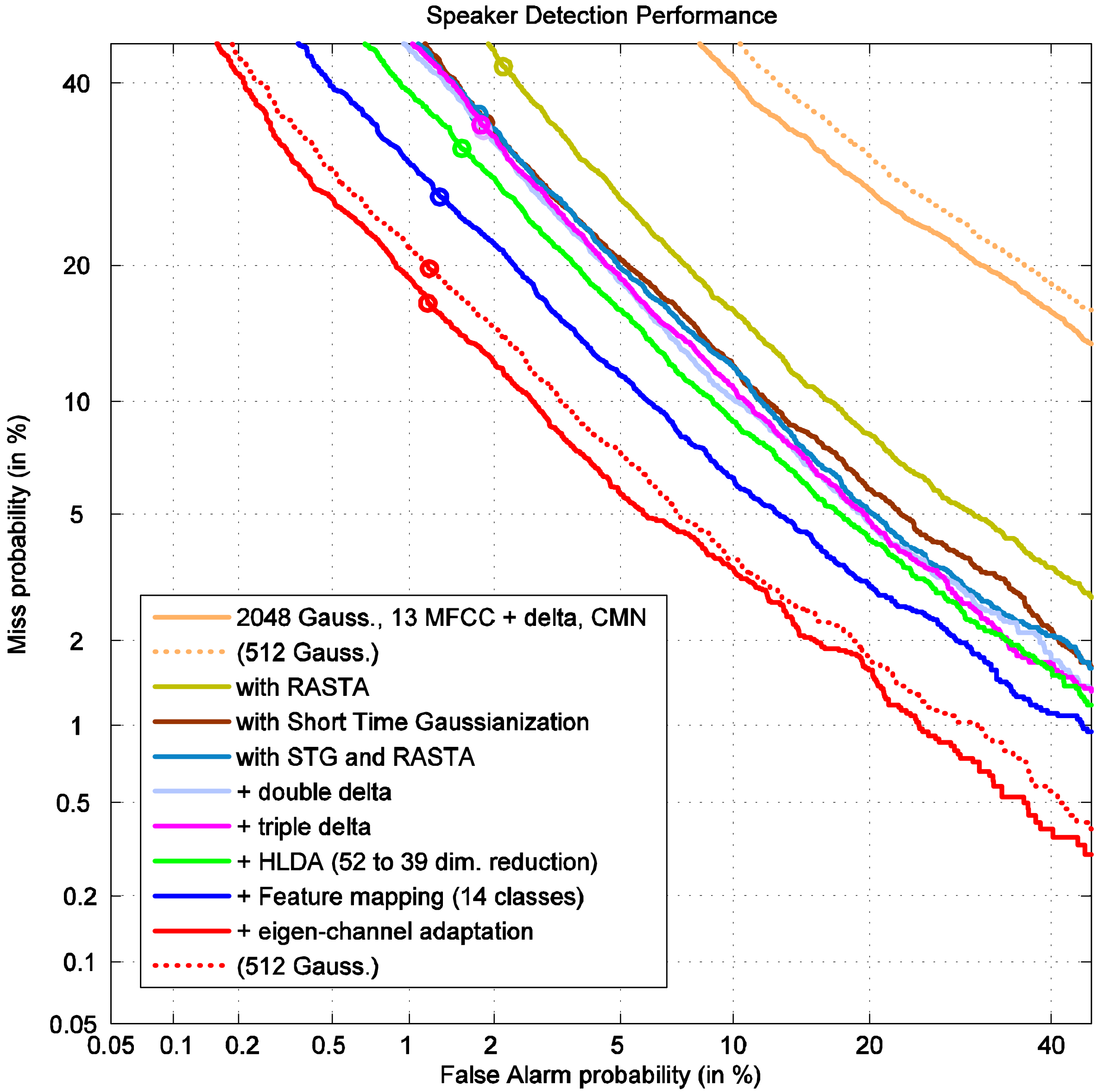
[Stolcke2005/6] A. Stolcke: MLLR Transforms as Features in SpkID, Eurospeech 2005, Odyssey 2006

[Brummer2004] N. Brummer: SDV NIST SRE'04 System description, 2004.

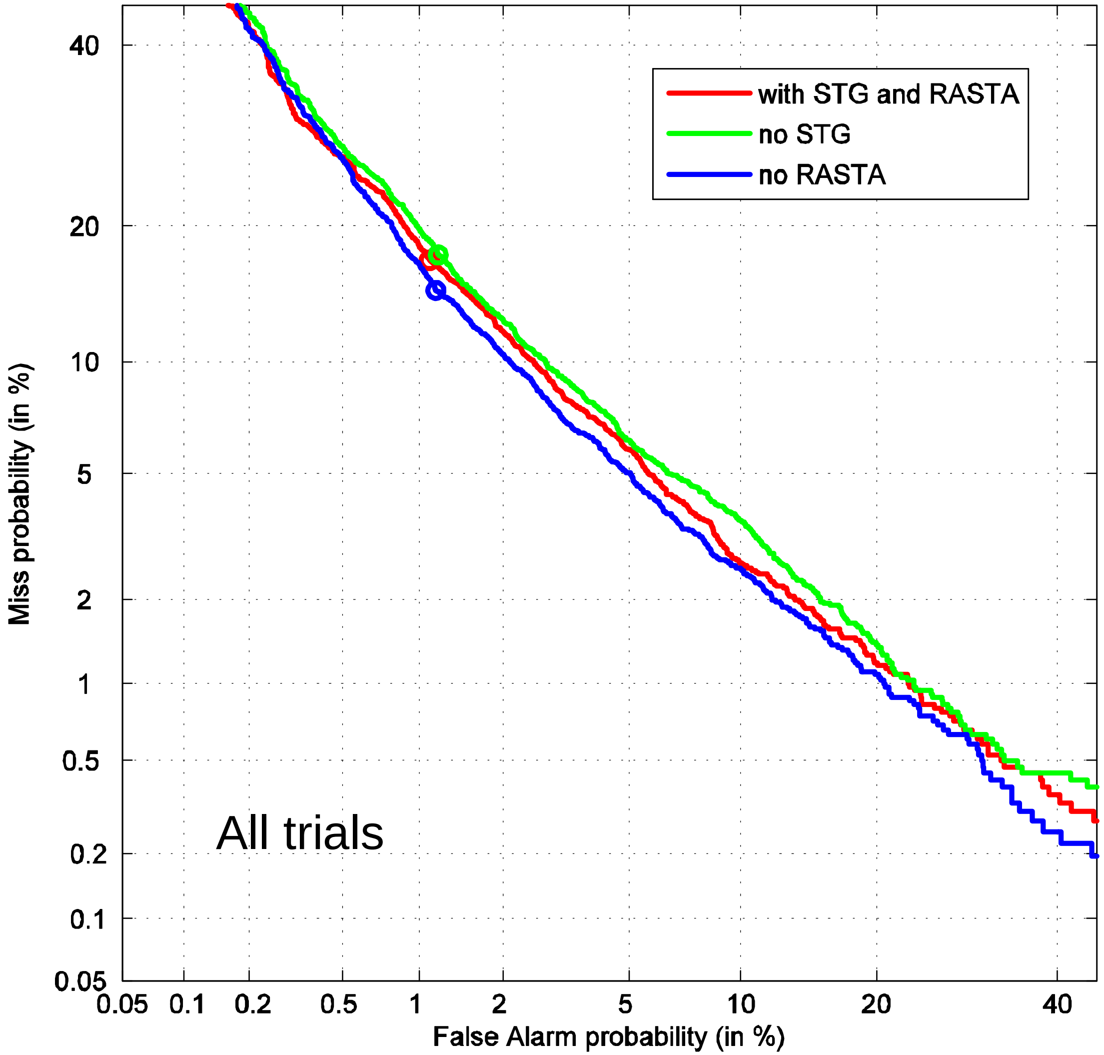
[Brummer_FoCal] N. Brummer: FoCal: Toolkit for fusion and Calibration, www.dsp.sun.ac.za/~nbrummer/focal

[Campbell2006] W. M. Campbell et al., "SVM Based Speaker Verification Using a GMM Supervector and NAP Variability Compensation," ICASSP 2006.

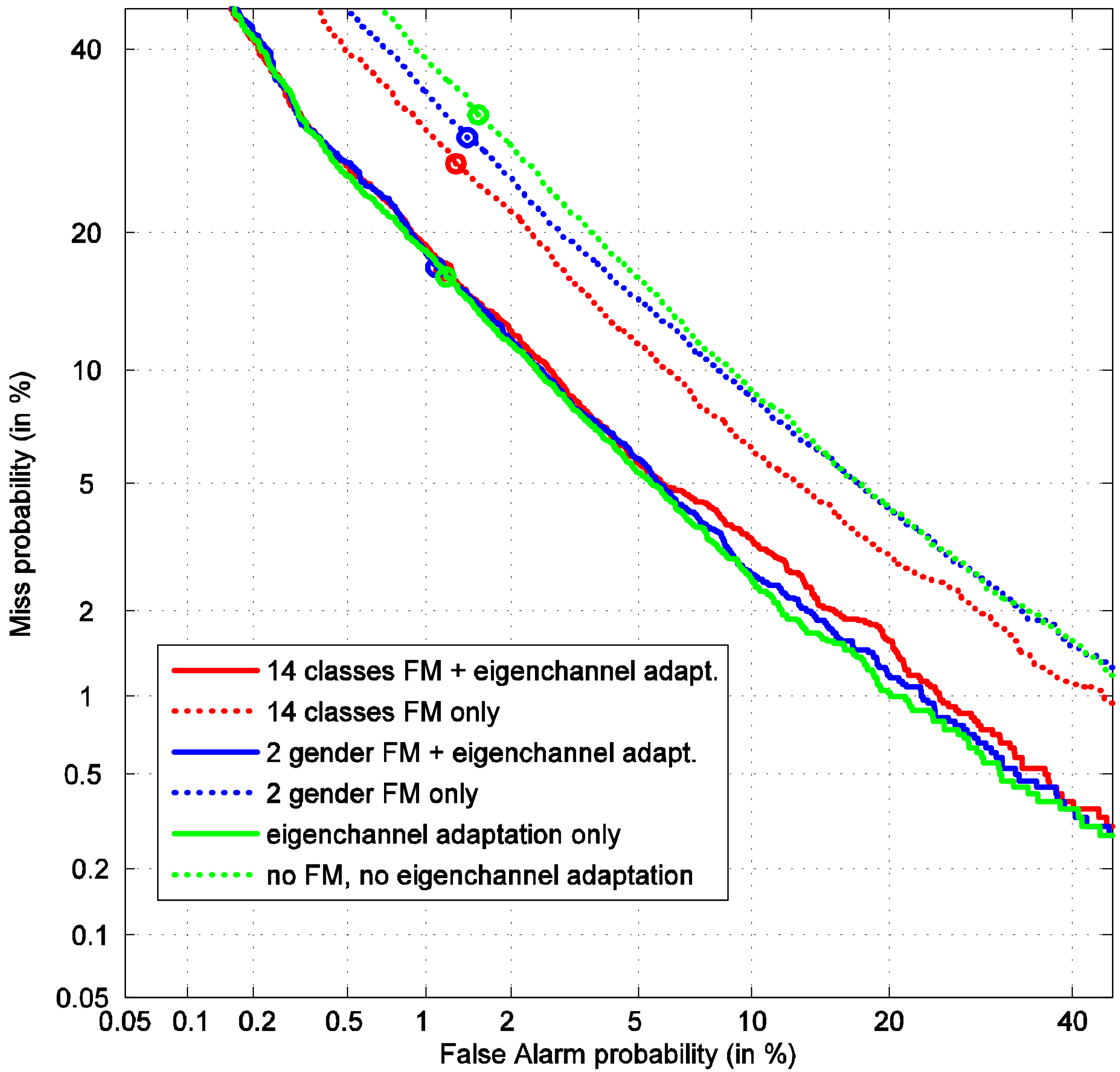
System Analysis 2006 all trials (det1)



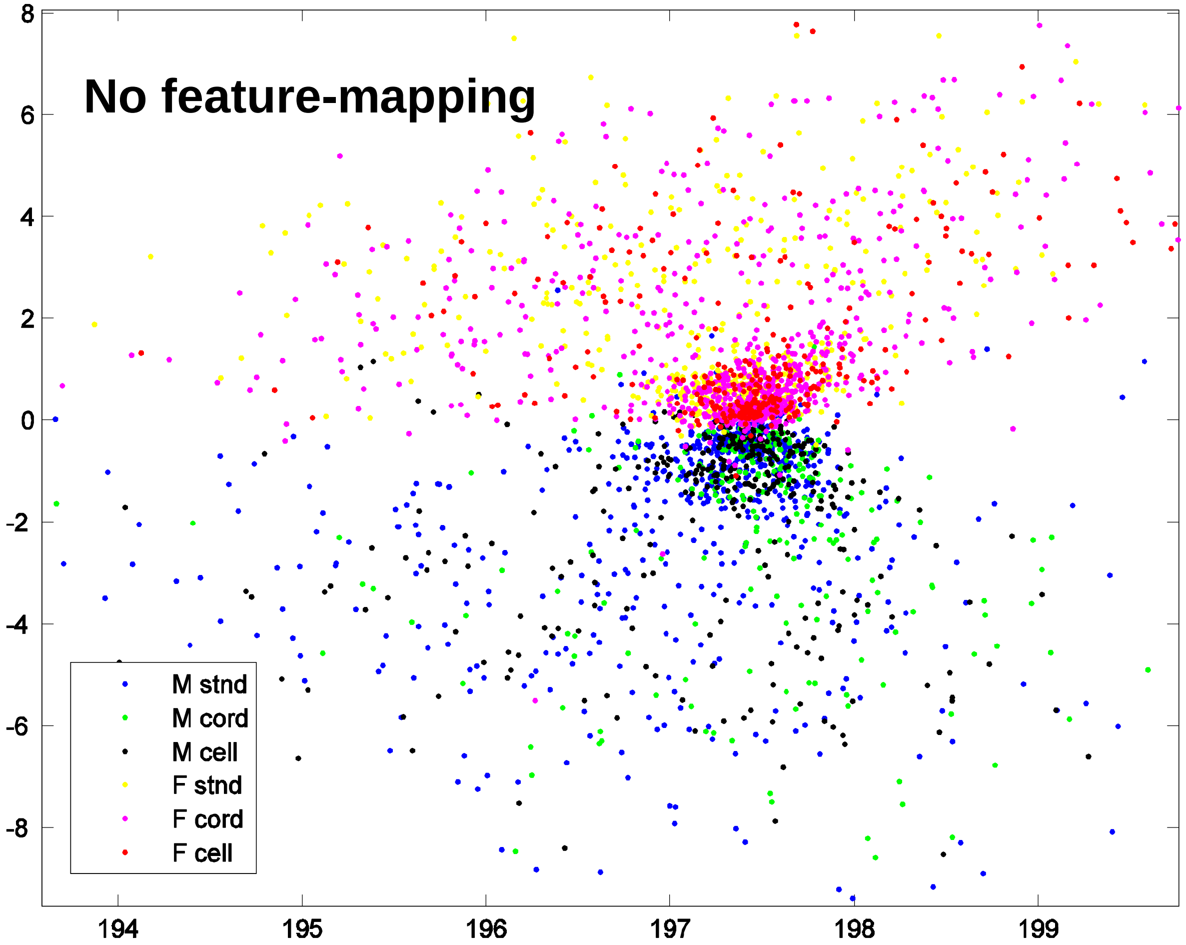
Importance of RASTA and STG -2006 –all trials



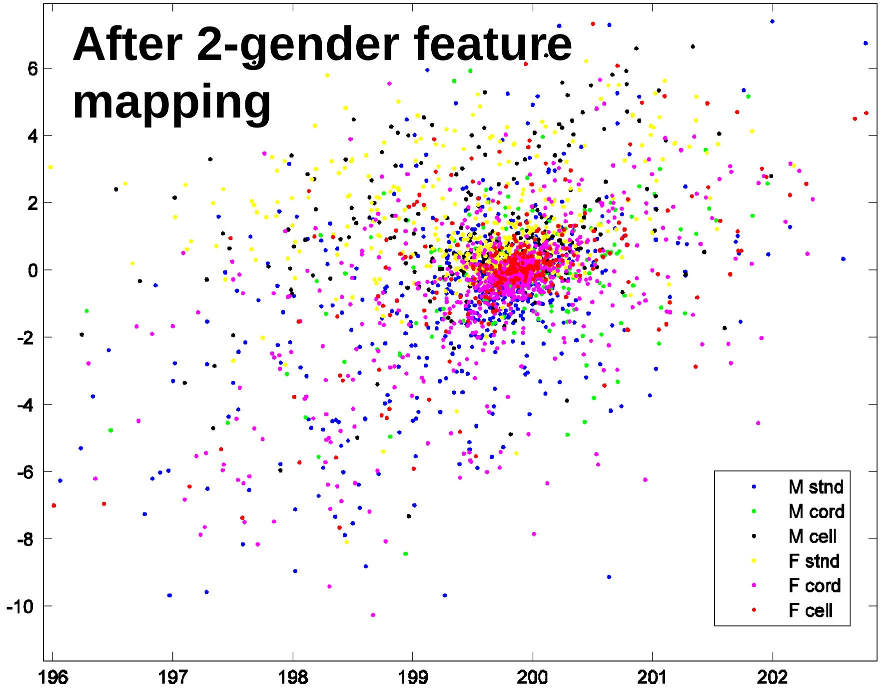
Feature mapping – 2006 all trials

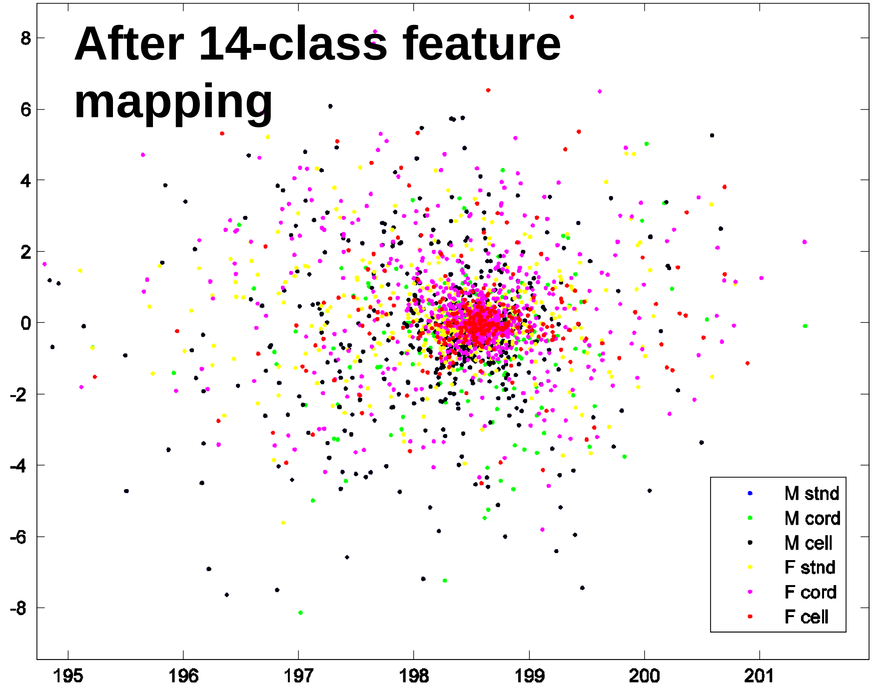


Projection of GMM super-vectors into first eigen-channel dimensions



Projection of GMM super-vectors into first eigen-channel dimensions – II.





=> No clusters visible after feature-mapping !