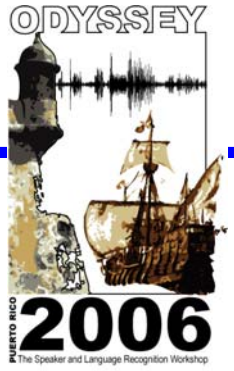


MIT Lincoln Laboratory Site Presentation

Bill Campbell, Doug Sturim, Wade Shen, Jiri Navratil*, and Doug Reynolds

NIST Speaker Recognition Workshop

26 June 2006

*IBM

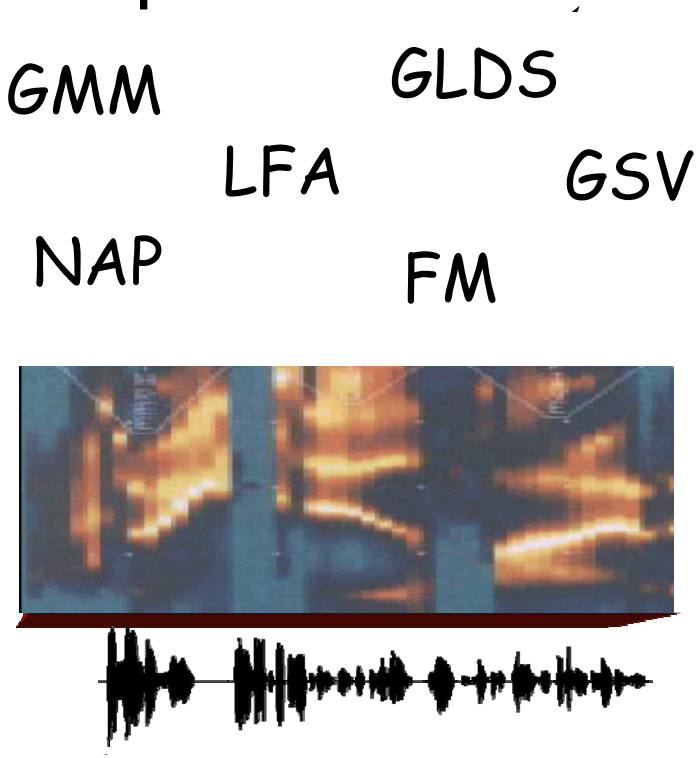
This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline

- System Overview
  - Theme: *Building the Base*
  - Core systems
  - Development data
- New for 2006
  - GMM with Latent Factor Analysis (LFA) Compensation
  - GMM SuperVector SVM
  - Multi-feature GLDS SVM
  - MLLR SVM with NAP Compensation
- Analysis
  - System breakout
  - Confidence score calibration
  - Final post-eval system and historic performance
- Conclusion

Theme for SRE 2006
Building The Base

- **At the 2002 JHU Summer Workshop, the SuperSID team demonstrated the power of exploiting multiple levels of speaker information in speech**
- **High-level features have shown incremental improvements in performance, but usually at substantial complexity and computational cost**
- **In keeping with our approach of making speaker recognition techniques *robust* and *portable* to new domains and platforms, we focused on *spectral based techniques***

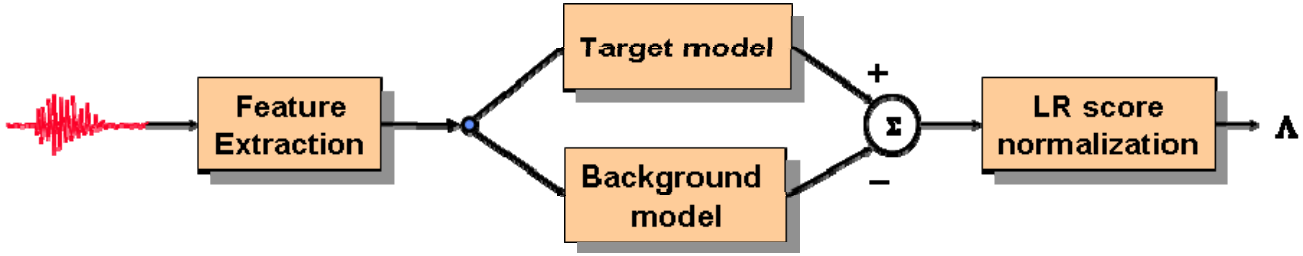


  - **direct attack on channel variability**
  - **robustness to language/dialect variability**
  - **computational speed**
  - **Small support infrastructure (e.g, no STT or phone rec)**

D. Reynolds, et al., "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," ICASSP 2003

System Overview

Core Detectors



New for 2006

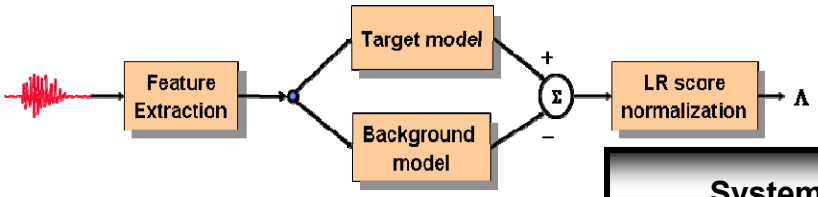
System	Features	Classifier	Znorm	Tnorm	Chan. Comp.
GMM-ATNORM	MFCC	GMM		55	FM
GMM-LFA	MFCC	GMM	200	300	LFA
SVM-GSV	GMM mean SuperVectors	SVM		300*	NAP
SVM-GLDS	MFCC+LPCC	SVM		300*	NAP
SVM-MLLR	MLLR coeff.	SVM		400*	NAP
SVM-WORD	Word lattice n-grams	SVM			
BT-WORD	Top-512 word occ.	Binary Tree	400 cnorm	400	
NGRAM-WORD	Word lattice	Lang. Model	400	400	
SVM-WORD_DUR	Word dur. stats.	SVM		400	

Word lattices and MLLR parameters are from BBN Byblos STT system

* Post-eval

System Overview

Development Data

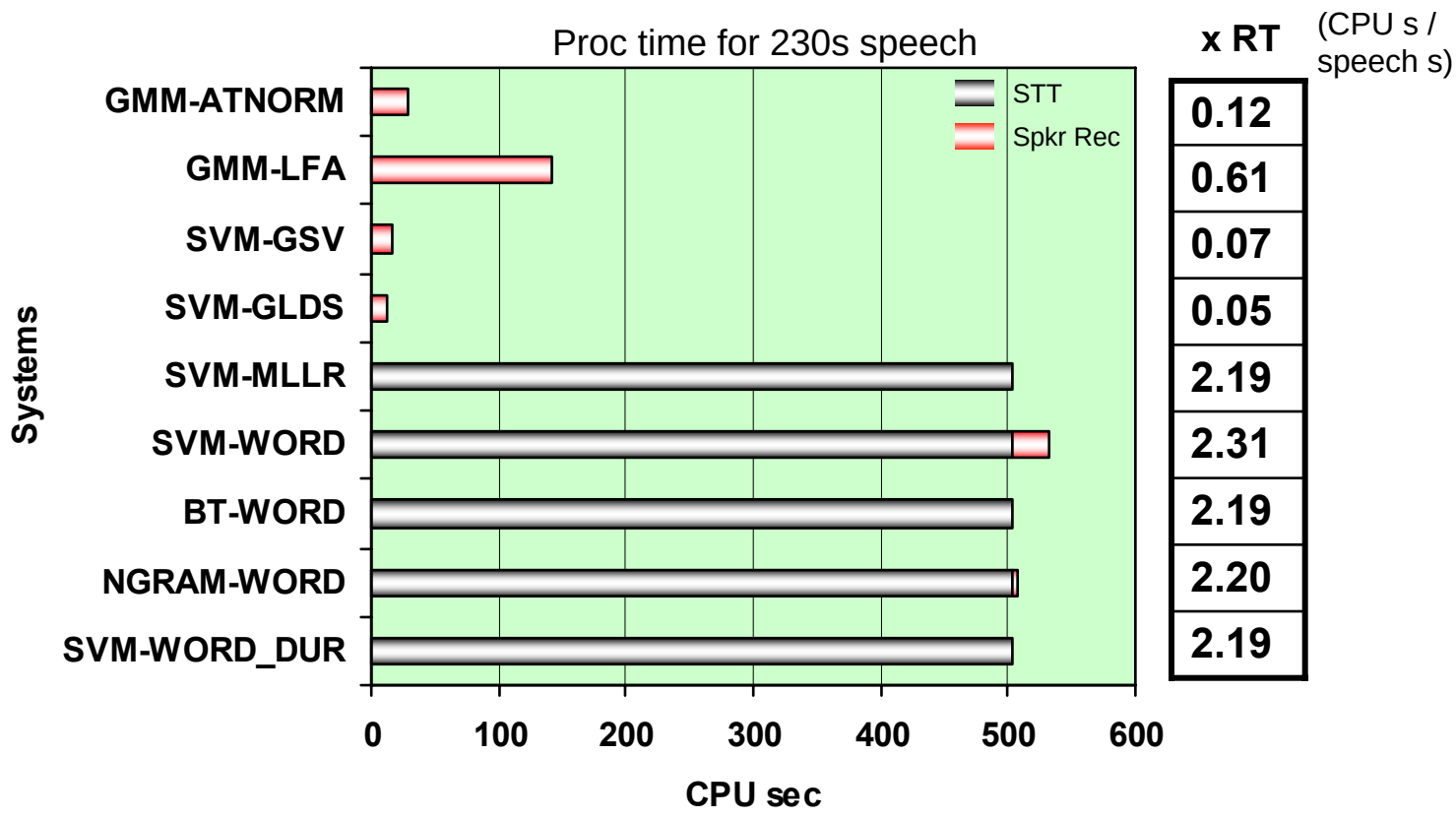


- Dev data focused on primary English condition
- Post-eval found further gains in using multi-language data

System	Background	Znorm	Tnorm	Chan. Comp
GMM-ATNORM	SWB2, SRE04		SRE04	SWB2, NatCell
GMM-LFA	SWB2, SRE04	SWB2	SRE04, FSH	SWB2
SVM-GSV	ubm=SWB2 svm=FSH		SRE04	SWB2
SVM-GLDS	FSH-ENG		SRE04	SWB2
SVM-MLLR	FSH-ENG		SRE04	SWB2
SVM-WORD	FSH-ENG			
BT-WORD	FSH-ENG	SRE05	SRE05	
NGRAM-WORD	FSH-ENG	SRE04	SRE04	
SVM-WORD_DUR	FSH-ENG		SRE04	
FUSION	Cross-Validation on system scores from SRE05			

System Processing Time*

- Input: ~230s speech
- Machine: Linux, Xeon 2-3GHz, 2 Gig memory
- All systems using STT output include STT time (~503s)



*Estimates from logfiles and reported times., Some times include model load and other I/O. Systems not run in optimized throughput mode.

Outline

- System Overview
  - Building the Base
  - Core systems
  - Development data
- **New for 2006**
  - **GMM with Latent Factor Analysis (LFA) Compensation**
  - **GMM SuperVector SVM**
  - **Multi-feature GLDS SVM**
  - **MLLR SVM with NAP Compensation**
- Analysis
  - System breakout
  - Confidence score calibration
  - Final post-eval system and historic performance
- Conclusion

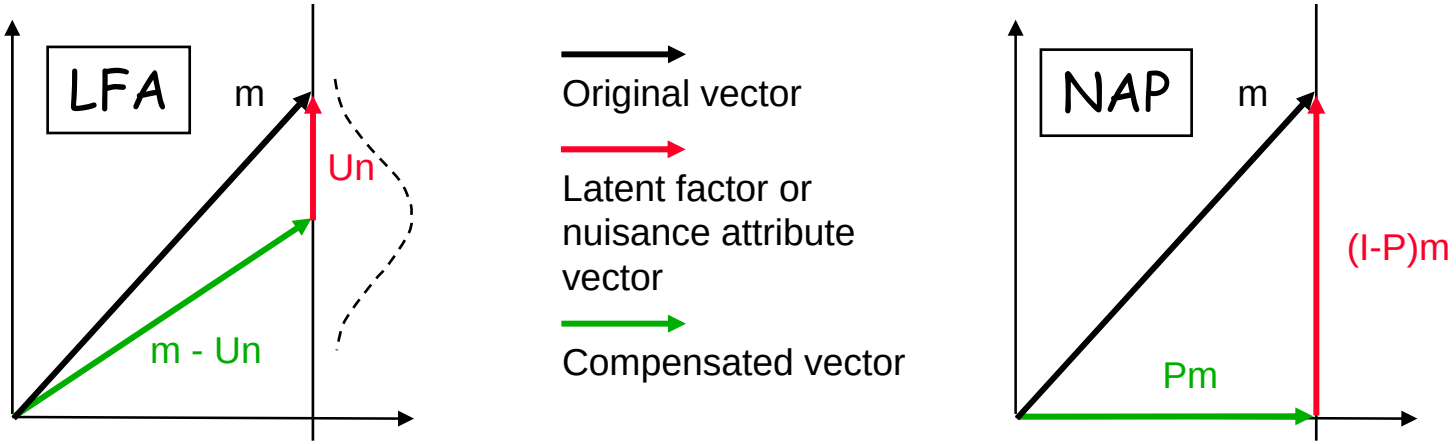
Spectral Systems
Motivations

- Concentration on session and channel variability compensation
  - Latent Factors Analysis (LFA)
    - Effective in SRE-2005 modeling session variation
    - Gaussian Mixture Models
  - Nuisance Attribute Projection (NAP)
    - Introduced in 2005 for SVM
    - Similarities to LFA for variation modeling
    - NAP is suited to high dimensional modeling (supervectors)
    - Support Vector Machines

- Combine best aspects of GMM and SVM systems
  - Gaussian Super Vector (GSV) SVM system
  - Hybrid of GMM-UBM distribution modeling with SVM discriminative classification

LFA and NAP Compensation

- **Both LFA and NAP attempt to remove undesired variation coming from a low-dimensional source**

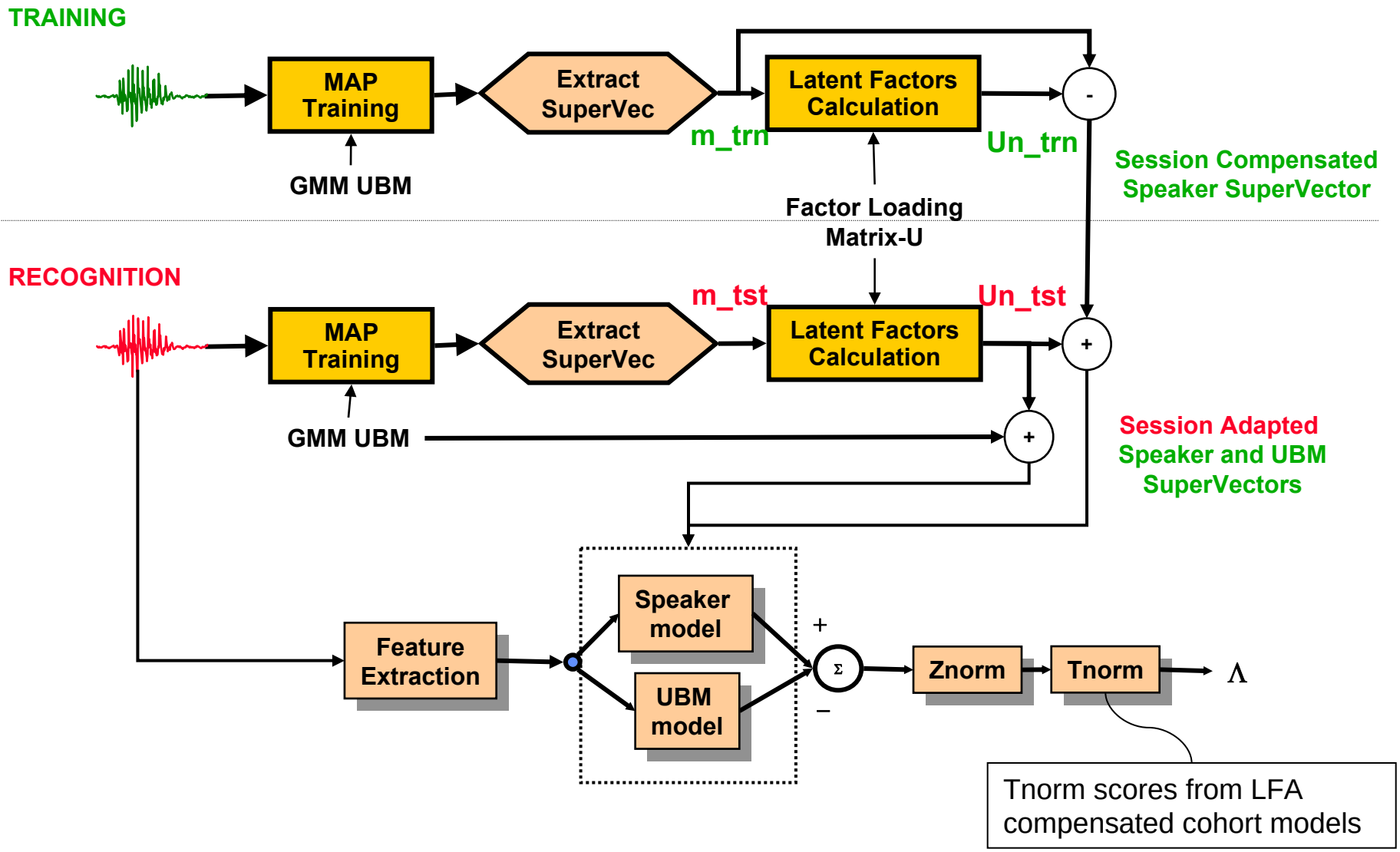


- **Amount of variation is hidden (latent) and described by source with a normal distribution (Bayesian type assumption)**
- **Estimates latent variables and subtracts out variation**
- **Applied in a GMM framework**
  - **LFA used on session variability**

- **Based on reducing a metric induced from SVM kernel**
- **Projects out nuisance space**
- **Applied in a SVM framework**
- **Handles channel, session, general nuisance**
  - **In 2005 NAP used on channel (cell, cb, elec) variability**
  - **In 2006 NAP used on session variability**

GMM with LFA Compensation

Training and Recognition

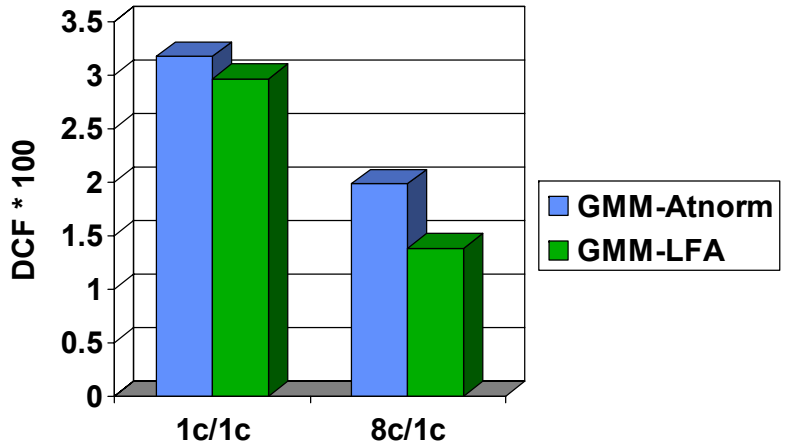


GMM with LFA Compensation

- **Details:**
  - **2048 Mixtures (512 mixtures in 1c)**
  - **Factor loading matrix calculated using kpca to calculate the eigenvectors**
  - **Znorm 200 utterances from SWB II**
  - **Tnorm drawn from Eval04 speakers**
    - **607 Cohorts on 8c 4-wire, 448 Cohorts on 3c {2,4}-wire, 394 Cohorts on 1c 4wire**
  - **Based on the model estimation algorithm presented in [Vogt06]**
  - **Did not use speaker factor estimation as in [Kenny05]**

- **Performance**
  - **GMM-LFA did slightly better than GMM-Atnorm at 1c/1c**
  - **At 8c/1c the GMM-LFA did significantly better than GMM-Atnorm**

SRE-2006 DCF versus Training condition-pooling all

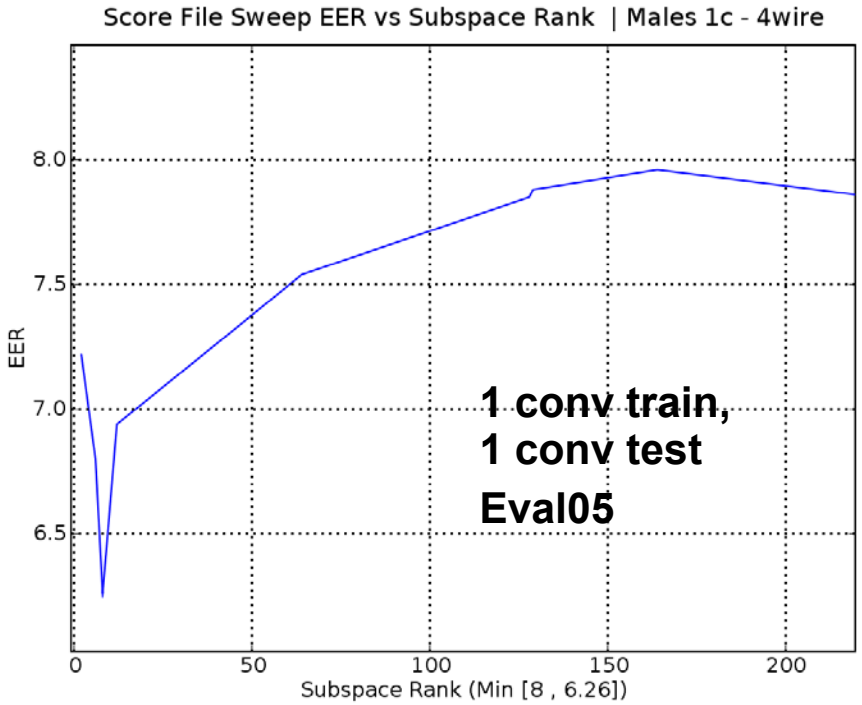


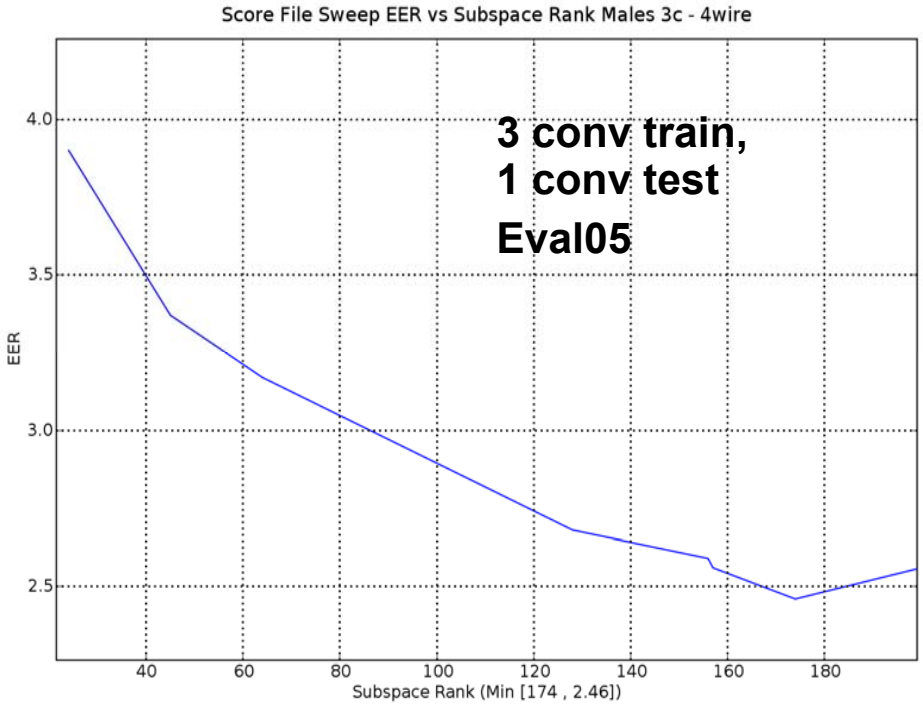
[Vogt06] R. Vogt, etal, Experiments In Session Variability Modeling For Speaker Verification, ICASSP 2006.
Kenny05] P, Kenny, etal, "Eigenvoice Modeling With Sparse Training Data" IEEE Trans SAP 2005.

GMM with LFA Compensation

Variation of Subspace Dimension

- Subspace dimension varies for with number of enrollment conversations
- Tuning critical to achieve good performance

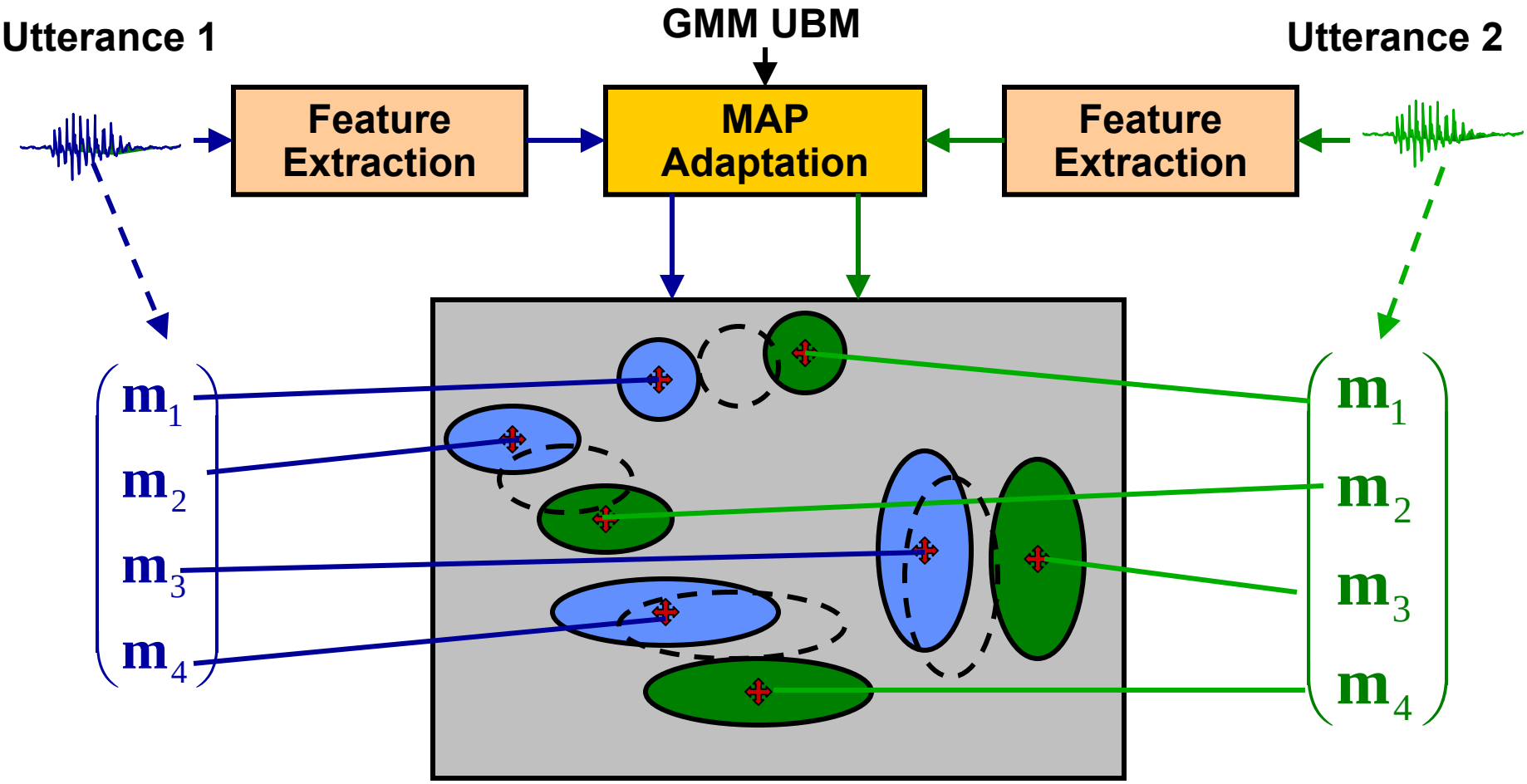




- Subspace dimension parameters were surprisingly stable from Eval05 to Eval06

GMM SuperVector SVM

Using Stacked Means

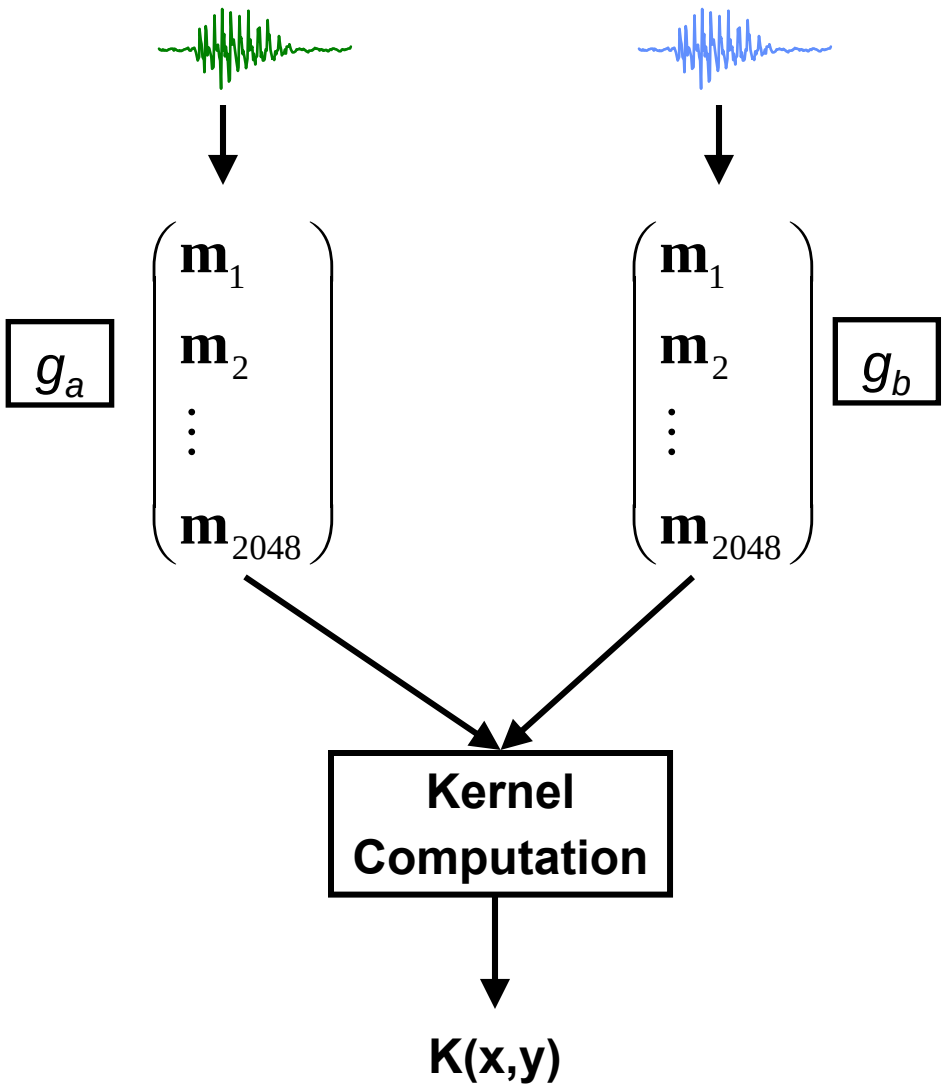


Different movements of means gives clues to speaker identity

UBM

GMM SuperVector SVM

- Use the GMM supervector in an SVM
- Supervectors are really just another way of describing a GMM
- Desirable to have a kernel that is computed directly from the supervectors



GMM SuperVector SVM

- Our approach:
  - KL divergence approximation
  - L^2 kernel
  - NAP session compensation
- Related Work:
  - (Wan-Sheffield) Fisher Kernels
  - (Ho/Moreno- HP Labs) KL divergence
  - (Campbell) SVM/GMM using GMM as a "decoder" to localize the scoring—but no stacking of means
  - CRIM
  - Persay
- References:
  - Campbell, W. M., D. Sturim, D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, vol 13, no. 5, pp. 308-311, 2006.
  - Campbell, W. M., D. Sturim, D. Reynolds, "SVM Based Speaker Verification using a GMM SuperVector Kernel and NAP Variability Compensation," ICASSP 2006.

GMM SuperVector SVM

Linear Kernel

- We want to look for comparisons of the MAP adapted models that involve GMM supervectors
- Indirectly: KL divergence
- "Linearize" to get final kernel
- Final kernel involves only operations with supervector

$$D(g_a \| g_b) = \int_{R^n} g_a(\mathbf{x}) \log \left(\frac{g_a(\mathbf{x})}{g_b(\mathbf{x})} \right) d\mathbf{x}$$

Upper Bound

$$D(g_a \| g_b) \leq \sum_{i=1}^{N} \lambda_i D\left(\mathcal{N}(\cdot; \mathbf{m}_i^a, \mathbf{\Sigma}_i) \| \mathcal{N}(\cdot; \mathbf{m}_i^b, \mathbf{\Sigma}_i) \right)$$

Compute

$$d(\mathbf{m}^a, \mathbf{m}^b) = \frac{1}{2} \sum_{i=1}^{N} \lambda_i (\mathbf{m}_i^a - \mathbf{m}_i^b) \mathbf{\Sigma}_i^{-1} (\mathbf{m}_i^a - \mathbf{m}_i^b)$$

Polarization

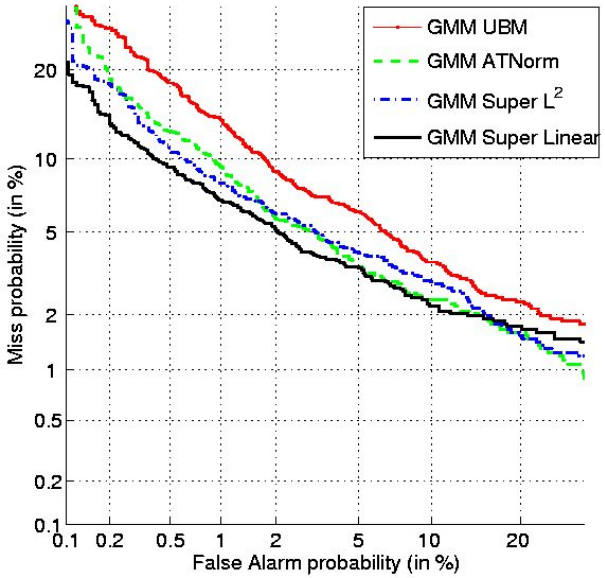
$$K(utt_a, utt_b) = \sum_{i=1}^{N} \lambda_i \mathbf{m}_i^a \mathbf{\Sigma}_i^{-1} \mathbf{m}_i^b$$

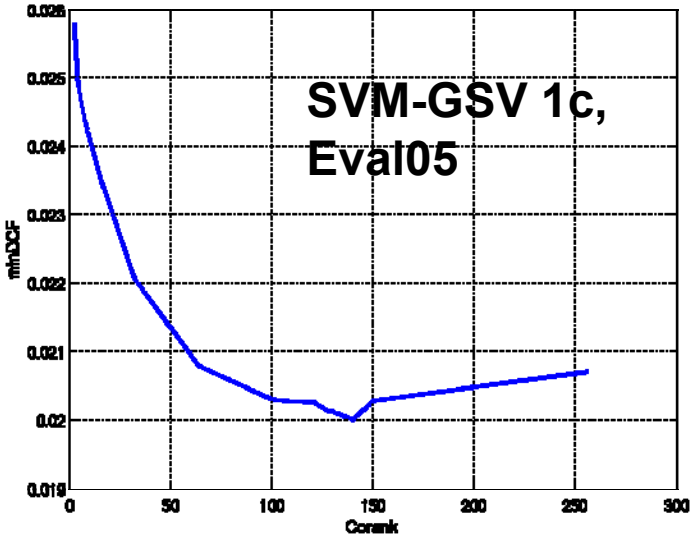
$$= \sum_{i=1}^{N} \left(\sqrt{\lambda_i} \mathbf{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_i^a \right)^t \left(\sqrt{\lambda_i} \mathbf{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_i^b \right)$$

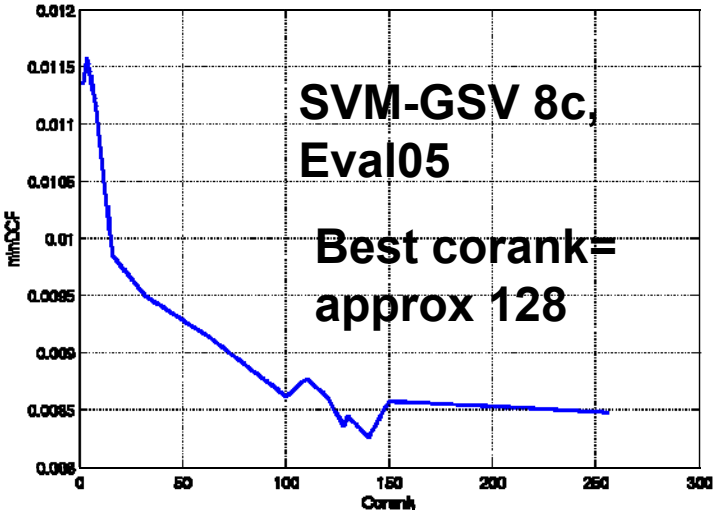
GMM SuperVector SVM

Tuning

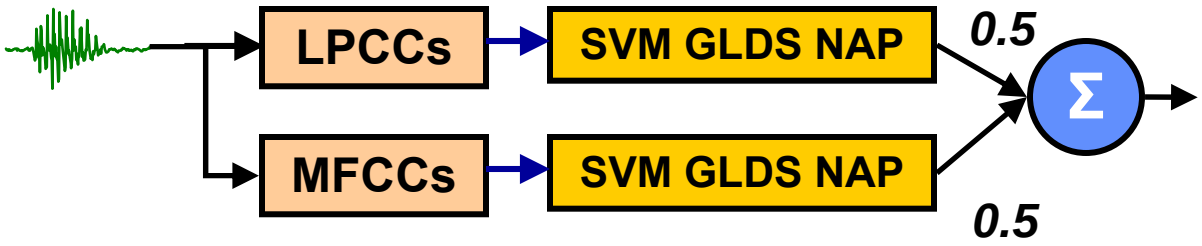
- **Kernel selection:**
  - 8c Eval05 example
  - L^2 kernel was based upon standard integral inner product
  - Conclusion: Linear kernel worked the best and was easiest to implement
- **Session NAP tuning:**
  - As we vary the dimension of the nuisance subspace (corank) the EER performance varies
  - Optimal NAP corank fairly consistent across different enrollment durations

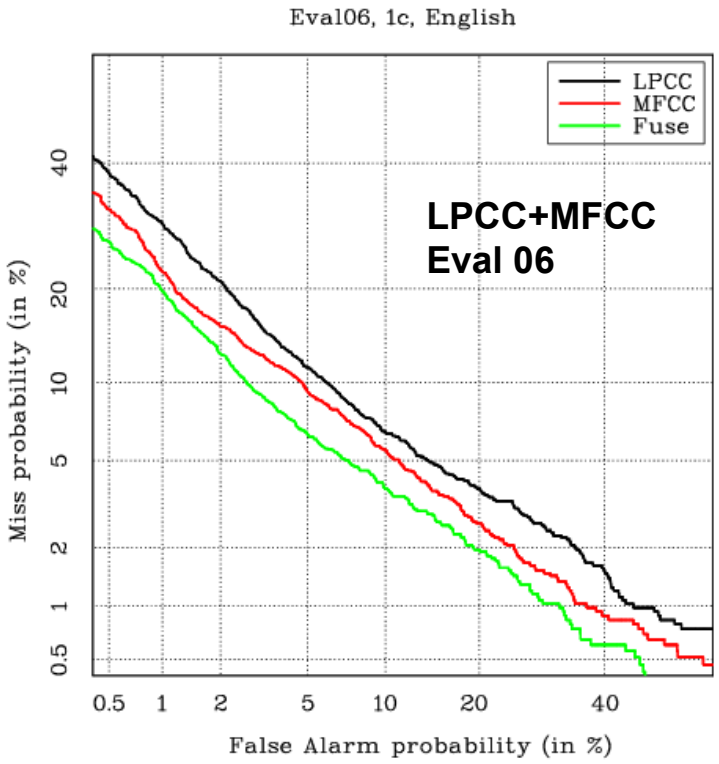


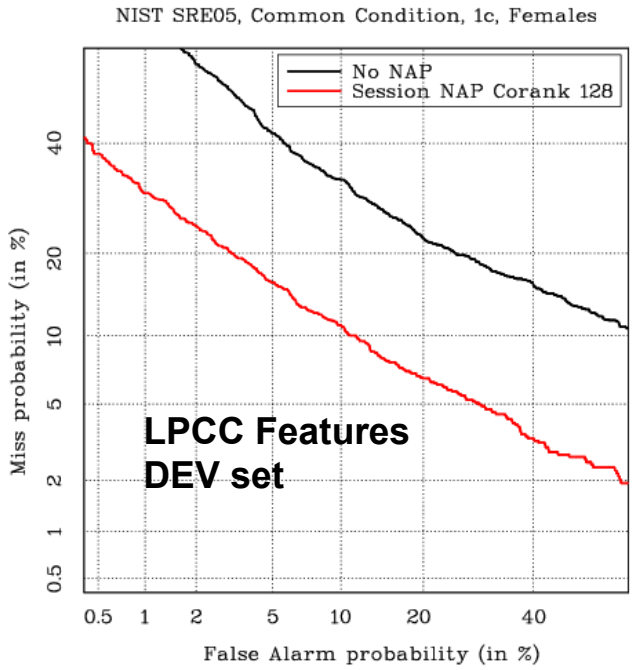




Multi-feature GLDS SVM





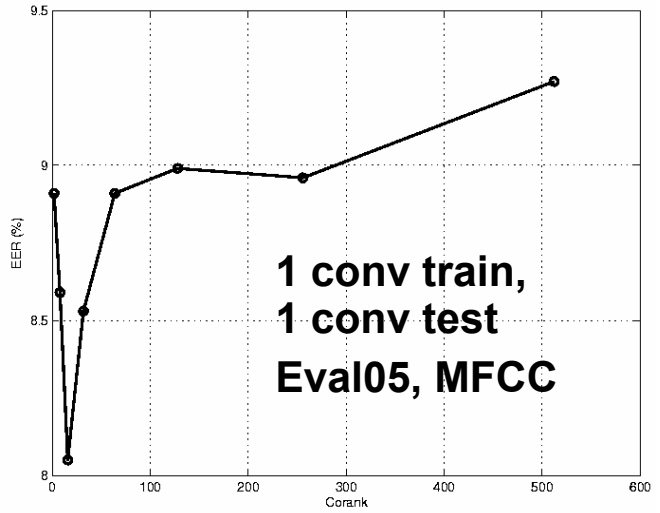


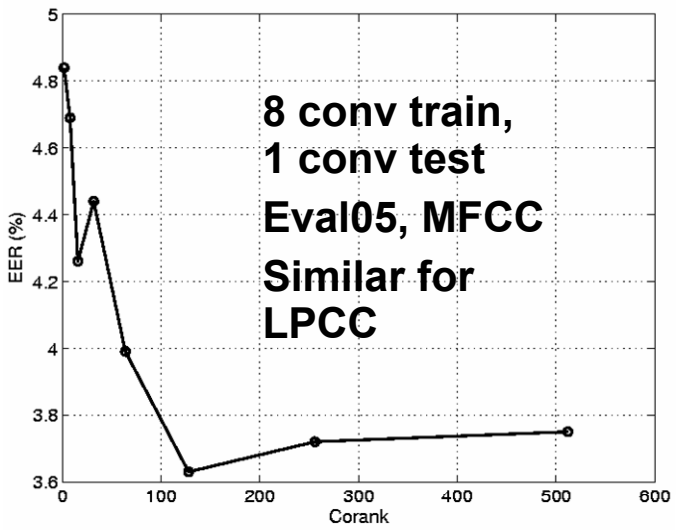
- Updated with new feature strategy
- Found that substantial gains could be obtained by applying NAP to LPCC features
- Resulting system had a fusion gain on 05 data and 06 data

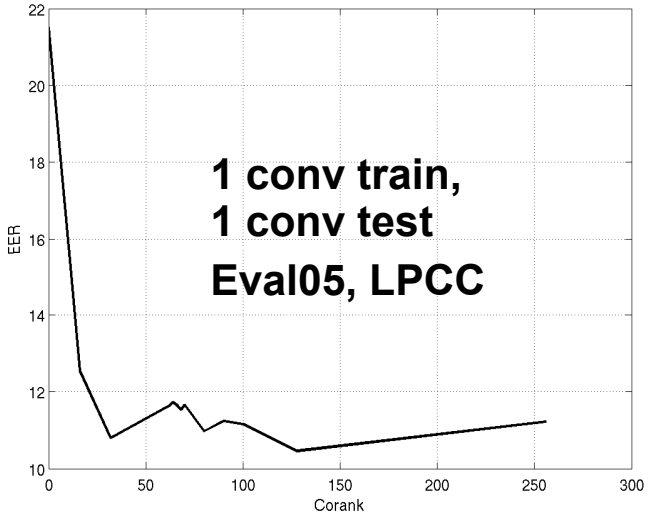
Multi-feature GLDS SVM

Variation of Session NAP with Corank

- Different behaviors for:
  - number of enrollment conversations
  - features
- NAP behaves differently for SVM-GSV versus SVM-GLDS
- Tuning critical to achieve good performance

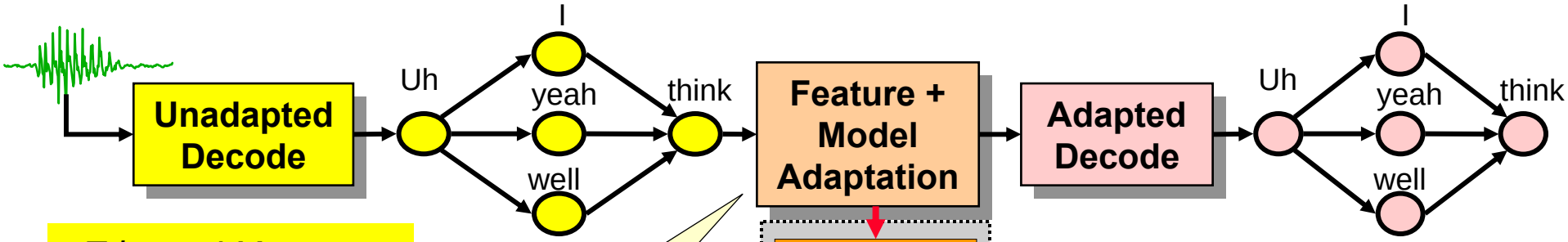






SVM MLLR with NAP Compensation

Applying Byblos STT to SID



- *Trigram LM*
- *GI non-xword Triphone models*

MLLR Coefficients

- Found Data! From our high-level systems
- 3 transforms (derived from data driven classes)
  - 1 root + 2 classes
  - |Transform| = 3600
- Only used Model Space Transforms
  - Other Feature space adaptations performed before MLLR

Adapted Lattices For High-level Systems

- *4-gram LM*
- *GD xword Quinphone models*

Used for

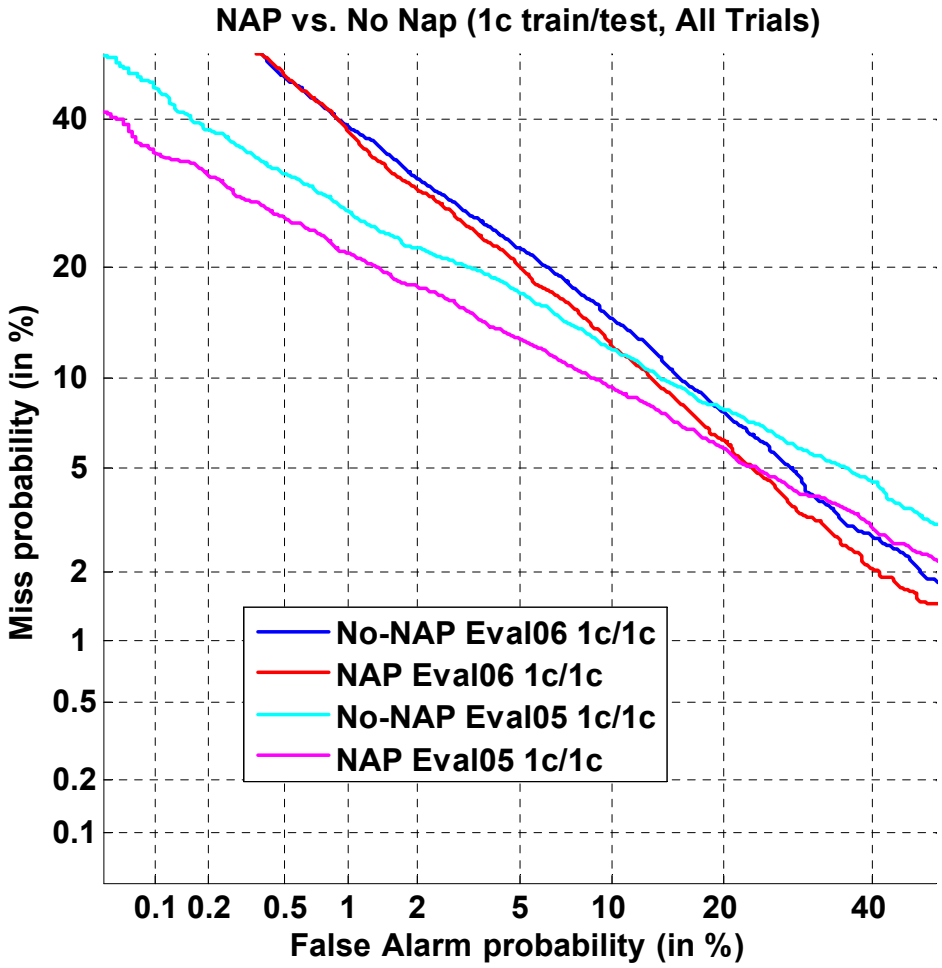
- *Word Duration*
- *N-gram systems*

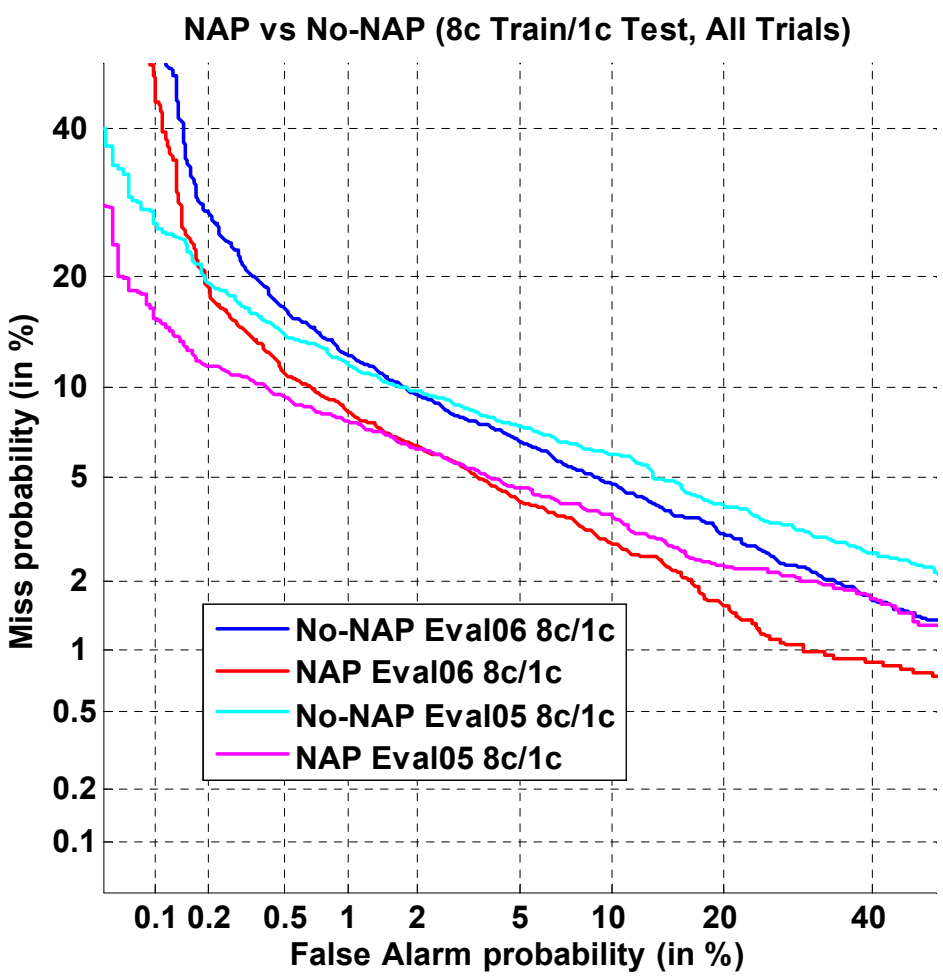
SVM MLLR with NAP

- 0-1 Normalization
- Applied NAP without modification
  - Switchboard training
- Fisher English Background
- Eval04 T-norm models

SVM MLLR with NAP Compensation

Results





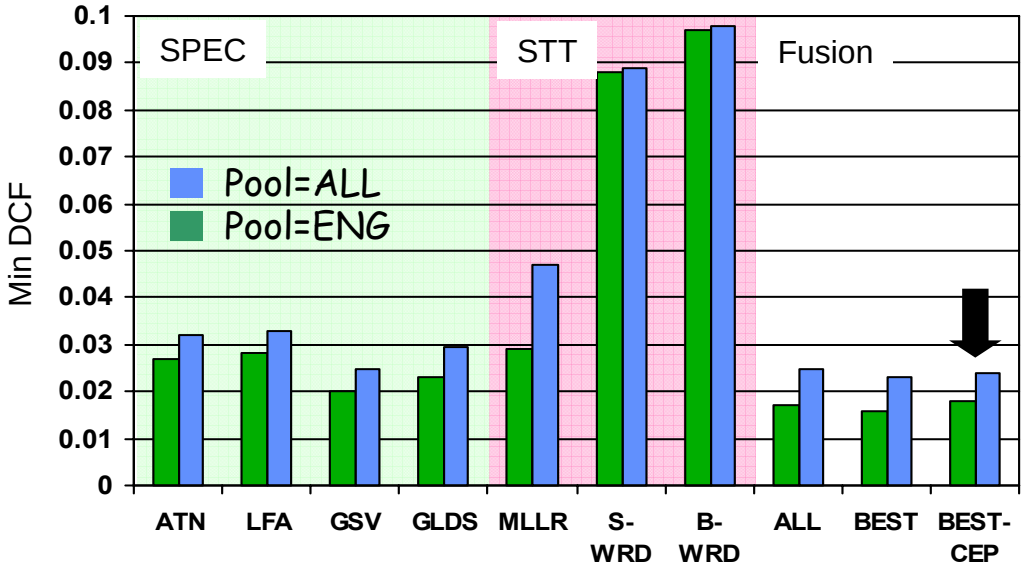
Outline

- System Overview
  - Building the Base
  - Core systems
  - Development data
- New for 2006
  - GMM with Latent Factor Analysis (LFA) Compensation
  - GMM SuperVector SVM
  - Multi-feature GLDS SVM
  - MLLR SVM with NAP Compensation
- **Analysis**
  - **System breakout**
  - **Confidence score calibration**
  - **Final post-eval system and historic performance**
- Conclusion

System Breakout

Min DCF for 1c/1c and 8c/1c

1conv4w/1conv4w

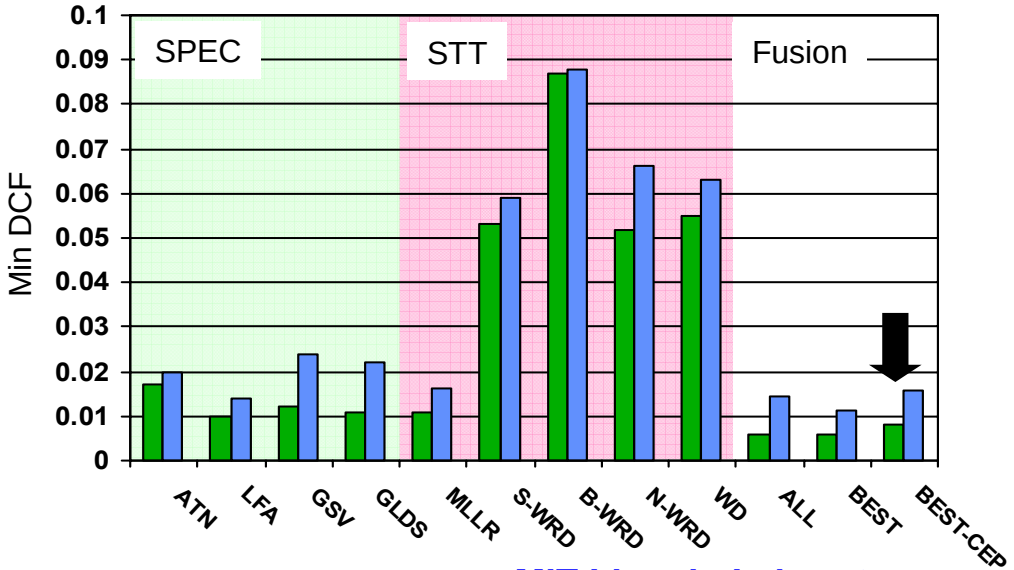


- Very low error rates for new data set (EER < 2% for 8c)
- Spectral based systems generally outperform STT based systems
  - MLLR is exception
  - But this is a spectral space transform

- Small accuracy loss from ENG to ALL pooling
- Fusion within spectral systems has performance similar to all fusion

ENG	1c/1c		8c/1c	
	EER	DCF	EER	DCF
Best	3.5	0.016	1.5	0.0056
Best cep	4.0	0.019	2.0	0.0080

8conv4w/1conv4w*

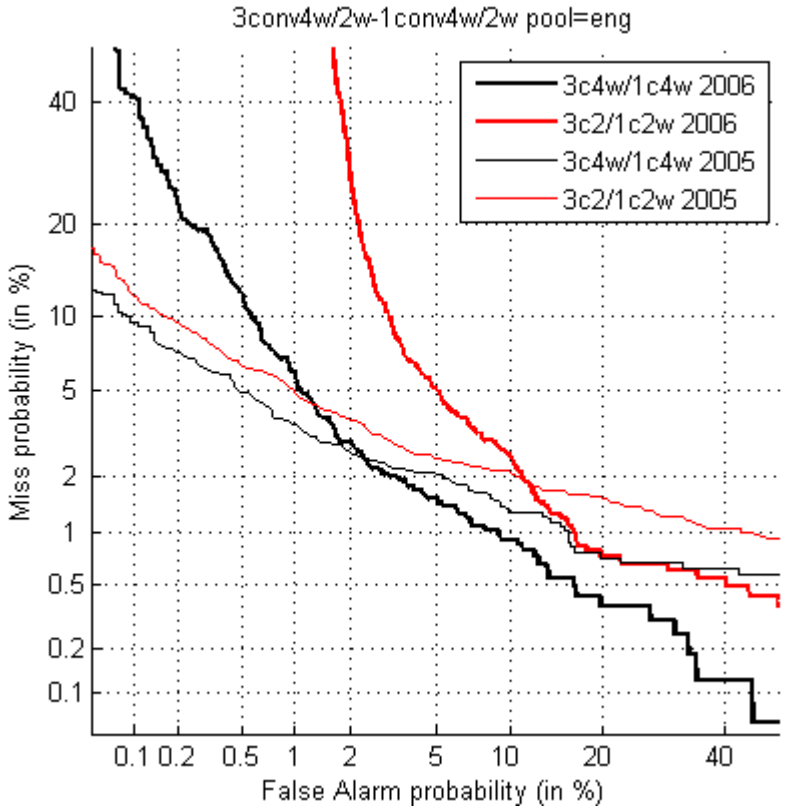


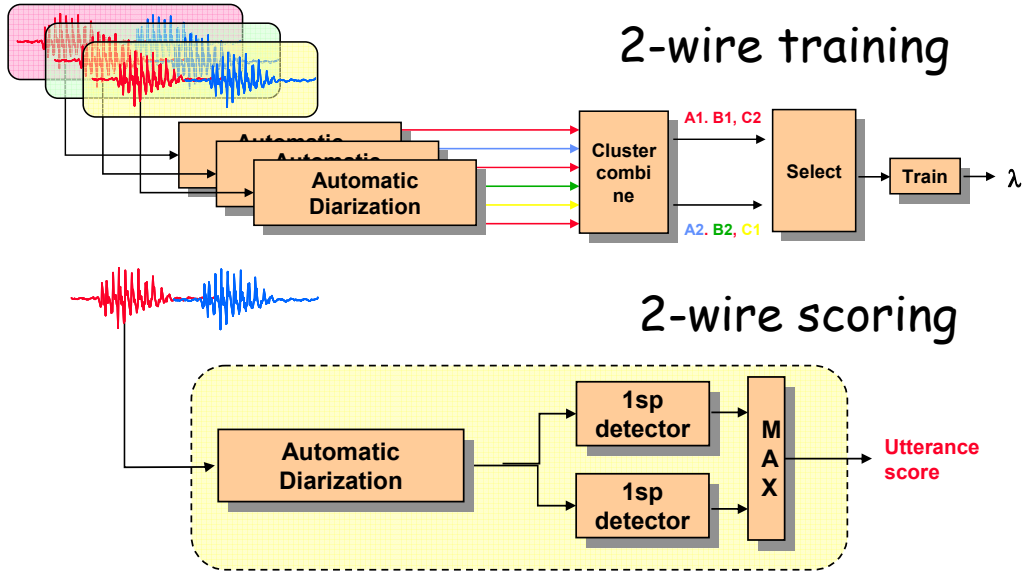
*8conv4w/1conv4w: not all systems T-normed

System Breakout

3c 2-wire Processing

- **Divide and conquer approach**
  - **Allows application of optimized detection systems**
- **Purification is critical step when using summed data**



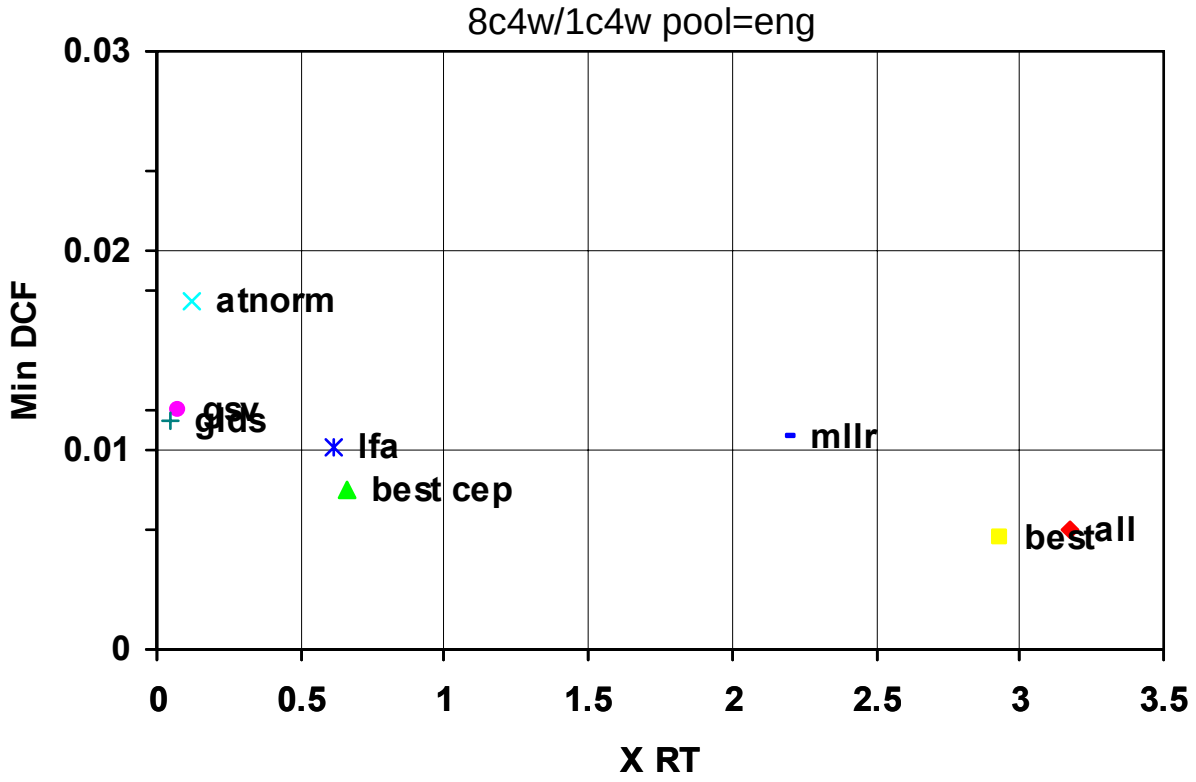


- **Loss of ~2.5% in EER between 2w and 4w processing**
- **Odd DET curve shape on 2006 data**
  - **Problem in key?**

System Breakout

Accuracy / Computation Tradeoff

- **High-level features provide gains … but at a cost**
  - **Computation and reliance on particular language (e.g. English)**
- **Most practical when STT is also needed in an application**
  - **May not allow speaker recognition 'tweaks' (e.g., MLLR classes)**
- **Are there less costly ways to extract the same information?**



*RT for combinations does not assume any computation sharing which would likely occur in front-end processing.

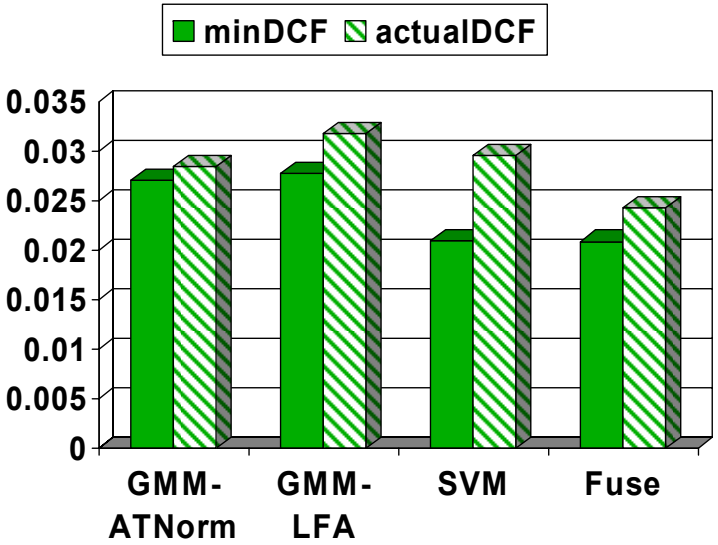
Score Calibration Analysis

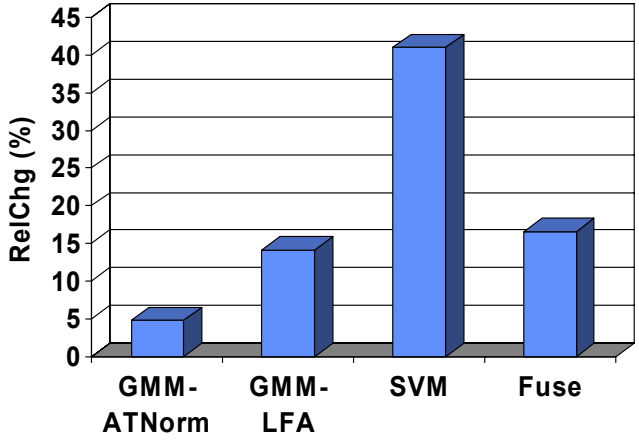
Thresholds

- minDCF, actualDCF disparity
- Systems:
  - GMM LFA
  - GMM ATNorm
  - SVM-GLDS
  - SVM-GSV
  - SVM-MLLR
- SVM = Fusion of all SVMs
- Fuse = Fuse all 5 systems
- Measuring stability:

  RelChg = (actDCF-minDCF)/minDCF
- Plots show submission systems
- Problem is worse for all trials

English Trials

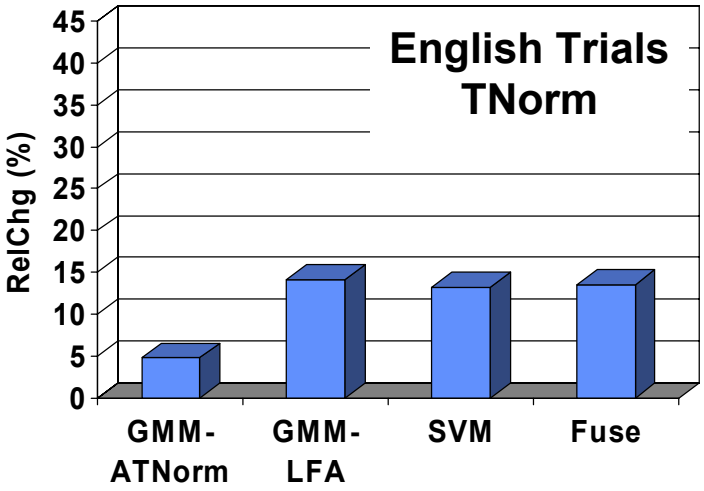


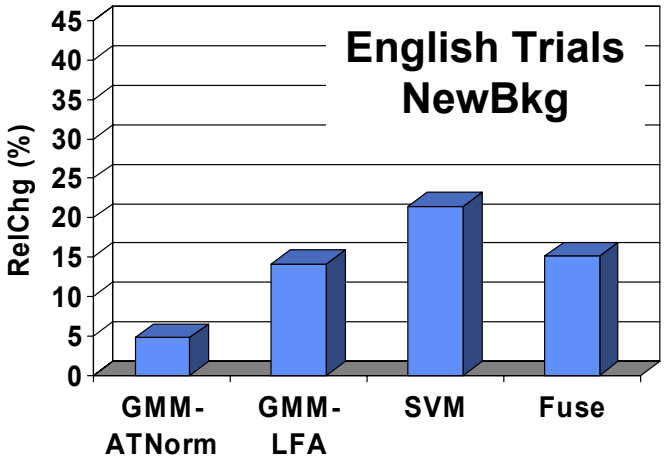


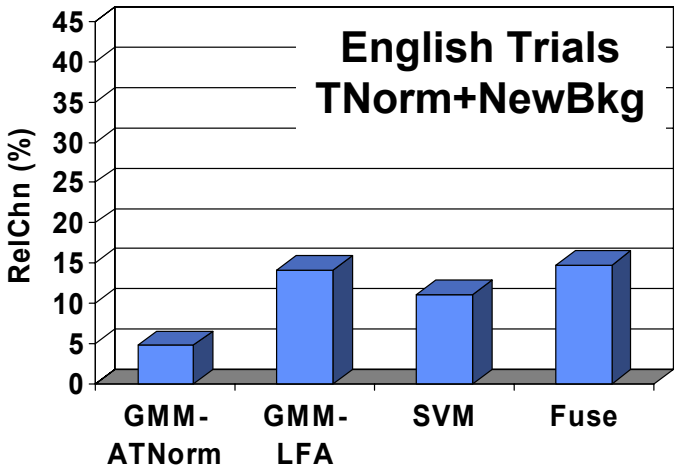
Score Calibration Analysis

TNorm + Background

- **TNorm:**
  - Added TNorm to all SVM systems
  - TNorm speakers from Eval04
- **New Background:**
  - Added non-English Fisher data (Arabic/Mandarin) to SVM backgrounds
  - SVM-GLDS, SVM-GSV
- **Both: TNorm+Bkg**



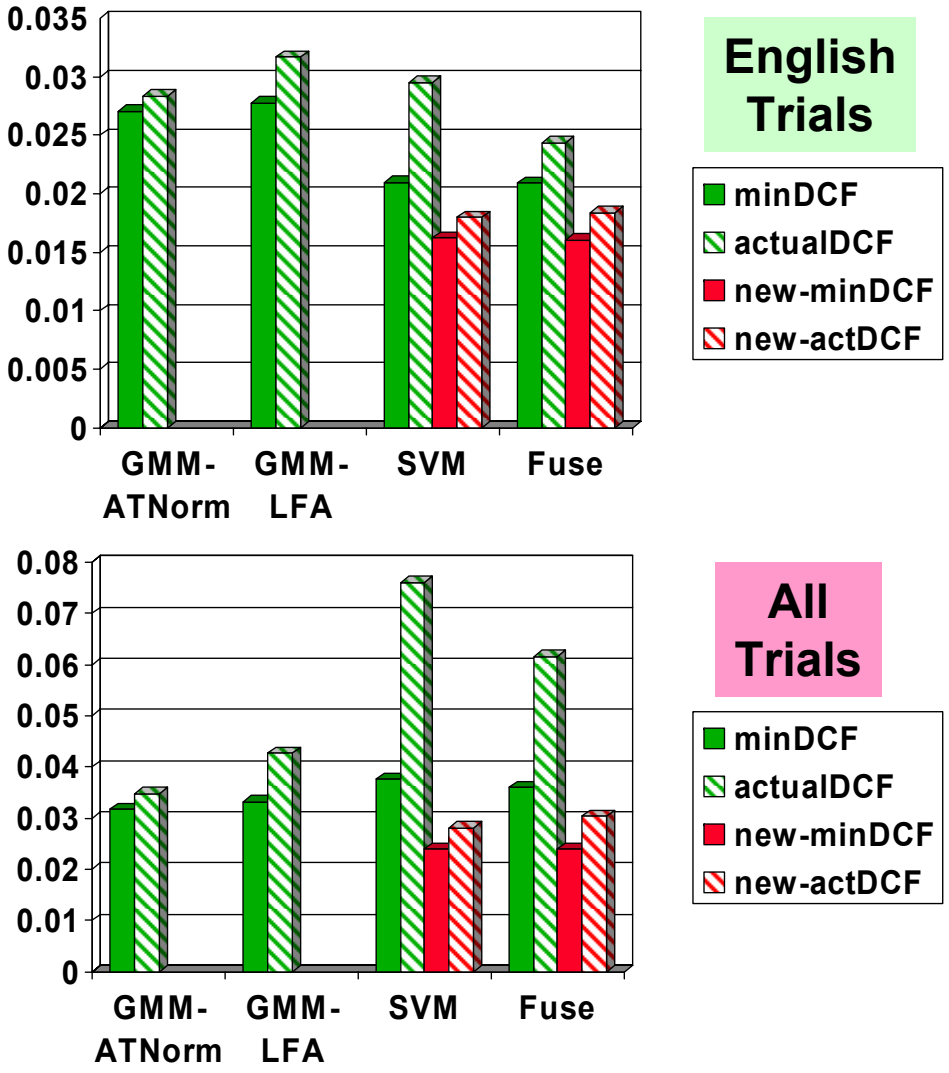




Score Calibration Analysis

Stabilization Results

- Final results show both English trials and all trials
- TNorm is a huge win for stabilizing thresholds for the SVMs—haven't seen this behavior before
- Stabilizing thresholds is possible
- All trials still is a challenge
- RelChg improvement:
  - English: 16.5% → 14.7%
  - All: 71.5% → 27.0%



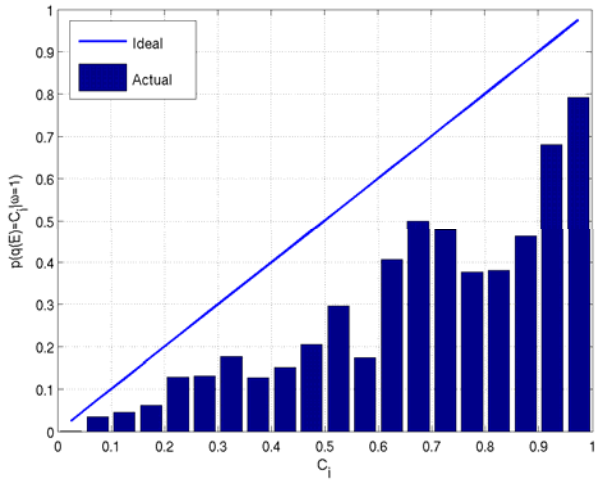
Score Calibration

Calibration & Refinement

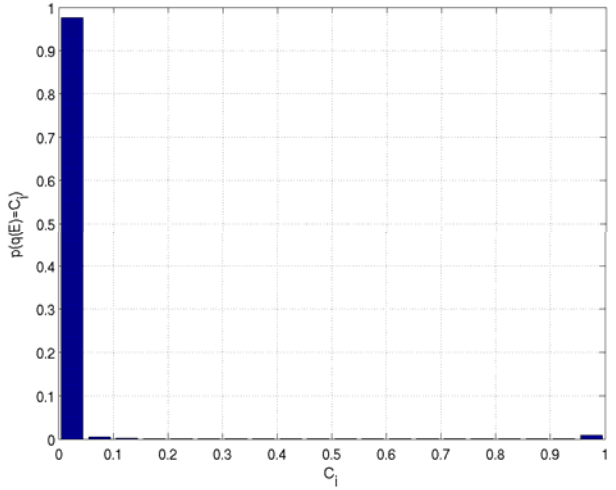
- English trials, new fusion system
- Calibration=How well does the output approximate a posterior?
- Refinement=Does the system produce scores near 0 & 1?

- Calibration still not good across all thresholds
- Chance:
  - $h(P_{tgt})=h(0.01)=0.081$ bits
- Cross-entropy (CE):
  - 0.038 bits
- NCE = (Chance-CE)/Chance
  - 53.0%, this is reasonable
- CE=calib+refinement
- Calibration error: 0.017 bits
- Refinement: 0.021 bits
- Calibration is a large part of the cross-entropy; ideally should be zero

Calibration



Refinement



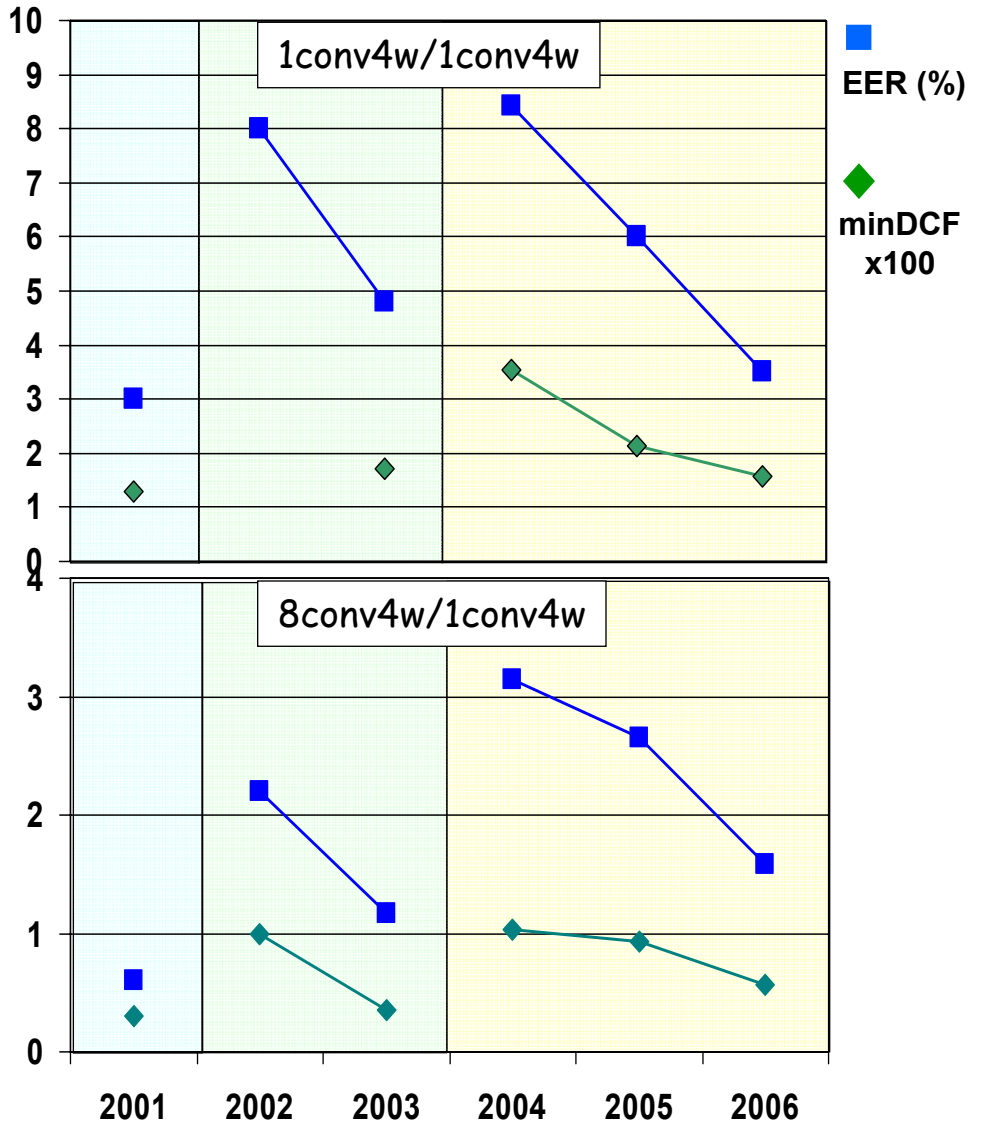
MITLL Submissions
Historic Performance

- Consistent and steady improvement for data/task focus

2001	2002	2003	2004	2005	2006
SWB1	SWB2		MIXER2-3		

- New data sets designed to be more challenging
- New features, classifiers and compensations drive error rates down over time

2001	Text-const GMM, word-ngram
2002	SuperSID : High-level features
2003	Feature Mapping, SVM-GLDS
2004	Phone/Word-SVM, GMM-ATNORM
2005	NAP, TC-SV, word/phone lattices
2006	SVM-GSV, GMM-LFA, MultiFeat SVM-GLDS, SVM-MLLR+NAP



Conclusions

- Excellent progress in 2006
  - Many sites independently demonstrating effectiveness of new features, classifiers and compensations
- MITLL focus was on spectral based systems
  - Direct attack on channel variability
  - Robustness to language/dialect variability
  - Computational speed
  - Minimal support infrastructure
- Highlights of new items for 2006
  - GMM with Latent Factor Analysis (LFA) Compensation
  - GMM SuperVector SVM
  - Multi-feature GLDS SVM
  - MLLR SVM with NAP Compensation
- Threshold analysis highlighted need for Tnorm
  - Unexplained calibration bias (mixer2 – mixer3)
- Retrospective look at performance shows a consistent and steady improvement for data/task focus
  - High-level SuperSID features brought attention to extended data task
  - Main drivers in performance improvement have been new spectral based systems and channel compensations

Selected References

Cepstral GMM

- D.A. Reynolds, T.F. Quatieri, R.B. Dunn. “Speaker Verification using Adapted Gaussian Mixture Models,” Digital Signal Processing, 10(1--3), January/April/July 2000
- Sturim, D.E. Reynolds, D.A. “Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification”, ICASSP '05.

Cepstral SVM

- W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," ICASSP 2002

GMM SuperVector SVM

- Campbell, W. M., D. Sturim, D. Reynolds, “SVM Based Speaker Verification using a GMM SuperVector Kernel and NAP Variability Compensation”, ICASSP 2006.
- Campbell, W. M., D. Sturim, D. Reynolds, “Support vector machines using GMM supervectors for speaker verification”, IEEE Signal Processing Letters, vol 13, no. 5, pp. 308-311, 2006.

NAP

- A. Solomonoff, W. Campbell, I. Boardman, “Advances In Channel Compensation For SVM Speaker Recognition,” ICASSP 2005
- Alex Solomono, William M. Campbell, Carl Quillen, ‘’Nuisance Attribute Projection’’, To appear in Speech Communications.

Latent Factor Analysis

- R. Vogt, S..Sridharan, Experiments In Session Variability Modeling For Speaker Verification, ICASSP 2006.
- P, Kenny, G. Boulianne, P. Dumouchel, “Eigenvoice Modeling With Sparse Training Data” IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING VOL. 13, NO. 3, MAY 2005.
- M.Tipping, C. Bishop,”Mixtures of Probabilistic Principal Component Analyzers”, Neural Computation 11, 443-482 (1999).

MLLR

- Stolcke, A. et al, "MLLR Transforms as Features in Speaker Recognition,“ in the Proceedings of Eurospeech 2005, Lisbon, Portugal 2005.

Word and Phonetic SVM and LLR

- A. Hatch, B. Peskin, A. Stolcke, “Improved Phonetic Speaker Recognition Using Lattice Decoding,” ICASSP 2005
- W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, T. R. Leek, "High-Level Speaker Recognition with Support Vector Machines," Proceedings of ICASSP, 2004

Confidence Esitmation and Fusion with metadata

- W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. J. Brady, “Estimating And Evaluating Confidence For Forensic Speaker Recognition,” ICASSP 2005

Diarization for Speaker Recognition

- Sue Tranter and Douglas Reynolds, "An Overview of Automatic Speaker Diarisation Systems," Special Issue on Rich Transcription, IEEE Trans on Speech and Lang. Processing, to appear October 2006
- D. A. Reynolds and P. Torres-Carrasquillo, “Approaches and Applications of Audio Diarization,” ICASSP 2005
- R.B. Dunn, D.A. Reynolds, T.F. Quatieri. “Approaches to Speaker Detection and Tracking in Multi-Speaker Audio,” Digital Signal Processing, 10(1--3), January/April/July 2000.