Information Technology Laboratory
Information Access Division (IAD)

NIST
National Institute of Standards and Technology

NIST 2006
Speaker Recognition Evaluation

Evaluation Results

Alvin Martin and Audrey Le
www.nist.gov/speech/tests/spk

June 26-27th, 2006
San Juan, Puerto Rico

Outline

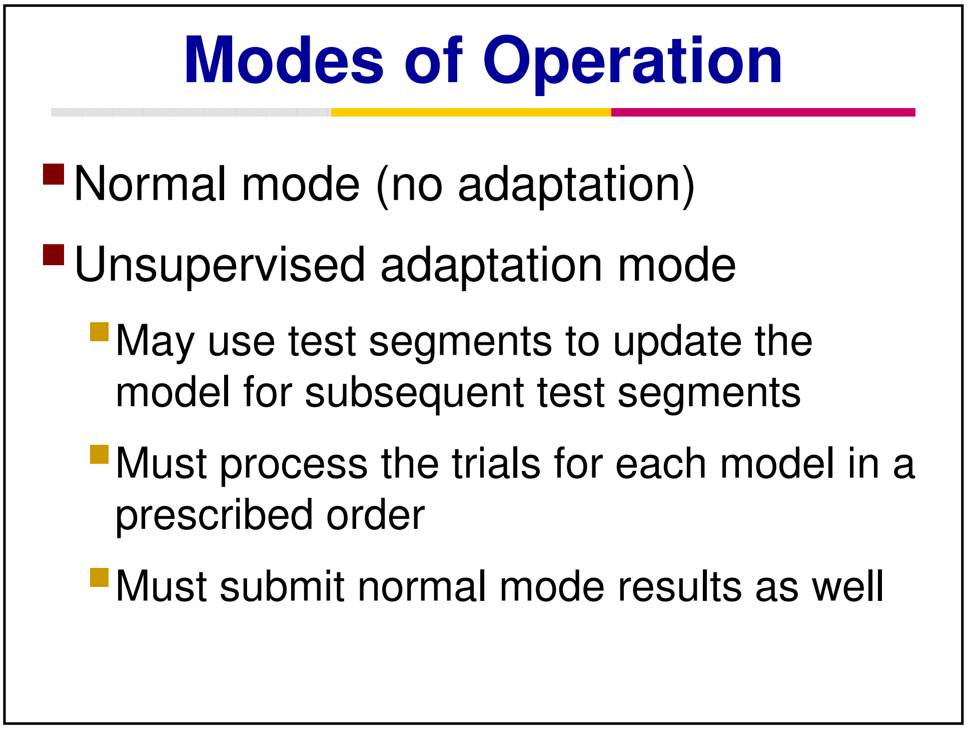
- Today
  - Evaluation Review
  - Evaluation Results
  - Mothballed Systems and History Plots
  - Language Effects
  - Summary
- Tomorrow
  - Cross-channel Results

Evaluation Review

- Task
- Modes of Operation
- Conditions
- Rules
- Data
- Changes from Last Year
- Metric and Performance Representation

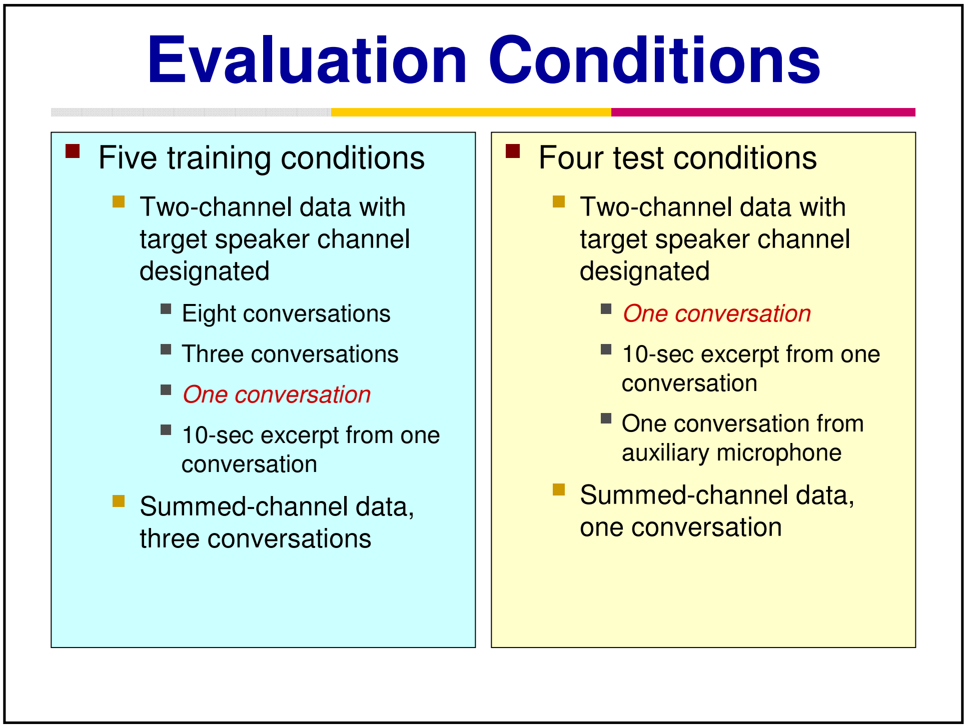
Speaker Detection Task

- Given a model speaker and side information, determine if that speaker is speaking in a given test segment
  - A model and a test segment define a *trial*
  - Permitted side information
    - Gender of the model speaker
    - ASR transcripts



Modes of Operation

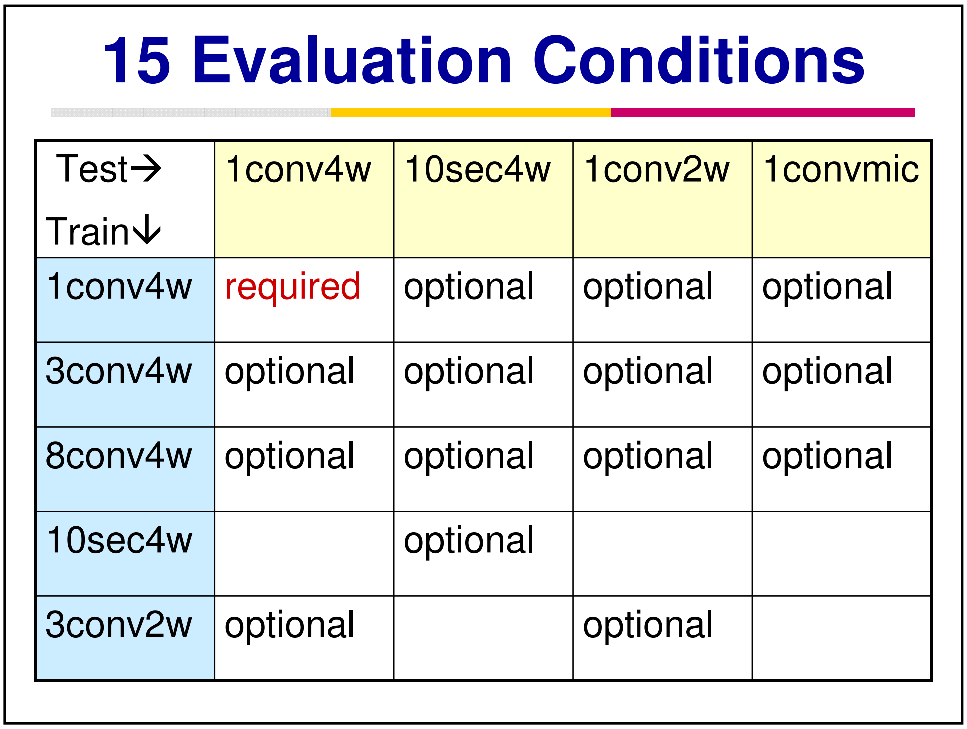
- Normal mode (no adaptation)
- Unsupervised adaptation mode
  - May use test segments to update the model for subsequent test segments
  - Must process the trials for each model in a prescribed order
  - Must submit normal mode results as well



Evaluation Conditions

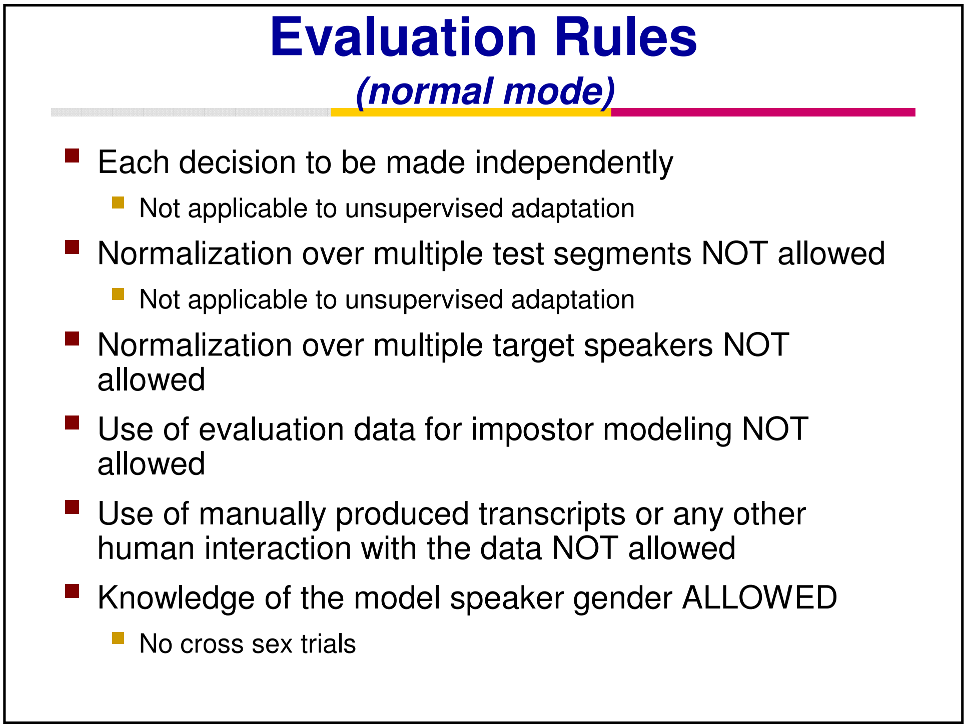
- Five training conditions
  - Two-channel data with target speaker channel designated
    - Eight conversations
    - Three conversations
    - *One conversation*
    - 10-sec excerpt from one conversation
  - Summed-channel data, three conversations

- Four test conditions
  - Two-channel data with target speaker channel designated
    - *One conversation*
    - 10-sec excerpt from one conversation
    - One conversation from auxiliary microphone
  - Summed-channel data, one conversation



15 Evaluation Conditions

Test→ Train↓	1conv4w	10sec4w	1conv2w	1convmic
1conv4w	required	optional	optional	optional
3conv4w	optional	optional	optional	optional
8conv4w	optional	optional	optional	optional
10sec4w		optional		
3conv2w	optional		optional	



Evaluation Rules

(normal mode)

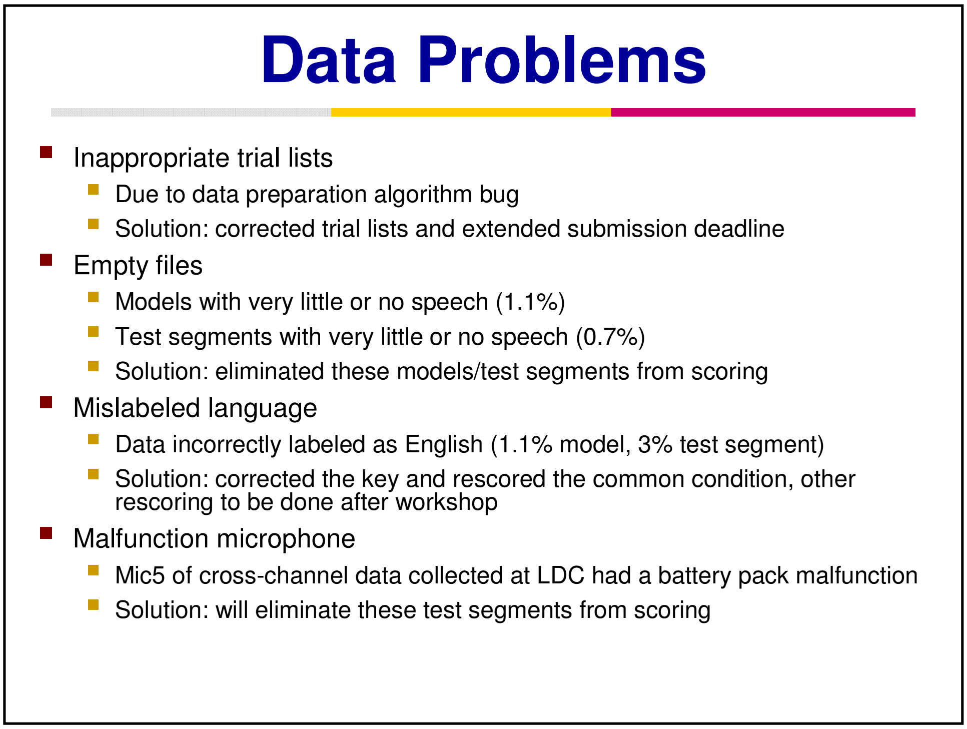
- Each decision to be made independently
  - Not applicable to unsupervised adaptation
- Normalization over multiple test segments NOT allowed
  - Not applicable to unsupervised adaptation
- Normalization over multiple target speakers NOT allowed
- Use of evaluation data for impostor modeling NOT allowed
- Use of manually produced transcripts or any other human interaction with the data NOT allowed
- Knowledge of the model speaker gender ALLOWED
  - No cross sex trials

Evaluation Data

- MIXER3
  - 528 new speakers
    - 139 native English speakers, 389 bilingual speakers
  - recordings made from Dec, 2005 to Feb, 2006
- Cross-channel
  - 85 unexposed MIXER2 speakers
    - 57 had other non-cross-channel calls
  - collected at LDC & ICSI
- SRE05 Data
  - 398 speakers from SRE05 for 8-conv training condition
- 16558 test segments
- 3459 models
  - 1484 male, 1975 female
- 514706 trials

Data Processing

- Processed with the Mississippi State provided echo canceller
- The “10 second” training and test segments had 7-13 seconds of actual speech
- ASR transcripts created for training and test data
  - Processed at BBN with a 1x real-time system
  - English recognizer run on all data in all languages
  - ASR produced no transcripts for some segments

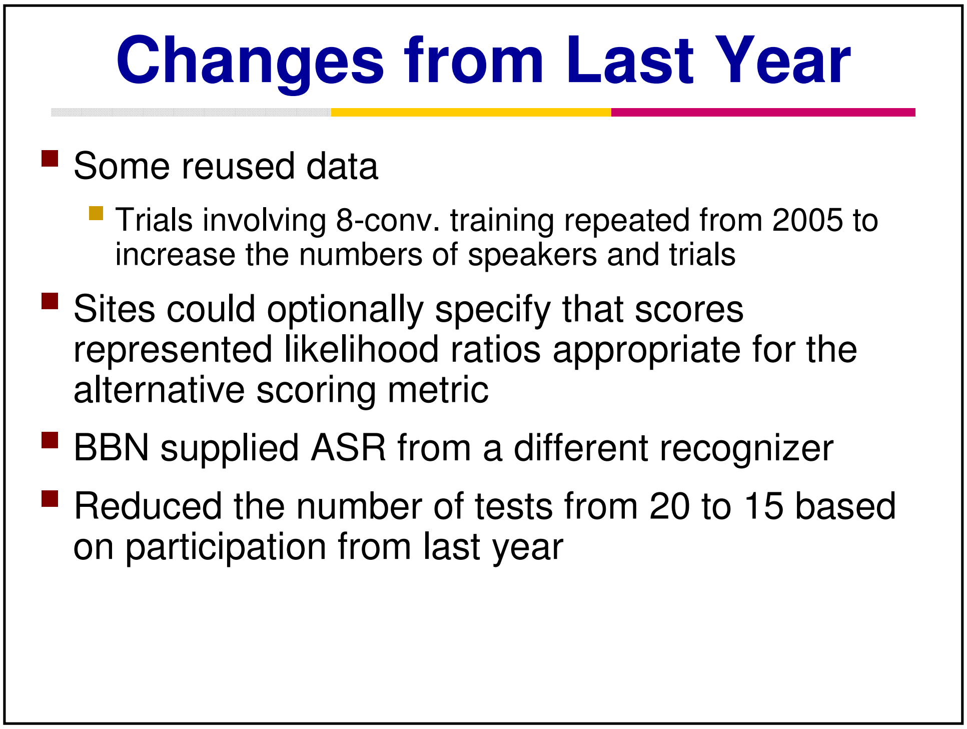


Data Problems

- Inappropriate trial lists
  - Due to data preparation algorithm bug
  - Solution: corrected trial lists and extended submission deadline
- Empty files
  - Models with very little or no speech (1.1%)
  - Test segments with very little or no speech (0.7%)
  - Solution: eliminated these models/test segments from scoring
- Mislabeled language
  - Data incorrectly labeled as English (1.1% model, 3% test segment)
  - Solution: corrected the key and rescored the common condition, other rescoring to be done after workshop
- Malfunction microphone
  - Mic5 of cross-channel data collected at LDC had a battery pack malfunction
  - Solution: will eliminate these test segments from scoring

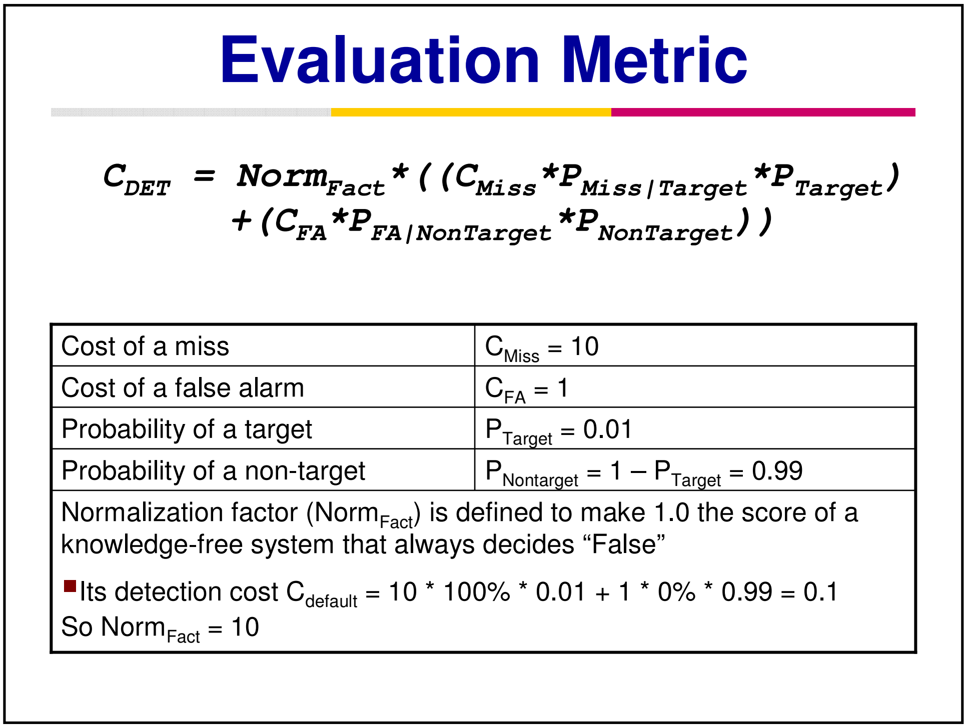
Introducing

Anton Filip Reynolds Feb 28, 2006



Changes from Last Year

- Some reused data
  - Trials involving 8-conv. training repeated from 2005 to increase the numbers of speakers and trials
- Sites could optionally specify that scores represented likelihood ratios appropriate for the alternative scoring metric
- BBN supplied ASR from a different recognizer
- Reduced the number of tests from 20 to 15 based on participation from last year



Evaluation Metric

$$C_{DET} = Norm_{Fact} * ((C_{Miss} * P_{Miss|Target} * P_{Target}) + (C_{FA} * P_{FA|NonTarget} * P_{NonTarget}))$$

Cost of a miss	$C_{Miss} = 10$
Cost of a false alarm	$C_{FA} = 1$
Probability of a target	$P_{Target} = 0.01$
Probability of a non-target	$P_{Nontarget} = 1 – P_{Target} = 0.99$
Normalization factor ($Norm_{Fact}$) is defined to make 1.0 the score of a knowledge-free system that always decides “False” ■Its detection cost $C_{default}$ = 10 * 100% * 0.01 + 1 * 0% * 0.99 = 0.1 So $Norm_{Fact}$ = 10	

Alternative Metric

$$C_{llr} = = 1 \,/\, (\ (2 * \log 2) * \\ (\ (\sum \log(1+1/lr)/NTT) + \\ (\sum \log(1+lr)/NNT)\ )\ )$$

lr	$P_{Data\|Target} / P_{Data\|NonTarget}$
Number of target trials	NTT
Number of non-target trials	NNT

- Reference
  - “Application-Independent Evaluation of Speaker Detection” in Computer Speech & Language, volume 20, issues 2-3, April-July 2006, pp. 230-275, by Niko Brummer and Johan du Preez

Performance Representation

- DET Plots
  - Shows the tradeoff of False Alarm and Miss error rates on a normal deviate scale
  - Actual decision points marked with a triangle, minimum detection point marked with a circle
  - Actual decision points often have a 95% confidence box around them
- Bar Graphs
  - Shows the contribution of two error types to C_{DET} values

Participants

- 36 submitting sites
  - Australia
  - Canada
  - China (6)
  - Czech Republic
  - Denmark
  - Finland
  - France (8)
  - Germany (2)
  - Israel
  - Italy
  - Lebanon
  - Singapore (2)
  - South Africa
  - Spain (2)
  - Switzerland
  - United Kingdom
  - United States (6)
- 90 systems
  - 10 unsupervised adaptation systems
  - 2 “mothballed” systems
- 283 test condition/system combinations

NIST
ODYSSEY
2006
Speaker Recognition Workshop
San Juan Puerto Rico
PUERTO RICO
2006
The Speaker and Language Recognition Workshop
26-27 June 2006
MIT Lincoln Laboratory
Universidad Politécnica Puerto Rico
QUT
BRNO UNIVERSITY OF TECHNOLOGY
FIT
FACULTY OF INFORMATION TECHNOLOGY
JOENSUUN YLIOPISTO
ATVS
NIST Speech Group
TNO
Institute for Infocomm Research
A*STAR
d-Ear
得意音通技术
d-Ear Technologies
CRIM
Laboratoire d'Informatique
Université d'Avignon
IRISA
Loquendo
VOCAL TECHNOLOGY AND SERVICES
TELECOM PARIS
école nationale supérieure des télécommunications
LRDE
EPITA
UT DCRSS
University of Balamand
ICSI
INTERNATIONAL COMPUTER SCIENCE INSTITUTE
SRI International
IMK
Fraunhofer Institute for Media Communication
IBM
Instituto de Investigación en Ingeniería de Aragón
Universidad de Zaragoza
UNIVERSITAS FRIBURGENSIS
Center for Speech Technology

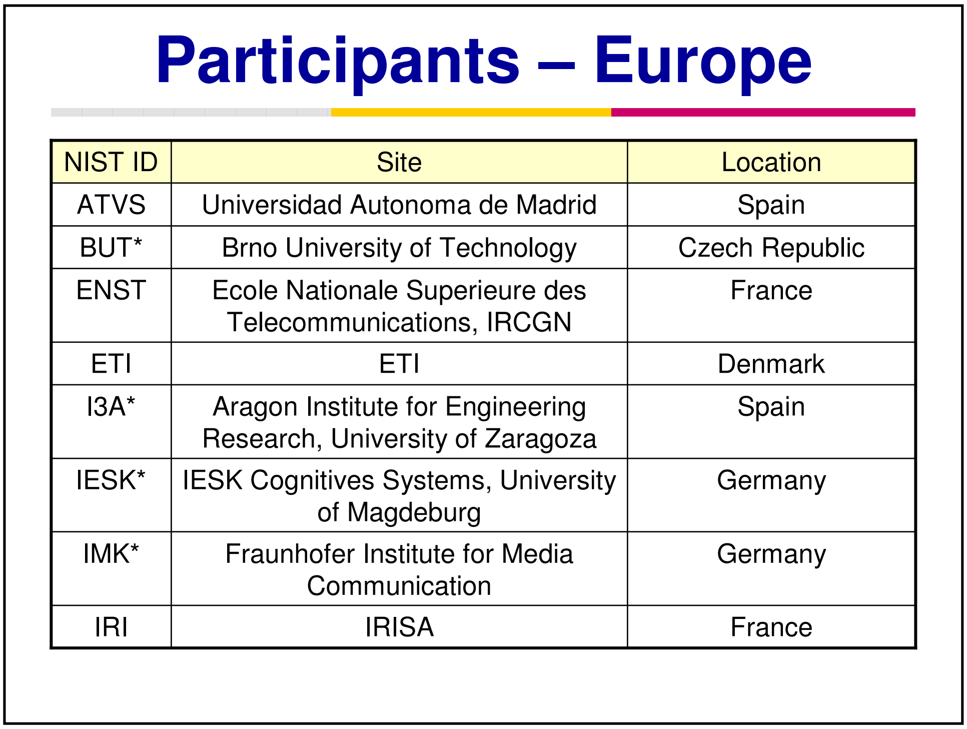
Participants – Asia

NIST ID	Site	Location
CST*	Center for Speech Technology, Tsinghua University	China
DEAR*	Beijing d-Ear Technologies Co. Ltd	China
FTRD*	France Telecom Research and Development Beijing	China
IIR	Institute for Infocomm Research	Singapore
IIRJ	Institute for Infocomm Research & University of Joensuu*	Singapore
IOA	Institute of Acoustics, Chinese Academy of Sciences	China
USTC	University of Science and Technology of China	China

* denotes first time participant

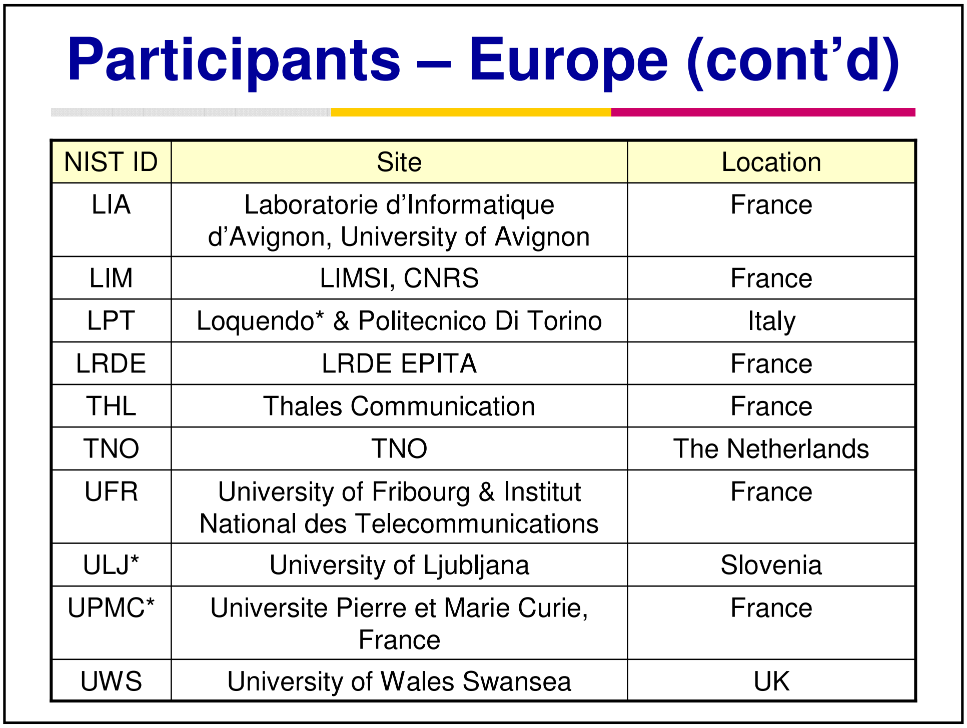
Participants – Australia

NIST ID	Site	Location
QNI	Queensland University of Technology & IBM	Australia



Participants – Europe

NIST ID	Site	Location
ATVS	Universidad Autonoma de Madrid	Spain
BUT*	Brno University of Technology	Czech Republic
ENST	Ecole Nationale Superieure des Telecommunications, IRCGN	France
ETI	ETI	Denmark
I3A*	Aragon Institute for Engineering Research, University of Zaragoza	Spain
IESK*	IESK Cognitives Systems, University of Magdeburg	Germany
IMK*	Fraunhofer Institute for Media Communication	Germany
IRI	IRISA	France



Participants – Europe (cont’d)

NIST ID	Site	Location
LIA	Laboratorie d’Informatique d’Avignon, University of Avignon	France
LIM	LIMSI, CNRS	France
LPT	Loquendo* & Politecnico Di Torino	Italy
LRDE	LRDE EPITA	France
THL	Thales Communication	France
TNO	TNO	The Netherlands
UFR	University of Fribourg & Institut National des Telecommunications	France
ULJ*	University of Ljubljana	Slovenia
UPMC*	Universite Pierre et Marie Curie, France	France
UWS	University of Wales Swansea	UK

Participants – Middle East

NIST ID	Site	Location
PRS*	Persay Ltd	Israel
UOB*	University of Balamand	Lebanon

Participants – N. America

NIST ID	Site	Location
CRIM	CRIM	Canada
CRSS*	Center for Robust Speech Systems, University of Texas at Dallas	USA
HEC	HEC, Air Force Research Laboratory	USA
ICSI	International Computer Science Institute	USA
MIT	MIT Lincoln Laboratory & IBM	USA
SRI	SRI International	USA

Evaluation Systems Collaborations

- MIT/IBM
- QUT/IBM
- SRI/ICSI
- IIR/University of Joensuu
- SDV/TNO/BUT/SUN
- ENST/LRDE/UFR/UPMC
- …
- There were numerous site collaborations in this year’s evaluation. This list is not exhaustive.

Outline

- Today
  - Evaluation Review
  - Evaluation Results
  - Mothballed Systems and History Plots
  - Language Effects
  - Summary
- Tomorrow
  - Cross-channel Results

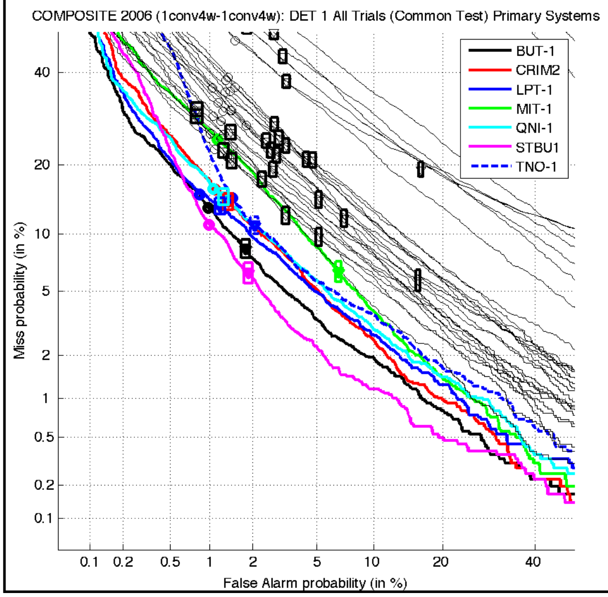
Core Test Condition

- 1conv4w-1conv4w
- Required of all participants
- Restrictions
  - None, but we removed many trials involving models or test segments in error

Targets			Non-Targets			
Trials (segs)	Speakers	Models	Trials (segs)	Model Speakers	Models	Segment Speakers
3612 (2410)	608	810	47836 (2456)	608	810	614

Core Test DET Plot *(all trials)*

1conv4w-1conv4w



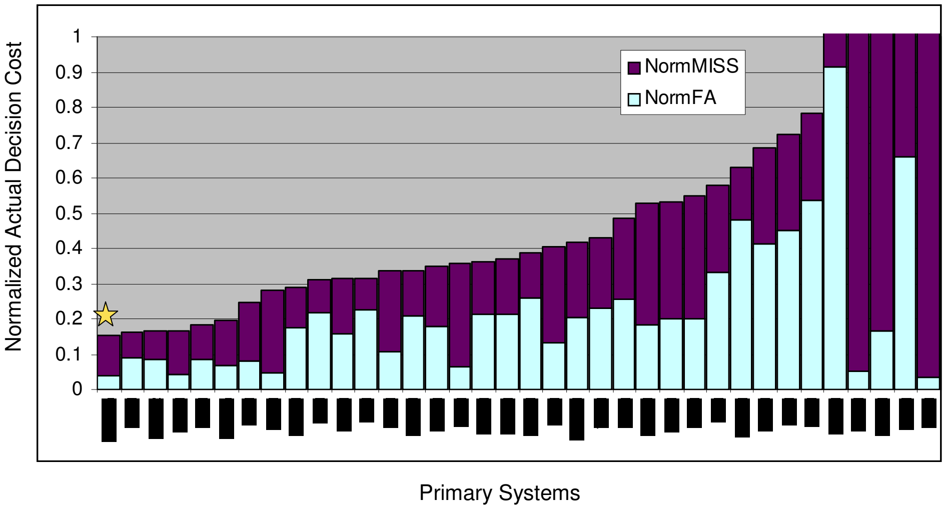
- The “required” test
- 35 participants submitted results, overwhelming MATLAB’s legend maximum
- Only several leading sites are identified
- Note that the best DET curve depends on which part of the plot one examines

Common Test Condition

- Subset of the core test condition with restrictions
  - English only data for training and test
  - Pooled gender
- Treated as the official evaluation outcome

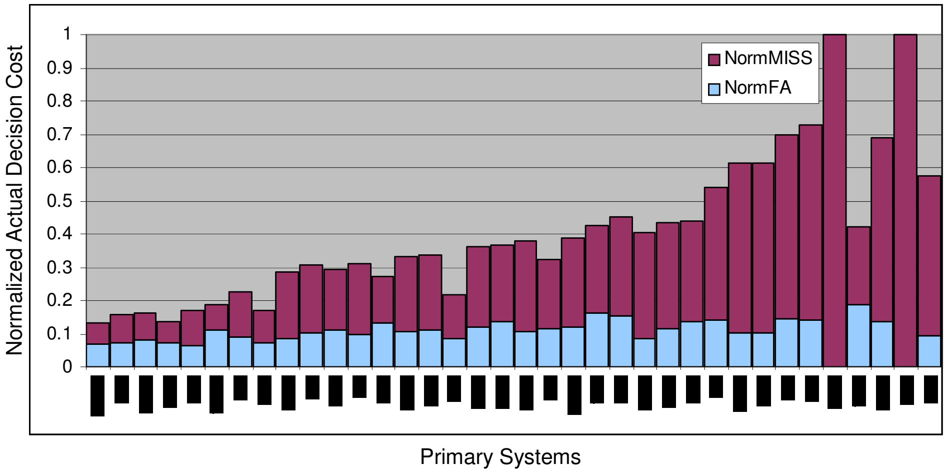
Targets			Non-Targets			
Trials (segs)	Speakers	Models	Trials (segs)	Model Speakers	Models	Segment Speakers
1854 (1691)	476	476	22159 (1862)	517	517	554

Common Condition
Actual Decision Costs



Common Condition

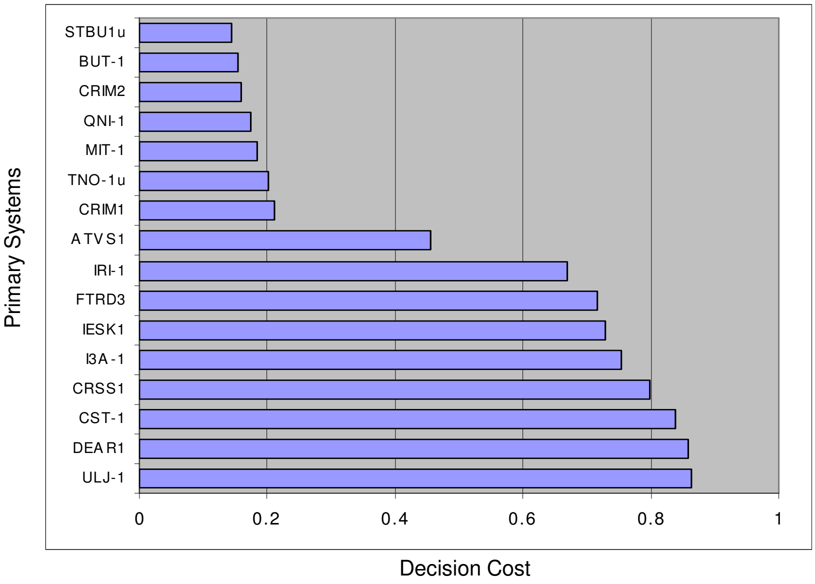
Minimum Decision Costs



- Systems ordered by increasing Actual Decision Cost (same order as in the previous slide)

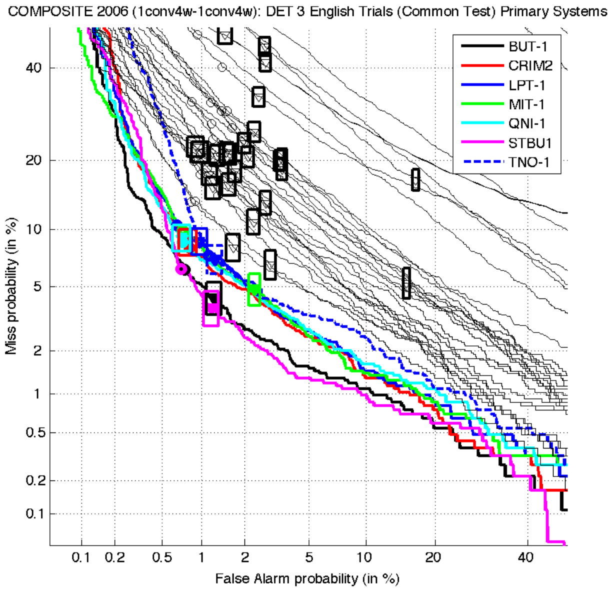
Common Condition

Cllr Scores



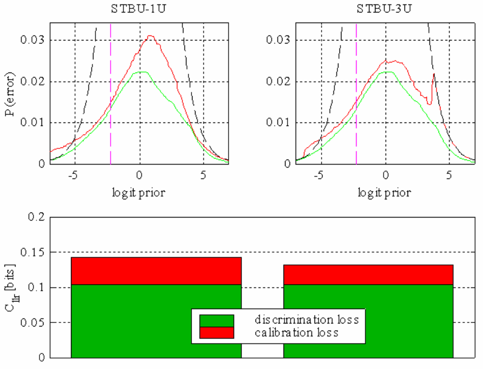
Common Condition DET Plot

1conv4w-1conv4w, English only trials



- Most systems exhibit improved performance from the “All Trials” condition, but system ordering shows little change

APE Curves



- Red curve is for system as submitted, green for optimally calibrated one
- Bar graph heights are proportional to areas under curves
- Equal error rate corresponds to curve maxima, CDet to value at -2.29
- Thanks to Niko Brummer, who will explain APE curves further

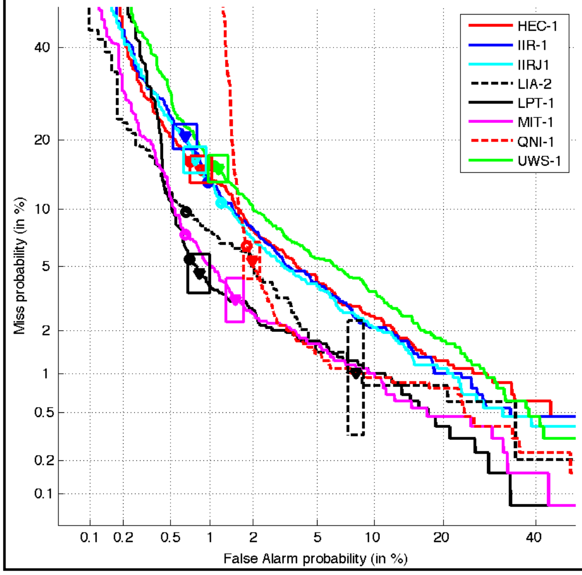
- Plot error rate against a range of llr values, where the error rate Pe is

  P1*Pmiss(-r) + (1-P1)*Pfa(-r), P1 = probability corresponding to the llr

3conv4w-1conv4w DET Plot

(English only trials)

COMPOSITE 2006 (3conv4w-1conv4w): DET 3 English Trials (Common Test) Primary Systems

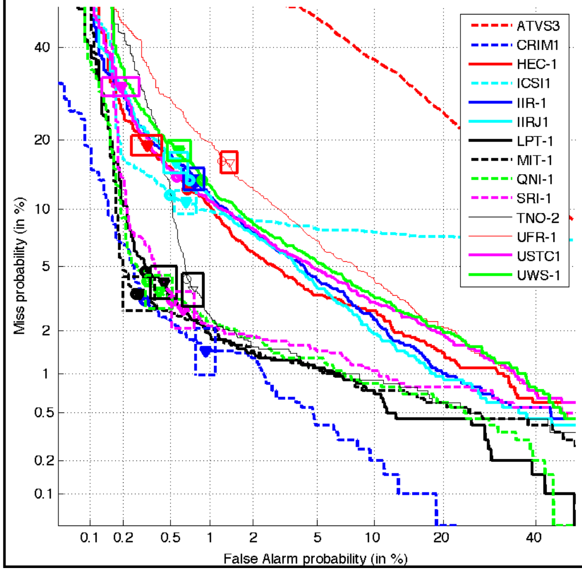


- 8 participants
- Four different systems contributed to the overall best DET

8conv4w-1conv4w DET Plot

(English only trials)

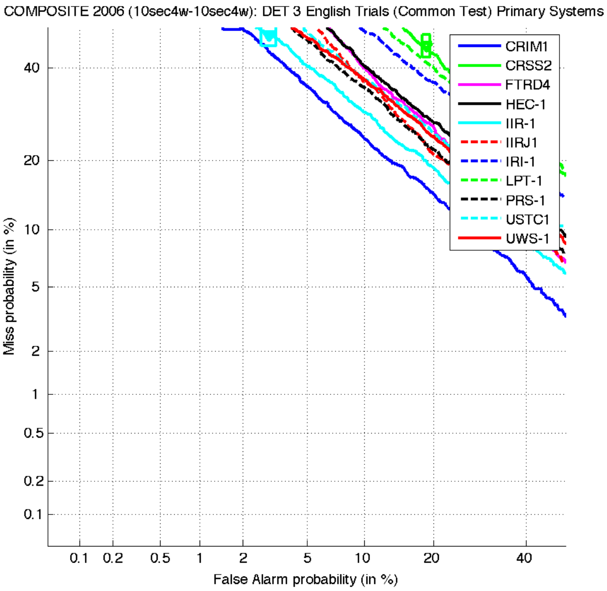
COMPOSITE 2006 (8conv4w-1conv4w): DET 3 English Trials (Common Test) Primary Systems



- 14 participants
- Condition with best overall performance (previously denoted extended data condition)
- CRIM and MIT contribute to best DET regions

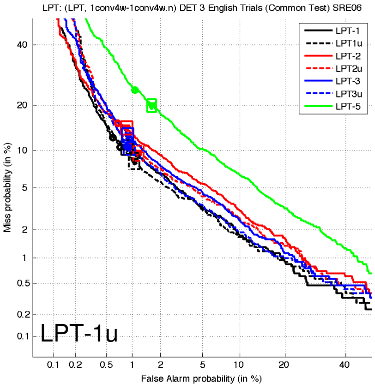
10sec4w-10sec4w DET Plot

(English only trials)

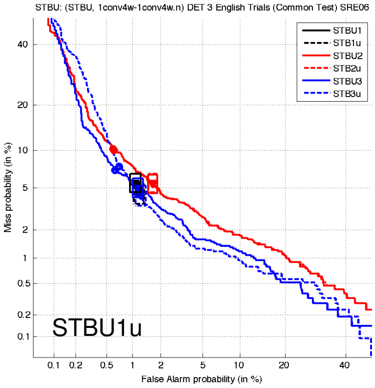


- 11 participants
- Difficult task, important for commercial applications
- Still plenty of room for improvement

Unsupervised Adaptation

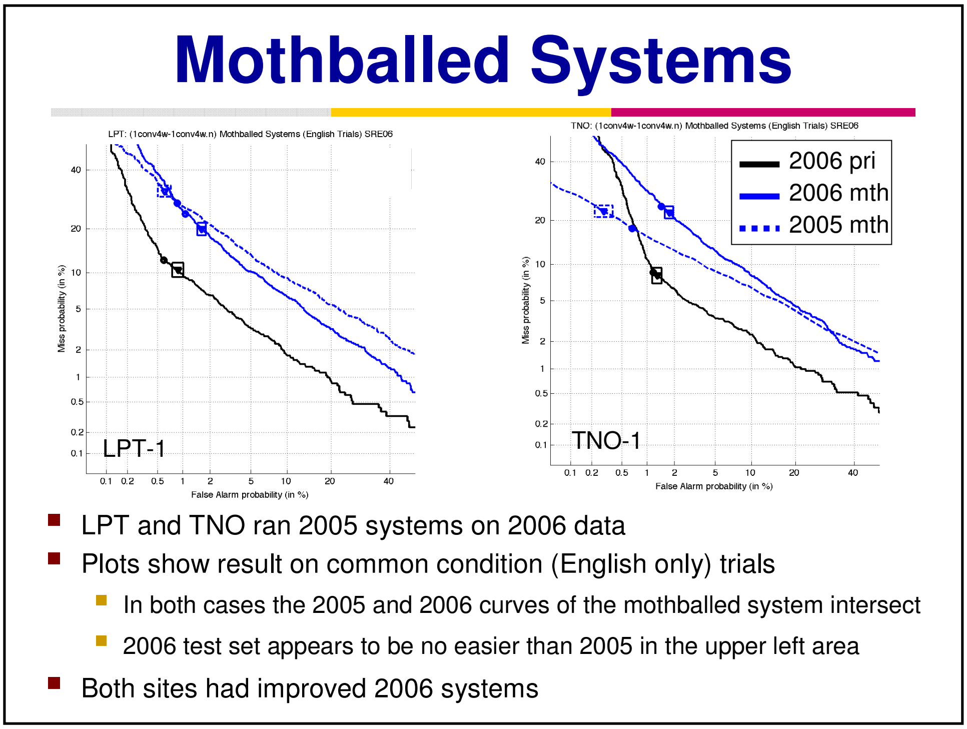


LPT-1u



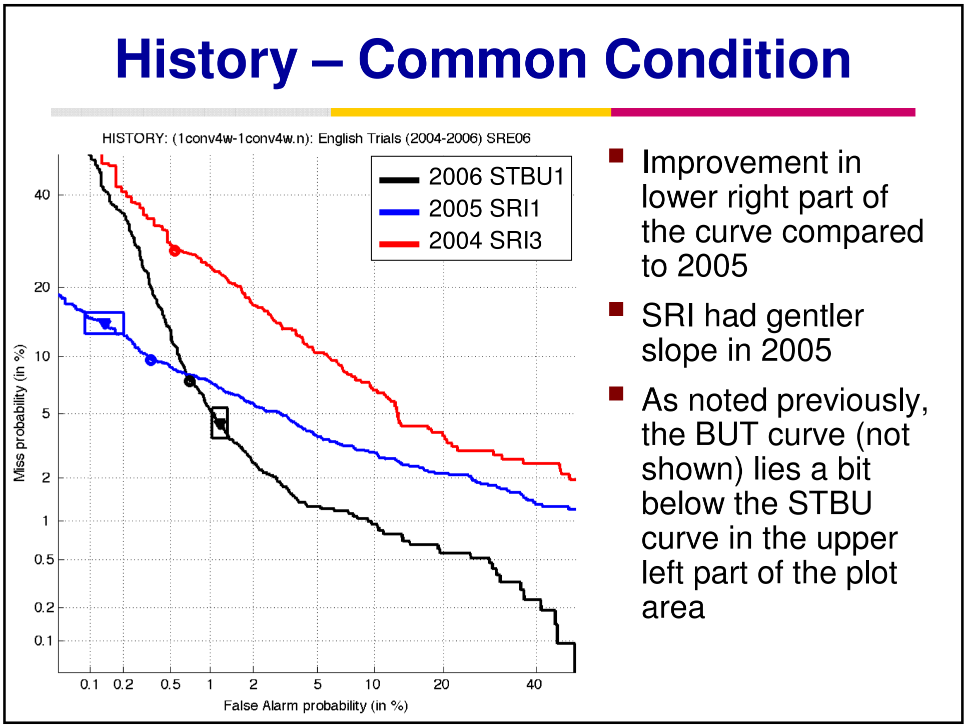
STBU1u

- LPT achieved some gains with unsupervised adaptation in the actual decision region
- Other sites had mixed results with gains only in some regions of the DET curve



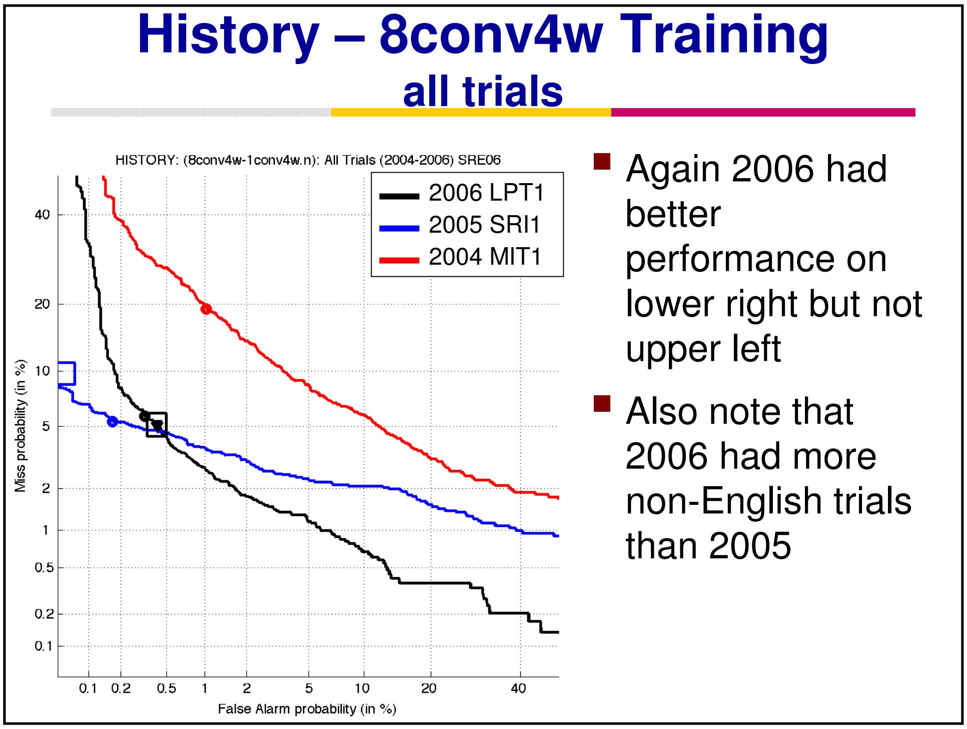
Mothballed Systems

- LPT and TNO ran 2005 systems on 2006 data
- Plots show result on common condition (English only) trials
  - In both cases the 2005 and 2006 curves of the mothballed system intersect
  - 2006 test set appears to be no easier than 2005 in the upper left area
- Both sites had improved 2006 systems



History – Common Condition

- Improvement in lower right part of the curve compared to 2005
- SRI had gentler slope in 2005
- As noted previously, the BUT curve (not shown) lies a bit below the STBU curve in the upper left part of the plot area

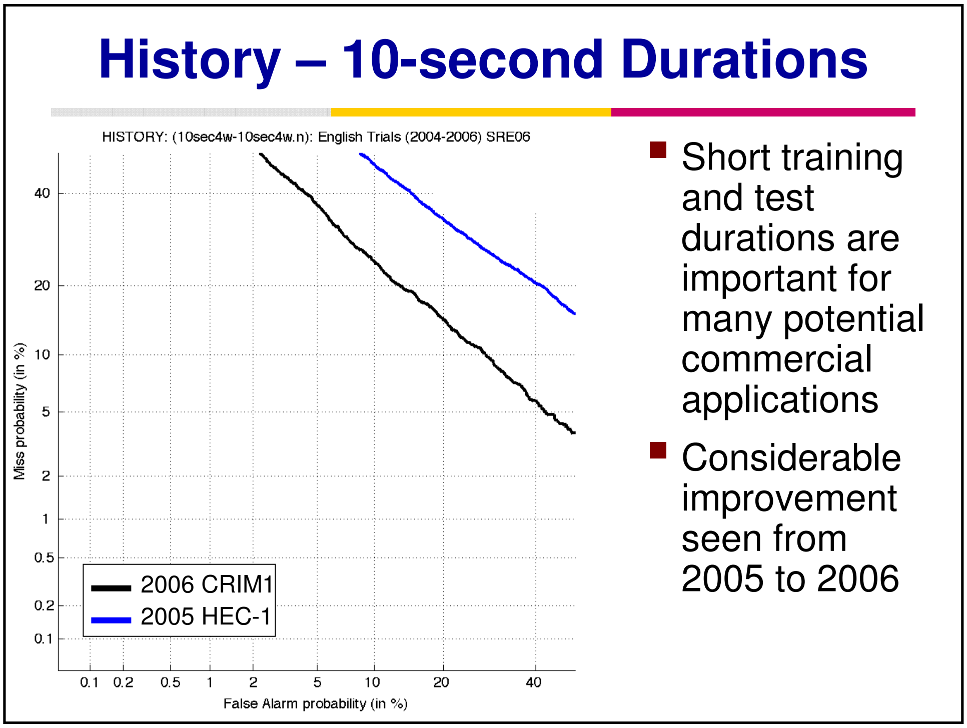


History – 8conv4w Training

all trials

HISTORY: (8conv4w-1conv4w.n): All Trials (2004-2006) SRE06

- Again 2006 had better performance on lower right but not upper left
- Also note that 2006 had more non-English trials than 2005

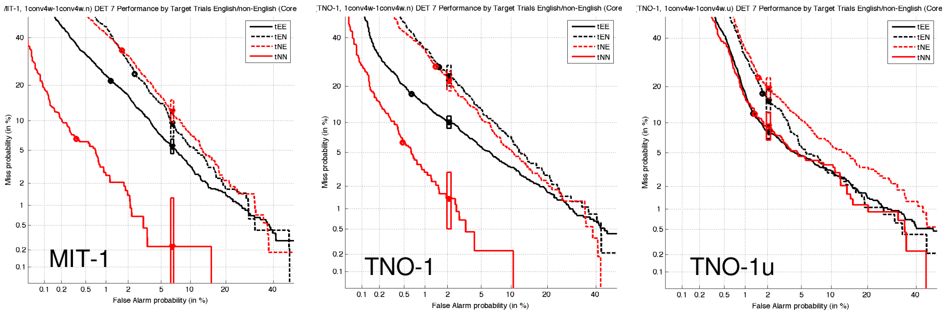


History – 10-second Durations

HISTORY: (10sec4w-10sec4w.n): English Trials (2004-2006) SRE06

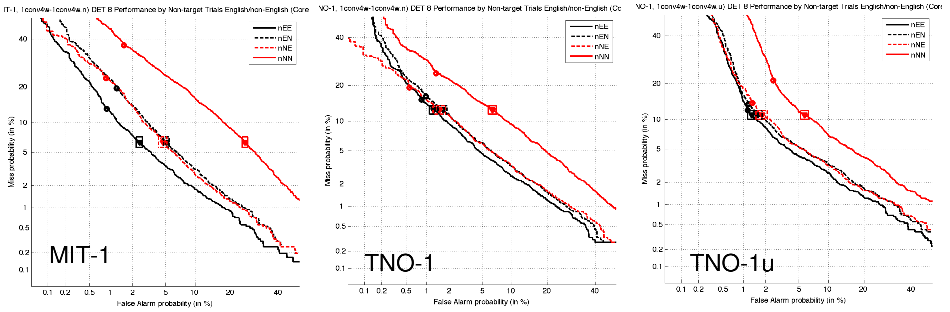
- Short training and test durations are important for many potential commercial applications
- Considerable improvement seen from 2005 to 2006

Language Effects – Target Trials



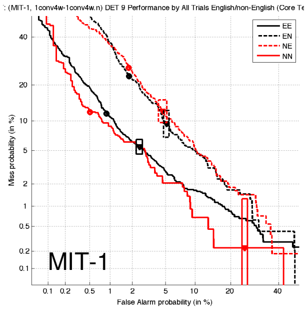
- Charts restrict target trials to four English/non-English train/test combinations, include all non-target trials
- Matched conditions give better performance, particularly non-English train and test
  - But unsupervised adaptation greatly limits this advantage, while helping a bit with English train, non-English test

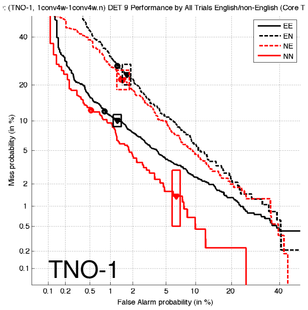
Language Effects – Non-Target Trials

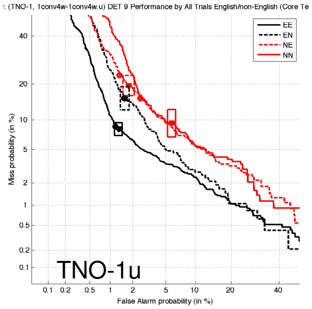


- Charts restrict non-target trials to four English/non-English train/test combinations, include all target trials
- Here the matched non-English train/test condition performs worst
  - MIT unusual in doing rather better on matched English train/test condition than mixed conditions

Language Effects – All Trials







- Charts restrict target and non-target trials to four English/non-English train/test combinations
- Putting both effects together, performance is best for matched conditions generally
- Unsupervised adaptation hurts matched non-English condition, but helps for English train, non-English test

Summary

- Record number of participants
- Increased size and complexity of the evaluation has overloaded the infrastructure and led to the data problems this year noted previously
  - More time and effort needed to audit the data
- New scoring metric works well for sites providing likelihood ratios
  - Should such scores be further encouraged, or required?
- Some notable performance improvements