NIST 2005 Speaker Recognition Evaluation Workshop

Analysis of Results

Alvin Martin & Mark Przybocki
NIST/ITL/IAD/SPEECH
www.nist.gov/speech/tests/spk

June 7-8th 2005
Montreal, Quebec, Canada

National Institute of Standards and Technology

Outline

- Evaluation Review
  - Task
  - Conditions and tests
  - Data and performance measures
- Participants
- Basic Results
  - Core test condition
  - Common evaluation condition
- History Plots
- Effects of Unsupervised Adaptation

Outline (cont'd)

- Results by Condition
  - Training and test durations
  - Single channel vs. summed channel
  - Cross channel microphones
- Other Performance Factors
  - Transmission type by target trials
  - Time between training and test
- Conclusion

Evaluation Review

- The NIST Year 2005 Speaker Recognition Evaluation operated under the protocols specified in the evaluation plan:
  www.nist.gov/speech/tests/spk/2005/SRE-05_evalplan-v6.pdf

- This evaluation plan defines the
  - Task,
  - Test conditions,
  - Rules,
  - Data, and
  - Scoring

Evaluation Task

- Speaker Detection Task
  - Given a model speaker, determine if that speaker is speaking in a given test segment
- Two modes of operation
  - Normal mode, no adaptation
  - Unsupervised adaptation mode
    - Trials for each model are ordered
    - May use test segment data to update the system for subsequent trials
    - Must do normal mode submission as well

Evaluation Conditions and Tests

Five Training Conditions	
8conv4w	8 conversations 4-wire (separate channels, both provided)
3conv4w	3 conversations 4-wire
1conv4w	1 conversation 4-wire
10sec4w	10 seconds of speech 4-wire
3conv2w	3 conversations 2-wire (summed channels)

Four Test Conditions	
1conv4w	1 conversations 4-wire
10sec4w	10 seconds of speech 4-wire
1conv2w	1 conversation 2-wire
1convmic	1 conversation (auxiliary microphone, in μ-law, on preferred channel) 4-wire

- Twenty Tests – all combinations of training and test conditions

Evaluation Rules *(normal mode)*

- Each decision is to be made independently
  - Based on the specified segment and the speaker model
  - Use of other segments or other models is NOT allowed
- Normalization over multiple test segments is NOT allowed
- Normalization over multiple target speakers is NOT allowed
- Use of evaluation data for impostor modeling is NOT allowed
- Use of manually produced transcripts or any other human interaction with the data is NOT allowed
- Knowledge of the model speaker gender IS allowed

Evaluation Data – MIXER Corpus

- Designed to support evaluation using test microphones different from training
- Includes large numbers of speakers with eight training conversations from a single handset
- Limited number of bilingual speakers
  - Arabic, Mandarin, Russian, or Spanish, along with English
  - Very few Non-English speakers and test segments used in 2005
- Speakers asked to
  - Record transmission and handset type for each call (often did not)
  - Speak in common non-English language when paired
- Includes multi-channel microphone data collected at the Linguistic Data Consortium and at Mississippi State University
  - Recorded on eight microphone channels in addition to the usual telephone channel

Evaluation Data - SRE Test Set

- 2nd NIST SRE using data from the MIXER corpus
  - All new speakers this year
  - Recordings made between Nov. 2003 and Oct. 2004
  - 8831 conversations involving 1857 speakers
- Whole conversations were distributed
  - Processed with an echo canceller
  - The category “10 second” segments contains variable length segments that include 7-13 seconds of actual speech
- ASR transcripts for almost all training and test data
  - Processed at BBN with a 1x real-time system
  - Estimated WER of 20%
  - English recognizer run on foreign language data

Protocol changes from last year

- Whole conversations released
  - To support dialog analysis and speaker turn detection
  - May allow better echo suppression
- BBN supplied output from a different (*English*) recognizer
  - 1x realtime @ about 20% WERR
  - Did not include ASR scores or Lattices
- Common condition included “cordless” handsets
- Reduced the number of tests
  - Dropped 16 conv. training and 30 sec. training and test
  - Added cross-channel microphone test data
- Sites limited to three system submissions per test (mothballed versions excluded)

The Performance Measures

- Detection Cost Function

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

- Parameter Values

Cost of a miss	$C_{Miss} = 10$
Cost of a false alarm	$C_{FalseAlarm} = 1$
Probability of a target	$P_{Target} = 0.01$
Probability of a non-target	$P_{NonTarget} = 1 - P_{Target} = 0.99$

- Normalization factor 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 $Norm_{Fact} = 10$

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

NIST ID.	SITE	City/Country
Asia		
INP	Israel National Police / Bar Ilan University	Jerusalem, Israel
PKU	Center for Information Science, Peking University	Beijing, China
PRS	Persay, LTD.	Tel Aviv, Israel
USTC	Speech Signal Processing Lab., Univ. of Science and Technology of china	Anhui Prov., China

Participants

NIST ID.	SITE	City/Country
Australia		
QUT	Speech Research Laboratory, Queensland University of Technology	Brisbane, Australia
Africa		
SDV	Spescom Datavoice	Stellenbosch, South Africa

Participants

NIST ID.	SITE	City/Country
Europe		
ATVS	Universidad Autonoma de Madrid	Madrid, Spain
DIVA	Department of Informatics Document Image Voice Analysis University of Fribourg	Fribourg, Switzerland
ENST	Ecole Nationale Superieure des Telecommunications	Paris, France
ETI	ETI	Denmark
IRI (IRISA)	Institut de Recherche en Informatique et Systemes Aleatoires	Rennes, France
IRIT	Institut Recherche Informatique Toulouse	Toulouse, France
LIA	Laboratoire Informatique d'Avignon	Avignon, France

Participants

NIST ID.	SITE	City/Country
Europe (cont'd)		
LIM (LIMSI-CNRS)	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur - Le Centre National de la Recherche Scientifique	Orsay, France
PDT	Politecnico di Torino	Torino, Italy
THL	THALES Communication France	Colombes, France
TNO	TNO - Human Factors Nederlandse Organisatie voor Toegepast-Natuurwetenschappelijk Onderzoek	Soesterberg, The Netherlands
UWS	School of Engineering, University of Wales Swansea	Swansea, Wales, UK

Participants

NIST ID.	SITE	City/Country *(or state)*
North America		
CRIM	Centre De Recherche Informatique De Montreal	Montreal, Quebec, Canada
GFS	Golden Finger Systems	Arcadia, CA
HEC	Air Force Research Laboratory/HEC	WPAFB, OH
IBM	IBM	Yorktown Heights, NY
ICSI	International Computer Science Institute	Berkeley, CA
MITL	MIT-Lincoln Laboratory / Oregon Graduate Institute	Lexington, MA
R64	R-64 research group	Fort Meade, MD
SRI	SRI International	Menlo Park, CA
UNB	University of New Brunswick	Fredericton, N.B. Canada

Site	Training																			
	8-conv (4w)				3-conv (4w)				1-conv (4w)				10 sec (4w)				3-conv (2w)			
	Test				Test				Test				Test				Test			
	10s	1s	1c	M	10s	1s	1c	M	10s	1s	1c	M	10s	1s	1c	M	10s	1s	1c	M
ATVS		3							2	2			1							
CRIM										3										
DIVA		3								3										
ENST										3										
ETI										1										
GFS										1	1									
HEC	1	1	1		1	1	1		1	1	1		1	1	1		1	1	1	
IBM										2										
ICSI		1								1										
IRI									1	3	2		1	1						
IRIT										2										
INP										2										
LIA					2	3			1	3										
LIMSI										2										
MITLL		4		2		4	1	2		3		2						3	2	
Persay									3	3	3		3							
PDT	1	1	1		1	1	1		1	1	1		1	1	1		1	1	1	
PKU		1								1		1	1						1	
QUT		2				2				2										
R64	1	1	1	1	1	1	1	1	1	1	1	1								
SDV										6										
SRI		3								3										
THL										2										
TNO	3	4								4		3	3							
UNB									1	2			1							
USTC	1	2			1	3			2	3			2	2				1	1	
UWS	3	2	2	2	3	2	2	2	3	3	2	2	3	3	2	2	3	2	2	2

27 Sites submitted 69 systems = 241 condition/system combination scoring

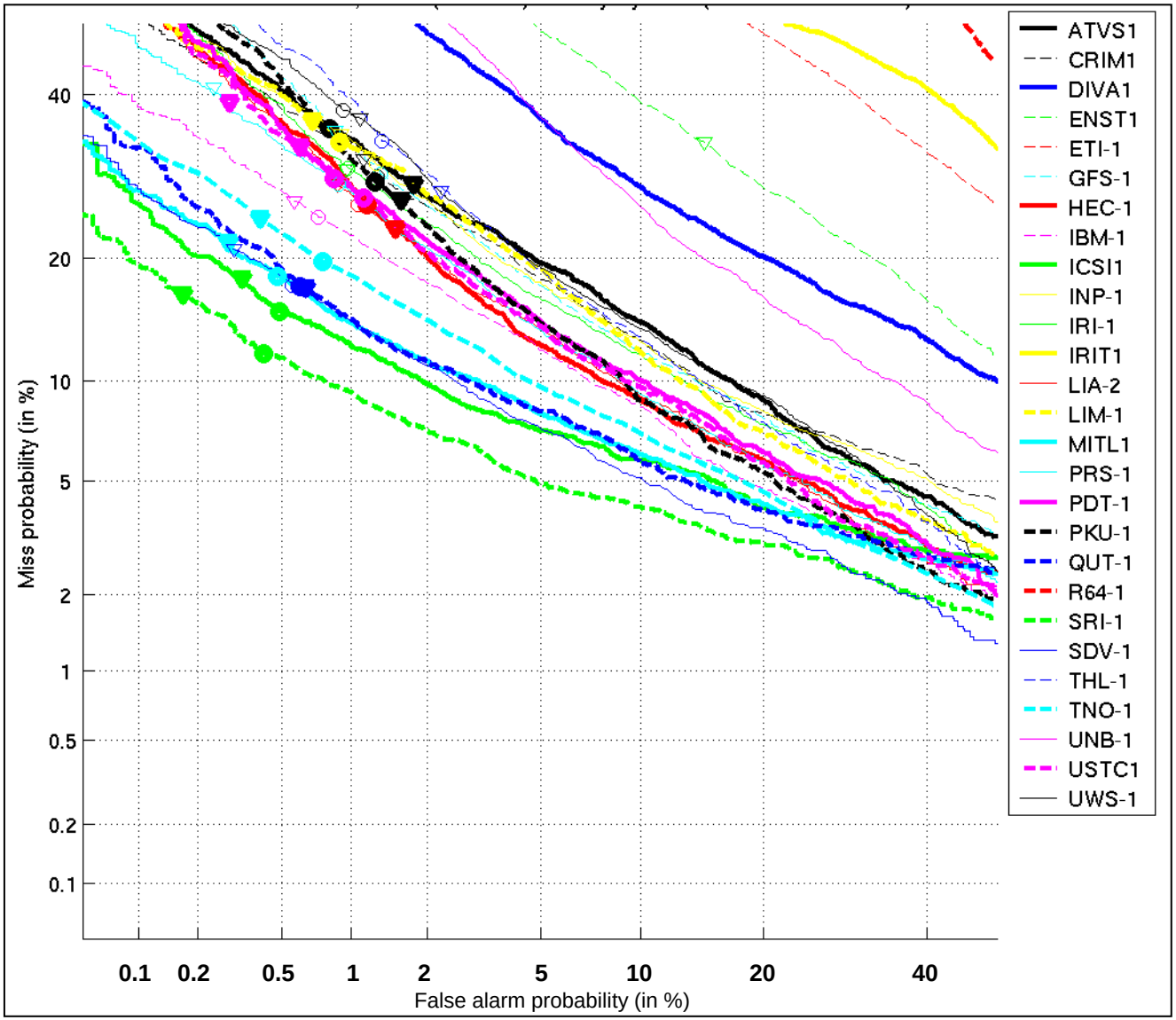
Results

- Core Test Condition
  - 1conv4w-1conv4w (no adaptation)
  - Required of all participants
- Restrictions
  - None, but we removed a handful of trials involving models or test segments in error

Targets			Non-Targets			
Trials (segs)	Speakers	Models	Trials (segs)	Model Speakers	Models	Segment Speakers
2771 (2133)	366	432	28,472 (2206)	564	634	373

Core Test DET Plot *(all trials)*

1conv4w-1conv4w



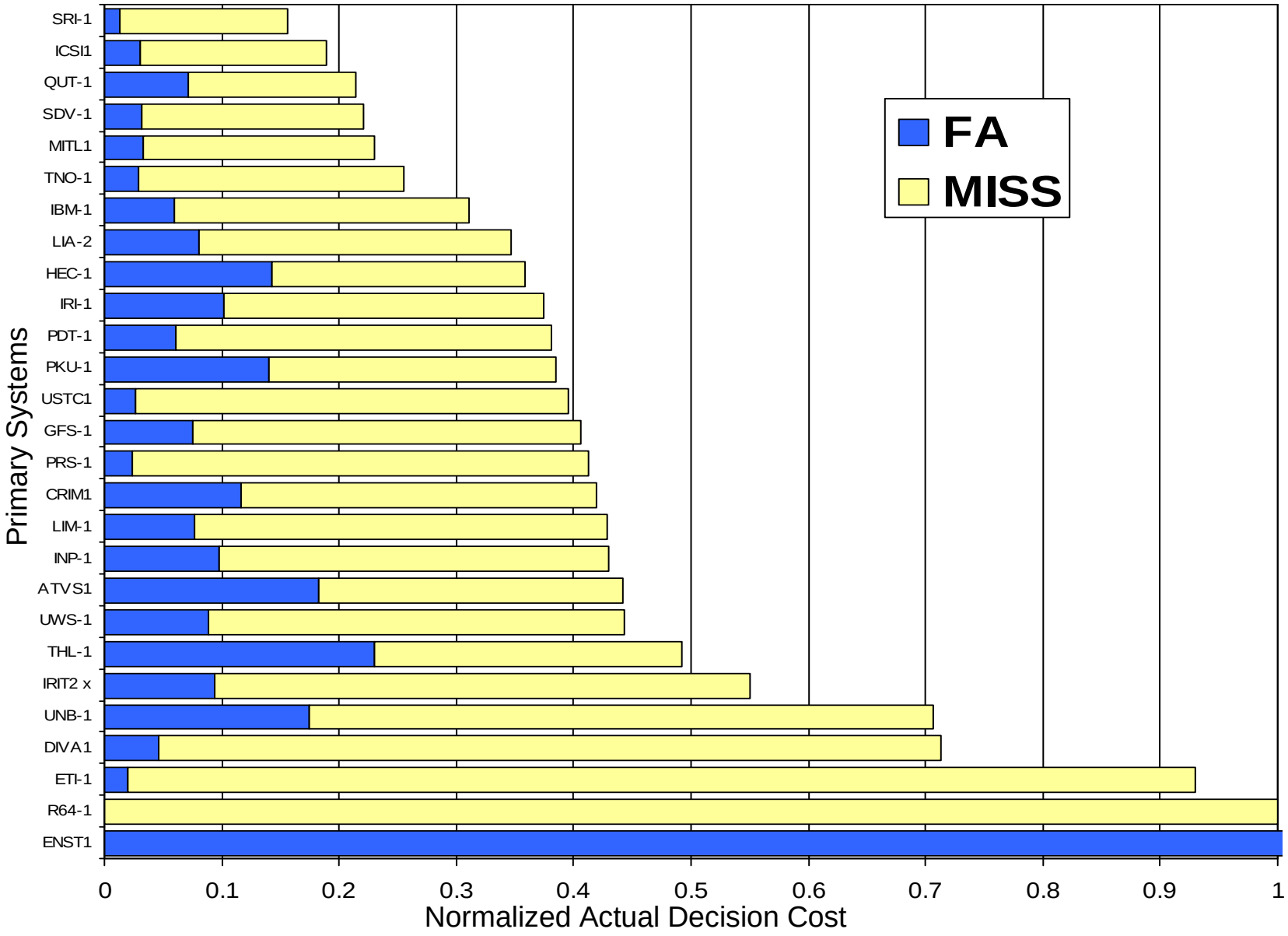
- The one "required" test
- 27 participants submitted results
- SRI and ICSI worked cooperatively
- THALES and LIA worked cooperatively using the LIA open-source system

Common Evaluation Condition

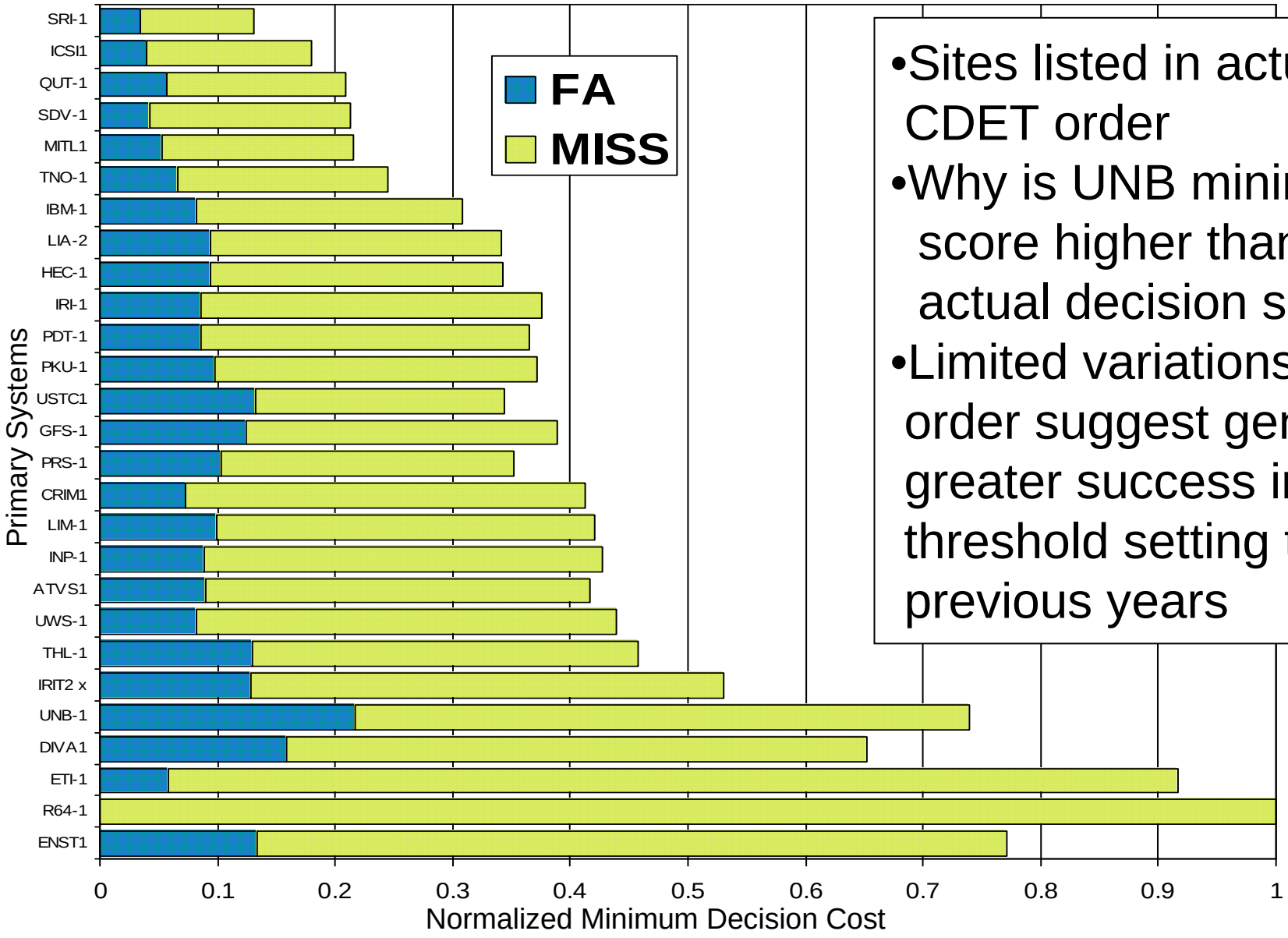
- Treated as the official evaluation outcome
- Restrictions on the core test *(1conv4w-1conv4w)*
  - English only data for training and test
  - Pooled across gender
  - Training and test involve:
    - A handheld microphone instrument

Targets			Non-Targets			
Trials (segs)	**Speakers**	**Models**	**Trials (segs)**	**Model Speakers**	**Models**	**Segment Speakers**
2148 (1777)	**314**	**344**	**18,782 (1957)**	**472**	**505**	**359**

Common Condition Actual Decision Costs

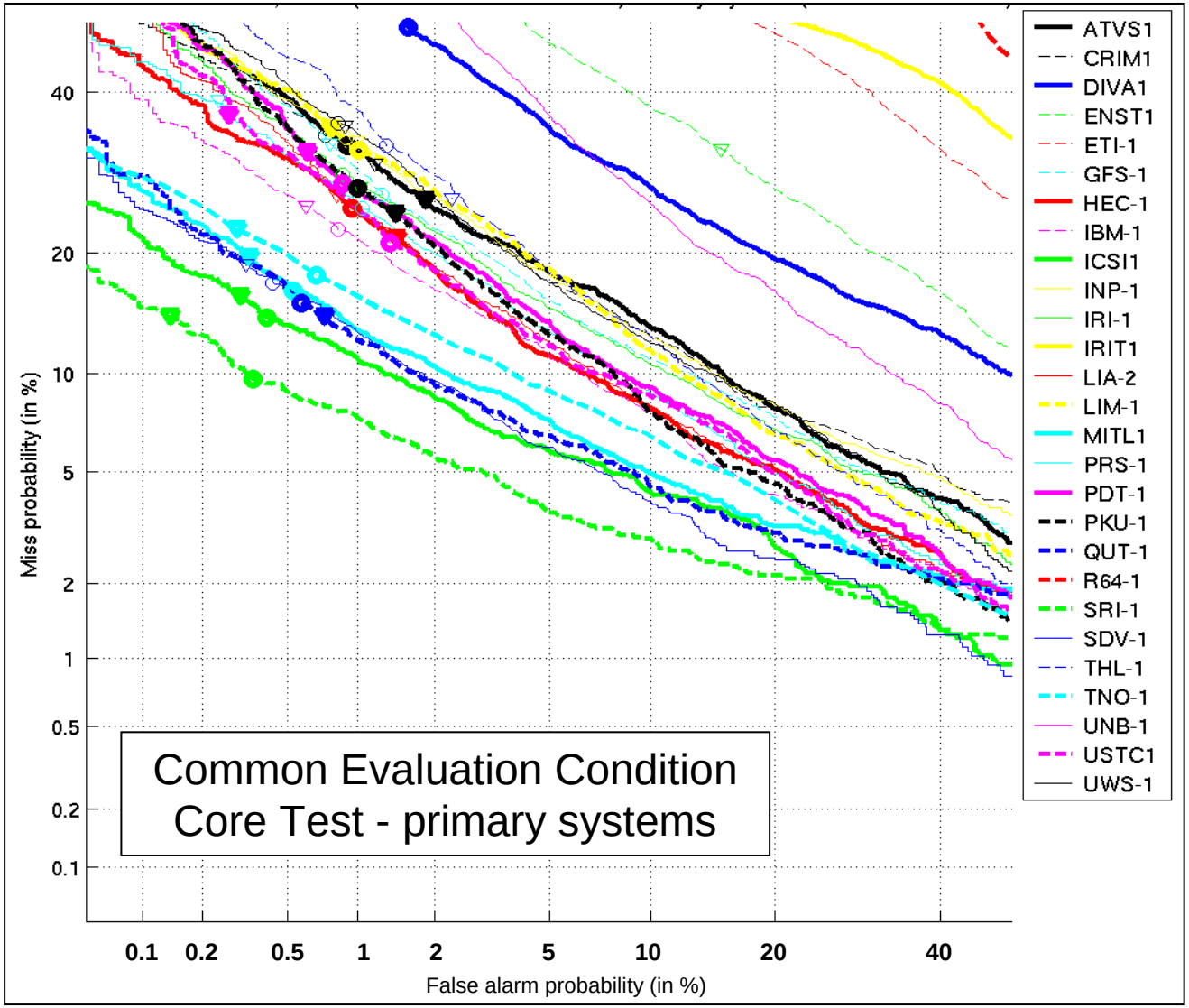


Common Condition Minimum Decision Costs



- Sites listed in actual CDET order
- Why is UNB minimum score higher than its actual decision score?
- Limited variations in order suggest generally greater success in threshold setting than in previous years

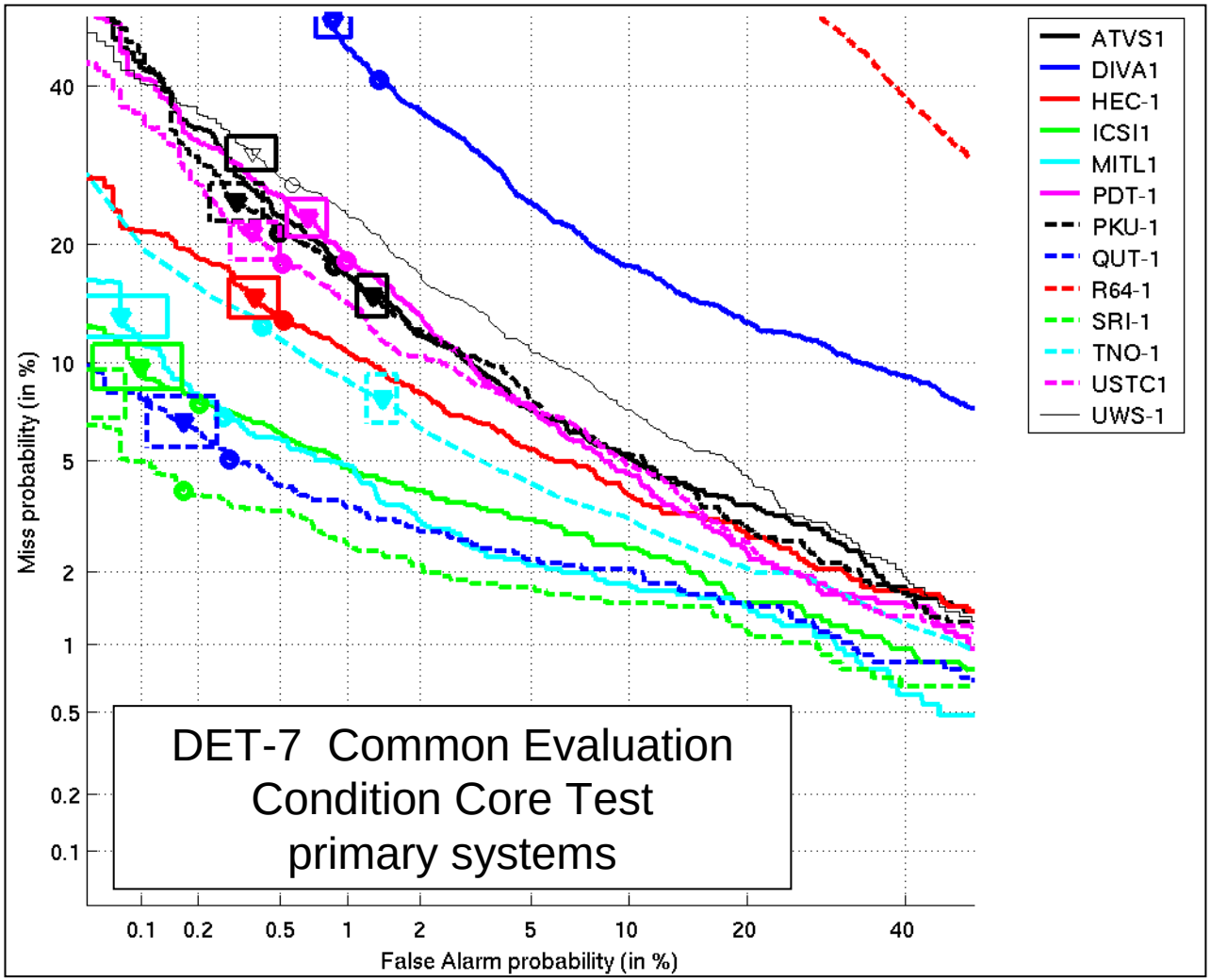
Common Condition DET Plot
1conv4w-1conv4w



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

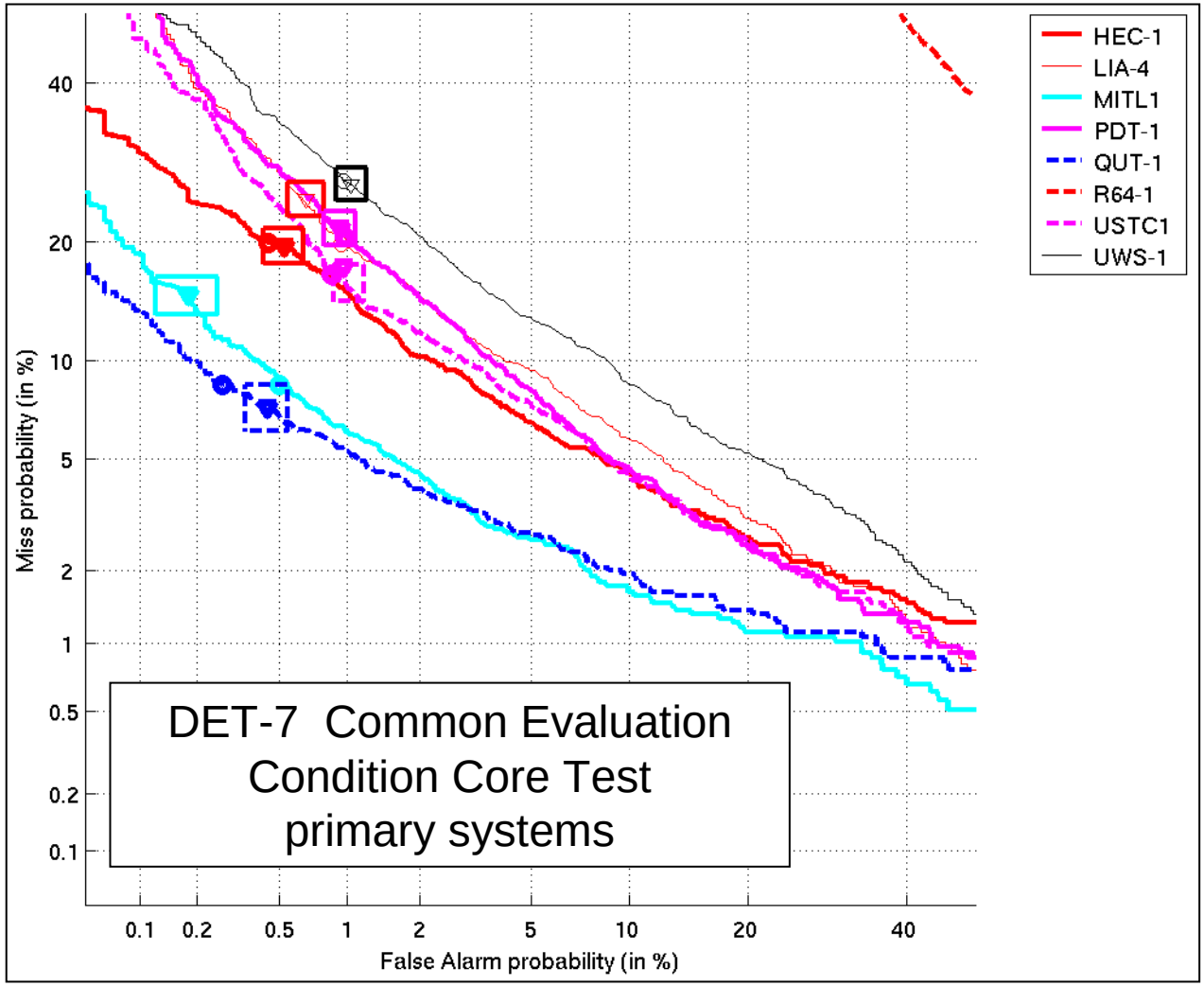
Common Condition DET Plot

8conv4w-1conv4w



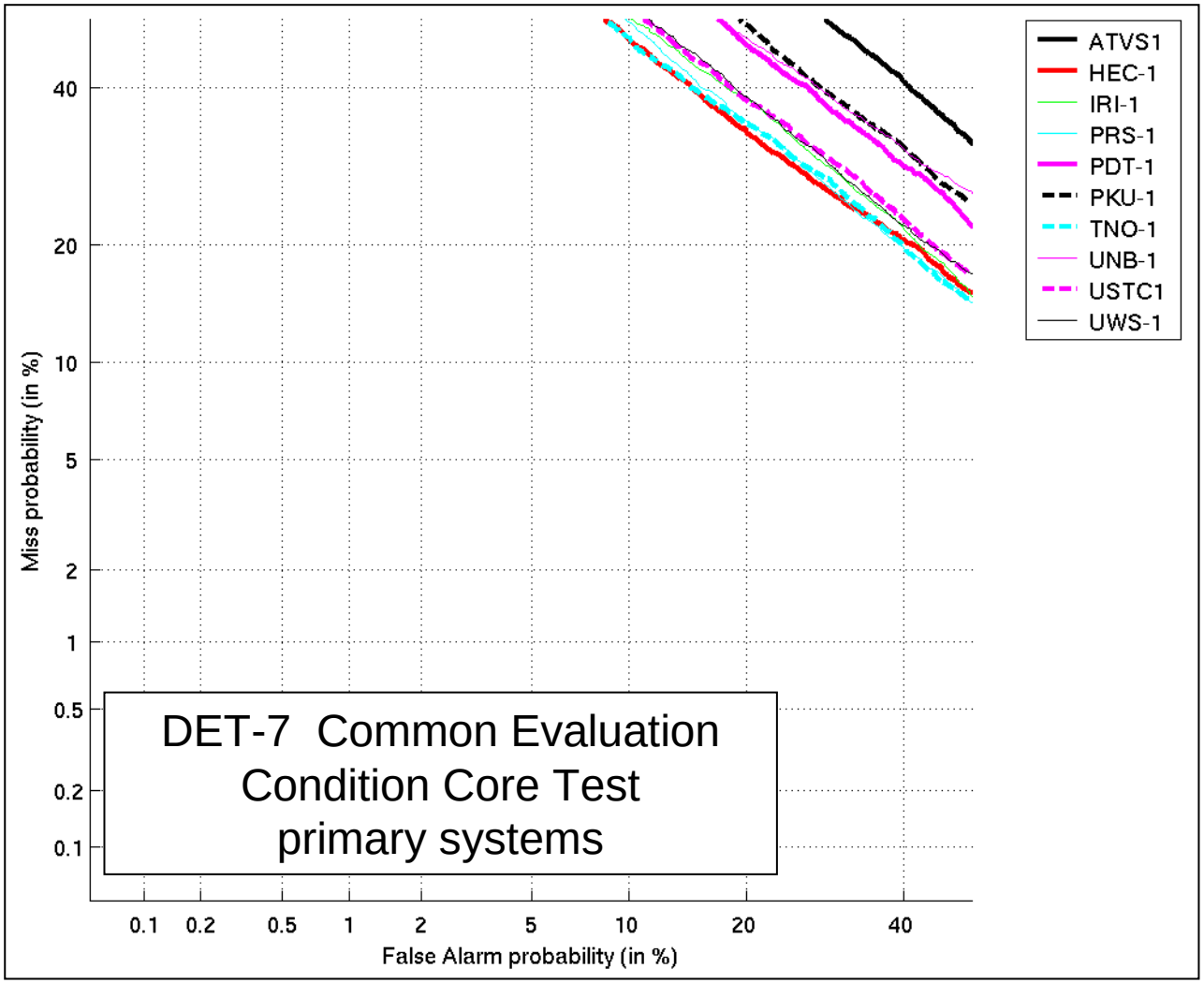
- 13 participants
- Performance on trials with the same restrictions as the common condition

Common Condition DET Plot
3conv4w-1conv4w



- 8 participants
- Performance on trials with the same restrictions as the common condition

Common Condition DET Plot
10sec4w-10sec4w



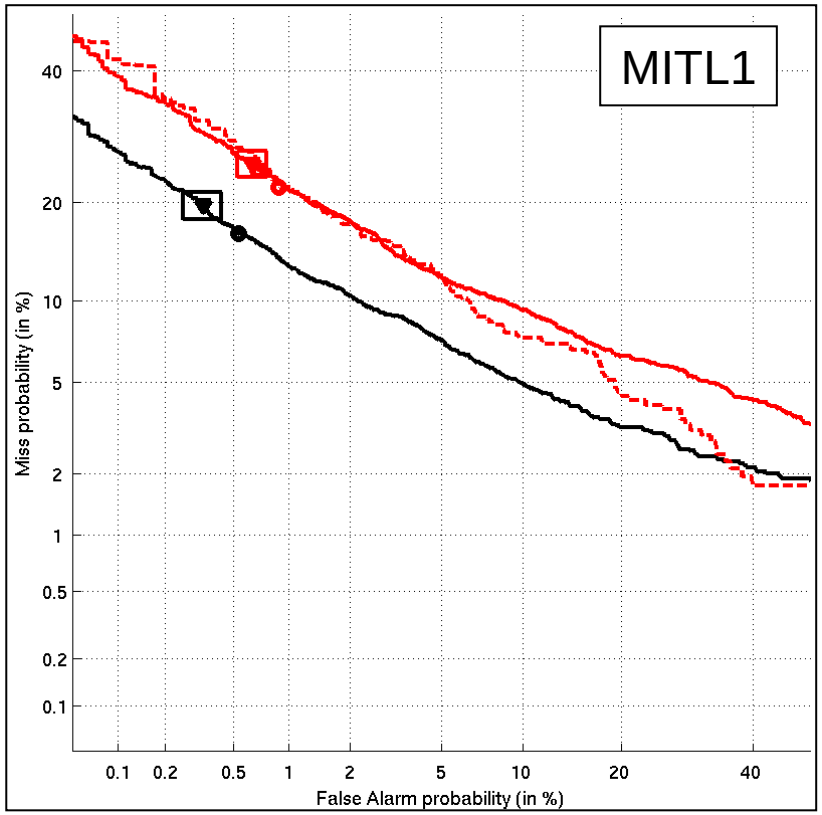
- 10 participants
- Performance on trials with the same restrictions as the common condition

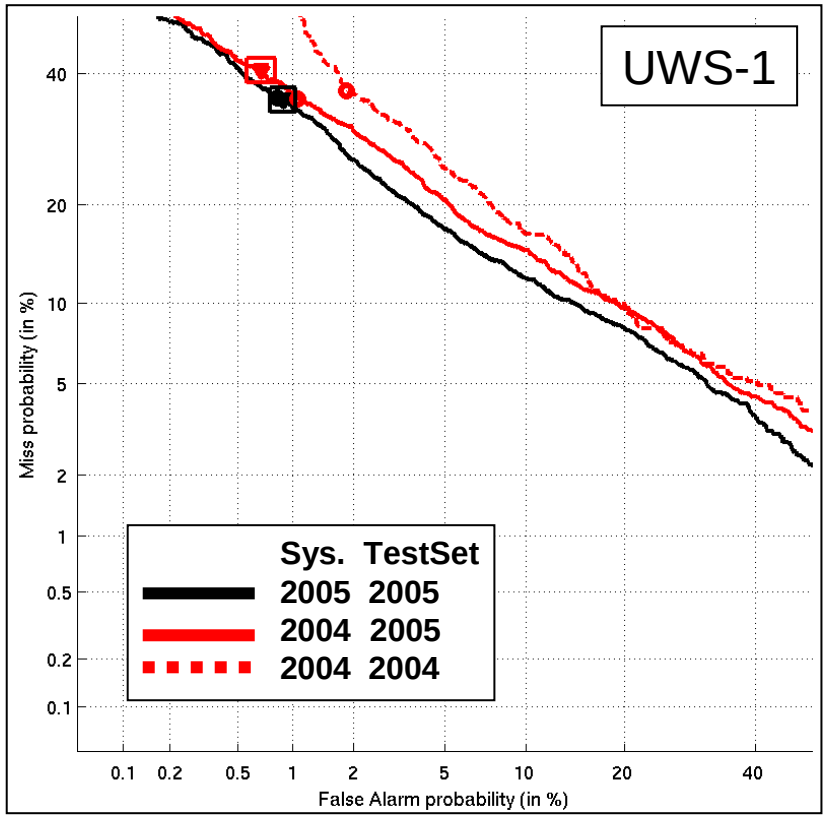
Mothballed Systems

- A few sites ran mothballed 2004 evaluation systems on this year's evaluation data
  - MITLL
  - SDV
  - University of Wales Swansea
- Comparison of performance by the same system on the 2004 and 2005 test sets indicates differences in the relative difficulty of the test sets.
- Equivalent performance to last year suggests test set comparability
- We show performance restricted to the "common condition"

Mothballed Systems

1conv4w-1conv4w for the common condition

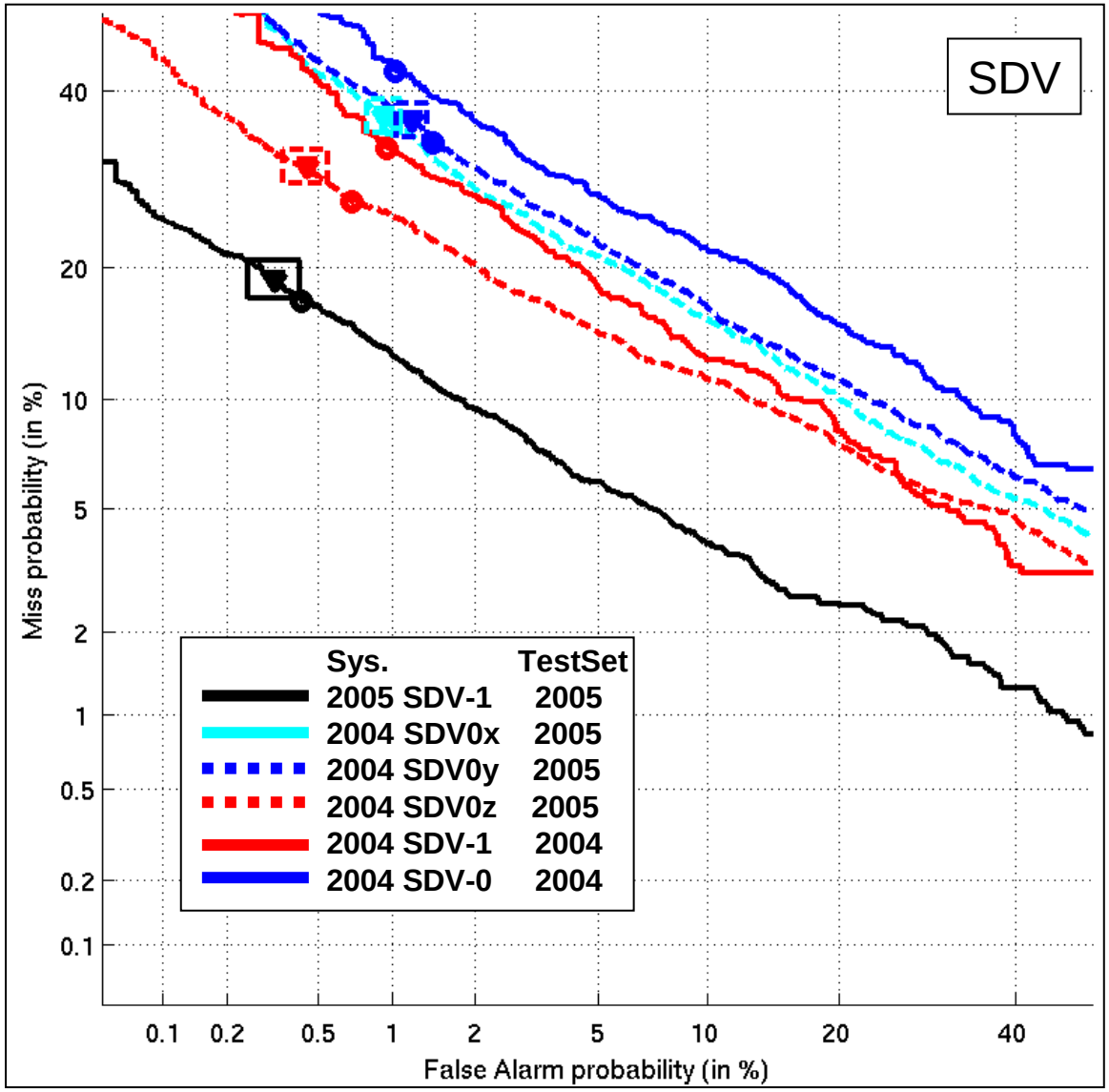




- Distance between the red lines indicate differences in test set difficulty
- Distance from black line indicates real system improvement

SDV Mothballed Systems

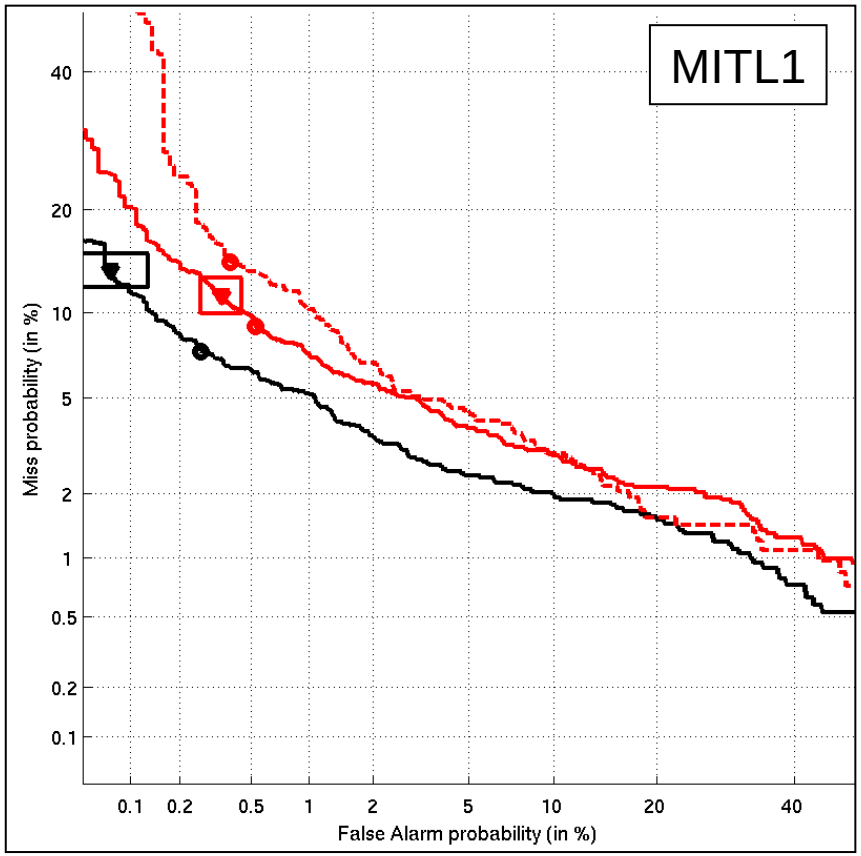
1conv4w-1conv4w for the common condition

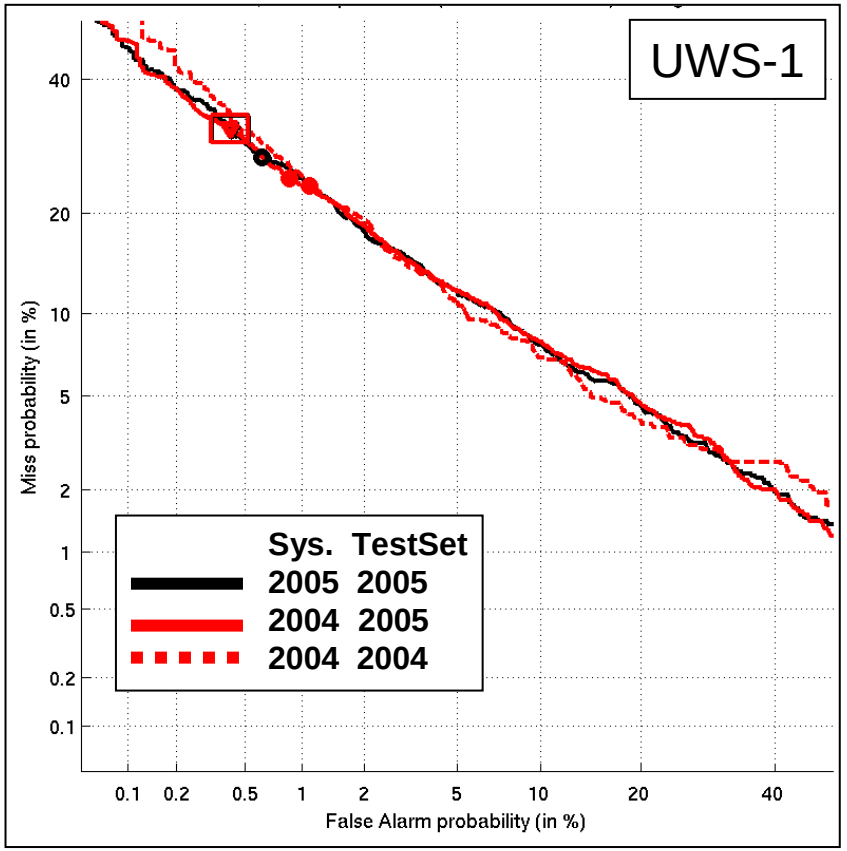


- Two mothballed systems
- SDV found the 2005 test set a bit easier, but produced real improvement
- SDV0x adds cross-channel squelch to SDV0y
  - This gave minor improvement

Mothballed Systems

8conv4w-1conv4w for the common condition

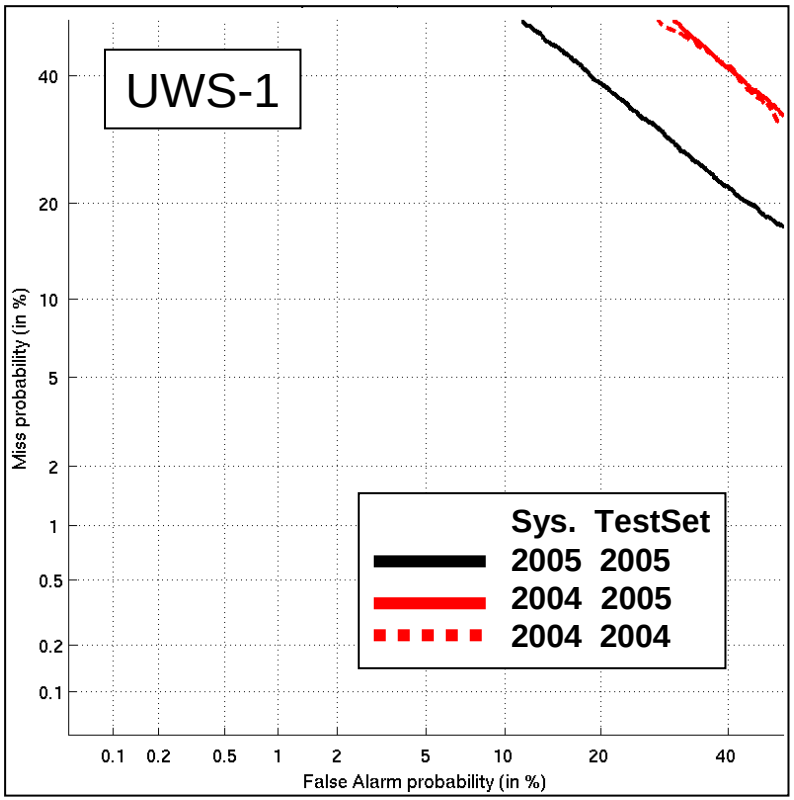




- MITLL found the 2005 data perhaps easier, but achieved real system improvement
- UWS concentrated on system speed up and score fusion, which helped performance with limited training data

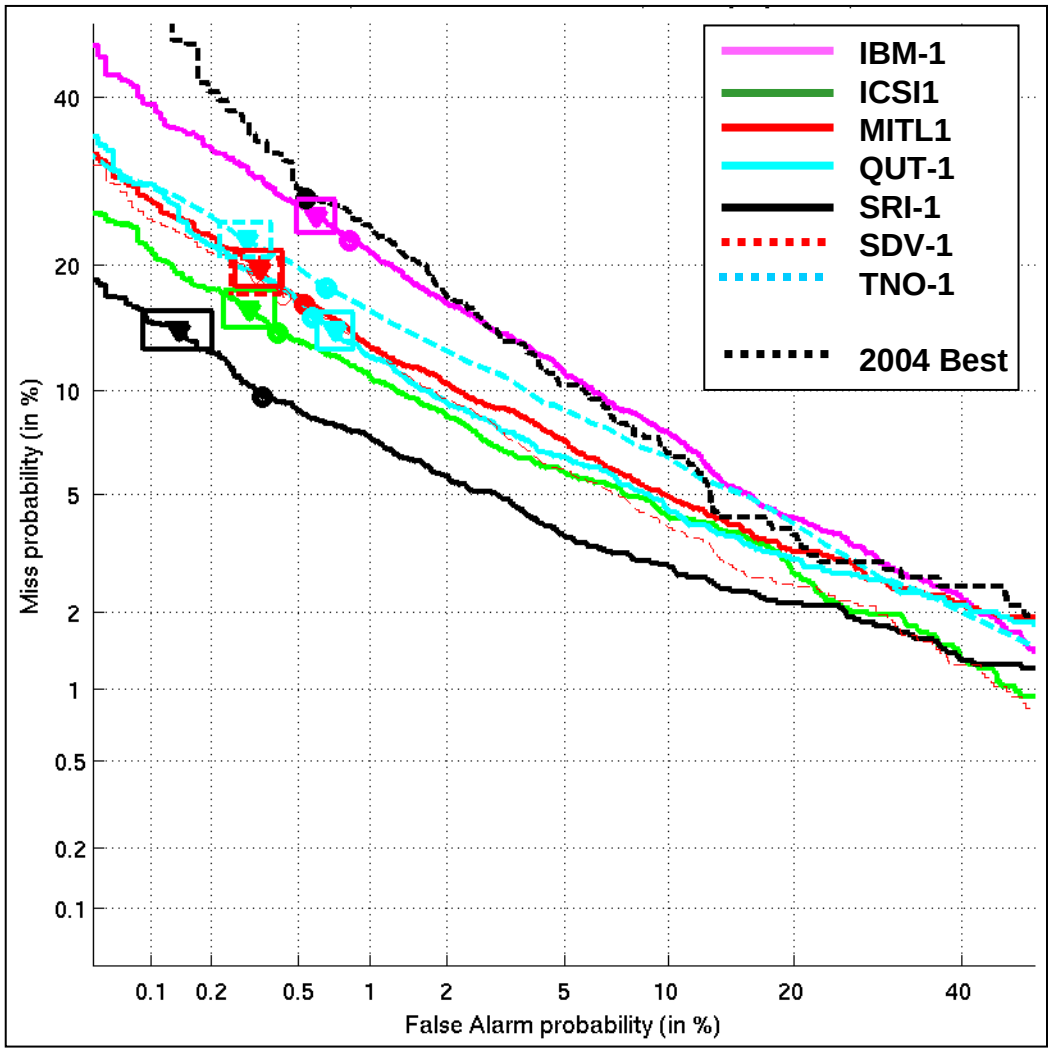
Mothballed Systems

10sec4w-10sec4w for the common condition



- UWS achieved real improvement for this difficult condition

Seven Sites Show Improvement Over 2004 Best Performance

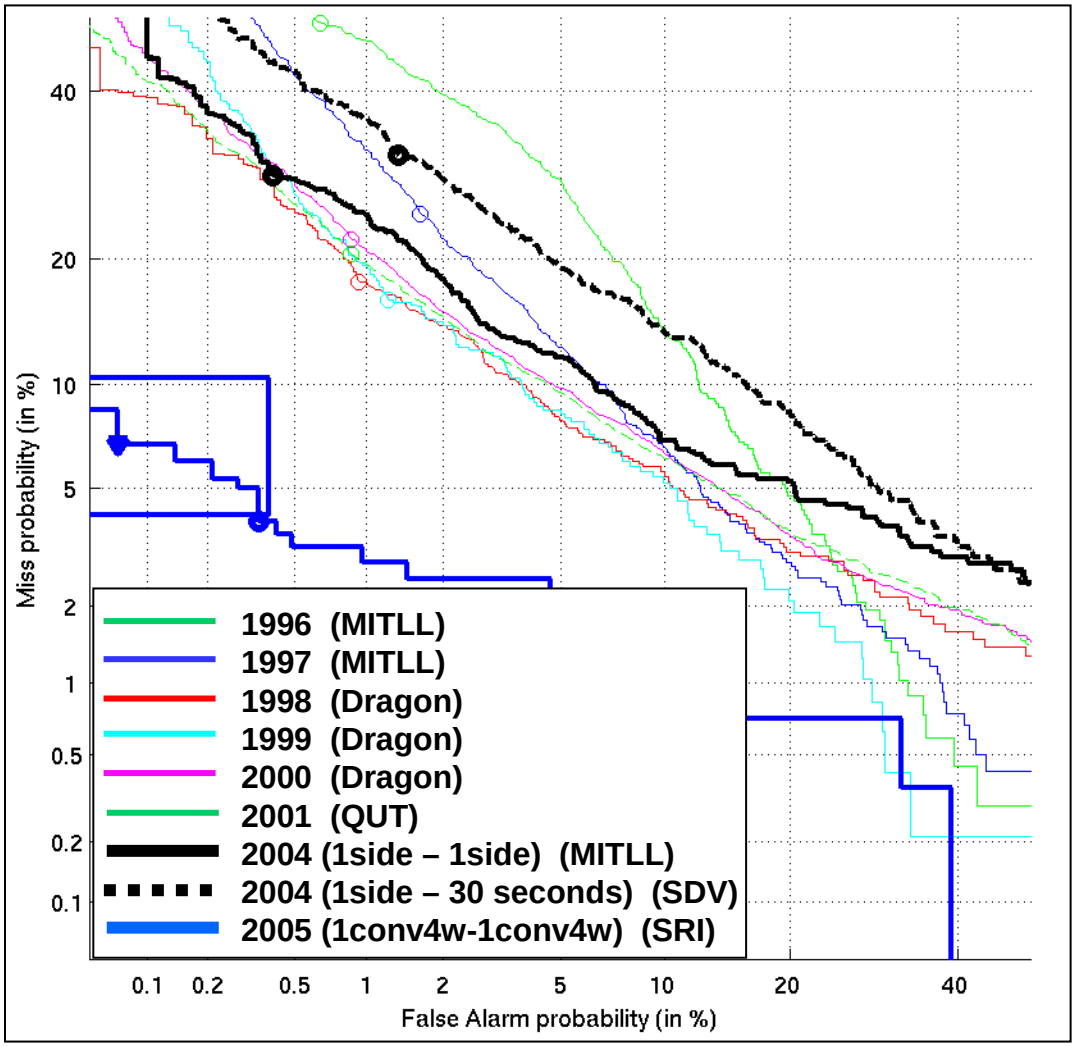


- 1conv4w-1conv4w common condition
- At a 10% miss rate the best performance this year has about a 0.4% FA rate, whereas last year it had about a 6% FA rate

Traditional 1-Speaker Detection History

- Each year we strive to match similar evaluation conditions with previous year's results in an attempt to track progress
  - Difficult as protocols change
- This year the protocols were quite similar to last year, with a couple of notable differences
  - Both channels were provided for the 4-wire tasks
  - BBN recognizer for ASR output changed

1 Speaker Detection Landline History



- A solid gain for 2005

- 1996-2001
  - Different number tests
  - 2 min. train, 30s test (variable length in '01)
- 2004-2005
  - 1 side training

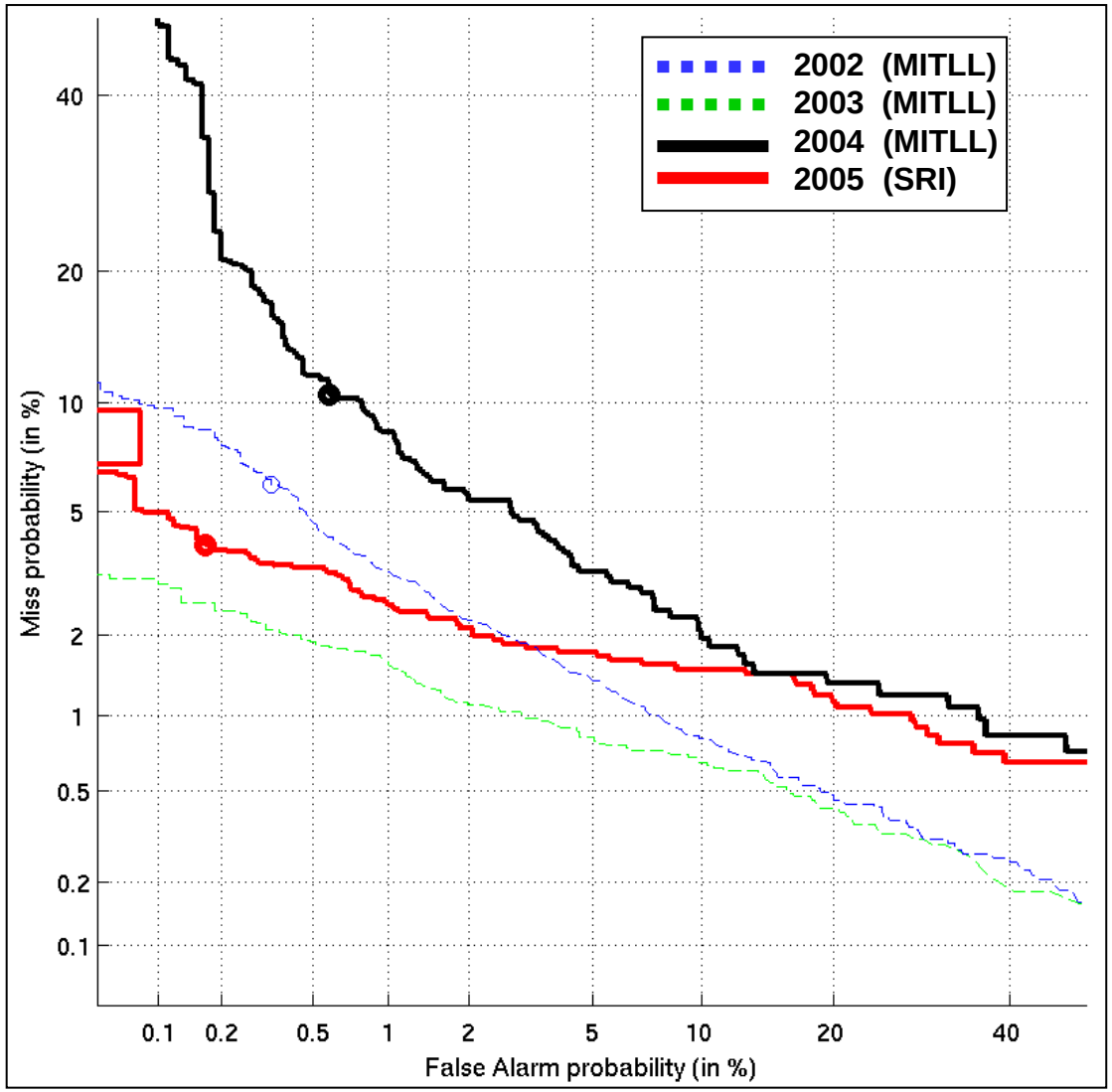
Year	Targets		Non-Targets	
	Trials	Spks	Trials	Spks
1996	677	40	46253	40
1997	1192	254	20383	280
1998	936	290	7157	424
1999	479	233	16247	489
2000	4209	804	42519	923
2001	4211	804	56091	923
04 1s	744	197	7019	239
04 30s				
2005	280	37	1437	109

Extended Data History

- 8sides-1side condition (02-04), 8conv4w-1conv4w (05)
- 2002-2003 used the same data sets
  - ASR transcripts at about 50% WER
  - Most target trials were same number tests
- 2004
  - ASR transcripts at about 20-30% WER
  - N-best Lists & ASR scores available
- 2005
  - ASR transcripts at about 20% WER

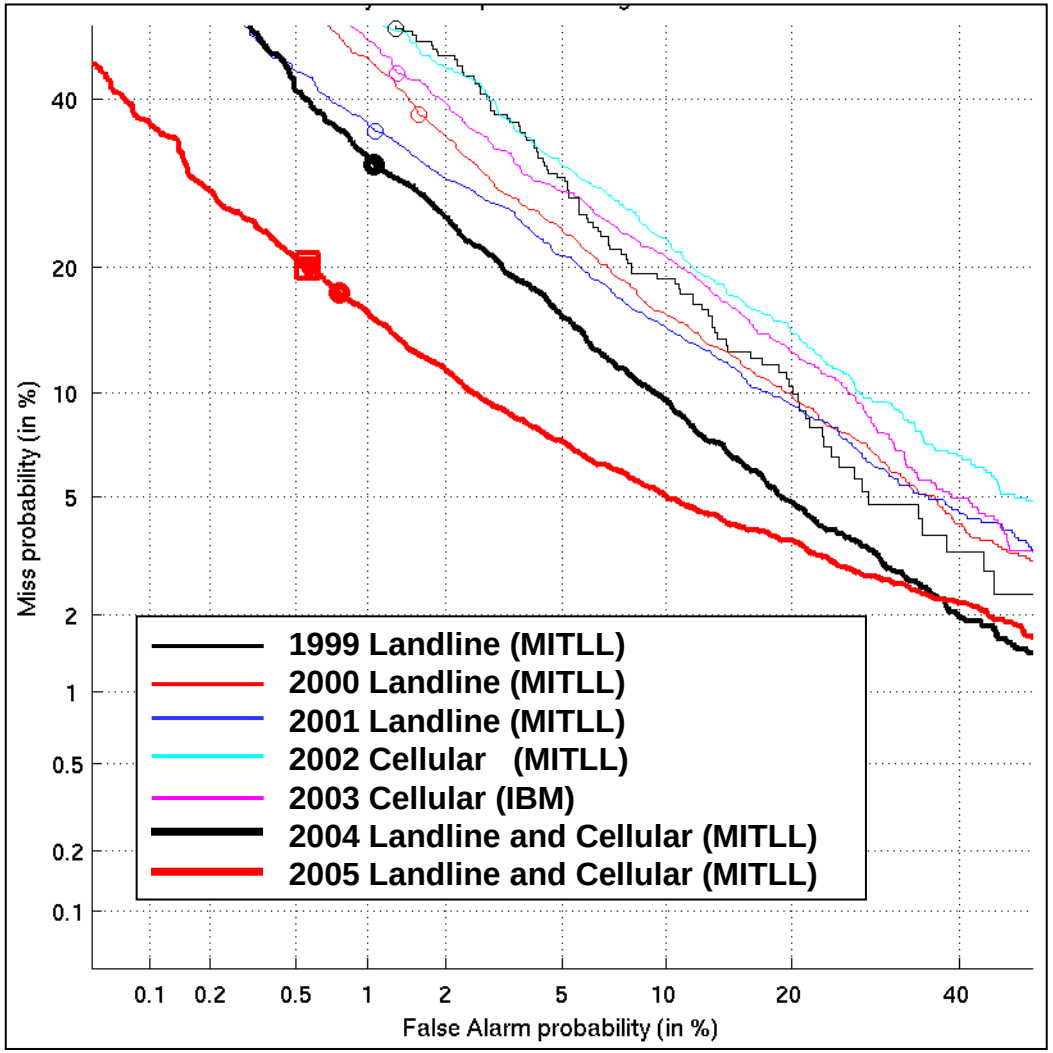
Year	Targets			Non-Targets		
	Trials	Speakers	Models	Trials	Models	Segment Speakers
2002	6127	291	613	6984	613	291
2003						
2004	830	243	272	8059	270	304
2005	1672	264	264	14300	384	359

DET Plot, History 8conv4w-1conv4w



- Unlike in earlier years, 2004 and 2005 trials contained
  - Entirely different number target trials
  - Fewer handsets used in training
  - Landline and cellular data
- A solid gain in 2005 over 2004

Two-Speaker History Plot



- 1999-2001
  - Referred to test segments only (1 min.)
- 2002-2004
  - Referred to training and test segments
- 2004-2005
  - Results combine transmission types
  - Whole conversation test segments

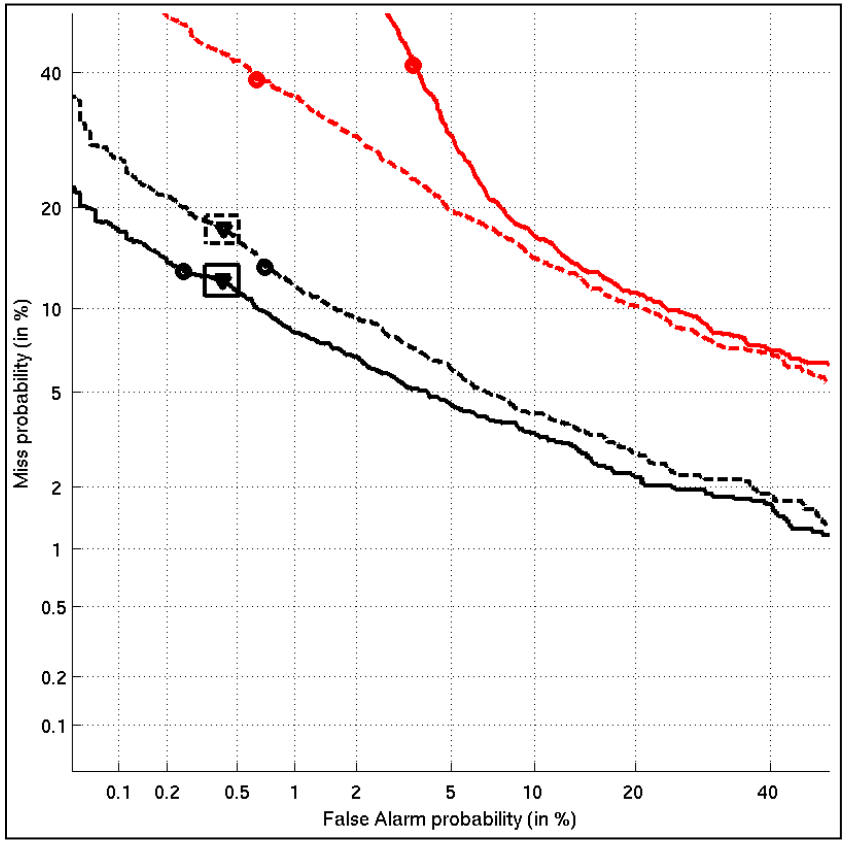
- Steady progress 2002-2005

Unsupervised Adaptation Mode

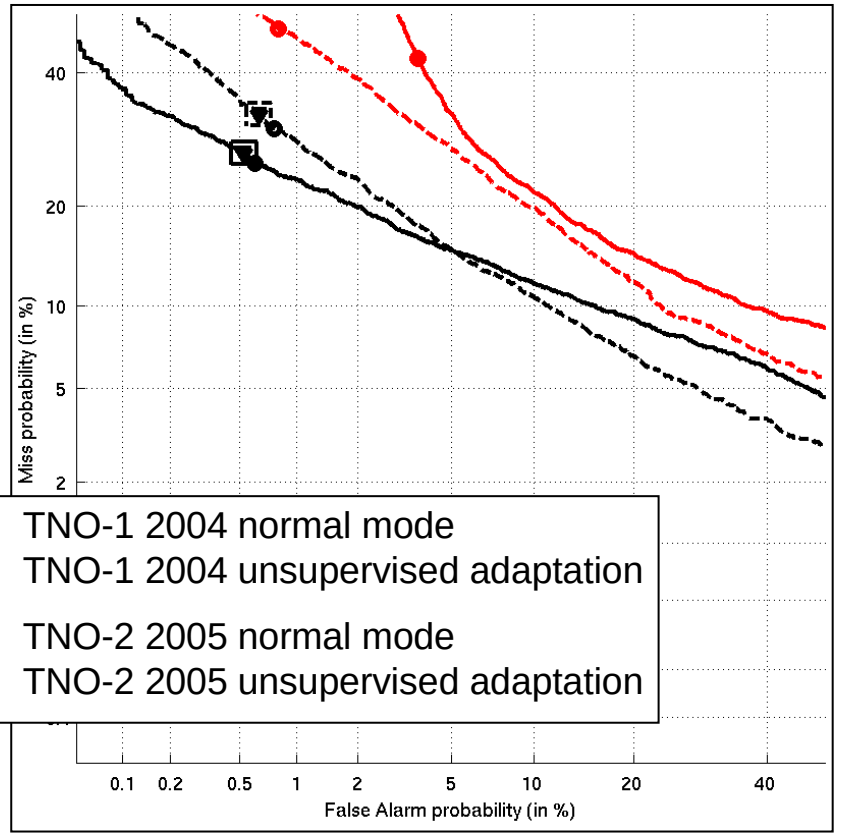
- Allowed models to be updated based on test segments processed in previous trials
  - Trials had to be processed in order
  - Systems had to be run without adaptation as well
- Only TNO attempted unsupervised adaptation this year
- The following plots compare TNO systems with and without adaptation in 2004 and 2005

TNO Unsupervised Adaptation

8conv4w-1conv4w



1conv4w-1conv4w

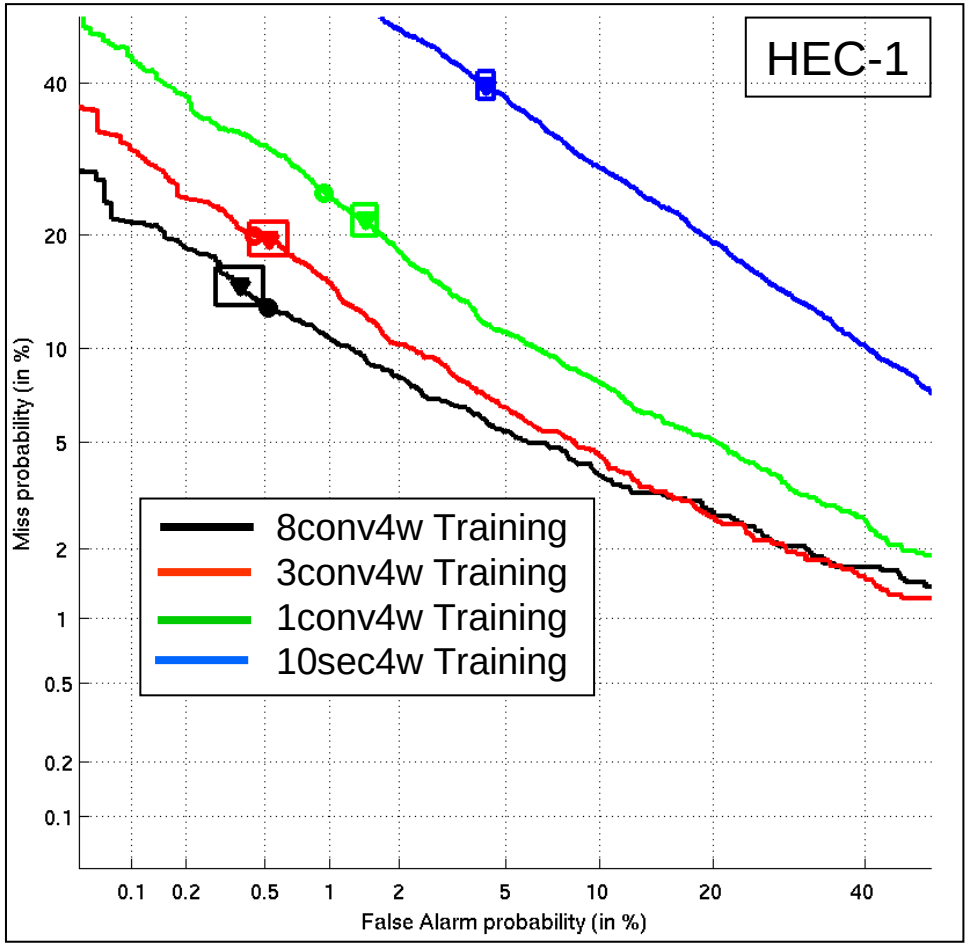


- Attempts to improve system performance using unsupervised adaptation techniques were unsuccessful in 2004, but TNO made it work this year

Varying the Training Duration

- Training condition may be
  - 8conv4w — 1conv4w
  - 3conv4w — 10sec4w
- Test duration is fixed at 1conv4w
- Plot all trials restricted to:
  - All English training data
  - All English test data
  - Single handset used in training

DET Plots by Training Duration
For 1conv4w Test Duration

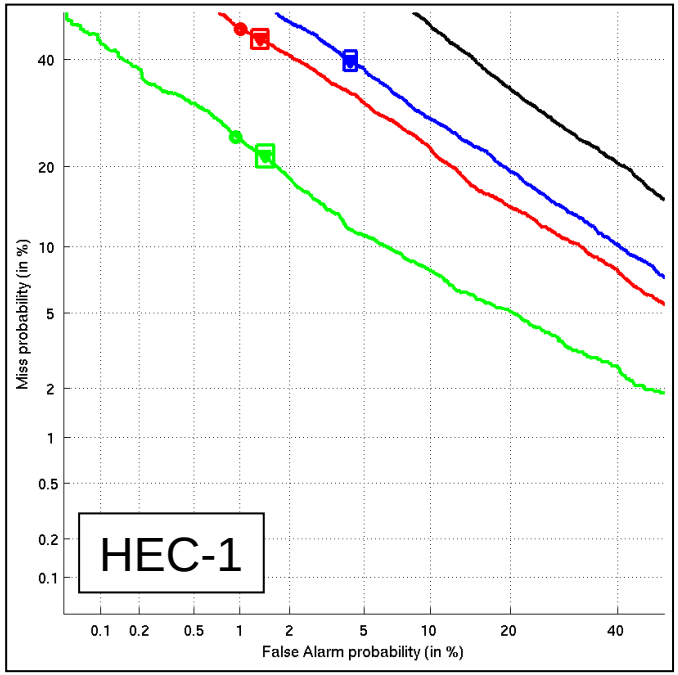


- Performance level varies directly with training duration
- Trend holds generally for other systems

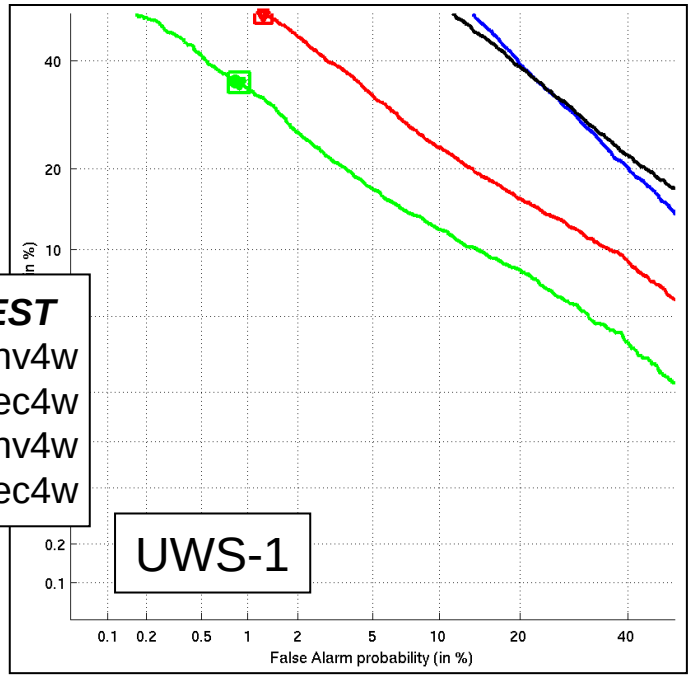
Varying the Training and Test Duration

- Training duration may be:
  - 1 conversation side (4-wire)
  - 10 seconds (4-wire)
- Test duration may be:
  - 1 conversation side (4-wire)
  - 10 seconds (4-wire)
- Plot all trials with
  - All English training and test data,
  - Single handset used in training

Varied Training and Test Durations



TRAIN	*TEST*
1conv4w	1conv4w
1conv4w	10sec4w
10sec4w	1conv4w
10sec4w	10sec4w

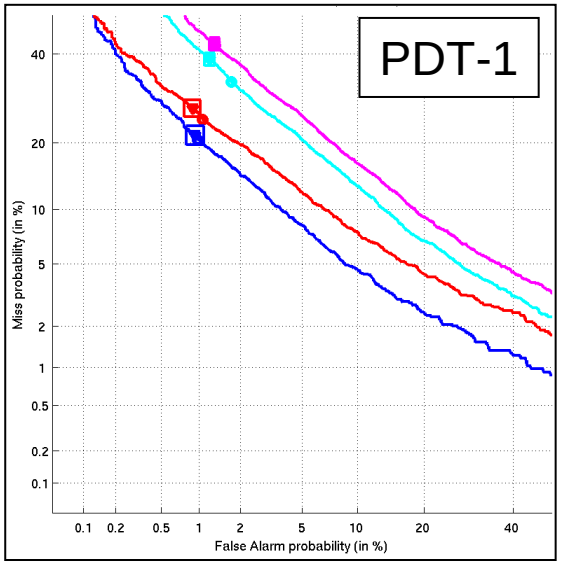


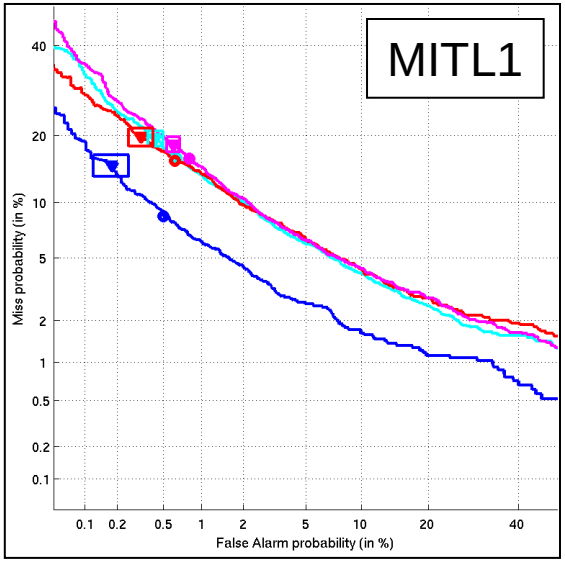
- Performance varies directly with both training and test durations
- Longer training and shorter test outperforms the reverse, at least for systems doing all these test conditions

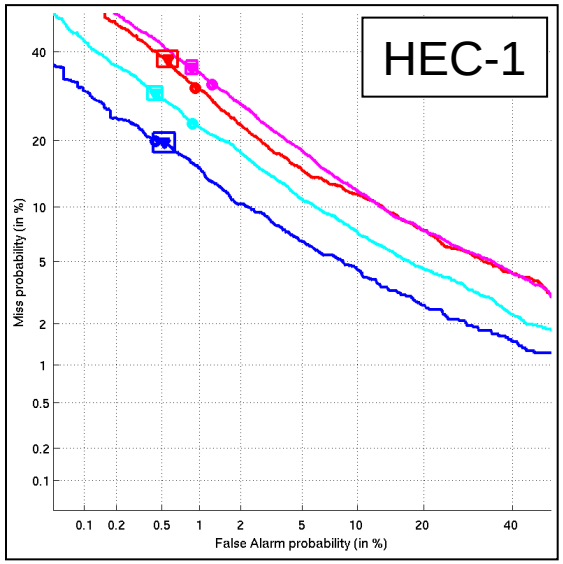
Single Channel vs. Summed Channel

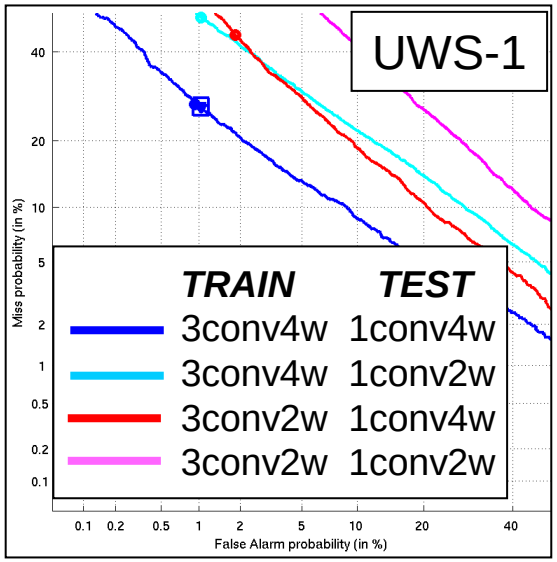
- Training condition may be:
  - 3conv4w
  - 3conv2w
  - 1conv4w
- Test condition may be
  - 1conv4w
  - 1conv2w
- Plot all trials with
  - All English training data
  - All English test data
  - Single handset used in training

Varied Test Segment Type and Training Type *(single channel or summed)*



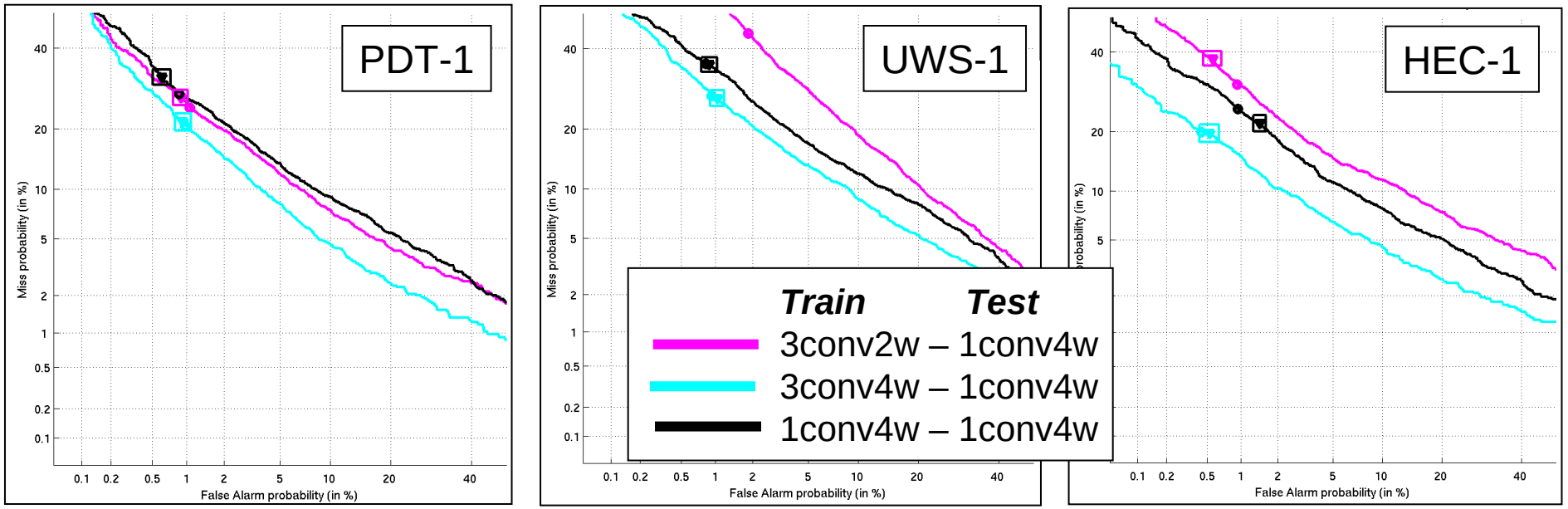






- All 4-wire is always best, and all 2-wire is always worst, but
- Where in the middle the mixed conditions place, and how the mixed conditions compare, varies by system

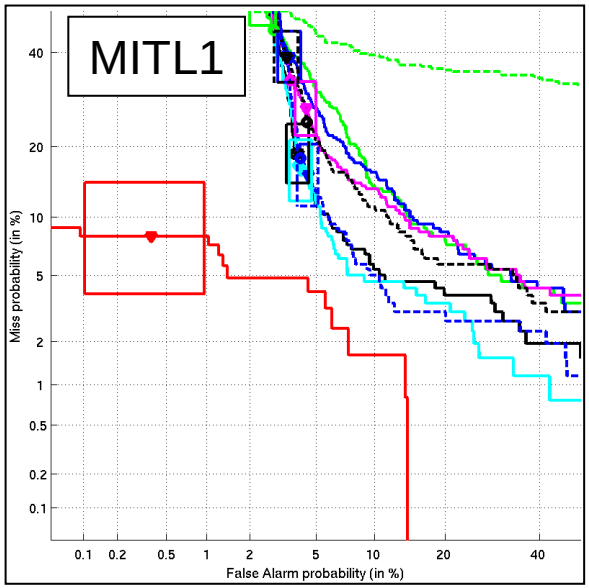
Varied Training for Fixed Test Data

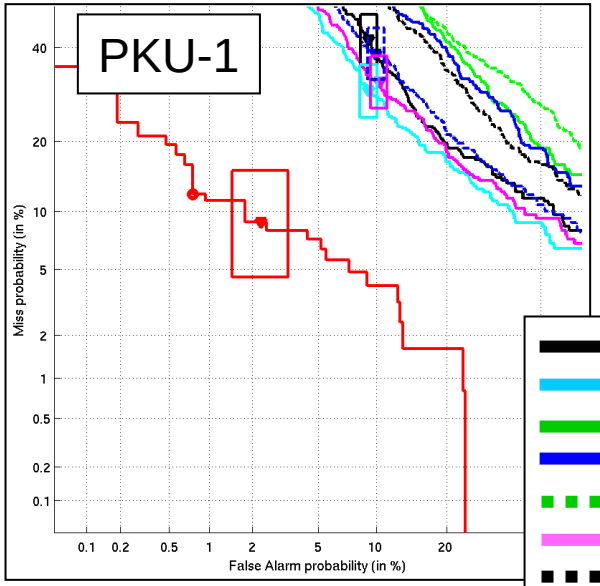


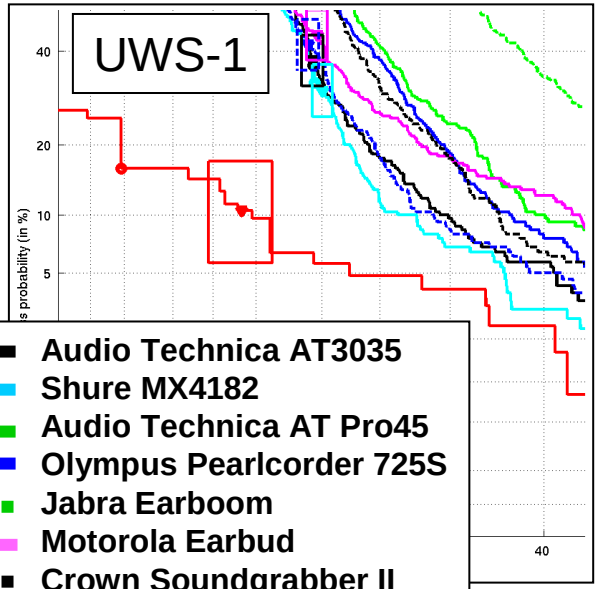
- The tradeoff between 3conv2w and 1conv4w training varies among systems

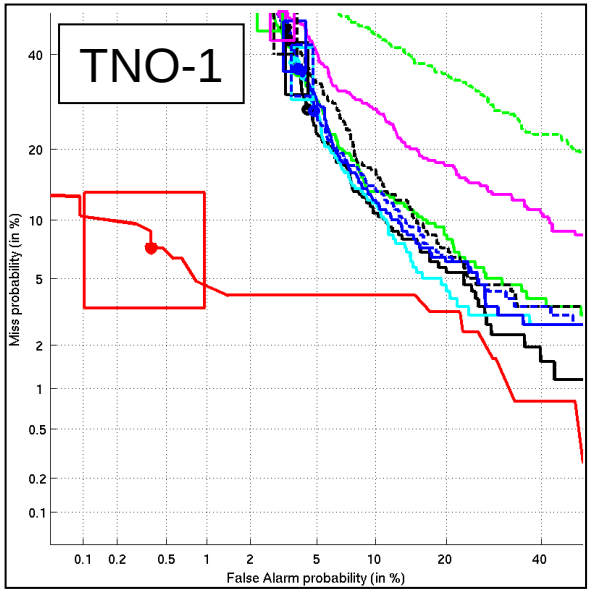
Cross Channel Results

1conv4w-1convmic









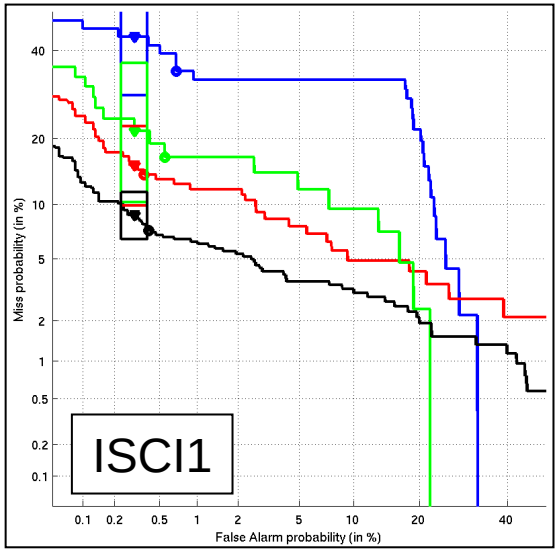
- Four sites attempted this condition
- Microphone performance lags telephone performance considerably
- Was there a problem with the Jabra Earboom?

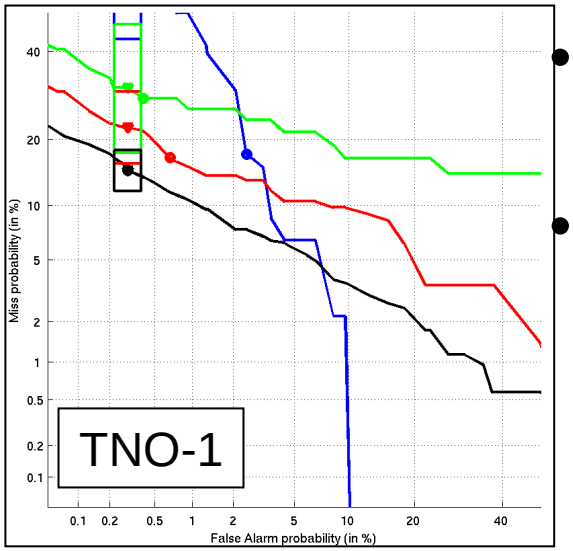
Transmission Type

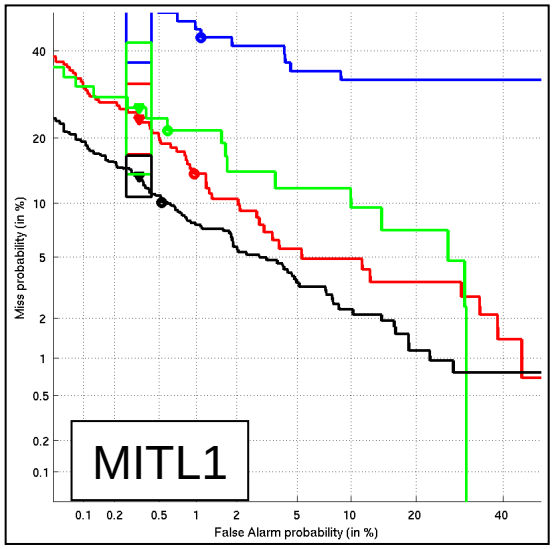
- MIXER subjects were to report the transmission type being used as either:
  - Cellular
  - Standard Landline, or
  - Cordless (landline)
- Unfortunately the most common reported transmission type is “NA” (not available)
  - Omitted from following plots

Targets			
Train – Test (transmission type)	**Trials**	**Speakers**	**Models**
Cellular – Cellular	46	10	12
Cellular – Landline	143	29	33
Landline – Landline	523	84	93
Landline – Cellular	42	14	14

Transmission Type Plots







TRAIN	*TEST*
Cellular	Cellular
Cellular	Landline
Landline	Cellular
Landline	Landline

- Fixed Non-Target trials to common condition
- Target Trials partitioned by model/segment transmission type:

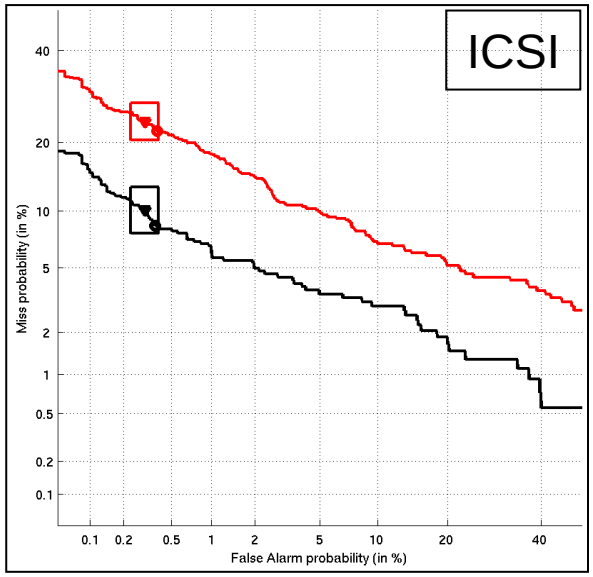
Model	Segment
Cellular	Cellular
Cellular	Landline
Landline	Landline
Landline	Cellular

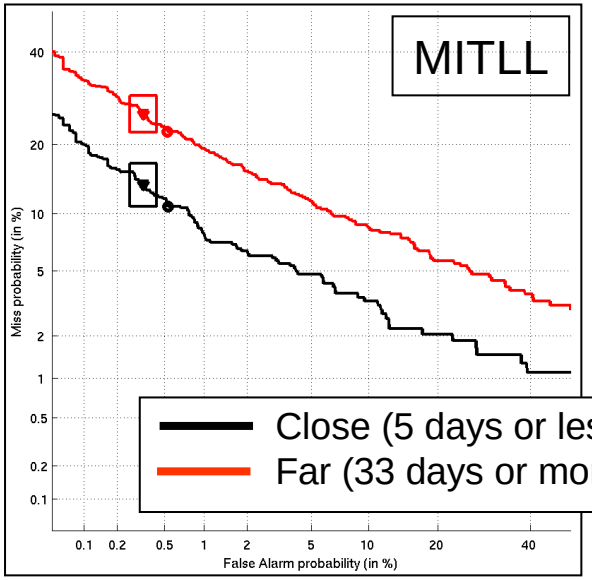
- Landline enhances performance, especially in the test segment, but data is limited

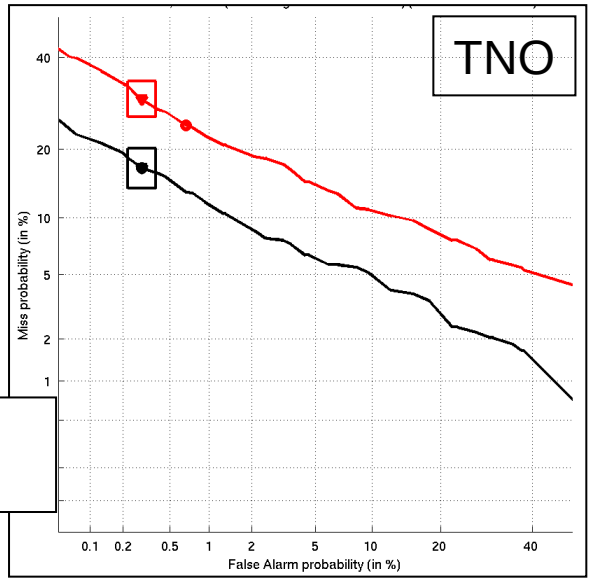
Time between “training” and “test”

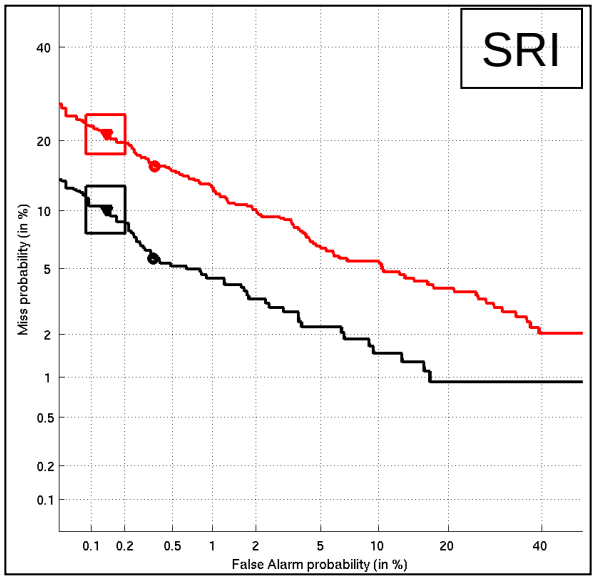
- Performance is shown by the amount of elapsed time between when the training data was recorded and when the test segment was recorded for the same speaker (target) trials.
  - CLOSE: 25% of target trials with smallest time difference
  - FAR: 25% of target trials with largest time difference

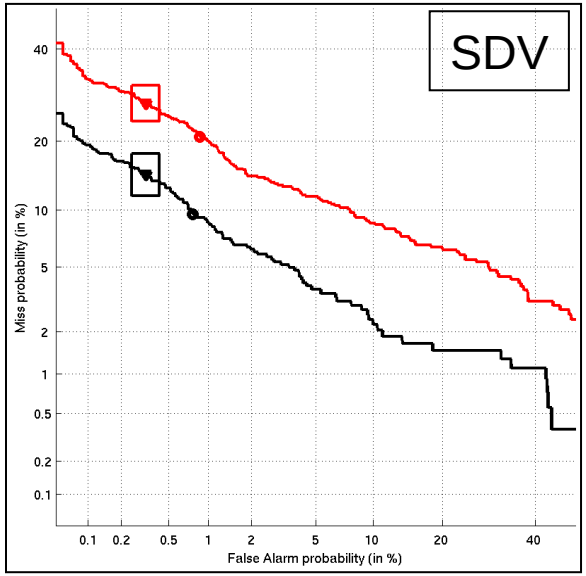
Time between “training” and “test”











- Is there an alternative explanation other than time between recordings?

Back to the 2005 Campaign
Best Actual Decision *(all systems)*

		Training				
		8conv4w	3conv4w	1conv4w	10sec4w	3conv2w
Test Conditions	1conv4w	SRI-2	QUT-1	SRI-1	USTC1	MITL1
	10sec4w	TNO-1	USTC1	HEC-1	TNO-1	PDT-1
	1conv2w	HEC-1	MITL1	PRS-1	HEC-1	MITL1
	Mic	MITL1	MITL2	MITL1	UWS04	UWS-1

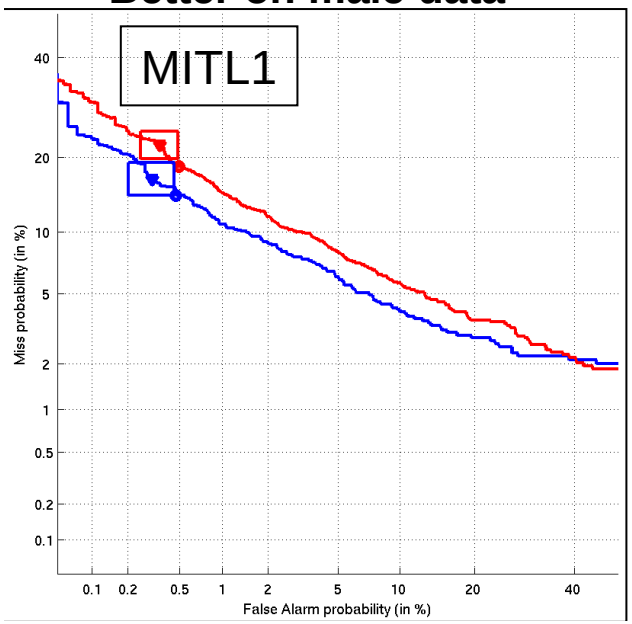
Quite a few “winners”

All tests restricted to the “common evaluation condition”

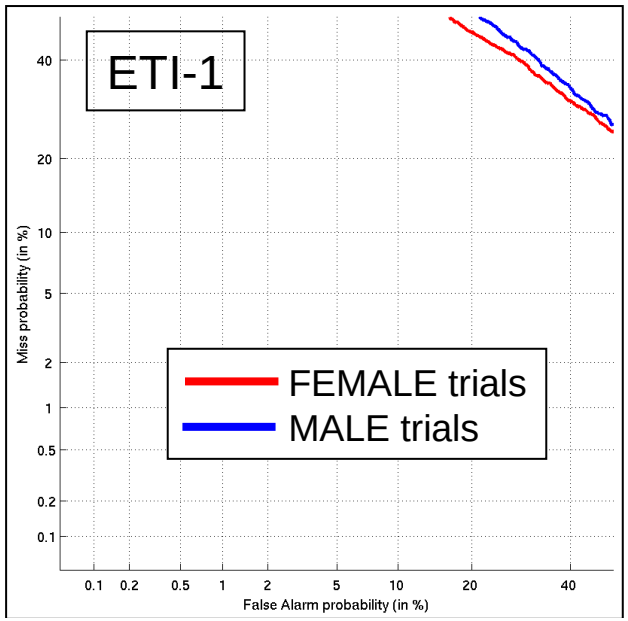
Performance by Speaker Sex
1conv4w-1conv4w (*common condition*)

- In general bias this year was toward better performance on the male data than on the female data
  - 18 primary systems performed better on the male data
  - 1 primary system performed better on the female data
  - 8 primary systems did not have a noticeable difference

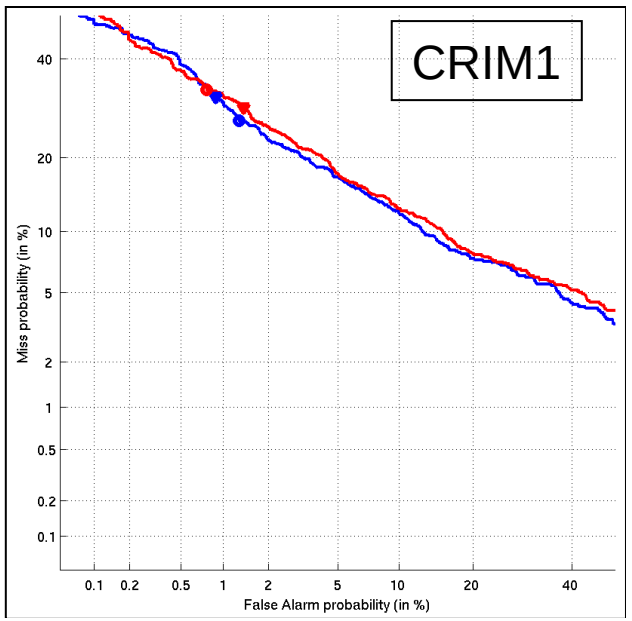
Better on male data



Better on female data



No difference



Conclusions

- Evidence of significant improvement in the past year
- Cross channel performance tested – much room for improvement
- One site made unsupervised adaptation work
- NIST will make available “publishable” anonymous plots as indicated in section 7 of the evaluation plan
- MITLL identified possible gender errors that we still need to verify