Human Assisted Speaker Recognition (HASR)

ICASSP 2011 Special Session

May 26, Prague, Czech Republic

ALL speaker recognition applications involve humans

Different applications involve different human roles, e.g.,

- physical entry and account access require meta-data judgments
- forensic and intelligence applications require speech judgments

SO the best outcomes require man-machine collaboration

- Need to understand relative strengths, limitations of both

BUT why a NIST eval?

- Leverage the data, metrics, and discipline of NIST approach
- No change in policy: NOT testing applications or products
- Testing assumptions underlying many applications
- Perhaps a basis for new research

Human Assisted Speaker Recognition (HASR) Report on A Pilot Experiment

Craig Greenberg*, Alvin Martin*,
George Doddington, John Godfrey^
*NIST Multimodal Information Group
^US Department of Defense

ICASSP 2011
May 26, Prague, Czech Republic

HASR Introduction

How can human experts effectively utilize automatic speaker recognition technology?

Little discussed … even less tested

- SRE10 included HASR (*Human Assisted Speaker Recognition*) evaluation to begin addressing this question – a pilot test

The HASR Task:

Given two different speech segments determine whether they are both spoken by the same speaker

- HASR included two tests:
  - HASR1 – 15 trials & HASR2 – 150 trials
- HASR systems may use human listeners, machines, or both
  - Participation open to all who might be interested

HASR Protocols

- Trials consist of two speech segments: same speaker or different?
  - Unlimited listening permitted
- Trials to be processed separately and independently
  - Trials processed in sequence, one trial at a time
- Systems provide a same-speaker/different speaker decision and a score
- Evaluation
  - Count numbers of Misses and False Alarms
    - A ***miss*** is deciding the segments were spoken by ***different*** speakers when they were spoken by the ***same*** speaker
    - A ***false alarm*** is deciding the segments were spoken by the ***same*** speaker when they were spoken by ***different*** speakers

Trial Selection

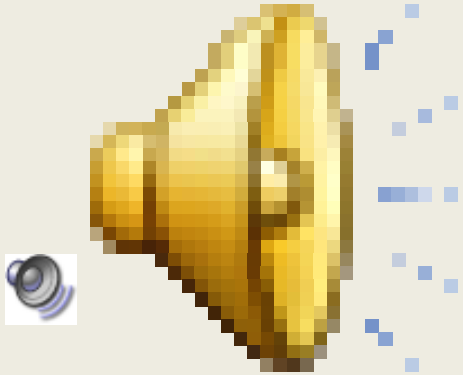
- Difficult trials were needed
  - First segment from interviews, various microphones (3 min)
  - Second segment from telephone calls (5 min)
- HASR1 selection procedure:
  - Segment-pair similarity judged by machine
    - Most similar different-speaker pairs selected for different-speaker trials
    - Least similar same-speaker segments selected for same-speaker trials
  - Segment pairs screened to select the most difficult trials and to eliminate content cues
- HASR2 selection procedure:
  - Same as HASR1 but screening was limited to removing only those pairs with content cues

HASR1 Trial Examples

- Example 1
  - Segment 1

Segment 2

- Example 2
  - Segment 1



Segment 2

HASR1 Trial Examples

- Example 1
  - Segment 1 Segment 2

- Example 2
  - Segment 1 Segment 2

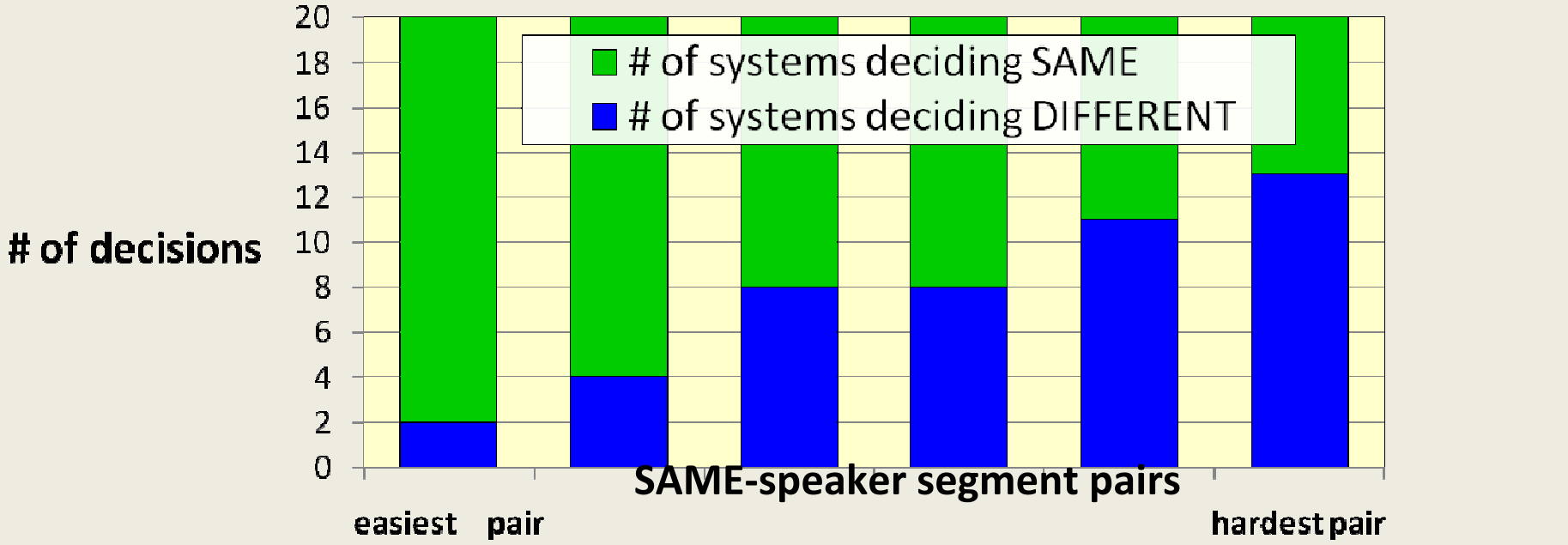
HASR Participation

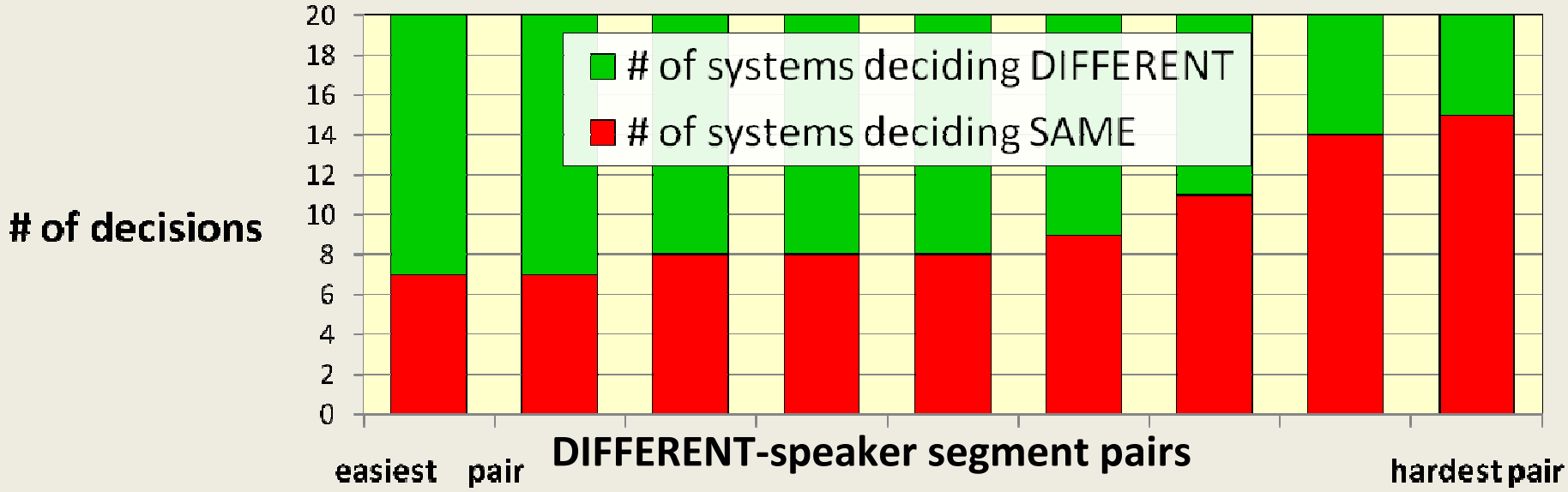
HASR1 -- 20 systems from 15 sites (six countries)

HASR2 -- 8 systems from 6 sites

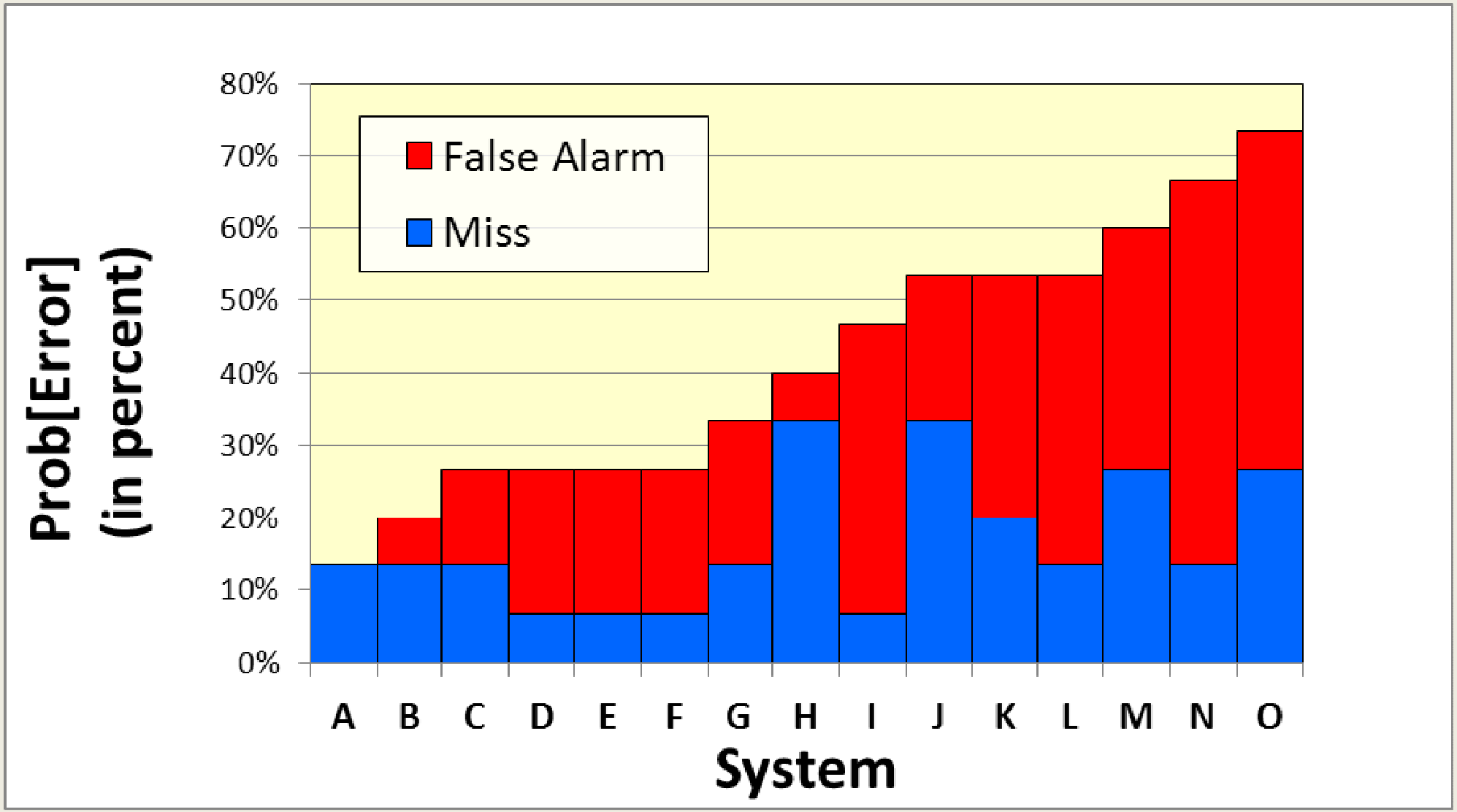
•Most sites also participated in the SRE10 automatic system evaluation

Difficulty of HASR1 segment pairs

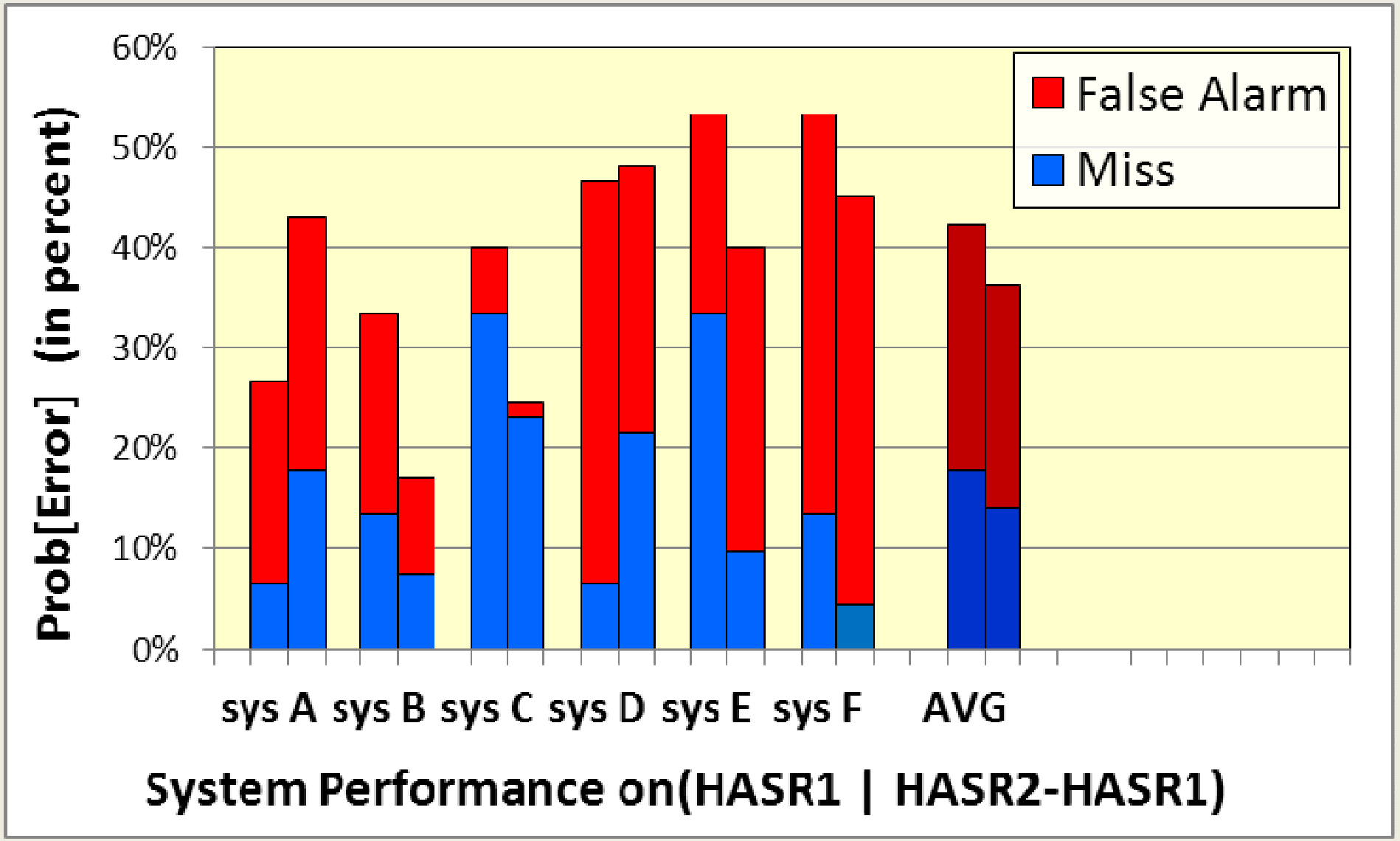




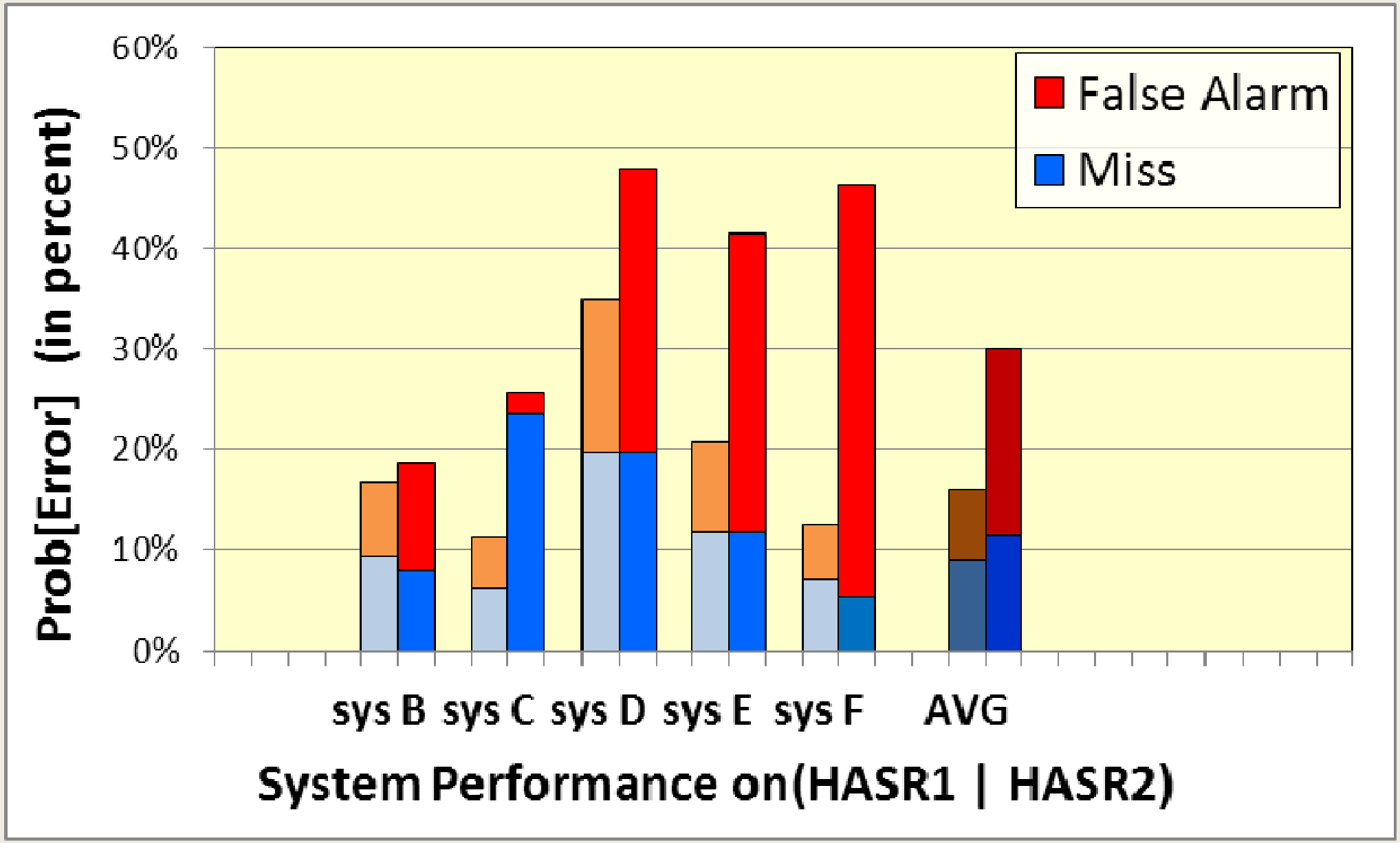
HASR1 System Performance



System Performance on HASR1 and HASR2



System Performance on HASR2



Summary

- HASR system performance did not compare favorably with that of automatic systems
- The test set was challenging
  - Half the systems got more trials wrong than right in HASR1
- Was this data appropriate?
- Whither HASR research?

Our Plans

- Another HASR evaluation in conjunction with SRE12
  - 20 trials for HASR1 and 200 trials for HASR2
  - Trial selection similar to HASR10
  - English Only
  - ~4 month evaluation period
- Desired feedback for improved evaluation
  - **Statistical significance**
  - More rigorous selection process
- Website:

  http://nist.gov/itl/iad/mig/hasr.cfm

Thank You Very Much!

Questions?

Trial Selection - HASR1 *(15 Trials)*

- Sought "difficult" cross-channel trials from the Mixer 6 Corpus
  - Training data from interviews included various room mic channels
  - Test data from phone calls included some with high or low vocal effort
- An automatic system* processed the "full matrix" of trials

on-target Speaker Pairs: Ran full matrix of possible interview-train, interview-test non-target trials over all speaker pairs

speaker pairs identified using a threshold of 6 scores (of 9 possible) in the top 1% of scores for trials run against the specific target

on-Target trials: Listened to all potential interview-train/phone-call-test trials for each pair

uch trials judged most similar were selected

rget trials: Ran full matrix of potential interview-train/phone-call-test target trials over all speakers

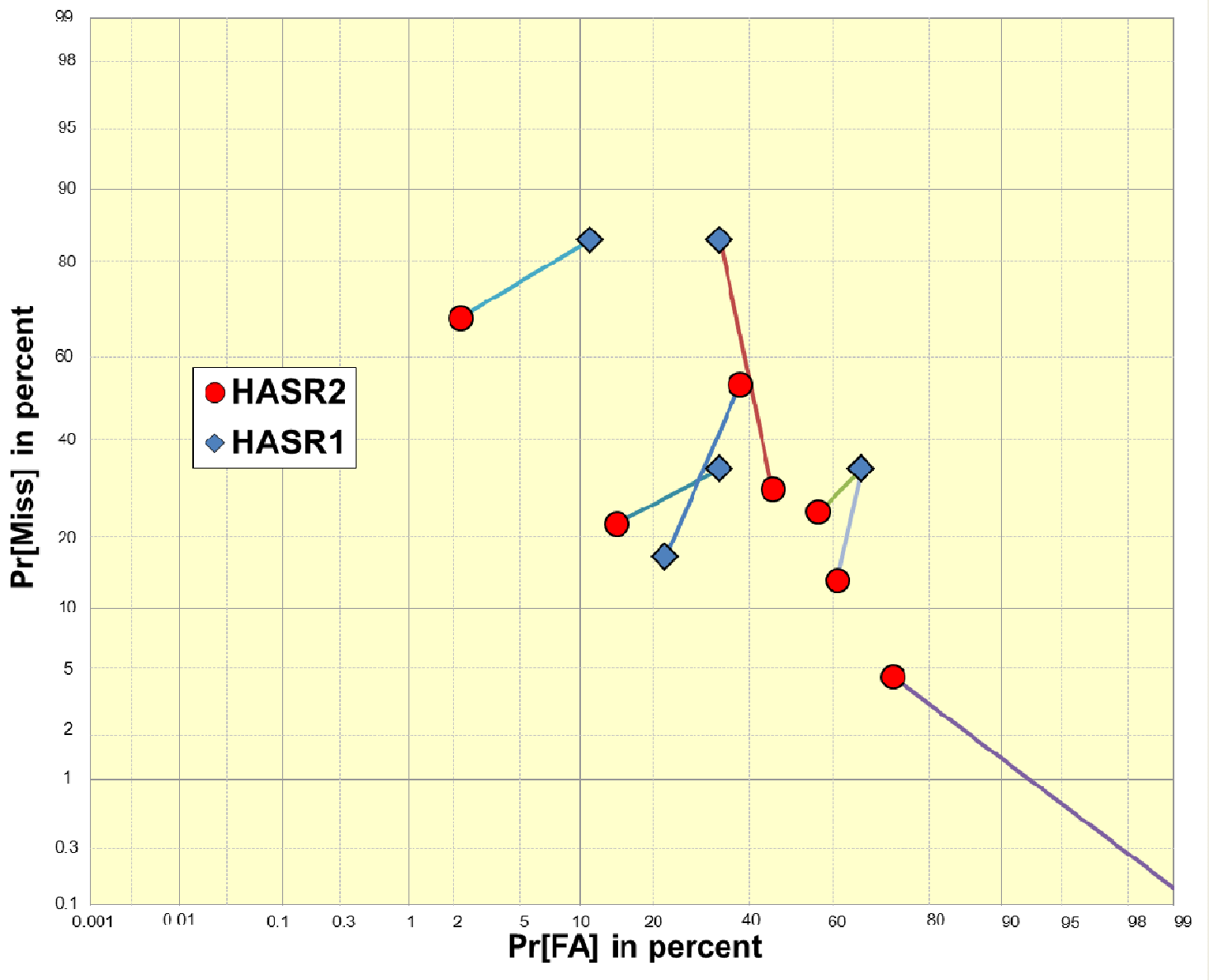
such trials with lowest scores were selected and listened to, and the 6 such trials judged most dissimilar were selected

* Thanks! to Howard Lei at ICSI for assisting us with processing the full matrix of trials

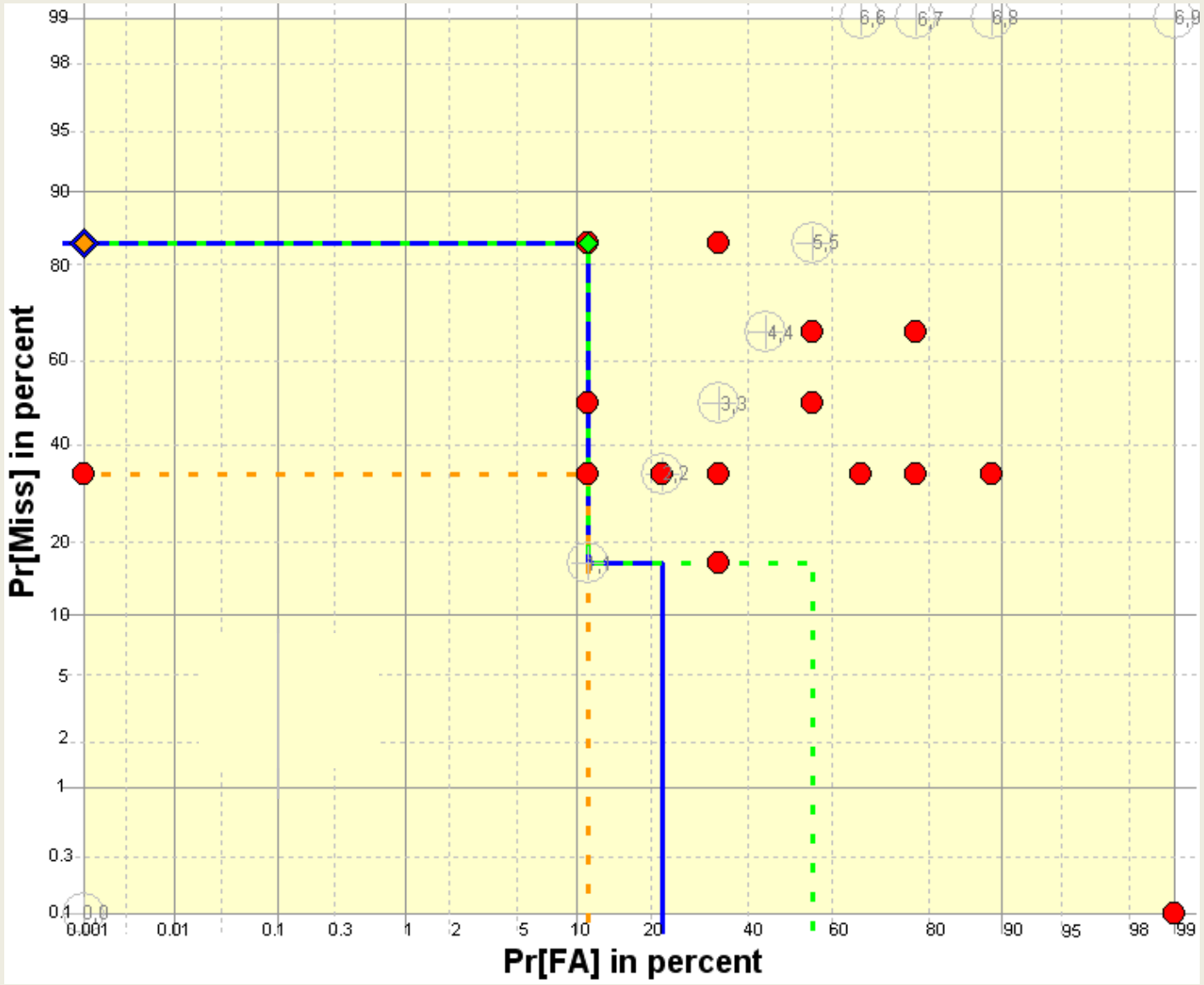
HASR1 Participants & Results Summary

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Misses	FAs	Total
System 1	t	f	f	f	f	f	t	f	f	t	f	f	f	t	f	2	-	2
System 2	t	t	f	f	t	f	t	t	f	t	f	f	t	f	t	1	3	4
System 3	t	t	f	f	t	t	f	f	f	t	t	f	f	t	f	2	3	5
System 4	t	t	f	f	t	t	f	f	f	t	t	f	f	t	t	1	3	4
System 5	t	t	f	f	t	f	t	t	f	t	f	f	t	f	t	1	3	4
System 6	t	f	t	t	f	t	f	f	t	f	t	f	f	t	f	4	5	9
System 7	f	t	f	t	f	f	f	t	f	f	f	f	f	t	f	5	3	8
System 8	f	t	t	t	f	t	f	t	t	t	t	f	f	t	f	4	7	11
System 9	t	t	f	t	t	f	f	f	t	t	t	t	t	t	f	2	6	8
System 10	t	t	f	t	t	f	f	f	t	t	t	t	t	t	f	2	6	8
System 11	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	-	9	9
System 12	f	f	t	f	t	t	t	t	t	t	t	t	f	t	t	1	6	7
System 13	f	t	t	f	t	t	t	f	t	t	t	t	t	t	f	2	7	9
System 14	f	t	t	f	t	t	t	f	t	t	t	t	t	t	f	2	7	9
System 15	t	f	f	f	f	f	t	f	f	t	t	f	f	t	f	2	1	3
System 16	f	t	f	f	f	f	t	f	f	t	t	f	f	t	f	3	2	5
System 17	t	t	t	t	f	t	f	f	f	t	t	f	f	t	f	3	5	8
System 18	t	t	t	t	t	t	f	f	t	t	t	t	t	f	t	2	8	10
System 19	f	f	f	f	t	f	f	t	f	t	t	f	f	t	t	2	2	4
System 20	f	f	f	f	f	t	f	f	f	t	f	f	f	f	f	5	1	6
KEY	**T**	**F**	**F**	**F**	**T**	**F**	**T**	**F**	**F**	**T**	**F**	**F**	**F**	**T**	**T**	-	-	-
Number of Errors	***8***	***14***	***8***	***8***	***8***	***11***	***11***	***7***	***9***	***2***	***15***	***7***	***8***	***4***	***13***	***46***	***87***	***133***

HASR2 Systems
HASR1/HASR2 DET Points



HASR1 DET Points and Curves

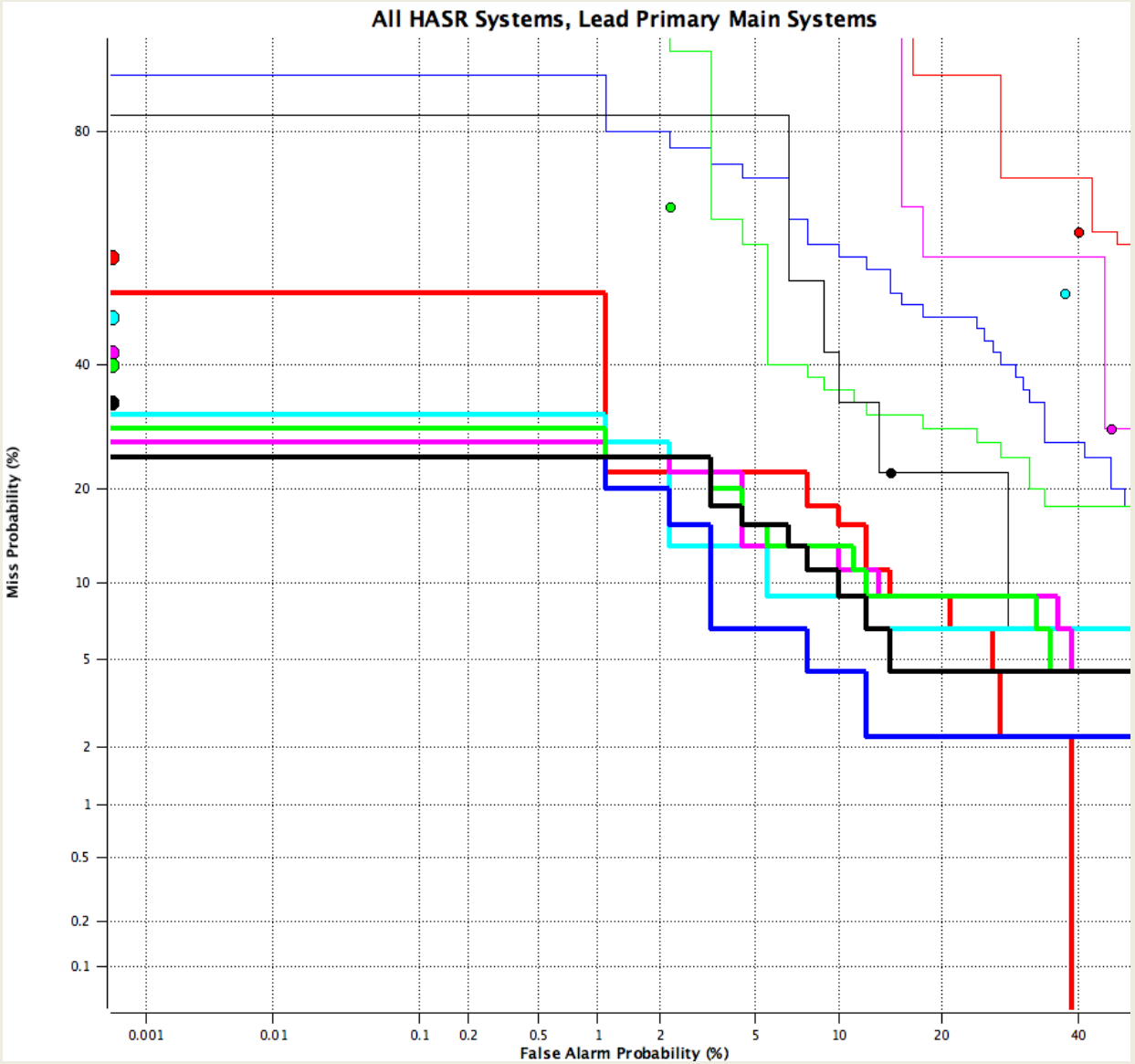


Red dots represent decision points (on DET plot) of HASR1 systems

Note that extreme points represent systems tuned to different tradeoffs between miss and FA rates

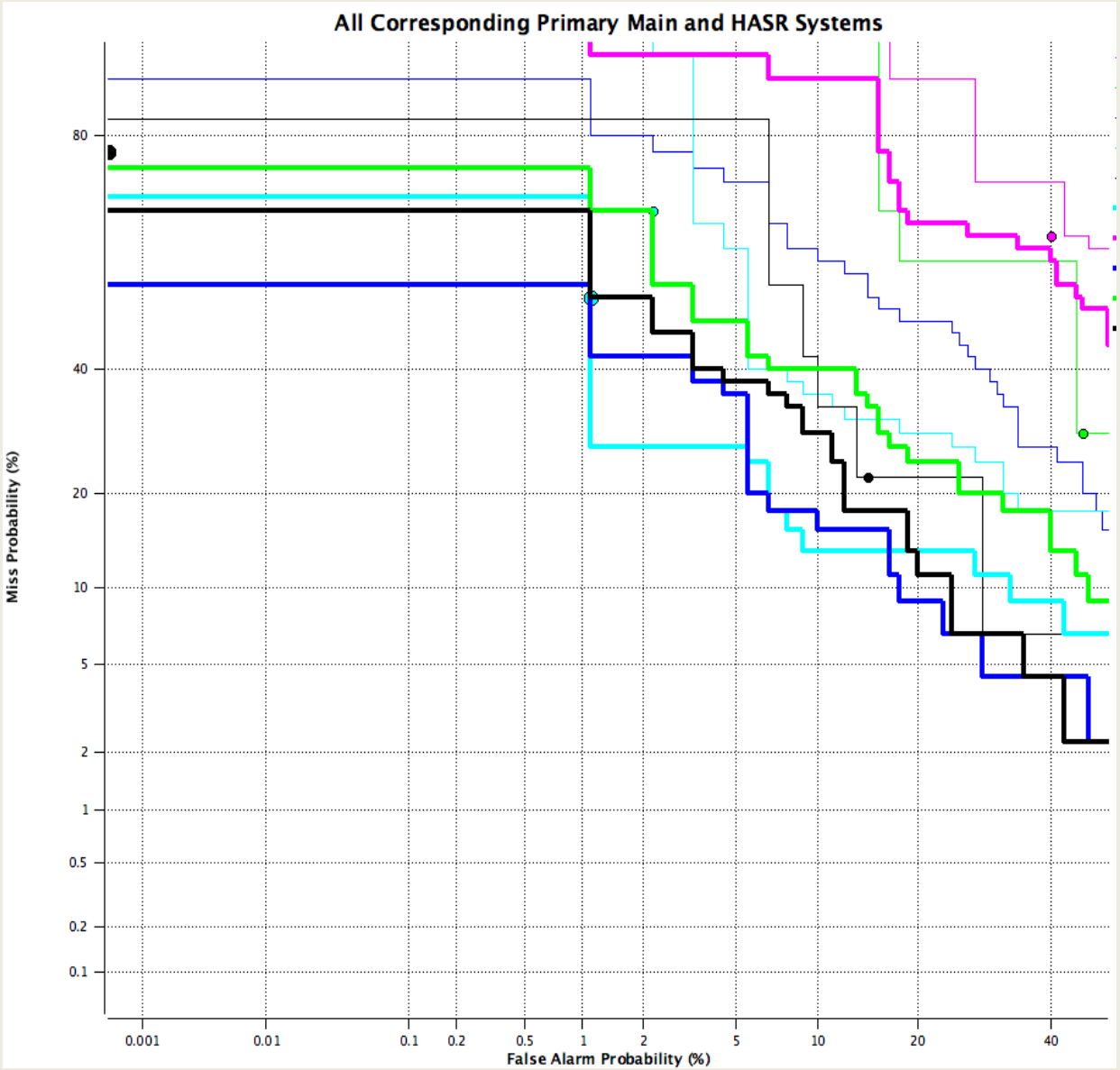
Leading automatic systems' DETs compare favorably with HASR decision points on HASR1 trials

HASR2 and Leading SRE10 Automatic Systems



- 135 HASR2 trials
- Six HASR systems (thin lines)
  one system = decision only
- Six Automatic systems (thick lines)

HASR2 and Corresponding Automatic Systems



- 135 HASR2 trials
- Five HASR systems (thin lines)
- Five Corresponding Automatic systems (thick lines)