Human Assisted Speaker Recognition (HASR) in NIST SRE10

Craig Greenberg[*], Alvin Martin[*], Linda Brandschain[¦], Joseph Campbell[#], Christopher Cieri[¦], George Doddington, John Godfrey[^]

[*]NIST Multimodal Information Group

[¦]Linguistic Data Consortium

[#]MIT Lincoln Laboratory

[^]US Department of Defense

Odyssey 2010

The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Introduction

How can human experts effectively utilize automatic speaker recognition technology?

Much discussed .. little tested

- SRE10 included HASR (*Human Assisted Speaker Recognition*) tests 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 a subset of trials from the SRE10 core test
  - ➢ HASR1 – 15 trials & HASR2 – 150 trials
- HASR implies human listening to assist in making a decision
  - ➢ System descriptions provided to describe processing techniques
- Participation open to all who might be interested, ranging from "experts" to "naïve" listeners

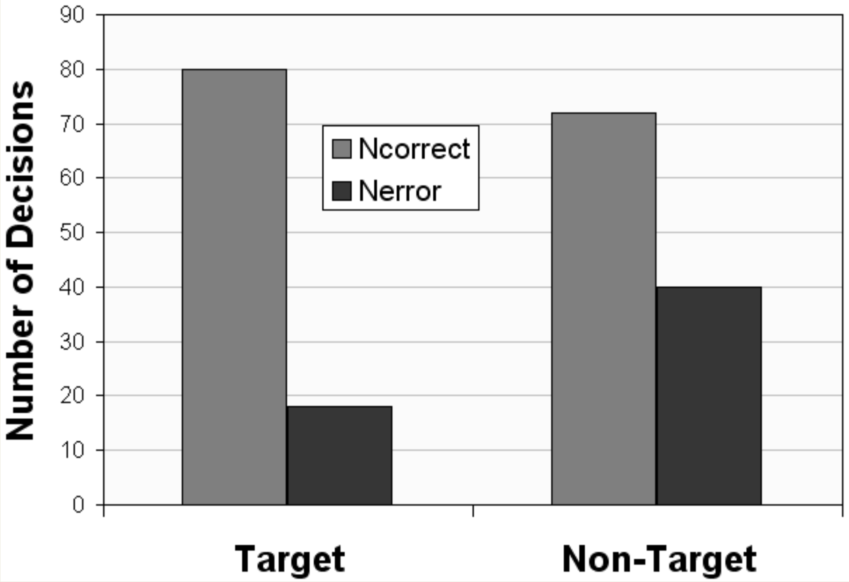
Preliminary Experiment

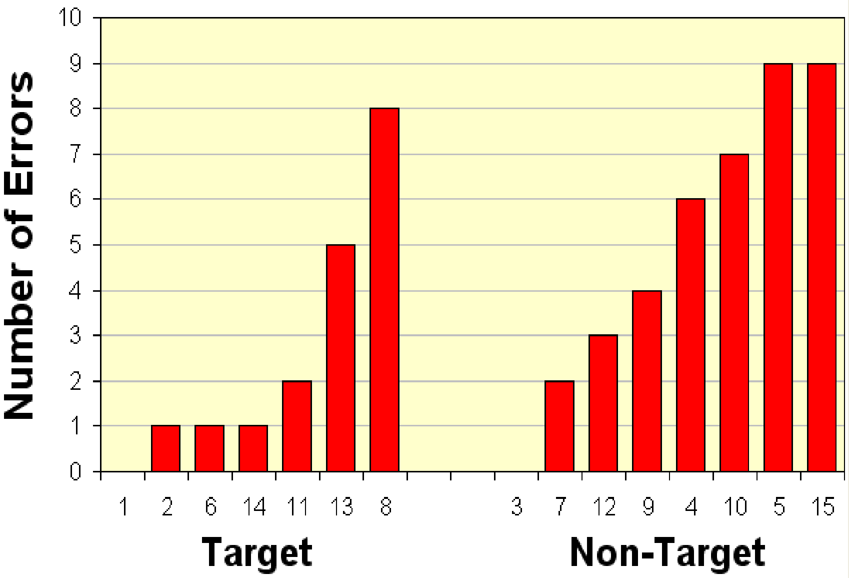
- A preliminary experiment was designed to test protocols for a proposed HASR study
- Identified confusable speaker pairs from SRE08
  - Used results from the full matrix runs of 150 speakers
  - Automatic systems identified speaker pairs (using ROVER)
    - 47 pairs found with multiple system errors
  - Listened to interview sessions of all these pairs
    - Selected 10 pairs as most difficult to distinguish
- (8) Non-target trials were selected from these 10 pairs
- (7) Target trials were selected from among the speakers included in the 47 identified pairs
  - Selected target trials whose two segments sounded most different
- Subject lavalier channel used in all cases

Preliminary Experiment (cont'd)

- 14 human evaluators listened to these trials
  - Volunteers involved in the project (Gov./Data providers)
  - Permitted unrestricted listening to both train and test
- Evaluators provided
  - Actual decision (required)
  - A confidence score (optional)

Preliminary Experiment – Results



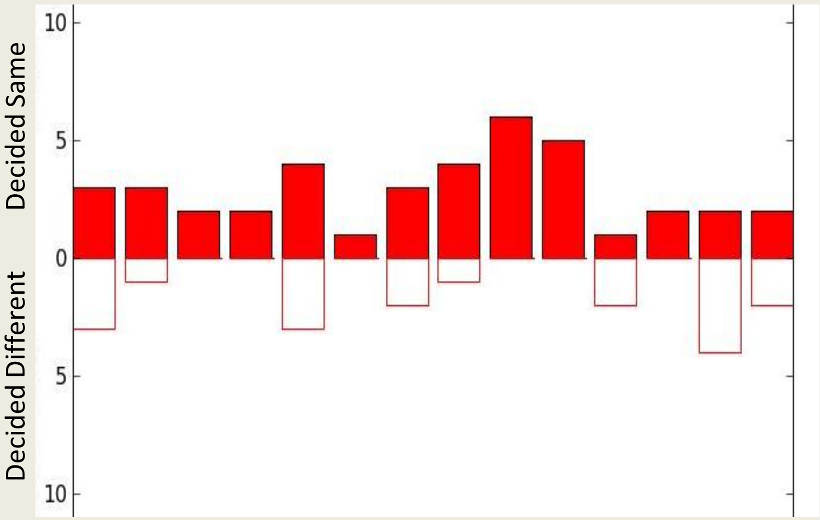


- Correct and incorrect decisions over all evaluators
  - Overall miss rate > 18%
  - Overall false alarm rate > 36%

- Some trials were very challenging
  - (3) had more errors than correct decisions
  - (3) others had more than 1/3 errors
  - Only (1) target and (1) non-target trial had no errors

Preliminary Experiment – Results

Errors by Evaluator



- Shows FA’s (solid red) and misses (outline) for each evaluator
  - All had FA errors
  - Half had total error rate of 1/3 or more

- Most trials proved quite challenging
- Experiment supported the idea a meaningful HASR test could be created with as few as 15 trials
  - Several potential sites were reluctant to do more
  - Limited statistical significance recognized
  - Decided to have both 15 trial (HASR1) and 150 trial (HASR2) tests

The SRE10 HASR Evaluation

HASR Protocols

- Trials consist of “training” and “test” speech segments
- Trials to be processed separately and independently (*Humans memory makes this difficult*)
  - Automated email used to submit each trial’s output before next trial was accessible
  - Unlimited listening (in whatever order) permitted for training and test data
- Human listeners could be one person or a panel
- A decision and a likelihood score were required for each trial
- Decisions could be made from:
  - A combination of automatic processing and human expertise, or
  - Solely based on human listening
- Scoring
  - Count number of Misses and False Alarms

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

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

37 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

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

9 such trials judged most similar were selected

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

30 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

Trial Selection - HASR2

- HASR1 trials were the first 15 HASR2 trials
- Speaker pairs selected as in HASR1, except that:
  - A threshold of 4 high scores was used (rather than 6)
  - Specific trial segments were then chosen at random for the 90 same-sex speaker pairs selected
- 45 low scoring target trials selected from the "full matrix" run
  - Human listening eliminated all trials with anomalous segments

HASR Participation

- 20 systems from 15 sites for HASR1
- 8 systems from 6 sites HASR2
- Most sites also participated in the main SRE10 evaluation
- Academic and government organizations from six countries participated

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 Participants & Results Summary

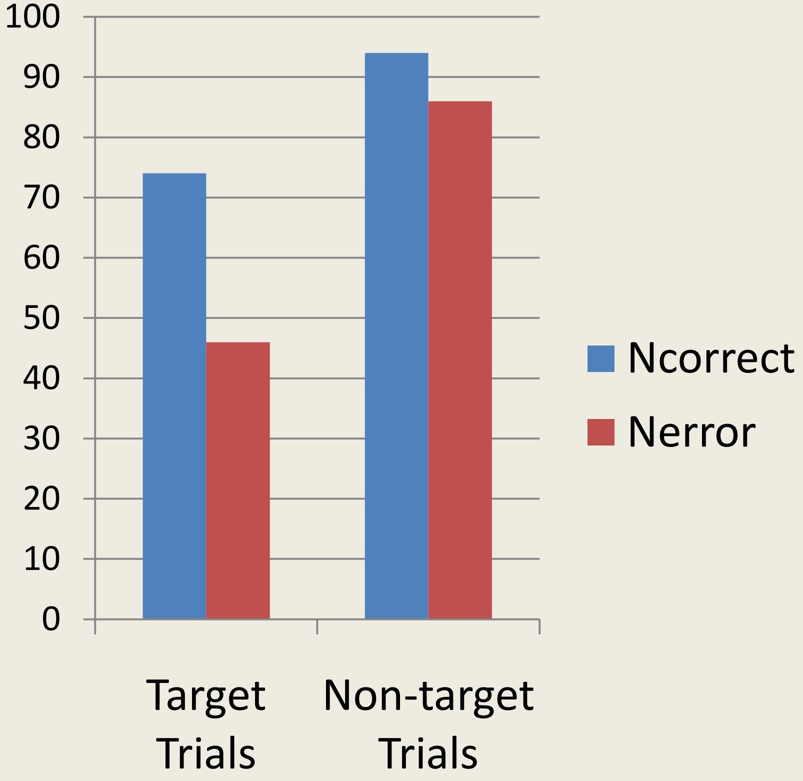
Site	Misses (51 trials)	FAs (99 trials)	Total (150 trials)
System 1			
System 2			
System 3	12	16	28
System 4			
System 5	25	37	62
System 6			
System 7	18	44	62
System 8			
System 9	8	61	69
System 10	13	57	70
System 11	2	75	77
System 12	30	42	72
System 13			
System 14			
System 15			
System 16			
System 17			
System 18			
System 19			
System 20	36	3	39
Total errors	*144*	*335*	*479*

Correct and Incorrect Decisions

Across All Participating Systems

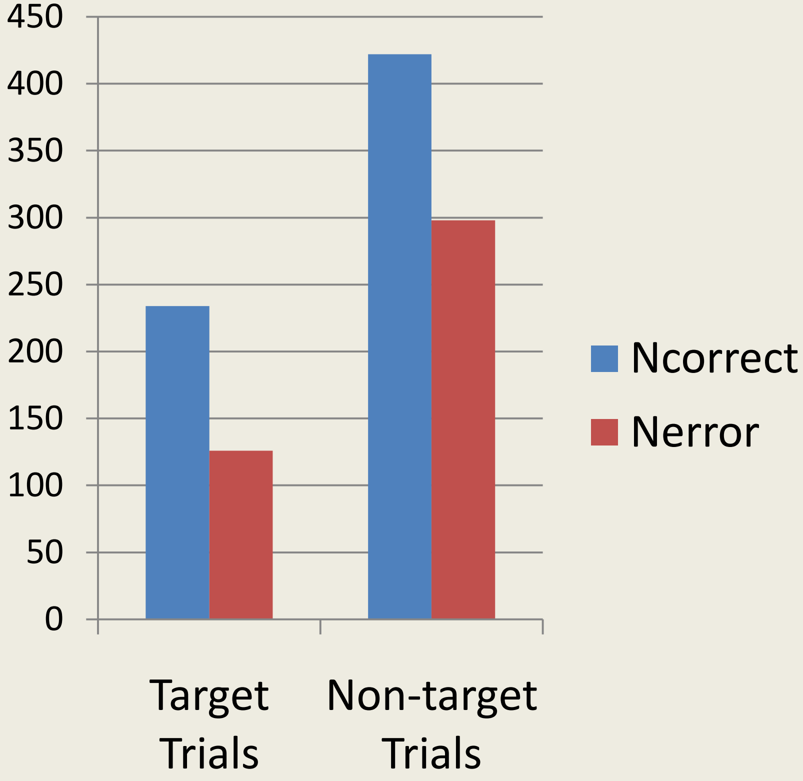
15 HASR1 Trials – 20 Systems

- Cumulative Miss Rate ≈ 38%
- Cumulative FA Rate ≈ 47%



135 Other HASR2 Trials – 8 Systems

- Cumulative Miss Rate ≈ 35%
- Cumulative FA Rate ≈ 41%

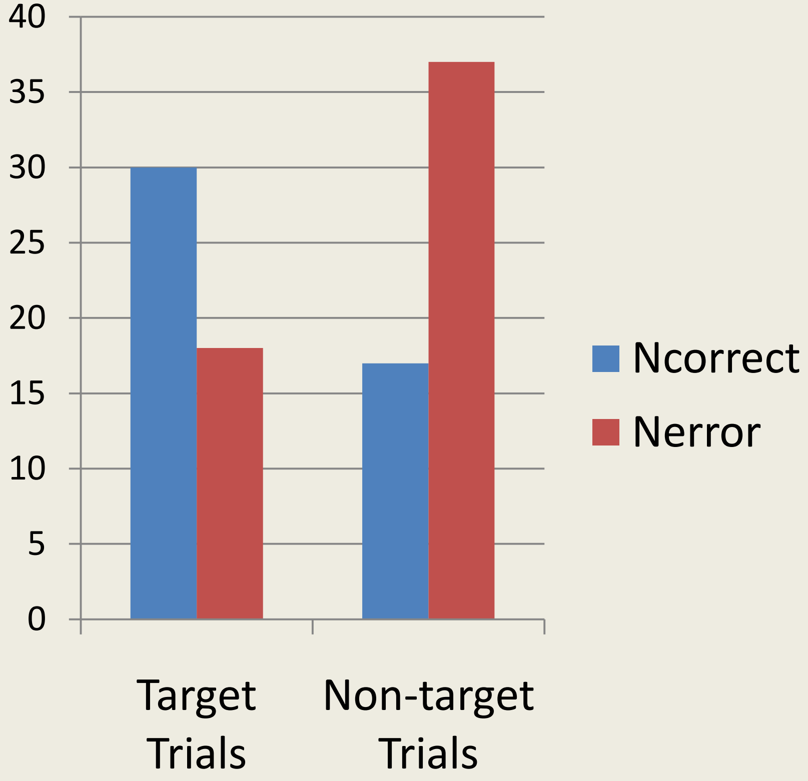


Correct and Incorrect Decisions

Across All Participating Systems

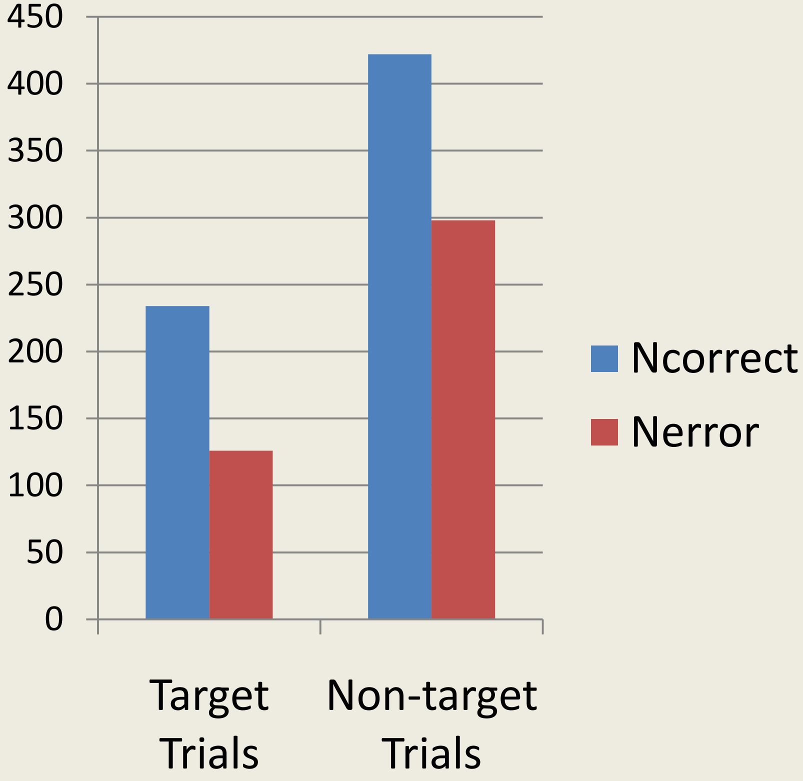
15 HASR1 Trials – 8 Systems

- Cumulative Miss Rate ≈ 38%
- Cumulative FA Rate ≈ 69%



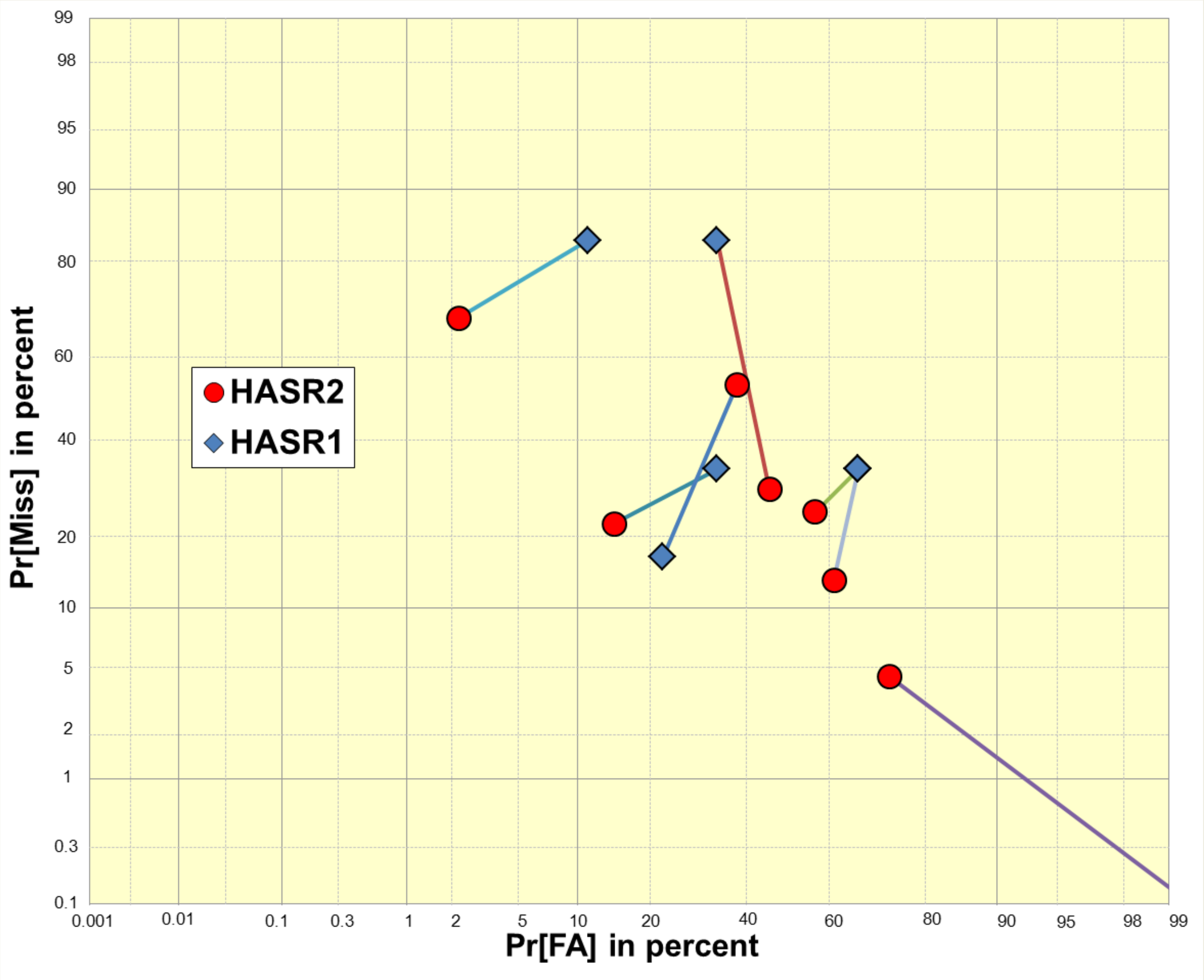
135 Other HASR2 Trials – 8 Systems

- Cumulative Miss Rate ≈ 35%
- Cumulative FA Rate ≈ 41%

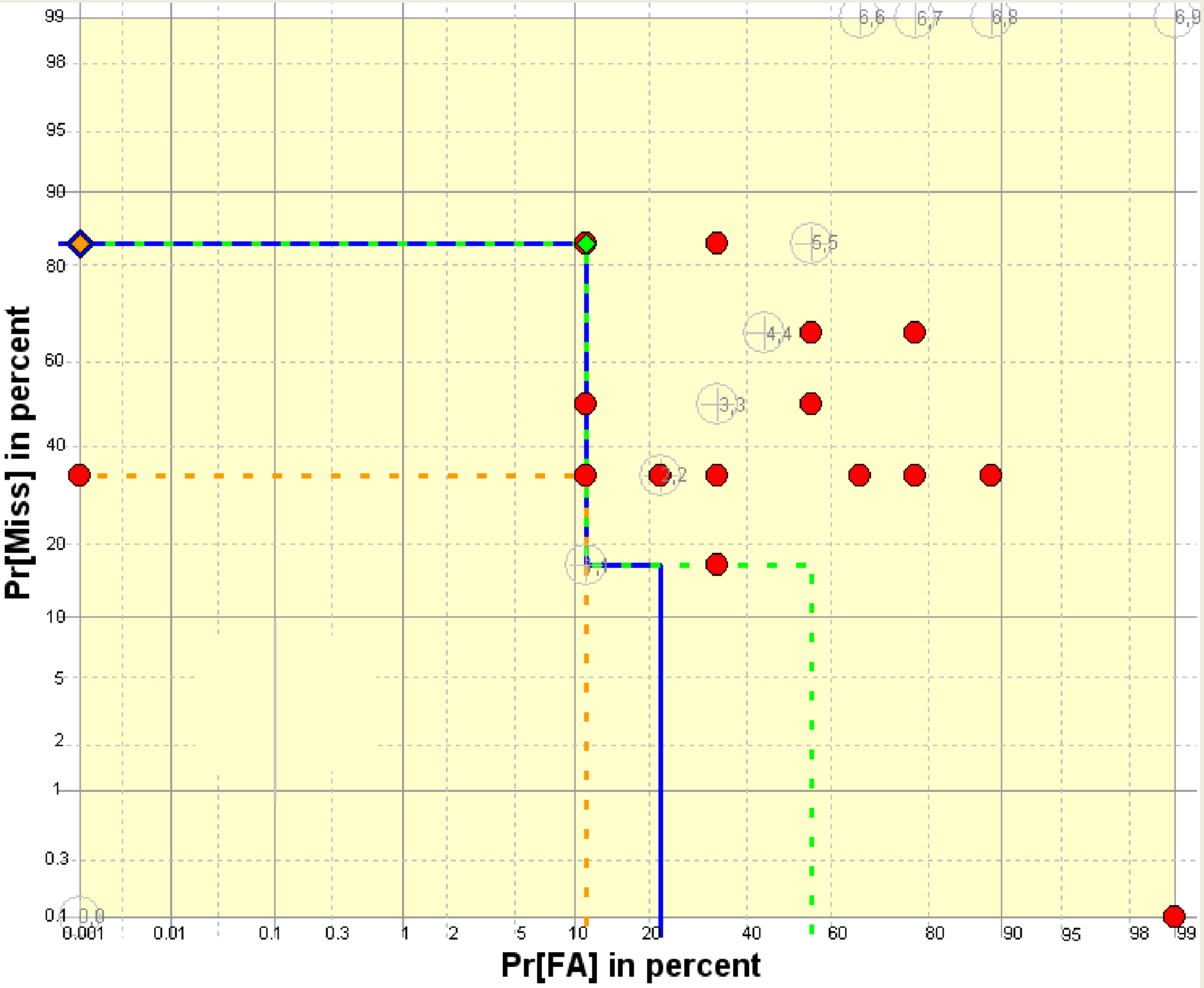


HASR2 Systems

HASR1/HASR2 DET Points



HASR1 DET Points and Curves

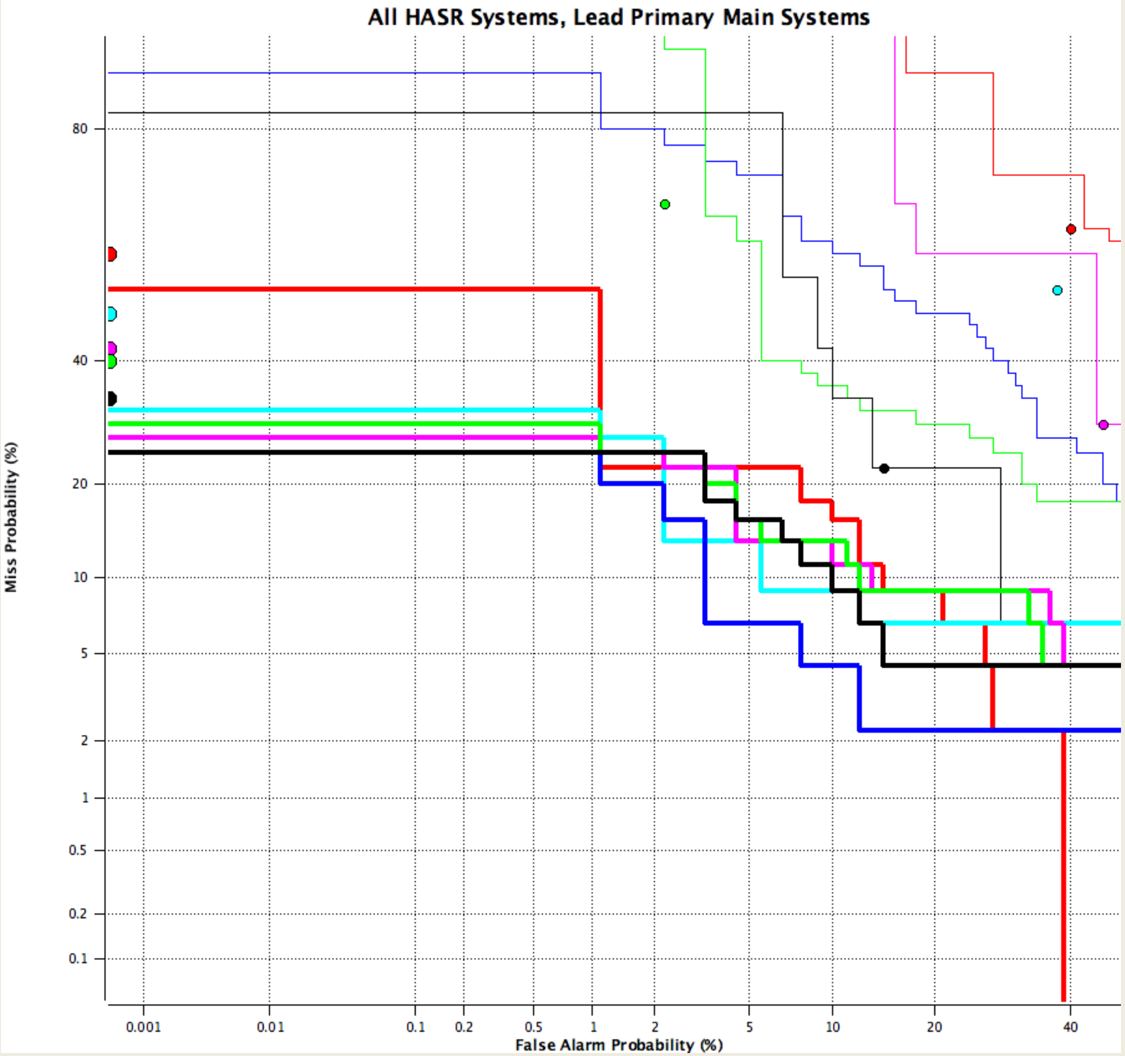


Red dots represent decision points (on DET plot) of HASR1 systems

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

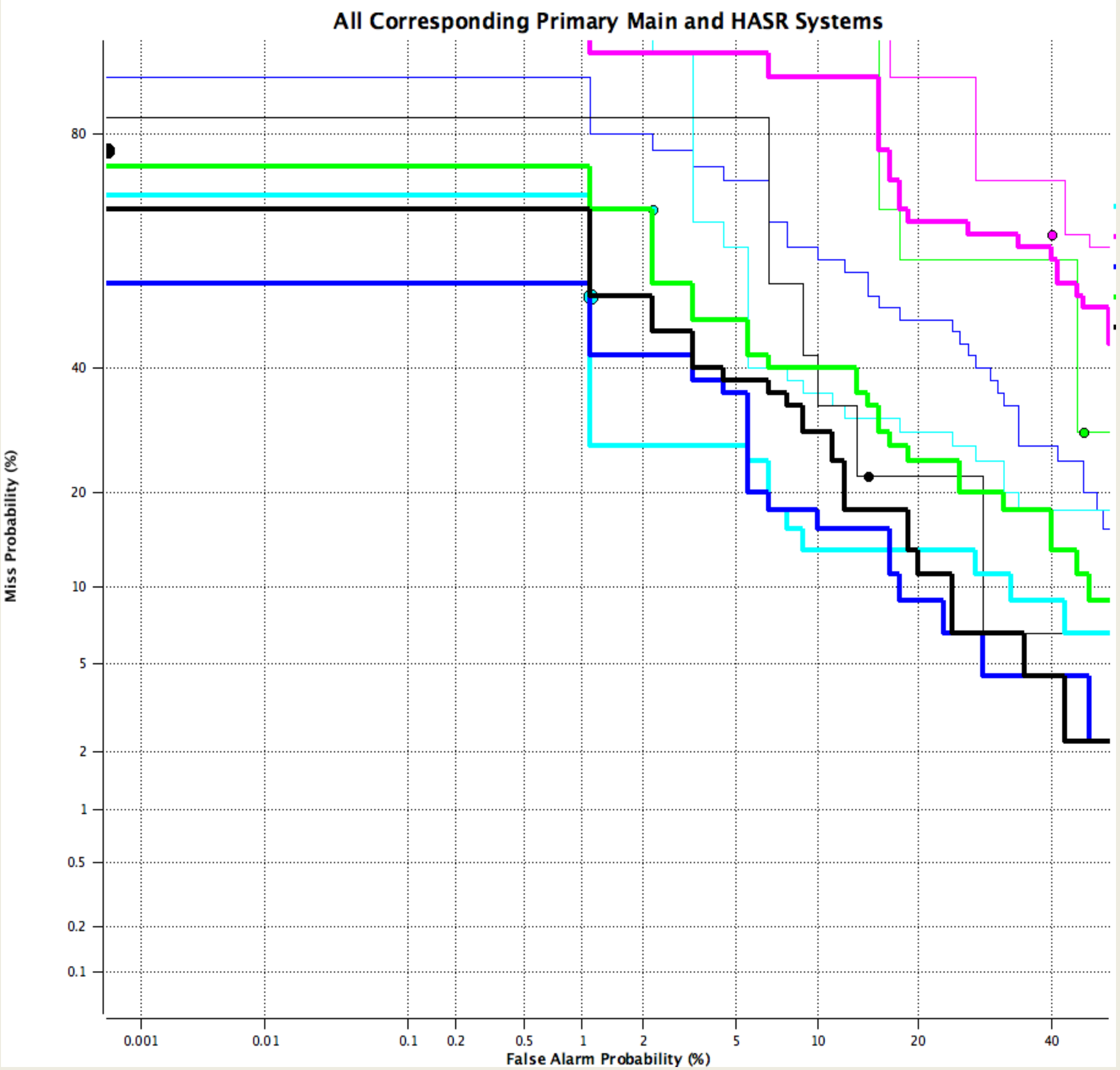
Leading automatic systems' DETs compare favorably with HASR decision points on HASR1 trials

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)

Summary

- Effort to select challenging HASR trials was successful
  - HASR2 trials only somewhat less challenging than HASR1 trials
- Performance of HASR systems did not compare favorably with that of automatic systems on HASR trials
- If HASR evaluation is deemed valuable, how should this pilot be extended?
  - Test protocol?
  - Trial selection?
  - Statistical significance?