# Human Assisted Speaker Recognition

CRAIG S. GREENBERG, ALVIN F. MARTIN,
MARK A. PRZYBOCKI

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY,
INFORMATION TECHNOLOGY LABORATORY,
INFORMATION ACCESS DIVISION

# NIST Speaker Recognition Evaluations (SRE)

NIST SRE measures speaker detection performance of state-of-the-art research systems on common test data

- Since 1996: sponsored by DoD, managed by NIST
- Open to participants worldwide
- Machine only: no listening or other human interaction allowed
- Recorded samples compared– may differ in channel and style, as in forensic/biometric apps:
  - Interviews and telephone conversations, many microphones

# The Speaker Detection Task

Given pairs of speech recordings:

- A "training" recording of 10sec, 5min, 8 min…
- A "test" recording of any such length
  - Telephone or microphone, conversation or interview
- Prior probability, and cost of miss and false alarm

System response, for each pair:

- Same voice: Y/N?
- How likely? (log likelihood)

# SRE 10 Evaluation Test Conditions

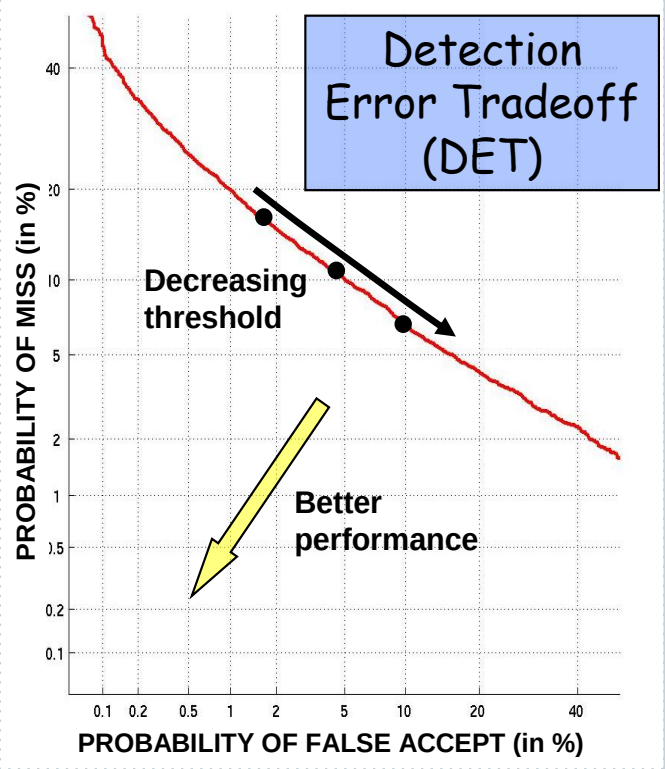
| | | Test Conditions | | |
|---|---|---|---|---|
| | | **10sec** | **5min (tel/mic)** | **summed channels** |
| Training Conditions | **10sec** | optional | - | - |
| | **5min (tel/mic)** | optional | required | optional |
| | **8conv** | optional | optional | optional |
| | **8conv summed channels** | - | optional | optional |

Number of trials: 31,387 - 610,748 per test condition
Number of speakers: 596
Data from the Linguistic Data Consortium (LDC)

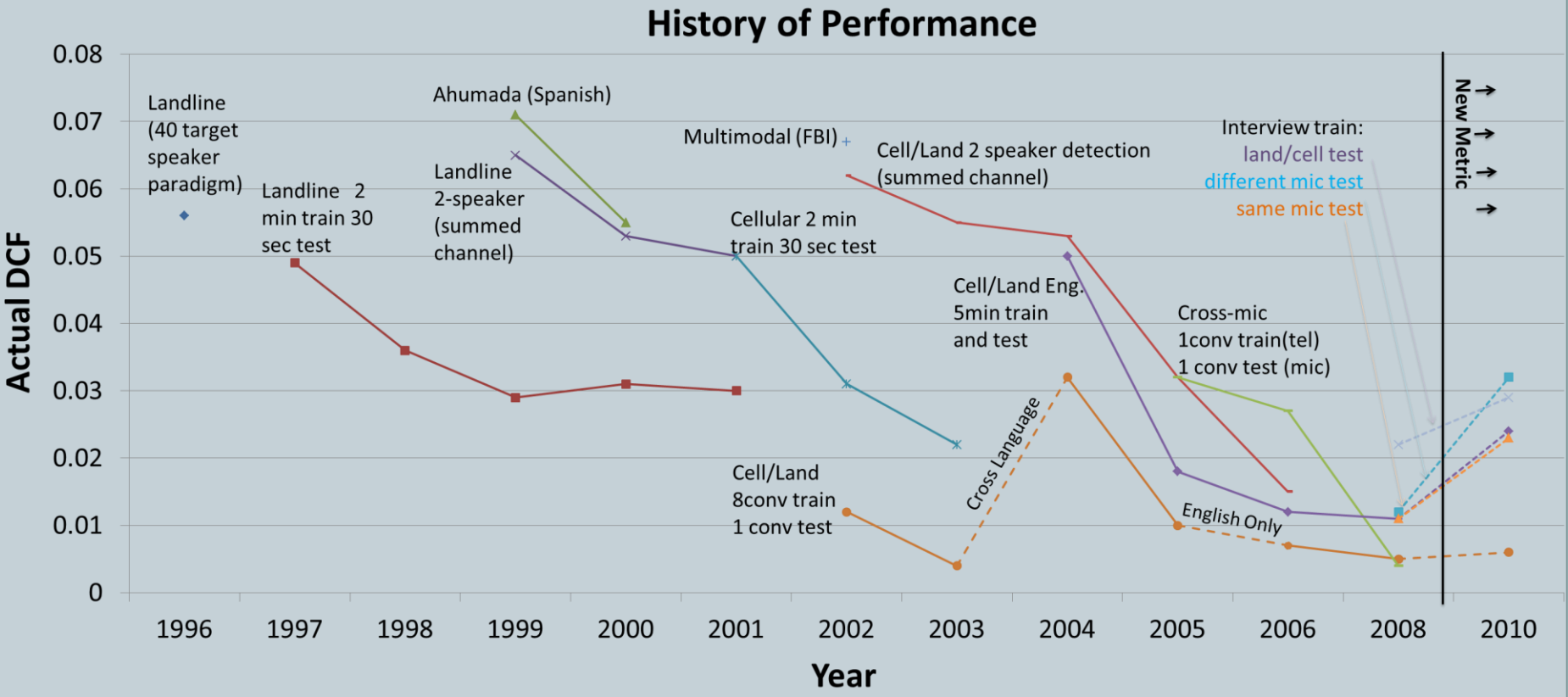
# Performance Metrics

Detection (not identification)

- False reject (miss): incorrectly reject a speaker
- False accept (false alarm): incorrectly accept a speaker
- Tradeoff made by decision threshold
- Measures:
  - Equal-error-rate (EER)
  - DCF
- DET Curve w/ all tradeoff points
- Example Figures of Merit:
  - %EER (easy to explain)
  - %FR @ .01%FA (forensic, military)
  - %FA @ 10%FR (access control)



# Why evaluate?
# SRE Performance History on Similar Tasks




Original Chart provided by Douglas Reynolds of MIT-Lincoln Laboratory

# Wow, that's great! Do humans even matter any more?

## **ALL** Speaker Recognition Applications Involve Humans!

- Forensic
- Biometric
- Watchlist
- …

# How can human experts effectively utilize speaker recognition technology?

- HASR (*Human Assisted Speaker Recognition*) began addressing this question – a 2010 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 | HASR2 |
|---|---|
| 15 trials | 150 trials |

- HASR systems may use human listeners, machines, or both
  - Participation open to all who might be interested

# Trial Selection

Trial: Pair of Speech Recordings (1 train, 1 test)

- Used “difficult” cross-channel trials
  - Training data from interviews included various room mic channels
  - Test data from phone calls included some with induced high or low vocal effort
- In-house baseline automatic system processed all possible cross-channel trials and the most difficult of those were selected for perception based sub-selection

# Sample Trials

Trial 1:

Trial 2:

# Sample Trials

Trial 1: DIFFERENT SPEAKER

Trial 2: SAME SPEAKER

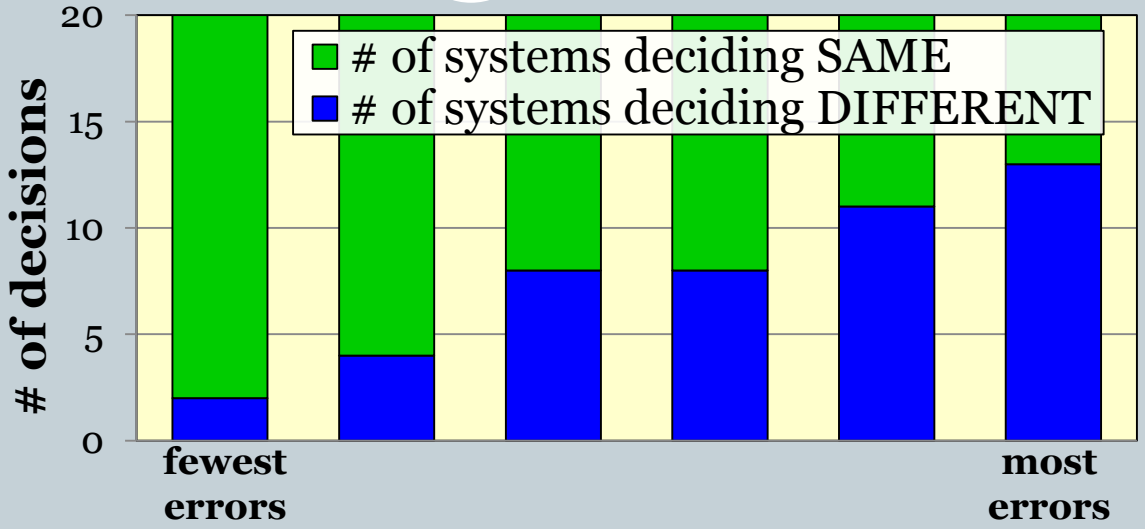
# HASR1 Results Summary

| | 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*** |

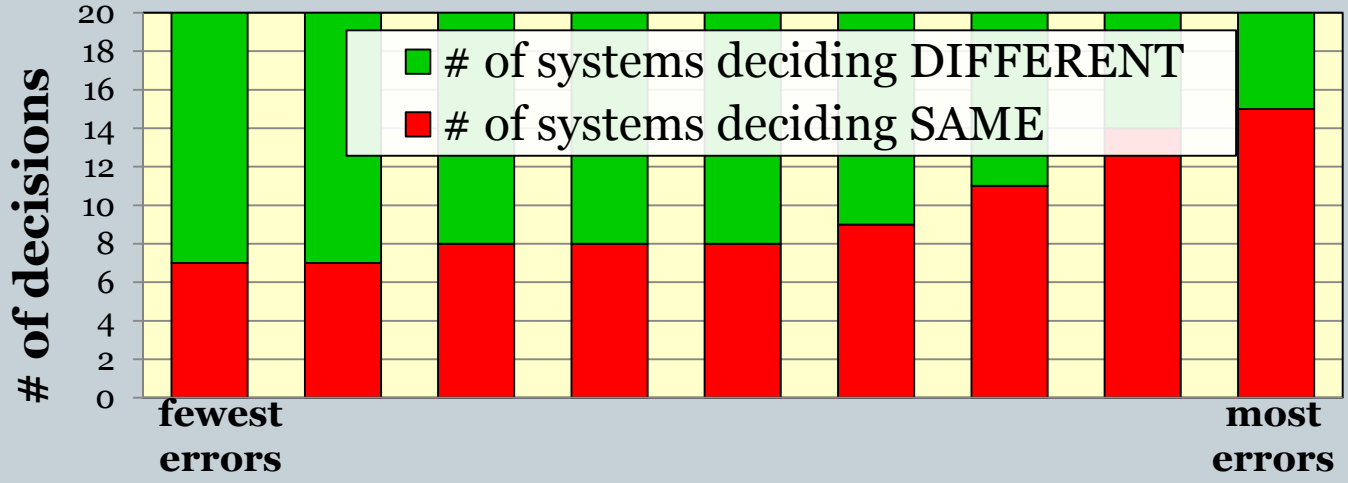
- Correct Accept
- Correct Reject
- Misses
- False Alarms

# Difficulty of 2010 HASR1 Trials

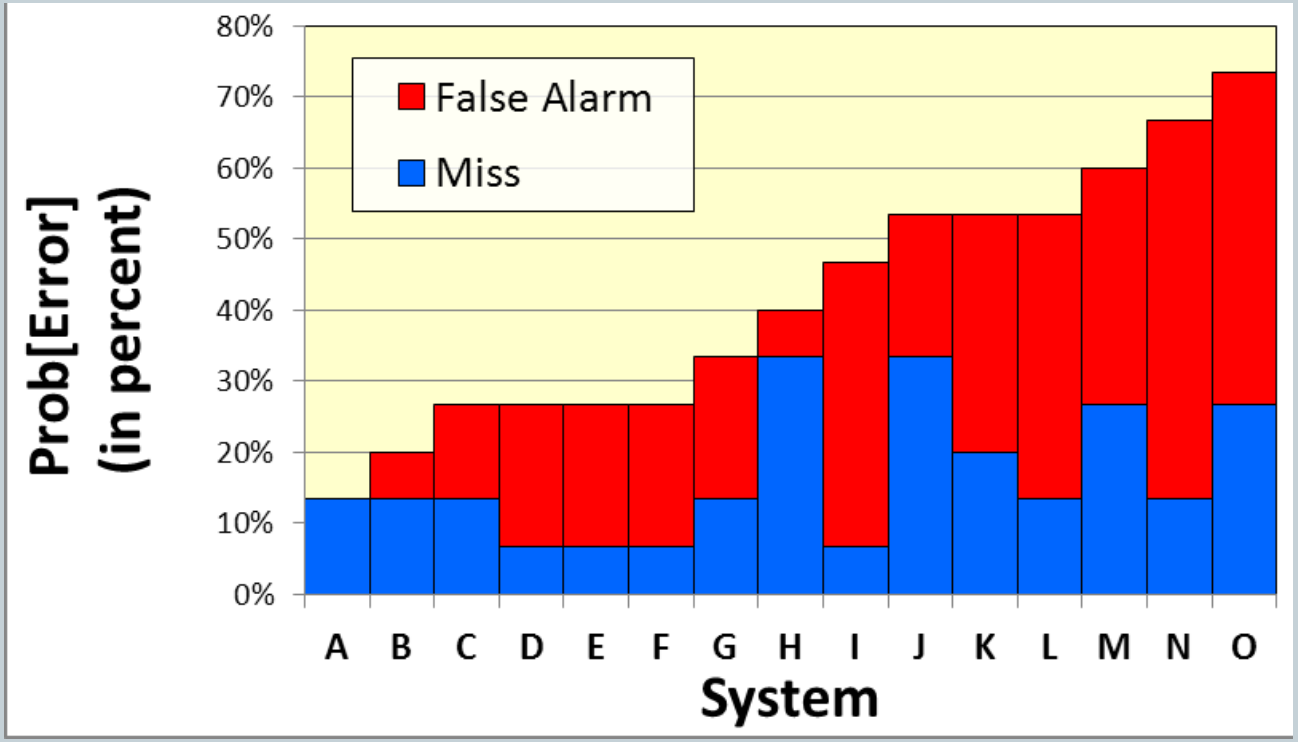
Target Trials




Non-target Trials



# HASR1 System Performance

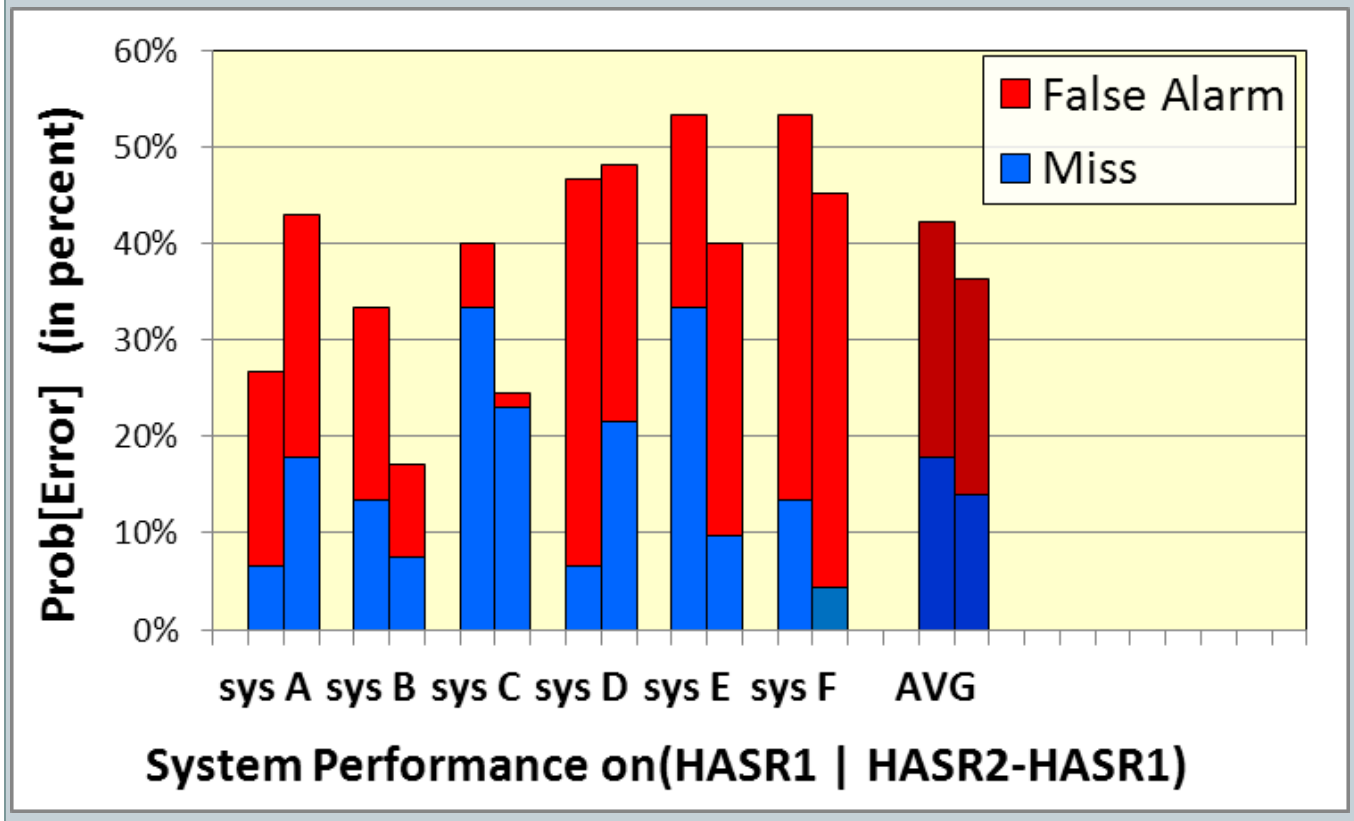



**fewest errors**

**Most errors**

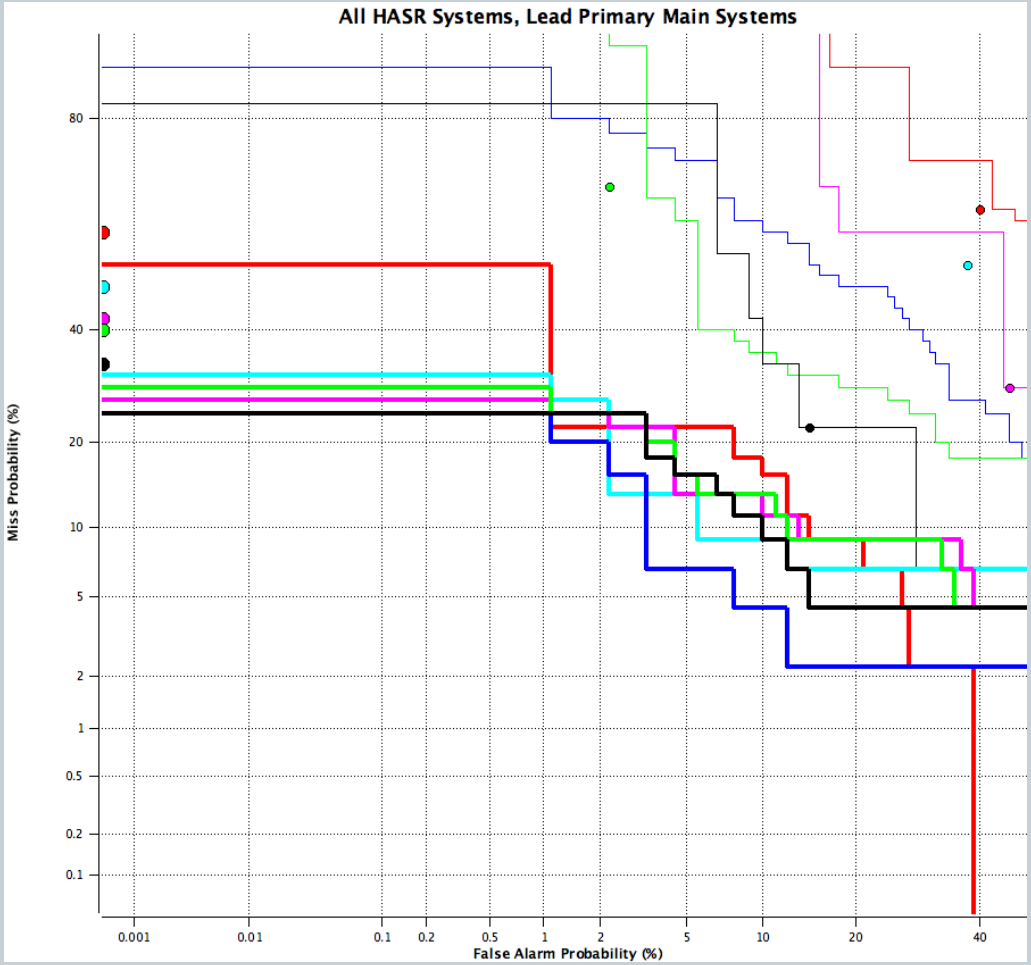
- All HASR1 Trials
- Best system per site

# System Performance on HASR1 and HASR2




- Bar on left shows HASR1 Performance
- Bar on right shows HASR2 – HASR1 Performance
- Results similar for HASR1 and HASR2

# HASR2 and Leading SRE10 Automatic Systems

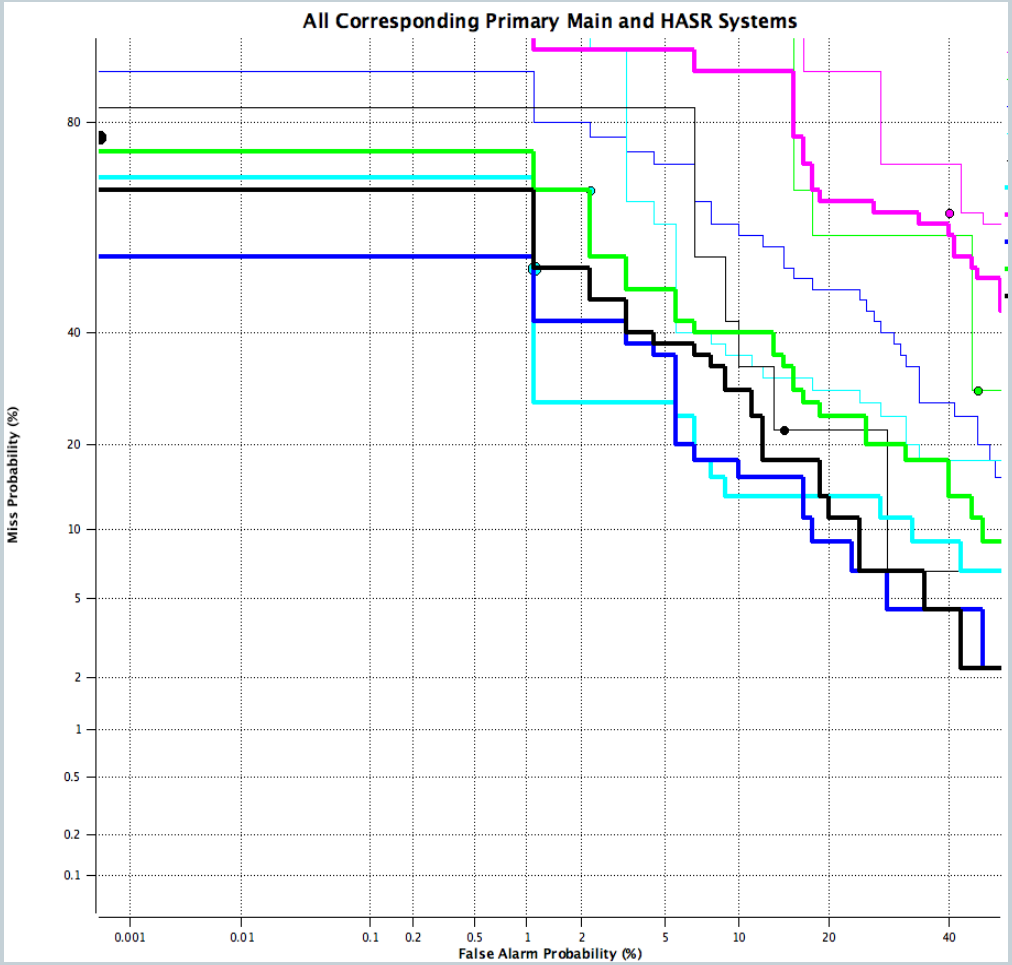



135 HASR2 trials

Six HASR systems (thin lines)

Six Automatic systems (thick lines)

# HASR2 and Corresponding SRE10 Automatic Systems




135 HASR2 trials

Five HASR systems (thin lines)

Five Corresponding Automatic systems (thick lines)

# Conclusions

- Humans are part of all speaker recognition applications
  - Understanding their capabilities and limitations is important
- Strong machine performance does not imply ready for deployment in any particular application
- The assumption that humans are superior to machines at speaker id needs to be qualified
- Spun off a whole line of research within the community
- More experiments planned