No Program?   No Problem!!

LDC, NIST evals, and a Golden Age of progress in Speaker ID

John J. Godfrey, PhD

Human Language Technology Center of Excellence

Johns Hopkins University

Outline

- Speaker ID (intro for non-specialists)
- The government's role in Speaker ID research
- The "virtuous cycle" of NIST evals/campaigns
- A report card on the last decade
- Reflections on LDC's role
- Discussion, questions

Terms of reference: SpeakerID – applications*

- Text dependent ("Open Sesame") v. *Independent*
- Identification *v. verification* (detection paradigm)
- Cooperative v. *uncooperative* speakers
- Limited v. *extensive* data (minutes v. seconds)
- Biometric v. forensic v. *surveillance* applications
  - Boundaries among these are poorly defined
  - Priors differ wildly, and matter a lot
  - Scale is also important; large numbers much harder

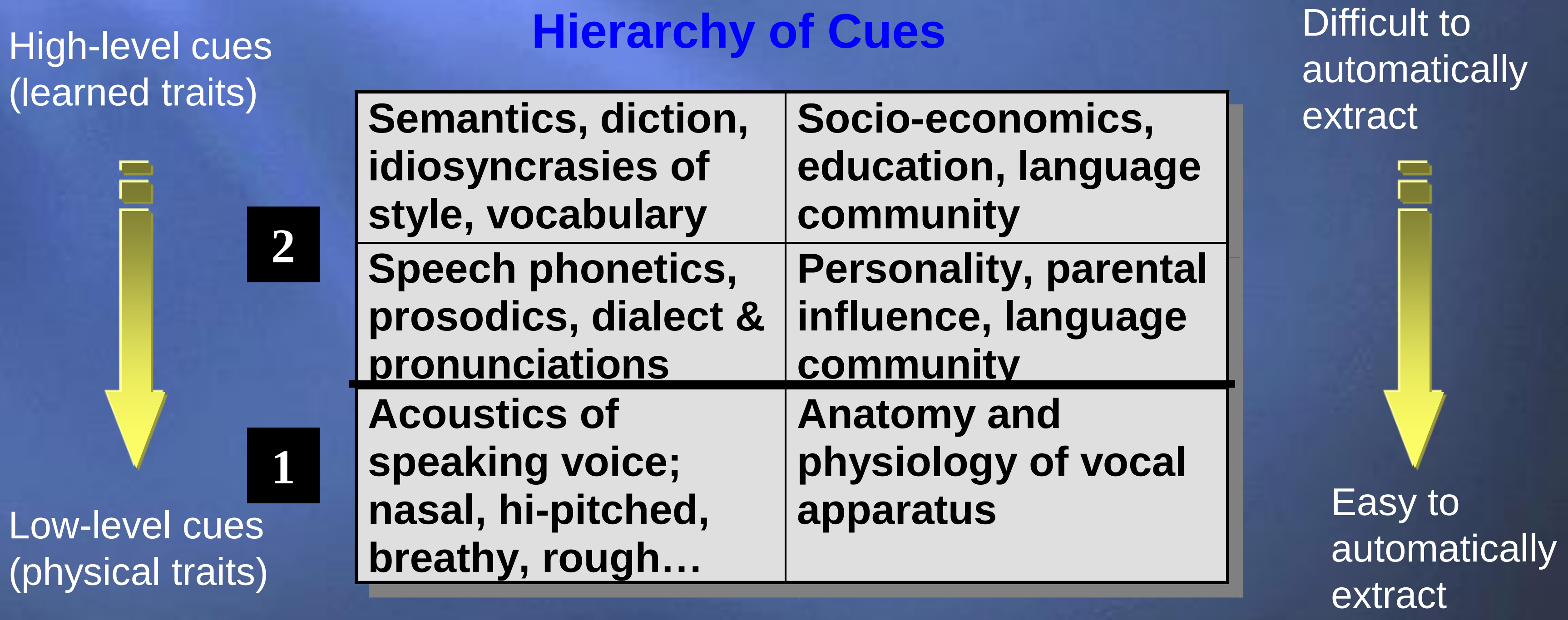
* Underlined terms are of greater interest for most government apps

Terms of reference: SpeakerID – sources of variance

- Research goal is to find features and algorithms to:
  - Maximize interspeaker variance
  - Minimize intraspeaker variance
  - In circumstances that *represent* some application(s)
- Sources of variance are either/or:
  - Extrinsic – due to environment, noise, reverb, channel, coding...
    - Hard engineering problems, but relatively well understood
  - Intrinsic – due to anatomy, physiology, psychic state, behavior…
    - Less understood; may require basic research
- For applications to succeed, both must be addressed

Terms of Reference: SpeakerID – sources of information

- **Humans use many perceptual cues for speaker recognition**



Hierarchy of Cues

High-level cues (learned traits)

Low-level cues (physical traits)

Difficult to automatically extract

Easy to automatically extract

2	**Semantics, diction, idiosyncrasies of style, vocabulary**	**Socio-economics, education, language community**
	Speech phonetics, prosodics, dialect & pronunciations	**Personality, parental influence, language community**
1	**Acoustics of speaking voice; nasal, hi-pitched, breathy, rough…**	**Anatomy and physiology of vocal apparatus**

- **Speaker cues mostly inseparable from speech cues**
- **Low-level cues most effective in current automatic systems**

Sources of Information (2)

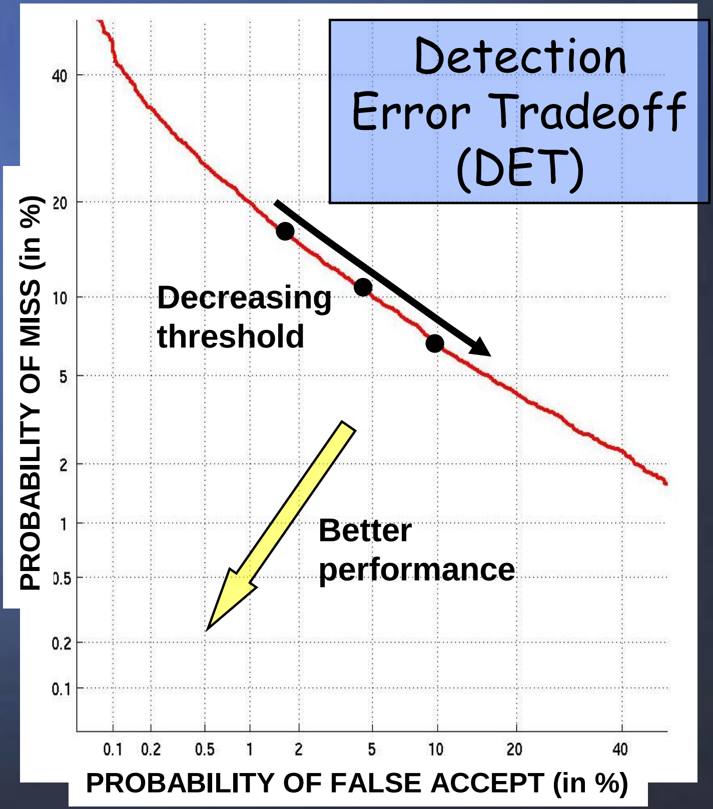
A few words about "high level" features (lexical, prosodic, phonetic)

- Lots of research last decade (since 2002 "SuperSID")
- ASR-based methods are now affordable, available
  - Even lexical features can be used
- Theoretically, can be more robust to signal problems
- But still not viable w/out low-level "acoustic" features
- Fusion succeeds (in research systems)

Research: Performance Metrics

Detection (verification, 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) (%FR = %FA)
  - DCF (C1*%FR + C2*%FA)
- Usually plot DET Curve w/ all tradeoff pts
- Examples of research Figures of Merit:
  - %EER (the PM's friend)
  - %FR @ .01%FA (forensic, military)
  - %FA @ 10%FR (access control)



Outline

- Some terms of reference (for non-specialists)
- The “virtuous cycle” of NIST evals/campaigns
- A report card on the last decade
- Needs, opportunities, seedlings
- Discussion, questions

The Government Role

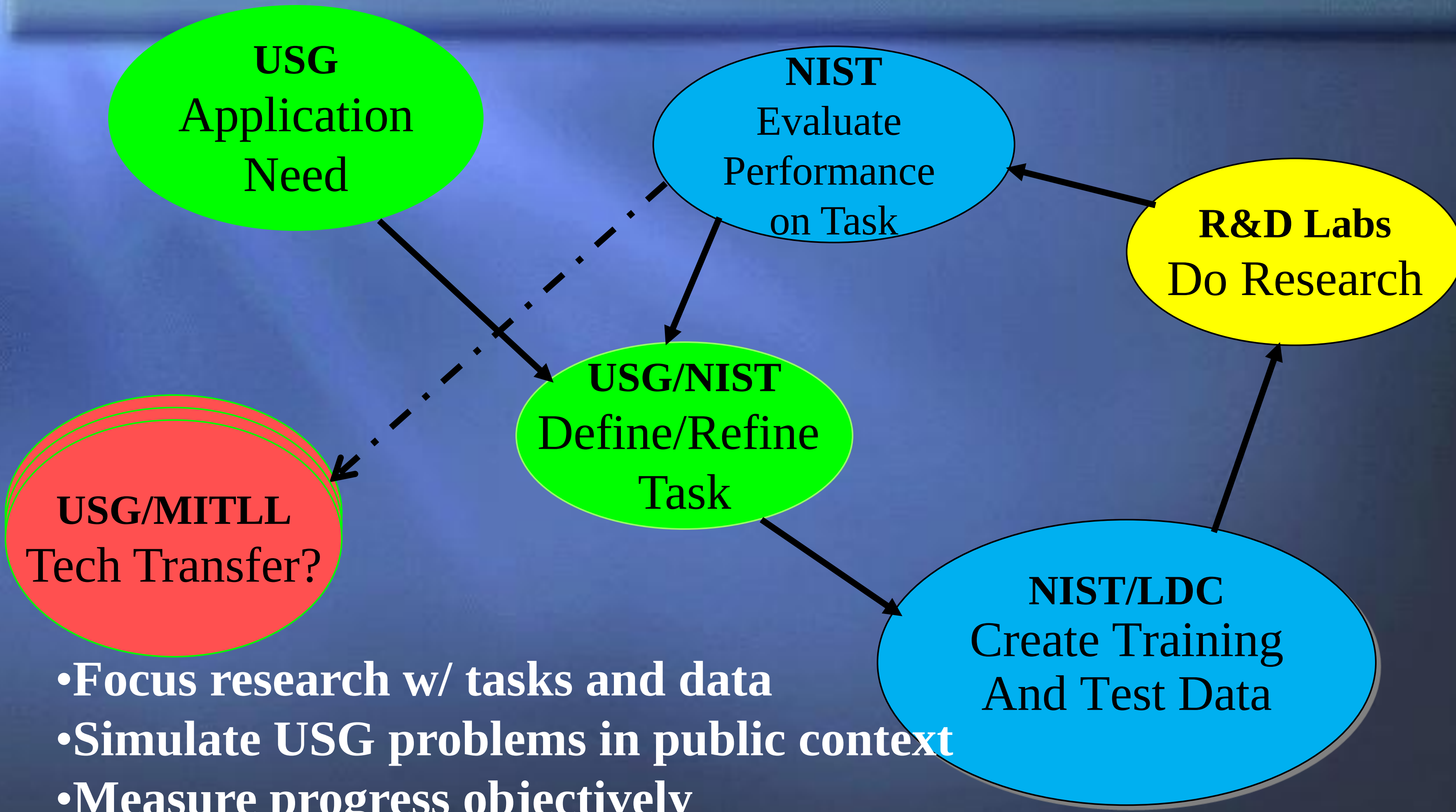
- Mission: Technology for the common good
- R&D model: Top-down (DoD) vs. Bottom-up (NSF)
- DoD often uses sponsored programs (e.g., DARPA or IARPA) to develop prototypes or demos
- For SID/LID, ever in the shadow of ASR, only 2 such programs in 20 years☹
- Yet application needs exist at several agencies
- How to accelerate progress without big $$?

No Program? No Problem! Just hold a bake-off (ie, a NIST Eval*)!

- NIST Evaluates: *NOT* products or apps; but solutions to problems *abstracted from* real apps
  - balance realism vs. generality
  - Ex: Forensic, “biometric,” surveillance, watchlist, ….
- Hard tasks, free data, good metrics, hope for $$
- Prestige earned from years of ASR, MT, TREC, etc
- Frequency + rules of participation elicit friendly competition, sharing by academics and industry

*http://www.itl.nist.gov/iad/mig//tests/sre

The NIST Evaluation Paradigm: a virtuous cycle



USG
Application
Need

NIST
Evaluate
Performance
on Task

R&D Labs
Do Research

USG/NIST
Define/Refine
Task

USG/MITLL
Tech Transfer?

NIST/LDC
Create Training
And Test Data

- **Focus research w/ tasks and data**
- **Simulate USG problems in public context**
- **Measure progress objectively**
- **Demonstrate SOA to govt**

A typical cycle: 2008

- New Application Need: forensics/biometrics have mismatch in channel + style, e.g., hi-quality mike interview vs. cell phone conversation
- New LDC Data: 1350 spkrs; 14 mikes + phone calls
  - Includes interviews and "captive" phone calls
- Task: for each pair of audio segments, answer "same/different", and assign a probability
  - Six training, 4 test conditions; 13 tests; ~.5M trials

Training/Test Conditions

Given:

- A “training” segment of length10sec, 5min, 8 min, or more
- A “test” segment of any such length
- From telephone or microphone, conversation or interview
- Prior probability, and cost of miss

Respond, for each such pair:

- Same voice: Y/N?
- How likely?

Number of trials: ~100,000 per test condition

Number of speakers: ~ 1350

13 Evaluation Test Conditions

Test→ Train↓	10-s tel	5-min tel or mic	8 min mic	1 tel conv summed
10-s tel	optional			
5-m tel/mic	optional	required		optional
3conv tel		optional		optional
8conv tel	optional	optional		optional
8-m mic		optional	optional	
3conv,sum		optional		optional

A phone conversation trial

– Segment 1 Segment 2

🔊 ↔ 🔊

Interview Train, Phone Test trial

– Segment 1

🔊

Segment 2

🔊

Participation in SRE08

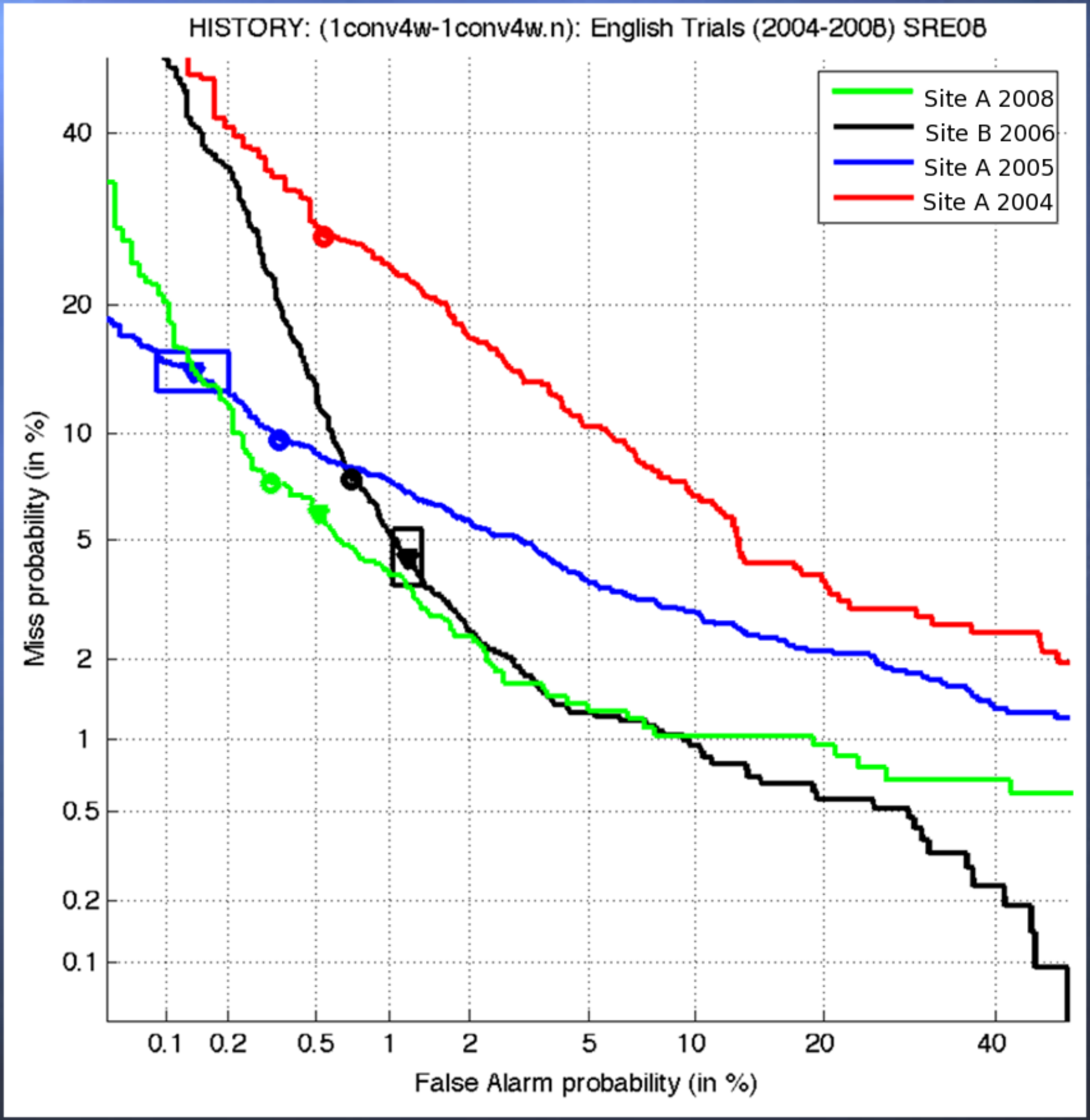
- 46 sites from 5 continents
- 107 systems
- 246 test condition/system combinations
- 1 “mothballed” system

See:

http://www.itl.nist.gov/iad/mig//tests/sre/2008/official_results/

Performance example (2008 vs 2004)

One sites results, on one condition, with previous years for comparison

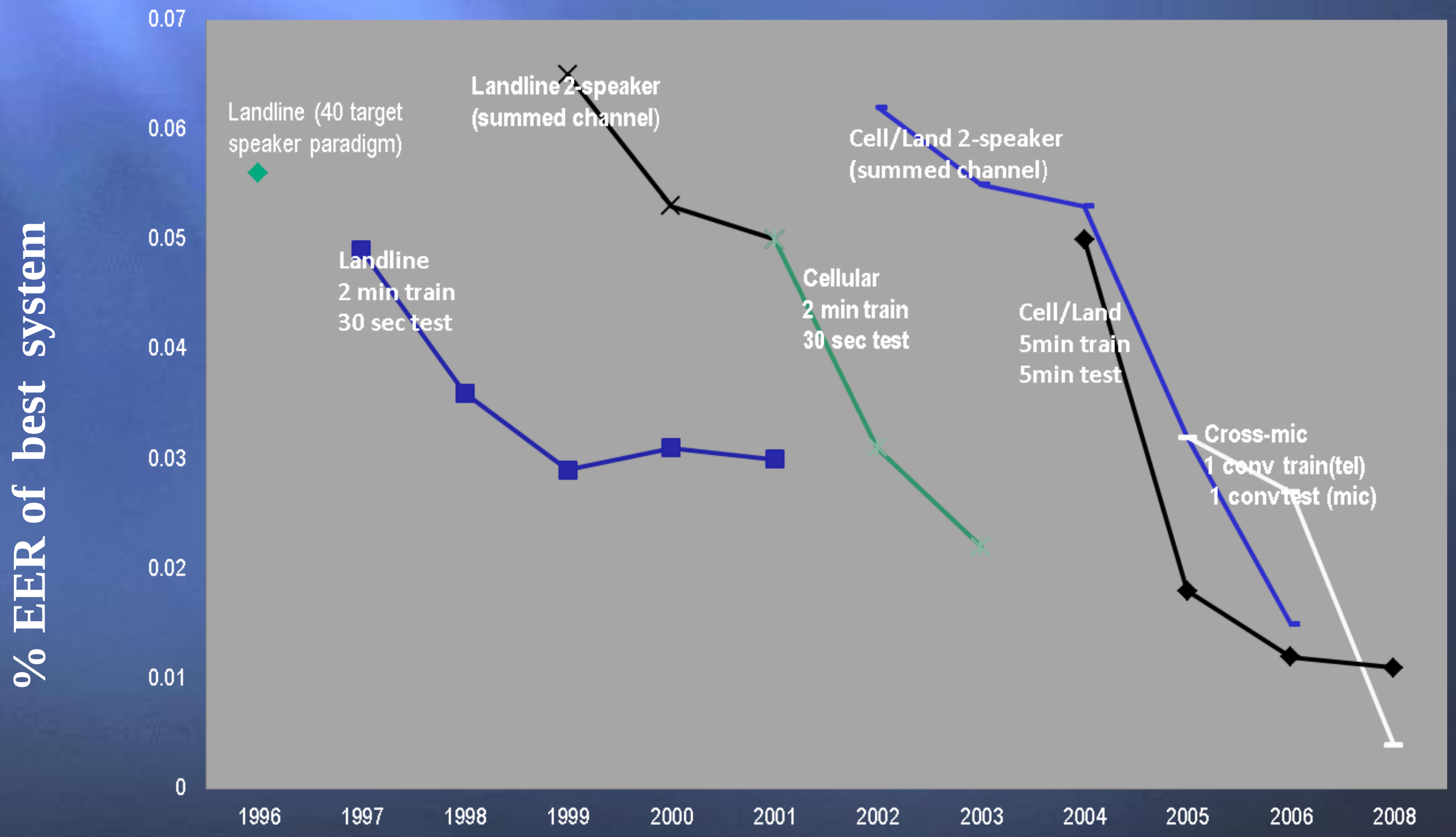


Circle:
Min DCF

Triangle:
Actual DCF

Box:
95% CI

The NIST view: 15 yrs of new conditions, continuous progress



A researcher's view: A decade of MIT-LL "results and reasons" on SID-CTS

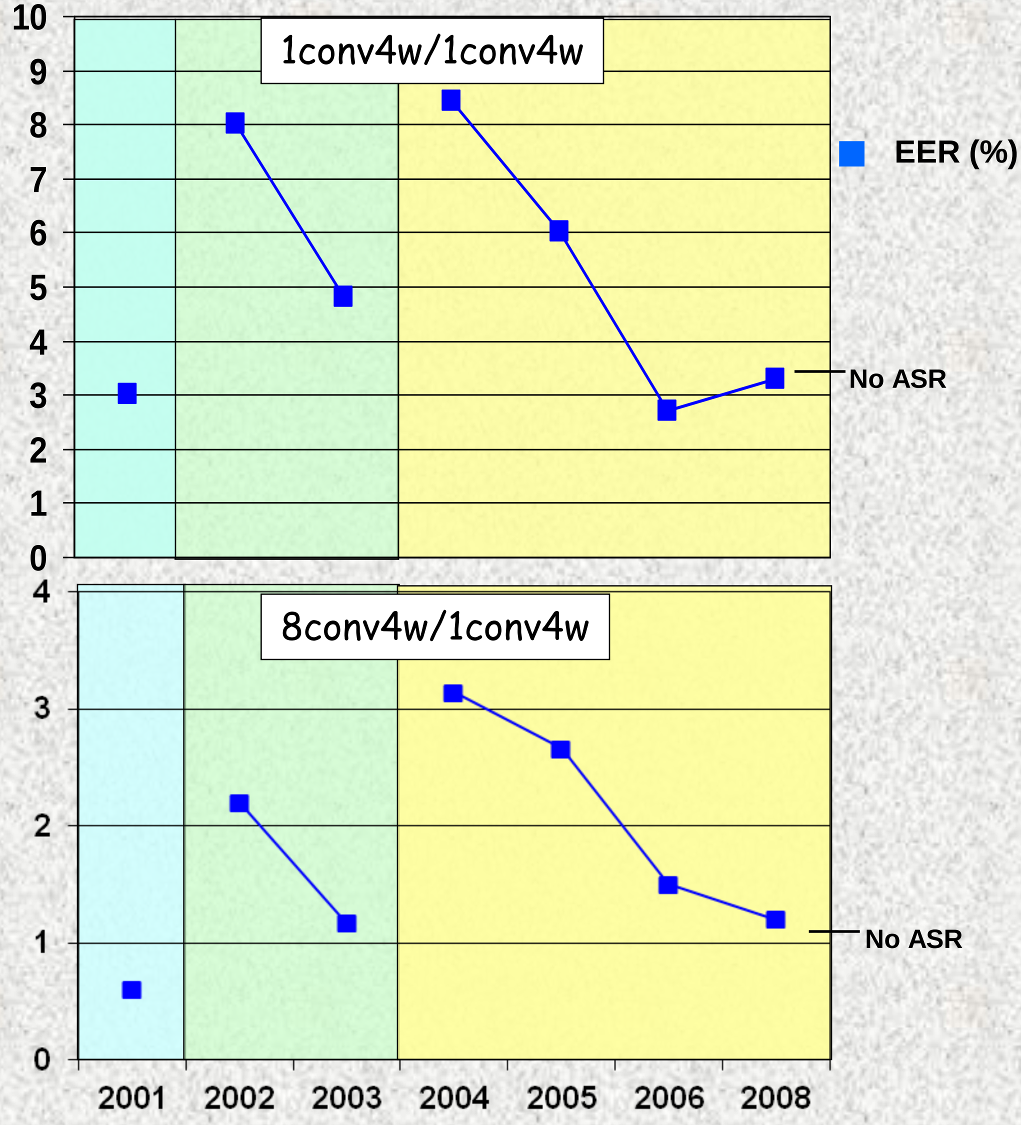
Consistent and steady improvement for data/task focus

2001	2002	2003	2004,2005,2006,2008
SWB1	SWB2		MIXER2-3

New data sets with more challenging conditions

New features, classifiers and compensations drive error rates down over time

2001	Text-const GMM, word-ngram
2002	SuperSID : High-level features
2003	Feature Mapping, SVM-GLDS
2004	Phone/Word-SVM, GMM-ATNORM
2005	NAP, TC-SVM, word/phone lattices
2006	SVM-GSV, GMM-LFA, MultiFeat SVM-GLDS, SVM-ASR-MLLR+NAP
2008	SVM-GSV, GMM-LFA, SVM-TOK-MLLR, SVM-TOK, SVM-KW



A word about common corpora

- Typically, too expensive for any one site
  - Large and realistic – for meaningful statistics
  - Truth-marked to a high standard (1 in 10^6?)
  - With proper evaluation sets and controls
- NEW: Multi-phase, backward-compatible corpora and dual-use corpora have enormous impact over years (e.g., SWITCHBOARD 1-5; MIXER 1-7)
  - Thousands of voices, publicly available, truth marked, in many languages & recording conditions
  - Open up new avenues of research (age; menagerie)

Outline

- Some terms of reference (for non-specialists)
- The “virtuous cycle” of NIST evals/campaigns
- A report card on the last decade
- Needs, opportunities, seedlings
- Discussion, questions

SRE Report Card

- Research progress, by all metrics, has been steady and impressive (a “golden age”?)
  - Errors halved about every 2.5 yrs for the last 15 yrs
  - EERs < 2% on cell phone speech; also crosschannel
- For so little investment, this is an A+ outcome!
- What about real applications?
  - USG has developed applications that work
    - But BEST program for “biometrics” was ended early
  - Companies have sprung up, but outcomes not yet clear

Where's Waldo: Why big apps are hard

- NIST Evals give detection rates *for any one target speaker,* over a realistic but controlled population
- BUT for SID: every 10x in #targets *doubles* EER ☹
  - Aminzadeh and Reynolds, 2008
- AND: important sources of variability still untested
  - Noise, reverberation, etc.
  - Stress, speaking range, physiological state, age
  - Vocal modalities (whisper, shout, disguise)
- AND: assumptions about humans – do they matter?

Do humans matter? – the HASR results

- SRE10 had a new “fun” task – let people listen
- Three motives:
  - – Forensic needs, claims (small sample, high accuracy)
  - – Large apps w/ human in the loop (triage, e.g.)
  - – Inspiration (how, and how well, it is done by humans)
- The results shocked most people
  - – Headline: “Machines better than humans!”
  - – Actually, there’s a lot more to learn

HASR: Trial Selection

- Difficult trials from SRE08 chosen (forensic-like?)
  - One segment: from interview, good microphone (3 min)
  - Another: from telephone call (~5 min)
- HASR selection procedure
  - Segment-pair similarity per SRE08 eval scores
    - Most-similar different-speaker pairs selected for “different” trials
    - Least-similar same-speaker segments selected for “same” trials
  - Pairs screened aurally to eliminate content cues

HASR1 Trial Examples

- Example 1
  - Segment 1 Segment 2
- Example 2
  - Segment 1 Segment 2

HASR1 Trial Examples

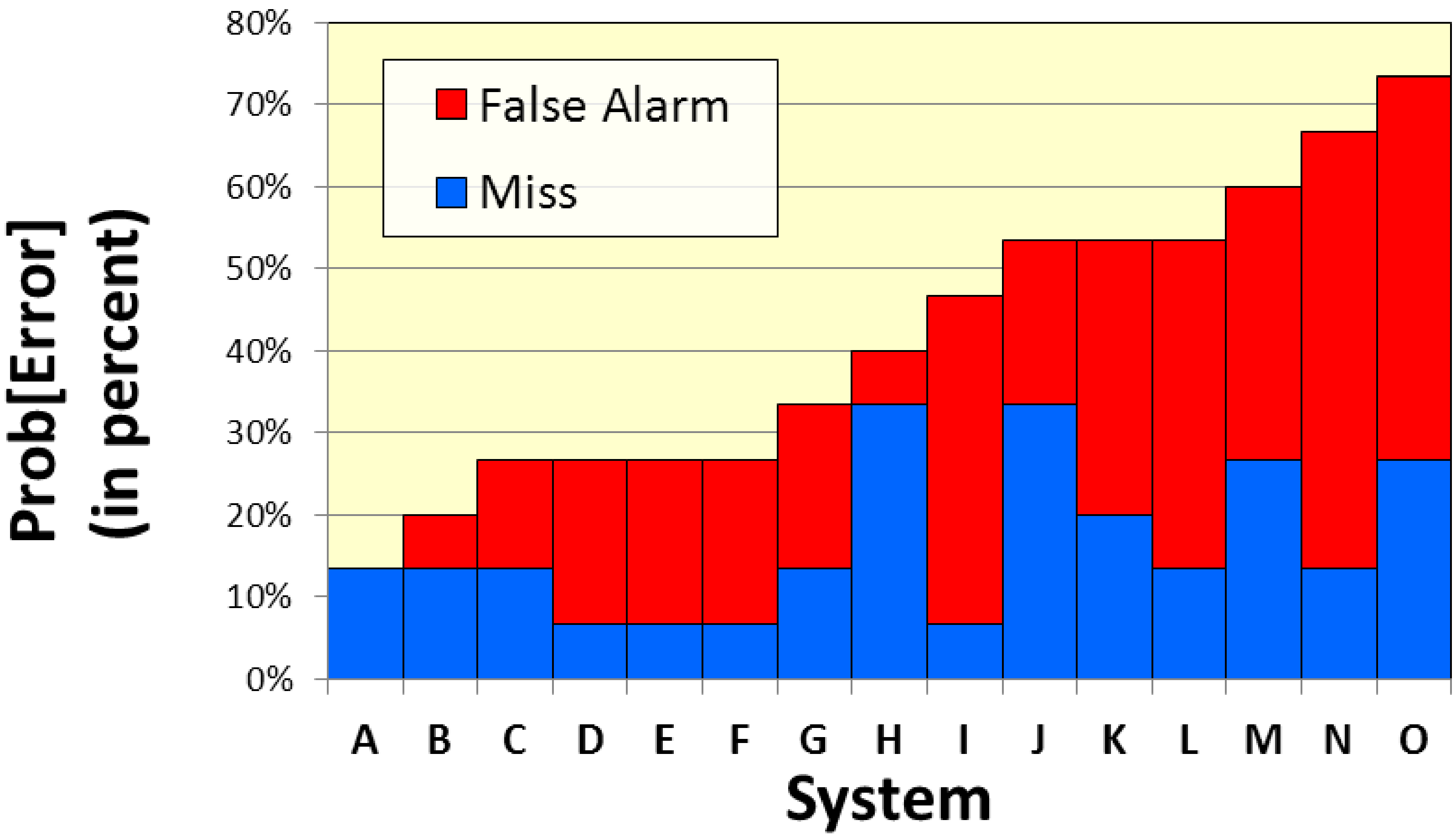
- Example 1
  - Segment 1 Segment 2

Different Speaker

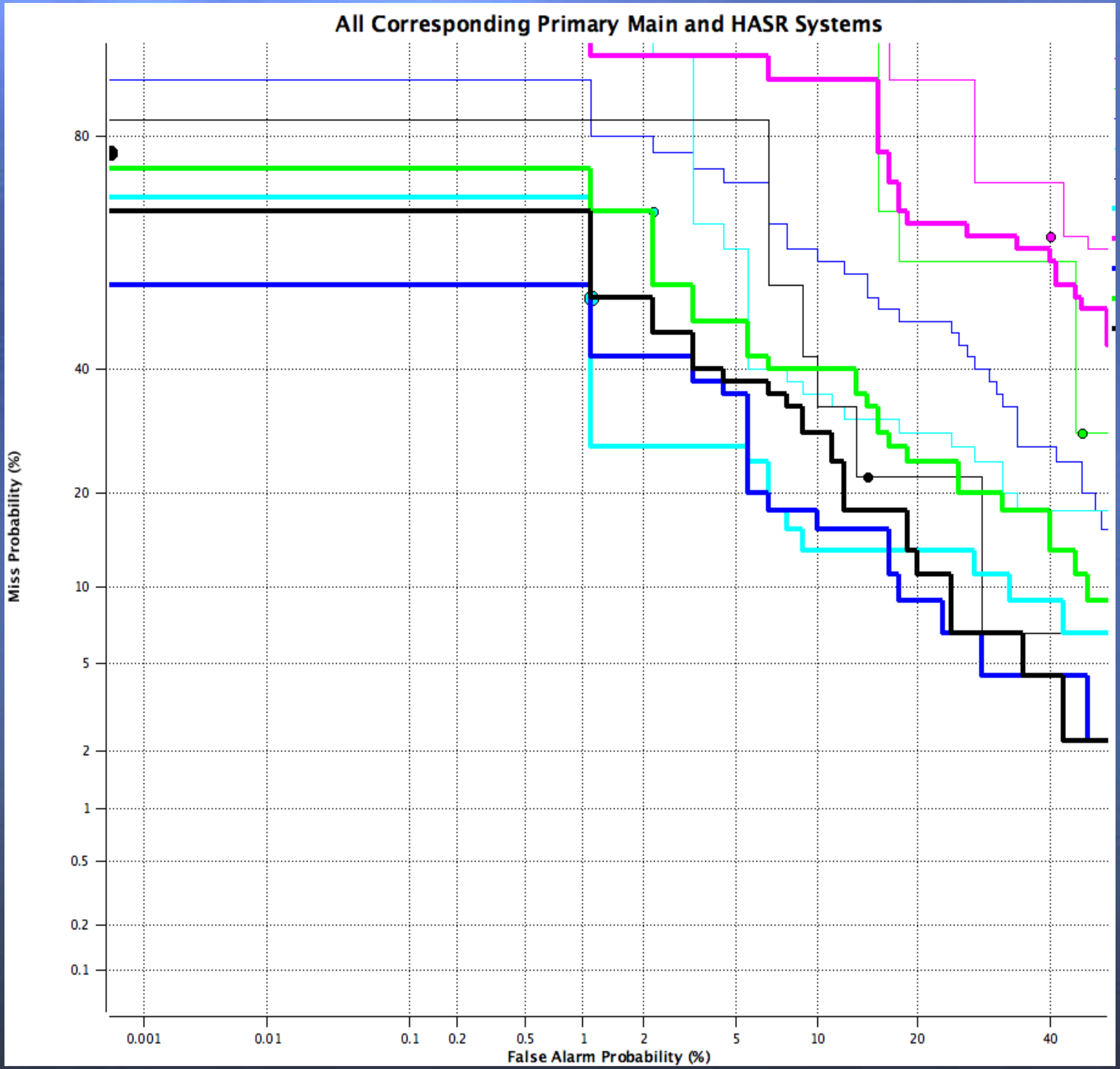
- Example 2
  - Segment 1 Segment 2

Same Speaker

HASR1 System Performance



HASR2 and Corresponding Automatic Systems



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

HASR Summary

- HASR system performance (human, or man-machine) did not compare favorably with that of automatic systems
  - Half the systems got more trials wrong than right in HASR1
- The test set was challenging, but not unrealistic
- Many questions about man-machine performance
- Another HASR planned for SRE 2012:
  - http://nist.gov/itl/iad/mig/hasr.cfm
- Let’s get this right before someone gets hurt!!

Outline

- Some terms of reference (for non-specialists)
- The “virtuous cycle” of NIST evals/campaigns
- A report card on the last decade
- **Reflections on LDC’s role**
- Discussion, questions

Big Data and SID Research

- Since SWITCHBOARD, the community has known the importance of large, well-documented SID data sets
  - # of voices; # of trials; # of conditions; etc.
- Unlike longer-term high-dollar programs, sponsor and NIST can change the task as often as progress demands
  - But the sponsor needs continuity, too!
- The new data, with new technical challenges almost annually, becomes the driver of research and progress
- The Challenge: balance new and old requirements, and get the contract to LDC on time!!

LDC and SID Research

- The SID corpora have been a collaborative effort: LDC, NIST, sponsors, other experts
- The MIXER collection(s)
  - An attempt to preserve continuity across a decade
  - New task definitions: introduced languages, speech styles (interview, telephone, read), demographics, mikes, channels, stress (Lombard), noise, reverb….
  - Kept basic conversational paradigm, recruiting, documentation, etc., nearly unchanged
  - Grew speaker population to ~1000; most recordings still comparable in evaluations

LDC Data and SID Research (2)

- Above and Beyond applications:
  - Technology development pays the bill
  - Many scientific questions whose answers would be useful to SID technology
  - The SWB, FISHER, MIXER, GREYBEARD, etc., corpora are documented, and in many cases transcribed, so
  - They can support scientific research on acoustic, phonetic, linguistic, aspects of speaker identity as well as other spoken language research
    - Examples: idiolect; age; PRLM; accent; read v spontaneous; interview v telephone

Summary: Speaker ID vs. ASR

- Speaker and Language ID have always stood in the shadow of ASR -- there have only been two modest USG research programs in the last 20 years. Nevertheless, progress has been truly remarkable, especially in the last decade, with multiple satisfied DoD customers.
- While not exactly "The Tortoise and the Hare," a story can be told that modest but steady funding was unusually successful in Speaker ID because of a "virtuous cycle" involving NIST and LDC (among others) playing different roles than in DARPA programs.
- There was also more freedom to explore new territory than in big programs, and occasional lapses into science while developing the desired technology. Examples are HASR, forensic SID, crosslingual SID, and PPRLM.

Discussion

- BACKUP SLIDES