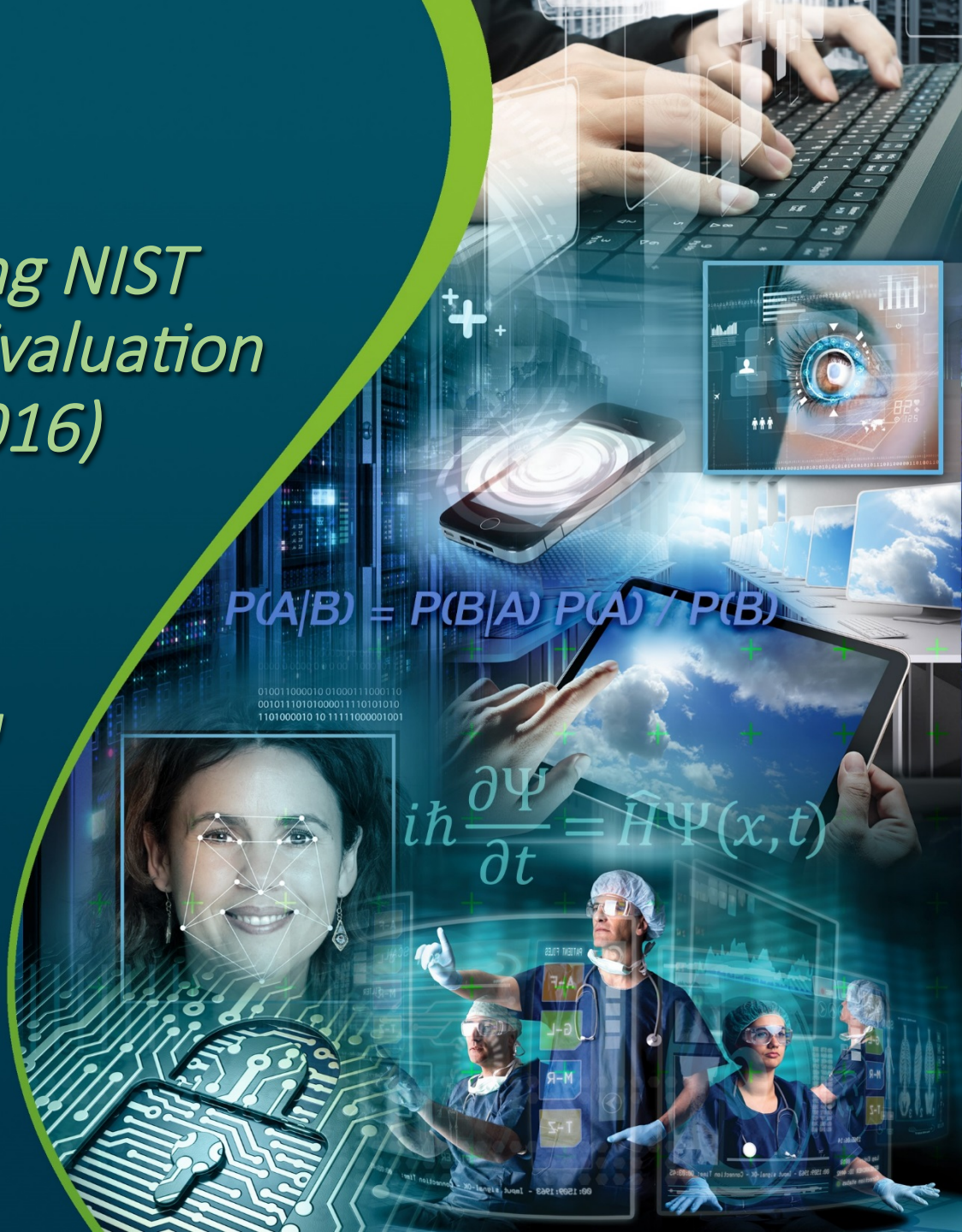
INFORMATION TECHNOLOGY LABORATORY

Design of the upcoming NIST Speaker Recognition Evaluation Vendor Test (SREVT 2016)



Vince Stanford, Oleg Aulov, Greg Fumara, Wayne Salamon,

& Peter Fontana

Multimodal Information & Image Groups

May 3rd, 2016

100 Bureau Drive

Gaithersburg, Maryland 20899

NIST Speaker Recognition Evaluation Vendor Test (SREVT) Pilot of 2016

- New ***Sequestered Data Evaluation*** for speaker recognition
- In ***Open Speaker Recognition Evaluations (Open SRE's)*** NIST provided speech data to labs, which returned scores for given train/test trial schedules
- In ***SREVT*** vendors provide systems running under the *NIST Biometric Toolkit API* via a native C++ shared-library, or under a wrapper layer as needed
- Modern recognition systems use Python, Java, Perl, or C/C++, and compute libraries, e.g. BLAS, LAPACK, or NumPy
  - These run under wrappers for Train/Test scripts
  - This interface is defined by files handled by the biometric toolkit wrapper

Historical Background

SREVT evolved out of twenty years of NIST Open SRE's

- Early ***Open SRE's*** showed benefits of providing standard data sets, and evaluation metrology, to the speaker recognition research community
- Early corpora were collected in 1990's for speaker recognition e.g.: TIMIT, KING, YOHO, and especially the Switchboard Corpora*
- NIST introduced standard metrology functions:
  - 1996 - Detection Cost Function (DCF)
  - 1997 - Detection Error Tradeoff (DET) Curves
  - Equal Error Rate (EER), Receiver Operating Characteristic (ROC) Curves also popular in the classifier literature.
- Over years, this helped the speaker recognition research technology mature

*See Linguistics Data Consortium at http://ccl.pku.edu.cn/doubtfire/CorpusLinguistics/LDC_Corpus/available_corpus_from_ldc.html

Milestones from Early NIST Open SRE's

- Many NIST Open SRE evaluations: 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2008, 2010, 2012, with 2016 ongoing. But forerunner events were as early as 1992.
- Speaker Recognition problems investigated by SRE Community include: *access control*, *speaker detection*, and *forensic matching.*
- Key milestones within this time frame:
  - 1992 – An early evaluation had several sites as part of DARPA program
  - 1994 – "Public Databases for Speaker Recognition and Verification" published
  - 1995 – Evaluation with 6 sites using Switchboard-1 data
  - 1998 – TIMIT data used, but forensic usefulness debated
  - 2001 – First Odyssey: More emphasis on evaluation
  - Further Odyssey workshops 2006-2012, and continuing through today
- But generally NIST SRE "…concentrated upon the speaker spotting task (speaker detection), emphasizing the low false alarm region of performance curve."* So speaker detection with a low target trial prior probability was the key task in many SRE's.

**The NIST Speaker Recognition Evaluations.* Alvin F. Martin, Odyssey 2012 Singapore, June 27, 2012

The SREVT 2016 Pilot

System test for upcoming SREVT Series

- Sequestered data evaluation is implemented, and will be tested at scale:
  - NIST Biometric Research Lab (BRL) has data installed – Processor blades, Petabyte data storage, and switching fabric, with data security, will run the participating systems for evaluation trial schedules
  - NIST Biometric Toolkit distributed processing layer has been implemented
  - The traditional evaluation metrics are implemented (e.g.: DCFs', DET Curves, ROC Curves, EER's, etc.)
  - Ports of the participant systems are now under development (LLSpeech is complete)
- Stakeholder communities have been consulted: But this process is continuing, particularly regarding more operational scenarios
- Four vendor Speaker recognition engines are in process, but the systems for the pilot are not cutting edge technology from all the vendors
- Data sets are incrementally larger than previous Open SRE data sets, but they contain some data previously seen by the community
- Thus the SREVT 2016 Pilot will not compare performance levels system-to-system

Tentative SREVT 2016 Pilot Evaluation Schedule

Sign up	Until May 15, 2016
Submission Verification Phase	May through July 15, 2016
Finalized executable systems installed on the BRL	July 15, 2016
SREVT Pilot Evaluation execution and Reporting	July through October 2016
Lessons learned reporting/planning, for full-scale Data and Scenario design	November 2016 through 2017 SREVT cycle.

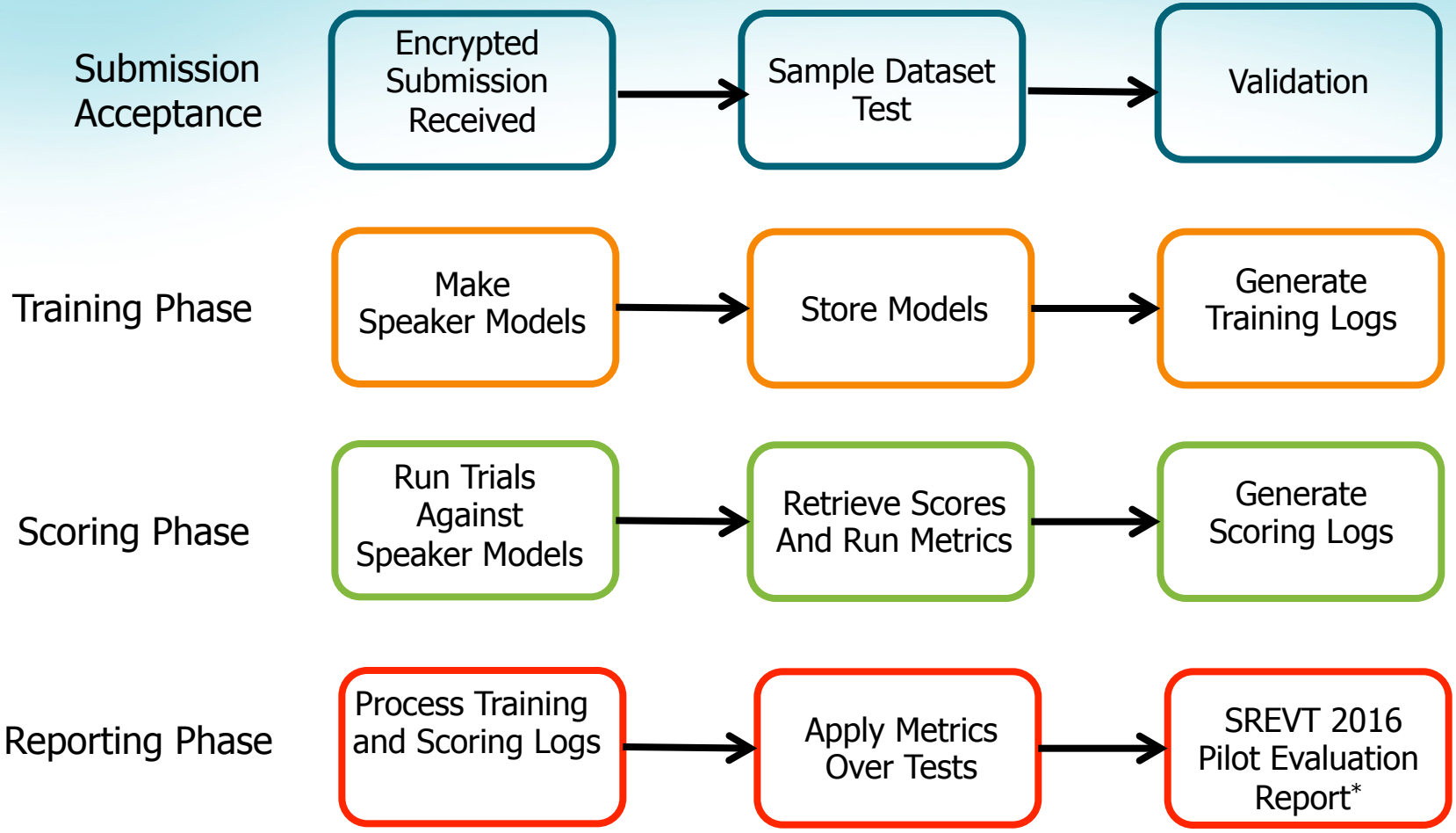
SREVT 2016 Pilot Evaluation is the *system test* for full-scale SREVT sequestered data evaluations:

- **Currently four participating systems**
- **Data includes:**
  - **Previously seen by Open SRE community in 2010 for validation purposes**
  - **Additional data not widely exposed to the community including ~3,000 additional subjects.**

SREVT Pilot Corpora

- Currently using three major corpora
- Approximately 4000 distinct subjects – Larger than any single Open SRE series evaluation to date.
- Several accents, native and non-native English speakers, and some additional languages
- Various:
  - Microphones
    - Telephony: Landline, mobile phone,
    - Room: Lapel, tabletop
  - Various Vocal effort levels (speakers in noise)
  - Speech styles – Read transcripts, near/far interview, conversation
  - Environments: Soft/hard rooms

Evaluation Process



*Will report on validation, runtimes, operating points, systems engineering lessons learned for full scale evaluation, etc.

Possible Training/Testing Combinations for SREVT 2016 Pilot

May include any type of test/train mismatch

		Test Segment		
		NSecTel	NSecMic	SingleChannel
Training	NCore	X	X	X
	NSingle	X	X	X

- ***Ncore***: 1 to N K-channel (K >1) telephone, interview, meeting conversational excerpts, duration (300>N>15 sec)
- ***Nsingle***: 2 to N single-channel telephone, interview, meeting conversations, duration (300>N>15seconds)
- ***NSecTel***: two-channel excerpt from a telephone conversation
- ***NSecMic***: two or more channels, mic-recorded excerpt, conversation, interview, or meeting, duration (300>N>15sec from target speaker)
- ***SingleChannel***: conversation excerpt (tel or mic), interview, or meeting, with one or more interlocutors, duration (300>N>15sec by target speaker)
- ***The exact train/test schedules and data sets to be presented will not be pre-announced to the vendors.***

Performance Measures

Detection Cost Functions (DCFs)

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NoTarget} \times (1 - P_{Target})$$

$$C_{Default} = \min \begin{cases} C_{Miss} \times P_{Target}, \\ C_{FalseAlarm} \times (1 - P_{Target}) \end{cases}$$

and

$$C_{Norm} = \frac{C_{Det}}{C_{Default}}$$

Note: C_{Det} is minimized over the range of detection thresholds

The DCF parameters are the relative costs of detection errors, C_{Miss} and $C_{FalseAlarm}$, and the a priori probability of the target speaker, P_{target}.

Example of possible Forensic vs. Investigatory DCF Parameters

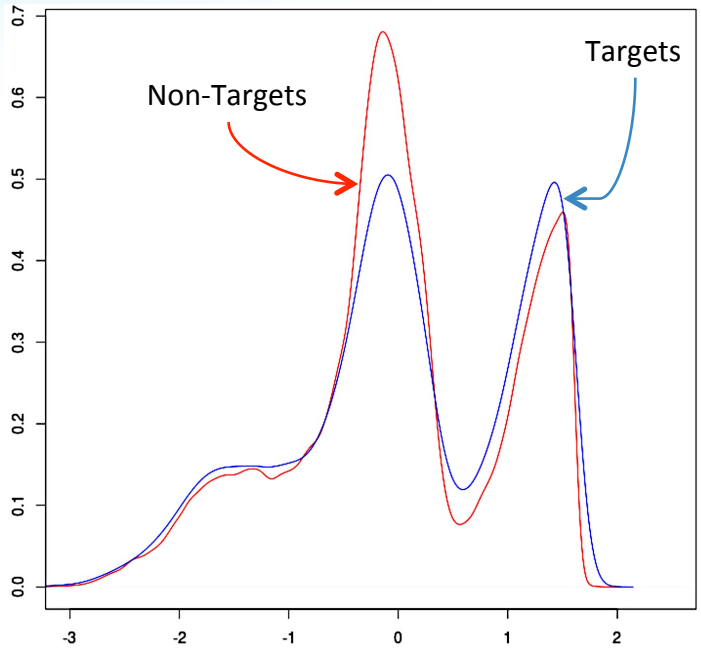
C_{Miss}	$C_{FalseAlarm}$	P_{Target}
1	10	0.01

C_{Miss}	$C_{FalseAlarm}$	P_{Target}
10	1	0.001

Note: Actual prior probabilities of target trials will not be published.

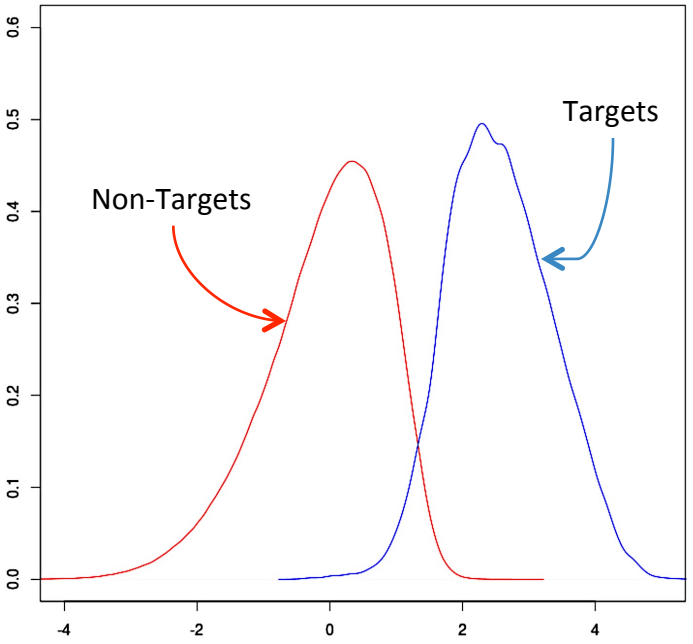
Examples of Target vs. Non-Target Score Distributions from SRE-12 systems

Speaker Match Scores



Uninformative SRE-12 System

Speaker Match Scores

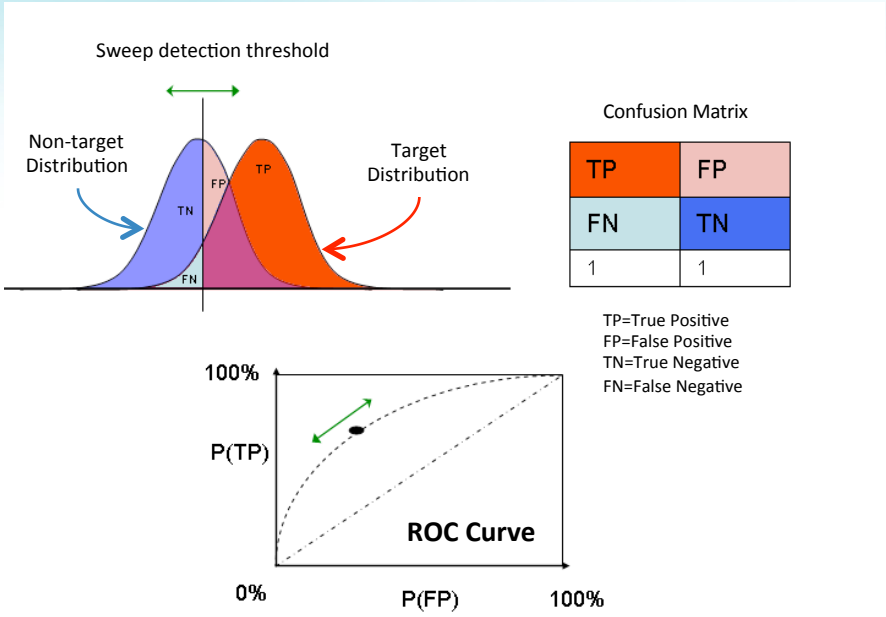


Informative SRE-12 System

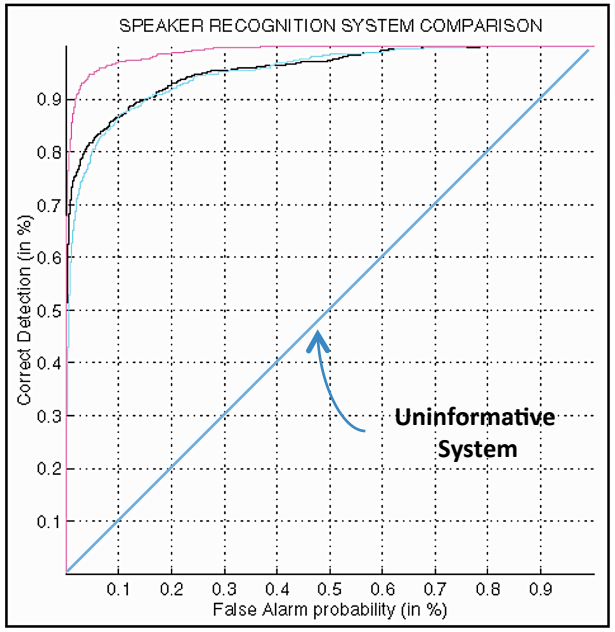
Performance Measures

Receiver Operating Characteristic (ROC) Curves

ROC Curve Quantities



ROC Curve for Three Systems



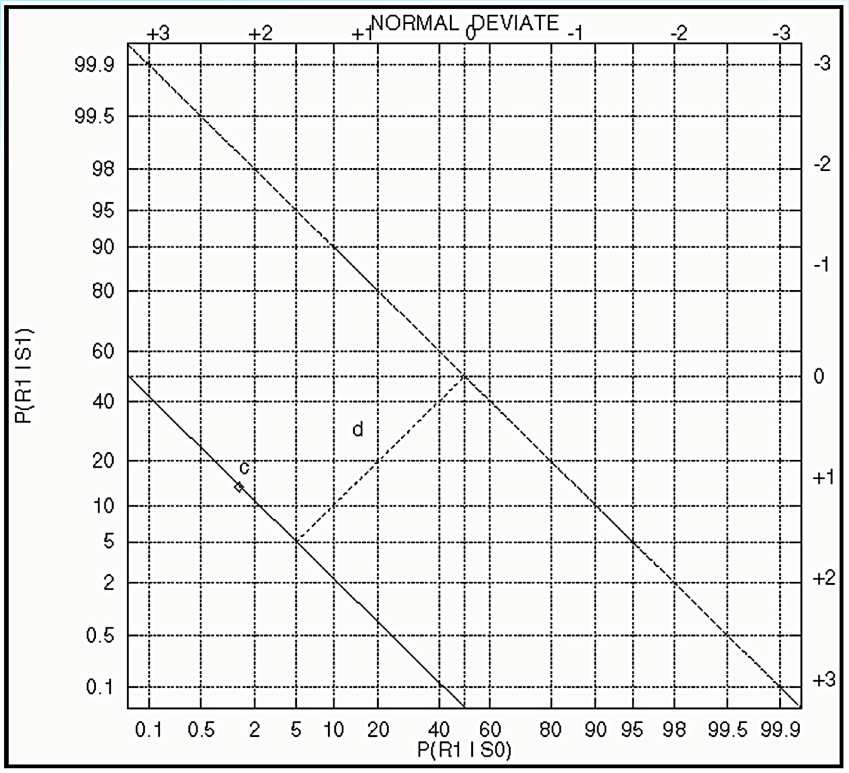
Doesn't treat the types of error, *False Negative*, and *False Positive* equally

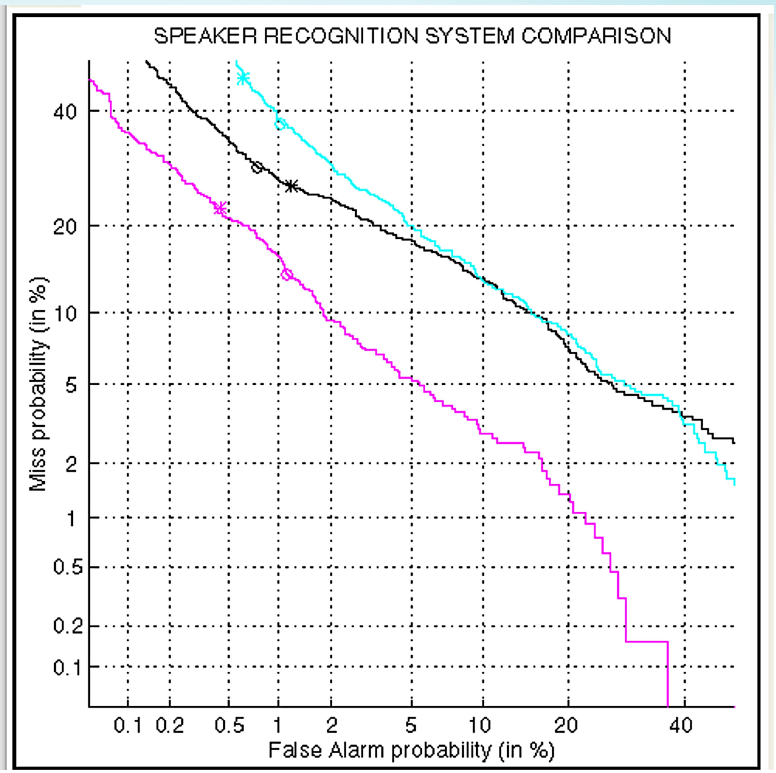
Two of the three curves are hard to distinguish, and most of the chart is empty

Classic ROC detection chart is hard to read and interpret, especially with many systems measured

Performance Measures

*From ROC Curves to Detection Error Tradeoff (DET) Curves**





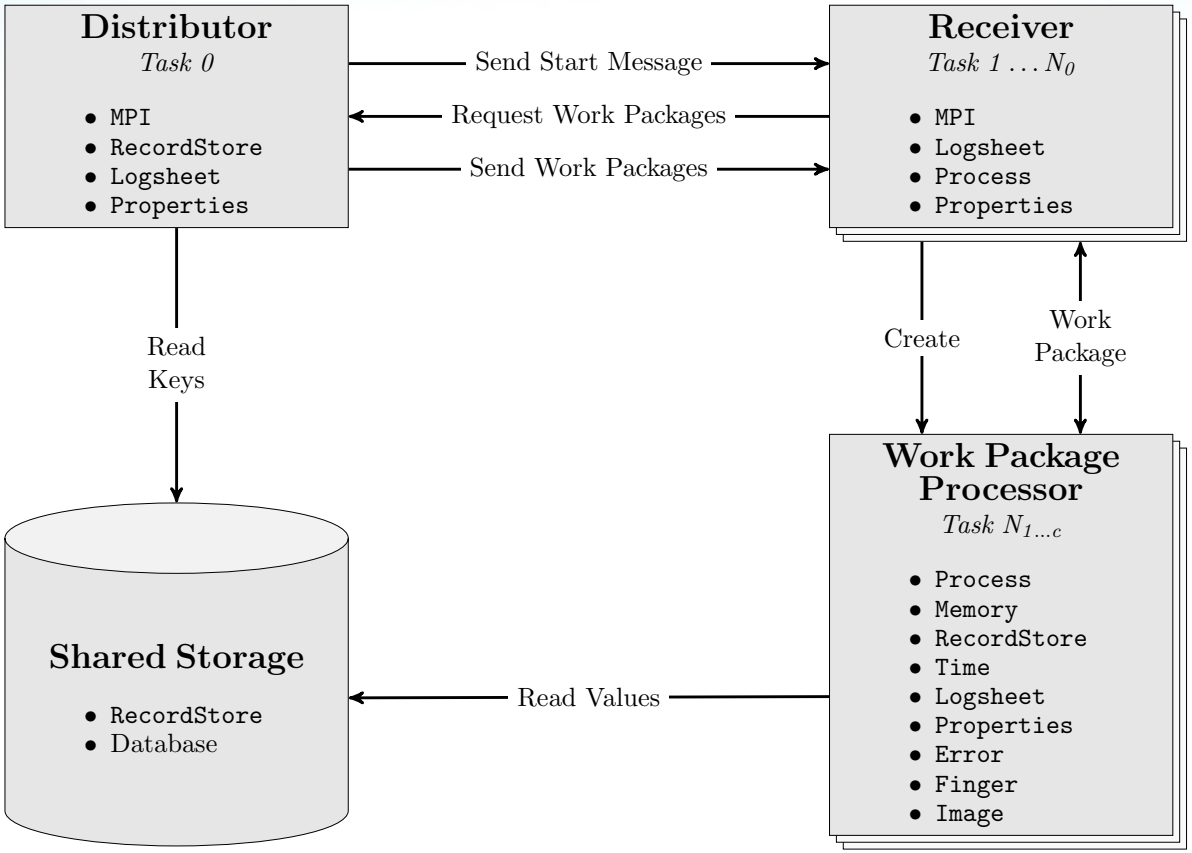
Plots normal quantiles of error rates
Treats False Alarm and Miss probabilities symmetrically
Separates multiple systems more effectively

*Alvin F Martin et al. “The DET Curve in Assessment of Detection Task Performance", Proc. Eurospeech '97, Rhodes, Greece, September 1997

INFORMATION TECHNOLOGY LABORATORY

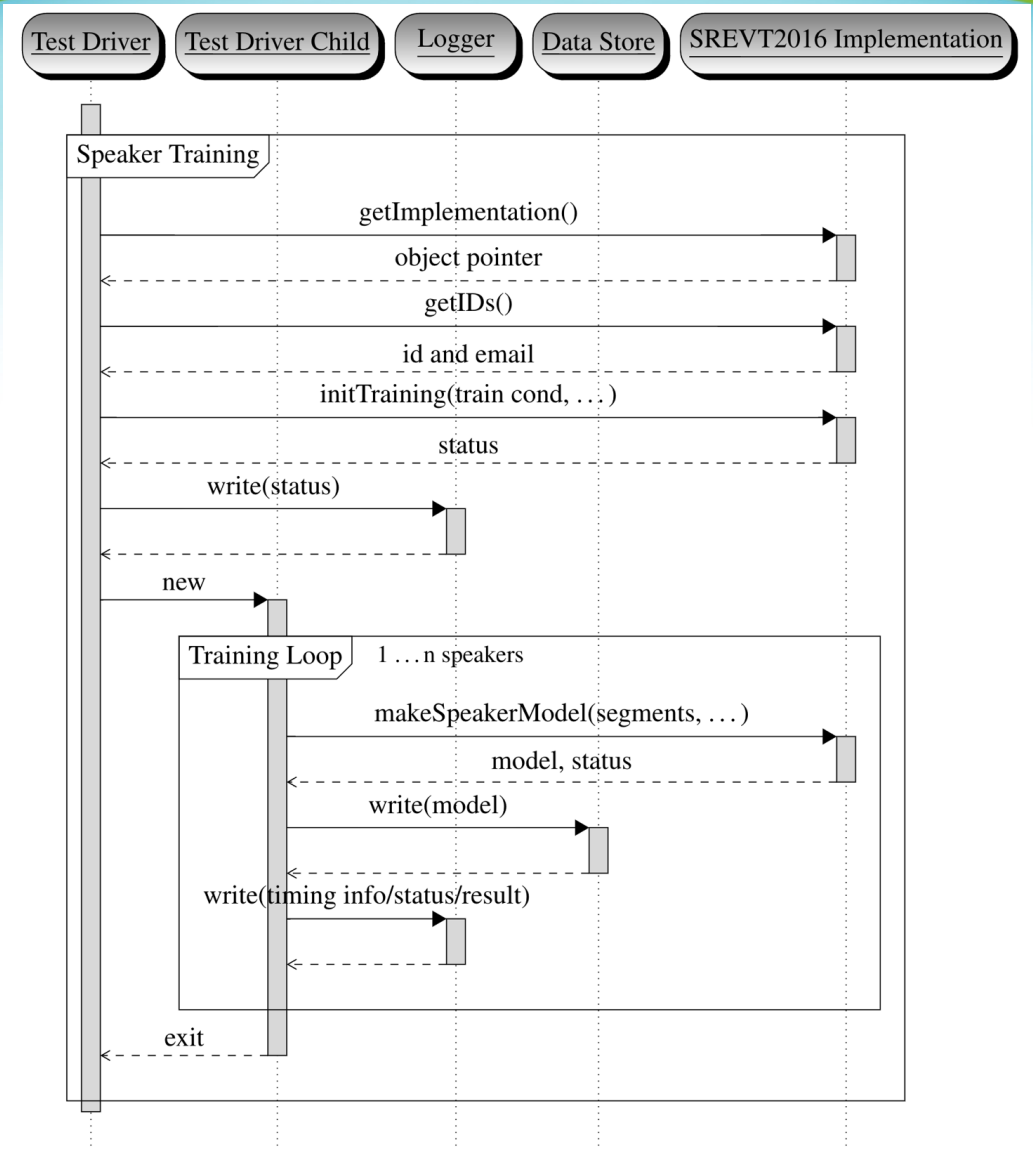
NIST National Institute of Standards and Technology U.S. Department of Commerce

Biometric Evaluation Framework Overview

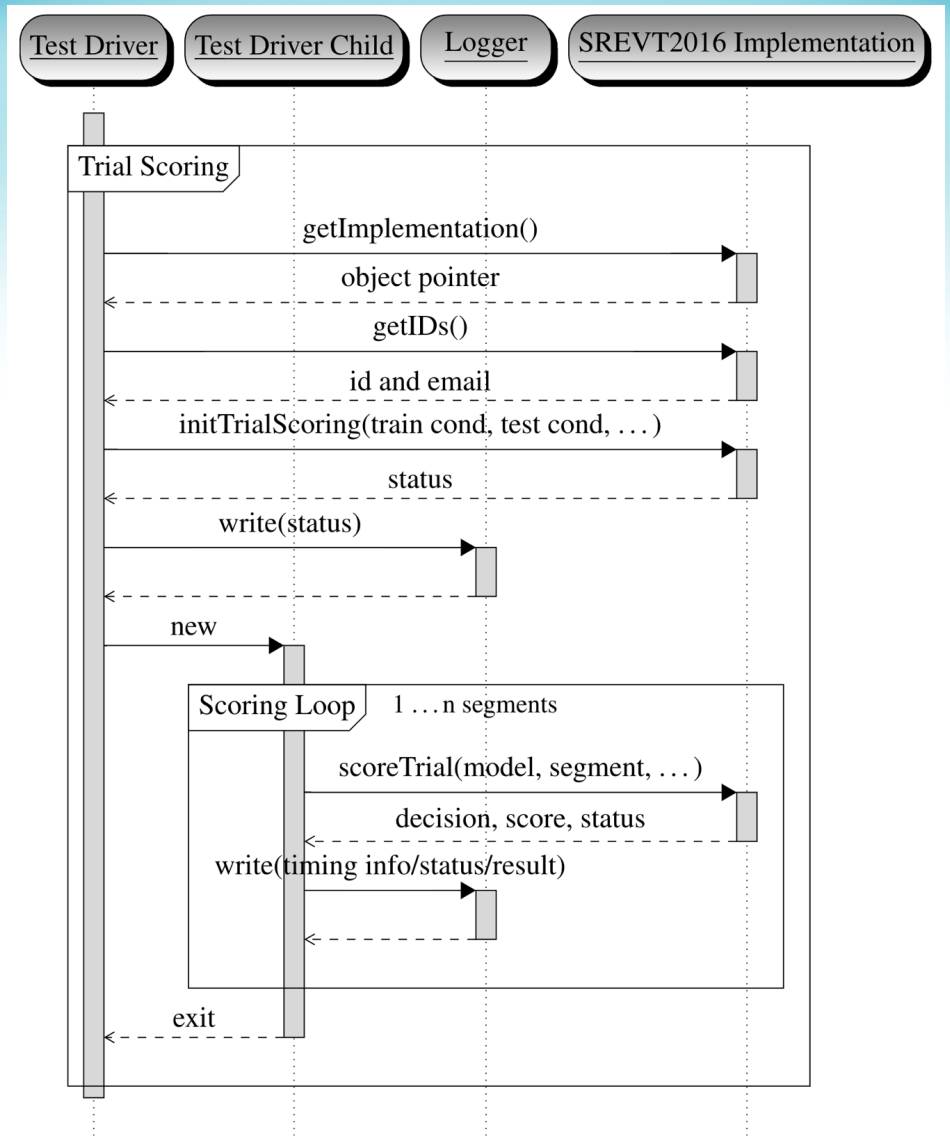


System Submission Options

- Shared Library Implementation with native C++ API calls for systems to obtain trial data, and return results. Suitable for use in high performance C++ binary systems
- Multi-language scripts (e.g. Java, Python, Perl, C, and high performance libraries) running under the wrapper that write prescribed files
- Runtime performance does matter.



Model Training



Trial Scoring

Wrapper Executable Syntax

train.sh --trainfiles <training list> --configdir <configuration directory> --tempdir <temp_directory> --modelfile <model file>

score.sh --testlist <testlist> --scorefile <score file> --configdir <configuration directory> --tempdir <temp directory>

<training list> - four fields per line: <model_ID> <gender> <full-path-to-audio>.<channel>
<configuration directory> - full path to the configuration directory
<temp directory> - full path to the temp directory
<model file> - path to the file where the model will be stored
<test list> - file with four fields per line: <full-path-to-audio>.<channel> <gender> <modelID>
<score file> - path to score file, contains: <test file> <model ID> <score> <decision>

NIST Biometric Research Lab Cluster

- **Storage:**
  - **SAN: 50TB (80 Drives); 240TB (415 Drives); 550TB (756 Drives)**
  - **Stornext FS (Windows/Linux shared Access)**
  - **Storage Duplicated for COOP**
  - **Tape Backup System (80TB)**
- **Blade Farm**
  - **10 blades (2 CPUs/4 cores, 64GB)**
  - **48 blades (2 CPUs/4 cores, 48GB) - NGI**
  - **16 blades (2 CPUs/6 cores, 96GB)**
  - **32 blades (2 CPUs/6 cores, 192GB)**
  - **32 blades (4 CPUs/4 cores,192GB)**
  - **8 blades (2 CPUs/8 cores, 256GB)**

  Totals
  - **146 blades -> 356 CPUs -> 1680 cores**
- **Private Network with OMB C&A of security**
- **20 Ton A/C and 100KVA UPS**
- **Temperature Monitoring/Alert System (email/phone)**
- **Hardware Failure Alerts (email)**

Important Points for the Future

- Data must be relevant to operational needs: Noise levels, Channels, Recording conditions, Subject demographics, accents, dialects, languages etc.
- Computation of meaningful likelihood ratios requires large corpora collections that represent populations and collection modes/conditions
- Investigatory and forensic use cases are different, data must support both, operating points must be understood
- In the U.S., Daubert Criteria require:
  - Empirical testing
  - Peer reviewed publication
  - Known or potential error rate
  - Standards and controls for operation
  - Generally accepted by relevant scientific community
- All of the above point to the need for systematic large-scale data collection, to establish operating points, and error rates.

Discussion and Questions

Contact: Vince Stanford (vincent.stanford@nist.gov)