OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

Babel – Addressing the Language Deluge

LEADING INTELLIGENCE INTEGRATION

The overall classification of this briefing is UNCLASSIFIED

Mary P. Harper
Incisive Analysis Office, IARPA
Babel Program Overview

Babel – Addressing the Language Deluge

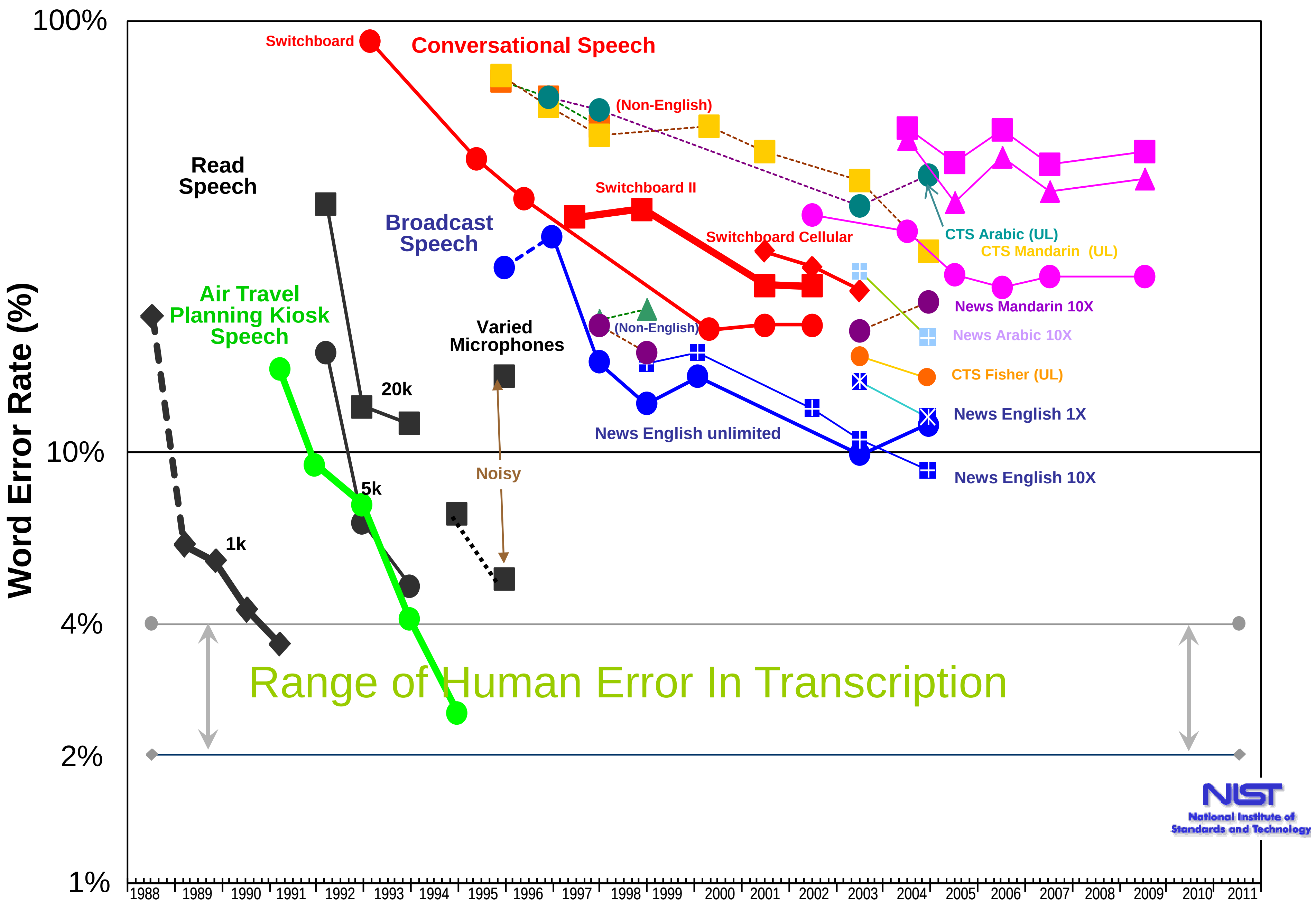
Goal:

- Develop agile and robust speech technology that:
  - can be rapidly applied to any human language
  - will provide effective keyword search capability for analysts to efficiently examine massive amounts of real-world recorded speech

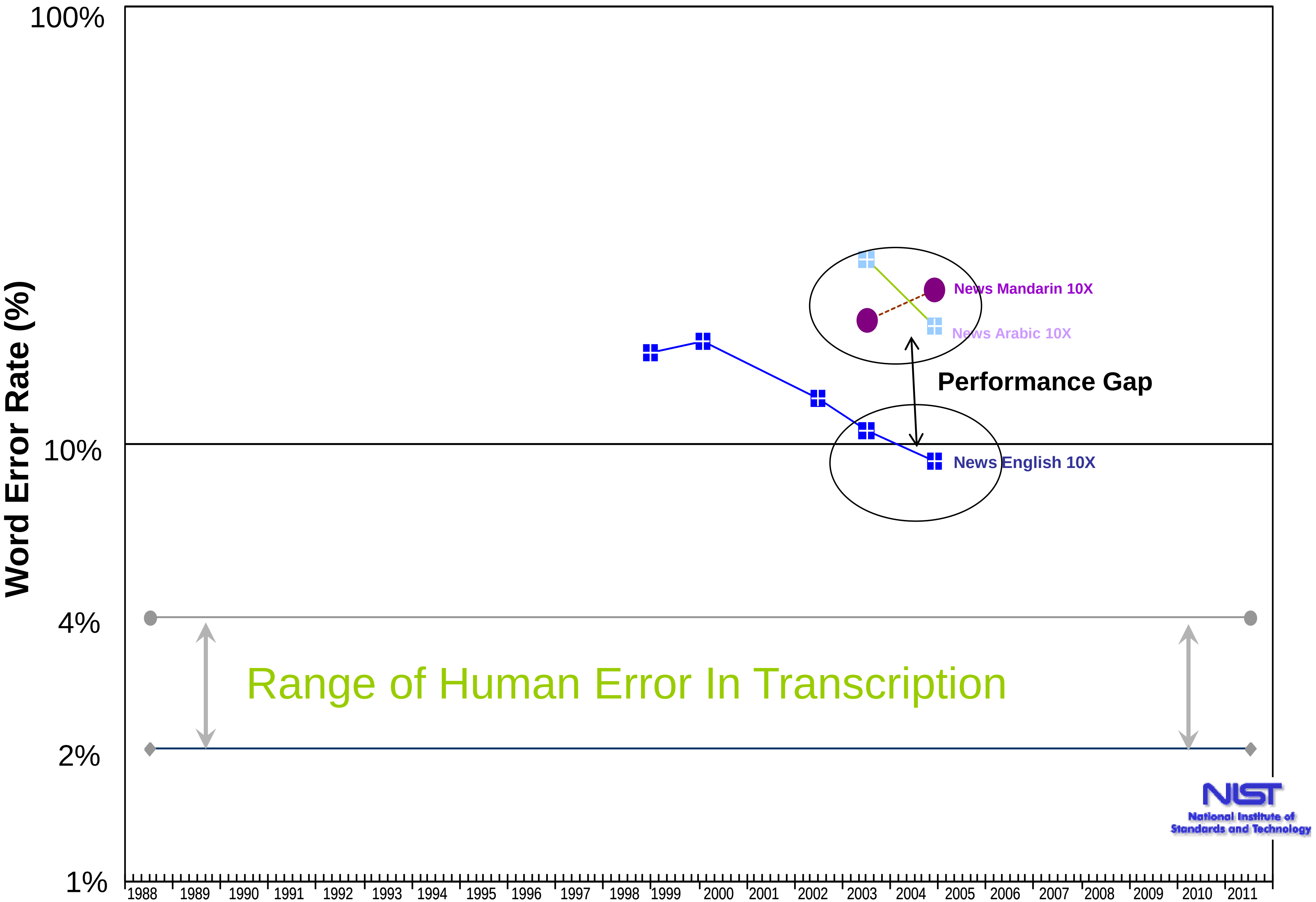
State-of-the-Art/Practice:

- 7,000+ languages, 330 have 1M+ speakers, but **only a few studied**
- Today's systems were originally developed for English on fairly clean speech with **significantly lower performance** when the technology is:
  - ➢ extended to other languages
  - ➢ applied to speech collected in real-world conditions
- System development time for a new language takes **months to years.**

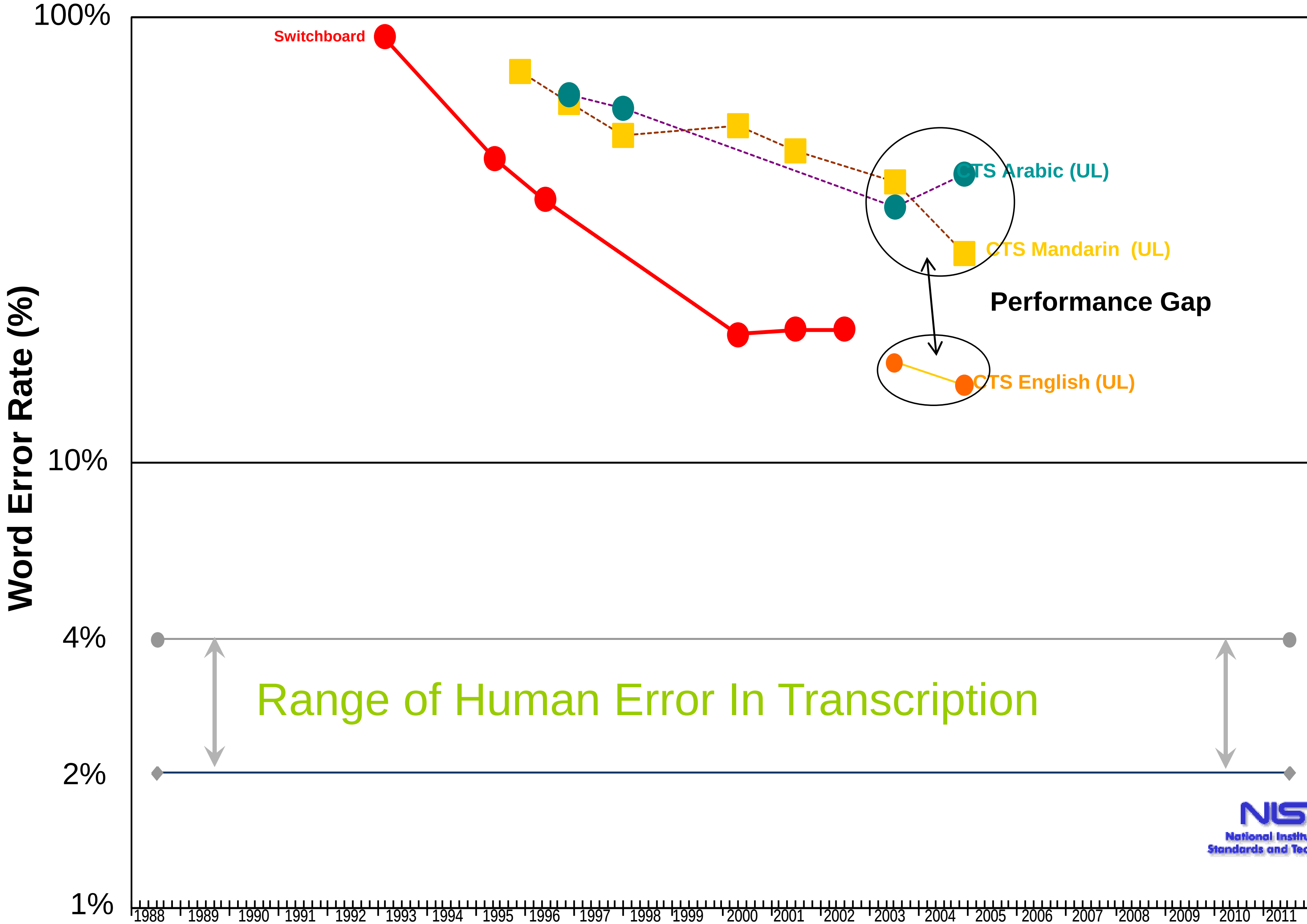
NIST Benchmark Speech Test History – May ’09



NIST Broadcast Speech



NIST Conversational Speech



NIST
National Institute of Standards and Technology

Example Use Case

- Thousands of hours of speech is acquired in a language of emerging importance to the IC.
- Few IC analysts have the ability to understand the language.
- There is no existing speech technology for the language.
- We must rapidly develop effective triage capabilities for the analysts.

Approach

- Broad language portfolio:
  - Languages from a variety of language families (e.g., Afro-Asiatic, Niger-Congo, Sino-Tibetan, Austronesian, Dravidian, Altaic)
  - Mixed language typology (i.e., with different phonotactic, morphological, syntactic characteristics)
- Researchers will:
  - work with development languages to create new methods
  - be evaluated annually on a surprise language with development time and training size constraints
- Annual evaluation:
  - On the set of development languages and the surprise language
  - Progress will be measured using:
    - NIST Spoken Term Detection Evaluation (see http://www.itl.nist.gov/iad/mig//tests/std/2006/index.html)
    - Word Error Rate (WER) when appropriate for the technology

Language Packs

- Each language will be provided to researchers in a “pack” that will contain speech data and language information
- Speech data may include:
  - Calls made from quiet and challenging environments including public places (bar, restaurant, shopping mall ...), street/roadside, quiet location (home, office ...), moving vehicle (in-car, train, bus ...) with handheld phone or hands-free device.
  - Mixed telephony recordings
    - Scripted speech to ensure baseline coverage of the language’s phoneme inventory
    - Conversational speech
  - Metadata for all recordings (e.g., gender, age, handset type, environment)
  - Transcription of conversational audio (the amount depends on program year)
- Language information may include:
  - Description of the language (e.g., dialect regions, phoneme set definitions)
  - Lexicon entries for words appearing in transcribed data

Program Organization

Phase 1 (27 months)

- Provide system build packs for a broad selection of languages of interest to force a general approach.
  - **Base Period (15 mos.)**: 4 language build packs, each with 100% transcription and a pronunciation lexicon for the words in the transcription. Only telephone channel data will be included.
  - **Option Period One (12 mos.):** 5 language build packs, each with no more than 75% transcription with correspondingly limited pronunciation lexicon. Telephone and non-telephone channels will be included.
- Government evaluation will use key word search metrics to test
  - each development language
  - a surprise language with limited system development time and different training set sizes (e.g., 10 or fewer hours, 80 hours)
  - using data from challenging environments, including some that do not match training environments

Program Organization

Phase 2 (24 months)

- Provide more limited system build packs for a greater number of languages.
  - **Option Period Two (12 mos.)**: 6 language build packs, each with no more than 50% transcription with correspondingly limited pronunciation lexicon. Telephone and non-telephone channels will be included.
  - **Option Period Three (12 mos.):** 7 language build packs, each with no more than 50% transcription with correspondingly limited pronunciation lexicon. Telephone and non-telephone channels will be included.
- Government evaluation will use key word search metrics to test
  - each development language
  - a surprise language with limited system development time and different training set sizes (e.g., 10 or fewer hours, 80 hours)
  - using data from challenging environments, including some that do not match training environments

Performance Goals

	Phase 1		Phase 2	
Year	Base	Option 1	Option 2	Option 3
Transcribed %	100%	≤ 75%	≤ 50%	≤ 50%
Pronunciation Lexicon (transcription coverage)	100%	≤ 75%	≤ 50%	≤ 50%
Language Packs Development+Surprise	4+1	5+1	6+1	7+1
Development Time for Surprise	4 weeks	3 weeks	2 weeks	1 week
Minimum NIST Actual Term Weighted Value (ATWV)[1]	0.3	0.3	0.3	0.3

NOTE: All evaluations will include data from challenging environments.
There will also be alternative evaluations with different amounts of transcribed audio.

1. See http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf

Before and After Babel

	Today	After Babel
Language Coverage	Limited	Any spoken human language
Resources	100's-1000's of hours of transcribed training data	Limited amounts of transcribed data
Channel	Homogeneous	Mixed
Time to develop	Months to years	1 week

Program Roles and Responsibilities

- Performer R&D
  - In Scope:
    - Multilingual speech modeling
    - Novel use of machine learning, data resource gathering, linguistics, etc.
    - Computational methods to limit running time and memory
    - Keyword search
  - Out of Scope:
    - Human user interface
    - Machine translation
- Government Support
  - Government Furnished Information (GFI):
    - Training data for diverse set of development languages
    - Development data to measure interim progress
    - Evaluation data
  - Testing and Evaluation:
    - Evaluation framework to measure performer progress on development languages
    - Evaluation framework to measure performer progress on an annual surprise language

Questions?

NIST Spoken Term Detection Evaluation

- Two methods used in 2006 for assessing performance:
  - Value function
    - An application model that assigns value to correct output and negative value for incorrect output
    - A weighted linear combination of Missed Detection and False Alarm Probabilities
  - Detection Error Tradeoff (DET) curve
    - A graphical representation of the tradeoff between Missed Detections and False Alarms

Value Functions

- Value assigned to correct and incorrect output
  - Value = 1 is perfect output
  - Value = 0 is the score for no output
  - Value < 0 is possible
- Term Weighted Value (TWV)
  - Restricted to terms with reference occurrences
  - V of 1.0 is the value (benefit) of a correct response
  - C of 0.1 is the cost of an incorrect response

C/V weights the error types

Adjusts for the priors

$$\text{Value}_{\text{T}}(\theta) = 1 - \underset{term}{average}\left\{ \text{P}_{\text{Miss}}(term,\theta) + \frac{C}{V}(\text{P}^{-1}_{\text{term}} - 1)\cdot \text{P}_{\text{FA}}(term,\theta) \right\}$$

- Choosing Θ
  - Developers choose decision threshold for their “Actual Decisions” to optimize term-weighted value: All “YES” system occurrences
    - Called the “**Actual Term Weighted Value**” (ATWV)
  - The evaluation code searches for the system’s optimum decision score threshold
    - Called the “**Maximum Term Weighted Value**” (MTWV)

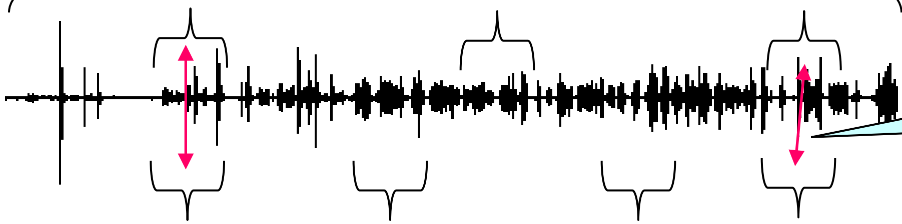
Error Counting Illustration

10 seconds

Reference Occ.

System Occ.

(with Actual Dec.)



YES 0.7 | YES 0.6 | NO 0.4 | NO 0.5

Aligned occurrence

Counting Errors

Value uses Actual Decision

Corr | FA | Miss | Miss

"NO" decisions ignored

Counting Errors

DET Curve uses Θ

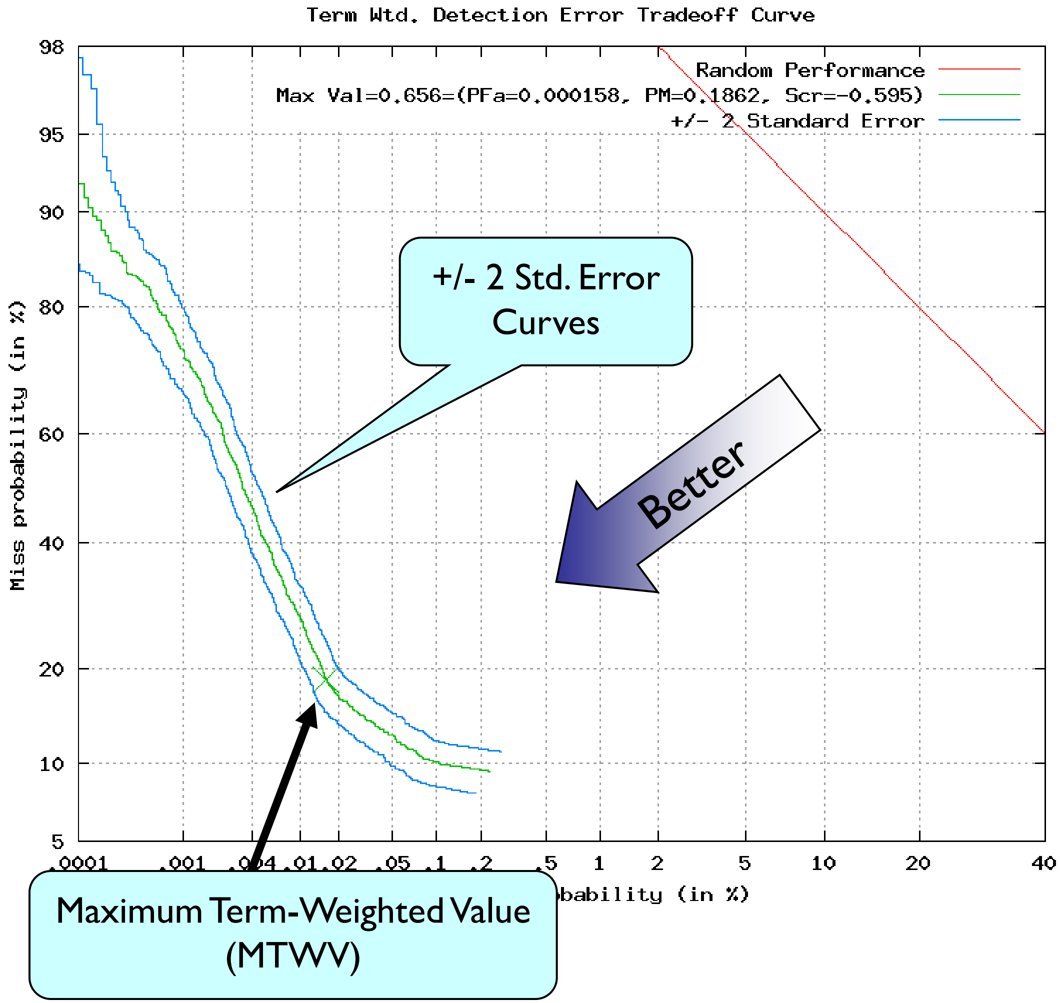
Corr | FA | Miss | Corr

For Θ = 0.5

Estimates of Error Rates

- Given the aligned occurrence, compute the following for each term:
  - Probability of Missed Detections
    - $P_{Miss}(\text{term},\theta) = 1 - N_{correct}(\text{term},\theta)/N_{True}(\text{term},\theta)$
      - not defined when $N_{True}(\text{term},\theta) = 0$
  - Probability of False Alarms
    - $P_{FA}(\text{term},\theta) = N_{Spurious}(\text{term},\theta)/N_{NT}(\text{term},\theta)$
- NT is the number of Non-Target trials
  - But this is NOT countable so used an Estimate
    - $N_{NT}(\text{term},\theta) = \text{TrialsPerSecond} * \text{SpeechDuration} - N_{True}(\text{term},\theta)$
    - TrialsPerSecond arbitrarily set to 1 for all languages
  - In hind sight, they would have used False Alarm rate (e.g., False Alarms per hour).
- Θ is a decision criteria that is set to differentiate likely vs. unlikely putative term occurrences

Detection Error Tradeoff (DET) Curves

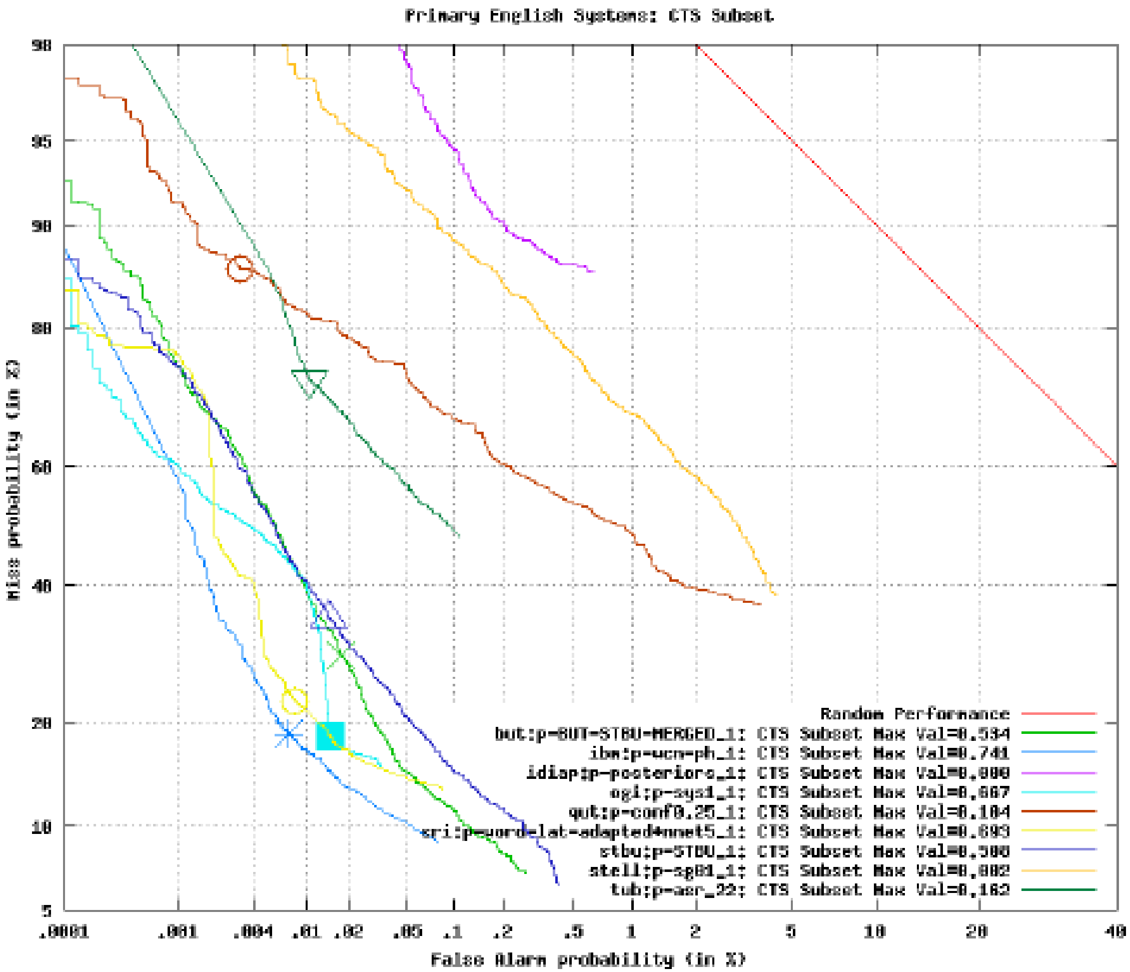


- Plot of P_{Miss} vs P_{FA}
- Axis is warped to the normal deviate scale
- Term-Weighted DET Curve
  - Created by computing term-averaged P_{Miss} and P_{FA} over the full range of a system's decision score space
  - Confidence estimates used for confidence curves

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

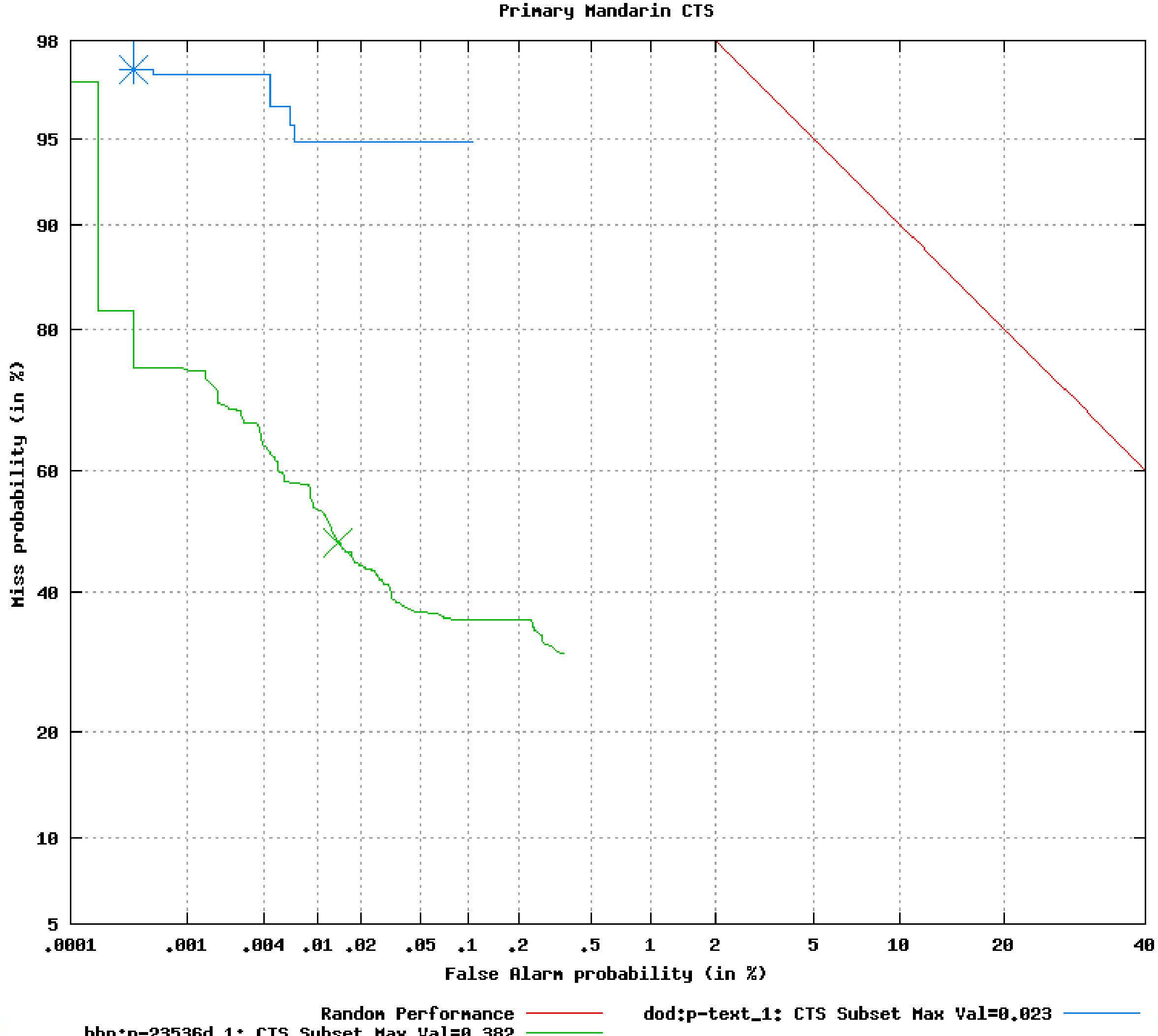
LEADING INTELLIGENCE INTEGRATION

English Conversational

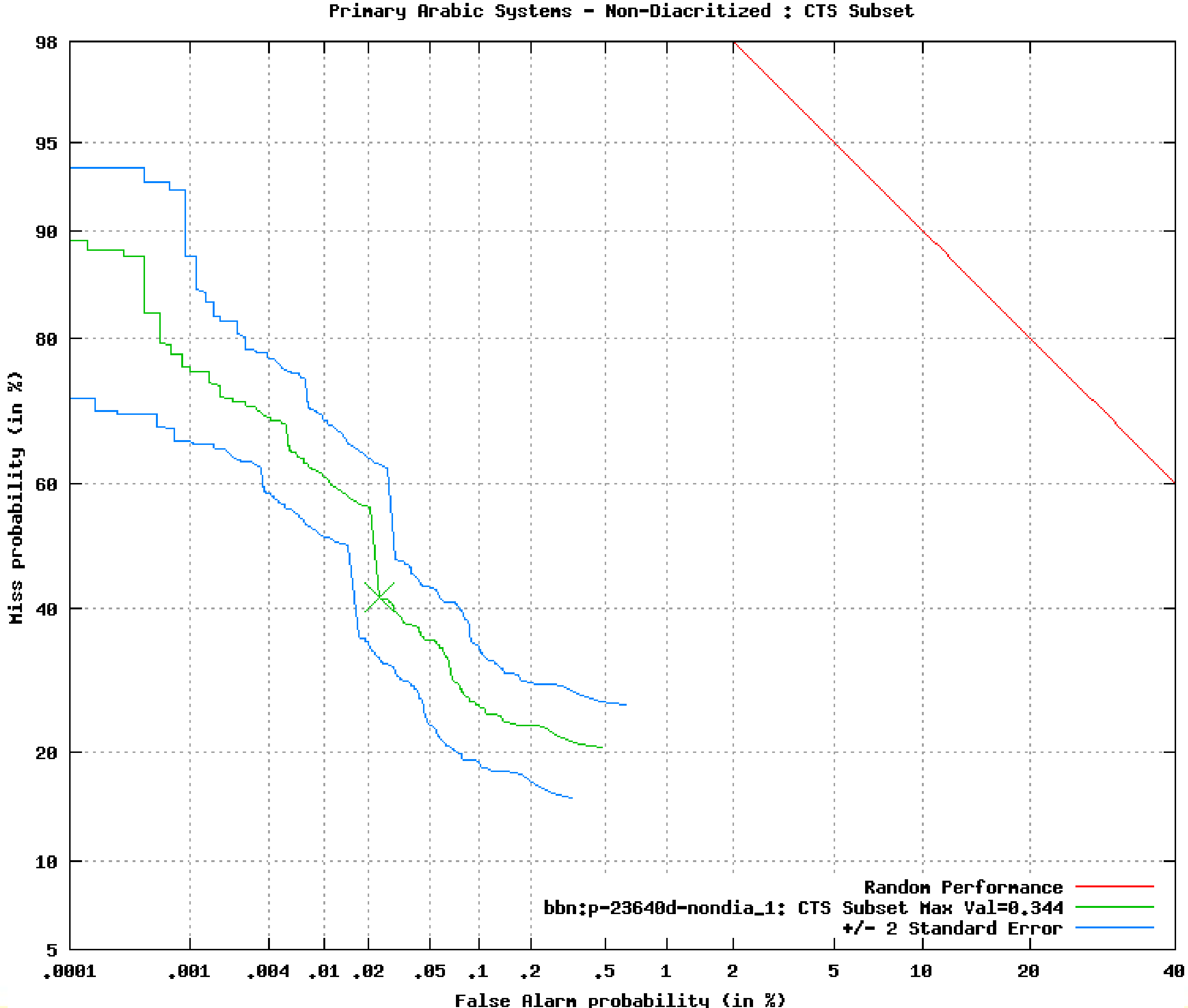


NIST National Institute of Standards and Technology

Mandarin Conversational



Non-Diacritized Arabic Conversational



NIST Evaluation Take-Aways

- Focus more on performance across genre rather than language
- **Few researchers participated in the non-English**
- Trade-offs between Phoneme and Word approaches

	Phoneme	Word
Pros	• Smaller amounts of training data needed (~20 hrs) • No vocabulary constraints • Compact models • Faster Recognition (50xFTRT)	• Higher precision (fewer false alarms) • Higher recognition accuracy • Faster search
Cons	• Lower precision (many false alarms) • Lower recognition accuracy • Slower search	• Larger amounts of training data needed • Finite vocabulary • Large models • Slower Recognition (10xFTRT)

Word Error Rate

- Word Error Rate (WER) is calculated using SCLITE on system (sys) and reference (ref) transcripts.
- Errors: insertions (I), deletions (D), substitutions (S)
- WER = (I+D+S)/ # ref words

ref:	it is	hard	to	recognize			speech
sys:	it 's	hard	to	wreck	a	nice	beach.
ERRORS:				**S**	**I**	**I**	**S**

WER = 4/6 = 66.67%