OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

Babel Program

Intelligence Advanced Research Projects Activity (IARPA)

LEADING INTELLIGENCE INTEGRATION

Dr. Mary P. Harper
Incisive Analysis Office
IARPA

Babel

- The Program
- Program Goals
- Program Structure
- Test and Evaluation

BABEL

Babel Pronunciation Guide

- How is Babel pronounced?
  - http://itre.cis.upenn.edu/~myl/languagelog/archives/004232.html

1. ["beIb@l] rhymes with *table*
2. ["b{b@l] influenced by
   - similarity to English word *babble*,
   - toponym *Babylon* with [{] rather than [eɪ] pronunciation
3. [bA"vEl] Hebrew word בבל, meaning Babel, Babylon, and Babylonia.

The Challenge

- Thousands of hours of speech are acquired in a language of emerging importance to the IC with varied audio quality.
- Few IC analysts have the ability to understand the language.
- There is no existing speech technology for the language.
- We must be able to rapidly develop effective triage capabilities to assist those few analysts.

Babel – Addressing the Language Deluge

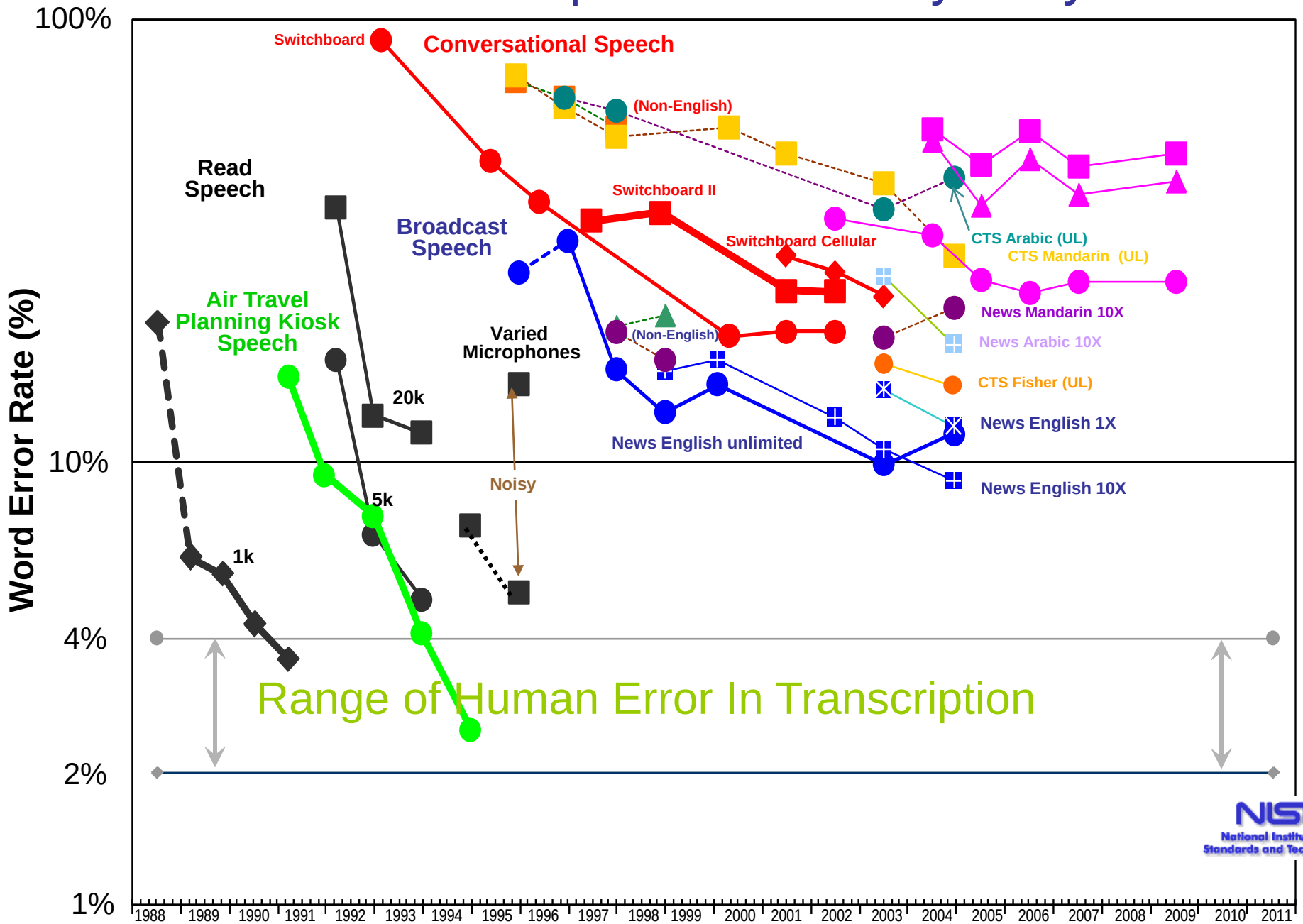
Goal:

- Develop agile and robust speech methods
  - Rapid application to any human language
  - Effective keyword search capability over massive amounts of real-world recorded speech

State-of-the-Art/Practice:

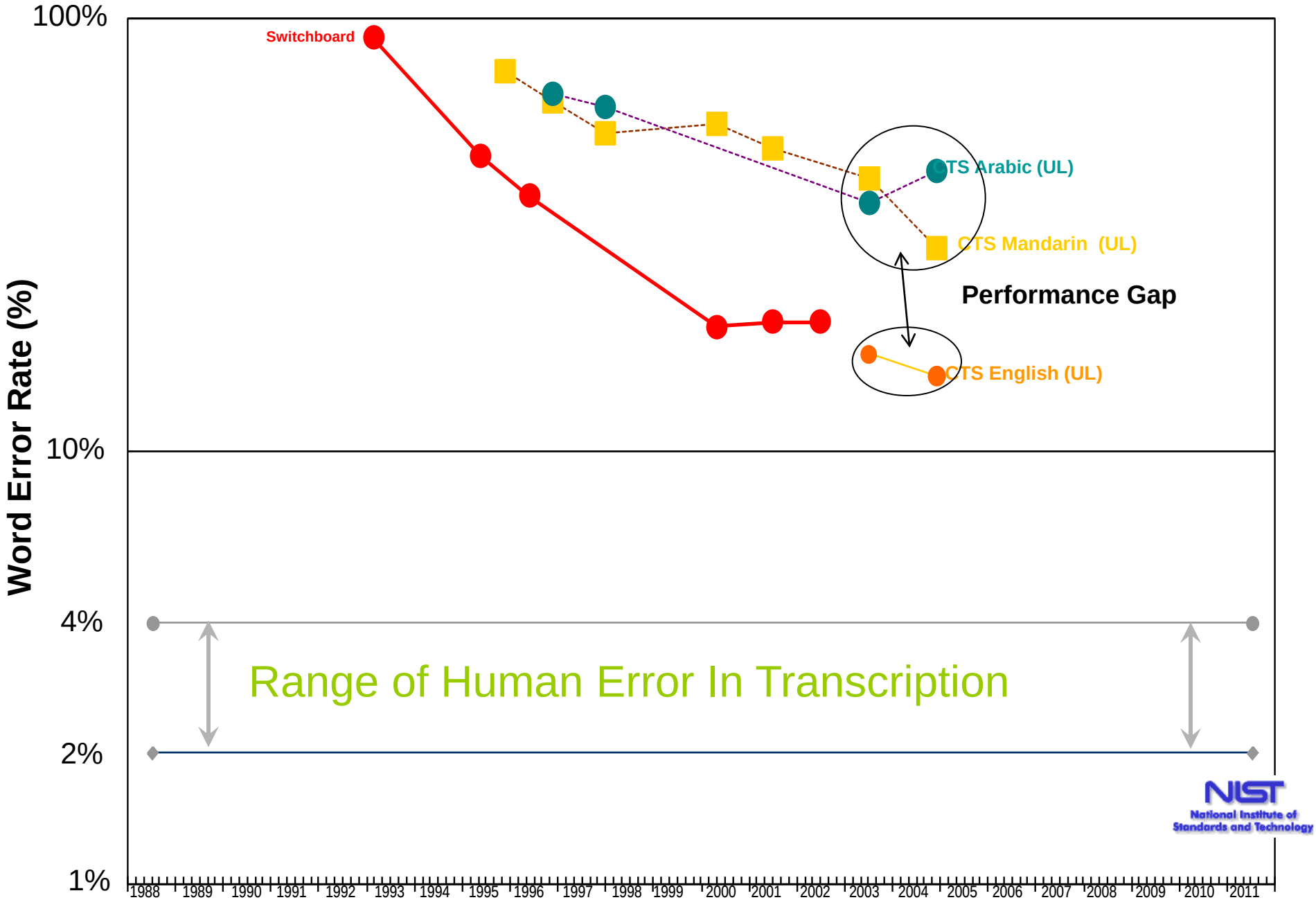
- 7,000+ languages, 330 have 1M+ speakers, but **only a few studied**
- Today's systems were originally developed for English on fairly clean speech with **significantly lower performance**:
  - On other languages
  - On speech collected in real-world conditions
- System development for a new language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

NIST Benchmark Speech Test History – May '09



NIST National Institute of Standards and Technology

NIST Conversational Speech



NIST National Institute of Standards and Technology

Babel's Approach

- **Work with diverse languages from the outset**
  - Acquire speech data in-country for languages from a broad set of language families (e.g., Afro-Asiatic, Niger-Congo, Sino-Tibetan, Austronesian, Dravidian, Altaic), chosen for coverage of language types
  - *Study multiple languages each program period*
- **Handle real recording conditions from the outset**
  - Acquire data in a variety of conditions (e.g., in a moving car, in a café, on the street) and use different recording devices (e.g., cell phone, hands free, table top microphone)
  - *Evaluate using diverse conditions each period*
- **Constrain resources and system development time each period**
  - *Reduce the amount of transcribed speech for use in system development*
  - *Reduce system development time*
- **Rigorous evaluation**
  - *Use a "surprise" language for system evaluation each program period*

Speech Search

Query: نهر الكلب

Hit: نهر الكلب

Indexer

Search Repository of Indexed Audio

نهر الكلب كلب بوليسي

Actual Term Weighted Value (ATWV)

Actual Term Weighted Value (ATWV): Developers choose Θ, the decision threshold for their "Actual Decisions" to optimize term-weighted value

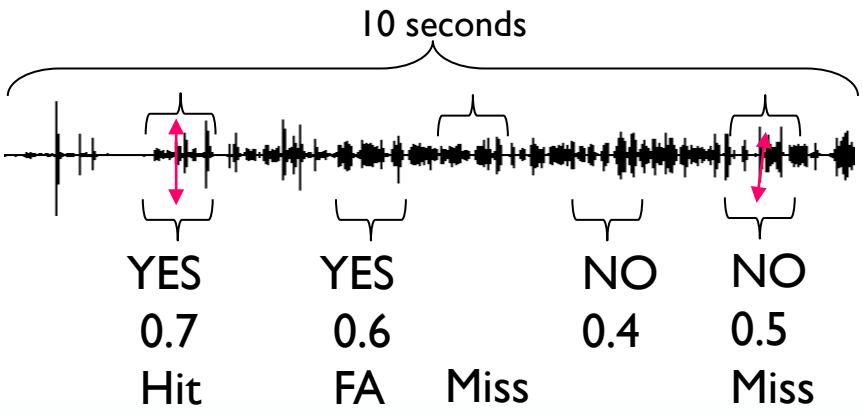
- Restricted to terms with reference occurrences
- V of 1.0 is the value (benefit) of a correct response
- C of 0.1 is the cost of an incorrect response

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

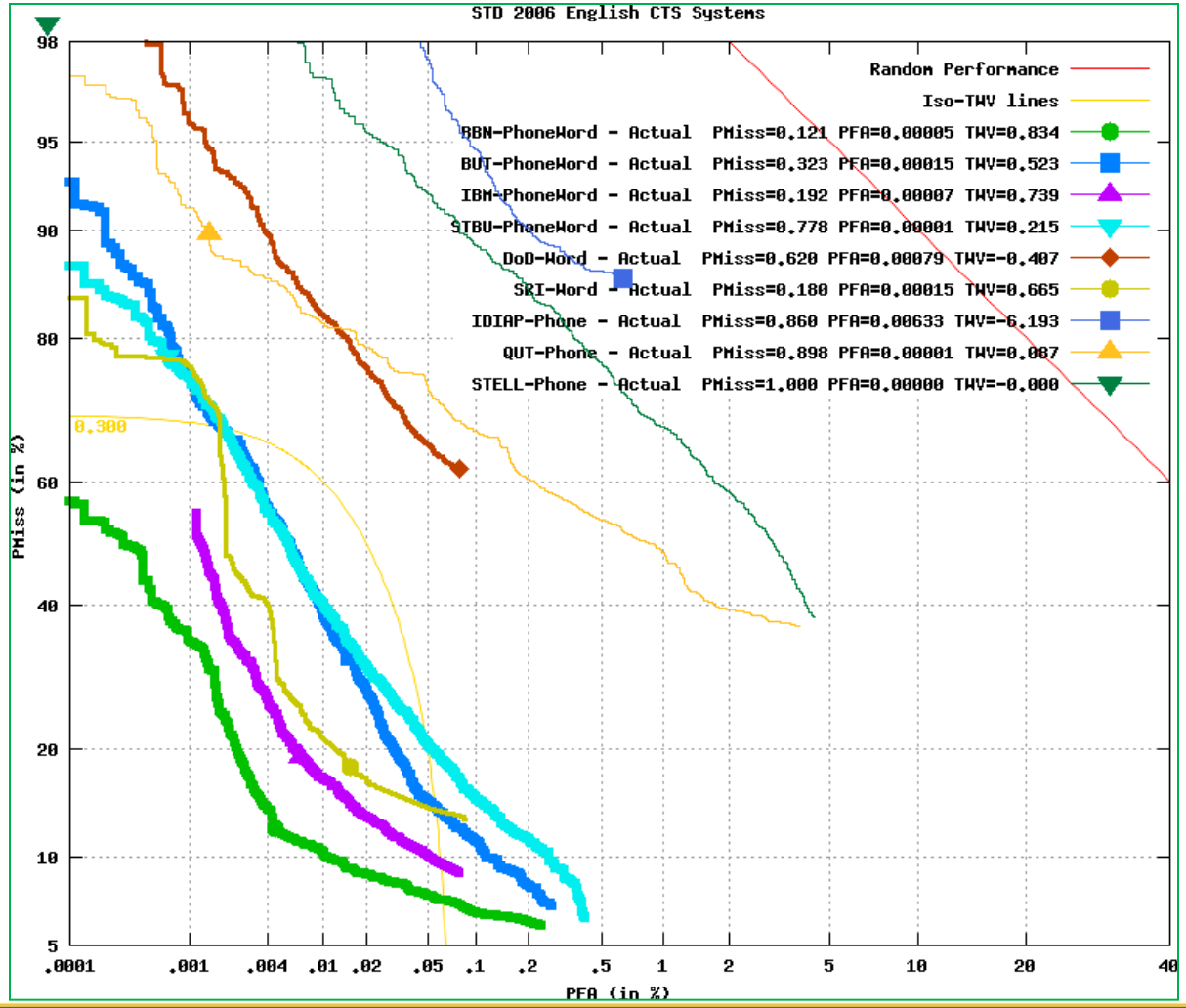
Reference Occ.

System Occ.

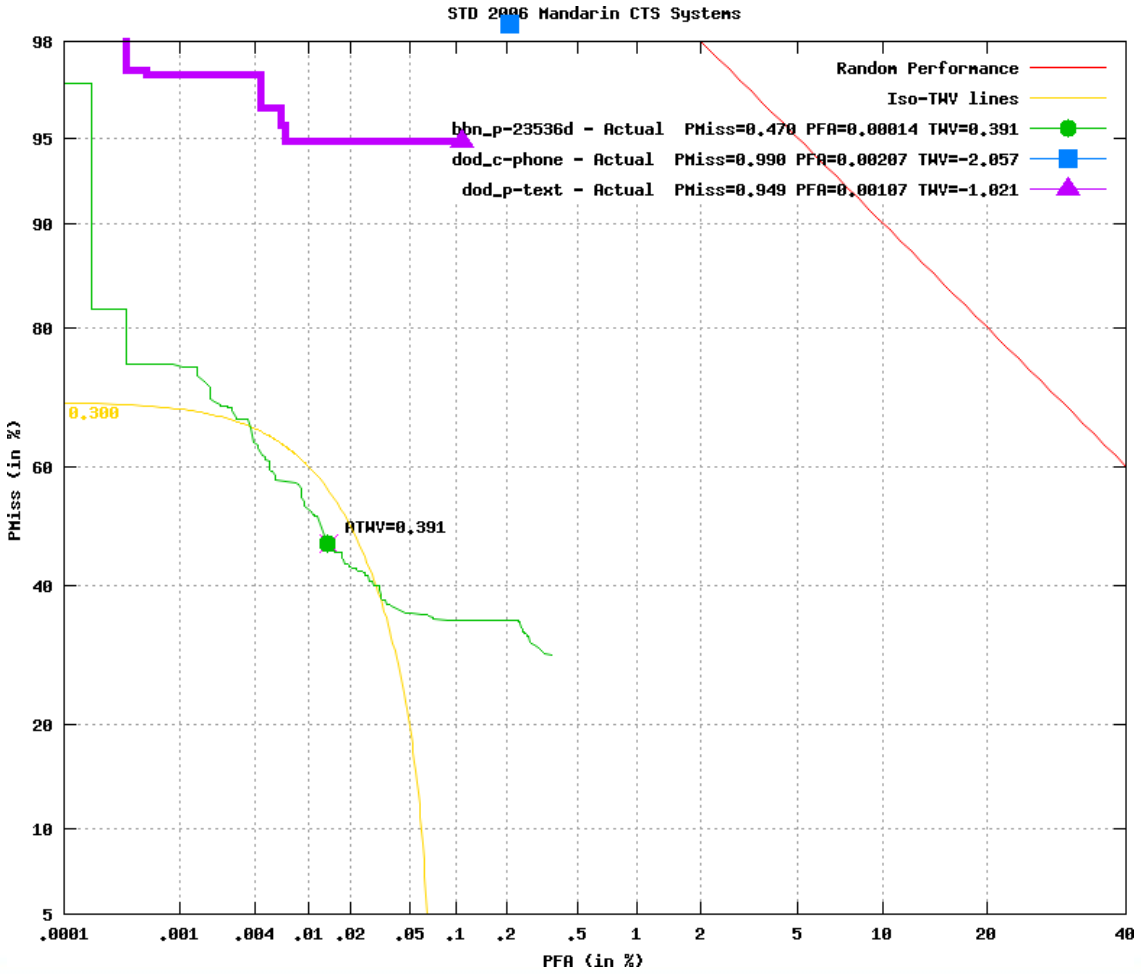
(with Θ = 0.6)



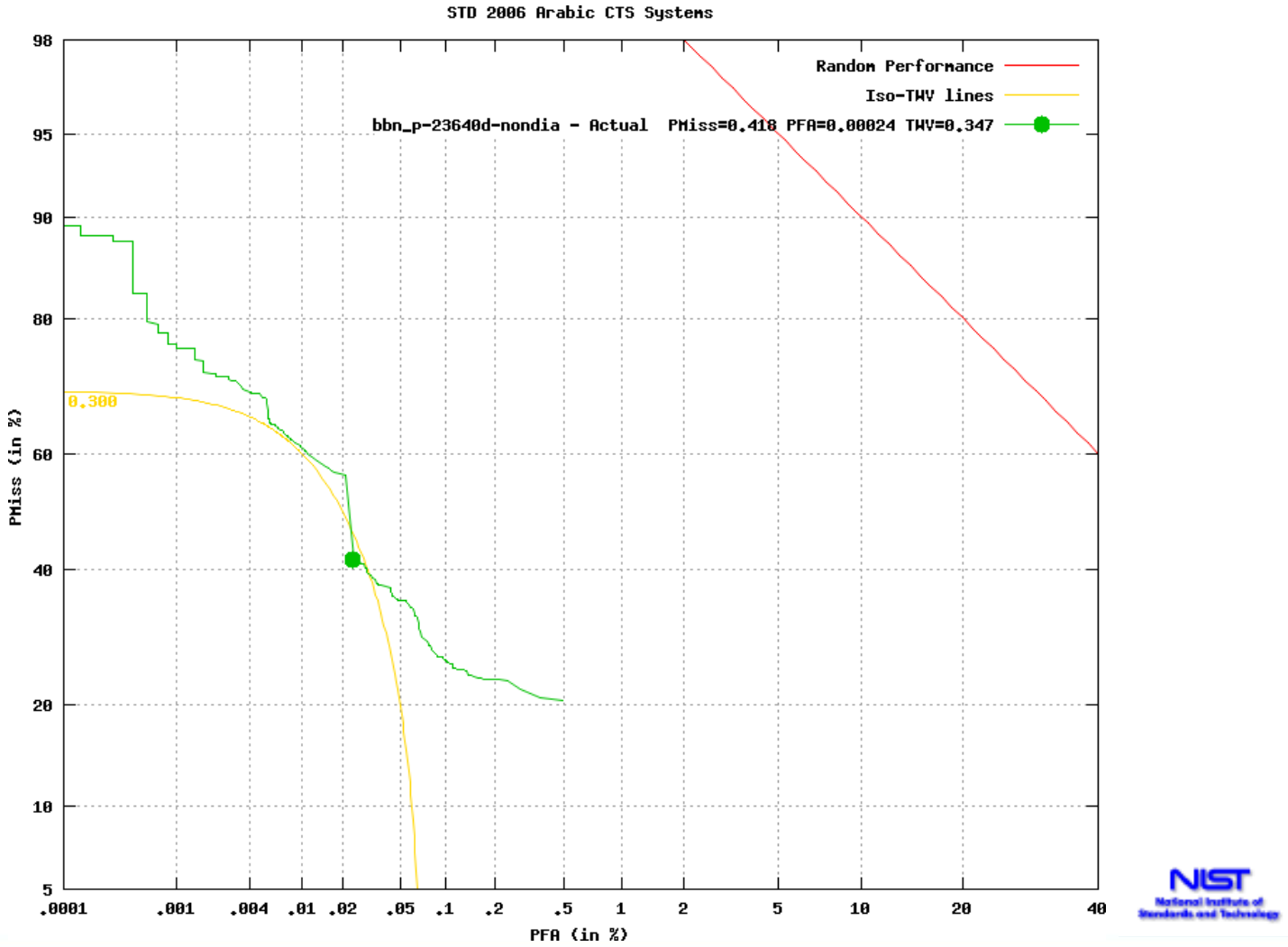
STD06: English CTS



STD06: Mandarin CTS



STD06: Non-Diacritized Arabic CTS



Snapshot of a Babel Program Period

Develop New Methods for N New Languages (~9 months)	Keyword Search Evaluation on the N languages (1 month)	Create Speech System for a Surprise Language (X weeks)	Keyword Search Evaluation on the Surprise (1 week)

Time

N= 4, 5, 6, and 7 Languages over the Program Periods
X = 4, 3, 2, and 1 Weeks over the Program Periods

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%
Channels	telephone	telephone and non-telephone	telephone and non-telephone	telephone and non-telephone
Languages Investigated Development+Surprise	4+1	5+1	6+1	7+1
Build Time for Surprise	4 weeks	3 weeks	2 weeks	1 week
Minimum NIST Actual Term Weighted Value (ATWV)	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.

T&E (Test and Evaluation)

- The Data
- The T&E Team

Test & Evaluation

MITRE

NIST National Institute of Standards and Technology

UNIVERSITY OF MARYLAND CASL CENTER FOR ADVANCED STUDY OF LANGUAGE

MIT Lincoln Laboratory

Civil Liberties, Privacy Protections, and Human Subjects

- The Civil Liberties and Privacy Office has reviewed the data collection process and the Babel program was found to have no CLPO issues.

- For each language, approximately 2000 subjects will sign consent forms and participate by speaking into a telephone, first to an automated system then with a friend of their choosing. The speech will be collected in a non-US country where the language is widely spoken. The government will not receive any PII (personally identifiable information). Additionally, the PII and the collected speech data are not at any point stored in the same computer system.

- The collection delivered to the government will contain audio recordings and transcription. Speech will be annotated with coarse-grained metadata, including speaker characteristics (gender, age, dialect spoken), channel (e.g., landline telephone, cell telephone, table top microphone at a distance), and environment (e.g., in a bar, at a restaurant, in a shopping mall, on the street/roadside, in an office, at home, in a moving vehicle).

- The data collection company has registered with the U.S. Department of Health and Human Services, Office for Human Research Protections, and has received Federal Wide Assurance (FWA) for the Protection of Human Subjects (Reference Number FWA00015539 with expiration date March 24, 2013). They have also registered with an approved Institutional Review Board (IRB) for review of the proposed collection method using human subjects.

The Data

- Audio data for the languages of the program is collected in-country. We anticipate collecting a total of 26 languages for the Program.
- The Languages of Base Period
  - Cantonese
  - Turkish
  - Pashto
  - Tagalog
  - And the Surprise!
- Data is being delivered in 2 deliveries: a partial A delivery and a full B delivery

Cantonese

- Sino-Tibetan, related to but not mutually intelligible with Modern Standard Chinese (*Putonghua* or *Guoyu*)
- Spoken in southern China: Babel data is from China's Guangdong & Guangxi Provinces (not Hong Kong or Macao)
- Written in Chinese characters: Babel data uses *simplified* characters
  - Vernacular writing not highly conventionalized: Morphemes that have no equivalent in Modern Standard Chinese may be represented with:
    - (Roughly) homophonous characters from the Modern Standard Chinese canon
    - (More or less) idiosyncratic vernacular characters
  - Word boundaries not marked in orthographic texts; word tokenization is problematic
- Phonology: intermediate complexity. According to Babel description:
  - Eight vowels, ten diphthongs
  - 19 consonants
  - Seven tones (some descriptions six to nine), limited tone *sandhi*
- Limited derivational morphology, *very* limited inflectional morphology

Cantonese Area

Cantonese Romanization & Tone

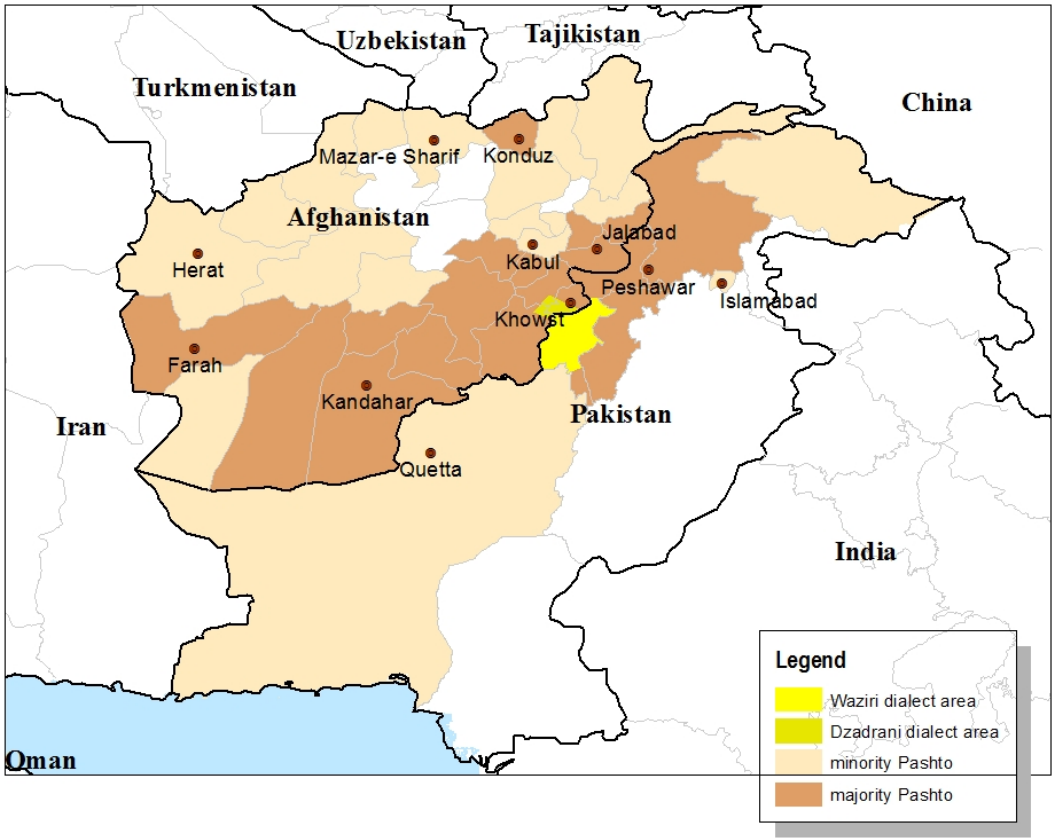
- Babel materials use a modified Yale Romanization:
  - Segments are identical to Yale (Huang & Kok 1970)
  - Yale represents tone with diacritics over the vowel and <h> after the syllable nucleus
  - Babel materials omit diacritics and <h>, and represent tone by numbers from 0 to 6 at the end of the syllable

Tone Contour	Yale	Babel
53	à	0
55, 5 + stop	ā, ā + /p, t, k/	1
35	á	2
33, 3 + stop	a, a + /p, t, k/	3
21	àh	4
13	áh	5
22, 2 + stop	ah, ah + /p, t, k/	6

Pashto

- Indo-Iranian (related to Iranian languages, but influenced by Indo-Aryan languages)
- An official language in Afghanistan, also spoken in Pakistan
- Written in Perso-Arabic script
  - Some characters are not in Arabic, recommend a wide-coverage Arabic font (e.g. SIL's Scheherazade)
  - Not all vowels are written in Perso-Arabic script, some phonemes have > 1 representation (e.g. four letters with same /z/ sound). SAMPA transcriptions also to be provided.
- Phonology
  - Large number of consonant phonemes: 30
    (cf. English ~24)
  - Contrasts dental vs. retroflex (+ palato-alveolar fricatives and affricates)
  - Smaller number of vowel phonemes: 7
  - *Some* speakers *might* add "elegant" (= Arabic) consonants *sometimes*

Pashto Area


Uzbekistan
Tajikistan
Turkmenistan
China
Mazar-e Sharif
Konduz
Afghanistan
Jalabad
Herat
Kabul
Peshawar
Islamabad
Khowst
Farah
Kandahar
Pakistan
Iran
Quetta
India
Oman
Legend
Waziri dialect area
Dzadrani dialect area
minority Pashto
majority Pashto

Pashto

- Dialectal variation, esp. in phonology
- Example (IPA) – for the letters ښ ږ څ ځ

– *Southwestern/ Kandahar “soft”	/ʂ/	/ʐ/	/ts/	/dz/
– Southeastern	/ʃ/	/ʒ/	/ts/	/dz/
– Northeastern/ Peshawar “hard”	/x/	/g/	/s/	/z/
– Northwestern	/ç/	/ɟ/	/s/	/z/
– Middle (including Waziri)	/ʃ/	/ʒ/	/s/	/z/

Pashto Morphology

- Nouns and Adjectives
  - Multiple declension classes
  - Suffixes mark case, number, gender/ animacy
  - Some stem allomorphy

    ɣal ‘thief’ (direct case, singular)
    ɣlo ~ ɣlúno (oblique/ ablative plural)
- Verbs
  - Prefixes and suffixes marking tense, aspect, mood, subject person/ number/ gender
  - Some stem allomorphy

    raség-əm ‘I arrive/am arriving’,
    wə-rased-əy ~ wə-rased-əl-əy ‘you-all arrived’

Turkish

- Turkic language family
- Official language of Turkey
- Written in Latin script
- Intermediate number of phonemes (8 vowels, 23 consonants)
  - Includes front rounded and back unrounded vowels
- Some dialectal variation; “standard” dialect is that of Istanbul

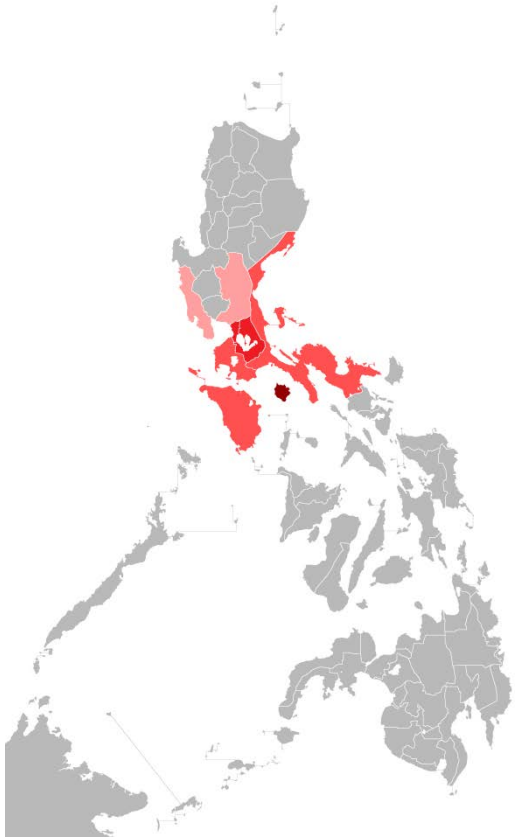
Turkish Morphology

- Agglutinating, strictly suffixal; very little irregularity
- Vowel harmony in suffixes
- Some phonological processes affect stem (devoicing)
- Nouns mark case, number, person of possessor (optional);
  ev-ler-in-izin ‘of your houses’
  ağaç-lar-ın-ızın ‘of your trees’
- Adjectives don’t decline (unless acting as noun)
- Verbs mark tense, aspect, mood/ evidential, negation, and subject person
  gel-mez-se-ler ‘if they did not come’
  oku-maz-sa-lar ‘if they did not study’
- Abundant morphological resources available

Tagalog (aka Filipino)

- Central Philippine language (Austronesian family)
- An official language of the Philippines (with English)
- Written in Latin script
- Intermediate phonology (six vowels, nine diphthongs,19 consonants)
  - Word-final voiceless stops often unreleased
  - Vowel length can be contrastive, but not written in orthography (nor is word-final /h/)
    aso /a:soh/ ‘dog’
    aso /asoh/ ‘smoke’
- Some dialectal variation (loss of glottal, [r] ~ [d], some morphology and lexical differences)

Tagalog



Tagalog Morphology

- Nouns not inflected, but “case”-marked by preceding particles
- Verbs are complex: marked by prefixes, suffixes, infixes, reduplication for “focus”, aspect, mode, voice

nag-*sabi*	‘say’	(actor focus, completed)
mag-sa-*sabi*		(actor focus, “contemplated”)
sa-*sabi*-hin		(object focus, “contemplated”)
s-um-*ayaw*	‘dance’	(actor focus, completed)
sa-*sayaw*		(actor focus, “contemplated”)
sa-*sayaw*-in		(object focus, “contemplated”)

Audio Data Statistics for Cantonese A Delivery

- Conversational Cantonese
  - 420 unique files
  - 207 unique sessions
    - 192 paired sessions with single 10-minute call
    - 7 multi-call paired sessions
    - 8 unpaired sessions (i.e. no corresponding inLine or outLine)
  - 67.7 total hours of audio
    - About half of it containing speech

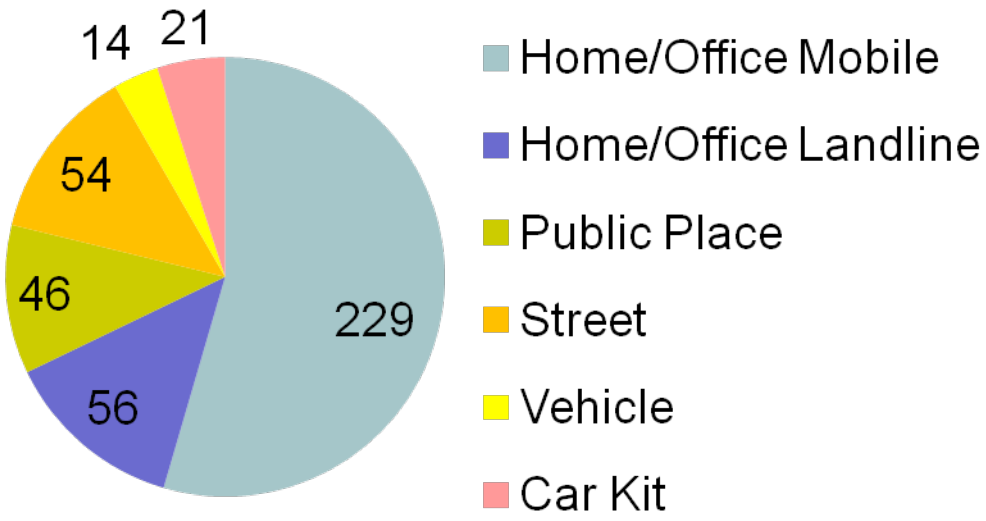
Audio Data Statistics for Cantonese A Delivery

- Scripted Cantonese
  - 6844 unique files
  - 167 unique sessions
    - Each matched to inLine speaker of conversational session with same ID
    - Each containing 40 or 41 audio files
  - 14.1 total hours of audio

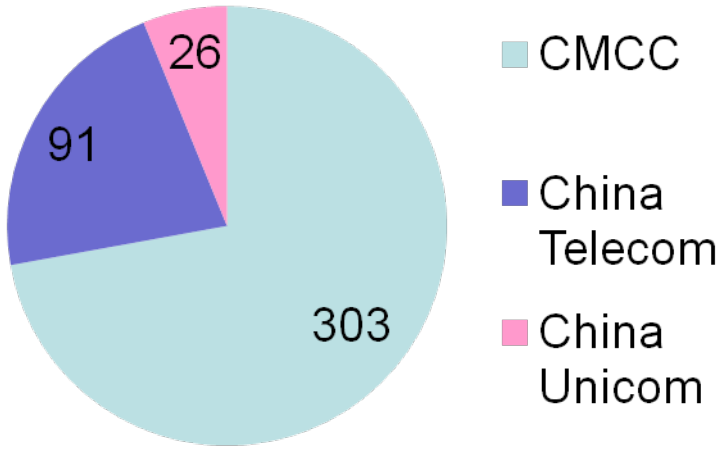
Varied Environments and Channels

- Calls recorded in a variety of environments and channels
  - 3 primary networks
  - Both landline and mobile calls
  - Many different phone models (including hands-free)
  - Multiple environments: office, street, in-vehicle, etc.

Call Sides per Environment



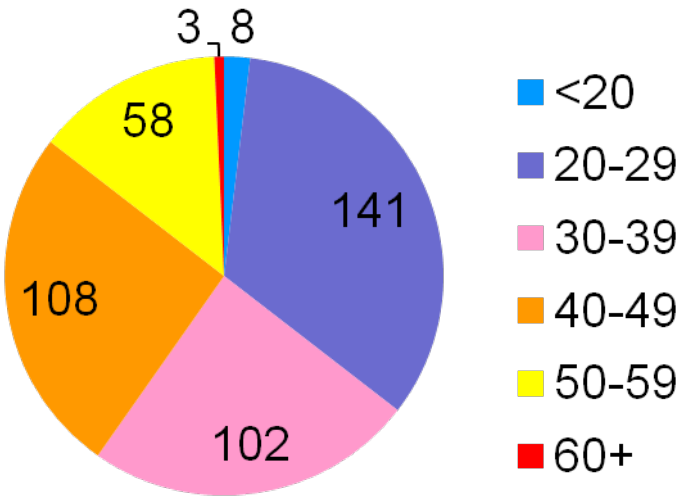
Call Sides per Network



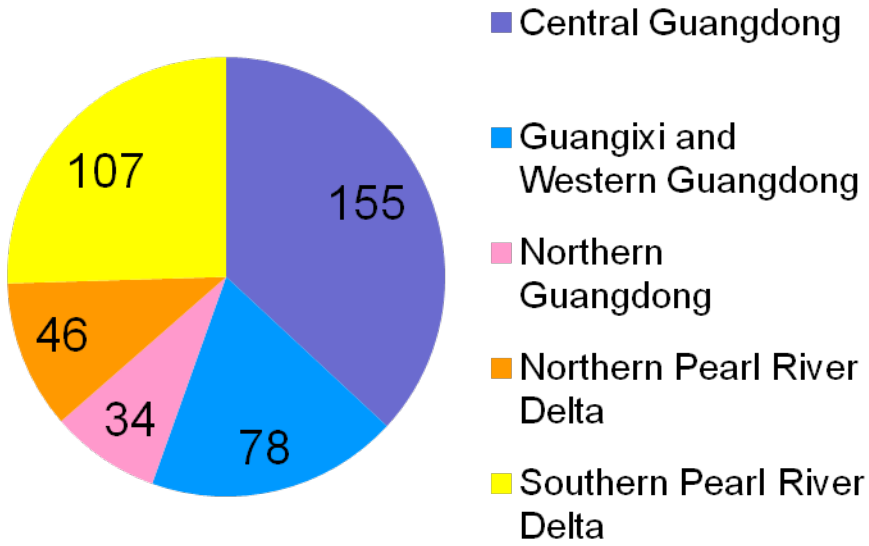
Many Speakers and Multiple Dialects

- Wide variety of speakers
  - 263 females / 157 males
  - Ages between 16 and 65
  - Collected in five regions to ensure variation in dialects

Call Sides per Age Group



Call Sides per Network



Environment Audio Issues

- Background speech

  有 啊 <int> 有 啊 (()) 工夫 (()) 冻烂 啦 (()) 啊 吓 系 咪

  [*Yes <int> yes (()) effort (()) ??? (()) huh huh yes*]

- In vehicle recording

  听唔听到 哦 听到 啊 系 嘛 有 啊 我 问 你 系 咪 去 香港 买- 买 佳能 相机

  [*Can you hear me? I can hear you. I want to ask you if you went to Hong Kong to buy the camera.*]

- In vehicle recording (hands free)

  唔 系 买 部 通风 三 零 六 定 三 零 七 咩

  [*Don't buy 3 6 3 7.*]

Cell Phone Audio Issues

- Cell phone coding errors

  啊 有时 **睇 埋 有** 冇 **睇 埋** 嘸 啲 **咩 新**闻 **啊** 啲

  [*Sometimes I watch some news.*]

- Cell phone drop outs

  哦 (()) **哦 而家 系** <hes> **打包 系 嘛**

  [*Oh. (()) Right now, yes. <hes> Pack it to go.*]

- Clipping

  <sta> **你 你 等等 先** 妈咪 **等等 先** 妈咪 倾 紧 电话 倾 紧 电话 **你 放**
  喺 **果度 先 放 系 碗 果** 头 **先 咯 哦 咁**样 **系 咪 好 啊**

  [*Wait. Wait for mommy to speak on the phone. Put it where the bowl is. Yes.*]

T&E Team Roles

UNIVERSITY OF MARYLAND CASL CENTER FOR ADVANCED STUDY OF LANGUAGE

- Center for the Advanced Study of Language (CASL) are providing guidance on the languages, data quality, and aspects of the language that affect search terms for evaluation.

  POCs: Anton Rytting and Mike Maxwell

NIST National Institute of Standards and Technology

- The National Institute of Standards and Technology (NIST) are serving as the evaluation lead for the T&E team. They will design and administer the test sets via a test server.

  POC: Jon Fiscus

T&E Team Roles

MIT Lincoln Laboratory

- MIT Lincoln Laboratory (LL) will be supporting Babel with a wide array of technical support functions in support of sound evaluations. They will receive the collected data and partition the files into training, development, and evaluation sets.

  POCs: Wade Shen and T. J. Hazen

MITRE

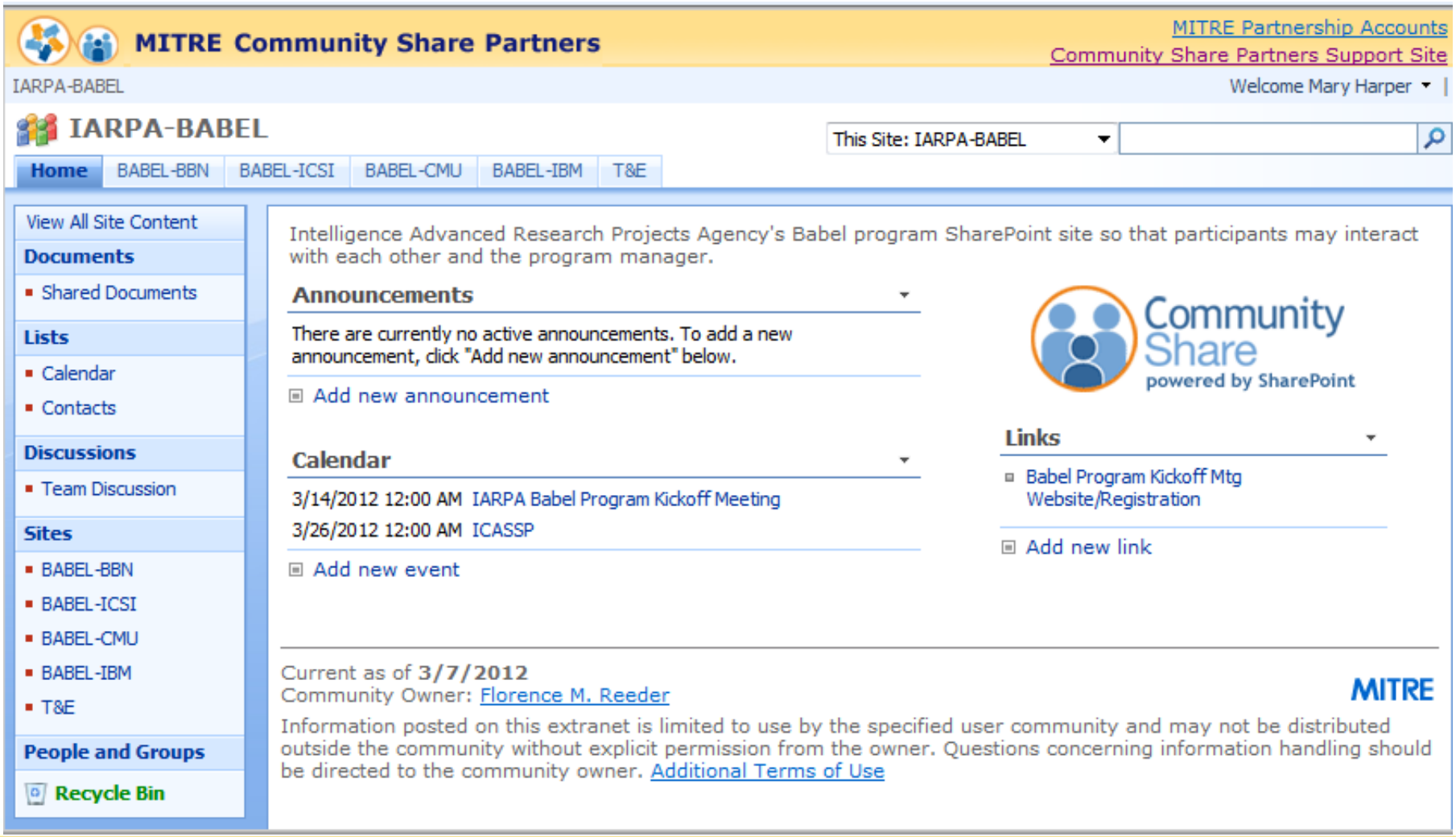
- MITRE will be helping to coordinate the T&E pipeline to ensure timely delivery of data to performers, as well as other program-level support.

  POC: Evelyne Tzoukermann

Babel Research SharePoint Site

- The Babel SharePoint site will facilitate collaboration and information sharing between the Program Office and the Research Teams.
  - The general information site will be accessible by all authorized Babel stakeholders and performers.
  - Each performer will have a private sub-site accessible only by its members and the Babel Program Office staff.
- Performers and government stakeholders will receive an email invitation to join the Babel SharePoint site.
- https://partners.mitre.org/sites/IARPA-BABEL/default.aspx

Snapshot of SharePoint Site



BABEL Program IARPA POCs

- Program Manager
  - Dr. Mary Harper, Mary.P.Harper@ugov.gov
  - IARPA, Incisive Analysis Office
  - Office of the Director of National Intelligence
    Intelligence Advanced Research Projects Activity
    Washington, DC 20511
  - Fax: 301-851-7672
  - Phone: 301-851-7429
  - mary.p.harper@ugov.gov
- Technical SETA
  - Ms. Amy Jarrett, Amy.Jarrett@ugov.gov
- Programmatic SETA
  - Ms. Kimiko Crummedy, Kimiko.B.Crummedy@ugov.gov