

CLASR: Cross-Language Automatic Speech Recognition

John Henderson

781-271-2849 • jhndrsn@mitre.org

MITRE Sponsored Research

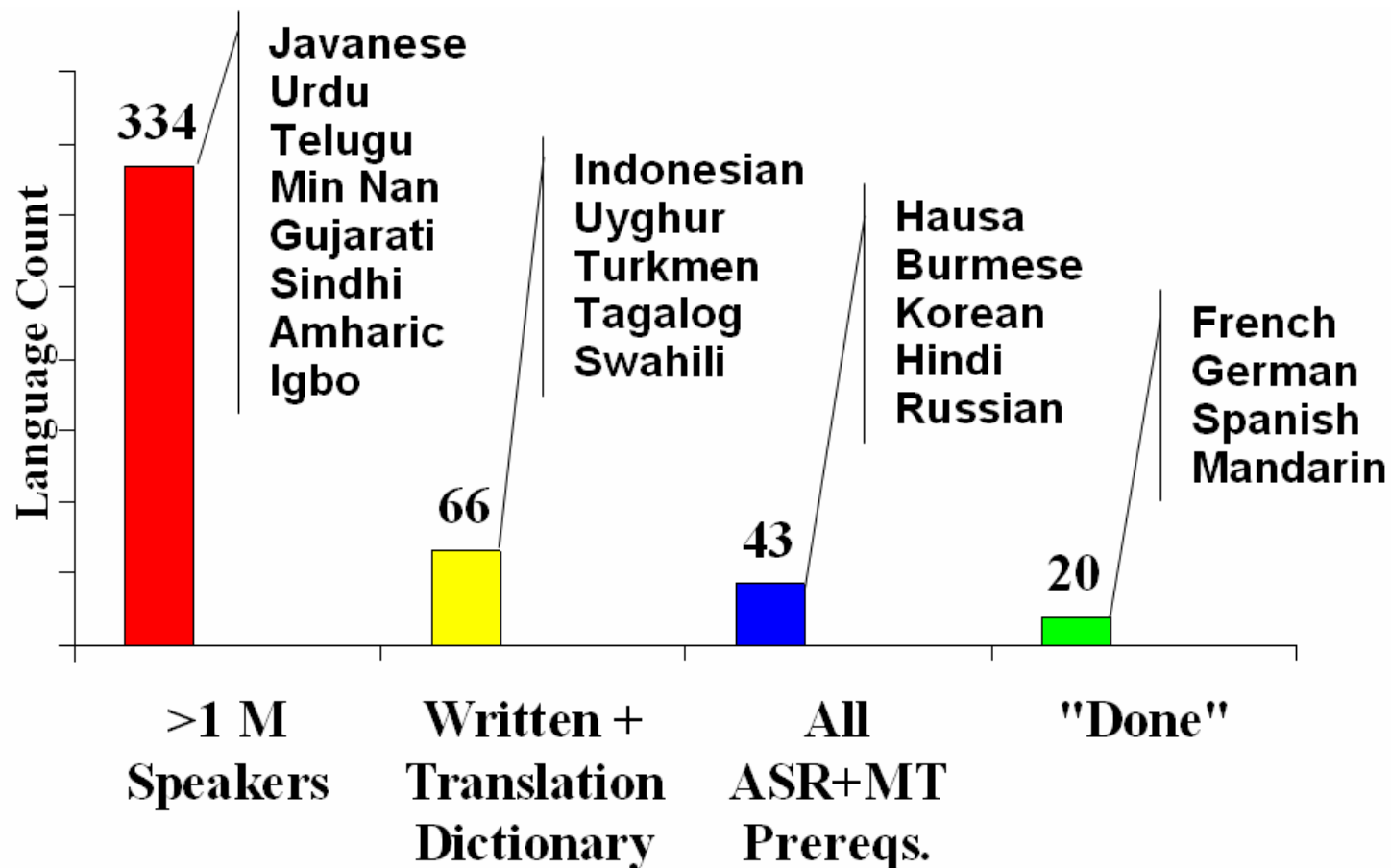


Problem

- **More than 300 languages have a million speakers.**
- **Fewer than 70 languages have standardized written forms and extensive written texts.**
- **There are too many languages and too few linguistic resources to create enough speech translation systems.**

Background

Resources needed to develop speech recognition and machine translation systems are not common for rare, high-value languages.



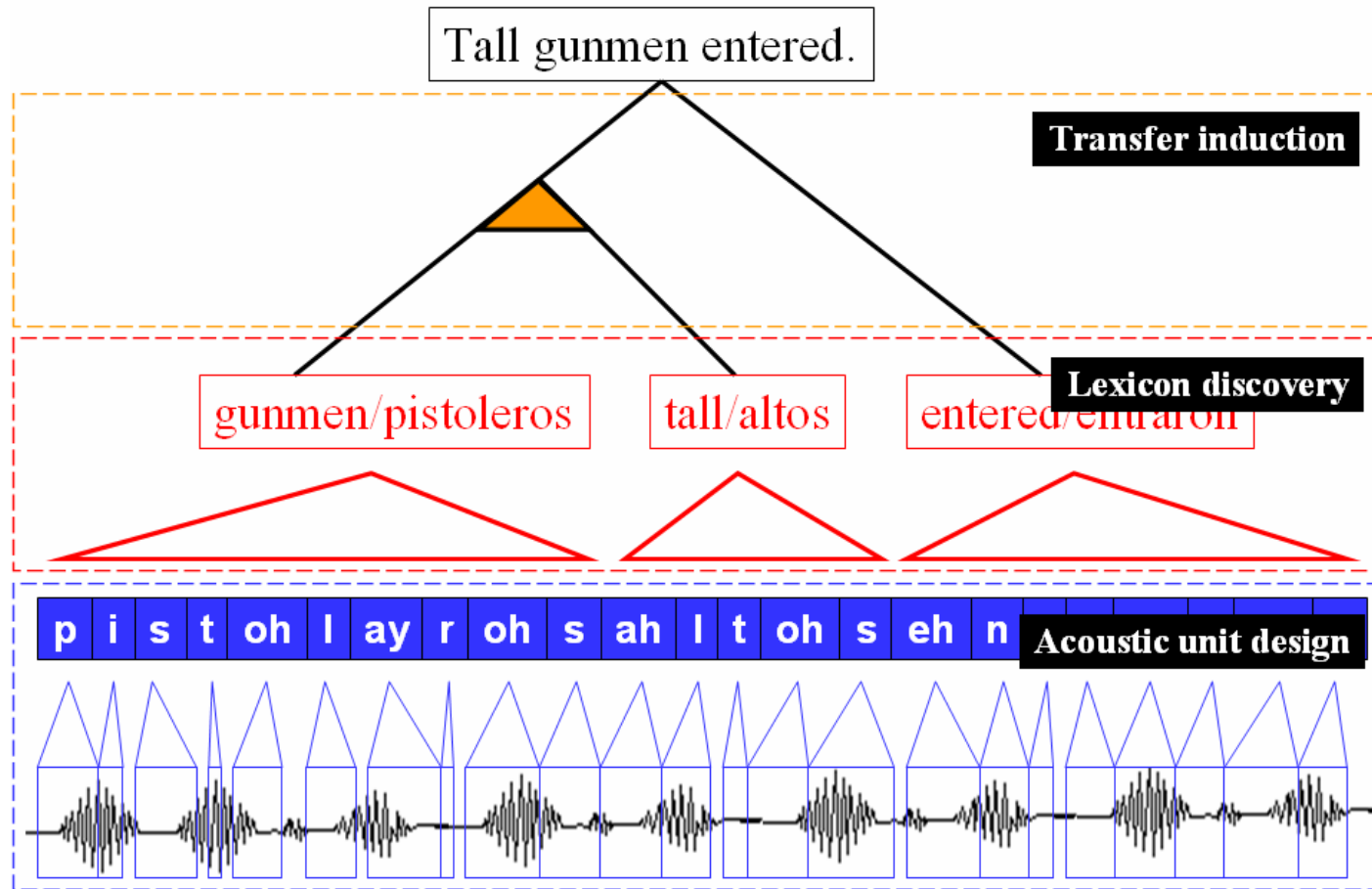
Objective

- **We are developing a system for spoken language translation of languages that lack significant written resources.**
 - **Acoustic units must be induced.**
 - **Bilingual pronunciation dictionaries must be derived.**
 - **Word reordering must be learned.**

Activities

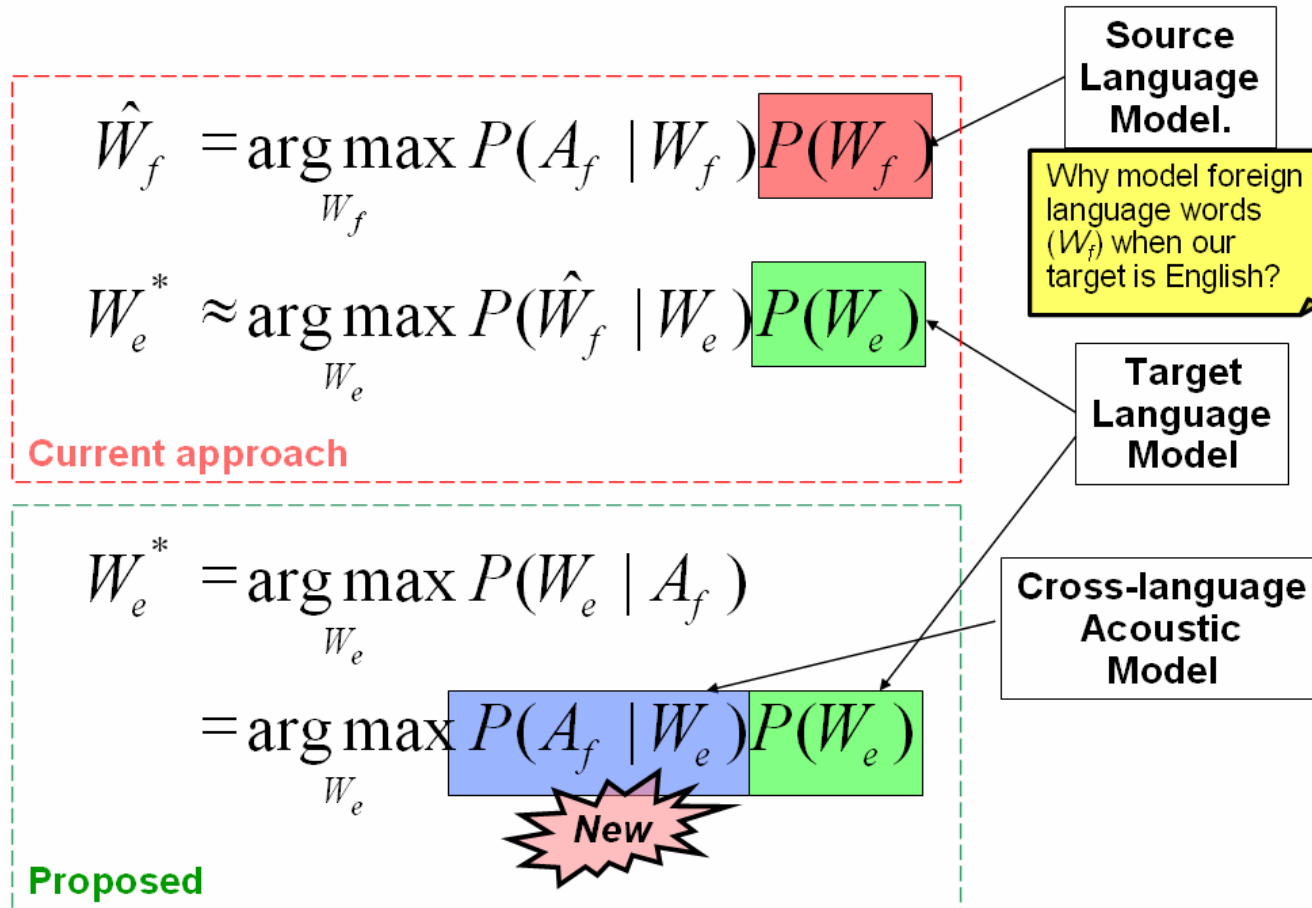
- **Year 1: Preliminaries**
 - **Develop dataset (translate and align)**
 - **Assemble relevant research code**
 - **Establish traditional, two-stage baseline system**
- **Year 2: Prototype of novel joint model system**
- **Year 3: Incorporate best available MT and ASR models into prototype**

Highlight



Three separate stochastic models are used to describe the process of producing foreign language audio from English text. Decode in reverse.

Highlight

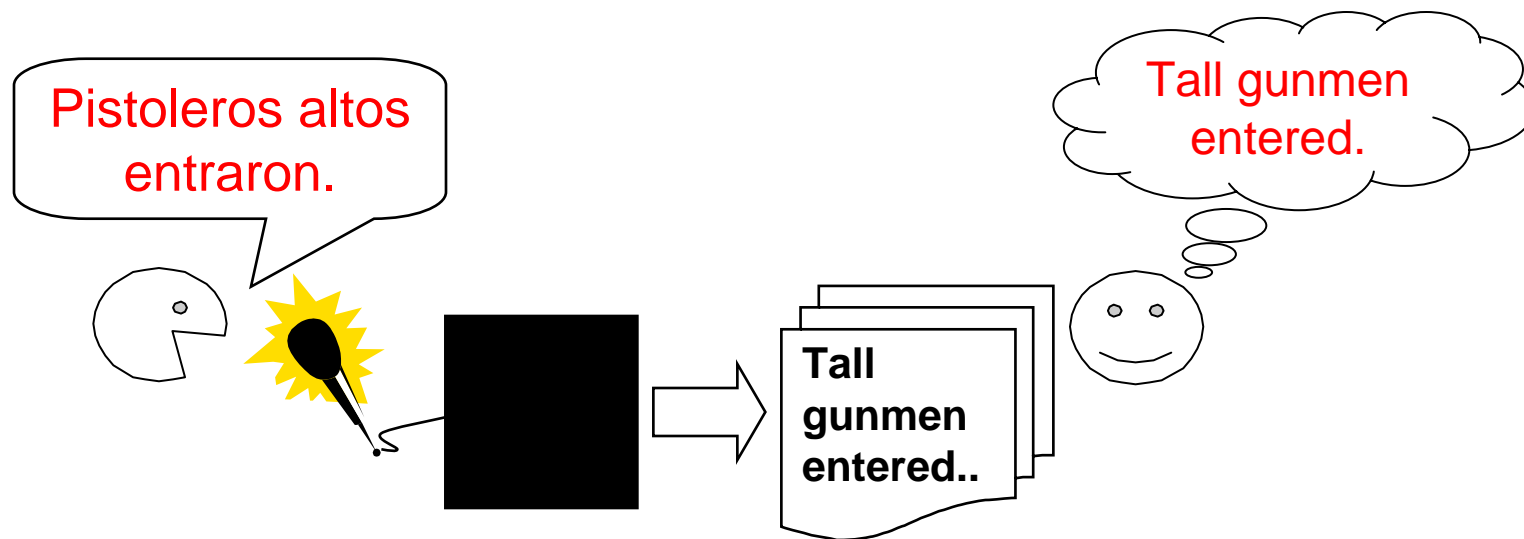


The new approach will reduce errors compounded in two-stage decoding and utilize signal information missing in the foreign language written form.

Impacts

- **Universal access** to massive audio collections
 - Support audio triage: coarse, rough characterization of foreign language audio
 - Enable global infectious disease monitoring
- **New approach** to an old problem
 - Joint model of ASR & MT is closer to the real optimization problem for many scenarios.
 - Results will widen the bridge between the ASR and MT research communities.

Future Plans



In the second year of this project we will produce the first single-stage spoken language translation systems for Spanish and Mandarin.