Speech and Audio Technology for Enhanced Understanding of Cognitive Radio Users and Environments

Scott M. Lewandowski, Joseph P. Campbell,
William M. Campbell, Clifford J. Weinstein
{scl, jpc, wcampbell, cjw}@ll.mit.edu
MIT Lincoln Laboratory
Lexington, MA

Software Defined Radio Forum Technical Conference
Phoenix, AZ
15-18 November 2004

This work was sponsored by the Defense Advanced Research Projects Agency under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Government.

Outline

- Introduction & Motivation: Cognitive Radio
- Speech Technologies:
  - Speaker Recognition
  - Language Identification
  - Text-to-Speech
  - Speech-to-Text
  - Machine Translation
  - Background Noise Suppression
  - Adaptive Speech Coding
  - Speaker Characterization
  - Noise Characterization
- Conclusions

Cognitive Radio and the Mobile Land Warrior

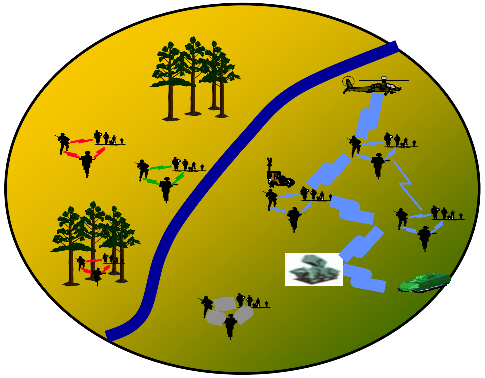
Sense & understand the user's state and needs

- Personalization, adaptation, authentication (PAA)
- Health state, stress

Provide robust radio comm.

Sense & understand the situation

- Friends, resources
- Foes, threats



Provide plan & decision assistance

- Team plan including rendezvous
- Continuous planning of actions/alternatives

Features & benefits

- Automated learning & reasoning about user & environment
- User focus on mission
- Enhanced mission effectiveness

> "If you know the enemy and know yourself, you need not fear the result of a hundred battles." Sun Tzu

Today and Tomorrow: Example Scenarios

Without Cognitive Radio

User manually adjusts

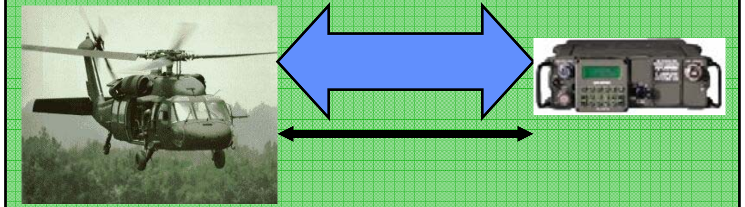
User Aware:

Speech technologies provide state, identity, and interface to the user.

RF Aware: **Links are established automatically by reasoning. The radio is aware of other networks and radios.**

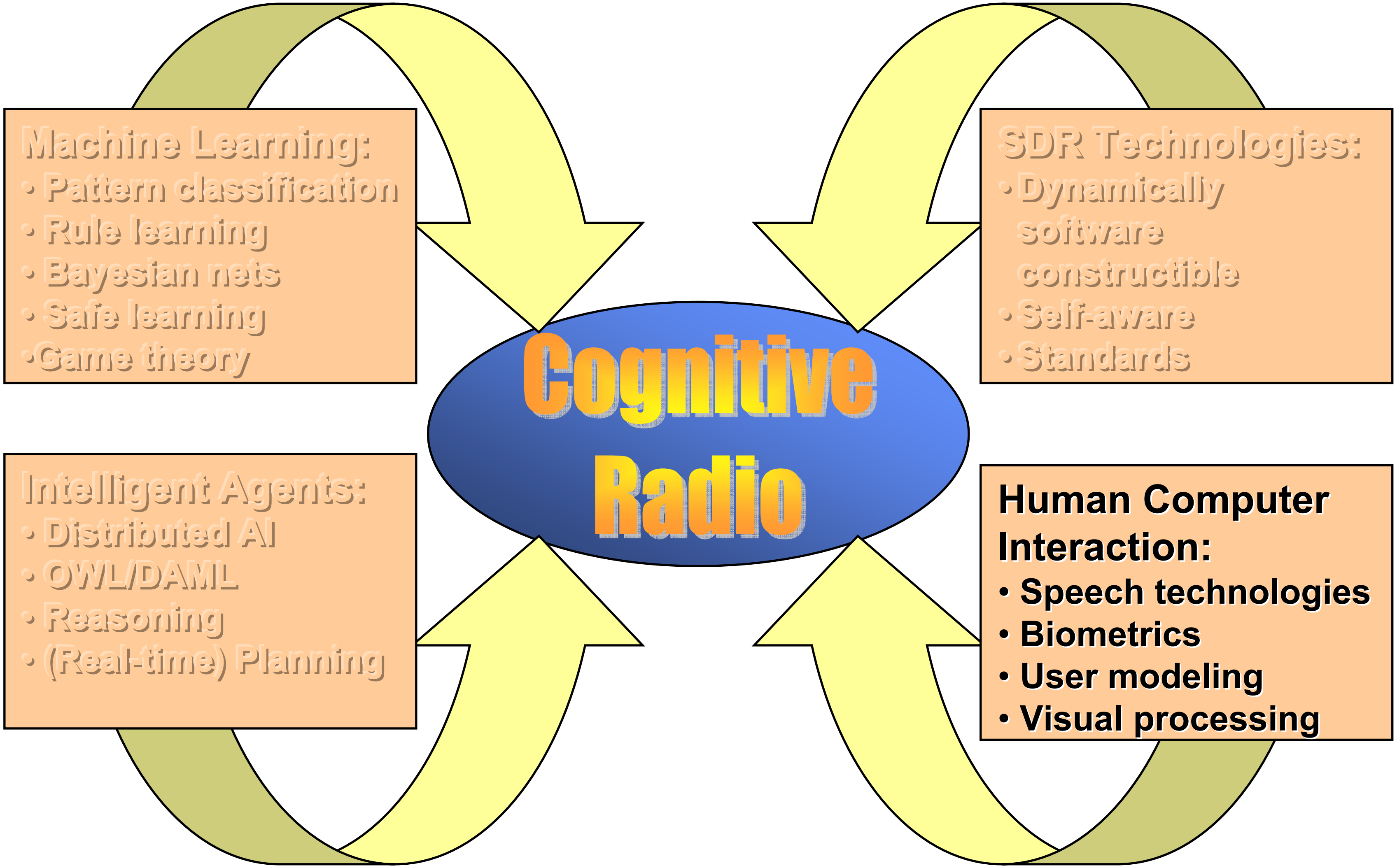
Environment Aware: **Situationally aware radio assists the user and understands rendezvous, location, and enemy & friendly forces.**

With Cognitive Radio



N
ENEMY X
FRIEND O
Assault Deploy
O
X
Defensive Peak
Upper Ridge
Center Hill
Defense Deploy
Assault Objective
Tunnel Convoy
Lower Ridge
X
Assault Objective
Blast Site Envoy

Cognitive Radio Technologies



Machine Learning:

- Pattern classification
- Rule learning
- Bayesian nets
- Safe learning
- Game theory

SDR Technologies:

- Dynamically software constructible
- Self-aware
- Standards

Cognitive Radio

Intelligent Agents:

- Distributed AI
- OWL/DAML
- Reasoning
- (Real-time) Planning

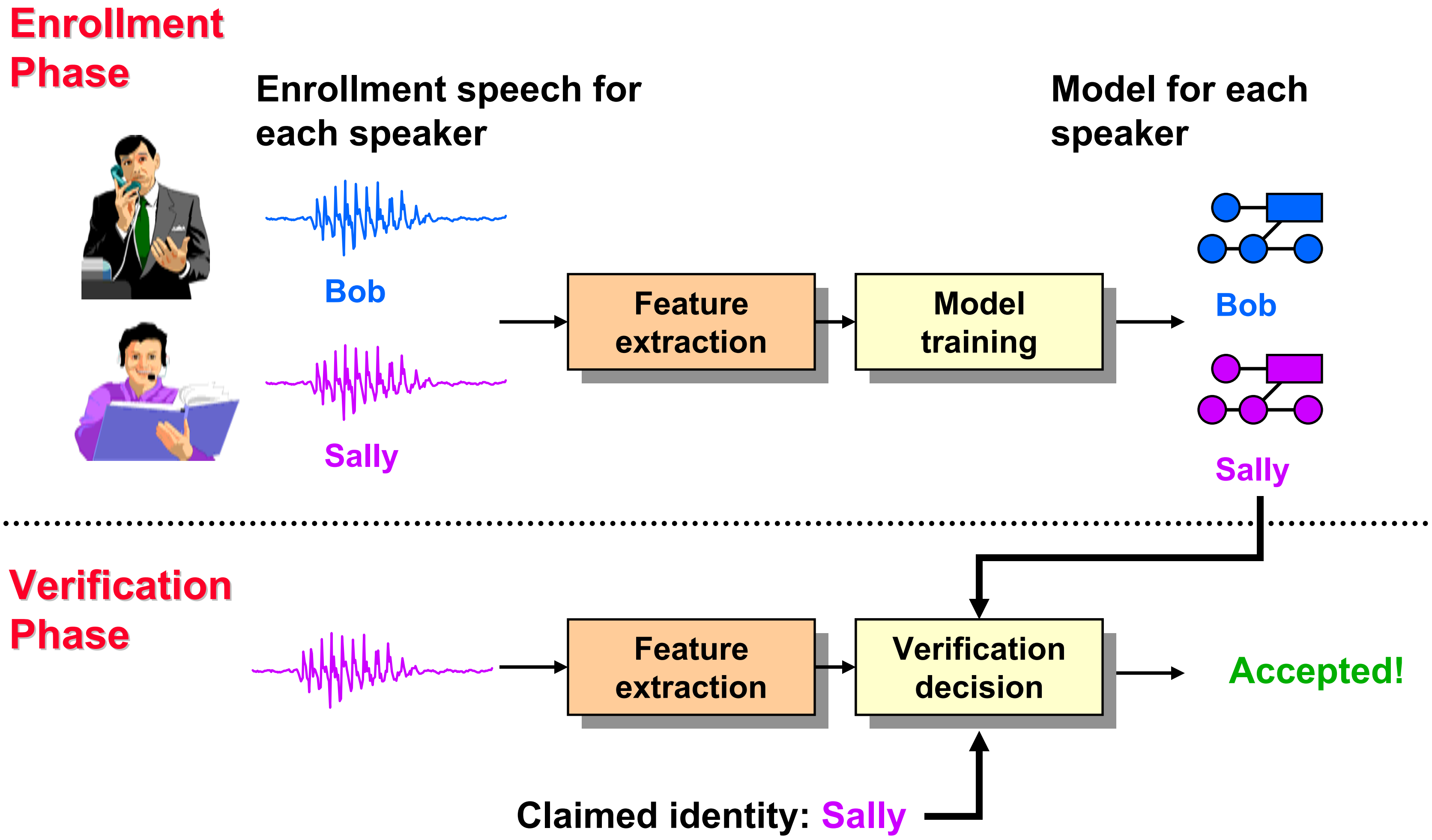
Human Computer Interaction:

- **Speech technologies**
- **Biometrics**
- **User modeling**
- **Visual processing**

Speaker Recognition

Phases of a Speaker Verification System

Two distinct phases to any speaker verification system



Enrollment Phase

Enrollment speech for each speaker

Model for each speaker

Verification Phase

Claimed identity: Sally

Speaker Recognition and Cognitive Radio

Cognitive Radio applications:

- Personalization (e.g., recalling user preferences or accomodating a user's unique workflow)
- Adaptation (e.g., simplifying the user interface based on the current task, or modifying radio parameters according to environmental factors)
- Authentication (e.g., detecting captured/stolen/lost devices, or providing "hands-free" biometric authentication)

References:

- Campbell, J. P., Campbell, W. M., Jones, D. A., Lewandowski, S. M., Reynolds, D. A., and Weinstein, C. J., "Biometrically Enhanced Software-Defined Radios," in Proc. Software Defined Radio Technical Conference in Orlando, Florida, SDR Forum, 17-19 November 2003.
- D.A. Reynolds, T.F. Quatieri, R.B. Dunn. "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, 10(1--3), January/April/July 2000.
- Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., and Leek, T. R., "High-Level Speaker Verification with Support Vector Machines," in Proc. International Conference on Acoustics, Speech, and Signal Processing in Montréal, Québec, Canada, IEEE, pp. I: 73-76, 17-21 May 2004.

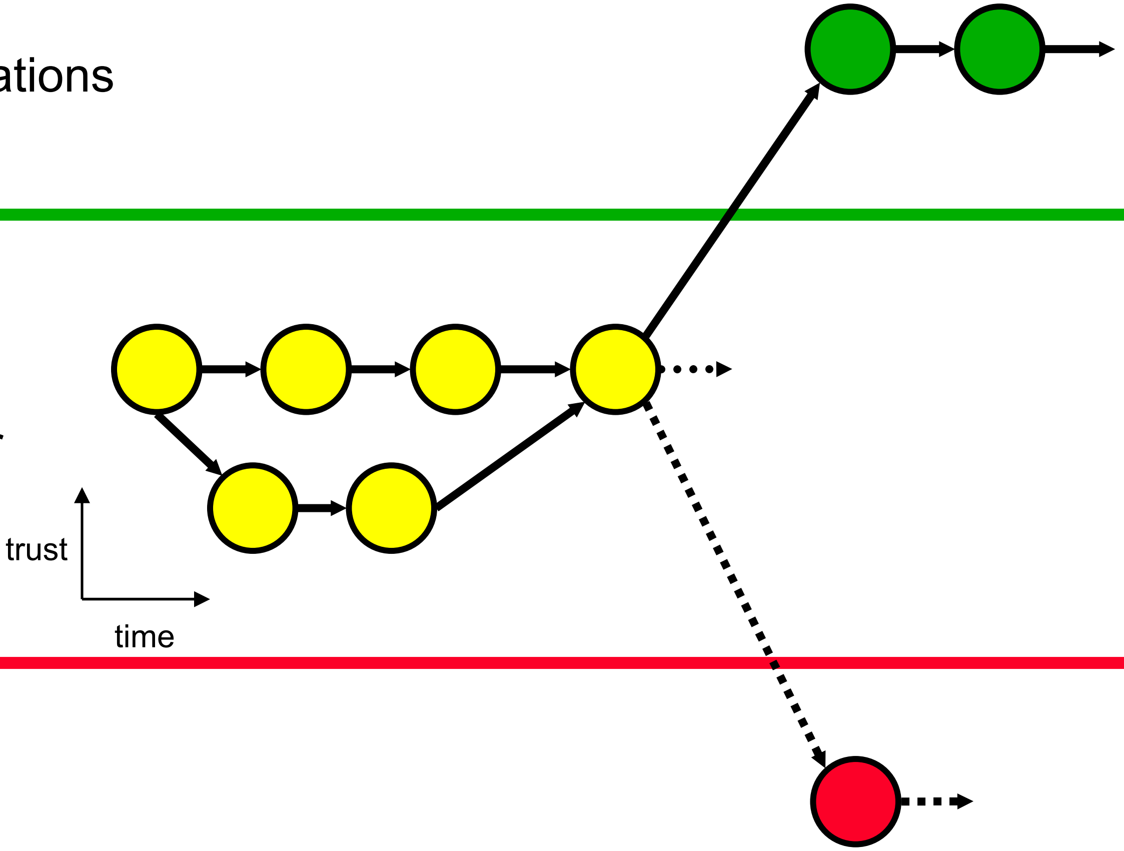
Continuous Authentication via Behavior & Voice Recognition

Trusted State

Required for sensitive operations

Provisional Trust

Continue interaction, gather behavioral & voice samples



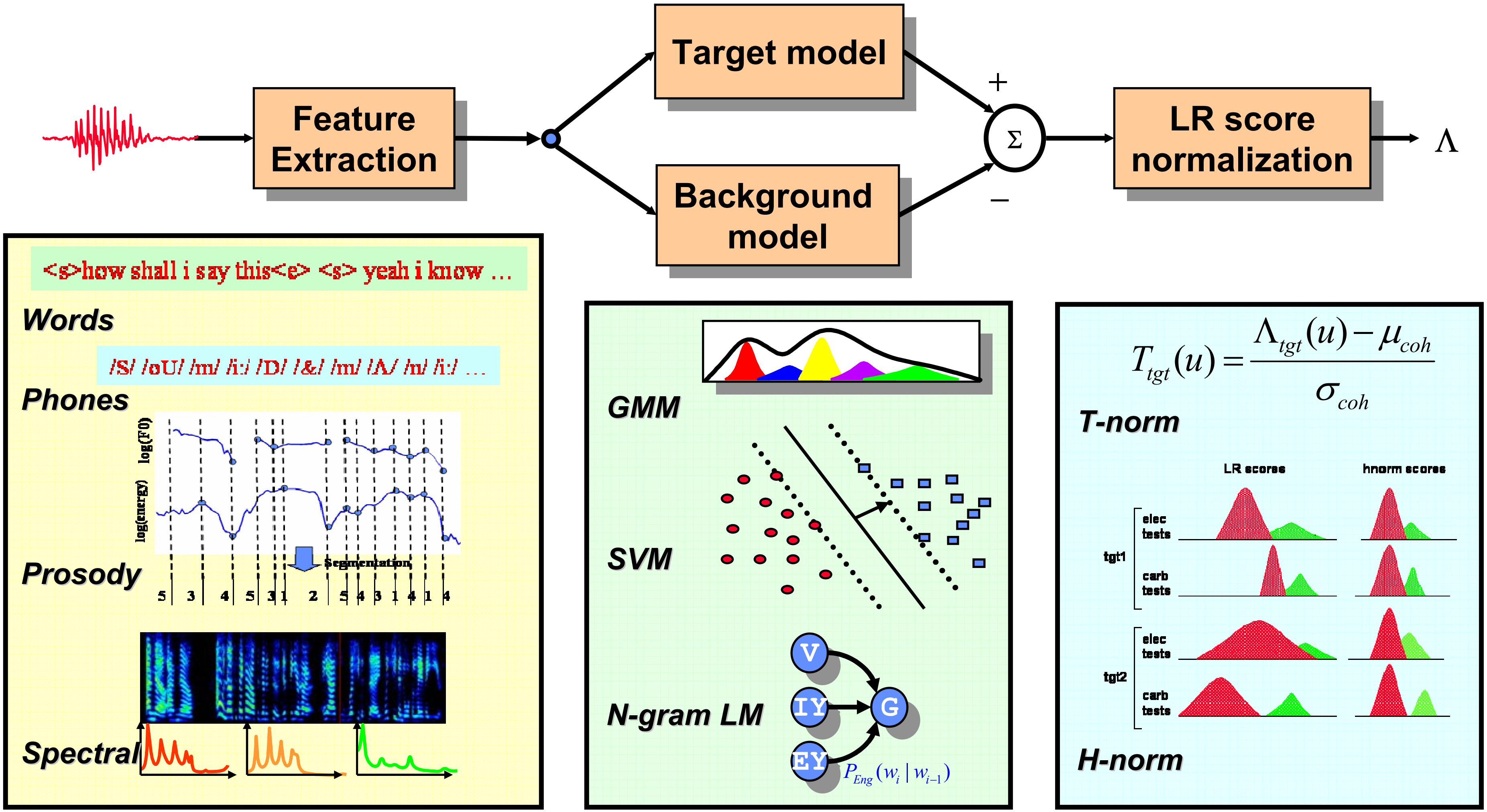
Untrusted State

Interrupt interaction

T. J. Hazen, D. Jones, A. Park, L. Kukolich, D. Reynolds, “Integration of Speaker Recognition into Conversational Spoken Dialogue Systems,” *Eurospeech,* 2003.

Speaker Recognition Core Technologies

- Basic decision statistic in core detectors is the likelihood-ratio



Feature Extraction

Target model

Background model

+

–

Σ

LR score normalization

Λ

<s>how shall i say this<e> <s> yeah i know …

Words

/S/ /oU/ /m/ /i:/ /D/ /&/ /m/ /A/ /n/ /i:/ …

Phones

Prosody

Spectral

GMM

SVM

N-gram LM

$P_{Eng}(w_i \mid w_{i-1})$

$$T_{tgt}(u) = \frac{\Lambda_{tgt}(u) - \mu_{coh}}{\sigma_{coh}}$$

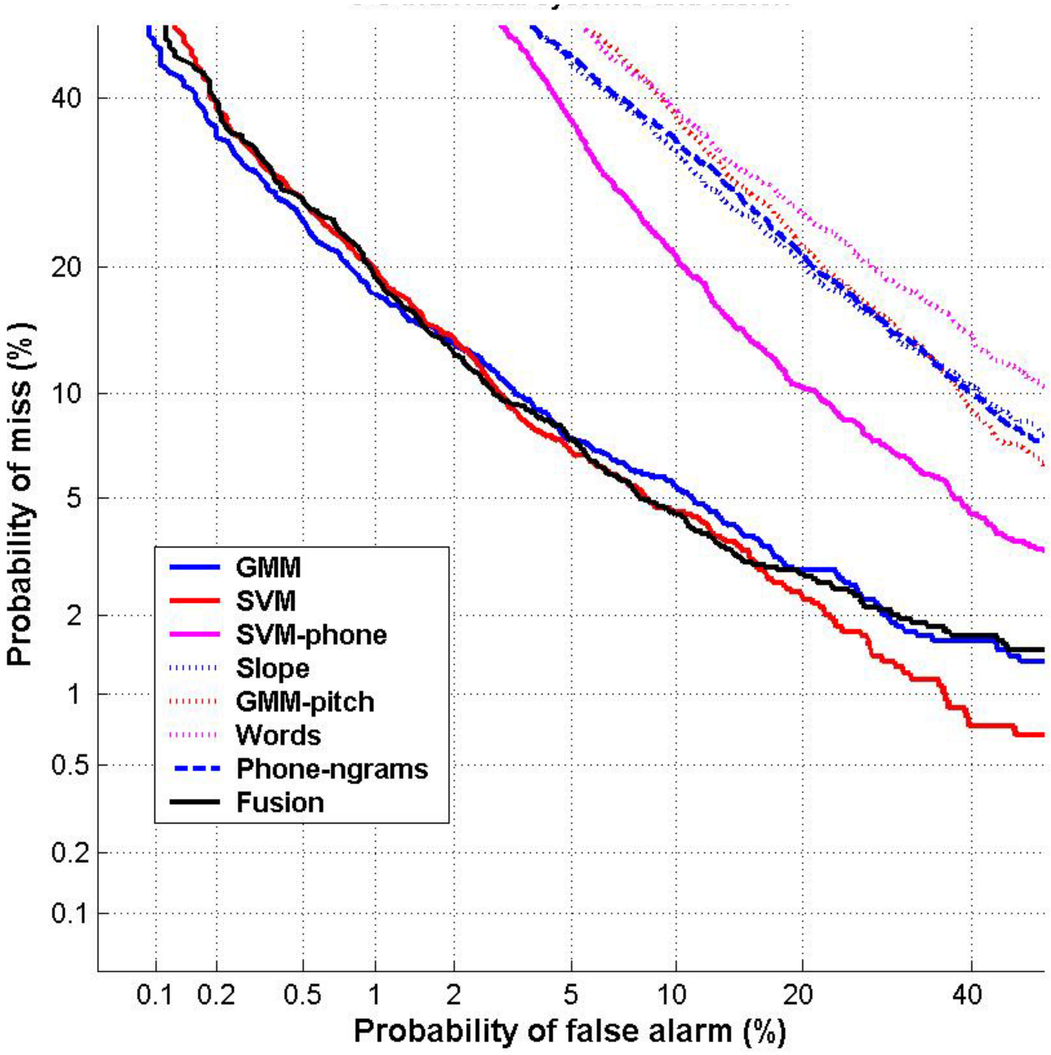
T-norm

H-norm

Speaker Recognition Performance

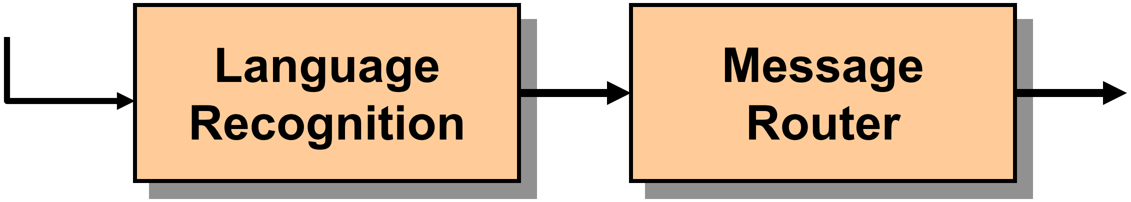
NIST 2004 Speaker Recognition Evaluation

- Miss and false alarm rates for a large corpora
- 8 conversation enrollment
- 1 conversation test
- Results show the use of high-level features, different classifier types, and fusion



Language Recognition Applications:
Front-end Routing for Human Operators

German-Speaking Caller



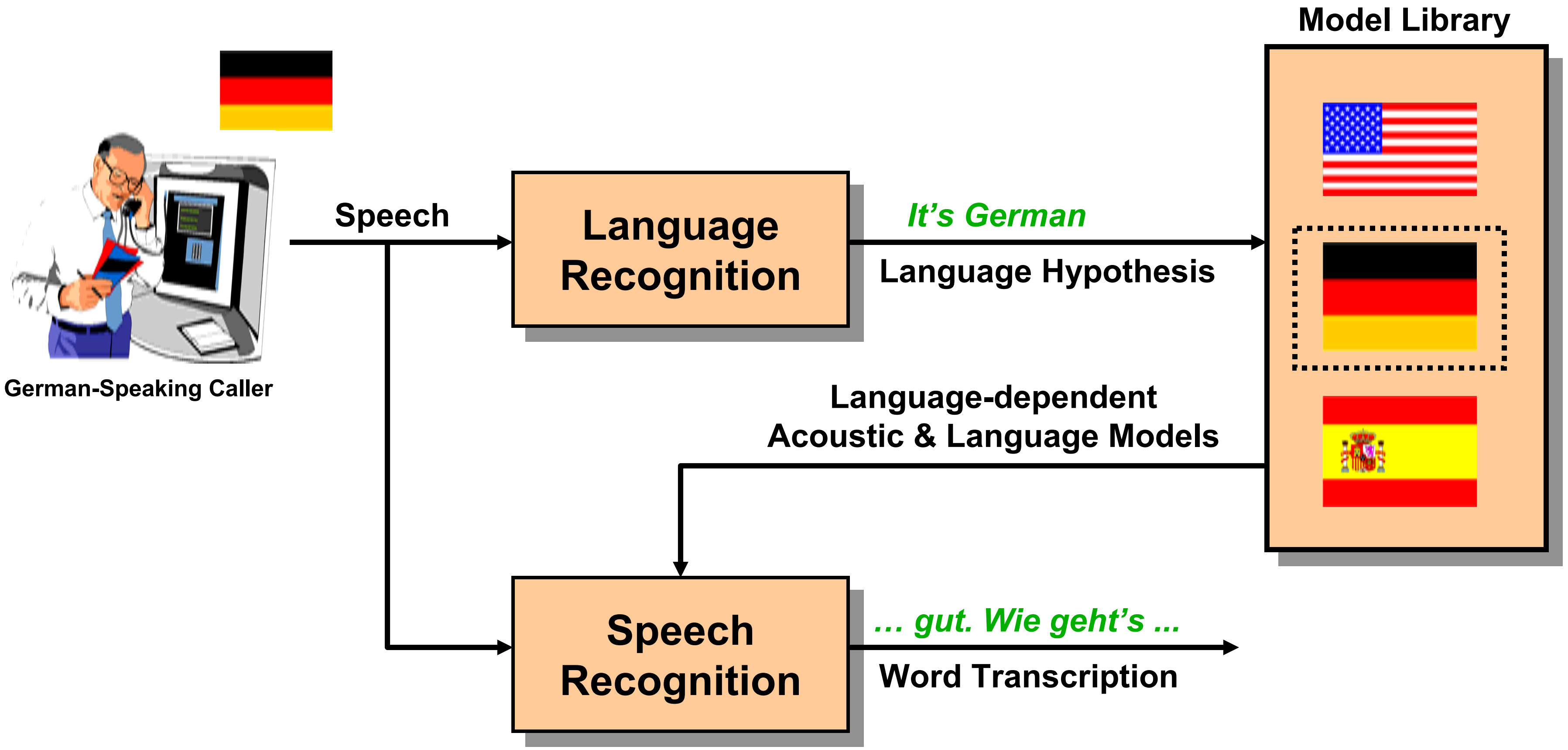
English-Speaking Operator

German-Speaking Operator

Spanish-Speaking Operator

- Language recognition system routes call to operator fluent in the speaker’s language

Language Recognition Applications:
Front-end for Automatic Speech Recognition



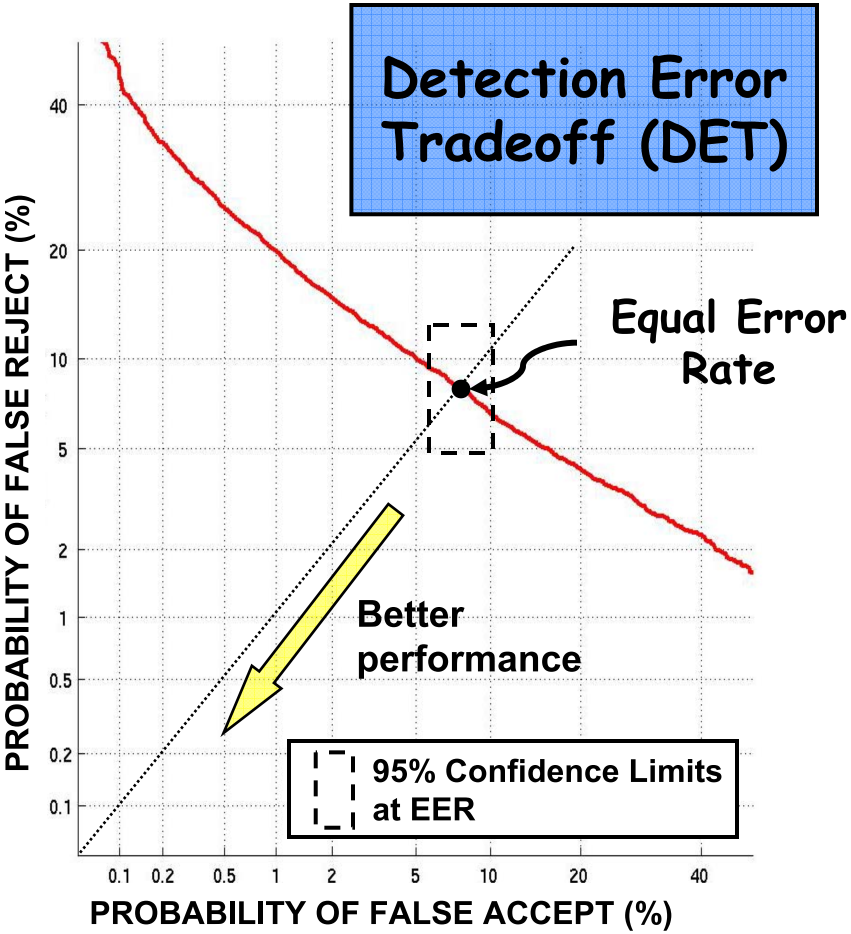
- Language recognition system selects models to be loaded into speech recognition system

Language Recognition Evaluation Metric

Detection Error Tradeoff

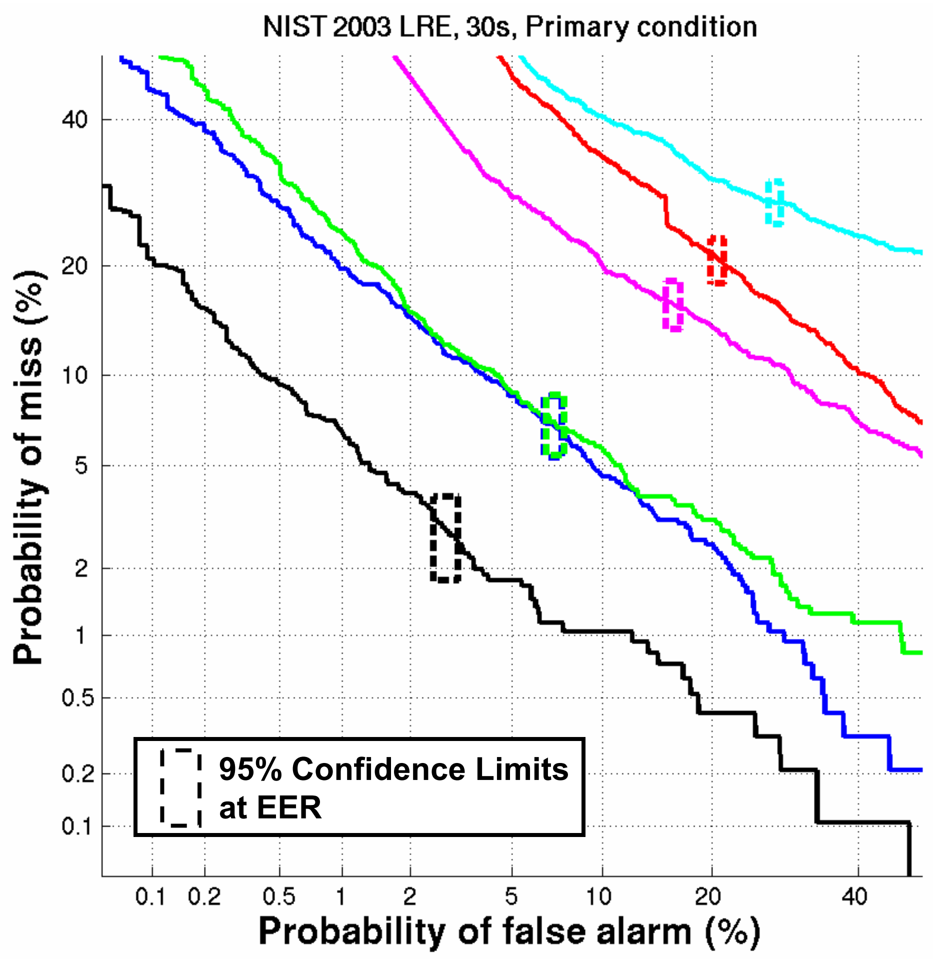
- For all language hypotheses
  - Sort scores
  - Label scores based on truth
  - Compute false accept and false reject error rates at every score threshold

Score	Truth
0.252	Target
0.208	Target
0.203	Non-target
⋮	⋮
-0.221	Target
-0.226	Non-target



NIST 2003 LRE Results

- NIST 2003 Language Recognition Evaluation (LRE)
- Six sites submitted results to NIST 2003 LRE
- Testing duration: 30s
- Languages:
  - Arabic, English, Farsi, French, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese



Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M. and Reynolds, D.A., “Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition,” in Proc. Eurospeech, pp. 1345-1348, 1-4 September 2003.

Text-to-Speech (TTS)

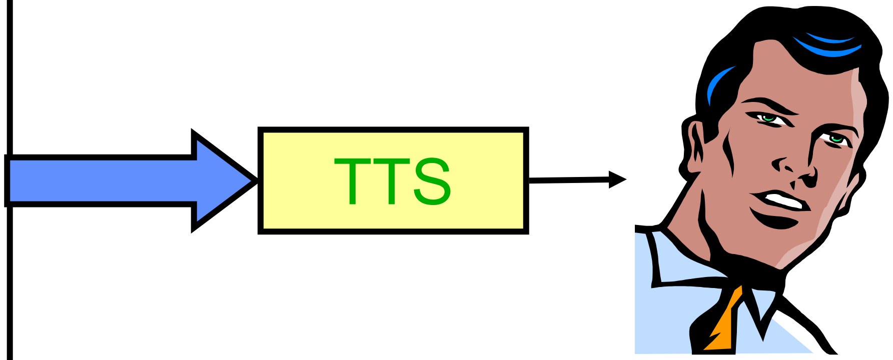
Cognitive Radio

Enable eyes-free use of systems

Effectively use modalities according to the environment

Choose speaking style and voice according to the situation

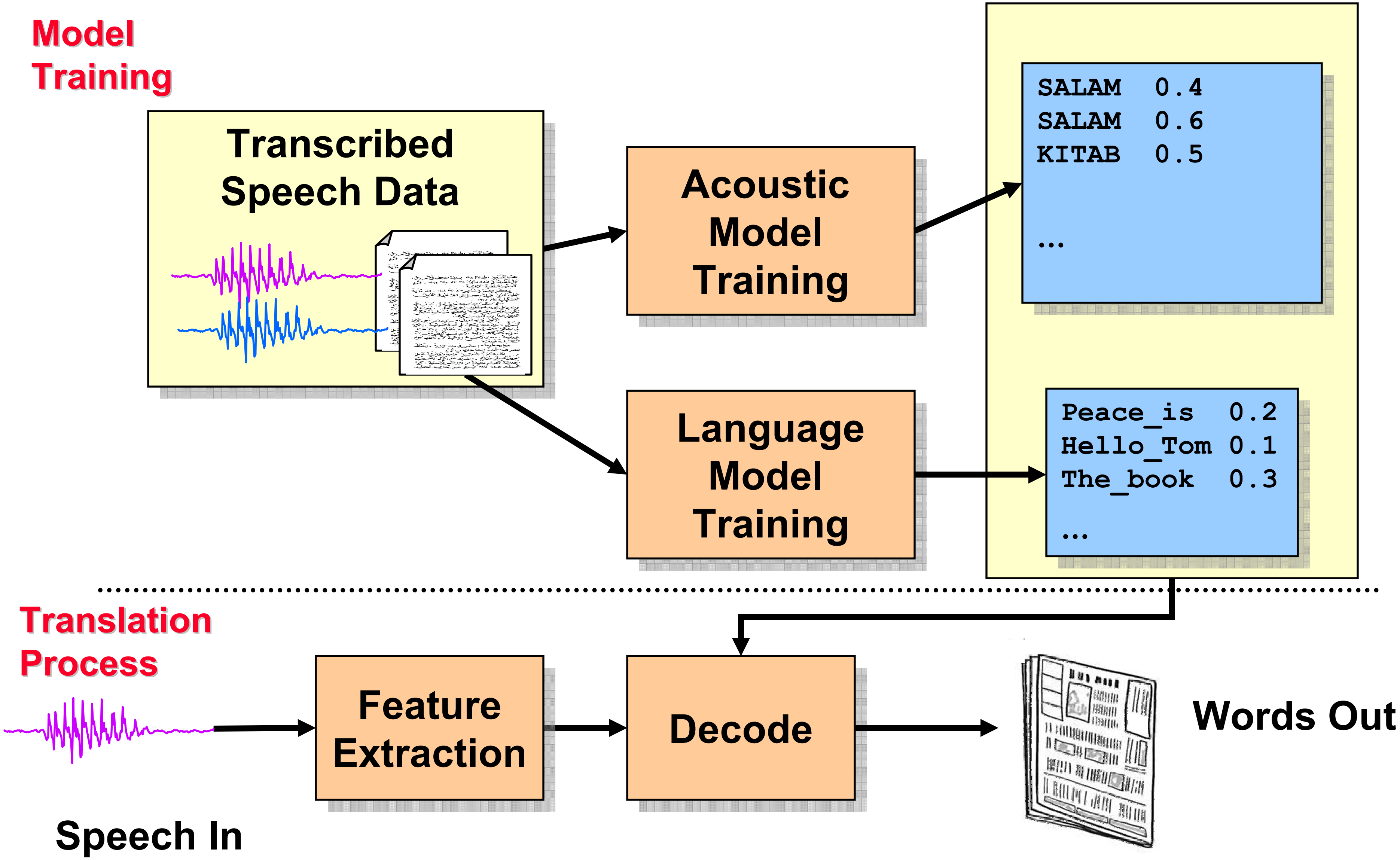
Integration with speech-to-text (STT) and machine translation (MT)



ATT_NaturalVoices.wav

Elan_SaysoUS1.wav

Speech-to-Text (STT) Architecture



Model Training

Transcribed Speech Data

Acoustic Model Training

```
SALAM   0.4
SALAM   0.6
KITAB   0.5

…
```

Language Model Training

```
Peace_is   0.2
Hello_Tom  0.1
The_book   0.3

…
```

Translation Process

Feature Extraction

Decode

Words Out

Speech In

Applications of STT to Cognitive Radio

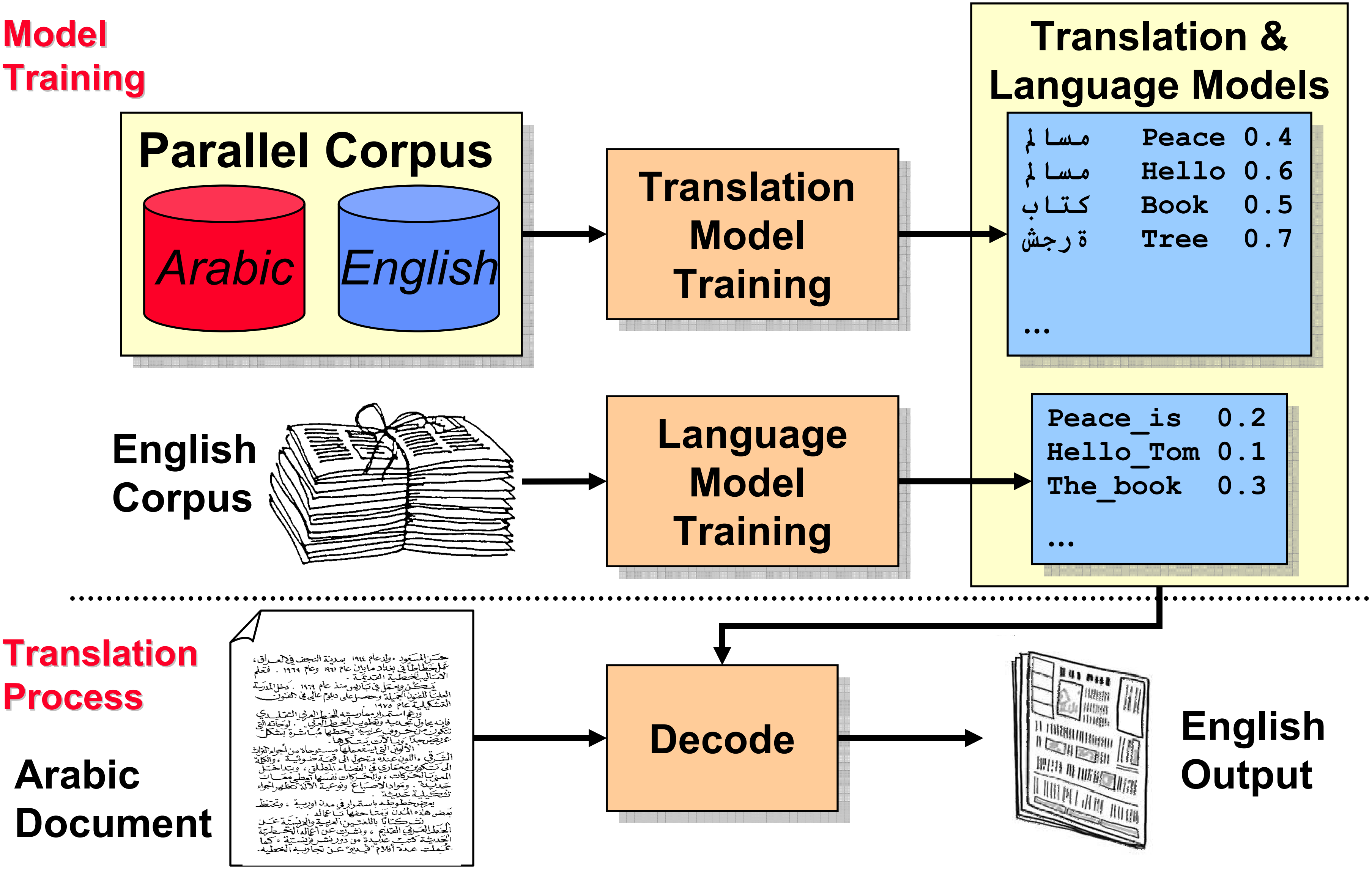
- Gisting: rather than having a user listen to the complete conversation, a summarized version of the output could be produced

- Routing: STT can be used to route certain conversations to appropriate users

- Data Mining: radio communication can processed by STT and stored, then text-retrieval techniques (such as those used to search documents on the internet) can be a quick and efficient way of searching content

- Command-and-Control (C2): a speech interface can free up tactile and visual modalities so that the user can more effectively multitask; the speech interface can be used to control various aspects of the cognitive radio (e.g., radio modes, sensor interfaces, sensor analysis, etc.)

Machine Translation
Statistical MT Architecture



Model Training

Parallel Corpus

Translation Model Training

Translation & Language Models

مسا لم	Peace	0.4
مسا لم	Hello	0.6
كتا ب	Book	0.5
ة رجش	Tree	0.7
…		

English Corpus

Language Model Training

Peace_is	0.2
Hello_Tom	0.1
The_book	0.3
…	

Translation Process

Arabic Document

Decode

English Output

Using Government Standards of Foreign Language Proficiency for MT Evaluation

Defense Language Proficiency Test (DLPT)
- **“High Stakes” test for DOD linguists**

We are proposing an <u>MT-DLPT</u>
- **Replace Arabic passages with English MT**
- **Enable monolingual to analyze texts**

Sponsors / Collaborators :
- **Defense Language Institute**
- **DARPA TIDES Program**

Sample Arabic Level 1 Test Item

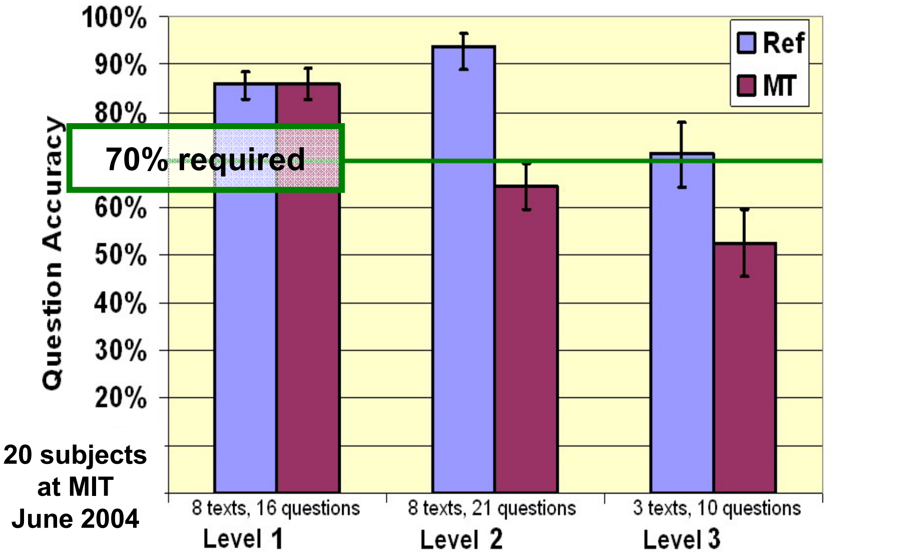
<u>From a society section in a newspaper</u>

في حفل بهيج ضم الأهل والأصدقاء تم حفل زفاف الشاب
إبراهيم رشيد أحمد
علـى
الآنسة ايفلين نداء ظاظا
أسرة المجلة تتقدم للعروسين بأسمى آيات التهنئة وتتمنى لهما السعادة والهناء. مبروك
للعروسين

Questions:
1. What is the purpose of this article?
2. What message does the magazine’s staff add?

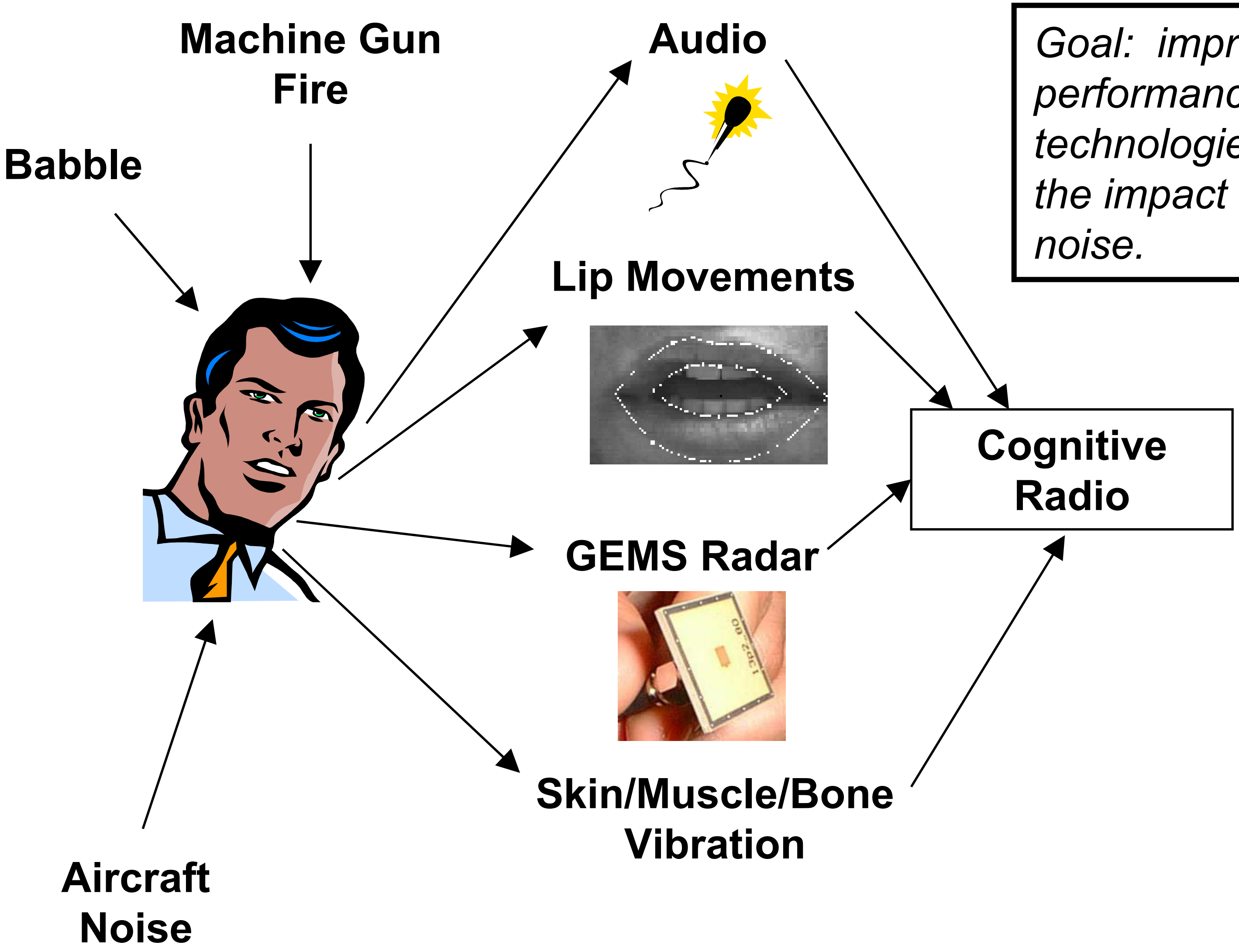
<u>Proficiency</u> measures the ability to perform <u>tasks</u>, such as:
- **Level 1: Extract Named Entities**
- **Level 2: Translate Newswire Texts**
- **Level 3: Analyze Argumentation (Goal is Level 3)**

“Smoke Test” suggests current MT Passes Level 1



See: Ray Clifford, Neil Granoien, Douglas Jones, Wade Shen, Clifford Weinstein. 2004.The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean. LREC 2004, Lisbon, Portugal.

Background Noise Suppression



Machine Gun Fire

Babble

Aircraft Noise

Audio

Lip Movements

GEMS Radar

Skin/Muscle/Bone Vibration

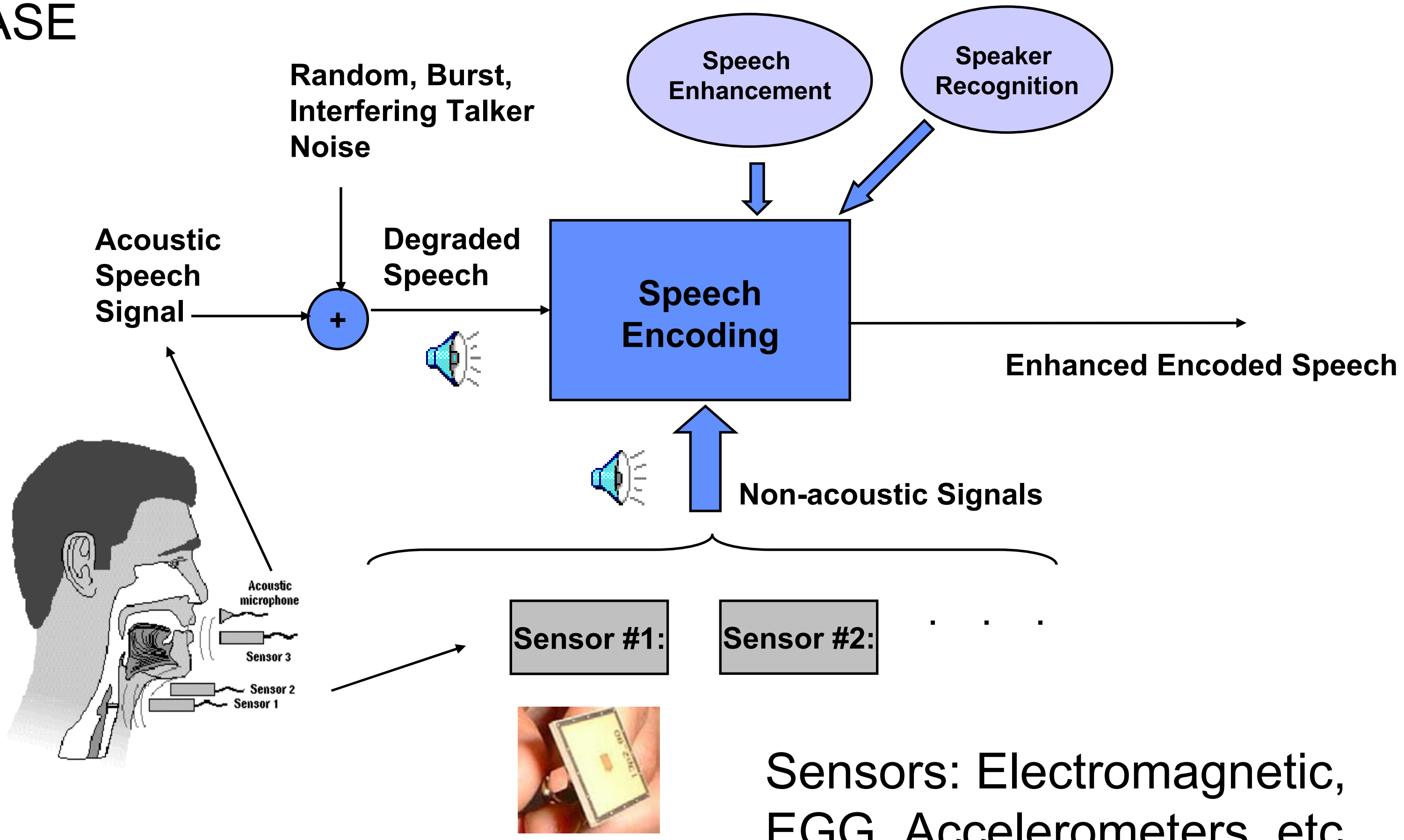
Cognitive Radio

Goal: improve the performance of speech technologies by reducing the impact of ambient noise.

Multisensor Noise Suppression

Objective: Use non-acoustic sensors to improve performance of speech encoding algorithms with speech that is degraded by severe additive noise backgrounds

DARPA ASE Program



Random, Burst, Interfering Talker Noise

Acoustic Speech Signal

Degraded Speech

Speech Enhancement

Speaker Recognition

Speech Encoding

Enhanced Encoded Speech

Non-acoustic Signals

Sensor #1:

Sensor #2:

. . .

Sensors: Electromagnetic, EGG, Accelerometers, etc.

Quatieri, T. F., Messing, D. P., Brady, K., Campbell, W. M., Campbell, J. P., Brandstein, M. S., Weinstein, C. J., Tardelli, J. D., and Gatewood, P. D., "Exploiting Nonacoustic Sensors for Speech Enhancement," in Proc. Workshop on Multimodal User Authentication, pp. 66-73, December 2003.

Other Speech Technologies With Applications to Cognitive Radio

- Adaptive Speech Coding
  - Required to fully exploit varying, limited channel capacity while achieving the goals of speech coding
  - Enhances radio performance by balancing between quality, intelligibility, LPI, LPD, etc.
- Speaker Characterization
  - Allows the “state” of a user to be determined by using voice processing techniques
  - Determines stress level, provides “reinforcement” feedback to cognitive radio, and improves user experience
- Noise Characterization
  - Allows the noise environment to be understood and interpreted
  - Provides situational awareness to radio operators

Conclusions and Implications for Cognitive Radio

- Speech technology is a critical part of cognitive radio
  - Speech is the primary input modality for radios
  - Provides natural user interaction
  - Provides situational awareness (e.g., intelligent analysis of communications)
- Many exciting speech technologies are available
  - Speaker recognition
  - Language recognition
  - Noise suppression
  - Etc.
- These technologies continue to improve in performance and are available now for prototyping in Cognitive Radios