Speech Recognition and Coversational Interfaces

Larry Rudolph
(content adapted from Jim Glass)
April 2006

The Space of Recognition

	Domain	
Speaker	Dependent	Independent
Dependent	not interesting	Transcription (training)
Independent	We are here	Ultimate Goal (requires knowledge)

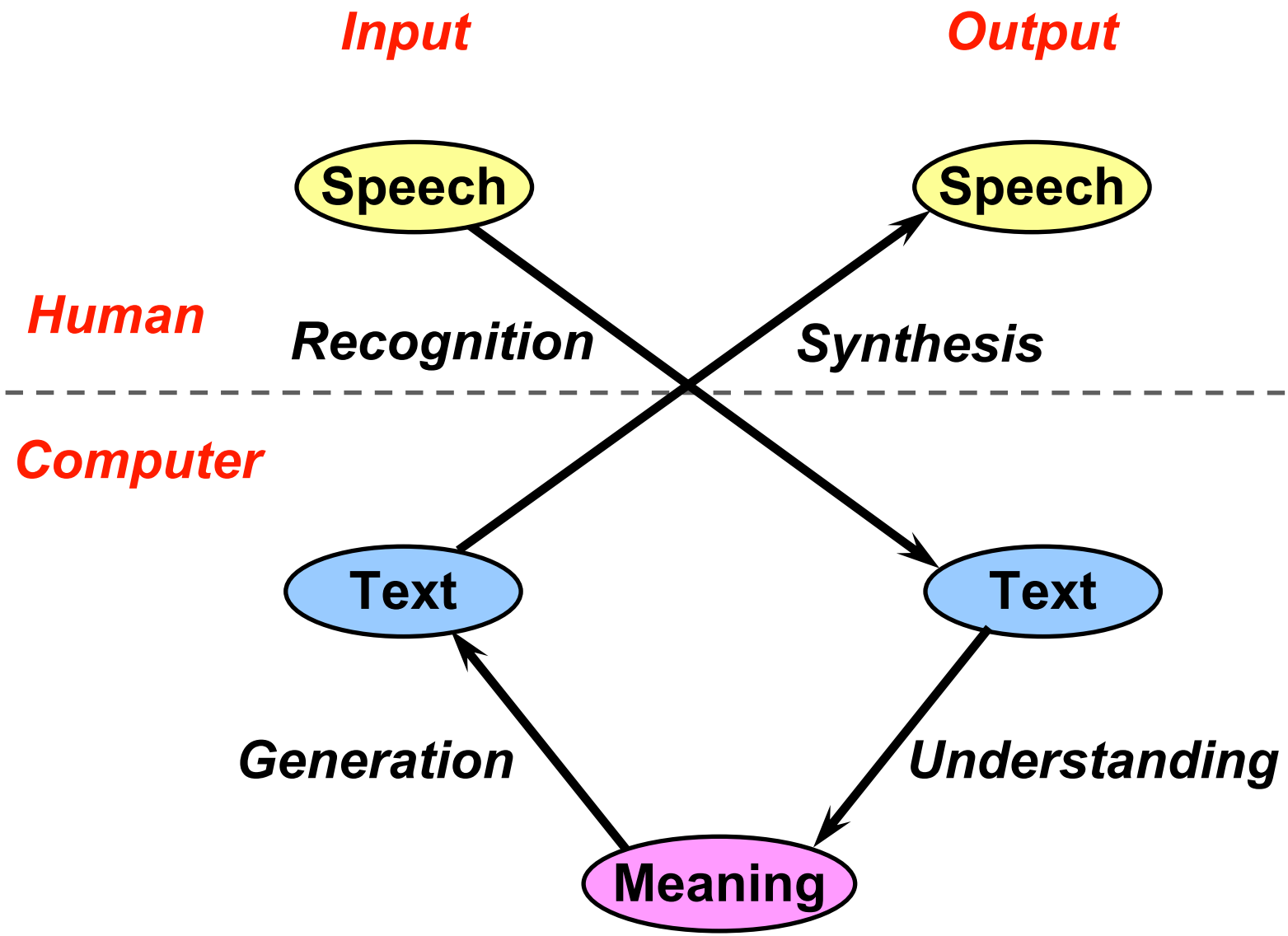
The Space of Recognition

- Speaker dependent
  - First train the system to recognize your speaking
  - Better recognition rates -- can learn idiosynchroses
- Domain dependent:
  - Only recognize what is in the domain
    - Better recognition rates
  - Domain can be large. How is it specified?

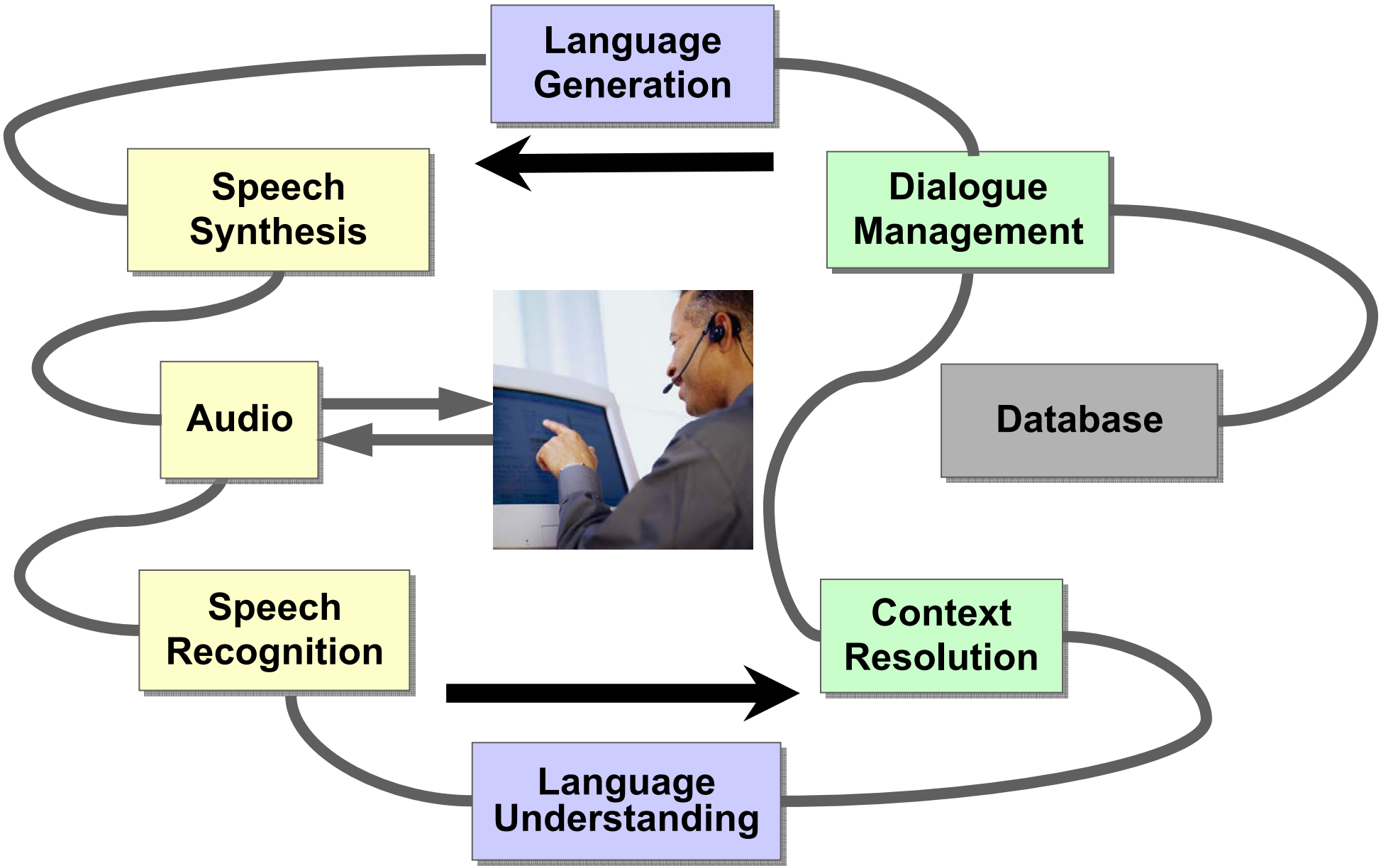
Why Speech?

- No special training -- naive users(?)
- Leaves hands and eyes free -- but must know when to start recognition
- High data rate -- assuming low errors
- Inexpensive I/O -- microphone, speaker, button
  - speaker needed for feedback
- Some things are easier to specify with speech

Communication via Spoken Language



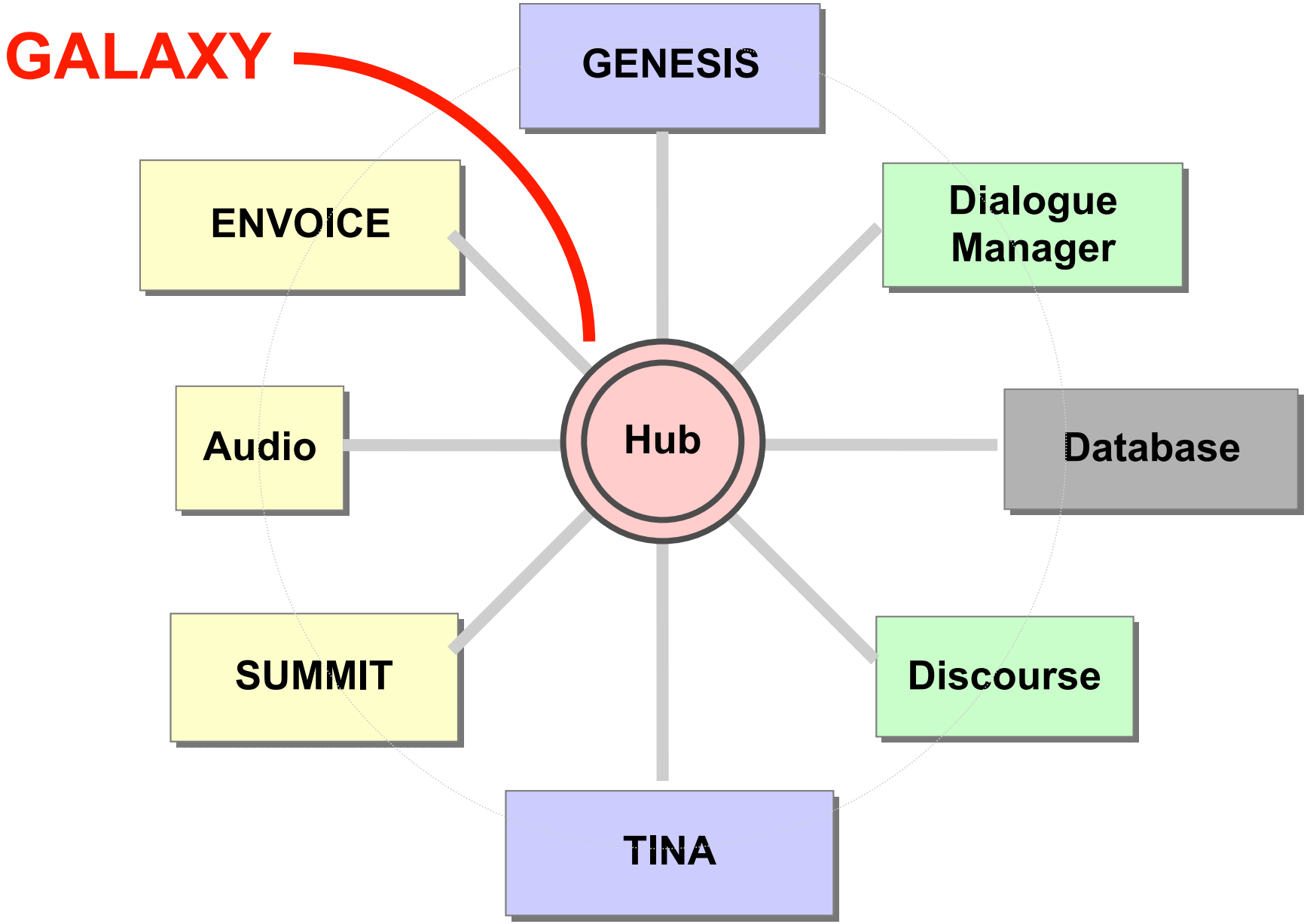
Components of Conversational Systems



Galaxy -- MIT SLS group

- SLS: Spoken Language Systems
- We will be making use of some of there technology
  - There are similar components developed by other groups (and some are public domain).
- The Galaxy System is organized around this cycle for conversational interfaces

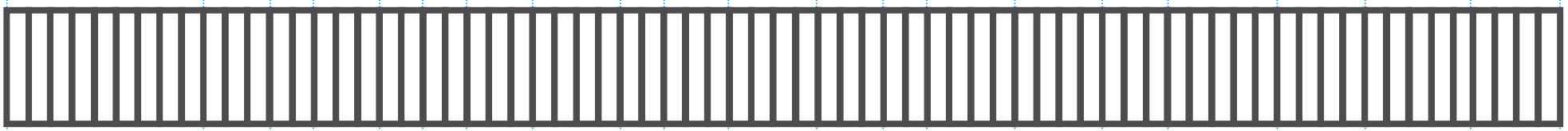
Components of MIT Conversational Systems



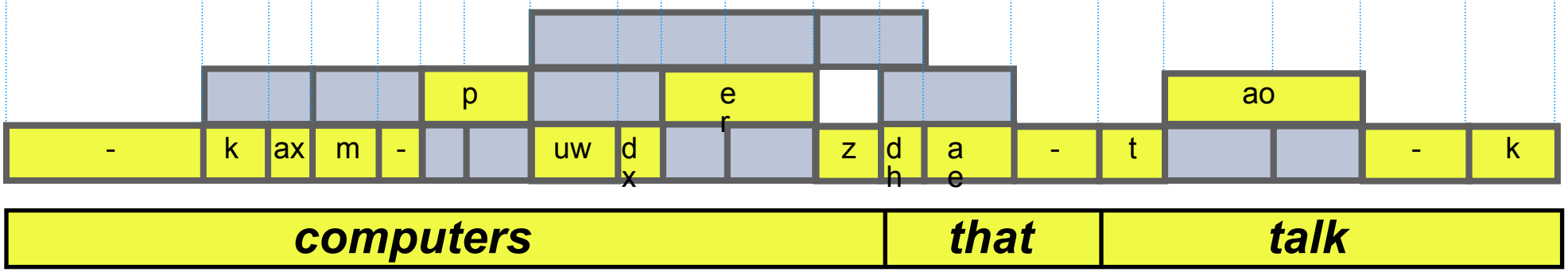
Segment-Based Speech Recognition

Waveform

Frame-based measurements (every 5ms)

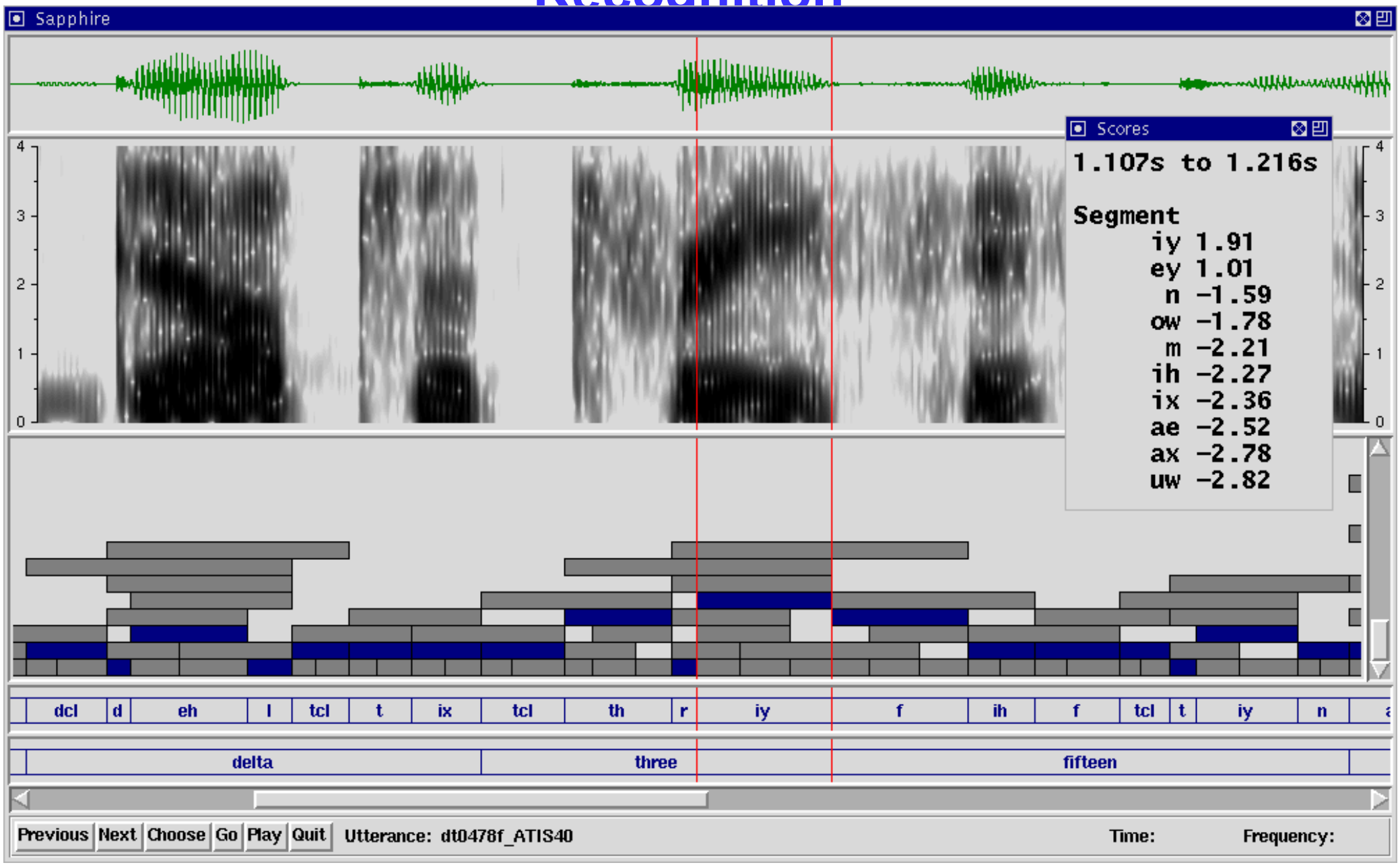


Segment network created by interconnecting spectral landmarks

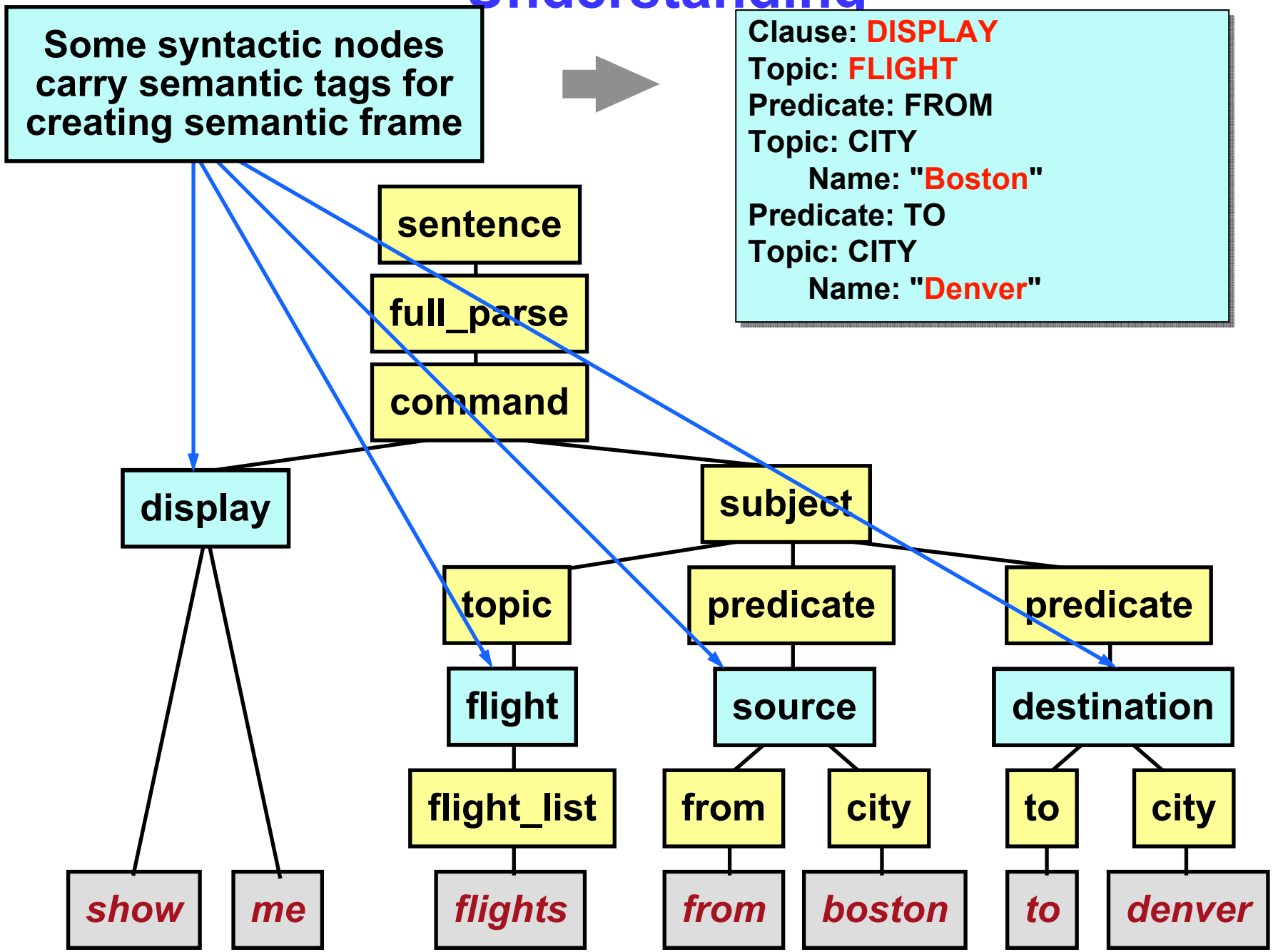


Probabilistic search finds most likely phone & word strings

Segment-Based Speech Recognition

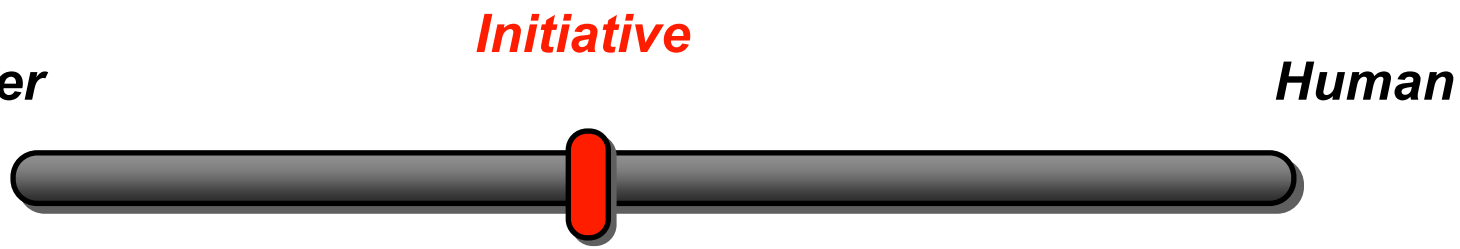


Natural Language Understanding



Dialogue Modeling Strategies

- **Effective conversational interface must incorporate extensive and complex dialogue modeling**
- **Conversational systems differ in the degree with which human or computer takes the initiative**



- **Computer maintains tight control**
- **Human is highly restricted**

C: *Please say the departure city.*

- **Human takes complete control**
- **Computer is totally passive**

H: *I want to visit my grandmother.*

- **The Galaxy System use a *mixed initiative* approach, where both the human & the computer play an active role**

Different Roles of Dialogue Management

- **Pre-Retrieval: Ambiguous Input => Unique Query to DB**

U: I need a flight from Boston to San Francisco
C: Did you say Boston or Austin?
U: Boston, Massachusetts
C: I need a date before I can access Travelocity
U: Tomorrow
C: Hold on while I retrieve the flights for you

Clarification (recognition errors)

Clarification (insufficient info)

- **Post-Retrieval: Multiple DB Retrievals => Unique Response**

C: I have found 10 flights meeting your specification. When would you like to leave?
U: In the morning.
C: Do you have a preferred airline?
U: United
C: I found two non-stop United flights leaving in the morning…

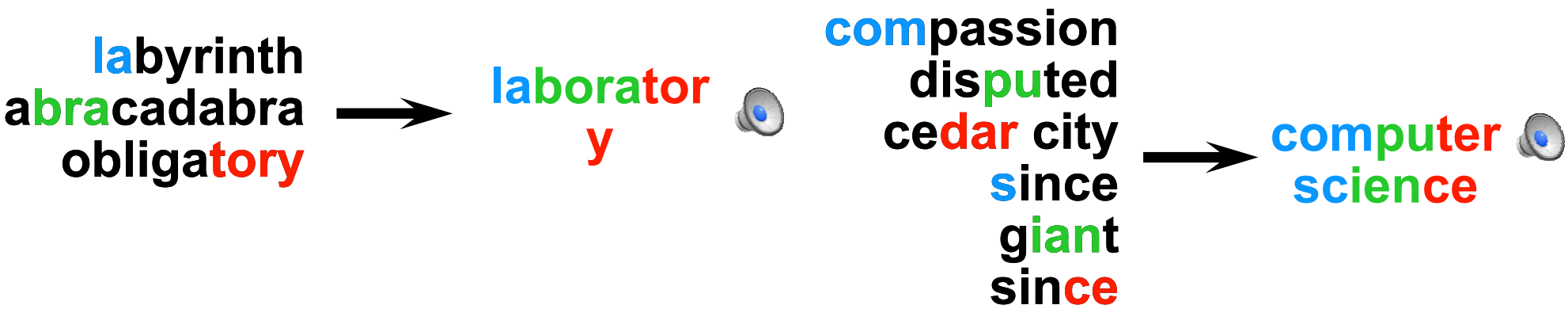
Help the user narrow down the choices

Concatenative Speech Synthesis

- Output waveform generated by concatenating segments of pre-recorded speech corpus.
- Concatenation at phrase, word or sub-word level.

Synthesis Examples

The third ad is a 1996 black Acura Integra with 45380 miles.
The price is 8970 dollars. Please call (404) 399-7682.

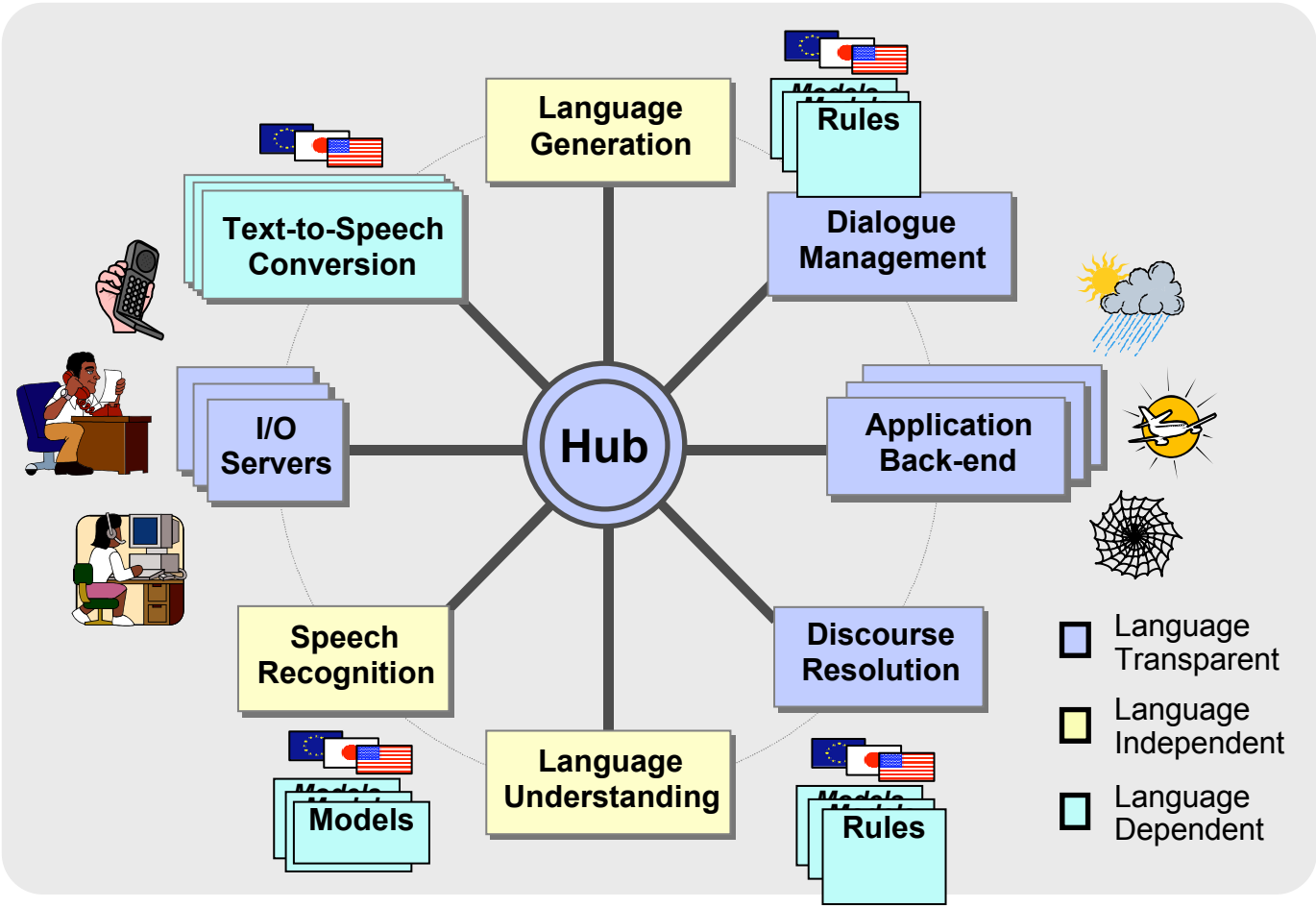


Continental flight 4695 from Greensboro is expected in Halifax at 10:08 pm local time.

Multilingual Conversational Interfaces

- Adopts an *interlingua* approach for multilingual human-machine interactions

- **Applications:**
  - MuXing: Mandarin system for weather information
  - Mokusei: Japanese system for weather information
  - Spanish systems are also under development
  - New speech-to-speech translation work (Phrasebook)

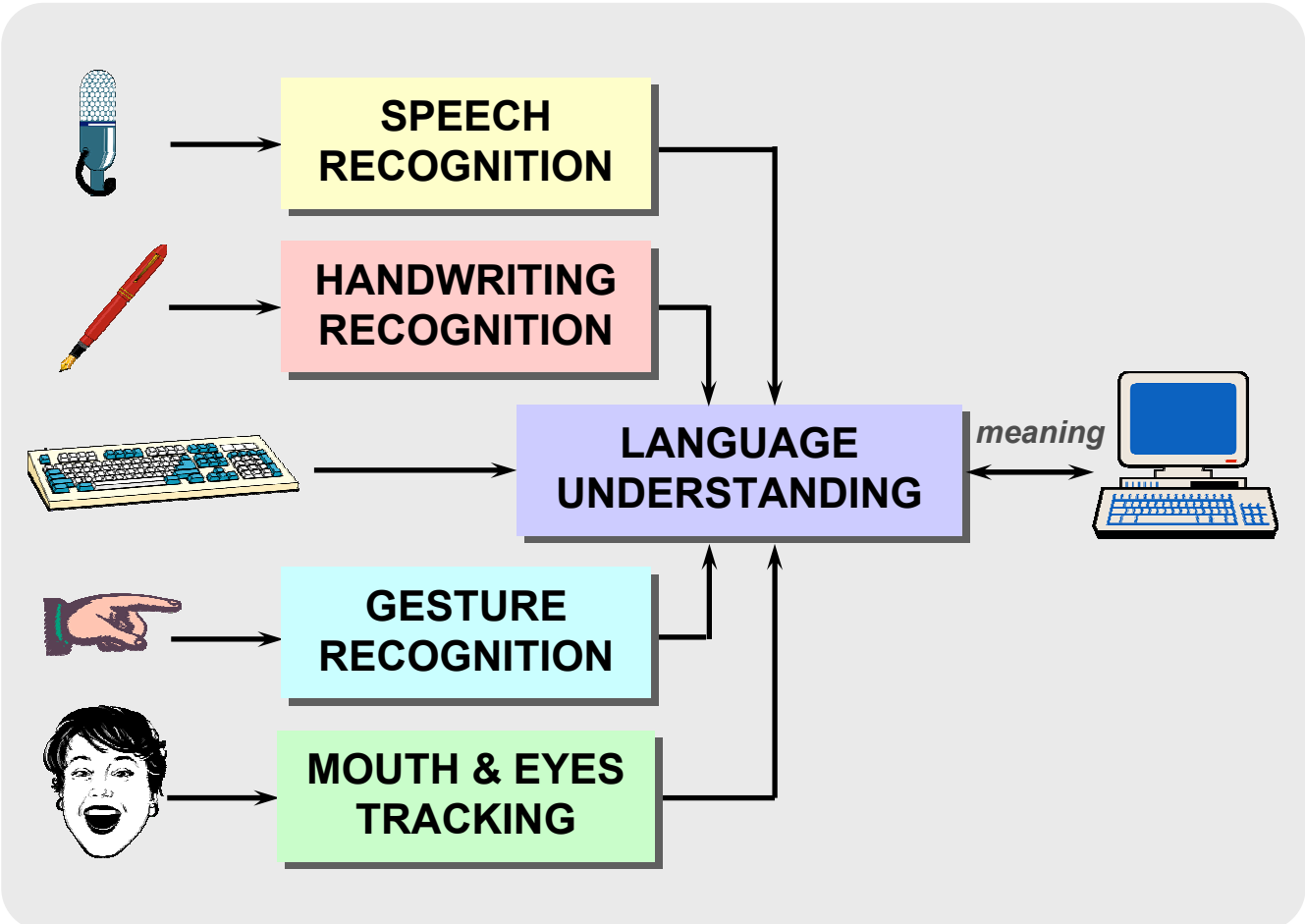


Bilingual Jupiter Demonstration

Multi-modal Conversational Interfaces

- Typing, pointing, clicking can augment/complement speech
- A picture (or a map) is worth a thousand words

- **Applications:**
  - WebGalaxy
  - Allows typing and clicking
  - Includes map-based navigation
  - With display
  - Embedded in a web browser
  - Current exhibit at MIT Museum



WebGalaxy Demonstration

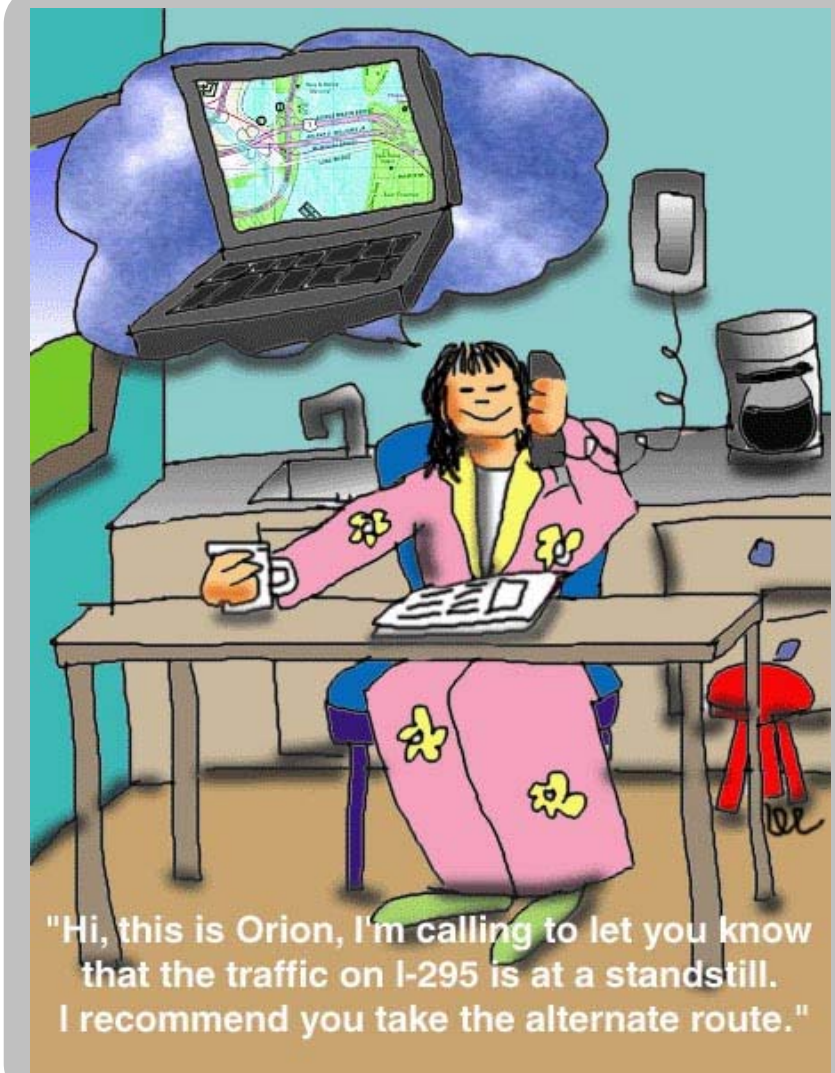
Delegating Tasks to Computers

- **Many information related activities can be done off line**
- **Off-line delegation frees the user to attend to other matters**

- **Application: Orion system**
  - Task Specification: User interacts with Orion to specify a task

    "Call me every morning at 6 and tell me the weather in Boston."

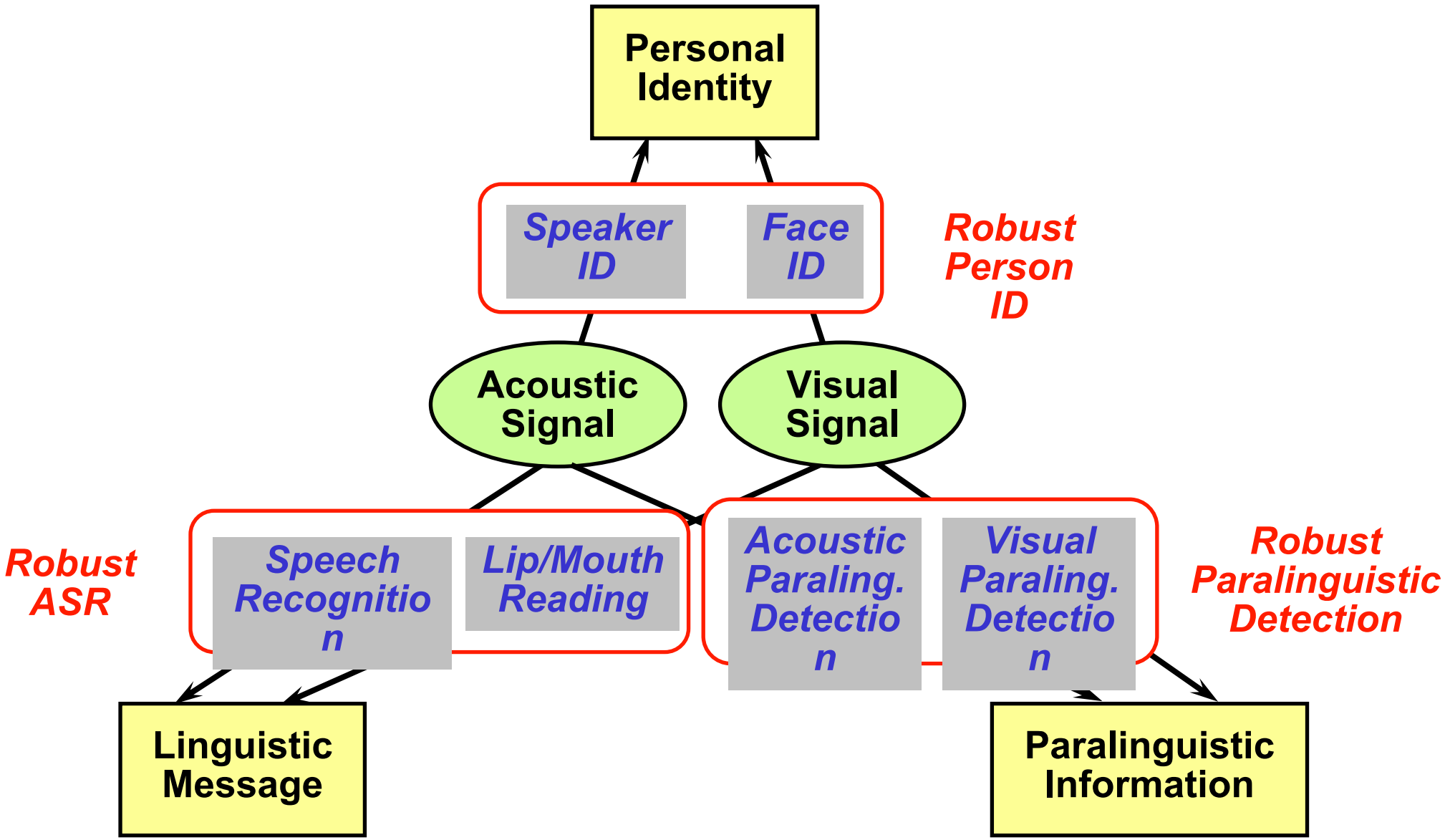
    "Send me e-mail any time between 4 and 6 p.m. if the traffic on Route 93 is at a standstill."
  - Task Execution: Orion leverages existing infrastructure to support interaction with humans
  - Event Notification: Orion calls back to deliver information



Audio Visual Integration

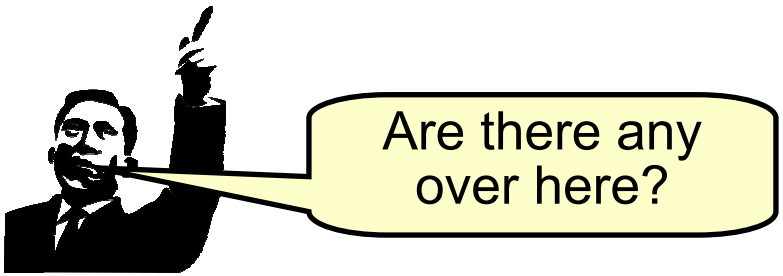
- **Audio and visual signals both contain information about:**
  - Identity of the person: *Who is talking?*
  - Linguistic message: *What's (s)he saying?*
  - Emotion, mood, stress, etc.: *How does (s)he feel?*
- **The two channels of information**
  - Are often inter-related
  - Are often complementary
  - Must be consistent
- **Integration of these cues can lead to enhanced capabilities for future human computer interfaces**

Audio Visual Symbiosis



Multi-modal Interfaces: Beyond Clicking

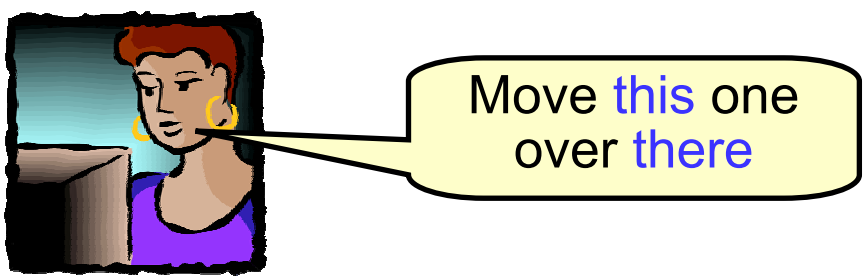
- **Inputs need to be understood in the proper context**



What does he mean by “any,” and what is he pointing at?

Does this mean “yes,” “one,” or something else?

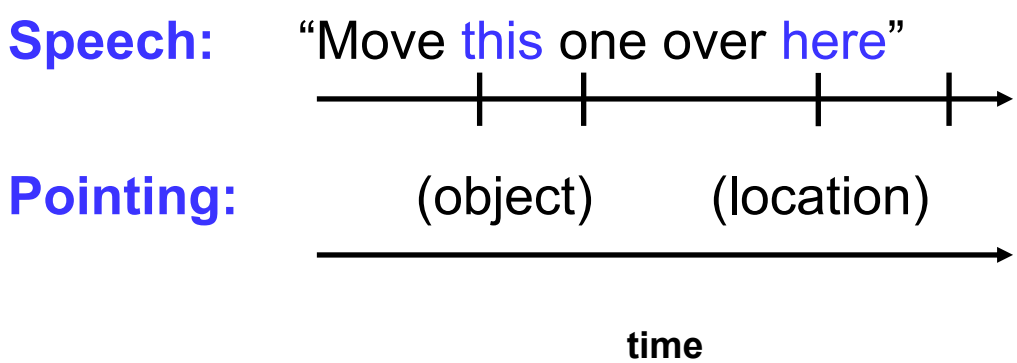
- **Timing information is a useful way to relate inputs**



Where is she looking or pointing at while saying “this” and “there”?

Multi-modal Fusion: Initial Progress

- **All multi-modal inputs are synchronized**
  - Speech recognizer generates absolute times for words
  - Mouse and gesture movements generate {x,y,t} triples
  - Network Time Protocol (NTP) is used for msec time resolution
- **Speech understanding constrains gesture interpretation**
  - Initial work identifies an object or a location from gesture inputs
  - Speech constrains what, when, and how items are resolved
  - Object resolution also depends on information from application



Multi-modal Demonstration

- **Manipulating planets in a solar-system application**
- **Created w. SpeechBuilder utility with small changes**
- **Gestures from vision (Darrell & Demirdjien)**

Multi-modal Demonstration

- **Manipulating planets in a solar-system application**
- **Created w. SpeechBuilder utility with small changes**
- **Gestures from vision (Darrell & Demirdjien)**