Voice Control and Artificial Intelligence

H. Shawn Xu
May 03, 2011

CIS II Paper Presentation

ERC | CISST

LABORATORY FOR Computational Sensing + Robotics
THE JOHNS HOPKINS UNIVERSITY

My Project

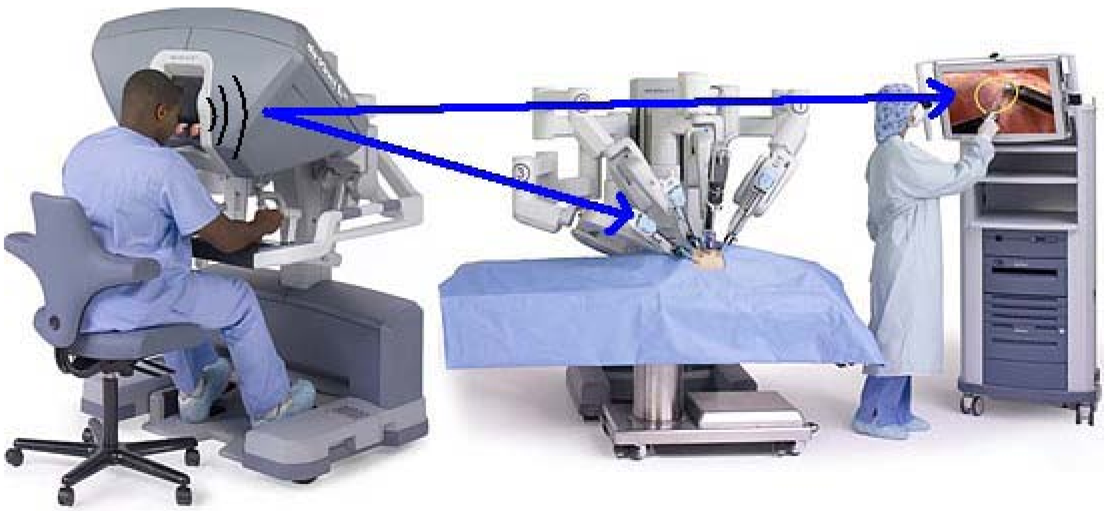
Allow surgeon to control certain parts of the *da Vinci*® system via voice

- The *da Vinci*® is a robotic teleoperated surgical system
- Controlled by surgeon at HD workstation with hands and feet

PROBLEM:

Complex gestures, stop-start procedures

My Project (continued)



Thinking Beyond My Project…

- We demoed how voice might be used to interact with the *da Vinci*® system to Intuitive Surgical
- We are adding some additional functionality to our demo
- All very nice, but what else is possible?

p2nSpeech

A cognitive architecture approach to robot voice control and response

- Siddtharth Patel (Masters student at Pace University)
- Published in 2008
- http://support.csis.pace.edu/CSISWeb/docs/MSThesis/PatelSiddtharth.pdf

Summary

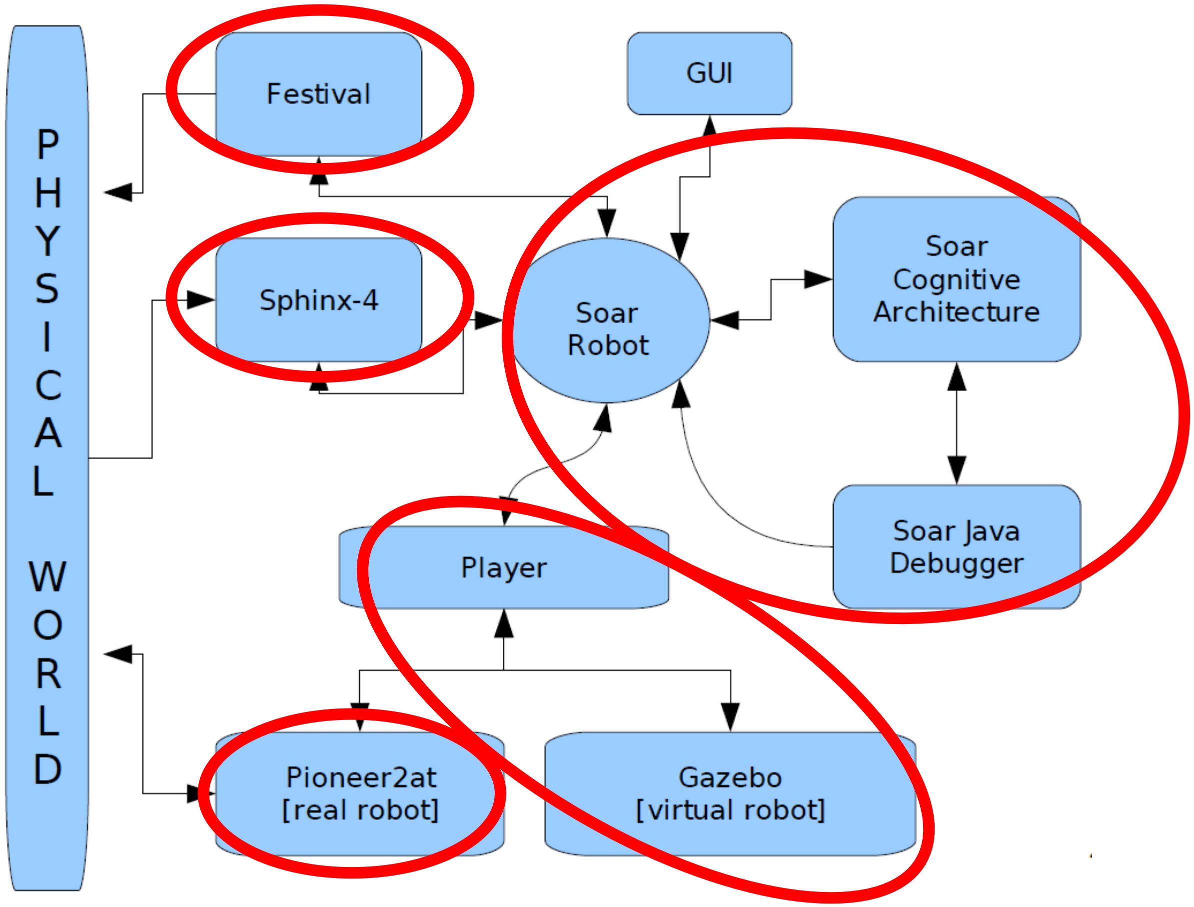
- p2nSpeech is a project that “aims to explore the ways to command a robot using human voice and also enable the robot to exhibit cognitive behavior.”

Design Overview

- Pioneer robot:

- Player–Gazebo: robot simulator
- Festival: speech synthesis
- **Sphinx 4: speech recognition**
- **SOAR: unified cognitive architecture**

Design Overview (Continued)



Pioneer Robot

- Mobile robot with solid rubber tires, a two-wheel differential, reversible drive system and a rear caster for balance

- Interacts with the real-world
  - multiple sonar sensors
  - stereo camera
  - position/speed encoders

Player–Gazebo

- Open–source robot simulator

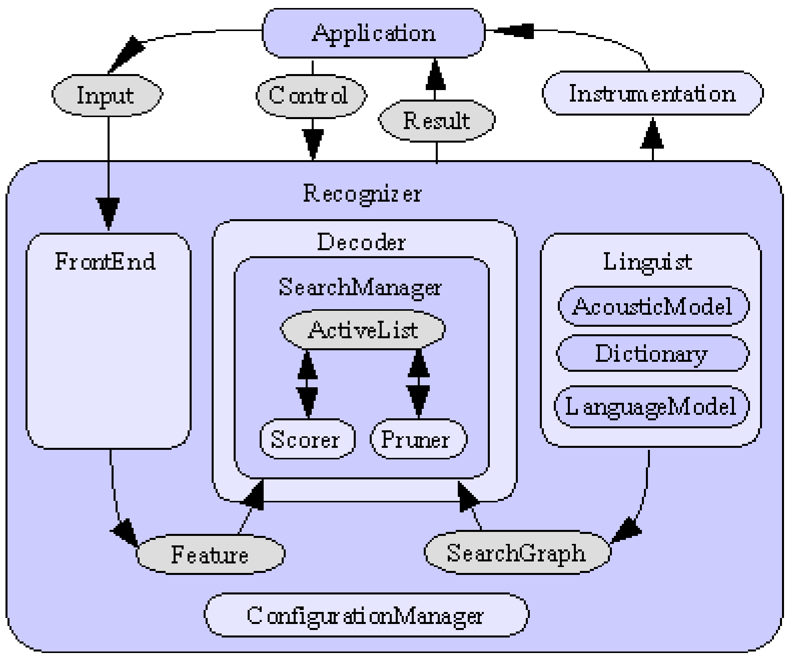
- Player gets real–world information from sensors of robot
- Gazebo builds 3D virtual environment with robot in it

Festival

- Full text-to-speech support for multiple languages (English is most advanced)

Sphinx 4

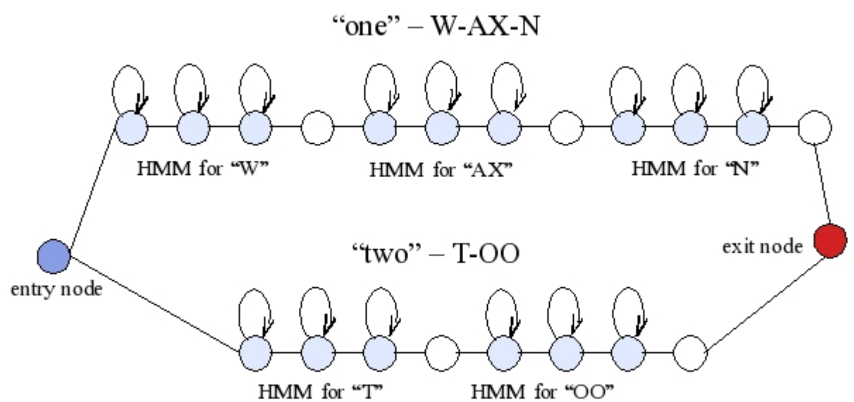
- Speech recognition package developed in Java
- WE USED THIS PACKAGE FOR OUR PROJECT
- Architecture overview:



ERC | CISST
LABORATORY FOR Computational Sensing + Robotics
THE JOHNS HOPKINS UNIVERSITY

How It Works

- A word, is many phonemes (unit of sound) strung together
- Short time periods of speech can be treated as stochastic process, specifically a Hidden Markov Model



How It Works (Continued)

- User defines a grammar, which is converted into a search graph
- Live audio is converted into a stream of sound features
- Search manager runs the the stream of features through the search graph, and determines what was said based on a running statistical score that is kept

SOAR

- Unified cognitive architecture that “[allows] knowledge to be encoded and used to produce action in pursuit of goals.”
- Rule-based

- Terminology
  - State: representation of current situation
  - Operator: transforms a state
  - Goal: desired outcome

Sample SOAR Code

```
sp {R1

   (State <s> ^forecast-info RAIN)

-->

   (<s> ^output BRING-AN-UMBRELLA)

}

sp {R2

    (State <s> -^forecast-info RAIN)

-->

   (<s> ^output DO-NOT-BRING-AN-UMBRELLA)

}
```

How It Works

- User defines "if–then" rules and assigns preferences to them (long–term knowledge)
- When the program is running
  - The system examines the conditions of every rule and determines a subset that matches the current state
  - Using a conflict resolution based on preference, the system determines the corresponding operator
  - Repeat until desired outcome is achieved and send output commands to external environment

More In-Depth

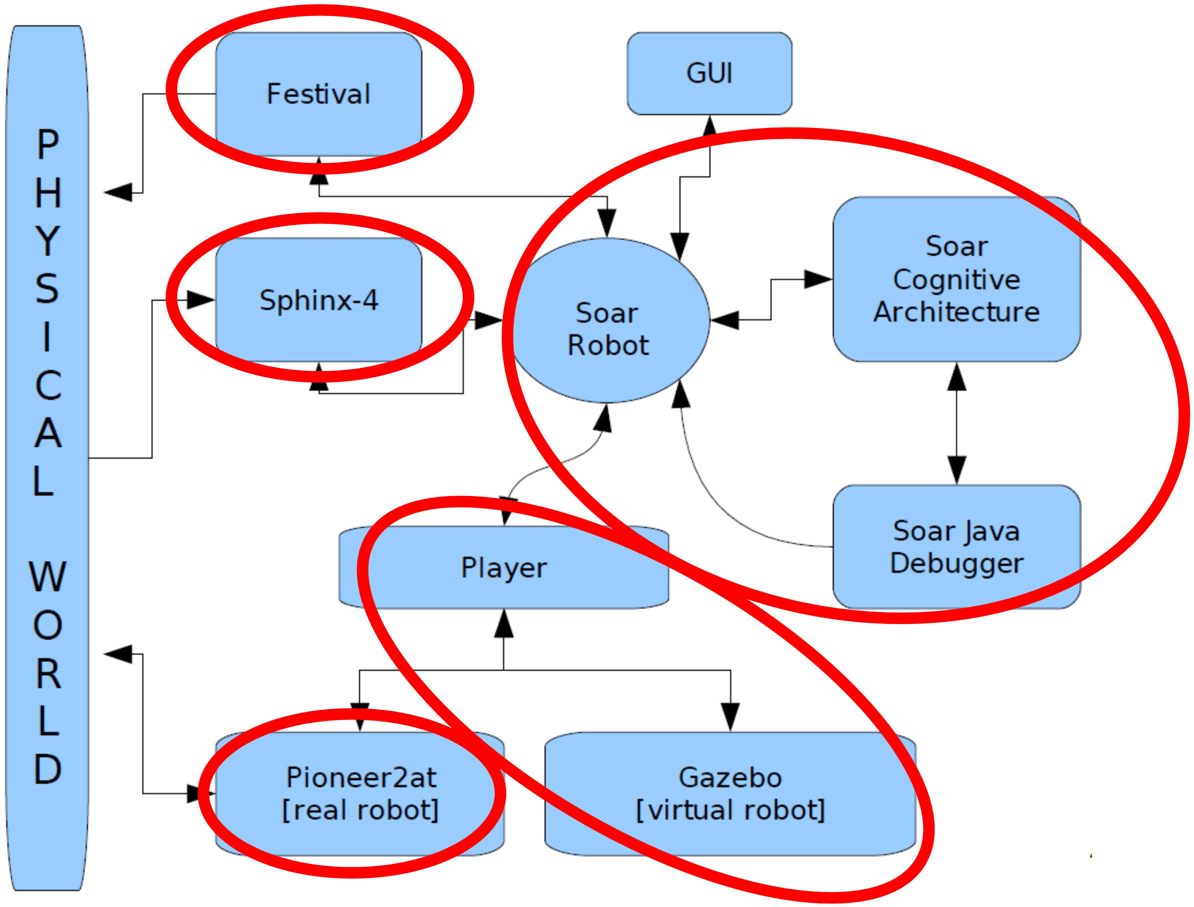
- Five-phased active cycle
  1. Input: new data comes into working memory (short-term memory)
  2. Proposal: interpret new data, propose operator for rules with a matching condition, determine would-be changes to state, and repeat
  3. Decision: compare and decide on next operator
  4. Application: apply new operator, and if there are changes to state, go back to phase 1
  5. Output: send output commands

Learning and Long-Term Memory

- At decision stage, preferences could be incomplete or insufficient, and thus system is at an impasse and doesn't know what to do

- User resolves impasse with input
- System creates substate and remembers the processing as new pseudo-rules (learning)
- This new information is stored in system's working memory (long-term memory)

Design Revisited



p2nSpeech in Action

- A basic application:
  - SOAR-Robot receives a command to move forward (from Sphinx 4)
  - Based on sonar sensors (from Player), SOAR determines if there is an obstacle directly in front
  - If so, tell robot to turn
  - If not, move robot forward, update information, repeat until there is an obstacle in front
- Possibilities for further exploration
  - Pre-command model information from Player-Gazebo
  - Learning

Back to My Project

- **How might we use these ideas to better integrate voice control with a surgical robot?**
- Cognitive decision-making allows for much more natural interaction between human and machine
    - Improve existing functionality (e.g.: measurement)
    - Opens door for more complex functionality to be added with voice (e.g.: control of patient-side)
- But we must always be aware of maintaining full and precise master control
    - Natural and intuitive vs. precise and accurate

Thank You