Advanced Multimodal Interfaces: Lessons Learned from Project Oxygen

Victor Zue (zue@csail.mit.edu)
MIT Computer Science and Artificial Intelligence Laboratory

CSAIL Formed in July 2003

- Merger of former Artificial Intelligence Laboratory and the Laboratory for Computer Science (= Project Mac July 1963)
- About 820 members
  - 93 principal investigators
    * 72 active teaching faculty
    * EECS, Math, Brain and Cognitive, Aero/Astro, Mechanical Eng, Planetary Sciences, Harvard-MIT HST
  - 450 graduate students
  - 90 research staff and research visitors
  - 80 postdoctoral fellows
  - 60 staff
  - 50 undergraduates
Our New Home: The Stata Center

How Do We Organize?

• Four Directorates (Research Directors)
  – Architecture, systems, and networks (Srini Devadas)
  – Theory (David Karger)
  – Language, learning, vision, and graphics (Leslie Kaelbling)
  – Physical, biological, and social systems (Randy Davis)

• Cross-cutting organizations within
  – T-Party
  – Project Oxygen
  – Center for Information Security and Privacy

• Funding
  – DARPA, NSF, NASA, CIA, NIH, ONR, AFOSR, ARDA
  – Singapore-MIT Alliance, Cambridge-MIT Institute, ITRI
  – Quanta Computer, Hewlett Packard, Microsoft, NTT, Nokia, Philips, Delta Electronics, Acer, Shell, Ford, Sun, IBM, Intel, Toyota, Honda, Cisco, ABB
• **Diverse**
  - From 7 departments
  - From 22 countries
  - 13 women

• **Recognized worldwide**
  - 16 NAS/NAE members,
  - 6 MacArthur Foundation Genius Awards,
  - 4 Turing Awards,
  - 2 Japan Prizes,
  - 1 Knight of the British Empire,
  - 1 Millennium Technology Award,
  - ….
What is Oxygen?

- A research project to radically change the ways people deal with information related activities
  - Launched in 2000
  - In collaboration with six world-class partners (Acer, Delta Electronics, HP, NTT, Nokia, Philips)
  - Develop technology on many fronts
  - Experiment with prototypes

The Oxygen Vision

- Computing and communication (C&C) will soon be “freely” available to everybody, everywhere
  - Need to accommodate a new, nomadic lifestyle
- Computers should serve people, rather than the other way around
  - Need to create a new mode of interaction

☞ To bring the abundance of computation and communication within easy reach of humans by
  - Blending computation into peoples’ lives through natural, perceptual interfaces, using sound, sight, gesture, etc.
  - Enabling them to easily do the tasks they want to do – collaborate, access knowledge, automate, and customize
Some Technical Challenges

- Devices
- Networks
- Security and Privacy
- Interfaces
- Collaboration
- Operating Systems
- ...

We Need Anthropomorphic Interfaces

- Computers should interact with humans in the same ways humans interact with humans
- Next-generation interfaces should be human centric:
  - Carry on a conversation with the user
  - Permit multi-modal interactions using speech, vision, gesture and sketching,
  - Create an integrated, immersive environment
Challenges for Multimodal Interfaces

- Input needs to be understood in the proper context
  - “What about that one”
- Timing information is a useful way to relate inputs
  
  **Speech:**  “Move this one over here”
  
  **Pointing:**  (object)  (location)

- Handling uncertainties
- Need to develop a unifying linguistic framework

**Map-Based Interactions**

**Conversational Interactions**
*(Seneff, Glass, Zue)*
Speech-Based Interfaces

- Communicate with users through a **conversational** paradigm
- **Understand** verbal input
  - Speech recognition
  - Language understanding (in context)
- **Verbalize** response
  - Language generation
  - Speech synthesis
- Engage in **dialogue** with a user during the interaction
Mono-lingual ⇔ Multi-lingual

- Use an Interlingua to capture meaning
- Keep technology components language-independent when possible
- Encode language specific information in external data structures

Visual Conversational Cues
(Darrell)
Visual Conversational Cues

- Locate and track people in a scene
- Determine who is speaking
- Determine to what/whom the speaker is speaking
- Perceive head-nod agreement gestures
- Track body position and pose
- Recognize arm gestures
- Provide pose-invariant features for audio-visual speech recognition
- Infer audio-visual activities
- ...

Head Pose Tracking

- Head pose provides important visual cues in dialogue:
  - Determine focus of attention
  - Acknowledge information and provide answers to Yes/No questions
- Real-world challenges
  - Variable background
  - Subtle movements
  - User independent
- Solutions
  - Stereo processing
  - Motion estimation
  - Online model acquisition
Body Pose Tracking

- Body pose can provide information about indirect object reference, gestures, etc.
- Current approach uses example-based learning
  - Match an image to a large example corpus based on shape or appearance
  - Incrementally update pose by aligning matched model to subsequent frames

Audio-Visual Integration
Importance of Audio-Visual Integration

- The audio and visual signals both contain information about:
  - Identity/location of the person: Who is talking? Where is he?
  - Linguistic message: What’s she saying?
  - Emotion, mood, stress, etc.: How does he feel?
- Proper utilization of these two channels of information can lead to robust and enhanced capabilities, e.g.,
  - Locating and identifying the speaker
  - Speech understanding augmented with facial features
  - Speech, gesture, and sketching integration
  - Audio/visual information delivery

Audio-Visual Symbiosis (for Input)

- Personal Identity
- Speaker ID
- Face ID
- Microphone Arrays
- Camera Arrays
- Speech Understanding
- Lip/Mouth Reading
- Acoustic Paraling. Detection
- Visual Paraling. Detection
- Acoustic Signal
- Visual Signal
- Robust Locating
- Robust Understanding
- Robust Person ID
- Paralinguistic Information
- Linguistic Message
Audio-Visual Symbiosis (for Output)

Robust Delivery

Audio/Video Spotlight

Multiple Personas

Robust Avatar Rendering

Acoustic Signal

Visual Signal

Robust Paralinguistic Generation

Speech Synthesis

Facial Animation

Acoustic Paraling. Generation

Visual Paraling. Generation

Locating and Identifying a Speaker

(Darrell, Fisher, Hazen)
Audio-Visual Synchrony Detection

- The “Cocktail Party” problem: which audio to listen to?
  - A visual link between user and device can help
    - Detect and recognize user
    - Confirm that utterance was from user

Audio-Visual Synchrony Detection (cont’d)

- Find most informative subspace projection between audio and video features using mutual information
Audio-Visual Fusion for Talker Tracking

Audio-only tracking is often ambiguous in reverberant environments.

Adding vision-based location cues makes tracking unambiguous!

Audio-Visual Person Verification

Combined audio-visual inputs reduces equal error rate by 90%
Audio Visual Speech Recognition
(Hazen, Glass)

Lip-reading significantly helps human recognition of speech in noisy environments.

Preliminary studies in automatic audio-visual speech recognition show the same trend.
AVSR Data Collection & Feature Extraction

Integrating Speech with Gesture and Sketching (Darrell, Davis)
Gesture and Speech

- Integrating speech with visual cues for multimodal interaction
  - Un-tethered audio/video processing
  - Pointing and manipulation gestures (e.g., “that one,” or “this big.”)

Sketching & Pen Gesture Understanding

- Sketching is a natural means of communication between people
- We have begun to develop a toolkit to make it a natural means of communication with legacy software systems
- We have used it in several applications, including
  - Mechanical design
  - Software development
- Integration with speech input is often desirable (e.g., “draw two parallel lines”)

Victor Zue — MIT Computer Science and Artificial Intelligence Laboratory

5/18/2005

Victor Zue — MIT Computer Science and Artificial Intelligence Laboratory
Audio Visual Information Delivery
(Poggio, Glass)

New, data-driven approach can produce very natural and intelligible synthetic speech
We can now produce photo-realistic animations
These animated agents can speak and sing in different languages
We can combine speech synthesis and facial animation to produce realistic avatars
Summary

- Oxygen strives for pervasive, human-centered computing
- Next-generation interfaces should be human centric:
  - Carry on a conversation with the user
  - Permit multi-modal interactions using speech, vision, gesture and sketching,
  - Create an integrated, immersive environment
- Many faculty members, researchers and students at CSAIL are actively conducting research in these areas
- Systems with limited capabilities are beginning to emerge
- But much research remains

The Team

- Vision:
  - Trevor Darrell, John Fisher, Tony Ezzat, Tommy Poggio
- Speech and Language:
  - Jim Glass, TJ Hazen, Stephanie Seneff, Chao Wang, Victor Zue
- Sketching:
  - Randy Davis
- Plus many of their talented students
Example of Content Processing

**English:** Some thunderstorms may be accompanied by gusty winds and hail

- **clause:** weather_event
- **topic:** precip_act, name: thunderstorm, num: pl
- **quantifier:** some
- **pred:** accompanied_by
- **adverb:** possibly
- **topic:** wind, num: pl, pred: gusty
  - **and:** precip_act, name: hail

**Japanese:** ところどころ、強いかぜそしてあられを伴う雷雨

**Spanish:** Algunas tormentas posiblemente acompañadas por vientos racheados y granizo

**Chinese:** 一些雷雨可能會伴有陣風和冰雹

*Multilingual Weather Information*