W3C standards for Multimodal Interaction

Dave Raggett
W3C/Canon
W3C Activity lead for Voice & Multimodal

Introduction

- 10 years ago, the Mosaic browser unleashed the Web, stimulating massive growth on an unprecedented scale
- Today the Web is pervasive and enables an incredible range of services
- Mobile networks and falling hardware costs make in-car systems increasingly attractive
- W3C is developing the vision and core technical standards for the multimodal Web

The Need for Standards

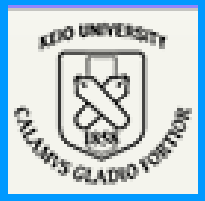
Bluetooth headset

SonyEricsson P800

Lexus console

Multimodal – Our Dream

- Adapting the Web to allow multiple modes of interaction:
  - *GUI, Speech, Vision, Pen, Gestures, Haptic interfaces*
- Augmenting human to computer and human to human interaction
  - *Communication services involving multiple devices and multiple people*
- Anywhere, Any device, Any time
  - *Services that dynamically adapt to the device, user preferences and environmental conditions*
- Accessible to all



W3C Multimodal Interaction Working Group

- Initial requirements study in W3C Voice Browser working group, followed by a joint workshop with the WAP Forum in 2000
- Working Group was formed in February 2002, and the following organizations are participating:

  Access, Alcatel, Apple, Aspect, AT&T, Avaya, BeVocal, Canon, Cisco, Comverse, EDS, Ericsson, France Telecom, Fraunhofer Institute, HP, IBM, INRIA, Intel, IWA/HWG, Kirusa, Loquendo, Microsoft, Mitsubishi Electric, Motorola, NEC, Nokia, Nortel Networks, Nuance Communications, OnMobile Systems, Openstream, Opera Software, Oracle, Panasonic, ScanSoft, Siemens, SnowShore Networks, Sun Microsystems, Telera, Tellme Networks, T-Online International, Toyohashi University of Technology, V-Enable, Vocalocity, VoiceGenie Technologies, Voxeo
- Largest working group in W3C Interaction Domain
- Due to be rechartered in early 2004

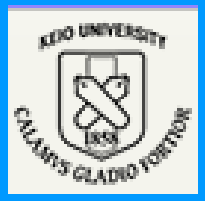
Benefits of Open Standards

Although costly to develop, the benefits are huge

- Benefit from helping to drive standards
  - You are no longer limited by the resources available within your own organization
  - The open standards process provides an opportunity to share ideas and experiences
  - Stronger specifications through broader review
  - Share costs of developing test suites
- Open standards stimulate greater innovation, increased competition, and greater choice
- Faster growth through increased market confidence
  - customers feel more secure through greater choice

Multimodal Devices

- High end – Desktops and Kiosks
  - Large vocabulary speech recognition, full sized keyboard and large high resolution display
- Mid-range – PDA's, Cars and upmarket phones
  - Smaller display, and limited memory
    - Large emerging market for in-car devices
      - Temperature and humidity issues for cars
    - Convergence of PDA, Phone and Media player
- Low end – mass market mobile phones
  - Long battery life, limited processing power, low memory, small display, keypad and audio
  - Application distributed between phone and network
  - Limited local recognition for speech



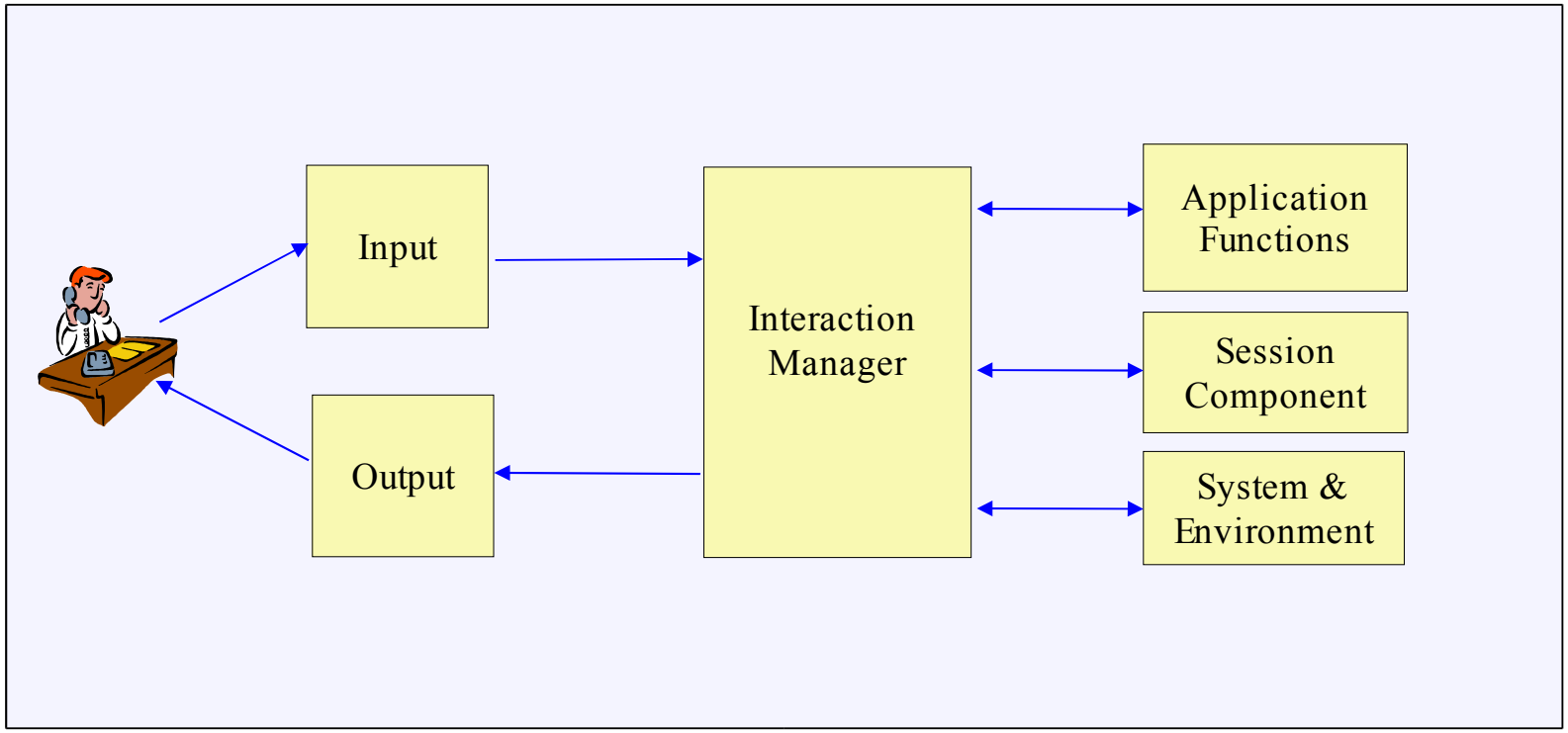
Automotive Applications

Web technology will reduce costs and increase flexibility

- Control of secondary systems
  - e.g. air conditioning
- Status and diagnostics
- Entertainment (radio, music, video, games)
- Journey planning and traffic management
  - Satellite and Inertial navigation
  - Online access to traffic and other information
    - Combine local and remote maps, photo's etc.
- Telephony and related services
  - Address book and name dialing
  - Messaging (voice mail, text messaging etc.)

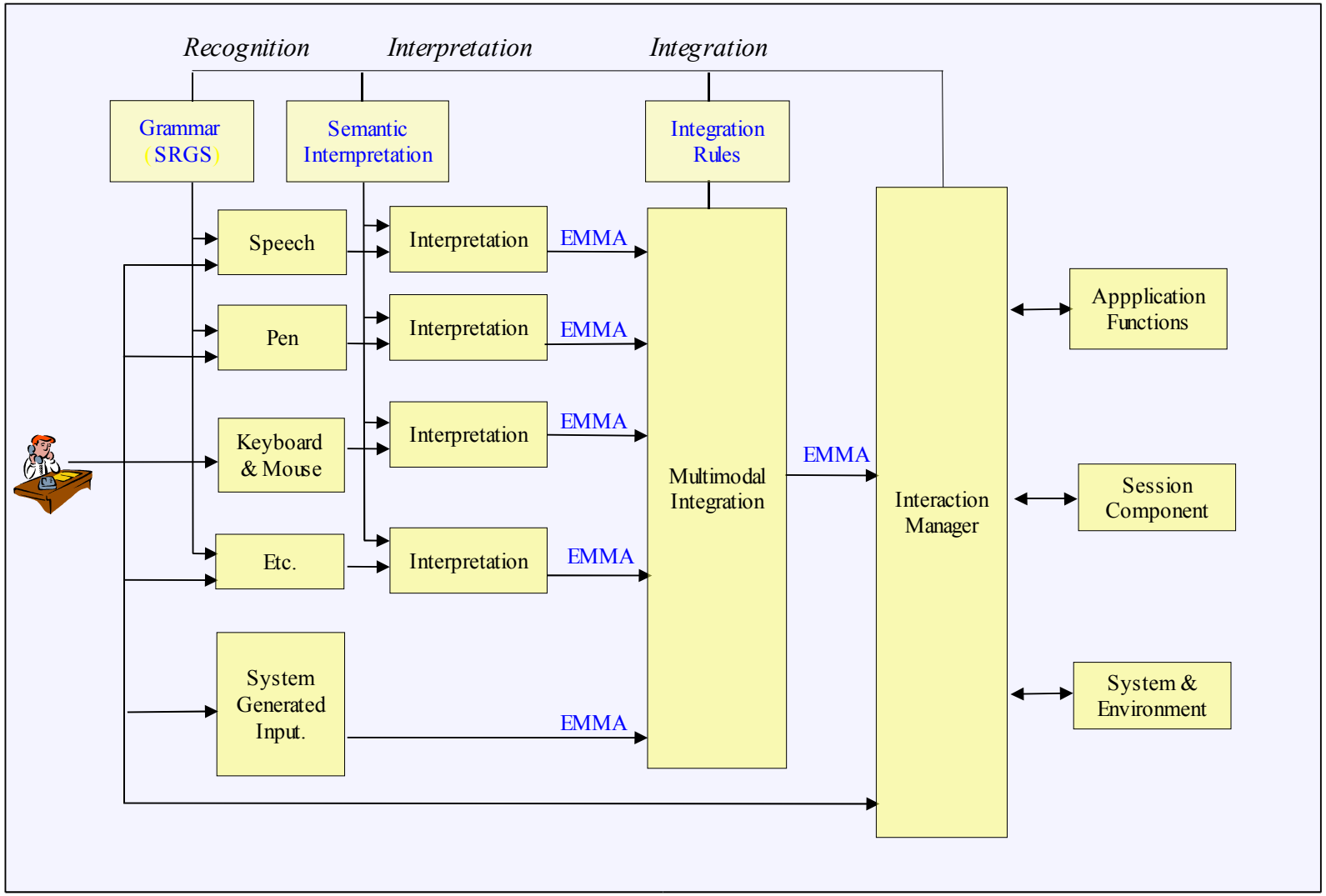
W3C Multimodal Interaction Framework

- Basis for creating applications in terms of markup, scripting, style sheets and other resources
- Architecture neutral: standalone or distributed



May involve multiple devices and users

Input Details



W3C MMI Activity
Current Work Items

- Multimodal Interaction Framework
  - Modality Interfaces (DOM)
  - Interaction Management
  - System and Environment
  - Composite Input
  - Session support
- EMMA – Extensible Multi-Modal Annotations
  - XML based interface between input components and interaction management
- Ink markup language
  - Enable server-side processing of electronic ink
  - Application to representing pen gestures

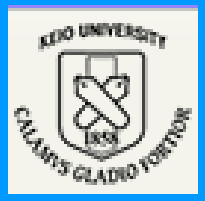
Modality Interfaces

- Modality components
  - Simple or complex input and/or output components
    - Voice, Pen, Display, Keypad, ...
- Modality components plug into Host environments
  - Markup language such as XHTML, SMIL or SVG
  - Scripting environment such as EMAScript
- We are defining a common set of interfaces in terms of the W3C Document Object Model (DOM)
  - Loading/Unloading a component
  - Events
  - Read/Write interfaces

EMMA

XML based transfer format between input components and interaction management

- Natural language input
  - Recognition
    - Speech grammars + acoustic models
    - Ink grammars + stroke models for handwriting
  - Extracting semantic results via grammar rule annotations (W3C Semantic Interpretation spec)
  - Results expressed as XML in app specific markup
- EMMA = annotated interpreted user input
  - Input mode, e.g. speech, pen or keypad
  - Confidence scores
  - Time stamps
  - Alternative recognition hypotheses
  - Sequential and partial results
  - Compositions from multiple modes



EMMA – an example

- User says: *"I want to fly to Boston"*
  - N-best list of recognition hypotheses
    - Destination is Boston with confidence of 0.6
    - Destination is Austin with confidence of 0.4
  - <destination> city name </destination>

```
<emma:emma xmlns:emma="http://www.w3.org/2002/emma">
  <emma:one-of>
    <emma:interpretation confidence="0.6">
      <destination>Boston</destination>
    </emma:interpretation>
    <emma:interpretation confidence="0.4">
      <destination>Austin</destination>
    </emma:interpretation>
  </emma:one-of>
</emma:emma>
```

- RDF as alternative syntax for annotations

Interaction Management

Coordinating data and execution flow

- Study underway of existing practices
  - State transition and plan based approaches
  - Role of markup versus scripting
  - Usability and Authorability
- User initiative vs System directed dialogs
  - Event driven interaction
    - Tap on link, or say link caption, to follow that link
    - Tap on a field to hear field's audio prompt
    - You immediately see what the system recognized
  - Dialog driven interaction
    - Speech and handwriting recognition are imperfect
    - Author can provide system directed dialog
      - Speech dialog triggered by event, e.g. tapping on a field
      - System asks you a sequence of questions
  - Adapting to hands and eyes free operation
    - e.g. when driving off in a car

System and Environment

Usability will be critical to Multimodal Applications

- Enabling applications to dynamically adapt to device, user preferences and environmental conditions, e.g.
  - Aural noise (window down)
  - Car in motion or parked
    - Cognitive load on driver – critical safety factor
  - Current location
  - Network connectivity (loss of signal)
- Object model
  - Query/Update properties
  - (Un) Subscribe to events
- Relationship to CC/PP and Device Independence

Integration of Composite Input

- Composite Input involving multiple modes
  - You tap on several files and say “print these”
  - You say “Are there any restaurants in this area” and draw a rough circle on a map with the pen
- Aggregation of sequential input on same mode
- Different kinds of constraints on composition:
  - Semantic type, number, input mode, spatial, temporal, logical, etc.
- One approach is for author to markup deictic references from speech and for the integration component to search for matching gestures
- We are exploring several approaches ...

Session Management

- Important for multitasking, distributed and multi-user applications
  - Temporary and Persistent sessions
  - Support for data synchronization (without blocking)
  - Resource broker and resource descriptions
    - Making it easier for authors to create distributed apps
      - Hide details of specific naming schemes and protocols
    - Bridging the gap between Web, Instant Messaging, conferencing and email based applications
- Still at a very early stage of consideration

Pen Input and InkML

- Pen input can be used for
  - Gestures,
  - Drawings
  - Note taking
  - Handwriting
  - Signature verification
  - Specialized notations: math, music, chemistry, ...
- Why develop a standard ink format?
  - Use of ink together with speech
  - Added flexibility through server-side interpretation
  - Passing ink to other people (drawings and notes)
- XML based transfer format – InkML
  - Possible application to describing pen gestures

In conclusion ...

- Work is underway to extend the Web to realize the opportunities for mobile and automotive applications
- Come and join us and help realize the benefits
- More information about W3C can be found at:
  - http://www.w3.org/
- *Thank You for listening!*