Multimodal Interaction Standards at the World Wide Web Consortium

User interfaces for natural interaction

Deborah Dahl
Chair, W3C Multimodal Interaction Working Group
Conversational Technologies

LDC Institute
December 4, 2015

Natural Interaction

Most human-computer interaction is semantically very simple

"I want that"

- Check a radio button
- Select a choice from a pulldown menu
- Fill in a field in a form
- Select text
- Scroll

What about more complex semantics?

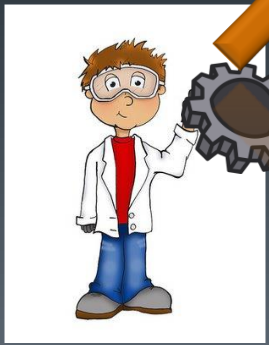
  - *We want to go from Philadelphia to Miami on January 15 for the three of us in the morning.*
  - *Will this be one-way or round trip?*
  - *Round trip, and we'll need a car, too*
- Express an intent including several fields at once
- Engage in an interactive dialog

Multimodal Interaction

Natural language, speech, pointing, camera…

Standards for Multimodal Interaction
Why?

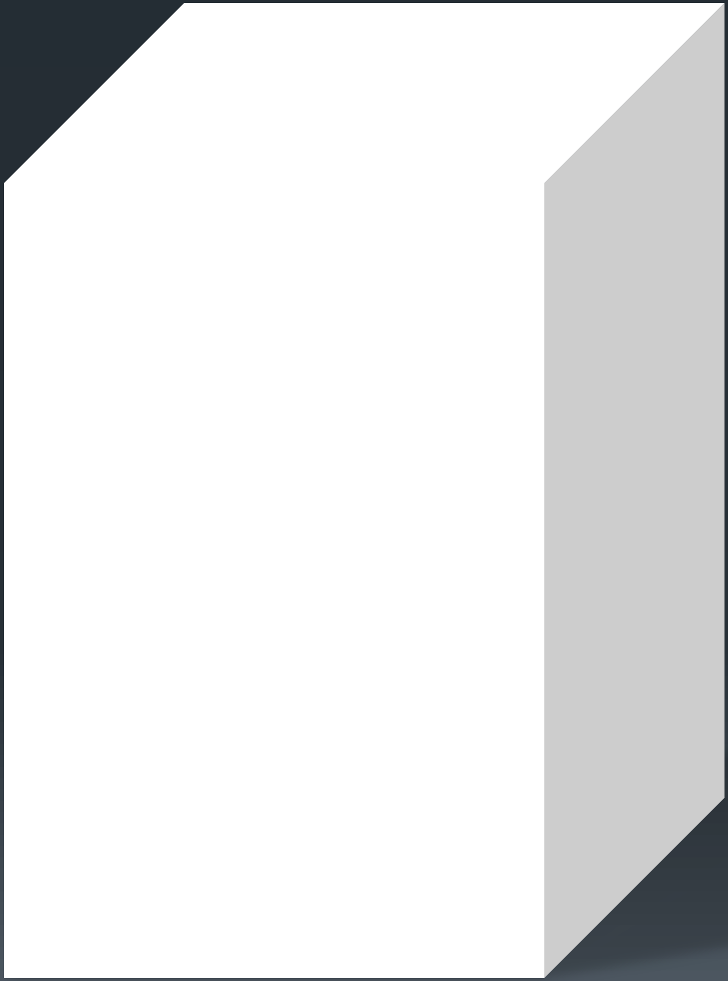
- Complex technologies
- Interoperability
- New modalities should be pluggable



Standards improve component interoperability:

- Even a small organization can contribute to a complex project that uses standards
- Keep focus on innovation and key features

Proprietary systems prevent other developers from contributing



Standards leverage thinking of many participants and points of view

Standards stimulate tools development

Standards help developers take advantage of other people's wor

Codify best practices and avoid repeating work

Standards reduce the need to learn multiple proprietary interfaces

Technical Standards Organizations

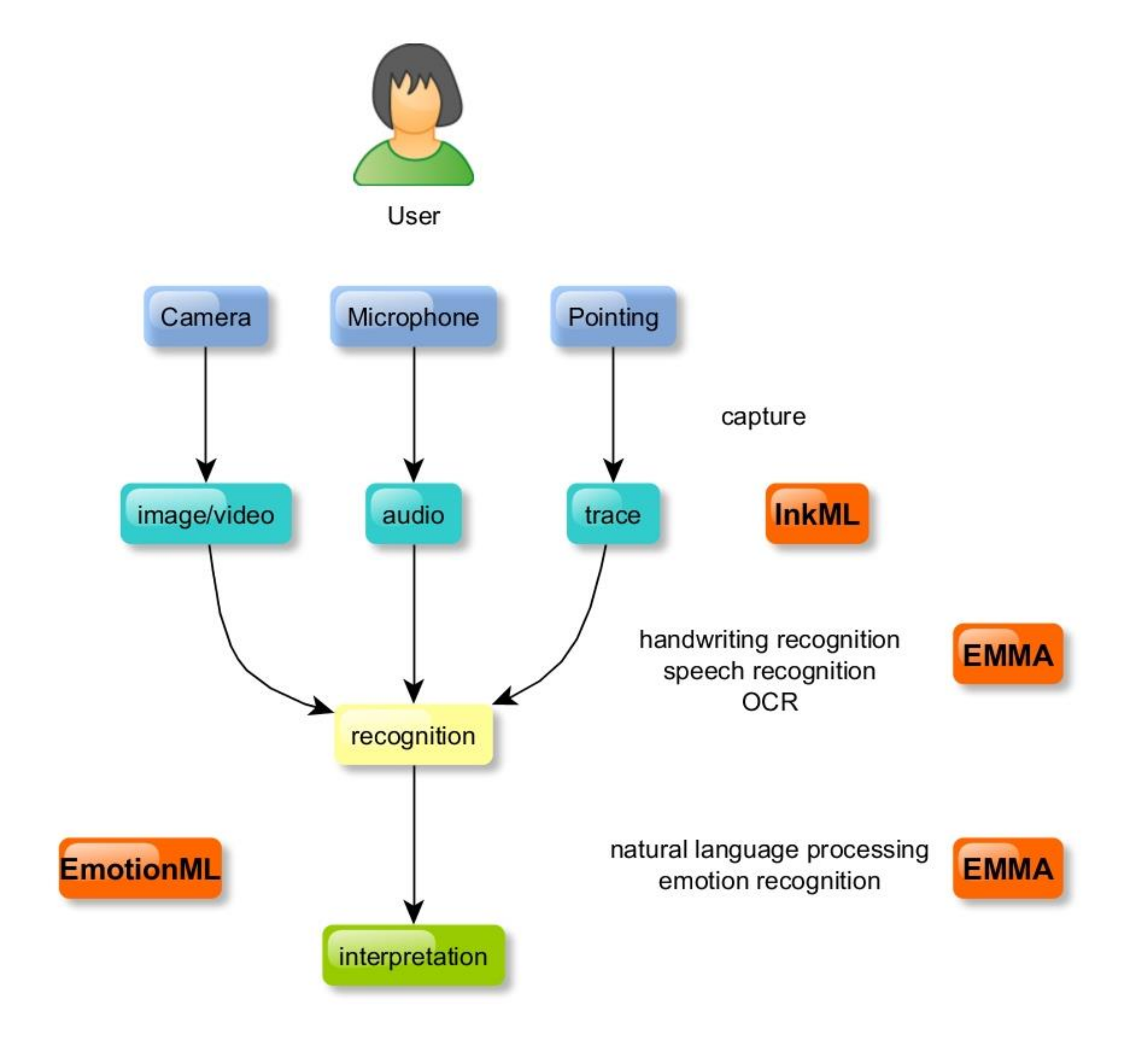
- The World Wide Web Consortium
  - Develops standards for the Web
  - Founded in 1994 by Tim Berners-Lee, the inventor of the World Wide Web
  - 407 member organizations
- Other relevant organizations
  - ISO
  - IETF
  - IANA
  - Unicode Consortium
  - TEI
  - DAISY Consortium

W3C Multimodal Interaction Working Group

- Started in 2002
  - Standards for user input
  - Communication among components
  - Technology has changed a lot, but users haven’t
- Today we’ll focus on user input

Multimodal Interpretation Process

- Capture
- Recognition
- Interpretation
- Response



Three standards for user interaction

- Extensible Multimodal Annotation (EMMA)
- Emotion Markup Language (EmotionML)
- Ink Markup Language (InkML)

EMMA Use Cases

- User input to multimodal dialog systems
- Uniform logging and analysis language for tuning multimodal dialog systems

EMMA Requirements

- Modality independence
- Extensible to new modalities and new semantic representations
- Able to represent most common metadata
  - Tokens
  - Timestamps
  - Confidence
  - Alternatives
- Widely available syntax (XML)
- Composite multimodal inputs
- Derivation history

EMMA 1.0 : Extensible Multimodal Annotation

- EMMA standard provides a common XML language for representing the interpretation of inputs to spoken and multimodal systems
- XML markup for capturing and annotating various processing stages of user inputs
  - Container elements / Annotation elements and attributes
- W3C Recommendation February 2009 http://www.w3.org/TR/emma
- Implementations
  - AT&T (APIs), Microsoft, Nuance, Loquendo, Conversational Technologies, DFKI, Openstream…

EMMA 1.0 Major Features

- Semantic interpretation (application-specific format)
- Metadata
  - Confidence
  - Alternatives (nbest or lattice)
  - Timestamps
  - Tokens of input
  - Medium and Mode
  - Function (dialog, transcription, translation)
  - Pointer to signal
  - Process
  - Grammar
  - Language
  - Verbal
  - Grouping of related inputs
  - Dialog turn
- Application-specific information (<emma:info>

EMMA Example

```
<emma:interpretation emma:confidence="0.885" emma:process="wit.ai"
emma:tokens="put 5 pounds of granny smith apples on my list" id="interp41">
       <nlResult>
          <outcomes>
         <measure>
            <value>pounds</value>
         </measure>
             <number>
            <value>5</value>
         </number>
             <shopping_item>
           <value>granny smith apples</value>
             </shopping_item>
         <intent>shoppingList</intent>
        </outcomes>
       </nlResult>
 </emma:interpretation>
```

EMMA Composite Example for Multimodal Fusion

“zoom in here”

```
<emma:interpretation emma:medium="acoustic tactile" emma:mode="voice ink"
emma:process="http://example.com/myintegrator.xml">
        <emma:derived-from
resource="http://example.com/voice1.emma/#voice1" composite="true"/>
        <emma:derived-from resource="http://example.com/pen1.emma/#pen1"
composite="true"/>
        <command>
                <action>zoom</action>
                <location>
                        <type>area</type>
                        <points>42.1345 -37.128 42.1346 -37.120 ... </points>
                </location>
        </command>
</emma:interpretation>
```

EmotionML

- Represent emotions
- Three use cases
  - Support for emotion recognition (from language, voice, behavior)
  - Support for presenting emotions (avatar facial expression, TTS, robot posture)
  - Annotation of natural data for research
- Separate emotion metadata from vocabulary

EmotionML Major Features

- Categorize emotions
- Intensity of emotions
- System confidence in category
- How the emotion is expressed (face, text, gaze…)
- Timing (start, end, duration, relative time)
- Timing in media
- Accommodate multiple vocabularies

EmotionML Examples

A simple emotion

```
<emotion category-set="http://www.w3.org/TR/emotion-voc/xml#everyday-categories">
            <category name="satisfied"/>
</emotion>
```

- A mixture of emotions

```
<emotion category-set="http://www.w3.org/TR/emotion-voc/xml#big6">
            <category name="sadness" value="0.3"/>
            <category name="anger" value="0.8"/>
            <category name="fear" value="0.3"/>
</emotion>
```

Emotion within Media

```
<emotion category-set="http://www.w3.org/TR/emotion-voc/xml#big6">
        <category name="happiness"/>
        <reference uri="myVideo.avi#t=clock:2009-07-26T11:19:01Z,2009-07-26T11:20:01Z"/>
    </emotion>
```

EMMA/EmotionML Example

```
<emma:interpretation emma:confidence="0.15075567228888181"
emma:tokens="i'm feeling very cheerful today" id="interp35">
          <emma:derived-from composite="false" resource="#initial2"/>
          <emotion category-set="http://www.w3.org/TR/emotion-
voc/xml#everyday-categories">
            <category confidence="0.15075567228888181"
name="happy"/>
          </emotion>
        </emma:interpretation>
        <emma:derivation>
          <emma:interpretation emma:device-type="keyboard"
emma:end="1448922106115" emma:expressed-through="language"
emma:function="emotionClassification" emma:lang="en-US"
emma:medium="tactile" emma:mode="keys" emma:tokens="i'm feeling very
cheerful today" emma:verbal="true" id="initial2">
            <emma:literal>i'm feeling very cheerful today</emma:literal>
          </emma:interpretation>
        </emma:derivation>
      </emma:emma>
```

EMMA and EmotionML Demo

- http://www.proloquia.com/processLanguage.html

InkML – Representing ink traces

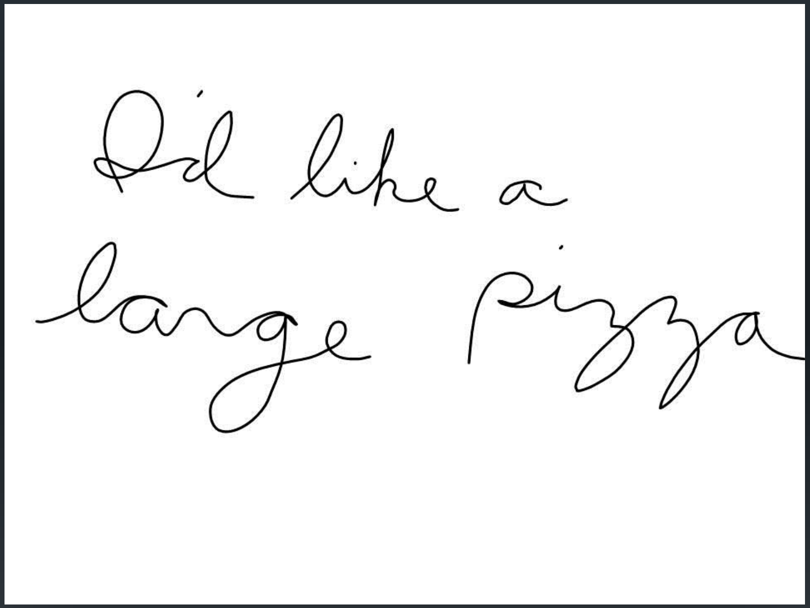
Use Cases

- Ink Messaging
- Ink and SMIL coordinating ink with a spoken commentary
- Ink Archiving and Retrieval
- Electronic Form-Filling
- Pen Input and Multimodal Systems

InkML Major Features

- Representation of an ink trace
- Pen orientation
- Pen color
- Stroke width
- Brushes
- Timing
- Hardware used for ink capture
- Canvas
- Annotation (open-ended, for example to describe the user or the type of content, such as math or music)

InkML Example



I'd like a large pizza

InkML for “I’d”

(Microsoft Office Implementation)

<inkml:trace brushRef="**#br0**" contextRef="**#ctx0**" timeOffset="**-16454.2572**">
1556 257,'0'0,"0"0,0 0,0 0,0 0,0 0,-14-15,-1-16,1 1,-1 0,1-30,-30-31,-14-45,-29-60,-1 0,30 60,-14-
60,-1-15,58 60,30-60,-1 0,30 0,14 0,-14 90,14-15,-14 46,14 14,14-29,1-1,0 46,0 15,58-16,-59
31,1 15,29 0,29 30,-15 0,-14 30,-15 1,0 14,1 45,-1 16,-29 0,0 15,-29-16,-29 31,0-15,-14 15,-30
75,-14-30,0-60,0 0,-1-16,-13 1,-59 15,-15-1,45-59,-16-1,30-30,29-14,0-1,-30-15,-43 15,1-30,-1
0,29-15,15-15,-30-31,45 16,28 15,0 0,1-15,14-1,14 1,0 0,1 0,-1-1,15 31,15-15,-1 15,1 0,14 0,15
0,-1 0,1 15,14 15,0 0,-14 0,29 15,43 30,0 1,1-16,14 15,0-29,-15-1,15 0,0-30,0-15,0-15,-15-1,29
1,-14 0,-14-15,-30 15,0 0,102-31,-43 31,-88 15,0 0,-14 0,-1 15,1 0,-15 0,0 0,0 0,0 0,0 15,0 0,-14
0,-1 0,1 0,0-15,-1 0,1 0,-1 0,1 0,-1 0,1 0,-1 0,1 0,-15 0,0 0,0 0,0 0,0 0,0 0,0 0,0 0,-15 0,-14 15,0
0,-14 0,-1 16,0-1,1 15,-1 0,0 1,1-1,-1 0,30 0,-1 1,0-16,1 0,14 0,29 0,-14 0,28 1,1-16,-1-15,16 0,-
1-15,29-31,15-14,0-31,58-44,0-1,-59 45,-13-29,28-46,-29-30,-28 75,-1-15,-29 1,0 44,-15 31,1-
46,0-30,-15 31,0 60,0-1,-15 1,15 0,0 15,-15 15,1 15,-1-16,1 16,-1 16,-14-16,-14 30,-16 30,-42
46,-16 14,30 76,-15-15,59-45,-1 0,0 15,15 0,15-1,-1 47,15 14,15-76,-1 1,15-45,44-1,43 16,-14-
16,-15-44,-28-16,28 15,44 0,72-15,379 1
</inkml:trace>

EMMA for “I’d”
(Microsoft Office Implementation)

```
<emma:emma version="1.0" xmlns:emma="http://www.w3.org/2003/04/emma">
		<emma:interpretation id="{F141D774-6A5B-483A-BD75-D517116EE936}" emma:mode="ink“
				emma:medium="tactile">
				<msink:context type="inkWord" rotatedBoundingBox="2603,2930 7412,2340 7873,6105
3065,6694" xmlns:msink="http://schemas.microsoft.com/ink/2010/main"/>
		</emma:interpretation>
		<emma:one-of id="oneOf1" disjunction-type="recognition">
				<emma:interpretation emma:lang="en-US" id="interp5" emma:confidence="1">
						<emma:literal>I'd</emma:literal>
				</emma:interpretation>
				<emma:interpretation emma:lang="en-US" id="interp6" emma:confidence="0">
						<emma:literal>I‘ll</emma:literal>
				</emma:interpretation>
				<emma:interpretation emma:lang="en-US" id="interp7" emma:confidence="0">
						<emma:literal>Ide</emma:literal>
				</emma:interpretation>
				<emma:interpretation emma:lang="en-US" id="interp8" emma:confidence="0">
						<emma:literal>Ice</emma:literal>
				</emma:interpretation>
				<emma:interpretation emma:lang="en-US" id="interp9" emma:confidence="0">
						<emma:literal>Idol</emma:literal>
				</emma:interpretation>
		</emma:one-of>
</emma:emma>
```

InkChat

- http://inkchat.org/index.html

Next steps

- EmotionML and InkML basically finished
- EMMA 2.0

EMMA 2.0*

- EMMA 2.0 Working Draft published 9/8/2015
- http://www.w3.org/TR/emma20/

*Slides courtesy of Michael Johnston, Interactions, W3C Invited Expert and the Editor of EMMA

EMMA 2.0 Summary

- System output
- Incremental results
- Location annotations
- Extensions to emma:tokens
- Non XML semantic payloads
- Support for annotations
- Multiple grammars
- Specification of body part used to express an input
- Mechanisms to reduce document size
- Specification of partial content

1. System Output

- EMMA 1.0 focused on capturing inputs to multimodal systems

  <emma:interpretation>

- In interactive systems components also share messages specifying system output
- EMMA 2.0 adds support for containing and annotating the various stages of processing of system output (Section 3.3)

  <emma:output>

- For example, dialog manager generates semantic representation in <emma:output>, natural language generation yields <emma:output> converting this to a prompt, which is then passed to TTS

<emma:output> example

- Using emma:group for multimodal output combining graphics and speech
- <emma:one-of> for spoken generation variants

```
<emma:emma version="2.0"
    xmlns:emma="http://www.w3.org/2003/04/emma"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/04/emma
     http://www.w3.org/TR/2009/REC-emma-20090210/emma.xsd"
    xmlns="http://www.example.com/example">
<emma:group
    emma:process="http://example.com/multimodal_presentation_planner">
    <emma:one-of id="ooo1"
         emma:medium="acoustic"
         emma:mode="voice"
         emma:function="dialog"
         emma:result-format="application/ssml+xml"
         emma:process="http://example.com/nlg">
        <emma:output emma:confidence="0.8" id="tts1">
             <speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
                          xml:lang="en-US">
             I found three flights from Boston to Denver.
             </speak>
        </emma:output>
        <emma:output emma:confidence="0.7" id="tts2">
             <speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
                          xml:lang="en-US">
             There are flights to Boston from Denver on United, American,
             and Delta.
             </speak>
        </emma:output>
     </emma:one-of>
     <emma:output id="gui1"          emma:medium="visual"
         emma:mode="gui"
         emma:result-format="text/html"
emma:function="dialog"
         emma:process="http://example.com/gui_gen">
           <html xmlns="http://www.w3.org/1999/xhtml">
            <body>
                <table>
                    <tr><td>United</td><td>5.30pm</td></tr>
                  <tr><td>American</td><td>6.10pm</td></
                  <tr><td>Delta</td><td>7pm</td></tr>
                </table>
          </body>
         </html>
      </emma:output>
</emma:group>
</emma:emma>
```

Spoken alternatives

Graphical presentation

EMMA 2.0 Output Demo

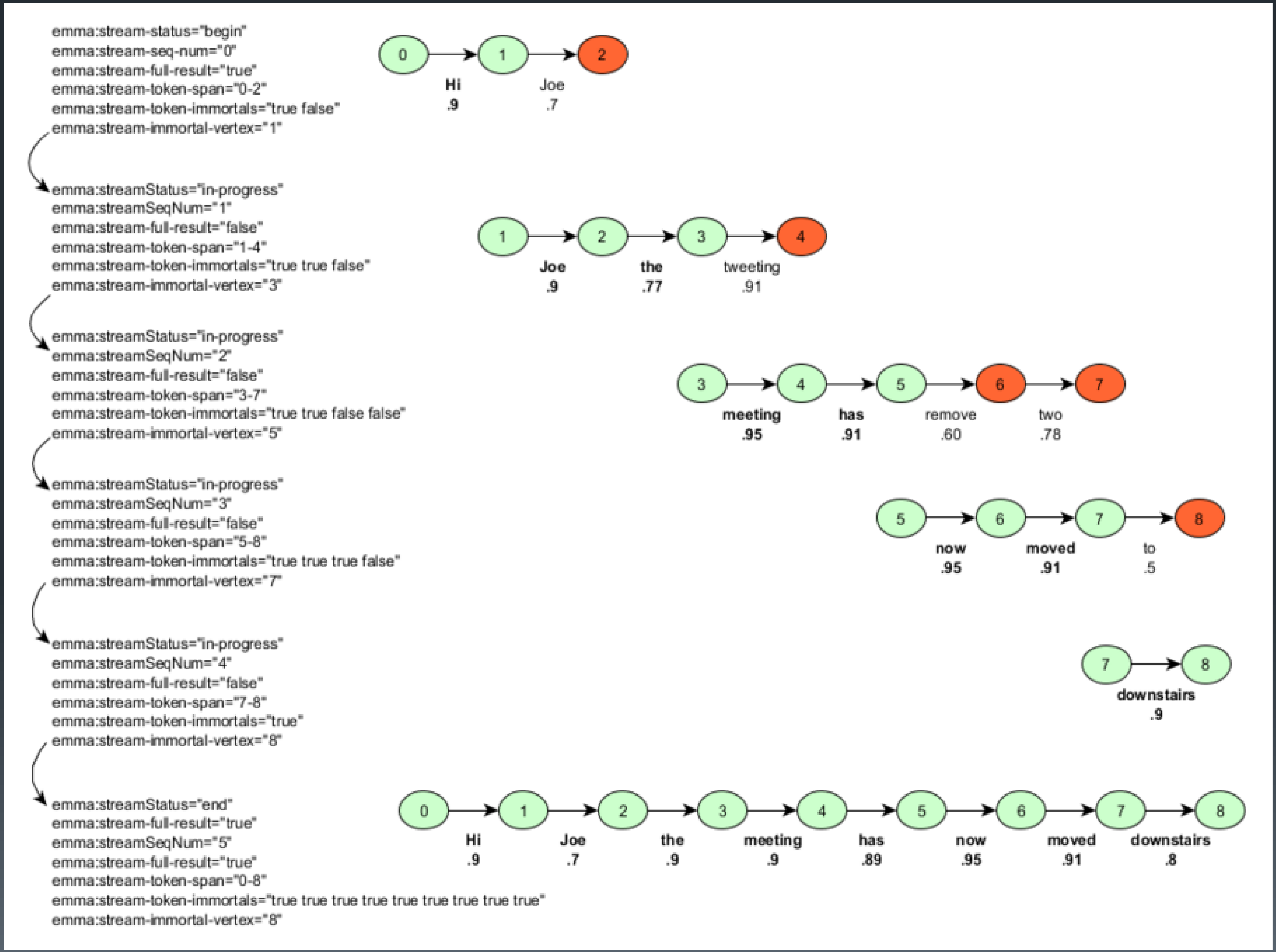
http://www.proloquia.com/NYCEventsDemo.html

http://findeventsnyc.net

Incremental Results in EMMA (Sec 4.2.23)

- e.g. ASR partial results are one use case:
  - “piazza ...”
  - “pizza in ...”
  - “pizza in santa ...”
  - “pizza in santa cruz”
- Key to providing a responsive user experience for mobile speech applications
- Also for handwriting, gesture, computer vision applications

Incremental Results in EMMA (cont.)



```
emma:stream-status="begin"
emma:stream-seq-num="0"
emma:stream-full-result="true"
emma:stream-token-span="0-2"
emma:stream-token-immortals="true false"
emma:stream-immortal-vertex="1"

emma:streamStatus="in-progress"
emma:streamSeqNum="1"
emma:stream-full-result="false"
emma:stream-token-span="1-4"
emma:stream-token-immortals="true true false"
emma:stream-immortal-vertex="3"

emma:streamStatus="in-progress"
emma:streamSeqNum="2"
emma:stream-full-result="false"
emma:stream-token-span="3-7"
emma:stream-token-immortals="true true false false"
emma:stream-immortal-vertex="5"

emma:streamStatus="in-progress"
emma:streamSeqNum="3"
emma:stream-full-result="false"
emma:stream-token-span="5-8"
emma:stream-token-immortals="true true true false"
emma:stream-immortal-vertex="7"

emma:streamStatus="in-progress"
emma:streamSeqNum="4"
emma:stream-full-result="false"
emma:stream-token-span="7-8"
emma:stream-token-immortals="true"
emma:stream-immortal-vertex="8"

emma:streamStatus="end"
emma:stream-full-result="true"
emma:streamSeqNum="5"
emma:stream-full-result="true"
emma:stream-token-span="0-8"
emma:stream-token-immortals="true true true true true true true true true"
emma:stream-immortal-vertex="8"
```

Incremental Results in EMMA (cont.)

New attributes supporting incremental:

- emma:stream-id
- emma:stream-seq-num
- emma:stream-status
- emma:stream-full-result
- emma:stream-token-span
- emma:stream-token-span-full
- emma:stream-token-immortals
- emma:stream-immortal-vertex

Incremental Results in EMMA (cont.)

```
<emma:emma version="2.0"
    xmlns:emma="http://www.w3.org/2003/04/emma"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/04/emma
     http://www.w3.org/TR/2009/REC-emma-20090210/emma.xsd"
    xmlns="http://www.example.com/example">
  <emma:interpretation id="int12"
       emma:medium="acoustic"
       emma:mode="voice"
       emma:confidence="0.8"
       emma:tokens="Joe the tweeting"
       emma:token-type="word"
       emma:token-score="0.9 0.77 0.91"
       emma:stream-id="s1"
       emma:stream-seq-num="1"
       emma:stream-status="in-progress"
       emma:stream-full-result="false"
       emma:stream-token-span="1-4"
       emma:stream-token-immortals="true true false"
       emma"stream-immortal-vertex="3">
      <emma:literal>Joe the tweeting</emma:literal>
  </emma:interpretation>
</emma:emma>
```

Location annotation (Sec 4.1.10)

- Many mobile devices and sensors are equipped with geolocation capabilities
- Information about where an event occurred can be very useful both for interpretation and logging
- Annotating interpretations with location information in EMMA 2.0 is achieved with the <emma:location> element
- The emma:location attributes are based on the W3C Geolocation API specification, with the addition of attributes for a description of the location and address information.

Location annotation (cont.)

```
<emma:emma version="1.1"
    xmlns:emma="http://www.w3.org/2003/04/emma"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/04/emma
    http://www.w3.org/TR/2009/REC-emma-20090210/emma.xsd"
    xmlns="http://www.example.com/example">
   <emma:location
      latitude="42.361860"
       longitude="-71.091840"
       altitude="6.706"
       accuracy="20.5"
       altitudeAccuracy="1.6"
       heading=""
       speed=""
       description="W3C MIT office"
       address="32 Vassar Street, Cambridge, MA 02139 USA"/>
   </emma:location>
   <emma:interpretation id="nlu1"
      emma:medium="acoustic"
      emma:mode="voice"
      emma:tokens="flights from boston to denver">
         <origin>Boston</origin>
         <destination>Denver</destination>
   </emma:interpretation>
 </emma:emma>
```

Geolocation specification

EMMA additions

Possible Additions: Semantic information about locations

- Spatial Data on the Web WG
  - Collaboration with the Open Geospatial Consortium
- IETF RFC 5139 Civic Addresses – XML format to represent addresses

Additional EMMA 2.0 Extensions (cont.)

- Specifying multiple grammars and indicate which are active:
  - emma:grammar-active
  - emma:active
- Support for semantic payloads beyond XML:
  - emma:result-format=
- Multiple info elements for vendor/app specific metadata
  - <emma:info>
  - emma:info-ref
- Specification of modality used to express an input
  - emma:expressed-through

Future work

- Simplified JSON version of EMMA
- More specific work on location
- Working on details of output

Summary

- Discussed three W3C standards that support multimodal interaction
- EMMA: a modality-independent format for expressing user input and system output
- EmotionML: a way to represent emotions
- InkML: a representation for ink traces

References

- Multimodal Working Group
  - http://www.w3.org/2002/mmi/
- Specs
  - InkML http://www.w3.org/TR/InkML/
  - EMMA http://www.w3.org/TR/emma20/
  - EmotionML http://www.w3.org/TR/emotionml/
- Other publications of interest
  - MMI use cases https://www.w3.org/wiki/MMI/Use_Cases

Time permitting

More about EMMA 2.0
Extensions to emma:tokens

- Support for per token scores e.g. word scores
- Ability to specify type of token

```
<emma:emma version=”2.0"
    xmlns:emma="http://www.w3.org/2003/04/emma"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3.org/2003/04/emma
     http://www.w3.org/TR/2009/REC-emma-20090210/emma.xsd"
    xmlns="http://www.example.com/example">
  <emma:interpretation id="int1"
      emma:tokens="From Cambridge to London tomorrow"
      emma:token-type="word"
      emma:token-score="0.9 0.7 0.7 0.9 0.8"
      emma:medium="acoustic" emma:mode="voice">
    <origin emma:tokens="From Cambridge">Cambridge</origin>
    <destination emma:tokens="to London">London</destination>
    <date emma:tokens="tomorrow">20030315</date>
  </emma:interpretation>
</emma:emma>
```

Additional EMMA 2.0 Extensions

- Support for addition of annotations:
  - <emma:annotation>
  - emma:annotated-tokens=
- Specification of process parameters (e.g. asr beamwidth)
  - <emma:parameters>
  - <emma:parameter>
  - emma:parameter-ref=
- Specification of models used in processing beyond grammars:
  - <emma:process-model>
  - emma:process-model-ref=

Additional EMMA 2.0 Extensions (cont.)

- Mechanisms for reducing size of EMMA documents
  - ref= added to number of elements allowing for use of URIs to point to content outside the document
  - (<emma:one-of>, <emma:sequence>, <emma:group>, <emma:info>,<emma:parameters>, <emma:lattice>)
- Partial Content:
  - emma:partial-content="true”
  - Along with ref= allows for document to contain partial content of element
  - e.g. N-best one-of and retrieval of full element through URI in ref=

Additional EMMA 2.0 Extensions (cont.)

- References to document on server
  - <emma:emma>
    - doc-ref (reference to where this document can be picked up from server)
    - prev-doc= (previous document in sequence)

EMMA Analysis Example: Speech Rate

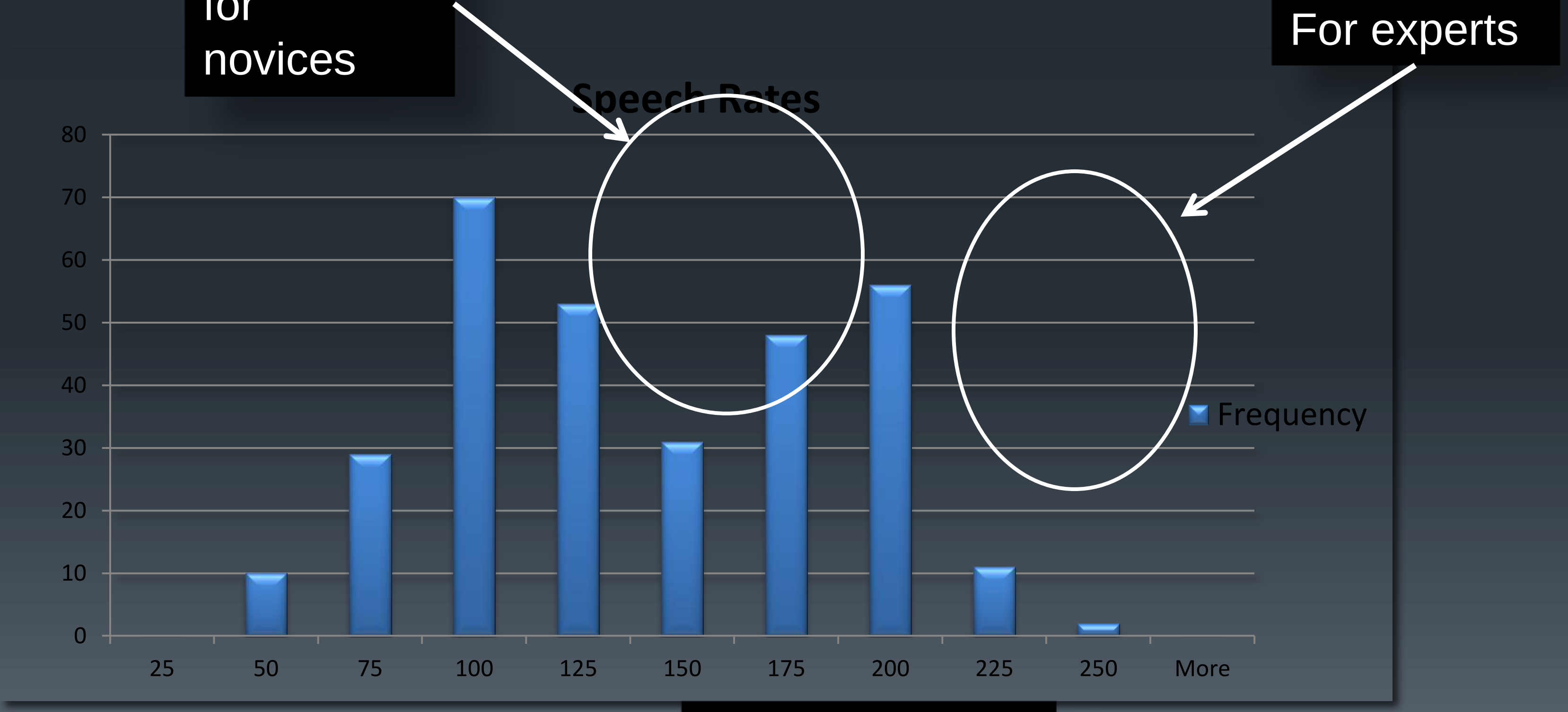
- Some users think that the application speaks too slowly
- Other users think that it speaks too fast
- If we dynamically adjust the system’s speech rate to the user’s speech rate, we can accommodate both kinds of users

Hypothetical Speech Rate Data

- Bimodal distribution may point to two kinds of users
- Match UI to different kinds of users

Slow down for novices

Speed up For experts



Words per minute

Calculating Users’ Speech Rate in real time from EMMA

- EMMA provides the words that the user spoke and the duration of the user’s speech
- # tokens/duration in minutes = words per minute
- We can measure the user’s speech rate and match the system’s speech rate in real time