Voice on the Web: Input/Output Modality Challenges
Why Speech? Fast data entry

- Multiple concepts per utterance
- Visual confirmation

Boston to London on Tuesday April 25, in the evening
Why Speech? User driven dialog = direct access

- Short-cut through application hierarchy

Show me books on wine
Why Speech? Efficient device control

- Simplified selection from long lists

Play Beautiful by James Blunt
Why Don't I have Speech Already?

Complexity and Cost

Error Handling

Visual has Priority
W3C Standards

Jerry Carter, 11 May 2007
Media Resources for Generated Audio and Video

Basic Process:
• Markup provided by the application
• Markup is processed and the context analyzed
• A media stream is generated and controlled
W3C: Speech Synthesis Markup Language

- [http://www.w3.org/TR/speech-synthesis](http://www.w3.org/TR/speech-synthesis)
- Provides rich text-to-speech markup and supports embedded audio recordings
  - Allows for TTS 'fallback' when audio is unavailable
- Targets regular dialogs
  - Does not handle emotion or non-verbal responses
- Allows for multi-voice and mixed language use

```xml
<prompt>
  <voice name='costello'>What's the guy's name on first base?</voice>
  <voice name='abbott'>No. What is on second?</voice>
  <voice name='costello'>I'm not asking you who's on second.</voice>
  <voice name='abbott'>Who's on first.</voice>
</prompt>

<prompt>
  <s xml:lang='en-UK'>Repeat after me. I will learn French today.</s>
  <s xml:lang='fr-FR'>J'apprendrai le français aujourd'hui.</s>
</prompt>
```
Document processing can be tricky…

- Abbreviations (St., ETA)
- Homographs (read, coax, conduct, dove, …)
- Pronunciation based on derivation (Jesus, Richard)
- Tokenization (123)

SSML provides `<say-as>`, `<token>`, etc. but can't solve all problems.
W3C: Pronunciation Lexicon Specification

- PLS provides a dictionary format
  - [http://www.w3.org/TR/pronunciation-lexicon/](http://www.w3.org/TR/pronunciation-lexicon/)
- Pronunciations may be specified using IPA or other alphabets
- Roles may be associated to assist with markup processing
- PLS helps achieve consistency
  - SSML and grammars may reference the same PLS lexicon

```xml
<?xml version="1.0"?>
<lexicon version="1.0"
  alphabet="ipa" xml:lang="en-US">
  <lexeme>
    <grapheme>INXS</grapheme>
    <phoneme>n ek ses</phoneme>
    <phoneme>n ık ses</phoneme>
  </lexeme>

  <lexeme role="ex:locale">
    <grapheme>Lima</grapheme>
    <phoneme>lima</phoneme>
  </lexeme>

  <lexeme role="ex:food">
    <grapheme>lima</grapheme>
    <phoneme>lai me</phoneme>
  </lexeme>
</lexicon>
```
Media Resources for User Input Recognition

**Basic Process:**
- Application offers constraints using markup or by reference
- Incoming media stream is compared against constraints
- Semantic meaning is attached
- Result is generated
W3C Specifications for Grammars

- **Speech Recognition Grammar Specification (SRGS 1.0)**
  - [http://www.w3.org/TR/speech-grammar/](http://www.w3.org/TR/speech-grammar/)
  - Recommendation published 16 March 2004
  - Finite state grammars
  - Speech and key entry (DTMF) supported
  - Allows for mixed language use

- **Semantic Interpretation for Speech Recognition (SISR 1.0)**
  - [http://www.w3.org/TR/semantic-interpretation/](http://www.w3.org/TR/semantic-interpretation/)
  - Recommendation published 5 April 2007
  - ECMA Script expressions for attaching semantic meanings
I Don't Care What You Said

- Normalizing information collected in grammars allows the application to focus on the meanings.

```xml
<?xml version='1.0'?>
<grammar version='1.0' root='digits' xml:lang='x-klingon' mode='voice' xmlns="http://www.w3.org/2001/06/grammar"
tag-format="semantics/1.0-literals">

  <rule id='digits'>
    <one-of xml:lang='en-SG'>
      <item> one <tag>'1'</tag></item>
      <item> two <tag>'2'</tag></item>
      <item> three <tag>'3'</tag></item>
      <item> four <tag>'4'</tag></item>
    </one-of>
    <one-of xml:lang='zh-SG'>
      <item> 一 <tag>'1'</tag></item>
      <item> 二 <tag>'2'</tag></item>
      <item> 三 <tag>'3'</tag></item>
      <item> 四 <tag>'4'</tag></item>
    </one-of>
  </rule>
</grammar>
```
User: I want to fly from London to Paris.
System: { depart: LHR , arrive: CDG }
W3C: Grammars (Limitations)

The W3C grammar formats work best when user responses are taken from a limited set or have a well characterized structure. Not all interactions fall into this category:

- **System:** What color are your walls?  
  **User:** I'd say they're not an aqua or a sky blue, but more of a turquoise.

- **History:** User has selected a date for a (golf) tee time but the application shows a lightning warning.  
  **User:** Change that to Friday if the weather is better.

- **System:** Animal control. How can we help?  
  **User:** I have a bat trapped in my car. The engine is still running and it's eating my lunch. Can you get a team here before I run out of gas?
W3C: Extensible Multi-Modal Annotation language

- [http://www.w3.org/TR/emma/](http://www.w3.org/TR/emma/)
  - Last Call Working Draft; moving toward Candidate Recommendation
  - Technical contributions from Avaya, AT&T, IBM, Nuance, Volantis, and others
  - Seven years mature; work started in 2000 as NLSML

- Provides rich annotation set for recognition or data collection using speech, tactile, keyboard, and other inputs

- Contains application data which may include strict schema validation and provide a data model
  - Helps catch invalid results that might result from unexpected inputs or constraint errors

- Supports derivation chains for multilevel semantic processing
Structured Result from Speech Grammar

```xml
<?xml version="1.0"?>
<emma:emma version="1.0" xmlns:emma="http://www.w3.org/2003/04/emma"
            xmlns="http://www.example.com/example">

<emma:grammar id="grammar1" href="http://www.example.com/simple_speech.grxml"/>

<emma:one-of id="nbest" emma:grammar-ref="grammar1" emma:mode="speech"
             emma:signal="http://www.example.com/recoding/1234.ulaw"
             emma:media-type="audio/x-ulpaw; rate:8000" emma:duration="375"
             emma:function="dialog">
  <emma:interpretation id="int1" emma:tokens="today" emma:confidence="0.821">
    <day>1</day>
    <month>9</month>
    <year>2006</year>
  </emma:interpretation>
  <emma:interpretation id="int2" emma:tokens="Tuesday" emma:confidence="0.282">
    <day>5</day>
    <month>9</month>
    <year>2006</year>
  </emma:interpretation>
</emma:one-of>
</emma:emma>
```
I want to fly from Boston to Dallas Tuesday
from Austin to Dallas Today
from Alden to Dulles

…hmm…

<table>
<thead>
<tr>
<th>Lattice Result Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>I want to fly</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<emma:lattice initial="1" final="6">
<emma:arc from="1" to="2" emma:confidence="1.0"><action>airline</action></emma:arc>
<emma:arc from="2" to="3" emma:confidence="0.7"><origin>BOS</origin></emma:arc>
<emma:arc from="2" to="3" emma:confidence="0.5"><origin>AUS</origin></emma:arc>
<emma:arc from="2" to="3" emma:confidence="0.1"><origin>IL05</origin></emma:arc>
<emma:arc from="3" to="4"/>
<emma:arc from="4" to="5" emma:confidence="0.6"><dest>DFW</dest></emma:arc>
<emma:arc from="4" to="5" emma:confidence="0.2"><dest>IUD</dest></emma:arc>
<emma:arc from="4" to="5" emma:confidence="0.1"><dest>DLH</dest></emma:arc>
<emma:arc from="5" to="6" emma:confidence="0.4">
<date><month>8</month><day>15</day><year>2006</year></date></emma:arc>
<emma:arc from="5" to="6" emma:confidence="0.3">
<date><month>8</month><day>10</day><year>2006</year></date></emma:arc>
</emma:lattice>
“How do I drive to here from here?”

EMMA: Simultaneous Input Example
User Interfaces for Mobile Speech

Jerry Carter, 11 May 2007
Directed Dialog

**Interrogator**: Your full name please.
**Prisoner**: William Robert Livingstone.
**Interrogator**: Your rank?
**Prisoner**: Sergeant.
**Interrogator**: And your serial number, please.
**Prisoner**: Three, nine, five, five, two, five, seven, five.

-- From an oral history of Stalag VII A
   http://www.moosburg.org/info/stalag/liveng.html
Dave Bowman: Open the pod bay doors, HAL.
HAL 9000: I'm sorry Dave, I'm afraid I can't do that.
Dave Bowman: What's the problem?
HAL 9000: I think you know what the problem is just as well as I do.
Dave Bowman: What are you talking about, HAL?
HAL 9000: This mission is too important for me to allow you to jeopardize it.

-- From the movie, 2001: A Space Odyssey
The Two Faces of Mobile Users

- **Right Brain**
  - Voice primary
  - User glances as screen episodically
  - Closer to the two-way paradigm of traditional dialog

- **Left Brain**
  - Visual primary
  - Data entry may use keyboard, touch screen, or voice
  - Audio prompts very rarely used

Users move back and forth between these modes
Surely Dates are Easy?
Just Filling in the Blank

- Users may speak to any field and in any order
- Responses may fill fields, correct entries, or do both

Grammars authors must anticipate a variety of potential interactions
The Curse of the Long Tail

- Grammar coverage gaps are the dominant source of errors today
- Noise is generally less significant

- NL dialog options are not easily enumerated
- 'Unexpected' responses often constitute a significant fraction
Looking Forward

Where We Are

• Solid specifications from the W3C
  – SSML, PLS, SRGS, SISR, EMMA for audio output and speech recognition
  – (X)HTML, SVG, CSS, … for document authoring targeted at the desktop
  – VoiceXML for single mode, speech dialogs

• Speech interfaces are needed to address input limitations of mobile devices

Where We're Going

• Increasingly powerful grammar formats able to extract semantics from rich natural language

• Authoring language: ???
  – C++ / Java
  – XHTML
  – Others…