Trends in Speech Standards

Daniel C. Burnett
Director of Speech Technologies and Standards

Mobile Voice 2011

Why do we have standards?

- Wall socket vs mobile device connectors
- USB chargers

Why do we have standards?

Why do we have standards?

Standards trends follow market trends

voxeo

- HTML for simple document creation, then self-service
- Back-ends adjusted to use “web model”
- Proprietary IVRs → Telephony/IVRs using web model
- Voice traffic over IP, web on phones

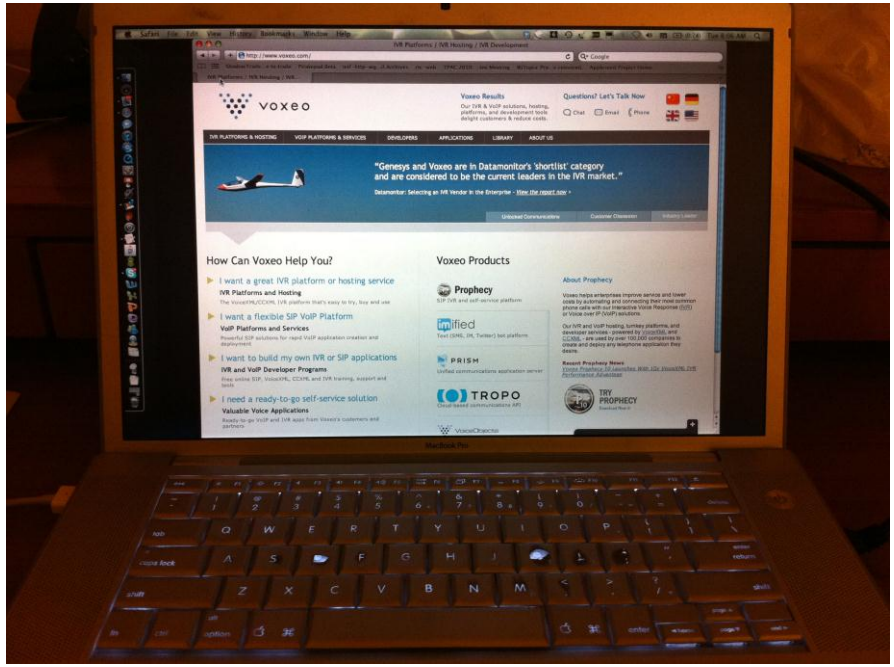
Speech Standards Pre-2000

- Voice standards are programming language/API-based
  - Java Speech API
  - Microsoft SAPI
  - JSML, VML, PML, VoxML, etc.

Speech Standards 2000-2010

- **IVR** standards come to the forefront, with speech processing as major component
- Author-level are XML-based (W3C)
  - VoiceXML (1)/2/2.1
  - SRGS 1
  - SSML 1/1.1
  - EMMA 1
- Implementer-level are protocol-based (IETF)
  - MRCP (1)/2

Speech Standards Today



- IVR standards incorporate visual web
- Visual web standards incorporate Voice (and maybe telephony?)

Speech Standards Today

- IVR standards better support visual web paradigm
  - VoiceXML 3
  - Related multimodal standards (EMMA 1.1, MMI Arch)
- Visual web standards incorporate Voice
  - HTML Speech Incubator Group
- Protocol-level enablers
  - RTC-Web (IETF, thus Internet-based)
  - Device access (W3C, also Internet-based)
  - 3GPP – telephony-based

VoiceXML 3

- 2.1 represents vast telephony speech experience, but not enough like HTML
- V3 Participants: Voxeo, Microsoft, Nuance, Alcatel, Loquendo, IBM, et al.
- V3 adds
  - Some new features, and
  - DOM event-based core
- Opens up to whole new world of web programmers

Example (v2 → v3)

```
<vxml version="2.1">

  <form id="main">
    <field …>
      What's your zip code?
      <grammar uri="builtin:digits"/>
      <noinput>
        <!-- Do my thing here -->
      </noinput>
    </field>
  </form>
</vxml>
```

Example (v2 → v3)

```
<vxml version="3.0" onload="load();">
  <script>
    function load() {
      var el = document.getElementById("main");
      el.addEventListener('noinput', doMyThing, false);
    }
    function doMyThing() {
      /*   Do my thing here */
    }
  </script>
  <form id="main">
    <field …>
      What's your zip code?
      <grammar uri="builtin:digits"/>
    </field>
  </form>
</vxml>
```

V3 and HTML together

```
<html onload="load();">
  <script>
    function load() {
      var el = document.getElementById("main");
      el.addEventListener('noinput', doMyThing, false);
    }
    function doMyThing() {/*   Do my thing here */}
  </script>
  <vxml version="3.0">
    <form id="main">
      <field …>
        What's your zip code?
        <grammar uri="builtin:digits"/>
      </field>
    </form>
  </vxml>
</html>
```

V3 new features

- <media> element w/ begin/end, volume, speed controls
- Real-time controls for volume, speed, perhaps others
- <transition> controllers for FIA control

HTML Speech XG

- Goal: add simple speech processing to HTML
- Participants: Voxeo, Microsoft, Google, Mozilla, AT&T, OpenStream, et al.
- Requirements gathering complete
- Proposals expected imminently
- Recommendations to HTML group in August

Top requirements

- Speech reco: barge-in, EMMA for results, processing control, language/grammars configuration
- Speech synthesis: completion notification, processing control
- Privacy/security: end-user consent required for audio capture

RTC-Web

- IETF group charter: protocols for interactive real-time voice/video/collaboration/gaming between browsers and other devices
- W3C group charter: access to and control of these protocols via web author interface, e.g., <audio> attributes

Device access

- Issues:
  - Microphone access
  - Barge-in audio cutoff
- Relevant groups
  - W3C DAP
  - W3C Audio
  - WHATWG (Device API proposal(s))
  - Likely others

Trends in Speech Standards

Daniel C. Burnett
Director of Speech Technologies and Standards

Mobile Voice 2011

MMI Architecture

- W3C Multimodal Interaction WG
- Loosely-coupled architecture for MM UIs
- Primarily standardized events for coordinating among modality components
- Working on interoperability
- Last Call WD expected shortly

EMMA 1.1

- W3C Mulitmodal Interaction WG
- EMMA 1.0 – XML format for describing human input (speech, pen, gesture, etc.)
- EMMA 1.1 – minor extensions and improvements over EMMA 1.0, better multiple modality integration
- First draft expected this quarter

Links

- V3 http://www.w3.org/Voice
- HTML Speech XG
  - http://www.w3.org/2005/Incubator/htmlspeech
- RTC-Web http://rtc-web.alvestrand.com/
- Device API http://www.w3.org/2009/dap/
- W3C Audio http://www.w3.org/2005/Incubator/audio/
- HTML 5 http://www.w3.org/html/wg/
- WHATWG http://www.whatwg.org/
- MMI, EMMA http://www.w3.org/2002/mmi/