New Directions for Voice and the Web

• W3C TPAC Breakout Session
• October 19 and 20, 2021
• Kaz Ashimura, W3C
• Deborah Dahl, Conversational Technologies
Agenda

• Review existing standards and requirements for voice and the web, presented at the AC meeting (Kaz)
  • Existing standards
  • Requirements for ensuring correct Text to Speech pronunciation
  • Interoperable Speech to Text and Text to Speech in the browser

• Interoperability of voice agents (Debbie)
  • Current voice agents are proprietary and are not based on web standards
  • How can we get them to interoperate?
  • How can they take advantage of existing Web infrastructure?
  • Voice Interaction Community Group

• Voice Workshop (Kaz)
Existing mechanisms for speech interfaces

• **VoiceXML 2.1 REC** (2007), **VoicXML 3.0 WD** (2010)
• **Speech Synthesis Markup Language (SSML) 1.1 REC** (2010)
• CSS Speech Module
  • **CSS Level 2 Aural Style Sheets W3C REC** (1998)
  • **CSS Level 3 Speech module Candidate REC** (2020)
• **Web Speech API** – not a W3C REC but a CG Report
• **Specification for Spoken Presentation in HTML WD** (2021)
Voice agents getting popular

However, emerging needs for improved voice agents:
• Accurate pronunciation and presentation
• flexible speech styles
• various voices
• conversational interactions
• etc.

Also...
Interoperability of voice agents

• Current voice agents are proprietary and are not based on web standards
  • Smart speakers, phone OS, desktop or mobile apps
• How can we get them to interoperate?
• How can they take advantage of existing Web infrastructure?
• Voice Interaction Community Group is working on interoperability – will meet next week
• https://www.w3.org/community/voiceinteraction/
Intelligent Personal Assistants and Web Pages—not so different!

The same

<table>
<thead>
<tr>
<th>Web Page</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helps users to find information, perform actions, learn, or be entertained</td>
<td>Helps users to find information, perform actions, learn, or be entertained</td>
</tr>
<tr>
<td>Communicates with servers over the Internet</td>
<td>Communicates with servers over the Internet</td>
</tr>
</tbody>
</table>

Intrinsic differences

<table>
<thead>
<tr>
<th>Web Page</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>User interaction mainly based on graphical interface</td>
<td>User interaction mainly based on voice and natural language</td>
</tr>
</tbody>
</table>

Arbitrary differences

<table>
<thead>
<tr>
<th>Web Page</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rendered in a browser</td>
<td>Rendered with a proprietary platform (smart speaker, app, or phone OS)</td>
</tr>
<tr>
<td>Applications run on any browser</td>
<td>Applications only run on their own platform</td>
</tr>
<tr>
<td>Found with a URL and/or search engine</td>
<td>Found through the proprietary platform</td>
</tr>
</tbody>
</table>
Potential voice workshop

We'd like to organize a W3C workshop to:

1. summarize the current status of the existing voice agents
2. see the potential pain points for global deployment, e.g., interoperability of the multiple platforms
3. then discuss how we could solve the issues, e.g., by providing a forum for joint discussion by related stakeholders

⇒ See also the initial proposal on GitHub 😊

https://github.com/w3c/strategy/issues/221
Feedback from the 1st breakout

• Existing standards within W3C
  • Specifications including recent work like natural language interface requirements
• The other related technologies, e.g., speaker recognition
  • Any value to including that in a standard?
• Pain points in a11y
  • What would be valuable to do in voice?
• Other topics
  • Emotions expressed in text or speech
  • Common sense database related to people's perception rate and user experience

Further comments welcome:
➢ on the GitHub issue at: https://github.com/w3c/strategy/issues/221
➢ or email at <ashimura@w3.org>