Pilot Deployment of HTML5 Video Descriptions

Masatomo Kobayashi
Hironobu Takagi
Kentarou Fukuda
Reiko Nagatsuma
(IBM Research – Tokyo)
## Video Accessibility Status in Japan

<table>
<thead>
<tr>
<th></th>
<th>Captions</th>
<th>Audio Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>5.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Movies</td>
<td>12.0%</td>
<td>0.9%</td>
</tr>
<tr>
<td>TV (Private)</td>
<td>43.8%</td>
<td>0.6%</td>
</tr>
<tr>
<td>TV (Public)</td>
<td>56.2%</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

*1: Our informal survey (2009)  
*2: Media Access Support Center (2008)  
*3: Ministry of Internal Affair and Communication (2010)
Problems: Workload, Skills and Cost

Captions

Audio Descriptions

Recording

Describing

Narrating calls for an experienced narrator and recording equipments

Scriptwriting calls for a special expertise to better describe the scenes
Approach: Text-to-Speech
Problems: Workload, Skills and Cost

Captions

Audio Descriptions

Recording

Describing

Speech synthesis reduce the cost

Cost

Transcribing
Advantage of Text-to-Speech

- No recording
- Previewing the results
- Voice quality
Long-tail Videos

Possible Cost

$100,000,000  Hollywood

TV programs

Internet Videos

Video Contents

$0
Project History

Feasibility Studies

- Online survey (U.S., Feb 2011; with WGBH)
- F2F survey (Japan, Feb 2011)
- F2F survey (Japan, Nov 2010)
- Focus group sessions (U.S., Mar 2010; with WGBH)
- Online survey (U.S., Mar 2010; with WGBH)
- In-depth studies (Japan, Feb 2010)
- F2F survey (Japan, Nov 2009)
- Informal interviews (Japan, Sep 2009)

Tool Development

- To be updated (March 2012)
- Updated (October 2011)
- User evaluation (Japan, Mar 2011)
- Professional evaluation (July-August 2010)
- Open-sourced (v0.1.0 Apr 2010)
- User evaluation (Japan, Mar 2010)
- Development started (Sep 2009)

* The results of 2009-2010 studies are reported in:
Feasibility Study Results
Synthesized audio descriptions are acceptable

How was the listening experience?

Education
Information
Entertainment
Information

Comfortable
Acceptable
Neutral
Slightly uncomfortable
Uncomfortable

Online survey in the U.S. (Mar 2010)
236 respondents

F2F survey in Japan (Nov 2010)
120 respondents
Extended descriptions improve the comprehension

Comprehension Rates

In-depth study in Japan (Feb 2010)
24 participants

*Extended descriptions*: pause the playback for presenting a long description
Novices can describe a video (if extended descriptions are allowed)

Subjective Effectiveness

Score

Novice (Normal)  Novice (Extended)  Expert (Normal)  Expert (Extended)

In-depth study in Japan (Feb 2010)
24 participants
Authoring Tool
Authoring Tool | Script Editor

http://www.eclipse.org/actf/downloads/tools/ScriptEditor/
Hiroshima City

www.hiroshima-navi.or.jp
Hiroshima Movie Channel

Videos for culture introduction, public announcement, etc.
## Target Videos

<table>
<thead>
<tr>
<th>#</th>
<th>Content</th>
<th>Language</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sightseeing Guide</td>
<td>English</td>
<td>4 m 06 s</td>
</tr>
<tr>
<td>2</td>
<td>Sightseeing Guide</td>
<td>Japanese</td>
<td>4 m 18 s</td>
</tr>
<tr>
<td>3</td>
<td>Report (<em>Educational Event</em>)</td>
<td>Japanese</td>
<td>2 m 09 s</td>
</tr>
<tr>
<td>4</td>
<td>Culture (<em>Seafood</em>)</td>
<td>Japanese</td>
<td>5 m 21 s</td>
</tr>
<tr>
<td>5</td>
<td>Culture (<em>Japanese Pickles</em>)</td>
<td>Japanese</td>
<td>4 m 22 s</td>
</tr>
<tr>
<td>6</td>
<td>Culture (<em>Craftwork</em>)</td>
<td>Japanese</td>
<td>4 m 43 s</td>
</tr>
<tr>
<td>7</td>
<td>Culture (<em>Food</em>)</td>
<td>Japanese</td>
<td>4 m 54 s</td>
</tr>
<tr>
<td>8</td>
<td>Culture (<em>Food</em>)</td>
<td>Japanese</td>
<td>3 m 56 s</td>
</tr>
<tr>
<td>9</td>
<td>Culture (<em>Japanese Sake</em>)</td>
<td>Japanese</td>
<td>1 m 18 s</td>
</tr>
</tbody>
</table>
## Who Describes Videos?

<table>
<thead>
<tr>
<th>Role</th>
<th>Quality</th>
<th>Cost</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Creator</td>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Professional Descriptor</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>External Volunteers</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>
Describing Videos

15 adult volunteers described videos using Script Editor
Assessments by Volunteers | Difficulties

Subjective difficulties for each subtask

Easy  Normal  Difficult

Decide timings
Write text
Preview
Revise timings
Revise text
Final check
Opinions of Volunteers

What will encourage your volunteer work?

- **User feedback** will motivate me to describe videos: 100% (14/14)

- Wants to compare my descriptions with others’ work: 79% (11/14)

- Wants my descriptions to be corrected by experts: 71% (10/14)

- Wants to describe videos at home: 64% (9/14)
Steps towards Delivery

1. Volunteers
   Make audio descriptions

2. Expert users
   (People with visual impairments)
   Review descriptions

3. Expert describer
   Revises descriptions

4. Hiroshima City
   Uploads described videos
How to deliver
HTML5 supports **text-based** audio descriptions

HTML5 supports **synchronized** audio tracks

---

**Text-based**

```
<video src="foo.ogv">
  <track kind="descriptions" src="bar.vtt" ...>
</video>
```

**Synchronized**

```
<video src="foo.ogv" mediagroup='baz'>
  ...
</video>
```

---

http://dev.w3.org/html5/spec/media-elements.html
Text Format (WebVTT and TTML)

WebVTT

```
00:00:00.015 --> 00:00:02.035 A:start D:vertical
Smarter Energy, Wind Power and Electric Vehicles. Born
00:00:03.070 --> 00:00:06.017 A:start D:vertical
Let's build a smarter planet. IBM.
00:00:07.011 --> 00:00:11.024 A:start D:vertical
Tall wind turbines are rotating in bright sunlight with a cle
00:00:13.007 --> 00:00:17.029 A:start D:vertical
Anders Quitzau, Innovation Executive, IBM Denmark.
```

Expected to be a primary format for HTML5

TTML

```
<p begin="00:00:00:15" dur="00:00:02:20" tvd:extend="Energy, Wind Power and Electric Vehicles. Born<br/>
Let's build a smarter planet. IBM.\n</p>
<p begin="00:00:07:11" end="00:00:11:24">Tall wind<br/>
in bright sunlight with a clear blue sky.\n</p>
<p begin="00:00:13:07" dur="00:00:04:22">Anders Q<br/>
uitzau, Innovation Executive, IBM Denmark.\n</p>
<p begin="00:00:25:27" tvd:extended="true" dur="00:1<br/>
Europe. Zooming into a small island in the Baltic S<br/>
</p>
```

Easy to extend to support extended descriptions, etc.
Delivery Methods

Text Track + Client-side Syntheses
- One video
- One text
- Quality: Medium
- Network Load: Low
- Flexibility: High or Low

Text Track + Pre-recorded Syntheses
- One video
- Text + audio fragments
- Quality: High
- Network Load: High
- Flexibility: Medium

Synchronized Audio Track
- One video
- One audio
- Quality: High
- Network Load: High
- Flexibility: Low
* No extended descriptions
### HTML5 Video Player in Practice

#### 3 JavaScript layers

<table>
<thead>
<tr>
<th>vd-player.js</th>
<th>Supports <strong>extended descriptions</strong> and <strong>pre-recorded audio fragments</strong> and provides <strong>playback control buttons</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>vd-compat.js</td>
<td><strong>Fills the gap</strong> between the <strong>HTML5 spec</strong> and the <strong>implementation</strong> of the latest Firefox</td>
</tr>
<tr>
<td>vd-compat-ie.js</td>
<td><strong>Fills the gap</strong> in implementations between the latest <strong>Firefox</strong> and old <strong>Internet Explorers</strong> (IE 8 or older)</td>
</tr>
</tbody>
</table>
Pilot Deployment

Announced to a local community of people with visual impairments

http://www.city.hiroshima.lg.jp/riyou/movie/hiroshima_tourism.html (client-side TTS)
http://www.city.hiroshima.lg.jp/riyou/movie/hiroshima_tourism_p.html (pre-recorded)

* experimental pages: subject to change

See also: WGBH NCAM’s experiment with HTML5 text-based audio descriptions:
http://ncamftp.wgbh.org/ibm/dvs/
User Feedback | Helpful or Uncomfortable?

11 respondents (8 used pre-recorded only, 3 used both)

**Helpful?**
- Yes: 8
- Neutral: 2
- No: 1

**Uncomfortable?**
- Yes: 5
- Neutral: 2
- No: 4

Generally helpful for understanding videos
Sometimes uncomfortable
Needs a “brief or explicit” switch
User Feedback | Easy to Control?

11 respondents (8 used pre-recorded only, 3 used both)

Both client-side TTS and pre-recorded were easy to control.
User Feedback | Other Comments

[Screen Reader vs Pre-recorded]
- Screen reader: 2
- Pre-recorded: 1

[Novice Describers]
- Some redundant or inconsistent descriptions
- Generally sufficient and helpful

[Description Placement]
- People with low vision preferred descriptions strictly synchronized with visual events on the screen
- People who are blind requested descriptions during a sufficient pause of original narrations
User Feedback | Other Comments

[Description Voice]
- TTS (female) voice was sometimes confusing for videos originally narrated by a female

[Volume Control]
- Volume up/down operations were confusing (changes video's or descriptions' volume?)

[Shortcut Keys]
- Shortcut keys were helpful
- Some keys were conflicted with keys used by some screen readers
Summary

Tested HTML5 videos with text-based descriptions in a practical context.

On the HTML5 spec, text-based descriptions are easy to add.

Some extra features are necessary to make better descriptions:
- extensions of browsers, extensions of text formats, JavaScript, ...
- extended descriptions, pre-recorded audio fragments, ...

By the end of March:
- The city’s HTML5 video pages are planned to be publicly announced.
- The updated Script Editor and player scripts are planned to be released.
Acknowledgements

Hiroshima City
Voice Information Center (Volunteer Circle @ Hiroshima)
WGBH - NCAM
All participants in the experiments

* This research is partly funded by
  National Institute of Information and Communications Technology (NICT), Japan
Contact

Masatomo Kobayashi
IBM Research – Tokyo
mstm@jp.ibm.com
http://www.research.ibm.com/trl/people/mstm/