Unifying Video Captions and Text-Based Audio Descriptions

Masatomo Kobayashi (IBM Research – Tokyo)
Online Videos Support Our Lives

News
Entertainment
Education
Announcements
Historical Archive
Home
How Accessible are Current Online Videos?
Percentage of Described Online Videos

0%

Percentage of Captioned Videos: 5%

* Based on our informal survey of the websites of government agencies and local governments in Japan (2009)
## Status of Captions and Audio Descriptions

<table>
<thead>
<tr>
<th></th>
<th>Captions</th>
<th>Audio Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>5.0% *1</td>
<td>0.0% *1</td>
</tr>
<tr>
<td>Movies</td>
<td>12.0% *4</td>
<td>0.9% *4</td>
</tr>
<tr>
<td>TV (Private)</td>
<td>42.3% *2</td>
<td>0.4% *3</td>
</tr>
<tr>
<td>TV (Public)</td>
<td>49.4% *2</td>
<td>5.6% *3</td>
</tr>
</tbody>
</table>

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

Captions can be provided using automatic transcription programs and is lower cost than video description.

Audio descriptions are more expensive because they require special expertise for scriptwriting to better describe the scenes. Narrating must be done by an experienced narrator using special recording equipment.
Solution: Using the Power of Community

The usual method of providing captions or video description is for the video owner to do it.

Using the power of community, we can enlist external volunteers and experts to describe and caption videos.
cf. Using the Power of Community for Website Accessibility Improvements

Social Accessibility Project (2008-)

1. **Encounter an inaccessible webpage**
   - Visually-Impaired Users

2. **Report a problem on the webpage**
   - Social Accessibility Server

3. **Browse the reported problems on webpages**
   - Sighted Volunteers

4. **Fix the problem on the webpage (make accessibility metadata)**
   - Social Accessibility Server

5. **Browse the accessible webpage fixed with the accessibility metadata**
   - Visually-Impaired Users
Solution: Using Text-to-Speech (TTS)

Use text to speech to create transcriptions of audio content. Allow people to preview the result.
Approach: Unifying Captions & Audio Descriptions

- Development of **easy-to-use authoring tools** to allow novices to caption and describe videos
- Development of a **common, standard platform** to allow collaborations for captioning and describing videos
- **Evaluation** of the **acceptance, effectiveness and characteristics** of synthesized audio descriptions
- Produce **common authoring interfaces and distribution formats**
Evaluations

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

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

**How was the listening experience?**

- **Comfortable**
- **Acceptable**
- Neutral
- Slightly uncomfortable
- Uncomfortable

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

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

![Comprehension Rates](image)

*In-depth study in Japan (Feb 2010)*
24 participants
Result: Novices can describe a video (with the support of an authoring tool and the use of extended descriptions)

Subjective Effectiveness

In-depth study in Japan (Feb 2010)
24 participants
Latest Results: Design Implications

Support extended descriptions
• Benefit to the understandability
• Allow novice describers to make effective audio descriptions

Provide easy-to-use interfaces for customization
• For both describers and listeners
• Parameters: level of descriptions, voices, rate, volume, ...

Support step-by-step quality improvement
• Synthesized audio descriptions of minimal quality can be provided quickly
• Rewriting the script, using better voices, etc. can improve the quality
Latest Results: Authoring Tools – ScriptEditor v1
Authoring Tools – ScriptEditor v1

- Interactive visual interface
- Previewing the resulting narrations
Feedback from Practitioners

- Interview sessions with 3 organizations (July – August 2010)
  - City Lights (Volunteer Organization)
  - Cinema Access Partners (NPO)
  - Media Access Support Center (NPO)

- 50+ requirements from the sessions
  - Enhancements of user interface: 20
    - support both of captions and audio descriptions in a single list
    - add “speaker” column to support captions and scripts
    - improve usability
  - Enhancements for input/output support for script edit: 11
    - enable to mark time period of caption, audio description, etc. while watching the video
    - expand coverage of file formats (text, csv, srt, tttml, etc.)
    - enable to reuse proper names (speaker name, etc.)
  - Requests to support diffusion of audio description: 21
    - consider to use the tool for education of audio description
    - add audio description check function against guidelines
    - support social editing (share audio description, captions, scripts, proper names, etc. with others)
  - Others: 5
    - enable to record audio description
Authoring Tools – ScriptEditor v2
Authoring Tools – ScriptEditor v2

- Side-by-side editing of the original scenario, captions and descriptions
- Improvements based on the feedbacks from practitioners
Platform

Online Videos

Website

Audio Description / Caption Scripts

Script Repository

Refer

Post

Fetch

Browse

Authoring Tool

Player / Browser

http://www.eclipse.org/actf/downloads/

```html
<video src="foo.ogg">
  <track kind="subtitles" src="bar.srt" ...>
  <track kind="captions" src="baz.html" ...>
  <track kind="descriptions" src="qux.srt" ...>
</video>
```

The `<track>` element allows authors to specify explicit external timebase captions and audio descriptions.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtitles</td>
<td>Subtitles</td>
<td>Transcription or translation of the content</td>
</tr>
<tr>
<td>captions</td>
<td>Captions</td>
<td>Transcription or translation of the content</td>
</tr>
<tr>
<td>descriptions</td>
<td>Descriptions</td>
<td>Textual descriptions of the content</td>
</tr>
<tr>
<td>chapters</td>
<td>Chapters</td>
<td>Chapter titles, intended to be used</td>
</tr>
<tr>
<td>metadata</td>
<td>Metadata</td>
<td>Tracks intended for use from script</td>
</tr>
</tbody>
</table>
Text Format (WebVTT, TTML, etc.)

Basic parameters:
Start time, End time (Duration), Text

Caption-specific parameters:
Layout, Font size, Font color, etc.

Description-specific parameters:
Speed, Volume, “Extended” flag, etc.

WebVTT

00:00:00,015 --> 00:00:02,035 A:start D:vertical
Smarter Energy. Wind Power and Electric Vehicles. Bornholm, De
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

TTML

<p begin='00:00:00:15' dur='00:00:02:20' actftvd:extended='true' actftvd:speed='80'>Smarter Energy. Wind Power and Electric Vehicles. Bornholm, De</p>
<p begin='00:00:03:07' end='00:00:06:17' actftvd:speed='80'>Let’s build a smarter planet. IBM.</p>
<p begin='00:00:07:11' end='00:00:11:24' actftvd:speed='80'>Tall wind turbines are rotating in bright sunlight with a clear blue sky. </p>
<p begin='00:00:13:07' dur='00:00:04:22' actftvd:speed='80'>Anders Quitzau, Innovation Executive, IBM Denmark</p>
Proof-of-Concept Demo | HTML5+JavaScript
## Distribution Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Narration Quality</th>
<th>Authoring Cost</th>
<th>Network Cost</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional model – human narrator</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>hears audio on the server and provides</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>audio to a user</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-recorded synthesized audio – uses a</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>TTS to translate audio on the server and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>provide audio to a user</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server-side synthesizer - text on a server</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>is funneled through a TTS to provide audio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to a user</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client-side synthesizer take text from</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High or Low</td>
</tr>
<tr>
<td>the server, funnels the text through the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTS on the client side and provides audio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>directly from the TTS to a user</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Workshop on Video Accessibility - Dec 2010
Related Activities

**VDManager/VDPlayer** (CRIM) and **CapScribe** (Inclusive Media and Design) using TTS-base audio descriptions on pre-recorded synthesized audio.

**LiveDescribe** (Ryerson University) using the power of community on human voice.
Summary

Authoring Tool

HTML-based Player

Feasibility Studies

Sample Video

Sample Video

Graph showing feasibility studies: 100% Education, 80% Information, 60% Entertainment, 40% Information, 20% Education, 0% Entertainment.
Acknowledgements

WGBH Educational Foundation
Japan Broadcasting Corporation (NHK)
Cabinet Office, Government of Japan
Cocoon Co., Ltd.
Orange Page Inc.

WGBH – NCAM
Japan Braille Library
Cinema Access Partners (CAP)
City Lights
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 (at) jp.ibm.com
http://www.research.ibm.com/trl/people/mstm/