H.761 Support of a News <input> Node and a New "recognition" Node-Event to Enable Multimodal User Interactions

Alan L. V. Guedes
TeleMídia - PUC-Rio
alan@telemidia.puc-rio.br

Sergio Colcher
Informatics Departament - PUC-Rio
colcher@inf.puc-rio.br

ABSTRACT

Multimedia languages traditionally, they focus on synchronizing a multimedia presentation (based on media and time abstractions) and on supporting user interactions for a single user, usually limited to keyboard and mouse input. Recent advances in recognition technologies, however, have given rise to a new class of multimodal user interfaces (MUIs). In short, MUIs process two or more combined user input modalities (e.g. speech, pen, touch, gesture, gaze, and head and body movements) in a coordinated manner with output modalities . An individual input modality corresponds to a specific type of user-generated information captured by input devices (e.g. speech, pen) or sensors (e.g. motion sensor). An individual output modality corresponds to user-consumed information through stimuli captured by human senses. The computer system produces those stimuli through audiovisual or actuation devices (e.g. tactile feedback). In this proposal, we aim at extending the NCL multimedia language to take advantage of multimodal features.

KEYWORDS

NCL, Ginga, Multimodal Interactions, SSML, SRGS

1 BACKGROUND

H,761[2] traditionally focus on synchronizing a multimedia presentation (based on media and time abstractions) and on supporting user interactions for a single user, usually limited to keyboard and mouse input. Recent advances in recognition technologies, however, have given rise to a new class of multimodal user interfaces (MUIs). In short, MUIs process two or more combined user input modalities (e.g. speech, pen, touch, gesture, gaze, and head and body movements) in a coordinated manner with output modalities. An individual input modality corresponds to a specific type of user-generated information captured by input devices (e.g. speech, pen) or sensors (e.g. motion sensor). An individual output modality corresponds to user-consumed information through stimuli captured by human senses. The computer system produces those stimuli through audiovisual or actuation devices (e.g. tactile feedback). In this proposal, we aim at extending the NCL multimedia language to take advantage of multimodal features.

2 PROPOSAL

For the representation of new output modalities, we propose new support for a TTS (Text-To-Speech) content using the W3C SSML[1].

[1]http://w3.org/TR/speech-synthesis11/

In: Future of Interactive Television Workshop (V WTVDI), Rio de Janeiro, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2019.
ISSN 2596-1683

For the representation of input modalities, we propose the new `<input>` element, which is a type of NCL Node and can be used in Link relationships. The *src* of a `<input>` is also a collection of information.

Different from the <media>, however, the `<input>` information is expected to be captured, not presented. Some examples of `<input>` contents include: W3C SRGS[2] used for speech recognition, such as recognizing words and phrases spoken by the user(s); W3C InkML[3] used for pen writing ("ink") recognitions. Since `<input>` is indeed a specialization of NCL Node, it is also possible to define <area> in it. This <area> element specifies a portion of the recognition content. For instance, an anchor may refer to expected speech tokens defined in an SRGS file. This <area> is associated to a "recognition" event. The "recognition" event indicates that the system has recognized the expected information defined in a `<input>`. This proposal is detailed in the Guedes thesis "Extending multimedia languages to support multimodal user interactions"[1].

3 USE CASE

To illustrate our proposal, we introduce a scenario called "Multimodal Sightseeing of Today". In this scenario, during some time window opportunity during a video presentation, the application asks the user to interact via voice commands and choose which touristic place to visit next (i.e., which video object to play next). More precisely, if the user says the word "downtown", the downtown video is started; otherwise, if the user says "beach", then the beach video is started. The following codes listing below are descriptions used in the following instantiations. The first two show SRGS and SSML files.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<grammar>
 <rule id="downtown">downtown</rule>
 <rule id="beach">beach</rule>
</grammar>
```

Listing 1: SRGS code fragment .

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<speak>
 <s id="downtownOrBeach">Do you want to visit Rio de Janeiro's
    downtown or the Copacabana beach?</s>
</speak>
```

Listing 2: SSML code fragment.

The following NCL code illustrates the "Multimodal Sightseeing of Today".

[2]http://w3.org/TR/speech-grammar/
[3]https://www.w3.org/TR/InkML/

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<ncl xmlns="http://www.ncl.org.br/NCL3.0/EDTVProfile">
 <head>
   <connectorBase>
     <causalConnector id="onRecognizeStart">
       <simpleCondition role="onRecognize"/>
       <simpleAction role="start"/>
     </causalConnector>
   </connectorBase>
 </head>
 <body>
   <port id="p1" component="intro">
   <media id="intro" src="intro.mp4">
     <area id="credits" begin="30s"/>
   </media>
   <media id="videoBeach" src="videoBeach.mp4">
   <media id="videoDowntown" src="videoDowntown.mp4">
   <media id="choiceQuestion" src="choiceQuestion.ssml"/>
   <input id="voiceRec" src="voiceRec.srgs"/>
   <link xconnector="onBeginStart">
     <bind role="onBegin" component="intro" interface="credits"/>
     <bind role="start" component="choiceQuestion"
interface="downtownOrBeach"/>
     <bind role="start" component="voiceRec"/>
   </link>
   <link xconnector="onRecognizeStart">
       <bind role="onRecognize" component="voiceRec" interface="
         downtown"/>
       <bind role="stop" component="intro"/>
       <bind role="stop" component="voiceRec"/>
       <bind role="start" component="videoBeach"/>
   </link>
   <link xconnector="onRecognizeStart">
     <bind role="onRecognize" component="voiceRec" interface="beach"
       />
     <bind role="stop" component="intro"/>
     <bind role="stop" component="voiceRec"/>
     <bind role="start" component="videoDowntown"/>
   </link>
 </body>
</ncl>
```

Listing 3: NCL code fragment using the proposed approach.

REFERENCES

[1] Álan Lívio Vasconcelos Guedes, Roberto Gerson de Albuquerque Azevedo, and Simone Diniz Junqueira Barbosa. [n. d.]. Extending Multimedia Languages to Support Multimodal User Interactions. 76, 4 ([n. d.]), 5691–5720. https://doi.org/10.1007/s11042-016-3846-8 00003.

[2] ITU. 2009. *H.761: Nested Context Language (NCL) and Ginga-NCL for IPTV Services.* Technical Report. ITU, Geneva, Switzerland. https://www.itu.int/rec/T-REC-H.761