A Browser-based Multimodal Interaction System

Kouichi Katsurada
Toyohashi Univ. of Tech.
1-1 Hibarigaoka, Tempaku-cho,
Toyohashi 441-8580,JAPAN
+81-532-44-6884
katurada@tutkie.tut.ac.jp

Teruki Kirihata
Toyohashi Univ. of Tech.
1-1 Hibarigaoka, Tempaku-cho,
Toyohashi 441-8580,JAPAN
+81-532-44-6884
kirihata@vox.tutkie.tut.ac.jp

Masashi Kudo
Toyohashi Univ. of Tech.
1-1 Hibarigaoka, Tempaku-cho,
Toyohashi 441-8580,JAPAN
+81-532-44-6884
kudo@vox.tutkie.tut.ac.jp

Junki Takada
Toyohashi Univ. of Tech.
1-1 Hibarigaoka, Tempaku-cho,
Toyohashi 441-8580,JAPAN
+81-532-44-6884
takada@vox.tutkie.tut.ac.jp

Tsuneo Nitta
Toyohashi Univ. of Tech.
1-1 Hibarigaoka, Tempaku-cho,
Toyohashi 441-8580,JAPAN
+81-532-44-6884
nitta@tutkie.tut.ac.jp

ABSTRACT

In this paper, we propose a system that enables users to have multimodal interactions (MMI) with an anthropomorphic agent via a web browser. By using the system, a user can interact simply by accessing a web site from his/her web browser. A notable characteristic of the system is that the anthropomorphic agent is synthesized from a photograph of a real human face. This makes it possible to construct a web site whose owner's facial agent speaks with visitors to the site. This paper describes the structure of the system and provides a screen shot.

Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User Interfaces – *Voice I/O.*

General Terms

Design.

Keywords

Multimodal interaction system, web-based system.

1. INTRODUCTION

Many multimodal interaction (MMI) systems have been proposed [1][2][3]. Although these systems resulted in significant outcomes regarding such things as system architecture and authoring, not many are widely used as human-computer interfaces. One reason for this is complexity of installation, compilation, and so on, to use the system. To avoid this, we designed a web browser-based MMI system. The system enables users to interact with an anthropomorphic agent simply by accessing a web site via a common web browser. A notable characteristic of the system is that the agent is synthesized from a photograph of a real human face. Therefore a web site owner can construct a site in which his/her facial agent speaks with visitors to the site. In the following, we outline the structure of the MMI system as well as provide a screen shot.

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.
ACM 978-1-60558-198-9/08/10.

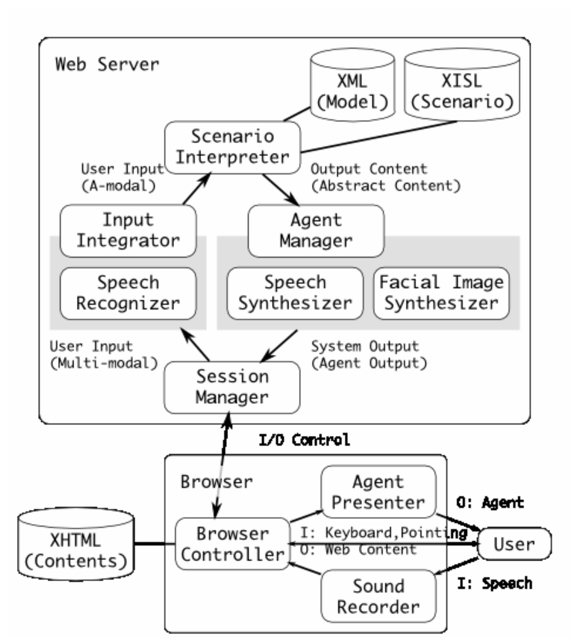

Figure 1. Structure of multimodal interaction system.

2. SYSTEM STRUCTURE

Since a web browser does not have enough computing power, we divided the system into two components: a server component and a browser component. Figure 1 shows the structure of the system. The server component, coded using Java language, is the main component that controls interaction flow, speech recognition, speech synthesis, and facial image synthesis. Meanwhile, the browser component, coded using JavaScript, merely plays a synthesized facial Flash movie, records speech input, and handles a web page. These two components communicate with each other using AJAX technology. The speech recognition engine (Julius [4]), the speech synthesis engine (gtalk [5]), the facial image synthesis engine (FSM [6]), and the interaction manager are deliverables from the Japanese research project that developed anthropomorphic spoken dialogue agents (the Galatea

```
<xisl  version="1.0">
  <body>
    <dialog  id="order">
      <exchange>
        <prompt>
          <output  type="agent"  event="speech">
            <![CDATA[ <param  name="text">
                          Please select an item.
                      </param>      ]]>
          </output>
        </prompt>
        <operation  comb="alt">
          <input  type="speech"  event="recognize"
                  match="./grammar.txt#items"
                  return="item"/>
        </operation>
        <action>
                  :
```

Figure 2. A fragment of XISL description.

Project [7]), and the later ISTC (Interactive Speech Technology Consortium [8][9]).

Here, we explain the flow of interaction. First, the system accepts a user's inputs. If a user's input consists of speech, it is recorded by the sound recorder. The browser controller then sends the inputs (speech, pointing, and keyboard) to the session manager on a web server. The speech input is recognized by Julius at the input integrator. After input integration, the inputs are sent to the scenario interpreter. The scenario interpreter manages dialogue flow based on scenario description written in XISL (eXtensible Interaction Scenario Language [10]). System outputs are generated by gtalk and FSM at the agent manager. They are then sent to the web browser through the session manager, and are played on the web browser.

Figure 2 shows a fragment of XISL description and Figure 3 is a screen shot of the system.

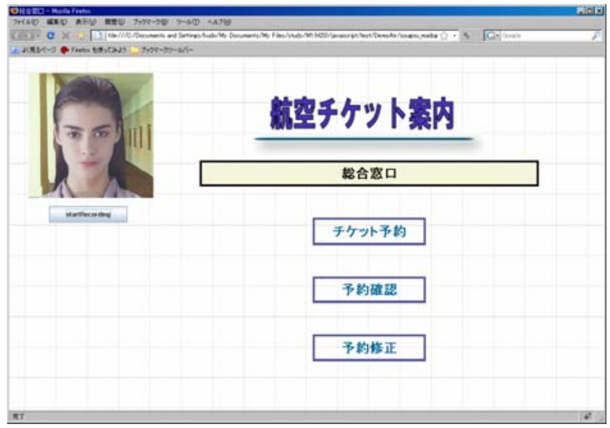

Figure 3. A screen shot of the browser-based MMI system.

3. CONCLUSIONS

This paper discussed a web browser-based MMI system. An advantage of the system is that it can be executed on any type of web browser that can handle JavaScript, Java applets, and Flash. This means that the system can be executed not only on a PC but also on a PDA, smart phone, etc. We believe this characteristic will help boost the use of multimodal interaction systems by the average web user.

4. ACKNOWLEDGMENTS

This work was supported by the Hori Information Science Promotion Foundation, the Telecommunications Advancement Foundation, and the Grant-in-Aid for Young Scientists (B) 20700156 2008 from MEXT - Japan.

5. REFERENCES

[1] N. Reithinger, et al., "SmartKom - Adaptive and Flexible Multimodal Access to Multiple Applications", Proc. of ICMI'03, pp.101-108 (2003).

[2] M. Johnston, et al., "MATCH: An architecture for multimodal dialogue systems", Proc. of the Annual Meeting of the Association for Computational Linguistics, pp.376-383 (2002).

[3] D. Gibbon, et al. (ed.), "Handbook of Multimodal and Spoken Dialogue Systems", Kluwer Academic Publishers, (2000).

[4] T. Kawahara, et al., "Sharable software repository for Japanese large vocabulary continuous speech recognition", Proc. ICSLP'98, pp.3257-3260 (1998).

[5] T. Yoshimura, et al., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", Proc. EUROSPEECH'99, pp.2347-2350 (1999).

[6] T. Yotsukura, et al., "An open source development tool for anthropomorphic dialog agent -face image synthesis and lip synchronization-", Proc. IEEE Fifth Workshop on Multimedia Signal Processing, 03_01_05.pdf (2002).

[7] S. Kawamoto, et al., "Galatea: Open-source software for developing anthropomorphic spoken dialog agents", in Life-Like Characters, ed. H. Prendinger and M. Ishizuka, pp.187-212, Springer-Verlag (2004).

[8] http://www.astem.or.jp/istc/index_e.html

[9] T. Nitta, et al., "Activities of interactive speech technology consortium (ISTC) targeting open software development for MMI systems", Proc. RO-MAN'04, 2B4 (2004).

[10] K. Katsurada, et al., "XISL: A language for describing multimodal interaction scenarios", Proc. ICMI'03, pp.281-284, (2003).