Anthropomorphic Agent as an Integrating Platform of Audio-Visual Information

Shigeki Sagayama, Takuya Nishimoto

Graduate School of Information Science and Technology, The University of Tokyo
Hongo, Bunkyo-ku, Tokyo 113-8656 Japan / {sagayama,nishi}@hil.t.u-tokyo.ac.jp

1 Introduction

In integration of audio-visual information input from sensors and control output to actuators and robotic systems, anthropomorphic spoken-dialog agent can be a good platform combining them under a unified concept such as “virtual human.” This talk will focus on an anthropomorphic spoken-dialog agent and its future possibility as a integrating platform.

2 Spoken Dialog Agent

2.1 Galatea Toolkit

One of ultimate human-machine interfaces is anthropomorphic spoken dialog agent which behaves like humans with facial animation and gesture and make speech conversations with humans. Among numerous efforts devoted for such a goal, Galatea Project conducted by 17 members from 12 universities is developing an open-source license-free software toolkit [1] for building an anthropomorphic spoken dialog agent under a financial support from IPA during fiscal years of 2000–2002. The authors are members of the project. The features of the toolkit are as follows: (1) high customizability in text-to-speech synthesis, realistic face animation synthesis, and speech recognition, (2) basic functions to achieve incremental (on-the-fly) speech recognition, (3) mechanism for “lip synchronization”; synchronization between audio speech and lip image motion, (4) “virtual machine” architecture to achieve transparency in module to module communication. The Galatea Toolkit for UNIX/Linux and Windows operating systems will be publicly available from August 22, 2003, at http://hil.t.u-tokyo.ac.jp/~galatea/.

2.2 Toolkit Components

The Galatea Toolkit consists of five functional modules: speech recognizer, speech synthesizer, facial animation synthesizer, agent manager which works as an inter-module communication manager, and task (dialog) manager. Fig. 2.2.1 shows the basic module architecture of the Galatea toolkit. The Galatea Project members newly created these components or modified existing components of their own or publicly available. The outline of some of these functional modules are stated below.

2.2.1 Common features

Galatea employs model-based speech and facial animation synthesizers whose model parameters are adapted easily to those for an existing person if his/her training data is given. Synthesized facial images and voices are customizabile easily depending on the purposes and applications of the toolkit users. This customizability is achieved by employing model based approaches where basic model parameters are trained or determined with a set of training data derived from an existing person. Once

the model parameters are trained, facial expressions and voice quality can be controlled easily.

2.2.2 Speech recognition module (SRM)

SRM consists of three submodules: the command interpreter, the speech recognition engine, and the grammar transformer. Based on a speech recognition engine “Julian” developed by Kyoto University and others, it accepts the grammar to represent sentences to recognize and has been modified to accept multiple formats for grammar representation and output incremental recognition results. It can change the grammar by request from external modules during dialog sessions. It also produces N-best recognition candidates for sophisticated use of multiple results.

2.2.3 Speech synthesis module (SSM)

This module is the first open-source license-free Japanese Text-to-Speech conversion system consisting of four sub-modules. The command interpreter receives an input command from the agent manager and invokes sub-processes according to the command. The text analyzer decomposes arbitrary Japanese input texts containing Kanji, Kana, alphabetic, numeric characters, and optionally embedded tags according to the JEIDA-62-2000 [3] typically specifying the speaking style, and extracts linguistic information including pronunciation, accent type, part of speech, etc., partly utilizing ChaSen[2] and newly developed dictionaries for Japanese morphological analysis. The waveform generation engine is an HMM-based speech synthesizer, that simultaneously models spectrum, $F_0$ and duration in a unified framework of HMM capability. The speech output sub-module outputs the synthetic speech waveform.

2.2.4 Facial image synthesis module (FSM)

FSM is a module for high quality facial image synthesis, animation control and precise lip-synchronization

\[ ^1 \text{Information-Technology Promotion Agency} \]
with synthetic and natural voice. To customize the
face model, a graphical user interface is equipped to
fit a generic face wire frame model onto a frontal face
shot image. Face action units are defined on
this generic model and prototype facial expression
be synthesized by combination of these action
units. Also autonomous actions such as blinking and
nodding can be generated. Lip movement in an utter-
cease is controlled by “viseme” and duration. Facial
animation is expressed easily by a simple script.

2.2.5 Task manager (TM)
The task of user-agent dialog management can be de-
described in VoiceXML. TM consists of translator, from
VoiceXML documents to the intermediate language
(Primitive Dialogue Operation Commands, PDOC),
and the dialogue controller that interprets the PDOC
documents. We extended the original specification of
VoiceXML to add some commands, including the fac-
ial expression controls of anthropomorphic dialogue
agents.

3 Integrating Platform

3.1 Galatea Architecture
In the Galatea Toolkit, the functional units are in-
dependently modularized, input/output devices are
directly managed in the module, and the agent man-
ger controls inter-module communication. In order
to easily integrate additional modules, all modules
are modeled as virtual machines having a simple com-
mon interface and connected to each other through a
broker (communication manager). The Agent Man-
ger (AM) works as a hub through which all modules
communicate with each other. For example, issuing
a command to the speech synthesis module means
setting voice synthesis of a given text right now “set
Speak = Now”.

3.2 Integration of Audio-Visual Infor-
mation
Recently, we added a body animation module to
the Galatea Toolkit. First, we built a cartoon-type
computer-graphics(CG)-based human-image anima-
tion (Fig. 3.2) and connected it to the Galatea agent
system. By making use of the simple Galatea archi-
tecture, connection was relatively easy. Second, we
combined the Galatea face and the body part of the
CG animation by overlapping the two animated im-
ages in the same window with synchronization. Now,
the agent system has face and body both animated by
commands received from the agent manager (AM).

Extending the above idea, we can connect informa-
tion inputs and control outputs with the same agent
architecture and operate the whole system utilizing
the new input and output. In this sense, anthro-
 morphic spoken-dialog agent is one of excellent plat-
forms for integrating audio-visual information from
sensors and actuators. For example, visual informa-
tion may provide the user’s information of position,
face expression, and gesture. Sound-source separa-
tion will be helpful not only in speech recognition in
noise, but also in face/body direction control. Com-
bined with mechanical robots, intelligent agent will
move and work physically in the real world and com-
municate with human users by speech, gestures and
facial expressions.

4 Acknowledgement
Authors would like to thank Shin-ichi Kawamoto
(JAIST) for providing materials for this article, and
other colleagues of the Galatea Project, Yasuharu
Den (Chiba Univ), Keikichi Hirose (UT), Katsunobu
Ito (Nagoya U), Atsuhiko Kai (Shizuoka U), Takao
Kobayashi (TIT), Akinobu Lee (NAIST), Nobuaki
Minematsu (UT), Shigeo Morishima (Seikei U),
Satoshi Nakamura (ATR), Tsuneo Nitta (TUT), Hi-
rosi Shimodaira (JAIST), Keiichi Tokuda (NITech),
Takahito Utsuro (Kyoto U), Atsushi Yamada
(ASTEM), and Yoichi Yamashita (Ritsumeikan U),
for their contribution to the Galatea Project. Also,
we are grateful to Masayuki Nakazawa and Christer
Lunde for their additional development (body anima-
tion) of the agent. We look forward to future collab-
oration with Profs. Shigeki Ando, Ayumu Matani
and Koichi Ishikawa in audio-visual information in-
tegration.

References
An	hropomorphic Spoken Dialogue Agent Toolkit,”
IPSJ Technical Report, 2002-SLP-45-10, pp. 57-
[3] Japan Electronic Industry Development Associa-
tion, “Standard of Symbols for Japanese Text-to-