Chapter 3
User Generated Dialogue Systems: uDialogue

Keiichi Tokuda, Akinobu Lee, Yoshihiko Nankaku, Keiichiro Oura, Kei Hashimoto, Daisuke Yamamoto, Ichi Takumi, Takahiro Uchiya, Shuhei Tsutsumi, Steve Renals and Junichi Yamagishi

Abstract This chapter introduces the idea of user-generated dialogue content and describes our experimental exploration aimed at clarifying the mechanism and conditions that makes it workable in practice. One of the attractive points of a speech interface is to provide a vivid sense of interactivity that cannot be achieved with a text interface alone. This study proposes a framework that spoken dialogue systems are separated into content that can be produced and modified by users, and the systems that drive the content, and seek to clarify (1) the requirements of systems that enable

K. Tokuda (✉) · A. Lee · Y. Nankaku · K. Oura · K. Hashimoto · D. Yamamoto · I. Takumi
T. Uchiya · S. Tsutsumi
Nagoya Institute of Technology, Nagoya, Japan
e-mail: tokuda@nitech.ac.jp

A. Lee
e-mail: ri@nitech.ac.jp

Y. Nankaku
e-mail: nankaku@nitech.ac.jp

K. Oura
e-mail: uratec@nitech.ac.jp

K. Hashimoto
e-mail: hashimoto.kei@nitech.ac.jp

D. Yamamoto
e-mail: yamamoto.daisuke@nitech.ac.jp

I. Takumi
e-mail: takumi@nitech.ac.jp

T. Uchiya
e-mail: t-uchiya@nitech.ac.jp

S. Tsutsumi
e-mail: tsutsumi.shuhei@nitech.ac.jp

S. Renals
University of Edinburgh, Edinburgh, UK
e-mail: s.renals@ed.ac.uk

J. Yamagishi
National Institute of Informatics, Tokyo, Japan
e-mail: jyamagis@nii.ac.jp

© Springer Japan KK 2017

T. Nishida (ed.), *Human-Harmonized Information Technology, Volume 2*,
DOI 10.1007/978-4-431-56535-2_3

the creation of attractive spoken dialogue, and (2) the conditions for the active generation of attractive dialogue content by users, while attempting to establish a method for realizing them. Experiments for validating user dialogue content generation were performed by installing interactive digital signage with a speech interface in public spaces as a dialogue device, and implementing a content generation environment for users via the Internet. The proposed framework is expected to lead to a breakthrough in the spread of using speech technology.

Keywords User-generated content · Spoken dialogue system · Speech recognition · Speech synthesis

3.1 Introduction

A human-centered information environment is an environment where everyone is a source of information, and is able to enjoy information naturally. Since speech is the most basic form of communication for humans, it is one of the ideals of modern society to realize the widespread availability of speech communication environments where people can naturally and freely interact other people wherever they may be, using advanced ubiquitous network telecommunication equipment. Although the fundamental technologies of speech recognition, speech synthesis and dialogue processing are making progress towards the sort of level needed for practical applications, it cannot yet be said that this sort of speech communication environment is available in the real world. There are other issues can be addressed, such as improving the accuracy of speech recognition, but in general it will probably not be possible to solve every issue simply by accumulating more technology. One such issue concerns the "attractiveness" of spoken dialogue systems to users. The ability to take part in a realistic interactive conversation is one of the important "draws" of speech interfaces that cannot be achieved with text processing alone. However, this can only be achieved by entering into regions where high-level human speech processing capabilities are required, such as facial expressions, gestures, voice quality and timing. The hardware and software limitations of current dialogue systems tend to make them rather inflexible and lifeless.

The aim of this study is to separate spoken dialogue systems into content that can be produced and modified by users and the systems that drive the content. In this way, we hope to clarify what sort of requirements must be met to produce "attractive" content and systems so that speech technology can spread widely among people (Fig. 3.1). However, attractiveness is created through the combination of human feelings and knowledge, and is not something that can easily be evaluated mechanically. Therefore, we have to inductively clarify the essential qualities of attractiveness by establishing a framework that makes it easy for users to create and evaluate large quantities of dialogue content. The basic strategy of this study is to construct a "circulatory system" of content generation as shown in Fig. 3.1 and empirically analyze

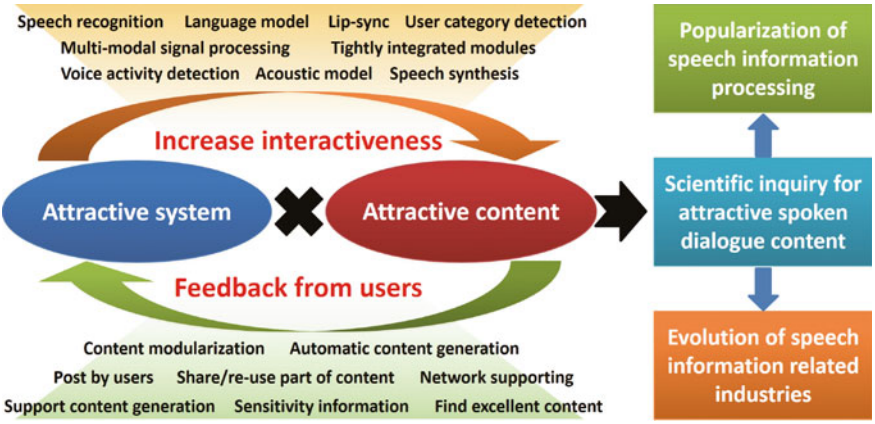

Fig. 3.1 A circulatory system of content generation

the factors achieving a loop gain of more than 1 in order to establish techniques for constructing the framework of user-generated dialogue content. s

3.2 Background and Purpose of the Study

3.2.1 The Relationship with Industrial Structures

To establish a framework to facilitate the creation and evaluation of dialogue content, we need to separate the content from spoken dialogue system software, and make them widely available to creators and ordinary users without software engineering skills. This process resembles the evolution of industrial structures shown in Fig. 3.2. The telecommunications industry started out with the creation of electronic equipment, but as it grew larger, the software production parts and then the content creation parts were separated out, and ended up forming major industry fields. Mobile game production is a typical example, where content creation is progressively separated from development of game engine software. To bring about this sort of change with regard to speech technology, it will be necessary to accelerate the creation of attractive content by getting as many creators and users as possible involved in content creation.

3.2.2 Relationship with the UGC Approach

Today, attention is focused on content created by users as referred to by terms such as CGM (Consumer Generated Media)and UGC (User Generated Content). This is a

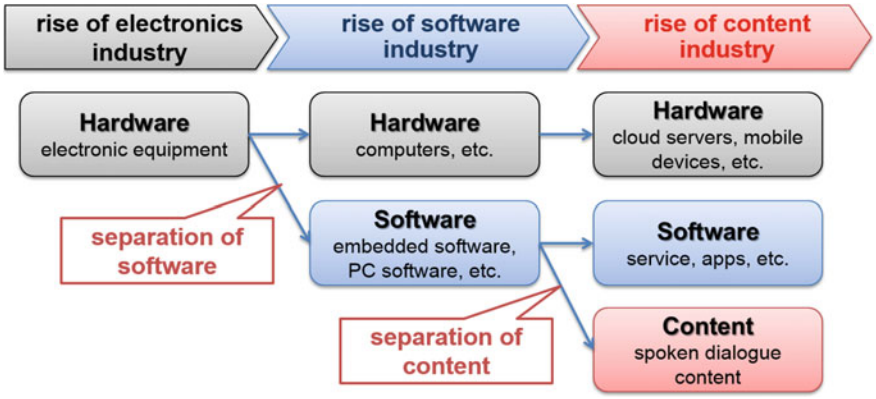

Fig. 3.2 The evolution of industry structure related to telecommunications

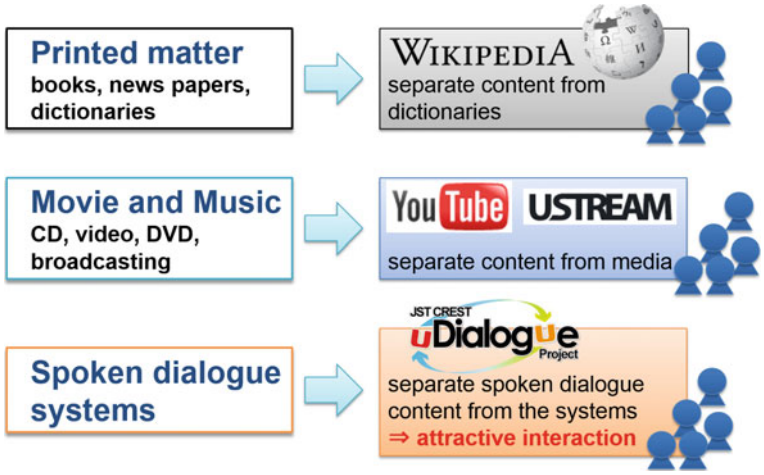

Fig. 3.3 Separation of content

system where users are mainly responsible for the creation of content—well-known examples include Wikipedia, YouTube, Facebook, Twitter and Instagram. The main features of these systems are that new content is continuously created by users, and that the users' assessments and wishes are directly reflected in the content. Our approach in this study can be regarded as a version of dialogue content, which aims to implement a users' information transmission environment based on a new type of media content (Fig. 3.3).

3.2.3 *Devices for Implementing a Ubiquitous Speech-Based Information Environment*

Devices for implementing a ubiquitous speech-based information environment can take many forms, such as ordinary PCs, information appliances or smartphones, and the created dialogue content will also vary according to the device characteristics and usage environments. In this study, we focused on digital signage placed in public spaces as one style of demonstration experiment. Digital signage is a medium for the delivery of information and advertising using digital communication technology and display technology such as large-sized liquid crystal displays. Its benefits include the ability to display rich media content such as audio and video, and change content at any time by using digital communication. In recent years, attempts have also been made to use technologies such as proximity sensors, touch panels and face recognition to control the display interactively. By further developing this idea to add voice interaction functions, it should become possible to produce a natural and impressive level of interactivity. In this study, digital signage devices equipped with speech processing functions were installed in various places such as a university campus, a tourist office and a city hall, and were used to perform demonstration experiments involving various cooperative mechanisms via the Internet.

3.3 MMDAgent: A Toolkit for Building Voice Interaction Systems

To comprehensively research the various elements of voice interfaces such as their level of engagement with users and their implicit attractiveness against other user interface, these systems need to develop into areas where human assessment and advanced processing are required, such as expressions, gestures, tone of voice and timing. To do this, we need a platform that is closely integrated with not just the voice processing system but also the image processing and virtual agent representation. Furthermore, since various data has to be collected for various tasks and situations, we require an advanced and flexible system where both users and system developers can work freely on every part of the system and dialogue content.

So far, we have continuously developed and released open-source research platform software tools that cooperate with speech technology including HTS (an HMM-based speech synthesis system) [1], Open JTalk (a Japanese text-to-speech system based on HTS) [2], and Julius (a speech recognition engine) [3]. Based on this group of software, we built "MMDAgent" toolkit for building voice interaction systems [4] by incorporating speech recognition, HMM-based flexible speech synthesis, embodied 3D agent rendering with simulated physics, and dialog management based on a finite state transducer (FST) and we released it as an open-source software toolkit (Fig. 3.4). The inter-module architecture is fairly simple: a single message queue is shared among all modules, and an output from a module will broadcast to all

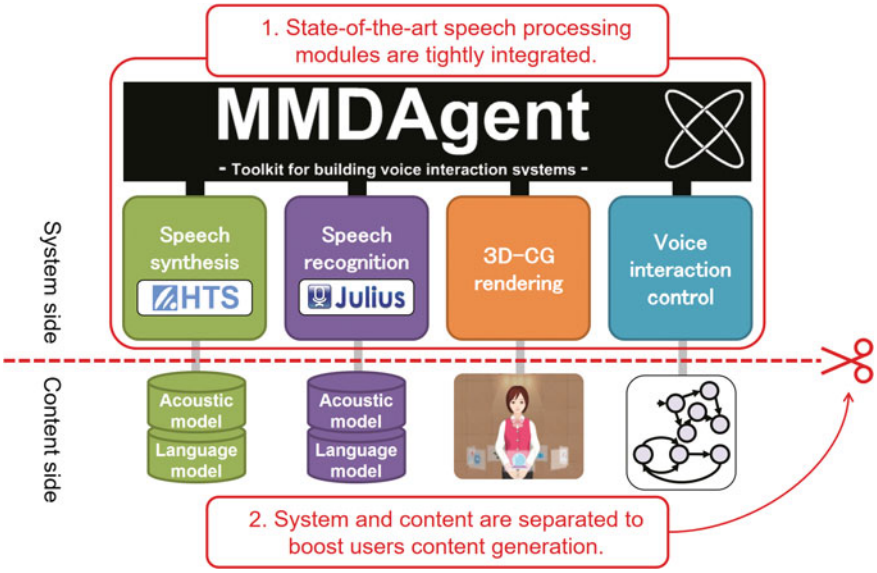

Fig. 3.4 A toolkit for building voice interaction systems (MMDAgent)

modules. The interaction control script is written in FST [5], a generic automaton representation converting the received messages (recognition result, sensor status, timer, etc.) into output messages (synthesis text, motion trigger etc.). This system uses open formats not only for the speech recognition and speech synthesis, but also for the virtual agent's 3D models and the motion data and FST definitions that drive these models. This makes it possible for users and developers to freely create, edit and replace any of the system's components using existing tools.

MMDAgent adopts a design policy that is geared strongly towards enabling not only speech technology experts but also ordinary people everywhere to enjoy creating systems using speech technology, and aims to support the continuous creation of attractive dialogue content. This is the main feature of MMDAgent. MMDAgent allows richly expressive dialogue and computer graphics to be produced with high performance, and is used as the technical platform of this study.

In this study, we use the MMDAgent toolkit for building voice interaction systems as a platform to separate spoken dialogue systems into the content provided to users and the systems that drive this content. Our aim is to clarify which factor is essential to create systems and content that are sufficiently engaging and attractive to utilize speech technology widely to many people. There are many issues that need to be addressed in order to achieve this goal. They can be summarized in the following three categories:

1. Enhancement of the baseline technology and software
2. Creating a framework for content creation
3. Public experiments of content creation

By addressing these issues, we established a framework that makes it easy for users to create and share dialogue content, and we clarified the mechanism whereby attractive dialogue content is created from the creation/sharing/evaluation of dialogue content by users. In the following three sections, the above three items will be described.

3.4 Enhancing the Baseline Technology

To create attractive dialogue content, it is essential to provide a technology infrastructure where spoken dialogue systems are able to engage users. This section describes our efforts to enhance the speech recognition, speech synthesis and other underlying technologies of the spoken dialogue system, and to optimize the underlying software for building the spoken dialogue system, and the design of the corpus/agents.

3.4.1 Underlying Technology

To optimize the underlying technology, we performed a lot of basic research for processing speech information, such as speech recognition and speech synthesis. Some typical examples are listed below.

- Integration of feature extraction and modeling for HMM based speech synthesis [6]
  In recent years, speech synthesis has been performed using statistical models called hidden Markov models (HMMs) to model the acoustic relationships between speech features and linguistic features. In this study, by integrating the extraction of features from speech into the HMM training process, we were able to directly model the speech waveforms with a unified framework. This improved the system's overall modeling capabilities, and greatly improved the quality of synthesized speech.
- Improvement of spectral modeling with an additive structure [7]
  We proposed an additive model for speech synthesis by assuming additive structures in the relationship between linguistic features and speech features. We also proposed a training algorithm for the efficient training of additive models. It was shown that the synthesized speech quality with an additive model was greatly improved.
- Conversational speech synthesis method [8]
  Dialogue between people often contains hesitations and filler pauses (like "ah" and "um"). We therefore examined the impact of hesitations and fillers on dialogue, and we studied how to automatically insert them into synthesized speech. This enabled us to make accurate predictions about where they will be inserted, and to perform speech synthesis in a more natural conversational style.
- Improving the precision of speech recognition using a language model based on a recurrent neural network [9]

For high-performance speech recognition, studies are being performed where recurrent neural networks are used for language models that model linguistic features. In this study, we proposed a learning method based on a neural network-based language model suitable for speech recognition, and we improved the speech recognition accuracy. We were also able to perform speech recognition with greater precision by making it possible to use acoustic features as additional information.

- Improvement of speech synthesis based on deep neural networks [10–12]
  It was recently reported that the performance of speech synthesis can also be greatly improved through the use of neural networks. In this study, we proposed methods for integrating multiple neural networks and methods for learning in neural networks that are suited to the problem of speech synthesis, demonstrated their effectiveness.
- Construction of text-to-speech systems for unknown-pronunciation languages [13]
  Ordinary speech synthesis consists of a text analysis part that predicts how text should be read, and a waveform generation part that creates the corresponding speech waveforms, but it is not possible to construct a text analysis part for languages whose pronunciation information is unknown. We therefore proposed a text-to-speech system construction method for unknown-pronunciation language, which involves the use of speech recognition systems for different languages. Using this construction method, we were able to construct text-to-speech systems for a wide variety of languages.
- Analysis of dialogue content using a topic model [14]
  The creation and management of dialogue content requires a framework where it is possible to perform operations such as searching for topics, detecting similar content, and recommending popular content. There are many different kinds of dialogue content, and it is difficult to perform procedures such as rule-based tagging. In this study, we proposed a method for automatic statistical classification of dialogue content by applying a topic model (a kind of language model).

3.4.2 Underlying Software

We summarized the results obtained in the advancement of underlying technology as research platform software, and we published the following open source software.[1]

- MMDAgent: Toolkit for building voice interaction systems
  (http://www.mmdagent.jp/)
  (61,000 downloads)
- Julius: Open-source large vocabulary continuous speech recognition engine
  (http://julius.sourceforge.jp/)
  (216,000 downloads)

[1]The number of downloads is the cumulative total from October 2011 to March 2016.

- HTS: HMM speech synthesis toolkit
  (http://hts.sp.nitech.ac.jp/)
  (371,000 downloads)
- hts_engine API: HMM speech synthesis engine
  (http://hts-engine.sourceforge.net/)
  (41,000 downloads)
- Open JTalk: Japanese text-to-speech system
  (http://open-jtalk.sourceforge.net/)
  (47,000 downloads)
- SPTK: Speech signal processing toolkit
  (http://sp-tk.sourceforge.net/)
  (42,000 downloads)
- Sinsy: HMM-based singing voice synthesis system
  (http://sinsy.sourceforge.net/)
  (1,400 downloads)

The development of this open-source software is ongoing, and new versions continue to be released. In particular, the MMDAgent toolkit for building voice interaction systems was developed as cross-platform software that can run on Windows, Mac OS, Linux, Android OS and iOS. It can run by itself on smartphones and tablet PCs, and makes it possible for a spoken dialogue system with small response delays to be used in smartphones and tablets (Figs. 3.5 and 3.6). It seems to be the first open-source implementation of a spoken dialogue system with a 3D agent capable of running on a stand-alone smartphone or tablet PC.

These software platforms include state-of-the-art technology, and as the number of downloads show, they have already achieved the status of de facto standards. In fact, as shown in Fig. 3.7, it is being widely used in many different situations, including academic papers, software development and events.

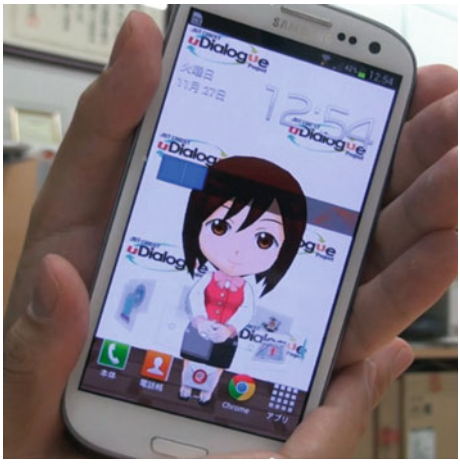

Fig. 3.5 Android-compatible MMDAgent

Fig. 3.6 iOS-compatible MMDAgent

Fig. 3.7 Examples of the usage of open-source software

3.4.3 Corpora and Agents

In order to obtain universal knowledge that is not language-dependent, it is necessary to perform validation experiments in parallel with various different languages. When doing so, it is necessary to give full consideration to cultural differences as well as linguistic differences. We therefore developed a spoken dialogue system targeting Japanese and English, and by performing experiments with the Japanese and English

spoken dialogue systems, we verified which points are dependent on language and culture, and which are not.

First, we constructed a Japanese speech corpus and a British English speech corpus, which were needed to develop the spoken dialogue system.

- Japanese male voice actor corpus
  Recordings of 3,000 sentences in Japanese spoken by a male voice actor. The recorded utterances are selected from multiple domains including newspaper, etc. There are 5 speaking styles consists of 1800 normal utterances and 300 emotional utterances for angry, sad, happy and whisper, respectively.
- CSTR VCTK corpus [15]
  This corpus is a large-scale speech corpus containing about 60 hours of speech by 109 speakers with various British English accents. The recorded speech consists of about 400 sentences from each speaker, selected from multiple domains including newspaper articles. This corpus is 50% larger than the WSJCAM0 corpus (available from LDC for a fee), which is the current standard speech database for research on British English. It is expected that it will be used for diverse applications in many different fields in the future.
- Corpus of British English (Edinburgh) spoken by a female
  Recordings of 4,600 sentences in British English spoken by a female voice actor with an Edinburgh accent. Speech was recorded at two different speeds, with 800 sentences spoken at high speed, and 800 sentences spoken at low speed. 800 sentences spoken at low speed consist of 2 speaking styles: talking to a hearing impaired person and talking to a computer. There are also over 100 minutes of spontaneous conversation recordings.

Next, we created a conversation agent suited to dialogue (Fig. 3.8). To make the spoken dialogue system acceptable to users, it was necessary to make the system by taking cultural differences into consideration. In Japan, people are thought to have little resistance to animated 3D agents, so we created a female agent with a height of 2.5 heads, and a male agent. On the other hand, from the results of various discussions, it was considered that a realistic 3D agent would be more acceptable to a European audience, so we created a dialogue agent for the British English version of the spoken dialogue system.

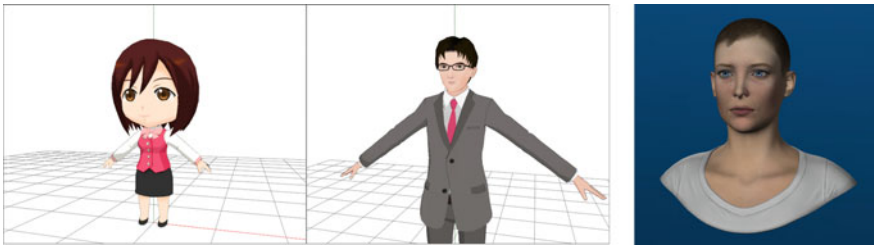

Fig. 3.8 Animated dialogue agent and realistic dialogue agent

3.5 Building a Mechanism for the Creation and Sharing of Content

In this section, we discuss a mechanism for creating and sharing dialogue content by introducing the concept of user-generated dialogue content.

As shown in Fig. 3.9, it consists of a three-level hierarchy. The material layer contains specialized model data such as voice models and language models, and binary files such as images, music, 3D models and motion data. The action layer contains short action sequences, such as dialogue patterns for simple greetings, or for displaying weather forecast panels. The scenario layer combines the actions of the action layer to produce more complex dialogue scenarios. The scenario layer and action layer are scripted in FST format, while each material in the material layer is stored in its own format.

As shown in Fig. 3.10, we also assumed that the spoken dialogue system may cooperate with other systems instead of operating as a closed stand-alone system. In particular, the smartphone version of MMDAgent implements a mechanism that makes it easy to link up with networks and other smart phones [16].

3.5.1 Tools for Dialogue Content Creation

In the MMDAgent toolkit for building voice interaction systems, the dialogue scenarios are written in FST. However, it is difficult for most users to create FST scripts manually. We therefore implemented a mechanism that allows dialogue content to be edited easily.

Dialogue Script Editing Tool

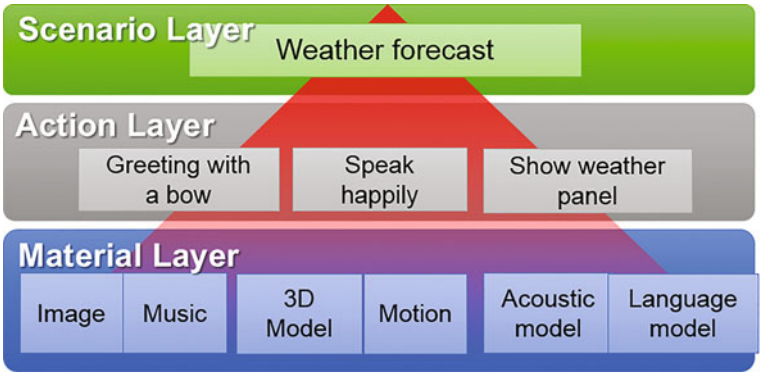

Fig. 3.9 Hierarchy of dialogue content

Fig. 3.10 Linking spoken dialogue systems with a network

To facilitate the creation of user-generated dialogue content, it is essential that users can easily create dialogue content. We therefore developed the EFDE dialogue content creation tool (Fig. 3.11) and the MMDAE dialogue content creation tool for advanced content creators using web browsers (Fig. 3.12).

- EFDE (Extended FST based Dialogue Editor) [17]
  A tool targeted at beginners using Android devices, where FST scripts displayed as state transition diagrams can be edited using a touch interface. The edited scripts can be run on demand so that users can easily edit scripts while checking the action of the dialogue content. Also, by providing typical parts of dialogue scripts as templates, we have made it possible to implement complex dialogues with a small number of steps. Furthermore, users can also input sentences and keywords in a dialogue content by speaking, thus a dialogue content can be created easily in a multi-modal manner.
- MMDAE (MMDAgent Editor) [18]
  A web-based FST editor for a detailed editing of FST scripts. Editing can be performed easily and reliably by active input suggestion for message formats and

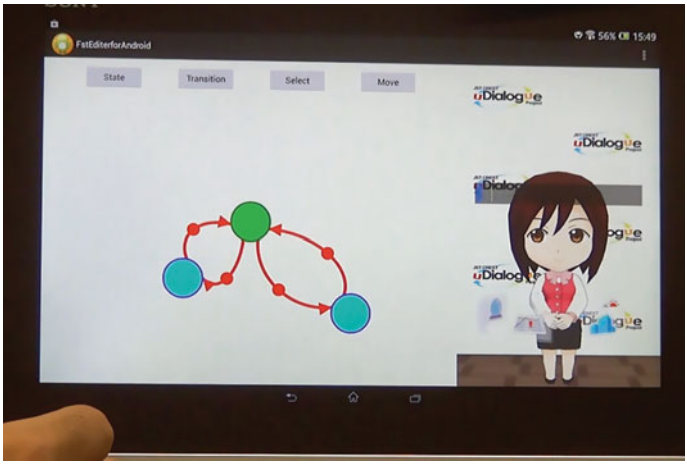

Fig. 3.11 Tablet interface for EFDE

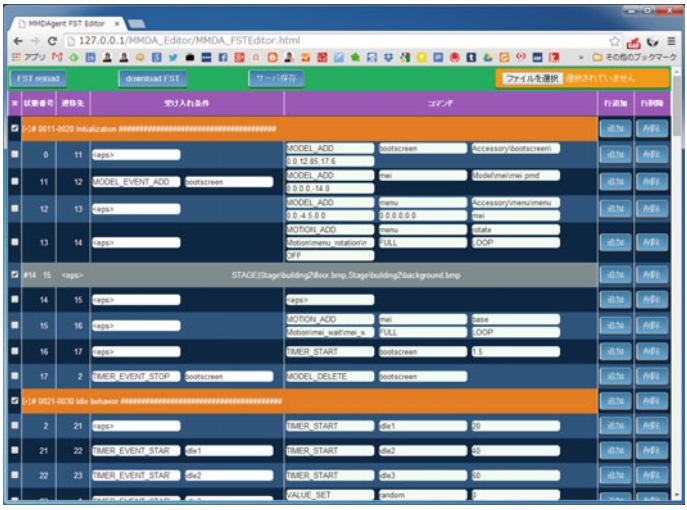

Fig. 3.12 Web browser interface for MMDAE

structured FST input view. Since this tool runs on a web browser, it can be used on a wide variety of platforms.

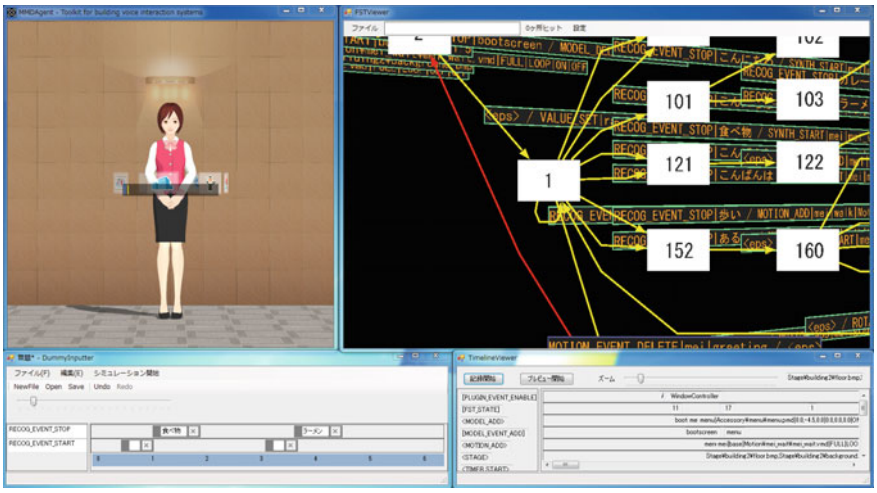

Fig. 3.13 Voice interaction builder

Voice Interaction Builder

To promote the creation of attractive high-quality content by users, a creator-oriented environment must be provided where users can exercise detailed control over the addition of dialogue content according. We therefore made a prototype voice interaction builder as a development environment where interactive dialogue content can be created with detailed control over the speech timing and other details, and where the action of this content can be verified (Fig. 3.13). This voice interaction builder consists of three components: (1) a function for grasping the structure of FST scripts by visualizing and browsing the state transition diagram in 3D space, (2) a function for tracing the operation of a script by means of an event input simulation, and (3) a function for verifying time-series interactions by storing and playing back a series of input/output events. In subjective tests, users reported that it was easier to create content than in the existing environment.

3.5.2 Construction of a Cloud Environment for Collaborative Content Editing

One of the characteristics of user-generated media is that it is frequently built as a collaborative effort with other users. In dialogue content, by constructing an environment where dialogue content can be created and edited over a network, it becomes easier to create complex content in collaboration with other users, and to incorporate

and extend existing content. In this section, we discuss the construction of a collaborative editing environment for dialogue content based on a cloud environment.

Collaborative Editing System for Building Dialogue Content with Dialogue Context

One of the simplest way for the collaborative construction of dialogue content is to make system that gathers a set of questions and answers from users via Web. However, this system can only handle simple question-and-answer exchanges because it retains no context of previous utterances, and is thus clearly inadequate as a system intended to provide fun and attractive dialogue with continuity and situational awareness. Also, it is sometimes hard to build a system that covers all the responses that a user might give during a lengthy conversation.

We have therefore developed a user-generated spoken dialogue system makes it easy to collaboratively construct conversations with dialogue context (Fig. 3.14). Conversations are stored in units of keyword/response sets as same as simple QA systems, but can also include parent-child relationships, whereby users are able to collaboratively construct multiple consecutive question-and-answer type interactions. The whole dialogue content is stored in a tree structure consisting of keyword/response pairs as its building blocks. For instant registration of a dialogue content, we built an SNS (social networking service) chat-style Web interface where it is possible for anyone to intuitively grasp and edit the flow of a conversation (Fig. 3.15). This system can be used to create long continuous conversations, and allows dialogue registered by other users to be branched off into other conversations by another users. In subjective evaluation tests, it was found that this approach generates much more anticipation and interest among users than the conventional method, and enables the construction of user-generated systems with attractive conversations.

Construction of a Cloud-based Dialogue Content Editing Environment

To increase the variety of conversations in MMDAgent, it is generally necessary to describe a larger number of conversation scenarios. However, it is difficult for a person to create conversation scenarios on a large scale. We therefore developed a crowd sourcing system specialized for the creation of dialogue scenarios (Fig. 3.16). In this system, multiple users receive orders for dialogue scenario creation tasks based on a crowd sourcing scheme. This allows scenarios to be created by sharing the workload among multiple users. To make it easy for a user to create a scenario after receiving a dialogue scenario creation request, we also strengthened the functions for collaboration between MMDAE and a Skype version of the tool to debug the system easily. In an experimental evaluation, we confirmed its validity for the creation of dialogue scenarios based on crowd sourcing [19, 20].

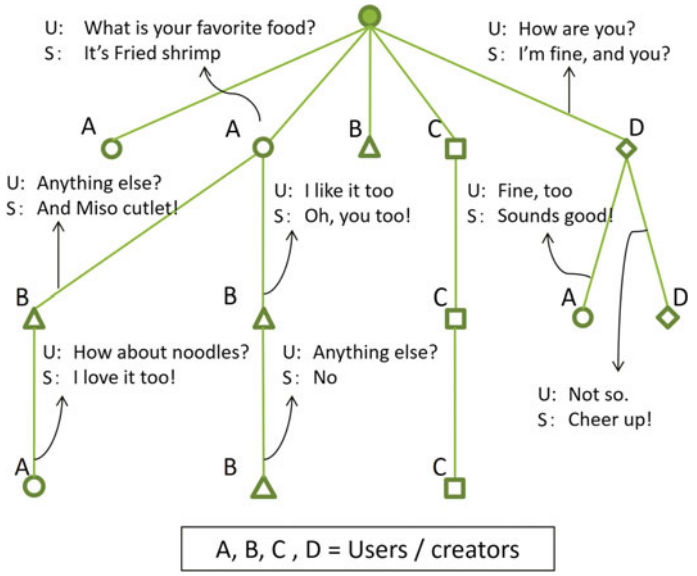

Fig. 3.14 Concept of interactive content with history

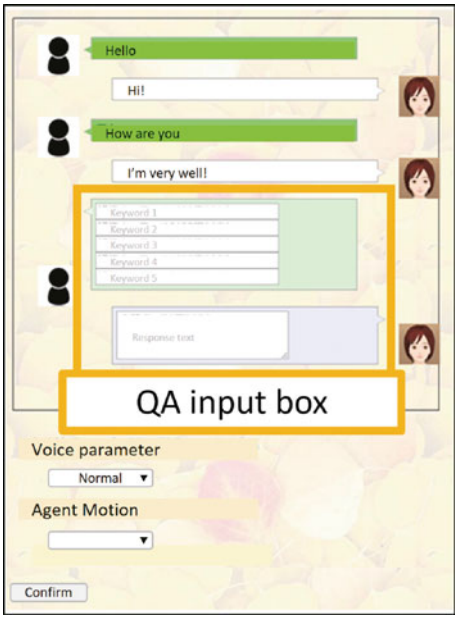

Fig. 3.15 Registration interface for interaction with history

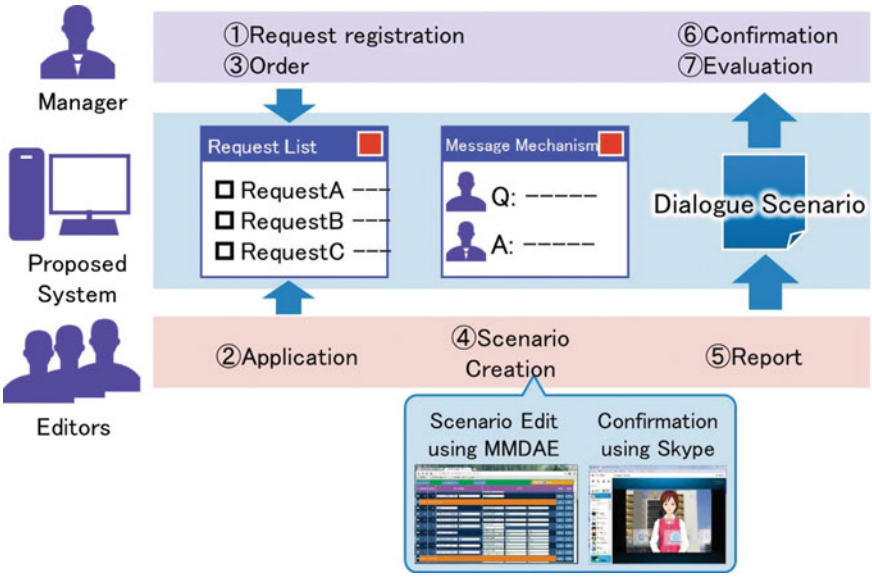

Fig. 3.16 Crowd sourcing system specialized for the creation of dialogue scenarios

3.5.3 Modularization, Networking and Inter-agent Collaboration

Packaging Dialogue Content

Most of the current dialogue systems are "Haute couture": they are built for a specialized task, tuning all the components especially for the task, and runs solely for the task. Their availability or portability toward various tasks or users as a media content is not well studied yet. We therefore proposed a framework for distributing and sharing user-generated dialogue content over a network via a server. First, we studied a packaging method by extending FST scripts for modularization and parallelization. This made it possible to update a part of FST script, or function-based script delivering, while at the same time making FST scripts easier to maintain. We also studied a framework for circulating content in package units by creating a prototype delivery framework as shown in Fig. 3.17.

We also proposed and built a user-generated spoken dialogue system architecture with the aim of implementing an environment where the user is free to reassemble, select and construct not only dialogue scripts but also each constituent element of dialogue content such as word dictionaries, voice data, 3D models or motion data (Figs. 3.18 and 3.19). By defining the part of dialogue content as general-purpose modules for the integrated handling of speech recognition/synthesis, dialogue management and all content including agent models, we made it possible to achieve consistency in the operation of packages and handle dependencies between mod-

ules. In practice, we proposed a mechanism for managing modules in packages for MMDAgent, and we implemented a framework for applying scopes to module name specifications (e.g., inside dialogue scripts), and for automatically detecting dependencies and conflicts between packages based on their external declarations.

A Platform for Cooperation Between Spoken Dialogue Systems Using Network Agent Technology

Spoken dialogue systems for existing smart phones either work as stand-alone applications or communicate with a remote server, and it has not been possible to build complex dialogue scenarios that efficiently link up with multiple terminal devices. We therefore introduced the system collaboration platform described below.

- We built an environment where MMDAgent can run cooperatively on multiple smartphone devices. Specifically, we developed a network connection mechanism for spoken dialogue systems based on agent/NFC/Bluetooth technology (Fig. 3.20)

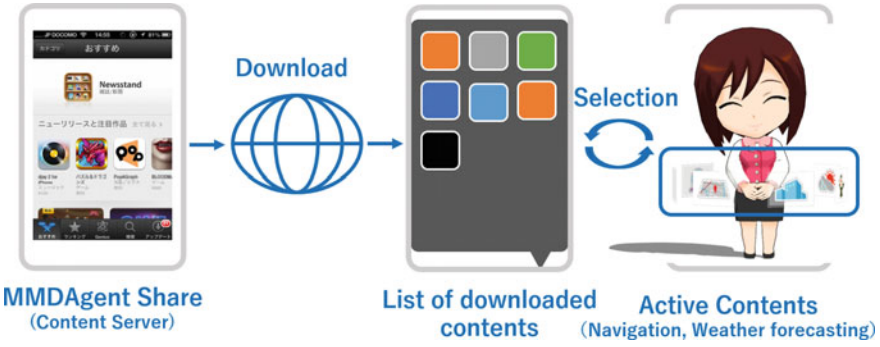

Fig. 3.17 Packaging of dialogue content

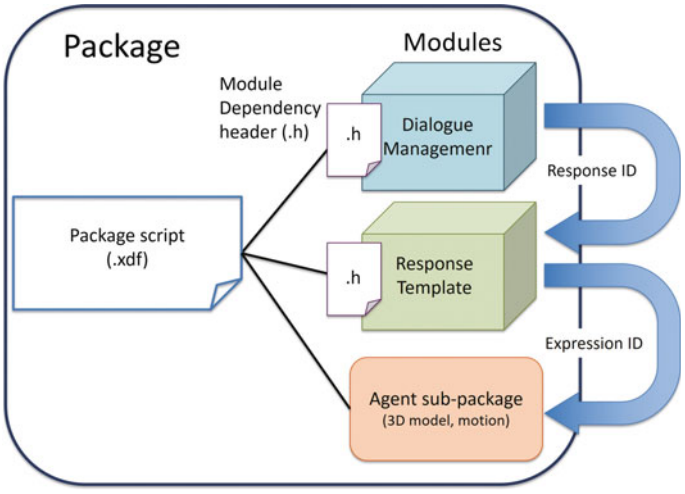

Fig. 3.18 Converting dialogue modules into packages

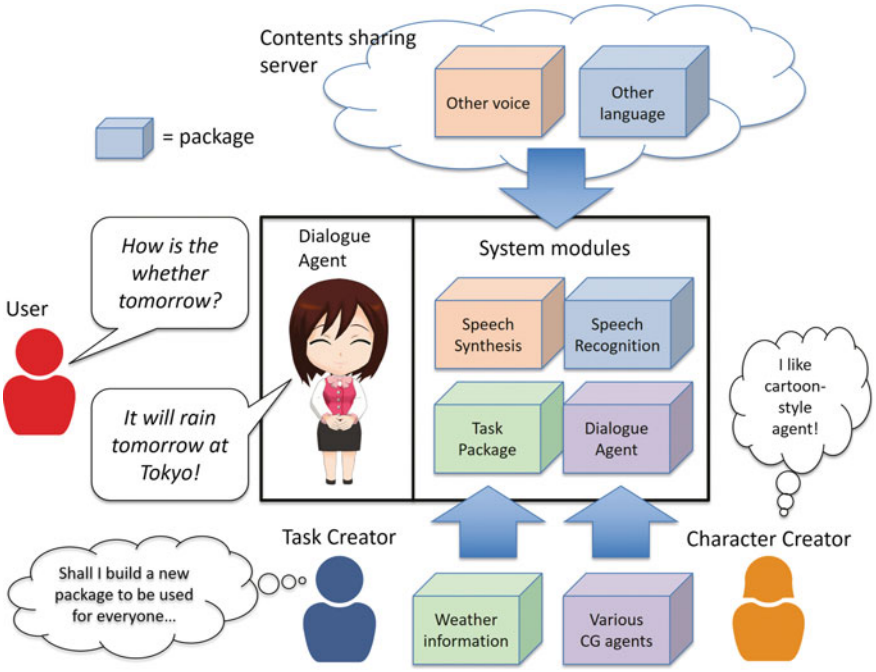

Fig. 3.19 Module-based sharing of dialogue content

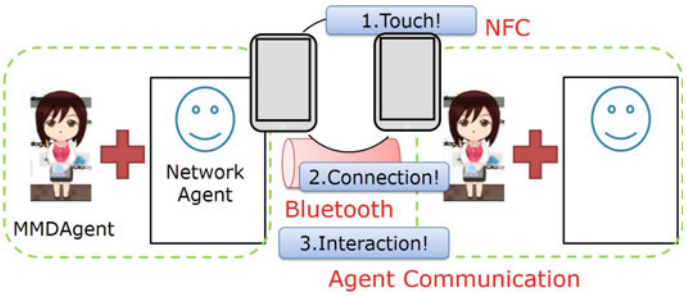

Fig. 3.20 P2P network connection mechanism

[21]. We made a prototype scheduling system to confirm the validity of services involving cooperation between multiple spoken dialogue systems.

- We built an environment for running MMDAgent cooperatively on multiple signage devices. This environment was designed to be highly scalable in order to accommodate large numbers of signage devices. We confirmed the validity of system collaboration between signage devices by making a prototype system to share the number of dialogues.

3.6 Public Experiments of Content Creation

The studies in the previous sections are essentially aimed for re-defining dialogue system as a player executing dialogue content, and we have worked on several aspect of both system and content creation/management, which are essential to make the dialog content applicable as a user generated media.

In this section, several public tests performed using the proposed system are summarized for discussion.

3.6.1 Building a Framework to Promote the Use of Content

In user-generated dialogue content, it can be expected that a large quantity of diverse dialogue content will be created by users. In such situation, it is important to maximize the usefulness of this content by organizing it into a suitable taxonomy. Here, we discuss the results of a study for the analysis, categorization and correlation of dialogue content groups in order to support the creation and use of spoken dialogue systems.

Related Word Recommendations Based on the Usage History of a Spoken Dialogue System

Fig. 3.21 Related word presentation method. (*red frame box* related words, *blue frame box* conversation history)

With the aim of promoting the use of spoken dialogue systems by beginners with no experience in the use of spoken dialogue systems, we studied a framework for presenting the user with a group of word candidates that can be said next. We then implemented a technique for acquiring relevant words from user's conversation data stored on the server using an information recommendation technique (Fig. 3.21). In user testing, we obtained results suggesting that the presentation of related keywords helps users make effective use of spoken dialogue systems. We also conducted a comparative study of four different related word recommendation methods (history-based recommendations, content-based recommendations, ranking recommendations and random recommendations), and a composite recommendation method that integrates all of these. In user testing, we confirmed that the composite recommendation method produces better overall results than individual content-based or random recommendation.

Strengthening of Mutual Incentives by Sharing Usage Histories Between Users

In user-generated media, sharing activity histories is essential in that it will cause mutual stimulation among users, leading to increased usage of the system and increased content creation. Thus, also in dialogue content, it is expected that sharing interactions with the spoken dialogue system and content usage histories between users can mutually increase the willingness of users to use the system, which is required to make the dialogue content as user-generated media. Therefore, in an open question-and-answer style spoken dialogue system in which any user can add any QA to the system, we examined several methods that provide user-oriented mutual incentives for interaction by (1) on-line sharing what other users are saying now, and (2) displaying information of "hot" content, which are extracted according

Fig. 3.22 Sharing conversation histories and displaying dialogue content rankings

to the frequency ranking and its registration time. As shown in Fig. 3.22, when this information was presented to the users and creators of dialogue content, users can share other user's activity and will be stimulated to use the system more. We also implemented a feedback from users to content creators by gathering user's out-of-domain utterances and give the statistics to the creators. In user testing, we confirmed that these frameworks can lead to an improvement in the motivation (incentivization) of both users and content creators.

3.6.2 Demonstration Experiments in Public Spaces

To demonstrate the proposed system's suitability for public installations, we installed and operated it at various locations including in front of the main gate at the Nagoya Institute of Technology, and at the tourist information office in Handa city. The dialogue content collected in these validation experiments will shed light on methods for making dialogue content more attractive. Some typical experiments are listed below.

Demonstration Experiment in a University Campus

To test the applicability of our system to public, an all-weather voice enabled digital signage system ("Mei-chan") was built by using MMDAgent and installed in an open public space at the main gate (Fig. 3.23). This system supports various functions including using multiple cameras and face recognition technology to control the line of sight of an animated character, and having the character actively address people detected using pyroelectric sensors. Content that integrates together not only the displayed text, images and dialogue text, but also the character's movements and speaking style when offering guidance can be updated dynamically from a server, and is used to display timely guidance ranging from information about events on campus that are recorded in a database at the information platform center to weather forecasts and other information optionally selected depending on the dialogue timing and content (Fig. 3.24).

The first version of the system went into operation on April 6, 2011, and on June 15, 2011, the content registration system was made open to be used by anyone in the university. By November 15, 2011, more than 100 items of content had been registered, and the number of user utterances captured per day was about 350 on average, even including holidays. There were 243 submissions of dialogue content from students and faculty staff (of which 4 were from students). Figure 3.25 shows examples of the submitted panel displays. Since September 2014, the system has also been installed in the open space ("Yume Room") at the university (Fig. 3.26). The guidance system installed in the student space allows dialogue content to be submitted easily not only from a PC but also from a smart phone with near field communication (NFC). During the six-month period following installation, a total of 437 dialogue content contributions were made by students and others, totally spontaneously.

Fig. 3.23 "Mei-chan": Digital signage in Nagoya Institute of Technology main gate

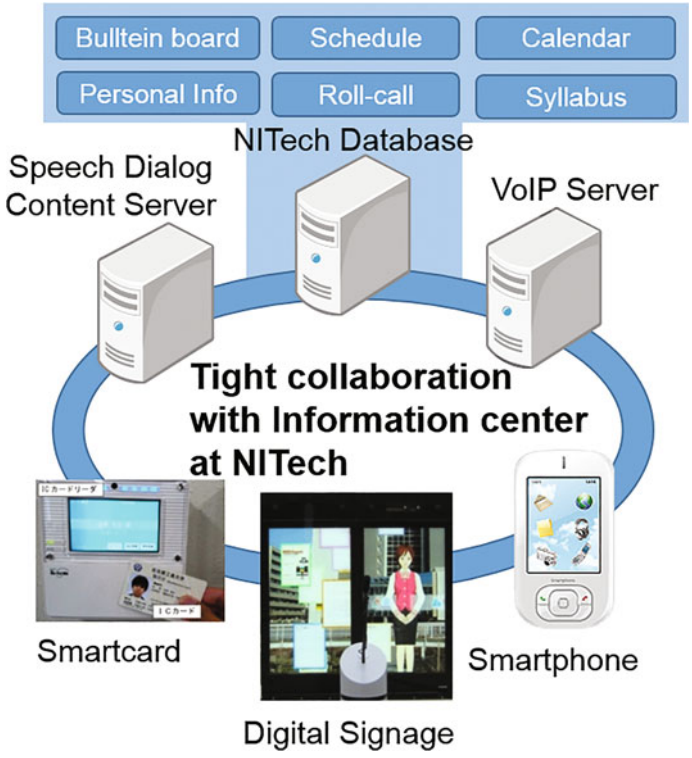

Fig. 3.24 Cooperation with the information infrastructure system

Fig. 3.25 Panel display

Fig. 3.26 Open space at the Nagoya Institute of Technology

Fig. 3.27 Handa city tourist information office (inside view)

Fig. 3.28 Handa city tourist information office (outside view)

A network questionnaire was held to conduct a survey of Mei-chan's popularity and frequency of use. Of the 262 valid responses, nearly everyone (99%) were aware of this system. A high percentage (77%) had also attempted conversations, showing that the spoken dialogue system using MMDAgent and its user-oriented way of making dialogue content attracts people inside the university. The free questionnaire drew a response rate of 33%, showing that one out of three users actively offered suggestions on how to improve the spoken dialogue system, etc.

Demonstration Experiment Outside the University

Next, the system was experimented on several places outside the campus for public test. An indoor digital signage prototype was installed at a tourist information office of a sightseeing spot in Handa city (Figs. 3.27 and 3.28). Since the room for this system was confined, the screen layout and expressions were optimized for close use.

Fig. 3.29 Handa city hall

We also implemented a localization whereby users (in this case, the tourist office staff) could add and update the dialogue content with ease. This enabled them to freely modify and adapt the content to the varying daily needs on the site by their own, which could made this system more applicable and attractive to tourists. In fact, a large amount of tourist information content was registered by the staff at the tourist office. This system was featured on television and in the newspapers, and appeared on the front page of Handa city's official newsletter, generating interest from other regions in connection with the use of this technology for tourism and PR.

Another spoken dialogue systems were also installed at other locations including the Handa city hall building, the National Institute of Informatics, and the NHK's Nagoya broadcasting station (Figs. 3.29, 3.30 and 3.31). For the National Institute of Informatics, we built a spoken dialogue system using a 3D model of their mascot (a cartoon dog called "Bit") (Fig. 3.30). At the NHK Nagoya broadcasting station, we performed a demonstration experiment where the equipment was used in TV broadcasts and events. This system was extended to include extra features such as being able to operate multiple characters by linking them together. We also constructed a speech database matched to our character, and a speech synthesis system that uses the character's voice.

Other Validation Experiments

More experiments has been conducted to explore the use of spoken dialogue systems and user generated dialogue content creation, namely on mobile environments.

- A video streaming version of our system has been developed (Fig. 3.32) [22]. This system was implemented by linking MMDAgent with Skype. A public trials and a questionnaire survey was conducted at the annual conference of IPSJ to serve as a conversation-based guidance for the sessions and venues. From the results

Fig. 3.30 National institute of informatics

Fig. 3.31 NHK Nagoya broadcasting station (from NHK news program "Hot Evening" on August 28, 2015)

of 120 questionnaires, it is found that it can provide information in a friendly conversational manner, whereas its network response delay might annoys users.

- A mobile campus tour support system based on dialogue was developed and used for junior high school students (Fig. 3.33) [23]. In this system, a dialogue agent on a smart phone provides each of them with information about the Nagoya Institute of Technology campus and its facilities. Subjective evaluation tests demonstrated the effectiveness of the system.
- It would be useful to link it up with systems such as the public telephone network for more availability. We therefore developed a framework that links MMDAgent with a VoIP client [24]. We created a middleware to connect the system to Skype for business. At the same time, to connect with the internal VoIP phone system of the internal unified communication system at Nagoya Institute of Technology,

Fig. 3.32 MMDAgent skype version

Fig. 3.33 Campus tour support system

we developed middleware for connecting a software phone, and we subjected it to interconnection tests.

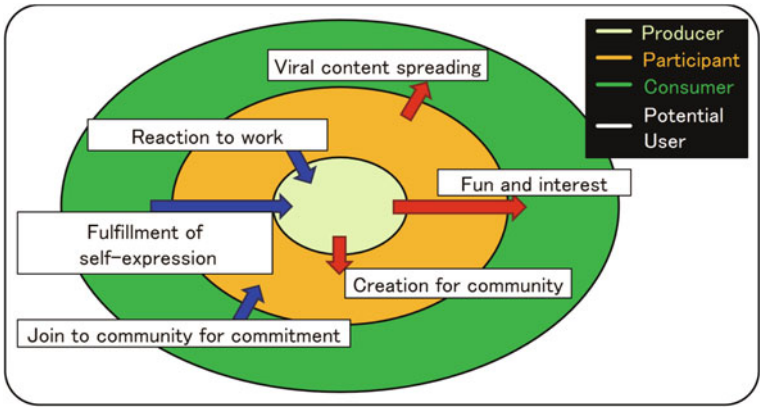

Fig. 3.34 User layer categories and incentives

3.6.3 Demonstration Experiment in a Network Environment

To obtain cues on the implementation of user-generated dialogue content in a broader social environment, we performed demonstration experiments involving a general network environment.

First, we analyzed the motivations and draw factors (incentives) that persuade users to engage with the system, which is an essential requirement for the growth of user-generated dialogue content in society.

We classified factors that motivate and attract users into the following four types according to the user's degree of involvement with the content, and we analyzed their respective interrelationships and movements (Fig. 3.34).

- Potential user—someone who has never used the system.
- Consumer—someone who is using or has used the system. Passive user.
- Participant—someone who actively engages with the system while commenting, favoriting and evaluating. Active user.
- Producer—someone who creates and posts content.

The requirements that should be met by a spoken dialogue system so that dialogue content can be created in a network environment as user-generated media for our system are as follows:

- From potential users to consumers: expanding the dialogue content playback environment to multiple devices including smart phones.
- From consumers to participants: stimulating interest in the overall system by providing periodic information by SNS or by designing user flow lines
- From participants to producers: using facilities reachable for users to promote them to create dialogue content on a particular theme (Fig. 3.35).

We built a multi-platform question-and-answer type dialogue content distribution/registration system incorporating the above improvements, and we used it to perform social experiments (Fig. 3.36). The content server on which the dialogue content was stored was placed, and a multiplatform (Windows/Linux/Android) system was provided to facilitate access. In this experiment, we prepared several flow line to draw users into the usage of the system and support system to attract ordinary users, and we performed public testing for one month. About 30,000 people found out about the system via tweets on Twitter, and we obtained 6,300 dialogues and 232 newly registered dialogue content registrations. 55% of the users participated from Android devices, and 15% performed light content registration by participating in the provision of "themes" using the Twitter library functions. In this way, we were able to confirm the effectiveness of the proposed method and accumulate valuable materials relating to user trends.

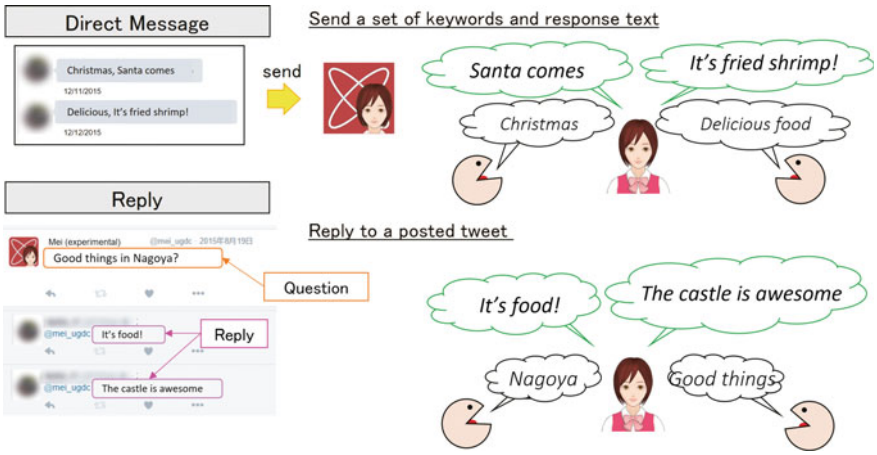

Fig. 3.35 Themed dialogue content creation

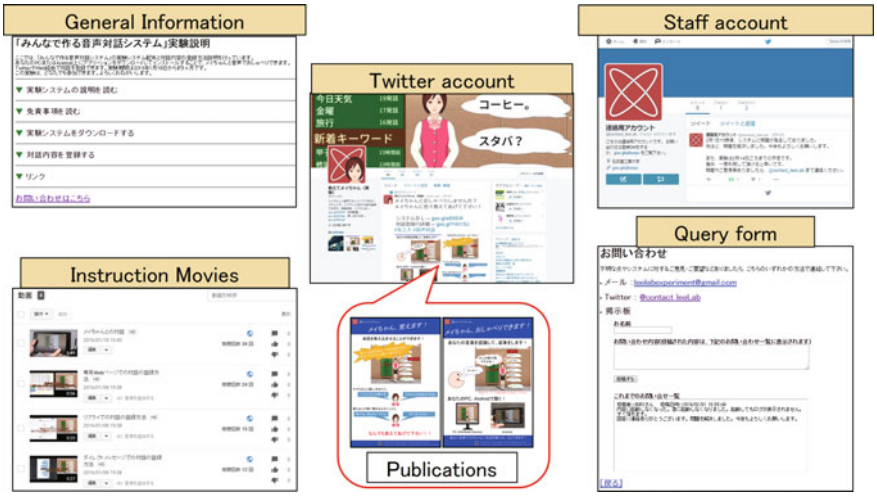

Fig. 3.36 Design of participation flow lines and description pages for ordinary users in the social experiment

3.6.4 Dialogue Content Sharing Service: MMDAgent SHARE

No framework was available for the sharing of dialogue content between the creators and users of this content, and the created dialogue content ended up being distributed to various different locations. We therefore considered that providing a place where dialogue content can be presented and shared would lead to more dialogue content being produced. Specifically, we launched a dialogue content sharing ser-

Fig. 3.37 Dialogue content sharing service

vice “MMDAgent SHARE” as a framework to facilitate sharing of dialogue content between users (Figs. 3.37 and 3.38). In general, there are various difficulties that may need to be addressed in order to share dialogue content easily, but the proposed system solves these problems by using the methods described below.

1. Since dialogue content often consists of a set of many files and is cumbersome to work with, we defined our own MMDA format that allows them to be used as a single file.
2. We created a function whereby MMDA format files can be produced easily on the server.
3. We developed a new MMDAgent installer for Windows to make it possible to run MMDA format content with a single click.
4. To make the service easy to use, we enabled cooperation with Open IDs services such as Google accounts.
5. We developed a framework that automatically detects script errors and warnings when content is posted.

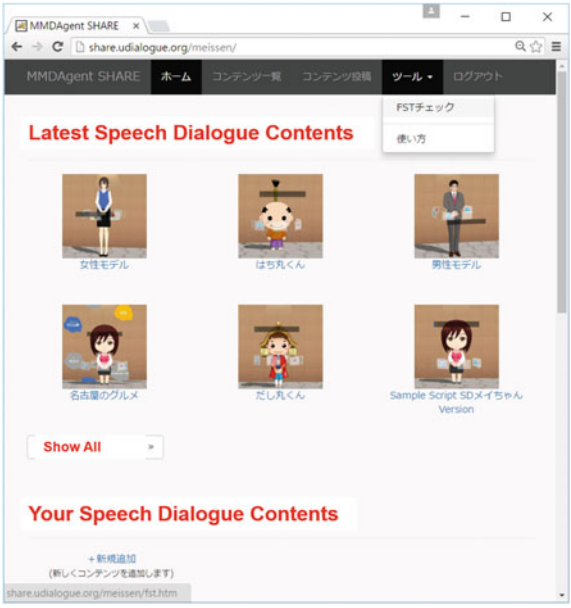

Fig. 3.38 Shared dialogue content

6. We are carefully preparing a user agreement and privacy policy for this service.
7. We developed some features that are very useful for copyright holders, such as a framework whereby all the attached text included in the content (README files, etc.) can be automatically checked from this service.

We also built a prototype service that makes it easy to create Web pages containing dialogue content. This framework facilitates the creation of dialogue content by allowing the user to input data such as images, keywords and text via an ordinary Web browser (Fig. 3.39).

3.7 Encyclopedia MMDAgent

So far, in addition to developing the dialogue content production support tool and underlying software, including MMDAgent, we have also conducted numerous demonstration experiments both on and off campus, written review papers, held MMDAgent workshops, and published tips on an Internet blog, and we have completed a set of materials on the use of MMDAgent and the production of dialogue content including guide books/tutorials, lecture slides, reference manuals, and sample scripts. By integrating these achievements and carrying out further expansion and maintenance, we constructed an all-in-one "Encyclopedia MMDAgent" package that includes a set of software, manuals, basic dialogue libraries and dialogue content design guidelines. This package includes the following items:

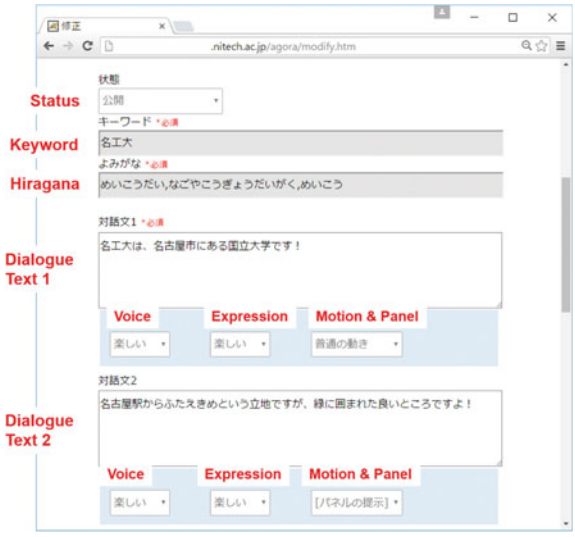

Fig. 3.39 Dialogue content creation service

1. The MMDAgent toolkit for building interaction systems (Windows, Mac OS, Linux, Android OS and iOS compatible)
2. MMDAgent Primer (Japanese and English)
3. MMDAgent creators reference manual (Japanese and English)
4. MMDAgent developers reference manual (Japanese and English)
5. MMDAgent lecture slides (Japanese and English)
6. MMDAgent lecture videos (MOOC, OCW)
7. Voice interaction content creation support tools (Web application for editing dialogue content, tablet application for editing dialogue content, voice interaction builder, etc.)
8. Dialogue content library (basic dialogue library, sample 3D models, speech synthesis models, etc.)
9. Dialogue content design guidelines

This package comprises a coordinated collection of multi-platform software, user-oriented content production support tools, dialogue content design guidelines based on the results of long-term demonstration experiments, reference manuals and lecture slides for content creators and developers, tutorial videos for MOOC/OCW courses, and a library of dialogue content, providing an environment where users can easily create their own content.

3.8 Future Prospects

The essential differences between the services offered by the leading IT companies and the user content creation concept discussed in this project are tabulated in Fig. 3.40. The services of leading IT companies have a top-down structure where users can only use content within a framework provided by the company. Recently leading IT companies often have teams responsible for the creation of spoken dialogue content, giving rise to a workflow whereby the construction of the spoken dialogue system is treated as a content production issue. However, commercial services are only able to progress within their own closed worlds, and it is not possible to share the generated content over as wide a range as user-generated content in the true sense. On the other hand, when taking the user content generation approach, it becomes possible to share content without having to rely on any particular service, and any type of content can be produced and shared without any constraints. The content and knowledge that is shared and integrated in this sort of way becomes a form of common property, and it is expected that this shared content will be divided into different content types, and that new types of content will be discovered. This sort of content and knowledge may of course be used to provide feedback to services offered by the leading IT industries, and in addition is also expected to fuel the growth of completely new modes of use and business applications (bottom-up approach).

To facilitate user creation in the abovementioned true sense over a broad range, including industrial applications, it is necessary to have a platform for this purpose, which could perhaps be fulfilled by the encyclopedia MMDAgent that brings together the MMDAgent of this project and the related achievements. Furthermore, if the user content creation shown in Fig. 3.40 can be deployed while expanding into industrial applications, then it should be possible to generate a diverse range of technical research topics as shown in Fig. 3.41. In other words, it is liable to open up the new academic field of spoken dialogue content studies, and the research achievements of this project are expected to provide a platform for this sort of expansion.

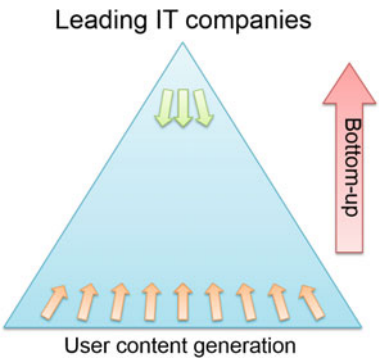

Fig. 3.40 Bottom-up approach—opening up voice interaction systems to the people

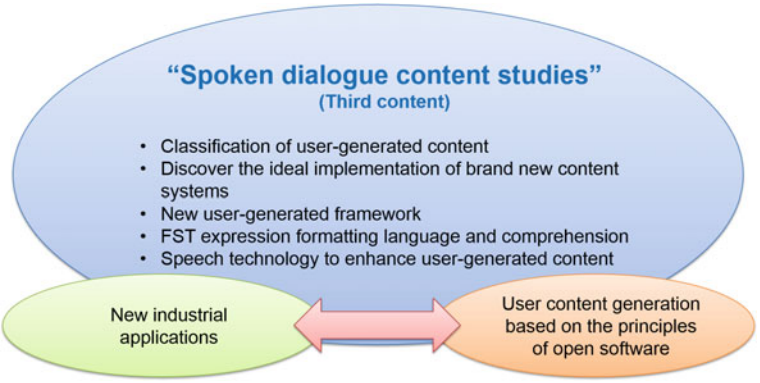

Fig. 3.41 What sort platform will emerge from the next stage of research based on the current research achievements?

3.9 Conclusion

In this study, by separating spoken dialogue systems into the content provided by users and the systems that drive this content, we have created a content creation framework based on enhanced technology where users and creators with abilities close to those of users can be expected to create a continuous supply of attractive dialogue content. We have also conducted demonstration experiments to show how this system can be used to create content. Based on the number of times the software tools have been downloaded and on the diverse uses of this software on the Internet, we have seen a lot of content being created. We have created and published an "Encyclopedia MMDAgent" that integrates all the results obtained so far. Furthermore, we have built a content sharing server in order to analyze this content. In the future, we hope that by operating this content sharing server, we will encourage the creation of more attractive content while gaining further insights into the system's use.

In this study, we set out to consider speech technology from the new perspective of building an environment for the creation of user-generated dialogue content. We are hopeful that it will not only yield useful insights into the creation of dialogue interfaces, but will also lead to future breakthroughs in the spread of voice interfaces. In fact, the latest spoken dialogue systems are increasing their attractiveness by engaging in witty exchanges with users, based on content produced by people who could be called scenario writers. This sort of situation is consistent with the outlook presented in this chapter. Also, the implementation and testing of digital signage in public spaces is an embodiment of a new kind of ubiquitous information environment, which could soon become more widespread and more commercialized. In the future, if it becomes possible to collect large numbers of actual dialogue samples and examples of dialogue content, then it could become possible to perform statistical modeling of dialogue based on these large data sets.

References

1. HTS: HMM speech synthesis toolkit, http://www.hts.nitech.ac.jp/
2. Open JTalk: Japanese text-to-speech system, http://open-jtalk.sourceforge.net/
3. Julius: Open-source large vocabulary continuous speech recognition engine, http://julius.sourceforge.jp/
4. MMDAgent: Toolkit for building voice interaction systems, http://www.mmdagent.jp/
5. T. Funayachi, K. Oura, Y. Nankaku, A. Lee, K. Tokuda, A simple dialogue description based on finite state transducers for user-generated spoken dialog content, in *Proceedings of ASJ 2013 Autumn Meeting, 2-P-28*, pp. 223–224, 25–27 Sept 2013. (in Japanese)
6. K. Nakamura, K. Hashimoto, Y. Nankaku, K. Tokuda, Integration of spectral feature extraction and modeling for HMM-based speech synthesis. IEICE Trans. Inf. Syst. **E97-D**(6), 1438–1448 (2014)
7. S. Takaki, Y. Nankaku, K. Tokuda, Contextual partial additive structure for HMM-based speech synthesis, in *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, pp. 7878–7882, 2013
8. R. Dall, M. Tomalin, M. Wester, W. Byrne, S. King, Investigating automatic & human filled pause insertion for speech synthesis, in *Proceedings of Interspeech*, 2014
9. S. R. Gangireddy, S. Renals, Y. Nankaku, A. Lee, Prosodically-enhanced recurrent neural network language models, in *Proceedings of Interspeech 2015*, Dresden, Sept 2015
10. K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, The effect of neural networks in statistical parametric speech synthesis, in *Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, pp. 4455–4459, 19–24 Apr 2015
11. S. Takaki, S. Kim, J. Yamagishi, J.J. Kim, Multiple feed-forward deep neural networks for statistical parametric speech synthesis, in *Proceedings of Interspeech*, vol. 2015, pp. 2242–2246, 2015
12. K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, Trajectory training considering global variance for speech synthesis based on neural networks, in *Proceedings of 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)*, Shanghai, China, pp. 5600–5604, 20–25 Mar 2016
13. K. Sawada, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Evaluation of text-to-speech system construction for unknown-pronunciation languages," Technical Report of IEICE, vol. 115, no. 346, SP2015-80, pp. 93–98, 2–3 Dec 2015
14. S.R. Gangireddy, Q. Huang, S. Renals, F. McInnes, J. Yamagishi, in *Topic Model Features in Neural Network Language Models*, (UK Speech Meeting, 2013)
15. CSTR VCTK Corpus, http://www.udialogue.org/ja/download-ja.html
16. D. Yamamoto, K. Oura, R. Nishimura, T. Uchiya, A. Lee, I. Takumi, Keiichi Tokuda, Voice interaction system with 3D-CG human agent for Stand-alone smartphones, in *Proceedings of the 2nd International Conference on Human Agent Interaction* (ACM digital library, 2014), pp. 320–330
17. K. Wakabayashi, D. Yamamoto, N. Takahashi, A voice dialog editor based on finite state transducer using composite state for tablet devices, computer and information science 2015. Stud. Comput. Intell. **614**, 125–139 (2016)
18. R. Nishimura, D. Yamamoto, T. Uchiya, I. Takumi, Development of a dialogue scenario editor on a web browser for a spoken dialogue system, in *Proceedings of the Second International Conference on Human-agent Interaction*, pp. 129–132, 2014
19. Y. Matsushita, T. Uchiya, R. Nishimura, D. Yamamoto, I. Takumi, Crowdsourcing environment to create voice interaction scenario of spoken dialogue system, in *Proceedings of the 18th International Conference on Network-Based Information Systems (NBiS-2015)*, pp. 500–504, 2015
20. Y. Matsushita, T. Uchiya, R. Nishimura, D. Yamamoto, I. Takumi, Experiment and evaluation of crowd sourcing model for creation of voice interaction scenario. Proc. IEEE GCCE **2015**, 321–322 (2015)

21. T. Uchiya, R. Nakano, D. Yamamoto, R. Nishimura, I. Takumi, Extension with intelligent agents for the spoken dialogue system for smartphones. Proc. IEEE GCCE **2015**, 298–299 (2015)
22. T. Uchiya, D. Yamamoto, M. Shibakawa, M. Yoshida, R. Nishimura, I. Takumi, Development of spoken dialogue service based on video call named "Mobile Meichan". Proc. JAWS2012 (2012). (in Japanese)
23. T. Uchiya, M. Yoshida, D. Yamamoto, R. Nishimura, I. Takumi, Design and implementation of open-campus event system with voice interaction agent. Int. J. Mob. Multimed. **11**(3, 4), 237–250 (2015)
24. R. Nishimura, K. Sugioka, D. Yamamoto, T. Uchiya, I. Takumi, A VoIP-based voice interaction system for a virtual telephone operator using video calls. Proc. IEEE GCCE **2014**, 529–532 (2014)