User Generated Dialogue Systems: uDialogue

Keiichi Tokuda[1] tokuda@nitech.ac.jp
Akinobu Lee[1] ri@nitech.ac.jp
Yoshihiko Nankaku[1] nankaku@nitech.ac.jp
Keiichiro Oura[1] uratec@nitech.ac.jp
Kei Hashimoto[1] hashimoto.kei@nitech.ac.jp
Daisuke Yamamoto[1] yamamoto.daisuke@nitech.ac.jp
Ichi Takumi[1] takumi@nitech.ac.jp
Takahiro Uchiya[1] t-uchiya@nitech.ac.jp
Shuhei Tsutsumi[1] tsutsumi.shuhei@nitech.ac.jp
Steve Renals[2] s.renals@ed.ac.uk
Junichi Yamagishi[3] jyamagis@nii.ac.jp

[1]Nagoya Institute of Technology
[2]The University of Edinburgh
[3]National Institute of Informatics

Abstract

This article introduces the idea of user-generated dialogue content and a framework that makes it workable in practice, and describes our experiments aimed at clarifying the conditions that must be met by such systems. One of the appealing points of a spoken interface is that it provides a vivid sense of interactivity that cannot be achieved with a text interface alone. In this study, we will separate dialogue systems into content that can be produced and modified by users, and the systems that drive this content, and we will seek to clarify (1) the requirements of systems that enable the creation of engaging dialogue systems, and (2) the conditions for the sequential generation of engaging dialogue content by users, while at the same time establishing a framework for doing so. For bidirectional digital signage with voice guidance installed in a public space as a dialogue device, we implement an Internet-based content generation environment, and we perform a validation experiment where dialogue content is generated by users. The framework for dialogue content generation proposed in this study is expected to lead to a breakthrough in the spread of spoken content.

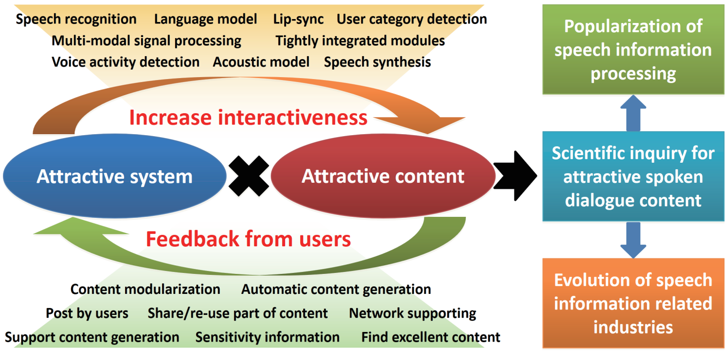


Fig. 1. Circulatory system of content creation

1 Introduction

A human-centered information environment is an environment where everyone is a source of information, and is able to enjoy information naturally. Speech is the most basic form of communication for humans. Due to the ubiquitous nature of advanced network telecommunication equipment, we are able to speak naturally and freely to other people wherever they may be. The widespread availability of environments where people can interact is one of the ideals of modern society. But 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 information 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 “appeal” of 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. But this can only be achieved by entering into regions where high-level human speech processing capabilities are required, such as 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 dialogue systems into content that can be produced and modified by users and the systems that drive this content. In this way, we hope to clarify what sort of requirements must be met to produce “engaging” content and systems so that speech technology can spread widely between people (Fig. 1). However, engagement is created through the combination of human feelings and knowledge, and is not something that can easily be evaluated mechanically. Therefore, after establishing a framework that makes it easy for users to create and evaluate large quantities of dialogue content, we have to inductively ascertain these qualities. In this study, the basic strategy is to construct a “circulatory system” of content generation as shown in Fig. 1 and empirically clarify the factors for achieving a loop gain of at least 1 in order to establish techniques for the construction of frameworks that make it easy to achieve this sort of state.

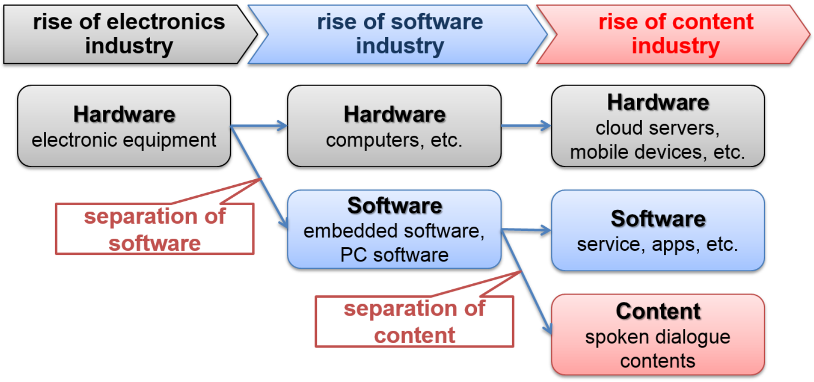

rise of electronics industry
rise of software industry
rise of content industry
Hardware electronic equipment
Hardware computers, etc.
Hardware cloud servers, mobile devices, etc.
separation of software
Software embedded software, PC software
Software service, apps, etc.
separation of content
Content spoken dialogue contents

Fig. 2. The evolution of industry structure related to telecommunications

2 Background and purpose of the study

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 production parts from the software production parts, and make them widely available to creators and ordinary users. This process resembles the evolution of industrial structures shown in Fig. 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 in their own right. The mobile games industry is a typical example, where the separation of content creation is progressing. To bring about this sort of change with regard to speech technology, it will be necessary to accelerate the creation of engaging content by getting as many creators and users as possible involved in content creation.

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 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 is that it can be grasped as a version of dialogue content, which to aims implement an information environment where users disseminate information (Fig. 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

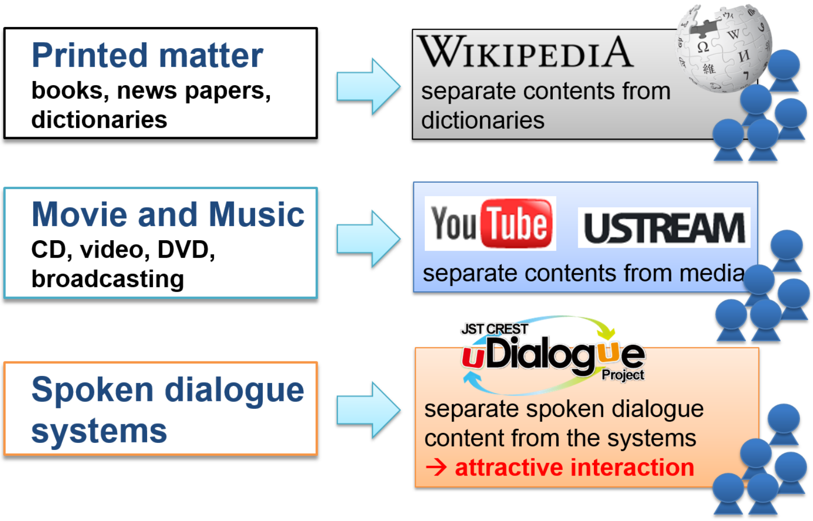

Printed matter
books, news papers, dictionaries
Wikipedia
separate contents from dictionaries
Movie and Music
CD, video, DVD, broadcasting
YouTube
USTREAM
separate contents from media
Spoken dialogue systems
JST CREST uDialogue Project
separate spoken dialogue content from the systems
→ attractive interaction

Fig. 3. Separation of content

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 content such as video that can be changed at any time by 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 bidirectional 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 mechanisms for cooperation via the Internet.

3 MMDAgent: A toolkit for the construction of voice interaction systems

To comprehensively research the various elements of voice interfaces such as their level of engagement with users, these systems need to develop into areas where human judgment 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 agent representation units. Furthermore, since data has to be collected performing demonstration experiments of various tasks in various situations, we require an advanced and flexible system where users and system developers can each work

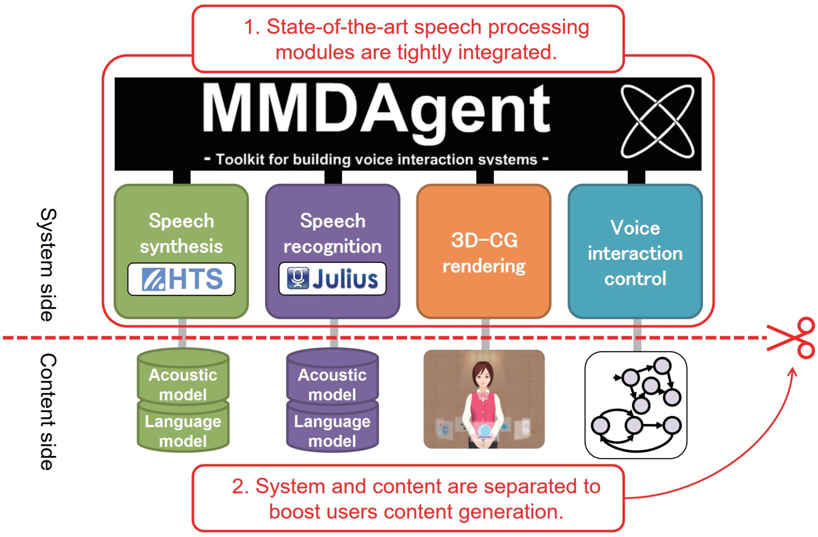

1. State-of-the-art speech processing modules are tightly integrated.
MMDAgent
- Toolkit for building voice interaction systems -
System side
Speech synthesis
HTS
Speech recognition
Julius
3D-CG rendering
Voice interaction control
Content side
Acoustic model
Language model
Acoustic model
Language model
2. System and content are separated to boost users content generation.

Fig. 4. Voice interaction system construction toolkit (MMDAgent)

freely on each part of the system and dialogue content.

So far, we have continuously developed and published open-source research platform software that cooperates with speech technology including HTS (an HMM-based speech synthesis system), Open JTalk (a Japanese text-to-speech system based on HTS), and Julius (a speech recognition engine). Based on this group of software, we built a voice interaction system construction toolkit (MMDAgent) by implementing and combining a new graphic display unit that can draw an agent's expressions in real time with advanced 3D computer graphics including physical computations, and an interaction description unit based on a finite state transducer (FST), and we published it as free open-source software (Fig. 4). This system uses open formats not only for the speech recognition and speech synthesis dictionaries and models, but also for the 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 engaging dialogue content. This is the main feature of MMDAgent. MMDAgent allows richly expressive dialogue adn computer graphics to be produced with high speed and precision, and is used as the technical platform of this study.

In this study, we use MMDAgent as a platform for the creation of a voice interaction system construction toolkit, and we separate dialogue systems into the content provided to users and the systems that drive this content. Our aim is to clarify what is needed to create systems and content that are sufficiently engaging to ensure that speech technology can spread widely to many people.

There are many issues that need to be addressed in order to achieve this goal, but they can be summarized in the following three categories:

1. Enhancement of the underlying technology and software
2. Creating a framework for content creation
3. Performing content creation demonstration experiments

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 engaging dialogue content is created from the creation/sharing/evaluation of dialogue content by users.

4 Enhancing the underlying technology

To create engaging dialogue content, it is essential to provide a technology infrastructure where 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 dialogue system, and to optimize the underlying software for representing the dialogue system, and the design of the corpus/agents.

4.1 Underlying technology

To optimize the underlying technology, we performed a lot of basic research on techniques 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 speech synthesis [10]
  In recent years, speech synthesis has been performed using statistical models called hidden Markov models (HMMs) to model the acoustic relationships between speech characteristics and language characteristics. In this study, by integrating the extraction of features from speech into the HMM learning process, we were able to directly model the speech waveforms with a unified standard. This improved the system's overall modeling capabilities, and greatly improved the quality of synthesized speech.

- Improvement of spectral modeling with an additive structure [11]
  We proposed an additive model for speech synthesis by assuming additive structures in the relationship between language characteristics and speech characteristics. We also proposed a learning algorithm for the efficient learning of additive models. The quality of speech synthesis performed by spectral modeling with an additive model was greatly improved.

- Conversational speech synthesis method [12]
  Dialogue between people often contains hesitations and filler words (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 model [13]

For high-performance speech recognition, studies are being performed where recurrent neural networks are used for language models that model linguistic characteristics. In this study, we proposed a learning method based on a neural network 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 characteristics as additional information.

- Expansion of conversation declaration language in the MMDAgent toolkit for constructing dialogue interaction systems [20]
  3 dialogue descriptions in MMDAgent are made using a finite state transducer. A finite state transducer is represented by state transitions and symbol inputs and outputs, but if the specification is expanded so it can handle variables and matching regular expressions, it can be used to describe complex conversations with simple scripts.
- Improvement of speech synthesis based on deep neural networks [14, 15, 16]
  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.
- Methods for the construction of text-to-speech systems in languages where the information being spoken is unknown [17]
  Ordinary speech synthesis consists of a text analysis unit that predicts how text should be read, and a waveform generation unit that creates the corresponding speech waveforms, but it is not possible to construct a text analysis unit for languages whose pronunciation information is unknown. We therefore proposed a method for the construction of text-to-speech systems for speech information in an unknown 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.
- Improvement of speech synthesis based on an analysis of factors expressing diverse tones of voice [18]
  A method involving the introduction of factor analysis in speech synthesis has previously been proposed for the representation of diverse tones of voice. In this study, we improved the quality of synthesized speech by making efforts to improve the selection of model learning and model structures in speech synthesis methods based on factor analysis.
- Analysis of dialogue content using a topic model
  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).

We have also worked on improving various fundamental technologies that are needed by dialogue systems, such as speech synthesis that can express diverse emotions, and acoustic model adaptation methods.

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].

- Voice interaction system construction toolkit MMDAgent
  (http://www.mmdagent.jp/)
  (61,000 downloads)
- Speech recognition engine Julius
  (http://julius.sourceforge.jp/)
  (216,000 downloads)
- HMM speech synthesis toolkit HTS
  (http://hts.sp.nitech.ac.jp/)
  (371,000 downloads)
- HMM speech synthesis engine hts_engine API
  (http://hts-engine.sourceforge.net/)
  (41,000 downloads)
- Japanese speech synthesis system Open JTalk
  (http://open-jtalk.sourceforge.net/)
  (47,000 downloads)
- Speech signal processing toolkit SPTK
  (http://sp-tk.sourceforge.net/)
  (42,000 downloads)
- Japanese singing speech synthesis system Sinsy
  (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 voice interaction system construction toolkit 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 dialogue system with small response delays to be used in smartphones and tablets (Figs. 5, 6). This is thought to be the first open-source implementation of a 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 download figures show, they have already achieved the status of de facto standards. In fact, as shown in Fig. 7, it is being widely used in many different situations, including academic papers, software development and events.

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

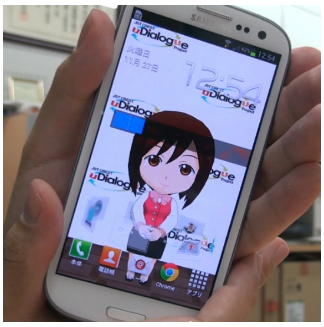


Fig. 5. Android-compatible MMDAgent

Fig. 6. iOS-compatible MMDAgent

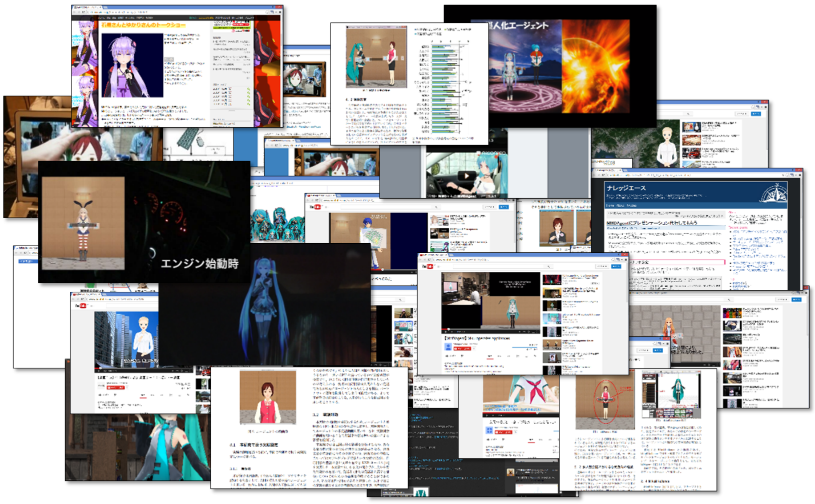


Fig. 7. Examples of the usage of open-source software

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 dialogue system targeting Japanese and British English, and by performing experiments with the Japanese and English 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 dialogue system.

- Japanese male voice actor corpus

- CSTR VCTK corpus [19]
  This corpus is a large-scale speech corpus containing about 60 hours of speech by 109 speakers with various British English accents. The

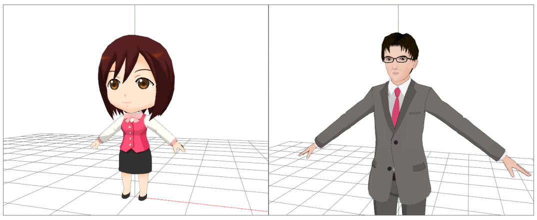


Fig. 8. Animatic dialogue agent

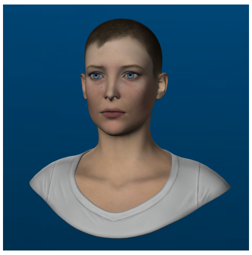


Fig. 9. Realistic dialogue agent

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. There are also over 100 minutes of spontaneous conversation recordings. Furthermore, there are 400 recordings of the same sentence spoken in four different styles such as normal speaking and fast speaking.

Next, we created a conversation agent suited to dialogue. To make the 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 an agent with a height of 2.5 heads, and a male agent (Fig. 8). 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 dialogue system (Fig. 9).

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. 10, 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.

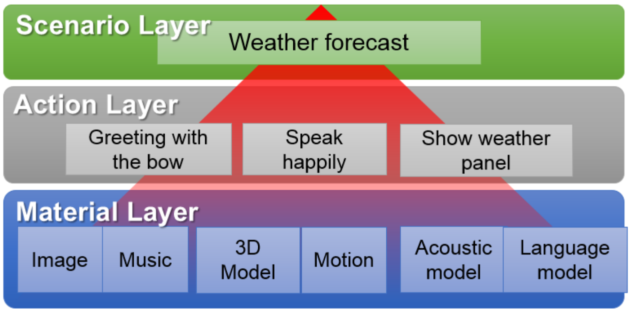


Fig. 10. Hierarchy of dialogue content

As shown in Fig. 11, we also envisaged that the 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 smartphones [9].

5.1 Building content creation support tools

10 In the MMDAgent toolkit for the construction of dialogue interaction systems, FST script files are used to control the dialogue scenarios. However, it is difficult for ordinary users to create FST script files directly. We therefore implemented a mechanism that allows dialogue content to be edited easily.

Interactive script creation tool

To facilitate the creation of user-generated dialogue content, it is essential that users can easily create dialogue content via a network. We therefore developed the EFDE dialogue content creation tool (Fig. 12) and the MMDAE dialogue content creation tool for advanced content creators using web browsers (Fig. 13).

- EFDE dialogue content creation tool [7]
  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 ordinary users can easily edit scripts while checking the action of the dialogue content. Also, by providing commonly used FST scripts as templates, we have made it possible to implement complex dialogues with a small number of states. Furthermore, by adding functions that allow users to input dialogue sentences and keywords by speaking, we arrived at a framework where dialogue content can be created easily by speaking and using a touch interface.
- MMDAE dialogue content creation tool [1]
  An interface for advanced users, where it is possible to use a Web browser to perform detailed edits on the FST scripts. By supporting the input of FST scripts, editing can be performed easily by assembling these scripts. Also, since this tool is accessed via a Web

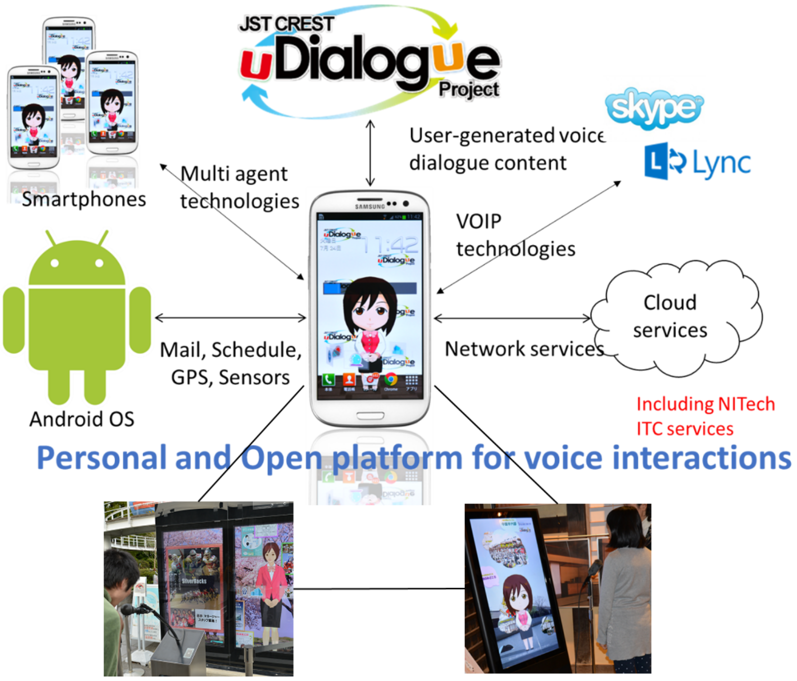


Fig. 11. Linking dialogue systems with a network

browser, it can run on a wide variety of platforms, and enables the provision of an environment where content can be created easily.

Voice interaction builder

To promote the creation of engaging 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

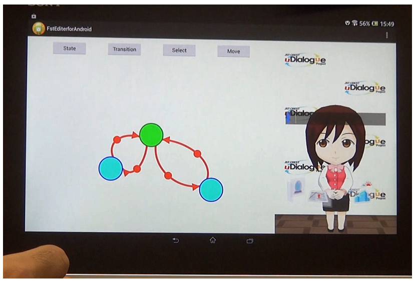


Fig. 12. Tablet interface for editing dialogue content

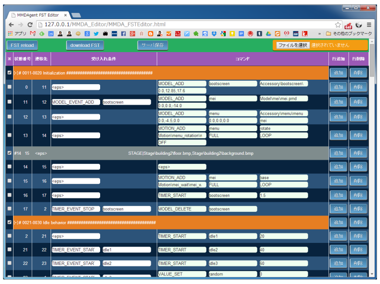


Fig. 13. Web browser interface for editing dialogue content

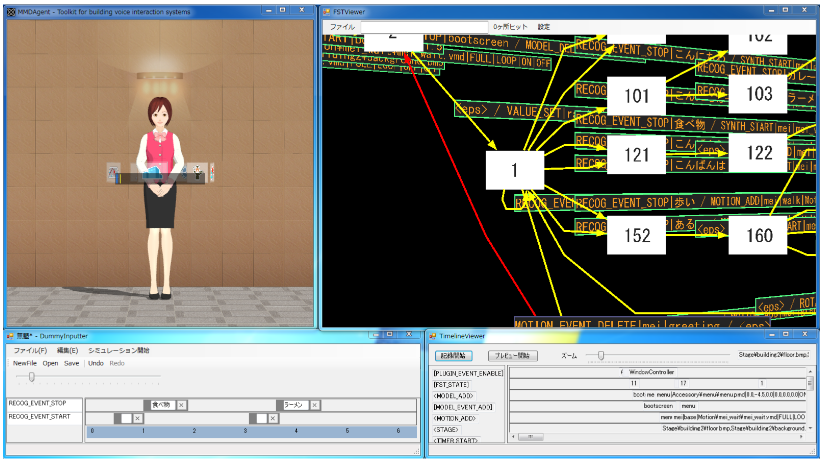


Fig. 14. Voice interaction builder

action of this content can be verified (Fig. 14). 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 checking 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.

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 collaboratively 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.

Construction of a collaborative editing system for dialogue content with a history function

When one considers the construction of environments and systems for the collaborative construction of dialogue content, one of the simplest forms is a sort of “chatterbot” that operates based on a set of questions and answers recorded by the user and shared via the Web. However, this format can only handle simple question-and-answer exchanges because it retains no history of previous questions, and is thus clearly inadequate as a system intended to provide fun and engaging dialogue with continuity and situational awareness. On the other hand, it is difficult to anticipate all the responses that a user might give during a lengthy conversation.

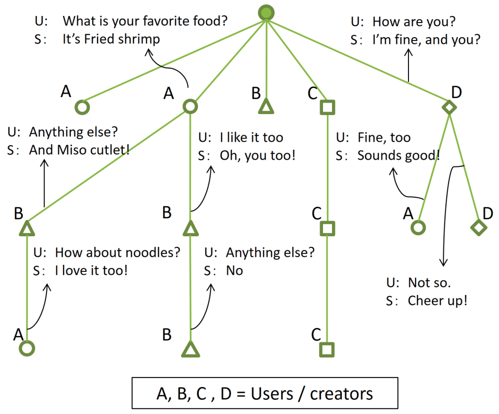
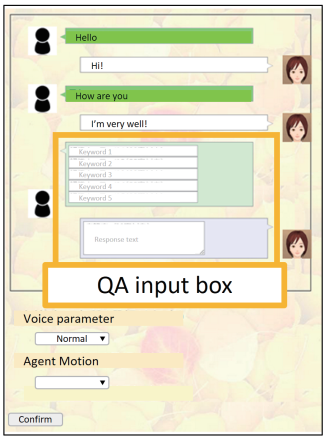


Fig. 15. Concept of interactive content with history


Hello

Hi!

How are you

I'm very well!

Voice parameter: Normal ▾

Agent Motion: ▾

Fig. 16. Registration interface for interaction with history

We have therefore developed a user-generated dialogue system makes it easy to collaboratively construct conversations with a history function (Fig. 15,16). Conversations are recorded in units of keyword/response sets, but can also include parent-child relationships, whereby users are able to collaboratively construct multiple consecutive question-and-answer type interactions. The recorded dialogue content is all stored in a tree structure consisting of keyword/response pairs as its building blocks. In the recording window, we added features such as an SNS chat-style Web interface where it is possible for anyone to intuitively grasp and edit the flow of a conversation. This system can be used to create long continuous conversations, and allows dialogue recorded by other users to be branched off into other conversations or augmented at a later date. 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 original engaging conversations.

Construction of a cloud-based dialogue content editing environment

To increase the vocabulary of conversations in MMDAgent, it is generally necessary to describe a larger number of conversation scenarios. However, it is difficult for humans to create conversation scenarios on a large scale. We therefore developed a crowd sourcing system specialized for the creation of dialogue scenarios (Fig. 17). In this system, multiple users receive orders for dialogue scenario creation tasks based on a crowd sourcing concept. 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 for operating and checking agents. In an experimental evaluation of this system, we confirmed its validity for the creation of dialogue scenarios based on crowd sourcing [2, 3].

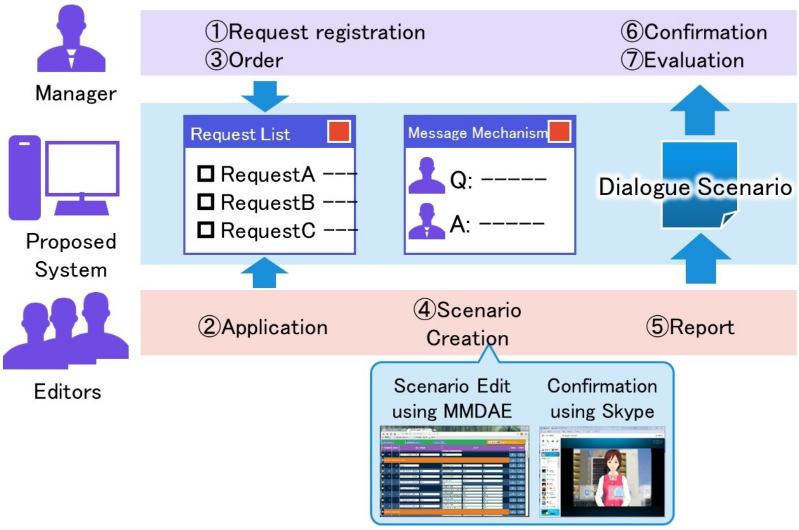


Fig. 17. Crowd sourcing system specialized for the creation of dialogue scenarios

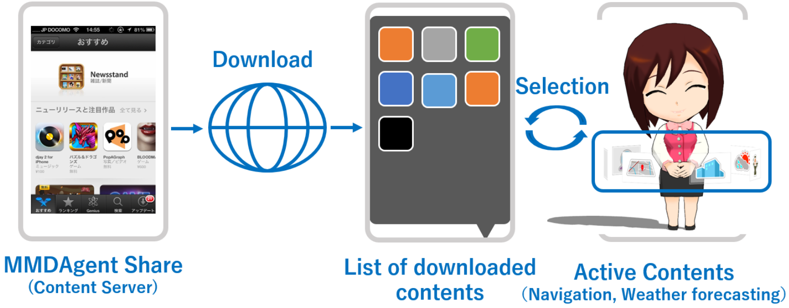


Fig. 18. Packaging of dialogue content

5.3 Constructing a platform for modularization and inter-agent collaboration

Packaging dialogue content

In general, dialogue content is only created individually by users and engineers, and there has not been much consideration of how they can be circulated effectively. 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 based on multiple concurrency by extending FST scripts. This made it possible to deliver partial updates and functional units of FST scripts (which was hitherto difficult), 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 that packages dialogue content as shown in Fig. 18.

We also proposed and built a user-generated dialogue system architec-

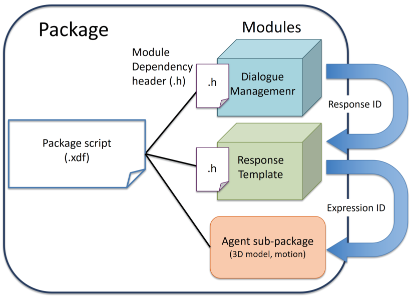


Fig. 19. Converting dialogue modules into packages

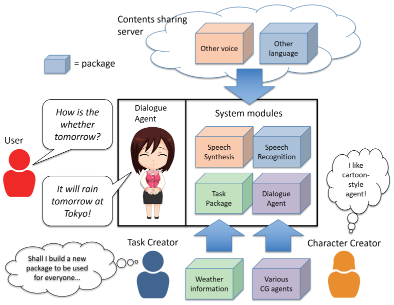


Fig. 20. Module-based sharing of dialogue content

ture 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 (Fig. 19, 20). By using 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 modules. 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 header declarations.

A platform for cooperation between dialogue systems using network agent technology

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 dialogue systems based on agent/NFC/Bluetooth technology (Fig. 21) [4]. We made a prototype scheduling system to confirm the validity of services involving cooperation between multiple 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.

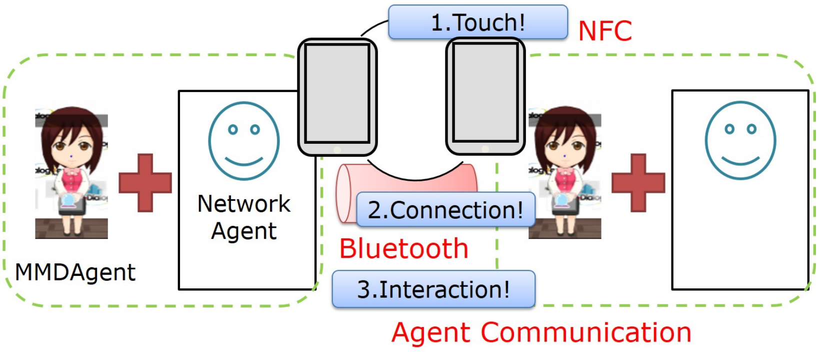


Fig. 21. P2P network connection mechanism

6 Content creation validation experiment

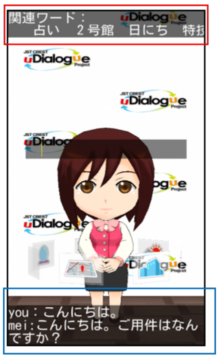
In this section, we discuss a validation experiment performed using the proposed system.

6.1 Building a framework to promote the use of content

In user-generated dialogue content, although it can be expected that a large quantity of diverse dialogue content will be created by users, it is difficult to maximize the usefulness of this content by organizing it into a suitable taxonomy. Here, we discuss the results of a study in which we implemented techniques such as content recommendation and searching to research the analysis, categorization and correlation of dialogue content groups in order to support the creation and use of dialogue systems.

Related word recommendations based on the usage history of a dialogue system

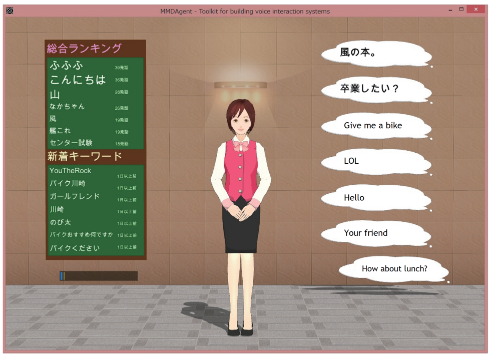
With the aim of promoting the use of dialogue systems by beginners with no experience in the use of dialogue systems, we studied a framework for presenting the user with a group of candidates for the next keyword that should be said. We then implemented a technique for acquiring relevant words by analyzing the user's conversation history stored on the server by means of an information recommendation technique or the like (Fig. 22). In user testing, we obtained results suggesting that the presentation of related keywords helps users make effective use of 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.



関連ワード：
占い　２号館　日にち　特

you：こんにちは。
mei:こんにちは。ご用件はなんですか？

Fig. 22. Related word presentation method


MMDAgent - Toolkit for building voice interaction systems

総合ランキング

ふふふ	39発話
こんにちは	36発話
山	28発話
なかちゃん	26発話
風	19発話
艦これ	19発話
センター試験	18発話

新着キーワード

YouTheRock	1日以上前
バイク川崎	1日以上前
ガールフレンド	1日以上前
川崎	1日以上前
のび太	1日以上前
バイクおすすめ何ですか	1日以上前
バイクください	1日以上前

風の本。
卒業したい？
Give me a bike
LOL
Hello
Your friend
How about lunch?

Fig. 23. Sharing conversation histories and displaying dialogue content rankings

Strengthening of mutual incentives by sharing usage histories between users

In user-generated media, a phenomenon is observed whereby sharing activity histories between users causes mutual stimulation leading to increased usage of the system and increased content creation. In dialogue content, sharing interactions with the dialogue system and content usage histories between users is expected to mutually increase the willingness of users to use the system. Therefore, in a question-and-answer style dialogue system with user participation, we built a framework that provides stronger incentives for interaction by (1) continuously providing online information about what users are saying, and (2) recording and displaying rankings of the number of times a conversation is played back and the number of times keywords are uttered. As shown in Fig. 23, when this information was presented to the users and creators of dialogue content, it became possible for the creators to learn the keywords that are needed by users. We also implemented a framework that makes it easy for content creators to record the conversations that users need based on the ranking results. In user testing, we confirmed that these frameworks can lead to an improvement in the motivation (incentivization) of both users and content creators.

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 was also made available as a data set that can be used to shed light on methods for making dialogue content more engaging. Some typical examples are listed below:

Demonstration experiment in a university campus

Using our MMDAgent toolkit for the construction of voice interaction systems, we constructed an all-weather bidirectional voice guidance digital signage system ("Mei-chan") and installed it in an open space just

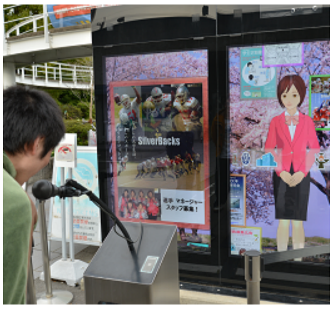


Fig. 24. "Mei-chan": Digital signage with two-way voice guidance

Fig. 25. Nagoya Institute of Technology main gate

outside the university's main gate (Fig. 24, 25). 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. 26). The users are assumed to be graduates, undergraduates and faculty staff.

The system went into operation on April 6, 2011, and on June 15, 2011, the content registration system was published within the campus. 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 27 shows an example of the submitted panel display. Since September 2014, the system has also been installed in the open space ("Yume Room") at the Nagoya Institute of Technology (Fig. 28, 29). 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.

We also used a Moodle questionnaire to conduct a survey of Mei-chan's name recognition and frequency of use among students. Of the 262 valid responses, nearly everyone (99%) were aware of Mei-chan. A high percentage (77%) had also attempted conversations, showing that the dialogue system using MMDAgent is highly practical. The free response questionnaire drew a response rate of 33%, showing that one out of three users actively offered suggestions on how to improve the dialogue system, etc.

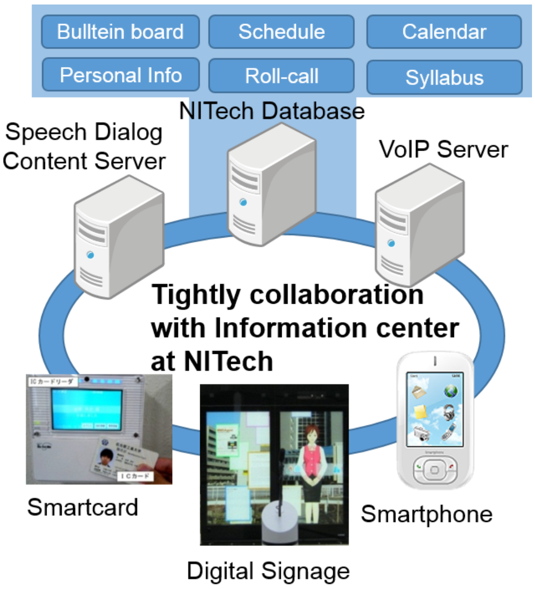


Fig. 26. Cooperation with the information infrastructure system

Demonstration experiment outside the university

During the 2013 school year, we produced a new indoor digital signage prototype with bidirectional voice guidance, and we installed it at a tourist information office in Handa city (Fig. 30, 31). A small character is placed at the bottom of the screen, and information panels and the like are shown at the top of the screen. This makes the equipment comfortable to use even in a confined space. The users were ordinary visitors to the premises, who were mostly tourists.

We also implemented a framework whereby users (in this case, the tourist office staff) could use their smart phones or Web browsers to add and update the dialogue content with ease, and this enabled them to use their own ingenuity to make the tourist information more interesting and suited to their location. In fact, a large amount of tourist information content was registered by the staff at the tourist office. This validation experiment 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.

Similar dialogue systems were also installed at other locations including the Handa city hall building, NHK's Nagoya broadcasting station, and the National Institute of Informatics (Fig. 32, 33, 34). 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. For the National Institute of Informatics, we built a dialogue system using a 3D model of their mascot (a cartoon dog called "Bit") (Fig. 34). We also constructed a speech database matched to this character, and a speech synthesis system that

Fig. 27. Panel display

uses the character's voice.

Other validation experiments

We performed a study of dialogue systems and user-generated dialogue content creation environments in mobile environments.

- We developed a system that enables spoken conversations with a CG character using a smart phone such as an iPhone (Fig. 35)[5]. This system was implemented by linking MMDAgent with the video phone capabilities of Skype. We also conducted public trials and a questionnaire survey at the 74th National Convention of the Information Processing Society of Japan. From the results of 120 questionnaires, we found that although mobile-oriented dialogue systems have dialogue response speeds inferior to those of digital signage systems, they are regarded as being more friendly.
- In January 2014, we conducted a demonstration experiment of a mobile in-school field trip support system based on dialogue at a junior high school (Fig. 36) [6]. In this system, a dialogue agent on a smart phone provides the junior high school students with information about the Nagoya Institute of Technology campus and its facilities. The dialogue content used in this experiment was created by two Nagoya Institute of Technology staff members responsible for arranging visits. Subjective evaluation tests demonstrated the effectiveness of spoken guidance based on the proposed system.

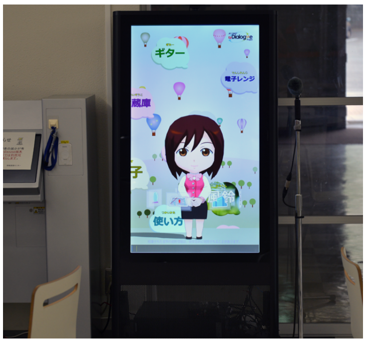


Fig. 28. Open space (Dream Room) at the Nagoya Institute of Technology

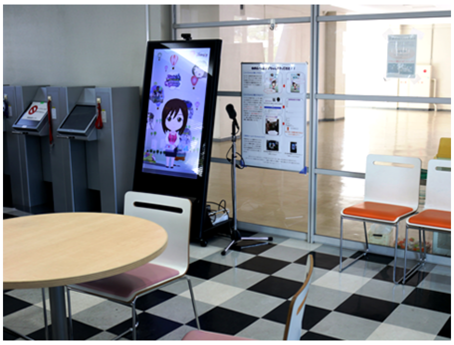


Fig. 29. Open space (Dream Room) at the Nagoya Institute of Technology

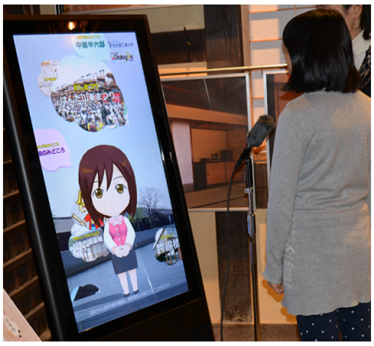


Fig. 30. Tourist information office in Handa city

Fig. 31. Handa city tourist information office (outside view)

- To increase the use of the dialogue system by the general public, it would be useful to link it up with systems such as the public telephone network. We therefore developed a framework that links MMDAgent with a VoIP client[8]. We created and middleware to connect the system to Skype for business (the new version of Lync), which was installed and deployed throughout the campus. At the same time, to connect with the new internal VoIP phone system that forms part of the Nagoya Institute of Technology Unified Communications System, we developed middleware for connecting a software phone, and we subjected it to interconnection tests. While developing and evaluating this framework, we also examined the framework of cooperation with the campus telephone network.

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. As a preliminary stage,

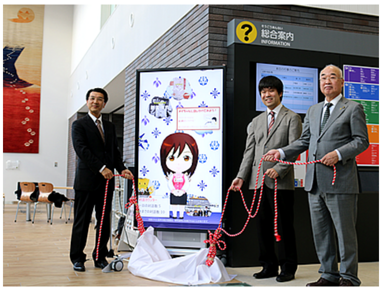


Fig. 32. Handa city hall

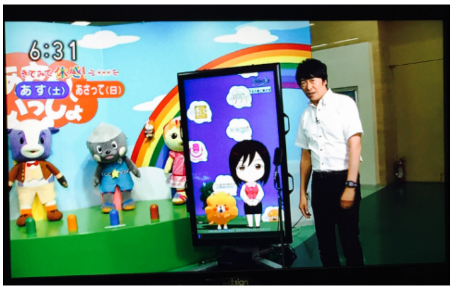


Fig. 33. NHK Nagoya broadcasting station

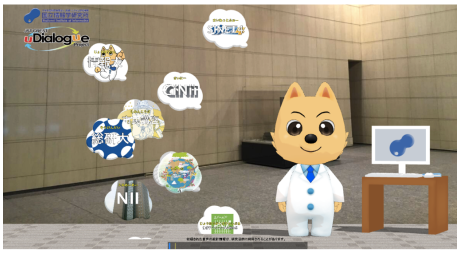


Fig. 34. National Institute of Informatics

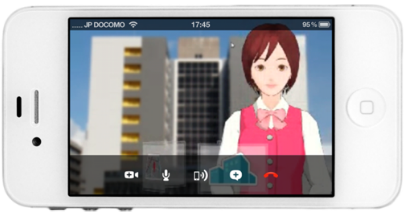


Fig. 35. Skype

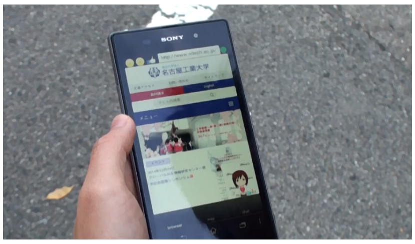


Fig. 36. Mobile campus tour support system

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

First, 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. 37).

- 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 must be met by a dialogue system so that dialogue content can be created in a network environment as user-generated media are as follows:

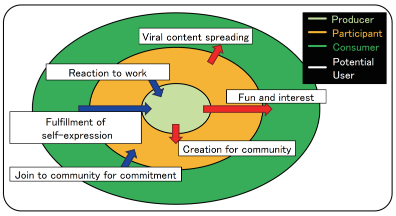


Fig. 37. User layer categories and incentives

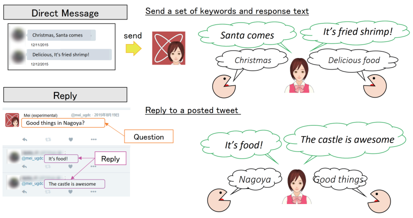


Fig. 38. Themed dialogue content creation

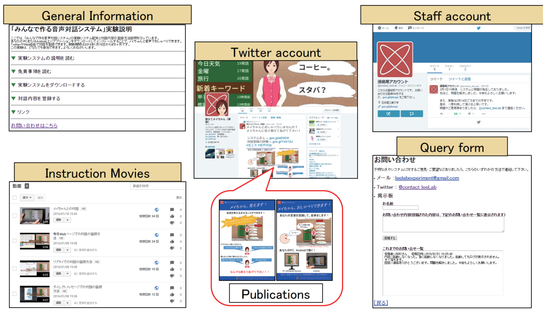


Fig. 39. Design of participation flow lines and description pages for ordinary users in the social experiment

- From potential users to consumers: expanding the dialogue content playback environment for compatibility with 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 the Twitter library functions to create dialogue content on a particular theme (Fig. 38).

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. 39). The content server on which the dialogue content was stored was published on the university's external server, and a multiplatform (Windows/Linux/Android) system was provided to facilitate access. In this experiment, we used a flow line design 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.

6.4 Dialogue content provision service

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 service as a framework to facilitate sharing of dialogue content between users (Fig. 40, 41). 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.

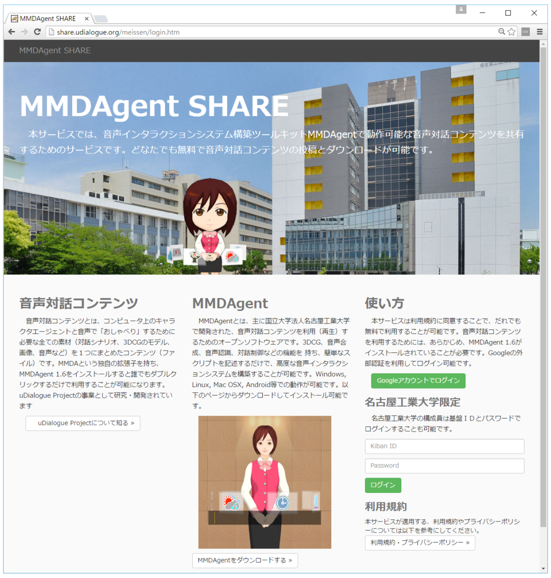
1. Since dialogue content often consists of a set of many files and is cumbersome to work with, we defined our own MMDA format that allows them to be used as a single file.
2. We created a function whereby MMDA format files can be produced easily on the server.
3. We developed a new MMDAgent installer 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.
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. 42).

7 Encyclopedia

So far, in addition to developing the dialogue content production support tool and underlying software centered on the MMDAgent toolkit for the construction of dialogue interaction systems, 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, slides, reference manuals, 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 design guidelines. This package includes the following items:

1. The MMDAgent toolkit for the construction of voice interaction systems (multi-platform compatible)
2. MMDAgent Primer (Japanese and English)
3. MMDAgent creators reference manual (Japanese and English)


MMDAgent SHARE
share.udialogue.org/meissen/login.htm
MMDAgent SHARE
Googleアカウントでログイン
Kiban ID
Password
ログイン
uDialogue Projectについて知る »
MMDAgentをダウンロードする »
利用規約・プライバシーポリシー »

MMDAgent SHARE

本サービスでは、音声インタラクションシステム構築ツールキットMMDAgentで動作可能な音声対話コンテンツを共有するためのサービスです。どなたでも無料で音声対話コンテンツの投稿とダウンロードが可能です。

音声対話コンテンツ

音声対話コンテンツとは、コンピュータ上のキャラクタエージェントと音声で「おしゃべり」するために必要な全ての素材（対話シナリオ、3DCGのモデル、画像、音声など）を１つにまとめたコンテンツ（ファイル）です。MMDAという独自の拡張子を持ち、MMDAgent 1.6をインストールすると誰でもダブルクリックするだけで利用することが可能になります。uDialogue Projectの事業として研究・開発されています

MMDAgent

MMDAgentとは、主に国立大学法人名古屋工業大学で開発された、音声対話コンテンツを利用（再生）するためのオープンソフトウェアです。3DCG、音声合成、音声認識、対話制御などの機能を 持ち、簡単なスクリプトを記述するだけで、高度な音声インタラクションシステムを構築することが可能です。Windows, Linux, Mac OSX, Android等での動作が可能です。以下のページからダウンロードしてインストール可能です。

使い方

本サービスは利用規約に同意することで、だれでも無料で利用することが可能です。音声対話コンテンツを利用するためには、あらかじめ、MMDAgent 1.6がインストールされていることが必要です。Googleの外部認証を利用してログイン可能です。

名古屋工業大学限定

名古屋工業大学の構成員は基盤ＩＤとパスワードでログインすることも可能です。

利用規約

本サービスが適用する、利用規約やプライバシーポリシーについては以下を参考にしてください。

Fig. 40. Dialogue content sharing service

4. MMDAgent developers reference manual (Japanese and English)
5. MMDAgent training 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 conversation 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 ex-

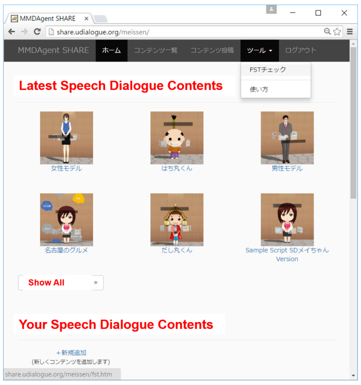
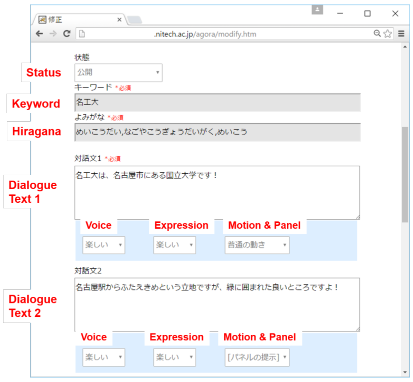

MMDAgent SHARE
share.udialogue.org/meissen/
MMDAgent SHARE
ホーム
コンテンツ一覧
コンテンツ投稿
ツール
ログアウト
FSTチェック
使い方
Latest Speech Dialogue Contents
女性モデル
はち丸くん
男性モデル
名古屋のグルメ
だし丸くん
Sample Script SDメイちゃん Version
Show All »
Your Speech Dialogue Contents
＋新規追加
(新しくコンテンツを追加します)
share.udialogue.org/meissen/fst.htm

Fig. 41. Shared dialogue content



状態: 公開

キーワード *必須: 名工大

よみがな *必須: めいこうだい,なごやこうぎょうだいがく,めいこう

対話文1 *必須: 名工大は、名古屋市にある国立大学です！

楽しい / 楽しい / 普通の動き

対話文2: 名古屋駅からふたえきめという立地ですが、緑に囲まれた良いところですよ！

楽しい / 楽しい / [パネルの提示]

Fig. 42. Dialogue content creation service periments, 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.

8 Conclusion

In this study, by separating 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 engaging 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 engaging 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 dialogue systems are increasing their appeal 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 article. 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] R. Nishimura, D. Yamamoto, T. Uchiya, and I. Takumi, "Development of a Dialogue Scenario Editor on a Web Browser for a Spoken Dialogue System," Proceedings of the Second International Conference on Human-agent Interaction, pp. 129–132, 2014.

[2] Y. Matsushita, T. Uchiya, R. Nishimura, D. Yamamoto, and I. Takumi, "Crowdsourcing Environment to Create Voice Interaction Scenario of Spoken Dialogue System," Proc. of the 18-th International Conference on Network-Based Information Systems (NBiS-2015), pp.500–504, 2015.

[3] Y. Matsushita, T. Uchiya, R. Nishimura, D. Yamamoto, and I. Takumi, "Experiment and Evaluation of Crowd Sourcing Model for Creation of Voice Interaction Scenario," Proc. of the IEEE GCCE 2015, pp.321–322, 2015.

[4] T. Uchiya, R. Nakano, D. Yamamoto, R. Nishimura, and I. Takumi, "Extension with Intelligent Agents for the Spoken Dialogue System for Smartphones," Proc. of the IEEE GCCE 2015, pp.298–299, 2015.

[5] T. Uchiya, D. Yamamoto, M. Shibakawa, M. Yoshida, R. Nishimura, I. Takumi, "Development of Spoken Dialogue Service based on Video Call named "Mobile Meichan"," Proceedings of JAWS2012, 2012. (in Japanese)

[6] T. Uchiya, M. Yoshida, D. Yamamoto, R. Nishimura, and I. Takumi, "Design and Implementation of Open-Campus Event System with Voice Interaction Agent," International Journal of Mobile Multimedia, Vol.11, No.3,4, pp.237–250, 2015.

[7] K. Wakabayashi, D. Yamamoto, and N. Takahashi, "A Voice Dialog Editor Based on Finite State Transducer Using Composite State for Tablet Devices," Computer and Information Science 2015, Studies in Computational Intelligence, Vol.614, pp.125–139, 2016.

[8] R. Nishimura, K. Sugioka, D. Yamamoto, T. Uchiya, and I. Takumi, "A VoIP-based Voice Interaction System for a Virtual Telephone Operator Using Video Calls," Procceedings of the IEEE GCCE 2014, pp.529–532, 2014.

[9] D. Yamamoto, K. Oura, R. Nishimura, T. Uchiya, A. Lee, I. Takumi and Keiichi Tokuda, "Voice Interaction System with 3D-CG Human Agent for Stand-alone Smartphones," Proceedings of the 2nd International Conference on Human Agent Interaction, ACM digital library, pp.320–330, 2014.

[10] K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Integration of spectral feature extraction and modeling for HMM-based speech synthesis," IEICE Transactions on Information and Systems, vol.E97-D, no.6, pp.1438–1448, 2014.

[11] S. Takaki, Y. Nankaku, and K. Tokuda, "Contextual partial additive structure for HMM-based speech synthesis," 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.7878–7882, Vancouver, Canada, 2013.

[12] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King. "Investigating automatic & human filled pause insertion for speech synthesis." Proceedings of Interspeech, 2014.

[13] S. R. Gangireddy, S. Renals, Y. Nankaku, and A. Lee, "Prosodically-enhanced Recurrent Neural Netwrok Language Models." Proceedings of Interspeech 2015, Dresden, Sep. 2015.

[14] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015), pp.4455–4459, Brisbane, Australia, April 19–24, 2015.

[15] S. Takaki, S. Kim, J. Yamagishi, and J.J. Kim, "Multiple feed-forward deep neural networks for statistical parametric speech synthesis," Proceedings of Interspeech 2015, pp.2242–2246, 2015.

[16] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Trajectory training considering global variance for speech synthesis based on

neural networks," Proceedings of 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016), pp.5600–5604, Shanghai, China, March 20–25, 2016.

[17] 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, December 2–3, 2015.

[18] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Simultaneous optimization of multiple tree structures for factor analyzed HMM-based speech synthesis," Proceedings of Interspeech 2015, pp.1196–1200, Dresden, Germany, September 6–10, 2015.

[19] CSTR VCTK Corpus
`http://www.udialogue.org/ja/download-ja.html`

[20] T. Funayachi, K. Oura, Y. Nankaku, A. Lee, and K. Tokuda, "A simple dialogue description based on finite state transducers for user-generated spoken dialog content," Proceedings of ASJ 2013 autum meeting, 2-P-28, pp.223–224, September 25–27, 2013. (in Japanese)