THE HINGE between Input and Output: Understanding the Multimodal Input Fusion Results In an Agent-Based Multimodal Presentation System

Yong Sun
National ICT Australia
Australian Technology Park,
Eveleigh NSW 1430, Australia
yong.sun@nicta.com.au

Fang Chen
National ICT Australia
Australian Technology Park,
Eveleigh NSW 1430, Australia

Helmut Prendinger
National Institute of Informatics, Japan
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430 Japan

Vera Chung
School of IT,
The University of Sydney,
NSW 2006, Australia

Yu (David) Shi
National ICT Australia
Australian Technology Park,
Eveleigh NSW 1430, Australia

Mitsuru Ishizuka
Graduate School of Information Science and Technology,
University of Tokyo,
1-18-13, Sotokanda, Chiyoda-ku,
Tokyo 101-0021, Japan

CHI 2008, April 5–10, 2008, Florence, Italy.
ACM 978-1-60558-012-8/08/04.

Abstract
A multimodal interface provides multiple modalities for input and output, such as speech, eye gaze and facial expression. With the recent progresses in multimodal interfaces, various approaches about multimodal input fusion and output generation have been proposed. However, less attention has been paid to how to integrate them together in a multimodal input and output system. This paper proposes an approach, termed as THE HINGE, in providing agent-based multimodal presentations in accordance with multimodal input fusion results. The analysis of experiment result shows the proposed approach enhances the flexibility of the system while maintains its stability.

Keywords
Multimodal interfaces, multimodal input fusion, input understanding, discourse representation

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces---theory and methods

Introduction

A multimodal interface allows input and/or output to be conveyed over multiple modalities. With the recent progresses in multimodal interfaces, exciting new types of interactive entertainment applications are being created such as audience-guide movies, virtual travel guides and tutors [5]. Since multimodal interfaces support both multimodal input and output, rich interaction experiences become possible. Our interest within multimodal interfaces is in a multimodal infotainment (information and entertainment) system, where life-like animated agents act in the role of virtual presenters that understand multimodal input from a user and convey information with their multimodal expressiveness in a convincing and entertaining way. In the system, the interactive presentation content is automatically generated from a text. A user can also ask questions with his/her multiple input modalities such as speech, gesture, eye gaze and so on. Researches on several tasks have achieved initial results such as the Multimodal Presentation Markup Language 3D (MPML3D) [6], the Polynomial Unification-based Multimodal Parsing Processor (PUMPP) [9, 10] for multimodal input fusion and the method for generating multimodal content automatically from text [7]. To integrate them together, we try to devise a systematic approach to hinge multimodal fusion result with dialog management based on discourse information. This paper presents our approach, termed as THE HINGE, for this task. In this paper, a multimodal utterance refers to a set of multimodal inputs expressing a user's intention.

Related works

Customized formalisms were used in some researches. In [4], the semantic meaning of multimodal utterances was represented with "a simple logical representation with predicates pred(....) and lists [a,b,...]". There was not a separate dialog management component to hinge semantic meaning of multimodal utterances with discourse information. Feature structures were adopted in other researches. [3] applied typed feature structures to represent semantic meaning. There was not a clear separation between syntactic and semantic information in fusion results. [8] used unification on feature structures to integrate multimodal inputs; however, how to utilize the fusion result in dialog management was not addressed.

The overview of our system

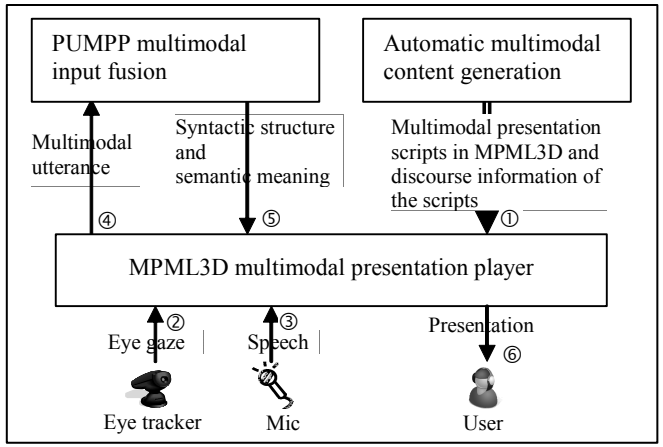

figure 1. The structure of our system.

As shown in figure 1, the multimodal presentation scripts in MPML3D and their discourse information are generated from texts. After the MPML3D player accepts the multimodal presentation content from the content generation module, it delivers the specified multimodal presentation to a user. During the presentation, it also

listens to the signals captured by an eye tracker and a Mic. The captured eye fixations are interpreted to the entities being fixated on. A speech recognizer attached to the MPML3D player interprets the captured speech signal to speech strings. The MPML3D player constructs a multimodal utterance with these information, and passes it to the PUMPP multimodal input fusion module which returns a syntactic structure and semantic meaning of the utterance as the fusion result. With the semantic meaning and the discourse information, THE HINGE in the MPML3D player decides the content to deliver to a user in the next step.

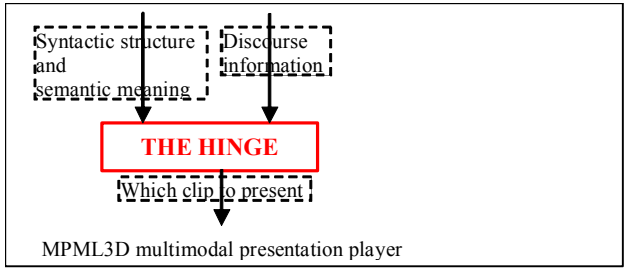

figure 2. The input and output of THE HINGE.

As shown in figure 2, the focus of this paper is on THE HINGE between multimodal input and output. Specifically, they are the multimodal input fusion result and the presenting decision. In the discourse information, each presentation clip is assigned a *triggering condition*.

Subsumption in hybrid logic

Feature structures and subsumption

Feature structures were used in some of previous researches to represent semantic meaning. They are simply sets of attribute-value pairs, where attributes are atomic symbols, and values are either atomic symbols or feature structures themselves. In figure 3, feature structure (1) has an attribute "*Agreement*" whose value is another feature structure, which has an attribute "*Person*" and value "*3rd*".

$$\left[Agreement \quad \left[Person \quad 3rd\right]\right] \qquad (1)$$

$$\left[Agreement \quad \left[Number \quad single\right]\right] \qquad (2)$$

$$\left[Agreement \quad \begin{bmatrix} Person & 3rd \\ Number & single \end{bmatrix}\right] \qquad (3)$$

figure 3. Three feature structures.

When two feature structures are unified, they are compared with every attribute in anyone of them. 1) If one's value is a variable or not specified, another's value is a constant; the variable will be bound to the other one. 2) When both of them have constant values; if their values are equal then unification succeeds, otherwise unification fails. The result of unification is two identical feature structures or fail. In figure 3, the unification result of (1) and (2) is (3). Intuitively, unifying two feature structures produces a new one which is more specific (having more attributes) than, or is identical to, either of the input feature structures. We say a more general feature structure **subsumes** an equally or more specific one. In figure 3, both (1) and (2) subsume (3). If feature structure A subsumes B, B satisfies all constraints outlined by A. Although feature structures can be used to represent semantic meaning, hybrid logic is more suitable to capture meaning.

Hybrid logic and feature structures

A hybrid logic [1] formula can be viewed as a flat conjunction of the heads and dependents inside it.

Hybrid logic provides an internal means to refer to propositions. For example, the formula in figure 4 is a hybrid logic formula. Its head is *m1*. The dependents specify the value of the head and its properties. "*<HasProp>s1:proposition*" specifies that the *m1* has an attribute "*HasProp"* whose value is a nominal *s1*.

```
@m1 (@s1:proposition(storage) ^
     @m1:phys-obj(medium) ^
     @m1:phys-obj(<num>sg) ^
     @m1:phys-obj(<HasProp>s1:proposition))
```

figure 4. A hybrid logic formula.

Although feature structures are essentially a two-dimensional notation for hybrid logic [2], by making use of *nominals*, hybrid logic allows adjuncts to insert their semantic import into the meaning of the head. This flexibility makes it amenable to represent a wide variety of semantic phenomena in a propositional setting, and it can furthermore be used to formulate a discourse theory [1]. The track from grammar to discourse can be covered with a single meaning formalism. That is one of the reasons why we chose hybrid logic to represent semantic meaning of multimodal utterances and discourse information.

Hinging semantic meaning of multimodal utterances and multimodal presentations

To determine the clip in corresponding with a multimodal utterance, every clip is assigned a *triggering condition* in discourse information, which is the loosest semantic meaning to select this clip. For example, for the clip "*Flash memory is the memory medium of this EasyMP3Pod*", the *triggering condition* in figure 5 is defined. It subsumes the semantic meaning of multimodal utterances such as "*what is the storage medium of this easy one"* while fixating on *"EasyMP3Pod*", "*what is the memory medium of this easy one"* while fixating on *"EasyMP3Pod*", and so on. To verify if the semantic meaning of a multimodal utterance ($S_{utterance}$) satisfies the *triggering condition* of a clip (T_{clip}), THE HINGE checks if T_{clip} subsumes $S_{utterance}$ with the process described in figure 7 (at the end of this paper).

```
@q (@q:quantification(what) ^
    @q:quantification(<Body>e:state) ^
    @e:state(be) ^
    @e:state(<Arg>x:phys-obj) ^
    @x:phys-obj(medium) ^
    @x:phys-obj(<Modifier>m) ^
    @m(<Ref>y:appliance) ^
    @y:appliance(EasyMP3Pod))
```

figure 5. A triggering condition in hybrid logic.

Experiment and Analysis

To analyze the performance of THE HINGE, an experiment was conducted.

Setup and Scenario

In the experiment, there is a virtual sales scenario where a team of two 3D animated agents present MP3 players (EasyMP3Pod and MP3Advance) to a human user. Each of the two agents (female and male) can perform body and facial gestures (emotional expressions). A user is seated in front of the monitor screen, as shown in figure 6. Agents and environment are controlled by the MPML3D player attached with eye tracking and speech recognition function.

figure 6. Experiment setup.

After an interactive presentation between the two agents, a user can ask some questions based on the presentation with his/her speech and eye gaze. The system would present a corresponding clip as a response to the question. Sample multimodal utterances asked by users are listed in table 1.

#	Speech Input	Eye gaze fixation
1	How big is its storage	EasyMP3Pod
2	How many songs can it hold	MP3Advance
3	How many songs can this powerful one hold	MP3Advance
4	Does this simple one have FM tuner	MP3Advance
5	What functions does this big one come with	EasyMP3Pod
6	What is the storage medium of this easy one	EasyMP3Pod
7	What is the storage medium of this simple one	MP3Advance
8	Does this lovely one have a screen	EasyMP3Pod
9	How many buttons does it have	EasyMP3Pod

table 1. Sample multimodal utterances with speech and eye gaze as input modalities in the experiment.

Early Observations and Analysis

More than half multimodal utterances triggered corresponding responses. Others fell into the following categories. 1) Incorrect Responses Observed. When a user asked "How big is its storage for pictures" while fixating on "*EasyMP3Pod*", the system answered the whole storage size of the EasyMP3Pod rather than the size for pictures. Because the *triggering condition* of EasyMP3PodStorageClip ($T_{EasyMP3PodStorage}$) is more general than that of EasyMP3PodPicStorageClip ($T_{EasyMP3PodPicStorage}$), and $T_{EasyMP3PodStorage}$ firstly subsumed the semantic meaning of the utterance. That implies that *triggering conditions* should not be compatible (one can subsume another one) with others. Or, if multiple *triggering conditions* can subsume the semantic meaning of an utterance, a confirmation from the user should be pursued. 2) No Reply. There are several causes. Firstly, it is due to the instability of eye gaze fixation and eye tracking. After an intentional fixation, a user's eye gaze may lie on or be recognized as laying on another entity in the screen. Therefore, an incorrect multimodal utterance was constructed for multimodal input fusion. Secondly, a user uses words which are out of the vocabulary of the presentation. That implies speech recognition and/or multimodal input fusion should be able to at least skip/ignore them. During multimodal content generation, the key words should be used repeatedly; therefore, a user will prefer to use them in his/her multimodal questions.

Conclusions and future work

This paper proposes a systematic approach—THE HINGE to hinge multimodal input fusion and output generation in an agent-based multimodal presentation system. In it, the subsumption on feature structure is adapted to hybrid logic to check the generalization of

one hybrid logic formula over another one. That enables a system to respond to multimodal utterances flexibly. It adopts a single formalism—hybrid logic to represent semantic meaning of multimodal utterances and discourse information so that the relationship between them can be described directly. The preliminary experiment result supports its flexibility on the system performance. We also observed the overall performance of the system is closely related to other modules in the system. In the future, the compatibility between hybrid logic formulas (the *triggering conditions*) should be further investigated. The approach of communication between fusion results and discourse information can also be a pending topic.

A hybrid logic formula can be formulized as:

head
dependent1
dependent2
…

A dependent can be one of the following:

- Nominal : Proposition
- Nominal : <Mod>Proposition
- Nominal : <Mod>Nominal2

✶ Nominal2 is another nominal which refers other dependents.

References

[1] Baldridge, J. and Kruijff, G. M. Coupling CCG and Hybrid Logic Dependency Semantics In *Proc. ACL 2002*.

[2] Blackburn, P. Representation, Reasoning, and Relational Structures: a Hybrid Logic Manifesto. *Journal of the Interest Group in Pure Logic*, 8(3):339–365.

[3] Holzapfel, H., Nickel, K. and Stiefelhagen, R. Implementation and Evaluation of a Constraint-based Multimodal Fusion System for Speech and 3D Pointing Gestures. In *Proc. ICMI 2004*, ACM Press (2004).

[4] Johnston, M. and Bangalore, S. Finite-state Multimodal Integration and Understanding. *Natural Language Engineering* Volume 11, Issue 2 (June 2005).

[5] Maybury, M., Stock, O., and Wahlster, W. Intelligent Interactive Entertainment Grand Challenges. *IEEE Intelligent Systems* 21, 5, 14-18.

[6] Nischt, M., Prendinger, H., André, E., and Ishizuka, M. MPML3D: a Reactive Framework for the Multimodal Presentation Markup Language. In *Proc. IVA 2006*.

[7] Prendinger, H., Piwek, P. and Ishizuka, M. A Novel Method for Automatically Generating Multi-modal Dialogue from Text. *International Journal of Semantic Computing, 2007, Vol. 1, No. 3, Sept. 2007*.

[8] Rudzicz, F. Put a Grammar Here: Bi-Directional Parsing in Multimodal Interaction. *Ext. Abstracts CHI 2006*, ACM Press (2006).

[9] Sun, Y., Shi, Y., Chen, F. and Chung, V. An Efficient Multimodal Language Processor for Parallel Input Strings in Multimodal Input Fusion. In *Proc. ICSC 2007*.

[10] Sun, Y., Shi, Y., Chen, F. and Chung, V. An Efficient Unification-based Multimodal Language Processor in Multimodal Input Fusion. In *Proc. OZCHI 2007*.

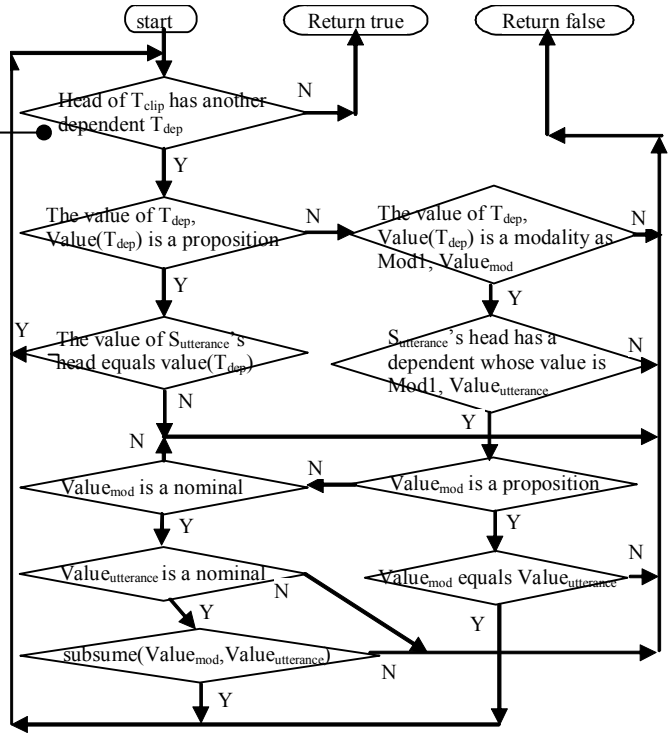

figure 7. Flowchart for subsume(Head of Tdep, Head of Sutterance).