NICE project (IST-2001-35293)

Natural Interactive Communication for Edutainment

NICE Deliverable D3.6-2

Multimodal Input Understanding Module for the Second Prototype

15 December 2004

Authors

LIMSI-CNRS : Jean-Claude Martin, Guillaume Pitel, Stéphanie Buisine

NISLab : Niels Ole Bernsen

Teliasonera : Johan Boye, Joakim Gustafson

Project ref. no.	IST-2001-35293
Project acronym	NICE
Deliverable status	Restricted
Contractual date of delivery	1 November 2004
Actual date of delivery	16 December 2004
Deliverable number	D3.6-2
Deliverable title	Multimodal input understanding module for the second prototype.
Nature	Report
Status & version	Final
Number of pages	32
WP contributing to the deliverable	WP3
WP / Task responsible	LIMSI-CNRS
Editor	Jean-Claude Martin
Author(s)	Jean-Claude Martin, Guillaume Pitel, Stéphanie Buisine, Niels Ole Bernsen, Johan Boye, Joakim Gustafson
EC Project Officer	Mats Ljungqvist
Keywords	Input fusion, multimodal input, temporal fusion, semantic fusion, gesture interpretation
Abstract (for dissemination)	This report, Deliverable 3.6-2 from the HLT project Natural Interactive Communication for Edutainment (NICE), describes the Input Fusion module used for the 2nd prototype including requirement and analysis of input fusion, specification of input fusion and application to both the “HCA Study” prototype and the “Fairy Tale World” prototype.

Table of Contents

Frequently used acronyms:

- Gesture Recognition module (GR)
- Gesture Interpretation module (GI)
- Input Fusion module (IF)
- NISLab Character Module (CM)
- Teliasonera Dialogue Manager module (DM)
- Hans Christian Andersen (HCA)
- Fairy Tale World (FTW)

1 Introduction

Input Fusion in the NICE project aims at integrating children's speech and 2D gestures when conversing with virtual characters about 3D objects. It shares some general requirements of multimodal input systems such as the need to manage and represent timestamps of input events, multi-level interpretation, composite input, confidence scores (Avaya et al. 2004). Yet, the conversational goal of the NICE system and the fact that it aims at being used by children makes it different from current research on multimodal systems which studies speech and gestures for task-oriented applications (Johnston et al. 2002; Kaiser et al. 2003).

Section 2 of this document explains the requirement and analysis of input fusion in the context of the NICE project. Section 3 describes the specifications of input fusion at a general level. Section 4 describes how it is applied to both the "HCA Study" and the "Fairy Tale World" prototypes.

2 Requirement and analysis of input fusion

The global requirements on the IF module were to:

- work with 2 different NLU modules: NISLab NLU and Telia NLU,
- work with the NISLab Character module and the Telia Dialogue module,
- work in the 3 conditions : HCA Study, Cloddy Hans in HCA Study, Cloddy Hans in Fairy Tale World.

The design of semantic fusion in IF PT2 was driven by:

- multimodal behaviour observed in PT1 user tests,
- requirements and directions identified in D1.1-2a (Martin et al. 2004) which are recalled in appendix 1,
- multimodal task analysis for PT2.

2.1 HCA Study

2.1.1 Multimodal behaviours observed during PT1 user tests

Only 8 multimodal behaviours were observed in videos which covered only part of the PT1 user tests (Martin et al. 2004). They are recalled in Table 1.

Succession of modalities	Delay* between modalities	Object gestured to	Shape of gesture	Spoken utterance + NLU frame	Cooperation between modalities
Gesture – speech	2 sec.	Picture of colosseum	Circle	"What's this?"	Complementarity
Simultaneous	0 sec.	Picture of HCA mother	Circle	"What's that picture?"	Complementarity
Simultaneous	0 sec.	Hat	Circle	"I want to know something about your hat."	Redundancy
Gesture – speech	4 sec.	Statue of 2 people	Circle	"Do you have anything to tell me about these two?"	Complementarity
Simultaneous	0 sec.	Statue of 2 people	Point	"What are those statues?"	Complementarity
Gesture – speech	4 sec.	Picture above book-case	Circle	"Who is the family on the picture?"	Complementarity
Gesture – speech	3 sec.	Picture above book-case	Circle	"Who is in that picture?"	Complementarity
Simultaneous	0 sec.	Vase	Circle	"How old are you?"	Concurrency

Table 1. Description of multimodal sequences observed in the PT1 video corpus. * The delay between modalities was measured between end of first modality and end of second modality.

These examples provide illustrative semantic combinations of modalities:

- Deictic: « What's this? » + circle on picture colosseum
- Type of object mentioned in speech :
  - « What's that picture? » + circle on picture HCA Mother
  - « I want to know something about your hat? » + circle on the hat

- Incompatibility between internal singular representation of objects and their plural/singular perceptual « affordance » (e.g. a single object is referred to in the user's speech as a plural object): « Do you have anything to tell me about these two?" » (or « What are those statues? ») with a circle on the statue of two characters.
  Several objects might elicit such plural/singular incompatibility. They represent several entities of the same kind but they are internally represented as a single object. They could be referred to as a single or as several objects, the number of which can be planned for some of them: books (>2) ; boots : 2 ; papers (>2) ; pens : 2 ; statue : 2.

Symmetrically (although not observed as such in the PT1 video user tests), several objects of similar type and in the same area might be perceived as a single "perceptual group" (Landragin et al. 2001) and elicit a plural spoken reference and a singular gesture on one of the items of the group:

- the pictures on the wall above the desk.
- the "clothes group": coat – boots – hat – umbrella,
- the furniture: table and chairs,
- the small objects on the small shelf,

Multimodal examples from the log files include:

- linguistic reference to concepts related to the graphical object (e.g. « dad » and gesture on a picture) instead of direct reference to the object type or name (« picture »),
- plural/singular inconsistency between speech and gesture (e.g. « your dad and your mom and your grandpa » while gesturing on pictureJonasCollin).

What we keep from the study of PT1 user tests videos and log files for specifying semantic fusion required in such a 3D conversational application is that the richness of the 3D environment and the conversational intuitive context might lead to ambiguous references to objects.

2.1.2 Task analysis

The user is asked to select one or several object(s) in HCA's study in order to get information about it. A task analysis identified communicative acts related to this scenario (Table 2).

	Communicative acts
1.	Ask for task clarification
2.	Ask for initial information about the study
3.	Select one referenceable object
4.	Select one non referenceable object
5.	Select several referenceable objects
6.	Select an area
7.	Explicitly ask information about selected object
8.	Negatively select an object (e.g. "I do not want to have information on this one")
9.	Negatively select several objects
10.	Confirm the selection
11.	Reject the selection
12.	Correct the selection
13.	Interrupt HCA
14.	Ask HCA to repeat the information on the currently selected object
15.	Ask HCA to provide more information on the currently selected object
16.	Comment on information provided by HCA
17.	Comment on another object than the one currently selected
18.	Select another object while referring to the previous one
19.	Select another object of the same type than the one currently selected
20.	Move an object (user may try to do that although not possible and not explicitly related to the task)
21.	Compare objects
22.	Thank

Table 2. The list of communicative acts identified during task analysis.

The following step was to identify utterances that the user might speak for each of these communicative acts, possibly involving also gestures (Table 3).

	Communicative act	Illustrative example of spoken utterance (possibly combined with gesture)
1.	Ask for task clarification	"Do you want me to select an object?"
2.	Ask for initial information about the study	"Is this your study?" "Can I select an object anywhere in the study?"
3.	Select one referenceable object	"What is this?" "Who is this?" "Who is in that picture ?" "I want some information on this picture" "Who is this lady?" "Who is the lady?" "I am selecting this picture" "What is represented on the top right picture above your desk?" "I want to know something about your hat." "Tell me about your father" "Explain." "Can you tell me about these two? (and gesture on the statue)" "Is it one of these books? (and gesture on a group of books)"
4.	Select one non referenceable object	Idem as above
5.	Select several referenceable objects	"Are these pictures your family?" "Do you have anything to tell me about these ?" "Do you have anything to tell me about these two?" "Tell me about this one and this one" "Tell me about this one, this one and this one"
6.	Select an area	"Show me all objects in this part of the study"
7.	Explicitly ask information about selected object	"I want some information about the object I have selected"
8.	Negatively select an object (e.g. "I do not want to have information on this one")	"I do not want to have information on this one" "I already know about this fairy tale" "This is not the little mermaid"
9.	Negatively select several objects	"I do not want to have information on these ones"
10.	Confirm the selection	"Yes, I want to know about it"
11.	Reject the selection	"No, I did not mean that one" "Sorry, I did wrong."
12.	Correct the selection	"I meant that one"
13.	Interrupt HCA	"Please stop"
14.	Ask HCA to repeat the information on the currently selected object	"What did you say?" "Are you talking about the top right picture above your desk?" "Can you describe it again?"
15.	Ask HCA to provide more information on the currently selected object	"Can you tell me more about it?"
16.	Comment on information provided by HCA	"I like it" "I like that one"
17.	Comment on another object than the one currently selected	"I prefer this fairy tale"
18.	Select another object while referring to the previous one	"And this one?"
19.	Select another object of the same type than the one currently selected	"And this one?"
20.	Move an object	"Can I move this picture here?"
21.	Compare objects	"Are they from the same fairy tale?" "Is this woman the same as this woman ?"
22.	Thank	"Thanks"

Table 3. A list of possible utterances for each communicative act.

Combinations of speech and gesture can display inconsistencies regarding the name of the object, the type of the object, or the number of selected objects in speech and gesture. Such inconsistency need to be detected, and the conditions under which the corresponding constraints need to be relaxed have to be studied. Examples of such complex cases are listed in Table 4.

Combination	**Speech**	**GI output**
Combination of several communicative acts in a single turn (reject selection followed by correct selection)	"I did not mean that one, I meant that one" "It is not this one but this one"	Select (current object) Followed by: Select (new object)
Combination of several communicative acts in a single turn (interrupt HCA and select an object)	"Stop this story and tell me about your hat"	Select (HCA) Followed by: Select (hat)
Concurrency (no reference in speech)	"How old are you?"	Select (desk)
Inconsistency between gesture and speech (name of object)	"I want to know about the little mermaid" "This is the little mermaid" "Is this the little mermaid?"	Select (JennyLind)
Inconsistency between gesture and speech (singular reference in speech and plural gesture)	What is this ?" this boot this book(s), « is this your family ? »	ReferenceAmbiguity (several objects)
Inconsistency between gesture and speech (plural reference in speech and singular gesture)	These boots These books These (two) statues	Select (single object)

Table 4. Complex cases of gesture and speech combinations.

2.2 Fairy Tale World

The different objects as well as the scenario of the Fairy Tale World prototype are described in D1.2-2b (Boye et al. 2004). We provide a small summary in this section of the issues related to multimodal fusion.

2.2.1 Informal task analysis

The key device in the laboratory is a fairy-tale machine, which nobody except Andersen himself is allowed to touch. On a set of shelves beside the machine, various objects, such as a key, a hammer, a diamond and a magic wand, are located. By removing objects from the shelves, putting them into suitable slots in the machine and pulling a lever, one lets the machine construct a new fairy-tale in which the objects come to life. The first scene thus develops into a kind of "put-that-there" game, where it is the task of the user to instruct Cloddy Hans; tell him where to go, which objects to pick up and where to put them down, etc. If the user does not understand what to say, Cloddy Hans will encourage him or her, give suggestions, and eventually take matters into own hands. Because the initial scene is task-

oriented in a straightforward way, the system is able to anticipate what the user will have to say to solve it. The real purpose is not to solve the task, but to engage in a collaborative grounding conversation where the user learns what the fairy-tale objects can be used for and how they should be referred to. This process also lets the players find out (by trial and-error) how to adapt in order to make it easier for the Cloddy Hans to understand them, e.g. by using multimodal input in certain contexts. The intention is to make the interaction smoother in the subsequent scenes in the fairy-tale world, since the objects that appear in it already have been grounded in the initial scene.

The user's means of action in the world are: speaking to other characters in the game, pointing and gesturing at arbitrary characters, objects and locations. The following dialog acts are expected from the user: confirm a proposition, disconfirm a proposition, ask for an explanation, request the character to do something, ask the character a question. The request include pickup an object, put down an object.

At the beginning of the second scene, Cloddy Hans encourages the player to explore the immediate surroundings on the small island. While wandering about and looking around, the player discovers that the objects that were put in the fairy-tale machine in the preceding scene are now lying scattered in the grass. Although it is not completely clear to the player at this point, these objects will actually constitute valuable assets when solving various tasks in the world. Cloddy Hans is able to refer multimodally to object found in the grass, and if the user tells him he will pick them up. Thus, it is the task of the player to find the appropriate object, and use this object to bargain with another character, Karen. It turns out that what she is especially interested in jewels. There are three jewels (a diamond, a ruby and an emerald) lying in the grass on the island. In this phase it is possible to encourage graphical gesture references by letting Cloddy Hans say that he doesn't know what a ruby looks like, and if the user says "pick up the red jewel" he might state that he cannot see the difference between green and red. Another possibility is to have more than one ruby. When the users has identified which jewel Karen wants, gotten Cloddy Hans to fetch that jewel to the drawbridge, and promised Karen that they will give it to her when they get over, she will lower the bridge, and let the player and Cloddy Hans pass. As in the first scene, Cloddy Hans will provide the appropriate hints if the user does not understand what to do.

2.2.2 Multimodal behaviours observed during PT1 user tests for Cloddy Hans in HCA Study

Behaviours observed during user tests were already described in D2.2b (Gustafson et al. 2004). The PT1 user tests were not videotaped and there are therefore limitations on what can be interpreted from log files (semantics of a gesture shape, intended gestured object).

Some examples of observed behaviours:

- Speech only behaviour : "Let's take that sword", "put it into the machine",
- Gesture only behaviour: {clicks on the diamond}
- Multimodal behaviour: "We could take one of these couples" {clicks on the figurine portraying the prince}

The observed behaviours were classified by Teliasonera as the following categories:

- Gestural reference + verbal accept. The user accepts a suggestion from the system by saying "yes", or something to the same effect, and pointing at an object.
- Gestural reference + verbal correction. The user verbally rejects a suggestion from the system, and points at another object instead.

- Gestural reference + verbal deictic pronoun. The user says a deictic phrase (like "this one") and points at an object.
- Gestural reference + verbal redundant reference. The user issues a request and backs it up with a pointing gesture (e.g. "take the knife" while pointing at the knife).
- Gestural reference + verbal contradicting reference. The user gives contradictory information in the two input channels (e.g. "take the knife" while pointing at the axe).

3 Specification of input fusion

Fusion of gestures and speech requires considering temporal and semantic dimensions. Regarding semantic fusion we have decided to focus on 1) semantic compatibility between gestured and spoken object implemented via semantic distance computation (which is less strict that object type unification and should be more appropriate for conversational systems for children), and 2) the plural/singular property of objects. This section describes the process of input fusion at a general level. The next section will describe how it is applied / adapted to the two NICE prototypes.

3.1 Analysing semantic combinations of speech and gesture

We will limit ourselves to one reference per NLU frame. We identified 16 semantic combinations of speech and gesture (Table 5).

GI **NLU**	**No message from GI**	**1 message from GI but "noObject"**	**1 object detected by GI** **"select"**	**Several objects detected by GI "referenceAmbiguity"**
No message from NLU	1	2	3	4
1 message from NLU but no explicit reference in NLU frame	5	6	7	8
1 message from NLU with 1 singular reference	9	10	**11**	**12**
1 message from NLU with 1 plural reference	13	14	**15**	**16**

Table 5. Analysing 16 combinations of speech and gesture along the singular/plural dimension of references.

Only cases 11, 12, 15, 16 possibly lead to fusion. We systematically analysed each of these 16 cases. We specify hereafter the instructions to be executed by the IF and the output it will produce. Such instructions consider the following features of speech and gesture references: singular/plural, reference/no reference, semantic distance.

No message from NLU (cases 1 – 4):

1. NLU : no message, GI : no message
Example: The user is exploring the screen
IF output (there is no fusion): no output

2. NLU : no message, GI : noObject
Example: The user points on an area where there is no referenceable object
IF output (there is no fusion): GI NoObject

3. NLU : no message, GI : select
Example: the user points on an object but says nothing.
IF output (there is no fusion): GI select (object name)

4. NLU : no message, GI : referenceAmbiguity
Example: the user surrounders several objects, but says nothing.
IF output (there is no fusion): GI referenceAmbiguity (object names)

1 message from NLU but no explicit reference in NLU frame (cases 5 – 8):

5. NLU : no explicit reference to object, GI : no message
Example: An utterance with no reference and no associated gesture.
IF output (there is no fusion): NLU frame

6. NLU : no explicit reference to object, GI : noObject
Example: The user points on an area where there is no referenceable object while saying something without any reference (e.g. "information").
IF output (there is no fusion): NLU frame + GI No Object
(both frames are forwarded for information to higher level modules)

7. NLU : no explicit reference to object, GI : select
Example: The user points on an object and says "pretty cool".
IF output (inconsistency): NLU frame + GI select (object name)
(The IF signals inconsistency since there is no explicit reference in the speech)

8. NLU : no explicit reference to object, GI : referenceAmbiguity
Example: The user surrounders several objects. Utterance with no reference such as “information”.
IF output (inconsistency): NLU frame + GI referenceAmbiguity (object names)
(The IF signals inconsistency since there is no explicit reference in the speech)

1 message from NLU with 1 singular reference (cases 9 - 12):

9. NLU 1 ref singular to object, GI : no message
Example: "What is this ?" but no gesture.
IF output (there is no fusion): NLU Frame

10. NLU 1 ref singular to object, GI : noObject
Example: "What is this ?" | "Is this your coat ?", gesture on an empty area .
IF output (there is no fusion): NLU frame + GI No Object

11. NLU 1 ref singular to object, GI : select
Example: "Who is this?" + Click on a picture
IF output:
IF GI object is semantically compatible with reference in NLU
THEN (there is fusion) resolve reference in NLU Frame and forward it
ELSE signal inconsistency

12. NLU 1 ref singular to object, GI : referenceAmbiguity
Example: "What is that picture" + selection of paintings with a surrounder gesture
IF output:
IF at least one of the GI Objects is compatible with NLU,
THEN // Case 12 A (there is fusion)
resolve reference in NLU Frame with compatible GI objects and send it
ELSE
IF the NLU Object can also be considered as a plural element (e.g. “the group”)
THEN // Case 12 B (there is fusion)
resolve reference in NLU frame with compatible GI Objects and send it
ELSE signal inconsistency

1 message from NLU with 1 plural reference (cases 13 - 16):

13. NLU 1 ref plural to objects, GI : no message
Example: "Tell me about those pictures"+ gesture on an empty area
IF output (there is no fusion): NLU frame
(Another possibility would be to send all compatible candidates visible on screen)

14. NLU 1 ref plural to objects, GI : noObject
Example: "Tell me about those pictures"+ gesture on an empty area
IF output (there is no fusion): NLU frame
(Another possibility would be to send all compatible candidates visible on screen)

15. NLU 1 ref plural to objects, GI : select
Example: "What are those statues ?" + single click on the statue representing 2 characters
IF output:
IF the single GI Object is semantically compatible with the NLU objects,
THEN /* release plural constraint */ (there is fusion)
Resolve the NLU plural reference with the single GI Object and forward it
ELSE signal inconsistency

16. NLU 1 ref plural to objects, GI : referenceAmbiguity
Example: "Are those members of your family on those pictures" + surrounder around pictures
IF output:
IF at least one of the GI Object is compatible with the NLU Objects
THEN (there is fusion)
Resolve NLU plural reference with all compatible GI Objects and forward it
ELSE signal inconsistency

Suggestions for the management of 2 references in NLU are also proposed in appendix 2.

3.2 Temporal algorithm

A main issue for input fusion is to have a newly detected gesture wait for possibly related spoken utterance. How long should the gesture wait before deciding that it was indeed a mono-modal behaviour ? Default values for delays drive the IF to have gestures wait a little for speech and have speech not wait (or for a very short while) for gestures since this is compatible with the literature and PT1 user tests observations (Martin et al. 2004). We have also introduced the management of startOfSpeech and startOfGesture messages sent to the IF in order to enable a more adequate waiting behaviour from the IF than in PT1. Four temporal parameters of the IF have been defined to answer the following questions:

- How long should a NLU frame wait in the IF for a gesture when no *StartOfGesture* has been detected (*Speech-waiting-for-gesture-short-delay*) ?
- How long should a NLU frame wait in the IF for a gesture when a *StartOfGesture* has been detected (*Speech-waiting-for-gesture-long-delay)* ? default value 6 seconds
- How long should a GI frame wait in the IF for a NLU frame when no *StartOfSpeech* has been detected (*Gesture-waiting-for-speech-short-delay*) ?
- How long should a GI frame wait in the IF for a NLU frame when a *StartOfSpeech* has been detected (*Gesture-waiting-for-speech-long-delay)?*default value 6 seconds

The part of the IF algorithm managing temporal behaviour is specified with the instructions to be executed for each event that can be detected by the IF:

- init the values of the temporal parameters,
- a new NLU frame is received by the IF,
- a new GI frame is received by the IF,
- a StartOfSpeech message is received by the IF,
- a StartOfGesture message is received by the IF,
- a Speech-waiting-for-gesture time out is over,
- a Gesture-waiting-for-speech time out is over.

The IF behaviour is described informally below for each of these events.

Init()

//----------
// Starts with "short" delays when no start of speech or gesture has been received
// When start of speech/gesture will be received, these will be set to longer delays
// since there is a high probability that an associated speech or gesture frame
// will be received afterwards
//----------
Speech-waiting-for-gesture-delay = Speech-waiting-for-gesture-**short**-delay
Gesture-waiting-for-speech-delay = Gesture-waiting-for-speech-**short**-delay

When a new NLU frame is received by the IF

//----------
// Test if a gesture was already waiting for this NLU frame
//----------
If the timeout *Gesture-waiting-for-speech* is running
Then
//----------

// A GI frame was already waiting for this NLU frame
//----------
Call semantic fusion on the NLU and the GI frames
Stop-Timer(*Gesture-waiting-for-speech*)

Else

//----------
// This new NLU frame will wait for incoming gesture
//----------
Start-Timer(*Speech-waiting-for-gesture*)

When a new GI frame is received by the IF

//----------
// Test if a NLU frame was already waiting for this GI frame
//----------
If the timeout *Speech-waiting-for-gesture* is running
Then

//----------
// A NLU frame was already waiting for this GI frame
//----------
Call semantic fusion on the NLU and the GI frames
Stop-Timer(*Speech-waiting-for-gesture*)

Else

//----------
// This new GI frame will wait for incoming speech
//----------
Start-Timer(*Gesture-waiting-for-speech*)

When a *startOfSpeech* message is received

//----------
// A new NLU frame will soon arrive.
// Ensure that the GI frame that is already waiting waits longer
// or that if a new GI frame arrives soon (since a StartOfGesture was received)
// it will wait for the NLU frame
//----------
*Gesture-waiting-for-speech-delay = Gesture-waiting-for-speech-**long**-delay*

If *Gesture-waiting-for-speech* is running
Then
Restart-Timer(*Gesture-waiting-for-speech*)

When a *startOfGesture* message is received

//----------
// A new GI frame will soon arrive.
// Ensure that the NLU frame that is already waiting waits longer
// or that if a new NLU frame arrives soon (since a StartOfSpeech was received)
// it will wait for the GI frame
//----------
*Speech-waiting-for-gesture-delay = Speech-waiting-for-gesture-**long**-delay*

If *Speech-waiting-for-gesture* is running
Then
Restart-Timer(*Speech-waiting-for-gesture*)

When timeout *Speech-waiting-for-gesture* is over

//----------
// A NLU frame has waited for a GI frame which did not arrive.
//----------
Build and send an IF frame containing only the NLU frame
Stop-Timer(Speech-waiting-for-gesture)
Init()

When timeout *Gesture-waiting-for-speech* is over

//----------
// A GI frame has waited for a NLU frame which did not arrive.
//----------
Build and send an IF frame containing only the GI frame
Stop-Timer(Gesture-waiting-for-speech)
Init()

3.3 Semantic fusion algorithm

Semantic compatibility between gestured and spoken objects is evaluated with a graph of concepts connected with a "is-related-to" relation.

Each concept is represented by:

- a name (e.g. « feather Pen », « _Family »),
- a plural boolean (e.g. « true » for the statue of 2 characters),
- a singular boolean (e.g. « true » for the feather Pen),
- a boolean describing if it is an object in the study (true for «pictureColoseumRome) or an abstract concept (false for "_Mother"),
- the set of semantically related concepts (generic relation "isRelatedTo").

A reference in speech is represented by: a boolean stating if it is solved, a boolean stating if it is plural/singular, a boolean stating if it is numbered (if yes, an attribute gives the number of referred objects : « two » in the reference « these two pictures»).

A perceptual group is represented by the same attributes as a single concept, and by the set of concepts which might be perceived as a group (e.g. the set of pictures above the desk). Perceptual groups can be used for driving the fusion or for studying user's behaviour in the logged files.

The identified cases of semantic combinations described above are integrated in a single algorithm for semantic fusion. The informal algorithm below only details cases for which one message has been sent by the NLU and one by the GI (cases 6-7-8, 10-11-12, 14-15-16 of the analysis).

Algorithm Semantic Fusion (NLU frame, GIFrame)

```
//-------------------------------------------------------------------------------
// Manage each multimodal combination case
// We suppose that one NLU frame and one GI frame have been received by the IF
//-------------------------------------------------------------------------------
IF there is no reference in the NLU frame
THEN
        //-------------------------------------------------------------------------------
        // CASES 6-7-8
        //-------------------------------------------------------------------------------
        Group both frames and send them

ELSE
        IF there is only one reference in the NLU frame
        THEN
                IF the reference is singular
                THEN call Semantic Fusion Singular NLU
                ELSE call Semantic Fusion Plural NLU
```

Semantic Fusion Singular NLU (NLU frame, GI frame)

```
//-------------------------------------------------------------------------------
// The Referential Expression in the NLU frame  is singular
// CASES 10- 11 -  12A (singular)
//-------------------------------------------------------------------------------
IF      there is at least one object selected by GI,
        which is semantically compatible with the NLU reference
THEN
        // Do semantic fusion (possibly not considering plural constraint
        // if there was several gestured objects)
        Resolve the NLU reference with the compatible gestured object(s)
        Send the modified NLU frame
ELSE
        // No gestured object revealed compatible with the NLU reference
        Signal inconsistency
        Send NLU frame and GI frame
```

Semantic Fusion Plural NLU (NLU frame, GI frame)

```
//-------------------------------------------------------------------------------
// The Referential Expression is plural
// CASES 14 - 15 – 16 – 12B (reference can be plural)
//-------------------------------------------------------------------------------
IF more than one object from GI is semantically compatible with the NLU reference
THEN
        // Do semantic fusion
```

Resolve the plural NLU reference with the compatible gestured object(s)
Send the modified NLU frame

ELSE

//----------

// Manage perceptual groups

//----------

IF

there is only one object from GI compatible with NLU reference
and this object belongs to a perceptual group

THEN

// Do semantic fusion
Resolve the plural NLU reference with the perceptual group
Send the modified NLU frame

ELSE

IF the GI object is compatible with the NLU reference but does not belong to a perceptual group

THEN

// Do semantic fusion (not considering plural constraint)
Resolve NLU reference with the compatible gestured object
Send the modified NLU frame

ELSE

// No gestured object revealed compatible
// with the NLU plural reference
Signal inconsistency
Send NLU frame and GI frame

Compatible (GI object, NLU reference)

Two objects are compatible if they are both
- Number Compatible and
- Semantically Compatible

Semantically Compatible (GI object, NLU reference)

IF the NLU referential expression holds a concept C
THEN
Compute distance between this NLU concept and the GI object in the ontology
Return true if this distance is not infinite

Number Compatible (GI object, NLU reference)

// The value of the number feature of theNLU reference could also be used

IF the plural feature of the object from GI is true,
and the number feature of the NLU reference is plural
THEN Return true
ELSE
IF the singular feature of the object from GI is true,
and the number feature of the NLU reference is singular
THEN Return true
ELSE Return false

4 Application to the two NICE prototypes

The IF module consists of an internal Fusion Module which include parameters tuned for each of the two NICE prototypes. Two parsers have been written for the two NLU modules. Two generators of IF Frames have been developed for the Character Module (NISLab) and the dialogue module (Teliasonera). Two simple ontology's have been designed to enable the IF to detect if two concepts are related.

The IF parameters are the temporal parameters (Speech-waiting-for-gesture-short-delay, Speech-waiting-for-gesture-long-delay, Gesture-waiting-for-speech-short-delay, Gesture-waiting-for-speech-long-delay), the name of the file describing the ontology, the module to which the IF frame should be sent (Dispatcher for the FTW prototype, the CM for the HCA Study prototype). The internal architecture of the IF is displayed in Figure 1.

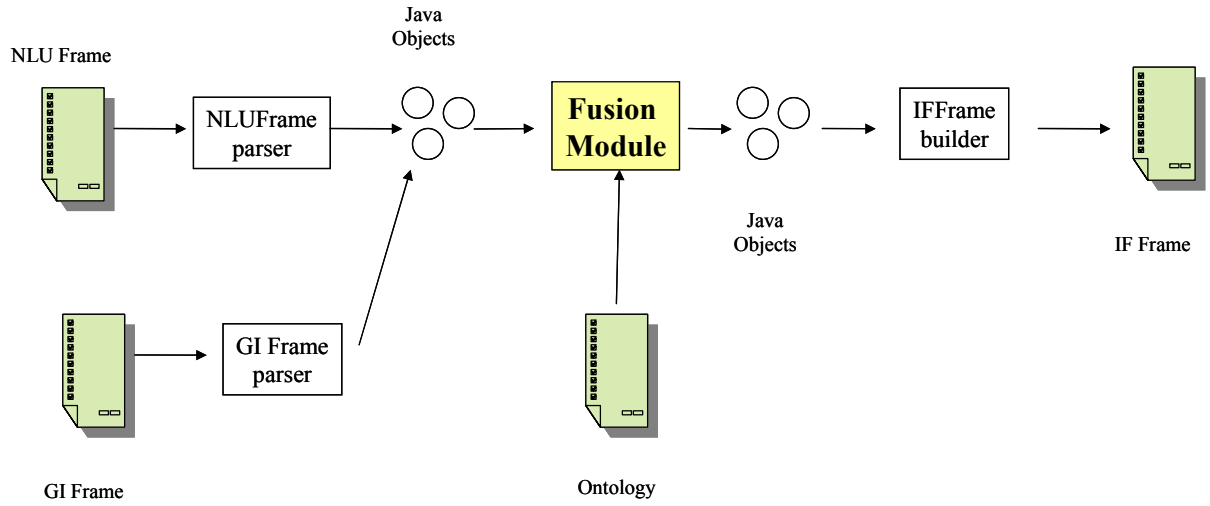

Figure 1. Internal architecture of the IF module.

4.1 Integrating IF in "HCA Study"

A NLU frame parser has been developed to parse XML frames produced by the NISLab NLU module into Java objects that can be manipulated by the fusion module.

Figure 2 represents a section of the ontology that has been defined for enabling semantic processing in the IF. This ontology informs the system that some objects such as the DeskBooks can be referred to by singular or plural reference. Perceptual groups can be defined in the ontology (e.g. the pictures above the desk) in order to study the plural behaviour of users. This ontology will be adapted to match the final list of object for PT2.

After fusion, an IF frame is sent to the NISLab CM module. Appendix 3 provides illustrative examples of the IF output in the HCA Study prototype for each of the 16 identified cases during analysis. An attribute called "fusionStatus" is used in the IF frame to indicate if the input was mono-modal ("none"), successful ("ok") or unsuccessful ("inconsistency").

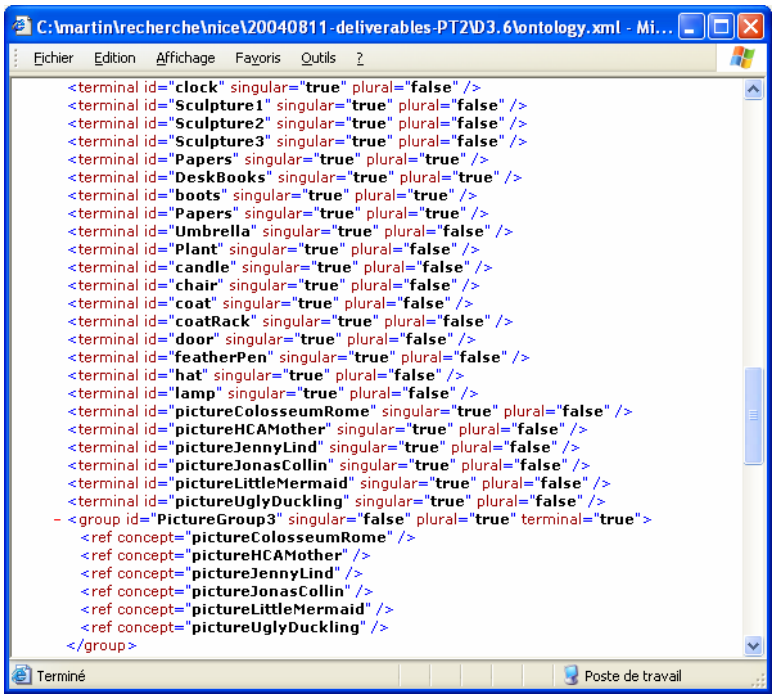

Figure 2. Part of the ontology used by the IF for the HCA Study.

4.2 Integrating IF in "Fairy Tale World"

Regarding the Fairy Tale World prototype, the Teliasonera NLU and the IF modules cooperate for managing fusion at different levels. Early fusion (a single gestured object followed by a single reference in NLU such as « Pick this thing up » followed by a gesture) is processed by the Teliasonera NLU. The IF manages "late fusion" for these simple cases by considering compatibility between the gestured object and the spoken object, as well as fusion of more complex cases involving constraints on types and plural/singular property of objects and cases for which gesture comes after speech.

Similarly to the HCA Study prototype, a simple ontology, an NLU frame parser and an IF Frame generator have been designed.

This section gives a brief overview of the system architecture which has been updated for the "Fairy Tale World" prototype since PT1. It shortly describes the servers used in NICE fairy-tale game system. The fairy-tale game involves a number of embodied conversational fairy-tale characters. To make the animated fairytale characters appear lifelike, they have to be autonomous, i.e. they must do things even when the user is not interacting with them. At the same time they have to be reactive and show conversational abilities when the user is interacting with them. To build a system that is both autonomous and reactive at the same time has led to the choice of the event driven, asynchronous system architecture that is shown in Figure 3.

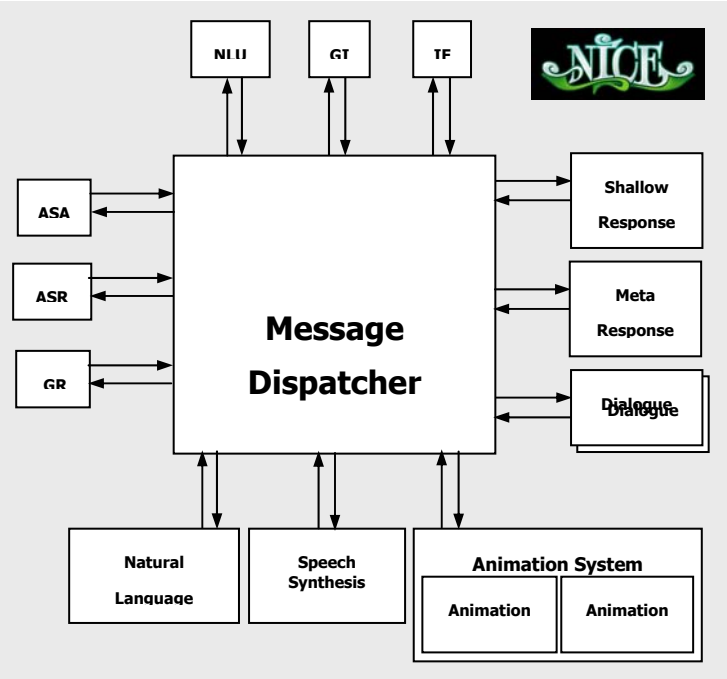

Figure 3. Updated system architecture for the FTW prototype

The modules in the system are briefly described in Table 6.

Since the dialogue system components are written at different sites and in different programming languages we have chosen a modular architecture, where modules communicate via central Message Dispatching server, and where all messages are sent in text form over a TCP/IP socket. The Message Dispatcher has two parts:

1) A low-level Broker part that handles the hand shaking and message routing on the TCP/IP socket level.
2) A higher-level Message Dispatcher part that handles the information flow and timings in the system

Since the low-level Broker part has a very basic functionality that is independent of the typology and information flow in the system architecture, we have chosen to use a publicly available Broker system developed at KTH (http://www.speech.kth.se/broker). There are two advantages with this package: 1) it does not impose a specific system architecture, but simply handles the socket communication and message routing, 2) there are brokerClient packages in Java/c++/tcl/perl/prolog that simplify the implementation of message handling in the dialogue component modules.

MessageDispatcher	Events from all servers are sent to a central Dispatcher server, that timestamps and logs them, and then routes them to the appropriate destination(s). It also generates timeouts in the system.
AcousticSpeechAnalyzer	Acoustically analyses the user spoken input and generates e.g. EndOfTurn & SpeechTooLoud events.
AutomaticSpeechRecognizer	Generates a text string from the user's spoken input, accompanied with recognition confidence scores.
NaturalLanguageParser	Parses the text string generated by the ASR server.
ShallowResponseGenerator	Generates simple responses to a number of out-of-domain user requests.
MetaResponseGenerator	Generates error handling turns like "you talk to loud" or "didn't you hear me, do you want me to."
GestureRecognizer	Recognizes the users Gestural input.
GestureInterpreter	Interprets the output from the Gesture Recognizer.
InputFusion	Takes the result of NLP and GI and generates a combined interpretation.
DialogueManagers (one per character)	Performs reference resolution, decides on user speech act, updates dialogue state and goal agenda, chooses an appropriate next system action.
NaturalLanguageGenerator	Multimodal Surface Generation of the system output.
Synthesizer	Creates sound file with corresponding viseme animation track, as well as a time-stamped animation request track sends it all back to the AnimationHandler.
AnimationHandler	Translates action requests from the DM into requests to the AnimationRenderer. Has an internal animation queue and can construct complex actions. Will ask the synthesizer to generate a sound file and an animation track that it will send to the AnimationRenderer.
AnimationRenderer	Performs the animation and action that the AnimationHandler asks for and informs the Dispatcher when it is ready. Is able to generate trigger event when a character enters a trigger area.

Table 6. List of the modules.

The high-level part of the Message Dispatcher gives the system an event-driven information flow, where the communication between the servers is coordinated by the Message Dispatcher. All messages are asynchronous, i.e. the sending module sends the message and proceeds to its next task without waiting for an answer. The Message Dispatcher is responsible for coordinating input and output events in the system, by time-stamping all messages from the various modules. The behaviour of the Message Dispatcher is controlled by a set of simple rules, specifying how to react when receiving a message of a certain type from one the modules. Since the Message Dispatcher is connected both to the input channels and the output modalities, it can increase the system's responsiveness by giving fast but simple feedback on input events (for example by sending a request for a eyebrow raise animation to the Animation System as soon as it receives a StartOfSpeech event from the ASR Module).

The Message Dispatcher can also increase the system stability using timeouts. For example, if it has sent an ASR string to the NLU and has not received a NLU event within one second, it can take certain actions. Lastly, the architecture above is very modular, in the sense new modules can be added without having to change the previous modules. For example if we would like to add a topic predictor that can use both the ASR string and the NLU analysis, neither the ASR module nor the NLU module have to be updated with information on where to send their results, as all communication goes via the Message Dispatcher. A list of messages that the Messages Dispatcher handles, their sources and their routing destinations is shown in Appendix 4.

5 References

Avaya, W. C., Dahl, D., Johnston, M., Pieraccini, R. and Ragget, D. (2004). EMMA: Extensible MultiModal Annotation markup language. W3C Working Draft 14 December 2004., W3C.http://www.w3.org/TR/emma/

Boye, J., Gustafson, J., Bell, L., Wirén, M., Martin, J.-C., Buisine, S. and Abrilian, S. (2004). NICE Deliverable D1.1b-2. Requirements specification for domain information, personality information and dialogue behaviour for the second NICE fairy-tale prototype.

Gustafson, J., Boye, J., Bell, L., Wirén, M., Martin, J.-C., Buisine, S. and Abrilian, S. (2004). NICE Deliverable D2.2b. Collection and analysis of multimodal speech and gesture data in an edutainment application.

Johnston, M., Bangalore, S., Visireddy G., et al. (2002). MATCH: An Architecture for Multimodal Dialog Systems. 40th Annual Meeting of the Association for Computational Linguistics (ACL) Philadelphia, July. http://www.research.att.com/%7Ejohnston/matchacl02.pdf

Kaiser, E., Olwal, A., McGee, D. and Benko, H., Corradini, A., Li, X., Cohen, P., Feiner, S. (2003). Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. Fith International Conference on Multimodal Interfaces (ICMI'03) Vancouver, British Columbia, Canada, ACM Press. http://www1.cs.columbia.edu/~aolwal/projects/maven/maven.pdf

Landragin, F., Bellalem, N. and Romary, L. (2001). Visual Salience and Perceptual Grouping in Multimodal Interactivity. First International Workshop on Information Presentation and Natural Multimodal Dialogue Verona, Italy. http://www.loria.fr/~landragi/publis/ipnmd.pdf

Martin, J.-C., Buisine, S. and Abrilian, S. (2004). NICE Deliverable D1.1-2a Part 2. Requirements and Design Specification for Gesture and Input Fusion in PT2 HCA Study.

6 Appendixes

6.1 Appendix 1: Requirements on IF in PT2 HCA Study from D1.1-2a

12. Rapid prototyping in order to quickly have a first full-gesture-chain input processing solution, including input fusion.

=> *A first temporal algorithm was developed.*

13. Adequate solutions to any timing issues arising from the need to potentially fusion gesture input and natural language input. The timing solutions must not jeopardise the system's real-time performance.

=> *StartOfSpeech and StartOfGesture timeout behaviours have been added.*

14. Adequate and reliable input fusion for the two tasks mentioned in (10) above, fusioning natural language module output concepts with gesture information.

=> *Ok: cf. specifications (finally only one scenario was kept in agreement with NISLab)*

15. No forwarding of input fusion tasks to the character module, all tasks must be solved by the input fusion module. The input fusion module should either deliver a 1-best semantic fusion solution to the character module or report to the character module the nature of any problem, of referential ambiguity or otherwise, it may have. Character module meta-communication will then take care of the problem.

=> *Ok*

4.4.1 Expects as input
The IF expects to receive:
" From NLU: nluFrames semantic representation containing one or several references (some intra-utterance co-references might be already solved by the NLU)
" From GI: giFrames
=> *Ok*

4.4.2 Provides as output
The IF has to produce and send to the CM a multimodal semantic representation where the semantic frame sent by NLU is completed at a semantic level with semantic information sent by GI. This include resolving some co-references and eventually leaving some unresolved co-references to be handled by the CM with the help of dialog context (e.g. no gesture was detected, or possible fusion solution had very low confidence score).
=> *Ok*

4.4.3 Forwards mono-modal behaviours to the CM

In the case of mono-modal behaviours, the nluFrame or the giFrame will be embedded in an ifFrame and forwarded to the CM. Several sequential giFrames will be grouped by the IF and sent to the CM as a single sequential selection of several objects (waiting duration between sequential giFrames to be fixed).
=> Ok

4.4.4 Merges related nluFrames and giFrames

Various kinds of " clean " multimodal behaviour should be managed:
- "Put this one in the bag" + one gesture
- "Who is this person ?" + one gesture
- "Who is the lady ?" + one gesture
- "Is this woman the same as this woman ?" + two gestures
- "I want information on these objects" + sequential gesture on several objects
- conjunction of references: "this one and this one" + two gestures
- disjunction of references: "this one or this one" + two gestures

The IF expects a single nluFrame per turn. It should detect multimodal behaviour and produce an output frame in which the nluFrame is completed with interpretations from gesture. This " late " fusion (since based on mono-modal semantic interpretation) will be based on both temporal and semantic criteria. The possibility to have early fusion eventually leading to the decision of cancelling previously merged hypotheses does not seem adequate for the NICE project as it would require complex maintenance of hypotheses throughout all the modules.
The adequacy of existing unification based techniques and algorithm will be investigated for computing compatible combinations of spoken and gestured objects attributes (type, number, name, time and rank of reference). Inconsistency and contradiction between speech and gesture will be dealt differently in different dialog acts (question, confirmation, negation, correction) detected by the NLU ("Is this the little mermaid?" + gesture on HCAmother's picture vs. "This is the little mermaid" + gesture on HCAmother's picture).
We will consider extending the list of physical referenceable objects in the GI/IF with a list of abstract objects and their possible mapping to physical objects in the GI/IF (e.g. " fairy tales world " concept bound to the door, " ugly duckling " fairy tale linked to the corresponding picture, relation between the object pictureHCAMother and the family concept). This would enable the GI/IF to detect whether the user is speaking of the object itself or of what it represent and enable fusion of verbal description of abstract concept with gesture on physical objects.
=> Type, number and time constraints are tackled. The IF the algorithm which is based on unification principles similar to other multimodal input systems but enables the individual management of combinatorial ambiguous cases including releasing some constraints (cf. IF algorithm).

4.4.5 Use internal confidence score computation for selecting a 1st best candidate for input fusion or detecting and signalling a limited typology of problematic cases

The IF will internally use confidence scores associated to gesture and input fusion hypotheses to select a 1st best candidate for input fusion that will be forwarded to the CM without confidence score. Possible appropriate schemes will have to be studied for multimodal score computation including the consideration of selected but limited solutions in other multimodal systems (e.g. multiplication/addition of mono-modal scores which have limitations).
The IF will use these scores to detect and signal to the CM the following " problematic " cases: referential ambiguity (e.g. different objects having high and equal scores due to gesture at equal distance between two objects, gesture on two objects having overlapping bounding

boxes and similar compatibility with speech), inconsistent behaviour (e.g. affirmative utterance referring to an object and gesture on another object), gesture object unknown (e.g. the user might have gestured on a non-referenceable object), noisy input (e.g. garbage gesture). Character module meta-communication will then take care of the problem.

=> *The scores are kept internally by the GI but can be produced in the output on demand (one of the parameters of the IF configuration file). An output score from IF revealed not necessary.*

4.4.6 Manage incompatible number of gesture and referring expression

Several potential incompatibilities between speech and gesture can be expected regarding the number of referenced objects (e.g. no referring expression in speech and one gestured object, two referring expressions in speech and only one gestured object) such as in "What is this ?" + gesture on 2 objects. The IF will manage singular / plural combinations (e.g. " these statues " + click on the single object representing two statues, this/these boot(s), this/these book(s), " is this your family ? " + encircling several pictures).

=> *Ok.*

4.4.7 Manage incompatible object type in speech and gesture

Incompatible types of object might be observed in gesture and speech (e.g. " I tried to get a fairy tale by clicking on his hat " mentioned in D7.2a part 1 page 12). This type incompatibility error will be detected by the IF and signalled to the CM.

=> *Ok.*

4.4.8 Avoid time-consuming solutions to input fusion

One time-consuming issue in input fusion is that in the case a gesture is detected, the input fusion module needs to wait for eventually following spoken utterance for potential fusion before deciding that the gesture is indeed a mono-modal behaviour. The same might apply to speech before gesture combination if it is observed in user's behaviour where the IF has to delay forwarding of NLU interpretation to the CM to check if no gesture occurs shortly after speech.

In the case of sequential selection of different objects (resulting in sequential giFrames), the IF needs to reset temporal delays after each received giFrame in order to wait until the end of the gesture sequence and send this as a single but sequential selection of several objects, or merge it with any nlu frame compatible with the number and types of gestured objects.

These real time performances are even more difficult to achieve since some individual users might display disfluent or "unusual" temporal behaviours such as unexpected long delay between associated gestures, long delay between subparts of a single multimodal pattern (e.g. "this one"+ <gesture> … "and this one" + <gesture>), unexpected long delay between gesture and speech (D7.2a page 16 reports " 1.4 % of errors were due to the intriguing gesture/speech timing behaviour of a single user "). Intrusive solutions might involve to elicit/prompt synchronised behaviour which is easier to process (e.g. HCA saying " speak and gesture at the same time otherwise I will not understand ") or discourage synchronised behaviour by fast feedback (including on mono-modal behaviour).

In PT1, it was decided that 1) after a gesture detection, the IF would wait 3 seconds for any nluFrame, 2) the IF would not wait for any gesture after receiving an nluFrame. These two durations could be modified in a text file.

There is a limited set of solutions to this problem which will be considered in PT2:

1) NLU and GR send "startOfSpeech" and "startOfGesture" messages to the IF

=> *Ok.*

2) Different temporal delays values are set to different settings : WOZ vs. speech recognition
=> *Revealed unnecessary since real speech recognition is integrated (Yet, the StartofSpeech can also be used in a WOZ setting and temporal parameters can be modified in configuration file).*

3) If the co-references in the nluFrame have been solved by the NLU ("This picture, is it about the little mermaid"), the IF does not wait for gestural input and sends a NLUframe only to the character module. If the IF module receives a further gesture on the same referred object, it decides that this is already managed by the CM and ignores the gesture.
=> *Not done.*

4) Have different delays duration for different users
=> *Not done.*

5) If the nluFrame contains several references and one of them has been solved by the IF, segment the nluframe, send solved reference to the CM, keep unresolved reference(s) in the IF.

=> Not done.

6.2 Appendix 2: Processing of 2 references in the NLU frame

	One object detected by GI "select"	**Several objects detected by GI "referenceAmbiguity"**
One message from NLU with two singular reference	**IF** the gestured object and the spoken objects belong to the same perceptual group **THEN** Send elements of the perceptual group **ELSE** Signal inconsistency	**IF** at least two gestured objects are compatible with the two spoken objects **THEN** Resolve reference with the two compatible objects (1st and 2nd best if there are more than two) **ELSE** **IF** there is only one compatible object, and it can be considered as a plural element, and the two spoken objects are both compatible with this object **THEN** Resolve reference with this object **ELSE** Signal inconsistency
One message from NLU with two plural reference **OR** **One message from NLU with one singular and one plural reference**		**IF** at least two gestured objects are compatible with each of the NLU Objects **THEN** Resolve reference with **ELSE** **IF** for any of the 2 spoken objects, there is only one compatible gestured object, **IF** it can be considered as plural **THEN** Resolve reference with it **ELSE** Signal inconsistency

Corresponding instructions:

FOR each Referential Expression **in** the NLU Referential Expression List

Resolve this Referential Expression as if it was the only one
with Objects selected from the GI

Rebuild NLU frame

Remove selected gestured objects from the list of gestural candidates
// Temporal information from GI and NLU could also be used to resolve successive References in NLU

6.3 Appendix 3 : Examples of fusion output for each cases

The table below provides illustrative examples of IF output for each of the 16 identified cases (one case per row). Each box include the category of the module output (e.g. “select” for GI) and an example of associated value (e.g. “picturejennylind”).

	Message from NLU	**Message from GI**	**relation**	**IF output**	**expected fusion status**
1.	none	none		none	none
2.	none	noObject floor		GI NoObject	none
3.	none	select picturejennylind		GI select (object name)	none
4.	none	referenceAmbiguity picturejennylind + picturelittlemermaid		GI referenceAmbiguity (object names) picturejennylind + picturelittlemermaid	none
5.	no reference to object hello	none		NLU frame	none
6.	no reference to object hello	noObject floor		NLU frame + GI No Object	none
7.	no reference to object hello	select picturejennylind		NLU frame + GI select (object name)	inconsistent
8.	no reference to object hello	referenceAmbiguity picturejennylind + picturelittlemermaid		NLU frame + GI referenceAmbiguity (object names) picturejennylind + picturelittlemermaid	inconsistent
9.	1 ref singular to object what's this	none		NLUFrame	none
10.	1 ref singular to object what's this picture	noObject floor		NLU frame + GI No Object	none

	Message from NLU	Message from GI	relation	IF output	expected fusion status
11.	1 ref singular to object what's this what's this picture	select picturejennylind picturejennylind	compatible	NLU frame with reference resolved	ok
	1 ref singular to object what's this picture	select featherpen	incompatible	NLU frame + GI select (object name)	inconsistent
12.	1 ref singular to object what's this what's this picture	referenceAmbiguity picturejennylind + featherpen picturejennylind + featherpen	at least 1 compatible	NLU frame with reference resolved picturejennylind + featherpen picturejennylind	ok
	1 ref singular to object what's this picture	referenceAmbiguity featherpen + candle	incompatible	NLU frame + GI referenceAmbiguity (object names) featherpen + candle	inconsistent
13.	1 ref plural to objects what are these	none		NLUFrame	none
14.	1 ref plural to objects what are these	noObject floor		NLUFrame	none
15.	1 ref plural to objects what are these	select picturejennylind	compatible	NLU frame with reference resolved	ok
	1 ref plural to objects what are these pictures	select featherpen	incompatible	NLU frame + GI select (object name)	inconsistent
16.	1 ref plural to objects what are these pictures what are these pictures what are these	referenceAmbiguity picturejennylind + featherpen picturejennylind + picturelittlemermaid picturejennylind + picturelittlemermaid	at least 1 compatible	NLU frame with reference resolved picturejennylind picturejennylind + picturelittlemermaid picturejennylind + picturelittlemermaid	ok
	1 ref plural to objects what are these pictures	referenceAmbiguity featherpen + candle	incompatible	NLU frame + GI referenceAmbiguity (object names) featherpen + candle	inconsistent

6.4 Appendix 4: Messages exchanged in FTW prototype

Source	Messages	Explanation	Destination(s)
ASR	StartOfSpeech	The Recognizer has detected that the user started speaking.	ASA, IF, AnimationHandler
ASR	EndOfSpeech	The Recognizer has detected that the user stopped speaking.	ASA, IF, AnimationHandler
ASR	AsrResult	The speech recognizer has been able to generate a speech recognition result.	NLP, SRG, MRG
ASR	RecognitionFailed	The speech recognizer has failed to generate a speech recognition result.	DM, MRG, AnimationHandler
ASA	SpeechTooLoud	The user speech input was too loud.	MRG
ASA	SpeechTooLow	The user speech input was too low.	MRG
ASA	BackchannelPlace	The speaker indicates prosodically at a pause that her turn has not ended - a possible back-channel place.	AnimationHandler
ASA	EndOfTurn	The speaker indicates prosodically that her turn has ended - a possible turn-taking place.	DM, AnimationHandler
NLP	NlpResult	The NLP has arrived at an analysis of the latest utterance.	DM, IF, Dispatcher
NLP	NluFailed	The NLP failed to deliver an analysis of the ASR result.	DM, Dispatcher
SRG	ShallowResponseDone	The SRG has been able to generate a response to the latest utterance.	DM, Dispatcher
SRG	ShallowResponseFailed	The SRG has not been able to generate a response to the latest utterance.	Dispatcher
MRG	MetaResponseDone	The MRG has been able to generate a response to the latest utterance.	DM, Dispatcher
MRG	MetaResponseFailed	The MRG has not been able to generate a response to the latest utterance.	Dispatcher
GR	GrResult	The Gesture Recognizer has been able to produced a gesture recognition result	DM, AnimationHandler
GR	GrFailed	The Gesture Recognizer has not been able to produced a gesture recognition result.	DM, AnimationHandler
GI	select	The Gesture Interpreter has been able to interpret the gesture as a select object.	NLU, DM, IF AnimationHandler
GI	referenceAmbiguity	The Gesture Interpreter has been able to single out only one object in the select gesture, thus there is a reference ambiguity.	DM, IF, AnimationHandler
GI	noObject	The Gesture Interpreter has not been able to identify which object the user selected.	DM, AnimationHandler
IF	IfResult	The Input Fusion been able to Fusion the GI and NLP results.	DM, AnimationHandler
IF	IfFailed	The Input Fusion not been able to Fusion the GI and NLP results.	DM, AnimationHandler

Table 7. List of messages in the NICE fairy-tale game system.

Source	Message	Explanation	Destination(s)
Dispatcher	Timeout	A certain amount of time has passed since the last user input and/or system output. Used by MRG to fire meta utterances, by DM to drive the dialogue forward and by the AnimationHandler to manage the idle behaviour.	DM, MRG, AnimationHandler
DM	ConveyRequest	The DM has generated a verbal turn that it wants the character to say .	AnimationHandler
DM	PerformRequest	The DM has generated an action that it wants the character to perform .	AnimationHandler
AnimationHandler	PerformDone	The AnimationHandler has completed all actions required to fulfil a convey or perform request from the DM.	DM,
Synthesizer	SynthesisGenerated	The synthesizer has been able to generate a sound file with corresponding viseme animation track, as well as a time-stamped animation request track.	AnimationHandler
AnimationRenderer	RequestStarted	The AnimationRenderer has received and started request for an animation or action.	AnimationHandler
AnimationRenderer	RequestFailed	The AnimationRenderer has perform a request for an animation or action.	AnimationHandler
AnimationRenderer	AnimationRequestDone	The AnimationRenderer has finished a request for an animation or action.	AnimationHandler
AnimationRenderer	SlotEvent	An object has been inserted into one of the slots of the fairy-tale machine.	DM, AnimationHandler
AnimationRenderer	TriggerEvent	The animation system has detected that the character has moved into a trigger.	DM, AnimationHandler

Table 8. Continued list messages in the NICE fairy-tale game system.