Multimodal Fusion: A New Hybrid Strategy for Dialogue Systems

Pilar Manchón Portillo
University of Seville
pmanchon@us.es

Guillermo Pérez García
University of Seville
gperez@us.es

Gabriel Amores Carredano
University of Seville
jgabriel@us.es

ABSTRACT

This is a new hybrid fusion strategy based primarily on the implementation of two former and differentiated approaches to multimodal fusion [11] in multimodal dialogue systems. Both approaches, their predecessors and their respective advantages and disadvantages will be described in order to illustrate how the new strategy merges them into a more solid and coherent solution. The first strategy was largely based on Johnston´s approach [5] and implies the inclusion of multimodal grammar entries and temporal constraints. The second approach implied the fusion of information coming from different channels at dialogue level. The new hybrid strategy hereby described requires the inclusion of multimodal grammar entries and temporal constraints plus the additional information at dialogue level utilized in the second strategy. Within this new approach therefore, the fusion process will be initiated at grammar level and will be culminated at dialogue level.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and presentation]:** User Interfaces - *Theory and methods, User interface management systems, Voice I/O.*

General Terms

Algorithms, Performance, Theory.

Keywords

Multimodal fusion, dialogue systems, NLP

1. INTRODUCTION

The ultimate goal for human-computer interfaces is the achievement of the maximum level of flexibility, efficiency, naturalness and usability. According to the literature, multimodal interfaces offer a number of advantages over uni-modal interfaces, namely, a more flexible and efficient environment that enables users to interact more freely and on their own terms with automated systems. It seems quite clear that at least in complex environments, multimodal interfaces are the future of human-computer interaction.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.
Copyright 2006 ACM 1-59593-541-X/06/0011...$5.00.

Multimodal systems have been largely studied since the 80´s and a number of approaches to multimodal events representation and fusion strategies have been proposed. The following proposals are to some degree related to the present strategy:

- Johnston and Cohen [3] [4] proposed the transformation of multimodal inputs into typed feature structures with semantic content combined by unification.
- Nigay and Coutaz [13] proposed the "melting pots", which contain types of structural parts of a multimodal event. These melting pots are the result of the fusion of simple input events through microtemporal, macrotemporal or contextual fusion procedures.
- Vo, Waibel and Wood [14] [15] [16] implemented partial action frames combined by a domain independent frame-merging algorithm (multi-state mutual information network).
- Pérez, Amores and Manchón [11] proposed two different strategies to multimodal fusion, the first one based on Johnston´s approach with some additional considerations, and the second one based on the inclusion of dialogue level information in the fusion process with some parallelisms with [13] and [14]

All of these proposals are precedents to the fusion strategies discussed here and have points in common with the arguments in favour of the new hybrid strategy. It is nonetheless important to note that this new strategy is a direct outcome of the Pérez, Amores and Manchón [11] proposal and implementation.

In order to understand the new strategy and its implementation, it is essential to understand the philosophy of the systems where it can be implemented.

The system used is MIMUS, the multimodal sibling of Delfos. It is based on the ISU (Information State Update) approach and consists of a collaborative dialogue manager linked to a Natural Language Understanding Module. It is a mixed-initiative dialogue system driven both by the information provided by the user and by the expectations within the dialogue manager.

The system kernel consists of:

- A Natural Language Understanding (NLU) module
- A Dialogue Manager

The former performs the lexical-syntactical analysis and generates information states. The latter updates and handles the information states by means of a number of dialogue rules.

The Information State handled in this system is a feature-value structure called DTAC [10]. The main features are DMOVE,

TYPE, ARG and CONT. The example in figure 1 illustrates the information state structure.

All the examples presented will be related to the Smart Home scenario, which is the target implementation of the project within which this research is being developed.

TURN ON THE KITCHEN LIGHT

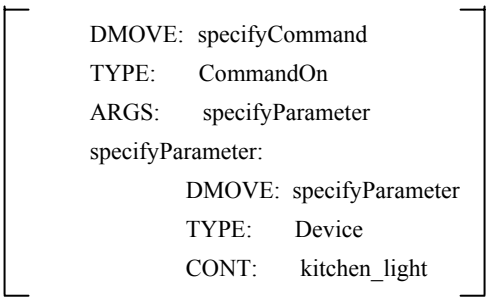

Figure 1: DTAC Information State

As mentioned above, the dialogue manager operates by means of update rules. An example of such rules is available in figure 2.

```
(RuleID:    MAKECALL;
    PriorityLevel:    15;
    TriggeringCondition:
        (DMOVE:specifyCommand,
            TYPE:MakeCall);
    DeclareExpectations: {
        Dest <= (DMOVE:specifyParameter,
               TYPE:Name|PhoneNumber);
    }
    SetExpectations: {
        Confirm <= (DMOVE:answerYN);
    }
    ActionsExpectations: {
        [Dest] => {
               ExecuteDMFunction(MakeCallDest);
        }
        [Confirm] => {
               ExecuteDMFunction(MakeCallDisam);
    }
    PostActions: {
        @if ((@is-MAKECALL.Confirm.TYPE == "YES")
&& (@RecoveredState() == "True")) @then {
        ExecuteDMFunction(MakeCallDisam);
        }
        (…)  })
```

Figure 2: Dialogue Update Rule

The dialogue manager is driven both by its own expectations and the user's input. The previous example illustrates how the dialogue manager handles the information at different levels:

- What dialogue move (DMOVE) activates the rule (TriggeringConditions)
- What additional information is required by rule (DeclareExpectations).
- What additional dialogue moves are required to fulfil the rule (SetExpectations).
- What to do if the information required in either of the former sections is not available (ActionExpectations).
- What to do when all the pre-requisites have been fulfilled (PostActions)

The current system set-up consists of a touch-screen where the user can visualize their home, a microphone to receive speech, and speakers to produce speech. The users can therefore utter spoken commands, click on the screen or do both at the same time, and the system may speak, display on the screen or do both at the same time.

Once the general system architecture has been illustrated, the previous fusion strategies on which the new one is based will be described in section two. In section three the new hybrid strategy will be presented. Section four offers the pros and cons of the strategies described. A summary of the conclusions and future work will be presented in section five.

2. PREVIOUS FUSION STRATEGIES

2.1 Strategy 1: Grammar-driven strategy

This strategy was analogous to Johnston's proposal [3] [4]. It implies the use of a unification based parser and the inclusion of modality and temporal constraints at unification level. The Pérez, Amores and Manchón proposal [11] adds however a higher degree of flexibility.

This multimodal dialogue system is very versatile and allows the full spectrum of multimodal inputs, that is, from speech-only events to click-only events, including mixed-modality events such as an utterance like "Turn this on" complemented by a click on the screen. One of the main issues to be taken into account when advocating for a certain fusion strategy is the way a multimodal event is perceived, i.e., as a single communicative act or as several complementary acts. In the first case, it seems intuitive to have one single grammar capable of coping with the combined modalities. This grammar would therefore include productions with components from different modalities:

```
Command-> CommandX(Speech) Parameter(Click)
```

The problem may arise if the system allows for multiple simultaneous tasks, where the ambiguity generated could not possibly be handled without additional information. The graph in figure 3 illustrates this strategy.

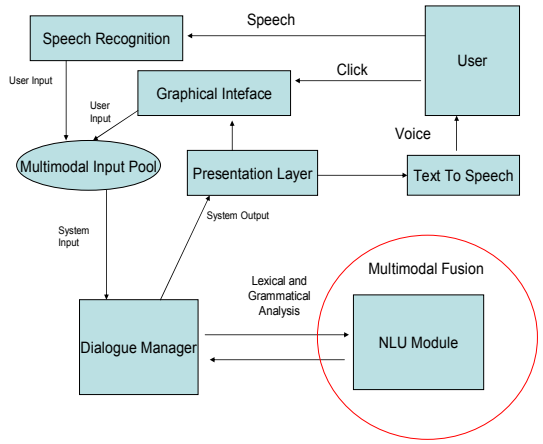

Figure 3: Grammar-driven strategy

The NLU module parses the input (whatever the input modality/ies may be) and adds additional features in the corresponding slots:

MOD: Modality of the event

INIT: Time at which the event started

END: Time at which the event ended

Together with a set of logical operators, this additional information defines the time and modality constraints in the grammar. Figure 4 illustrates an example of what the unification rules looks like.

```
(Rule 1: Command -> ComdOn DevSpec)
        { @up = @self-1;}
(Rule 2 : Command -> DevSpec ComdOn)
        @up.DevSpec =a @self-1;
 @if((@self-1.MOD == CLICK) &&
        (@self-2.MOD == VOICE))
 @then {
        @if ((@self-1.INIT-@self-2.INIT <= 5) &&
                (@self-1.INIT-@self-2.INIT <=-5))
        @then { @break();}
        @else { @up.MOD =a [VOICE,CLICK];

        @if((@self-1.INIT <= @self-2.INIT))
        @then { @up.INIT =a @self- 1.INIT;}
        @else { @up.INIT =a @self-2.INIT;}

        @if((@self-1.END >= @self-2.END))
        @then { @up.END =a @self-1.END;}
        @else { @up.END =a @self-2.END;}
        }
            }
        @else {break();}
}
```

Figure 4: Unification rules

The unification rules presented above define under what conditions unification will happen. It is important to note that the unification pre-requisites can be defined on a per case basis. However, macros are helpful when this level of granularity is not really necessary (Figure 5).

```
@assign_modality(@self-1,@self-2,@self-n)
   check if the modality of all the
   constituents is the same, otherwise,
   assign MODALITY:[MIXED] to the
   mother node.

@assign_time_init(@self-1,@self-2,@self-n)
   Get the lowest time init and assign it
   to the mother node.

@assign_time_end(@self-1,@self-2,@self-n)
   Get the highest time end and assign it
   to the mother node
```

Figure 5: Unification rule macros

It is important to illustrate the grammar-based strategy with an example:

Given the following set of user inputs:

- "Turn this on" (speech); INIT: 00:01; END: 00:03
- "Lamp_1" (click); INIT: 00:02; END: 00:02

The parser result would be as illustrated in figure 6.

```
DMOVE:  specifyCommand
TYPE:   CommandOn
ARGS:   specifyParameter
MOD:    speech
INIT:   00:01
END:    00:03
specifyParameter:
        DMOVE:  specifyParameter
        TYPE:   Device
        CONT:   Lamp_1
        MOD:    click
        INIT:   00:02
        END:    00:02
```

Figure 6: Parser result example for strategy 1

The combination of inputs with their respective modalities is a valid construction according to the grammar constraints.

2.2 Strategy 2: Dialogue-driven strategy

Unlike in the former strategy, in this case the fusion process takes places at dialogue level.

Although the grammar contemplates entries for different modalities, it does not include mixed-modality entries such as the example presented in figure 4. It only considers single-modality entries. In consequence, multimodal user inputs such as "Turn this on" followed by a click on the screen would be parsed as independent unrelated events, although the same time and modality information would be included (MOD, INIT and END). Two different DMOVES would be generated.

All inputs are sent to the multimodal input pool (where all user inputs are stored), where the dialogue manager will retrieve them at due time.

The time stamp added in the previous step is essential in order to determine a possible relationship between the DMOVEs stored in the multimodal input pool. When the DMOVE time stamps indicate a certain level of proximity in time, they are considered as simultaneous or pseudo-simultaneous, which implies further analysis to confirm or discard whether they are complementary or not.

The timeframe within which different inputs can be considered as potentially related is determined empirically. This is somewhat similar to the Microtemporal and Macrotemporal fusions in Nigay and Coutaz´s Melting pots [13].

This strategy is illustrated in figure 7.

Once the timeframe of co-occurrence has been analysed, if the result is positive, additional information will be taken into account to determine whether the events are complementary or not.

- If one triggers a Dialogue Rule, and the other one is part of the expectations.
- If both are expectations of an already active Dialogue Rule.
- If there is no other parallel dialogue history whose active Dialogue Rules may conflict with the previously identified one.

If the events are deemed as complementary, they are merged into a single information state, so the resulting information state ends up being identical to the resulting information state in the approach previously described. The dialogue manager uses the parser unification module, therefore similar rules to the ones illustrating strategy 1 are used. The basic difference is that instead of operating with grammatical symbols, DTACs will be used.

If the previous timeframe analysis renders a negative result, that is, given the distance in time between them the events are unlikely to be related, different possibilities are available:

- One completes a previous task and the other initiates a new task.
- Two dialogue histories may be active and each of them completes different tasks.
- Two dialogue histories may be active and each of them may complete both tasks, which implies overt disambiguation.
- Two new unrelated tasks are initiated simultaneously.

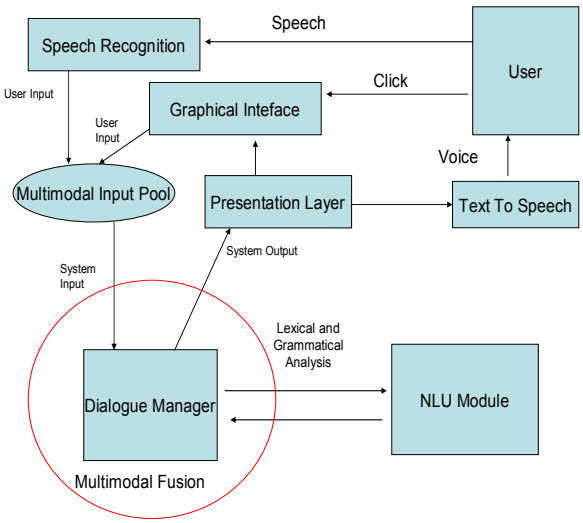

Figure 7: Dialogue-driven fusion

The high level algorithm for the dialogue-driven strategy is illustrated in figure 8.

The factors taken into account in order to determine whether the events are complementary are the following:

1. Dialogue Moves generated
2. Modality
3. Inter-Input timing
4. Dialogue Move order
5. Existing Dialogue Moves
6. Existing Dialogue Histories
7. Scenario and contextual factors

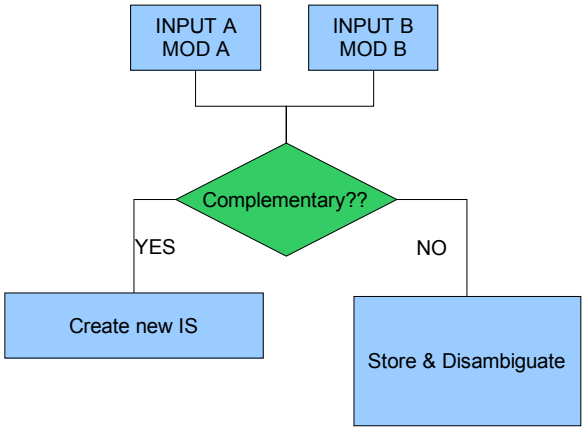

Figure 8: Dialogue-driven fusion algorithm

In the dialogue-based strategy and given the same set of user inputs as in the previous example illustrated in section 2.1:

- "Turn this on" (speech); INIT: 00:01; END: 00:03
- "Lamp_1" (click); INIT: 00:02; END: 00:02

The parser result would be as illustrated in figure 9.

DMOVE:	specifyCommand
TYPE:	CommandOn
ARGS:	specifyParameter
MOD:	speech
INIT:	00:01
END:	00:03

DMOVE:	specifyParameter
TYPE:	Device
MOD:	click
INIT:	00:02
END:	00:02

Figure 9: Parser result example for strategy 2

Here the grammar allows for uni-modal inputs in different modalities, but there is no valid construction that combines both inputs.

It would be the Dialogue Manager that would determine whether the inputs are or not to be combined into a valid construction.

3. THE NEW HYBRID STRATEGY

In the previous section, the preceding strategies implemented in the MIMUS system have been described. The new hybrid strategy combines the advantages of both of them while avoiding introducing a higher degree of complexity.

As explained before, strategy 1 is grammar driven. The grammar contains multimodal entries and temporal and modal constraints. There is no additional decision process involved.

Strategy 2 however is dialogue driven and implies uni-modal grammar entries, temporal and modal constraints and a dialogue level decision process based on the seven factors listed above. It seems sensible to think that adding information and therefore intelligence to the fusion strategy will render better results.

Comparing the pros and cons of both strategies, it could be concluded that:

1. Strategy 1 is more coherent in terms of the definition of a communicative act as a single event that may be more or less complex (single vs. multiple modalities).
2. Nonetheless, strategy 1 implies a significant computational load and is more dependent on time measures, which is not the case in strategy 2. This dependency and precision need for strategy one implies as well larger amounts of real user data to tune the multimodal grammar.
3. When dealing with additional or alternative modalities, the inter-modality disambiguation will no longer be between pairs (one or the other), but would imply the generation of full disambiguation lattices. In this case strategy 2 would reach a significant degree of complexity whereas strategy 1 could handle it more easily. Then again, there would be a significant computational overload with strategy 1.
4. Strategy 2 can handle independent simultaneous tasks in different modalities (multimodal multitasking), which would not be possible with strategy 1. Nonetheless, strategy 2 presents a potential theoretical problem that arises from the assumption that every uni-modal input can generate a dialogue move. No examples of this case have been found, but the opposite has not been proven either.

Multimodal Multitasking refers to the possibility of accomplishing independent unrelated tasks simultaneously, sparing continuous system disambiguation. The research indicates that humans are often able and even prefer to accomplish several tasks at once, as long as they are familiar with the tools and/or environment and none of the tasks imply too heavy a cognitive load.

The goal of this new strategy is taking advantage of the positive side of each of the previous strategies: including multimodal grammar entries and temporal and modal constraints as in strategy 1, but delegating the decision to the dialogue manager, in order to take into account the additional information involved in the strategy 2 decision process. Fortunately, this can be done in the MIMUS system without much effort at all, once strategy 2 is implemented.

The parser within MIMUS renders all possible parsing chunks. In previous versions of the system (speech-only versions), the most likely chunk would be selected by an internally developed criterion based on empirical data. However, when this selection strategy is deactivated, all possible parsing results are outputted. This basically means that given a grammar where simultaneous or pseudo-simultaneous multimodal entries may or not be related, the parser will output all possibilities: two unrelated events in different modalities, or one complex multimodal event. It will then be up to the dialogue manager to select which option is more likely to be appropriate, instead of having to build the most appropriate construction by a post-parsing unification process.

The graph in figure 10 illustrates the process: two pseudo-simultaneous inputs in different modalities are analysed by the parser. The latter generates two valid possible results. It is then the dialogue manager that, taking into account the relevant additional information, determines that the events are related and correspond to a single communication act.

Combining the two solutions allows the system to optimise the process; in the cases where the parsing that includes the temporal and modal constraints is not ambiguous, there will be only one valid parsing result, which spares the dialogue manager from performing any additional processing regarding this issue. When ambiguity exists, it will be the more knowledgeable agent which selects the most appropriate result.

Given the example presented in previous sections ("Turn this on" + click-Lamp_1), since there is no reasonable ambiguity regarding whether the inputs are or not related, the parser would produce the same result as in figure 6 (strategy 1). Given a more

ambiguous example, for instance in cases where the INIT time-frames of the inputs are borderline, the grammar would allow for two possible outputs (figures 6 and 9), and it would then be the Dialogue Manager that would select the most appropriate option.

4. ADVANTAGES AND DISADVANTAGES

In previous sections different advantages and disadvantages about adopting strategy 1 or strategy 2 have been discussed.

From that previous analysis it is quite clear that:

1. The hybrid strategy is coherent with the perception of a multimodal event as a single communicative act (as it was strategy 1).
2. There is a significant computational load but the amount of data available is not as relevant a problem.
3. The complexity regarding additional modalities can be handled more easily than in strategy 2.
4. Multimodal multitasking is also possible since the complex decisions would be made at dialogue level, rather than at grammar level.
5. The potential theoretical problem which arises in strategy 2 does no longer exist, since the grammar can handle it.

In a nutshell, it seems that this strategy combines all the advantages of the former strategies and also overcomes most of their problems, with the exception of the significant computational load.

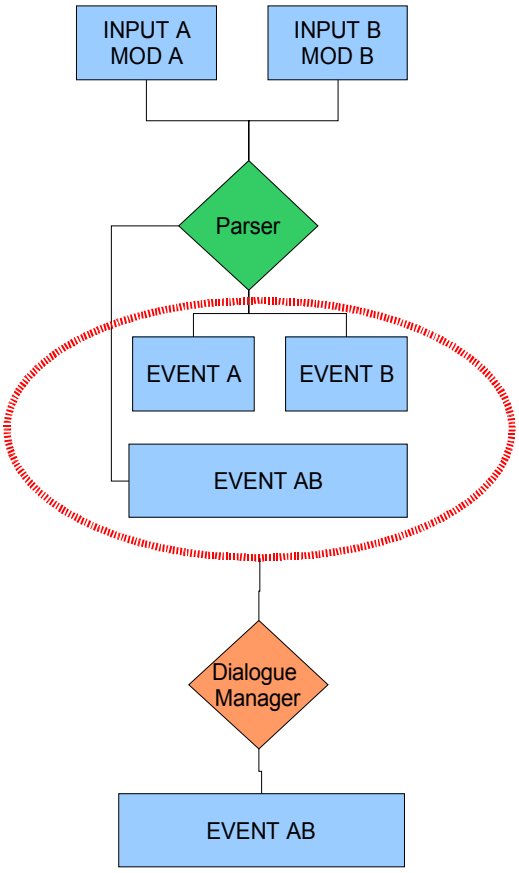

Figure 10: Hybrid fusion

5. CONCLUSIONS AND FUTURE WORK

The new hybrid strategy presented here represents a significant development with respect to the previous strategies. It allows for maximum flexibility while respecting additional theoretical considerations.

One of the advantages of this new strategy is that it makes full use of the former implementations for strategies 1 and 2, while introducing a new perspective.

This new strategy integrates the elegance of the grammar-based strategies with the functionality and intelligence of the dialogue-based ones.

Future work in this area will help determine and optimize how much of the processing should be performed at grammar level and when the dialogue manager needs to be involved in the decision process.

6. ACKNOWLEDGEMENTS

This work was carried out under the "TALK" research project, funded by EU´s FP6 [ref. 507802] and the "Multilingual Management of Spoken Dialogues" project, funded by the Spanish Ministry of Education under grant TIC2002-00526.

7. REFERENCES

[1] Gabriel Amores & José Francisco Quesada (1997) Episteme Procesamiento del Lenguaje Natural 21. pp 1-16.

[2] Gabriel Amores, José Francisco Quesada (2000) Diseño e Implementación de Sistemas de Traducción Automática. Sevilla: Secretariado de Publicaciones de la Universidad de Sevilla.

[3] Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman, Ira A. Smith (1997): Unification-based Multimodal Integration. ACL 1997: pp 281-288

[4] Michael Johnston (1998): Unification-based Multimodal Parsing. COLING-ACL 1998, 624-630.

[5] Michael Johnston, Srinivas Bangalore (2000). Finite State Multimodal Parsing and Understanding. Proceedings of the 18th conference on Computational linguistics - Volume 1. pp 369-375.

[6] Michael Johnston, Srinivas Bangalore (2001). Finite-state Methods for Multimodal Parsing and Integration, Finite State Methods in Natural Language Processing, August 2001.

[7] Pilar Manchón, Guillermo Pérez, Gabriel Amores (2005). WOZ experiments in Multimodal Dialogue Systems. Proceedings of Dialor'05, Nancy, France, pp 131-135.

[8] Sharon Oviatt (1999). Ten myths of multimodal interaction, Communications of the ACM, Vol. 42, No. 11, pp. 74-81.

[9] Sharon Oviatt (2003). Multimodal interfaces. In The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. J. Jacko & A. Sears, Eds. Lawrence Erlbaum Assoc., Mahwah, NJ, chap.14, pp 286-304

[10] José Francisco Quesada, Doroteo Torre, Gabriel Amores (2000). Design of a Natural Command Language Dialogue System. Deliverable 3.2, Siridus Project

[11] Guillermo Pérez, Gabriel Amores & Pilar Manchón (2005) Proceedings of ICMI'05 Workshop on Multimodal

Interaction for the Visualisation and Exploration of Scientific Data. Trento, Italy.

[12] Dafydd Gibbon, Inge Mertins & Roger Moore Eds. (2000). Handbook of Multimodal and Spoken Dialogue Systems. Kluwer Academic Publishers, N orwell, MA.

[13] Laurence Nigay & Joëlle Coutaz (1995). A generic platform for addressing the multimodal challenge. International Conference on Human-Computer Interaction, pp 98-105, Denver (CO). ACM

[14] Minh Tue Vo (1998). A Framework and Toolkit for the Construction of Multimodal Learning Interfaces. PhD. Thesis, Carnegie Mellon University, Pittsburgh, USA.

[15] Minh Tue Vo & Alex Waibel (1997). Modelling and interpreting multimodal inputs. A semantic integration approach. Technical Report CMU-CS-97-192, Carnegie Mellon University, Pittsburgh, USA

[16] Minh Tue Vo & Wood, C. (1996). Building an application framework for speech and pen input integration in multimodal learning interfaces. International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA. IEEE.