IntelliMedia TourGuide: understanding reference at the language/vision interface

Paul Mc Kevitt[*]
Center for PersonKommunikation (CPK)
Fredrik Bajers Vej 7-A2
Institute of Electronic Systems (IES)
Aalborg University
DK-9220, Aalborg
DENMARK, EU.
pmck@cpk.auc.dk

Abstract

One of the most important issues in developing Intelligent MultiMedia (IntelliMedia) or the real-time computer processing, understanding and integration of perceptual input from speech, textual and visual sources is that of the semantics of communication between the various modules. We provide here such a semantics in terms of frames and give a worked example of how it can be used to process a sample query where the application is an IntelliMedia TourGuide giving information about building plans on an IntelliMedia Workbench. The worked example focusses on of exophoric reference ("Who's office is this?" + pointing) where reference is one of the most prominent phenomena in IntelliMedia. This is one application of our general CHAMELEON platform for performing IntelliMedia through integration of speech and image processing.

1 Introduction

The area of MultiMedia is growing rapidly internationally and it is clear that it has various meanings from various points of view. MultiMedia can be separated into at least two areas: (1) (traditional) MultiMedia and (2) Intelligent MultiMedia (*IntelliMedia*). The former area is the one that people think of as being MultiMedia, encompassing the display of text, voice, sound and video/graphics with possibly touch and virtual reality linked in. However, the computer has little or no understanding of the meaning of what it is presenting.

IntelliMedia, which involves the computer processing and understanding of perceptual input from speech, text and visual images and reacting to it is much more complex and includes research from Engineering, Computer Science and Cognitive Science (see Mc Kevitt 1995/96, 1997). This is the newest area of MultiMedia research which has seen an upsurge over the last three years and one where most universities internationally do not have expertise. Aalborg University, Denmark has already initiated IntelliMedia 2000+ (see Mc Kevitt and Dalsgaard 1996) which involves research with the production of a number of real-time demonstrators showing examples of IntelliMedia applications, establishing a new Master's degree in IntelliMedia and a nation-wide MultiMedia Network which is concerned with technology transfer to industry. More details can be found on WWW: http://www.cpk.auc.dk/CPK/MMUI/.

Four research groups exist within the Faculty of Science and Technology in the Institute of Electronic Systems, each of them covering expertise which together is necessary for building up IntelliMedia systems. The four research groups are Computer Science (CS), Medical Informatics

[*] Paul Mc Kevitt is also a British Engineering and Physical Sciences Research Council (EPSRC) Advanced Fellow at the University of Sheffield, England for five years under grant B/94/AF/1833 for the Integration of Natural Language, Speech and Vision Processing.

(MI), Laboratory of Image Analysis (LIA) and Center for PersonKommunikation (CPK) each of them contributing knowledge to platforms for specification, learning, integration and interactive applications, expert systems and decision taking, image/vision processing, and spoken language processing/sound localisation.

2 CHAMELEON

The results from the research groups have hitherto to a large extent been developed within the groups themselves. However, there is no doubt that the establishment of future and widespread use of IntelliMedia systems requires collaborations among the groups in order to integrate their results in new user-friendly applications. Some of the results may be integrated within a short term perspective as some of the technologically based modules are already here, others on the longer term as new results become available.

In general, applications within IntelliMedia may conceptually be divided into a number of broad categories such as intelligent assistant applications, teaching, information browsers, database-access, command control and surveillance, and transaction services (banking). Examples of applications which may result within a short term perspective are enhanced reality (e.g. library guide), newspaper reader for blind/near-blind people, intelligent manuals, dedicated personal communicator (DPC), diagnosis systems (e.g. medical data processing) and mixed reality (e.g. surgery support systems).

The four groups of IntelliMedia 2000+ are developing an IntelliMedia computing platform called CHAMELEON which will be general enough to be used for a number of different applications.

2.1 System architecture

The general architecture of CHAMELEON consists of the following modules:

Speech recognizer for recognizing input speech and mapping it into written text. The recognizer has the capability to determine intentions (query?, instruction!, declarative) of the user from pitch contour.

Speech synthesizer for providing speech output to give acknowledgements and verbal responses to queries and also for asking questions.

Gesture recognizer for determining the coordinates (X,Y) of a pointing gesture.

Image recognizer for conducting 3D image processing of objects on an IntelliMedia Workbench.

Dialogue/NLP module for parsing the output of the speech recognizer and determining a meaning representation. We use frames as a syntax for meaning representation. The dialogue model has a stored representation of the anticipated dialogue interaction with the user using a Dialogue Description Language (DDL) (see Baekgaard 1996, Dalsgaard and Baekgaard 1994, Larsen 1997).

Domain model for modelling the domain of application (e.g. Architectural Layout/Plans of building(s)/tenants)

Topsy, a synchronizer which detects and learns co-occurrences in the input and synchronizes output. Topsy uses learned co-occurrences to conduct reasoning and decision-taking over the frame semantics.

Microphone array for determining and localising on the coordinates (X,Y,Z) of a sound source.

Figure 1: IntelliMedia Workbench

An IntelliMedia Workbench has 2D architectural plans (or 3D model) on it. Initially we are working with 2D plans of a building at Aalborg University. Cameras are used to interpret the plans on the workbench and the user's pointing. The system points using a laser pointing device and in more advanced scenarios will even give route descriptions for some destination. Microphone arrays perform sound localisation which aids speech processing. A computer monitor is linked in so that internet/WWW data about the building and the domain model, which has a database record of offices and their functionality/tenants, can be accessed.

We have implemented the image processing software, laser pointing driver, and domain model for this application and are currently tuning topsy and the spoken dialogue system for the application. The microphone arrays will be integrated in due course. The sample dialogue to be processed by the first prototype includes two of the most common phenomena which occur in IntelliMedia: (1) reference ("Who's office is this?") and (2) spatial relations ("Who's in the office beside him?"). Examples of other interesting problems to be solved as part of this application are resolution of ambiguity where a user says "Who's office is this?" but where the pointing-gesture is ambiguous since the person points sloppily between two rooms rather than into one. The system can then become active in the dialogue and ask the user for a clarification as to which office he/she means. Other interesting interactions are "Point to Paul's office" where there are two Pauls and again the system would ask a question.

2.3 Sample dialogue

We present here a sample dialogue which the first prototype of the system will process. The example is pregnant with phenomena such as instructions, questions, statements/declaratives, ambiguity (CHAMELEON becomes active), ellipsis, reference (endophoric/ pronominal, exophoric/ demonstrative), spatial relationships (beside), and learning.

```
USER:      Point to Hanne's office.
CHAMELEON: [points]
           This is Hanne's office.
USER:      Where is Jorgen's office?
CHAMELEON: [points]
           This is Jorgen's office.
USER:      Show me Paul's office.
CHAMELEON: [points(twice)]
           This is Paul Dalsgaard's office and
           this is Paul Mc Kevitt's office.
or/
CHAMELEON: There are two Paul's.
           Do you mean Paul Dalsgaard or Paul Mc Kevitt?
USER:      Paul Dalsgaard!
CHAMELEON: [points]
           This is Paul Dalsgaard's office.

USER:      Who's in the office beside him?
CHAMELEON: [points]
           Boerge, Jorgen and Hanne's offices
           are beside Paul Dalsgaard's office.
USER:      [points]
           Who's office is this?
CHAMELEON: [points]
           This is Ipke's office.
USER:      No, that's Tom's office!
CHAMELEON: [points]
           I've updated Ipke's office to Tom's office.
/or
CHAMELEON: [points]
           Ipke and Tom are in the same office!
```

3 Frame semantics

The frame language consists of three types of frames: (1) *input* (F-in), (2) *output* (F-out) and (3) *integration* (F-int). Input frames are those which come from modules conducting perceptual input, output frames are those produced by modules generating system output and integration frames are those which are processed by modules accessing the blackboard.

3.1 Input frames

An input frame (F-in) takes the general form:

```
[MODULE
 INPUT: input
 INTENTION: intention-type
 TIME:      timestamp]
```

where MODULE is the module producing the frame, INPUT can be at least UTTERANCE or GESTURE, *input* is the utterance or gesture and intention-type includes different types of utterances and gestures. An utterance input frame is where intention-type can be at least (1) query?, (2) instruction! and (3) declarative. An example of an utterance input frame is:

```
[SPEECH
 UTTERANCE: (Point to Hanne's office)
 INTENTION: instruction!
 TIME:      timestamp]
```

A gesture input frame is where intention-type can be at least (1) pointing, (2) signal-1, (3) signal-2 where the meaning of signals are common knowledge between the user and system. An example of a gesture input frame is:

```
[GESTURE
 GESTURE:  coordinates (3, 2)
 INTENTION: pointing
 TIME: timestamp]
```

3.2 Output frames

An output frame (F-out) takes the general form:

```
[MODULE
 INTENTION: intention-type
 OUTPUT: output
 TIME: timestamp]
```

where INTENTION is at least UTTERANCE or GESTURE, intention-type is the different types of utterance or gesture and output is the utterance or gesture. An utterance output frame is where intention-type is (1) query? (2) instruction!, and (3) declarative. An example utterance output frame is:

```
[SPEECH-SYNTHESIZER
 INTENTION: declarative
 UTTERANCE: (This is Hanne's office)
 TIME: timestamp]
```

A gesture output frame is where intention-type is (1) description (pointing), (2) description (signal-1), (3) description (signal-2) where signal meaning is common to user and system. An example utterance output frame is:

[LASER
INTENTION: description (pointing)
LOCATION: coordinates (5, 2)
TIME: timestamp]

3.3 Integration frames

An integration frame (F-int) takes the general form:

[MODULE
INTENTION: intention-type
LOCATION: location
OUTPUT: output
TIME: timestamp]

where intention-type can be (1) query?, (2) instruction!, and (3) declarative, location is a specification of a location and OUTPUT is an UTTERANCE or GESTURE. An example utterance integration frame is:

[DIALOGUE/NLP
INTENTION: description (pointing)
LOCATION: office (tenant Hanne) (coordinates (5, 2))
UTTERANCE: (This is Hanne's office)
TIME: timestamp]

Things become even more complex with the occurrence of references and spatial relationships:

[MODULE
INTENTION: intention-type
LOCATION: location
LOCATION: location
LOCATION: location
SPACE-RELATION: beside
REFERENT: person
LOCATION: location
TIME: timestamp]

An example of a of such an integration frame is:

[DOMAIN-MODEL
INTENTION: query? (who)
LOCATION: office (tenant Hanne) (coordinates (5, 2))
LOCATION: office (tenant Jorgen) (coordinates (4, 2))
LOCATION: office (tenant Boerge) (coordinates (3, 1))
SPACE-RELATION: beside
REFERENT: (person Paul-Dalsgaard)
LOCATION: office (tenant Paul-Dalsgaard) (coordinates (4, 1))
TIME: timestamp]

There are input and output gestures (G-in, G-out) and input and output utterances (U-in, U-out). Input modules are SPEECH-RECOGNIZER (U-in), IMAGE-GESTURE (G-in), and IMAGE-WORKBENCH (W-in). In our initial prototype the workbench images (2D building plans) are preprocessed by the system. Output modules are LASER (G-out) and SPEECH-SYNTHESIZER (U-out). Most modules give and take frames to/from the blackboard database and process them (F-int).

3.4 Exophoric reference

Here, we present all the steps and frames involved in processing a query "Who's office is this?" + pointing presented to CHAMELEON. Although we show the various modules acting in a given sequence here, since CHAMELEON is intended to work in a completely distributed manner, then module processing and frames may not necessarily run in this order. The frames given are placed on the blackboard as they are produced and processed.

USER(G-in,U-in): [points]
Who's office is this?

PROCESSING(1):
SPEECH-RECOGNIZER:
(1) wakes up when it detects registering of U-in
(2) maps U-in into F-in
(3) places and registers F-in on blackboard

FRAME(F-in)(1):
[SPEECH
UTTERANCE: (Who's office is this ?)
INTENTION: query?
TIME: timestamp]

PROCESSING(2):
DIALOGUE/NLP:
(1) wakes up when it detects registering of F-in
(2) maps F-in into F-int
(3) places and registers F-int on blackboard:

FRAME(F-int)(1):
[DIALOGUE/NLP
INTENTION: query? (who)
LOCATION: office (tenant person) (coordinates (X, Y))
REFERENT: this
TIME: timestamp]

PROCESSING(3):
DOMAIN-MODEL:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees its from DIALOGUE/NLP
(3) cannot update F-int as doesn't have a name or coordinates
(4) goes back to sleep

PROCESSING(4):
GESTURE:
(1) wakes up when it detects registering of G-in
(2) maps G-in into F-in
(3) places and registers F-in on blackboard

```
FRAME(F-in)(2):
[GESTURE
 GESTURE: coordinates (3, 2)
 INTENTION: pointing
 TIME: timestamp]

PROCESSING(5):
TOPSY:
(1) wakes up when it detects registering of F-in(1) and F-in(2)
(2) reads F-in(1) and F-in(2) and
    sees they are from SPEECH-RECOGNIZER and GESTURE
    that they have same/close timestamp,
    there is a query? (with referent) + pointing,
    in a rhythmic way (synchronized)
(3) dials and fires DIALOGUE/NLP to read GESTURE

PROCESSING(6):
DIALOGUE/NLP:
(1) woken up by TOPSY and reads F-in(2)
(2) sees F-in(2) is from GESTURE
(3) determines referent of "this" to be (coordinates)
(4) produces updated F-int (coordinates)
(5) places and registers updated F-int on blackboard:

FRAME(F-int)(2):
[DIALOGUE/NLP
 INTENTION: query? (who)
 LOCATION: office (tenant person) (coordinates (3, 2))
 REFERENT: this
 TIME: timestamp]

PROCESSING(7):
DOMAIN-MODEL:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees its from DIALOGUE/NLP
(3) produces updated F-int (tenant)
(4) places and registers updated F-int on blackboard:

FRAME(F-int)(3):
[DIALOGUE/NLP
 INTENTION: query? (who)
 LOCATION: office (tenant Ipke) (coordinates (3, 2))
 REFERENT: this
 TIME: timestamp]

PROCESSING(8):
DIALOGUE/NLP:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's from DOMAIN-MODEL
(3) produces updated F-int (intention + utterance)
(4) places and registers updated F-int on blackboard:

FRAME(F-int)(4):
[DIALOGUE/NLP
 INTENTION: declarative (who)
 LOCATION: office (tenant Ipke) (coordinates (3, 2))
 REFERENT: this
 UTTERANCE: (This is Ipke's office)
 TIME: timestamp]
```

```
PROCESSING(9):
LASER:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's from DOMAIN-MODEL
(3) produces F-out (pruning + registering)
(4) places and registers F-out on blackboard:

FRAME(F-out)(1):
[LASER
 INTENTION: description (pointing)
 LOCATION:  coordinates (3, 2)
 TIME: timestamp]

PROCESSING(10):
SPEECH-SYNTHESIZER:
(1) wakes up when it detects registering of F-int
(2) reads F-int and sees it's from DIALOGUE/NLP
(3) produces F-out (pruning + registering)
    places and registers F-out on blackboard:

FRAME(F-out)(2):
[SPEECH-SYNTHESIZER
 INTENTION: description
 UTTERANCE: (This is Ipke's office)
 TIME: timestamp]

PROCESSING(11):
TOPSY:
(1) wakes up when it detects registering of F-out(1) and F-out(2)
(2) reads F-out(1) and F-out(2) and sees they are from
    LASER and SPEECH-SYNTHESIZER
(3) dials and fires LASER and SPEECH-SYNTHESIZER
    in a rhythmic way (synchronized)
    (1) LASER reads G-out and fires G-out
    (2) SPEECH-SYNTHESIZER reads U-out and fires U-out

CHAMELEON(G-out): [points]
CHAMELEON(U-out): This is Ipke's office.
```

4 Conclusion

We have presented here a semantics for communication between various modules in our CHAMELEON platform being applied as an IntelliMedia TourGuide. The application is one where a system gives advice on building usage and integrates speech and image processing, synchronization, and laser pointing technology. Frames are created by various modules and placed on a blackboard where they can be read, written, and processed by other modules. We show that there are different types of frames in the semantics, the general form these take, and specific instances of them. A worked example with all associated frames and module interactions demonstrating the understanding of exophoric reference ("Who's office is this?" + pointing) is given. Future work will involve augmenting the frame semantics to handle more complex situations and testing various methods of communication and interaction between the modules.

Mobile computing aspects of the IntelliMedia TourGuide become evident if we consider the user walking in the building represented by the plans/model with a wearable computer (see Bruegge and Bennington 1996, Rudnicky et al. 1996, and Smailagic and Siewiorek 1996) and head-up display. This research could eventually be incorporated into more advanced scenarios involving multiple speakers in a VideoConferencing environment, say planning building and institution layout.

Intelligent MultiMedia will be important in the future of international computing and media development and IntelliMedia 2000+ at Aalborg University, Denmark brings together the necessary ingredients from research, teaching and links to industry to enable its successful implementation. Particularly, we have research groups in spoken dialogue processing, image processing, and radio communications which are the necessary features of this technology. Our IntelliMedia TourGuide application which focusses on giving help on building usage is an ideal one for testing integration of various modules.

Acknowledgements

I would like to acknowledge Tom Broendsted, Lars Bo Larsen, Mike Manthey, Thomas Moeslund, Kristian Olesen, and Helge Vad from IntelliMedia 2000+ who have provided various comments on the semantics and frames given here.

References

Baekgaard, Anders (1996) Dialogue Management in a Generic Dialogue System. In *Proceedings of the Eleventh Twente Workshop on Language Technology (TWLT), Dialogue Management in Natural Language Systems*, 123-132, Twente, The Netherlands.

Bruegge, Bernd and Ben Bennington (1996) Applications of wireless research to real industrial problems: applications of mobile computing and communication. In *IEEE Personal Communications*, 64-71, February.

Dalsgaard, Paul and A. Baekgaard (1994) Spoken Language Dialogue Systems. In *Prospects and Perspectives in Speech Technology: Proceedings in Artificial Intelligence*, Chr. Freksa (Ed.), 178-191, September. München, Germany: Infix.

Fink, G.A., N. Jungclaus, H. Ritter, and G. Sagerer (1995) A communication framework for heterogeneous distributed pattern analysis. In *Proc. International Conference on Algorithms and Architectures for Parallel Processing*, V. L. Narasimhan (Ed.), 881-890. Brisbane, Australia: IEEE.

Larsen, L.B. (1996) Voice controlled home banking - objectives and experiences of the Esprit OVID project. In *Proceedings of IVTTA-96*, New Jersey, USA, September (IEEE 96-TH-8178).

Leth-Espensen, P. and B. Lindberg (1996) Separation of Speech Signals Using Eigenfiltering in a Dual Beamforming System. In *Proc. IEEE Nordic Signal Processing Symposium (NORSIG)*, Espoo, Finland, September, 235-238.

Manthey, Mike (1997a) *Distributed computation, the twisted isomorphism, and auto-poiesis.* Technical Report R-97-5007, Department of Computer Science, Aalborg university, Denmark, June.

Manthey, Mike (1997b) *The phase web paradigm and anticipatory systems, two short papers.* Technical Report R-97-5006, Department of Computer Science, Aalborg university, Denmark, June.

Mc Kevitt, Paul (Ed.) (1995/1996) *Integration of Natural Language and Vision Processing (Vols. I-IV).* Dordrecht, The Netherlands: Kluwer-Academic Publishers.

Mc Kevitt, Paul (1997) SuperinformationhighwayS. In *"Sprog og Multimedier" (Speech and Multimedia)*, Tom Broendsted and Inger Lytje (Eds.), 166-183, April 1997. Aalborg, Denmark: Aalborg Universitetsforlag (Aalborg University Press).

Mc Kevitt, Paul and Paul Dalsgaard (1996) INTELLIMEDIA-2000+ at Aalborg University, Denmark. In *'Vistas in Astronomy', Special Issue on Strategies and Techniques of Information for Astronomy (STIA), Dec.*, Workshop Proceedings of European Science Foundation (ESF) Network on Converging Computing Methodologies in Astronomy (CCMA) Workshop on Strategies and Techniques of Information for Astronomy (STIA), European Science Foundation (ESF), Strasbourg, France, June, Andre Heck and Fionn Murtagh (Guest Eds.), Vol. 40, Part 3, 385-392. Exeter, England: Pergamon (Elsevier Publishers).

Rudnicky, Alexander I., Stephen D. Reed, Eric H. Thayer (1996) SpeechWear: a mobile speech system. In *Proceedings of International Symposium on Spoken Dialogue (ISSD 96), October 2-3, Wyndham Franklin Plaza Hotel, Philadelphia, USA*, Fujisaki, Hiroya (Ed.), 161-164. Tokyo, Japan: Acoustical Society of Japan (ASJ).

Smailagic, Asim and P. Siewiorek (1996) Matching interface design with user tasks: modalities of interaction with CMU wearable computers. In *IEEE Personal Communications*, 14-25, February.