Linking Speech and Gesture in Multimodal Instruction Systems

Joerg C. Wolf, *Student Member, IEEE,* and Guido Bugmann

***Abstract*—This paper analyses the timing of gesture and speech acts in a corpus (MIBL) of free-flowing human-to-human instruction dialogues. From there, an algorithm is proposed to establish the pairing between speech and gesture of the instructor. It is shown that correct pairing requires timing and semantic information. Further work will explore the use of this algorithm in unconstrained free flowing multimodal instruction dialogues between human and robot. A brief overview of a robotic system is given, that is able to learn a card game from a human teacher.**

I. Introduction

In the future, service robots should be programmable by anybody interacting with them. There are far too many possible tasks for the robot to be pre-programmed completely and users want to change the robots behaviour to their individual preference [1]. Users may not be experts in programming Therefore "programming" of service robots should be done in the language of humans. Humans teach by step-by-step task instructions. So the robot becomes a student and the human an instructor who teaches it. What do humans do when they teach? Humans teach by speaking and demonstrating. Therefore a service robot must be designed to understand natural language and demonstrations, in the domain where it is going to be taught. Looking at examples how humans teach is best done by collecting a corpus. This approach is called "corpus-based robotics" [1,2]. In corpus-based robotics, the interaction between human-teacher to human-student is analysed and the human-student is then replaced by a robot-student. A corpus provides the researcher with all the information required to design the robot to cope with unconstrained flow of speech and gesture (demonstrations).

So far our research group has investigated two corpora, one in the Instruction Based Learning project (IBL) and one in the Multimodal IBL project (MIBL). There are only few multimodal corpora aimed at human-robot interaction studies [3]. The IBL corpus contained route instructions mainly composed of sequences of actions. In the current project (MIBL), we focus on instructions also containing rule specifications. These are found frequently in game instructions. Using the same corpus-based method, we started with recording card game instruction dialogues between a teacher and a student. The teacher could demonstrate actions using a touch screen (figure 1) and all movements on the screen were recorded. The aim of this work is to develop a system capable of understanding such game instructions, build them into an internal representation and subsequently play the game with the user/teacher. In the chosen setup, the robot needs neither artificial vision nor effectors, as it can "see" cards moved on a touch screen and can play by moving cards on the touch screen. It allows concentrating the research on the learning process.

The plan of this paper follows the main development stages of the MIBL card-game learning system. Section II describes the corpus collection and its analysis. This includes an identification of functions referred to in utterances. It also includes an analysis of the type of gestures found in the corpus. Section III focuses on the process of determining which speech events correspond to which gesture events. In free flowing human-to-human instructions, these events start and end at different times and a combination of timing and semantic rules are required to achieve perfect pairing. Section IV proposes a system implementation based on the current findings. Section V concludes.

II. Corpus collection and analysis

A. Corpus Collection

The MIBL corpus was collected by recording dialogues between a person who already knows a card game (teacher) and another person who doesn't know the card game, using the setup shown in the figure below.

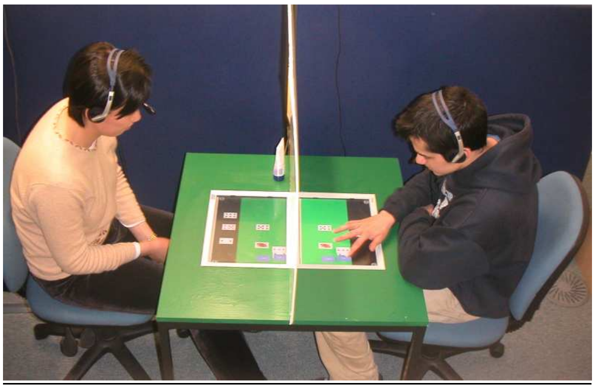

Fig. 1: Corpus collection setup. The instructor on the right moves a card on the touch screen. The learner sees a copy of the move on her screen.

This work was supported by a studentship of the University of Plymouth.

J. C. Wolf and Guido Bugmann are with the School of Computing Communications and Electronics, University of Plymouth, Drake Circus, Plymouth, PL48AA U.K., (e-mail: Joerg.wolf a_t plymouth.ac.uk, gbugmann a_t plymouth.ac.uk).

In card games, gestures can be pointing gestures, gestures moving cards from one place to another (e.g. stack to table, hand to table), re-arranging gestures (making a group of cards look tidier) and turning-over gestures. The separation panel between instructor and learner force the gesture component of the communication to take place via the touch screen and can easily be recorded. Each screen has a small "private" black band representing the hand of the player. The larger green area represents the table and is shared by the two players.

The dialogue was unconstrained; the participants were allowed to describe the card game at their own pace in their own words.

Transcription was done using dedicated multimodal transcription software called MuTra [4]. MuTra generates XML files with utterances and gesture content and timing. Table 1 shows information extracted from transcription files. U_i are the utterances and G_i are the gestures (from the touch screen). So far we have only analysed explanations of the dealing phase of the game.

TABLE I
EXAMPLE DIALOGUE (Session 03 from MIBL CORPUS)

No	Time in 10th sec.	utterance text or gesture semantics
U0	2396-2428	"I will just explain how you deal the cards"
U1	2431-2477	"er what you do first of all is.. er you deal three cards for yourself"
G1:	2416-2477	move(D/5,C/2,H/QQ, Stock , Temp2)
U2	2486-2513	"face down and I will take three"
G2	2482-2522	move(D/QQ,D/KK,D/AA, Stock , Temp1)
U3	2540-2563	"you take these into your black area"
G3	2531-2577	move(D/QQ,D/KK,D/AA,Temp1,Hand1)
U4	2564-2575	"so you can drag them down"
U5	2606-2627	"and then er turn them over"
G5	2599-2644	turn(D/5,C/2,H/QQ)
U6	2660-2675	"so you can see them and I can't see them"
U7	2680-2704	"and then what we do next is er"
U8	2708-2760	"put four cards face up on the table"
G8	2668-2753	move(H/2,D/3,C/KK,D/JJ,stock,table)
U9	2772-2778	"yep just four on the table"
U10	2801-2815	"yeh and three for each player"
U11	2821-2834	"so I will just turn those over"
G11	2808-2844	turn(H/2,D/3,C/KK,D/JJ)
U12	…	…

see text for explanations

B. Analysis of speech transcriptions

The corpus provides all information required to write a grammar and tune speech-recognition software. Currently a statistical language model has been trained with the corpus using NUANCE 8.5, a user independent speech recognition system.

Analysing the utterances of the transcriptions also reveals primitive procedures that the robot has to be able to carry out (the robot's "prior knowledge"). Such "language primitives" are specific to the level at which humans communicate with each other. They can constitute complex robot procedures which may require the use of micro planers (see section IV).

The following language primitive have been identified in the dealing phase:

```
start_of_sequence(name)
end_of_sequence()
deal(objects,amount,target)
move(objects,amount,source,target)
turn(objects)
owner(objects,player)
visible(objects,player)
count(objects,amount)
```

For instance U3 from the example above would need to be mapped onto a function call of the form:

```
move(objects=these?,num_of_cards=3,target=hand2)
```

Many of these primitives can only be completely specified and resolved using a combination of speech and gesture information. For instance the primitive function of U3 contains a "these" which can be resolved by the objects identified in the gestures.

C. Analysis of gesture transcriptions

Raw gesture data are a trail of X, Y coordinates of where the card is positioned on the touch screen. In case of a real robot, such tracking data could be the output the robot's vision system. The "analogue" trail of X, Y data of a cards position is then registered as a movement from a start area to a destination area ,e.g.

move(H/2,D/3,C/KK,D/JJ,stock,table).

Where "stock" is the source screen area and "table" is the target area of the cards. These areas, namely: stock, table, hand1, hand2, temp1 and temp2 divide the screen. The areas numbers and their boundaries are defined from observations of where the movements of the players usually end. These areas are currently simple squares and gesture labelling is straightforward [7]. If a vision system were to be used, the added uncertainty could call for the use of more complex probabilistic methods [5,6].

In general, gestures taken alone do not constitute a complete specification of the instruction. This is probably not true for the dealing phase where simply copying the gestures (without language) would be sufficient for the robot to deal correctly. However, in later phases of game instructions, such as in winning a trick, gestures only constitute examples, where objects of action are to be specified in general terms by the content of the spoken instructions. Therefore it is important to determine which speech act corresponds to which gesture.

Sections II B) and II C) have argued that language or gestures alone do not carry a complete message. Speech and gesture are acquired through different channels and must be re-associated to reconstruct or determine the complete meaning of a message. In the next section we exploit the idea of using temporal synchronization of speech and gesture.

III. Linking Speech and Gesture

A. *Pairing of speech and gesture*

A detailed analysis was carried out measuring the timing between gesture and speech of the teacher [7]. The MIBL corpus shows that verbal instructions are always in the same order as the corresponding gestures. Timing histograms (Figure 2,3) suggest the design of a pairing algorithm based on the maximum time-difference between start-of-speech/end-of-speech and start-of-gesture.

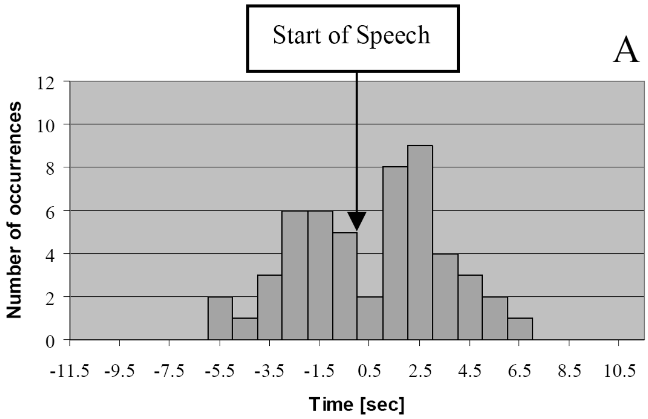

Fig 2: Histogram of the time intervals between start of speech and start of gesture.

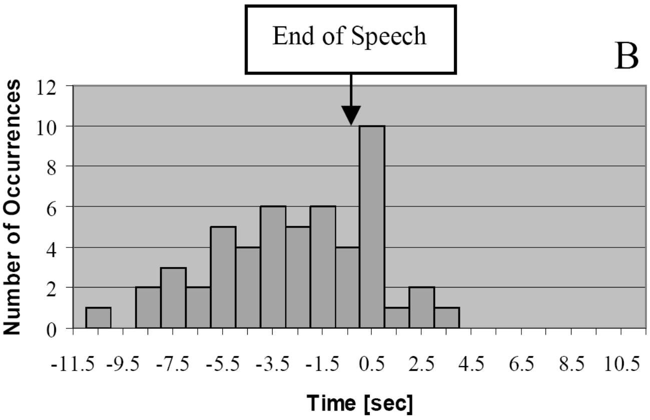

Fig 3: Histogram of time intervals between the end of speech and the start of gesture. Only gestures associated with speech events are plotted.

Figure 2 shows that gestures never start more than 5.5 sec before speech starts. Figure 3 shows that gestures never start later than 4 sec after corresponding speech ends. These observations suggest that a time window around the speech duration could be used to group speech and gesture (Figure 4). The time window borders are based on the maximum extend of the histogram.

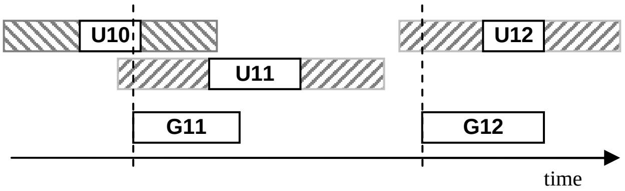

Fig. 4: This figure shows three incoming utterances and two incoming gestures. The grey areas show the maximum pairing range of the utterance. If a start-of-gesture falls within that range, it is a candidate for pairing with the utterance.

However, care must be taken with such grouping rules because time windows generally overlap.

Therefore, filters designed for grouping utterance-gesture groups can often only narrow down the candidates for grouping, but not solve the grouping problem completely.

In the example figure 4 U12 is clearly a candidate for G12, there is no confusion. U10 and U11 however could both belong to G11. In this ambiguous case, semantics must be used. In a first attempt the Gesture is assigned to the nearest neighbour utterance. In the MIBL corpus, this results in 83% correctly grouped cases.

The analysis of the 17% erroneous groupings revealed that they occur systematically with utterances which point to incompatible language primitives. For instance, in figure 4, when trying to pair G11, U10 is a reply to a question from the student and therefore not related to G11. U10 does not refer to the primitive `turn(objects)`.

Using timing alone pairs U10 with G11, while semantic filtering, as just described, eliminates U10 from the pairing candidates. Inspection of the corpus indicates that this algorithm can achieve perfect pairing.

B. *Semantic Integration of Speech and Gesture*

Once speech and gesture is paired, semantic integration must take place. Work is currently underway to develop first-order predicate logic statements that carry out the unification, although temporal logic could be considered as well. A Prolog rule that compares the parameters of the language primitive to the parameters of the gestures is at the core of the mechanism. The following 4 cases can occur as a result of pairing:

1) *Completion:*

A gesture and an utterance are individually incomplete, but complete each other.

$n_s = 1$, all variables are resolved

2) *Confirmation:*

A gesture and an utterance are individually complete. When combining they match.

$n_s = 1$, no variables exist

3) Contradiction:

The gesture supplies contradicting semantics when compared to the utterance.

$n_s = 0$

4) Under-specification:

The gesture and language combined are still semantically underspecified. Therefore several possible candidates are returned.

$n_s > 1$

Where n_s is the number of solutions of the Prolog rule.

Note that the completion-case can be used to do reference resolution. In the MIBL corpus, a specific set of cards is often co-referred with “them”,”these” or “those”. This part of the system is not discussed in this paper.

IV. Proposed System

Shown in figure 5 is an overview of the system implementing concepts described in previous sections. Interestingly, Perzanowski [8,9] produced a similar system proposal independently.

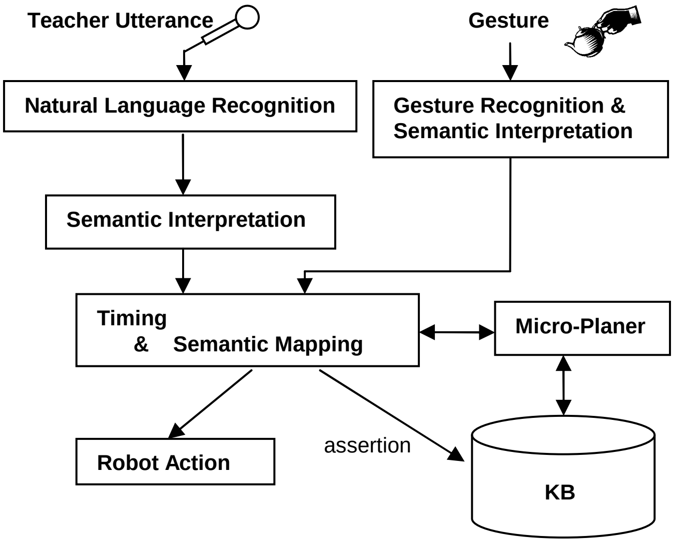
Teacher Utterance
Gesture
Natural Language Recognition
Gesture Recognition & Semantic Interpretation
Semantic Interpretation
Timing & Semantic Mapping
Micro-Planer
Robot Action
assertion
KB

Fig. 5. Multi-modal input processing in the MIBL robot.

We are currently using a statistical language model for the language recognition. A robust interpretation grammar extracts the semantics.

A multithreaded application (one for gesture and one for speech recognition) forwards information to the Timing & Semantic Mapping process. The semantics are unified and the micro-planer is consulted.

The micro-planer produces a detailed plan of what the robot should do. Sentences such as “take out all the eights, nines and tens from the deck”, are one primitive to a human, but require a variety of robot-actions to be carried out at the low level (i.e. moves and comparisons). The micro-planer is a problem solver which returns the steps required for the robot to achieve the language-primitive. If a single solution-path is returned, the problem is solved. The path is executed if needed, and stored if the robot is in its learning phase.

The resultant plan can be a robot action or a change in the knowledge base. Robot actions range from moving cards to replying to the user via a text-to-speech processor.

V. Conclusion

The work shows that it is possible to pair speech and gesture as occurring in unconstrained human-to-human instruction dialogue. The proposed pairing algorithm combines timing and semantic information. Further work will explore if this algorithm allows unconstrained free flowing multimodal instruction from human to robot.

References

[1] Bugmann G., Wolf J. C., Robinson P. The Impact of Spoken Interfaces on the Design of Service Robots. *Industrial Robot, 32:6,* 2005, pp 499-504,

[2] Bugmann,G., Klein, E., Lauria, S., Bos, J. and Kyriacou T., “Corpus-Based Robotics: A Route Instruction Example” *in Proceedings of IAS-8*, 10-13 March 2004, Amsterdam, pp. 96-103.

[3] Green A., Hüttenrauch, Topp E.A., and Eklundh K.S.,”Developing a Contextualized Multimodal Corpus for Human-Robot Interaction”, *in the Proc. 5th International Conference on Language Resources and Evaluation (LREC 2006),* Genoa, Italy 2006

[4] Wolf J.C., Bugmann G. Multimodal Corpus Collection for the Design of User-Programmable Robots. *Proc. Taros 2005*, London, pp. 251-255.MuTra link: (http://www.swrtec.de/swrtec/mibl/mutra/index.php)

[5] Naphade, M.R.,Kristjansson T.,Frey B. and Huang T.S., “Probabilistic Multimedia Objects Multijects: A novel Approach to Indexing and Retrieval in Multimedia Systems” *in Proc. IEEE International Conference on Image Processing*, Volume 3, pages 536-540, Oct 1998, USA, Chicago

[6] Roy Deb, "Semiotic schemas: A framework for grounding language in action and perception", *Artificial Intelligence, ELSEVIER,* Vol 167(1-2), pp. 170-205 , (Sept 2005)

[7] Wolf J.C., Bugmann G., “Integration of visual and spoken input in robot instructions” *in the Proceedings of the European Robotics Symposium,* Italy, Palermo, 2006

[8] Perzanowski, D., Shultz A.C. and Adams W., “Integrating Natural Language and Gesture in a Robotics Domain” *in Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISAS Joint Conference, Gaithersburg, MD: National Institute of Standards and Technology*, 247-252, September 1998.

[9] Perzanowski, D., Adams W., Shultz A.C. and Elaine Marsh “Towards Seamless Integration in a Multi-modal Interface.” *in Proceedings of the Workshop on Interactive Robotics and Entertainment*, Carnegie Mellon University: AAAI Press, 3-9, April 2000.