View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universidad Carlos III de Madrid e-Archivo

A multi-agent architecture to combine heterogeneous inputs in multimodal interaction systems

David Griol, Jesús García-Herrero, José Manuel Molina

Applied Artificial Intelligence Group (GIAA),
Computer Science Department,
Carlos III University of Madrid,
Spain
{david.griol,jesus.garcia,josemanuel.molina}@uc3m.es

Abstract. In this paper we present a multi-agent architecture for the integration of visual sensor networks and speech-based interfaces. The proposed architecture combines different techniques related to Artificial Intelligence, Natural Language Processing and User Modeling to provide an enhanced interaction with their users. Firstly, the architecture integrates a Cooperative Surveillance Multi-Agent System (CS-MAS), which includes several types of autonomous agents working in a coalition to track and make inferences on the positions of the targets. Secondly, the proposed architecture incorporates enhanced conversational agents to facilitate human-computer interaction by means of speech interaction. Thirdly, a statistical methodology allows to model the user conversational behavior, which is learned from an initial corpus and posteriorly improved with the knowledge acquired from the successive interactions. A technique is proposed to facilitate the multimodal fusion of these information sources and consider the result for the decision of the next system action.

Keywords: Software agents, Multimodal fusion, Visual sensor networks, Surveillance applications, Spoken interaction, Conversational Agents, User Modeling, Dialog Management.

1 Introduction

Research on multimodal interaction has grown considerably during the last decade as a consequence of the advent of innovative input interfaces, as well as the development of research fields such as speech interaction and natural language processing [1–3]. Speech and natural language technologies allow users to communicate in a flexible and efficient manner, making possible to access applications in which traditional input interfaces cannot be used (e.g. in-car applications, access for disabled persons, etc). Also speech-based interfaces work seamlessly with small devices and allow users to easily invoke local applications or access remote information. For this reason, multimodal conversational agents

are becoming a strong alternative to traditional graphical interfaces which might not be appropriate for all users and/or applications [4, 5].

In human conversation, speakers adapt their message and the way they convey it to their interlocutors and to the context in which the dialog takes place. The performance of a multimodal conversational agent also depends highly on its ability to adapt to the environmental conditions, such as other people speaking near the system or noise generated by other devices. This way, information related to the environment and users presence and location is essential to achieve this adaptation [6, 7].

Adaptation can play a much more relevant role in speech-based applications [8]. For example, users have diverse ways of communication. Novice users and experienced users may want the interface to behave completely differently, such as maintaining more guided versus more flexible dialogs. In these cases, processing context is not only useful to adapt the systems' behavior, but also to cope with the ambiguities derived from the use of natural language [9, 10]. For instance, contextual information can be used to resolve anaphoric references depending on the context of the dialog or the user location.

In order to acquire this information, visual sensor networks (VSN) present a number of benefits. Firstly, the use of these networks is growing rapidly as powerful public safety and security tools (for instance, in airports [11], sea environments [12], railways or undergrounds [13], and other critical environments). Secondly, the use of agents to develop VSNs provides important advantages, like "reactivity" (agents can perceive and respond to a changing environment), "social ability" (by means of which agents interact with other agents), and "proactivity" (through which agents behave in a goal-directed way). In addition, VSNs allow to know users current position (also considering users specific speeds, directions or even specific behaviors or physical features), but also to estimate users intentions and future actions (e.g., by detecting one or more users getting closer or moving away, looking at specific places, etc.).

In this work we present a novel architecture for the integration of visual sensor networks and speech-based interfaces. Our proposal is based on the multiagent framework for deliberative camera-agents forming visual sensor networks described in [14]. In this framework, each camera is represented and managed by an individual software agent, called a surveillance-sensor agent [15]. In addition, a visual fusion agent guarantees that objects of interest are successfully tracked across the whole area, assuring continuity and seamless transitions.

As far as we are concerned, there are not previous works proposing the integration of the information provided by visual sensor networks to improve human-machine interaction by means of conversational agents. To integrate speech interaction and visual sensor networks, we propose the incorporation of enhanced conversational agents [5, 4]. This kind of agents can be defined as computer programs that accept natural language as input and produces natural language as output, engaging in a conversation with the user. To successfully manage the interaction with users, conversational agents usually carry out five main tasks: automatic speech recognition (ASR), natural language understanding (NLU), di-

alog management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS). These tasks are usually implemented in different modules.

In this paper we propose to incorporate two additional modules to generate enhanced conversational agents acting in conjunction with visual sensor networks. The first module, that we have called User Modeling Module, generates a prediction of the next user action by taking into account the previous interactions with the conversational agent. User profiles are considered in this module for a better prediction. The second module, that we have called Multimodal Fusion Module, generates the next input for the dialog manager by considering the spoken interaction and the information provided by the VSN.

The main contributions of this work are: (i) To provide a detailed architecture that considers heterogeneous information generated by cooperative surveillance multi-agent systems (CS-MAS) and conversational agents; (ii) To describe a multimodal fusion methodology that takes these information sources into account to generate and encode the input of the dialog manager in the conversational agent; (iii) To propose a statistical user modeling methodology to predict the current task of the dialog and the next user action; (iv) To provide a statistical methodology for dialog management that considers the data generated by the multimodal fusion and user modeling methodologies for the selection of the next system action.

2 Proposed architecture

As described in the previous section, the proposed architecture to integrate visual sensor networks and speech interaction is based on [14]. As Figure 1 shows, different types of autonomous agents interact to fulfill this integration. The *Surveillance-Sensor Agent* tracks all the targets moving within its local field of view (FoV) and sends data to the *Visual-Fusion Agent*. It also sends information to the *Context Agent*. This agent is coordinated with other agents in order to improve surveillance quality. It can play different roles (individualized agent, object recognition agent, face recognition agent), each with different specific capabilities, but only one role at a time.

The *Visual-Fusion Agent* integrates the information sent from the associated surveillance-sensor agents. It analyzes the situation in order to manage the resources and coordinate the surveillance-sensor agents. This agent has the global view of the environment being monitored by all the surveillance-sensor agents. It is in charge of creating the dynamic coalitions of surveillance-sensor agents using contextual information and the prediction of certain situations requiring a cooperative fusion process. This agent also integrates the information from the different cameras and assures continuity and seamless transitions.

The *Recorder Agent* belongs to a specific camera with recording features only [14]. The *Planning Agent* has a general vision of the whole scene. It makes inferences on the targets and the situation. The *Context Agent* provides monitored context-dependent information. This agent indicates the semantic distance between different surveillance-sensor agents. The context agent stores information

about static objects that could provoke partial conclusions of the tracked targets but it also stores dynamic information about the scene [16]. The *Interface Agent* provides a graphical user interface that shows the evolution of the targets that are being tracked.

As described in [14], the coordination among Surveillance-Sensor Agents makes possible to jointly achieve a surveillance task. This way, the proposed CS-MAS architecture improves trajectory tracking by fusing data from several neighboring surveillance-sensor agents (camera agents in a visual sensor network), which are in a coalition.

In this paper, we propose the use of the information provided by the visual sensor network to facilitate the interaction with users by means of enhanced *Conversational Agents*. As Figure 1 shows, two main modules has been incorporated to enrich the general architecture of a conversational agent previously described. As stated in the previous section, the User Modeling module considers the previous dialog interactions and specific users features (defined by means of user profiles) to calculate a prediction of the next user action. The Multimodal Fusion module takes as input this prediction, the current user utterance, and the information provided by the surveillance sub-system. Using this information this module generates the input of the dialog manager, which selects the next system action. The following subsections describe the statistical methodologies proposed for the development of these modules.

2.1 The User Modeling module

Research in techniques for user modeling has a long history within the fields of language processing and speech technologies [17]. The main purpose of a user intention model in this field is to improve the usability of a conversational agent through the generation of corpora of interactions between the system and the user model [18].

Our proposed technique for user modeling simulates the user intention level by means of providing the next user dialog act in the same representation defined for the natural language understanding module. The lexical, syntactic and semantic information (e.g., words, part of speech tags, predicate-arguments structures, and name entities) associated to speaker u's ith clause is denoted as c_i^u.

Our model is based on the proposed in [19]. In this model, each user clause is modeled as a realization of a user action defined by a subtask to which the clause contributes, the dialog act of the clause, and the named entities of the clause. For speaker u, DA_i^u denotes the dialog label of the ith clause, and ST_i^u denotes the subtask label to which the ith clause contributes. The dialog act of the clause is determined from the information about the clause and the previous dialog context (i.e., k previous utterances) as shown in Equation 1.

$$DA_i^u = \underset{d^u \in \mathcal{D}}{\operatorname{argmax}} P(d^u | c_i^u, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \tag{1}$$

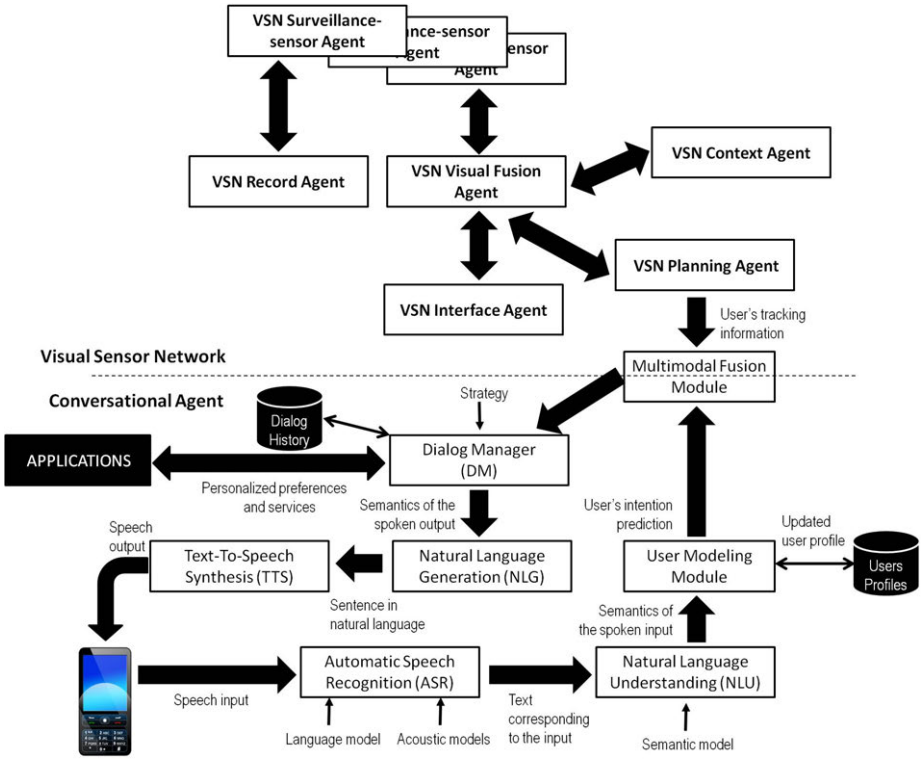

Fig. 1. Proposed multi-agent architecture to combine visual sensor networks and spoken interaction

In a second stage, the subtask of the clause is determined from the lexical information about the clause, the dialog act assigned to the clause according to Equation 1, and the dialog context, as shown in Equation 2.

$$ST_i^u = \underset{s^u \in \mathcal{S}}{\operatorname{argmax}} P(s^u | DA_i^u, c_i^u, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \tag{2}$$

In our proposal, we consider both static and dynamic features to estimate the conditional distributions shown in Equations 1 and 2. Dynamic features include the dialog act of each utterance and the task/subtask of each utterance. Static features include the words in each utterance (unigrams, bigrams, and trigrams), the part of speech tags in each utterance (unigrams, bigrams, and trigrams), supertags in each utterance (unigrams, bigrams, and trigrams), and a set of features that has been included in a user profile. This profile is comprised of user's:

- Id, which he can use to log in to the system;
- Gender;

- Experience, which can be either 0 for novel users (first time the user calls the system) or the number of times the user has interacted with the system;
- Skill level, estimated taking into account the level of expertise, the duration of their previous dialogs and the time that was necessary to access a specific content and the date of the last interaction with the system. A low, medium, high or expert level is assigned using these measures;
- Most frequent objective of the user;
- Reference to the location of the previous interactions and the corresponding objective and subjective parameters for the user.

2.2 Multimodal Fusion and Dialog Management

When dealing with multiple input sources, fusion of these input sources is a necessary feature of multimodal interaction creation tools. In fact, fusion of input data can be considered as one of the distinguishing features of multimodal interaction. Typical algorithms for decision-level fusion are frame-based fusion, unification-based fusion, and hybrid symbolic/statistical fusion [20]. Symbolic/statistical fusion [21] is an evolution of standard symbolic unification-based approaches, which adds statistical processing techniques to the fusion techniques previously described. These kinds of "hybrid" fusion techniques have been demonstrated to achieve robust and reliable results.

The methodology that we propose to develop the multimodal fusion module considers the set of information sources (spoken interaction, user modeling, and video tracking) by using different machine-learning techniques. The main objective of this module is to successfully associate the visual situation detected by the VSN and the user interaction with the conversational agent.

As described in [19], the conditional distributions shown in Equations 1 and 2 can be estimated by means of the general technique of choosing MaxEnt distribution that properly estimates the average of each feature in the training data [22]. This can be written as a Gibbs distribution parameterized with weights λ as Equation 3 shows, where V is the size of the label set, X denotes the distribution of dialog acts or subtasks (DA_i^u or ST_i^u) and Φ denotes the vector of described features for user modeling.

$$P(X = st_i|\phi) = \frac{e^{\lambda_{st_i}\cdot\phi}}{\sum_{st=1}^{V} e^{\lambda_{st_i}\cdot\phi}} \tag{3}$$

Each of the classes can be encoded encoded as a bit vector such that, in the vector for class, the ith bit is one and all other bits are zero. Then, one-versus-other binary classifiers are used as Equation 4 shows.

$$P(y|\phi) = 1 - P(\overline{y}|\phi) = \frac{e^{\lambda_y\cdot\phi}}{e^{\lambda_y\cdot\phi} + e^{\lambda_{\overline{y}}\cdot\phi}} = \frac{1}{1 + e^{-\lambda'_{\overline{y}}\cdot\phi}} \tag{4}$$

where $\lambda_{\overline{y}}$ is the parameter vector for the anti-label $\overline{y}$ and $\lambda'_{\overline{y}} = \lambda_y - \lambda_{\overline{y}}$.

Once the users action prediction has been calculated, a prediction of the system action can also been generated using a similar process. Each system

action is also defined in terms of the subtask to which it contributes and the dialog act to be performed. The determination of the system action, therefore, also proceeds in two stages: prediction of the system subtask (Equation 5), and prediction of the dialog act (Equation 6).

$$ST_i^a = \underset{s^a \in \mathcal{S}}{\operatorname{argmax}} P(s^u | ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \tag{5}$$

$$DA_i^a = \underset{d^a \in \mathcal{D}}{\operatorname{argmax}} P(d^a | ST_i^a, ST_{i-1}^{i-k}, DA_{i-1}^{i-k}, c_{i-1}^{i-k}) \tag{6}$$

The dialog manager decides the next action of the conversational agent. In addition, it updates the dialog history, provides a context for interpreting the sentences, and coordinates the other modules of the multimodal system. Thus, the dialog manager has to deal with different sources of information such as the semantic interpretations of the users utterances, database queries results, application domain knowledge, knowledge about the users and the dialog history.

A conventional dialog manager maintains a state n such as a form or frame and relies on two functions for control, G and F. For a given dialog state n, $G(n) = a$ decides which system action to output, and then after observation o has been received, $F(n, o) = n_0$ decides how to update the dialog state n to yield n_0. This process repeats until the dialog ends.

In a statistical approach, the conventional dialog manager is extended in three respects: firstly, its action selection function $G(n) = a$ is changed to output a set of one or more (M) allowable actions given a dialog state n, $G(n) = \{a_1, a_2, \cdots, a_M\}$. Next, its transition function $F(n, o) = n_0$ is extended to allow for different transitions depending on which of these actions was taken, $F(n, a, o) = n_0$.

In order to control the interactions with the user, our proposed statistical dialog management technique represents dialogs as a sequence of pairs (A_i, U_i), where A_i is the output of the dialog system (the system answer) at time i, and U_i is the semantic representation of the user turn (the result of the understanding process of the user input) at time i; both expressed in terms of dialog acts [23]. This way, each dialog is represented by:

$$(A_1, U_1), \cdots, (A_i, U_i), \cdots, (A_n, U_n)$$

where A_1 is the greeting turn of the system, and U_n is the last user turn. We refer to a pair (A_i, U_i) as S_i, the state of the dialog sequence at time i.

In this framework, we consider that, at time i, the objective of the dialog manager is to find the best system answer A_i. This selection is a local process for each time i and takes into account the previous history of the dialog, that is to say, the sequence of states of the dialog preceding time i:

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | S_1, \cdots, S_{i-1}) \tag{7}$$

where set $\mathcal{A}$ contains all the possible system answers.

Following Equation 7, the dialog manager selects the following system prompt by taking into account the sequence of previous pairs (A_i, U_i). The main problem to resolve this equation is regarding the number of possible sequences of states, which is usually very large. To solve the problem, we define a data structure in order to establish a partition in this space, i.e., in the history of the dialog preceding time i). This data structure, which we call *Interaction Register* (IR), contains the following information:

- sequence of user dialog acts provided by the user throughout the previous history of the dialog (i.e., the output of the NLU module);
- predicted user dialog act (generated by means of Equation 1);
- predicted user subtask (generated by means of Equation 2);
- predicted user position (provided by the agents in the virtual sensor network as explained in [14]);
- predicted system dialog act (generated by means of Equation 5);
- predicted system subtask (generated by means of Equation 6);

After applying these considerations and establishing the equivalence relation in the histories of dialogs, the selection of the best A_i is given by Equation 8.

$$\hat{A}_i = \underset{A_i \in \mathcal{A}}{\operatorname{argmax}} P(A_i | IR_{i-1}, S_{i-1}) \tag{8}$$

We propose the use of a classification process to decide the next system action following the previous equation. Specifically, we propose a multilayer perceptron (MLP) for the classification, where the input layer receives the current state of the dialog, which is represented by the term (IR_{i-1},A_i). The values of the output layer can be viewed as the a posteriori probability of selecting the different user intention given the current situation of the dialog. Figure 2 summarizes the operation of the proposed multimodal fusion and dialog management methodologies. As it can be observed, the user modeling module provides predictions of the next user dialog act and the current subtask of the dialog. Then, the system prediction module considers this information to generate the corresponding estimations for the system. The complete set of predicted values and the user position prediction provided by the planning agent are inputs of the fusion module to generate the interaction register. The dialog manager considers this register and the current user turn for the selection of the next system action.

3 Conclusions

In this paper we have described an architecture to develop multi-agent systems that considers the information generated by cooperative surveillance systems to provide user-adapted spoken interaction. To do this, we propose the integration of enhanced conversational agents in the CS-MAS architecture described in [14]. Two main modules have been incorporated in the classical architecture of a conversational agent to achieve the integration between visual sensor networks and conversational agents. These modules respectively allow to predict the next

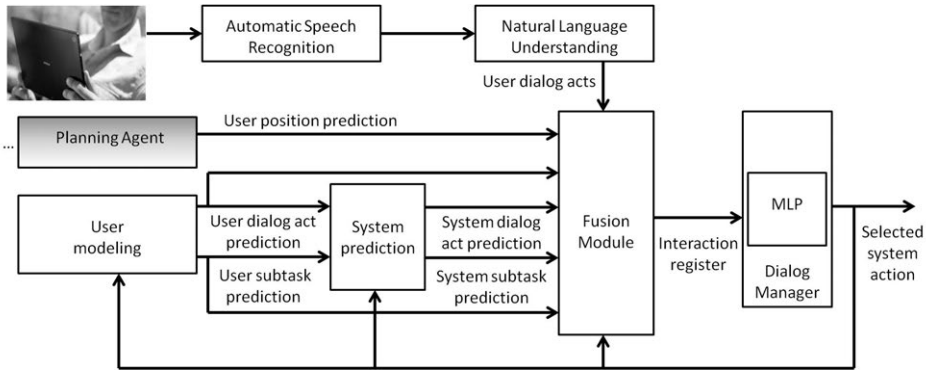

Fig. 2. Proposed multimodal fusion and dialog management methodologies for the development of conversational agents

user response for the conversational agent and carry out the fusion of visual and spoken information. The proposed multimodal fusion and dialog management techniques allow considering these heterogeneous information sources to select the next system action according to the current dialog and visual situations. Although the different methodologies proposed to develop the described modules have been evaluated in previous works [14, 24, 19], as a future work we propose the application of the described architecture to develop and evaluate a practical system in a real environment.

Acknowledgements

This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

References

1. D. Gibbon, I. Mertins, and R. K. Moore(Eds.), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation.* Kluwer Academic Publishers, 2000.
2. T. Heinroth and W. Minker, *Introducing Spoken Dialogue Systems into Intelligent Environments.* Springer, 2012.
3. O. Lemon and O. Pietquin(Eds.), *Data-Driven Methods for Adaptive Spoken Dialogue Systems. Computational Learning for Conversational Interfaces.* Springer, 2012.
4. R. López-Cózar and M. Araki, *Spoken, Multilingual and Multimodal Dialogue Systems.* John Wiley & Sons Publishers, 2005.
5. R. Pieraccini, *The Voice in the Machine: Building Computers that Understand Speech.* The MIT Press, 2012.

6. P. Osland, B. Viken, F. Solsvik, G. Nygreen, J. Wedvik, and S. Myklbust, "Enabling Context-Aware Applications," in *Proc. of ICIN'06*, 2006, pp. 1–6.
7. T. Lech and L. W. M. Wienhofen, "AmbieAgents: A Scalable Infrastructure for Mobile and Context-Aware Information Services," in *Proc. of AAMAS'05*, 2005, pp. 625–631.
8. P. Strauss and W. Minker, *Proactive Spoken Dialogue Interaction in Multi-Party Environments*. Springer, 2010.
9. S. Seneff, M. Adler, J. Glass, B. Sherry, T. Hazen, C. Wang, and T. Wu, "Exploiting Context Information in Spoken Dialogue Interaction with Mobile Devices," in *Proc. of IMUx'07*, 2007, pp. 1–11.
10. J. McCarthy, "Generality in Artificial Intelligence," *Communications of the ACM*, vol. 30, no. 12, pp. 1030–1035, 1987.
11. M. E. Weber and M. L. Stone, "Low altitude wind shear detection using airport surveillance radars," in *Record of IEEE Radar Conference*, 1994, pp. 52–57.
12. P. Avis, "Surveillance and canadian maritime domestic security," *Canadian Military Journal*, vol. 1, no. 4, pp. 9–15, 2003.
13. B. P. L. Lo, J. Sun, and S. A. Velastin, "Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems," *Acta Automatica Sinica*, vol. 29, no. 3, pp. 393–407, 2003.
14. F. Castanedo, J. García, M. A. Patricio, and J. M. Molina, "Data fusion to improve trajectory tracking in a Cooperative Surveillance Multi-Agent Architecture," *Information Fusion*, vol. 11, pp. 243–255, 2010.
15. M. Wooldridge and N. R. Jennings, "Surveillance and Canadian maritime domestic security," *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995.
16. A. M. Sánchez, M. Patricio, J. García, and J. M. Molina, "Video tracking improvement using context-based information," in *Proc. of 10th Int. Conference on Information Fusion*, 2007, pp. 1–7.
17. J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies," in *Knowledge Engineering Review*, vol. 21(2), 2006, pp. 97–126.
18. D. Griol, J. Carbó, and J. M. Molina, "Agent Simulation to Develop Interactive and User-Centered Conversational Agents," *Advances in Intelligent and Soft Computing*, vol. 91, pp. 69–76, 2011.
19. S. Bangalore, G. D. Fabbrizio, and A. Stent, "Learning the Structure of Task-Driven HumanHuman Dialogs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1249–1259, 2008.
20. D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J.Ladry, "Fusion engines for multimodal input: a survey," in *Proc. of ICMI-MLMI'09*, 2009, pp. 153–160.
21. L. Wu, S. L. Oviatt, and P. R. Cohen, "From members to teams to committee-a robust approach to gestural and multimodal recognition," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 972–982, 2002.
22. A. Berger, S. Pietra, and V. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist*, vol. 22, no. 1, pp. 39–71, 1996.
23. D. Griol, L. F. Hurtado, E. Segarra, and E. Sanchis, "A statistical Approach to Spoken Dialog Systems Design and Evaluation," *Speech Communication*, vol. 50, no. 8-9, pp. 666–682, 2008.
24. D. Griol, J. Molina, and Z. Callejas, "Bringing together commercial and academic perspectives for the development of intelligent AmI interfaces," *Journal of Ambient Intelligence and Smart Environments*, vol. 4, no. 3, pp. 83–207, 2012.