# "What's that called?": A Multimodal Fusion Approach for Cultural Heritage Virtual Experiences

Marco Grazioso[1], Maria Di Maro[2], and Francesco Cutugno[1]

[1] Department of Electrical Engineering and Information Technology, Università degli Studi di Napoli 'Federico II', Italy
[2] Department of Humanities, Università degli Studi di Napoli 'Federico II', Italy
{marco.grazioso,maria.dimaro2,cutugno}@unina.it

**Abstract.** In this paper, a multimodal dialogue system architecture is presented. The cultural heritage application of the software makes it important to use different channels of communication to enable museum visitors to naturally interact with it and still enjoying the artistic environment, whose exploration is supported by the system itself. A question answering system for the 3D reconstruction of the Absis and Presbytery of the San Lorenzo Charterhouse (Padula, Salerno) is considered as a case study to demonstrate the capabilities of the proposed system. The implemented multimodal fusion engine will be described along with the strategies adopted to involve multiple users in an immersive, interactive environment supporting queries and commands expressed through speech and mid-air gestures. The collected feedback shows that the system was well received by the users.

**Keywords:** multimodal dialogue · cultural heritage · fusion engine

## 1 Introduction

Human beings function multimodally. The use of gestures accompanied the history of men from the beginning. According to the Gestural Theory of language, human language developed from gestures used to communicate [1]. The alleged situational ambiguity-based incompleteness of both gestural and vocal channels can explain their common joint adoption. According to McNeill's terminology [16], the typical movements that can be recognised in gestures are of four types: i) deictics (or pointing gestures), which connect speech to concrete or abstract referents; ii) iconics, which depict concrete objects or events in discourse; iii) metaphorics, which put abstract ideas into more concrete forms; iv) beats, which have no semantic meaning, but are used to structure the discourse. Interestingly, the same cognitive aspects, lying behind the production of these visual signals, govern the production of the acoustic counterparts. In fact, words are used to refer to external referents, both concrete and abstract ones, to describe events and

objects, to explain abstract ideas, and, together with super-segmental items, to organise the discourse and modulate it. The two communicative codes depend, indeed, on similar neural systems.

Among the aforementioned gesture types, one is for us of particular interest: the pointing gesture. First of all, deictic gestures are the ones used to refer to something, which can be the topic of the communicative exchange representing the basis for the mutual knowledge being the skeleton of the interaction itself [3]. Therefore, they have a *referent identification* function and are a *grounding* tool. Secondly, these gestures are used as an embodiment tool for cognition [2]. According to the embodiment cognition approach, cognitive processes have their roots in motor behaviour. This means that cognition relies on a physical body acting on the environment in which it is immersed [22]. In this perspective, the design of dialogue systems whose interaction is based on the exchange of information about specific referents cannot overlook the use of such natural cognitive-motor means of communication. Multimodality becomes, therefore, the ultimate goal of this work, which aims to show how different modalities can be fused within a single system architecture.

The systems we are interested in are multimodal conversational agents, whose application is gaining more and more importance. Different studies show how technologies of this kind are being adopted for one of the most traditional among human experiences, which is the museums visit. In fact, the introduction of technological devices offering a virtual experience in the exploration of cultural contents can create more memorable exhibitions [12], and at the same time it can change the way museums are perceived and, consequently, the expectation of users [14]. For this reason, new studies are committed to the exploration of the ways museum visitors can be better engaged via these new devices [18], for what both visual stimuli and engaging communication strategies are concerned.

Concerning multimodality, some scholars [11, 25], in dealing with real environmental issues, contrive strategies which restrict the way users can freely interact. Therefore, we are interested in investigating and testing alternative approaches to model small groups interactions in real contexts, allowing users to communicate with both verbal and non-verbal actions. Specifically, the main purpose of the presented software architecture is to allow each member of a group of museum visitors to express their request to our system by interacting multimodally. With the exclusive use of natural human means of communication, i.e. voice, language and gestures, the virtual agent, projected on a curved screen, understands multimodal dialogue acts by users asking for information about artworks and architectural structures contained in a 3D scene.

In the next session, the architecture of the system will be explained starting from the language modelling for both understanding and generation purposes (Section 2.1). Afterwards, we will focus our attention on the pointing interpretation (Section 2.2) and on the active speaker detection (Section 2.3), before explaining the way the different signals were fused together to give a single interpretation of the user turn (Section 2.4). To conclude, the results of the system evaluation will be presented (Section 3).

## 2 System Architecture

In this section, the modules used to develop our multimodal system are described in detail. The entire setup of the interaction environment aims at creating an immersive experience for the users. Therefore, the interactive area consists of a 2,5m high and 4,4m long curved screen , used to project a realistic 3D environment representing the interactive scene. To track users movements and their speech signals in real time, a Microsoft Kinect 2 [29] sensor is placed on the floor, at the centre of the screen. Speech recognition is performed using the grammars discussed in Section 2.1 and the Microsoft Speech Platform. Acquired data concerned with user signals represent the input of the Game Engine (Unreal Engine 4[3]) used to model the 3D environment. An input recogniser communicate with the Multimodal Dialogue System in charge of understanding user intentions and providing the related responses. OpenDial is the framework adopted to implement this component [13], which communicate with the Knowledge Base designed through a graph database (Neo4j[4] [27]). Once a response is retrieved, the Game Engine synthesises the machine utterance by using Mivoq[5], a TTS engine. In the next subsections, the different knowledge bases (i.e. for conceptual representation of spaces, natural language understanding and generation) are structurally described. In the construction of meanings uttered during the interaction, whose context is shared by the interlocutors, other signals gain importance, specifically the pointing interpretation and the active speaker detection. In fact, we are finally going to focus on the fusion of the different signals in the modelling of our multimodal system.

### 2.1 Dialogue Modelling

In this section, the dialogue organisation, as far as the knowledge-base is concerned, is presented. In fact, the interactional engine of the system is supported with content and linguistic knowledge of the domain under consideration. The knowledge with which the system is provided comprises both a corpus-based grammar for extracting the topic of interest from a user question and a domain-dependent corpus to extract the proper answer as a feedback given to the human interlocutor by the system.

First of all, a collection of possible questions that a user could pose was carried out. The *ad hoc* structured survey enabled us to collect about 800 spoken questions divided into 10 different categories. Each category was then modelled in a Speech Recognition Grammar. The choice of this methodology depends on the fact that i) the speed of computation is higher in detecting the right class of the belonging question without the need of running complex algorithms on raw data; ii) the restricted domain of application can be better modelled with a rule-based approach [15, 23]; iii) the process of hand-crafting rules was simplified by

[3] www.unrealengine.com
[4] https://neo4j.com/
[5] www.mivoq.it

the use of a linguistic ontology by means of which semantic related words could be automatically included [5]. Specifically, the Speech Recognition Grammar Specification (SRGS) [9] W3C standard has been developed to allow Automatic Speech Recognition (ASR) engines to output the semantic interpretation of the matched pattern instead of the raw transcription. This is an important advantage for spoken dialogue systems as they can instruct ASR modules to *expect* specific word patterns and to present a structured interpretation of the obtained input to be provided to the dialogue manager. When the uttered word pattern is not included in one of the rules, the ASR finds the most similar pattern selected with a specific confidence[6]. This results in a reduced latency, as linguistic analysis chains working on raw strings are avoided. This methodology was already tested for a preliminary application for a different case study [5].

In more detail, a speech recognition grammar is a finite set of rules, where each rule, associated to a semantic label, generates a set of utterances. As far as the lexical enrichment of the grammar is concerned, given a collection of semantic relations (i.e. synonymy, hyperonymy and meronymy), the system is capable of expanding each rule in a grammar, in order to produce a new grammar, where a new set of generated utterances includes the previous ones plus the additional lexical and morpho-syntactic information [5].

For the lexical extension of the differentiated syntactic realisations of questions in our grammar, we made use of a graph database, named MultiWordNet-Extended [17, 5] which contains the Italian lexicon and the semantic relationships between words taken from MultiWordNet [20].

For the Natural Language Generation module we used another graph database containing textual nodes describing different points of the shown 3D scene. The nodes correspond to the concepts identified in the question classification process and are related to each other by inheritance (in-depth) relationships, when possible. Each node is, moreover, related to another node containing the textual answer to be given by the system.

In order to generate the response, the system needs to verify three conditions. Being $A$ a user interacting with the system:

- $A$ is the last speaker.
- $A$ asked a question recognised by the ASR.
- $A$ pointed at one relevant object in the scene.

When all the conditions are true with a considerable probability, the concept relative to the pointed object is used to query the graph database, retrieving the needed information in accordance with the semantic interpretation provided by the natural language understanding module. A more in depth explanation of multimodal fusion is provided in section 2.4.

[6] If the confidence is too low, the system can be modelled to ask for further clarifications.

**Fig. 1.** Reproducing users and their movements in the 3D environment.

### 2.2 Pointing Interpretation

The interpretation of the referential function of a message is strictly connected to the uttered entities which represent extralinguistic domain-related objects in the 3D scene, namely, in our scenario, the artworks or the structural items of the 3D model. These entities can be sometimes ambiguous, since users could have a minor expertise of the domain or since different items of the 3D model can be similarly uttered within the same domain. To overcome this interpretation problem, the non-verbal behaviour, such as pointing gestures, can help the disambiguation process. Therefore, in this subsection, we are going to present the way we modelled the pointing recognition function for enabling our system to interpret pointing gestures in the multimodal construction of intents by users. The pointing recognition (PR) task can be divided into two subtasks: a) user positioning in the virtual environment, b) gesture recognition.

In our system, the Kinect sensor returns the set of points representing the joints of human body. Using the points provided is possible to represent the user and his movements in the virtual environment by mapping kinect joints to a virtual avatar's joints. as shown in Figure 1. Note that virtual avatar in Figure 1 is displayed only for demonstration purposes while, in a real interaction, its visibility is turned off. The kintenct *base of spine* joint position has been used to estimate the user position in relation to the screen. Furthermore, the user height has been also taken into account in order to improve the pointing precision.

After the user representation is obtained, the next step is to recognise pointing activities. We realised this task through a geometrical approach based on Unreal Engine 4 geometrical functions and collision detection. Using the shoulder and hand positions we emit an invisible light ray that ends on an object surface generating a collision event. The event has been managed using the semantic maps mechanism combined with the Art and Architecture Thesaurus [8] in order to retrieve the concept label with relative relevance value, associated with the collided point. In this way, the collision event is enriched with conceptual meaning to enable the system to understand what the user is referring to. To avoid wrong pointing recognition event triggered by transit area we considered the arm movement speed to distinguish between transit area and fixation points.

### 2.3 Active Speaker Detection

Since museums are generally visited in groups, the system also needs to identify the speaker in order to address the answer to the right interlocutor. The active speaker detection module (ASD) has the responsibility to recognise the user that is actually speaking in a group. Specifically, the objective of this module is to distinguish between environment noises and speaking acts, and to compute the probability that a user is effectively talking. In this way the system is able to take into account the gestures produced by the user with the highest probability.

Several approaches were proposed in literature using visual features [24] and both video and audio ones [7]. In order to avoid problems deriving from data-driven approaches (data collection, computational complexity), we adopted a technique that computes speaking probability considering only the current loudest sound source location and users positions. More in depth, we define $\alpha$ as the angle between the Kinect forward vector and the vector that point to the sound source, and $\beta$ as the angle between the Kinect forward vector and the vector that point to the user. We also define $\Delta(\alpha, \beta)$ as the difference $\alpha - \beta$. Normalising $\Delta(\alpha, \beta)$ in the range $[0, 100]$ and dividing it by 100 we obtain a probability measure formalised as :

$$P(U_i = True \mid \theta_S, L_i) \mid U_i = \{True, False\}$$

Where $U_i$ represents the $i-th$ user, $\theta_S$ represents the sound source direction and $L_i$ represents the current position of the $i-th$ user.

### 2.4 Multimodal Fusion Engine

In this subsection, we present the approach used to tackle the fusion of all the previously explained modules in order to provide multimodality. This module receives asynchronous messages from input modules ASD, NLU and PR, handled by specific receivers. The messages are respectively:

- an ASD message: current speaker probability for each user.
- a NLU message: user sentence recognised by NLU module with a confidence value.
- a PR message: pointed object's semantic labels with relevance values for each user.

Messages received cause the update of the corresponding Bayesian network input variables (current speaker variable, verbal act variable, pointing variable)

The input fusion process is activated as soon as an user dialogue act is recognised. Input variables are synchronised and propagated through the probabilistic network to derive a common interpretation. To obtain multimodal unification, different formalisms and approaches are adopted in literature, i.e. statistical approaches [28], salience-based [6], or rule-based approaches [10], and biologically motivated approaches [26]. Here we propose a strategy that defines random variables validation rules based on the study discussed in [19]. Several modules collaborate in charge of performing this task. The *Multimodal Input Integrator* aims

at combining input variables coherently. In particular this module analyses verbal actions, pointed objects, and speakers in order to understand the current request. Since the variables evolve in real-time, the *Multimodal Time Manager* is used to check consistency and prune out-of-date variables. In more detail, starting from time-stamps related to the input variables, once a new speech signal is captured, the module compares its time intervals with those computed for each pointing variable, pruning off pointing gestures whose occurrence was concluded more than 4 seconds before the start of the current speech signal. As input variables come asynchronously, the *State Monitor* directs the entire operation by observing changes in dialogue state. Therefore, the unification methods are called by this component according to dialogue progresses.

Next operations are in charge of the *Dialogue Manager*. This has been implemented using the OpenDial framework [13]. Here, the *Dialogue State* manages variables and their connections encoded in a Bayesian network, while the *Dialogue System* provides the APIs to check and update the model. Once the system has derived the current request, this level provides services to select the most appropriate machine action to be performed.

## 3 System Evaluation

In order to evaluate various aspects of the proposed system an experimental setting was adopted. As this is an ongoing work, a humanoid virtual conversational agent is not yet present in the setup. Specifically, a fixed 3D scene was designed in Unreal Engine 4 showing the user a part of San Lorenzo Charterhouse, namely the Absis and the Presbitery of the Church. A simple question-answering based interaction was modelled making users capable of multimodally interact with the system in order to obtain information about a small set of objects in the 3D environment. The evaluation was conducted in our laboratory by analysing interactions between the designed system and two users simultaneously. In order to avoid wrong interpretations caused by background noises, the evaluation was conducted in a room where other persons besides the observer and the evaluated group were not admitted. Moreover, a threshold relative to the input sound signal intensity was established in order to cut off environment noises. The entire process was sliced up into the following steps:

1. **System presentation**: system functionality are presented to participants.
2. **Training session**: a video-clip presentation is shown and a first guided interaction is performed, in order to allow users to get acquainted with the multimodal interface.
3. **Task-oriented interaction**: users are asked to cooperate in completing a set of assigned tasks to test system functionality. Interactions are recorded through the Kinect and data-log in order to subsequently compute the success rate of each input recognition. Recorded data can be used to compose a training set to automatically tune the parameters of the probabilistic network.

| ID | TI | ASD | PR | NLU |
|---|---|---|---|---|
| 1 | 23 | 21 (91.3%) | 22 (95.6%) | 18 (78.3%) |
| 2 | 19 | 18 (94.7%) | 19 (100%) | 15 (78.9%) |
| 3 | 20 | 19 (95%) | 20 (100%) | 12 (60%) |
| 4 | 21 | 19 (90.4%) | 21 (100%) | 15 (71.4%) |
| 5 | 24 | 21 (87.5%) | 23 (95.8%) | 15 (62.5%) |
| 6 | 26 | 19 (73%) | 24 (92.3%) | 20 (76.9%) |
| **TOT** | **133** | **117 (88%)** | **129 (97%)** | **95 (71.4%)** |

**Table 1.** Success rate computed for Active Speaker Detection (ASD), Pointing Recognition (PR) and Natural Language Understanding (NLU) during group interactions

4. **Free session**: users interact with the system for an arbitrary time interval. This phase was useful to collect further data concerned with the way users would freely interact with such systems. Nevertheless, they were not yet evaluated. The time spent by users during this phase could be used as implicit estimation of their satisfaction.

In order to evaluate the system and to identify principal causes of irregular multimodal fusion, the success rate $SR_i$ was computed for each module $i$ as follows:

$$SR_i = \frac{SU_i}{TI} \cdot 100$$

Where $SU_i$ is the total number of successful interpretations reported by the module recogniser $i$ and $TI$ is the total number of users interactions.

A total of 6 groups, composed by 2 persons, was involved in the evaluation, recording the data shown in Table 1. Specifically, a total of 41'05" of task interactions and a total of 76' of training interactions were analysed. In particular, the 133 tasks interactions were used to estimate the success rate. Results (Table 1) show that, starting from a correct recognition of each input signal, the probabilistic network designed for the fusion engine is able to derive user requests during multimodal group interactions. Most relevant cases of erratic inferences are caused by a wrong input recognition. The Pointing Recognition module shows the highest result, with a success rate of 97%. The module that shows the worst behaviour is the Active Speaker Detection module. In particular, this result can be described as related to the users tendency to overlap themselves during collaborative interactions. Anyway, ASD performances may be improved by combining sound source angle and users locations in 3D environment with further features like users gaze direction and/or lips movements, similar to what is discussed in [21].

## 4 Conclusion

The paper aims at showing a multi-channel inputs' fusion approach applied in the development of a multimodal dialogue system for cultural heritage virtual

experiences. The promising results show that this approach is valuable for further investigations. For this reason, starting from the architecture proposed in this paper, our purpose is to further improve the performances extending the system functionality by enriching the content and linguistic knowledge and applying the pointing recognition to other objects in a more extended 3D scene. Furthermore, we aim at processing new input signals and modelling a multi-party dialogue to improve and promote collaborative interactions between users.

## Acknowledgment

This work is funded by the Italian ongoing PRIN project CHROME - *Cultural Heritage Resources Orienting Multimodal Experience* [4] #B52F15000450001.

## References

1. Armstrong, D.F.: The gestural theory of language origins. Sign Language Studies **8**(3), 289–314 (2008)
2. Ballard, D., Hayhoe, M., Pook, P., Rao, R.: Deictic codes for the embodiment of cognition (1996)
3. Clark, H.H., Marshall, C.R.: Definite knowledge and mutual knowledge (1981)
4. Cutugno, F., Dell'Orletta, F., Poggi, I., Savy, R., Sorgente, A.: The chrome manifesto: integrating multimodal data into cultural heritage resources. In: Fifth Italian Conference on Computational Linguistics, CLiC-it (2018)
5. Di Maro, M., Valentino, M., Riccio, A., Origlia, A.: Graph databases for designing high-performance speech recognition grammars. In: IWCS 2017—12th International Conference on Computational Semantics, Short papers (2017)
6. Eisenstein, J., Christoudias, C.M.: A salience-based approach to gesture-speech alignment. In: HLT-NAACL 2004: Main Proceedings. pp. 25–32. Association for Computational Linguistics, Boston, Massachusetts, USA (May 2 - May 7 2004), https://www.aclweb.org/anthology/N04-1004
7. Gebru, I.D., Ba, S., Evangelidis, G., Horaud, R.: Tracking the active speaker based on a joint audio-visual observation model. In: IEEE International Conference on Computer Vision Workshop (ICCVW) (2015)
8. Grazioso, M., Cera, V., Di Maro, M., Origlia, A., Cutugno, F.: From linguistic linked open data to multimodal natural interaction: A case study. In: 2018 22nd International Conference Information Visualisation (IV). pp. 315–320. IEEE (2018)
9. Hunt, A., McGlashan, S.: Speech recognition grammar specification version 1.0. Tech. rep., W3C (2003)
10. Johnston, M.: Unification-based multimodal parsing. In: Proceedings of the 17th International Conference on Computational Linguistics - Volume 1. pp. 624–630. COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA (1998). https://doi.org/10.3115/980451.980949, https://doi.org/10.3115/980451.980949
11. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide–design and evaluation of a real-world application. In: Intelligent virtual agents. pp. 329–343. Springer (2005)
12. Lepouras, G., Vassilakis, C.: Virtual museums for all: employing game technology for edutainment. Virtual reality **8**(2), 96–106 (2004)

13. Lison, P., Kennington, C.: Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules. Proceedings of ACL-2016 System Demonstrations pp. 67–72 (2016)
14. Marty, P.F., Jones, K.B.: Museum informatics: People, information, and technology in museums, vol. 2. Taylor & Francis (2008)
15. McGlashan, S., Fraser, N., Gilbert, N., Bilange, E., Heisterkamp, P., Youd, N.: Dialogue management for telephone information systems. In: Proceedings of the third conference on Applied natural language processing. pp. 245–246. Association for Computational Linguistics (1992)
16. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago press (1992)
17. Origlia, A., Paci, G., Cutugno, F.: Mwn-e: a graph database to merge morpho-syntactic and phonological data for italian. Proc. of Subsidia, page to appear (2017)
18. Othman, M.K., Petrie, H., Power, C.: Engaging visitors in museums with technology: scales for the measurement of visitor and multimedia guide experience. In: IFIP Conference on Human-Computer Interaction. pp. 92–99. Springer (2011)
19. Oviatt, S.L., DeAngeli, A., Kuhn, K.: Integration and synchronization of input modes during multimodal human-computer interaction. In: Proceedings of Conference on Human Factors in Computing Systems CHI '97 (March 22-27, Atlanta, GA). ACM Press, NY. pp. 415–422 (1997)
20. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: developing an aligned multilingual database, pp. 293–302 (2002)
21. Richter, V., Carlmeyer, B., Lier, F., Meyer zu Borgsen, S., Schlangen, D., Kummert, F., Wachsmuth, S., Wrede, B.: Are you talking to me?: Improving the robustness of dialogue systems in a multi party hri scenario by incorporating gaze direction and lip movement of attendees. In: Proceedings of the Fourth International Conference on Human Agent Interaction. pp. 43–50. ACM (2016)
22. Schneegans, S., Schöner, G.: Dynamic field theory as a framework for understanding embodied cognition. In: Handbook of Cognitive Science, pp. 241–271. Elsevier (2008)
23. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems: The journal version. Dialogue & Discourse **9**(1), 1–49 (2018)
24. Stefanov, K., Sugimoto, A., Beskow, J.: Look who's talking: Visual identification of the active speaker in multi-party human-robot interaction. Association for Computing Machinery (ACM) pp. 22–27 (2016)
25. Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., Swartout, W.: Ada and grace: Direct interaction with museum visitors. In: International Conference on Intelligent Virtual Agents. pp. 245–251. Springer (2012)
26. Wachsmuth, I.: Communicative rhythm in gesture and speech. In: Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction. pp. 277–289. GW '99, Springer-Verlag, London, UK, UK (1999), http://dl.acm.org/citation.cfm?id=647591.728724
27. Webber, J., Robinson, I.: A programmatic introduction to neo4j. Addison-Wesley Professional (2018)
28. Wu, L., Oviatt, S.L., Cohen, P.R.: Multimodal integration - a statistical view. IEEE Trans. Multimedia **1**, 334–341 (1999)
29. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia **19**(2), 4–10 (2012)