1. TOWARDS BELIEVABLE INTERACTIVE EMBODIED AGENTS

Among the goals of research on Autonomous Agents one important aim is to build Believable Interactive Embodied Agents, ones apt to be applied to friendly interfaces in e-commerce, tourist and service query systems, entertainment and as pedagogical Agents.

A Believable Agent is one able to show (even, may be, to feel?) emotions and one who has a definite personality. An Interactive Agent has to take into account the particular User and the particular Context where the interaction takes place, and therefore has to make up its own model of the User and the Context, and interact with it by following the rules of face-to-face interaction, like turn-taking, back-channel and so forth. An Embodied Agent must be able to interact with the User not only through written text, but in all the modalities a human Agent may use: through words, voice, gesture, gaze, facial expression, body movements, body posture (sometimes, maybe, even through touch?). But it also must be able to conceive, represent and convey all the possible meanings that natural language and multimodal interaction may convey in Humans.

The list of capacities required by a Believable Interactive Embodied Agent allows us to sketch the outline of the steps to move in this field of research; some of them have already been moved in recent work, and are fairly represented among papers in this workshop.

Research must include three phases. First, a phase of empirical research aimed at finding out the regularities in the mind and behavior of Human Agents, and at constructing models of them. Second, a phase of modelling of Believable Interactive Embodied Agents, where the rules found out are formalized, represented and implemented in the construction of Agents. Third, a phase of evaluation of the implemented Agents, aimed at testing how they fit the User’s needs and how similar they look to a real Human Agent.

2. EMPIRICAL RESEARCH

Let us see what has been done already and what should yet be done to complete this program, by listing the topics that have been tackled by previous research.

a. Emotion. A lot of precious work has been done in the realm of Emotions, not only in the endless literature in the Psychology of Emotions [63, 67, 15], but also in the field of Emotion simulation [45, 18, 21, 48, 46]. Models of Emotions have been proposed that can be implemented in Emotional Agents [3, 39].

b. Personality. Models of Personality in Computer simulation have mainly followed routes quite different from the classical models in Psychology (Freud, Yung...). After the brilliant work by Carbonell, more recent research has interestingly modeled and implemented different possible personalities in Artificial Agents [7, 68] and taught them to guess the User’s personality in User Models research [24, 8].

c. Multimodal Communication. Research on what is presently called Multimodality heavily leans on much of the research on so-called Nonverbal Communication. In this field, work has been done both on single modalities and on the relationship among the different modalities. Milestones that are worthy quoted recurrently in all papers, both in the simulation era and before, are for example the names of Kendon and McNeill for gestures [27, 40], Ekman and Friesen [15] and Chovil [10] and Fridlund [20] for facial expression, Argyle and Cook [2] and Beattie [4] for gaze, Schellen [62] and Mehrabian [41] for body posture, Hall [22] for proxemic behavior. About the synchronization of different modalities, especially word and gesture, important works are ones by Kendon [30], Duncan [13], Rimé and Scharlatura [60], Knapp [34], Condon and Ogston [11], Poggi and Magno Caldognetto [53]. More recent work is being done to analyse specific multimodal behavior in view of creating Autonomous Multimodal Agents. Some good examples of this kind are the works by Lester [37], André and Rist [1] and, in this workshop, the one by Cassell et al. But more work has to be done yet: in the next section we outline what in our view should be the steps of research to reach in order to construct an Interactive Multimodal Agent.

3. MULTIMODALITY IN HUMANS

To state the research steps needed to reach a model of multimodal communication, we can see the human body as...
composed by various parts, each of which bears its specific communicative repertoire. In our view, in fact, each part of our body that may use to communicative purposes may be seen as something that is depository of its own communication system.

Let us see how many productive organs a human has, that is, how many communication systems it can use. If we consider simply the visual and the acoustic modality, we can distinguish a fair amount of body parts that produce communicative signals. Starting from the head, we have the head itself, and then face, hands, trunk and legs. Not only each of these parts of the body, but even each of their subparts (say, the regions of face, forehead, eyes, nose, mouth) produces its own system of signals. The head produces head movements, the eye region (eyebrows, eyelids, eyes) all the signals of gaze; the nose can wrinkle or dilate nostrils; the mouth produces, in the acoustic modality, words and prosodic and paralinguistic signals, while in the visual modality it produces visemes but also smiles and grimaces; shoulders, arms and hands produce gestures, while trunk and legs produce postures, movements, orientations and proxemics signals.

Now, all these signals are exhibited at once in multimodal communication. Therefore, two things are worth studying in order to produce Multimodal Agents:

1. the structure of each communication system by itself, that is, it is necessary to provide descriptions of the different mode-specific languages of the body.

2. the rules stating how the Agent chooses to convey meanings via one or the other modality, and how it synchronizes signals in different modalities.

### 3.1 Mode-specific languages

Let us start with the first issue. In our hypothesis, each system of signals makes up a “mode-specific communication system”, that is, a set of rules to link signals to meanings. The links between signals and meanings may be either “codified” or “creative” [38, 53]. In the former the signal-meaning link is coded in memory (for example in words or symbolic gestures) on a biological or cultural basis (say, both the biologically determined gesture of shaking fist up and the “Churchill” gesture for Victory mean happiness for an achievement), and a whole set of these links makes a “lexicon”. In the latter, what is coded in memory is only a small set of inference rules about how to create a new signal starting from a given meaning, or about how to retrieve a meaning from a given signal (like in pantomime or in the creation of new words in natural languages).

Moreover, a lexicon usually also includes an “alphabet”, that is, a set of sub-lexical components, the rules of production of signals, that, variously combined simultaneously or in sequence, form all the possible signals of the lexicon.

According to our hypothesis, many of the systems of communication above are of a lexical kind: not only words or symbolic gestures, as it is generally accepted, but also other kinds of gestures (say, batons or affect displays) and even gaze, facial expression, posture shifts, do have each a precise meaning that is coded in the Speakers’ memory: in our opinion, it is only thanks to this that we can understand each other! In fact, each signal in each modality can be attributed some meaning, and this meaning can be restated in a verbal language. For instance, the gesture of shaking fists up means “I am exulting”, a baton, “this is the comment f my sentence”, a frown may mean either “I am worried” or “I am angry at you”, a posture shift means something like “I am shifting to another topic”, and so forth.

A first task to accomplish in order to make Artificial Agents that really communicate multimodally is then to find out the rules that link signals to meanings. For “creative” communication systems one has to find out the inference rules that state how new signals may be created by a Speaker and understood by an Addressee. Some studies accomplishing this task are for example McNeill [40], Magno Caldognetto and Poggi [38] and Sowa and Wachsmuth [64]. For “codified” systems, that is, “lexicons”, the task is to compile lexicons of the systems in all modalities. Some examples in this field are the dictionaries of Sign Languages and, more recently, the flourishing of dictionaries of symbolic gestures for many different cultures [42] for gestures all over the world; Morris et al. [43] for the Mediterranean area; Tumarkin [69] for Japanese gestures, Kreidlin [35] for the Russian, Payratò [47] for the Catalan, Posner and Sereñari [59] for Berlin gestures, Poggi [51] for the Italian.

But also for other systems can we write down lexicons; Ekman and Friesen’s FACS [17], for instance, can be considered a lexicon of the face; a sketch of a lexicon of gaze is Poggi, Pezzato and Pelachaud [58]; and also lexicons of very specific systems may be written, like the fragment of lexicon of “performative faces” by Poggi and Pelachaud [56], the lexicon of the orchestra Conductor’s face [55], the lexicon of deictic gestures and gaze [26].

A task somehow included in the construction of lexicons is the discovery of “alphabets” of nonverbal systems. In this view, important examples are Laban’s notation [36] and Birdwhistell’s [5] system. One seminal work is Stokoe [66], who proposed the notion of “formational parameters” of signs in Sign Languages of the Deaf. He found out that each sign is produced by a particular Handshape, a Movement, a Location, to which Orientation was then added. Formational parameters were since then found out in different gestural systems of the Hearings: Calbris [6] found them for French gestures, Kendon [29] for the Australian Aboriginal Sign Language, Sparhawk [65] for Persian gestures, Romagna [61] for Italian gestures, Ekman et al. [16, 25] for the American gestures.

More recently, the notion of formational parameters has been applied to gaze [58], by singling out the parameters and values that pertinently describe each item of gaze: eye direction, eye opening, humidity, eyebrow movements and so on. And finally, formational parameters were found also to describe the communicative gestures of touch [54].

Other systems for coding and annotating nonverbal items are MPEG-4 notation [12], Kipp’s system (this volume) and Martin’s system (this volume).

### 3.2 How to construct multimodal lexicons

To make a lexicon of a particular modality requires

a. an extensive work of collecting all the items of that lexicon, and

b. an intensive work of semantic analysis for each item.

For both kinds of work, three methods can be used in order to find out the meaning of the items to write down in the lexicon. One is the Chomskian method of the Speaker’s
judgments: it consists in judging if the item under analysis is semantically acceptable in one or another context, if it is ambiguous (has more than one meaning), how it can be paraphrased in the verbal language, which other items in the lexicon of the same or other modalities may be synonyms of it, and so on. This may be done even through judgements of the researcher alone. The resulting lexicon will be in this case the representation of his/her single communicative competence; but from a theoretical point of view this is yet a good way to discover the mechanisms of that communication system, since each single competence is obviously a self-consistent system. This method has proved useful for both the extensive overview of a whole lexicon [53], and for the intensive analysis of single items [49, 53].

Of course, if one aims at a real dictionary of nonverbal items that can represent the lexical nonverbal competence shared by all people in a culture, another method is necessary: one based on questionnaires through which the researcher can verify to what extent his/her own intuitions are shared. This is perhaps the most used method in gesture literature [43, 59, 47].

Finally, when one goes into the semantic analysis of single items, a detailed intensive analysis is needed. The third one is then an observational method: the researcher collects several video-recorded occurrences of a single nonverbal item used in real-life situations, and tries to single out, first its meaning in each occurrence, then the core meaning that is common to all occurrences. This method is generally used for detailed analyses of single gestures, for instance, by Kendon [31], Mueller [44], and Poggi [50].

But how can we find the meanings of nonverbal items? In compiling dictionaries of natural languages, Linguists have generally started by collecting words in a language and then tried to outline their meanings. To use such a method looks quite obvious for verbal languages, since signals in these communication systems are fairly segmentable, and introspection of their meaning is made easier by their being used with total awareness. Neither condition, though, always holds for nonverbal communication systems, where the signals are seldom produced at a high level of awareness, and the job of finding out “lexical” units has not yet been accomplished thoroughly. For this reason, in this case it is both easier and more heuristic to start the other way around: first try to guess what in principle may be all the possible kinds of information an Agent may need to provide other Agents for its adaptive goals; and then to wonder if, which and how those kinds of information are generally conveyed in such or such modality, such or such communication system. This “deductive” method for finding meanings has proved useful, for instance, in building a lexicon of gaze [58]. We started from the idea that in whatever communication system (animal and human communication, verbal and nonverbal human communication systems), it is possible, and often useful, to distinguish at least three classes of meanings: Information on the World, Information on the Speaker’s Identity, and Information on the Speaker’s Mind.

**Information on the World**. When we talk we provide information on the concrete or abstract events we communicate about, their actors and objects, and the time and space relations among them. This is provided, of course, mainly through the words of sentences and their syntactic structure; but often also by deictic, iconic and symbolic gestures. In fact, a deictic gesture indicates something in the surrounding environment: a way to set the reference of our discourse, then a way to explain what, in the external world, we are going to talk about. An iconic gesture instead describes (with a literal or metaphorical sense) the shape, size or movements of some referent we are mentioning. Finally, some symbolic gestures directly mention some object, feature or action. But not only gesture can indicate or describe; sometimes this is done also through gaze, voice, head or body movements: we may point at things or persons in the context even by eye or chin direction, and we may refer to some feature of some word or person also by gaze, prosody and body movement: we squeeze our eyes to refer to something small or difficult, open eyes wide to refer to something huge, lengthen a vowel to say something is long, or speak in a staccato way to indicate precision; we may mime another person’s movement by moving as she does.

**Information on the Speaker’s Identity**. While talking, even if generally not out of a conscious and deliberate goal of communication, we provide information on our Identity: with physiognomic traits of our face, eyes, lips, the acoustic parameters of our voice, and often our posture, we provide information on our sex, age, socio-cultural roots, and personality.

**Information on the Speaker’s Mind**. While we are talking of events of the external world, we also communicate why we want to talk of those events, what we think and feel about them, how we plan to talk of them and so on: we provide information on beliefs we’re mentioning, our own goals concerning how to talk about them, and the emotions we are feeling while talking.

More specifically, among information concerning our own beliefs, we may inform:

1. on the degree of certainty of the beliefs we are mentioning, by words like perhaps, certainly, or the conditional or conjunctive verb mode, but also by frowning, which means: “I am serious in stating this”, or by opening hands, which means “this is self-evident”;  
2. on the source of the beliefs we mention, whether they come from memory, inference, or communication [9]: we look up when trying to make inferences, snap fingers while trying to remember, we make the gesture “quote” with index and middle fingers curved twice to mean that we are quoting other people’s words for which we are not responsible.

The goals of ours that we inform about while talking concern:

1. the performative of our sentence, that may be conveyed by performative verbs, but also through intonation or through performative facial expression [57];  
2. the topic-comment distinction within a sentence or discourse, which may be conveyed by batons, by eyebrow raising, by intensity or pitch of tonic vowel;  
3. the discourse rhetorical relationships: a list may be scanned by words (first, second, third...), but also by
studying in depth, through careful analyses of real data.

tle and intertwined than one could think, and then worth
of different modalities. These links are in fact more sub-
build "mode-specific" lexicons, that is, to single out the cor-
tention, facial expression, gaze and posture.

Again, we may inform on the emotions we are feeling while
talking, not only by affective words, but with gestures, emo-
tional intonation, facial expression, gaze and posture.

This semantic taxonomy is, in our view, a useful tool to
build “mode-specific” lexicons, that is, to single out the cor-
respondences between signals and meanings in the systems
of different modalities. These links are in fact more sub-
tile and intertwined than one could think, and then worth
studying in depth, through careful analyses of real data.

3.3 Signal synchronization

The other important issue before constructing Multimodal
Agents is to assess how the signals coming from the commu-
nication systems of the different modalities mix up in real
interaction. Just thanks to the fact that we are endowed by
communication systems in different modalities, we can use
multiple signals at the same time: we may in the same in-
stant utter a word, move our trunk towards our interlocutor,
look at him and while raising our eyebrows, and open and
drop hands. What determines which signals we will perform
at each moment of our discourse; how do we choose to com-
municate some content through words or other signals, or
through both?

Various scholars have dealt with the relationship between
speech and other modalities: see Kendon [33, 28, 31, 32];
Freedman [19]; Rimé and Schiaratura [60]; McNeill [40]; Ek-
man [14].

Some factors affecting the choice and synchronization of
signals are: the presence-absence in the different modalities
of a signal apt to convey the intended meaning; the likeli-
ness of occurrence of that signal as opposed to others; the
appropriateness of a specific signal in the particular context
at hand.

3.4 How to study mode-specific languages
and multimodal synchronization

Both the task of constructing lexicons of mode-specific
languages and the task of studying choice and synchroniza-
tion rules require a wide use of annotation systems and tools
to analyze real videotaped data. Some useful tools to tran-
scribe real data are Media Tagger (Max Planck Institute
for Psychologists); Com-Trans [23]; Poggi and Magno Cal-
dogueto [52]; some new ones are presented in this Workshop
(Martin et al., this volume; Kipp, this volume). These tools
are useful because they allow studying the precise timing of
signals against each other, they may provide relevant infor-
mation about the planning and distribution of multimodal
signals.

4. A WORKSHOP ON MULTIMODAL EM-
BODIED AGENTS

The contents of the present Workshop follow the topics
outlined above.

Many of the papers tackle more than one of these topics,
because they present a complete system architecture. Yet,
we have decided to cluster the papers based on the workshop
topics.

The papers by Martin et al., Kipp and Cassell et al. are
good representatives of basic empirical research specifically
devoted to the simulation of particular multimodal behav-
iors. Cassell et al. present a research on posture shifts,
concluding that these behaviors clearly tend to occur dur-
ing turn shifts and especially in correspondence with topic
shifts; from this research they draw a model of posture shifts
and implement it within the Agent REA. They provide a
good example of how one should collect, analyse and ex-
plot empirical data from human-human interaction with
regard to verbal / nonverbal inter-dependencies (timing, co-
ocurrence, context). Kipp and Martin et al. present coding
schemes for annotating gestures, designing tools for coding,
viewing and analysing multimodal interactions; in partic-
ular, Martin et al. propose a standardized coding scheme
to encode multimodal corpora producing XML descriptions.
This approach allows the exchange of annotated documents
in an easy way, which responds to today’s requirements.
Kipp describes the design of a tool, ANVIL, for coding,
viewing and analysing multimodal interactions. The author
provides a categorization scheme of gesture and an analysis
of the communicative functions of gestures aiming at getting
information on human gestural behavior in conversation set-
ings in order to generate synthetic presentation teams.

The papers by Allbeck and Badler, Guerrin et al., Prendinger
and Ishizuka, and Traum and Rickel are interesting exam-
ple of how one can set up a dialog model and cognitive
model of the Agent to make it more believable. The topic
of what are the requirements to make an Agent Believable
is tackled by Allbeck and Badler, who illustrate the multi-
ple variables entering in the communication process. They
are setting up the background to drive consistent behavior
from a cognitive model of the Agent. The social relationship
between Agent and User is taken into account by Guerrin
et al. and by Prendinger and Ishizuka; a topic which is
relevant in determining how to assess, interpret, and affect
responsiveness to the context in which the interaction takes
place. Traum and Rickel’s dialog model is able to simulate
dialog party between two or more participants. For each dis-
cussion among Agents the authors propose several conver-
sation types, such as initiating a speaking turn, negotiating
about something with other Agent(s), or even getting the
attention of other(s). Guerrin describes a system where an
agent, a shopkeeper, is able to interact with a user, to take
initiative by presenting the user objects of its virtual shop,
or, for example, by pointing at objects in its shop.

Another aspect toward the creation of a Believable Agent
is the relevant issue of emotion, its modeling and its com-
municative output, that is exploited in Marsella et al. and
in Silva et al.. Marsella et al. propose an approach where
emotion appraisal, influence of the emotion on the Agent’s
verbal and nonverbal behavior, and the communicative in-
tention of the agent are encompassed in a single and complex
model. Another approach taken by Silva et al. is to script
the text a storyteller is saying with emotion tags and other
communicative tags. In another level, Huang et al. propose
a cognitive model of soccer player based on an evaluation of
the context of the match, where the soccer is able to com-
pute its next move.
Once represented a complex dialog planner and an emotion model, one needs to address the problem of computing the behavior (facial expression, gaze and gesture type, body movement) the agent should display. To avoid a simplistic approach that adds all the behaviors computed by the system to each other, Ruttkay and Noot have developed a constraint system that allows expressions to appear, remain on the face, and disappear in a non-systematic way, but nevertheless, always within pre-defined limits. For example, simultaneous expressions may differ, for example, in their degree of symmetry, on their onset time, thus creating a more natural animation. Paradiso and L’Abbate derive a computational model to combine facial expressions in a more sophisticated manner. Several computation formulas are proposed by the authors.

Finally the problem of how to evaluate multimodal interaction with an embodied agent has to be faced. Heylen and Nijholt draw conclusions from their experience in the creation of several embodied agents and propose some technical considerations to follow when creating Embodied Agent Systems. McBreen et al. designed an empirical setting to evaluate the impact of different types of agents (male vs female, casual vs formal dressing), in several applications (a cinema box-offices, a bank and a travel agency). The authors address the delicate problem of trust: does the user trust the agent, is the user ready to let the agent make some decisions and perform some transactions on its behalf? Baljko designs an iterative process where evaluation is performed and its findings give rise to the refinement of the Agent model; as iterations grow it will offer to the system a mean to compute more coordinated and communicative behaviors for a more Believable Agent.

5. REFERENCES

6. REFERENCES
[26] W. Johnson, J. Rickel, and J. Lester. Animated pedagogical agents: Face-to-face interaction in


[61] M. Romagna. L’alfabeto dei gesti, parametri formazionali nel lessico dei gesti simbolici usati dagli


