J Multimodal User Interfaces (2010) 3: 249–261
DOI 10.1007/s12193-010-0048-y

ORIGINAL PAPER

Generating distinctive behavior for Embodied Conversational Agents

Maurizio Mancini · Catherine Pelachaud

Received: 19 November 2009 / Accepted: 16 August 2010 / Published online: 15 September 2010

Abstract In this paper we describe an algorithm for generating distinctive behavior for Embodied Conversational Agents. To this aim, we introduce the concepts of agent's general behavior tendency, named Baseline, and local behavior tendency, called in turn Dynamicline. Depending on the communicative intentions of the agent, the Baseline is modulated. The obtained behavior tendency corresponds to the Dynamicline which is then used to determine the nonverbal signals and their expressivity the agent will produce to communicate its intentions. We also propose a system to extract the movement expressivity of a human user standing in front of a camera. The extracted characteristics are then used to characterize the agent's Baseline. We end the paper by presenting an evaluation study of our model.

Keywords ECA · Behavior · Multimodal · Distinctive · Expressive

1 Introduction

Embodied Conversational Agents (ECAs) are a kind of Human-Computer Interface that are embodied and have conversational skills [10]. They have a human-like aspect, in both appearance and behavior, capable of exhibiting conversational functions, of showing emotional states, personality traits and so on. In general people tend to deal with computers as if they were humans, as demonstrated by Reeves and Nass in their book "The Media Equation" [38]. These results are particularly verified when the Human-Machine Interface integrates an ECA. Using ECAs tends to increase the quality of communication between human and computer as they are designed to communicate and interact in a human-like manner [10].

The first systems implementing ECAs aimed mainly at reproducing the basic aspects of human-human conversation: ECAs had schematic bodies, exhibited monotonic speech, and produced few gestures and facial expressions [10, 22, 24, 25]. In recent years, developers focused on refining ECAs by modeling key aspects of human-human interaction, involving the verbal and nonverbal behaviors that are produced in human communication: the words we utter constitute the verbal part, while the nonverbal part includes a very large set of behaviors, going from speech intensity and intonation to facial expressions, hand gestures, head and torso movements, posture changes and so on [2, 3, 15, 19, 28].

In their work about bodily communication, Argyle [3] and Gallaher [16] state that there is an underlying tendency which is constantly present in each person's behavior: for example people that look more at their interlocutors tend to do so in most situations, that is, there is a certain amount of consistency with the person's general tendency. The paper by Wallbott and Scherer in [44] illustrates a study on actors' body movements during the expression of several emotions. Some behavior characteristics seemed independent from the emotion: for example the number of head movements and total activity. All of the above and other studies [2, 40, 43] suggest that the behavior of a person does not depend only on what the person is communicating, that is, their *communicative intention*, but also on the person's *general behavior tendency*, that is, their personal way of behaving.

M. Mancini (✉)
InfoMus Lab, DIST, Università di Genova, Viale Causa 13,
16145 Genova, Italy
e-mail: maurizio.mancini@dist.unige.it

C. Pelachaud
CNRS-LTCI, TELECOM ParisTech, 37 rue Dareau,
75014 Paris, France
e-mail: catherine.pelachaud@telecom-paristech.fr

On the other hand, the way in which people perform nonverbal behavior depends not only on their general behavior tendency, but also on other changing *dynamic factors*: e.g., emotional states, disposition and/or relation with other person and/or event and/or object, social roles.

What we propose in this paper, is the implementation of ECAs exhibiting nonverbal behavior that is driven by the agent's *general* behavior tendency modulated by *dynamic* factors: we call them agents exhibiting *distinctive behavior*.

For the purposes of the work presented in this paper, after illustrating some related works in Sect. 2, we describe the implementation of our *distinctive behavior algorithm* in Sect. 3, a computational module that, starting from a description of the behavior tendency of an ECA and the list of the agent's communicative intentions, computes the nonverbal behaviors the agent has to perform. Section 4 describes a system extracting automatically the agent's Baseline expressivity from the behavior of a user standing in front of a camera. In Sect. 5 we present an evaluation study in which we tested how is perceived the agent's behavior tendency.

We developed and tested our algorithm in the Greta agent framework developed by Pelachaud et al. [26, 27, 34], that is, we used the Greta animation module to generate the animation corresponding to the nonverbal signals determined by our model.

2 State of the art

Other research has addressed the problem of capturing human nonverbal behavior variability in creating ECAs.

2.1 Variability depending on the behavior repertoire

In [20, 21, 32] Kipp et al. present their gesture animation system based on statistical models of human speakers gestures. The goal of their work is the creation of *gesture profiles* and *gesture lexicons* of human speakers, that are then used in combination with *general rules* to generate the gestures (arm and torso movements) of a virtual character.

Kallmann and Marsella [18] propose to dynamically assemble and blend a pre-designed repertoire of gestures with automatically generated gestures. The characters are able to react to events and, for example, to interrupt the ongoing pre-designed gesture to produce another gesture that depends on an event (e.g., a person entering the room).

Ruttkay et al. [39] propose the idea of behavior style, defined in terms of when and how the ECA uses certain gestures. Agents are differentiated by a *style dictionary* that defines the agent's gesture repertoire.

Poggi et al. [35] propose a model of a *reflexive* agent. At first, the agent has to decide if a given communicative intention has to be communicated or not. The decision depends on several factors: the agent's relationship with its interlocutor, the agent's knowledge about its interlocutor's personality and the agent's goals. For example, if the information to be conveyed is related to the agent's emotional state, the facial modality has a higher priority in the array. On the other hand, if the physical context of communication is a noisy place (e.g., disco, stadium) then the gesture modality is privileged.

2.2 Variability depending on movement expressivity

Allbeck and her colleagues created a system to select the most appropriate nonverbal behaviors (gestures and facial expressions) and to control the movement quality of the Jack agent [4] depending on its personality and emotional state [1].

The way in which the agent performs its movements is influenced by a set of high level parameters, embedded in the Expressive Motion Engine (EMOTE), an implementation of the *Effort* and *Shape* components defined by the Laban Movement Analysis system [23].

The influence of Laban-like parameters on pre-stored facial expression is also implemented in the FacEMOTE system [6]. By varying these parameters the agent's set of static facial expressions can appear and disappear in a quicker/slower way and show several degrees of intensity (muscle contraction).

Neff et al. [29–31] discovered some key movement properties by reviewing arts and literature, such as theater and dance. They found that body and movement characteristics such as balance, body silhouette (contour of the body), position of torso and shoulder, etc. influence the way in which people perceive others. They have implemented three motion properties into animated characters: the pose of the character, the timing of movements and the transition from one pose to another.

2.3 Variability depending on emotion and personality

Egges et al. [14] as well as Ball and Breese [5] implemented models in which the agent personality traits and dynamic emotional state are used to determine the updated agent's emotional state.

DiPaola et al. [13] implemented a system for simulating the agent's *mood* depending on a music source. The agent moves its head and updates its facial expression as the emotional content of the music performance changes. The agent has a personality, that is, a set of rules that determines how the agent tends to respond to the external stimuli.

2.4 Discussion

The works of Kipp et al. [20, 32] and Ruttkay [39] present both similarities and complementarities with our system.

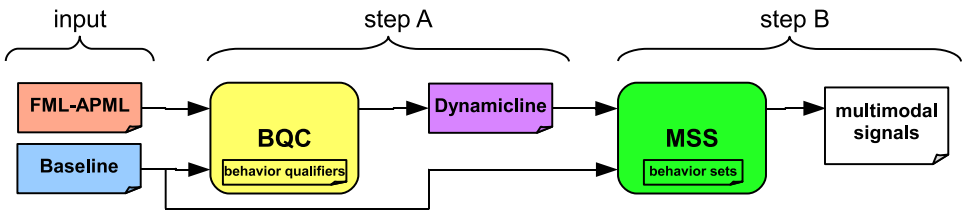

Fig. 1 The *distinctive behavior* algorithm diagram

They aim at defining conversational agents that exhibit a distinguishable behavior: that is, agents with a recognizable way of performing nonverbal behavior.

With Kipp's work, we share a same goal: to model the visible effects of individuality; however we do not investigate the origins of these influences (which could be cultural, social, etc.) as Ruttkay et al. do. Similarly to us, Ruttkay considers the agent's static definition and computes its behavior by evaluating the current communicative intention. Complementarities emerge as we look at other aspects of Kipp and Ruttkay's work: they both implement behavior variability by defining the agent repertoire of gestures; instead, in our work, we aim to define only the agent behavior habits in terms of preferred communicative modalities, we do not define repertoires; our system will determine the agent's behavior from a common set of behaviors. Our agents do not differ in the gestures they use, but in the way they tend to use their modalities, and in the quality (amplitude, speed, energy, etc.) of movement.

Our set of expressivity parameters shares similarities with the EMOTE and FacEMOTE parameters [1, 6] as they allow one to alter the agent's behavior with a few high level descriptors. At the same time, their systems are complementary to the work we present in this paper: they explain how static body/face poses are modulated by expressive parameters; instead we present a method for modulating the expressive parameters values depending on the agent's communicative intention.

Compared to Neff et al. [30, 31], we do not take into account modulations of behavior due to the character's physical constraints (for example balance constraints, e.g., the posture adjustments caused by the variation of the position of the agent's center of gravity). Moreover, Neff et al. describe movement expressivity at a lower level, e.g., at limbs rotation angle level. That is, they could compute the animation of agents exhibiting different knee quantity of rotation or relation between wrist position in space and pelvis rotation. Our parameters work at a higher level, allowing us to model agents exhibiting "larger" gestures, with no need to take care of explicit limb rotation angles [31].

Finally, the works in [5, 13, 14] are more concerned with the agent's current emotional/mood state, presenting algorithm to update it. Instead, in our system we assume that the agent's emotional state is an input data and we provide a technique to determine how it influences the agent's visible behavior.

3 Distinctive behavior algorithm

To create an algorithm implementing agents exhibiting distinctive behavior we aim at performing several steps:

1. modeling the agent's *general* behavior tendency and *communicative intention*;
2. modeling how *dynamic* factors could modulate the agent's general tendency;
3. calculating the agent's *local* behavior tendency, that is, the tendency that endows both the general tendency and the dynamic modulation factors;
4. computing the nonverbal *signals* the agent has to display, given its communicative intention and local behavior tendency.

Our algorithm for distinctive behaving agents is composed of sequential modules. It uses different representation languages to ensure the data flow within these modules. We will introduce our algorithm by illustrating the process represented in Fig. 1 step-by-step.
The *input* part of the algorithm corresponds to point 1 of the above list; *step A* refers to points 2 and 3; *step B* allows us to implement point 4. The algorithm takes as input both the agent's communicative intentions specified by an FML-APML file and the agent's general behavior tendency encoded in the Baseline. The final output of the algorithm is the multimodal behavior of the agent, described as a list of multimodal signals, i.e., facial expressions, torso and head movements, gestures and so on. We describe how the algorithm works in the following sections.

An example of the results of the algorithm computation can be seen at:
http://www.mauriziomancini.org/downloads/jmui-demo.mpg.

The above video shows two agent defined by different Baselines: the one on the left prefers the gesture modality and performs quick and large movements; the one on the right prefers to use the head and face modalities with smaller and smoother movements.

3.1 Input

This is the input data of our algorithm for implementing agents exhibiting distinctive behavior. It does not involve computation, instead it includes the definition of two representations, one for the agent's communicative intention and one for the agent's general behavior tendency.

3.1.1 FML-APML

The FML-APML tags are an extension of the ones defined by APML, an XML-based markup language for representing the agent's communicative intention and the text to be uttered by the agent [12].

FML-APML tags endow different types of data:

- *communicative intention*: it is the information an agent may seek to communicate [35, 37], that is, information on the world or on the agent's mind. For example we could use FML-APML to describe that the agent has the intention to communicate that it is approving what the user said. With respect to APML, the FML-APML tags also include information about:
  - *emotional state*: in FML-APML we can represent the agent's emotional state, that is, the emotional state the agent aims to express. We can also represent complex expressions, for example, if the agent has a certain emotional state but it hides it by showing another one, fake [33]. We base our extension on the EARL language[41].
  - *world*: when communicating with others, we could have the intention of communicating about some physical or abstract properties of objects, persons, events. For example, we can accompany speech with hand shapes that mimic the shape of an object, or perform large arm movements to give the idea of an "amazing" event.
- *timing*: each tag contains explicit timing data, similarly to what happens, for example, in the BML language [42]. It allows us to define time markers, that is, point in time to which we can attach, for example, the starting and ending time of one or more FML-APML tags.
- *importance*: not all the communicative intentions we communicate to others have the same level of importance since, as explained in [11, 36], different people may attribute a different importance to the same goal. In our system the importance attribute allows us to choose the multiplicity of multimodal behaviors, as it will be clarified later. If the importance raises, we increment the number of modalities on which the agent's intentions are communicated.

3.1.2 Baseline

We now introduce one of the key concepts of our work: the agent *Baseline*. In Sect. 1 we referred to researchers who demonstrated that there is an underlying tendency in the way each one of us behaves: a person performing wide movements while gesturing will probably walk with large steps, write large letters; people that tend to look more and perform a lot of gestures will continue to do so in most situations [2, 3, 16, 43]. In our work we provide a system that allows the modeling of such characteristics of the external visible behavior of a person. This is the idea we want to capture with the concept of *Baseline* for ECAs.

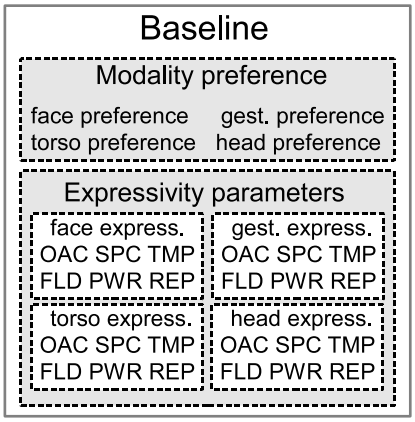

Fig. 2 The Baseline structure. It is composed of two subsets of parameters: the modality preference and the expressivity parameters. The first one describes the tendency the agent has to use each of the available communicative modalities; the second one describes the agent's behavior expressivity of each modality

The Baseline of an agent is defined as a set of fixed parameters that represent the agent's general, underlying behavior tendency. Figure 2 represents the Baseline structure, that consists of two main subsets, that is:

1. *Modality preference*. In our system we define the modality preference to represent the agent's degree of preference for each available modality. If for example we want to specify that the agent has the tendency to mainly use hand gestures during communication, we assign a high degree of preference to the gesture modality; if it uses mainly the face, the face modality is set to a higher value, and so on. For every available modality (face, head movement, gesture, posture), we define a value between 0 and 1 which represents its preferability. Agents can also use two or more modalities with the same degree of preference. This means that the agent communicates with these modalities equally.
2. *Expressivity parameters.*[1] With the terms *expressivity of behavior* we identify the external, visible qualities of movement, like its speed, amplitude, fluidity and so on. Expressivity is an integral part of the communication process as it can provide information on the emotional state, mood and personality of the person [44].

   In the following description we talk about "movement" in a general sense, but the same description can be applied to arm/hand movements, head movements, facial muscles movements and so on. In our algorithm we

[1] We are very thankful to Björn Hartmann for defining the Greta's set of expressivity parameters, and for implementing them in the Greta's gesture generation engine [17].

use the following 6 *expressivity parameters*, as defined by Hartmann et al. [17]:

- *Overall Activity—OAC*: amount of activity (e.g., passive/static versus animated/engaged). For example, as this parameter increases, the number of head movements, facial expressions, gestures and so on, increases.
- *Spatial Extent—SPC*: amplitude of movements (e.g., expanded versus contracted). This parameter determines the amplitude of, for example, head rotations and gestures.
- *Temporal Extent—TMP*: speed of movements (e.g., quick versus sustained actions). Agent's movements are slow if the value of the parameter is low, or fast when the parameter is high.
- *Fluidity—FLD*: smoothness and continuity of movement (e.g., smooth, graceful versus sudden, jerky). Higher values allow smooth and continuous execution of movements while lower values create discontinuity in the movements.
- *Power—PWR*: dynamic properties of the movement (e.g., weak/relaxed versus strong/tense). Higher (resp. lower) values increase (resp. decrease) the acceleration of the head or limbs rotation, making the overall movement look more (resp. less) powerful.
- *Repetitivity—REP*: this parameter permits the generation of rhythmic repetitions of the same rotation/expression/gesture. For example, a head nod with a high repetitivity becomes a sequence consisting of very fast and small nods.

3.2 Step A

Given the agent's characteristic and its communicative intentions and emotional state to convey, our algorithm goes through several steps (see Fig. 1).

Step A is the first step of computation of our algorithm. Input data is both the agent's communicative intention and the agent's general behavior tendency described by its Baseline. The goal of the present step of computation is to determine how the behavior tendency of the agent is modulated by the information to communicate (e.g., agent's emotion, intention and so on). That is, it determines how the agent *general* behavior tendency becomes the agent's *local* behavior tendency. This idea is similar to the *character sketch* concept implemented by Neff and Fiume in [30]: different sketches characterize, for example, the agent speed of movement.

As reported in Fig. 1 the present step of computation is performed by a module called *Behavior Quality Computation* (BQC). Internally, the BQC module contains a description of how all the possible agent's communicative intentions modulate its Baseline. This description is represented by a set of rules called *Behavior qualifiers*. The result of the modulation of the Baseline depending on the behavior qualifiers is the agent's *local* behavior tendency, that we call *Dynamicline*.

3.2.1 Behavior qualifiers

Our communicative intention influences on the choice of modalities used to display them: e.g., to communicate an emotional state of sadness usually our body movements are reduced and performed slowly. We call *behavior qualifier* the set of *modulations* that, given a communicative intention, act on the general behavior tendency of an agent. A *modulation* is defined as a variation over one of the parameters contained in the agent's Baseline and is defined as the quadruple:

$$(name, destination, operation, terms); \tag{1}$$

where:

- *name*: is the name of the qualifier and is used as a one-to-one correspondence between communicative intentions and behavior qualifiers. For example the qualifier with the name "emotion-anger" represents the behavior modulation that should be applied when the agent is communicating anger.
- *destination*: specifies where the modulation acts on and its parameter indicates where the result of the modulation is stored. It can be one of the modality preference, or an expressivity parameter (OAC, SPC, TMP, PWR, FLD, REP) of a given modality. For example if destination is equal to "face" and parameter is "SPC.value" then the current modulation result is stored into the Spatial Extent (SPC) parameter of the face modality.
- *operation*: specifies which operation should be performed among the terms listed in the modulation definition. The operators currently implemented in our system are simple mathematical operations like addition, subtraction, multiplication, division, scaling. We have also defined an assignment operator to copy values between parameters. Moreover, new operators can easily be added by defining their $C++$ implementation in the source code of the system.
- *terms*: are the terms of the modulation, where each term is either one of the Baseline parameters (modality preference or expressivity) or a numeric value. The number of terms depends on the operator, for example for a simple assignation (like $parameterX = value$) we need just one term (that is the *value* to be assigned), while for a sum (like $parameterY = term1 + term2$) we need two terms.

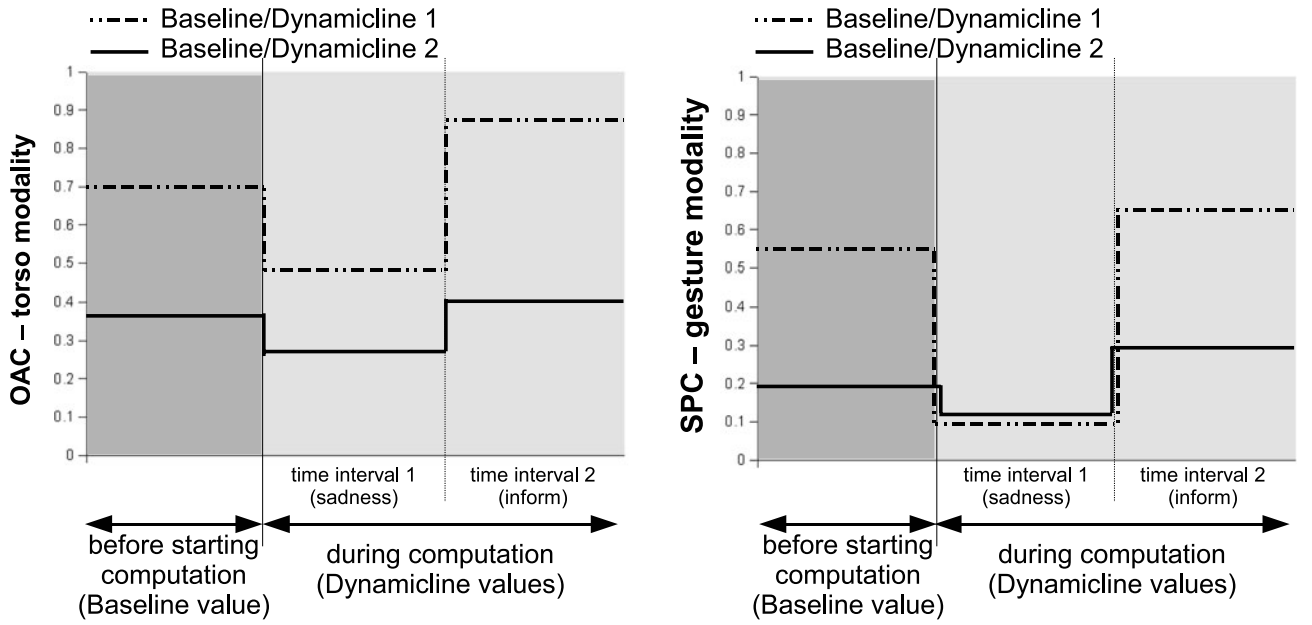

Fig. 3 Evolution over time of two different Baselines and the corresponding Dynamiclines. One of the two Baselines and the corresponding Dynamicline is represented by a *continuous line*, the other one and the corresponding Dynamicline is represented by a *dashed line*. The *diagram on the left* represents the variation of Overall activation (*OAC*) of the torso modality, the *one on the right* represents the variation of Spatial (*SPC*) for the gesture modality

As an example, let us see how we define a behavior qualifier that represents the following description:
"a sad state (i) decreases the degree of bodily activation and at the same time (ii) the speed, amplitude and energy of movements are very low".

The behavior variation described in the sentence above are of two kinds: *relative* and *absolute*. In the example, part (i) of the sentence indicates that we may decrease the degree of bodily activation. This is a relative variation because we give an indication of the current behavior tendency (Dynamicline) in terms of the general tendency (Baseline). Instead, part (ii) of the sentence indicates that speed, amplitude and energy of movement should be very low: in this case we talk about absolute values, that is, the current behavior tendency (Dynamicline) are explicitly defined, and we do not refer to general tendency (Baseline).

In our algorithm we manually defined behavior qualifiers by collecting data from the literature [2, 7, 9, 37, 44]. For example, communicative intentions regarding the regulation of conversation (turn taking, giving the turn) tend to increase the agent's preference and activation of the head and posture modalities. Moreover the literature reports also that the face is an important mean to display emotional states. So when communicating emotional states the degree of preference for the face modality is raised. Other behavior qualifiers can be defined. They could be obtained from video corpora analysis (see Sect. 4).

3.2.2 *Behavior Quality Computation and Dynamicline*

Each time the BQC module receives as input the current agent's communicative intention and Baseline it applies the modulations described by the Behavior qualifiers and produces the agent's *local* behavior tendency, that we call the agent's Dynamicline. That is, the Dynamicline is a set of parameters that derive both from the agent's Baseline and its current communicative intention. The Dynamicline has the same structure of the Baseline, illustrated in Fig. 2, but the meaning of the contained parameters is different.

For each communicative intention, the BQC module computes a new Dynamicline for the agent. The communicative intentions implemented by our system are those described by the FML-APML language: the agent may aim to communicate its emotional state, information about the world and about its mind (see Sect. 3.1.1 and [37]). The diagram in Fig. 3 illustrates an example of how two Dynamiclines corresponding to a sequence of FML-APML tags are computed starting from two Baselines.

To simplify the explanation we focus on two expressivity parameters: Overall activation (*OAC*) for the torso modality (diagram on the left) and Spatial (*SPC*) for the gesture modality (diagram on the right). The continuous and dashed line in each diagram represents the same parameter value for the two Baselines and Dynamiclines. On the x axis we represent time. In both diagrams we have three time spans: the first (bars on the left) is the starting time in which the Baselines values are set to their initial values; then (bars in

the middle) the agent aims at communicating *sadness*, so the Baselines are modulated into the Dynamiclines corresponding to the sadness communicative intention; finally (bars on the right) the agent's communicative intention is to *illustrate* something to the user and the Baselines are again modulated into the corresponding Dynamiclines. From the diagrams we may notice that the sadness state (bar in the middle) sets the *SPC* parameter (diagram on the right) of both Dynamiclines (the continuous and dashed lines) to the same value (even if for the sake of clarity, in Fig. 3, the two lines are represented as not perfectly coincident), as indicated by the corresponding qualifier definition.

3.3 Step B

When the agent's local behavior tendency and communicative intention is provided as input to the second step of our algorithm, the module called *Multimodal Signal Selection* (MSS) selects which nonverbal signals, for example gestures, facial expressions and torso movements, the agent has to perform.

In human behavior, a given intention can be communicated in a great variety of ways [2]. E.g., to communicate a state of joy we can smile, stretch our arms upwards, jump, run, scream, or produce any combination of these signals together. As represented in Fig. 1 the MSS module contains a set of rules representing such correspondence between intentions and signals, called *Behavior sets*. These sets, that are described in the next section, together with the communicative intention and Dynamicline allow the algorithm to choose the most appropriate signals the agent has to produce.

The MSS process ensures that agents with different baseline will communicate differently. For example, an agent with a very passive behavior tendency, in a joy state could produce just a light smile, without moving the rest of the body. Instead, a very expressive agent could produce a combination of signals: smiling, stretching the arms and body.

3.3.1 Behavior sets

Behavior sets model the correspondence between the agent's communicative intention and nonverbal behaviors. For example, to *greet* someone, the agent could raise the palm of its hand, showing a smile. Or, to *emphasize* a word representing important information, it could produce a head nod, raising its eyebrows. Each one of these nonverbal behaviors is referred to with the term *signal*: raising hand is a signal, smiling is a signal and so on.

We define a *multimodal signal* as a combination of signals (on single modalities) produced simultaneously on different modalities to convey a certain communicative intention. For example, the action of raising the hand palm and smiling at the same time is a multimodal signal expressing the greet communicative intention. In this paper, when a signal on a single modality is produced to communicate a certain intention, we still call it a multimodal signal, with a multiplicity of 1. Every multimodal signal acts to convey a given communicative intention.

The definition of a *behavior set* BS is a quadruple:

$$(name, sigs, core, implications); \tag{2}$$

where:

- *name*: is the name of the behavior set; with this parameter we build the one-to-one correspondence between the behavior set and the communicative intention.
- *sigs*: is a set of signals emitted on single modalities; this set represents the widest set of signals which can be used to convey the meaning specified in the parameter *name* of the behavior set. The *sigs* set does not specify *how* and *if* these signals can be combined. The next two parameters specify this information.
- *core*: it is a subset of *sigs*, representing those signals which have to appear in the multimodal signals communicating the given intention. With the *core signals* we impose the presence of one or more of these signals in the final selection. For example, we may aim at specifying that in *denying* something, the head shake signal must (always and necessarily) be used.
- *implications*: it is a set of implication rules that allow us to conditionally constrain the presence of a signal of the *sigs* set depending on the presence of the other signals. For example combined signals do not always convey the same meaning as the meaning associated separately with each of the signals: The act of shaking the head can have a different meaning when associated with an angry or a happy face. Some arm gestures look less natural if they are performed without some rotation/movement of the torso. So, the *implications* rules are used to describe constraints on the possible combinations of signals in a behavior set.

3.3.2 Multimodal Signal Selection

The Multimodal Signal Selection (MSS) process takes as input the lexicon of nonverbal behavior sets defined in the previous section. It also considers the Dynamicline of a given agent. The output of MSS is the multimodal behavior that best represents the current agent's communicative intention taking into account the agent's modality preference and the core and implication rules of the behavior sets.

The diagram in Fig. 4 shows the process of Multimodal Signal Selection. The algorithm considers three main elements: (i) the set of modalities which are currently in use by the agent to convey other communicative intention(s); (ii) the agent's current communicative intention; (iii) the agent's

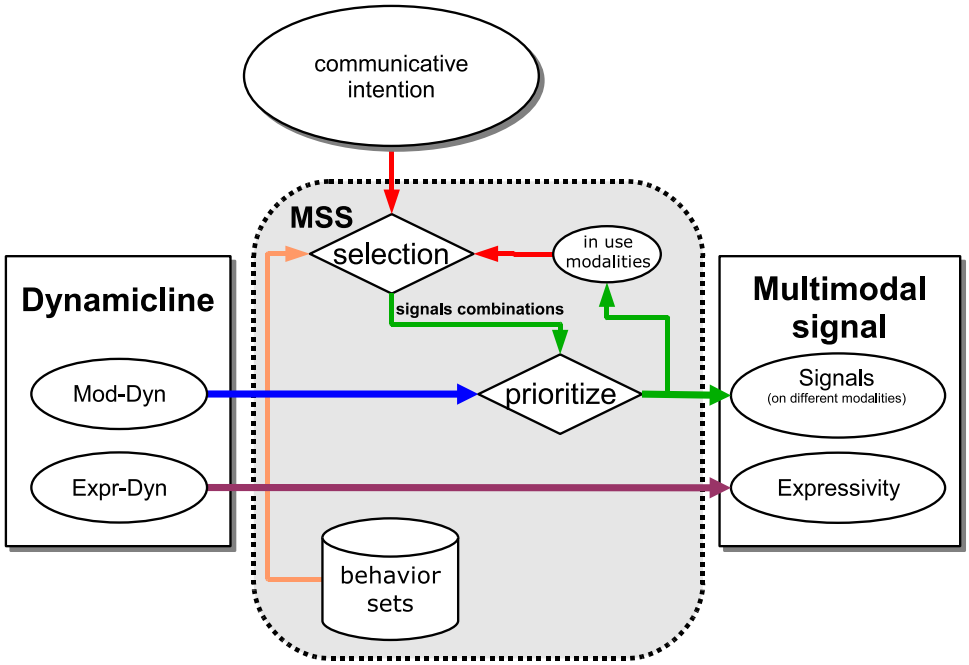

Fig. 4 The Multimodel Signal Selection (MSS) process. By taking as input the agent's Dynamicline and communicative intention we compute which signals the agent has to perform

local behavior tendency, that is represented by the agent's Dynamicline, as previously explained.

For example, let us suppose that the agent's current communicative intention is to *deny* what its interlocutor is saying. Let us also suppose that in the MSS module there is the following behavior set defined for the "deny" communicative intention:

$$(deny, S, C, I); \tag{3}$$

where:

- $S =$ ("*head shake*", "*torso forward*", "*wavefinger gesture*", "*frown eyebrows*");
- $C =$ ();
- $I =$ ("*torso forward*" $\rightarrow$ "*frown eyebrows*");

First of all, the algorithm must ensure that the multimodal signal the agent is going to produce will not create a conflict on the modalities which are already used by the agent to convey other meanings. In the above example suppose that our agent is already performing a head movement to communicate another meaning while it aims to *deny*. This means that in choosing which multimodal signal the agent has to produce the algorithm must exclude all the signals involving the use of the head modality. That is, it must choose a combination between the signals "frown eyebrows", "torso forward" and "wavefinger gesture". Moreover, the implication set of rules I impose that the signal "torso forward" can not be performed without the "frown eyebrows" signal. The result is that the algorithm can choose to perform one among the following multimodal signals:

- "*wavefinger gesture*";
- "*frown eyebrows*";
- "*torso forward*" plus "*frown eyebrows*";
- "*wavefinger gesture*" plus "*frown eyebrows*";
- "*wavefinger gesture*" plus "*torso forward*" plus "*frown eyebrows*";

At this point, the algorithm considers the agent's Dynamicline. In particular, it compares the agent's *Overall Activation* (OAC) parameter for the different modalities. The higher this parameter is for a given modality, the more the agent tends to be active on that modality, and vice-versa. So the algorithm filters out those multimodal signals which involve the use of modalities whose OAC parameter does not reach a certain threshold. In the above example, let us say that the agent's OAC on the torso modality if lower than this threshold. The set of possible multimodal signals is further reduced to:

- "*wavefinger gesture*";
- "*frown eyebrows*";
- "*wavefinger gesture*" plus "*frown eyebrows*";

The final step of our algorithm considers the agent's modality preference contained in the agent's Dynamicline, provided as input to the algorithm. Basically the algorithm assigns a score to each possible multimodal signal depending of the agent's preference for each modality involved in the signal production. The "wavefinger gesture" is assigned the preference of the gesture modality, the "frown eyebrows" is assigned the preference of the face modality and the "wavefinger gesture" plus "frown eyebrows" is assigned the

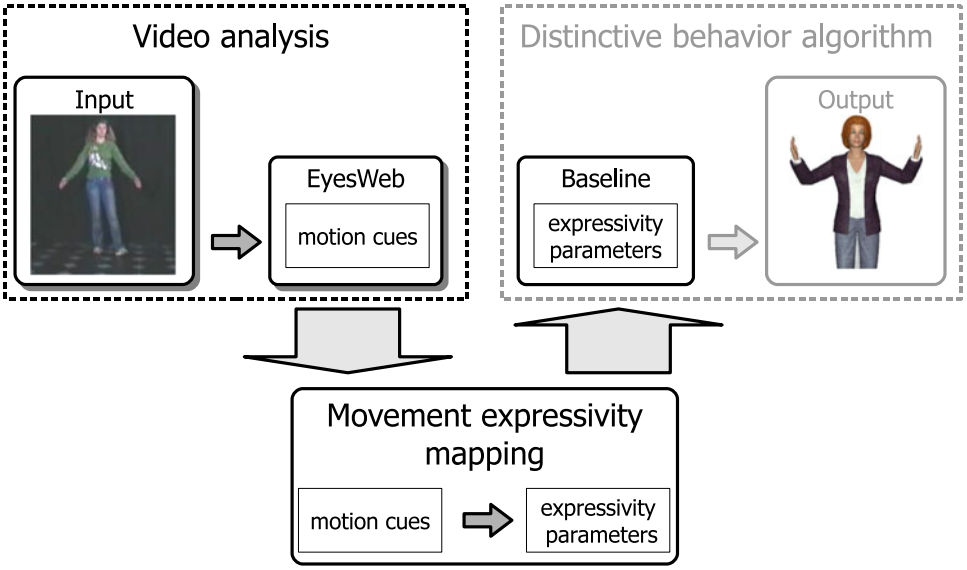

Fig. 5 Overview of the automatic expressivity extraction process. Input video is analyzed in order to extract some motion cues that are mapped into the agent's Baseline expressive parameters

highest of the gesture and face modality preferences. Then the algorithm selects the signal with the highest score.

In Sect. 3.1.1 we have introduced the *importance* attribute of a communicative intention. At this step of the MSS process, if more than one signals have the same score the algorithm performs a selection based on the importance attribute of the agent's current communicative intention: the higher is the importance attribute the higher is the number of modalities involved in the multimodal signal production. In the above example, let us suppose that both the "wavefinger gesture" and the "wavefinger gesture" plus "frown eyebrows" have the same score. If the agent's intention to "deny" what the user says has a low importance than only the "wavefinger gesture" is chosen; if the importance to "deny" is high than the "wavefinger gesture" plus "frown eyebrows" signal is selected.

4 Extracting Baseline expressivity automatically

In this section we present how the expressive information of a person's Baseline can be extracted automatically using video analysis technique.[2] We have applied such a technique to analyze and extract automatically the expressivity parameters and modalities hierarchy.

The analysis technique we have applied, does not use either invasive or very expensive equipment. It uses common hardware such as normal desktop or laptop computer equipped with a camera or webcam. It is based on EyesWeb system (http://www.eyesweb.org) [8] for video tracking and analysis of human movement.

4.1 Process description

Figure 5 shows the process of extracting expressivity automatically from a person's behavior. First the person's movements are analyzed in order to extract some *motion cues* related to the extraction algorithms reported in [45]. Then they are mapped onto expressivity parameters and finally these parameters are copied into the expressivity part of the agent's Baseline that is used to compute the agent's behavior, as we explained in the previous sections. Let us illustrate the two main steps of the process more in detail:

- *Video analysis*: we extract the user body silhouette and the hand position using EyesWeb XMI and the Expressive Gesture Processing Library [8]. The video analysis is performed on 2D video data, so the computation results illustrated in the following are accurate only if movements occur on the plane parallel to the image plane. We compute the following motion cues:
  - *Contraction Index—CI*: it measures, the degree of contraction and expansion of the body. The algorithm compares the area covered by the minimum rectangle surrounding the body with the area currently covered by the silhouette. If the body is contracted and the limbs are attached to the body the CI is high, whereas if the limbs are fully stretched the CI is low.
  - *Velocity—VEL*: given the coordinates in a 2D plane of sampled points in a motion trajectory (here the coordinates of the user's right or left hand's barycenter) we compute the first derivatives dx and dy of the 2D coordinates and velocity is equal to: $\sqrt{dx^2 + dy^2}$.

[2]The work presented in this section has been realized in collaboration with G. Castellano of Queen Mary University of London (previously at InfoMus Lab, DIST, University of Genova, Italy). It has been supported by the EU funded Human-Machine Interaction Network on Emotion Network of Excellence (http://emotion-research.net).

Fig. 6 Two examples of automatic expressivity extraction

- *Acceleration—ACC*: it is calculated in the same way as the Velocity. The first derivative is computed on the velocity values dx and dy and the absolute acceleration is equal to: $\sqrt{ddx^2 + ddy^2}$.
- *Directness—DI*: it is a measure of how much a trajectory is direct or flexible. It is computed as the ratio between the length of the shortest path between two points a and b and the length of the effective trajectory between the same two points. If trajectory is straight these two lengths are almost equal so Directness is almost one; if trajectory is very complex its length is much higher than the one of the shortest path, so Directness tends to approximate zero.

– *Movement expressivity mapping*: we defined the following correspondence between the automatically extracted motion cues and agent's Baseline expressivity parameters: Contraction Index is mapped onto Spatial Extent, since these provide a measure of the amplitude of movements; Velocity onto Temporal Extent, as these refer to the velocity of movements; Acceleration onto Power, as both are indicators of the acceleration of the movements; Directness onto Fluidity, as these refer to the degree of the smoothness of movements.

Figure 6 illustrates 2 examples of the automatic expressivity extraction process: in A we extract the expressivity of a low aroused person so the agent performs a gesture with small amplitude; in B the person movements are larger and the agent's gestures are modified accordingly.

5 Evaluation of algorithm Step A

The aim of the study presented in this section is to evaluate the correctness of the distinctive behavior algorithm step A, described in Sect. 3.2. To do that, we must verify that subjects correctly perceive the agent's behavior variations induced by the Baseline. We conducted a subjective evaluation, in which subjects had to rate the characteristics of the behavior generated by our algorithm.

Table 1 The Baselines of the four different agents we used in our evaluation studies. We created such Baselines manually. For each we list: the level of activity of the head (number of facial expressions and head movements) and of the body (number of torso movements and hand/arm gestures); the expressivity of the head (amplitude/speed/energy/fluidity/repetitivity of facial expressions and head movements) and of the body (amplitude/speed/energy/fluidity/repetitivity of torso movements and hand/arm gestures)

Baseline n.	Head activity	Head expressivity	Body activity	Body expressivity
1	high	medium	low	medium
2	low	medium	high	medium
3	low	low	low	low
4	high	high	high	high

5.1 Setup

We manually defined the Baselines of four different agents, representing the agents' general behavior tendency. We summarize their characteristics in Table 1.

We split the body into two regions: the *head* region, including head movements, gaze and facial expressions; the *body* region, including torso movements and hand gestures. The four Baselines have been set up to obtain four different behavior tendencies: an agent very active in the head region, very inactive with the body and with medium expressivity; an agent very active in the body region, very inactive with the head and with medium expressivity; an agent very static and inexpressive; a very active and expressive agent.

5.2 Participants

A total of 75 Italian participants (29 women and 46 men, aged between 19 and 60) took part in the evaluation study. Their background experience with computers varied from those who use their computers solely to surf the web (the great majority of the subjects) to experienced computer science students. None of them had ever seen or interacted with a conversational agent before performing the test.

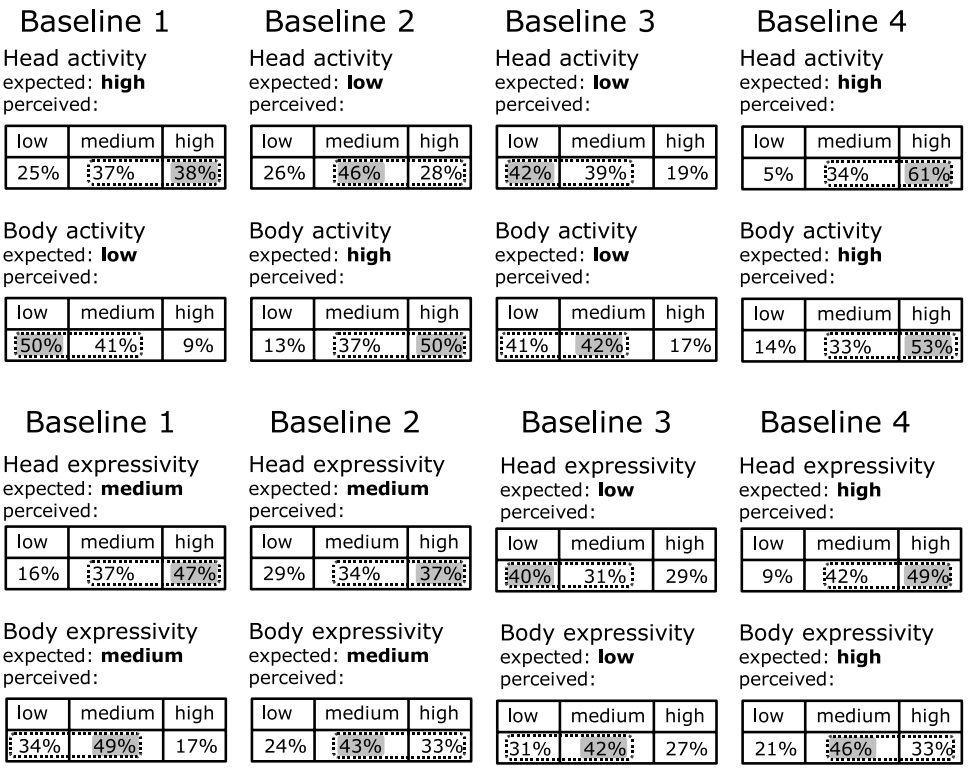

Fig. 7 Evaluation of the head and body region *activity* and *expressivity*. Results refer to the agent's four different Baselines. We highlight the prevailing answer (*grey box*) and the next two prevailing answers (*dashed box*)

5.3 Procedure

Each participant was instructed by email to reach a certain url with his/her browser. Once the website had been reached, participants could read a document explaining the experience in Italian. We asked participants to carefully watch each animation once without interruptions. At the end of the playback they could quantify the agent's head and body activity and expressivity by choosing values between 0 and 4 corresponding to the activity and expressivity they perceived in the video. We told participants to take all the time they needed to think before answering, but we asked them never to use the *back* button of their browser to return to a previous video.

5.4 Results and discussion

The tables in Fig. 7 report the evaluation of the *activity* and *expressivity* for the head and body region. During the test, participants could evaluate each parameter by choosing a value between 0 and 4. The results are groups by the number of users who selected values 0 and 1 under the label *low*, the number of users who selected value 2 under the label *medium* and the number of users who selected values 3 and 4 under the label *high*.

Results show that participants did not unambiguously distinguish between the terms *activity* and *expressivity*. In general, when the expected values for the two parameters were identical (e.g., high activity and high expressivity), participants recognized them as expected. On the other hand, when these two parameters had different expected values (e.g., low activity and medium expressivity), participants had difficulties in recognizing them. In the presentation of the evaluation tests, we gave to participants only a brief definition of the concepts "activity" and "expressivity". In everyday language, these terms can be confusing. When looking at an agent producing many head movements, one could hesitate between judging the head as "very activated" or "very expressive". Or, when viewing an agent with medium body expressivity and low body activity, participants evaluated it as having low-medium expressivity, since the low activity value influenced the perception of the expressivity parameter. We observed this sort of "co-variation" in the perception of the activity and expressivity parameters both in the head and the body regions.

6 Conclusion

In this paper we have described an algorithm for generating distinctive behavior for Embodied Conversational Agents, i.e. agents we can differentiate from their communicative behavior tendency: we have introduced the concept of Baseline to encompass this general behavior tendency. The Baseline is then modulated by the agent's communicative intentions

and emotional states and the result is the agent's local behavior tendency, modeled by the agent's Dynamicline. The Dynamicline influences the agent's behavior at two levels: the selection of multimodal signals to display as well as the specification of the behavior execution quality. As a possible method for defining the agent's Baseline we have described a system that automatically extracts the movement expressivity of a human user.

We conducted an evaluation study to verify whether the first step of our algorithm worked correctly. We asked participants to watch several examples of an agent and to evaluate the behavior characteristics for each of the examples. The results of this study show that distinctiveness in the agent's behavior is perceived by subjects. The second step of our algorithm, that is, the choice of the signals the agent has to produce is not yet evaluated. To overcome some of the limitations of our work, we propose several directions to pursue. At first we aim to continue working on extracting automatically not only the agent's Baseline but also behavior qualifiers from corpora of real data. Regarding behavior sets, the rules by which we select the multimodal signals the agent has to produce, should be extended to include more sophisticated constraint types. For example they could include temporal information: if the communicative intention X is displayed over a "short" time span, then use the behavior set $BS1$; if the communicative intention X has a "long" duration, then use the behavior set $BS2$.

Acknowledgements The presented work has been mostly realized in the framework of the EU-IST Project *HUMAINE* (Human–Machine Interaction Network on Emotion), a Network of Excellence (NoE) in the EU 6th Framework Programme (2004–2007).

References

1. Allbeck J, Badler N (2003) Representing and parameterizing agent behaviors. In: Prendinger H, Ishizuka M (eds) Life-like characters: tools, affective functions and applications. Springer, Berlin
2. Allwood J (2002) Bodily communication - dimensions of expression and content. In: Granstrom B, House D, Karlsson I (eds) Multimodality in language and speech systems. Kluwer, Dordrecht, pp 7–26
3. Argyle M (1988) Bodily communication, 2nd edn. Methuen & Co, London
4. Badler N, Phillips C, Webber B (1993) Simulating humans: computer graphics animation and control. Oxford University Press, Oxford
5. Ball G, Breese J (2000) Emotion and personality in a conversational agent. In: Cassell J, Sullivan J, Prevost S, Churchill E (eds) Embodied conversational characters. MIT Press, Cambridge
6. Byun M, Badler N (2002) Facemote: qualitative parametric modifiers for facial animations. In: Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on computer animation. ACM, New York, p 71
7. Caldognetto EM, Poggi I (2000) Dall'analisi della multimodalità quotidiana alla costruzione di facce parlanti. In: IV Conference of GFS (Gruppo di Fonetica Sperimentale). Padova, Italy
8. Camurri A, Mazzarino B, Volpe G (2004) Analysis of expressive gesture: The eyesweb expressive gesture processing library. In: Camurri A, Volpe G (eds) Gesture-based communication in human-computer interaction. LNAI, vol 2915. Springer, Berlin
9. Cassell J, Nakano YI, Bickmore TW, Sidner CL, Rich C (2001) Annotating and generating posture from discourse structure in embodied conversational agents. In: Workshop on representating, annotating, and evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents. Montreal, Quebec
10. Cassell J, Sullivan J, Prevost S, Churchill E (2000) Embodied conversational agents. MIT Press, Cambridge
11. DeCarolis B, Pelachaud C, Poggi I (2000) Verbal and nonverbal discourse planning. In: Workshop on achieving human-like behaviors. Autonomous Agents
12. DeCarolis B, Pelachaud C, Poggi I, Steedman M (2004) APML, a mark-up language for believable behavior generation. In: Prendinger H, Ishizuka M (eds) Life-like characters, cognitive technologies. Springer, Berlin, pp 65–86
13. DiPaola S, Arya A (2004) Affective communication remapping in musicface system. In: Electronic imaging & visual arts
14. Egges A, Kshirsagar S, Magnenat-Thalmann N (2003) A model for personality and emotion simulation. In: Knowledge-based intelligent information and engineering systems. Springer, Berlin, pp 453–461
15. Ekman P, Friesen W, O'Sullivan M, Scherer K (1980) Relative importance of face, body, and speech in judgments of personality and affect. J Pers Soc Psychol 38:270–277
16. Gallaher PE (1992) Individual differences in nonverbal behavior: Dimensions of style. J Pers Soc Psychol 63(1):133–145
17. Hartmann B, Mancini M, Buisine S, Pelachaud C (2005) Design and evaluation of expressive gesture synthesis for embodied conversational agents. In: Third international joint conference on autonomous agents & multi-agent systems. Utretch
18. Kallmann M, Marsella S (2005) Hierarchical motion controllers for real-time autonomous virtual humans. In: Intelligent virtual agents. Springer, Berlin, pp 253–265
19. Kendon A (2004) Gesture: visible action as utterance. Cambridge University Press, Cambridge
20. Kipp M (2006) Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In: Intelligent virtual agents, pp 230–242
21. Kipp M, Neff M, Kipp KH, Albrecht I (2007) Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In: Proceedings of the 7th international conference on intelligent virtual agents. Lecture notes in computer science, vol 4722. Springer, Berlin, pp 15–28
22. Kopp S, Wachsmuth I (2002) Model-based animation of coverbal gesture. In: Proceedings of computer animation, pp 252–257
23. Laban R (1971) The mastery of movement Plays. Inc, Boston
24. Lebourque T, Gibet S (1999) High level specification and control of communication gestures: the GESSYCA system. In: Proceedings of computer animation, vol 99
25. Lester J, Voerman J, Towns S, Callaway C (1997) Cosmo: A life-like animated pedagogical agent with deictic believability, pp 61–69
26. Mancini M, Pelachaud C (2007) Dynamic behavior qualifiers for conversational agents. In: Lecture notes in computer science, vol 4722. Springer, Berlin, p 112
27. Mancini M, Pelachaud C (2008) Distinctiveness in multimodal behaviors. In: Proceedings of the 7th international joint conference on autonomous agents and multiagent systems, pp 159–166
28. McNeill D (1992) Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago
29. Neff M, Fiume E (2004) Artistically based computer generation of expressive motion. In: Proceedings of the AISB symposium on language, speech and gesture for expressive characters, pp 29–39

30. Neff M, Fiume E (2005) AER: Aesthetic exploration and refinement for expressive character animation. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on computer animation. ACM Press, New York, pp 161–170
31. Neff M, Kim Y (2009) Interactive editing of motion style using drives and correlations. In: SCA '09: Proceedings of the 2009 ACM SIGGRAPH/Eurographics symposium on computer animation. ACM, New York, pp 103–112
32. Neff M, Kipp M, Albrecht I, Seidel H (2008) Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Trans Graph 27(1)
33. Niewiadomski R, Pelachaud C (2007) Intelligent expressions of emotions. In: Affective computing and intelligent interaction. Lecture Notes in Computer Science, vol 4738. Springer, Berlin, pp 12–23
34. Pelachaud C (2005) Multimodal expressive embodied conversational agents. In: MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on multimedia. ACM Press, New York, pp 683–689
35. Poggi I (2007) Mind, hands, face and body. A goal and belief view of multimodal communication. Weidler, Berlin
36. Poggi I, Pelachaud C (2000) Performative facial expressions in animated faces. In: Embodied conversational agents. MIT Press, Cambridge, pp 155–188
37. Poggi I, Pelachaud C, Caldognetto EM (2003) Gestural mind markers in ECAs. In: Proceedings of the second international joint conference on Autonomous agents and multiagent systems. ACM Press, New York, pp 1098–1099
38. Reeves B, Nass C (1996) The media equation: How people treat computers, television and new media like real people and places. CSLI Publications, Stanford
39. Ruttkay Z, Pelachaud C, Poggi I, Noot H (2008) Exercises of style for virtual humans. In: Canamero L, Aylett R (eds) Animating expressive characters for social interactions. Benjamins, Amsterdam
40. Scherer KR, Wallbott HG (1985) Analysis of nonverbal behavior. Handbook of discourse analysis, vol 2, pp 199–230
41. Schröder M, Pirker H, Lamolle M (2006) First suggestions for an emotion annotation and representation language. In: Devillers, L, Martin, JC, Cowie, R, Douglas-Cowie, E, Batliner, A (eds) Proceedings of the international conference on language resources and evaluation: workshop on corpora for research on emotion and affect, pp 88–92. Genova, Italy
42. Vilhjalmsson H, Cantelmo N, Cassell J, Chafai NE, Kipp M, Kopp S, Mancini M, Marsella S, Marshall AN, Pelachaud C, Ruttkay Z, Thorisson KR, van Welbergen H, van der Werf R (2007) The behavior markup language: Recent developments and challenges. In: 7th international conference on intelligent virtual agents
43. Wallbott HG (1998) Bodily expression of emotion. Eur J Soc Psychol 28:879–896
44. Wallbott HG, Scherer KR (1986) Cues and channels in emotion recognition. J Pers Soc Psychol 51(4):690–699
45. Zhao L, Badler N (2005) Acquiring and validating motion qualities from live limb gestures. Graph Models 67(1):1–16