style has been evaluated by friends, and by self-evaluation. In a first study, many characteristics of behavior have been evaluated: tendency to use body, face, head, gestures; qualities of movement, like fast-slow, small-large, smooth-jerky, etc. The person's behavior tendency has been shown to be an innate individual characteristic that the author claims to be a personality trait. In the second study she investigated consistency of person's behavior across time and situations. Results have demonstrated this consistency: people that are quick in writing are also quick while eating; if a person produce wide gestures then she also walk with large steps; energy of movements is also an enduring characteristic, constant over time. Wallbott and Scherer [16] describe a study on actors' body movements during the expression of several emotions. A group of people judged the actors' behaviors and annotated them. In the study, authors have found that the way actors portrayed emotional states seemed being actor dependent, that is it depended on the actor's personal way of expressing those emotions. Some behavior characteristics seemed also independent from the portrayed emotion: for example the number of head movements and overall behavior activity. Some actors seemed to show some emotions better than others: some people were more capable to show happiness, just because their behavior style was similar to the one shown during an happy emotional state. Similar results have been proposed by Gross et al. [6]. They found that the capacity of people in expressing their emotions depends on, among other, dispositional expressivity of a person. Low-expressivity individuals tend to inhibit negative emotions, when high-expressivity individuals do not.

2.2 Communicative intention

In this paper we focus on how a conversational agent can convey a particular communicative intention or emotional state to the user by producing multimodal signals. According to Poggi[13], the intentions that humans aim to convey while communicating with others belong only to one of the following main classes: information on the speaker's mind and information on the world. In our work, we refer to *Mind Markers*, defined by Poggi [13], which constitute a taxonomy of the first class of intentions. They are the ones we aim to communicate with our agent.

3. RELATED WORK

Several researchers have addressed the problem of defining conversational agents that exhibit distinctive behaviors.

Michael Kipp presents a gesture animation system based on statistical models of human speakers gestures [8]. Videos of interviewed people have been manually annotated in terms of gestures types (iconic, deictic, etc. [11]), together with their frequency of occurrence and timing (that is the synchronization between the gesture stroke and the emphasized syllable of the occuring utterance). The statistics on the speaker's gestures are then used to model the agent's set of preferred gestures (the probabilities of their occurrence is computed from the annotated gesture frequency) and synchronization tendency (for example an agent can perform gesture strokes always synchronized with speech emphasis). In a more recent work [9], the agent's gestures selection can be human authored or automatically learned using machine learning algorithms on the basis of previously annotated scripts. In our work we mainly look at which modalities are used, and which are the qualities of movement of the produced signals. Kipp's approach instead aims to find the gesture types which characterize a person. His approach and ours are thus complementary, each one looking at a different aspect of the production of multimodal signals in conversation. Similarly to our work, M. Kipp does not model the possible causes of visible variations in behaviors.

Ruttkay et al. [14, 12] propose the idea of behavior style, defined in terms of when and how the ECA (Embodied Conversational Agent) uses certain gestures. Styles are implemented by selecting gestures from a *style dictionary* that defines both which gestures an agent has in his repertoire and its habits in using them. The style dictionaries are written in GESTYLE. This language specifies which modalities should be used to display non-verbal behaviors and is also used to annotate the text that the agent has to utter. Ball and Breese [4] have applied psychological theories to the creation of models that simulate personality, mood and emotion. They propose a model for individualization for virtual agents in which the final behavior is computed depending on the agent's actual emotional state and personality by choosing the most appropriate style. The PAR model of Allbeck et al. [1] offers a parameterization of actions. The actions that the agent is able to carry out are defined together with the conditions that need to be true in order to perform the actions. Conditions can refer to the state of other agents or objects in the agent's environment.

In André et al. [2] the agent's behavior depends both on a script that describes what the agent has to communicate to the user (for example how to do a reservation for a room in a hotel's website) and its personalized behavior. The last one includes idle movements like for example tapping with its foot while the user does nothing or jumping when the mouse passes over the agent's figure.

4. BASELINE AND DYNAMICLINE

In our model we want to capture the idea that people have tendencies that characterise globally their behavior, but these tendencies may change in situations or rise after some particular events. To encapsulate the global and local qualities we have introduced the concepts of Baseline and Dynamicline. They represent the agent's behavior tendency on different time span: while the Baseline is the overall definition of how the agent behaves in most situations, the Dynamicline is the *local* specification of the actual agent's behavior (for example during a given agent's emotional state).

In our model, Baseline and Dynamicline do not only differ by their meaning (global vs local behavior tendency) but also by the fact that the Baseline is an input parameter. The Baseline of each agent has to be defined manually before running the system. On the other hand, the Dynamicline is automatically computed by the system at runtime, depending on the agent's current communicative intention and/or emotional state.

We define the Baseline by the pair *(Mod,Expr)* where:

- *Mod*: this parameter represents the modalities preference. So, if an agent has the tendency to mainly use hand gestures during communication a high degree of preference is assigned to the *gesture* modality. But if the face is the main active modality, the face modality is set to a higher value. For every available modality (face, head movement, gesture, posture), we define a

value between 0 (no used) and 1 (used a lot) which represents its preferability. An agent can also use two or more modalities with the same degree of preference: in this case then the agent will communicate with these modalities equally.

- *Expr*: this is a set of values that represents the base behavior tendency of the agent. Starting from results reported in [5], we have defined and implemented [7] a set of parameters that affects the qualities of the agent's behavior such as its speed (TMP parameter), spatial volume (SPC parameter), energy (POW parameter), fluidity (FLD parameter), and repetitivity (REP parameter). This set of parameters enables to differentiate an agent gesturing slowly and smoothly from an agent moving in a fast and jerky manner. A set of expressivity parameters can be specified for each modality separately.

Let us now see how the Dynamicline is computed at runtime (see Figure 1). The data provided as input is: the agent's Baseline and agent's communicative intentions or emotional states (i.e. what the agent wishes to communicate). For each new communicative intention and/or emotional state the system computes a new Dynamicline for the agent.

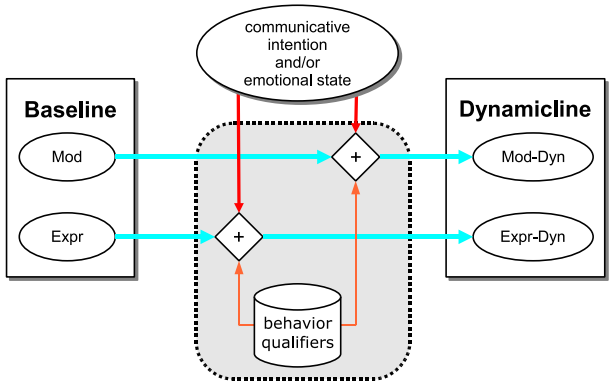

Figure 1: The agent's Baseline, communicative intention and/or emotional state determine the agent's Dynamicline

Dynamicline is modeled by the pair *(Mod-Dyn,Expr-Dyn)* where:

- *Mod-Dyn*: this is the agent's current modalities preferences. It is obtained by modulating the modalities preference *Mod* of the Baseline depending on the actual communicative intention and/or emotional state.

- *Expr-Dyn*: this the agent's current expressivity parameters. It is obtained by modulating for each modality the expressivity parameters *Expr* of the Baseline depending on the actual communicative intention and/or emotional state.

In previous work reported in [10] we have described the process of computing the agent's Dynamicline starting from the Baseline. In this paper we turn our attention to the selection of multimodal behavior to convey a given communicative intention/emotional state based on the current representation of the agent.

5. BEHAVIOR SET REPRESENTATION

We are interested in multimodal signals, that is signals produced on different modalities at the same time. The produced multimodal signals act as a whole to convey a given communicative intention. We define a multimodal signal *mms* as a set of mono-modal signals on different modalities:

$$mms = \{s_1, s_2, ..., s_n\} \qquad \bigcap_{i=0}^{i=n} s_i.modality = \emptyset \qquad (1)$$

where $s1, ..., s_n$ are signals produced on single modalities and $s_i.modality$ is the modality on which the signal s_i is produced.

In our system, communicative intentions and emotional states are associated with multimodal signals. Each of these associations represents one *entry* of a lexicon, called *behavior set*.

```
01 <behavior-set name="deny">
02   <signals>
03     <signal id="s1" name="shake" modality="head"/>
04     <signal id="s2" name="small_ahead"
05     modality="torso"/>
06     <signal id="s3" name="no" modality="gesture"/>
07     <signal id="s4" name="frown" modality="face"/>
08   </signals>
09   <constraints>
10     <core>
11       <item id="s3"/>
12     </core>
13     <rules>
14       <implication>
15         <ifpresent id="s2"/>
16         <thenpresent id="s4"/>
17       </implication>
18       <implication>
19         <ifpresent id="s1"/>
20         <thennotpresent id="s4"/>
21       </implication>
22     </rules>
23   </constraints>
24 </behavior-set>
```

Figure 2: Behavior set example.

The definition of a behavior set BS is a quadruple:

$$BS = (name, Sigs, Core, Implications); \qquad (2)$$

defined by:

- *name*: this is the name of the communicative intention associated to the behavior set; this parameter allows one to build the one-to-one correspondence between the behavior set and the communicative intention or emotional state. For example, the behavior set in Figure 2 is automatically used when selecting multimodal signals for communicating the intention of denying something. This association is defined by the `name` attribute of the root tag in line 01.

- *Sigs*: this is a set of signals produced 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. In the behavior set defined in Figure 2 the meaning *deny* is conveyed through one or a combination of the signals listed in the *signals* tag in lines 02-08 (which is the set *Sigs* defined above). In our example the given meaning can be conveyed with a combination of: shaking the head, producing a *no* gesture (index finger stretched up with the hand moving horizontally from left to right and vice-versa), moving torso, frowning. The *Sigs* set does not precise *how* and *if* these signals can be combined. The next two parameters will specify this information.

- *Core* and *Implications*: the first one is a subset of *Sigs*, representing those signals which have to appear in the multimodal signals communicating the given intention or emotional state; the second is a set of implication rules that allows one to conditionally constraint the presence of a signal of the *Sigs* set depending on the presence of the other signals. We will give extended definitions and examples about these two elements in the next two subsections, respectively.

The lexicon syntax has been defined with an XML Schema Definition (XSD). We have introduced validation rules that are applied as the lexicon is parsed and loaded in memory. The system is extensible. We can add new modalities (e.g. legs) by simply editing the behavior sets.

Core signals

When communicating a particular intention or emotional state some signals may have to be used. Among the possible multimodal signals communicating a given meaning there is a core subset of signals composing them. We have given the definition of an entry of our behavior sets: we specify which are the signals by which the multimodal signal corresponding to the actual communicative intention (or emotional state) may be composed. With the *core signals* we impose the presence of one or more of these signals in the final selection. For example, we may aim to specify that in *denying* something, the *no* gesture must be used, as shown in the behavior set of Figure 2. Lines 10-12 specify that the signal with id $s3$ (the no gesture) must be used to communicate the denying communicative intention.

Implication rules

A Behavior set describes the signals and all their combinations involved in communicating a given meaning/emotional state. But some combinations of signals of a same behavior set are not possible. It may be due to physical or other constraints. Combined signals do not always conveyed the same meaning as the meaning associated separately to each signal. The act of shaking the head can have a different meaning when associated with an angry or a happy face. We have defined a language to describe constraints on the possible combination of signals in a behavior set. The set of implications we have implemented in our system is:

- *if A then B* : if the signal A is selected for conveying a certain intention, then the signal B must be selected.
- *if A then not B* : if the signal A is selected for conveying a certain intention, then the signal B must not be selected.
- *A iff B* : with this condition we impose the simultaneous presence of two (or more) signals: if one of the two is selected, the other one must be selected at the same time.

Let us illustrate this language through an example. The behavior set in Figure 2 specifies that for communicating the intention to deny something the agent can produce a *shake* signal on the head modality; we also do not want this signal be produced together with the *frown* signal on the face modality. This condition is represented in lines 18-21 of the behavior set. These lines model the implication rule: *if the signal identified with the id* $s1$ *is selected for conveying the deny intention, then the signal identified with the id* $s4$ *must not be selected.*

6. MULTIMODAL SIGNAL SELECTION

In this Section we describe how the system selects multimodal behaviors to convey a given comunicative intention/emotional state. The Multimodal Signal Selection (MSS) process takes as input the lexicon of nonverbal behavior sets defined in the previous Section. It also considers the Baseline and Dynamicline of a given agent. The output of MSS is the multimodal behavior that better represents the actual agent's communicative intention or emotional state taking into account the agent's modalities preference and the core and implication rules of the behavior sets.

The diagram in Figure 3 shows the process of Multimodal Signal Selection. The MSS process is composed of 6 sequential sub-steps. The output of a previous sub-step serves as input to the next sub-step. We provide as input to the MSS:

- *lexicon*: this is composed by the agent's behavior sets (described in the Section 5), associating communicative intentions and emotional states to multimodal signals.
- *communicative intention or emotional state*: this is the communicative intention or emotional state for which we want to determine the multimodal signal to be produced.
- *used modalities*: this is a list of modalities that are already in use to convey a given meaning (e.g. the hand may already be in movement as part of a previous communicative gesture for signals emission); they cannot be considered in the MSS process.
- *Dynamicline*: this is the agent's Dynamicline, modeling the actual agent's modalities preference and expressivity (see Section 4). By including the Dynamicline in the MSS process we are considering the agent's behavior tendency, introducing variability depending on the agent's definition.

Let us now describe each of the 6 sub-steps of the MSS process.

Steps 0, 1 and 2: Parse, Pick behavior set and Apply rules

The lexicon of behavior sets is parsed from the corresponding XML file. The lexicon structure is validated by the lexicon XSD and computation continues only if the syntax is correct. The read entries are stored in memory together with

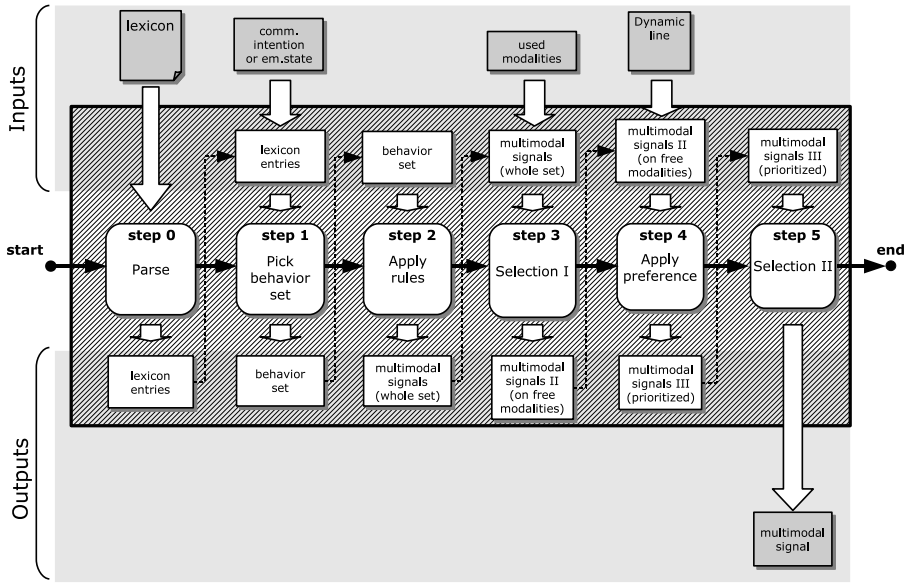

Figure 3: The Multimodel Signal Selection (MSS) process. Computation goes from left to right and is splitted in 6 substeps. Input data for each step is reported in the upper part of the diagram. Output data is in the lower part.

their constraints (as defined in core and implication rules). Once all the entries are loaded, we pick the behavior set corresponding to the communicative intention or emotional state provided as input. We illustrate these steps with an example. The communicative function *deny* is provided as input to the MSS that picks the behavior set represented in Figure 2.

Lines 03-07 specify which signals can be used to describe the *deny* communicative intention: shaking the head ($s1$) and/or doing a small *ahead* movement with torso ($s2$) and/or performing a *no* gesture ($s3$: the up extended index finger moving back and forth from left to right and viceversa) and/or *frowning* with the eyebrows ($s4$). Our system "expands" these mono-modal signals into their possible combinations to obtain multimodal signals. By composing the 4 signals of the example we obtain the following multimodal signals:

```
mms1 = (s1); mms2 = (s2); mms3 = (s3); mms4 = (s4);
mms5 = (s1,s2); mms6 = (s1,s3); mms7 = (s1,s4);
mms8 = (s2,s3); mms9 = (s2,s4); mms10 = (s3,s4);
mms11 = (s1,s2,s3); mms12 = (s1,s2,s4);
mms13 = (s1,s3,s4); mms14 = (s2,s3,s4);
mms15 = (s1,s2,s3,s4);
```

These signals are stored in a set, called $MMSign$:

$$MMSign = \{mms1, mms2, mms3, mms4, mms5, mms6, mms7, mms8, mms9, mms10, mms11, mms12, mms13, mms14, mms15\}$$

After this process of expansion, the sub-step *Apply rules is called* in which the *core* and *implication* rules of the behavior set are considered.

In the example, there is one signal which must always be present in the final selection, defined at lines 10-12. That is, the signal $s3$ has to be in the selected multimodal signal. This reduces the set of possible multimodal signals to:

$$MMSign' = \{mms3, mms6, mms8, mms10, mms11, mms13, mms14, mms15\}$$

Two implication rules are present at lines 13-22. The first one says that the presence of $s2$ implicates the presence of $s4$. *Deny* is displayed by a slight forward movement of the torso while the face should also show a frown: these two signals, torso movement and frown, work in conjunction for this given intention. That is, if we are denying by slightly moving torso ahead we can do this only in conjunction with an eyebrows frown. The second rule says that $s1$ implies the absence of $s4$, meaning that a head shake and a frown can not happen at the same time. By applying these rules to the actual set of multimodal signals we obtain the reduced set:

$$MMSign'' = \{mms6, mms10, mms11, mms14\}$$

This is the widest set of multimodal signals among which we can select the one that better conveys the intended meaning (deny). We will examine the rest of the selection process in the following subsections.

Step 3: Selection I

The input of this sub-step is the set of multimodal signals $MMSign''$ computed in the previous steps and the set of modalities which are actually in use by the agent to convey another communicative intention/emotion. In our example, suppose that the set of used modalities is:

$$Mod_{USED} = \{face\}$$

Since the *face* modality is currently used, multimodal signals in $MMSign''$ which use this modality ought to be eliminated as they can not potentially be selected to convey the current communicative intention/emotional state. In the behavior