Constraint Based Model for Synthesis of Multimodal Sequential Expressions of Emotions

Radosław Niewiadomski, Sylwia Julia Hyniewska and Catherine Pelachaud

Abstract—Emotional expressions play a very important role in the interaction between virtual agents and human users. In this paper, we present a new constraint-based approach to the generation of multimodal emotional displays. The displays generated with our method are not limited to the face, but are composed of different signals partially ordered in time and belonging to different modalities. We also describe the evaluation of the main features of our approach. We examine the role of multimodality, sequentiality and constraints in the perception of synthesized emotional states. The results of our evaluation show that applying our algorithm improves the communication of a large spectrum of emotional states, while the believability of the agent animations increases with the use of constraints over the multimodal signals.

Index Terms—H.5.2.f Graphical user interfaces, H.5.1.b Artificial, augmented, and virtual realities

1 Introduction

Virtual agents are used as partners in human-computer interactions. As such they need to be endowed with communicative capabilities. In particular they ought to be able to convey their emotional states. Displaying few expressions, typically the six "basic" expressions [1], is not enough. The agents appear too stiff and repetitive with a very limited repertoire. Moreover these six expressions may not always be the most adequate in a human-computer interaction. Emotional states such as satisfaction, frustration, annoyance or confusion may be more relevant (e.g. [2], [3]). Thus, virtual agents ought to be endowed with a large palette of expressions allowing them to display subtle and varied expressions.

In this paper we propose a new approach to the generation of emotional expressions. It allows a virtual agent to display *multimodal sequential expressions* **(MSE)** i.e. expressions that are composed of different nonverbal behaviors (called in this paper *signals*) partially ordered in time and belonging to different nonverbal communicative channels. Few models have been proposed so far for creating dynamical multimodal expressions in virtual agents (e.g. [4], [5], see also section 2). More often agents use only stereotypical facial displays which are defined at their apex and then interpolated. Instead our model generates a variety of multimodal emotional displays of an arbitrary duration. Each of them is composed of a sequence of nonverbal behaviors that are displayed not only by face but also with the use of other modalities like gaze, gesture, head and torso movements. With MSE the repetitiveness of the emotional expressions is avoided by introducing diversity in the signals choice, order and timing. This variability is obtained by probability of appearance and temporal constraints which are defined separately for each signal. In our model a high-level symbolic representation of the behavior emotional displays are generated from samples described in literature and from annotated videos. Thus, captured data is not directly reproduced, but different plausible expressions of emotions are generated. They are composed from the same signals as the original ones.

Our approach is coherent with recent research in psychology. It was shown that several emotions are expressed by a set of different nonverbal behaviors which include *different modalities*: facial expressions, head and gaze movements [6], gestures [7], torso movements and posture [8], [9]. Thus emotional expressions may be composed of several behaviors. Interestingly, these signals do not have to occur simultaneously. Dacher Keltner and colleagues (e.g. [7], [10]) showed that in the case of some emotions, like embarrassment, the signals occur in a sequence. The *sequentiality* of signals in emotional expressions is also postulated in Scherer's Component Process Model [11]. Signals in multimodal expressions do not occur by chance. In the embarrassment sequence [7] some temporal relations between the signals were observed that may be represented in the form of *constraints*.

In this paper we describe our model of multimodal sequential expressions. For this purpose we defined a representation scheme that encompasses the dy-

- *R. Niewiadomski works at Telecom ParisTech, France.*
- *S. Hyniewska works at Université de Geneve, FPSE, Switzerland and Telecom ParisTech, France.*
- *C. Pelachaud works at CNRS-Telecom ParisTech, France.*

namics of emotional displays and we called it *multimodal sequential expressions language (MSE-language)*. It ensures the description in a formal way of the configurations of signals as well as of the relations that occur between them. For a given emotional label our algorithm chooses a coherent set of multimodal signals and orders them in time. For this purpose we define two data structures for each emotional state: a *behavior set* contains signals through which the emotion is displayed while a *constraint set* defines the relations between the signals of each behavior set. The final animation is an ordered sequence of signals in time i.e. a subset of signals from the behavior set with their durations, which is consistent with all the constraints of the corresponding constraint set.

In the second part of the paper we present the results of an evaluation of three main features of MSE: *multimodality*, *sequentiality*, and *constraints*. First of all, we check if the agent that uses the MSE algorithm is able to communicate its emotional states properly, i.e. if its multimodal sequential expressions are recognized by humans. We also examine whether using multimodality and sequentiality influences the recognition rate. Finally, we verify the importance of the constraints in the perception of believability of the agent's behavior.

The remaining part of this paper is structured as follows. The next section describes related research of multimodal emotional displays. Section 3 is dedicated to an overview of computational models of multimodal and/or sequential expressions. In section 4 our algorithm is explained while in section 5 the results of evaluations studies of MSE are presented. Section 6 concludes the paper.

2 Human emotional expressions

An emotion is a dynamical episode that produces a sequence of response patterns on the level of body movement, posture, voice and face [11]. Although the face is often the major focus of attention, the changes in the other modalities are more than complementary to the facial expression. Thus, not only body movements have an impact on the interpretation of the facial expression [12], but some of them seem also to be specific expressions of particular emotional states (e.g. [8], [9]). Several studies show also that emotional expressions are often composed of signals arranged in a sequence [7], [10], [11], [13]. Keltner, for example, showed that it is the temporal unfolding of the nonverbal behaviors that enables one to differentiate the expressions of embarrassment and amusement, which in some studies (e.g. [14]) tend to be confused by judges as they have a similar set of signals involving smiling, numerous sideway gaze and head shifts [7].

Expressions of several emotional states were analyzed, among others, amusement [6], [10], anxiety [15] awe [10], confusion [16], embarrassment [7] shame [6], [7] and worry [16]. They have explored the complexity of emotional expressions in terms of their dynamics and/or multimodality. Various multimodal signals were observed by Shiota and colleagues in expressions of awe, amusement and pride [10]. They showed that these three emotions could be expressed by a set of possible signals, sometimes with asynchronous onsets, offsets and apices. Not all signals have to be present at the same time, for the expressions to be recognized as a display of a particular emotional state. In the expression of pride, for example, Shiota and colleagues [10] observe a mild smile with a contraction of the eyelids causing crow's feet (AUs 6 + 12) with lips pressed together (AU 24) and some straightening of the back. They note that pride can also be often accompanied by some pulling back of the shoulders to expose the chest and by a slight head lift. Also anxiety [15] is displayed by various signals as partial facial expressions of fear (i.e. the expression of fear by the open mouth), mouth stretching movements, eyes blinking, and non-Duchenne smiles (AU12 without AU6 that is the smile without crow's feet).

Another study goes further, describing also the sequences of multimodal signals. Keltner [7] studied the expressions of embarrassment by analyzing the appearance frequencies of signals in audio-visual data. The typical expression of embarrassment starts from a downward gaze or gaze shifts which are followed by "controlled" smiles, i.e. smiles accompanied by pressed lips. At the end of the expression a movement of the head to the left was often observed, as well as some face touching gestures [7].

Summing up, several emotions are expressed by a set of different nonverbal behaviors, all relying on the use of more than one modality ([6], [7], [8], [9], [13]), such as facial expressions, head and gaze movements, hand and arm position changes, torso movements and posture. We try to model these expressions with our algorithm.

3 Synthesized emotional expressions

Two different approaches are usually used to create emotional expressions in virtual agents: the motion capture-based and the procedural one. The first one is often used in commercial applications, e.g. in the movie industry. In this approach the synthesized expressions are characterized by a very high level of details and a great realism. This approach is however very time and resource consuming. It may also lack some flexibility and variability - two important issues in agent's behavior synthesis.

In the second approach, an emotional display is generated from a symbolic description. This description is used to define the key-frames of the animations, which are then generated using an interpolation. Usually a facial expression is presented in its apex (maximal intensity moment is defined as a key-frame), while the animation is interpolated for the rest

of frames. In this approach animations can be of any arbitrary duration, but the generated animations are schematic and stereotypical. It is difficult to generate animations of subtle emotional states. Often only six facial expressions in their apex, which were described in [1], are implemented in virtual agents.

A few models have been proposed recently for creating dynamical facial expressions. Ruttkay [17] proposed a system that permits, for any single facial parameter, to define manually the course of the facial animation. The plausibility of the final animation is assured by a set of constraints that are defined on the key-points of the animation. Stoiber et al. [18] propose an interface for the generation of both realistic still images and fluent sequences of facial expressions. Using a 2D custom control space the user may deform both the geometry and the texture of a facial model. The approach is based on the principal component analysis of the database containing a variety of facial expressions of one subject.

Other researchers were inspired by Scherer's Component Process Model [11], which states that different cognitive evaluations lead to specific micro-expressions. Paleari and Lisetti [19] and Malatesta et al. [20] focus on the temporal relations between different facial actions predicted by this theory. In [19] the different facial parameters are activated at different moments and the final animation is a sequence of several micro-expressions while in [20] the expression is derived from the addition of a new AU to the former ones.

Clavel et al. [21] found that the integration of the facial and postural changes affects users' perception of basic emotions. In particular an improvement of the emotion recognition was observed when facial and postural changes are congruent [21]. Nevertheless only some models for multimodal emotional expressions have been created so far. Lance and Marsella [5] model head and body movements in emotional displays using the PAD dimensional model. A set of parameters describing how the multimodal emotional displays differ from the neutral ones was extracted from the recordings of acted emotional displays. Consequently, emotionally neutral displays of head and body movements are transformed to multimodal displays expressing e.g. low/high dominance and arousal. Pan et al. [4] proposed an approach to display emotions by sequences of signals (facial expressions and head movements). From real data, they built a motion graph in which the arcs are the observed sequences of signals and the nodes are possible transitions between them. New animations are generated by reordering the observed displays. Mana and Pianesi [22] use Hidden Markov Models to model the dynamics of emotional expressions during speech acts.

In comparison to the solutions presented above our system generates a variety of multimodal emotional expressions automatically. It is based on a high-level symbolic description of nonverbal behaviors. It is built on observational data but contrary to many other approaches which use captured data for behavior reproduction, in this approach the observed behaviors are interpreted by a human (i.e. a FACS expert) who defines constraints. The sequences of nonverbal displays are independent behaviors that are not driven by the spoken text. The system allows for the synthesis of any number of emotional states and is not restricted by the number of modalities.

Our algorithm does not define an animation a priori as a set of key-frames but it dynamically generates a number of animations which satisfies a manually defined set of constraints. These constraints ensure the correct order of behaviors in the sequence. It generates a variety of animations for one emotional label avoiding the repetitiveness in the behavior of a virtual agent. On the other hand using procedural approaches it is difficult to generate different animations for subtle emotional states. Introducing the sequences of signals we aim at enlarging the set of emotional states that can be communicated by virtual agents.

Last but not least, while the algorithm uses a discrete approach in its use of labels to refer to emotions, it is also linked to the componential approach by the underlined importance of the sequence of signals.

4 Multimodal sequential expressions in virtual agents

The main task of our algorithm is to generate the multimodal sequential expressions of emotions, i.e. expressions that are composed of different signals partially ordered in time and which involve different nonverbal communicative channels. Our algorithm is based on the following criteria:

- the emotional displays are sequences of behaviors on different modalities,
- the animations are not predefined but are created dynamically,
- there is variability in the created animations,
- the sequences are built in real-time allowing the instantaneous execution of the animation,
- the sequences may have an arbitrary duration,
- the algorithm uses human-readable descriptions of behaviors and constraints.

In the following subsections, we present the details of our approach starting from the observation to the synthesis of emotional expressions with the virtual agent.

4.1 Data collection

We base our work on observational studies of human emotion [6], [7], [10], [16], as well as on the annotation realized in our laboratory on nonverbal behavior.

Videos from the EmoTV corpus [23], the Belfast Naturalistic Emotional Database [24] and the HUMAINE database [25] as well as some extracts from French TV live shows have been chosen in order to observe behavior expressed in highly emotional situations by non-actors. An annotation scheme was developed to describe low and high levels of information: from signals to emotional states. On the low level, the signal level, we are using FACS (Facial Action Coding System, [26]) developed by Paul Ekman and colleagues to describe visible facial muscular activity. The extracts have been annotated by a certified FACS coder, with two to six video extracts per state. For annotating other nonverbal behaviors such as hand, arm and torso movements, a free textual description was used. An emotional label was attributed in each extract, based on observed expression and the context, e.g. a woman describing the happiest day of her life and using vigorous movements was labeled as cheerful. Although only a very short extract (between 4 and 50 seconds) was annotated, limited strictly to the emotional expression, a longer part of the video clip was viewed to enable the comprehension of the context. A detailed description of our annotation can be found in [27].

4.2 MSE-language

To go beyond agents showing simply static facial expressions of emotion (i.e. expressions at their apex), at first, we gathered information on the signals (see sections 2 and 4.1) involved in the emotional expressions as well as on the temporal constraints regulating them. Consequently, we have designed a representation scheme that is based on these observational studies. It encompasses the dynamics of emotional behaviors by using a symbolic high level notation.

The issue of processing temporal knowledge and temporal reasoning found many solutions in the domain of artificial intelligence. In particular, James Allen [28] proposed a time interval based deduction technique based on constraint propagation. He proposed a set of time relations that can represent any relationship that holds between any two intervals. Its interval relation reasoner is able to infer consistent relations between events with some time constraints posed. This method was applied then to a classical planning problem [29]. More recent planning algorithms that deal with temporal knowledge such as TGP [30] allow for efficient plan construction from actions of different duration.

For the purpose of generating multimodal sequential expressions we define a new XML-based language in two steps: a *behavior set* and a *constraint set*. Single signals like a smile, shake or bow belong to one or more behavior sets. Each emotional state has its own behavior set, which contains signals that might be used by the agent to display that emotion. The relations that occur between the signals of one behavior set are more precisely described in the constraint sets. The appearance of each signal s_i in the animation is defined by two values: its start time, $start_{s_i}$ and its stop time $stop_{s_i}$. During the computation the constraints influence the choice of values $start_{s_i}$ and $stop_{s_i}$ for each signal to be displayed.

Comparing to the solution proposed in [28] we use "exists" operator that influences our inference algorithm. We also use less operators that are more suitable for the manual video annotation. Our operators are sufficient to describe relations between nonverbal behaviors. We also propose ad-hoc algorithm to infer on both temporal and interval duration relations.

4.2.1 Behavior set

The behavior set contains a set of signals of different modalities e.g. head nod, shaking-hand gesture or smile to be displayed by a virtual agent. Let us present an example of such a behavior set. In [7], a sequence of signals in the expression of embarrassment is described. The behavior set based on Keltner's description [7] of embarrassment (see section 2) may contain the ten signals:

- two head movements: *head down* and *head left*,
- three gaze directions: *look down*, *look right*, *look left*,
- three facial expressions: *smile*, *tensed smile*, and *neutral expression*,
- *open flat hand on mouth* gesture, and
- a *bow* torso movement.

A number of regularities occur in expressions that concern signals duration and their order of displaying. Consequently for each signal in the behavior set one may define the following five characteristics:

- *probability_start* and *probability_end* - probability of occurrence at the beginning (resp. towards the end) of a multimodal expression (a value in the interval $[0..1]$),
- *min_duration* and *max_duration* - minimum (resp. maximum) duration of the signal (in seconds),
- *repetitivity* - possibility that the signal might be repeated.

For instance, in the embarrassment example the signals *head down* and *gaze down* occur much more often at the beginning of the multimodal expression [7] than later. Thus their values of *probability_ start* are much higher than the value of *probability_end*.

4.2.2 Constraint set

The signals in multimodal expressions often occur in some relations like "two signals s_i and s_j occur contemporarily", or that "the signal s_i cannot start (end) the display" etc. Each emotional state can be characterized by a constraint set that describes reliable configurations of signals. This set introduces a set of limitations on the occurrence and on the duration (i.e. on the values for $start_{s_i}$ and $stop_{s_i}$) of the signal s_i

in relation to others signals. We introduced two types of constraints:

- *temporal constraints* define relations on the start time and end time of a signal using arithmetic relations: $<$, $>$ and $=$;
- *appearance constraints* describe more general relations between signals like inclusion or exclusion e.g. "signals s_i and s_j cannot co-occur" or "signal s_j cannot occur without signal s_i".

The constraints of both types are composed using the logical operators: and, or, not. The constraints take one or two arguments.

Three types of temporal constraints are used *morethan*, *lessthan*, and *equal*. These arithmetical relations may involve one or two signals: for example the observation: "signal s_i cannot start at the beginning of animation" will be expressed as following $start_{s_i} > 0$, while "signal s_i starts immediately after the signal s_j finishes" will be $start_{s_i} = stop_{s_j}$.

In addition, five types of appearance constraints were introduced for the more intuitive definition of relations between signals:

- exists(s_i) - is true if the s_i appears in the animation;
- includes(s_i, s_j) - is true if s_i starts before the signal s_j and ends after the s_j ends;
- excludes(s_i, s_j) - is true if s_i and s_j do not occur at the same time t_k i.e.: if $start_{s_i} < \mathrm{t}_k < stop_{s_i}$ then $stop_{s_j} < \mathrm{t}_k$ or $start_{s_j} > \mathrm{t}_k$ and if $start_{s_j} < \mathrm{t}_k < stop_{s_j}$ then $stop_{s_i} < \mathrm{t}_k$ or $start_{s_i} > \mathrm{t}_k$;
- precedes(s_i, s_j) - is true if s_i ends before s_j starts;
- rightincludes(s_i, s_j) is true if s_i starts before the signal s_j ends, but s_j ends before s_i ends.

For example, using two appearance constraints of the type *includes* and *exists* one may define that $signal_2$ occur only if $signal_1$ has started before and it will end before the ending of $signal_2$:

$$\text{exists(signal}_1\text{) and includes(signal}_1\text{, signal}_2\text{)}$$

During the computation of the animation, constraints are instantiated with the appearance times (i.e. $start_{s_i}$ and $stop_{s_i}$) of the signals. By convention, the constraints that *cannot* be *instantiated* (i.e. one of the arguments does not appear in the animation) are ignored. An animation is consistent if all *instantiated* constraints are satisfied.

4.3 Algorithm

Let A be the animation to be displayed by a virtual agent. A can be seen as a set of triples $A = \{(s_i, start_{s_i}, stop_{s_i})\}$, $start_{s_i}$, $stop_{s_i} \in [0..t]$, $start_{s_i} < stop_{s_i}$ where s_i is the name of the signal, $start_{s_i}$ is the start time of the signal s_i and $stop_{s_i}$ is its stop time.

At the beginning A is empty. In the first step the algorithm chooses the behavior set $BS_e = \{s_k\}$, the constraint set $CS_e = \{c_m\}$ corresponding to the emotional state e, and the number n of uniform intervals, time stamps, for which t is divided.

Next, at each time step, t_j, ($j = 0..n-1, t_n = t$), the algorithm randomly chooses a signal-candidate s_c from the signals of the behavior set BS_e. For this purpose it manages a table of probabilities that contains, for each signal s_k, its current probability value $\mathrm{p}_{k_{(tj)}}$. At the first time stamp, $t_0 = 0$, the values of this table are equal to the values of the variable *probability_start*, while at the last time stamp t_{n-1} the probabilities are equal to the *probability_ end*. At each time stamp, t_j, the probabilities $p_{k_{(tj)}}$ of each signal $s_k \in BS_e$ are updated. The candidate for a signal to be displayed s_c in a turn t_j is chosen using the values $p_{k_{(tj)}}$.

Next, the start time $start_c$ is chosen from the interval [t_j , t_{j+1}] and the consistence of CS_e with the partial animation $A(t_{j-1}) \cup (s_c, start_{s_c}, \text{ø})$ is checked.

If all the constraints are satisfied the stop time $stop_c$ is chosen in the interval defined by minimum and maximum duration of s_c. Otherwise, if not all constraints can be satisfied, another signal from BS_e is chosen as candidate. The consistency of the triple (s_c, $start_{s_c}$, $stop_{s_c}$) with the partial animation A(t_{j-1}) is checked again.

If all the constraints are satisfied the signal s_c starting at $start_{s_c}$ and ending at $stop_{s_c}$ is added to A. The table of probabilities is updated and the algorithm chooses another signal, moves to the next time stamp, or finishes generating the animation.

A $\leftarrow$ { ø}
choose BS_e, CS_e, and n
for $j = 0$ to $n - 1$ **do**
  choose $s_c \in BS_e$
  choose $start_c \in [t_j, t_{j+1}]$
  if $A(t_{j-1}) \cup (s_c, start_{s_c}, \text{ø})$ consistent **then**
    choose $stop_c \in$ [min-dur$_{s_c}$, max-dur$_{s_c}$]
    if $A(t_{j-1}) \cup (s_c, start_{s_c}, stop_{s_c})$ consistent **then**
      $A(t_j) \leftarrow A(t_{j-1}) \cup (s_c, start_{s_c}, stop_{s_c})$
      update $\mathrm{p}_{k_{(tj)}}$
    end if
  end if
end for

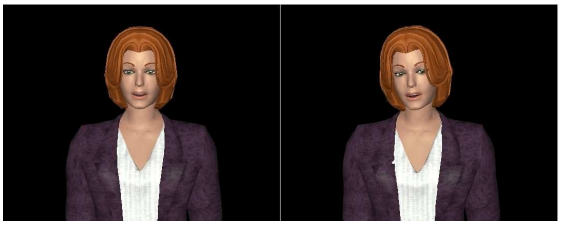
Main steps of MSE-algorithm.

In our approach we do not scale the timing of an observed sequence of behaviors to t. Rather the algorithm chooses between the available signals of a behavior set. The choice of our approach is motivated by research results showing that the duration of signals is related to their meaning. For example, spontaneous facial expressions of felt emotional states are usually not longer than four seconds, while the expression of surprise is much shorter [1], [31]. Similarly, gestures have also a minimum duration. Moreover the same gesture performed with different velocity might convey different meanings. It is worth to notice

that in each computational step the algorithm adds a new signal that starts not earlier than the previous one. Consequently a partial animation A(t_{j-1}) can be generated and displayed at t_j.

4.4 Examples of animations

The MSE-algorithm enables us to generate a number of animations of different duration that is consistent with the constraints. We obtain a variety of animations, each of which is consistent with the observation, but which go beyond the set of observed cases. In this way, we avoid the repetitiveness of the agent's behavior.



a *b*

Fig. 1. Multimodal expression of relief (SEQ1).

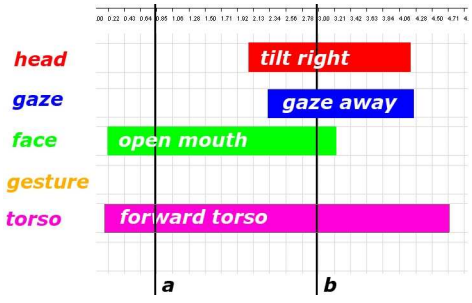


Fig. 2. Duration of signals in SEQ1.

We present here some examples of animations generated with the MSE algorithm. In first example we generate different relief sequences. The duration of the first animation is 4 seconds. In Figure 1a, relief is expressed by an open mouth and a forward torso movement, which is then accompanied on the second image (Figure 1b) by a head tilt. It is interesting to notice that these signals do not start and end simultaneously (see Figure 2).

Next we generate two longer expressions of relief. For a 10 second animation of relief, the algorithm generates a sequence of behaviors. Some signals that occurred in the expressions in Figure 1 are used again in longer animations (Figure 3 and 5), but they are accompanied by some new ones. Figures 4 and 6 illustrate the variability of animations that can be generated with the algorithm. Two sequences are composed of different signals chosen from the same *behavior set*. In Figure 3 the following behaviors are displayed: two head movements: *head up*, *till right*, two facial expressions: *open mouth* and *smile*, *upwards hands thrust* gesture, *gaze away* and *torso backward* movements. Then Figure 5 presents another 10 seconds sequence which is composed of: two facial expressions *open mouth* and *smile*, hands towards exterior gesture, torso backward, gaze away and head up movements.

5 Evaluation

We carried out two studies to validate our approach to the generation of emotional displays for a virtual agent. In the first study, we checked whether people are able to recognize the emotions expressed by the agent. Then, in the second study, we verified if the multimodal sequential expressions are recognized more than static images of emotional displays and dynamical single signal emotional expressions. In the same evaluation the role of constraints in the perception of multimodal sequential expressions was also checked.

For the purpose of these studies eight emotional states were chosen: anger (ANG), anxiety (ANX), cheerfulness (CHE), embarrassment (EMB), panic fear (PFE), pride (PRI), relief (REL) and tension (TEN). This arbitrary choice is motivated by the following:

- C1) we want to differentiate between several positive emotional states. Usually in literature all the positive emotions are described with the general label "joy" and are associated with the Duchenne smile [31]. In this study we evaluate: cheerfulness, pride and relief.
- C2) we want to differentiate expressions in which different types of smiles (Duchenne and non-Duchenne) might occur. Smiles are used to display positive emotions (e.g. in joy) but they also occur in negative expressions like embarrassment or anxiety.
- C3) we want also to differentiate negative states to be used by the virtual agent like anxiety, tension, panic fear and we want to compare them with the expression of anger.

The behavior and constraint sets for pride, embarrassment and anxiety were defined from the literature (see section 2). The sets of other 5 emotional states: anger, cheerfulness, panic fear, relief, tension were based on the annotation study [27] (see section 4.1).

5.1 Set-up

All evaluation studies have a similar setup. For the generation of animations the Greta agent [32] was used. Participants accessed the evaluation studies through a web browser. Each study session was made of a set of web pages, each page presenting one question a_i. The participants could not come back to the preceding question a_{i-1} and they could not jump to the question a_{i+1} without providing answer to the current one. No time constraint was put on the

Fig. 3. An example of the sequence of relief (SEQ2).

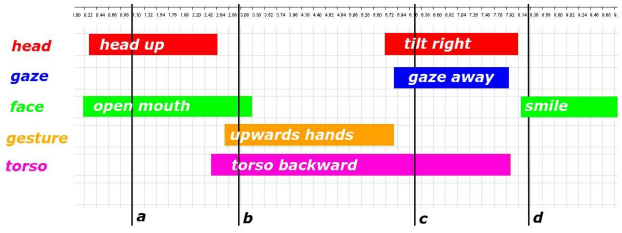


Fig. 4. Duration of signals in SEQ2.

task. The questions were displayed in a random order, the emotional labels were ordered alphabetically. The participation in the studies was anonymous.

5.2 Recognition of emotional states

First, we were interested in checking if the emotional states expressed with multimodal sequential expressions are recognized by the participants. For this purpose we show the participants a set of animations of the Greta agent displaying the eight emotional states and we asked them to attribute to each animation one emotional label.

In this study our hypotheses were the following:

- H1.1) each of the intended emotions is more often correctly recognized on the corresponding animation than chance level,
- H1.2) for each animation the proper label is attributed more often than any other label.

We were also interested in the habituation effect (H1.3) i.e. if showing the same set of animations more than once influences the recognition rate.

5.2.1 Procedure of the recognition study

Eight animations presenting different emotional displays were used in the study. Participants were asked to recognize the emotions displayed by the virtual agent. Each video shows the agent displaying one emotional state. The agent is not speaking. The duration of each video is about 10 seconds. After watching an animation the participants have to attribute one emotional label to the perceived emotional state from an 8-element list before they can pass to another page with a new animation. Participants were told that they could use each label more than once, or not at all.

Each study session consists of seeing twice the same set of eight videos presented in a random order. Each subject has to see all eight videos (turn 1) before seeing any of them for the second time (turn 2). They cannot replay the animation.

5.2.2 Results

Fifty three participants (25 women, 28 men) with a mean age of 28 years mainly from France (21%), Poland (21%) and Italy (15%) took part in the study. None of them works in the domain of virtual agents.

The attribution of correct answers (number of hits) (see Table 1) for each emotional expression in both turns is above chance level (which is 12,5%). In each turn, the greatest amount of hits was for the emotion of Anger (93% both turns mean) while the least correctly attributed was Embarrassment (41% both turns mean). The number of hits vs. alternative answers in turn 1 and turn 2 was compared and the improvement was not significative (univariate ANOVA, $p>.05$). Therefore, although the analyses for each of the two turns have been realized, the means for both turns are stated for reference in the text when not otherwise specified.

In general, the proper label was attributed more often than any other label. For the animations of Anger, Cheerfulness, Panic Fear and Relief, the correct labels were significantly more often attributed than any other ones in both turns (McNemar test, $p<.05$ in each turn). For the remaining animations of Anxiety, Embarrassment, Pride and Tension the proper label was found but some confusions occurred. The strongest confusion occurred between Anxiety and Embarrassment. For the Anxiety animation, the number of attributions of the Anxiety (43% both turns means) and of the Embarrassment (36% both turns means) labels did not differ significantly (McNemar test, $p>.05$). In the Embarrassment animation, Embarrassment (41% both turns means) was confused with Anxiety (36% both turns means) ($p>.05$). In turn 2, Embarrassment (40%) was also labeled Tension (28%)

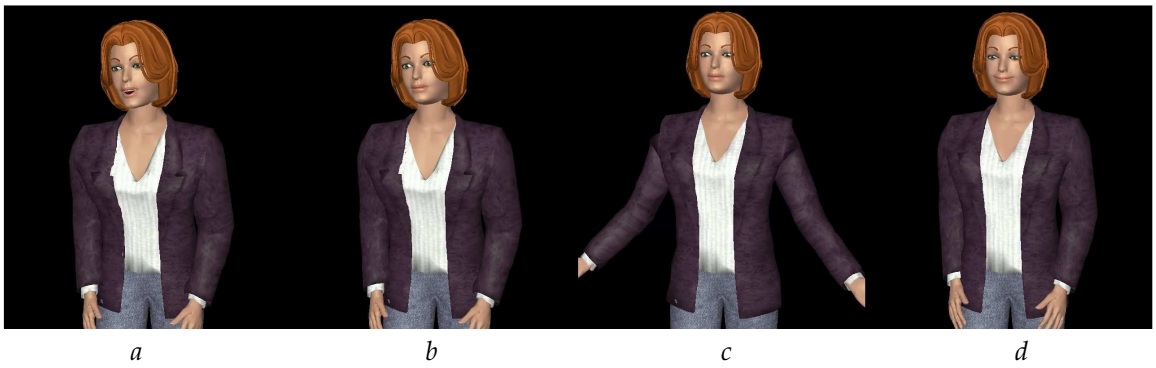


Fig. 5. Another example of the sequence of relief (SEQ3).

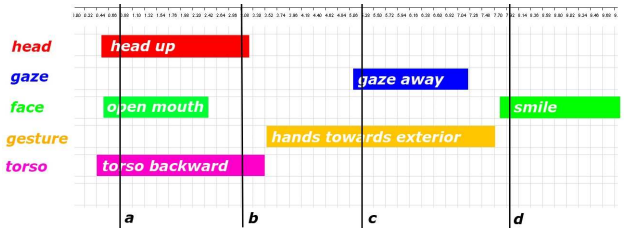


Fig. 6. Duration of signals in SEQ3.

(p>.05) (while in turn 1 it was labeled tension by 17%). Although on the limit of a significant difference (p=.066) some other confusions were found: Pride (45% both turns means) was labeled Relief (26% both turns means) in both turns and Tension (49%) was labeled Embarrassment (25%) in turn 2.

As it might be argued [33] that a correct recognition of a particular emotional expression may not only be considered in terms of correct attributions of a label, but also of rejections of that label for expressions not related to that emotion, we also calculated *unbiased hit rate*. For this purpose we use a Kappa score (κ), as outlined by [34]:

$$\kappa = \frac{(h + cr) - r_{exp}}{(i * j) - h_{exp}} \quad (1)$$

where h is number of hits, cr is number of correct rejections, r_{exp} - chance expected number of responses, i - presented items, j - number of judges, h_{exp} - chance expected number of hits. κ may vary from 0 (in the case of totally aleatory attribution) to 1 (if a label was always correctly attributed and correctly rejected, i.e. absence of false alarms).

One κ was calculated for each emotion (see Table 1). It was satisfactory for all emotions when the eight labels were counted in, with the highest κ for Anger (0.870) and the lowest for Embarrassment (0.702). Indeed, Embarrassment had also the lowest hit rate (37%) and the greatest number of false alarms (17%), showing a general tendency to attribute this label more often to our agent's behavior than any other label. The incorrect attributions of embarrassment were aimed at negative emotions other than Anger (Anxiety, Panic Fear, Tension).

Since "false alarms" are more likely to occur between similar emotions we also compare each emotion (summed attribution form the two turns) against the others from each conditions C1, C2 and C3. In C1, each of the three emotions was compared against two more labels. Relief had the highest unbiased score (κ = 0.503), then Cheerfulness (κ = 0.494) and Pride (κ = 0.356). In C2, each emotion was compared against four others: Cheerfulness and Relief had the highest recognition (κ = 0.697), then Pride (κ = 0.671), Anxiety (κ = 0.548) and Embarrassment (κ = 0.513). In C3, against 4 other labels, Anger was most recognized (κ = 0.807); than Panic Fear (κ = 0.715), Tension (κ = 0.612), Anxiety (κ = 0.558) and Embarrassment (κ = 0.547).

	Video							
	ang	anx	che	emb	pfe	pri	rel	ten
ang	**93**	0	0	0	0	2	0	5
anx	0	43	0	36	3	0	1	18
che	0	0	**70**	0	0	25	6	0
emb	0	36	0	41	1	0	0	23
pfe	2	11	0	17	**61**	0	7	2
pri	0	4	14	6	0	45	26	5
rel	0	0	23	1	0	8	**69**	0
ten	4	21	1	24	3	2	0	46
κ	0.87	0.71	0.80	0.70	0.83	0.77	0.80	0.75

TABLE 1
Matrix of confusions presented as percentages of attributions of the eight emotional labels and κ values (means for both turns, sign. values in bold, p<.05).

5.2.3 *Discussion*

The main aim of this evaluation study was to check if the multimodal sequential expressions are recognized by the participants. The hypothesis H1.1 was verified: the simple recognition rate (41% - 93% both turns means) exceeds strongly chance level and the unbiased hit rate measured by κ is satisfactory, also when the chance level is brought to in-group comparison instead of all the eight labels. The hypothesis H1.2 was only partially verified: although the number of

attributions of correct labels was higher than that of alternatives, the difference was not significative for some emotions. Finally, we observed that the effect of habituation (hypothesis H1.3) is not significative and consequently multimodal sequential expressions may be used straight away, in short period interactions with the user.

While the recognition rate is quite high, we believe it could have been higher if behavior expressivity was considered. In the videos used for this perception study, emotions were conveyed through signals defined in the behavior set. Behavior execution did not vary, that is behaviors had the same expressive qualities in all the videos. However, body expressivity is an important cue to convey emotional states as claims Wallbott [8] and as we can infer from our corpus annotation. The non adaptation of the behavior expressivity to the particular states might have influenced their perception and might have created a general bias. For instance, it appears that participants have a higher tendency to attribute embarrassment when judging the behavior of our agent, particularly when the emotional expressions have a negative aspect and do not portray anger. Thus, we believe that our MSE model should be extended in the future by a number of expressivity features.

The emotions that received the highest recognition rate - anger, cheerfulness, panic fear and relief - are those that are described by facial expressions as well as by specific body and arm movements (e.g. anger with the hands on the hips and cheerfulness with raised arms). It might be that expressions of emotions that make use of the full body were better perceived compared to expressions of emotions conveyed mainly with the face (such as embarrassment and tension). However this effect may also be explained by the framing used in this study. The animations showed half-body of the agent and consequently the face was quite small. The use of multimodality in communicating emotions should be more carefully analyzed (for example by studying how each modality contributes to the recognition of the internal state).

Nevertheless, our results show that even such subtly differentiated expressions like these of relief or of cheerfulness were recognized surprisingly well. One could argue that none of these expressions probably could have been recognized from still facial expressions in their apex or dynamical single signals, such as a hand or gesture movement. This claim is checked in the second evaluation.

5.3 The role of sequentiality, constraints and dynamical signals

In the previous section we showed that the emotional expressions generated with our algorithm are recognized. We also suggested that the MSE might be particularly useful to show subtly differentiated expressions. In the following studies we want to check which features of our approach permit a better recognition of emotions. First, we compare MSE animations to static emotional expressions presented in their apex and MSE animations to animations that do not respect the defined constraints (see section 5.3.1). Second, we look at signals that we have singled out from the sequenced sets of emotional behaviors and we present them one by one (see section 5.3.4). We check if the dynamic animations of short signals that contribute to multimodal expressions of emotions are sufficient per se for a particular emotion attribution.

a b c

Fig. 7. Three static images used for panic fear.

Our hypotheses are the following:

- H2.1) the recognition rate of multimodal sequential expression **is higher** than the recognition rate of static displays presented at the apex,
- H2.2) the recognition rate of multimodal sequential expression **is higher** than the recognition rate of single dynamical signals,
- H2.3) the animations generated using the constraint-based sequences **are more believable** than constraint-less sequences (i.e. animations not obeying the constraints).

5.3.1 Procedure of the MSE validation study

This study, accessible through a web browser, was divided into three sections:

- G1: 24 static images,
- G2: 16 MSE animations presented alone,
- G3: 8 MSE animations presented along with 8 "constraint-less" animations.

The images in section G1 show facial expressions, gaze and/or head movements (neither gestures nor torso movement were used). Since given facial expression has not been specified yet for some emotional states that are used in our evaluation, we have opted for three images presenting three different expressions chosen from the signals that occur in the MSE animations. For each image we used the key frame that corresponds to the apex. Each image was shown for 4 sec (see Figure 7).

For sections G2 and G3 three most dissimilar MSE animations were chosen from a set generated for each emotional state. The choice was based on the presence of different signals and/or their occurrence in time.

Section G2 of the study is composed of animations showing sequences of multimodal behaviors. For each emotion two different MSE animations were shown.

One animation was presented on each website. The agent was not speaking. The duration of each animation was about 10 seconds.

In section G3 16 animations are presented, eight of which were generated using our algorithm and which satisfied the defined constraints. The other eight present the same nonverbal behaviors but the order of appearance and duration of each nonverbal signal were chosen manually to be inconsistent with one or more constraints. We call these animations "constraint-less" animations.

Participants were asked to recognize the emotions displayed by the virtual agent. Each image or video shows the agent displaying one emotional state. The dimension of the agent head was kept constant in all the images and videos although a different framing was used (only the head for images and a half-body for the animations).

The study was constructed as follows. Each subject has to see all 24 images from section G1 before seeing the animations of the latter sections. In sections G1 and G2 after watching one image or animation the participants have to attribute one emotional label to the perceived emotional state from an 8-elements list before they can pass to another page with a new animation.

Section G3 checks the role of constraints in emotion recognition. On each website two animations of the same emotion were presented. Contrary to the other studies, the participants could start, stop and review the animations. In this section of the study (G3), participants have two different tasks: a recognition task was complemented by a ranking task. First, participants were asked to attribute a label to the depicted emotion, using a similar procedure to the previous sections (G1 and G2). But they were also asked to choose which animation is more believable between constraint-based and constraint-less animations.

In all studies participants were told that they could use each label more than once, or not at all.

5.3.2 Results of the MSE validation study

48 participants took part in sections G1 and G2 of the study (25 women, 22 men and one gender not stated) with a mean age of 29 years (SD=7.36), mainly from France (23%), Poland (21%) and Italy (12.5%). None of them works in the domain of virtual agents. Out of the 48 participants that finished the sections G1 and G2, 42 finished section G3 (20 women, 21 men and one gender not stated) with a mean age of 28 years (SD=3.98) mainly from France (21%), Poland (19%) and Italy (14%).

Images. In section G1 (still images), a repeated measures ANOVA was calculated to check for the impact of Emotions. Mauchly's test indicated that the assumption of sphericity had been violated and degrees of freedom were corrected using Greenhouse-Geisser estimates or Huyn-Feldt correction where appropriate. An effect of emotions on the number of correct recognitions (hits) of still images was observed ($F(3.80, 179) = 52.13$, $p<.05$). Overall hit rate interval is very large [4% - 90%] while the maximal hit rate varies from 25% (Pride) to 90% (Anger). Means, standard deviations and maximal hit rates for the three presentations of each emotion from section G1 (still images) are shown in Table 3.

Similarly to the previous experiment (see section 5.2.2) we also calculate for each image an *unbiased hit rate* by a Kappa score. We modified the calculation of the number of correct rejections, as we did not have an 8 x 8 design, but 8 emotional labels x 24 (3 images per emotion setting). The lowest κ value (0.595) was obtained for one Cheerfulness image and the highest value (0.865) - for one Anger image (see Table 3 for mean results).

A comparison brought down to labels from the same category was also realized. The lowest unbiased hit rate was of 0.179 for an image of Embarrassment in C3 and the highest of 0.761 for an image of Cheerfulness in C2 (C1 - [0.279-0.505]; C2 - [0.593.-0.761]; C3 - [0.179-0.741]). In C1, the most recognized image was of Relief ($\kappa = 0.505$), followed by one of Cheerfulness ($\kappa = 0.472$). In C2, the most recognized image was that of Cheerfulness ($\kappa = 0.761$), followed by an image of Embarrassment ($\kappa = 0.688$). In C3, the most recognized image was of Anger ($\kappa = 0.741$), followed by Panic Fear ($\kappa = 0.739$).

MSE Animations. The hit rate means of the animation presentations from G2 were compared with those from G3. Linear contrasts showed no difference between the grouped means of presentation one and two (section G2 of the study) and the means of presentation three (section G3) for 6 out of 8 emotions. Only in the case of Embarrassment the first presentation is recognized by 85%, while the second and third presentations are similar with a recognition mean of 42% and 40%. For Pride the third presentation is more recognized than the other two. Consequently, the Presentations displaying the MSE-sequence from sections G2 and G3 were considered together. The mean and maximal recognition rates for eight emotions are presented in Table 3.

A repeated measures ANOVA showed an effect of Emotions on the simple hit rate with $F(1.56, 30) = 14.71$, $p<.05$ after a Greenhouse-Geisser correction for sphericity.

Again, we calculate an *unbiased hit rate* by a Kappa score (κ). The scores show a non aleatory attribution for all MSE animations with min. value (0.719) for an Anxiety MSE animation and max. value (0.865) for a Panic Fear MSE animation (see Table 3 for mean results).

Finally, a comparison brought down to labels from the C1, C2 and C3 categories was realized for the animation presentations. The lowest unbiased hit rate was of 0.289 for an animation of Cheerfulness in C1

and the highest of 0.785 for an animation of Panic Fear in C3 (C1 - [0.289-0.590]; C2 - [0.569-0.747]; C3 - [0.396-0.785]). In C1, the most recognized animations was of Relief (κ = 0.590), followed by one of Pride (κ = 0.563). In C2, the most recognized animation was that of Embarrassment (κ = 0.749), followed by Pride (κ = 0.736). In C3, the most recognized animation was of Anger (κ = 0.825), followed by another animation of Anger (κ = 0.768) and by one of Panic Fear (κ = 0.768).

Effect of constraints. The emotional displays generated with and without the use of the constraints were compared. In section G3, the MSE animation was more often chosen as believable than the constraintless one for the emotions of Anger, Anxiety, Cheerfulness, Panic Fear, Pride and Relief (as measured with χ^2, p<.05). Only for Embarrassment and Tension the choice of the sequential animation was not above chance level (p<.05). The percentage of choices of the MSE animations is presented in Table 2.

5.3.3 Comparison of Images and MSE Animations (section G1 vs sections G2 and G3)

The mean hit rate for images is lower (0.402) than the mean for MSE animations (0.615), and the repeated measures ANOVA shows that the difference is significant (F(1, 41) = 91.64, p<.05, after a Huyn-Feldt correction).

In section G1 we obtained a very poor recognition of some images. Consequently, when comparing the hit rate between the images and the animations we relied only on the image and the video with the greatest hit rate for each Emotion. A repeated measures ANOVA calculated on the hit rates of the images and animations that were best recognized showed, after a Huyn-Feldt correction for sphericity, an effect of Emotions (F(6.96, 284)=15.14, p<.05) and of Dynamics (i.e. still images vs. animations) (F(1, 41) = 29.71, p<.05). An interaction effect was also observed for Emotions x Dynamics (F(7, 287) = 5.24, p<.05).

The mean hit rate for the best recognized videos and images is of 65%. When relying on the attributions to the images and animations with the best recognition proportion per emotion, t–tests were used to compare if the animations are more recognized than images for each emotional state. The number of participants was of 48 for Anger, Anxiety, Embarrassment, Panic Fear and Relief, while it was 42 for the others. The hit rate was higher for the Anxiety animation than that of the image, t= 2.44 (p<.05), higher for the Embarrassment animation than image, t = 5.08(47), p<.05, higher for the Panic Fear animation than image, t = 3.07(47), p<.05, higher for the Pride animation than image, t = 5.99(41), p<.05, higher for the Relief animation than image, t = 3.27(47).

However, it was higher for the Cheerfulness image (81%) than the video (64%), t = 2.44(41), p<.05. For Anger, there was no difference in the mean hit rate for the video (96%) and the image (90%), p>.05 (N=48), nor was there one between the Tension video (52%) and image (40%), p>.05 (N=42).

	MSE animations (%)	Std. Deviation
Anger	83	0,377
Anxiety	67	0,477
Cheerfulness	98	0,154
Embarrassment	55	0,504
Panic fear	74	0,445
Pride	83	0,377
Relief	90	0,297
Tension	52	0,505

TABLE 2
Means and standard deviations for the choice of the sequential animation as more believable than the random one in the eight emotions (N=42).

5.3.4 Procedure of the single signals study

In this study we evaluate all signals that appear in the behavior sets of eight emotions. Each animation shows Greta displaying only one signal. 61 different single signal animations (SS-animations) were created. The signals were extracted from MSE animations (sections G2 and G3, see section 5.3.1) and the duration and expressivity of each signal was kept the same as in the original animations.

In the single signals study participants were asked to recognize expressions of the virtual agent. Each participant was asked to watch 20 randomly chosen animations out of 61. However he could stop the study at any moment if he wished and he was told so in the instruction. The experimental procedure was similar to the one used in G1 and G2. Also in this study participants were told that they could use each label more than once, or not at all.

5.3.5 Results of the single signals study

89 participants took part this study, mainly from France (37%), Switzerland (9%) and Italy (9%). Each animation was evaluated by 23 to 27 participants.

SS-Animations. We looked at the attribution of expected labels to each signal. An animation of Anger showing frowning with parted lips that expose the teeth received the highest hit rate (96%). Out of 7 signals for Anxiety the highest hit rate is for the movement of hands coming down to a grasp of each other (48%). For Cheerfulness the most recognized signal was a slight smile with an open mouth (70%) and for Embarrassment - tensed lips (56%) and slightly spread hands (52%). For Panic Fear two facial expressions were the most recognized: eyebrows raised with mouth widely opened or with lips corners pulled down (both 62%). For Pride the best recognized signal is a head up movement (37%), while for Relief - a mouth opening (58%). The highest Tension attribution is for the fingers moving into a fist, arm hanging down

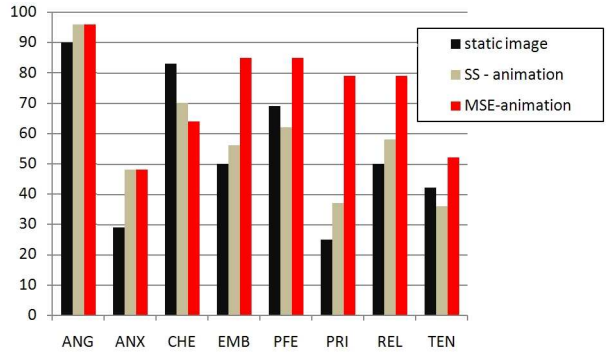


Fig. 8. Interaction effect of Emotions and Dynamics of the presented expressions on the hit rate of most recognizable items per category

along the body (36%). The mean and maximal hit rates for eight emotions are presented in Table 3.

5.3.6 Comparison of Images, SS Animations and MSE Animations

A repeated measures ANOVA was run on the hit rate of the most recognized images, single signals and MSE animations. Greenhouse-Geisser correction for sphericity was applied. An effect of Emotions ($F(2.605, 257.91)=73.55$, $p<.05$) and of Dynamics (i.e. still images, SS and MSE animations) ($F(1.77, 174.90) = 80.21$, $p<.05$) was found. An interaction effect was also observed for Emotions x Dynamics ($F(4.44, 439.643) = 26.11$, $p<.05$).

T–tests were used to compare if the most recognized MSE animations have a higher hit rate than the most recognized SS ones. The hit rate was higher for the Embarrassment MSE animation than the SS animation, $t=4.342(24)$, $p<.05$, higher for Panic Fear, $t=3.715(23)$, $p<.05$, higher for Pride, $t=4.914(26)$, $p<.05$ and for Relief, $t=4.053(23)$, $p<.05$. There was no difference for Anger, Anxiety, Cheerfulness nor Tension ($p>.05$).

5.3.7 Discussion

In the study different animations of multimodal sequential expressions were generally well recognized. The recognition rate was between 38% and 96% (to be compared with a chance level of 12.5%) which is similar result to the one obtained in the first study (see section 5.2). This result confirms again the utility of MSE to communicate emotional states.

We find that not only the recognition mean for images was lower than that of MSE animations, but also that this predicted effect was kept even when the best image and best video examples for each state were compared (H2.1 satisfied). What is more, the best example comparison showed that in most cases the use of multimodal sequential animations resulted in a higher recognition rate of illustrated states than in still expressions. Only in one case (Cheerfulness) the still face image was more recognized than MSE, while in two other cases (Anger, Tension) no significant difference was not observed.

Similarly for hypothesis H2.2 a significant improvement was observed for four emotions: Embarrassment, Panic Fear, Pride and Relief. In the case of Tension the improvement was not significant probably due to a small number of participants in the single signals study. For the remaining three emotions no improvement was observed.

We acknowledge that the improvement of an emotion recognition may be due to the dynamics of the animations or to any of three main features of the MSE-algorithm: the different modalities used, the sequentiality and/or the chosen constraints. The lack of improvement for some states could simply be due to an insufficient number of annotated videos and not to the insufficiency of the model. In some cases, the recognition is already very high for some still images and single signals, as for Cheerfulness or Anger. This could be due to the fact that the expression relies on a key signal, sufficient for the recognition of a given state. Indeed in the single signals study they obtained a high result for (96% and 70%). For others (e.g. Tension) the expressive qualities of a behavior may be guided mostly by expressivity characteristics, while in our animations these were kept constant.

In particular in the case of Cheerfulness adding supplementary information may not disambiguate the expression but may drive the attributional process away, increasing the chance of attributing alternatives. Indeed MSE animations of Cheerfulness had a very low κ score in C1, and much higher for still images. For Anger the recognition was already very high in the still images (90% correct recognitions) and the improvement was not significant (96%).

In the case of Anxiety and Tension, the improvement due to MSE may be less marked as annotations seem to show that this state may be particularly expressed by cues that are presented individually at longer periods of time. Moreover the expressive qualities of a behavior are very important while the expressivity of the agent was constant through the animations.

The presented results show that dynamical and multimodal expressions generated with our algorithm enable our agent to communicate many emotional states more efficiently than through static facial expressions or dynamical single modality expressions. We have also found that constraints play an important role in multimodal sequential expressions. Indeed in most of the cases the MSE were considered more believable than the animations not respecting any constraints (6 out of 8 cases).

5.4 Limitations

Our first challenge was to evaluate the MSE algorithm, without an a priori defined lexicon and without

	Static image (G1)				Single Signals Animation			MSE Animation (G2-G3)			
	Mean (%)	Max (%)	SD	κ	Mean (%)	Max (%)	SD	Mean (%)	Max (%)	SD	κ
Anger	70	90	0.22	0.84	39	96	0.3	88	96	0.32	0.85
Anxiety	24	29	0.04	0.67	31	48	0.095	43	48	0.50	0.72
Cheerfulness	55	83	0.29	0.72	32	70	0.2	54	64	0.50	0.80
Embarrassment	38	50	0.14	0.72	30	56	0.15	57	85	0.50	0.76
Panic fear	46	69	0.36	0,79	27	62	0.18	78	85	0.41	0.86
Pride	24	25	0.02	0.72	11	37	0.09	65	79	0.48	0.80
Relief	24	50	0.23	0.73	58	58	0.12	61	79	0.49	0.80
Tension	30	42	0.11	0.69	47	36	0.09	43	52	0.50	0.75

TABLE 3
The mean and maximal hit rate, the standard deviation, and mean κ score

any former verification of particular constraints. We realize that the results are dependent on the quality and quantity of processed and integrated information, whether from literature or from annotations. Indeed some emotions like tension had relatively worse scores than the other emotions (e.g. panic fear or relief). This may show that even though our approach is efficient for certain emotions it may not be sufficient to generate good displays for all emotional states. For instance in the case of tension and anxiety such cues as the expressive quality of the behavior are particularly important.

Another limitation of this study is the use of forced-choice. This procedure is often used in the perception studies on emotional displays (among others by Ekman in [1]) but it may force the user's interpretation of the expressive behaviors (see [6]). We also use only one virtual character in this study, while studies show that the interpretation of generated behavior may be influenced by the agent's physical characteristics (e.g. prominence of the eyebrows, gender, [35]). What is more, due to the limitations of the Greta agent we could not exploit all the possibilities of multimodal communication. For example, we did not use the posture to express emotional states which is an important channel to communicate emotions [36].

Last but not least in this work we focus on context-free emotional displays and we ask our participants to recognize emotions only from visual cues. However, the communication of subtle emotional states probably cannot be fully successful without considering situational context (e.g. nonverbal behavior of the receiver). Many other factors such as, e.g., position of the sender and receiver, available modalities or other communicative intentions to be communicated at the same time etc. may influence multimodal affective behaviors. These limitations may explain why some recognition rates are still somehow low.

6 Conclusion

In this paper we presented a novel approach to the generation of emotional displays in a virtual agent. The emotional expressions generated with the approach are not limited to the face but can be displayed using different modalities as well as by signals that occur in sequences. This approach allows for a high flexibility and variation of emotional displays. Avoiding the repetitiveness of nonverbal behavior is crucial in a long-term interaction with a virtual agent. Moreover, animations generated with the MSE-algorithm contain enough information to enhance the emotional communication, including the emotional states that are generally not considered.

In the second part of the paper we presented evaluation studies in which we verify three main features of our approach, which are the multimodality, sequentiality and the use of constraints. The results of our first study show that the recognition of the MSE animations is high. The second study enabled us to observe further, that multimodal sequential expressions are better recognized than static emotional displays in their apex and (at least for some emotional states) better than dynamical single signals. It showed also that the application of constraints increased the believability of the multimodal sequential expressions.

In the future we want also to improve the expressive quality of animations and link emotional states to some expressivity parameters. Furthermore, the role of different modalities in the perception of emotional states should be more deeply analyzed.

Acknowledgments

Part of this research is supported by the FP6 Project IP-CALLAS and French Projects ANR IMMEMO and ANR CECIL.

References

[1] P. Ekman and W. Friesen, *Unmasking the Face. A guide to recognizing emotions from facial clues*. New Jersey: Prentice-Hall, Inc., Englewood Cliffs, 1975.

[2] B. Jung, "Flurmax: An interactive virtual agent for entertaining visitors in a hallway," in *Proceedings of 4th International Workshop, IVA 2003*. Springer, 2003, pp. 23–26.

[3] M. Ochs, C. Pelachaud, and D. Sadek, "An empathic rational dialog agent," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 338–349.

[4] X. Pan, M. Gillies, T. M. Sezgin, and C. Loscos, "Expressing complex mental states through facial expressions," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII)*. Springer, 2007, pp. 745–746.

[5] B. Lance and S. Marsella, "Emotionally expressive head and body movements during gaze shifts." in *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*. Springer, 2007, pp. 72–85.
[6] J. Haidt and D. Keltner, "Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition," *Cognition and Emotion*, vol. 13, no. 3, pp. 225–266, 1999.
[7] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame," *Journal of Personality and Social Psychology*, vol. 68, pp. 441–454, 1995.
[8] H. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, pp. 879–896, 1998.
[9] F. Pollick, H. Paterson, A. Bruderlin, and A. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, pp. 51 – 61, 2001.
[10] M. N. Shiota, B. Campos, and D. Keltner, "The faces of positive emotion: Prototype displays of awe, amusement, and pride," *Annals of the New York Academy of Sciences*, 2003.
[11] K. R. Scherer and H. Ellgring, "Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal?" *Emotion*, vol. 7, pp. 113–130, 2007.
[12] H. K. M. Meeren, C. C. R. J. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proc Natl Acad Sci USA*, vol. 102, no. 45, pp. 16 518–23, 2005.
[13] J. A. Harrigan and D. M. O'Connell, "Facial movements during anxiety states," *Personality and Individual Differences*, vol. 21, pp. 205–211, 1996.
[14] R. J. Edelmann and S. E. Hampson, "The recognition of embarrassment," *Personality and Social Psychology Bulletin*, vol. 7, no. 1, pp. 109–116, 1981.
[15] J. A. Harrigan and D. M. O'Connell, "How do you look when feeling anxious? Facial displays of anxiety," *Personality and Individual Differences*, vol. 21, pp. 205–212, 1996.
[16] P. Rozin and A. Cohen, "High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans," *Emotion*, vol. 3, no. 1, pp. 68–75, 2003.
[17] Z. Ruttkay, "Constraint-based facial animation," *International Journal of Constraints*, vol. 6, pp. 85–113, 2001.
[18] N. Stoiber, R. Seguier, and G. Breton, "Automatic design of a control interface for a synthetic face," in *Proceedings of the 2009 International Conference on Intelligent User Interfaces*, C. Conati, M. Bauer, N. Oliver, and D. S. Weld, Eds. ACM, 2009, pp. 207–216.
[19] M. Paleari and C. Lisetti, "Psychologically grounded avatars expressions," in *First Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence*, Bremen, Germany, 2006.
[20] L. Malatesta, A. Raouzaiou, K. Karpouzis, and S. D. Kollias, "Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis," *Applied intelligence*, vol. 30, no. 1, pp. 58–64, 2009.
[21] C. Clavel, J. Plessier, J.-C. Martin, L. Ach, and B. Morel, "Combining facial and postural expressions of emotions in a virtual character," in *Proceedings of 9th International Conference on Intelligent Virtual Agents, IVA 2009*, ser. Lecture Notes in Computer Science, vol. 5773. Amsterdam, The Netherlands: Springer, 2009, pp. 287–300.
[22] N. Mana and F. Pianesi, "HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads," in *Proceedings of the 8th International Conference on Multimodal Interfaces, ICMI 2006*, F. K. H. Quek, J. Yang, D. W. Massaro, A. A. Alwan, and T. J. Hazen, Eds. Banff, Alberta, Canada: ACM, 2006, pp. 380–387.
[23] S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin, "EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces," in *11th International Conference on Human-Computer Interaction (HCII'2005)*, Las Vegas, Nevada, USA, 2005.
[24] E. Douglas-Cowie, N. Campbell, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1–2, pp. 33–60, 2003.
[25] E. Douglas-Cowie, C. Cox, J.-C. Martin, L. Devillers, R. Cowie, I. Sneddon, M. McRorie, C. Pelachaud, C. Peters, O. Lowry, A. Batliner, and F. Honig, "Humaine database," http://emotion-research.net/download/pilot-db/.
[26] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System. The Manual*. Salt Lake City, USA: A Human Face, 2002.
[27] R. Niewiadomski, S. Hyniewska, and C. Pelachaud, "Evaluation of multimodal sequential expressions of emotions in ECA," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. Amsterdam, Holland: Springer, 2009.
[28] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, pp. 832–843, November 1983.
[29] J. F. Allen and J. A. Koomen, "Planning using a temporal world model," in *Proceedings of the 8th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1983, pp. 741–747.
[30] D. E. Smith and D. S. Weld, "Temporal planning with mutual exclusion reasoning," in *Proceedings of the 16th international joint conference on Artifical intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 326–333.
[31] P. Ekman, "Darwin, deception, and facial expression," *Ann. N.Y. Acad. Sci.*, vol. 1000, pp. 205–221, 2003.
[32] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive ECA system," in *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*. IFAAMAS, 2009, pp. 1399–1400.
[33] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, pp. 102–141, 1994.
[34] D. M. Isaacowitz, C. E. Lckenhoff, R. D. Lane, R. Wright, L. Sechrest, R. Riedel, and P. T. Costa, "Age differences in recognition of emotion in lexical stimuli and facial expressions," *Psychology and Aging*, vol. 22, pp. 147–159, March 2007.
[35] U. Hess, R. B. J. Adams, and R. E. Kleck, "Facial appearance, gender, and emotion expression," *Emotion*, vol. 4, pp. 378–388, 2004.
[36] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 48–58.

Radosław Niewiadomski received his PhD degree in Computer Science at University of Perugia, Italy in 2007. Currently he is a post-doc researcher at Telecom ParisTech, France. His research interests include nonverbal communication between humans and computers, synthesis of emotional displays, and multimodal interfaces.

Sylwia Julia Hyniewska is a junior scientist with a background in affective psychology. Her research interests comprise nonverbal behavior, psychophysiological measures and affective computing. She is currently working as a PhD candidate at Telecom ParisTech and the University of Geneva on the attribution of appraisals and emotion labels to third parties and on embodied virtual agents' behavior.

Catherine Pelachaud is Director of Research at CNRS in the LTCI laboratory, Telecom ParisTech. Her research interest includes embodied conversational agents, representation language for agent, nonverbal communication, expressive behaviors and multimodal interfaces. She has been involved and is still involved in several national and European projects.