KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

INTRODUCING MULTIMODAL SEQUENTIAL EMOTIONAL EXPRESSIONS FOR VIRTUAL CHARACTERS

Radoslaw NIEWIADOMSKI, Sylwia HYNIEWSKA and Catherine PELACHAUD

LTCI-CNRS, Telecom ParisTech, France

ABSTRACT

In this paper we present a system which allows embodied conversational agent to display multimodal sequential expressions. Recent studies show that several emotions are expressed by a set of different nonverbal behaviors which include different modalities: facial expressions, head and gaze movements, gestures, torso movements and posture. Multimodal sequential expressions of emotions may be composed of nonverbal behaviors displayed simultaneously over different modalities, of a sequence of behaviors or of expressions that change dynamically within one modality. This paper presents, from the annotation to the synthesis of the behavior, the process of multimodal sequential expressions generation as well as the results of the evaluation of our system.

Keywords: **virtual characters, emotional expressions, multimodality**

1. INTRODUCTION

In this paper a novel approach to the generation of emotional displays of a virtual character is presented. The aim is to develop a model of multimodal emotional behaviors that is based on data from literature and on the annotation of a video-corpus. For this purpose a language was developed to describe the appearance in time of single signals as well as the relations between them.

We call *multimodal sequential expressions of emotions* emotional displays that go beyond the description of facial expressions of emotions in their apex. Dacher Keltner and colleagues (e.g.

[1, 2]) showed that several emotions are expressed by a set of different nonverbal behaviors which include different modalities: facial expressions, head and gaze movements [3], gestures [1], torso movements and posture [4, 5]. The expressions of emotional states are dynamic, composed of several nonverbal behaviors (called signals in this paper) and arranged in a certain interval of time. It is in line with the componential appraisal theory, which claims that an emotion is a dynamic episode that produces a sequence of response patterns on the level of gestures, voice and face [6]. The expressive complexity of emotions like anxiety [7], confusion [8], embarrassment [1] or worry [8] was analyzed in some observational studies. Among others three positive emotions: pride, awe and amusement were differentiated [2]. Their expressions go beyond the one well recognized expression of a positive emotional state i.e. associated to a smile. For example awe [2] may be expressed by raised inner eyebrows (AU 1), widened eyes (AU 5), an open mouth with a slight drop of the jaw (AU 26 + AU27). These facial expressions are completed by other dynamic behaviors across different modalities like forward head movements or deep inhalations. Another emotion displayed by multimodal behaviors is shame which is expressed by a coordinated sequence of a downward gaze and head movements [1, 3].

The remaining part of this paper is structured as follows. In the next Section different approaches to emotional displays in virtual characters are described. Then, Section 3 explains how two structures, behavior set and constraint set, are created. Section 4 describes the algorithm of multimodal sequential expressions as well as some examples of expressions synthesized with MPEG-4 compliant virtual character. In Section 5 the results of an evaluation study of multimodal sequential expressions are presented. We conclude the paper in Section 6.

2. RELATED WORKS

Several models of emotional expressions have been proposed to enrich virtual characters behavior. Most of them focus on facial expressions. A tool that allows one to modify manually the course of the animation of any single facial parameter was proposed in [9]. In that work to maintain plausibility of animations, the facial displays are limited by a set of constraints. These constraints are defined manually on the key-points of the animation and concern the facial animation parameters.

Other researchers were inspired by the appraisal theory [10], which states that different cognitive evaluations of the environment lead to specific micro-expressions. Paleari and Lisetti [11] and Malatesta et al. [12] focus on the temporal relations between different facial actions predicted by this theory. In [11] the different facial parameters are activated at different moments and the final animation is a sequence of several micro-expressions linked to cognitive evaluations. Also in Malatesta et al. [12] the emotional expressions are created manually from sequences predicted in Scherer's theory [10]. Differently from Paleari and Lisetti's work [11] each expression is derived from the addition of a new AU to the former ones. What is more, the authors [12] compared the additive approach with the sequential one. Results show an above chance level recognition in the case of the additive approach, whereas the sequential approach gives recognition results marginally above random choice [12]. The dynamics of emotional expressions is also modeled by Xueni Pan et al. [13]. In this approach a motion graph is used to generate emotional displays from sequences of signals like facial expressions and head

movements. The arcs of the graph correspond to the observed sequences of signals while nods are possible transitions between them. The data about emotional expressions were extracted from a video-corpus. Different paths in the graph correspond to different displays of non-Ekmanian emotions. Thus, new animations can be generated by reordering the observed displays.

The expressive multimodal behaviors of virtual characters are generated in the system proposed by Michael Kipp [14]. This system automatically generates nonverbal behaviors that are synchronized with the verbal content in four modalities using a set of predefined rules. These rules determine triggering conditions of each behavior in function of the text. Thus a nonverbal behavior can be triggered, for example, by a particular word, sequence of words, type of sentence (e.g. question) or when the agent starts a turn. The system offers also the possibility to discover new rules. Similarly Hofer and Shimodaira [15] propose an approach to generate head movements based on speech. Their system uses Hidden Markov Models to generate a sequence of behaviors. Data to train the model was manually annotated and it includes four classes of behaviors: postural shifts, shakes and nods, pauses, and movement.

Lance and Marcella [16] 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 form the recordings of acted emotional displays. For this purpose the head and body movements' data was captured through three motion sensors and evaluated by human coders. A set of proposed parameters contains temporal scaling and spatial transformations. Consequently, emotionally neutral displays of head and body movements can be transformed in this model to multimodal displays showing e.g. low/high dominance and arousal.

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. Contrary to many other approaches which use captured data for behavior reproduction, in this approach the observed behaviors are interpreted by a human 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. It is built on observational data. Last but not least it generates a variety of animations for one emotional label avoiding the repetitiveness in the behavior of a virtual character.

3. MULTIMODAL SEQUENTIAL EXPRESSIONS LANGUAGE

In this section we present the representation scheme that encompasses the dynamics of emotional behaviors. The scheme is based in observational studies. We use a symbolic high level notation. Our XML-based language defines multimodal sequential expressions in two steps: behavior set and constraint set. Single signals like a smile, shake or bow are described in the repositories of the character's nonverbal behaviors. Each of them may belong to one or more behavior sets. Each emotional state has its own behavior set, which contains signals that might by used by the character to display that emotion. According to the observational studies (e.g. [1]) the signals occurrence in an emotional display is not accidental. The relations that occur

between the signals of one behavior set are more precisely described in the *constraint sets*. In our algorithm 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.

3.1. Behavior set

The concept of behavior set was introduced in [17]. 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 character. All behaviors belonging to a behavior set are defined in a central database called *lexicon* [17]. We use behavior sets to describe the multimodal sequential expressions of emotions. Let us present an example of such a behavior set. In [1] the sequence of signals in the expression of embarrassment is described. The typical expression of embarrassment starts from a downward gaze or gaze shifts which are followed by "controlled" smiles (often realized with pressed lips). The expression of embarrassment often ends with the head movement to the left accompanied by face touching gestures [1]. Thus the behavior set based on Keltner's description [1] of embarrassment will contain the ten signals: two head movements: *head down* (signal 1) and *head left* (signal 2), three gaze direction: *look down* (signal 3), *look right* (signal 4), *look left* (signal 5), three facial expressions: *smile* (signal 6), *tensed smile* (signal 7), and *neutral expression* (signal 8), open flat hand on mouth gesture (signal 9), and a *bow* torso movement (signal 10).

A number of regularities occur in expressions that concern the signal duration and the order of displaying (see e.g. [1, 2]). Consequently for each signal in a 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* - number of repetitions during an expression. In the embarrassment example the signals *head down* and *gaze down* occur much more often at the beginning of the multimodal expression [1]. Thus their values of *probability start* are much higher than the value of *probability end*. For example, the definition of *head down* signal in *lexicon* is:

```
<signal id="1" name="head=head_down"
repetitivity="0" min_duration="2" max_duration="4"
probability_start="0.8" probability_end="0.3"/>
```

3.2. 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 arithmetical 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 co-occur at the same time t_k i.e.:

  if $start_{si} < t_k < stop_{si}$ then $stop_{sj} < t_k$ or $start_{sj} > t_k$ and if $start_{sj} < t_k < stop_{sj}$ then $stop_{si} < t_k$ or $start_{si} > t_k$;
- *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.

During the computation of the animation constraints are instantiated with signals appearance times (i.e. $start_{s_i}$ and $stop_{s_i}$). By the 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 there is no constraint that is not satisfied.

4. FROM ANNOTATION TO BEHAVIOR GENERATION

In this section we present how the definition of behavior and constraint sets are created from the manual annotation. We will show also some examples of animations generated with our algorithm [19] from this description.

4.1. Annotation

A corpus has been created by choosing, from different sources, audio-visual clips that captured expressions from non-actors behaving naturally during emotional situations. For each treated emotion two to six videos have been chosen. One coder annotated the modalities of the face, head, gaze and body movements. The facial changes have been annotated by a certified FACS expert [18], while the head, gaze and body movements were described verbally. For

practical reasons a *signal* is defined as a configuration of body actions that can occur at the same time in a particular modality. Thus one signal per modality is displayed at a time. Usually different body actions of one modality were defined as independent signals, e.g. *a hand touching the face* and *a hand hiding the mouth* gestures are two signals. The same body actions can be part of several signals, if they can occur in different configurations and with different co-occurrences, e.g. a smile is a signal, a smile with an open mouth is another one even if they have some AUs in common. Several signals, in different modalities, can be defined in the annotation at the same time. For a body action to be included in a signal, it has to be long enough to be clearly observed, e.g. when only the offset of a facial action unit is seen in a time interval, that element is not considered to occur, nor is it when the occurrence is only transitional.

Relief was one of the analyzed emotional states. Twelve different signals were identified by the annotator, among which 5 facial expressions, 3 torso and 2 head movements, 1 gaze and 1 gesture. In Figure 1 we can see the annotation of one of the videos. The following signals were individuated in this sample: raising hands gesture (*signal 1*), head movement to left (*signal 5*), backward torso movement (*signal 3)*, open mouth (*signal 2*) and smile (*signal 12*).

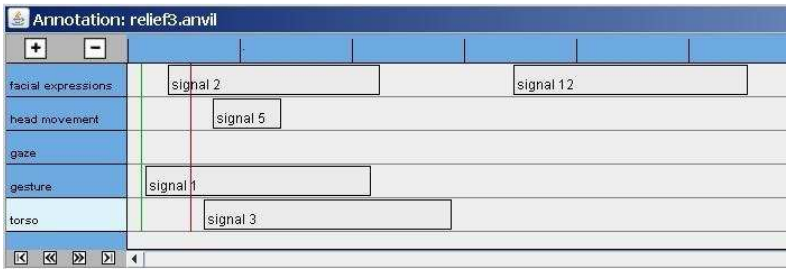

Fig. 1. An example of a multimodal annotation of a relief expression in Anvil [20].

By looking at the relief videos one could argue that *signal 5* (i.e. a head movement to the left) co-occurs with *signal 2* (i.e. facial expressions of mouth extremely opened) or with *signal 3* (i.e. backward torso movement associated with a strong exhalation). Indeed, in the example (see Figure 1) signal *5* starts after the *signal 2* starts and it also stops before the end of *signal 2*. In fact, *signal 5* by itself is considered not sufficient to convey any clear emotional meaning. It has been interpreted rather as an accentuation of a state expressed by *signal 2*. This information might be described in the constraint set by *appearance constraints* of the type *includes* and *exists*. When *signal 5* cannot appear without *signal 2* or without *signal 3* we obtain the composed constraint: (*exists(signal2) and includes(signal2,signal5)) or (exists(signal3) and includes(signal2,signal3)).*

.

4.2. Generation

In our model, the behavior and constraint sets are used to generate multimodal sequential expressions of emotions. The input to the system is one label *e* (e.g. *panic fear* or *embarrassment)* from a predefined set of emotional labels and its expected duration, *t*. Our system generates sequences of multimodal expressions, i.e. the animation *A* of a given duration *t*

composed of a sequence of signals $s_{i(j)}$ on different modalities. It does so by choosing a coherent subset of signals from the behavior set BS_e as well as their timing $start_{s_i}$, $stop_{s_i}$. More details in [19].

The algorithm is able to generate several animations that are consistent with the constraints. In this way we avoid the repetitiveness of the character's behavior and we obtain a variety of animations, each of which is consistent with the annotator's observation but which go beyond a set of annotated cases.

We used the Greta agent [21] to generate animations using our model. Two examples are presented below. In section 3.3 an expression of relief was discussed. In Figure 2 an animation generated by our algorithm from this description is presented. The following signals are displayed: 2a) *signal 1* with *signal 2* – raising hands gesture and mouth extremely open, 2b) *signal 1*, *signal 2* with *signal 5* – head movement to the left, 2c) *signal 2*, *signal 5* with *signal 3* – backward torso movement, 2d) *signal 12* - smile.

a b c d

Fig. 2. An example of a multimodal expression based on the annotation of relief.

In section 3.1 the behavior set of embarrassment was presented. Figure 3 shows an animation of the agent displaying this emotion. It is composed of the following signals: *signal 1* and *signal 3* - head and gaze down (Figure 3a) *signals 2* and *signal 5* - head and gaze left (Figure 3b) which are accompanied by a tensed smile (*signal 7*) on Figure 3c and a gesture of touching the mouth (*signal 9*) on Figure 3d.

a b c d

Fig. 3. An example of a multimodal expression of embarrassment.

5. EVALUATION

A study was run to evaluate the recognition of the expressions created with the algorithm and synthesized with the Greta agent [21]. 53 participants took part in the study. They evaluated eight animations displaying an affective state (anger, anxiety, cheerfulness, embarrassment, panic fear, pride, relief, and tension). Participants were asked to choose which label from the eight possible emotions described the expression best. The set of eight animations was evaluated twice and the order of presentations of the animations was random in both turns. For all the animations the intended emotional labels was expected to be attributed more often than any alternative one, and the recognition rate to be higher than by chance.

The number of correct vs. alternative answers in turn 1 and turn 2 was compared and the improvement was not significative (McNemar test, $p>.05$). The recognition level for each emotional expression in both turns is above chance level (which is 12.5%). The best recognized emotion was anger (94% both turns mean) while the least recognized one was embarrassment (41% both turns mean).

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, $p<.05$). For the remaining animations of anxiety, embarrassment, pride and tension the proper label was found but some confusion occurred. The strongest confusion occurs between anxiety and embarrassment. In both turns, we found that in the anxiety animation the number of attributions of the anxiety label (43% both turns means) and of the embarrassment label (33% both turns means) did not differ significantly (McNemar test, $p>.05$). In the embarrassment animation, embarrassment (41% both turns means) was confused with anxiety (37% both turns means) ($p>.05$). In turn 2 embarrassment (40%) was also labeled tension (28%) ($p>.05$) (more than in turn 1 with 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.

The strongest confusions were observed among similar emotions. This similarity can be cognitive, as in embarrassment and anxiety, which both share a certain amount of uncertainty. It can also be of a physical type, sharing similar features such as a tensed smile in tension and in embarrassment. Thus, our results do not deny the relevance of multimodal sequential cues in communicating emotional expressions. On the contrary, one has to keep in mind that the studies sustaining the universality of recognition of the most prototypical expressions often state only that the percentage of subjects who agreed with prediction, were greater than that to be expected by chance. Our results show that even such subtly differentiated expressions like these of relief or of cheerfulness were recognized surprisingly well. One could argue that these emotions probably would not have been recognized from still facial expressions in their apex. This claim needs however to be checked in future studies.

The effect of habituation is small and the improvement as seen in correct labeling between first and second turn is not significative. Consequently multimodal sequential expressions may be used straight away, in short period interactions with the user.

6. CONCLUSIONS

In this paper a multimodal sequential expressions model for a virtual character was introduced. These expressions go beyond static facial displays defined in their apex. For this purpose a language was proposed that allows formalizing the observational data and an algorithm that generates multimodal sequential expressions coherent with their descriptions. A perceptual study was conducted and the results show that multimodal sequential expressions enable the recognition of affective states, such as relief, that are not prototypical expressions of basic emotions. In the case of all eight emotions the recognition rate surpassed chance level.

The research on multimodal sequential expressions of emotions should be continued. In the videos used for the 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 [4]. We believe that the recognition rates might improve if expressivity parameters, such as fluidity and power of gestures, were modified in accordance with particular emotional states. Thus, we plan to include them in our model. We will also continue with the evaluation of model. In particular we would like to check if the recognition rates vary for different animations of the same emotion.

ACKNOWLEDGEMENT

Part of this research is supported by the EU FP6 Integrated Project CALLAS IP-CALLAS IST-034800.

REFERENCES

[1] Keltner, D. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame, *Journal of Personality and Social Psychology* 68, pp. 441–454, 1995.

[2] Shiota, MN, Campos, B, Keltner, D. The faces of positive emotion: Prototype displays of awe, amusement, and pride, Annals of the New York Academy of Sciences, 2003.

[3] Haidt, J, Keltner, D. Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition, *Cognition and Emotion* 13(3), pp. 225–266, 1999.

[4] Wallbott, H. Bodily expression of emotion. *European Journal of Social Psychology* 28, pp. 879–896, 1998.

[5] Pollick, F, Paterson, H, Bruderlin, A, Sanford, A. Perceiving affect from arm movement. *Cognition* 82, pp. 51–61, 2001.

[6] Scherer, KR, Ellgring, H. Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion* 7, pp. 158–171, 2007.

[7] Harrigan, JA, O'Connell, DM. Facial movements during anxiety states, *Personality and Individual Differences 21*, pp. 205–211, 1996.

[8] Rozin, P, Cohen, A. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans, *Emotion* 3(1), pp. 68–75, 2003.

[9] Ruttkay, Z. Constraint-based facial animation, *International Journal of Constraints 6*, pp. 85–113, 2001.

[10] Scherer, KR. Appraisal considered as a process of multilevel sequential checking, In Scherer, K, Schorr, A, Johnstone, T, eds., Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, pp. 92–119, 2001.

[11] Paleari, M, Lisetti, C. Psychologically grounded avatars expressions, In: First Workshop on Emotion and Computing at KI 2006, *29th Annual Conference on Artificial Intelligence*, Bremen, Germany, 2006.

[12] Malatesta, L, Raouzaiou, A, Karpouzis, K, Kollias, SD. Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis, *Appl. Intell*. 30(1), pp. 58–64, 2009.

[13] Pan, X, Gillies, M, Sezgin, TM, Loscos, C. Expressing complex mental states through facial expressions. In: *Second International Conference on Affective Computing and Intelligent Interaction*, Springer, pp. 745–746, 2007.

[14] Kipp, M. Creativity meets automation: Combining nonverbal action authoring with rules and machine learning, In: *Proceedings of 6th Conference on Intelligent Virtual Agents*. LNCS, Springer, pp. 230–242, 2006.

[15] Hofer, G, Shimodaira, H. Automatic head motion prediction from speech data, In: *Proc. Interspeech 2007*, Antwerp, Belgium, August 2007.

[16] Lance, B, Marsella, S. Emotionally expressive head and body movements during gaze shifts, In: *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, Springer, pp. 72–85, 2007.

[17] Mancini, M, Pelachaud, C. Distinctiveness in multimodal behaviors, In: *7th International Joint Conference on Autonomous Agents and Multiagent Systems* (AAMAS 2008), Estoril, Portugal, pp. 159–166, 2008.

[18] Ekman, P, Friesen, W. *Facial Action Coding System*. Consulting Psychologists Press, 1978.

[19] Niewiadomski, R, Hyniewska, S, Pelachaud, C. Modeling emotional expressions as sequences of behaviors. In: *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, Amsterdam, Holland, 2009.

[20] Kipp, M. Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370, 2001.

[21] Niewiadomski, R, Bevacqua, E, Mancini, M, Pelachaud, C. Greta: an interactive expressive ECA system In: *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems*, Budapest, 2009.