Modelling Expressive Embodied Conversational Agents

Catherine Pelachaud
Elisabetta Bevacqua
Sylwia Hyniewska
Radoslaw Niewiadomski

CNRS - TELECOM ParisTech
Greta: Embodied Conversational Agent able to display in realtime multimodal communicative and emotional nonverbal behaviors.

Presentation
- Interactive system: SAIBA compliant
- Multimodal sequential expressions
- Listener model
Interactive System

SAIBA Compliant
International initiative:

Aim: exchangeability of the modules and the animation scenes

- Examples:
  - SmartBody (USC)
  - CADIA BML Realizer (Reykjavik University)
SAIBA FRAMEWORK

**FML Function Markup Language:**
- description of the aspects that are relevant and influential in the planning of verbal and nonverbal behaviour, without reference to overt behaviour
- basic semantic units
- expressive, affective, discursive, epistemic, or pragmatic functions.

**BML Behavior Markup Language:**
- describes multimodal behaviors as they are to be realized by the final stage of the generation
- must provide means of describing behavior at a level of detail needed by the “optimal” realizer
- from the mere occurrence and relative timing of behaviors, to the detailed (yet player-independent) definition of form
Taxonomy of Communicative Functions

A semantic topology based on the information to be conveyed has been defined by I. Poggi:

- Information on the **Speaker's Identity**
  - Age, culture, gender...

- Information on the **World**
  - Description of object (iconic)
  - Reference to object (deictic)

- Information on the **Speaker’s Mind**
  - Information about Speaker’s belief (degree of certainty, belief relation)
  - Information about Speaker’s intention (performative, turn-taking, emphasis)
  - Information about Speaker’s affective state
  - Meta-cognitive information about Speaker’s mental action
Lexicon = (meaning, signal)

- **Expression meaning**
  - **deictic**: this, that, here, there
  - **adjectival**: small, difficult
  - **certainty**: certain, uncertain...
  - **performative**: greet, request
  - **topic comment**: emphasis
  - **belief relation**: contrast,...
  - **turn allocation**: take/give turn
  - **affective**: anger, fear, happy-for, sorry-for, envy, relief, ....

- **Expression signal**
  - **deictic**: pointing finger
  - **adjectival**: small eye aperture
  - **certainty**: *Certain*: palm up open hand; *Uncertain*: raised eyebrow
  - **performative**: *Suggest*: small raised eyebrow, head aside; *Assert*: horizontal ring
  - **topic comment**: raised eyebrows, head nod, beat
  - **belief relation**: *Contrast*: raised eyebrow
  - **affective**: *Sorry-for*: head aside, inner eyebrow up; *Joy*: raising fist up
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE fml-apml SYSTEM "fml-apml.dtd" []>
<fml-apml>
  <bml>
    <speech id="s1" start="0.0" language="english" text="Hello world."
     <description level="1" type="gretabml">
      <reference>tmp/from-fml-apml.pho</reference>
    </description>
    <tm id="tm1"/>
    Hello world!
    <tm id="tm2"/>
  </speech>
  </bml>
  <performative id="p1" type="greet" start="s1:tm1" end="s1:tm2"/>
  <emotion id="e1" type="joy" start="s1:tm1" end="s1:tm2"/>
  <world id="w1" ref_type="place" ref_id="away" start="s1:tm1" end="s1:tm2"/>
</fml-apml>
BEHAVIOR MARKUP LANGUAGE

- **BML:**
  - operates at signal level,
  - different channels of (non)verbal communication can be defined:
    - head movements, gaze, facial expressions, gestures, speech.
  - expressivity parameters.
Expressivity Dimensions

- Behaviors encode:
  - content information (the ‘What is communicating’)
  - expressive information (the ‘How it is communicating’)

- Behavior expressivity refers to the manner of execution of the behavior
  - **Spatial**: amplitude of movement
  - **Temporal**: duration of movement
  - **Power**: dynamic property of movement
  - **Fluidity**: smoothness and continuity of movement
  - **Repetitiveness**: tendency to rhythmic repeats
  - **Overall Activation**: quantity of movement across modalities
BEHAVIOR MARKUP LANGUAGE

expressivity parameters

unique name
duration
class and instance

standard
extensions
ARCHITECTURE

Audio & video input

Listener Intent Planner

FML-APML

Behavior Planner (dynamicline)

FML-APML BML

Behavior Realizer

BML

FAP-BAP

FAP-BAP

FAP-BAP

PSYCLONE

Central Clock

baseline

repositories

FAP-BAP Player

Audio & video input
BEHAVIOR PLANNER

Input: FML-APML files computes which signals should be used to display a communicative intention.

processes:

Agent definition:

*baseline* - information on the preference the agent has in using each communicative modality and on the expressive quality of each of these modalities;

*Repositories* – different modalities signals in described in a symbolical notation,

*Lexicon* – mapping between communicative intentions and signals

dynamicline:

depends on each communicative intention and baseline of the agent
INPUT: Communicative intentions in FML-APML

OUTPUT: set of signals in BML with expressivity parameters
ARCHITECTURE

Listener
Intent Planner

Behavior Planner
(dynamic line)

Behavior Realizer

FAP-BAP Player

Audio & video input

baseline

repositories

PSYCLONE

Central Clock

FML-APML

FML-APML

FAP-BAP

FAP-BAP

BML

BML
BEHAVIOR REALIZER

- Generation the animation in MPEG4 format:
  - input specified in *BML language*,
  - the *speech* is generated by an external TTS,
  - adds *lips movements*,
  - solves also eventual *conflicts* between the signals that use the same modality,
  - generates a smooth animation using interpolation algorithms.
Multimodal Sequential Expressions
Multimodal Sequential Expressions

- Go beyond static mono-modal expressions
- Many emotions are expressed by sequences (or combination) of multimodal signals rather than monomodal signals (e.g., static facial expressions)
- Data obtained from theory and literature: (Keltner (1995); Shiota et al (2003); Harrigan & O’Connell (1996); Rozin & Cohen (2003))

D. Keltner, B. N. Buswell
Embarrassment: Its Distinct Form and Appeasement Functions
Corpus of Emotional Displays

- **Difficulty:** lack of relevant research and video-corpora

- Annotation of audio-visual recordings from reality shows, hidden camera recordings, Belfast Naturalistic database, EmoTV corpus.

- The observed people are non actors in emotional situations: natural and not stereotyped multimodal behaviour is displayed

- 20 video clips (3 to 14 seconds each):
  - relief (2), tension (6), cheerfulness (2), sadness (2), anger (3), anxiety (1), panic fear (4)
Multimodal Annotation Scheme

Annotation on Anvil v4.7.6 (Kipp, 2001) with 5 tracks:
- Emotion (inferred from the situation)
- Facial expression (FACS coding)
- Head movement
- Gaze movement
- Gestures

Description in natural language
Video annotations with the Anvil software: multimodal display of cheerfulness from the Belfast Naturalistic Emotional Database (Cowie et al., 2003)
Extraction of Signals

- **Expressive signals**: clustering of all micro-expressions across modalities that co-occur.

- The independent expressive signals from each modality can be combined with expressive signals from other modalities.

- Some are sufficient for the display/recognition of cheerfulness (e.g. open mouthed smile), some are complementary (e.g. rotations of the torso).
Example: Panic Fear

- Emotion present: panic fear
- Duration of the annotated extract: 2 sec
- Annotation of the facial and body actions
Example: Panic Fear
Example: Panic Fear

**Signal 1:**
Eyebrows very raised and drawn together, Eyes extremely open, Mouth extremely open

**Signal 2:**
Eyebrows slightly raised, slightly drawn together, Mouth slightly open with lowered mouth corners

**Signal 3:**
Upper lid raised widening the eye, Mouth open

**Signal 4:**
Eyebrows drawn together, upper lid raised to widen the eye, lower lid raised (squint)

**Signal 5:**
Outer eyebrows raised, Eyebrows drawn together, Mouth open with lowered corners
While looking at the sequence, particularly of the mouth, one can see a repeated pattern of:

1. a prototypical expression of fear (eyes wide open, eyebrows extremely raised and mouth wide open) followed by
2. a second more moderate negative expression (eyes less open, eyebrows moderately drawn together and the mouth corners falling down)
Example: Panic Fear

<table>
<thead>
<tr>
<th>Face</th>
<th>1+25</th>
<th>1+2+4+5+25+27</th>
<th>1+4+5+15+25</th>
<th>5+25</th>
<th>4+5+7+25+27</th>
<th>2+4+15+25</th>
</tr>
</thead>
<tbody>
<tr>
<td>gaze</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>away from stimulus (left)</td>
</tr>
<tr>
<td>head</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>away from stimulus (left)</td>
</tr>
<tr>
<td>body</td>
<td>hands</td>
<td>touch face</td>
<td>R in the air, L hides mouth than touches chest</td>
<td>both on chest</td>
<td>L hides mouth</td>
<td></td>
</tr>
<tr>
<td>torso</td>
<td>shoulders up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


**Example: Panic Fear**

<table>
<thead>
<tr>
<th>Face</th>
<th>1+25</th>
<th>1+2+4+5+25+27</th>
<th>1+4+5+15+25</th>
<th>5+25</th>
<th>4+5+7+25+27</th>
<th>2+4+15+25</th>
</tr>
</thead>
<tbody>
<tr>
<td>gaze</td>
<td></td>
<td>away from stimulus (left)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>head</td>
<td></td>
<td>away from stimulus (left)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>body</td>
<td>hands touch face</td>
<td>R in the air, L hides mouth than touches chest</td>
<td>both on chest</td>
<td>L hides mouth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>torso</td>
<td>shoulders up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Signal 1**

**Signal 2**

**Signal 3**

**Signal 4**

**Signal 5**
Expression Representation

- Definition of a representation scheme for emotional expression:
  - signals description across modalities
  - their partial temporal order
  - their constraints

- **Behavior set**: set of signals (frown, nod, torso backward, ...) 

- **Constraint set**: describes the temporal and spatial constraints among the signals in the behavior set
  - Temporal constraint: start-signal $s_i <$ end-signal $s_j$
  - Spatial constraint: signals $s_i$ and $s_j$ cannot co-occur
<emotion id="e1" type="embarrassment" start="1.0" end="4.0">
Embarrassment
Panic Fear
Relief
Evaluation Study

- Validation of the computational model of multimodal sequential expressions
- 41 graduate and post-graduate participants between 23 and 34 years
- 8 videos of emotional expressions
  - C1: distinction between positive emotions (cheerfulness, pride and relief)
  - C2: distinction between Duchenne and non-Duchenne smiles (cheerfulness and embarrassment - anxiety)
  - C3: distinction between negative emotions (anger, anxiety, embarrassment, panic fear, tension)
- Tasks: view the videos and attribute one emotional label from a closed list
## Evaluation Result: Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>anger</th>
<th>anxiety</th>
<th>cheerful</th>
<th>embarra</th>
<th>panic fear</th>
<th>pride</th>
<th>relief</th>
<th>tension</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Cheerful</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Embarra</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>panic fear</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>26</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Pride</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Relief</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>Tension</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
</tbody>
</table>

Confusion between:
- anxiety and embarrassment
- cheerfulness and pride
- embarrassment and tension
- pride and relief
- tension and anxiety
Evaluation Result

- Recognition rate well above chance level (93% for anger – 37% for embarrassment)
- Subtly differentiated expressions like these of relief or of cheerfulness were recognized surprisingly well
- Distinctions happen between:
  - Positive expressions
  - Duchenne vs non-Duchenne smiles
  - Negative emotions
Listener Model

SEMAINE
THE SENSITIVE AGENT PROJECT
Interaction

- **Interaction** is by no means a one way communication channel between parties.

- Speaker and listener adapt their behaviors to each other
  - Speaker monitors addressee's attention and interest in what she has to say
  - Listener selects feedback behaviors to show the speaker that he is paying attention, agreeing, understanding, etc

- Tight dynamic coupling between both interactants
Listener model

- build a listener ECA able to display backchannel signals according to
  - its mental state and attitude:
    - assertive/not assertive,
    - believing/not believing,
    - interested/not interested
    - liking/not linking
  - its behavior distinctiveness

- define a set of backchannel signals (Backchannel Library) that users are able to interpret and understand
Backchannels functions [Allwood '93, Poggi '05, Heylen '07]:
- Signals: positive/negative; any combination
  - Agree / disagree, accept / refuse, interested / not interested, believe / disbelieve, understand / don't understand, like / dislike

Signals:
- vocalization
- head movement
- facial expression
- gaze
- posture
- gesture (ie mimicry)
SAL agents are characterized by

- distinct **emotional traits** along the valence and arousal dimensions
- **communicative functions**: agree, accept...
  - determined manually (or automatically, in the future, U Twente)
- distinct **baseline**
  - determined manually through the observation of videos of real people
Agent’s mental state:
- agent's intentions towards the content of the speaker's speech (agree/disagree, like/dislike...)
- emotional state

We link the agent's mental state to the SAL agents' emotional characteristics
- each SAL agent shows backchannel signals that are compatible with its emotional traits

For the moment: defined manually and/or via WoZ
SAL characters

- **Baseline:**
  - agent's modality preference
    - head, gaze, face, gesture and torso
  - agent's behaviour expressivity
    - defined by 6 parameters that

- **influence the quality of the agent's movements**

- **We determined manually a baseline for each SAL agent through the observation of videos of real people**
Examples from Naturalistic Database, Queens University of Belfast

Poppy

Spyke
LISTENER INTENT PLANNER

- computes the backchannels of the agent while listening to a user.
- processes:
  - *when* a backchannel should be triggered
  - probabilistic model based on user's signals analysis [Maatman '05, Ward & Tsukahara '00]
  - *which* communicative intentions are conveyed
  - according to the agent's mental state (if it agrees or disagrees, likes or dislikes...)

Diagram:
- Backchannel plannification
  - User's signals
  - User's estimated interest level
  - Agent's mental state
  - Trigger
  - Threshold
  - Selection
  - FML
LISTENER INTENT PLANNER

Watson

Video input (webcam)
detected user signals

Listener Intent Planning

backchannel signal (FML-APML)

Greta

MPEG4 animation
GRETA

- SAIBA compliant,
- standard languages
  - BML, FML-APML, MPEG-4 animation
- interactive applications
- open-source (Licence GPL)

Can be downloaded at: http://tsi.enst.fr/~pelachau/Greta/