15

An Emotional Talking Head for a Humoristic Chatbot

Agnese Augello[1], Orazio Gambino[1], Vincenzo Cannella[1], Roberto Pirrone[1], Salvatore Gaglio[1] and Giovanni Pilato[2]

[1]*DICGIM - University of Palermo, Palermo*
[2]*ICAR - Italian National Research Council, Palermo*
Italy

1. Introduction

The interest about enhancing the interface usability of applications and entertainment platforms has increased in last years. The research in human-computer interaction on conversational agents, named also chatbots, and natural language dialogue systems equipped with audio-video interfaces has grown as well. One of the most pursued goals is to enhance the realness of interaction of such systems. For this reason they are provided with catchy interfaces using humanlike avatars capable to adapt their behavior according to the conversation content. This kind of agents can vocally interact with users by using Automatic Speech Recognition (ASR) and Text To Speech (TTS) systems; besides they can change their "emotions" according to the sentences entered by the user. In this framework, the visual aspect of interaction plays also a key role in human-computer interaction, leading to systems capable to perform speech synchronization with an animated face model. These kind of systems are called Talking Heads.
Several implementations of talking heads are reported in literature. Facial movements are simulated by rational free form deformation in the 3D talking head developed in Kalra et al. (2006). A Cyberware scanner is used to acquire surface of a human face in Lee et al. (1995). Next the surface is converted to a triangle mesh thanks to image analysis techniques oriented to find reflectance local minima and maxima.
In Waters et al. (1994) the DECface system is presented. In this work, the animation of a wireframe face model is synchronized with an audio stream provided by a TTS system. An input ASCII text is converted into a phonetic transcription and a speech synthesizer generates an audio stream. The audio server receives a query to determine the phoneme currently running and the shape of the mouth is computed by the trajectory of the main vertexes. In this way, the audio samples are synchronized with the graphics. A nonlinear function controls the translation of the polygonal vertices in such a way to simulate the mouth movements. Synchronization is achieved by calculating the deformation length of the mouth, based on the duration of an audio samples group.
BEAT (Behavior Expression Animation Toolkit) an intelligent agent with human characteristics controlled by an input text is presented in Cassell et al. (2001). A talking head for the Web with a client-server architecture is described in Ostermann et al. (2000). The client application comprises the browser, the TTS engine, and the animation renderer. A

coarticulation model determines the synchronization between the mouth movements and the synthesized voice. The 3D head is created with a Virtual Reality Modeling Language (VRML) model.
LUCIA Tisato et al. (2005) is a MPEG-4 talking head based on the INTERFACE Cosi et al. (2003) platform. Like the previous work, LUCIA consists in a VRML model of a female head. It speaks Italian thanks to the FESTIVAL Speech Synthesis System Cosi et al. (2001). The animation engine consists in a modified Cohen-Massaro coarticulation model. A 3D MPEG-4 model representing a human head is used to accomplish an intelligent agent called SAMIR (Scenographic Agents Mimic Intelligent Reasoning) Abbattista et al. (2004). SAMIR is used as a support system to web users. In Liu et al. (2008) a talking head is used to create a man-car-entertainment interaction system. The facial animation is based on a mouth gesture database.
One of the most important features in conversations between human beings is the capability to generate and understand humor: "Humor is part of everyday social interaction between humans" Dirk (2003). Since having a conversation means having a kind of social interaction, conversational agents should be capable to understand and generate also humor. This leads to the concept of *computational humor*, which deals with automatic generation and recognition of humor.
Verbally expressed humor has been analyzed in literature, concerning in particular very short expressions (jokes) Ritchie (1998): a one-liner is a short sentence with comic effects, simple syntax, intentional use of rhetoric devices (e.g., alliteration, rhyme), and frequent use of creative language constructions Stock & Strapparava (2003). Since during a conversation the user says short sentences, one-liners, jokes or gags can be good candidates for the generation of humorous sentences. As a consequence, literature techniques about computational humor regarding one-liners can be customized for the design of a humorous conversational agent.
In recent years the interest in creating humorous conversational agents has grown. As an example in Sjobergh & Araki (2009) an humorous Japanese chat-bot is presented, implementing different humor modules, such as a database of jokes and conversation-based jokes generation and recognition modules. Other works Rzepka et al. (2009) focus on the detection of emotions in user utterances and puns generation.
In this chapter we illustrate a humorous conversational agent, called *EHeBby*, equipped with a realistic talking head. The conversational agent is capable to generate humorous expressions, proposing to the user riddles, telling jokes, ironically answering to the user. Besides, the chatbot is capable to detect, during the conversation with the user, the presence of humorous expressions, listening and judging jokes and react changing the visual expression of the talking head, according to the perceived level of humor. The chatbot reacts accordingly to the user jokes, adapting the expression of its talking head. Our talking head offers a realistic presentation layer to mix emotions and speech capabilities during the conversation with the user. It shows a smiling expression if it considers the user's sentence "funny", indifferent if it does not perceive any humor in the joke, or angry if it considers the joke in poor taste. In the following paragraphs we illustrate both the talking head features and the humorous agent brain.

2. EHeBby architecture

The system is composed by two main components, as shown in figure 1, a reasoner module and a Talking Head (TH) module. The reasoner processes the user question by means of the A.L.I.C.E. (Artificial Linguistic Internet Computer Entity)engine ALICE (2011), which has been extended in order to manage humoristic and emotional features in conversation. In

particular the reasoner is composed by a humoristic area, divided in turn in a humoristic recognition area and in a humoristic evocation area, and an emotional area. The first area allows the chatbot to search for the presence of humoristic features in the user sentences, and to produce an appropriate answer. Therefore, the emotional area allows the chatbot to elaborate information related to the produced answer and a correspondent humor level in order to produce the correct information needed for the talking head animation. In particular prosody and emotional information, necessary to animate the chatbot and express emotions during the speech process, are communicated to the Talking Head component. The TH system relies on a web application where a servlet selects the basis facial meshes to be animated, and integrates with the reasoner to process emotion information, expresses using ad hoc AIML (Artificial Intelligence Markup Language) tags, and to obtain the prosody that are needed to control animation. On the client side, all these data are used to actually animate the head. The presented animation procedure allows for considerable computational savings, so both plain web, and mobile client have been implemented.

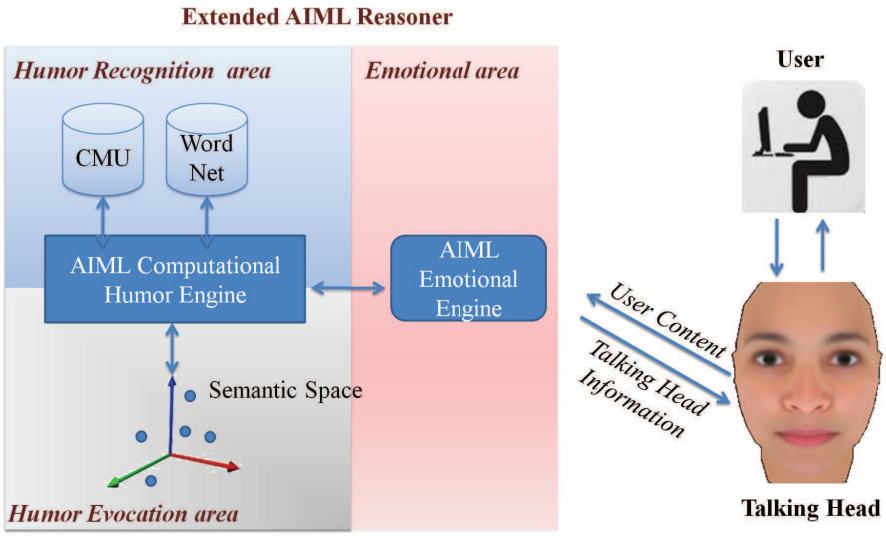


Fig. 1. EHeBby Architecture

3. EHeBby reasoner

The chatbot brain has been implemented using an extended version of the ALICE ALICE (2011) architecture, one of the most widespread conversational agent technologies.
The ALICE dialogue engine is based on a pattern matching algorithm which looks for a match between the user's sentences and the information stored in the chatbot knowledge base. Alice knowledge base is structured with an XML-like language called AIML (Artificial Intelligence Mark-up Language). Standard AIML tags make possible for the chatbot understanding user questions, to properly give him an answer, save and get values of variables, or store the context of conversation. The basic item of knowledge in ALICE is the *category*, which represents a question-answer module, composed a *pattern* section representing a possible user question, and a *template* section which identifies the associated chatbot answer. The AIML

reasoner has been extended defining *ad hoc* tags for computational humor and emotional purposes.
The chabot implements different features, by means of specific reasoning areas, shown in figure 1. The areas called *Humor Recognition Area* and *Humor Evocation Area*, deal with the recognition and generation of humor during the conversation with the user. A set of AIML files, representing the chatbot KB are processed during the conversation. Humor recognition and generation features are triggered when the presence of specific AIML tags is detected. The humorous tags are then processed by a *Computational Humor Engine*, which in turn queries other knowledge repositories, to analyze or generate humor during the conversation. In particular the *AIML Computational Humor Engine* exploits both WordNet MultiWordNet (2010) and the a pronouncing dictionary of the Carnegie Mellon University (CMU) CMU (2010) in order to recognize humorous features in the conversation, and a semantic space in oder to retrieve humorous sentences related to the user utterances. The area called *Emotional Area* deals with the association of chabot emotional reaction to the user sentences. In particular it allows for a binding of a conversation humor level with a set of *ad hoc* created emotional tags, which are processed by the *AIML Emotional Engine* in order to send the necessary information to the Talking Head. In particular in the proposed model we have considered only three possible humor levels, and three correspondent emotional expressions.

3.1 AIML KB

The AIML knowledge base of our humorous conversational agent is composed of four kinds of AIML categories:

1. the standard set of ALICE categories, which are suited to manage a general conversation with the user;
2. a set of categories suited to generate humorous sentences by means of jokes. The generation of humor is obtained writing specific funny sentences in the template of the category.
3. a set of categories suited to retrieve humorous or funny sentences through the comparison between the user input and the sentences mapped in a semantic space belonging to the evocative area. The chatbot answers with the sentence which is semantically closer to the user input.
4. a set of categories suited to to recognize an humorous intent in the user sentences. This feature is obtained connecting the chatbot knowledge base to other resources, like the WordNet lexical dictionary MultiWordNet (2010) and the CMU pronouncing dictionary CMU (2010).
5. a set of categories suited to generate emotional expressions in the talking head.

3.2 Humour recognition area

The humour recognition consists in the identification, inside the user sentences, of particular humorous texts features. According to Mihalcea and Strapparava Mihalcea et al. (2006) we focus on three main humorous features: alliteration, antinomy and adult slang. Special tags inserted in the AIML categories allows the chatbot to execute modules aimed to detect the humorous features.

3.2.1 Alliteration recognition module

The phonetic effect induced by the alliteration, the rhetoric figure consisting in the repetition of a letter, a syllable or a phonetic sound in consecutive words, captures the attention of

people listening it, often producing a funny effect Mihalcea et al. (2006). This module removes punctuation marks and stopwords (i.e. word that do not carry any meaning) from the sentence, and then analyzes its phonetic transcription, obtained by using the CMU dictionary CMU (2010). This technique is aimed at discovering possible repetitions of the beginning phonemes in subsequent words. In particular the module searches the presence of at least three words have in common the first one, the first two or the first three phonemes.
As an example the module consider the following humorous sentences:

```
Veni, Vidi, Visa: I came, I saw, I did a little shopping
Infants don't enjoy infancy like adults do adultery
```

detecting in the first sentence three words having the first phoneme in common, and in the second sentence two pairs of words having the first three phonemes in common. The words infancy and infants have the same following initial phonemes *ih1 n f ah0 n* while the words adultery and adults begin with the following phonemes *ah0 d ah1 l t*.

3.2.2 Antinomy recognition module

This module detects the presence of antinomies in a sentence has been developed exploiting the lexical dictionary WordNet. In particular the module searches into a sentence for:

- a direct antinomy relation among nouns, verbs, adverbs and adjectives;
- an extended antinomy relation, which is an antinomy relation between a word and a synonym of its antonym. The relation is restricted to the adjectives;
- an indirect antinomy relation, which is an antinomy relation between a word and an antonym of its synonym. The relation is restricted to the adjectives.

These humorous sentences contain antinomy relation:

```
A clean desk is a sign of a cluttered desk drawer
Artificial intelligence usually beats real stupidity
```

3.2.3 Adult slang recognition module

This module analyzes the presence of adult slang searching in a set of pre-classified words. As an example the following sentences are reported:

```
The sex was so good that even the neighbors had a cigarette
Artificial Insemination: procreation without recreation
```

3.3 Humor evocation area

This area allows the chatbot to evocate funny sentences that are not directly coded as AIML categories, but that are encoded as vectors in a semantic space, created by means of Latent Semantic Analysis (LSA) Dumais & Landauer (1997). In fact, if none of the features characterizing a humorous phrase is recognized in the sentence through the humor recognition area, the user question is mapped in a semantic space. The humor evocation area then computes the semantic similarity between what is said by the user and the sentences encoded in the semantic space; subsequently it tries to answer to the user with a funny expression which is conceptually close to the user input. This procedure allows to go beyond the rigid pattern-matching rules, generating the funniest answers which best semantically fit the user query.

3.3.1 Semantic space creation

A semantic representation of funny sentences has been obtained mapping them in a semantic space. The semantic space has been built according to a Latent Semantic Analysis (LSA) based approach described in Agostaro (2005)Agostaro (2006). According to this approach, we have created a semantic space applying the truncated singular value decomposition (TSVD) on a $m \times n$ co-occurrences matrix obtained analyzing a specific texts corpus, composed of humorous texts, where each *(i, j)-th* entry of the matrix represents square root of the number of times the *i-th* word appears in the *j-th* document.
After the decomposition we obtain a representation of words and documents in the reduced semantic space. Moreover we can automatically encode in the space new items, such as sentences inserted into AIML categories, humorous sentences and user utterances. In fact, a vectorial representation can be obtained evaluating the sum of the vectors associated to words composing each sentence.
To evaluate the similarity between two vectors v_i and v_j belonging to this space according to Agostaro et al. we use the following similarity measure Agostaro (2006):

$$sim\left(v_i, v_j\right) = \begin{cases} \cos^2\left(v_i, v_j\right) & \text{if } \cos\left(v_i, v_j\right) \geq 0 \\ 0 & otherwise \end{cases} \tag{1}$$

The closer this value is to 1, the higher is the similarity grade. The geometric similarity measure between two items establishes a semantic relation between them. In particular given a vector **s**, associated to a user sentence s, the set CR(s) of vectors sub-symbolically conceptually related to the sentence s is given by the q vectors of the space whose similarity measure with respect to **s** is higher than an experimentally fixed threshold T.

$$CR(s) = v_i | sim(s, v_i) > T \quad with \quad i = 1 \ldots q \tag{2}$$

To each of these vectors will correspond a funny sentence used to build the space. Specific AIML tags called *relatedSentence* and *randomRelatedSentence* allow the chatbot to query the semantic space to retrieve respectively the semantically closer riddle to the user query or one of the most conceptually related riddles. Tha chatbot can also improve its own AIML KB mapping in the evocative area new items like jokes, riddles and so on introduced by the user during the dialogue.

3.4 Emotional area

This area is suited to the generation of emotional expressions in the Talking Head. Many possible models of emotions have been proposed in literature. We can distinguish three different categories of models. The first one includes models describing emotions through collections of different dimensions (intensity, arousal, valence, unpredictability, potency, ...). The second one includes models based on the hypothesis that a human being is able to express only a limited set of primary emotions. All the range of the human emotions should be the result of the combination of the primary ones. The last category includes mixed models, according to which an emotion is generated by a mixture of basic emotions parametrized by a set of dimensions. One of the earlier model of the second category is the model of Plutchik Ekman (1999). He listed the following primary emotions: acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise. Thee emotions can be combined to produce secondary emotions, and in their turn those can be combined to produce ternary emotions. Each emotion can be characterized by an intensity level. After this pioneering model, many other similar

models have been developed. An interesting overview can be found in Ortony (1997). Among the models cited in Ortony (1997), the model by Ekman have been chosen as basis for our work. According to Ekman's model, there are six primary emotions: anger, disgust, fear, joy, sadness, surprise. We have developed a reduced version of this model, including only three of the listed basic emotions: anger, joy, sadness. We selected them as basis to express humor.
At this moment our agent is able to express one of these three emotions at a time, with a variable intensity level. The emotional state of the agent is represented by a couple of values: the felt emotion, and its corresponding intensity. The state is established on the basis of the humor level detected in the conversation. As just said, there are only three possible values for the humor level. These levels have to correspond to a specific emotion in the chatbot, with an intensity level. The correspondence should to be defined according to a collection of psychological criteria. At this moment, the talking head has a predefined behavior for its humorist attitude useful to express these humor levels. Each level is expressed with a specific emotion at a certain intensity level. This emotional patterns represent a default behavior for the agent. The programmer can create a personal version of emotional behavior defining different correspondences between humor levels and emotional intensities. Moreover, he can also program specialized behaviors for single steps of the conversation or single witticisms, as exceptions to the default one.
The established emotional state has to be expressed by prosody and facial expressions. Both of them are generated by the *Emotional Area*. This task is launched by *ad hoc* AIML tags.

4. EHeBby talking head

Our talking head is conceived to be a multi-platform system that is able to speak several languages, so that various implementations have been realized. In what follows the different components of our model are presented: model generation, animation technique, coarticulation, and emotion management.

4.1 Face model generation

The FaceGen Modeler FaceGen (2010) has been used to generate graphic models of the 3D head. FaceGen is a special tool for the creation of 3D human heads and characters as polygon meshes. The facial expressions are controlled by means of numerical parameters. Once the head is created, it can be exported as a Wavefront Technologies .obj file containing the information about vertexes, normals and textures of the facial mesh. The .obj is compliant with the most popular high level graphics libraries such as Java3D and OpenGL. A set of faces with different poses is generated to represent a "viseme", which is related to a phoneme or a groups of phonemes. A phoneme is the elementary speech sound, that is the smallest phonetic unit in a language. Indeed, the spoken language can be thought as a sequence of phonemes. The term "viseme" appeared in literature for the first time in Fischer (1968) and it is equivalent to the phoneme for the face gesture. The viseme is the facial pose obtained by articulatory movements during the phoneme emission. Emotional expressions can be generated by FaceGen also. In our work we have implemented just 4 out of the Ekman basic emotions Ekman & Friesen (1969): joy, surprise, anger, sadness. The intensity of each emotion can be controlled by a parameter or mixed to each other, so that a variety of facial expressions can be obtained. Such "emotional visemes" will be used during the animation task. Some optimizations can be performed to decrease amount of memory necessary to store such a set of visemes. Just the head geometry can be loaded from the .obj file. Lights and virtual camera parameters are set within the programming code. A part of the head mesh can be loaded as a background mesh and after the 3 sub-meshes referred to face, tongue and teeth are loaded.

Indeed, these 3 parts of the head are really involved in the animation. The amount of vertexes can be reduced with a post-processing task with a related decrease of quality, which is not severe if this process involves the back and top sides of the head. Moreover, for each polygon mesh a texture should be loaded, but all the meshes can use the same image file as texture to save memory. A basic viseme can provide both the image texture and the texture coordinates to allow the correct position of the common texture for the other ones.

4.2 Animation

The facial movements are performed by morphing . Morphing starts from a sequence of geometry objects called "keyframes". Each keyframe's vertex translates from its position to occupy the one of the corresponding vertex in the subsequent keyframe. For this reason we have to generate a set of visemes instead of modifying a single head geometric model. Such an approach is less efficient than an animation engine able to modify the shape according to facial parameters (tongue position, labial protrusion and so on) but it simplifies strongly the programming level: First, the whole mesh is considered in the morphing process, and efficient morphing engines are largely present in many computer graphics libraries. Various parameters have to be set to control each morphing step between two keyframes, i.e. the translation time. In our animation scheme, the keyframes are the visemes related to the phrase to be pronounced but they cannot be inserted in the sequence without considering the facial coarticulation to obtain realistic facial movements. The coarticulation is the natural facial muscles modification to generate a succession of fundamental facial movements during phonation. The Löfqvist gestural model described in Löfqvist (1990) controls the audio-visual synthesis; such a model defines the "dominant visemes", which influence both the preceding and subsequent ones. Each keyframe must be blended dynamically with the adjacent ones. The next section is devoted to this task, showing a mathematical model for the coarticulation.

4.2.1 Cohen-Massaro model

The Cohen-Massaro model Cohen & Massaro (1993) computes the weights to control the keyframe animation. Such weights determine the vertexes positions of an intermediate mesh between two keyframes. It is based on the coarticulation, which is the influence of the adjacent speech sounds to the actual one during the phonation. Such a phenomenon can be also considered for the interpolation of a frame taking into account the adjacent ones in such a way that the facial movement appear more natural. Indeed, the Cohen-Massaro model moves from the work by Löfqvist, where a speech segment shows the strongest influence on the organs of articulation of the face than the adjacent segments. Dominance is the name given to such an influence and can be mathematically defined as a time dependent function. In particular, an exponential function is adopted as the dominance function. The dominance function proposed in our approach is simplified with respect to the original one. Indeed, it is symmetric. The profile of a dominance function for given speech segment s and facial parameter p is expressed by the following equation:

$$D_{sp} = \alpha \cdot exp(-\theta\,|\tau|^c) \tag{3}$$

where α is the peak for $\tau = 0$, θ and c control the function slope and τ is the time variable referred to the mid point of the speech segment duration. In our implementation we set $c = 1$ to reduce the number of parameters to be tuned. The dominance function reachs its maximum value (α) in the mid point of speech segment duration, where $\tau = 0$. In the present approach, we assume that the time interval of each viseme is the same of the duration of the respective phoneme. The coarticulation can be thought as composed by two sub-phenomenons: the pre- and post- articulation. The former consists in the influence of the present viseme on the

facial parameters to be used for interpolating the preceding keyframe towards the present one (τ<0). The latter regards the dominance of the next viseme on the parameters used morph the present keyframe towards the next one (τ>0). Our implementation doesn't make use of an animation engine to control the facial parameters (labial opening, labial protrusion and so on) but the interpolation process acts on the translation of all the vertexes in the mesh. The prosodic sequence S of time intervals $[t_{i-1},t_i[$ associated to each phoneme can be expressed as follows:

$$S = \{f_1 \in [0, t_1[\,; f_2 \in [t_1, t_2[\,; \ldots ; f_n \in [t_{n-1}, t_n[\} \tag{4}$$

A viseme is defined "active" when t falls into the corresponding time interval. The preceding and the following visemes are defined as "adjacent visemes". Due to the negative exponential nature of the dominance function, just the adjacent visemes are considered for computing weights. For each time instant, 3 weights must be computed on the basis of the respective dominance functions of 3 visemes at a time. The weights are computed as follows:

$$w_i(t) = D_i(t) = \alpha_i \cdot \exp(-\theta_i \cdot |t - \tau_i|) \tag{5}$$

where τ_i the mid point of the i-th time interval. The w_i must be normalized:

$$w_i^{'}(t) = \frac{w_i(t)}{\sum_{j=-1}^{+1} w_{i-j}(t)} \tag{6}$$

so that for each time instant the coordinates of the interpolating viseme vertexes $v_{int}^{(l)}(t) \in \{V_{int}(t)\}$ will be computed as follows:

$$v_{int}^{(l)}(t) = \sum_{k=i-1}^{i+1} w_i^{'}(t) \cdot v_k^{(l)}(t) \tag{7}$$

where the index l indicates corresponding vertexes in all the involved keyframes.
Our implementation simplifies also this computation. It is sufficient to determine the result of the coarticulation just for the keyframes, because the interpolation is obtained using directly the morphing engine with a linear control function. Once the dominance functions are determined, each coarticulated keyframe is computed and its duration is the same as in the corresponding phoneme.

4.2.2 Diphthongs and dominant visemes

A sequence of two adjacent vowels is called diphthong. The word "euro" contains one diphthong. The vowels in a diphthong must be visually distinct as two separate entities. The visemes belonging to the vowels in a diphthong mustn't influence each other. Otherwise, both the vowel visemes wouldn't be distinguishable due to their fusion. In order to avoid this problem, the slope of the dominance function belonging to each vocal viseme in a diphthong must be very steep (see Fig.2). On the contrary, the sequence vowel-consonant requires a different profile of the dominant function. Indeed, the consonant is heavily influenced by the preceding vowel: a vowel must be dominant with respect to the adjacent consonants, but not with other vowels. As shown in Fig.3, the dominance of a vowel with respect to a consonant is accomplished with a less steep curve than the consonant one.

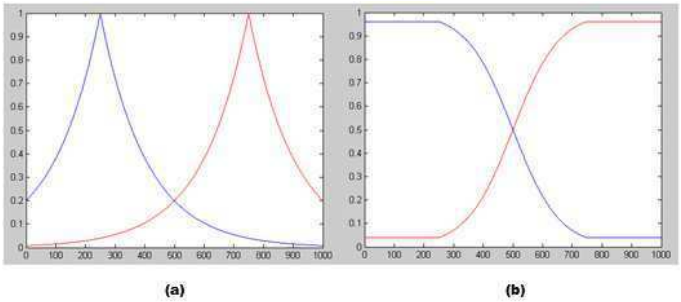


Fig. 2. The dominance function for the diphtong case (a) and the weights diagram (b) for the diphthong case

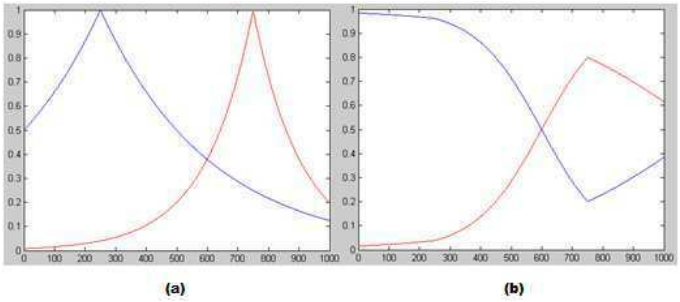


Fig. 3. The same of Fig.2 for the vowel-consonant case.

4.3 The emotional talking head

Emotions can be considered as particular visemes, called emotional visemes. They must be "mixed" with the phonetic visemes to express an emotion during the facial animation. Such a process can be performed in two different ways. FaceGen can generate also facial modification to express an emotion, so a phonetic viseme can be modified using FaceGen to include an emotion. As result, different sets of modified phonetic visemes can be produced. Each of them are different both as type and intensity of a given emotion. Such a solution is very accurate but it requires an adequate amount of memory and time to create a large emotional/phonetic visemes database. The second approach considers a single emotional viseme whose mesh vertexes coordinate are blended with a viseme to produce a new keyframe. Even though such a solution is less accurate than the previous one, it is less expensive on the computational side, and allows to include and mix "on the fly" emotional and phonetic visemes at run-time.

4.4 Audio streaming synchronization

Prosody contains all the information about the intonation and duration to be assigned to each phoneme in a sentence. In our talking head model, the prosody is provided by Espeak espeak (2010), a multilanguage and multiplatform tool that is able to convert the text into a .pho prosody file. The Talking Head is intrinsically synchronized with the audio streaming because the facial movements are driven by the .pho file, which determines the phoneme (viseme) and its duration. Espeak provides a variety of options to produce the prosody for the language and speech synthesizer to use. As an example it can generate a prosody control for the couple Italian/Mbrola, which is a speech synthesizer based on concatenation of diphones. It takes as

input a list of phonemes, together with prosodic information (duration and intonation), and produces an audio file .wav which is played during the facial animation.

5. Some example of interaction

5.1 Example of humorous sentences generation

The following is an example of an humorous dialogue

```
User: What do you think about robots?
EHeBby: Robots will be able to buy happiness,
         but in condensed chip form!!
```

obtained writing an *ad hoc* AIML category:

```
<category>
  <pattern>WHAT DO YOU THINK ABOUT ROBOTS</pattern>
  <template>Robots will be able to buy happiness,
          but in condensed chip form!!
  </template>
< /category>
```

The *pattern* delimits what the user can say. Every time the *pattern* is matched, the corresponding *template* is activated.

5.2 Example of humor recognition

The recognition of humorous sentences is obtained using specific tag inserted into the template, as shown in the following categories:

```
<category>
  <pattern>CAN I TELL YOU A JOKE</pattern>
  <template>Yes you can</template>
< /category>

<category>
  <pattern>*</pattern>
    <that>YES YOU CAN</that>
  <template>
    <srai> <humorlevel> <star/> </humorlevel> <srai>
  </template>
< /category>
```

The second category is activated if the previous answer of the chatbot was "Yes you can" (according to the *that* tag behavior), and the *humorlevel* tag evaluates the level of humor of the sentence matched with the * wildcard (i.e. what the user said). The humor level can assume three different values, *low*, *medium* and *high*. Depending on the humor level value, the category will recursively call, by means of the *srai* tag, another category, which will explicit an emotional tag, including the information needed to the talking head expression, and a *prosody* tag to produce the prosody file. In particular we have extended the AIML language to include three emotional tags *joy*, *anger* and *sadness*. Each of them also includes a mandatory *intensity* attribute. The value assigned to the attribute is a measure of how much that emotion combines to produce the overall emotional state of the chatbot. The called tag link the proper

.m3g emotional file to the talking head, with the corresponding intensity level. In particular a low humor level will be associated to the *anger* tag, a medium level to a *sadness* tag and finally a high level to a *joy* tag. In a default behavior the emotional intensity for each of these emotions is fixed to a specific value. The following example shows the category related to a high humor level.

```
<category>
  <pattern>HIGH *</pattern>
  <template> <think>
   <prosody>
<star/>
</prosody>
    </think>
<joy intensity="080" />
  </template>
</category>
```

This is a dialogue example that shows the chatbot reaction to a funny joke, while in figure 4 is shown the corresponding TH expression:

```
User: Doctor gave a man six months to live. The man couldn't pay
      his bill, so he gave him another six months.
EHeBby: Ehehe! You are very funny!! Good.. your joke is fantastic!!
```

Fig. 4. TH reaction to a funny joke

This other example shows the chatbot reaction to a non-humorous user sentence (see figure 5:

```
User: The book is on the table.
EHeBby: I don't enjoy myself.. !! Your joke is stupid!!
```

The last example shows the chatbot reaction to a rude joke, while in figure 6 is shown a disgusted talking head expression:

```
User: 100,000 sperm and you were the fastest?
EHeBby: What a disgusting joke!!
```

Fig. 5. TH reaction to a non-humorous sentence

Fig. 6. TH reaction to a rude joke

5.3 AIML categories for the evocation of humorous sentences

The semantic space of the humorous chatbot evocative area has been created by using 3835 funny sentences downloaded from internet one-liners repositories onliners (2011) as well as the standard categories stored in the AIML knowledge base. Their retrieval is obtained through the definition and the use of specifically created AIML tags. As an example, the *relatedSentence* tag in the category template triggers the comparison between the user inputs with the sentences mapped in the semantic space belonging to the evocative area by choosing the sentence semantically closer to the user question. The *randomRelatedSentence* is defined in order to make possible for the chatbot to answer with one of the sentences which is semantically related to the user input.
Below we report the following example of category:

```
<category>
  <pattern>What do you think about *</pattern>
  <template>
   <randomRelatedSentence> <star/></randomRelatedSentence>
```

```
  </template>
< /category>
```

If the user question is:

```
User: What do you think about marriage?
```

The chat-bot will answer randomically in one of the following ways:

```
EHeBby: Marriage is an expensive way of getting your laundry done
        for free.
EHeBby: Benjamin Franklin said "'Keep your eyes wide open before
        marriage, half shut afterwards"'.
EHeBby: Love may be blind but marriage is a real eye-opener.
```

While the standard Alice chat-bot answers:

```
Alice: The state of being married; relation between husband and wife;
       married life.
```

The following dialogue shows a comparison between the answers given with the traditional pattern matching by the chat-bot Alice and the answers given by the evocative chat-bot using the <relatedSentence> tag.

```
User: You are ignorant.
Alice: We can learn new things.
EHeBby: Everybody is ignorant, only on different subjects
User: what is the difference between woman and men?
Alice: I think you already know the answer. I will search for it.
EHeBby: Men always want to be a woman's first love. Women have a more
        subtle instinct; what they like to be is a man's last romance.
User: See you tomorrow.
Alice: Goodbye.
EHeBby: Tomorrow is a mistery...
User: Do you enjoy working?
Alice: I enjoy working 24 hours every day.
EHeBby: I only work to enjoy when I am not working
User: My computer has the windows operating system.
Alice: Where did he get it?
EHeBby: What do Windows and a handgun have in common? Both are
        harmless while not loaded.
User:Are you a pessimist or an optimist?
Alice: Am I a pessimist or an optimist? Yes.
EHeBby: A pessimist is one who makes difficulties of his
        opportunities and an optimist is one who makes
        opportunities of his difficulties. (Harry Truman)
```

5.4 AIML categories for targeting

The humorous chatbot is able to update its own sub-symbolic knowledge through a targeting process, which maps new acquired riddles in the semantic space. Targeting is obtained by means of the ad-hoc created AIML tag *addRiddle*, as shown in the following chunk of AIML code:

```
<category>
  <pattern>Listen this joke *</pattern>
  <template> <humorlevel><star/> </humorlevel>
    <think>
    <addRiddle> <star/> </addRiddle>
    < /think>
  < /template>
< /category>
```

The *addRiddle* tag has the effect to code a sentence introduced by the user as a vector in the semantic space by means of the folding-in procedure. The entire procedure is hidden to the user by the *think* tag. In this manner, the user will see only the chat-bot reaction to his joke.

6. Results

In order to validate the humor recognition, two appropriate data set have been created. The former, called *DataSet1*, is composed of 100 humorous phrases extracted by ad-hoc selected web sites, characterized by the presence of a percentual of humoristic features, as shown in table 1.

Features	Percentual in DataSet1
Alliteration	45%
Direct Antinomy	10%
Extended Antinomy	6%
Indirect Antinomy	5%
Adult Slang	6%
No Feature	28%

Table 1. Humoristic Features Distribution in DataSet1

The latter, called *DataSet2*, is composed of 200 phrases, equally distributed between humorous and non-humorous sentences, where the non-humorous examples are always extracted from Internet and chosen among titles of newspapers, proverbs and sets definitions. The implemented algorithms well performed in humor recongition, as shown in table 2.

DataSet	Correct Classification	No Recognition
DataSet1	66%	34%
DataSet2	73%	27%

Table 2. Humor Recognition in the analyzed DataSets

The results can be considered satisfactory. Moreover if with the humor recognition area cannot identify the humorous level of sentences written by the user, the dialogue can continue in a funny way. In fact, the chatbot exploits the semantic space, which allows retrieving amusing sentences so that the dialogue can continue in a funny manner.
The following results show a set of funny answers extracted by the chatbot exploiting the semantic space, and their semantic closeness to the user entry, compared to the answers given by standard Alice chatbot: Example 1:

```
User: What do you think about chocolate?
Alice: Cocoa flavoring used in a variety of food products.
EHeBby:
```

```
0.94 A quarter-ounce of chocolate = four pounds of fat.
    (Slick's Second Universe Law)
0.82 Equal amounts of dark chocolate and white chocolate,
     is a balanced diet.

User: What can you say about children?
Alice: I will think about it and get back to you.
EHeBby:
0.68 Praise your children openly, reprove them secretly.
    (W. Cecil)
0.68 When speaking to your children: say what you mean,
     mean what you say,mbut don't be mean when you say it.

User: What do you think of life?
Alice: Life is the opposite of death.
EHeBby:
0.987 Life is an echo.
0.986 Life is a roller coaster.

User: Have you ever seen Santa Claus?
Alice: No I haven't.
EHeBby:
0.983 The Bermuda Triangle got tired of warm weather.
      It moved to Finland. Now Santa Claus is missing.
0.746 There are 4 Santa Claus stages:1) When you believe in Santa.
      2) When you donŠt believe in Santa.3) When you are Santa.
      4) When you look like Santa.ve not seen it. What's it like?
```

7. Conclusion

A complete framework for an emotional talking head able to manage humor while conversing with the user has been presented along with its implementation. The whole architecture relies on a suitable AIML-based chatbot, and an animation engine for the talking head.

The chatbot reasoner module is based on an extended AIML architecture where both humor, and emotions can be dealt with using suitable tags. A computational humor engine is able both to detect and to generate humorous sentences. Humor detection relies on the presence of alliteration, antinomy, or adult slang in the user's utterances, which are searched for using suitable thesauri like CMU and WordNet. Generation of humor makes use of a LSA based semantic space where humorous sentences have been placed along with the conversation topics.

The system can also select the control parameters for the animation engine, regarding the mesh deformation due to the emotive state to be expressed, the prosody for controlling speech generation, and the coarticulation model that is used to morph a set of key visemes related to phonemes.

The whole system has been tested on the humor recognition task with satisfactory results. However, our system is currently under development and much work has to be done in order to improve the whole architecture. Humor recognition algorithms can be enhanced, in order to capture different grades of humor, and to fully exploit the different levels of intensity in Talking Head emotional expressions.

The emotion database has to be completed al least with all the six Ekman basic emotions. Moreover, the most recent emotion models Ekman (1999) use more than six basis emotional

states so we plan to investigate these models using compositions of our current emotion database visemes. Finally, also web technology is going along with emotions management, and new standards like the W3C EmotionML emotionML (2011) are going to be released. In consideration of this, we plan to modify our AIML extensions towards these standards in order to enable interoperability with other emotion-oriented web systems.

8. References

Abbattista F, Catucci G, Semeraro G, Zambetta F. SAMIR: A Smart 3D Assistant on the Web. Psychnology Journal, 2(1):43-60, 2004.

F. Agostaro, A. Augello, G. Pilato, G. Vassallo, S. Gaglio. A Conversational Agent Based on a Conceptual Interpretation of a Data Driven Semantic Space. Lecture Notes in Artificial Intelligence, Springer-Verlag GmbH, vol. 3673/2005, pp 381-392, ISSN: 0302-9743.

Francesco Agostaro. Metriche per l'Analisi della Semantica Latente finalizzata ai Modelli del Linguaggio. PhD thesis, Università degli Studi di Palermo. Dipartimento di Ingegneria Informatica, 2006. Supervisor: Prof. S. Gaglio.

Alice website: www.alicebot.org

Cassell J, Vilhjálmsson H H, Bickmore T. BEAT: the Behavior Expression Animation Toolkit. s.l. : Proceedings of the 28th annual conference on Computer graphics and interactive techniques (2001), pp. 477-486. doi:10.1145/383259.383315

CMU Dictionary: (2010) http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Cohen, M. M., and Massaro, D. W. (1993) Modeling coarticulation in synthetic visual speech. In N. M. Thalmann and D. Thalmann (Eds.) Models and Techniques in Computer Animation. pp 139-156. Springer-Verlag.

Cosi P., Tesser F., Gretter R., Avesani C. (2001). FESTIVAL Speaks Italian. In Proceedings Eurospeech 2001, Aalborg, Denmark, September 3-7 2001 (pp. 509-512)

Cosi P., Fusaro A., Tisato G. (2003). LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In Proceedings of Eurospeech 2003, Geneva, Switzerland, September 1-4, 2003 (pp. 2269-2272).

Heylen Dirk. (2003)Talking Head Says Cheese! Humor as an impetus for Embodied Conversational Agent Research. CHI-2003 WorkShop: Humor Modeling In the Interface.

Dumais Susan T. Thomas K. Landauer (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104(2)

Pawel Dybala, Michal Ptaszynski, Jacek Maciejewski, Mizuki Takahashi, Rafal Rzepka and Kenji Araki. Multiagent system for joke generation: Humor and emotions combined in humanagent conversation. Journal of Ambient Intelligence and Smart Environments 2 (2010) 31-48. DOI 10.3233/AIS-2010-0053. IOS Press

Ekman, P., and Friesen, W. V (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica, 1, 49Ð98.

Ekman, P., Basic Emotions, in Dalgleish, T., Power, M., Handbook of Cognition and Emotion, Sussex, UK: John Wiley and Sons, (1999)

http://www.w3.org/TR/2011/WD-emotionml-20110407/

espeak.sourceforge.net/download.html

Singular Inversions Inc., (2010) FaceGen Modeller: www.facegen.com/modeller.htm

G, Fisher C. Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 11(4):796-804.

Kalra P, Mangili A., Magnetat-Thalmann N, Thalmann D. Simulation of facial muscle actions based on rational free form deformations. SCA '06 Proceedings of the 2006 ACM SIGGRAPH Eurographics symposium on Computer animation ISBN:3-905673-34-7

Löfqvist, A. (1990) Speech as audible gestures. In W.J. Hardcastle and A. Marchal (Eds.) Speech Production and Speech Modeling. Dordrecht: Kluwer Academic Publishers, 289-322.

Lee Y, Terzopoulos D, Waters K. Realistic modeling for facial animation. Proc. ACM SIGGRAPH'95 Conference, Los Angeles, CA, August, 1995, in Computer Graphics Proceedings, Annual Conference Series, 1995, 55-62.

Liu K, Ostermann J. Realistic Talking Head for Human-Car-Entertainment Services. IMA 2008 Informationssysteme für mobile Anwendungen, GZVB e.V. (Hrsg.), pp. 108-118, Braunschweig, Germany

Mihalcea R. and C.Strapparava. (2006) Learning to laugh (automatically): Computational Models for Humor Recognition. Computer Intelligence, Volume 22, 2006

MultiWordNet (2010): http://multiwordnet.itc.it/english/home.php

http://www.oneliners-and-proverbs.com/ and
http://www.bdwebguide.com/jokes/1linejokes-1.htm.

Ortony, A. and Turner, T. J. (1990) What's basic about basic emotions? In Psychological Review, Vol. 97, pp. 315–331, ISSN 0033-295X

Ostermann J, Millen D. Talking heads and synthetic speech: an architecture for supporting electronic commerce.. ICME 2000. 2000 IEEE International Conference on Multimedia and Expo, 2000. 71 - 74 vol.1 ISBN: 0-7803-6536-4

Ritchie G. (1998). Prospects for Computational Humour. Pp. 283-291 in Proceedings of 7th IEEE International Workshop on Robot and Huma Communication (ROMAN-98), Takamatsu, Japan, October 1998.

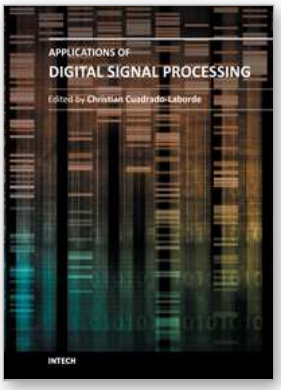
Rafal Rzepka, Wenhan Shi, Michal Ptaszynski, Pawel Dybala, Shinsuke Higuchi, and Kenji Araki. 2009. Serious processing for frivolous purpose: a chatbot using web-mining supported affect analysis and pun generation. In Proceedings of the 14th international conference on Intelligent user interfaces (IUI '09). ACM, New York, NY, USA, 487-488. DOI=10.1145/1502650.1502728 http://doi.acm.org/10.1145/1502650.1502728

Jonas Sjobergh and Kenji Araki. A Very Modular Humor Enabled Chat-Bot for Japanese. Pacling 2009

Stock O. and C.Strapparava. (2003). Getting serious about the development of computational humor. In proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-03) pp. 59-64, Acapulco, Mexico,2003.

Tisato G, Cosi P, Drioli C, Tesser F. INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. In CD Proceedings INTERSPEECH 2005 Lisbon, Portugal, 2005 (pp. 781-784).

Waters K, Levergood T M. An automatic lip-synchronization algorithm for synthetic faces. s.l. : MULTIMEDIA '94 Proceedings of the second ACM international conference on Multimedia ISBN:0-89791-686-7



Applications of Digital Signal Processing

Edited by Dr. Christian Cuadrado-Laborde

ISBN 978-953-307-406-1
Hard cover, 400 pages
Publisher InTech
Published online 23, November, 2011
Published in print edition November, 2011

In this book the reader will find a collection of chapters authored/co-authored by a large number of experts around the world, covering the broad field of digital signal processing. This book intends to provide highlights of the current research in the digital signal processing area, showing the recent advances in this field. This work is mainly destined to researchers in the digital signal processing and related areas but it is also accessible to anyone with a scientific background desiring to have an up-to-date overview of this domain. Each chapter is self-contained and can be read independently of the others. These nineteenth chapters present methodological advances and recent applications of digital signal processing in various domains as communications, filtering, medicine, astronomy, and image processing.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Agnese Augello, Orazio Gambino, Vincenzo Cannella, Roberto Pirrone, Salvatore Gaglio and Giovanni Pilato (2011). An Emotional Talking Head for a Humoristic Chatbot, Applications of Digital Signal Processing, Dr. Christian Cuadrado-Laborde (Ed.), ISBN: 978-953-307-406-1, InTech, Available from: http://www.intechopen.com/books/applications-of-digital-signal-processing/an-emotional-talking-head-for-a-humoristic-chatbot

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821