Cognitive Emotion in Speech Interactive Games

Tarashankar Rudra, Terry Bossomaier

Abstract—**This paper falls into the category of Human Computer Interaction. Here the focus lies in injecting emotions into the non-player characters (NPC) in reaction to the emotion of the player in computer games. The idea of injecting emotion into games has been a popular topic for discussion in the Games industry today. The paper reviews the emotional side of games from both the game side and the human player side, considering speech style and content, facial expressions and gestures. The emotional content is encapsulated within an XML and an associated finite state automaton.**

Index Terms— **Emotioneering^{TM}, prosody, cognition, XML.**

I. Introduction

The reasons for putting emotion in games are they get better player buzz, higher praise in the press and the development team put more passion into their work [1]. Emotion and cognitive models enter computer games both from the side of the human player and from the NPC's in the game itself. The human side is a matter of experimentation and modeling. We need to observe players, encode their actions and build models of how they act. Kavakli [2] has advocated the use of protocol analysis for developing an understanding of human cognitive models. Capturing emotion in human players is also an interesting area of research, which goes far beyond the computer games industry to monitoring speech for many purposes such as fraud detection, currently being investigated as a methodology by car insurers. Emotional cues arise from facial expressions, speech, and gestures. Each of these domains is an active research area. The thrust of this paper is from the other side. How do we build a cognitive model for an NPC in a game and, then, allow its emotional states to be communicated to the human player (and possibly in advanced games, to other NPCs)? There are two stages. Firstly, as a result of perceptual input and internal goals and states, an emotion is generated in the NPC. It then has to be transmitted to an external representation such as the manner of speech. This paper, then, falls into two such parts. In the first we propose a framework for the cognitive modelling of emotion in NPCs. In the second, we review the state of the art in emotional representation. There are many representations we could have chosen in this framework. Since the model has to ultimately be represented in software, UML [3] might appear a good choice. But it tends to be too low-level and is not comprehensible to non-programmers. A better choice is XML [4], now a universal interchange format across many domains, from mathematics (MathML), spatial information (GML) to predictive modeling and data mining (PMML). The advantage lies in being able to define a precise general specification as an XML Schema, which can be converted automatically to code, using, say, the Java SAXB or DOM models. Yet at the same time the individual examples (documents in XML) can contain specific textual information. In fact the scripting process for the actions of characters in a game morphs smoothly into an XML description and subsequent software agent.

The research area of emotion identification is also known as affect sensing, sentiment classification and emotion recognition [5]. Emotion is known to affect how people remember, learn and how they make decisions [6]. Earlier in 1987, Klatt had suggested a synthesis-by-rule concept in which speech is produced by matching phonetic and linguistic rules deduced from analysed speech data, consisting of phoneme sequences and prosodic features [7]. Emotive texts are an interesting intermediate between truly spontaneous emotion and neutral sentences overlaid with emotional expression [8].

Research in the field of robots to substitute for pets has a strong emphasis on the expression of their own emotions [9]. The problem of an emoting computer becomes more evident when we allow the human user to speak to the computer using speech recognition (that pays attention to what is said and ignores how it is said) in the "front end", which is unaware of the emotional state of the human user [10]. It may be of immense benefit in the field of human computer interaction where users receive bad news from a computer with apparent empathy and sympathy [11], Internet chatting becomes more enjoyable with the display of appropriate emotion tags along with text [12], more e.g. [5].

Emotioneering^{TM} is a set of techniques "that can create for a player or participant, a breadth and depth of emotions in games or other interactive experience, or that can immerse a game player in a world or role" [1]. Computer games move fast in real time [13] and the response from the game characters are expected to be immediate so as to maintain the flow of the game.

II. Cognitive modelling of emotion

The history of the modern empirical/cognitive psychology dates back to 1885 to the work of Ebbinghaus who set himself tasks of learning lists of nonsense words in order to study the processes underlying memory [14]. Cognition is a complex system that is highly interactive, where a weaving

This work was supported in full by the Australian Research Council under Grant No. LP0216837 (We wish to thank Ratbag Games Australia to sponsor this project).

T. Rudra is with the Department of Computer Science, Macquarie University, NSW – 2109, Australia. (e-mail: trudra@ comp.mq.edu.au).

T. Bossomaier, is with the Centre for Research in Complex Systems, Charles Sturt University, NSW – 2795, Australia (e-mail: tbossomaier@csu. edu.au).

together of inputs contributes significantly to defining the course of human thoughts and reasoning [15]. "Cognition is the behavioural manifestation of psychological processes" [16]. The relationship between cognition and emotion has been one of the stumbling blocks in defining the concept of emotion [10]. A methodological difficulty in this area of research is that despite the multitude of lists of terms indicating emotions, there is no widely accepted definition and taxonomy for emotion [17].

Presently, in studies of emotional speech, the term emotion is used with different meanings, and sometimes includes notions such as attitude, feeling, or intention [17]. There are three aspects to determine the emotional state of the speaker, speech recognition in the presence of emotional speech, synthesis of emotional speech and emotion recognition [18]. Past studies conducted by several researchers have shown high correlation between some statistical measures of speech and emotional state of the speaker; among them are pitch, energy, articulation and spectral shape [18]. Mozziconacci [17] asserts that there are two main tendencies in theories to define emotions: the first considers emotions as discrete categories where a distinction is made between the basic emotions and a combination of these basic ones; the second is to view emotions as characterized by progressive, smooth transitions where similarities and dissimilarities between emotions are characterized in terms of gradual distances on dimensions such as pleasant/ unpleasant, novel/old, consistent/ discrepant and control/ no-control. Pollermann [10] has argued that a unified model of cognition and emotion based on interactionist epistemology would allow an unambiguous definition of emotion and facilitate research in affective aspects of speech communication. The unified cognitive model of Pollermann that thrusts on three basic dimensions traditionally attributed to emotions namely, valence, arousal and power is applicable for organisms described as open systems [10].

III. Emotion in Speech

Expression of emotion

Dealing with speaker's emotion is one of the latest challenges in speech technologies [18] and identification of speaker's emotional state is one of the natural goals of research on vocal expression of emotion [8]. Research of emotions in speech primarily deals with the search for acoustic features of speech that distinguish a number of emotional states [19].

Emotions are known to be carried in two ways, firstly via emotional expressions and secondly via expressivity, that is by using linguistic structures [20]. According to Oudeyer [9], emotional expression takes place in the following sequence -

a) When one is in the state of joy, anger or fear –
   i) The sympathetic nervous system is aroused.
   ii) The heart rate and blood pressure increase.
   iii) The mouth becomes dry and there are occasional muscle tremors.
   iv) This is accompanied by loud and fast speech enunciated with strong high frequency energy.

b) When one is bored or sad –
   i) The parasympathetic nervous system is aroused.
   ii) Decrease in the heart rate and blood pressure.
   iii) Increase in salivation.
   iv) Speech production is slow, low-pitched and with little high frequency energy.

Prosody and emotion

The role of prosody in emotional expressions is still not well known but is supposed to be timed by emotional events and constrained by "linguistic prosody" [20]. Emotionally induced phonemetic and prosodic variations primarily have status of spontaneous, natural indicators or symptoms of physiological reactions automatically triggered by cognitive appraisal outcomes [10]. In speech synthesis prosodic breaks play an important role in structuring utterances and thus increasing their understandability [21]. Prosody clearly organizes spoken language, but how humans use prosodic information in language acquisition and language processing is not yet well established [22]. Generation of pleasant prosody parameters is very important for speech synthesis that can be modelled by time-delay recurrent neural network topologies by considering prosody generation unit as a dynamical system [23]. Their design of acoustic prosody unit is shown in figure 1 with permission.

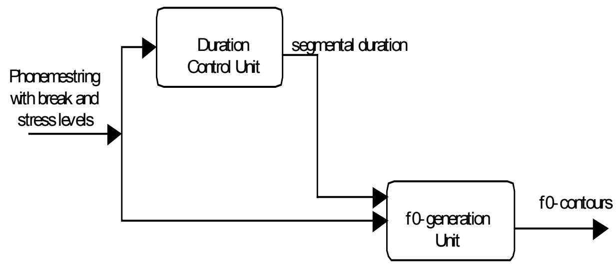

Figure 1 (reprinted from [23])

There is a general agreement among the scientific community that even though the existence of a syntax-prosody interface cannot be doubted, the mapping from syntax to prosody is non-trivial [21]. Prosody is the organization of prosodic material (what ever the relevant signals studied) specifically relevant to carry out the given function values and is directly built from the signal [20].

Automatic Statistical Summary of Elementary Speech Structures (ASSESS) is a pre-processing system, which automatically recovers a range of measures from the speech signal. Most of the measures considered here are statistical summaries of properties related to prosody. Speech samples can be classified into different emotional categories by applying various statistical learning rules to the ASSESS measures. An improved version of ASSESS is to rectify the problem with spectral measurement and extends the use of tune-type units that identify units bounded by appreciable silences at either end, thus producing statistical summaries for each tune. McGilloway *et al* [8], used ASSESS system to extract 32 potentially relevant prosodic features derived from contour tracing, the movement of intensity and pitch.

The features that have been used for the identification of emotion from word strings are [5] –

a) Position of utterance in dialogue [24].
b) Whether the utterance is a repeat/rephrase/ correction [25], [24].

c) Likelihood of utterance according to different language models e.g., class based trigram model [24] and word based unigram model [26].
d) Keywords: straightforward keyword spotting – e.g. [27] and spotting emotionally salient words [28].
e) Type of dialogue act/sequences of dialogue acts [25].
f) Latent Semantic analysis [29].

IV. Emotion from Facial Expression and Gestures

There are 44 muscles in our face most of which contract in a single way except for some like those in the forehead that can move in several ways [30]. These muscles are coded into various facial action codes for computers to recognise. Ekman's Facial Action Coding System (FACS) breaks down facial expressions into 46 individual motions, or action units. A technique called feature-based analysis, measures variables such as the degree of skin wrinkling at various points on the face [31]. Humans vary in their ability and tendency to produce facial expressions which is presumed to be related to underlying muscular, neurobiological, or social differences [32]. Facial expression is either interpreted as a human universal, with basic expressions represented in all human populations [33], or it is conceptualized as the natural outgrowth of cultural differences, with little overlap in expression from population to population [34].

Freeman [1] has talked about 1500 ways of Emotioneering™ for broadening, deepening and enhancing the quality of games and has grouped them into 32 categories. Two prior studies [35], [36] in analysing facial behaviour of emotional episodes had shown that game scenearios and manipulations are apt to induce specific emotional reactions, both positive (happiness, pride, relief) and negative (anger/irritation, anxiety/fear, sadness/ dissapointment, embarrassment/shame) [37]. Facial expression can also furnish information about ones personality and temparament like shyness and hostility [38]. In their study [39] of the dominance cues perceived from specific emotions, Knutson [40], discovered that "personality" traits are associated with specific emotions. His measure of dominance, which included the adjectives, "dominant," "self-assured," "self-confident," and "assertive," elicited very high ratings when participants were shown facial displays of anger, a "masculine" expression [41].

The combined results of two studies investigated emotion-related biases in selective attention for pictorial stimuli in non-clinical subjects (the stimuli included threatening, happy and neutral facial expressions) showed evidence of an emotion-related attentional bias for facial expressions (i.e. an interaction effect of dysphoria and the emotional valence of the facial expression on attentional bias). Non-dysphoric subjects (i.e. those with low levels of anxiety and depression) showed an attentional bias away from threatening facial expressions, relative to neutral expressions. The dysphoric subjects did not show significant evidence of attentional biases for emotional stimuli [42]. “Scientists have already found ways for distinguishing false facial expressions of emotion from genuine ones. In depressed individuals, they have also discovered differences between the facial signals of suicidal and non-suicidal patients. "The trouble,” Sejnowski [31] explains, is that some people don't wrinkle at all and some wrinkle a lot. It depends on age and a lot of other factors, so it's not always reliable.

The movements of the skin and other features of the face are the basis for determining both the particular action unit (AU) that has occurred and the intensity of the AU. These movements have a direction, and they produce bunching, bagging, pouching, and wrinkling between the origin of the muscle and its attachment, as well as flattening, stretching, and pulling in other areas [43]. NECA (Net Environment for Embodied Emotional Conversational Agents) aims at developing a cost efficient adaptation of web applications where two or more virtual human-like characters communicate with each other with emotionally rich speech, gesture and facial expression [44]. Progress has been made in games for capturing gestures. Sony has developed a game called EyeToy™ that tracks the movement of the hand and feet in different positions. This technology could be enhanced to capture emotions from facial expressions and gestures.

V. Architecture for Modelling Emotion in Game Characters

A proposed model for incorporating emotion into “speechy” games

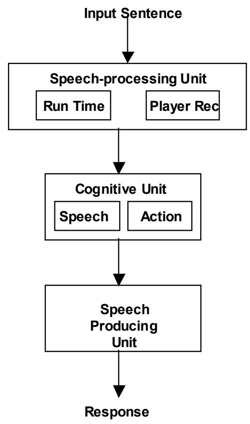

Figure 2: The Pidgin Game model

The Model of Figure 2 consists of the following -
a) A Speech-processing unit, a Cognitive unit and Speech-producing unit.
b) Speech-processing unit consists of two sub-modules – Runtime unit and an optional Player-recognition unit. The Run time unit is explained in Figure 3.

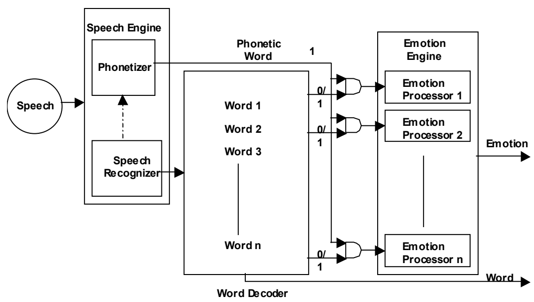

Figure 3: Run Time Unit of Speech-processing Unit

Run Time Unit of Speech-processing Unit in Figure 3 has the following -

a) A speech Recogniser that outputs the closest match of a word.
b) The word decoder will accept the word from the speech recogniser and check whether it exists in the game dictionary.
c) The Phonetizer will transform the signal into its phonetic equivalent irrespective of whether or not it is a valid word.
d) The Emotion processor has the phonetic word as input along with the word from the word decoder and generates an appropriate emotion state. Each word has its own emotion processor for deriving an emotional state consisting of Finite state Automaton as shown in figure 4.
e) The output is an emotion string along with the word.

An emotion processor is a finite state automaton, consisting of a number of states, input symbol and probability of transition. It extracts emotion from a phoneme [45] by analysing individual phonetic alphabet. The emotion processor is similar to a Hidden Markov Model [46] and is given below -

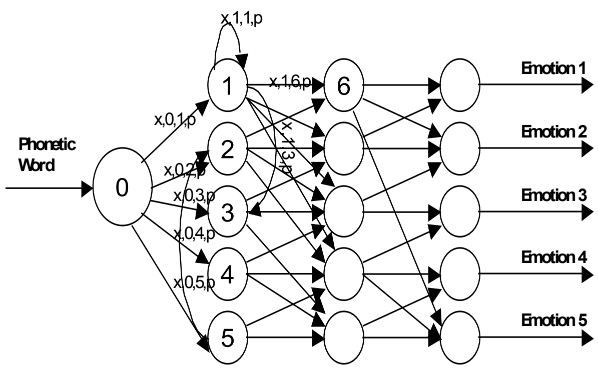

Figure 4: Emotion Processor. Here the transition is given by the symbols x, S_i, S_o, p, the first letter 'x' is the input phone, followed by current state, followed by next state, p gives the probability of transition.

The emotion Processor of figure 4 has the following -

a) The number of rows = the number of emotion states namely – joy, boredom, anger, sadness, fear, indignation and neutral [17]. The states are not layered and there can be any number of states in a row and hence it is left unlabelled.
b) The transition from one state to another depends upon –
    i) Input symbol (phone).
    ii) Present state (emotion).
    iii) Next input symbol (phone).
    iv) The weight/probability is a number between 0 and 1. of the link between any two states. The probabilities get multiplied with each transition. This may desist drastic change of emotion states while the input string is processed in cases of ambiguity. Prosodic [47] information of stress, rhythm, timing and intonation and proximity of the emotional states assist in the determination of this value.

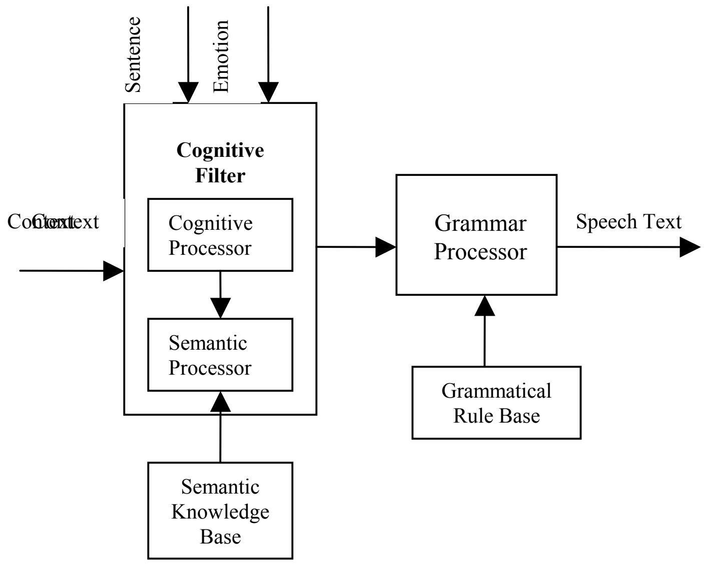

Figure 5: Cognitive Speech Unit

The context of the scene of the game will play a significant role in generating an appropriate oral response to the player's action. We have developed a pidgin [48] language for conversation between human players and game characters [13], [49]. The Speech (figure 5) and Action (figure 6) units complement each other and priority can be set to suppress a conflict if a spoken response is generated simultaneously in both. The output of this unit is a speech text in Pidgin Language. In figure 5, there are 3 inputs –

a) Sentence – What has been said.
b) Emotion – How it has been said.
c) Context – Why it has been said.

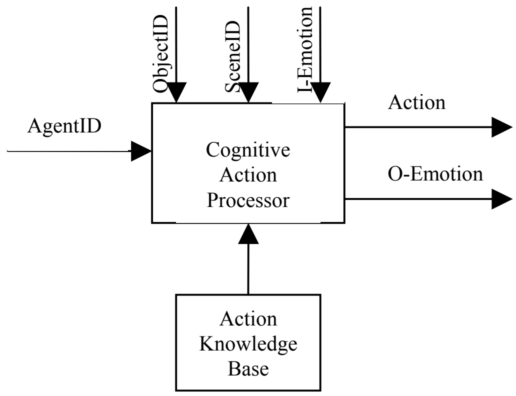

Figure 6: Cognitive Action Unit

Cognitive Action Unit of figure 6 generates an appropriate Physical or Perceptual action corresponding to an event in the scene. There are 4 inputs –

a) SceneID – An identifier that uniquely identifies a scene.
b) I-Emotion – The initial emotion state of the Player/agent if any. The emotion content is analysed from the words recognized by the Physical action called *Speak* of the Player/Agent.
c) AgentID – An identifier that uniquely identifies an Agent in the scene.

d) ObjectID - An identifier that uniquely identifies an object in the scene that may have triggered a reaction.

VI. Conclusion

In this paper we have discussed elaborately the various techniques to infer emotional states from faces, gestures and speech. We have presented an architecture for extracting emotion from speech in real time for computer games. We have also discussed how speech and emotion from the playing character can be transmitted to the NPC.

References

[1] D. Freeman, 2003, Creating Emotion in Games, New Riders Publishing, USA.

[2] M. Kavakli, J.R. Thorne, A Cognitive Modelling Approach To Capturing the Context of Complex Behaviour in Gameplay, ICITA, Australia, 2002.

[3] J. Rumbaugh, I. Jacobson, G. Booch, 1999, The Unified Modeling Language Reference Manual, Addison Wesley, USA.

[4] Williamson, H., 2001, XML: The Complete Reference, Tata McGraw Hill, New Delhi, India.

[5] A. Belz, Nov 2003, And Now with feeling: Developments in Emotional Language Generation, Information Technology Research Institute Technical Report Series, University of Brighton, UK.

[6] R.W. Picard, 1997, Affective Computing, MIT Press, Cambridge, Massachusetts.

[7] D. Jitca, H. N. Teodorescu, V. Apopei, F. Grigoras, Improved Speech Synthesis Using Fuzzy Methods, International Journal of Speech Technology 5227-235, 2002.

[8] S. McGilloway, R. Cowie, E.D. Cowie, S. Gielen, M. Westerdijk, S. Stroeve, Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark, ICSA Workshop on Speech and Emotion, Ireland, 2000.

[9] P. Oudeyer, Synthesis of Cartoon Emotional Speech, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[10] B. Z. Pollermann, A place for Prosody in a Unified Model of Cognition and Emotion, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[11] F. de Rosis, F. Graso, 2000, Affective Natural Language Generation. In Paiva, A.M.,, editor, Affective Interactions, number 1814 in Springer Lecture notes in AI, pp 204-218, Springer.

[12] L. Holzman, W., Pottenger, 2003, Classification of emotions in Internet Chat, Technical Report LU-CSE-03-002, Dept of Science and Engineering, Lehigh University.

[13] T. Rudra, M. Kavakli, T. Bossomaier, 2003: A Game Pidgin Language for Speech Recognition in Computer Games, COSIGN 2003, The 3rd International Conference on Computational Semiotics for Games and New Media, University of Teeside, Middlesbrough, United Kingdom, September 9-12.

[14] R.,P. Cooper, Modelling High-Level Cognitive Processes, 2002, Lawrence Erlbaum Associates, USA.

[15] A.T. Rappaport, 1997: Context, Cognition, and the Future of Intelligent Infostructures, In: Feltovich, Ford, Hoffman, Expertise in Context, AAAI Press, Menlo Park California, pp. 523-539.

[16] M., H. Johnson, Brain Development and Cognition: A Reader, Blackwell Publishers, Cambridge Mass, 1993.

[17] S. Mozziconacci, Prosody and Emotions, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[18] A. Nogueiras, A. Moreno, A. Bonafonte, J. B. Mariño, Speech Emotion Recognition Using Hidden Markov Models, Speech Prosody 2002, an International Conference, France, 11-13 April 2002.

[19] L. Bosch, Emotions: What is possible in the ASR Framework, ICSA Workshop on Speech and Emotion, Ireland, 2000.

[20] V. Aubergé, A Gestalt Morphology of Prosody Directed by Functions: the Example of Step by Step Model Developed at ICP, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[21] M. Atterer, 2002, Assigning Prosodic Structure for Speech Synthesis: A Rule-Based Approach, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[22] O. Bagou, C. Fougeron, U. H. Frauenfelder, Contribution of Prosody to the Segmentation and storage of "Words" in the Acquisition of a New Mini-Language, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[23] Ç. Erdem, H. G. Zimmermann, Duration Control by Asymmetric Causal Retro-Causal Neural Networks, Speech Prosody 2002 an International Conference, France, 11-13 April 2002.

[24] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, A. Stolcke, 2002, Prosody-Based automatic detection of annoyance and frustration in human-computer dialog. Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002), Volume 3, pp 2037-2040.

[25] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Nöth, 2000. Desperately seeking emotions: Actors, wizards and human beings. Proceedings of the ISCA workshop on Speech and Emotion: A conceptual framework for research, pp 195-200.

[26] L. Devillers, I. Vasilescu, L. Lamel, 2002, Annotation and detection of emotion in a task oriented human-human dialog corpus. In proceedings of ISLE Workshop.

[27] C. Elliot, 1992, The Affective Reasoner: A process model of emotions in a multi-agent system. PhD thesis, Institute for the Learning Sciences, Northwestern University. Available as Technical Report No. 32.

[28] C.M. Lee, S. Narayanan, R. Pieraccini, 2002, Combining acoustic and language information for emotion recognition. ICSLP 2002.

[29] T. K. Landauer, S. T. Dumais, 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction and representation of knowledge. Psychological Review, 104:211-140.

[30] P. Ekman, W. Friesen, 1978, The Facial Action Coding System, Consulting Psychology Press.

[31] N. R. Flanigan, 2001, Facing the Truth: A new Tool to Analyze our Expressions, HHMI Bulletin May, 2001, pp. 12-17, http://www.hhmi.org/bulletin/pdf/faces.pdf.

[32] K. L. Schmidt, J. F. Cohn, 2001, Human Facial Expressions as Adaptations: Evolutionary questions in Facial Expression Research, Year book of Physical Anthropology 44:3-24.

[33] D. Brown, 1990, Human Universals, McGraw-Hill Inc, New York.

[34] R. L. Birdwhistell, 1970, Kinesics and context, University of Pennsylvania Press, Philadelphia.

[35] S. Kaiser, T. Wehrle., Situated emotional problem solving in interactive computer games, 1996, Proceedings of the VIIIth conference of ISRE, ISRE Publication.

[36] S. Schmidt, T. Wehrle, S. Kaiser, Emotion eliciting situations, subjective feeling, and facial expression in an interactive computer game, 7th conference on Facial Expression – Measurement and Meaning, July 16-20, 1997, Salzburg.

[37] S. Kaiser, T. Wehrle, S. Schmidt, Emotional Episodes, facial expressions, and reported feelings in human-computer interactions, 1998, Proceedings of Xth conference of the International Society for research on Emotions (pp. 82-86), Würzburg, ISRE Publications.

[38] P. Ekman, Facial signs: Facts, fantasies and possibilities. T. Seboek (Ed.), Sight sound and sense, pp. 124-156. Bloomington: Indiana University press, 1978.

[39] S. B. Algoe, B.N. Buswell, J.D. DeLamater, Gender and Job Status as Contextual Cues for the Interpretation of Facial Expression of Emotion, Sex Roles: A Journal of Research, Feb, 2000.

[40] B. D. Knutson, (1993). A new view of character: Links between facial expression and disposition. Dissertation, Stanford University, Stanford, CA.

[41] E. A. Plant, J. S. Hyde, D. Keltner, P. G. Devine, (2000). The gender stereotyping of emotion. Psychology of Women Quarterly, in press.

[42] B.P. Bradley, K. Mogg, N. Millar, C. Bonham-Carter, E. Fergusson, J. Jenkins, M. Parr, 2004, Attentional Biases for Emotional Faces, Cognition and Emotion, Volume: 11 Number: 1 pp: 25 – 42, Psychology Press, part of the Taylor & Francis Group.

[43] P. Ekman, W. V. Friesen, J. C. Hager, The Manual: Facial Action Coding System, A Human Face USA, 2002.

[44] B. Krenn, H. Pirker, 2004, Defining the Gesticon: Language and Gesture Coordination for Interacting Embodied Agents, Proceedings of the AISB – 2004 symposium on Language, speech and gesture for expressive characters, University of Leeds, UK, pp 107-115.

[45] C.M. Lee, S. Yildrim, M. Bulut, A.K.C. Busso, Z. Deng, S. Lee, S. Narayanan, 2004, Emotion Recognition based on Phoneme Classes. ICSLP 2004.

[46] D. Jurafsky, J. H. Martin, 2000, Speech and Language processing, Prentice Hall, USA.

[47] A.D.N. Edwards, Speech Synthesis Technology for Disabled People, 1991, Paul Chapman Publishing Ltd.

[48] R.A. Hall, Pidgin and Creole Languages, 1966, Cornell University Press, London.

[49] T. Rudra, D. Tien, T. Bossomaier, 2005: Spoken Communication with Computer Game Characters, proceedings ICITA 2005, The 3rd International Conference on Information Technology and Applications, University of Technology, Sydney, Australia, July 4-7.