Published at Computer Animation and Social Agents CASA 2004 Conference, Geneva, Switzerland

Autonomous Speaker Agent

Karlo Smid
Ericsson Nikola Tesla
Krapinska 45, p.p. 93,
HR-10 002 Zagreb
Phone: +385 1 365 3053
Fax: +385 1 365 3610
Email:
karlo.smid@ericsson.com

Igor S. Pandzic
Faculty of electrical
engineering and
computing, Zagreb
University
Email:
Igor.Pandzic@fer.hr

Viktorija Radman
Faculty of electrical
engineering
and computing, Zagreb
University
Email:
Viktorija.Radman@fer.hr

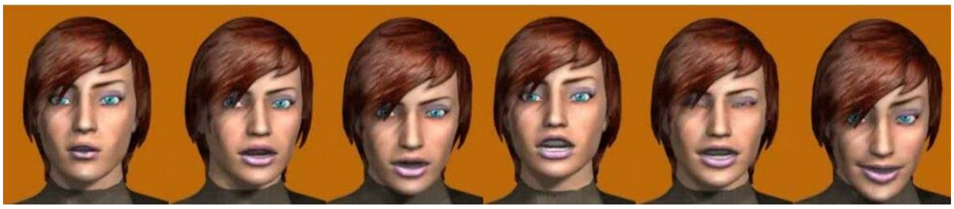

Figure 1: Sample images of an animated face with facial gestures

Abstract

Autonomous Speaker Agent is a graphically embodied animated agent (a virtual character) capable of reading plain English text and rendering it in a form of speech, accompanied by the appropriate, natural-looking facial gestures. The system uses lexical analysis and statistical models of facial gestures in order to generate the gestures related to the spoken text. It is intended for the automatic creation of the realistically animated virtual speakers, such as newscasters and storytellers, and incorporates the characteristics of such speakers captured from the training video clips. Autonomous Speaker Agent is based on a visual text-to-speech system which generates lip movement synchronized with the generated speech. This is extended to include eye blinks, head and eyebrow motion, and a simple gaze following behavior. The result is a full face animation produced automatically from the plain English text.

Keywords: facial animation, virtual character, visual text-to-speech, MPEG-4 FBA, embodied conversational characters

1. Introduction

Autonomous Speaker Agent is a graphically embodied animated agent (a virtual character) capable of reading plain English text and rendering it in a form of speech accompanied by the appropriate facial gestures. Our system uses lexical analysis of the English text and statistical models of facial gestures in order to automatically generate the gestures related to the spoken text. It is intended for the automatic creation of the realistically animated virtual speakers, such as newscasters and storytellers and incorporates the characteristics of such speakers.

Autonomous Speaker Agent is an extension of a Visual Text-to-Speech (VTTS) system. A classical VTTS system [1][2][3][4][5][6] produces lip movements synchronized with the synthesized speech based on the timed phonemes generated by the speech synthesis. Normally, it also solves the coarticulation problem [7][8][23][9]. A face, that only moves the lips, looks extremely unnatural because the natural speech always involves facial gestures: head movements, such as nods and swings, eyebrow raising, eyes movement and blinking. Very often the body gestures are involved too, mostly the hand movements. However, a VTTS system can only obtain phonetic information from the speech synthesis and has no basis for generating realistic gestures. Very often this problem is solved by introducing some partially random gestures triggered by a set of rules, e.g. a preset frequency of the eye blinks [23][10]. Another solution is recording one or more sequences of facial gestures from a real speaker using the face tracking methods, and then playing those tracks during the speech [11]. These methods produce much better visual results than a static talking face. However, the movements are generally too simplistic and look mechanical. Yet another approach is manually inserting tags or bookmarks into text from which the facial gestures or expressions are generated [23]. Obviously, this is time consuming and unsuitable for the fully automatic applications. The Eyes Alive system [12] introduces a full statistical model of eye movement based on the known theory of eye movement during speech, as well as the precise recordings of eye motion during speech. The system reproduces eye movements that are dynamically correct at the level of each movement, and that are also globally statistically correct in terms of frequency of movements, intervals between them and their amplitudes. The statistical model distinguishes between the speaking and the listening mode because eye movement patterns are different in these two modes. However, the movements are still unrelated to the underlying speech contents, punctuation, accents etc. In natural speech, most gestures are directly related to the lexical structure of speech and have distinct functions [13][14][15][16][17][18][19]. The "BEAT" system [20] uses linguistic and contextual information contained in a text to control the movements of the hands, arms and a face, and the intonation of a voice. The mapping from a text to the facial, intonational and body gestures is contained in a set of rules derived from a state of the art in nonverbal conversational behavior research. However, the system does not introduce a fully statistical model for the supported gestures.

In this paper, we propose a new approach that combines the lexical analysis of input text with a statistical model describing the dynamics, frequencies and amplitudes of facial gestures. The statistical models are obtained by analysing a training data set consisting of several speakers recorded on video and stenographs of their speech. A lexical analysis of the stenograph texts allowed to correlate the lexical characteristics of a text with the corresponding facial gestures and to incorporate this correlation into a statistical model. Using a lexical analysis of input text to trigger this statistical model, the Autonomous Speaker Agent can perform gestures that are not only dynamically correct, but also correspond to the underlying text (Figure 5).

As a basis for the statistical model, we use the repertoire of known facial gestures and their functions in speech based on the existing theory described in Section Background. In order to obtain the lexical structure of the input text, we use a lexical analysis system described in Section Lexical analysis of text. The statistical model of facial gestures, and how it was produced from the training data set, is described in Section Statistical model of facial gestures. The complete Autonomous Speaker Agent system is presented in Section The System. The final sections present results and conclusions.

2. Background

A conversation consists of two domains: verbal and nonverbal. These two domains are highly synchronized because they are driven by the same forces: the prosody and lexical structure of the uttered text as well as the emotions and personality of a person that is involved in a conversation [21]. The verbal domain deals with a human voice, while body and facial gestures (head, eyes and eyebrows movement) are part of the nonverbal domain. In this article our focus is on facial gestures and how they are synchronized and driven by the prosody and lexical structure of the uttered text.

Facial gestures are driven by [13]:

- **interactional function of speech**: we unconsciously use facial gestures to regulate the flow of speech, accent word or segments, and punctuate speech pauses.
- **emotions**: they are usually expressed with facial gestures.
- **personality**: it can often be read through facial gestures.
- **performatives**: for example, advice and order are two different performatives and they are accompanied with different facial gestures.

In this article we deal with the interactional function of speech. In this context, facial gestures can have several different roles, usually called determinants [1]. These determinants are:

- **conversational signals**: they correspond to the facial gestures that clarify and support what is being said. These facial gestures are synchronized with accents or emphatic segments. Facial gestures in this cathegory are eyebrow movements, rapid head movements, gaze directions and eye blinks [14].
- **punctuators**: they correspond to the facial gestures that support pauses; these facial gestures group or separate the sequences of words into discrete unit phrases, thus

reducing the ambiguity of speech [17]. The examples are specific head motions, blinks or eyebrow actions.

- **manipulators**: they correspond to the biological needs of a face, such as blinking to wet the eyes or random head nods because being completely still is unnatural for humans.
- **regulators**: they control the flow of conversation. A speaker breaks or looks for an eye contact with a listener. He turns his head towards or away from a listener during a conversation [22]. We have three regulator types: Speaker-State-Signal (displayed at the beginning of a speaking turn), Speaker-Within-Turn (a speaker wants to keep the floor), and Speaker-Continuation-Signal (frequently follows Speaker-Within-Turn). The beginning of *themes* (an already introduced utterance information) are frequently synchronized by a gaze-away from a listener, and the beginning of *rhemes* (a new utterance information) are frequently synchronized by a gaze-toward a listener.

Since we are currently concentrating on Autonomous Speaker Agent, which is not involved in a conversation but performs a presentation, this work focuses on conversational signals, punctuators and manipulators. All these functions are supported by a fairly broad repertoire of facial gestures. We distinguish three main classes of facial gestures [1]:

- Head movement
- Eyes movement
- Eyebrows movement

Within each class we distinguish specific gestures, each characterized by their particular parameters. The parameters that are important for the head and eyebrows movements are amplitude and velocity. Those two parameters are in inverted proportion. A movement with a big amplitude is rather slow. Table 1 shows the types of facial gestures as identified during our data analysis (Section Statistical model of facial gestures). This is an extension of the classification proposed in [19]. We introduce symbols incorporating both a gesture type and a movement direction.

Head	Nod	ˆ v > <	An abrupt swing of the head with a similarly abrupt motion back. We have four nod directions: up and down (ˆ), down and up (v), left and right (<) and right and left (>).
	Overshoot nod	~	Nod with an overshoot at the return, i.e. the pattern looks like an 'S' lying on its side.
	Swing	u d L R diag	Abrupt swing of the head without the back motion. Sometimes the rotation moves slowly, barely visible, back to the original pose, sometimes it is followed by an abrupt motion back after some delay. Five directions: up (u), down (d), left (l), right (R) and diagonal (diag).
	Reset	reset	Sometimes follows swing movement. Returns head in central position.
Eyes	Movement in various directions		Eyes are always moving. Parameters are: gaze direction, points of fixation, the percentage of eye contact over gaze avoidance, duration of eye contact.
	Blink		Periodic blinks keep the eyes wet. Voluntary blinks support conversational signals and punctuators.
Eyebrows	Raise	ˆˆ	Eyebrows go up and down.
	Frown	ˇˇ	Eyebrows go down and up.

Table 1: The specification of the facial gestures.

3. Lexical analysis of text

The speech analysis module performs linguistic and contextual analysis of a text written in English language with the goal of enabling the nonverbal (gestures) and verbal (prosody) behaviour assignment and scheduling.

Starting from plain a English text, it produces an XML document annotated with tags for each word. These tags allow the distinction of the newly introduced words, words known from the previous text and punctuation marks. Based on this knowledge, the process described in Section The System assigns and schedules the gestures.

The input text is first phrase-parsed, because the module needs to know the morphological, syntactic and part-of-speech information. In order to get the morphologic and semantic data about words in a sentence, we use Connexor's Machinese Phrase Tagger[1] (MPT). In the second step, we break the paragraphs (UTTERANCE) into clauses (CLAUSE). The largest unit is UTTERANCE, which represents an entire paragraph of input. The next unit is

[1] http://www.connexor.com/ 29/03/2004

CLAUSE, which is held to represent a proposition. In order to detect clauses in an utterance, the module is searching for the punctuation marks and the placement of verb inside a phrase.

The smallest unit is word with its **new** attribute. To determine the newness of each word, we keep track of all previously mentioned words in an utterance. We also use WordNet 1.7.1.[2] to identify sets of synonyms. We tagged each noun, verb, adverb and adjective as **new** if itself or its synonym has not been seen in an utterance before. The other word classes are not considered for the new parameter. We determine a word class based on the Connexor's word class tags. Since pronouns need to be tagged as **new** and WordNet does not process them at all, a special algorithm to deal with pronouns is proposed. The logic for this algorithm is based on knowledge and intuition, but that, of course, does not lead us to the universal solution. Here is the pseudocode for the algorithm:

&NH nominal head;

&>N nominal determiner or premodifier;

```
FOR each PRONOUN in the text
IF the PRONOUN is &NH and belongs to SET(any, anything,
anyone, anybody, some, somebody, someone, something, no,
nobody, no-one, nothing, every, everybody, everyone,
everything, each, either, neither, both, all, this, more, what, who,
which, whom, whose)
THEN mark the PRONOUN as NEW
ELSE IF the PRONOUN is &>N and belongs to SET(I, you,
he, she, it, we, they)
THEN mark the PRONOUN as NEW
ELSE mark the PRONOUN as OLD;
```

4. Statistical model of facial gestures

In this section we present the statistical model of facial gestures and the methods, tools and datasets used in order to build it.

As a training set for our analysis, we chose Ericsson's "5minutes" video clips. Those clips are published by LM Ericsson for internal usage and offer occasional in-depth interviews and reports on major events, news, or hot topics from the Telecom industry. They are presented by professional newscasters. We used the footage showing the newscasters (Figure 2). We investigated three female and two male Swedish newscasters.

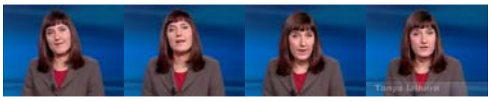

Figure 2: Tonya's nod with overshoot[3]

First, using a video editing tool, we extracted from the video news casting extracts according to their stenographs. Then we grouped those news extracts for every observed speaker. Observing those news casting clips, we marked the starting and ending frames for every eye blink, eyebrow raise and head movement (Figure 2). Analyzing those frames, the speakers Mouth-Nose Separation unit (MNS0) value, facial gesture amplitude value, facial gesture type (Table 3) and direction were determined. We used the following algorithm for the amplitude values: those values represent the difference between the speaker's nose top position at the end and the begining of a facial gesture (eyebrow raise or head movement).

Data values, that were gathered from the video clips, were statistically processed using Microsoft Excel and MatLab 5.3. That means that a number of pie charts (Figure 3) were produced by simply calculating how many times were facial gestures triggered/not triggered by words. Every gesture type has a corresponding pie chart. Amplitude values probabilities (Figure 4) were calculated using the histogram statistical function.

Table 2 presents an example of the gathered raw data for one news extract.

word	52		Three	arraignments
eyes	3		blink::cs	
head		|up;A=2	|d to n |d A=0.25	Id A =0.5
eyebrows	2		raise::cs A=1/4	
pitch	13		+	
lexical	44		new	new
			cs - conversational signal p – punctuator m - manipulator	
		~nod::A1=2:A2=0.5::cs		

Table 2: Example of data set gathered during the analysis.

The **word** row contains the analysed news extract separated word-by-word. The **eyes**, **head** and **eyebrows** rows hold data about facial motion that occurred on the corresponding word (separated with :: symbol): the type of motion and its direction (according to the

[2] http://www.cogsci.princeton.edu/~wn/ 29/03/2004

[3] Published with permission of LM Ericsson

notation summarised in Table 3), amplitude value (A stands for amplitude) and determinant code values (cs, p, m summarised in Table 2).

The head basic motions are mapped to head movements as described in the last column of Table 1. We used a rule that before and after every head movement type, the head stands still for three frames.

We replayed the newscaster footage to determine facial gestures type, direction and duration parameters. The last row in Table 2 contains the head movement facial gesture parameters.

The **pitch** row indicates which words were emphasised by voice intonation. The **lexical** row holds information about word's newness in the text context (Section Lexical analysis of text). Te second column in Table 2 represents the number of occurrences of a particular facial gesture and pitch accents, the number of words in current news extract, and the number of words that are new in the text context.

4.1 Determinant values of facial motions

A determinant value for a particular facial motion is determined as follows. If a facial motion occurred on a punctuator mark, then a determinant for that motion was the punctuator (p). If a facial motion accompanied a word that is new in the context of the uttered text, then a determinant was the conversational signal (cs). Otherwise, a determinant of a facial motion was the manipulator (m). The raw data tables were populated by manual analysis and measurement. All amplitude values were normalized to MNS0 for the particular speaker. MNS0 is Facial Animation Parameter Unit (FAPU) in MPEG-4 Face and Body Animation (FBA) standard [23]. Using MNS0 FAPU our model could be applied to every 3D model of speaker.

Facial gesture	Type of	Description
Head movement	| up	vertical up
	| up to n	vertical up to neutral
	| down	vertical down
	| down to n	vertical down to neutral
	-- to left	horizontal left
	-- to right	horizontal right
	-- to n	horizontal to neutral (centre)
	/ up	diagonal up from left to right[4]
	/ down	diagonal down from right to
	\ up	diagonal up from right to left[4]
	\ down	diagonal down from left to
Eyebrows	raised s	eyebrows going up to
	raised e	eyebrows going down to

Table 3: Basic facial motions triggered by words.

In our model, the basic unit which triggers facial gestures is a word. We chose not to subdivide further into syllables or phonemes for simplicity reasons. Since some facial gestures last through two or more words, this level of subdivision seems appropriate.

The raw data (Table 2) for the complete training set was statistically processed in order to build a statistical model of speaker behaviour. A statistical model consists of a number of components, each describing the statistical properties for a particular gesture type in a specific speech context. A speech context can be an old word, a new word or a punctuator. The statistical properties for a gesture type include the probability of occurrence of particular gestures and histograms of amplitude and duration values for each gesture. Figure 3 shows an example of a statistical data component for head gestures in the context of a new word. It is visible that in 51%, we have some kind of head movement. For example, probability of occurrence for the rapid head movements is 22%. Further, we have five directions: up (u), down (d), left (L), right (R) and diagonal (diag). In the end, we must determine the amplitude for a rapid movement. Figure 4 shows a linear approximation of the histogram that represents the frequency of occurrence of rapid head movement amplitude values.

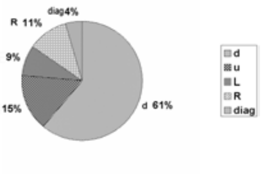

Figure 3: : Statistical data for head gestures occurrences in the context of a new word

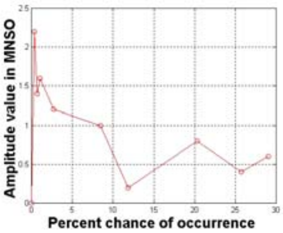

Figure 4: Frequency of occurrence of rapid head movement amplitude values.

[4] From the listener point of view.

Such statistics exist for each gesture type and for each speech context we treated. They are built into the decision tree (Figure 5) that triggers gestures. The process is described in the following section. Note that, in the context of punctuators, only eyes gestures are used, because the statistics show that other gestures do not occur on punctuators.

5. The system

Figure 6 shows the complete Autonomous Speaker Agent system. The input to the system is plain English text. It is processed by a lexical analysis (Section Lexical analysis of text) which converts it into an XML format with lexical tags (currently describing new/old words and punctuators). The facial gesture module is the core of the system – it actually inserts appropriate gestures into text in the form of special bookmark tags. These bookmark tags (Table 4) are read by the TTS/MPEG-4 Encoding module. While the Microsoft Speech API (SAPI) Text To Speech (TTS)[5] engine generates an audio stream, the SAPI notification mechanism is used to catch the timing of phonemes and bookmarks the containing gesture information. Based on this information, an MPEG-4 FBA bitstream is encoded with the appropriate viseme and facial gestures animation. For MPEG-4 FBA bitstream generation, we are using Visage SDK API[6] that uses SAPI 4.0 or 5.1. Visage SDK API uses information provided by the SAPI notification mechanism.

The facial gesture module is built upon the statistical model described in the previous section. The statistical model is built into the decision tree illustrated in Figure 5.

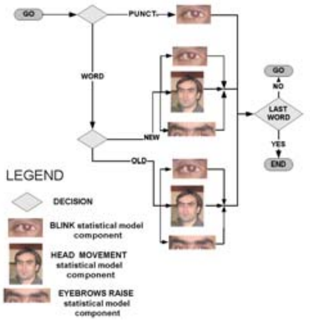
GO
PUNCT.
WORD
NEW
OLD
GO
NO
LAST WORD
YES
END
LEGEND
DECISION
BLINK statistical model component
HEAD MOVEMENT statistical model component
EYEBROWS RAISE statistical model component

Figure 5: Decision tree with components of the statistical model.

Let us follow the decision tree (Figure 5). The first branch point classifies the current context as either a word or a punctuation mark. Our data analysis showed that only eye blink facial gesture had occurred on the punctuation marks. Therefore only the blink component of the statistical model is implemented in this context. The words could be new or old (Section Lexical analysis of text) in the context of uttered text – this is the second branch point. All facial gestures occurred in both cases but with different probabilities. Because of that, in each case we have different components for facial gestures parameters. From Figure 5, it is obvious that a word could be accompanied by all three types of facial gestures at the same time. The facial gesture signals (eye blink, head movement, eyebrow raise) are generated separately, based on their statistical component data. They will be blended later in the TTS/MPEG-4 encoding component. The output from the facial gesture module is plain English text accompanied by bookmark pairs for facial gestures.

Bookmark code	Facial gesture
\Mrk=1\	conversational signal blink
\Mrk=2\	punctuator blink
\Mrk=300\	eyebrows raise
\Mrk=400\	nod ^
\Mrk=700\	nod V
\Mrk=1000\	nod <
\Mrk=1300\	nod >
\Mrk=9\	rapid reset
\Mrk=1600\	rapid d
\Mrk=1900\	rapid u
\Mrk=2200\	rapid L
\Mrk=2500\	rapid R
\Mrk=2800\	rapid diagonal

Table 4: SAPI bookmark codes of facial gestures.

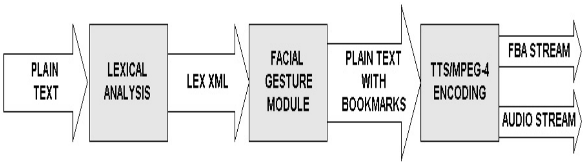

Figure 6: The data flow through the Autonomous Speaker Agent system

Every facial gesture has a corresponding pair of bookmarks: one bookmark marks the starting moment of a facial gesture and the other marks the ending moment. Table 4 shows values for each bookmark. The head and eyebrows movement bookmark values not only define type of facial gesture, but also contain the amplitude data of a facial movement. For example, bookmark value 2300 defines the rapid head movement to the left (symbol L) of

[5] Microsoft speech technologies http://www.microsoft.com/speech/ 29/03/2004

[6] Visage Technologies AB http://www.visagetechnologies.com/ 29/03/2004

amplitude 1 MNS0 (Section Statistical model of facial gestures). The function for amplitudes of facial gestures L is:

A=((Bmk_value –Bmk_code)/100). (1)

The interval for bookmark values for L is [2200,2500> because the statistical model showed that the maximal amplitude value for facial gesture L was 2.2 MNS0.

The head nods and eyebrows raises could last through two or more words. Statistics have shown that the maximum duration of a nod is five words, eyebrow raise could last through eleven words and the maximal duration for a nod with an overshoot is eight words.

We code a nod with an overshoot as two nods: a nod up immediately followed by a nod down. Every nod has its own amplitude distribution.

TTS/MPEG-4 encoding module using the bookmark information encodes an MPEG-4 FBA bitstream with an appropriate viseme and gestures animation. The animation model for head and eyebrows movement facial gestures is based on the trigonometry sine function. That means that our Autonomous Speaker Agent nods his head following the sine function trajectory.

We have implemented a simple model of gaze following, meaning that the eyes of our Autonomous Speaker Agent are moving in the opposite direction of a head movement. This gives an impression of an eye contact with the Autonomous Speaker Agent.

6. Results

We have produced a number of MPEG-4 FBA facial animation bitstreams and accompanying audio files using our system, and rendered these animations as high quality video sequences. To generate the example accompanying this paper, we have used the Siggraph 2003 press release as an input text. We have also used the transcripts of the original training set video footage, so we could compare the behavior of real speakers and the Autonomous Speaker Agent. We have played both the real speaker footage and the Autonomous Speaker Agent videos for a limited audience. The audience reaction was positive. They have noticed that the Autonomous Speaker Agent did not merely repeat facial gestures in some predefined or random manner and judged the gesturing as fairly realistic and convincing.

7 Conclusion

According to feedback that we have received from the audience, we can conclude that our statistical model of facial gestures can be used in a system that implements a fairly convincing Autonomous Speaker Agent. Also, with statistical data that we have gathered during our work, we have confirmed some of the conclusions of other papers. We confirmed that, on average, the amplitude of a faster head nod is lesser than the amplitude of a slower nod. Furthermore, we concluded that the words, that bring something new in the utterance context, are very often accompanied by some facial gesture. However, our system is not ready yet for the Turing test. An extension to Embodied Conversational Characters is a logical item for future work, extending the system to support natural gesturing during a conversation and not only for independent speakers. This will involve adapting and extending the statistical model to include more complicated gesturing modes and speech prosody that occur in a conversation. Modifying speech prosody [24][25][26] of the input text according to statistical prosody data of professional speakers would produce a much more convincing Autonomous Speaker Agent. In order to get more natural head movements, the velocity dynamics [27] of those movements must be implemented in the TTS/MPEG-4 encoding (Figure 6) module.

References

[1] Pelachaud, C., Badler, N., and Steedman, M. 1996. Generating Facial Expressions for Speech, *Cognitive, Science*, 20(1), 1–46.

[2] Kalra, P., Mangili, A., Magnenat-Thalmann N., and Thalmann, D. 1991. SMILE: A Multilayered Facial Animation System. In *Proceedings of Modeling in Computer Graphics,* KUNII, T.L. (Ed), Springer-Verlag, 189-198.

[3] Legoff, B. and Benoît, C. 1997. A French speaking synthetic head. In *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing* 1997, C. Benoît, and R. Campbell, Eds., Rhodes, Greece, 145–148.

[4] Lewis, J.P., and Parke, F.I. 1987. Automated lipsynch and speech synthesis for character animation. In *Proceedings of Human Factors in Computing Systems and Graphics Interface 1987,* J. H. Caroll and P. Tanner (Eds), 143-147.

[5] Waters, K., and Levergood, T. 1994. An automatic lip-synchronisation algorithm for

synthetic faces. In *Proceedings of ACM Multimedia*, 149-156.

[6] Smid, K., and Pandzic, I., S., 2002. A Conversational Virtual Character for the Web. In *Proceedings of Computer Animation* 2002, Geneva, Switzerland, 240 – 248.

[7] Beskow, J. 1995. Rule-based visual speech synthesis. In *Proceedings of ESCA - EUROSPEECH 1995*. 4th European Conference on Speech Communication and Technology, Madrid. vol.1, 299–302.

[8] Cohen, M. M., and Massaro, D. W. 1993. Modeling coarticulation in synthetic visual speech. In *Proceedings of Models and Techniques in Computer Animation*, M. Magnenat-Thalmann, and D. Thalmann, Eds., Springer-Verlag, Tokyo, 139–156.

[9] Lundeberg, M., and Beskow, J. 1999. Developing a 3D-agent for the August dialogue system. In *Proceedings from AVSP1999*, Santa Cruz, USA.

[10] Ostermann, J., and Millen, D. 2000. Talking heads and synthetic speech: An architecture for supporting electronic commerce. In *Proceedings of* ICME 2000, 71.-74.

[11] Pandzic, I. S. 2002. Facial Animation Framework for the Web and Mobile Platforms. In *Proceedings of Web3D Symposium* 2002, Tempe, AZ, USA, 27 – 34.

[12] Lee, S. P., Badler, J. B., and Badler, N. I. 2002. Eyes Alive. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques_2002* , San Antonio, Texas, USA, ACM Press New York, NY, USA, 637 – 644

[13] Ekman, P., and Friesen, W. 1969. The repertoire of nonverbal behavioral categories—Origins, usage, and coding. *Semiotica* 1:49–98.

[14] Ekman, P. 1979. About brows: Emotional and conversational signals. *Human ethology: Claims and limits of a new discipline*, M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, eds., New York: Cambridge University Press. 169–249.

[15] Cassell J., Sullivan J., Prevost S., and Churchill E. 2000. *Embodied Conversational Agents*. The MIT Press Cambridge, Massachusetts London, England.

[16] Argyle, M., and Cook, M. 1976. *Gaze and mutual gaze*. Cambridge University Press.

[17] Collier, G. 1985. *Emotional expression*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

[18] Chovil, N. 1992. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction* 25:163–194.

[19] Graf, H. P., Cosatto, E., Strom, V., and Huang, F. J., 2002. Visual Prosody: Facial Movements Accompanying Speech. In *Proceedings of AFGR 2002*, 381-386.

[20] Cassell, J., Vilhjálmsson, H., and Bickmore, T., 2001. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH 2001*, ACM Press / ACM SIGGRAPH, New York, E. Fiume, Ed., Computer Graphics Proceedings, Annual Conference Series, ACM, 477-486.

[21] Faigin, G. 1990. *The artist's complete guide to facial expression*. Watson-Guptill Publications, New York.

[22] Duncan, S. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, Oxford University Press , 23(2), 283-292.

[23] Pandzic, I. S., and Forchheimer R. 2002. MPEG-4 Facial Animation - The standard, implementations and applications. John Wiley & Sons.Lundeberg, M., and Beskow, J. 1999. Developing a 3D-agent for the August dialogue system. In *Proceedings from AVSP1999*, Santa Cruz, USA.

[24] Hiyakumoto, L., Prevost, S., and Cassell, J., 1997. Semantic and Discourse Information for Text-to-Speech Intonation. In *Proceedings of ACL Workshop on Concept-to-Speech Generation 1997,* Madrid. 47.-56.

[25] Parent, R., King, S., and Fujimura, O. 2002. Issues with Lip Synch Animation: Can You Read My Lips?, In *Proceedings of Computer Animation* 2002, Geneva, Switzerland, p.p. 3 – 10

[26] Silverman, K., Beckman, M., Pitrelli, J., Osterndorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Herschberg, J. 1992. ToBI: A Standard for Labeling English Prosody. In *Proceedings of Conference on Spoken Language*, 1992, Banff, Canada, 867-870.

[27] Hadar, U., Steiner, T., Grant, E., and Rose, F. C. 1983. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46.