Communicative and Statistical Eye Gaze Predictions

Massimo Bilvi
Department of Computer and Systems Science
University of Rome "La Sapienza"
bilvi@dis.uniroma1.it

Catherine Pelachaud
IUT of Montreuil, University of Paris 8
LINC, Paragraphe
c.pelachaud@iut.univ-paris8.fr

ABSTRACT

In this paper we propose an eye gaze model for an embodied conversational agent that embeds information on communicative functions as well as on statistical information of gaze patterns. This latter information has been derived from the analytic studies of an annotated video-corpus of conversation dyads. We aim at generating different gaze behaviors to stimulate several personalized gaze habits of an embodied conversational agent.

1. INTRODUCTION

Today state of the art 3D graphics hardware has higher performance than before and conversational agents begin to appear, in terms of appearance, more realistic. To be more believable the agents should be endowed with the communicative and expressive capacities similar to those exhibited by humans (speech, gestures, facial expressions, eye gaze, etc). In the context of the EU project MagiCster [1], we aim at building a prototype of conversational communication interface that makes use of non-verbal signals when delivering information, in order to achieve an effective and natural communication with humans or artificial agents. To this aim, we create an embodied conversational agent (ECA), Greta, that incorporates communicative conversational aspects. To determine speech-accompanying non-verbal behaviors the system relies on a taxonomy of communicative functions proposed by [19]. A communicative function is defined as a pair (meaning, signal) where meaning corresponds to the communicative value the agent wants to communicate and signal to the behavior used to convey this meaning. To control the agent we are using a representation language, called 'Affective Presentation Markup Language' (APML) where the tags of this language are the communicative functions [18]. Our system takes as input the text (tagged with APML) the agent should say. The system instantiates the communicative functions into the appropriate signals. The output of the system is the audio file and the animation file that drives the facial model (for further details see [18]).

After presenting related works, we present a gaze model that may generate various gaze behaviors. The model embeds a communicative functions model and a statistical models (Section 3). The gaze model is implementing using a Belief Network (BN) (Section **??**). In Sections 3.2 and **??** we present the algorithm and the parameters that are used to simulate several gaze pattern types.

[1]IST project IST-1999-29078, partners: University of Edinburgh, Division of Informatics; DFKI, Intelligent User Interfaces Department; Swedish Institute of Computer Science; University of Bari, Dipartimento di Informatica; University of Rome, Dipartimento di Informatica e Sistemistica; AvartarME

2. RELATED WORK

Of particular interest for our work are also approaches that aim to produce communicative and affective behaviors for embodied conversational characters (e.g., by Ball & Breese [2], Cassell et al. [6, 5], Lester et al. [15], Lundeberg & Beskow [16], Poggi et al. [20]). Some researchers concentrate on gaze models to emulate turn-taking protocols [3, 6, 7, 22, 5], or to call for the user's attention [23] to indicate objects of interest in the conversation [3, 15, 21], to simulate the attending behaviors of agents during different activities and for different cognitive actions [8]. On the other hand [9, 11, 14] use a statistical model to drive eye movements. In particular, the model of Colburn et al. [9] uses hierarchical state machines to compute gaze for both one-on-one conversation than multi-party interactions. On the other hand Fukayama et al. [11] use a two-state Markov model which outputs gaze points in the space derived from three gaze parameters (*amount of gaze*, *mean duration of gaze* and *gaze points while averted*). These three parameters have been selected based on gaze perception studies. While the previous researches focused more on eye gaze as communication channel, Lee et al. [14] an eye movement model based on empirical studies of saccades and statistical models of eye-tracking data. An eye saccade model is provided for both talking and listening modes. The eye movement is very realistic but no information on the communication functions of gaze drives the model. Most models presented so far concentrate either on the communicative aspects of gaze or on a statistical model. In this paper we propose a method that combines both approaches to get a more natural as well as meaningful gaze behavior.

3. THE EYE GAZE MODEL

The eye gaze during conversation plays an important role in regulating the comunication. It may have several functions in social interaction such as regulating the exchange of speaking turns, showing a point in focus in the speech [1, 12, 10, 13].

In previous work, we have developed a gaze model based on the communicative functions model proposed by Poggi et al. [20]. This model predicts what should be the value of gaze in order to have a given meaning in a given conversational context. For example if at a point of her speech, the agent wants to emphasize a given word, the model will out-

put that the agent should gaze at her conversant. But using only this model creates a very deterministic behavior model: at every communicative function associated with a meaning corresponds all the time the same signals. This model also does not take into account the duration that a given signal remains on the face. Indeed, this model is event-driven: it is only when a communicative function is specified that the associated signals are computed and that the corresponding behaviors may vary. Such a model used by itself has several drawbacks: first of all it does not take into account the past nor the current gaze behaviors to compute the new one, neither does it consider the duration gaze states by speaker and listener have lasted. To embed this model into temporal considerations as well as to compensate somehow missing factors in our gaze model (such as social and culture aspects) we have developed a statistical model (explained in the remaining of this paper). Thus, our model comprises two main steps:

1. *Communicative prediction*: First it applies the communicative function model as introduced in [18] and [20] to compute the gaze model so as to convey a given meaning.

2. *Statistical prediction*: The second step is to compute the final gaze behavior using a statistical model and considering information such as: what is the gaze behavior for the Speaker (S) and a Listener (L) that was computed in step one of our algorithm, in which gaze behavior S and L were previously, the durations of the current gaze of S and of L.

4. STATISTICAL MODEL

The first step of the model has already been described elsewhere [18, 20]. In the remaining of this section we concentrate on the statistical model. We use a *Belief Network* (BN) made up of several nodes (see Figure 1). Suppose we want to compute the gaze states of two agents, one being the speaker S and the other being the listener L, at time T_i the nodes are:

- **Communicative Functions Model**: these nodes correspond to the communicative functions occuring at time T_i. These functions have been further increased from the set specified in [20] to take into account Listener's functions such as back-channel and turn-taking functions. The nodes are:
  - *MoveSpeaker* S_{T_i}: the gaze state of S at time T_i. The set of possible states is $0, 1$ which correspond, respectively, to the states *look away* and *look at*.
  - *MoveListener* L_{T_i}: the gaze state of L at time T_i. The set of possible states is $0, 1$.

- **Previous State**: these nodes denote the gaze direction at time T_{i-1} (previous time). As previous nodes, the possible values are 0 and 1. We consider:
  - *PrevGazeSpeaker* $S_{T_{i-1}}$: the gaze state for S at time T_{i-1}.
  - *PrevGazeListener* $L_{T_{i-1}}$: the gaze state of L at time T_{i-1}.

- **Temporal consideration**: these nodes monitor for how long S (respectively L) has been in a given gaze state. They ensure that neither S or L will be blocked in a given state for too long.
  - *SpeakerDuration* S_D: this node is used to "force" somehow the speaker to change her current gaze state if she has been in this particular state for too long. The set of possibile states is $0, 1$ which correspond, respectively, to the states *less* than duration D (meaning that S has been for a lesser time than a given duration D in the current gaze state) and *greater* than duration D (S has been for a greater time than a given duration D in the current gaze state).
  - *ListenerDuration* L_D: same function as the SpeakerDuration node but for the listener.

- **NextGaze** (S'_{T_i}, L'_{T_i}): the gaze state for both agents at time T_i. The state is computed by setting the root nodes with the respective values and by propagating the probabilities to the leaf node. The set of possible states is: $\{ (0,0) , (0,1) , (1,0) , (1,1) \}$.

The transition from T_{i-1} to T_i is phoneme based that is at each phoneme the system instantiates the BN nodes with the appropriate values to obtain the next gaze state (S'_{T_i}, L'_{T_i}) by the BN. This model has been built using data reported in [4]. This data corresponds to interactions between two subjects lasting between 20 and 30 minute. A number of behaviors (vocalic behaviors, gaze, smiles and laughter, head nods, back channels, posture, illustrator gestures, and adaptor gestures) have been coded every 1/10th of second. Analysis of this data was done having in mind to establish two sets of rules: The first one, called 'sequence rules', refers to the time a behavior change occurs and its relation with other behaviors (does breaking mutual gaze happened by having both conversants breaking the gaze simultaneously or one after the other); while the second set of rules, called 'distributional rules' refers to probabilistic analysis of the data (what is the probability to have mutual gaze and mutual smile). The weights specified within each node of the BN have been computed using this empirical data as well as to follow the sequence and distributional sets of rules. For example, the BN has been built so that a change of state corresponding to 'breaking mutual gaze' may not happen by having both agents breaking the gaze simultaneously; that is our model does not allow that, given the previous states $S_{T_{i-1}} = 1$ and $L_{T_{i-1}} = 1$, to have the next gaze state sets to $S_{T_i} = 0$ and $L_{T_i} = 0$.

5. TEMPORAL GAZE PARAMETERS

We aim at simulating not a generic and unique gaze behavior type but to simulate personalized gaze patterns. The gaze behaviors ought to depend on the communicative functions one desires to convey as well as on factors such as the general purpose of the conversation (persuasion discours, teaching...), and personality, cultural root, social relations... Not having precise information on the influence each of such factors may have on the gaze behavior, we introduce parameters that characterized the gaze patterns. Thus we do not propose a gaze model for a given culture or social role; rather we propose parameters that control the gaze behavior itself. The parameters we are considering are:

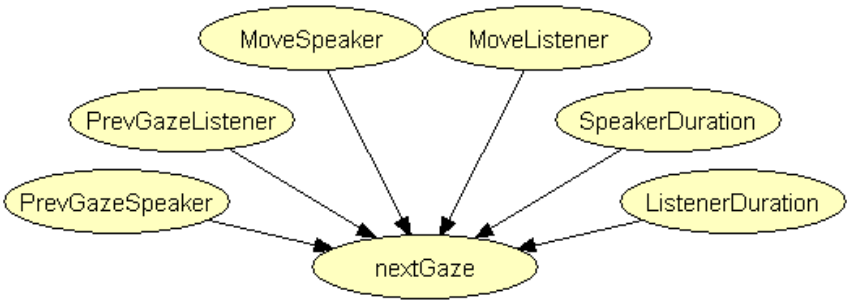

Figure 1: The belief network used for the gaze model

- $T^{max}_{S=1,L=1}$: maximun duration of mutual gaze state, $S = 1$ and $L = 1$; that is the maximum duration the mutual gaze state may remain active.
- $T^{max}_{S=1}$: maximum duration of gaze state $S = 1$.
- $T^{max}_{L=1}$: maximum duration of gaze state $L = 1$.
- $T^{max}_{S=0}$: maximum duration of gaze state $S = 0$.
- $T^{max}_{L=0}$: maximum duration of gaze state $L = 0$.

These parameters are used to set up dynamically the variables *SpeakerDuration* S_D and *ListenerDuration* L_D that are used in the BN. Their role is to control the gaze pattern in an overall manner rather than phoneme to phoneme as it is done in the BN. They provide further control on the overall gaze behavior. We notice that by varying the temporal gaze parameters we can simulate different gaze behaviors; Varying these parameters implies a change of S_D and/or L_D values thus changing the probability of having a change of gaze state. For example if we want to simulate a shy agent who glances very rapidly at the interlocutor, we can lower the values $T^{max}_{S=1}$ and $T^{max}_{S=1,L=1}$ and raise the value $T^{max}_{S=0}$. On the other hand, to simulate two agents that are very good friends and that look at each other a lot, we can raise the time value of mutual gaze $T^{max}_{S=1,L=1}$.

6. ALGORITHM FOR GAZE PERSONALIZATION

The first step is to compute the value of the BN nodes. The instantiation of the node values MoveSpeaker and MoveListener are provided by the model of communicative functions [20]. The PrevGazeSpeaker and PrevGazeListener get the values of the previous gaze state (i.e. at time T_i). SpeakerDuration S_D and ListenerDuration L_D are obtained as explained in section 3.2. NextGaze is computed by propagating the probabilities in the BN. The outcomes are probabilities for each of the four possible states of NextGaze, namely $G = \{G^{00}, G^{01}, G^{10}, G^{11}\}$. The final choice is obtained by applying a uniform distribution computation over the propabilities that are greater than a given treshold.

7. EYES ANIMATION

After applying the BN for every transition T_i (that is at every phoneme), the eyes animation for the MPEG-4 face model *Greta* [17] is computed. The output of the BN specifies only if the gaze state corresponds to "look at" or to "look away". We now have to determine the gaze direction associated to the resulting state. For a given eye state, the direction is chosen randomly with the constraints that the vertical and horizontal angles specifying the eye direction should fall off within a given range. The head direction is coordinated with the eye movements; when the eyes turn over a certain angle, the head follows the eyes movement (see Figure 3). Head movement (such as head nod) may be added to this move, allowing the head to perform movement whatever its direction is. The horizontal and vertical angles for the "look away" gaze are chosen by using a linear function ranging from 0 to d_{max}, where d_{max} is the maximum degree for the "look away" gaze, while the direction is taken randomly.

Figure 2: Close view of the eyes

Figure 3: Head-eye coordination

8. CONCLUSIONS

In this paper we have proposed a gaze model based on a communicative functions model [18, 20] and on a statistical model. These models are integrated within a *belief network* using data reported in [4]. The values of the BN nodes have been set up using results from the statistical analysis of conversation dyads [4]. To allow for the creation of personalized gaze behavior, temporal gaze parameters have been specified. The main purpose of this research is to build gaze behaviors for different agent characteristics and to investigate the effects on the quality of human-agent dialogues. Animations may be viewed at the URL:

http://www.iut.univ-paris8.fr/~pelachaud/AAMAS03

9. REFERENCES

[1] M. Argyle and M. Cook. *Gaze and Mutual gaze*. Cambridge University Press, 1976.

[2] G. Ball and J. Breese. Emotion and personality in a conversational agent. In S. Prevost, J. Cassell, J. Sullivan, and E. Churcill, editors, *Embodied Conversational Characters*. MIT Press, Cambridge, MA, 2000.

[3] J. Beskow. Animation of talking agents. In C. Benoit and R. Campbell, editors, *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, pages 149–152, 1997.

[4] J. Cappella and C. Pelachaud. Rules for Responsive Robots: Using Human Interaction to Build Virtual Interaction. In Reis, Fitzpatrick, and Vangelisti, editors, *Stability and Change in Relationships*, New York, 2001. Cambridge University Press.

[5] J. Cassell, T.W. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan. Embodiment in Conversational Interfaces: Rea. In *Proceedings of CHI99*, pages 520–527, Pittsburgh, PA, 1999.

[6] J. Cassell, C. Pelachaud, N.Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Computer Graphics Proceedings, Annual Conference Series*, pages 413–420. ACM SIGGRAPH, 1994.

[7] J. Cassell, O. Torres, and S. Prevost. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. In Y. Wilks, editor, *Machine Conversations*. Kluwer, The Hague, 1999.

[8] S. Chopra-Khullar and N. Badler. Where to look? Automating visual attending behaviors of virtual human characters. In *Autonomous Agents Conference*, Seattle, WA, 1999.

[9] A. Colburn, M. Cohen, and S. Drucker. The role of eye gaze in avatar mediated conversational interfaces. Technical Report MSR-TR-2000-81, Microsoft Corporation, 2000.

[10] P.C. Ellsworth and L.M. Ludwig. Visual behavior in social interaction. *Journal of Communication*, 22, 1972.

[11] A. Fukayama, T. Ohno, N. Mukawaw, M. Sawaki, and N. Hagita. Messages embedded in gaze on interface agents - Impression management with agent's gaze. In *CHI*, volume 4, pages 1–48, 2002.

[12] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[13] A. Kendon and M. Cook. The consistency of gaze patterns in social interaction. *British Journal of Psychology*, 60:481–494, 1969.

[14] S. Lee, J. Badler, and N. Badler. Eyes alive. In *ACM Transactions on Graphics, Siggraph*, pages 637–644. ACM Press, 2002.

[15] J.C. Lester, S.G. Stuart, C.B. Callaway, J.L. Voerman, and P.J. Fitzgeral. Deictic and emotive communication in animated pedagogical agents. In S. Prevost, J. Cassell, J. Sullivan, and E. Churcill, editors, *Embodied Conversational Characters*. MIT Press, Cambridge, MA, 2000.

[16] M. Lundeberg and J. Beskow. Developing a 3D-agent for the August dialogue system. In *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, Santa Cruz, USA, 1999.

[17] C. Pelachaud. Visual text-to-speech. In Igor S. Pandzic and Robert Forchheimer, editors, *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002.

[18] C. Pelachaud, V. Carofiglio, B. de Carolis, F. de Rosis, and I. Poggi. Embodied Contextual Agent in Information Delivering Agent. In *Proceedings of AAMAS*, volume 2, 2002.

[19] I. Poggi. Mind markers. In N. Trigo, M. Rector, and I. Poggi, editors, *Meaning and use*. University Fernando Pessoa Press, Oporto, Portugal, 2002.

[20] I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3D synthetic agent. *Special Issue on Behavior Planning for Life-Like Characters and Avatars, Journal of AI Communications*, 13(3):169–181, 2000.

[21] K.R. Thórisson. Layered modular action control for communicative humanoids. In *Computer Animation'97*, Geneva, Switzerland, 1997. IEEE Computer Society Press.

[22] K.R. Thórisson. Natural turn-taking needs no manual. In I. Karlsson, B. Granström, and D. House, editors, *Multimodality in Language and speech systems*, pages 173–207. Kluwer Academic Publishers, 2002.

[23] K. Waters, J. Rehg, M. Loughlin, S.B. Kang, and D. Terzopoulos. Visual sensing of humans for active public interfaces. Technical Report Technical Report CRL 96/5, Cambridge Research Laboratory, Digital Equipment Corporation, 1996.