Published in the Lecture Notes in Artificial Intelligence, 2008, Part II, LNAI 5178

Towards Natural Head Movement of Autonomous Speaker Agent

Marko Brkic[1], Karlo Smid[2], Tomislav Pejsa[1], Igor S. Pandzic[1]

[1] Faculty of electrical engineering and computing, Zagreb University, Unska 3,
HR-10000 Zagreb, Croatia
{marko.brkic, tomislav.pejsa, igor.pandzic}@fer.hr
[2]Ericsson Nikola Tesla, Krapinska 45, p.p. 93, HR-10 002 Zagreb
karlo.smid@ericsson.com

Abstract. Autonomous Speaker Agent (ASA) is a graphically embodied animated agent capable of reading plain English text and rendering it in a form of speech, accompanied by appropriate, natural-looking facial gestures [1]. This paper is focused on improving ASA's head movement trajectories in order to achieve facial gestures that look as natural as possible. Based on the gathered data we proposed mathematical functions that, using two input parameters (maximum amplitude and duration of the gesture) generate natural-looking head motion trajectory. Proposed functions were implemented in our existing ASA platform and we compared them with our previous head movement models. Our results were shown to a larger number of people. The audience noticed that results showed improvement in head motion and didn't detect any patterns which would suggest that animation was done with predefined motion trajectories.

1 Introduction

While animated motion pictures such as Shrek feature truly impressive and natural-looking animation of character's faces, they require tremendous effort by highly skilled and talented artists. Production of facial animations in a fully automated way, without any human intervention, while striving to reach similar levels of fidelity, is a very demanding process. In Section 2 we give a brief overview of research efforts in this field including our own research on the system we call the Autonomous Speaker Agent (ASA).

ASA (Fig. 1 shows snapshots of an ASA while talking) is an extension of a Visual Text-to-Speech (VTTS) system. A classical VTTS system produces lip movements synchronized with the synthesized speech based on the timed phonemes generated by the speech synthesis. Since VTTS system can only obtain phonetic information from the speech synthesis, it has no basis for generating realistic gestures. This problem is solved by an approach that combines the lexical analysis of input text with a statistical model describing the dynamics, frequencies and amplitudes of facial gestures [1].

ASA is based on HUman GEsturing (HUGE) [2] software architecture for production and use of statistical models for facial gestures based on inducement of arbitrary kind. An inducement is a signal that occurs in parallel to the production of facial gestures in human behaviour and that may have a statistical correlation with the occurrence of facial gestures, e.g. text that is spoken, audio signal of speech, bio signals, emotions etc. The correlation between the inducement signal and facial gestures is used to first build the statistical model of facial gestures based on a training corpus consisting of sequences of gestures and corresponding inducement data sequences. In the runtime phase, the raw, previously unknown inducement data is used to trigger (induce) the real time facial gestures of the agent based on the previously constructed statistical model.

Fig. 1. Autonomous Speaker Agent

In first version of ASA, head motion animation was implemented using sine function trajectory. This paper expands on this research by trying to improve head motion of facial gestures in order to get more natural-looking facial gestures.

In order to capture the dynamics of real gestures, we analyzed recorded video clips of real professional TV speakers as a starting point for extraction of head movement coordinates. Section 3 explains which facial gestures were analyzed and how measurements were done in order to obtain head movement trajectories.

The following two sections deal with proposing the facial gestures animation functions by fitting a mathematical model to the recorded motion trajectories. Section 4 explains the analysis of data gathered from nod gesture motion analysis and proposes a way to interpret this data using trigonometric functions. In a similar way Section 5 explains swing motion and proposes a way to interpret this gesture.

Section 6 elaborates achievements we made in this project and we conclude the paper with the conclusion and a discussion of future work.

2 Background

State of the art research on real-time natural facial gesture animation is very intensive field. Busso et al. [3] present a novel data-driven approach to synthesize natural human head motions driven by speech prosodic features. The problem was modeled as classification of discrete representations of head poses. Hidden Markov Models (HMM) [4] were used to learn the temporal relation between the dynamics of head motion sequences and the prosodic features. Using new speech material, the HMM works as a sequence generator for the most likely head motion sequences. In the end bi-grams and spherical cubic interpolation techniques are used to smooth the synthesized sequence. Hofer et al. [5] models longer units of motion and speech and to reproduce their trajectories during synthesis, they utilize a promising time series stochastic model called “Trajectory Hidden Markov Models” [6]. The BEAT system [7] uses linguistic and contextual information contained in a text to control the movements of 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 research in nonverbal conversational behavior. That mapping also depends on the knowledge base of an Embodied Conversational Agent ECA environment. That knowledge base is populated by a user who animates the ECA so the production of the facial gestures is not automatic. Furthermore, the system only provides output for animation systems but it does not introduce any animation model for ECA facial gestures. Also, in the current set of supported nonverbal behaviors, head nods and eyes blinks must be included.

In our approach facial gesture animation is driven by linguistic and contextual structure of uttered English text. The goal is to enable the ASA to perform gestures that are not only dynamically correct, but also correspond to the underlying text [1]. Input for animation models are facial gesture maximal amplitude and duration. Each gesture trajectory was nose tip movement that needed to be represented with some function (Fig. 2).

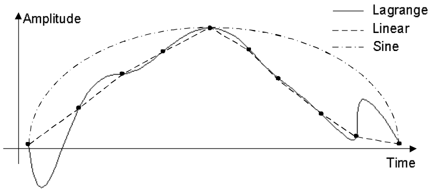

Fig. 2. Linear and Lagrange implementation

First approach is to connect neighboring coordinates with linear function so each gesture has its own subintervals (time frames) defined with those linear functions. If ASA animation module requests amplitude calculation for an exact time point, we have to call the linear function the time frame of which includes this time point.

Second approach is to use Lagrange interpolation. Theory says that a set of N points can be interpolated by a polynomial of degree at most N-1 [8]. If we had ten coordinates for one gesture (Fig. 2.) after using interpolation we would get one function of degree (at most) nine. Resulting polynomial is:

$$P(x) = \sum_{j=1}^{n} \left(y_j \prod_{\substack{k=1 \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k} \right) \qquad (1)$$

Those two approaches aren't good enough for several reasons: linear implementation has sharp edges and observer can notice that virtual character has discontinued movement at interval boundaries; Lagrange implementation has sudden large amplitude movement at the beginning and at the end of intervals; there is only one time interval for each gesture (with constant length and maximal amplitude) and after observing ASA after some time it is obvious that head always moves in the same way.

Trigonometry sine implementation solves the problem regarding trajectory constant length and amplitude. However, implementation still doesn't solve the problem of describing natural-looking gestures since we know that sine function is not an ideal representation of facial motion (refer to Fig 2.). This paper explains how sine implementation was improved.

3 Head Movement Analysis

Within head movement we distinguish two types of facial gestures: nods and swings. Nod is an abrupt swing of the head with a similarly abrupt motion back and swing is an abrupt swing of the head without the back motion with following directions: up, down, left, right and diagonal (divided into four quadrants). We have following nod directions: up and down, down and up, left and right, right and left and diagonal (divided into four quadrants). The parameters that are important for head movement are maximum amplitude and duration. To get these parameters we analyzed 56 video clips originating from Ericsson's “5minutes” web cast. To describe those videos we used HUGE [2] gesture XML files containing gesture type, maximum amplitude, start and end time for every identified head movement. The best way of tracking head movement is by observing the speaker’s nose tip position as a reference point. Obtained coordinates were used to get graphical representation of head movement trajectory. In order to ensure that all measurements taken from these videos are in the same proportion, we normalized them using the Mouth-Nose Separation unit (MNS0): a distance between (in pixels) nose tip and upper lip. Excel was used to create graphs from obtained data and those graphs were used for gesture motion analysis. In the next chapter we elaborate on nod head movements.

4 Synthesizing Nod Gestures

In this chapter we explain graphs of gathered data for nod gesture and propose a function that describes that specific gesture trajectory. After all measurements were done and everything was placed in tables, graphs were made using Microsoft Excel for each gesture. Since all measurements in nod gestures show same properties, graph of nod up gesture is enough to comment all nod gestures. The only difference between all nod motions is in motion direction.

Functions drawn in graphs using obtained data were used for gesture motion analysis. After graph analysis it was concluded that trigonometry functions were the best solution for describing nod trajectory functions. Since most of the functions on graph were not exact sine functions some statistic analysis needed to be done. After analysis we divided our proposed functions in three types where the first one is a function where its time of maximum amplitude value is to the left of the function duration midpoint (D/2), the second one is ideal sine function (time of maximum amplitude is exactly in the middle) and the third one with maximum positioned to the right of the duration midpoint.

First and third functions are divided in two intervals where time of maximum is the separation point. Fig. 3. shows all three functions with characteristic parameters.

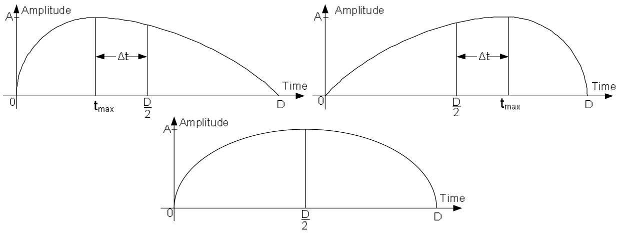

Fig. 3. Three characteristic functions of nod gesture.

Sine function is used to describe the ascending part of the function, and cosine function is used to describe the descending part of function. In order to calculate percentages by which three different functions appear we analyzed all data from obtained functions and compared maximum amplitudes of each function with amplitude value from half of the interval of the same function. Percentages for nod up are given in table 3.

During data analysis we couldn't find any connection between Δt value (a difference in milliseconds between function maximum and functions half-interval value) and facial gesture. We only concluded that Δt value can be inside [0, D/4] interval. This interval was used to randomly generate Δt value for each head motion.

Table 1. Half interval deviation percentage values for nod up.

	$t_{max} < D/2$	$t_{max} = D/2$	$t_{max} > D/2$
Nod up	41%	11%	48%

Function definitions shown in Table 4 are same for all nod gestures. With these three functions we tried to describe the behavior of node gesture. By implementing Δt we managed to describe different velocities of head movement in two different directions, one from the start position to the maximum and second one in the opposite direction. The next chapter explains swing gesture implementation.

Table 2. Function definitions for nod gesture.

	$0 < t <= t_{max}$	$t_{max} < t < D$
$t_{max} < D/2$	$f(t) = A\sin\left(\frac{t}{\frac{D}{2} - \Delta t} \cdot \frac{\pi}{2}\right)$	$f(t) = A\cos\left(\frac{t - \frac{D}{2} + \Delta t}{\frac{D}{2} + \Delta t} \cdot \frac{\pi}{2}\right)$
$t_{max} = D/2$	$f(t) = A\sin\left(\frac{t\pi}{2D}\right)$	
$t_{max} > D/2$	$f(t) = A\sin\left(\frac{t}{\frac{D}{2} + \Delta t} \cdot \frac{\pi}{2}\right)$	$f(t) = A\cos\left(\frac{t - \frac{D}{2} - \Delta t}{\frac{D}{2} - \Delta t} \cdot \frac{\pi}{2}\right)$

5 Synthesizing Swing Gestures

This chapter explains graphical results of swing gesture measurements and proposes a function that describes its motion trajectory. Swing gesture is a head movement without return motion to starting position (like in nod gestures). Previous sine implementation used first quarter of sine period for reaching maximum coordinate, but analysis done for this movement showed that after the first half of swing gesture duration maximum coordinate is reached with linear trajectory (Fig 4.).

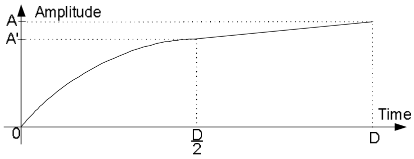

Fig. 4. Two characteristic functions of swing gesture.

Based on graph in Fig. 4. all swing functions are defined in Table 5:

Table 3. Function definitions for swing gesture.

0 < t <= D/2	$f(t) = A\sin\left(\frac{t\pi}{2D}\right)$
D/2 < t < D	$f(t) = \frac{2t(A-A')}{D} + 2A' - A,\ A' = A\sin\left(\frac{\pi}{4}\right)$

By dividing the function into two segments we tried to simulate the increase of head movement velocity in the second part of function duration.

6 Results

To determine the effect of new head gestures on naturalness of virtual characters' motion, we conducted a subjective test. We generated two animation sequences, one with old sine-based head movement and another with new trigonometric gestures. Both sequences featured the same virtual character presenting the same text and making the same facial and head gestures. The sole difference between the sequences was the mathematical model used for head movement.

The test sequences were shown to groups of subjects. There was a total 185 subjects. All of them are students in fields of Computer Science and Telecommunications and none of them are familiar with our area of research. After viewing a test sequence, the subjects were asked to score the virtual character by answering the following questions:

1. Did the character's head movements appear natural (5) or not (1)?

2. Did the character on the screen appear interested in (5) or indifferent (1) to you?
3. Did the character appear engaged (5) or distracted (1) during the conversation?
4. Did the personality of the character look friendly (5) or not (1)?

Note that possible scores are 1 to 5, where higher scores correspond to more positive attributes in the speaker.

Our testing has shown that new head movements have no measurable effect on subjects' subjective perception of the virtual character. These findings were confirmed when we performed one-way ANOVA on the results and determined that virtual characters in the two test sequences did not receive significantly different scores on any of the 4 questions.

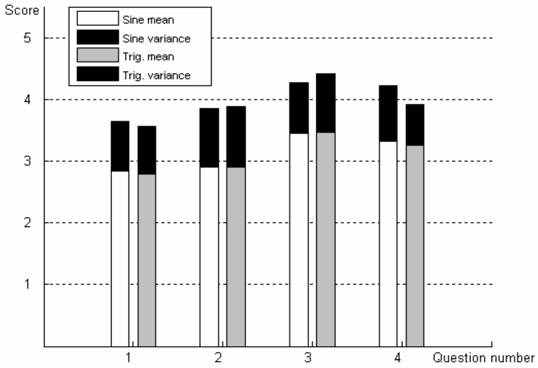

Fig. 5. Test scores for virtual characters with old and new head movements.

However, it must be noted that a small number of subjects were explicitly asked to compare head gestures in the two animation sequences and they invariably noted an improvement in naturalness for trigonometric head motion.

7 Conclusion and Future Work

ASA combines lexical analysis of English text and statistical models of facial gestures in order to generate the gestures related to the spoken text. Current state of the art work in the field of head movement animation trajectories is mostly based in correlating head movement with the prosody features of the speech. In our approach we use the correlation between the lexical structure of spoken text and head motion. Also, we introduce rather simple mathematical techniques as a base for animation model. Our goal is to test practical value of computationally low intensive models. Progress explained in this paper should be just a beginning in improving head gestures of virtual characters. We still have to analyze diagonal head motion and eyebrows motion in order to improve the basic sine trajectory model. Future work should include four main diagonal implementations both in nod and swing head motions and eyebrow raise and frown motion. Goal of this paper is to improve animation of virtual speakers in order to achieve better automatic generation of head gestures for applications such as newscasters and storytellers. We improved nod and swing motion and gave a solid base ground for further improvement and research. This future research should be based on a larger data corpus.

Acknowledgment

This work was partly carried out within the research project "Embodied Conversational Agents as interface for networked and mobile services" supported by the Ministry of Science, Education and Sports of the Republic of Croatia and within the research project "Summer Camp 2007" supported by Ericsson Nikola Tesla.

References

1. Smid K., Pandzic I. S., Radman V.: Autonomous Speaker Agent, in Proceedings of the Computer Animation and Social Agents Conference CASA 2004, Geneva, Switzerland.
2. Smid K., Zoric G., Pandzic I. S.: [HUGE] Universal Architecture for Statistically Based HUman GEsturing, Lecture Notes on Artificial Intelligence LNAI 4133, pp. 256-269 (Proceedings of the 6th International Conference on Intelligent Virtual Agents IVA 2006, Marina del Ray, USA, 2006.)
3. Busso C., Deng Z., Neumann U., Narayanan S: Natural head motion synthesis driven by acoustic prosodic features. Computer Animation and Virtual Worlds, Volume 16, Issue 3-4 pp. 283 – 290, 2005.
4. Rabiner L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77 (2), p. 257–286, February 1989.
5. Hofer G., Shimodaira H., Yamagishi J.: Speech driven head motion synthesis based on a trajectory model. International Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2007 posters, article no. 86, San Diego, USA, 2007
6. Zen H., Tokudaa K., Kitamura T.: Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. Computer Speech & Language Volume 21, Issue 1, January 2007, Pages 153-173
7. Cassell J., Vilhjálmsson H., Bickmore T.: BEAT: the Behavior Expression Animation Toolkit. SIGGRAPH 2001, ACM :477–486, 2001.
8. Abramowitz M., Stegun I. A.: (Eds.). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, pp. 878-879 and 883, 1972.