Published in Lecture Notes in Computer Science. 4012 / 2006; pp. 401 - 410; 2006.

Automated Gesturing for Embodied Agents

Goranka Zoric[1], Karlo Smid[2] and Igor S. Pandzic[1]

[1] Department of Telecommunications,
Faculty of Electrical Engineering and Computing, University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
[2] Ericsson Nikola Tesla
Krapinska 45, HR-10002 Zagreb, Croatia F-91405
{Goranka.Zoric, Igor.Pandzic}@fer.hr, karlo.smid@ericsson.com

Abstract. In this paper we present our recent results in automatic facial gesturing of graphically embodied animated agents. In one case, conversational agent is driven by speech in automatic Lip Sync process. By analyzing speech input, lip movements are determined from the speech signal. Another method provides virtual speaker capable of reading plain English text and rendering it in a form of speech accompanied by the appropriate facial gestures. Proposed statistical model for generating virtual speaker's facial gestures, can be also applied as addition to lip synchronization process in order to obtain speech driven facial gesturing. In this case statistical model will be triggered with the input speech prosody instead of lexical analysis of the input text.

1 Introduction

Conversational Agent is a graphically embodied animated agent capable of human-like behavior, most importantly talking and natural-looking facial gesturing. A human face can express lots of information, such as emotions, intention or general condition of the person. In this article, we concentrate on two ways for automatic facial animation - Lip sync and Visual Text-to-Speech (VTTS). The goal is to animate the face of a Conversational Agent in such a way that it realistically pronounces the given text. In order to appear realistic, the produced face animation should also include facial gesturing.We have achieved this in case of VTTS through statistical modeling of behavior and plan to extend this method to lip sync as well. Lip sync produces lip movements synchronized with the input speech. The most important issue in Lip Sync research is Audio to Visual mapping which consists of the speech analysis and classification in visual representatives of the speech (Fig. 1). We present a new method for mapping natural speech to lip shape animation in real time. The speech signal is classified into viseme classes using neural networks. The topology of neural networks is automatically configured using genetic algorithms.

We propose a new approach, Autonomous Speaker Agent that combines the lexical analysis of input text with the statistical model describing frequencies and amplitudes of facial gestures. Using a lexical analysis of input text to trigger

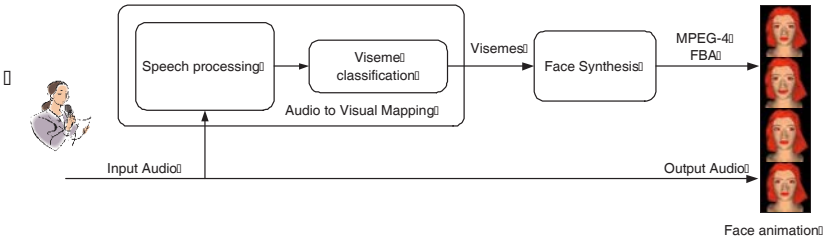

Fig. 1. Schematic view of a lip sync system

the statistical model, a virtual speaker can perform gestures that are not only dynamically correct, but also correspond to the underlying text. Fig. 2 depicts training and content production processes. The Section 2 and Section 3 describe

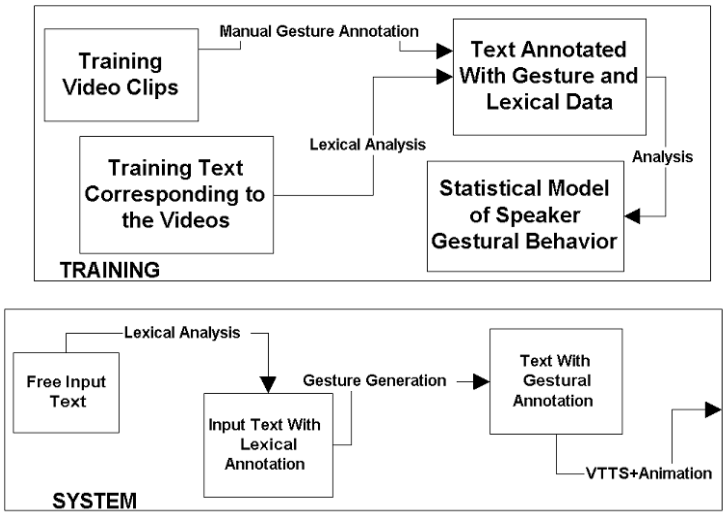

Fig. 2. Schematic view of an Autonomous Speaker Agent

respectively automatic lip sync system and autonomous speaker agent in more details. The paper closes with a conclusion and a discussion of the future work.

2 Automatic Lip Sync

Lip synchronization is the determination of the motion of the mouth and tongue during speech [1]. To make lip synchronization possible, position of the mouth

and tongue must be related to characteristics of the speech signal. There are many acoustic sounds that are visually ambiguous. Therefore, there is a many-to-one mapping between phonemes and visemes, where viseme is a visual representation of phoneme [2].

The process of automatic lip sync consists of two main parts (Fig. 1). The first one, audio to visual mapping, is key issue in bimodal speech processing. In this first phase speech input is analyzed and classified into viseme categories. In the second part, calculated visemes are used for animation of virtual character's face.

2.1 The proposed lip synchronization algorithm

Our system for automatic lip sync is suitable for real-time and offline applications. It is speaker independent and multilingual. We use visemes as the main classification target. Speech is first segmented into frames. For each frame most probable viseme is determined. Classification of speech in viseme classes is performed by neural networks. Then MPEG-4 compliant facial animation is produced. What follows is short description of the system components.

Phoneme database As training data, a set of phonemes is collected. For every phoneme, three different samples were recorded by nine test subjects. This gives 27 versions of each phoneme in our database. These phonemes are manually mapped onto MPEG-4 visemes, and in doing so the database is organized in 14 classes, each corresponding to one MPEG-4 viseme. On average, each viseme class is represented by 50 samples in the database.

For fine tuning of animation, phonemes specific for certain language might be added in the database.

Audio to Visual Mapping In order to synchronize the lips of a computer generated face with the speech, speech must be first preprocessed and then classified into visemes.

The Mel-Frequency Cepstrum Coefficients (MFCC) representation of the speech is chosen as first step in preprocessing the speech.

MFCC is audio feature extraction technique which extracts parameters from speech similar to ones that are used by humans for hearing speech, while, at the same time, deemphasizes all other information.

Additionally, Fisher linear discriminant transformation (FLDT) is done on MFCC vectors to separate classes. If there is no separation between classes before FLDT, transformation will not enhance separability, whereas if there is only slight distinction between classes, the FLDT will separate them satisfactory.

In order to use MFCCs on the speech signal, frame length and the dimension of the MFCC vectors must be determined. The choice is frame length of 256 samples and 12 dimensional MFCC vector. Overlapping of the frames is used to smooth transition from frame to frame.

The phoneme database is now used as a training set in order to train NN. Every frame of the speech is classified in the correct viseme class. When correct viseme is chosen, it can be sent to animated face model.

MPEG-4 Face Animation Face animation (FA) is supported in MPEG-4 standard [2]. MPEG-4 FA specifies a face model in its neutral state, a number of feature points (FPs) and a set of Facial Animation Parameters (FAPs). Each FAP corresponds to a particular facial action deforming a face model in its neutral state. The first group of FAPs contains high-level parameters, visemes and expressions. Only 15 static visemes are included in the standard set.

Facial animation can be generated for any parameterized face model for speech animation if the visemes are known.

2.2 Training neural networks for AV mapping using GA

Neural networks (NNs) are widely used for mapping between the acoustic speech and the appropriate visual speech movements [4]. Many parameters, such as weights, topology, learning algorithm, training data, transfer function and others can be controlled in neural network. A major unanswered question in NN research is how best to set a series of configuration parameters so as to maximize the network's performance. As training neural network is an optimization process where the error function of a network is minimized, genetic algorithms can be used to search optimal combination of parameters.

Genetic algorithms (GA) are a method for solving optimization or search problems inspired by biological processes of inheritance, mutation, natural selection and genetic crossover. A conventional GA consists of coding of the optimization problem and set of the operators applied on the set of possible solutions [5].

GAs might be used to help design neural networks by determining [6]:

- *Weights.* Algorithms for setting the weights by learning from presented input/output examples with given fixed topology often get stuck in local minima. GAs avoid this by considering many points in the search space simultaneously.
- *Topology* (number of hidden layers, number of nodes in each layer and connectivity). Determining a good/optimal topology is even more difficult - most often, an appropriate structure is created by intuition and time consuming trial and error.
- *A suitable learning rule.*

Training NNs In our approach, we use multilayer feedforward networks to map speech to lip movements. These kind of neural networks are widely used and operate in a way that an input layer of nodes is projected onto output layer through a number of hidden layers. Backpropagation algorithm is used as training algorithm for adjusting weights. 15 networks, each for every viseme

class, is trained since phonemes that are visual ambitious, do not need to be separated.

The 12-dimensional MFCC vectors are used as inputs to networks. For each viseme class, a NN with 12 inputs, a number of hidden nodes and 1 output is trained. The number of hidden layers and the number of nodes per each layer should have been determined for each network. This is laborious and time consuming work since the training session must be run until the result is satisfactory. In order to avoid time consuming trial and error method, we introduced simple genetic algorithm to help find suitable topology for our NNs.

GA and NNs in our approach Since the design of neural network is optimized for a specific application, we had to find suitable network for our lip sync application. As determining a good or optimal topology is even the most difficult task in design of NN, we tried to solve this problem with GAs.

In our example, given the learning rule, we used GA for training a back-propagation feedforward network to determine near optimal network topology, including the number of hidden layers and the number of units within each layer.

We use simple genetic algorithm [7], where number of genes specify the number of hidden layers (n). Gene maximum and minimum values are defined in the range from zero to m, determining the number of nodes per layer. If a value of the single gene is set to zero, the number hidden layers is decreased, so practically it ranges from zero to n.

By using genetic algorithms, the process of designing neural network is automated. Once the problem to be solved is coded and GA parameters are determined, the whole process is automated. Although it is still a time consuming work, much time is saved since all work is done automatically by computer.

2.3 Implementation

Constructing database and creation of 15 neural networks have to be done only once. In training process, network's biases and weights are extracted and saved for later use. Together with Fisher matrix (obtained by calculating FLDT), biases and weights matrix are loaded in the application.

Fig. 3 shows GUI (Graphical User Interface) of our application. Application captures speech from the microphone and segments it into frames of 256 samples. When a frame has been captured, data is stored and calculations are performed during capturing of the next frame. These calculations consist of MFCC extraction and simulation of 15 networks. The outputs are added to outputs from the previous frame. Every fourth frame, the viseme class that has the largest sum of output values from NNs is presented on the screen [3]. It is important that calculation time does not exceed time needed for recording of a frame (in case of 16 kHz, 16 bit coding takes 16 ms).

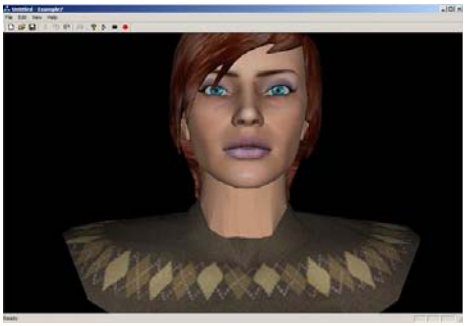

Fig. 3. GUI of our application

3 Autonomous Speaker Agent

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. We distinguish three main classes of facial gestures [8]: head, eyes and eyebrows movement. Within each class we distinguish specific gestures, each characterized by their particular parameters. Table 1 shows the types of facial gestures as identified during our data analysis. We introduce symbols incorporating both a gesture type and a movement direction.

3.1 Lexical analysis of English text

The speech analysis module [9] 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 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. 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 have made simplified version of morphologic and semantic analyzer extending WordNet 2.0 database [10]. In order to determine the correct word type based on the output queried from the extended WordNet 2.0 database, we must pass multiple times through whole sentence and apply various English grammatical rules.

3.2 Statistical model of gestures

As a training set for our analysis, we used the footage showing the newscasters presenting news. We investigated three female and two male Swedish newscasters. Observing those news casting clips, we marked the starting and ending

Head	Nod	An abrupt swing of a head with a similarly abrupt motion back. We have four nod directions: up and down (^), 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	An abrupt swing of a head without a back motion. Sometimes 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	Sometimes follows swing movement. Returns head in central position.
Eyes	Movement in various directions	The 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	Eyebrows go up and down (^^).
	Frown	Eyebrows go down and up ( ).

Table 1. The specification of facial gestures

frames for every eye blink, eyebrow raise and head movement. Analyzing those frames, the speakers Mouth-Nose Separation unit (MNS0) value, facial gesture amplitude value, facial gesture type and direction were determined. In our model, the basic unit which triggers facial gestures is a word. The raw data 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. Such statistics exist for each gesture type and for each speech context we treated. They are built into the decision tree (Fig. 4) 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.

3.3 The System

The input to the system is plain English text. It is processed by a lexical analysis which converts it into an XML format with lexical tags. The facial gesture module is the core of the system - it actually inserts appropriate gestures into text in the form of special bookmark tags that are read by the TTS/MPEG-4

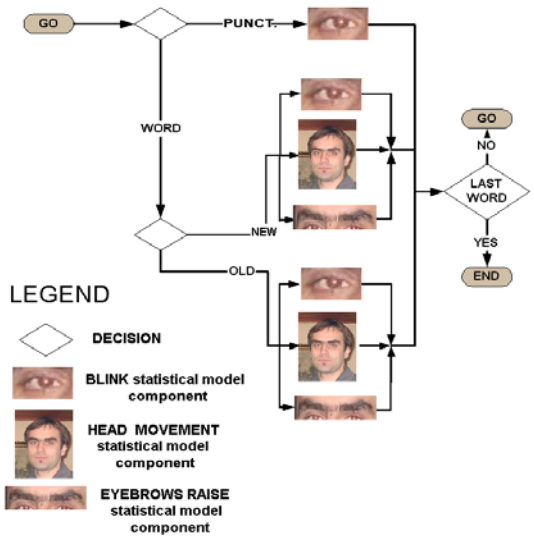

Fig. 4. Decision tree with components of the statistical model

Encoding module. While the Microsoft Speech API (SAPI) [11] Text To Speech (TTS) 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 bit stream is encoded with the appropriate viseme and facial gestures animation. 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 Fig. 4.

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 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.

Results We conducted a subjective test in order to compare our proposed statistical model to simpler techniques. We synthesized facial animation on our face model using three different methods. In the first (Type 1), head and eye movements were produced playing animation sequence that was recorded by tracking movements of a real professional speaker. In the second (Type 2), we produced a facial animation using the system described in this paper. In the third (Type 3), only the character's lips were animated.

The three characters (Type 1, Type 2 and Type 3) were presented in random order to 29 subjects.The subjects were asked the following questions:

Q1: Did the character on the screen appear interested in (5) or indifferent (1) to you?

Q2: Did the character appear engaged (5) or distracted (1) during the conversation?

Q3: Did the personality of the character look friendly (5) or not (1)?

Q4: Did the face of the character look lively (5) or deadpan (1)?

Q5: In general, how would you describe the character?

Note that higher scores correspond to more positive attributes in a speaker. For questions 1 to 4, the score was graded on a scale of 5 to 1. Fig. 5 summarizes

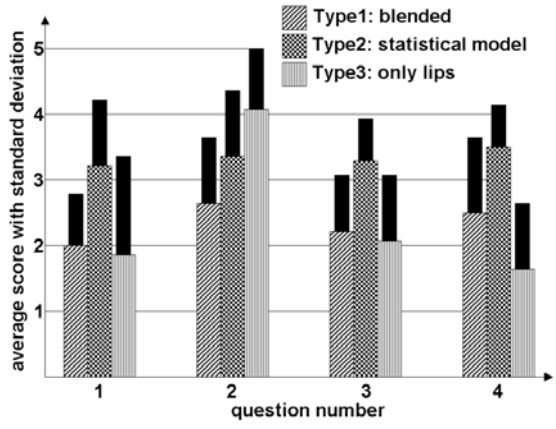

Fig. 5. Results of subjective evaluations

the average score and standard deviation (marked with a black color). From the figure, we can see that the character of type 2 was graded with the highest average grade for all questions except for the Q2. The reason for that is because type 3 character only moves its lips and its head is static. This gave the audience the impression of engagement in the presentation. According to general remarks in Q5, our character had a more natural facial gesturing and facial gestures were coarticulated to some extend. Head movements and eye blinks are related to the spoken text. However, eyebrow movements were with unnatural amplitudes and were not related to the spoken text.

4 Conclusion and Future Work

In this paper we present two methods for automatic facial gesturing of graphically embodied animated agents. One approach is an autonomous speaker agent with full facial animation produced automatically from the plain text. 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. An extension to Embodied Conversational Characters is a logical item for future work, adapting and extending the statistical model to include more complicated gesturing modes and speech prosody that occur in a conversation.

As well, we propose our approach for lip sync system by speech signal analysis. Speech is classified into viseme classes by neural networks and GA is used for obtaining optimal NN topology. By introducing segmentation of the speech directly into viseme classes instead of phoneme classes, computation overhead is reduced, since only visemes are used for facial animation. Automatic design of neural networks with genetic algorithms saves much time in the training process. Moreover, better results are achieved than with manual search of network configuration.

However, a face that only moves the lips, looks extremely unnatural because natural speech always involves facial gestures. Our next step will be to extend automatic lip sync system with the similar statistical model for facial gestures as proposed here in order to generate facial expressions in addition to lip movements from the speech signal. But in case of speech driven facial gesturing, statistical model will be based on the input speech prosody, instead of lexical analysis of the text.

5 Acknowledgment

The initial version of this lip sync system has been implemented by A. Axelsson and E. Bjrhall as part of their master thesis of Linkping University [3] and in collaboration with Visage Technologies AB, Linkping, Sweden. This work is also partly supported by Visage Technologies. Research on autonomous speaker agent is partly supported by Ericsson Nikola Tesla (ETK).

References

1. McAllister, D.F., Rodman, R.D., Bitzer, D.L., Freeman, A.S.: Lip synchronization for Animation, Proceedings of SIGGRAPH 97, Los Angeles, CA, 1997.
2. Pandic, I.S., Forchheimer, R., Editors, MPEG-4 Facial Animation - The Standard, Implementation and Applications, John Wiley & Sons Ltd, ISBN 0-470-84465-5, 2002.
3. Axelsson, A.,Björhall, E., Real time speech driven face animation, Master Thesis at The Image Coding Group, Dept. of Electrical Engineering at Linkping University, Linkping 2003.
4. Dávila, J.J., Genetic optimization of neural networks for the task of natural language processing, dissertation, New York, 1999.
5. Rojas, R., Neural networks, A Systematic Introduction, Springer-Verlag Berlin Heidelberg, 1996.

6. Jones, A.J., Genetic algorithms and their applications to the design of neural networks, Neural Computing & Applications, 1(1):32-45, 1993.
7. Black Box Genetic algorithm, http://fdtd.rice.edu/GA/
8. Pelachaud, C., Badler, N., Steedman, M., Generating Facial Expressions for Speech. Cognitive, Science, 20(1), 1-46, 1996.
9. Radman, V., Leksicka analiza teksta za automatsku proizvodnju pokreta lica, Graduate work no. 2472 on Faculty of Electrical Engineering and Computing, University of Zagreb, 2004.
10. http://www.cogsci.princeton.edu/ wn/
11. Microsoft Speech Technologies, http://www.microsoft.com/speech