Conversational Informatics: An Engineering Approach Edited by Toyoaki Nishida © 2007 John Wiley & Sons, Ltd

9

Facial Gestures: Taxonomy and Application of Non-Verbal, Non-Emotional Facial Displays for Embodied Conversational Agents

Goranka Zoric, Karlo Smid, Igor S. Pandzic

Abstract

Facial displays are an extremely important communication channel fulfilling a wide variety of functions in discourse and conversation. Humans use them naturally, often subconsciously, and are therefore very sensitive to the application of such displays in their computer-generated correspondents, Embodied Conversational Agents (ECA). In this chapter, we aim to provide an extensive survey of one class of facial displays, the facial gestures. Facial gestures include various nods and head movements, blinks, eyebrow gestures and gaze, i.e., all facial displays except explicit verbal and emotional displays (visemes or expressions such as smile). Consciously or subconsciously, facial gestures play an important role both in discourse and in conversation. They are instrumental in turn taking, emphasizing, providing rhythm, and can be connected to physiological functions. While verbal and emotional displays may be regarded as the explicit, perhaps even obvious, the facial gestures are less tangible - yet they are largely responsible for what we intuitively call natural behavior of the face. In other words, an ECA pronouncing a sentence using perfect coarticulation mechanism for the lips and displaying a carefully modeled expression of surprise will still look unnatural if the facial gestures are not right as well. It is therefore extremely important for an ECA to implement facial gestures well. While there is a large body of knowledge on this topic both from psychology and ECA literature, it is quite scattered. Existing ECA implementations typically concentrate on some aspects of facial gestures but do not cover the complete set. We attempt to provide a complete survey of facial gestures that can be useful as a guideline for their implementation in an ECA. Specifically, we provide a systematically organized repertoire of usual facial gestures. For each gesture class, we provide the information on its typical usage in discourse and conversation, conscious or subconscious causes, any available knowledge on

typical dynamics and amplitude of the gesture. Finally we provide an example of a practical system implementation as a case study.

4.1 Introduction

Communication is the information exchange process. Humans communicate in order to share knowledge and experiences, to tell who they are and what they think, or to cooperate with each other. Communication implies bidirectional interchange of messages. One-way transfer of information is inefficient. In one-way communication there is no proof that what is heard is what is intended. It is necessary to have interaction between people trying to communicate. Therefore, a feedback from the listener is required in order to avoid misunderstanding.

In attempt to transfer their meaning to another person, humans use different methods or channels. Communication can be verbal or nonverbal, and often mixed. Albert Mehrabian, a psychologist, in his studies in 1971 [1] comes to conclusion that the non-verbal accounts for 93% of the message while words account for 7%. The above, disproportionate influence of non-verbal becomes only effective when the communicator is talking about their feelings or attitudes, since experiments were conducted dealing with communications of feelings and attitudes (i.e., like-dislike). Anthropologist Ray Birdwhistell who studied non verbal communication extensively in the 1960's claimed that in conversation about 35% of the message was carried in the verbal modality and the other 65% in the non verbal [2].

Although psychologists still argue about the percentage of information non-verbally exchanged during a face-to-face conversation, it is clear that the non-verbal channel plays an important role in understanding human behaviour. If people rely only on words to express themselves, that can lead to difficulties in communication. Words are not always associated with similar experiences, similar feelings or even meaning by listeners and speakers. That is why non-verbal communication is important. A situation when there is no consistency, i.e. one signal is being said and another is shown can be very misleading to another person. When listeners are in doubt they tend to trust the non-verbal message since disturbances in nonverbal communication are "more severe and often longer lasting" than disturbances in verbal language [7] [3].

Non-verbal communication refers to all aspects of message exchange without the use of words. It includes all expressive signs, signals and cues (audio, visual, etc.) apart from manual sign language and speech. Non-verbal communication has multiple functions. It can repeat the verbal message, accent the verbal message, complement or contradict the verbal message, regulate interactions or substitute for the verbal message (especially if it is blocked by noise, interruption, etc.).

There are a number of categories into which non-verbal communication can be divided into:

- *Kinesics (body language).* Non-verbal behaviour related to movement of the body. Includes facial expressions, eye movements, gestures, posture, and the like.
- *Oculesics (eye contact).* Influence of visual contact on the perceived message that is being communicated.
- *Haptics (touch).* Touching behaviour.
- *Proxemics (proximity).* Concerned with personal space usage.
- *Paralanguage (paralinguistics).* Non-word utterances and other non-verbal clues relatively closely related to language use.
- *Chronemics.* Use of time, waiting, pausing.
- *Silence.* Absence of sound (muteness, stillness, secrecy).
- *Olfactics (smell).*
- *Vocalics (vocal features of speech).* Tone of voice, timbre, volume (loudness), speed (rate of speech).
- *Physical appearance and artifacts.* Physical characteristics of body, clothing, jewellery, hairstyle...

- *Symbolism (semiotics).* Meaning of signs and symbols.

Non-verbal communication can be both conscious and subconscious. Culture, gender and a social status might influence the way non-verbal communication is used. There are certain rules that apply to non-verbal communication. It has a form, a function and a meaning, all of which may be culturally specific [2] (e.g. some cultures forbid direct gaze while others find gaze aversion an offense [5]). Facial expressions of primary emotions (e.g. disgust, surprise) are universal across cultures as the psychologist Paul Ekman and his colleagues have shown [6].

Non-verbal cues may be learned, innate or mixed [7]. Some of them are clearly learned (e.g. eye-wink, thumbs-up) and some are clearly innate (e.g. eye-blink, facial-flushing). Most of non-verbal cues are mixed (e.g. laugh, shoulder-shrug), because they originate as innate actions, but cultural rules and environment shaped their timing, energy and use.

What arises from the above is that in a face to face conversation, much can be said even without words. This fact should be kept in mind when generating computer correspondents of humans – ECA. Just as a human-human interaction, a human-computer interaction should consist of two way communication channel for a verbal and a non-verbal message exchange. By building the human-computer communication on the rules of human-human communication the ECA behaviour will be more like human behaviour. Therefore, humans are more likely to consider computers human-like and use them with the same efficiency and smoothness that characterizes their human dialogs [8].

In this article we will attempt to give a complete survey of facial gestures that can be useful as guideline for their implementation in an ECA. Facial gestures include all facial displays except explicit verbal and emotional displays. A facial gesture is a form of non-verbal communication made with the face or head, used continuously instead of or in combination with verbal communication. There is a number of different facial gestures that humans use in everyday life. While verbal and emotional displays have been investigated substantially, existing ECA implementations typically concentrate on some aspects of facial gestures but do not cover the complete set. We will first describe existing systems introducing one or more aspects of facial gesturing and then we will attempt to specify a complete set of facial gestures with its usage, causes and typical dynamics.

4.1.1. Related work

The Autonomous Speaker Agent in [9] performs dynamically correct gestures that correspond to the underlying text incorporating head movements (different kinds of nods, swing), eyes movements (movement in various directions and blinking) and eyebrows movements (up and down and v.v.).

In [10], Albrecht et al. introduce a method for automatic generation of the following non-verbal facial expressions from speech: head and eyebrow raising and lowering dependent on the pitch; gaze direction, movement of eyelids and eyebrows, and frowning during thinking and word search pauses; eye blinks and lip moistening as punctuators and manipulators; random eye movement during normal speech. The intensity of facial expressions is additionally controlled by the power spectrum of the speech signal, which corresponds to the loudness and intensity of the utterance.

Poggi and Pelachaud in [11] focused on the gaze behaviour, analyzing each single gaze in terms of a small set of physical parameters like eye direction, humidity, eyebrow movements, blinking, pupil dilatation etc. In addition, they tried to find which are the meanings that gaze can convey.

The Eyes Alive system [12] reproduces eye movements that are dynamically correct at the level of each movement, and that are also globally statistically correct in terms of the frequency of movements, intervals between them and their amplitudes. Although that the speaking and the listening mode is distinguished, movements are unrelated to the underlying speech contents, punctuation, accents etc.

Cassel et al. in [13] automatically generate and animate conversations between multiple human-like agents including intonation, facial expressions, lip motions, eye gaze, head motion and hand gestures.

The BEAT system [14] controls movements of hands, arms and the face and the intonation of the voice, relying on rules derived from the extensive research in the human conversational behaviour.

Graf et al. in [19] analyze head and facial movements that accompany speech and investigate how they relate to the text's prosodic structure. They concluded that despite large variations from person to person, patterns correlated with the prosodic structure of the text.

Pelechaud et al. in [5] reports results from a program that produces animation of facial expressions and head movements conveying information correlated with the intonation of the voice. Facial expressions are divided by its function (determinant) and the algorithm for each determinant is described, with special attention to lip synchronization and coarticulation problems.

4.2 Facial gestures for embodied conversational agents

Facial gestures are driven by [15]:

- **Conversational function of speech:** we unconsciously use facial gestures to regulate the flow of speech, accent words 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.

The same facial gesture can have different interpretations in different conditions. E.g. blinking can be the signal of a pause in the utterance, or can serve to wet the eyes. In the context of conversational function of speech, all facial gestures can be divided into four categories according to the function (usually called determinant) they have [5]:

- **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 category are eyebrow movements, rapid head movements, gaze directions and eye blinks.
- **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. 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 and have nothing to do with the linguistic utterance.
- **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. 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 (already introduced utterance information) is frequently synchronized by a gaze-away from a listener, and the beginning of rhemes (new utterance information) is frequently synchronized by a gaze-toward a listener.

In addition to characterizing facial gestures by their function, we can also characterize them by the amount of time that they last [16]. Some facial gestures are linked to personality and remain constant across a lifetime (e.g. frequent eye-blinking). Some are linked to emotional state, and may last as long as the emotion is felt (e.g. looking downward and frowning in case of sadness). And some are synchronized with the spoken utterance and last only a very short time (e.g. eyebrow rising on the accented word).

All previously mentioned functions are supported by a fairly broad repertoire of facial gestures. What follows is detailed description of all facial gestures. Anatomic parts of the face we take into account are: head, mouth (including lips, tongue and teeth), eyebrows (inner, medial, outer), eyelids (upper, lower), eyes, forehead, nose and hair. Within each part, different gestures are recognized. Each facial gesture class is described with its attributes, parameters (aspects, actions, and presence/absence) and the function it can serve. According to the function single facial gesture can have, it is additionally described by its causes and usage, level of synchronization and typical dynamics and amplitudes. The overview of the facial gestures is shown in the Table 1.

The table is organized as follows.

The first column, named *Facial Region*, contains the anatomic part of the face that takes part in facial gesture creation. The values of this column are: head, forehead, eyebrows, eyelids, eyes gaze (representing the eyes), pupil, hair, nose, lips, tongue and teeth. The regions that are used during the speech, but not in any other facial gesture are not included in the table (e.g. cheek or chin). Similarly, the regions that are not connected with any facial gesture (at least according to the available literature) are not included (e.g. iris or ears) in the table.

Facial gestures recognized within each part of the face are given in the second column, named *Gesture*. For some facial regions, for example head there are several known gestures such as different kinds of nods, while some regions are characterized with only one gesture (e.g. hair with the hairline motion).

The following column, named *Description*, gives brief description of each facial gesture.

The fourth column, named *Attributes*, contains characteristics important for single gesture or the group of gestures. Some examples are: direction, amplitude, duration etc. If the single gesture is not characterized with any attribute, the cell is left empty and shaded (this is the case when the gesture is described unambiguously without any additional parameters – e.g. nose wrinkling).

For the given attributes, the fifth column, named *Parameters* provides values that certain attribute can achieve (e.g. a direction of the eyebrows raise might be up, central or down).

A function that facial gesture can have is given in the following column, *Function*. According to the previous division, **C** stands for conversational signal, **P** for punctuator, **M** for manipulator and **R** for regulators. Some facial gestures can have more than one function (e.g. frowning serves as the conversational signal as well as the punctuator). For several facial gestures, the function is not assigned and that field is left empty and shaded, since those gestures are connected to the personality (e.g. teeth gnashing while listening or talking).

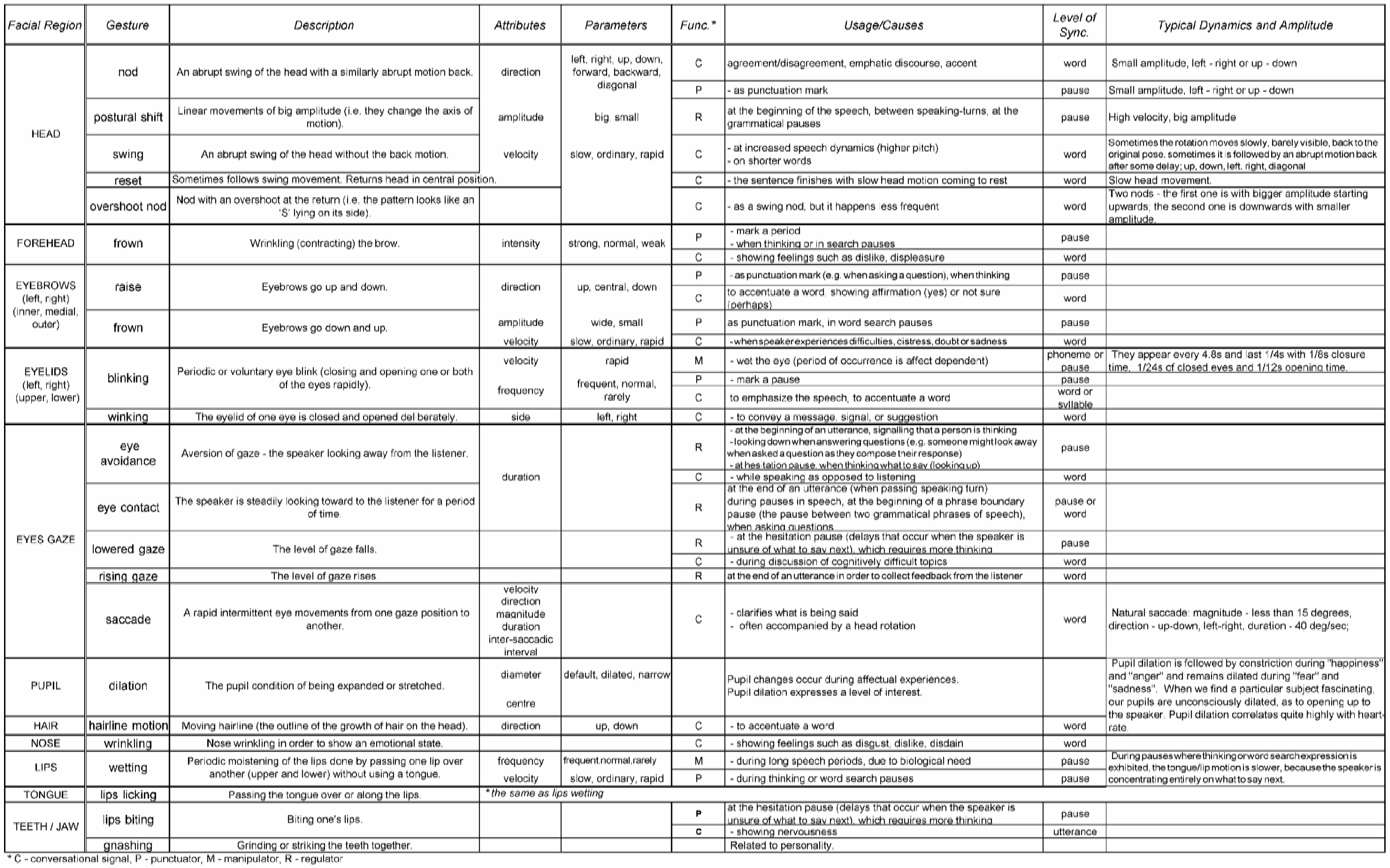
The seventh column, named *Usage/Causes*, provides information about each gesture, according to the function it has and concerning typical usage scenario. It gives information about situations in which certain facial gesture might appear and the meaning it has. Also, any available knowledge important for facial gesture understanding and using in ECA systems is given here.

Verbal and nonverbal signals are synchronized - synchrony occurs at all levels of speech: phonemic segment, word, syllable or long utterance, as well as at pauses. The eight column, named *Level of Synchronization*, specifies on which level is the gesture synchronized with the underlying speech.

The last column, named *Typical Dynamics and Amplitude*, gives available knowledge on typical dynamics and amplitudes of the gesture including known rules for their application in an ECA. For example head nods are characterized with the small amplitude and the direction up-down or left-right. Another example are saccadic eye movements which are often accompanied by a head rotation.

Cells left empty but not shaded in the table determine the fields which are not filled due to shortage or insufficiency of information in the studied literature. As the interest of this chapter lays in the non-verbal and non-emotional facial displays, the influence of the affects is not described in details. However, in some cases its presence is stated in order to make clear the usage and causes of the facial gestures.

Table 1: Facial gestures



Facial Region	*Gesture*	*Description*	*Attributes*	*Parameters*	*Func.**	*Usage/Causes*	*Level of Sync.*	*Typical Dynamics and Amplitude*
HEAD	nod	An abrupt swing of the head with a similarly abrupt motion back.	direction	left, right, up, down, forward, backward, diagonal	C	agreement/disagreement, emphatic discourse, accent	word	Small amplitude, left - right or up - down
					P	- as punctuation mark	pause	Small amplitude, left - right or up - down
	postural shift	Linear movements of big amplitude (i.e. they change the axis of motion).	amplitude	big, small	R	at the beginning of the speech, between speaking-turns, at the grammatical pauses	pause	High velocity, big amplitude
	swing	An abrupt swing of the head without the back motion.	velocity	slow, ordinary, rapid	C	- at increased speech dynamics (higher pitch) - on shorter words	word	Sometimes the rotation moves slowly, barely visible, back to the original pose, sometimes it is followed by an abrupt motion back after some delay; up, down, left, right, diagonal
	reset	Sometimes follows swing movement. Returns head in central position.			C	- the sentence finishes with slow head motion coming to rest	word	Slow head movement.
	overshoot nod	Nod with an overshoot at the return (i.e. the pattern looks like an 'S' lying on its side).			C	- as a swing nod, but it happens less frequent	word	Two nods - the first one is with bigger amplitude starting upwards; the second one is downwards with smaller amplitude.
FOREHEAD	frown	Wrinkling (contracting) the brow.	intensity	strong, normal, weak	P	- mark a period - when thinking or in search pauses	pause	
					C	- showing feelings such as dislike, displeasure	word	
EYEBROWS (left, right) (inner, medial, outer)	raise	Eyebrows go up and down.	direction	up, central, down	P	- as punctuation mark (e.g. when asking a question), when thinking	pause	
					C	to accentuate a word, showing affirmation (yes) or not sure (perhaps)	word	
	frown	Eyebrows go down and up.	amplitude	wide, small	P	as punctuation mark, in word search pauses	pause	
			velocity	slow, ordinary, rapid	C	- when speaker experiences difficulties, distress, doubt or sadness	word	
EYELIDS (left, right) (upper, lower)	blinking	Periodic or voluntary eye blink (closing and opening one or both of the eyes rapidly).	velocity	rapid	M	- wet the eye (period of occurrence is affect dependent)	phoneme or pause	They appear every 4.8s and last 1/4s with 1/8s closure time, 1/24s of closed eyes and 1/12s opening time.
			frequency	frequent, normal, rarely	P	- mark a pause	pause	
					C	to emphasize the speech, to accentuate a word	word or syllable	
	winking	The eyelid of one eye is closed and opened deliberately.	side	left, right	C	- to convey a message, signal, or suggestion	word	
EYES GAZE	eye avoidance	Aversion of gaze - the speaker looking away from the listener.	duration		R	- at the beginning of an utterance, signalling that a person is thinking - looking down when answering questions (e.g. someone might look away when asked a question as they compose their response) - at hesitation pause, when thinking what to say (looking up)	pause	
					C	- while speaking as opposed to listening	word	
	eye contact	The speaker is steadily looking toward to the listener for a period of time.			R	at the end of an utterance (when passing speaking turn) during pauses in speech, at the beginning of a phrase boundary pause (the pause between two grammatical phrases of speech), when asking questions	pause or word	
	lowered gaze	The level of gaze falls.			R	- at the hesitation pause (delays that occur when the speaker is unsure of what to say next), which requires more thinking	pause	
					C	- during discussion of cognitively difficult topics	word	
	rising gaze	The level of gaze rises.			R	at the end of an utterance in order to collect feedback from the listener	word	
	saccade	A rapid intermittent eye movements from one gaze position to another.	velocity direction magnitude duration inter-saccadic interval		C	- clarifies what is being said - often accompanied by a head rotation	word	Natural saccade: magnitude - less than 15 degrees, direction - up-down, left-right, duration - 40 deg/sec;
PUPIL	dilation	The pupil condition of being expanded or stretched.	diameter centre	default, dilated, narrow		Pupil changes occur during affectual experiences. Pupil dilation expresses a level of interest.		Pupil dilation is followed by constriction during "happiness" and "anger" and remains dilated during "fear" and "sadness". When we find a particular subject fascinating, our pupils are unconsciously dilated, as to opening up to the speaker. Pupil dilation correlates quite highly with heart-rate.
HAIR	hairline motion	Moving hairline (the outline of the growth of hair on the head).	direction	up, down	C	- to accentuate a word	word	
NOSE	wrinkling	Nose wrinkling in order to show an emotional state.			C	- showing feelings such as disgust, dislike, disdain	word	
LIPS	wetting	Periodic moistening of the lips done by passing one lip over another (upper and lower) without using a tongue.	frequency	frequent,normal,rarely	M	- during long speech periods, due to biological need	pause	During pauses where thinking or word search expression is exhibited, the tongue/lip motion is slower, because the speaker is concentrating entirely on what to say next.
			velocity	slow, ordinary, rapid	P	- during thinking or word search pauses	pause	
TONGUE	lips licking	Passing the tongue over or along the lips.	**the same as lips wetting*					
TEETH / JAW	lips biting	Biting one's lips.			P	at the hesitation pause (delays that occur when the speaker is unsure of what to say next), which requires more thinking	pause	
					C	- showing nervousness	utterance	
	gnashing	Grinding or striking the teeth together.				Related to personality.		

* C - conversational signal, P - punctuator, M - manipulator, R - regulator

4.2.1 Head

Head movements are frequently used facial gesture. Attributes and parameters that characterize head movements are: direction (left, right, up, down, forward, backward and diagonal), amplitude (wide, small) and velocity (slow, ordinary or rapid). Amplitude and velocity are in inverted proportion. Movement with big amplitude is rather slow. Different combinations of these parameters, define several head movement types [5][9]:

- *Nod.* An abrupt swing of the head with a similarly abrupt motion back.

Nod can be used as a conversational signal (e.g. nodding for agreement/disagreement or to accentuate what is being said), synchronized at the word level or as a punctuation mark. Typically, the nod is described as the rapid movement of the small amplitude with four directions: left and right, right and left, up and down and down and up.

- *Postural shifts.* Linear movements of wide amplitude often used as a regulator.

Postural shifts occur at the beginning of the speech, between speaking-turns and at grammatical pauses maintaining the flow of conversation. The synchronization with the verbal cues is generally achieved at the pauses of the speech.

- *Overshoot nod.* Nod with an overshoot at the return.

The pattern looks like an ‘S’ lying on its side. It is composed of two nods - the first one is with bigger amplitude starting upwards, while the second one is downwards with smaller amplitude.

- *Swing*. An 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. Possible directions are up, down, left, right and diagonal. It occurs at increased speech dynamics (when the pitch is also higher) and on shorter words.

- *Reset*. Sometimes follows swing movement; returns head in central position.

Reset is a slow head movement. It can be noticed at the end of the sentence – the sentence finishes with slow head motion coming to rest.

Another issue is the base head position or orientation which can be towards or away from a listener, up or down etc. The head direction may depend on affect, i.e. speaker-listener relationship or can be used to point at something. For example, if the utterance is a statement, head is positioned to look down as the speaker reaches the end of the sentence.

4.2.2 Mouth

Mouth, including lips, tongue, jaw and teeth take part in speech production. Their shape or position depends on the articulated phoneme and forms a visemes. The viseme is the visual representative of a phoneme [17]. According to the MPEG-4 standard [18] we can distinguish only 15 different visemes, including neutral face.

Openness of the lips is not only dependent on the articulated speech. Emotional state can also influence how wide the lips will be open. Intensity of the lip shape action decreases during fast speech-rate.

Besides in speech production, lips, tongue and teeth are used in generating facial gestures:

- *Lips wetting.* Periodic moistening of the lips done by passing one lip over another (upper and lower) without using a tongue.

It occurs during long speech periods due to biological need of the face (serving as manipulator), but also during thinking or word search pauses (serving as punctuator). Lips wetting is characterized by the frequency and velocity attributes. During pauses where thinking or word search expression is exhibited, the tongue/lip motion is slower, because the speaker is concentrating entirely on what to say next. Frequency of occurrence also depends on the personality and the outside conditions.

- *Lips licking.* Passing the tongue over or along the lips.

The function, usage and causes are the same as in lips wetting gesture (the same result is obtained in two different ways).

- *Lips biting.* Biting one's lips.

A teeth biting is often a sign of nervousness or insecurity. It might occur on the hesitation pauses, when the speaker is thinking what to say next (serving as punctuator).

- *Teeth gnashing.* Grinding or striking the teeth together.

This is a gesture related to the personality and not to the context of the speech.

4.2.3 Eyebrows

Eyebrow movements appear frequently as conversational signals or punctuators. When serving as punctuator, they are used to mark a period or thinking and word search pauses.

Eyebrow raise (eyebrows go up and down) is often used to accentuate a word or a sequence of words as well as to show affirmation (yes) or insecurity (perhaps) [11]. Eyebrow frown (eyebrows go down and up) might appear when speaker experiences difficulties, distress or doubt.

Besides direction, eyebrow movement is described with the amplitude and velocity, and closely related to pitch contour: eyebrows are raised for high pitch and lowered again with the pitch [10].

4.2.4 Eyelids

Eyelids determine the openness of eyes. Temporal closure of eyes happens quite frequently due to eye blinks. They are described as rapid closing and opening of one or both eyes which might happen in frequent, normal or rare periods. There are two types of eye blinks according to its function:

- *Periodic blinks.* They serve the physical need to keep the eyes wet. Periodic eye blinks are manipulators. On average, they appear every 4.8 sec, but their period of occurrence is dependent on affect [5]. Duration of the eye blink consists of three components: closure time, approx. $^1/_8$ sec, closed eyes, approx. $^1/_{24}$ sec. and opening time, approx. $^1/_{12}$ sec what makes duration last app. ¼ sec.
- *Voluntary blinks.* They appear in two roles, as punctuators (to mark a pause), synchronized with a pause or as conversational signals (to emphasize speech or to accentuate a word), synchronized with a word or syllable.

Eye openness varies also depending on the affect. For "surprise" and "fear" the eyes are wide open and they are partially closed during "sadness", "disgust" and "happiness" [5].

Another facial gesture performed by eyelids is eye winking (the eyelid of one eye is closed and opened deliberately). It is quite common in everyday communication and is used to convey a message, signal or suggestion. This conversational signal is synchronized with the word.

4.2.5 Eyes

Eyes play an essential role as a major channel of non-verbal communicative behaviour [12]. Different expressions can be reflected in eyes. Eyes can be in tears, red or dry, open or closed, showing clearly the state of our mind. For example, slightly narrow eyes during the talk when adding more precise information or wide open eyes when asking for speaking turn [11].

Pupil changes occur during emotional experiences. They may be dilated or narrow, centred or not. Pupil dilation is followed by constriction during "happiness" and "anger" and remains dilated during "fear" and "sadness". Pupil dilation also expresses a level of interest. When we find a particular subject fascinating, our pupils are unconsciously dilated, as to opening up to the speaker [5].

Eyes interact in the face-to-face communication through gaze direction or intensity and saccade. Saccade is a rapid intermittent eye movement from one gaze position to another executed voluntary by human. It is characterized with several attributes [12]:

- *Direction*,

- *Velocity*,
- *Magnitude or amplitude* (the angle through which the eyeball rotates as it changes fixation from one position to another),
- *Duration* (the amount of time that the movement takes to execute, typically determined using a velocity threshold) and
- *Inter-saccadic interval* (the amount of time which elapses between the termination of one saccade and the beginning of the next one).

Natural saccade movement (usually up-down, left-right) rarely have a magnitude greater than 15 degrees, while the duration and velocity are functions of its magnitude [12].

Eye gaze is used to signal the search for feedback during an interaction, look for information, express emotion, influence another person's behaviour or help regulate the flow of conversation [13]. However, some cultural differences are found in the amount of gaze allowed [5]. Gaze can be classified into four primary categories depending on its role in the conversation [13]:

- *Planning* - corresponds to first phase of a turn when the speaker organizes thoughts,
- *Comment* - accompanies and comments speech, by occurring in parallel with accent and emphasis,
- *Control* - controls the communication channel and functions as a synchronization signal and
- *Feedback* - used to collect and seek feedback.

Also, according to Poggi [11] two broad types of meanings of gaze can be distinguished: Information on the world and Information on the sender's mind. The first class includes places, objects and times to which we refer, while the second one contains sender's beliefs, goals and emotions (e.g. words giving information *of course, no* are shown with eyebrows central and down).

Aversion of gaze (the speaker looking away from listener) happens at the beginning of an utterance, signalling that a person is thinking, while speaking as opposed to listening or at hesitation pause, when thinking what to say. Eye contact (the speaker is steadily looking toward to the listener for a period of time) occurs at the end of an utterance (when passing speaking turn), during pauses in speech or at the beginning of a phrase boundary pause (the pause between two grammatical phrases of speech). Eye avoidance and eye contact follow the same rules as head movements for speaking turns. The level of gaze falls at the hesitation pause or during discussion of cognitively difficult topics, while it rises at the end of an utterance in order to collect feedback from the listener.

4.4.6 Forehead

Wrinkles often appear on the forehead. Vertical or horizontal, curved or oblique, they are positioned central, lateral or all along forehead, giving a special note to the one's personality. During the speech, depending on the context and the affect, wrinkles often deepen or change its direction and shape.

Another issue that is connected with the forehead is a frown. It is used during the search and thinking pauses and to mark a period (serving as a punctuator) or to show feelings such as dislike or displeasure. Intensity of the frown varies from strong to weak, depending on the context.

4.4.7 Other facial parts

Some people move their hairline (the outline of the growth of hair on the head) up and down to accentuate what is being said. This conversational signal is synchronized on the word level.

Nose wrinkling is used in order to show an emotional state and feelings such as disgust, dislike or disdain.

Color of the cheeks is connected with the emotional state, outside conditions or personality. Human face can loose color, blush or stay as it is.

There are some other facial gestures not directly affecting visual output, but as other non-verbal facial displays they complete the verbal output. Examples are: heavy gulp, rapid breathing, dry cough etc.

4.4.8 Combinations of facial gestures

This section attempts to give few representative examples how the group of facial gestures might be used together in order to convey different meaning. With this additional information, it is not meant to give complete overview, but only to illustrate an idea that facial gestures often come in groups. Due to complex problem involving various influences and lacking enough knowledge in particular fields, creating complete overview of facial gesture combinations would require further research. Since it is not in the scope of this work, in Table 2 are given only some frequently used combinations of facial gestures.

Table 2: Frequently used combinations of facial gestures

Combination of facial gestures	*Meaning / Function / Context*
– head aside – eyebrows up	Performative eyes [11] (I suggest)
– looking up – eyebrows raise – raise head – frowning	Thinking [12]
– avoidance of gaze – head of the speaker turns away from listener	Hesitation pause (speaker is concentrating on what is going to say)
– look down at the end of the sentence – head down	Statement [10]
– eyebrows raise – head raising at the end – gaze toward to listener – high pitch at the end of the utterance	Question [10]
– head nodding – eyebrows raise	Affirmation (yes) [11]

4.3 Case Study – Example of Practical System Implementation

In this section we are going to present an example of practical ECA system implementation as a case study of state of the art work in ECA domain. As our Autonomous Speaker Agent (ASA) system only implements a subset of a full set of ECA functionality, we will first give the scope of ASA system in the following subsection. Furthermore, after defining and elaborating system scope, we introduce the system overview and implementation. In this subsection technologies and architecture that we had used are explained in detail. In the end, in the subsection results, we will present empirical results of our ASA system along with the real users' impression of our system implementation.

4.3.1 System Scope

In this subsection we are defining the scope of our ASA implementation. ASA is a presentation system that pertains in the domain of ECA. Since its function is only presentation of content, it is developed using only subset of the ECA state of the art theory. First simplification is that we implemented in our ASA only a subset of facial gestures. That means that our ASA is presented with a model of a human head (Figure 1).

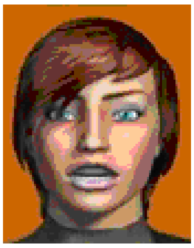


Figure 1 Model of a Autonomous Speaker Agent.

We are focusing our ASA implementation only on a subset of the interactional functions of speech. Our ASA supports conversational signals, punctuators and manipulators. The final simplification is a set of facial gestures that are supported in our ASA system. Table 3 shows implemented facial gestures. In the next subsection we are introducing ASA system overview along with its implementation.

Facial display	Type	Direction
Head	Nod	Up Down Left Right
	Overshoot nod	Up then down
	Swing	Up Down Left Right Diagonal
	Reset	To central
Eyes	Simple gaze	
	Blink	
Eyebrows	Raise	ˆˆ

Table 3: Implemented facial gestures

4.3.2 System overview

In this subsection we will first give the brief ASA system overview with basic explanation of data flow between system modules. After that, every system module is going to be described in detail along with technology and standards used in its implementation.

ASA system consists of several modules. System input is plain English text and output is real-time animation with appropriate facial gestures and audio. Figure 2 imparts ASA system modules (green colour) along with data (blue colour) flow through them. Plain English text is first processed using the Lexical Analysis module that generates as output English text in XML format with appropriate lexical tags (describing new/old words and punctuation marks). The core module of ASA is Facial Gesture module that inserts appropriate gestures into plain English text in the form of

Microsoft Speech API[1] (SAPI 5.0) XML tags. This text with appropriate gesture tags represents input for TTS/MPEG-4 Player module that is based on SAPI 5.0 Text To Speech (TTS) engine and Visage SDK API[2]. While SAPI 5.0 TTS engine generates an audio stream, the SAPI 5.0 notification mechanism is used to catch the timing of phonemes and XML tags containing gesture information. Gesture XML tags hold three gesture parameters: gesture type, amplitude and duration in word units. In order to produce real-time animation, we used the assumption that average word duration is one second.

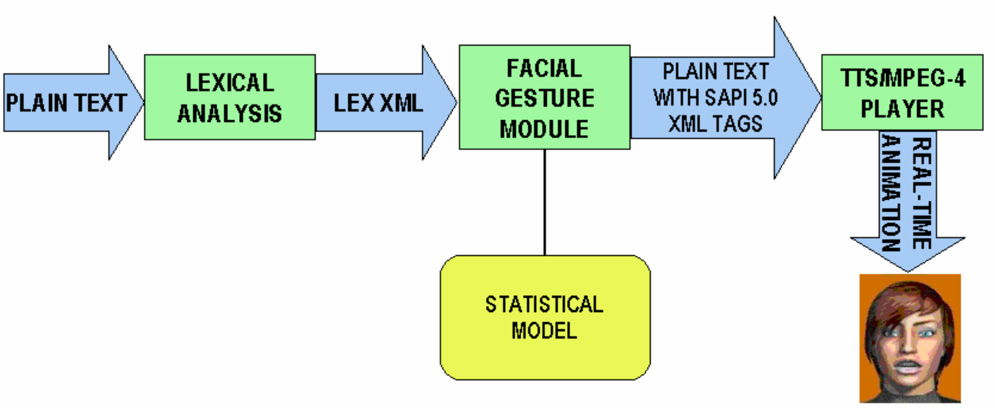


Figure 2: ASA system modules

Lexical 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. The main goal of this module is to determine if word (or group of words) is new in the utterance context (new), if extends some previously mentioned word or group of words (old) and to determine the punctuation marks. Punctuation mark determination is straightforward, but for the first two types there is need to know morphological, syntactic and part-of-speech word information. For that purpose we used WordNet 2.0[3] system. This is lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. First we need to determine word type (noun, verb, adverb or adjective) by querying WordNet 2.0 system, using English grammatical rules and parsing input text multiple times. Each noun, verb, adverb or adjective is tagged as new if itself or any of its synonyms (queried from WordNet 2.0 system) has not been mentioned in previous text. Other word classes are not considered as new. WordNet 2.0 does not handle pronouns, so we developed special algorithm for processing them. Every pronoun, that is not preceded by a noun in the sentence, and is part of the following 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") or any pronoun that is part of the following set: ("I", "you", "he", "she", "it", "we", "they"), gets the **new** tag assigned. All other pronouns, which do not fulfill above stated requirements, are tagged with **old**. Also, each pronoun, substituting a noun that appears after it, or a noun that does not appear in the text at all, needs to be tagged as **new.** Table 4 represents one example of input and output data of Lexical Analysis Module.

[1] Microsoft speech technologies http://www.microsoft.com/speech/

[2] Visage Technologies AB http://www.visagetechnologies.com/

[3] http://www.cogsci.princeton.edu/~wn/

Module input	Module output

However, figures presented by Business Unit Systems prompted more positive reactions.	<?xml version="1.0" encoding="UTF-8"?>
	<UTTERANCE>
	  <CLAUSE>
	   <WORD Text="However" New="Yes"/>
	   <WORD Text=","/>
	   <WORD Text="figures" New="Yes"/>
	   <WORD Text="presented" New="Yes"/>
	   <WORD Text="by"/>
	   <WORD Text="Business"/>
	   <WORD Text="Unit" New="Yes"/>
	   <WORD Text="Systems" New="Yes"/>
	   <WORD Text="prompted" New="Yes"/>
	   <WORD Text="more" New="Yes"/>
	   <WORD Text="positive" New="Yes"/>
	   <WORD Text="reactions" New="Yes"/>
	   <WORD Text="."/>
	  </CLAUSE>
	</UTTERANCE>

Table 4: Input and output data of Lexical Analysis Module

Next module in data flow through the ASA system is Facial Gesture module. Input to this module is XML tagged English text with appropriate lexical tags. Statistical Model of facial gestures is basic building block of this module. The decision tree algorithm uses data from Statistical Model and we will explain it in detail later. First, we will present Statistical Model of facial gestures along with methods, tools and datasets used to build it. Figure 3 represents the training process of ASA system. As an input for this process we used as a training data set a number of Ericsson's “5 minutes” video clips. These 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 and they are presented by the professional newscasters. As we have already stated, we identified three facial gesture parameters: gesture type (Table 3), amplitude and duration. So, Statistical Model of speaker gestural behaviour consists of components for those three facial gesture parameters for every lexical context (new, old and no lexical information). Next, we will explain the process of generating text annotated with gesture and lexical data.

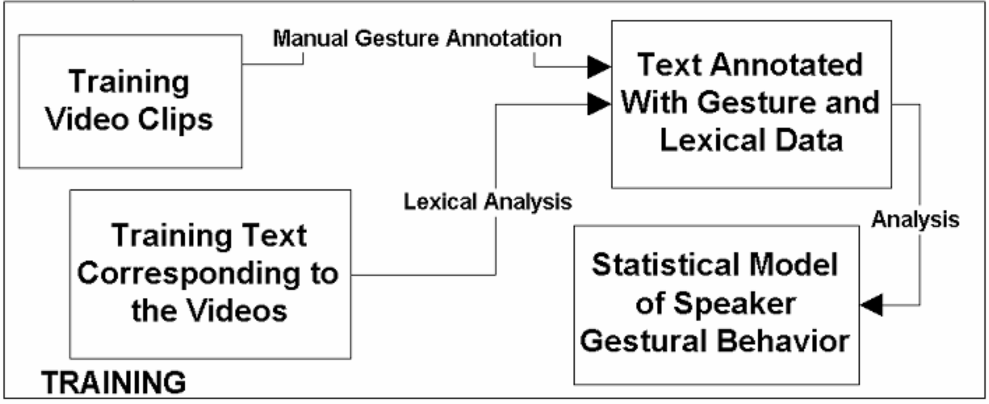

Training Video Clips
Manual Gesture Annotation
Text Annotated With Gesture and Lexical Data
Lexical Analysis
Analysis
Training Text Corresponding to the Videos
Statistical Model of Speaker Gestural Behavior
TRAINING

Figure 3: ASA training process of ASA system.

Table 5 presents an extract of a text annotated with gesture and lexical data.

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

Table 5: Text annotated with gesture and lexical data

<table>
<tr><th>Facial gesture</th><th>Type of motion</th><th>Description</th></tr>
<tr><td rowspan="11">Head movement</td><td>| up</td><td>vertical up</td></tr>
<tr><td>| up to n</td><td>vertical up to neutral (center) position</td></tr>
<tr><td>| down</td><td>vertical down</td></tr>
<tr><td>| down to n</td><td>vertical down to neutral (center) position</td></tr>
<tr><td>-- to left</td><td>horizontal left</td></tr>
<tr><td>-- to right</td><td>horizontal right</td></tr>
<tr><td>-- to n</td><td>horizontal to neutral (center) position</td></tr>
<tr><td>/ up</td><td>diagonal up from left to right[4]</td></tr>
<tr><td>/ down</td><td>diagonal down from right to left</td></tr>
<tr><td>¥ up</td><td>diagonal up from right to left</td></tr>
<tr><td>¥ down</td><td>diagonal down from left to right</td></tr>
<tr><td rowspan="2">Eyebrows movement</td><td>raised s</td><td>eyebrows going up to maximal amplitude</td></tr>
<tr><td>raised e</td><td>eyebrows going down to neutral position</td></tr>
</table>

Table 6: Basic facial motions triggered by words.

The **word** row contains an analyzed news extract separated word-by-word. The **eyes**, **head** and **eyebrows** rows hold data about facial motion that occurred on the corresponding word: type of motion, direction, amplitude and determinant, according to the notation summarized in Table 6.

These basic facial motions were mapped to the facial gestures described in Table 3. We replayed newscaster footage to determine the facial gestures type, direction and duration parameters. For example, the last row in Table 6 contains the head movement facial gesture parameters.

The **pitch** row indicates which words were emphasized by voice intonation. The **lexical** row holds information about a word's newness in the context of a text (output of Lexical Analysis module).

Data analysis was the most important part for the final proposal of our facial animation model. Because of that, it was very important to use good and valid analysis techniques. Our analysis was partly based on the work described in [12]. In that work, the authors proposed the eyes animation model. The animation was driven by the conversation between two people. The data was gathered using a sophisticated eyes movement capturing device. After capturing the eyes movement during the 9-minute conversation between two individuals, the authors analyzed the gathered data using well known statistical methods and functions. Based on the given results, the statistical eyes movement animation model was introduced and incorporated into the existing facial animation

[4] From the listener point of view.

model.

The first significant difference from the [12] work is that we were not analyzing a conversation between two individuals. Because our facial animation model was for the ASA, it means that we analyzed the recorded video of real speakers who were not involved in the conversation. Also, we didn't use any sophisticated equipment like in [12]. Our tools were Movie Maker and observing the recorded video. In Movie Maker we could analyze the facial and head movements watching the recorded video frame by frame. It was a tedious task, but the quality of analysis was not put in question.

It is important to state that in our model, the basic unit, which triggers head or facial movement, is word. Words can be divided into the phonemes or syllables, but for the simplicity of data analysis, we chose words as basic units for facial animation model.

In our data analysis (Table 5) we first populated the word row with words along with the punctuation marks from the news transcripts.

The first row we populated with observed data was the pitch row. Using headphones and playing the recorded video clips over and over again, we marked the words that the speaker had emphasized with her voice. Our primary concern was not the type of accent [20], only if a particular word had been accented or not. In our table, every word that had been accented was marked with an '+' in the pitch row. Every pitch accent had a duration of one word.

After analyzing the pitch information, we continued with eyes blinks. Eyes blinks started and ended on the same word or punctuation mark. Extracting data about the eyes blinks was the easiest part of the data analysis. We played the video clips in Movie Maker and marked the columns in eyes row with word blink if the speaker had blinked on the corresponding word or punctuation mark. It is important to state that eyes blinks were triggered, aside from words, also by punctuation marks.

After that, the head movement analysis followed and it was the most tedious task of the whole data analysis process because a head has the biggest freedom of movement. We can move a head vertically up and down, horizontally left and right, diagonally and even rotate it. The additional problem was that a head did not always start (or finish) it's basic movement element in the neutral position.

According to work [19], the following are the head movement patterns:

- Nod;
- Nod with overshoot;
- Swing (rapid);

Those three meta head movements have basic elements which are triggered by the uttered word or punctuation mark. Since words are the basic triggering units of our facial animation model, we had to propose the symbols for the basic elements of the head movement patterns. They are listed in Table 6.

Our task was to map basic elements into basic head movement patterns. First we observed the video clips (average duration was about 15 seconds) using Movie Maker playing functions, and, using the frame by frame analysis, we marked words with corresponding basic elements of head movements. When we finished with basic elements, we mapped them into the corresponding basic patterns according to Table 3.

The basic elements of head movements are triggered by the uttered words. Sometimes one word triggers two basic elements. In such cases we had a nod triggered by a word. That nod was very fast with a small amplitude. Nods usually lasted through two words. Nods with overshoot had the longest duration. They were not so frequent and they lasted through four or five words.

Finally, we analyzed the eyebrows movement. The basic eyebrows patterns are raises and frowns (Table 3). Eyebrows movements also had basic elements which are described in Table 6. Eyebrows movements also lasted through one or more words.

The determinant value for a particular facial motion is determined as follows: if the facial motion occurred on a punctuator mark, then the determinant for that motion was punctuator (p); if the facial motion accompanied a word that is new in the context of the uttered text, then the determinant was a conversational signal (cs); otherwise, the determinant of the facial motion was a manipulator (m).

During the data analysis we also extracted amplitudes of the head and eyebrows movements. Our goal was to propose a statistical model for the amplitudes as important components of the facial animation model. Also, amplitudes of the head and eyebrows movements were in close relation with the pitch amplitudes. The recorded video data that we had analyzed and the method used (that will be also described here), gave us a good starting point for the statistical model of amplitudes.

The raw data tables were populated during manual analysis and measurement. All amplitude values were normalized to Mouth-Nose Separation unit (MNS0) for the particular speaker. MNS0 is Facial Animation Parameter Unit (FAPU) in MPEG-4 Face and Body Animation (FBA) standard [22]. Using MNS0 FAPU our model could be applied to every 3D model of speaker. Algorithms used for extracting amplitude values from observed footages are described in details in [20].

In order to explain how Facial Gesture module works we will follow the decision tree (Figure 4).

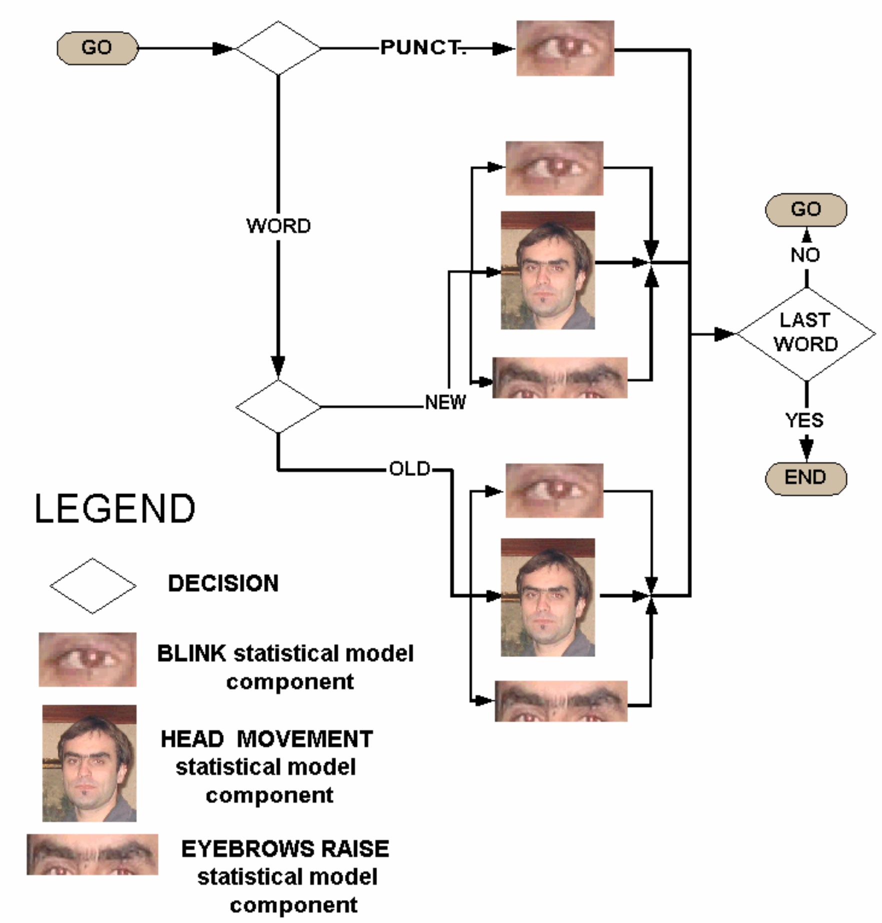

GO
PUNCT.
WORD
NEW
OLD
GO
NO
LAST WORD
YES
END
LEGEND
DECISION
BLINK statistical model component
HEAD MOVEMENT statistical model component
EYEBROWS RAISE statistical model component

Figure 4: The decision tree with components of statistical data.

The first branch point classifies the current input text context as either a word or a punctuation mark. Our data analysis showed that only eyes blink facial gesture had occurred on punctuation marks. Therefore, only the blink component of the statistical model is implemented in this context. A uniformly distributed random number between 0 and 100 is generated and 2 non-uniform intervals are assigned. That is, a random number between 0 and 33.68 is assigned to the eyes blink, and a number between 33.68 to 100.00 to the no blink motion. Thus, there is a 33.68% chance that an eyes blink will occur, and 66.32% chance that no eyes blink motion will be generated. Words could be new or old in the context of the uttered text – this is the second branch point. All facial gestures occurred in both cases but with different probabilities. Because of that we have different components for facial gestures parameters in both cases. In the case of a new word, we first compute an eyes blink motion. A uniformly distributed random number between 0 and 100 is generated and 2 non-uniform intervals are assigned. That is, a random number between 0 and 15.86 is assigned to the eyes blink, and a number between 15.86 and 100.00 to the no blink motion. Thus, there is 15.86% chance that an eyes blink will occur, and a 84.14% chance that no eyes blink motion will be generated. After eyes blink, we compute the eyebrows motion parameters. The two parameters are motion amplitude and duration. Again, a uniformly distributed random number between 0 and 100 is generated and 2 non-uniform intervals are assigned. That is, a random number between 0 and 18.03 is assigned to the eyebrows raise, and a number between 18.03 to 100.00 to the no eyebrows raise. Thus, there is a 18.03% chance that an eyebrows raise will occur, and a 81.97% chance that no eyebrows motion will be generated. After that we compute the eyebrows raise duration. A uniformly distributed random number between 0 and 100 is generated and 10 non-uniform intervals are assigned. That is, a random number between 0 and 77.31 is assigned to the duration of 1 word, a number between 77.31 to 86.57 to the duration of 2 words, and so on. The eyebrows raise amplitude is dependent on the duration parameter. The duration of 1 word is qualified as short duration, while all other durations are qualified as long durations. A random number between 0 and 100 is generated. The number corresponds to the x-axis (percentage of frequency) in Figure 5.

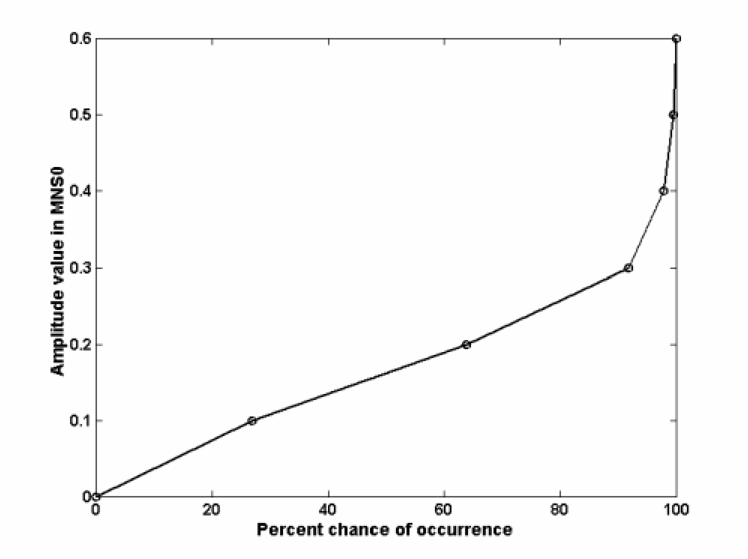


Figure 5: The linear approximation of the cumulative histogram for the amplitude of a short eyebrows raise.

Our decision tree algorithm is triggered for every word. However, an eyebrows raise motion could last through more than one word and because of that an eyebrows raise is not calculated for words that are already covered by previous calculation for an eyebrows raise. Finally, we calculate the head motion. First, the head motion type is determined. A uniformly distributed random number between 0 and 100 is generated and 5 non-uniform intervals are assigned (for every head motion subtype). That is, a random number between 0 and 1.34 is assigned to the overshoot nod, a number between 1.34 and 23.71 to the nod, a number between 23.71 and 46.25 to the rapid (swing) movement, a number between 46.25 and 51.43 to the reset movement, and a number between 51.43 and 100.00 to the no head movement. Thus, there is a 1.34% chance that an overshoot nod will occur, a 22.37% chance that a nod will occur, a 22.54% chance that a rapid (swing) movement will occur, a 5.18%

chance that a reset movement will occur, and a 48.57% chance that no head motion will be generated.

In case of an overshoot nod, the following calculations are performed. An overshoot nod has two amplitude parameters (for first nod up and second nod down) and one duration parameter. We begin by computing an overshoot nod duration. A uniformly distributed random number between 0 and 100 is generated and 7 non-uniform intervals are assigned. That is, a random number between 0 and 6.25 is assigned to the duration of 1 word, a number between 6.25 and 37.5 to the duration of 2 words, and so on. A random number between 0 and 100 is generated for the first and second amplitude. The random number corresponds to the x-axis (percentage of frequency) similar to Figure 5. Our decision tree algorithm is triggered for every word. However, an overshoot nod could last through more than one word and because of that, a head motion is not calculated for words that are already covered by the previous calculation for a head motion.

In case of a nod, the following calculations are performed. A nod has a type parameter, an amplitude parameter and one duration parameter. We begin by computing the nod type. A uniformly distributed random number between 0 and 100 is generated and 4 non-uniform intervals are assigned. That is, a random number between 0 and 60.15 is assigned to the nod up, a number between 60.15 and 89.3 to the nod down, a number between 89.3 and 97.42 to the nod right, and a number between 97.42 and 100.00 to the nod left. Thus, there is 60.15% chance that a nod up will occur, a 29.15% chance that a nod down will occur, a 8.12% chance that a nod right will occur, and a 2.58% chance that a nod left will be generated.

After the type parameter, we compute nod duration. A uniformly distributed random number between 0 and 100 is generated and 5 non-uniform intervals are assigned. That is, a random number between 0 and 62.31 is assigned to the duration of 1 word, a number between 62.31 and 85.45 to the duration of 2 words, and so on. Nod amplitude is dependent on the duration parameter. The duration of 1 word is qualified as short duration, and all other durations are qualified as long duration. A random number between 0 and 100 is generated for short and long amplitudes. The random number corresponds to the x-axis (percentage of frequency) similar as in Figure 5. Our decision tree algorithm is triggered for every word. However, a nod motion could last through more than one word and because of that a nod is not calculated for words that are already covered by previous calculation for a nod motion.

In case of a rapid (swing) movement, the following calculations are performed. A rapid movement has a type parameter and an amplitude parameter. First we compute the rapid movement type. A uniformly distributed random number between 0 and 100 is generated and 5 non-uniform intervals are assigned. That is, a random number between 0 and 61.11 is assigned to the rapid down movement, a number between 61.11 and 76.3 to the rapid up motion, a number between 76.3 and 84.82 to the rapid left motion, a number between 84.82 and 95.56 to the rapid right motion, and a number between 95.56 and 100.00 to the rapid diagonal movement. Thus, there is a 61.11% chance that a rapid down motion will occur, a 15.19% chance that a rapid up motion will occur, a 8.52% chance that a rapid left motion will occur, a 10.74% chance that a rapid right motion will occur, and a 4.44% chance that a rapid diagonal motion will be generated.

A random number between 0 and 100 is generated for the amplitude calculation. The random number corresponds to the x-axis (percentage of frequency) similar as in Figure 5. In case of an old word context, the generation process for every facial display is the same as in case of a new word context. The only differences are statistical data values.

Observing at Figure 4, it is obvious that a word could be accompanied with all three facial gesture kinds. The output from the Facial Gesture module is plain English text accompanied with SAPI5 bookmarks for facial gestures.

Bookmark code	Facial gesture
MARK=1	conversational signal blink
MARK=2	punctuator blink
MARK= 100000	eyebrows raise

MARK= 200000	nod ^
MARK= 300000	nod V
MARK= 400000	nod <
MARK= 500000	nod >
MARK=9	rapid reset
MARK= 600000	rapid d
MARK= 700000	rapid u
MARK= 800000	rapid L
MARK= 900000	rapid R
MARK= 1000000	rapid diagonal

Table 7: Our SAPI5 bookmark codes of facial gestures.

Every facial gesture has a corresponding bookmark value: Table 7 shows the boundary values for each bookmark. The head and eyebrows movement bookmark values not only define the type of facial gesture, but also contain the amplitude data and duration of the facial movement. For example, bookmark value 805120 (Bmk_value) defines the rapid head movement to the left (symbol L) of amplitude (A) 1.2 MNS0 and duration (D) of 5 words. The function for amplitudes of facial gestures L is:

D=(Bmk_value –Bmk_code)/1000. (1)

A=((Bmk_value –Bmk_code) - (D x 1000))/100. (2)

The interval for bookmark values for L is [800000, 900000> because the statistical data showed that the maximal amplitude value for facial gesture L was 2.2 MNS0, and duration was 1 word.

Head nods and eyebrows raises could last through two or more words. The statistics have shown that the maximum duration of a nod is five words, that an eyebrows raise can last through eleven words and the maximal duration for a nod with overshoot is eight words. We code a nod with overshoot as two nods: a nod up immediately followed by a nod down. Every nod has its own amplitude distribution.

TTS/MPEG-4 Playing Module plays in real-time, using the bookmark information, appropriate viseme and gestures model animation. The synchronization between the animation subsystem (MPEG-4 Playing) and the speech subsystem (Microsoft's TTS engine) can be realized in two ways: with time-based scheduling and event-based scheduling. Which synchronization method will be used depends on the underling TTS engine implementation. In time-based scheduling, a speech is generated before nonverbal behaviors. Event-based scheduling means that speech and nonverbal behaviors are generated at the same time. In our system we are using the event-based scheduling method. We have implemented simple animation models for eyes blink, simple gaze following and head and eyebrows movement. Our system implementation is open, so every user is able to easily implement its own animation models. The animation model for head movement and eyebrows movement facial gestures is based on the trigonometry sine function. That means that our ASA nods his head and raises eyebrows following the sine function trajectory. In gaze following animation model the eyes of our ASA are moving in opposite directions of a head movement if a head movement amplitude is smaller than the defined threshold. This gives the impression of eye contact with ASA.

4.3.3 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 Type 1 method, head and eye movements were produced playing animation sequence that was recorded by tracking movements of a real professional speaker. In Type 2 method, we produced a facial animation using the system described in this paper. Type 3 method animated only character's lips. We conducted a subjective test to compare those three methods of facial animation. The three characters were presented in random order to 29 subjects. All three characters presented the same text. The presentation was conducted in the Ericsson Nikola Tesla and all subjects were computer specialists. However, most of the subjects were not familiar with virtual characters, and none of the subjects were authors of the study. 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. Figure 6 summarizes the average score and standard deviation (marked with a black color) for the first four questions. From the figure, we can remark that the character animated using type 2 method was graded with the highest average grade for all questions except for the Q2. The reason for that is because type 3 method only animates character lips while its head remains still. This gave the audience the impression of engagement in the presentation. A Kruskal-Wallis ANOVA indicated that the three characters had significantly different scores (p = 0.0000).

According to general remarks in Q5, the subjects tended to believe the following:

1. Type 1 looked boring and uninteresting, it seemed to have cold personality. Also, implemented facial gestures were not related to the spoken text.

2. Type 2 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.

3. Type 3 looked irritating, stern and stony. However, it appeared to be concentrated and its lips animation was the best.

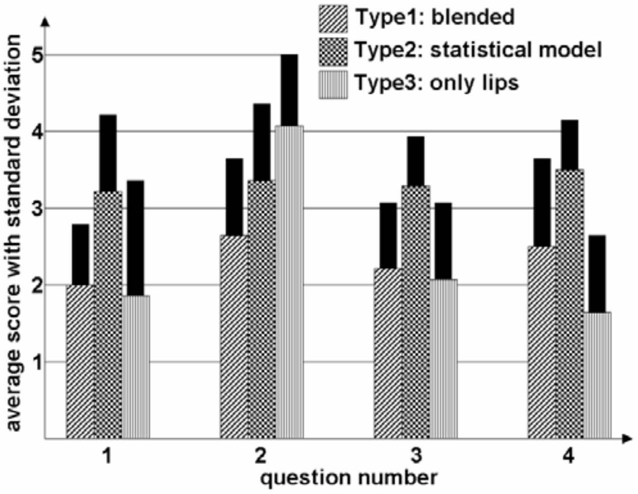


Figure 6: Results of subjective evaluations. Average score and standard deviation.

4.4 Conclusion

In this article we have tried to give a complete survey of facial gestures that can be useful as guideline for their implementation in an ECA. Specifically, we have concentrated on providing a systematically organized repertoire of usual facial gestures including the information on its typical usage, causes, any available knowledge on typical dynamics and amplitude of the gesture. We have studied face motions through its anatomic parts (head, mouth, eyebrows, eyelids, eyes, forehead, hair and nose) and within each part we have recognized different gestures. For each facial gesture class available knowledge is presented, including function it can serve and attributes with its parameters. According to the function single facial gesture can have, it is additionally described by its causes and usage, level of synchronization and typical dynamics and amplitudes. Finally we have provided an example of a practical system implementation as a case study. The system uses lexical analysis and statistical models of facial gestures in order to generate the facial gestures related to the spoken text. Gestures taken into consideration are head, eye and eyebrow movements and blinks.

Although we attempted to cover all facial gestures with complete data that would make possible immediate implementation of the facial gesture in an ECA system, there are still some insufficiently described gesture classes either because of the lack of the knowledge or due to complexity of problem. In order to finish this survey and to fulfil existing data, additional information needs to be added. However, since research on facial gestures belongs to multidisciplinary fields which are further investigated, information collecting for such survey could never end.

References

[1] Mehrabian, A. Silent messages, Wadsworth, Belmont, California, (1971).

[2] Knapp, M.L.: Nonverbal Communication in Human Interaction. 2nd edn. Holt, Rinehart and Winston Inc., New York, (1978).

[3] Ruesch, J. Nonverbal language and therapy. In Alfred G. Smith (Ed.), Communication and culture: Readings in the codes of human interaction (pp. 209-213). New York: Holt, Rinehart and Winston, 1966.

[4] Cassell, J., Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents, In J. Cassell, et al. (eds.), Embodied Conversational agents. Cambridge, MA: MIT Press. 1-27, 2000.

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

[6] http://www.paulekman.com/

[7] Nonverbal dictionary of gestures, signs and body language cues, http://members.aol.com/nonverbal2/diction1.htm

[8] Cassell, J., Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems, In S. Luperfoy (ed.), Spoken Dialogue Systems, MIT Press, 1989.

[9] Smid, K., Pandzic, I.S., and Radman, V., Autonomous Speaker Agent, Computer Animation and Social Agents Conference CASA 2004, Geneva, Switzerland.

[10] Irene Albrecht, Jorg Haber, and HansPeter Seidel. Automatic Generation of Non-Verbal Facial Expressions from Speech. In Proc. Computer Graphics International 2002 (CGI 2002), pages 283--293, July 2002.

[11] I. Poggi and C. Pelachaud. Signals and meanings of gaze in Animated Faces. In P. McKevitt, S. O' Nuallàin, Conn Mulvihill, eds.: Language,Vision, and Music,John Benjamins, Amsterdam (2002), 133-144.

[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] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douvillle, S. Prevost and M. Stone. *Animated Conversation: Rule-based Generation of Facial Expressions*, Jesture & Spoken Intonation for Multiple Conversational Agents. In Proceedings of SIGGAPH '94, 1994.

[14] 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.

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

[16] Cassell, J. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell, S. Prevost, J. Sullivan, and E. Churchill, editors, Embodied Conversational Agents. The MIT Press, 2000.

[17] Benoit, C., Lallouache, M.T., Abry, C., A set of French visemes for visual speech synthesis. in Talking Machines : Theories, Models and Designs, G. Bailly and C. Benoit, Eds., Elsevier Science Publishers, 1992, pp. 485-504.

[18] I. S. Pandžić, R. Forchheimer, Editors, “MPEG-4 Facial Animation - The Standard, Implementation and Applications “, John Wiley & Sons Ltd, ISBN 0-470-84465-5, 2002.

[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] 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.

[21] Smid, K., 2004. Simulation of a Television Speaker with Natural Facial Gestures. Master Thesis no. 03-Ac-10/2000-z on Faculty of Electrical Engineering and Computing, University of Zagreb.

[22] ISO/IEC IS 14496-2 Visual, 1999.