“May I talk to you? `:-)`*”* — Facial Animation from Text

Irene Albrecht[*], Jörg Haber[*], Kolja Kähler[*], Marc Schröder[†], and Hans-Peter Seidel[*]

[*] Max-Planck-Institut für Informatik, Saarbrücken, Germany
[†] Deutsches Forschungszentrum für künstliche Intelligenz, Saarbrücken, Germany
E-mail: {albrecht,haberj,kaehler,hpseidel}@mpi-sb.mpg.de, schroed@dfki.de

Abstract

We introduce a facial animation system that produces real-time animation sequences including speech synchronization and non-verbal speech-related facial expressions from plain text input. A state-of-the-art text-to-speech synthesis component performs linguistic analysis of the text input and creates a speech signal from phonetic and intonation information. The phonetic transcription is additionally used to drive a speech synchronization method for the physically based facial animation. Further high-level information from the linguistic analysis such as different types of accents or pauses as well as the type of the sentence is used to generate non-verbal speech-related facial expressions such as movement of head, eyes, and eyebrows or voluntary eye blinks. Moreover, emoticons are translated into XML markup that triggers emotional facial expressions.

Keywords: *facial animation, speech synthesis, speech synchronization, non-verbal facial expressions*

1 Introduction

In recent years, facial animation systems have reached a degree of realism that allows creation of photo-realistic full-feature movies. However, the process of generating the animations is still enormously time-consuming, especially when speech-synchronized facial animation is needed. A fully automatic method to generate facial animation from simple input text is thus a much desired goal. A major problem on the way to achieve this goal is that huge background knowledge about Life, the Universe and Everything is necessary to correctly interpret the meaning of written sentences, and to transfer this meaning to appropriate facial expressions. A question might be a rhetorical one, a remark can have an ironic touch. These subtleties should be reflected in the face of the speaker. Although promising attempts to solve this problem are currently being made, a fully automatic solution has not been found yet.

Fortunately, it is possible to synthesize audible speech from text input, taking into account intonation and syntactic importance of individual words and phrases. The linguistic analysis of text input, which is necessary to perform the task of speech synthesis, can be used to additionally drive a facial animation system that generates speech synchronized mouth movements as well as non-verbal speech-related facial expressions. Including additional emoticons in the text input allows for the display of emotions, which can be useful in applications such as the one depicted in Figure 2.
In this paper, we propose the following contributions:

- a fully automatic text-to-speech system that creates additional output suitable for generating facial expressions;
- a facial animation system that performs physically based animation including speech synchronization and non-verbal speech-related facial expressions.

2 Previous Work

2.1 Facial Animation

Over the past thirty years, a variety of different approaches to facial animation has been introduced. Most of these methods can be classified into expression-blending techniques, e.g. on face models generated from photographs [40, 9], direct parametric animation [35, 48, 34], or physically based methods [47, 30, 31]. A comprehensive overview of the field can be found in the textbook by Parke and Waters [36]. Speech synchronization is an important application area for facial animation, and several such systems have been proposed in the literature. A recent survey by Bailly [6] provides a very good overview. Lewis and Parke [32] have demonstrated automated speech synchronization for recorded speech using linear prediction, while Hill *et al.* [23] produce animations using speech synthesized by rules. Both approaches restrict animation exclusively to the movement of jaw and lips. Cohen and Massaro additionally included movement of the tongue and also introduced a technique for modeling coarticulation [16]. The integration of synchronized speech animation with facial expressions has been carried out by Pearce *et al.* [37] and by Ip and Chan [25], both using a script-based approach: the expression that should be displayed during speech is specified by the user in a domain-specific script language. Kalra *et al.* [28] describe a layered, script-based approach to specify facial animations. Similar to theses approaches, the RUTH system by DeCarlo *et al.* [18] takes as input text annotated with facial expressions. Audible speech is generated by a text-to-speech system, which also returns a timed phoneme string. Based on this phoneme representation, speech-synchronized mouth movements are computed as suggested by Cohen and Massaro [16].

In contrast to explicit scripting techniques, the image-based system proposed by Brand [10] learns the dynamics of real human faces during speech using original video footage. This information is then applied to create speech animations from novel audio input. The system generates mouth movements including coarticulation as well as additional speech-related facial animation, for instance, eyebrow movement.

A rule-based approach to create facial animations from speech has been presented recently by Albrecht *et al.* [1]. Prosodic parameters such as pitch contour, relative loudness, and pauses are extracted from pre-recorded speech using signal processing techniques. From this data, appropriate non-verbal facial expressions and head movements are generated by rules in addition to the speech-synchronized lip, jaw, and tongue movements [2].

Text-to-speech (TtS) techniques have been used by Pelachaud *et al.* [38] and by Cassell *et al.* [12] for synthesis of speech-synchronized animations of agents interacting with each other or with the user. The component that generates the text for the agent's speech has additional knowledge about content and structure of a piece of dialog, which is employed to generate appropriate gestures. In their more recent work, Poggi and Pelachaud [41] also include the dialogue situation into their animations. When suggesting something to your boss, your attitude will be completely different from when ordering your children not to do something. Using a similar approach as in [38, 12], Lundeberg and Beskow [33] have implemented a spoken dialog system featuring a virtual representation of the famous Swedish writer Strindberg. This agent is capable of communicating using bimodal speech augmented by simple generic punctuation gestures, for instance, nods or blinks. More complicated gestures have been explicitly designed for certain characteristic sentences.

In contrast to the above systems, which do not only generate the animations but also include a text generation module, the BEAT system by Cassell *et al.* [13] allows the user to input plain text which is then spoken by a virtual actor, adorned with additional facial expressions and body gestures. These include raising of the eyebrows, idle motor skills such as eye blinks, and either beat or iconic hand gestures from a knowledge data base. Unlike most other systems (including ours), the BEAT system performs linguistic and content analysis on its own instead of using the results from the analysis in the TtS system.

2.2 Paralinguistic Research

The relation of speech and eyebrow movement was systematically investigated by Ekman [22]. His pioneering research indicates that certain words and also greater parts of a sentence are often accompanied by raising or lowering of both the inner and the outer part of the brows. These facial gestures are called *batons*, when only one word is emphasized, or *underliners* for multiple words. The type of movement depends largely on the context: the brows will most probably be lowered in situations of perplexity, doubt or other difficulties. The situations in which raised eyebrows occur are not clearly defined. Eyebrow movements also serve as *punctuators*, i.e. they are used similarly to punctuation in written speech. Again, lowered brows indicate difficulties, doubt, or perplexity, but also seriousness and importance. To indicate that a question is being asked, eyebrows are often raised. During pauses caused by the speaker searching for words, raised brows occur accompanied by an upward gaze direction. Looking at a still object to reduce visual input is another typical behavior for word searches. Especially in conjunction with an '*errr*' sound, eyebrows may also be lowered in this situation.

According to Chovil [15], the largest amount of speech-related facial motion occurs with *syntactic displays*, i.e. batons, underliners, punctuators, etc. Again, movement of the eyebrows proves to be prevalent. Other facial motion of the speaker is, on the other hand, usually related to the *content* of the speech, e.g. a "facial shrug" while the speaker tries to remember something. Cavé *et al.* [14] investigated the link between eyebrow movement and pitch contour. In 71 % of the examined cases a correspondence was found, where rise and fall of the pitch of the speech signal coincided with raising and lowering of the speaker's eyebrows, respectively. The authors state that typically 38 % of the eyebrow movements occur during pauses or while listening. These movements are utilized to signal turn taking in dialogs, assure the speaker of the listener's attention, and mirror the listener's degree of understanding, serving as a back-channel. House *et al.* [24] found that both eyebrow and head movement contribute greatly to the perception of importance of speech. With respect to timing, perceptual sensitivity is in the range of 100–200 ms, coinciding with the average length of a spoken syllable. Investigating the relationship between questions and gestures, Cosnier [17]

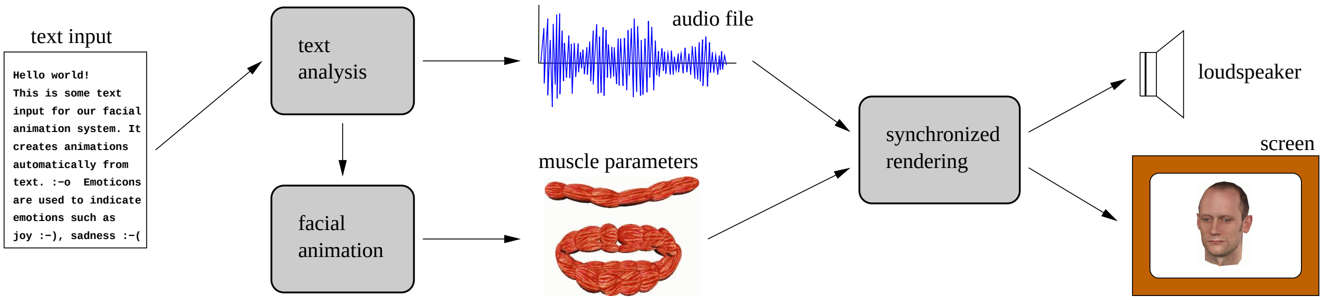


Figure 1. Components of our system: text input is converted into a speech signal using linguistic analysis, which additionally drives the muscle-based facial animation system. Rendering and audio output are synchronized in the final animation sequence.

found that for informative questions (i.e. not related to the interaction itself) head and eyebrow movements do not differ from normal informative conversation, with the exception of raising the head and possibly the eyebrows at the end of a question. However, the visual focus is more often on the listener than during statements.

2.3 Text-to-Speech Synthesis

Text-to-Speech synthesis [20] is a method for converting written text into audible speech. It consists of a text analysis part, generating a symbolic representation of a spoken utterance including a phonetic transcription of the words, followed by the actual speech synthesis part, in which the symbolic representation is converted into audible speech.

Early systems, such as the MITalk system [3], employed formant synthesis algorithms leading to relatively unnatural, "robot-like" voices. A major improvement in naturalness was brought about by concatenative synthesis techniques [21, 8], which produce synthetic speech by resequencing human recorded speech samples. These new synthesis techniques have increased the intelligibility of synthetic speech considerably. Naturalness, however, is still a prime issue. First attempts have been made to increase the expressive capabilities of synthetic voices by modeling vocal emotions [42]. These systems are starting to be used, for instance, in audio-visual speech synthesis [46].

Speech synthesis systems that are to be used in conjunction with facial animation need to provide intermediate processing results such as timing information in addition to the resulting speech. The most wide-spread research system for speech synthesis, the open-source FESTIVAL system [7], uses its own, relations-based data representation for this purpose. New systems using XML-based internal data representations, such as BOSS [29] and MARY [44], make the output of partial processing results a straightforward task. The XML data can be further analyzed by subsequent processing components using standard XML parsers.

3 System Overview

Our facial-animation-from-text system consists of three major components:

1. a text analysis module;
2. a facial animation module;
3. a module to synchronize rendering and audio output.

Figure 1 illustrates the connections between these modules.

The text analysis module performs linguistic analysis of the input text and creates a synthesized speech signal using a male or female voice. This process is described in more detail in Section 4. The results of the linguistic analysis are passed on to the facial animation module, which performs two different tasks: from the phoneme-based representation of the input text, speech-synchronized muscle contraction parameters of the facial muscles related to speech are generated. This process takes into account coarticulation and is further explained in Section 5. Additional high-level linguistic information such as different types of accents, pauses, and sentences is converted into non-verbal speech-related facial expressions, which are also represented as muscle contraction parameters, see Section 6. Finally, the animation sequence resulting from the muscle contraction parameters is rendered, in synchrony with the audio output of the speech signal. Using current PC hardware and graphics boards, we achieve real-time rendering frame rates of about 100 fps. Section 8 contains some more detailed information about this rendering step.

4 Generation of Speech from Natural Language Text

Our system uses Text-to-Speech (TtS) synthesis techniques (cf. Section 2.3) for the creation of the speech signal as well as for the description of the speech signal structure needed for the audiovisual synchronization. The MARY TtS system for German [44], which is publicly accessible through the URL `http://mary.dfki.de/`, was integrated into our system for performing these functions.

The MARY system creates speech from text in five major processing steps. In a first step, a shallow linguistic analysis of the plain text input is performed, using statistical algorithms trained on large text corpora [11]. This analysis component consists of a tokenizer identifying word and sentence boundaries including, in particular, the role of dots (abbreviation, ordinal number, or sentence-final); a part of speech ("noun", "adjective") tagger; and a local syntactic parser using statistically trained trigram models [45]. In the case of special text types such as poems (see Table 1), the tokenizer performs an additional segmentation at line breaks, needed for the line-based speech rhythm typical for poem reading. In a second step, a phonetic transcription is assigned to each word. This component performs a lexicon lookup for each of the words, and assigns the phonetic transcription to the known words. Unknown words, such as proper names, are transcribed by means of a set of transcription rules. Thirdly, an intonation contour including accents and boundaries is assigned to each sentence on the basis of the linguistic analysis. Accents are placed on content words, leading to a pitch excursion and thus to a perceptual prominence needed by listeners for understanding the meaning of the text. Intonation rises and falls at boundaries to reflect the sentence type (question vs. statement). In a fourth step, the symbolic information about phonetic transcription and intonation is used for determining the precise acoustic parameters of the utterance, each phoneme's duration (in milliseconds) and the shape of the intonation contour (using a sequence of target points with fundamental frequency expressed in Hertz). In a final step, these values are interpreted by a waveform synthesis algorithm to create a speech signal.

The MARY system was particularly well suited for being integrated into our facial animation system. It was designed to provide, in addition to the synthesized speech, as much explicit information as possible about the individual processing steps it runs through. This information is valuable for a number of reasons. Most obviously, exact timing information of the individual sound segments produced (determined at the second and the fourth processing step described above) is needed for a proper synchronization of lip movement with the sound. In addition to that, higher level information is available, such as the type (step three) and timing (step four) of accents, which correspond to the important bits of the sentence. Proper analysis of these accents allows the rendering of appropriate time-aligned facial gestures, thus conveying a truly multi-modal pattern of accentuation expression, contributing to both naturalness and intelligibility of the synthesized audio-visual speech [12]. A further type of high-level information that can be extracted from the MARY output is the type and location of boundaries (pauses), including a differentiation between sentence-internal and sentence-final pauses, as well as sentence type (determined in step one, and specified in steps three and four). The ability to make such distinctions is a prerequisite for assigning proper non-verbal facial expression, e.g. gaze behavior which differs between questions and statements, and between sentence-internal and sentence-final pauses.

A significant advantage of the MARY TtS system is that its data representation is based on XML. Among other things, this allows XML-based markup to be provided in the text input and to be passed on to subsequent processing components such as, in this case, the facial animation component, see also the example in Section 7. This property has been put to use for the automatic expression of emotions in the speech-synchronized facial animation.

As a simple means for the textual representation of emotions, so-called *emoticons* ("smileys", "frownies", etc.) are widely used, particularly in e-mails. We propose a simple but effective method for interpreting these emoticons for generating appropriate facial expressions. As a first list, the following emoticons are recognized.

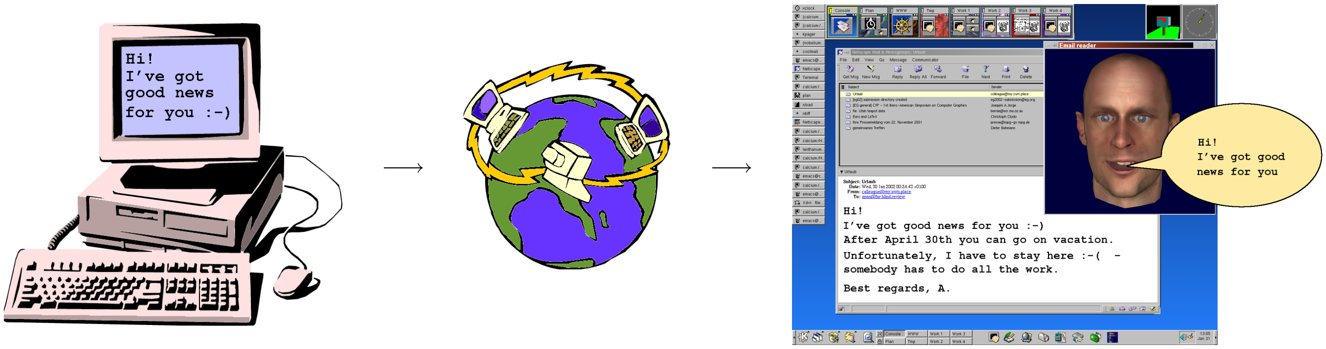
emoticon	emotion	emoticon	emotion
`:-)`	happy	`;-)`	kidding
`:-(`	sad	`>:-<`	angry
`:-o`	surprised	`:-|`	disgusted

These emoticons are automatically translated into XML-based emotion markup before the text is fed into the TtS system. Currently, the TtS system merely passes this information on to the visual generation component. The implementation of appropriate vocal changes, reflecting the emotion in the synthetic voice, is under development, see Section 9.

5 Speech Animation and Synchronization

The data required for the synchronization of lip movement to audio is provided by the TtS system, which generates the corresponding phonemes along with their durations from plain text input in a SAMPA representation. SAMPA (*Speech Assessment Methods Phonetic Alphabet*) is a machine-readable phonetic alphabet.

For realistic speech animation, the modeling of *coarticulation* is crucial. This term describes the coloring of a speech segment by surrounding segments. The visemes corresponding to the /k/ in 'cool' and the /k/ in 'cake', for instance, are very different, although described by the same phoneme.



Subject	Sender
Urlaub	colleague@my.own.place
[eg02] submission directory created	eg2002-submission@eg.org
[EG general] CfP – 1st Ibero-American Simposium on Computer Graphics	Joaquim A Jorge
Re: Utah teapot data	bernie@ecr.mu.oz.au
Euro and LaTeX	Christoph Clodo
Ihre Pressemeldung vom 22. November 2001	presse@mpg-gv.mpg.de
gemeinsames Treffen	Dieter Buhmann

Subject: Urlaub
Date: Wed, 30 Jan 2002 00:24:43 +0100
From: colleague@my.own.place
To: anon@for.blind.review

Hi!
I've got good news for you :-)
After April 30th you can go on vacation.
Unfortunately, I have to stay here :-(-
somebody has to do all the work.

Best regards, A.

Figure 2. Possible application scenario: a text message is embellished with additional emoticons, sent over the network as an e-mail, and read on the receiver's computer by a virtual character showing corresponding emotions.

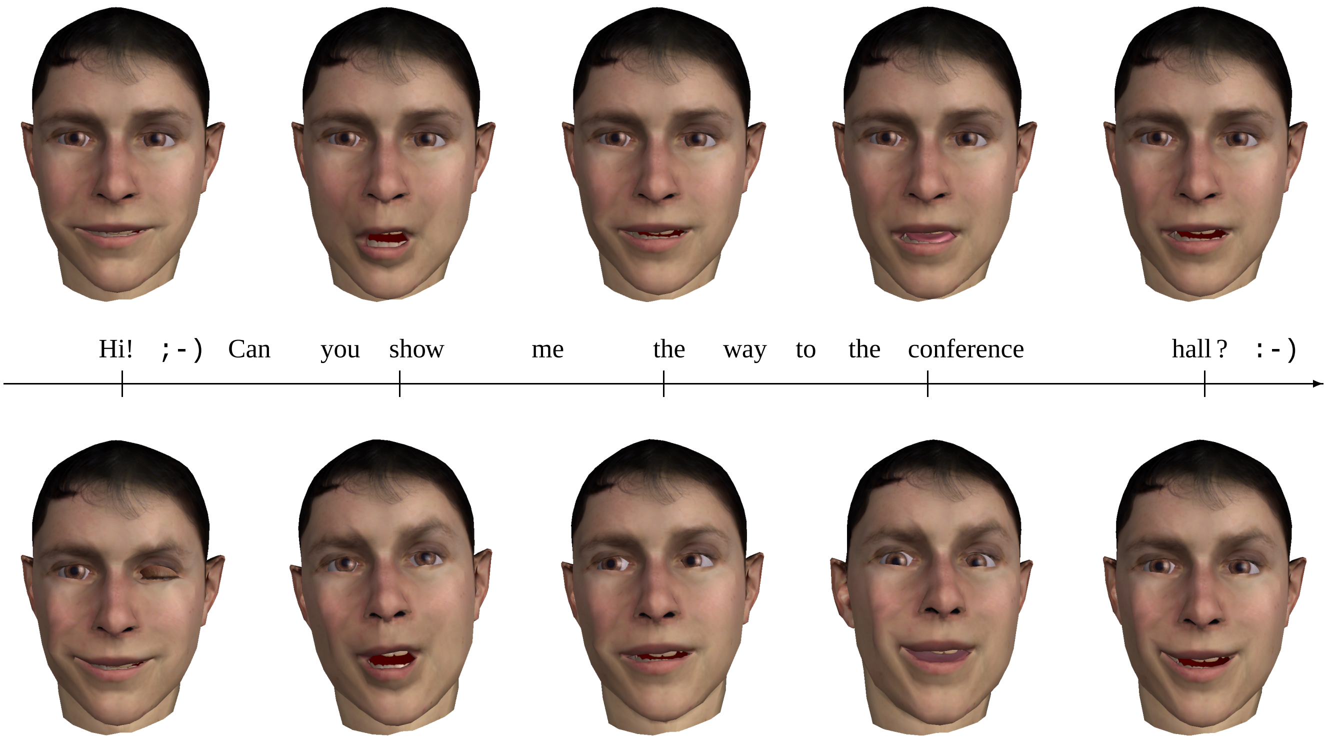


Figure 3. Snapshots from a facial animation sequence generated from the text input "*Hi! ;-) Can you show me the way to the conference hall? :-)***". Top row: movements of lips and jaw are synchronized to the audio signal created from the text input. Bottom row: additional non-verbal facial expressions are generated from the results of a linguistic analysis of the text input.**

The lip sync algorithm by Cohen and Massaro [16] generates lip movement for speech and considers coarticulation using a concept of *dominance* of speech segments over facial parameters. For each phoneme, every parameter has been assigned a target value. If a segment has a high dominance over a given parameter, it is important that the parameter comes close to its target value during this segment. This behavior is expressed via one dominance function for each segment-parameter pair. These functions do not necessarily vanish at the segment boundaries, but can well reach into neighboring segments, modeling coarticulation by the overlap of several such functions. The behavior of a parameter is described as the sum of all its targets in the segments forming the utterance, weighted by the values of the corresponding dominance functions at the given point in time.

The algorithm proposed by Cohen and Massaro [16] is well suited to handle time-based phonetic descriptions of speech such as the SAMPA representation. Since we adopted a recent approach to muscle-based facial animation [26] with some additional enhancements to obtain better results in speech animation [27], we have modified the Cohen/Massaro algorithm to directly control the contraction values of facial muscles, which are the animation parameters of our system. We also incorporated some recently proposed modifications [2] to the original algorithm: computations are simplified by considering coarticulation only over a finite range of phonemes. Moreover, to capture points of lip closure in the animation, additional frames are inserted at the beginning and end of the lip closure interval. This affects the bilabial stops, e.g. /p/ or /b/, and some other important consonants, such as /m/ and /f/. Very short phonemes are treated in a similar fashion to ensure they are not missed by the otherwise uniform sampling of the animation.

It turned out that the muscle-based facial animation approach we employ [26] is particularly well suited for automatically generating speech synchronized animation sequences. Since the virtual facial muscles are defined on a reference head and can be easily transferred to a new head model [27], no tedious parameter tuning is required for an individual head model: for each phoneme, the muscle contractions that have been found to result in the correct corresponding viseme on the reference head can simply be reused. Finally, since audio signal and muscle contractions are generated from the same time-based data, the resulting animations exhibit perfect synchronization of audio and video (see also Section 8).

6 Generating Facial Expressions

The decoration of speech with non-verbal facial expressions in daily face to face communication is so habitual to us that we are not even consciously aware of it. If, however, these facial expressions are missing from a facial animation, we perceive the presentation as boring. It even becomes more difficult to understand the meaning of the message due to the lack of visual structuring. Apart from providing structure, non-verbal facial motion also serves to accentuate words to indicate high importance, to facilitate turn-taking between speaker and listener, to underline that a question is being asked, to express emotions or opinions, and also to fulfill physical needs such as lip moistening or eye blinking.

6.1 Extraction of Speech-Related Facial Motion

Our animations include automatically generated speech-related eye movements, blinks, eyebrow gestures, and head movement. All movement is synchronized at the phoneme level, i.e. is realized at the phoneme boundary closest to the computed point of occurrence.

In a conversation between two or more real human beings, a large amount of turn-taking occurs where the flow of the dialog is controlled using facial gestures. Since we have only one virtual character talking to the user, we model just a small subset of these turn-taking gestures to emphasize certain parts of the text spoken by the virtual head, for instance, facial expressions during questions.

Eye blinks: Eye blinks are created as punctuators during pauses to structure the utterance. Additional eye blinks are inserted to keep up with the natural rate of cornea moistening, which occurs on average once every 4.8 seconds [39].

Questions: When posing a question, the eyes of the virtual character are directed at the user, who is assumed to sit directly in front of the screen. This kind of “making eye contact” increases the impression of the virtual face being aware of his vis-à-vis. Missing eye contact is unsettling, because the user feels unsure if he is really addressed by the question. Additionally, questions are marked by raising the eyebrows and the head on the last word, lasting over a potentially following pause. The TtS system provides information about word and sentence boundaries as well as on the type of the sentence, which ensures a correct placement of the corresponding facial behavior.

Expressions linked to intonation: We generate eyebrow and head movement from the intonation information produced by the TtS system. The intonation data is given in the form of `(time,value)` pairs, where the value indicates the fundamental frequency or pitch value at the given time. From this data, the local pitch extrema are extracted. In the animation, the head is raised at every local pitch maximum proportionally to the pitch value. For every local pitch minimum, the head returns to its rest position, independent from the actual pitch value at the minimum. To avoid repetitive movements, the head is randomly rolled slightly sideways in about 75 % of all head raises.

Eyebrow raising is implemented similarly, but occurs less often: head movement is more frequently used, because it is easier to detect over distances [24]. According to Cavé *et al.* [14], we specify the probability of occurrence of an eyebrow gesture at a pitch maximum as 0.71. The randomized distribution of actual eyebrow gestures also makes the animation less predictable. Following the observations made by Cavé, only the amplitude of the left eyebrow movement depends on the pitch value. Thus the right eyebrow is always raised by a constant amount.

It is also important that the speaker does not direct his gaze away from the listener at intonation maxima. Otherwise conflicting information would be conveyed: looking away from the listener could indicate low importance, while intonational stress, head, and eyebrow movement signal that important information is communicated. In fact, the speaker may look explicitly towards the listener in order to emphasize what he is saying [5]. In our implementation, the talking head glances at the listener during the utterance of about 75 % of all accented syllables, and does not shift his gaze further away from the listener in all other cases.

The bottom row of Figure 3 shows an example for intonation-related facial expressions. The word *conference* is emphasized. Therefore, the fourth snapshot in the bottom row, which was taken during articulation of this word, shows the talking head with eyebrows and head raised, glancing at the listener.

Word search: Explicit eye movement is also performed during prolonged pauses within a sentence, especially when accompanied by an '*errr*' sound. In this case it can be assumed that the virtual character is searching for words. We simulate this behavior by letting the character either look down and frown, or raise his eyebrows and stare at an imaginary ceiling. He also shifts his gaze slightly to the right: Andersen [4] reports that rightward lateral eye movements are associated with verbal and linguistic activity.

Gaze: Argyle and Cook [5, p. 121] found that the speaker looks at the listener during grammatical breaks both as a signal and to obtain visual feedback. From the viewpoint of turn-taking, Duncan and Fiske [19] call this phenomenon a *speaker within-turn signal*. We have implemented this behavior through glances during pauses that coincide with intonational phrase boundaries (cf. the definition in Section 6.2).

If, apart from the explicit eye movements described above, the eyes of the talking head are kept completely still over the course of the animation, the awkward impression of a "dead stare" is evoked. Hence, we vary the view direction randomly within a small range, making the character appear more lively. Together with the eyeball rotation, the eyelids open or close slightly when looking up or down, respectively. The amount of eyelid opening depends also on the tilt angle of the head: if the head is tilted downward, the eyelids are opened more widely to allow for a straight viewing direction, and vice versa.

6.2 Facial Expressions from Emotion Tags

Emotions are embedded in the text input via emoticons and translated to XML markup included in the final rich XML representation generated by the text analysis component (cf. Section 4). We use this information in the animation module to generate appropriate emotional facial expressions. The first snapshot in the bottom row of Figure 3 shows the wink and the smile that accompany the friendly greeting indicated by the `;-)` symbol.

The climax of the emotion is determined by the position of the emoticon in the text input. We assume an intensity of 100 % of the corresponding emotion at its climax. The area of influence of an emotion is the *intonational phrase* during which it occurs. An intonational phrase is a natural unit in speech production, which often comprises only part of a sentence and is typically surrounded by pauses. Towards the borders of these phrases, the intensity of the emotion decreases linearly to zero. If an emotion is specified between two phrases, the emotion stretches over both phrases. For instance, in the poem in Section 7, the `:-(` emoticon placed after the colon in the second line of the second verse represents such a case. In our approach, emotional expressions and speech animation are combined by summing up the respective muscle contraction values.

7 An Example

Currently, our text-to-speech component supports the German language only. An extension to the English language is actively being investigated. In the accompanying video, we show a facial animation sequence that has been generated automatically from the poem "*Die zwei Parallelen*" by the German poet Christian Morgenstern (see Table 1). The emoticons `;-)`, `:-(`, `:-o`, and `:-)` in the poem have been inserted manually.

In the XML representation of this poem (see Table 2), the original text tokens are highlighted in grey, e.g. `gingen`. The animation module uses the original text representation only to find question marks. Sentences are bracketed by `<div>` and `</div>`. Text analysis and speech synthesis information pertaining to a single word is stored in a `<t ...>` `...</t>` pair. The attribute `sampa` contains the MARY SAMPA phoneme representation of the word. This information is used to determine the temporal position of the word in the phoneme representation of the input text. In this way, non-verbal facial expressions derived from XML tags are synchronized to the speech-related mouth movements generated from the phoneme representation. The XML tags

Es gingen zwei Parallelen
ins ;-) *Endlose hinaus,*
zwei kerzengerade Seelen
und aus solidem Haus.

Sie wollten sich nicht :-(*schneiden*
bis an ihr seliges Grab: :-(
Das war nun einmal der beiden
geheimer Stolz und Stab.

Doch als sie zehn Lichtjahre
gewandert neben sich hin,
da wards dem einsamen Paare
nicht :-o *irdisch mehr zu Sinn.*

Warn sie noch Parallelen?
Sie wußtens selber nicht, –
sie flossen nur wie zwei Seelen
zusammen durch ewiges Licht.

Das ewige Licht durchdrang sie,
da wurden sie :-o *eins in ihm;*
die Ewigkeit verschlang sie
als wie zwei Seraphim. :-)

Table 1. "*Die zwei Parallelen*" by Ch. Morgenstern (1905)

for the beginning and end of an intonational phrase are color coded as `<phrase>` and `</phrase>`, respectively. During the second phrase, the emotion ;-) ("kidding") is specified as `<emotion type="kidding"/>`. The corresponding emotion is blended in at the beginning of "*ins*", reaches its maximum at the specified position, and is faded out at the end of "*hinaus,*".

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE maryxml SYSTEM "http://mary.dfki.de/MaryXML.dtd">
<maryxml>
<speaker gender="male">
<phonology nasal_assimilation="on" precision="precise"
    schwa_elision="on">
<div>
<phrase>
<t g2p_method="lexicon" pos="PPER" sampa="’?{s"
    syn_attach="1" syn_phrase="_">
Es
</t>
<t g2p_method="lexicon" pos="VVFIN" sampa="’gI-N@n"
    syn_attach="1" syn_phrase="_">
gingen
</t>
<t accent="l+h*" g2p_method="lexicon" pos="CARD"
    sampa="’tsvaI" syn_attach="1" syn_phrase="NP">
zwei
</t>
<t accent="l+h*" g2p_method="lexicon" pos="NN"
    sampa="pa-ra-’le:-l@n" syn_attach="0" syn_phrase="NP">
Parallelen
</t>
<boundary breakindex="4" tone="h-l%"/>
</phrase>
<phrase>
<t g2p_method="lexicon" pos="APPRART" sampa="’?Ins"
    syn_attach="1" syn_phrase="PP">
ins
</t>
<emotion type="kidding"/>
<t accent="l+h*" g2p_method="userdict" pos="ADJA"
    sampa="’?{nt-lo:z@" syn_attach="0" syn_phrase="PP">
Endlose
</t>
<t g2p_method="lexicon" pos="PTKVZ" sampa="hI-’naUs"
    syn_attach="1" syn_phrase="_">
hinaus
</t>
<t pos="$," syn_attach="2" syn_phrase="_">
,
</t>
<boundary breakindex="4" tone="h-l%"/>
</phrase>
⋮
</div>
⋮
</phonology>
</speaker>
</maryxml>
```

Table 2. XML representation of the first two lines of the poem "*Die zwei Parallelen*" (cf. Table 1 and Section 7).

8 Synchronized Rendering

In a real-time setting, it is important to achieve not only high rendering frame rates, but also accurate synchronization to audio. In our system, the animation is generated by a physics-based simulation, running in its own thread on a dual processor PC. This simulation thread performs numerical integration of the equations of motion for a mass-spring network, which represents the facial skin layer. Muscle contraction parameters drive the simulation by imposing external forces upon the mass-spring system. The resulting displacements of skin mesh vertices for one simulation time step are stored as a *simulation key frame* in a buffer along with the current simulation time, measured in wall-clock time. We typically obtain simulation frame rates of about 40 key frames per second. The second thread on the other CPU is responsible for rendering. Here, successive simulation key frames are interpolated according to the current rendering time, which is also measured in wall-clock time.

Simulation and audio are running at the same speed, because animation parameters and audio are generated from the same phonetic description. Synchronization is thus achieved by initiating the audio output when the first frame is rendered. Due to stable rendering frame rates of about 100 fps, we obtain a high consonance between audible speech and rendered images with a maximum inaccuracy of about 10 milliseconds.

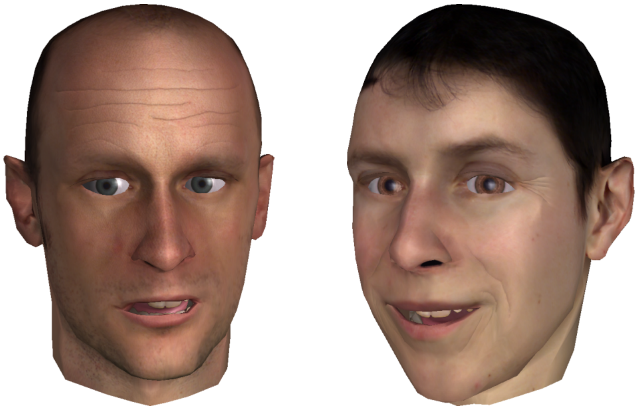


Figure 4. Two different facial expressions with automatically generated expressive wrinkles, rendered at 100 fps using hardware bump mapping.

The degree of realism of facial animations can be increased significantly by including expressive wrinkles with variable intensity. We make use of the *vertex program* and *register combiners* extensions of the NVidia GeForce3 graphics board to render bump mapped wrinkles at real-time frame rates. The bump map for the wrinkles is created from a “wrinkle height field” [49], which is in turn generated from the layout of the expressive wrinkles in the skin texture. The intensity of the wrinkles is controlled by the contraction values of the corresponding muscles. Contracting, for instance, the *frontalis* muscle, which is responsible for frowning, automatically results in wrinkles appearing on the forehead. Figure 4 shows two examples of different facial expressions with automatically generated expressive wrinkles.

9 Conclusion and Future Work

Our system provides an easy-to-use method to generate facial animation from text input with optionally included emoticons. Due to this simple interface and the full automation of the process after specifying the text, applications such as depicted in Figure 2 can be supported easily. The facial expression cues improve the quality of speech animation significantly. Naturally, this can be observed best when looking at videos that show a comparison of speech synchronized facial animation with and without additional non-verbal facial expressions. Such videos can be downloaded from `http://www.mpi-sb.mpg.de/resources/FAM/`. Figure 3 shows a comparison of some snapshots from such a facial animation sequence. Processing is fast enough to offer interactive response times in a dialog setting. More work needs to be done, though, to improve the naturalness of conversation in a human-machine dialog: the system would need to have knowledge about the emotional state of the user to be able to show appropriate reactions.

For emotion expression in the audible speech to become more convincing, vocal emotion cues must be delivered in addition to the facial emotion expression. The feasibility of emotional speech synthesis has been demonstrated [42], and an implementation using new ways of representing emotion for vocal expression [43] is currently under way. It is expected that the combined expression of emotion via both the visual and the auditory channel will improve the perceived naturalness.

Acknowledgments

Parts of this research are supported by the EC Project NECA IST-2000-28580.

References

[1] I. Albrecht, J. Haber, and H.-P. Seidel. Automatic Generation of Non-Verbal Facial Expressions from Speech. In *Proc. Computer Graphics International 2002*, pages 283–293, July 2002.

[2] I. Albrecht, J. Haber, and H.-P. Seidel. Speech Synchronization for Physics-based Facial Animation. In *Proc. WSCG 2002*, pages 9–16, 2002.

[3] J. Allen, S. Hunnicutt, and D. H. Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, 1987.

[4] P. Andersen. *Nonverbal Communication*. Mayfield Publishing Company, Mountain View, CA, 1999.

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

[6] G. Bailly. Audiovisual Speech Synthesis. In *Proc. ETRW on Speech Synthesis*, 2001.

[7] A. Black, P. Taylor, and R. Caley. Festival Speech Synthesis System, Edition 1.4. Technical report, Centre for Speech Technology Research, University of Edinburgh, UK, 1999.

[8] A. W. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Proc. Eurospeech*, volume 1, pages 581–584, 1995.

[9] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Computer Graphics (SIGGRAPH ’99 Conf. Proc.)*, pages 187–194, Aug. 1999.

[10] M. Brand. Voice Puppetry. In *Computer Graphics (SIGGRAPH ’99 Conf. Proc.)*, pages 21–28, Aug. 1999.

[11] T. Brants. TnT – a statistical part-of-speech tagger. In *Proc. 6th Conference on Applied Natural Language Processing*, Seattle, WA, USA, 2000.

[12] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. Animated Conversation: Rule-Based Generation of Facial Expression Gesture and Spoken Intonation for Multiple Conversational Agents. In *Computer Graphics (SIGGRAPH ’94 Conf. Proc.)*, pages 413–420, July 1994.

[13] J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *Proc. SIGGRAPH 2001*, pages 477–486, 2001.

[14] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. About the relationship between eyebrow movements and f0 variations. In *Proc. ICSLP '96*, 1996.

[15] N. Chovil. Discourse-Oriented Facial Displays in Conversation. *Research on Language and Social Interaction*, 25:163–194, 1991.

[16] M. M. Cohen and D. W. Massaro. Modeling Coarticulation in Synthetic Visual Speech. In N. M. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer–Verlag, 1993.

[17] J. Cosnier. Les gestes de la question. In Kerbrat-Orecchioni, editor, *La Question*, pages 163–171. Presses Universitaires de Lyon, 1991.

[18] D. DeCarlo, C. Revilla, M. Stone, and J. J. Venditti. Making Discourse Visible: Coding and Animating Conversational Facial Displays. In *Proc. Computer Animation 2002*, pages 11–16, June 2002.

[19] S. Duncan and D. Fiske. *Face-to-Face Interaction*. Lawrence Earlbaum Associates, 1977.

[20] T. Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, 1997.

[21] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken. The MBROLA project: Towards a set of high quality speech synthesisers free of use for non commercial purposes. In *Proc. 4th Int'l. Conf. of Spoken Language Processing*, pages 1393–1396, 1996.

[22] P. Ekman. About brows: emotional and conversational signals. In M. v. Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human Ethology: Claims and limits of a new discipline: contributions to the Colloquium*, pages 169–248. 1979.

[23] D. R. Hill, A. Pearce, and B. Wyvill. Animating Speech: An Automated Approach using Speech Synthesised by Rules. *The Visual Computer*, 3(5):277–289, Mar. 1988.

[24] D. House, J. Beskow, and B. Granström. Timing and Interaction of Visual Cues for Prominence in Audiovisual Speech Perception. In *Proc. Eurospeech 2001*, 2001.

[25] H. H. S. Ip and C. S. Chan. Script-Based Facial Gesture and Speech Animation Using a NURBS Based Face Model. *Computers & Graphics*, 20(6):881–891, Nov. 1996.

[26] K. Kähler, J. Haber, and H.-P. Seidel. Geometry-based Muscle Modeling for Facial Animation. In *Proc. Graphics Interface 2001*, pages 37–46, June 2001.

[27] K. Kähler, J. Haber, H. Yamauchi, and H.-P. Seidel. Head shop: Generating animated head models with anatomical structure. In *Proc. ACM SIGGRAPH Symposium on Computer Animation (SCA '02)*, pages 55–64, July 2002.

[28] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. SMILE: A Multilayered Facial Animation System. In *Proc. IFIP WG 5.10*, pages 189–198, Tokyo, Japan, 1991.

[29] E. Klabbers, K. Stöber, R. Veldhuis, P. Wagner, and S. Breuer. Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proc. Eurospeech*, volume 1, pages 521–524, 2001.

[30] Y. Lee, D. Terzopoulos, and K. Waters. Constructing Physics-based Facial Models of Individuals. In *Proc. Graphics Interface '93*, pages 1–8, May 1993.

[31] Y. Lee, D. Terzopoulos, and K. Waters. Realistic Modeling for Facial Animations. In *Computer Graphics (SIGGRAPH '95 Conf. Proc.)*, pages 55–62, Aug. 1995.

[32] J. P. Lewis and F. I. Parke. Automated Lip-Synch and Speech Synthesis for Character Animation. In *Proc. Graphics Interface '87*, pages 143–147, Apr. 1987.

[33] M. Lundeberg and J. Beskow. Developing a 3D-agent for the AUGUST dialogue system. In *Proc. Audio-Visual Speech Processing (AVSP) '99*, 1999.

[34] N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract Muscle Action Procedures for Human Face Animation. *The Visual Computer*, 3(5):290–297, Mar. 1988.

[35] F. I. Parke. Parameterized Models for Facial Animation. *IEEE Computer Graphics and Applications*, 2(9):61–68, Nov. 1982.

[36] F. I. Parke and K. Waters, editors. *Computer Facial Animation*. A K Peters, Wellesley, MA, 1996.

[37] A. Pearce, B. Wyvill, G. Wyvill, and D. R. Hill. Speech and Expression: A Computer Solution to Face Animation. In *Proc. Graphics Interface '86*, pages 136–140, May 1986.

[38] C. Pelachaud, N. Badler, and M. Steedman. Linguistic Issues in Facial Animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*. 1991.

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

[40] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing Realistic Facial Expressions from Photographs. In *Computer Graphics (SIGGRAPH '98 Conf. Proc.)*, pages 75–84, July 1998.

[41] I. Poggi and C. Pelachaud. Performative Facial Expressions in Animated Faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 155–188. MIT Press, 2000.

[42] M. Schröder. Emotional speech synthesis: A review. In *Proc. Eurospeech*, volume 1, pages 561–564, 2001.

[43] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proc. Eurospeech*, volume 1, pages 87–90, 2001.

[44] M. Schröder and J. Trouvain. The German Text-to-Speech synthesis system MARY: A tool for research, development and teaching. In *Proc. 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland, 2001.

[45] W. Skut and T. Brants. Chunk tagger – statistical recognition of noun phrases. In *Proc. ESSLLI Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany, 1998.

[46] J. Stallo. Simulating Emotional Speech for a Talking Head. Honour's thesis, School of Computing, Curtin University of Technology, Australia, 2000.

[47] D. Terzopoulos and K. Waters. Physically-based Facial Modelling, Analysis, and Animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, Dec. 1990.

[48] K. Waters. A Muscle Model for Animating Three-Dimensional Facial Expression. In *Computer Graphics (SIGGRAPH '87 Conf. Proc.)*, volume 21, pages 17–24, July 1987.

[49] C. Wynn. Implementing Bump-Mapping using Register Combiners. `http://www.nvidia.com/developer/`, 2001.