# Automatic Generation of Non-Verbal Facial Expressions from Speech

Irene Albrecht Jörg Haber Hans-Peter Seidel

Max-Planck-Institut für Infomatik
Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
E-mail: {albrecht, haberj, hpseidel}@mpi-sb.mpg.de

## Abstract

*Speech synchronized facial animation that controls only the movement of the mouth is typically perceived as wooden and unnatural. We propose a method to generate additional facial expressions such as movement of the head, the eyes, and the eyebrows fully automatically from the input speech signal. This is achieved by extracting prosodic parameters such as pitch flow and power spectrum from the speech signal and using them to control facial animation parameters in accordance to results from paralinguistic research.*

## 1. Introduction

While listening to another closely visible person, for instance in a dialogue or movie close-up, the main visual focus of the listener is on the mouth and the eyes of the speaker. But also non-verbal facial expressions such as movement of head and eyebrows or frowning play an important role during speech. They enhance understanding by emphasizing words and syllables of special importance. Speech synchronization for animated characters should thus not be restricted to mouth movements, but should rather include other speech-related facial expressions as well in order to render the animations more vivid and believable.

Speech-related facial animation is tightly coupled to the prosody of the utterance. Prosody, in turn, can be determined by the pitch of the signal. At the end of a question, the eyebrows raise and so does the pitch. In general, accented words and syllables come along with raised pitch.

The information that is not contained in the words themselves, but in the "acoustic packaging" of the utterance [2, p. 597], e.g. in prosody or frequency and duration of pauses, is referred to as *paralinguistic information*. Paralinguistic research provides valuable insights for the automatic generation of non-verbal facial expressions from speech: prosodic parameters of a speech signal such as slope and range of the fundamental frequency f0 are related to facial expressions. By extracting these prosodic parameters from

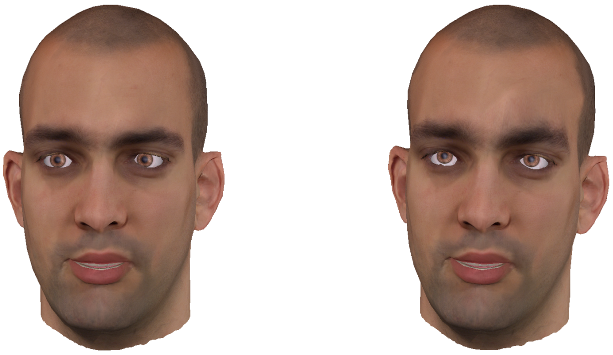

**Figure 1. Snapshot of a reflective moment during speech synchronized facial animation. Left: only mouth movement is generated from the speech signal. Right: additional movement of head, eyes, and eyebrows is generated automatically from prosodic parameters.**

the speech signal, we are able to automatically generate facial expressions that match the prosody of the utterance (see Figure 1).

We propose a method to automatically generate the following non-verbal facial expression 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 loudness and intensity of the utterance.

## 2. Previous Work

### 2.1. Facial Animation

Many different approaches for facial animation have been presented in the last thirty years. A comprehensive overview of the field can be found in the textbook by Parke and Waters [20]. In particular, several facial animation systems for speech synchronization have been proposed. Automated speech synchronization for recorded speech using linear prediction has been presented by Lewis and Parke [17], while an automatic approach to synthesize speech by rules has been proposed by Hill *et al.* [11]. Both approaches consider the movement of jaw and lips only. Cohen and Massaro additionally included movement of the tongue and introduced a technique for modeling coarticulation [7]. To combine speech synchronization with facial expressions, script-based approaches have been presented by Pearce *et al.* [21] and by Ip and Chan [13]. Here, the user specifies the facial expressions that should be displayed during speech in an application-dependent script language. Kalra *et al.* describe a script-based multi-layered approach to specify facial animation [16].

The image-based facial animation system introduced by Brand [3] takes speech as input and generates corresponding mouth movements including coarticulation as well as additional speech-related facial animation as, for instance, eyebrow movement. The system learns facial dynamics and states during speech from video footage of real humans and applies its "knowledge" to novel audio input.

Animation systems for agents that interact and communicate with the user or other agents have been developed by Pelachaud *et al.* [22] and by Cassell and Stone [4]. In these systems, text-to-speech techniques are used to synthesize the agent's speech. The same component that generates the text for the agent's speech also generates the accompanying gestures based on the content of the text and additional knowledge of the structure of the utterance and the course of the dialogue.

Lundeberg and Beskow [18] have developed a spoken dialogue system featuring a character that was designed to resemble the famous Swedish writer Strindberg. Similar to the above approaches, the agent is capable of communicating using bimodal speech accompanied by simple punctuation gestures like nods or blinks as well as by more complicated gestures especially designed for certain sentences.

## 2.2. Paralinguistic Research

Ekman [9] was one of the first to systematically investigate speech-related eyebrow movement. He noticed that raising of both the inner and the outer part of the brows is most often used for accentuating particular words (*batons*) and for emphasizing greater parts of a sentence (*underliners*). Ekman reckons that the choice of movement depends mainly on the context. When the speaker experiences distress, perplexity, doubt or other difficulties, the brows will probably be lowered, otherwise they will be raised. Raised or lowered brows are also used as *punctuation marks*, i.e. they are placed where in written speech some kind of punctuation mark would be placed. Again, lowered brows indicate some kind of difficulty, doubt, or perplexity as well as seriousness and importance. In addition to the context dependent use of eyebrow movement, eyebrow raising is often used to indicate that a question is being asked. Lowered eyebrows were also observed during word search pauses, especially together with an '*errr*'. Raised brows do occur as well in this context accompanied by an upward gaze direction. It is also typical for word searches that the eyes look at a still object to reduce visual input.

Chovil [6] reports that *syntactic displays* (batons, underliners, punctuators, etc.) occur most often. Among these, raising or lowering of brows are most prevalent. Other important movements of the speaker are related to the content of the talk, e.g. facial shrugs or expressions while trying to remember something. Cavé *et al.* [5] found that eyebrow movement during speech is often (in 71 % of the cases) linked closely to pitch contour. Raising pitch is usually accompanied by a raising of the eyebrows, while lowering of pitch and lowering of eyebrows also coincide. Typically, 38 % of overall eyebrow movements occur during pauses or while listening. They are used to indicate turn taking in dialogs, assure the speaker of the listener's attention, and mirror the listener's degree of understanding (back-channel). House *et al.* [12] examined the importance of eyebrow and head movement for the perception of significance. They observed that both movements are weighty here. Perceptual sensitivity to timing is about 100–200 ms, which is about the average length of a syllable. Cosnier [8] investigated the relationship between questions and gestures. He found that, apart from a head raising at the end of a question, for informative questions (as opposed to questions related to the interaction itself) the facial expression is not different from the head and eyebrow movement during normal informative conversation. The gaze, however, is directed more often towards the listener than during statements. Relations between emotions (joy, fear, anger, disgust, sadness, boredom) and prosodic parameters (f0 floor/range/slope, jitter, spectral energy distribution, number of accentuated syllables) have been investigated by Paeschke *et al.* [19] and by Johnstone and Scherer [14]. Accordingly, they report that most of the measured prosodic parameters are suitable to classify emotions.

## 3. Generating Non-Verbal Facial Expressions

We have implemented our method for automatic generation of non-verbal facial expressions from speech as a module in our facial animation system [10]. In this physics-based animation system, deformation of the skin is controlled by simulated contraction of facial muscles [15]. The set of muscles used in our animations is shown in Figure 2. Speech synchronized movement of lips and jaw is generated automatically from an input speech signal [1]. In addition to the contraction values of the facial muscles, we employ the following parameters for facial animations:

- rotation of the eyeballs / looking direction;
- variable opening/closing of the eyelids;
- rotation of the head (roll, pitch, yaw).

The generation of both speech synchronized mouth movement and non-verbal facial expressions is carried out in a preprocessing step, which takes about one minute for a speech signal of ten seconds duration on an 800 MHz Pentium III PC. Once the speech synchronized animation parameters have been generated, the simulation and rendering of the animation performs in real-time at about 30–40 fps on the same hardware with a GeForce2 graphics board.

### 3.1. Facial Expressions from Pitch

To automatically generate head and eyebrow movement from a speech signal, we first extract the pitch values of the utterance at a sampling distance of 10 ms. To this end, we use the Snack Sound Toolkit developed by Sjölander [23], which contains a variety of procedure calls for speech analysis. Since the production of unvoiced phonemes such as /p/ or /f/ does not involve vocal chord vibration, the notion of pitch does not exist for these sounds. Hence the pitch value is zero, which leads to a very rugged appearance of the pitch curve in turn. Therefore we eliminate these zero values and approximate the remaining pitch values using a B-spline curve. Next, the local minima and maxima of this curve are determined. Their positions and values, however, do not correspond exactly to

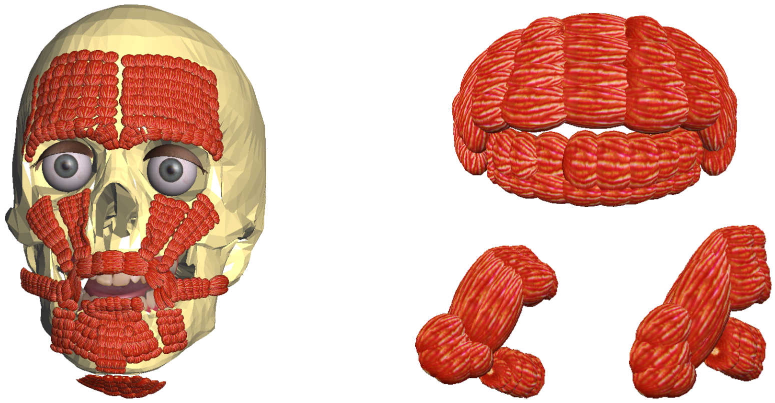

**Figure 2. Left: layout of the facial muscles used in our animations (skin surface removed); right: orbicularis oris muscle, front view (no contraction), and side views for protrusion and retraction.**

the minima and maxima of the original pitch curve. Thus, for every local maximum of the B-spline curve, position and value of the maximum of the original pitch data from the interval between the preceding and succeeding turning point of the B-spline curve are retrieved. An analogue process is performed for the local minima. The "reconstructed" original extrema are then used for the generation of head and eyebrow movement.

For each of the so determined maxima it is decided whether the differences between its value and the values of the preceding and succeeding minima exceed a given threshold. This threshold depends to a certain degree on the speaker due to the differences in voice melody: some people show greater pitch variations than others. For those maxima where the threshold is exceeded, the head is raised. The amount of head movement depends on the magnitude of the maximum. We typically generate head movements of at most three degrees rotation about the horizontal axis.

For each minimum, the difference values to the preceding and succeeding maxima are computed. Again, if the differences are larger than a given threshold, head movement is generated. In this case, the head is rotated back into its neutral position. This combination of upward and downward movement of the head supports accentuation of the speech. Figure 3 depicts different stages of the processing of the speech signal and the resulting animation parameters.

Both raising and lowering of the head are synchronized at the phoneme level, i.e. both beginning and end of head movements coincide with the closest phoneme boundary.

It is necessary to use only the most prominent maxima and minima, because otherwise too much head movement would be generated. In order to prevent monotony of the movement, the head is also randomly turned or tilted slightly to one side from time to time.

Head movement is often accompanied by analogue eyebrow movement: eyebrows are raised for high pitch and lowered again with the pitch. In our approach, eyebrow movement is generated using the same method as for the head rotations. Figure 3 (bottom) shows the eyebrow raising pertaining to the pitch.

According to Cavé *et al.* [5], only the magnitude of the left eyebrow's movement is related to the f0 pattern. This is taken into account by varying the degree of eyebrow raise for the left side according to the value of the current maximum. Cavé *et al.* also report that the duration of eyebrow raising is not correlated to the magnitude of the movement. This is inherently included in our implementation, since the duration depends only on the time step between the previous minimum and the maximum.

The only difference between questions and normal speech is that during questions the gaze of the speaker is directed towards the listener most of the time and always at the end [8], while for statements the speaker does not constantly look at the listener. If the speaker turns towards the listener, he either expects some feedback from the listener or wants him to take over the role of speaker. Since we do not model dialogues, this gaze behavior is not important for us.

### 3.2. Thinking and Word Search

During prolonged or filled pauses (e.g. " . . . errr . . . ") in a monologue, the speaker is typically either thinking about what to say next or searching for words. In both cases, similar facial expressions are exhibited: the gaze is directed at an immobile, fixed location to reduce input [9]. This location is usually either somewhere on the floor or up in the air. When people look up, they also raise their eyebrows. One possible explanation for this is an increase in the field of view when the eyebrows don't occlude part of the vision [9]. On the other hand, when people look at the floor while searching for answers, they often show a slight frown.

We have implemented this word search and thinking behavior during pauses. The duration of pauses that justify thinking and word search behavior seems to be speaker dependent and can hence be adjusted by a parameter.

### 3.3. Punctuators and Manipulators

Facial expressions that occur during speech at the same positions where punctuation marks occur in written text are called *punctuators*. They help to structure the flow of speech. A good example for such punctuators are eye blinks. In our implementation, we generate eye blinks at the beginning of pauses.

Movements that serve a physical need are called *manipulators* and are performed unconsciously. Eye blinks also serve the physical need to keep the cornea wet. Such additional blinks occur on average every 4.8 seconds [22]. To make our synthetic character more lifelike, we also include involuntary eye blinks: if the time elapsed between the previous and the next blink exceeds the threshold of 4.8 seconds, an additional blink is inserted.

As described by Pelachaud *et al.* [22], eye blinks consist of a closing interval (avg. duration: 1/8 s), the apex (avg. duration: 1/24 s), during which the eyes remain closed, and an opening interval (avg. duration: 1/12 s), where the eyes open again. Eye blinks are also synchronized to the speech: beginning of the closing, the apex, and the opening coincides with the nearest phoneme boundaries each. This behavior is simulated in our implementation.

Besides involuntary eye blinks, another example for a manipulator would be the moistening of the lips during extended speech periods. This can be implemented by letting the

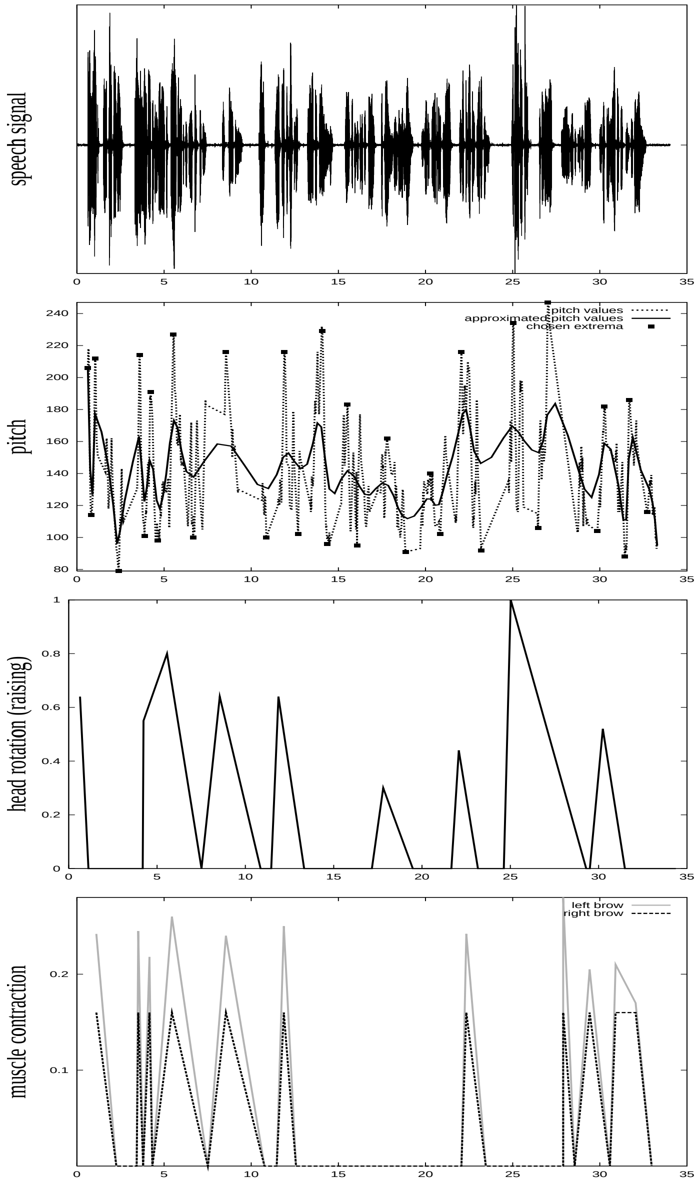

**Figure 3. Processing the speech signal (cf. Sec. 3.1). Top to bottom: input speech signal (about 33 s); original pitch values (stippled) and corresponding B-spline curve (solid); resulting head movement; muscle contractions of the frontalis muscles, which are responsible for the eyebrow movement (grey: left brow, black: right brow). Note that the movement of the left brow is scaled according to the magnitude of the pitch value.**

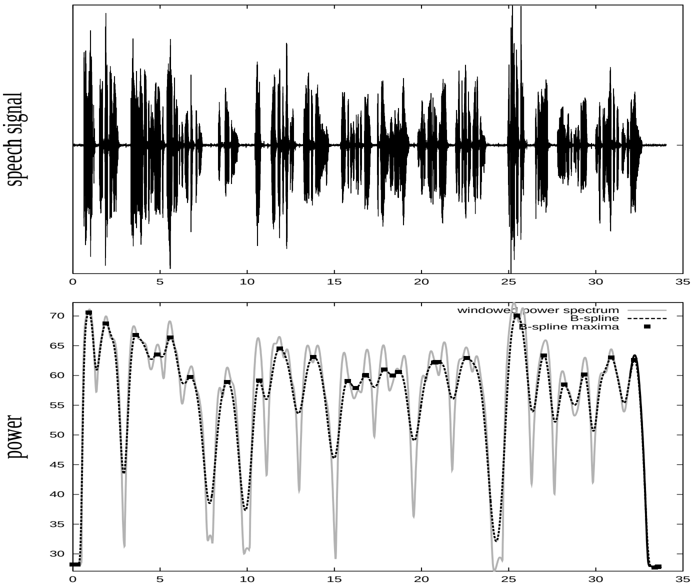

**Figure 4. The windowed power spectrum. Top: input speech signal. Bottom: windowed power spectrum (grey), approximating B-spline curve (black), and maxima of the B-spline (black squares). This information is used for scaling facial expressions with respect to loudness.**

synthetic character lick his/her lips during pauses that match the average lip moistening frequency best. During pauses where a thinking or word search expression is exhibited, the tongue motion should be slower, because the speaker is concentrating entirely on what to say next.

### 3.4. Random Eye Movement

During normal conversation, the speaker does not always look at the listener [8]. Moreover, eyes are almost constantly in motion. For an animated character lacking this behavior, the gaze is staring and dead. We have included additional random eye movement into our facial animations. Here it is important that the eye positions do not differ too much between two movements. Otherwise the movement would seem erratic and the character might seem agitated. As with all upward and downward eye movements, it is crucial that the lids accompany the eyeballs. If a person's gaze is directed downwards, the eyelids also close to a certain degree. Contrary, if one looks up, the eyelids open more to prevent an occlusion of the field of view.

### 3.5. Volume-controlled Intensity

Loudness primarily influences the magnitude of speech-related mouth movements for vowels. Additionally, it is also a good indicator for the distance of the person we are talking to. If somebody wants to pass on information to a person standing several meters away, he must speak louder in order to be understood. For the same reason he may also choose to intensify his speech-accompanying facial expressions. A very slight head movement, for example, is not perceivable at greater distances, so the speaker may want to nod more vigorously. Therefore we do not only scale the facial expressions pertaining directly to speech by the power of the signal, but it is also possible to do so for pitch related facial expressions. The extent to which this is done can be regulated by a parameter. This allows us to model differences in the behavior of the animated characters.

Using the Snack sound toolkit [23], we extract a windowed power spectrum of the speech signal and fit an approximating B-spline curve to it. An interpolating polynomial is fitted to the local maxima of this B-spline curve and normalized to a $[0, 1]$ range. It indicates the relative loudness of the speech signal. These relative loudness values are individually weighted for each animation parameter and used to scale the intensity of facial expressions. The weight for jaw rotation, for instance, is greater than the weight for eyebrow movement. As mentioned above, the weights can be modified to model characters with different tempers. Figure 4 shows the windowed power spectrum for an example sentence together with the approximating B-spline curve and their maxima.

## 4. Results

Incorporating non-verbal speech-related facial expressions into our facial animations definitely improved their naturalness and made them more convincing. Although the movements are generated by rules, random variations are taken into account to prevent the facial expressions from being entirely predictable. Some predictability, however, should remain indeed, since the accentuating facial expressions of humans tend to be predictable as well.

By specifying weights and frequencies for the movements of head, eyes, and eyebrows, different synthetic characters can be designed that exhibit different ways of visually accentuating their speech. This is also the case for real humans: some people tend to underline important parts of their utterances more by eyebrow movement, and some more by nodding. The frequency and amplitude of such movements depend highly on the temperament and culture of the individual as well. We would expect an Italian, for instance, to show much more facial and body gestures than a person from Northern Europe.

Figure 5 shows several snapshots from a facial animation sequence synchronized to a speech signal both with and without additional non-verbal facial expressions. The animation that includes non-verbal facial expressions looks clearly more convincing and lifelike. Since these enhancements are difficult to represent in still images, additional material including a movie can be found at `http://www.mpi-sb.mpg.de/resources/FAM`.

## 5. Conclusion and Future Work

We have presented a method to automatically generate non-verbal facial expressions from a speech signal. In particular, our approach addresses the movement of head, eyes,

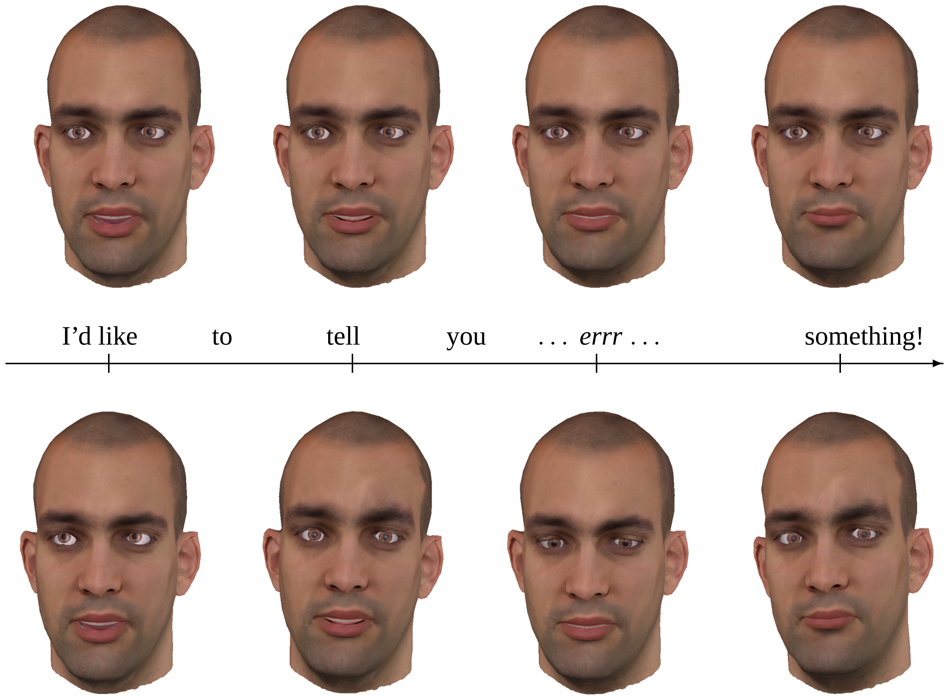

**Figure 5. Snapshots from a facial animation sequence synchronized to a speech signal with the textual representation: “I’d like to tell you** . . . *errr* . . . **something!”. Top row: movements of lips and jaw are generated from the speech signal. Bottom row: additional non-verbal facial expressions are created automatically from a paralinguistic analysis of the speech signal.**

eyelids, and eyebrows depending on prosodic parameters such as pitch, length and frequency of pauses, and the power spectrum of the input signal. These parameters are extracted automatically from the speech signal and control our facial animation parameters in accordance to results from paralinguistic research. Resulting animations exhibit a definitely more natural and vivid character compared to speech synchronized animations that control mouth movements only.

We are planning to extent our repertoire of prosodic movements further in order to increase the diversity of our animations. For instance, females often use nose wrinkling for prosodically motivated movements [9], and many people additionally employ tightening or widening of the eyes [6]. Incorporating facial expressions that match the emotion conveyed by the speech signal would enhance the realism of our system considerably. Some sophisticated signal processing should be sufficient to extract the f0 attributes that are characteristic of the basic emotions joy, sadness, fear, anger, disgust, and boredom.

## References

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

[2] D. Barr. Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In *Oralité et Gestualité. Actes du colloque ORAGE 2001*, pages 597–600, 2001.

[3] M. Brand. Voice Puppetry. In *Proc. SIGGRAPH '99*, pages 21–28, 1999.

[4] J. Cassell and M. Stone. Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. Technical Report FS-99-03, AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, 1999.

[5] 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.

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

[7] 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. 1993.

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

[9] 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.

[10] J. Haber, K. Kähler, I. Albrecht, H. Yamauchi, and H.-P. Seidel. Face to Face: From Real Humans to Realistic Facial Animation. In *Proc. Israel-Korea Binational Conference on Geometrical Modeling and Computer Graphics*, pages 73–82, Oct. 2001.

[11] 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.

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

[13] 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.

[14] T. Johnstone and K. Scherer. The Effects of Emotions on Voice Quality. In *Proc. XIVth International Congress of Phonetic Sciences*, 2000. in press.

[15] 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.

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

[17] 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.

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

[19] A. Paeschke, M. Kienast, and W. Sendlmeier. F0-Contours in Emotional Speech. In *Proc. International Congress of Phonetic Sciences '99*, pages 929–931, 1999.

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

[21] 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.

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

[23] K. Sjölander. The Snack Sound Toolkit. `http://www.speech.kth.se/snack/`, 1997–2001.