Generating Facial Expressions for Speech

Abstract

This paper reports results from a program that produces high quality animation of facial expressions and head movements as automatically as possible in conjunction with meaning-based speech synthesis, including spoken intonation. The goal of the research is as much to test and define our theories of the formal semantics for such gestures, as to produce convincing animation. Towards this end we have produced a high level programming language for 3D animation of facial expressions. We have been concerned primarily with expressions conveying information correlated with the intonation of the voice: this includes the differences of timing, pitch, and emphasis that are related to such semantic distinctions of discourse as "focus", "topic" and "comment", "theme" and "rheme", or "given" and "new" information. We are also interested in the relation of affect or emotion to facial expression. Until now, systems have not embodied such rule-governed translation from spoken utterance meaning to facial expressions. Our system embodies rules that describe and coordinate these relations: *intonation/information*, *intonation/affect* and *facial expressions/affect*. A meaning representation includes discourse information: what is *contrastive/background* information in the given context, and what is the "topic" or "theme" of the discourse. The system maps the meaning representation into how accents and their placement are chosen, how they are conveyed over facial expression and how speech and facial expressions are coordinated. This determines a sequence of functional groups: lip shapes, conversational signals, punctuators, regulators or manipulators. Our algorithms then impose synchrony, create coarticulation effects, and determine affectual signals, eye and head movements. The lowest level representation is the Facial Action Coding System (**FACS**), which makes the generation system portable to other facial models.

Generating Facial Expressions for Speech

Catherine Pelachaud, Norman I. Badler, Mark Steedman

Department of Computer and Information Science

University of Pennsylvania

Philadelphia, PA 19104-6389

Abstract

This paper reports results from a program that produces high quality animation of facial expressions and head movements as automatically as possible in conjunction with meaning-based speech synthesis, including spoken intonation. The goal of the research is as much to test and define our theories of the formal semantics for such gestures, as to produce convincing animation. Towards this end we have produced a high level programming language for 3D animation of facial expressions. We have been concerned primarily with expressions conveying information correlated with the intonation of the voice: this includes the differences of timing, pitch, and emphasis that are related to such semantic distinctions of discourse as "focus", "topic" and "comment", "theme" and "rheme", or "given" and "new" information. We are also interested in the relation of affect or emotion to facial expression. Until now, systems have not embodied such rule-governed translation from spoken utterance meaning to facial expressions. Our system embodies rules that describe and coordinate these relations: *intonation/information*, *intonation/affect* and *facial expressions/affect*. A meaning representation includes discourse information: what is *contrastive/background* information in the given context, and what is the "topic" or "theme" of the discourse. The system maps the meaning representation into how accents and their placement are chosen, how they are conveyed over facial expression and how speech and facial expressions are coordinated. This determines a sequence of functional groups: lip shapes, conversational signals, punctuators, regulators or manipulators. Our algorithms then impose synchrony, create coarticulation effects, and determine affectual signals, eye and head movements. The lowest level representation is the Facial Action Coding System (**FACS**), which makes the generation system portable to other facial models.

1 Introduction

Communication in face-to-face interactions is expressed through a number of channels, including the body, the voice, the face, and the eyes. While talking, people's faces are rarely still. They not only use their lips to talk, but raise their eyebrows, move or blink their eyes, or nod and turn their head [Ekman, 1979]. Facial signals seem to help to regulate the flow of conversation in much the same way as intonation does, signalling emphasis and contrast, as well as information related to turn-taking and control of the floor during an interaction. Facial signals also express affectual signals, which may be used communicatively, to influence the other participant's behavior [Argyle and Cook, 1976], [Collier, 1985]. While formal theories of the discourse semantics of such signals are generally lacking, in the area of spoken intonation, at least, there are more-or-less formalisable accounts of the way spoken intonation conveys discourse information associated with a speaker's messages [Halliday, 1967], [Bolinger, 1986], [Pierrehumbert and Hirschberg, 1990], [Steedman, 1991]. Similarly, some psychologists have claimed to find universal facial expressions linked to affects and attitudes [Ekman, 1979].

Animating the face by specifying every action manually is a very tedious task and often does not yield entirely appropriate facial expression. In order to improve facial animation systems, understanding linguistic semantics and its interaction with intonation is an important priority. It is likely that appropriate integration of gestural motions of the head and eyes, to accompany speech, will essential to effective simulation of human-like speaking agents.

The work reported here is exploratory, and was originally begun in advance of having a completely satisfactory theory of the association between discourse meanings, intonation, and facial expressions. We have been as much concerned with the development of such a fully formalised theory as with its exploitation in animation. Like most linguistic "fragments", the rules that we propose are incomplete. However, they are not arbitrary, and in every case have empirical motivations, whether in our own work or in the observations of others. While one goal of the research is to provide easily usable animation primitives, our intention is also to investigate the soundness of the discourse semantics the underpins their use, via a system in which their effects can be computationally generated by rule and investigated via controlled simulation.

Earlier systems for facial animation[Parke, 1982], [Magnenat-Thalmann and Thalmann, 1987] were based on a set

of parameters which affect not only the structure of the model (long nose, short forehead) but also its expressions (opening of the mouth, raising eyebrows). The separation between conformation parameters and expression parameters provides independence between facial features and the production of an expression.

More accurate facial motions can be obtained by simulating muscle actions [Platt, 1985], [Waters, 1987]. By integrating models of several layers of facial tissue with dynamic simulation of muscle movements, considerable realism in creating subtle facial actions may be achieved [Terzopoulos and Waters, 1991].

Automatic lip synchronization is included into animation systems by a multi-layer approach [Kalra et al., 1991] or by adding speech parameters [Nahas et al., 1988], [Hill et al., 1988]. A correspondence between each speech unit and a basic lip shape is established. Of particular relevance is the model of coarticulation proposed by [Cohen and Massaro, 1993]. It uses overlapping dominance functions. These functions specify for each viseme how close the lips reach their target value. Greater realism at the expense of synthetic control comes from techniques which extract information from live-animation [Williams, 1990], [Terzopoulos and Waters, 1991], [Essa and Pentland, 1994]. The computed movement information is interpreted as muscle contractions and is given as input to the animation system.

Texture mapping is used to enhance the realism of features and skin tone of the model [Nahas et al., 1988], [Williams, 1990], [Kurihara and Arai, 1991]. The consideration of wrinkles and of aging effects [Viaud and Yahia, 1992] adds much to the rendering of facial skin and expressions.

1.1 The Approach

Our main goal is to look at coordinated motion during a conversation. To do so, we need to understand the link between spoken intonation, the information transmitted in the given context, and facial movement. We use the **FACS** notation (Facial Action Coding System) created by P. Ekman and W. Friesen [Ekman and Friesen, 1978] to describe visible facial expressions[1]. This system is based on anatomical studies. Every facial action is due to muscular activity, relaxation or contraction. **FACS** describes temporary changes in facial appearance, how a feature is affected by specifying its new location, and the intensity of changes. An action unit **AU** corresponds to an action produced

[1] There has been some discussion lately (e.g., at a recent NSF facial workshop [Pelachaud et al., 1994]) that it may not suffice for fine description of the mouth region.

by one or a group of related muscles. Each **AU** describes the direct effect of a muscle plus eventual secondary motion due to the propagation of movement, and possible appearance of wrinkles or bulges. We refer the reader to [Ekman and Friesen, 1978] for a detailed description of each **AU**.

The highlights of this approach are the following:

- Our work resolves the difficulty of manipulating the action of each muscle at a keyframe level by offering to the user a higher level of animation by lip synchronization and automatic computation of facial expressions related to speech patterns (Fig. 1).

- We differentiate facial expressions linked to meaning from non-expressive ones. We have elaborated a repertory of such movements [Pelachaud et al., 1991]. We have included eye gestures as well as head gestures as part of the animation of the utterance. We view this partitioning of facial expression as an important tool for analyzing and exploring the significance or role of certain facial actions with spoken intonation and within the context of other facial movements.

- Facial actions are differentiated not only by their interpretations (highlighting a word or a pause, for example) but also by the facial features involved in the actions (such as eyebrow, head or eye movements). The separability of the dimensions of facial expressions allows us to decompose a facial expression along a number of independent dimensions of action. Indeed, the same facial action can be determined by different facial interpretations. Raised eyebrows can be the signal of "surprise" or can be used to highlight a word in an utterance. We define a *determinant* by its specific discourse-meaning interpretation relative to the flow of speech. A determinant can manifest itself by different facial actions (Fig. 1). For example, highlighting a word can be indicated by raised-brows, head nod, blink, eye flash and so on. A facial action can be unique to a determinant or it can be shared among different determinants (such as smile that can punctuate a pause or be a signal of "happiness"). The determinants are separated from each other even though they may have some facial actions in common. We list these determinants in Section 2.1. The final animation is obtained by summing together the individual sets of actions[2]. The mere fact that they merge together successfully to produce a unified animation is strong

[2]We assume for the moment that co-occurring actions are additive. That is in the case of smiling during speech, we simply add the **AU**s

evidence that this decomposition is reasonable.

- The different determinants for facial movements are intended to be experimentally manipulable in further research which will investigate their significance for human subjects. Our approach allows the contribution of any of the component determinants to be selected and the effects of the animation to be observed. For example, movements highlighting words could only be activated, to see what type of information they convey; or they alone could be disabled to determine if the facial animation conveys the same meaning. For the same reason, the system allows modular refinement of each determinant.

- The computation of each determinant of facial expressions is done by a set of rules which may be easily modified or augmented without altering any other part of the system. In particular, rules may be added or changed for one determinant without affecting the others.

- Individuals differ in the type of facial actions punctuating their speech (one individual may use mainly eyebrow movements, another may use nose wrinkling or eye flashes). They also differ in the number of displayed actions and their place of occurrence. We therefore define a facial action by two independent parameters. One parameter is the *type* of the facial action as a set of **AU**s; the other is the *time of occurrence* of the action in the spoken utterance. The user can therefore modify one parameter for one action without touching any other variable in the system. This process allows the user to independently vary the manifestation of the speaker's attitude (what is to be conveyed) and individuality in the computation of facial expressions [Moravetz, 1989].

- One important enhancement to the lip synchronization technique is the consideration of coarticulated effects where we examine how the action of a muscle is affected by temporal and spatial context. Indeed, the nature of muscle actions on the face, such as the contraction and release times, must be taken into account. This phenomenon is most apparent during rapid speech, when the mouth shapes created for sequences of phonemes is lost their characteristic shapes.

- The computation of the facial expressions linked to one particular utterance with its intonation and affect is done

corresponding to the smile and to the lip shapes.

independently of the facial model. Contrary to the technique of using a stored library of expressions which computes facial movements for one model only, this method works at the **AU** level for geometric independence. The facial expressions may be applied to any other facial model which uses **FACS** to drive its animation.

- The facial model we are presently using integrates the action of each muscle or group of muscles and propagates their movements though the skin [Platt, 1985]. It is programmed through **FACS AU**s.

1.2 Organization

In Section 2 we present the background of our system. We characterize the various determinants of facial expressions, describe head and eye movements computation, and present the intonational system we are using. We describe the overall system in Section 3, including the input representation, assumptions and properties. In Section 4 we describe the algorithms we have implemented for each determinant, with particular attention to lip synchronization and coarticulation problems. Finally in Section 5 we detail some examples where various affects and intonational meanings influence the final animation.

Throughout the exposition we illustrate our algorithms with examples.

2 Background

We present here the background and some definitions relevant to our study.

2.1 The Different Determinants of Facial Expressions

Among their other functions, facial movements are used to delineate items in a sequence as punctuation marks do in a written text [Argyle and Cook, 1976]. For example raising the eyebrows can punctuate a discourse. We consider the following determinants as defined in Section 1.1:

conversational signals : correspond to facial actions occurring on accented items or on emphatic segments; these actions clarify and support what is being said. Eyebrow movements appear frequently as conversational signals [Ekman, 1979], though rapid head movements, gaze direction and eyeblinks may also be involved.

punctuators : correspond to facial actions occurring on pauses; these actions can reduce the ambiguity of the speech by grouping or separating sequences of words into discrete unit phrases [Collier, 1985]. Specific head motions, a blink, or eyebrow actions may highlight a pause.

manipulators : correspond to the biological needs of the face (such as blinking to wet the eyes).

regulators : help the interaction between speaker-listener as they control the flow of speech. Breaking or looking for eye contact with the listener, and turning the head away or toward the listener are part of the elaborated interaction during a conversation [Duncan, 1974]. They are decomposed as a Speaker-State-Signal (displayed at the beginning of a speaking turn), a Speaker-Within-Turn (the speaker wants to keep the floor), and a Speaker-Continuation-Signal (frequently follows a Speaker-Within-Turn).

A more complete facial animation system is obtained if we include all these determinants. Nevertheless we presently exclude other facial functions not related directly to the pattern of the voice. We do not consider facial contortions (such as grimacing or twitching), nor do we compute automatically actions with a precise meaning (see Section 6). Indeed, a speaker may replace common verbal expressions by a specific facial expression, or may display part of the facial expression of an affect to mention it even though it is not actually being felt at the present moment [Ekman, 1979]. The appearance of such expressions is voluntary and depend on what is being said.

2.2 Specification of Head and Eye Movements

Each facial expression is expressed as a set of **AU**s. Since head and eye movements in our animation system are not coded as **AU**s but are defined by joint angles, we decided for the sake of simplicity and consistency to describe them separately. The final position of the head and of the eyes integrates the actions of all the facial determinants. Thus no generality is lost.

2.2.1 Head Movements

Continuous sequences of head movements support the verbal stream. Head movements may be associated with emblems (nodding or shaking for agreement/disagreement), or with maintaining the flow of conversation (turn taking

value deg/sec	**POS**	**RM**	**OM**	**SM**
max	125.0	77.0	59.2	28.8
min	77.0	59.2	28.8	3.4

Head movements are classified by their frequency and amplitude [Hadar et al., 1983].

POS: Postural Shift (high frequency, wide amplitude)

RM, OM, SM: small amplitude and various frequencies.

Table 1: Velocity values for each class of head movements

system). Head direction may depend upon affect ("sadness" is marked with a downward direction) (Table 10) or used to point at something.

Four classes of head motions are distinguished by their amplitude and frequency: slow movements (SM), ordinary movements (OM), and rapid movements (RM) [Hadar et al., 1983]. Postural Shifts (POS) are defined as linear movements of wide amplitude (i.e., they change the axis of motion) [Hadar et al., 1983] (Table 1). These speeds are affect-dependent (Table 12). Distinct patterns accompany linguistic features [Hadar et al., 1983] (Tables 4, 5, and 6). The occurrence of POS at the beginning of speech between Speaking-Turns [Duncan, 1974] and at grammatical pauses [Hadar et al., 1983] imply its involvement in speech production, regulation of turn-taking and finally, in marking syntactic boundaries inside clauses.

2.2.2 Eye Behavior

The eyes are always moving. Eye movements can be defined by the gaze direction, the point or points of fixation, the percentage of eye contact over gaze avoidance (with respect to another conversant), and the duration of eye contact [Argyle and Cook, 1976]. A common variable of eye behavior is *interest*. Eyes scan the objects of interest with longer glances. When looking at a picture of a person, viewers are found to look by saccade mainly at the eyes (58% of the time), then at the mouth (13%); the remaining regions of the face are each scanned just 1% of the time [Argyle and Cook, 1976].

2.2.3 Eye Contact

Eye contact is an important non-verbal process to establish relationship as well as to communicate with others:

1. Depending on the situation, eye contact or its avoidance can be variously interpreted [Argyle and Cook, 1976].

2. It plays an important role during social encounters to process information, to seek or send it, and to establish and synchronize the conversation [Argyle and Cook, 1976].

3. It is also linked with intonation. It is used to keep control of the communication process [Duncan, 1974].

4. It follows the same rules as head movements for speaking turns. Indeed, before ceasing to speak, eye contact is temporarily broken, then re-established, to signal the other's turn to speak [Duncan, 1974].

2.2.4 Eye Blinks

Eye blinks occur quite frequently. They serve not only to accentuate speech but also to wet the eye. Normally there is at least one eye blink per utterance. In this study we consider two types of blinks:

- Periodic blinks keep the eyes wet. On average, they appear every 4.8 sec. and last about 1/4 of a sec., with 1/8 sec. of closure time, 1/24 sec. of closed eyes, and 1/12 sec. of opening time [Argyle and Cook, 1976]. Nevertheless, their period of occurrence is affect-dependent [Collier, 1985] (Table 13).

- Voluntary blinks serve to emphasize speech, to accentuate a word, or to mark a pause [Ekman, 1979]. They are either synchronized to the word or to the syllable level [Condon and Osgton, 1971]. A blink is considered to be a conversational signal when it occurs on an accented word, a punctuator when it occurs on a pause.

2.3 Intonation

The intonational "melody" of an utterance can be viewed as conveying partial information of three kinds. The first is information about the syntax (and therefore the semantics) of an utterance [Selkirk, 1984], [Hirschberg and Pierrehumbert, 1986], [Pierrehumbert and Hirschberg, 1990]. We claim, following [Halliday, 1967], [Isard and Pearson, 1988], [Prevost and Steedman, to appear] that this information includes markers of questioning,

stating, and other speech acts, and markers of discourse information including topic or theme, comment or rheme, focus or new information, and background or given information. The second kind of information affecting intonation and prosody is affect or affectual attitude: involuntary aspects of the speaker's speech [Scherer et al., 1984]. A third of information concerns the conversational attitudes of the speaker: what stand the speaker takes towards the listener (e.g., politeness or irony which may be directly signaled or conversationally implied). Conversational attitudes may also include the conscious manipulation of affectual markers such as "calm" or "anger", but in our current research we are not considering such manipulations.

2.3.1 Vocal Parameters

A close relation between the syntax and the semantics of sentences and suprasegmental features has been suggested [Selkirk, 1984], [Steedman, 1991]. The latter reference claims that suprasegmental features are systematically related to discourse information units corresponding to topic or "theme" (that is, what the discourse segment is about) and comment or "rheme" (that is, what novel information the utterance supplies). Listeners may also detect the speaker's affect from prosodic features [Cahn, 1989]. Affects seem to be differentiated mainly by pitch (while frequency is a physical property of a sound, pitch is a subjective one), loudness (the perceived intensity of a sound), pitch contour (the global envelope of the pitch), tempo (rate of speech), and pause [Cahn, 1989].

2.3.2 Notational System

The notation for intonation contours that we use is derived from J. Pierrehumbert [Pierrehumbert, 1980]. We follow [Pierrehumbert and Hirschberg, 1990] [Prevost and Steedman, to appear] in assuming that different intonational "tunes" are used to convey various discourse-related distinctions of "focus", contrast and propositional attitude. We use the categories defined in [Prevost and Steedman, to appear]. Intonation contours serve to indicate the way that the current utterance relates to the context established by previous ones – for example, they may mark continuation of the same topic or theme, or the introduction of a new one.

We can represent informally the decomposition of an utterance into prosodic phrases using brackets (see below). The appropriate use of intonational bracketing is determined by the context in which the utterance is produced and on

the basis of what the speaker regards as the topic or theme of the utterance, and what he/she considers as requiring contrast as opposed to being background information. This bracketing is (partially) reflected in intonation.

Consider the sentence "Julia prefers popcorn" (the example is related to one discussed in [Steedman, 1991]).

context : I know that Harry prefers POTATO chips, but what does JULIA prefer?

bracketing: (JULIA PREFERS)(POPCORN).

accent : L+H* LH% H* LL%

The tune is also annotated more formally using Pierrehumbert's notation, in which (**H** and **L** denote high and low tones which combine in the various pitch accents and boundary tones. **L+H*** and **H*** are different kinds of pitch accent, and **LH%**, **LL%** and **L** are boundaries further decomposeable into phrasal and boundary tones.) The bracketing of the sentence, the placement of pauses and accents, and the type of the accents themselves vary with the context. Consequently the facial conversational signals and punctuators associated to each utterance differ also.

This system provides a case where intonation structure apparently departs from traditional surface structure. The speech generation component of our system is based on the Information Based Intonation Structure (IBIS) System of Prevost and Steedman [Prevost and Steedman, 1994], [Prevost and Steedman, to appear], which exploits a novel "flexible" approach to syntax and semantics based on Categorial Grammar to produce apparently appropriate intonation contours for spoken responses to database queries.

2.4 Underlying Property

An important property linking intonation and facial expression (in fact it extends to gesture and body movement in general) lies in the existence of synchrony between them [Condon and Osgton, 1971], [Unuma and Takeuchi, 1991], [Magnenat-Thalmann and Thalmann, 1987]. The face and the body do not move at random but in concert with the flow of speech. Thus, changes of body posture and orientation occurs at the beginning of a new topic of conversation. Similarly, a facial movement might be synchronized at the phoneme level such as blink, or at the word level such as eyebrow movement [Condon and Osgton, 1971]. Synchrony implies that changes occurring in speech and in body movements should appear at the same time. Thus facial synchrony is integrated in this body synchrony scheme as an

extension of this property. This is the basic principle which regulates the computations in our facial animation system.

In recent work with Justine Cassell this system and related assumptions about synchrony have been extended to manual gesture [Cassell et al., 1994].

3 Description of our System

The following Sections describe the system in more detail.

3.1 Input Assumptions

The input to the program is a file containing an utterance already decomposed and written in its phonological representation with its accents marked in its bracketed elements. That is to say that we sidestep the entire issue of recognition, leaving the integration of speech input for future work. Automatically finding the bracketing and intonational structure of a sentence is far from being a simple problem [Silverman, 1987].

The original work reported in [Pelachaud, 1991] used recorded natural speech to guide the animation. In this phase, after recording a sentence, the timing of each phoneme and pause was extracted from a spectrogram. In more recent work we use a query answering program including a sentence generation and a Bell Labs speech synthesizer to automate the determination of paralinguistic parameters and phoneme timing [Cassell et al., 1994].

At the beginning of the file, the user specifies the desired affectual parameters and their intensity (a number between 0 for minimum intensity and 1 for maximum intensity) (see Appendix B). Three levels of description are represented in the input. At the segmental level the sentence is specified (either by hand or by the generation program) as a list of strings corresponding to the phonetic representation of the utterance and whose notation is compatible with ascii-keyboard notation [DEC, 1985]. Pauses acting either as silence or as syntactic markers (such as comma or period) are included. Each segment and pause is followed by its duration expressed in seconds.

At the suprasegmental level a word is characterized either as a function word (‘F:’) (such as article or pronoun) or as a content word (‘C:’) (such as noun or verb). The modality of the utterance (such as declarative or interrogative) is noted by a boundary marker (‘.’, ‘?’, ‘!’).

At the linguistic level the accents and the bracketing of the utterance are defined. Following Pierrehumbert's notation we are using three types of markers: pitch accent, phrasal, and boundary tones. An utterance can be decomposed into intermediate or intonational phrases. Consider the example introduced in Section 2.3.2; it is represented by:

```
begin-intonational-phrase
C: JJ<0.08> (pitch-accent-L+H* UW<0.09>) LL<0.04> YY<0.078> AA<0.03>
C: PP<0.106> RR<0.026> AH<0.03> FF<0.098>
   (phrasal-tone-L (boundary-tone-H% ER <0.254>)) ZZ<0.088>
end-intonational-phrase
begin-intonational-phrase
C: PP<0.12> (pitch-accent-H* AO<0.118>) PP<0.096> KK<0.092> OXR<0.182>
(phrasal-tone-L (boundary-tone-L% NN<0.164>))
end-intonational-phrase
.<0.2>
```

In the recent phase of the work, this representation is derived entirely by rules from the output of the generation program and synthesizer [Prevost and Steedman, 1994], [Prevost and Steedman, to appear].

3.2 Organization of the Rules

The computation of facial expressions corresponding to each determinant (conversational signal, punctuator, regulator and manipulator) is entirely by rule[3]. Two parameters are used to define an action: its type and its time of occurrence. Our rationale is to allow the user to modify one of the parameters for one action without touching any other variable in the system. This is important, since the actions performed by a person while talking may vary. Most people show eyebrow movements to accentuate a word but other facial actions may be chosen such as nose wrinkling or eye flashes [Ekman, 1979]. The user just needs to modify the rules which describe the type of facial action and need not alter the rules of occurrence of facial action. Another unknown parameter is the frequency of occurrence of an action [Ekman, 1979]. A paralinguistic feature is not always accompanied by a facial movement (not every accented word

[3] Adults and children do not have the same systems of facial expression [Ekman, 1979]. Our focus is on an adult model.

is accompanied by an eyebrow movement, for example). Thus we need to have access to the timing of the occurrence of an action.

Not every rule involves the same level. Rules defining lip shapes work at the phoneme level; rules related to conversational signals are at the word level. The beginning and end of such a signal are computed by scanning the utterance word by word and not phoneme by phoneme. On the other hand, rules expressing affect alter the entire utterance. Moreover, affect is expressed vocally through the variation of paralinguistic parameters but does not modify the type or the placement of the accents relative to words in the utterance [Bolinger, 1986]. This is an important property that allows decomposition into **AU**s of the various facial patterns corresponding to either affect, accents, or other vocal parameters as well as simultaneous additive combination of all these facial actions. Therefore, given an intonational pattern, we can compute the corresponding facial actions through the defined set of rules, but the final occurrences and their types are affect-dependent.

3.3 Choice of the Rules

Facial actions such as conversational signals, punctuators and so on, have their intensity proportional to the speech-rate since their appearance follows the voice pattern. We compute the intensity of a facial expression in proportion to the speech-rate. We define two constants for each action, a minimal and a maximal one. Its actual value is a linear function of the speech-rate between the two constants:

$$\textit{intensity-AU} = \textit{minimum-AU} * \textit{speech-rate} + \textit{maximum-AU} * (1 - \textit{speech-rate}).$$

It should be noted that some rules of occurrence of a few facial actions may appear arbitrary. When there was a lack of information, we defined our own rules. We chose them intuitively (a “sad” person shows more hesitation in speech, uses more pauses and so on, thus we extrapolate that blinks for “sadness” might occur more frequently on pauses). As an example, for facial actions occurring in some accents, we define the term “first accent” as corresponding to the first accent of the utterances; such a choice follows Dittman’s results [Dittmann, 1974] which found that most movements happen at the beginning of the utterance. As new empirical results provide additional data, our rules can be refined to reflect this knowledge.

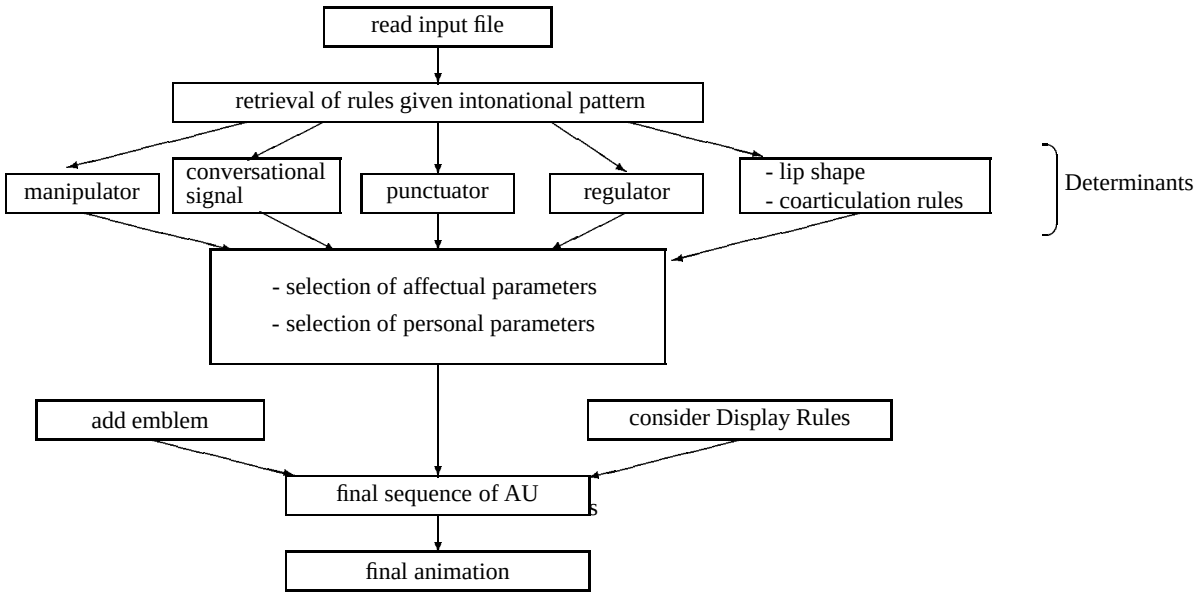

Figure 1: Algorithm of main program. Determinants are shown as parallel paths from input to parameters.

3.4 The Overall Algorithm

We first read the input file which contains the affect and its intensity, the phonemes and their timing, and the intonational structure (Fig. 1).

The next step is the computation of lip shapes which involves finding the set of phoneme clusters depending on the visual accuracy[4] and speech-rate, and then applying coarticulation rules. Each specified affect corresponds to a facial expression (set of **AU**s) which serves as a base with which the other facial actions are combined. Conversational signals and punctuators are then computed. The two parameters of those facial expressions (their type and their time of occurrence) are computed with respect to the given affect. Their lists of **AU**s are added to the one defining the affect and lip shapes. The beginning and end of any computed action follows the principle of synchrony. Any periodic blink is added. In the present stage of this study we consider only blinking in the manipulator determinant. Regulators are mainly characterized by head and eye motions.

The final list of **AU**s for each phoneme and pause is then obtained and the computation of the corresponding facial expressions can be performed using Steve Platt's program [Platt, 1985]. The script files describing the animation

[4] Visual accuracy is defined by the lighting conditions and visibility conditions of the speaker's lip [Jeffers and Barley, 1971].

are saved and the animation is played through *Jack*®[Badler et al., 1993], a 3-D human animation software system developed at the University of Pennsylvania.

3.5 Example

Before outlining each procedure of our algorithm, we present an example to clarify the process. Consider the utterance of *'(JULIA prefers) (POPcorn)'* with the affectual parameter *"disgust"* (see section 3.1 for detailed input description). (For examples involving affect like this one, the timings were derived by hand from read speech).

- First, the lip shapes are automatically computed for each phoneme.
- Phonemes are grouped into clusters depending on their lip shapes (see section 4.1). Fig. 2 depicts the lip shapes for the word *'popcorn'* in the case of fast and slow speech-rate[5]. During fast speech-rate, not only does the intensity of lip shape actions decrease but, in the case of deformable segments, their associated lip shapes lose their characteristic shape. The algorithm uses the value of the speech-rate so that the clustering of the phonemes depends on this parameter. As an example, lips open less for fast speech-rate.
- Then the program checks if a string belongs to a more deformable cluster (such as clusters containing 'n' or 't'). To these strings it applies forward and backward coarticulation rules. In this case, the considered item receives the same list of **AU**s as the vowel found by these rules. The phoneme /LL/ in the word *'Julia'* receives the same list of **AU**s with lower intensity as its preceding vowel /UW/. Indeed, /UW/ belongs to a less malleable cluster than /YY/. The algorithm applies the backward rule for /LL/ (Fig. 3).
- The next step considers the environment of each speech posture (i.e., the surrounding phonemes) and its relaxation and contraction times. As result, some lip shapes computed in the first two stages of the algorithm are automatically modified. This modification occurs either through the addition of new **AU**s to those already computed in the previous steps or through reduction in the intensity level of some of the **AU**s in the mentioned list. For the phoneme /YY/ in *'Julia'*, the program adds some pucker effect from the phonemes /UW/ and

[5]Fast speech-rate corresponds to more than 4 syllables per sec. Slow speech-rate corresponds to less than 2 syllables per sec.

/LL/. The lip shapes for /LL/ do not have enough time to relax completely from their puckered position to their extended lip shapes: some puckered effect remains, so the program has applied a temporal control. On the other hand, the pucker position of the item /AO/ from the syllable *'pop'* is altered due to its surrounding lip closures for the two /PP/s, so the program has applied a spatial control.

By the end of this part of the program, the lip shapes are computed. The program continues and goes through other procedures to compute the remaining facial movements. The remaining steps of the program are:

- The affect gives the overall orientation of the head. The list of **AU**s for the affect is computed and is added to the list of **AU**s for each item of the utterance (Tables 10 and 11).
- Conversational signals appear in this example, on pitch accents under various forms [Ekman, 1979]. Eyebrow movements coincide, for both actions, with the stressed syllables. Rapid movements around the actual position of the head characterize the head motion on the pitch accent. Moreover, blinks acting as conversational signals start at the beginning of the accented syllables and are synchronized at the phoneme level.
- From the specification of the affect *"disgust"*, no facial action is found between the two intonational phrases and only the juncture pause at the end of the utterance is considered (see Appendix C). A frown and a blink mark this pause. The blink begins and finishes at the same time as the pause.
- If the utterance is a statement, a Speaker-State-Signal (speaker looks away from listener) is emitted at the beginning and the head is positioned to look down as the speaker reaches the end of the sentence.
- The sentence finishes with slow head movement coming to rest.
- The last step is to look if more periodic blinks are needed. In our case, none are needed since already computed blinks occur at a sufficient rate.
- The facial expressions of each item of the utterance are computing using Platt's program [Platt, 1985]. Script files are output.
- The animation is done using the *Jack* software [Badler et al., 1993].

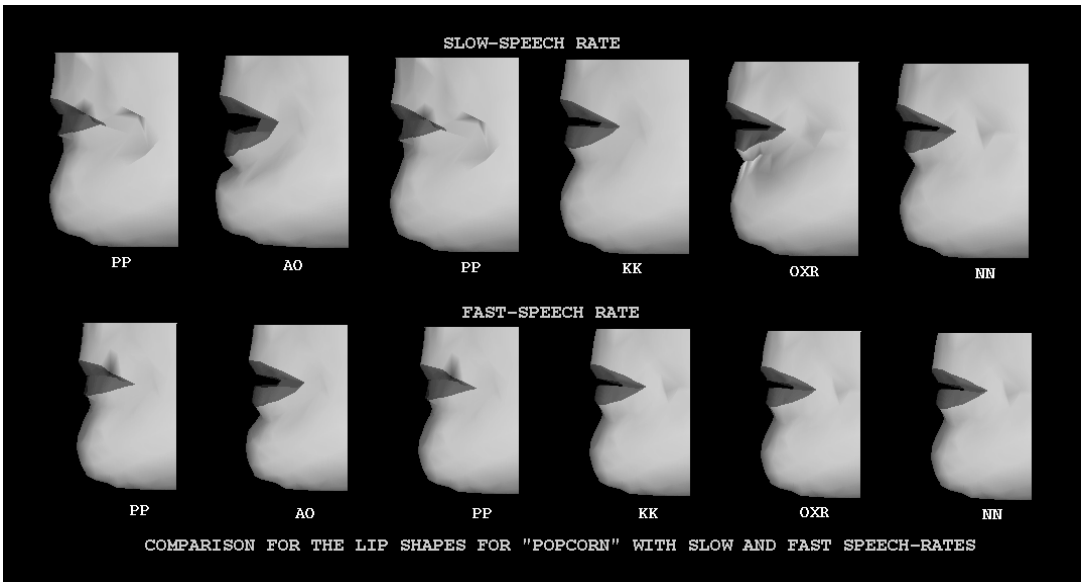

Figure 2: Comparison of the Lip Shapes for ‘popcorn’ with Fast and Slow Speech-rate

Figure 4 summarizes this sequence of coordinated expressions.

4 Details for each Determinant

So far, we have presented the main steps of the algorithm. We now show each of the determinants in more detail.

4.1 Lip Shape

Conventional cel animation resolves the problem of lip synchronization by defining a set of mouth shapes and timing for speech. Such techniques consider a small number of stereotyped speech postures to produce animation, including computer graphics [Emmett, 1985], [Kleiser-Walczak, 1988]. Even though they produce realistic animations, this technique requires a skilled animator and a considerable investment in time and manual effort. Other systems [Parke, 1982], [Lewis and Parke, 1987], [Magnenat-Thalmann and Thalmann, 1987], [Hill et al., 1988] offer a higher level of parameterization to their model. Parameters are grouped to represent the mouth shape of each phoneme. The user only needs to work at the phoneme level and not at the low level of facial parameters. But this technique has some drawbacks due to its lack of biological considerations.

Figure 3: Lip Shapes for ‘Julia prefers popcorn’ with Slow Speech-rate

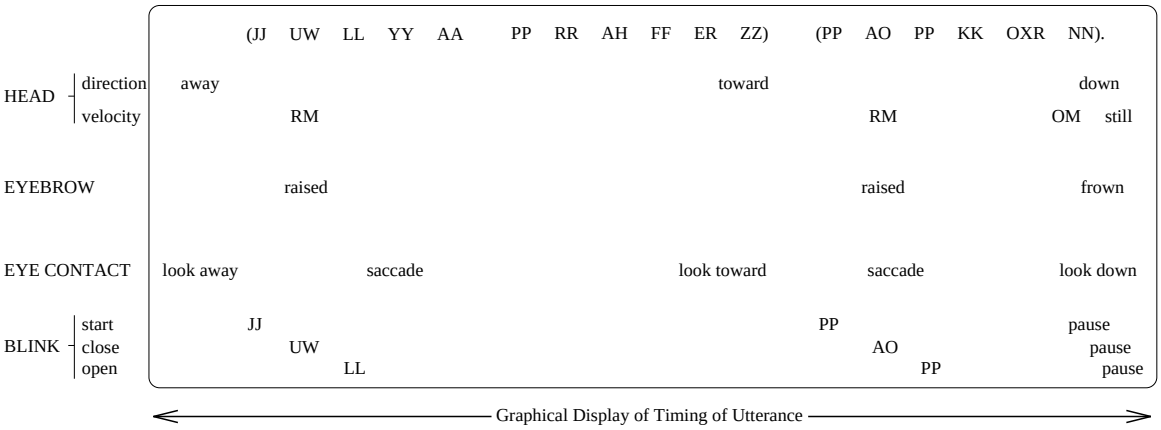

Figure 4: Coordination of Facial Movements for the example ‘Julia prefers popcorn’

cluster	**slow speech-rate**		**fast speech-rate**	
	context	list of **AU**s	context	list of **AU**s
{‘f’, ‘v’}	all	AU17+28L	all	AU17+28L
{‘iy’, ‘ixr’, ‘ih’, ‘ix’, ‘ey’, ‘exr’, ‘ah’}	prec. sgmt != ‘wh’, ‘ww’	AU11+12+25+28	prec. sgmt != ‘wh’, ‘ww’	AU25
	prec. sgmt = ‘wh’, ‘ww’	AU25		
{‘eh’, ‘ae’, ‘aa’, ‘axr’}	prev. sgmt != ‘wh’, ‘ww’	AU11+20+25	prec. sgmt = ‘wh’, ‘ww’	AU12+25
	prec. sgmt = ‘wh’, ‘ww’	AU12+25		

AU11: nasolabial furrow deepener AU12: lip corner puller AU17: chin raiser

AU20: lip stretcher AU25: lips part AU28: lip suck

Table 2: An example of rules to compute lip shapes

Speechreading techniques [Jeffers and Barley, 1971] offer a tool to interpret lip and facial movements to help the hearing-impaired understand speech. Clustering visible speech by facial expression enhances speech perception. In speechreading techniques, phonemes are grouped in the way that one lip shape corresponds to one cluster (called “visemes” [Benoit et al., 1990]).

These viseme groups may be ranked from the least deformable group (such as {‘f’,‘v’} cluster) to the most deformable one (i.e., very context dependent such as {‘l’, ‘n’} cluster) [Brooke, 1990], [Jeffers and Barley, 1971]. This clustering is speech-rate dependent (see Table 7). A person speaking fast moves lips much less than a person talking slowly. We decided to use this technique because of its reliability in describing visible lip movements. An example of the rules is given in Table 2.

The intonation of an utterance is the enunciation of a sequence of accented and non-accented segments. An accented vowel is differentiated acoustically from the remaining part of the utterance by its longer duration and increased loudness; visually, the jaw dropping motion is a characteristic of accented or emphasized segments. We shall draw on these observations in specifying our articulatory **AU**s.

4.2 The Coarticulation Problem

This phonemic notation, however, does not tell us how to deal with the difficult problem of coarticulation. Coarticulation arises from the temporal overlap of the articulatory actions realized by successive phonemic segments during their production. The realization of a consonant may be affected by anticipation of the following vowel (for example, the ‘r’ in ‘read’ and in ‘rat’ are articulatory and acoustically different). Similarly, the effect of a vowel may influence a succeeding consonant (for example, the ‘t’ in ‘rat’ looks and sounds different from the ‘t’ in ‘complete’).

A simple solution to the problem of coarticulation is to look at the previous, the current and the next segments to determine the mouth positions [Waters, 1987]. But in some cases this is not enough since the correct position can depend on up to five segments before or after the current one [Kent and Minifie, 1977]. Some rules look at the context of phoneme production to compute adequate lip positions [Brooke, 1990], [Kent and Minifie, 1977], [Cohen and Massaro, 1993]. Nevertheless no completely satisfactory set of rules solving every coarticulation problem exist.

We view lip movements corresponding to speech as a sequence of key positions (corresponding to phonemes belonging to non-deformable clusters) and transition positions (corresponding to phonemes belonging to deformable clusters). The problem is the shape computation of the transition position. We have implemented the look-ahead model. This model predicts that any articulatory adjustment starts just after a key-position and lasts until the next one. The transition position receives the same shape as the ‘strongest’ key-position (‘strongest’ meaning lip shapes belonging to the least deformable clusters). Two rules are considered: the forward and backward rules. They consider articulatory adjustment on a sequence of consonants followed or preceded by a vowel [Kent and Minifie, 1977]. Forward coarticulation arises in a sequence of consonants (not belonging to the low deformable clusters such as {‘f’, ‘v’} cluster) followed by a vowel, showing an articulatory adjustment. Respectively, backward coarticulation arises in a sequence of consonants (not belonging to the low deformable clusters such as {‘f’, ‘v’} cluster) preceded by a vowel, showing an articulatory adjustment (as cited in section 3.5, ’l’ of *Julia* receives the same lip shape as ’u’). Indeed, the lips show the influence of the vowel on the first consonant of the sequence. In the sequence *‘istrstry’* (French example taken from “sin*istre stru*cture” cited in [Kent and Minifie, 1977]) the influence of the ‘y’ is shown on the first

‘s’ (forward rule).

To solve particular problems (certain visual transitions between segments) which cannot be solved by these two rules, we consider a three-step algorithm. On the first step these coarticulation rules are applied to all clusters which have been defined as context-dependent. The next step considers relaxation and contraction time of a muscle. Finally, we look at the way two consecutive actions are performed. Therefore the speech and physical context is considered to yield a more physically-based model. In Section A we give an example of the effects of these coarticulation rules.

Given a segment belonging to a highly deformable cluster, the algorithm (Fig. 5) looks backward and forward for the first vowel member of a lower deformable cluster without consideration of word boundaries. The considered segment will take the lip shape of this vowel. Such a method ensures that each segment between the first one where the coarticulation rules applies and the low deformable vowel will have the same type of shape. (See example in Section 3.5).

After the first computation, we check that the current speech posture has time to contract after the previous speech posture (or, respectively, to relax before the next one). If the time between two consecutive articulatory configurations is smaller than the contraction time of a muscle [Bourne, 1973], the previous speech posture is influenced by the contraction of the current one. In a similar manner, if the time between two consecutive speech postures is smaller than the relaxation time, the current segment will influence the next segment when relaxing. Moreover, articulatory adjustments continue on the pause existing just after the considered word or is foreseen on the pause just before the word [Cathiard et al., 1991]. These influences are computed by simulating this muscular contraction and relaxation properties by two third-degree polynomial curves [Pelachaud, 1991].

Finally, we take into account the geometric relationship between successive actions. Lip closure is more easily performed from a slightly parted position than from a puckered position. The intensity of an action is rescaled depending on its surrounding context and on the cluster it belongs to (Table 3).

At the end of these steps, we obtain a list of **AU**s for each speech posture. The constraints between adjacent **AU**s are defined by a constant and are easily changed as is relaxation/contraction simulation. Moreover, lip shapes associated with each speech posture are determined by rules and are also easily modified. This is also a preliminary approach to solving the problem of correlating the acoustic and visual appearance of a phoneme, since we consider through

name/scaling factor	AU18	AU24
AU18	1	0.2
AU24	0.2	1

1: no deformation 0: high deformation

AU18: lip pucker AU24: lip presser

Table 3: Scaling factors for **AU**s

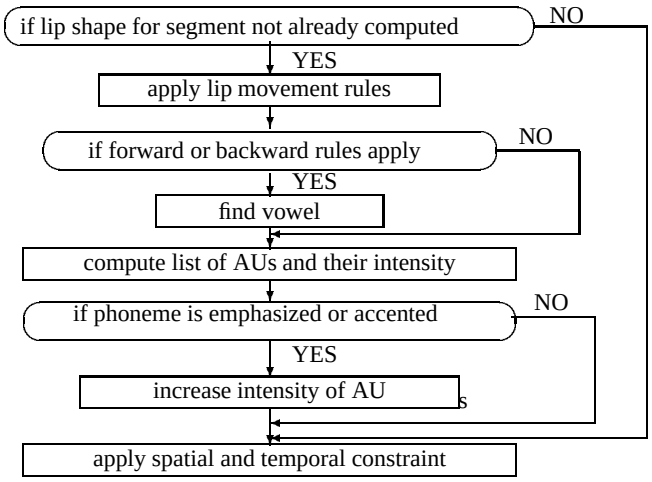

Figure 5: Algorithm of lip shapes computation and coarticulation rules

relaxation and contraction of a muscle, the notion of appearance and disappearance of lip actions. This approach may therefore eventually provide a tool for phoneticians to study coarticulation problems.

4.3 Conversational Signals

A stressed element is often accompanied not by a particular movement but by an accumulation of rapid movements (such as more pronounced mouth position, blinks or rapid head movements). Brow actions are frequently used as conversational signals [Ekman, 1979]. They can be used to accentuate a word or to emphasize a sequence of words. On accented words, actions may vary with the type of the pitch accent. The user has the possibility to choose the type of parameter defining the facial action. Nevertheless the final manifestation of actions is affect-dependent (Table 11).

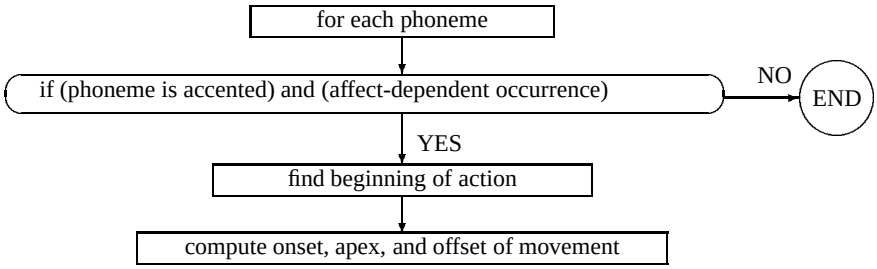

Figure 6: Algorithm of eyebrow actions

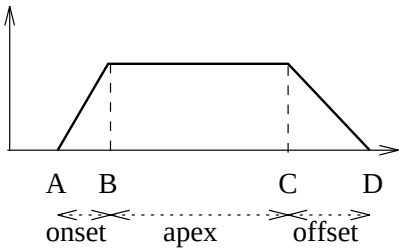

Figure 7: Computation of onset, apex and offset

Blink occurrence as conversational signals is affect-dependent (Table 14).

4.3.1 Algorithm

The program parses the sentence and looks at all possible movement occurrences considering the current affect.

When a facial movement appears on a stressed word or during emphasis the program computes the onset (time of appearance) and offset (time of disappearance) of the actions depending on the speech-rate (Fig. 6). If the speech-rate is slow, the movement will start on the beginning of the syllable, otherwise on the beginning of the word.

Knowing the location in time of these actions (points A and D in Fig. 7), the program computes their apex (time of maintenance of the action) (B and C). To find the starting point of the apex (B) it looks for the closest phoneme whose duration from the starting point of the action (A) is equal to the onset value. It does the same to find the ending point of the apex (C). It scans the sentence backward to look for the phoneme whose duration from the end of the action (D) is equal to the offset value. We can notice that the two points defining the apex will vary with the chosen affect since each affect has different onset and offset values.

Then the program computes all the blinks occurring as conversational signals (Fig. 8). The internal struc-

add blink on accent
final occurrence is affect-dependent

blink-close=begin on accented syllable
blink-start=depend on speech-rate
find the two other parameters

Figure 8: Algorithm for voluntary blinks occurring on accents

pitch accent	**phrase tone**	**boundary tone**	**emphasis**
RM (Rapid Movement)	OM (Ordinary Movement)	SM (Slow Movement)	RM (Rapid Movement)

Table 4: Head movements on accented segments

ture of an eye blink (closure time, time it remains closed, time of aperture) is synchronized with the articulation [Condon and Osgton, 1971]. To find such a timing, the program parses the utterance and looks at the phonemes which have the closest timing to the average speed of an eye blink. When occurring on an accent, for fast speech-rate, the blink starts at the beginning of the word and closes on the accented phoneme. For slow speech-rate the starting point is on the beginning of the syllable and the closing time remains the same.

Head movements are computed depending on the type of accents. The program scans the utterance and assigns the corresponding head movements to the considered segments (Table 4).

4.4 Punctuators

A boundary point (such as a comma) is underlined by slow movement and a final pause coincides with stillness [Hadar et al., 1983]. When occurring on an hesitation pause the type of accent varies with the affect (Table 15). But the type varies also with the type of pause. A question is often indicated by raised eyebrows, especially when the question is not verbalized, while a period might be marked by a frown.

4.4.1 Algorithm

When an action occurs on a pause, the starting and ending points of the apex coincide with the beginning and the end of the pause. The other values of the movement (onset, offset) are computed as for the conversational signals (Fig. 9).

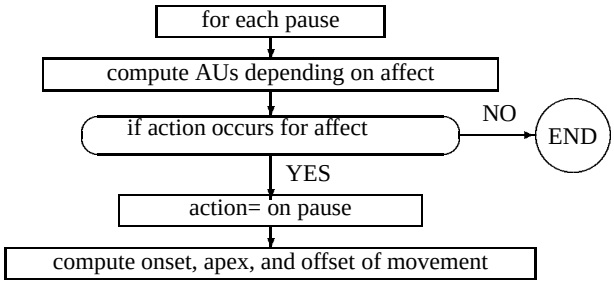

Figure 9: Algorithm of eyebrow actions

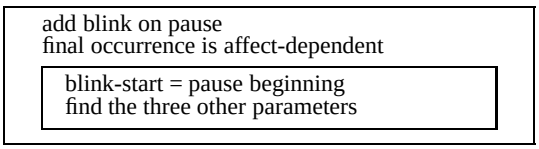

Figure 10: Algorithm for voluntary blinks occurring on pauses

Blinks occurring on a pause (Table 11) start and end with the pause if there is enough time for the blinks to be performed, otherwise the blinks are synchronized with the word following the pause (Fig. 10).

Finally the program computes the head movements. For each type of pause movement type is assigned (Table 5). The program runs through the utterance and computes the corresponding head action for each pause.

4.5 Regulators

Regulators correspond mainly to head and eye movements. Fig. 11 gives the algorithm. If the utterance is a question, the head is prepositioned to turn toward the listener (Speaker-Turn-Signal) and will raise up at the end of the utterance (Table 6). If the utterance is a statement, the head is prepositioned to look away from the listener (Speaker-State-Signal) and will look down at the end of the utterance. Speaker-Within-Turn appears between two intonational clauses

fluent pause	**hesitation pause**	**silence**
RM	OM	OM followed by SM (direction: up for ‘?’, down for ‘.’)

Table 5: Head movements on pauses

Speaker-State-Signal	**Speaker-Within-Turn**	**Speaker-Continuation-Signal**
POS on first word of the phrase direction: away	POS on pause direction: toward	POS on next word direction: away

Table 6: Summary of turn-taking system

belonging to same sentence. Speaker-Within-Turn (speaker turns his/her head towards listener) is normally followed by a Speaker-Continuation-Turn (turns head away from listener). But, if the utterance is followed by another utterance, only Speaker-Within-Turn occurs.

The program first scans the utterance and computes all the potential head movements. If no movement is specified for some phonemic items the head is forced to go back to its starting position with OM to insure head stability. This first step computes the various timings of head movements.

The second step is to find the direction of each motion. RM is associated with vertical motion of the head; the direction is changed at each cycle (if it was up at the previous cycle, it is down for the current one, and vice-versa). (A cycle corresponds to a phonemic segment for slow speech-rate or to a syllable for fast speech-rate.) The other types of head motion have sideways motions. The direction is sustained during each particular motion but it is inverted at its end for the next motion.

For each phoneme, the final head position is obtained by adding all the co-occurring one: if a stressed element occurs during a Speaker-State-Signal, the head will nod while turning away from the listener. Therefore, the two motions coincide if the corresponding supra-segmental parameters coexist. The program also applies an adjustment for co-occurring actions in the same fashion as for spatial constraints for lip shapes. At the end of the computation a specific and unique pattern is found for each clause. But we note in the case of successive utterances having in common certain features of intonational tunes, or part of the same topic, that they share the same type of head motion [Hadar et al., 1983].

The second part of the computation of regulators is the eye motion. For a first approach we decided to implement a simple simulation of eye movements (Fig. 12). We assume that when an action occurs (mutual gaze, breaking eye

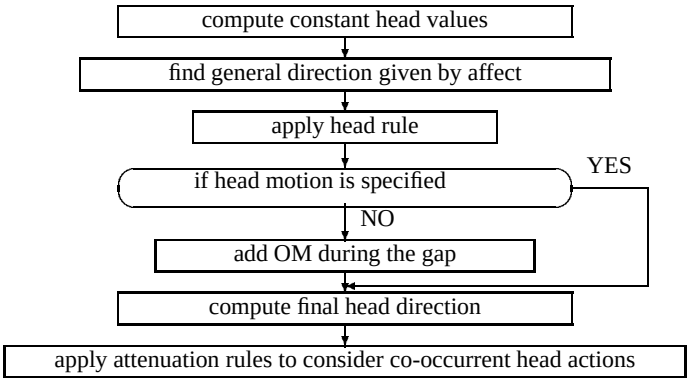

Figure 11: Algorithm of head motion

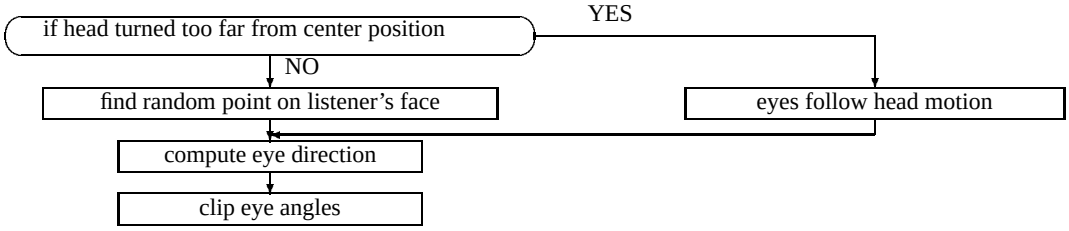

Figure 12: Algorithm of eye movements

contact, and so on) head and eyes follow the same behavior. This means we interpret eye position as head position. Otherwise, when no action is occurring, the eyes will scan in a saccadic motion the listener's face in a random way (e.g., 70% of the time is spent on the eyes, and 30% on the mouth). Nevertheless, the eyes are forced to follow head motion when the head position has passed a certain angle.

Pupil changes occur during affectual experiences. Pupil dilation is followed by pupil constriction during "happiness" and "anger" and remains dilated during "fear" and "sadness" [Hess, 1975]. Depending on the affect, eye openness also varies. For "surprise" and "fear" the eyes are wide open; they are partially closed during "sadness", "disgust" and "happiness" [Collier, 1985] (Table 10).

Many gaze patterns, such as mutual gaze or gaze avoidance, imply the existence of a listener/partner. This is beyond the scope of the present study. Eye movements not linked directly with speech (for example, when due to an external event) are not considered here either.

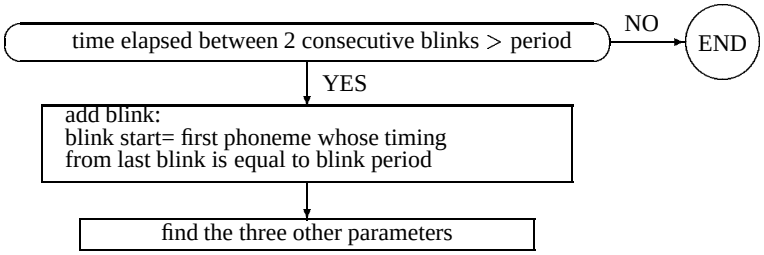

Figure 13: Algorithm for periodic blinks

4.6 Manipulators

Currently we consider only blinks as manipulators. Blinks should occur periodically. After having computed all the blinks appearing on accents and on pauses, the program adds any necessary ones (Fig. 13). It looks for the phoneme whose duration from the last blink is the closest to the period of blink occurrence. The remaining parts of the blink are computed as for voluntary blinks.

5 Some Examples

The following examples were run on the basis of measurements from read speech, rather than from the rule-based synthesis. We consider the utterance:

'Fred wanted to try to produce a play.'

This sentence was recorded with various speech-rates (fast and slow), different intonational patterns, and each of the six affects. In the following sections we analyze how these variations affect the final animation.

We choose to work with affects: specification of their facial expressions and other parameters used in our system can be found in the Appendix B. These values are not intended to be representative of every affect; rather, they demonstrate what our system can do.

5.1 Variation of the Speech-Rate

Let us see how the speech-rate parameter affects the computation of the lip shapes. The affect is neutral. The intonational pattern is:

speech-rate	Fred wan	/TT/	/IH/	/DD/	...
slow		forward rule less intens.	cluster: back	backward rule less intens.	
fast		forward rule less intens.	change cluster: relaxmoderate	backward rule less intens.	
speech-rate	to try to prod	/YU/	/SS/	/AX/	play
slow		cluster: narrow pucker	forward rule less intens.		
fast		same cluster: less intens.	forward rule less intens.	forward rule less intens.	

Table 7: The variations of speech-rate over lip shapes

(FRED WANTED TO TRY TO PRODUCE A PLAY).

(H*LL%)

Table 7 summarizes the lip shapes for slow and fast speech-rate. The results are different for both examples.

- Some speech postures differ only in the intensity of their actions. For slow speech-rate, the intensity is higher since the mouth has more time to perform each movement.
- Some other speech postures differ in their list of **AU**s (Table 2). For slow speech-rate, the segment /IH/ has its characteristic shape, an extension of the lip corners occurring with **AU11** and **AU12**, and a slight opening of the mouth is visible with **AU25**. When the lips extend, the outer part of the lips is less apparent: this is performed with **AU28**. For fast speech-rate, there are fewer different cluster, in fact, most of them gather into one type: moderate opening of the mouth performed by the **AU25**.
- From Table 7, we can notice that the forward rule is applied across word boundaries [Kent and Minifie, 1977]; the segment /AX/, even though it is another word, receives the same list of **AU**s as the segment /YU/ of the previous word.

context	Fred	/WW/	/AA/	/NN/	/TT/	/IH/	/DD/
1		onset	apex			offset	
2		onset	apex				
context	to	try to produce a	<silence>	play	<fluent pause>		
1							
2	offset		onset+apex	apex	apex+offset		

Table 8: The variations of eyebrow movements depending on context

5.2 Variations of the Intonational Pattern

We vary the type and placement of the accents in the utterance, together with that of intonational boundaries. The affect is neutral in all the examples considered here. We look at how the determinants of facial expressions (conversational signals and punctuators) vary along with the intonation.

We consider two contexts for the same utterance determining different partitions of the discourse information in the sentence.

Context 1:

Question: Why did Fred try to produce a play?

Answer: (FRED WANTED TO TRY TO PRODUCE A PLAY).

Accent : (H* LL%)

Context 2:

Question: Why did Fred try to produce a play rather than a movie?

Answer: (FRED WANTED)(TO TRY TO PRODUCE A PLAY).

Accent : (H* L)(L+H*LL%)

The answer in context 1 is pronounced with a slow speech-rate, but a fast speech-rate in context 2. In context 2 a pause appears between the words *'a'* and *'play'*.

In context 1 (Table 8), the program outputs an eyebrow raise (**AU1 + AU2**) on the accented syllable of the word *'wanted'*. The movement continues with decreasing intensity through the next two segments. Indeed the timing of the two succeeding segments is less than the offset.

In context 2 the speech-rate is fast, therefore the eyebrow action occurs on the whole word. The movement is also propagated to the segment of the next word *'to'*. The pause between the words *'a'* and *'play'* coincides with an eyebrow action. This movement is used as a punctuator. Since it is followed by a conversational signal for the accented segment and another punctuator (fluent pause), the movement of the eyebrows continues over the word and the end of utterance.

Finally we notice that the two examples differ not only in the number of brow movements occurring on the word *'play'* and on the pause occurring between the words *'a'* and *'play'* in context 2, but also on the time an action lasts relative to the speech-rate.

Similar considerations apply on head and eye motions and on blink occurrence.

5.3 Variations over the Affectual Parameters

In this section, we examine how the affectual parameters vary the occurrence of facial action. The context for each is the same, so that the utterances we are working on have identical intonational patterns.

affect	Fred	wanted	to try to produce	pause	a	pause	play	pause
anger		AU1+2					AU1+2	AU4
disgust		AU1+2				AU9+10		AU4
fear		AU1+2		AU1+2+4			AU1+2	
happiness		AU1+2					AU1+2	AU4
sadness		AU1+2						
surprise		AU1+2		AU1+2				AU4

AU1: inner brow raiser AU2: outer brow raiser AU4: brow lowerer

Table 9: Conversational Signals and Punctuator Occurrences for each Affect

Question: I know that FRED ended up trying to produce a MOVIE.

But what did Fred WANT to try to produce?

Answer: (FRED WANTED TO TRY TO PRODUCE)(A PLAY).

Accent : (L+H* LH%)(H*LL%)

For each affectual parameter, the existing pauses in each spoken utterance is given in Table 16, and the list of **AU**s corresponding to the type of pauses is computed (Table 10). The program also computes the constants defining the onset and offset of a movement and the constants specific to head motion. The “sadness” affect has very low values for its head motion, so the head moves a little and slowly. For “happiness”, it is the opposite. Movements appear and disappear very abruptly when the affect is “anger” or “fear”, but more smoothly for “sadness”.

The final appearance of conversational signals and punctuators is affect-dependent (Table 9). The facial expressions associated with punctuators can vary with an affect as being part of the characteristic expression of an affect.

6 Other Features of the System

In order to be able to refine some of the rules of occurrence of facial actions, the parameter set-up is interactive. There is a lack of empirical information on when an accent or other intonational components are accompanied by a facial action. The user can specify type and relative time of occurrence of a facial action as well as apex, onset and offset values [Pelachaud et al., 1993].

Given a context, an affect is associated with a particular facial expression. But there may be some cultural variability - some cultures forbid direct gaze while others find gaze aversion an offense; mourning in some cultures is over-acted while in others it should be masked by a smile. Display Rules [Ekman, 1979] refer to this problem of who can show which affect to whom and when. We have not taken them into consideration for automatic procedures since they are very difficult to handle and very little information is available. They may affect expression in various ways. They can amplify, de-amplify, or neutralize an expression [Ekman, 1979]; they may blend with other expressions or may even be masked by other facial expressions. The user can simulate these effects through the use of different functions. Amplify, de-amplify and neutralize affect the intensity of the facial changes. The blend is done by summing the effects together. Masking an expression A by another one B affects the timing of the parameters of B while some features of A remain [Pelachaud, 1993].

As we have seen, gaze behavior plays a great role in communication settings but some cultural differences are found in the amount of gaze allowed during a social encounter [Harper et al., 1978]. Gaze can be used to establish power relationships or it can act as a signal of liking [Argyle and Cook, 1976]. Personality and context are also parameters of the visual pattern. A submissive person more frequently breaks eye contact than a dominant one. Moreover, during a dialog situation the listener moves in synchrony with the speaker [Condon and Osgton, 1971]. To obtain a refinement of the computation of eye movements we add in the regulator group the auditor feedback determinant [Duncan, 1974] and consider various parameters (such as maximum gaze length, percentage of mutual gaze) [Cassell et al., 1994].

Another extension to the system is the integration of emblems and emotional emblems [Ekman, 1979]. The last ones are expressed by employing parts of the corresponding affect they refer to, while the first ones are used to replace and repeat verbal elements. Most of the time both are intentional, deliberate actions used to communicate. In general

they are produced consciously and are driven by the semantics of the utterance. They are conventionalized. Their encoding and decoding share a lot of appearances and meanings [Ekman, 1976]. Since they are discourse-driven their appearance is entered by the user. This is done by creating a library of possible emblems; Efron gives a large list of them [Efron, 1972] and Ekman proposes a set of words which have a corresponding emblem. Nevertheless the user can build his/her own emblems and add them to the library [Pelachaud, 1993]. When lying, the timing of an expression changes. An expression may appear too early or too late, too fast or too slow. Thus by having access to the value of the onset, apex and offset of an action, one can modify them in order to simulate lies.

Considering tongue movement when lip shapes are computed helps to make unambiguous obscure movements of some phonemic segments (since some speech postures are only differentiated by their tongue motions [Jeffers and Barley, 1971], [Kent and Minifie, 1977]) [Pelachaud et al., 1994].

7 Conclusion

We have proposed a method of characterizing any facial movements by separating them into phonemic, intonational, informational and affectual determinants. We believe that no previous computational model has taken into account all of these factors. Separately computing each determinant of facial expressions offers better control over facial animation. We are particularly interested in the facial actions which punctuate speech, their meaning, and their type and time of occurrence. An important factor in lip movement is the notion of coarticulation where temporal and spatial constraints over muscle actions are considered. The coordination of these various facial motions with the intonation is done completely automatically, by rule. Moreover, this method allows us to define various and individualized speaker characteristics by specifying particular sets of type and timing parameters for the facial actions [Calvert, 1990], [Ekman, 1979], [Moravetz, 1989], [Unuma and Takeuchi, 1991]. While the examples here are determined by measurement from real speech, we have reported a somewhat narrower range of examples generated entirely by rule from machine-generated semantic representations via a speech synthesizer [Cassell et al., 1994], [Prevost and Steedman, 1994], [Prevost and Steedman, to appear]. Our model can be expected to help further research of human communicative faculties via automatically synthesized animation. In particular, it offers to linguists and

cognitive scientists a tool to analyze, manipulate and integrate several different determinants of communication. Since our program allows the user to switch each determinant on and off, the function and the information that each of them provides can be analyzed. We expect by this mean to further refine the rather simplified theory of discourse information that we have assumed here. By providing a model of coarticulation, we hope that the program may eventually help to enhance speechreading techniques by providing hearing-impaired persons with a controllable animation capable of demonstrating the various effects on a phoneme of speech-rate and surrounding context.

8 Acknowledgments

We would like to thank Steve Platt for his facial model and for very useful comments. We would like to thank also Soetjianto and Khairol Yussof who have improved the facial model. We are also very grateful to Jean Griffin, Francisco Azuola and Mike Edwards who developed part of the animation software. All the work related to the voice synthesizer, speech and intonation was done by Scott Prevost. We are very grateful to him. Finally, we would like to thank all the members of the graphics laboratory, especially Cary Phillips and Jianmin Zhao for their helpful comments.

This research is partially supported by NSF IRI91-17110, CISE Grant CDA88-22719, ILI Grant USE-9152503, and the ARO AI Center of Excellence at the University of Pennsylvania through Grant DAAL03-89-C-0031.

9 Appendices

A Example of lip shapes

We run the program twice; once without using the coarticulation rules and once using them. In both case the speech-rate is slow. The results are for the string 'pref' of 'prefers' in 'Julia prefers popcorn'[6] (see also Fig. 14):

- no coarticulation rules applied:

```
Name of sentence PP
```

[6]The **AU**s definition is given in the caption of Table 10

```
name of AU = AU24, intensity = 0.800000
name of AU = AU23, intensity = 0.800000
Name of sentence RR
name of AU = AU11, intensity = 0.375000
name of AU = AU12, intensity = 0.225000
name of AU = AU25, intensity = 0.225000
name of AU = AU28, intensity = 0.150000
Name of sentence AH
name of AU = AU11, intensity = 0.500000
name of AU = AU12, intensity = 0.300000
name of AU = AU25, intensity = 0.300000
name of AU = AU28, intensity = 0.200000
Name of sentence FF
name of AU = AU17, intensity = 0.700000
name of AU = AU28L, intensity = 0.800000
```

- coarticulation rules applied:

```
Name of sentence PP
name of AU = AU24, intensity = 0.760000
name of AU = AU23, intensity = 0.760000
Name of sentence RR
name of AU = AU11, intensity = 0.300000
name of AU = AU12, intensity = 0.180000
name of AU = AU25, intensity = 0.180000
name of AU = AU28, intensity = 0.120000
name of AU = AU23, intensity = 0.455550
name of AU = AU24, intensity = 0.455550
Name of sentence AH
```

```
name of AU = AU11, intensity = 0.300000
name of AU = AU12, intensity = 0.180000
name of AU = AU25, intensity = 0.180000
name of AU = AU28, intensity = 0.200000
name of AU = AU17, intensity = 0.103906
name of AU = AU23, intensity = 0.227775
name of AU = AU24, intensity = 0.227775
Name of sentence FF
name of AU = AU17, intensity = 0.665000
name of AU = AU28L, intensity = 0.760000
```

In the first case, /RR/ receives the same list of **AU**s as /AH/ due to the forward rule.

In the second case, the coarticulation rules are applied on the results from the first case:

geometrical constraint rules : the intensities of successive antagonist **AU**s have decreased. These intensities are re-adjusted using the table of comparisons between **AU**s. As an example, the intensities of **AU23** and **AU24** for /PP/ go from 0.8 to 0.76

temporal constraint rules : /RR/ and /AH/ shows remains of lip actions of /PP/ since their duration altogether is less than the relaxing time of a muscle and the lip shape of /PP/ is propagated until /AH/ is pronounced. **AU23** and **AU24** appear for /RR/ and /AH/ with decreasing intensities. Identically, **AU17** appears for /AH/ due to the contraction time involved in the pronounciation of /FF/.

B Affect

Our modelling of affect is not meaning-based and is confined to those affects characteristically displayed on the face and through the voice. Body postures indicate essentially the intensity of affect [Collier, 1985]. “Happiness” is recognized by smile (corners of the mouth are drawn back and up) and raised cheeks creating wrinkles around the eyes. “Disgust” is characterized by nose wrinkling and raised upper lip. Six affects (“anger”, “disgust”, “fear”, “happiness”, “sadness”

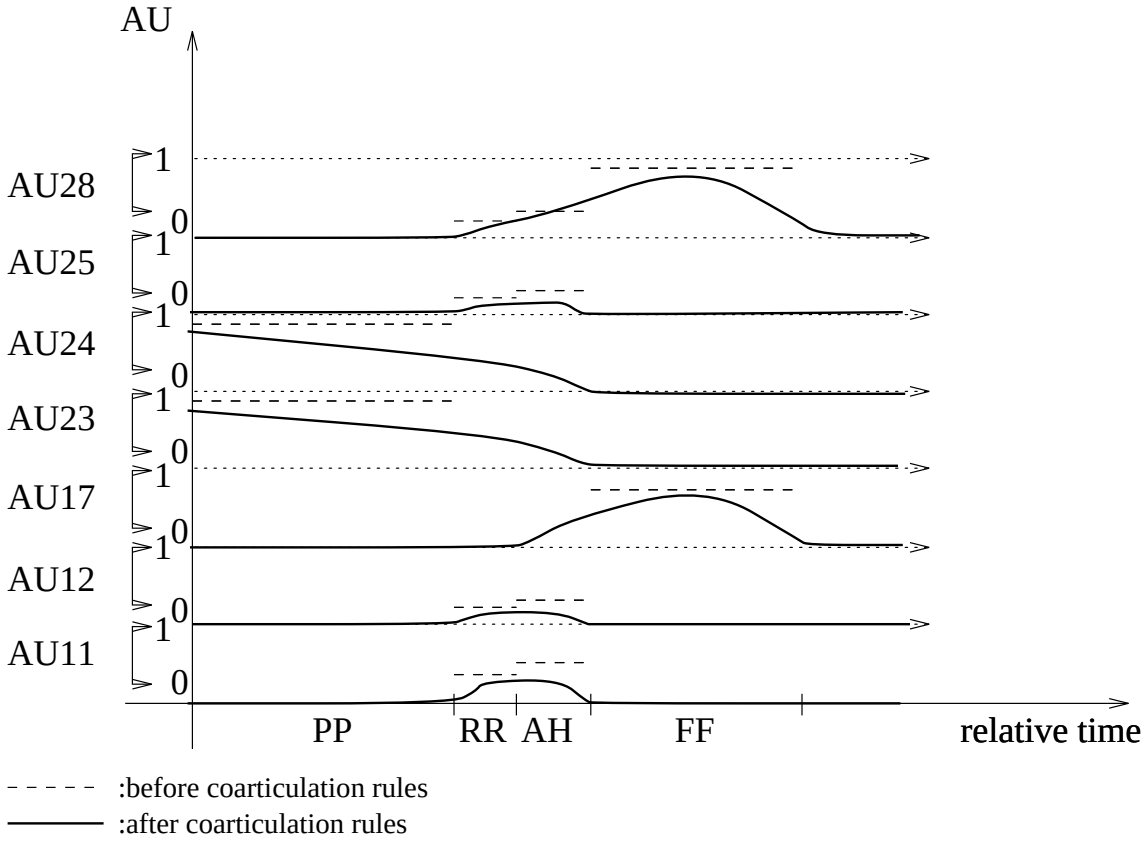

Figure 14: List of **AU**s for the lip shapes of ‘pref’ from ‘Julia prefers popcorn.’ The curves show the smoothing effect of coarticulation.

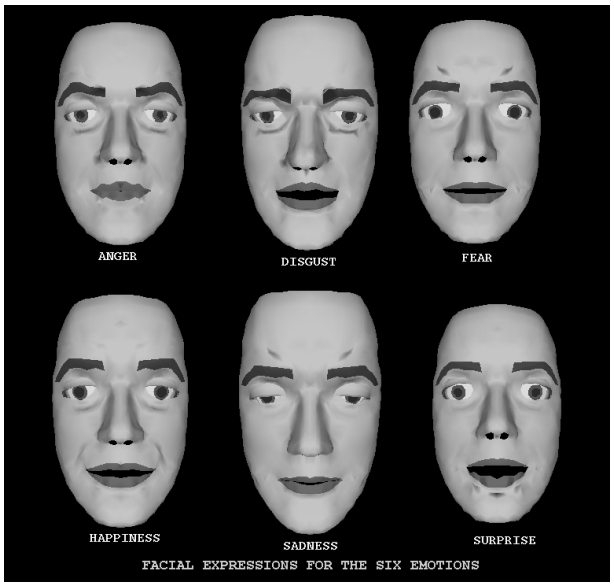

Figure 15: Facial Expressions for the Six Affects

and "surprise") have been claimed to have universal facial expressions [Ekman, 1979] corresponding to prototypes (Fig. 15). We have chosen to study these.

A person may feel an affect with different strength. If an affect is felt very lightly, not every facial movement corresponding to the affect will be visibly displayed, just the minimum requirement will appear with very little action; while in the case of very high intensity, the facial expression of the affect will be extreme. Thus, the intensity of an affect is a very important parameter that is reflected by the system.

To the extent that these expressions are conventionalized and under voluntary control, they resemble speech acts. Like speech acts, they can be used "indirectly". We leave the specification of these parameters to the animator.

B.1 Choice of the Rules for Affects

From the definition of affects presented in a previous section, and information stated in [Argyle and Cook, 1976], [Collier, 1985], [Ekman, 1979], we established the corresponding facial expression of affects in term of a set of **AU**s, eye openness, pupil size and types of movements (Fig. 15). "Sadness" is the least active affect and shows a few actions appearing and disappearing smoothly on the face. On the other hand an "angry" person moves in an aggressive

and rapid manner punctuating his/her speech with brusque facial actions. The amount and type of movements vary with the level of arousal [Collier, 1985]. For example, “fear” has varying types of action depending on whether the intensity of affect is low or high. “Fear” is expressed only with the eyes for low intensity, while for intense affect, the entire face is involved. These findings are summarized for each affect in Tables 10 and 11.

We compute the intensity of a facial action of affect proportionally to the intensity of the affect. Each action is defined by two pre-established constants, a minimal and a maximal one. Its actual value is a linear function of the intensity of affect between the two constants:

$$\textit{intensity-AU} = \textit{minimum-AU} * \textit{intensity-affect} + \textit{maximum-AU} * (1 - \textit{intensity-affect}).$$

It is possible to choose some affect not belonging to the basic set we considered. The user just needs to change the set of rules defining the occurrence of facial actions and the list of **AU**s for the facial expressions of the chosen affect. Having specified this new information, the program offers procedures that compute the exact location and manner of appearance (such as the onset, offset values, and so on) of the facial actions. For example, the user has to specify if blinks should occur on which type of pauses, on accents, and so on; then actual procedures can perform the computation.

C Examples

The sentence *(Fred wanted to try to produce) (a play)* was recorded, with the intonations indicated by one of the authors, with the six affects (“anger”, “disgust”, “fear”, “happiness”, “sadness” and “surprise”). From the spoken utterance we extracted timing including pauses. Table 16 gathers these values.

name	list of AUs	eye openness	pupil size	basic head position	active
anger	AU2+4+5+10+ AU23+24+28	small	dilate than constrict	forward	0.7
disgust	AU4+9+10+17	-	-	backward and up	-0.3
fear	AU1+2+4+5+7 (low intensity) AU1+2+4+5+7+ 15+20+25 (high int.)	wide	dilate	backward	0.8
happiness	AU1+2+5+6+10+11+ 12+13+20+25+28R	-	dilate then constrict	-	0.4
sadness	AU1+4+7+15	small	dilate	downward	-0.8
surprise	AU1+2+5+26	wide	-	backward	0.6

Our notation is:

active: For affect there is a corresponding activation level standing for the number and type (fast vs. slow) of facial, head and eye actions shown by the speaker.

AU1: inner brow raiser	AU2: outer brow raiser	AU4: brow lowerer
AU5: upper lid raiser	AU6: cheek raiser and lid compresser	AU7: lid tightener
AU9: nose wrinkler	AU10: upper lip raiser	
AU11: nasolabial furrow deepener	AU12: lip corner puller	AU13: sharp lip puller
AU15: lip corner depressor	AU16: lower lip depressor	AU17: chin raiser
AU20: lip stretcher	AU23: lip tightener	AU24: lip presser
AU25: lips part	AU26: jaw drop	AU28: lip suck

Table 10: Occurrence of facial actions

name	conversational signal			punctuator				regulator
						other		
	blink	**RM**	**othr**	**blink**	**OM, ST**	**occur.**	**type**	POS
neutral	yes	normal	every	si+fl	normal	fl+si	raised brow	normal
anger	yes	less	every	fl+he	less	fl+he	frown	more
disgust	no	less	first	si+fl	less	fl+si	nose wrinkling	less
fear	yes	more	every	si+he	more	he+si	brow fear	less
happiness	yes	more	every	si+fl	more	fl+he	smile	more
sadness	no	less	first	si+fl+he	less	-	-	less
surprise	no	more	first	si+fl	more	fl	raised brow	less

Pauses are classified by their function [Dittmann, 1974]:

fl: fluent pause occurs at boundary points

he: hesitation pause corresponds to false start, word finding problem

si: specified silence marked by the speaker

POS, RM, OM, SM: head movements (see explanation in Table 1)

every: occurrence on every accented segment

first: occurrence on the first accented segment of the utterance

Table 11: Occurrence of facial actions

head movement	anger	disgust	fear	happiness	sadness	surprise
POS	$/2$	$*1$	$/2$	$*2$	$/2$	$*1$
RM	$/1.5$	$/1.5$	$*2$	$*2$	$/2$	$*2$
OM	$/1.5$	$/1.5$	$*2$	$*2$	$/2$	$*1$
SM	$/1.5$	$/1.5$	$*2$	$*2$	$/2$	$*1$

Table 12: Scaling factors to compute head movements for each affect

anger	**disgust**	**fear**	**happiness**	**sadness**	**surprise**
shorter	no change	shorter	shorter	longer	no change

Table 13: The period of occurrence of blinks

	neutral	**anger**	**disgust**	**fear**	**happiness**	**sadness**	**surprise**
accent	yes	yes	no	yes	yes	no	no
emphasis	yes	yes	no	yes	no	no	no

Table 14: Occurrence of blinks as conversational signal

affect	**AU**
neutral	AU1+2
anger	AU4
disgust	AU9+10
fear	AU1+2+4
happiness	AU12
sadness	AU1+4
surprise	AU1+2

Table 15: **AU**s occurring during pauses

affect	Fred wanted to try to produce		a		play	
anger				<silence>		<fluent pause>
disgust				<silence>		<fluent pause>
fear		<fluent pause>				<fluent pause>
happiness						<fluent pause>
sadness		<fluent pause>				<fluent pause>
surprise		<fluent pause>				<fluent pause>

Table 16: The occurrence of pauses for each affect

References

[Argyle and Cook, 1976] Argyle, M. and Cook, M. (1976). *Gaze and Mutual gaze*. Cambridge University Press.

[Badler et al., 1993] Badler, N.I., Phillips, C., and Webber, B. (1993). *Simulating Humans: Computer Graphics Animation and Control*. Oxford University Press.

[Benoit et al., 1990] Benoit, C., Lallouache, T., Mohamedi, T., Tseva, A., and Abry, C. (1990). Nineteen (+- two) french *visemes* for visual speech synthesis. In *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans. ESCA.

[Bolinger, 1986] Bolinger, D. (1986). *Intonation and its Part*. Stanford University Press.

[Bourne, 1973] Bourne, G. (1973). *Structure and Function of Muscle*, volume III, Physiology and Biochemistry. Academic Press, second edition edition.

[Brooke, 1990] Brooke, N. (1990). Computer graphics synthesis of talking faces. In *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans. ESCA.

[Cahn, 1989] Cahn, J. (1989). Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

[Calvert, 1990] Calvert, T. (1990). Composition of realistic animation sequences for multiple human figures. In Badler, N.I., Barsky, B., and Zeltzer, D., editors, *Making Them Move: Mechanics, Control, and Animation of Articulated Figures*. Morgan Kaufmann Publishers Inc.

[Cassell et al., 1994] Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graghics Annual Conferences Series*, pages 413–420.

[Cathiard et al., 1991] Cathiard, M., Tiberghien, G., Cirot-Tseva, A., Lallouache, M., and Escudier, P. (1991). Visual perception of anticipatory rounding during acoustic pauses: A cross-language study. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, pages 50–53, Aix-en-Provence, France.

[Cohen and Massaro, 1993] Cohen, M. and Massaro, D. (1993). Modeling coarticulation in synthetic visual speech. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '93*. Springer-Verlag.

[Collier, 1985] Collier, G. (1985). *Emotional Expression*. Lawrence Erlbaum Associates.

[Condon and Osgton, 1971] Condon, W. and Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. In Horton, D. and Jenkins, J., editors, *The perception of Language*, pages 150–184. Academic Press.

[DEC, 1985] Dectalk (1985). *Dectalk and DTC03 Text-To-Speech System Owner's*. Digital Equipment Corporation.

[Dittmann, 1974] Dittmann, A. (1974). The body movement-speech rhythm relationship as a cue to speech encoding. In Weitz, editor, *Nonverbal Communication*. Oxford University Press.

[Duncan, 1974] Duncan, S. (1974). Some signals and rules for taking speaking turns in conversations. In Weitz, editor, *Nonverbal Communication*. Oxford University Press.

[Efron, 1972] Efron, D. (1972). *Gesture, Race, and Culture*. The Hague, Mouton.

[Ekman, 1976] Ekman, P. (1976). Movements with precise meanings. *The Journal of Communication*, 26.

[Ekman, 1979] Ekman, P. (1979). About brows: emotional and conversational signals. In von Cranach, M., Foppa, K., Lepenies, W., and Ploog, D., editors, *Human ethology: claims and limits of a new disipline: contributions to the Colloquium*, pages 169–248. Cambridge University Press, Cambridge, England; New-York.

[Ekman and Friesen, 1978] Ekman, P. and Friesen, W. (1978). *Facial Action Coding System*. Consulting Psychologists Press, Inc.

[Emmett, 1985] Emmett, A. (1985). Digital portfolio: Tony de peltrie. *Computer Graphics World*, 8(10):72–77.

[Essa and Pentland, 1994] Essa, I. and Pentland, A. (1994). A vision system for observing and extracting facial action parameters. *Proceedings of Computer Vision and Pattern Recognition (CVPR 94)*, pages 76–83.

[Hadar et al., 1983] Hadar, U., Steiner, T., Grant, E., and Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46.

[Halliday, 1967] Halliday, M. (1967). *Intonation and Grammar in British English*. Mouton, The Hague.

[Harper et al., 1978] Harper, R., Wiens, A., and Matarazzo, J. (1978). *Nonverbal Communication: The State of the Art*. J. Wiley and Sons, New York.

[Hess, 1975] Hess, E. (1975). The role of the pupil size in communication. *Scientific American*, pages 113–119.

[Hill et al., 1988] Hill, D., Pearce, A., and Wyvill, B. (1988). Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer*, 3:277–289.

[Hirschberg and Pierrehumbert, 1986] Hirschberg, J. and Pierrehumbert, J. (1986). The intonational structuring of discourse. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144.

[Isard and Pearson, 1988] Isard, S. and Pearson, M. (1988). A repertoire of British English intonation contours for synthetic speech. In *Proceedings of Speech '88, 7th FASE Symposium*, pages 1233–1240, Edinburgh.

[Jeffers and Barley, 1971] Jeffers, J. and Barley, M. (1971). *Speechreading (lipreading)*. C.C. Thomas.

[Kalra et al., 1991] Kalra, P., Mangili, A., Magnenat-Thalmann, N., and Thalmann, D. (1991). Smile: A multilayered facial animation system. In Kunii, T., editor, *Modeling in Computer Graphics*. Springer-Verlag.

[Kent and Minifie, 1977] Kent, R. and Minifie, F. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, 5:115–135.

[Kleiser-Walczak, 1988] Kleiser-Walczak (1988). Sextone for president. *ACM SIGGRAPH '88 Film and Video Show*, issue 38/39. Kleiser Walczak Construction Co.

[Kurihara and Arai, 1991] Kurihara, T. and Arai, K. (1991). A transformation method for modeling and animation of the human face from photographs. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '91*, pages 45–58. Springer-Verlag.

[Lewis and Parke, 1987] Lewis, J. and Parke, F. (1987). Automated lip-synch and speech synthesis for character animation. *CHI + GI*, pages 143–147.

[Magnenat-Thalmann and Thalmann, 1987] Magnenat-Thalmann, N. and Thalmann, D. (1987). The direction of synthetic actors in the film *rendez-vous à montrèal*. *IEEE Computer Graphics and Applications*, pages 9–19.

[Moravetz, 1989] Moravetz, C. (1989). A high level approach to animating secondary human movement. Master's thesis, School of Computing Science, Simon Fraser University.

[Nahas et al., 1988] Nahas, M., Huitric, H., and Saintourens, M. (1988). Animation of b-spline figures. *The Visual Computer*, 3:272–276.

[Parke, 1982] Parke, F. (1982). Parametrized models for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–68.

[Pelachaud, 1991] Pelachaud, C. (1991). *Communication and Coarticulation in Facial Animation*. PhD thesis, Computer and Information Science Department, University of Pennsylvania, Philadelphia, Pennsylvania.

[Pelachaud, 1993] Pelachaud, C. (1993). Consideration of facial and audio channels for a facial animation system. In *SCAN '93*, Philadelphia.

[Pelachaud et al., 1991] Pelachaud, C., Badler, N.I., and Steedman, M. (1991). Linguistic issues in facial animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '91*, pages 15–30. Springer-Verlag.

[Pelachaud et al., 1994] Pelachaud, C., Badler, N.I., and Viaud, M. (1994). Final report to NSF of the standards for facial animation workshop. Technical report, NSF, University of Pennsylvania.

[Pelachaud et al., 1994] Pelachaud, C., van Overveld, C., and Seah, C. (1994). Modeling and animating the human tongue during speech production. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '94*. Springer-Verlag.

[Pelachaud et al., 1993] Pelachaud, C., Viaud, M., and Yahia, H. (1993). Rule-structured facial animation system. In *IJCAI' 93*.

[Pierrehumbert, 1980] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology. Distributed by Indiana University Linguistics Club, Bloomington, IN.

[Pierrehumbert and Hirschberg, 1990] Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*, pages 271–312. MIT Press, Cambridge, MA.

[Platt, 1985] Platt, S. (1985). *A Structural Model of the Human Face*. PhD thesis, Computer and Information Science Department, University of Pennsylvania, Philadelphia, Pennsylvania.

[Prevost and Steedman, 1994] Prevost, S. and Steedman, M. (1994). Information based intonation synthesis. In *Proceedings of the ARPA Workshop on Human Language Technology*, Princeton.

[Prevost and Steedman, to appear] Prevost, S. and Steedman, M. ((to appear)). Specifying intonation from context for speech synthesis. *Speech Communication*.

[Scherer et al., 1984] Scherer, K., Ladd, D., and Silverman, K. (1984). Vocal cues to speaker affect: testing two models. *Journal of Acoustical Society of America*, 76:1346–1356.

[Selkirk, 1984] Selkirk, E. (1984). *Phonology and Syntax*. MIT Press, Cambridge, MA.

[Silverman, 1987] Silverman, K. (1987). *The structure and processing of fundamental frequency contours*. PhD thesis, University of Cambridge.

[Steedman, 1991] Steedman, M. (1991). Structure and intonation. *Language*, 67:260–296.

[Terzopoulos and Waters, 1991] Terzopoulos, D. and Waters, K. (1991). Techniques for realistic facial modelling and animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '91*, pages 45–58. Springer-Verlag.

[Unuma and Takeuchi, 1991] Unuma, M. and Takeuchi, R. (1991). Generation of human motion with emotion. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '91*, pages 45–58. Springer-Verlag.

[Viaud and Yahia, 1992] Viaud, M. and Yahia, H. (1992). Facial animation with wrinkles. In *3nd Workshop of animation, Eurographics' 92*, Cambridge.

[Waters, 1987] Waters, K. (1987). A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 21(4):17–24.

[Williams, 1990] Williams, L. (1990). Performance-driven facial animation. *Computer Graphics*, 24(4):235–242.