CHAPTER

Facial Animation 10

Realistic facial animation is one of the most difficult tasks that a computer animator can be asked to do. Human faces are familiar to us all. Facial motion adheres to an underlying, well-defined structure but is idiosyncratic. A face has a single mechanical articulator but has a flexible covering capable of subtle expression and rapid, complex lip movements during speech.

The face is an important component in modeling a figure, because it is the main instrument for communication and for defining a figure's mood and personality. Animation of speech is especially demanding because of the requirement for audio synchronization (and, therefore, is often referred to as *lip-sync animation*). In addition, a good facial model should be capable of geometrically representing a specific person (called *conformation* by Parke [24], *static* by others, e.g., [28]).

Facial models can be used for cartooning, for realistic character animation, for telecommunications to reduce bandwidth, and for human–computer interaction (HCI). In cartooning, facial animation has to be able to convey expression and personality that is often exaggerated. In realistic character animation, the geometry and movement of the face must adhere to the constraints of realistic human anatomy. Telecommunications and HCI have the added requirement that the facial model and motion must be computationally efficient. In some applications, the model must correspond closely to a specific target individual.

In addition to the issues addressed by other animation tasks, facial animation often has the constraint of precise timing with respect to an audio track during lip-synching and some expressions. Despite its name, lip-synching involves more than just the lips; the rigid articulation of the jaw and the muscle deformation of the tongue must also be considered.

10.1 The human face

10.1.1 Anatomic structure

The human face has an underlying skeletal structure and one main skeletal articulatory component—the jaw. In addition to the jaw, the other rigid articulatory components are the eyes.

The surface of the skull is covered with muscles, most of which (at least indirectly) connect areas of the skin to positions on the skull. These muscles tug on the skin to create movement, often recognizable as expressions. Features of the skin include eyelids, mouth, and eyebrows. The skin has various expressive wrinkles such as on the forehead and around the mouth.

As with the rest of the body, muscles provide the driving force for the face (Figure 10.1, Color Plate 7). However, unlike the rest of the body, the muscles mainly move the skin into interesting

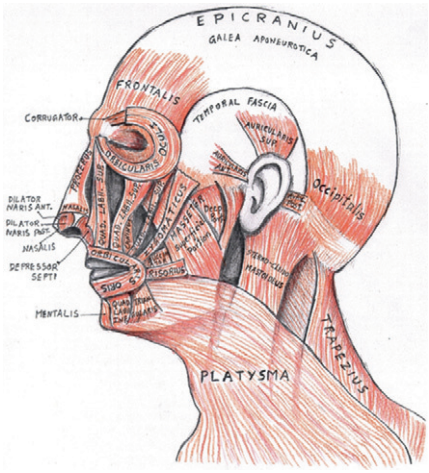

FIGURE 10.1

Muscles of the face and head [30].

(Image courtesy of Arun Somasundaram.)

FIGURE 10.2

Muscles of the face that are significant in speech and linguistic facial expressions [30].

(Image courtesy of Arun Somasundaram.)

positions, producing recognizable lip and skin configurations and, in speech, modify the sounds emanating from the labial opening. In particular, the muscles around the mouth are extremely flexible and capable of producing a variety of shapes and subtle movements. A muscle of particular importance is the orbicularis oris (Figure 10.2, muscle number 7; Color Plate 8), which wraps around the mouth and is connected to several muscles that are connected to the skull.

10.1.2 The facial action coding system

The facial action coding system (FACS) is the result of research conducted by the psychologists Ekman and Friesen [13] with the objective of deconstructing all facial expressions into a set of basic facial movements. These movements, called action units, or AUs, when considered in combinations, can be used to describe all facial expressions.

Forty-six AUs are identified in the study, and they provide a clinical basis from which to build a facial animation system. Examples of AUs are brow lowerer, inner brow raiser, wink, cheek raiser, upper lip raiser, and jaw drop. See Figure 10.3 for an example of diagrammed AUs. Given the AUs, an animator can build a facial model that is parameterized according to the motions of the AUs. A facial animation system can be built by giving a user access to a set of variables that are in one-to-one correspondence with the AUs. A parametric value controls the amount of the facial motion that results from the associated AU. By setting a variable for each AU, the user can generate all of the facial expressions analyzed by Ekman and Friesen. By using the value of the variables to interpolate the degree to which the motion is realized and by interpolating their value over time, the user can then animate the facial model. By combining the AUs in nonstandard ways, the user can also generate many truly strange expressions.

While this work is impressive and is certainly relevant to facial animation, two of its characteristics should be noted before it is used as a basis for a facial animation system. First, the FACS is meant to be descriptive of a static expression, not generative. Second, the FACS is not time based, and facial movements are analyzed only relative to a neutral pose. This means that the AUs were not designed to animate a facial model in all the ways that an animator may want to control a face. In addition, the FACS describes facial expressions, not speech. The movements for forming individual phonemes, the basic units of speech, were not specifically incorporated into the system. While the AUs provide a good starting point for describing the basic motions that must be in a facial animation system, they were never intended for this purpose.

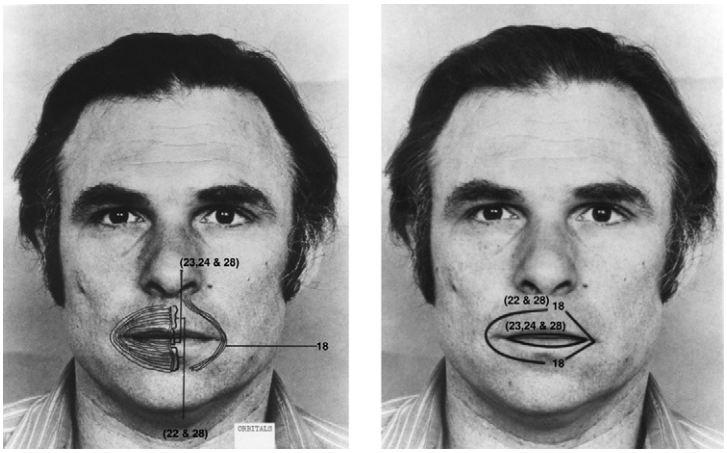

FIGURE 10.3

Three AUs of the lower face [13].

10.2 Facial models

Depending on the objective of the animation, there are various ways to model the human face. As with all animation, trade-offs include realism and computational complexity.

If a cartoon type of animation is desired, a simple geometric shape for the head (such as a sphere) coupled with the use of animated texture maps often suffices for facial animation. The eyes and mouth can be animated using a series of texture maps applied to a simple head shape (see Figure 10.4). The nose and ears may be part of the head geometry, or, simpler still, they may be incorporated into the texture map.

Stylized models of the head with only a pivot jaw and rotating spheres for eyeballs that mimic the basic mechanical motions of the human face may also be used. Eyelids can be skin-colored hemispheres that rotate to enclose the visible portion of the eyeball. The mouth can be a separate geometry positioned on the surface of the face geometry, and it can be animated independently or sequentially replaced in its entirety by a series of mouth shapes to simulate motion of deformable lips (see Figure 10.5, Color Plate 9). These approaches are analogous to techniques used in conventional hand-drawn and stop-motion animation.

For more realistic facial animation, more complex facial models are used whose surface geometry more closely corresponds to that of a real face, and the animation of these models is correspondingly more complex. For an excellent in-depth presentation of facial animation see the book by Parke and Waters [25]. An overview is given here.

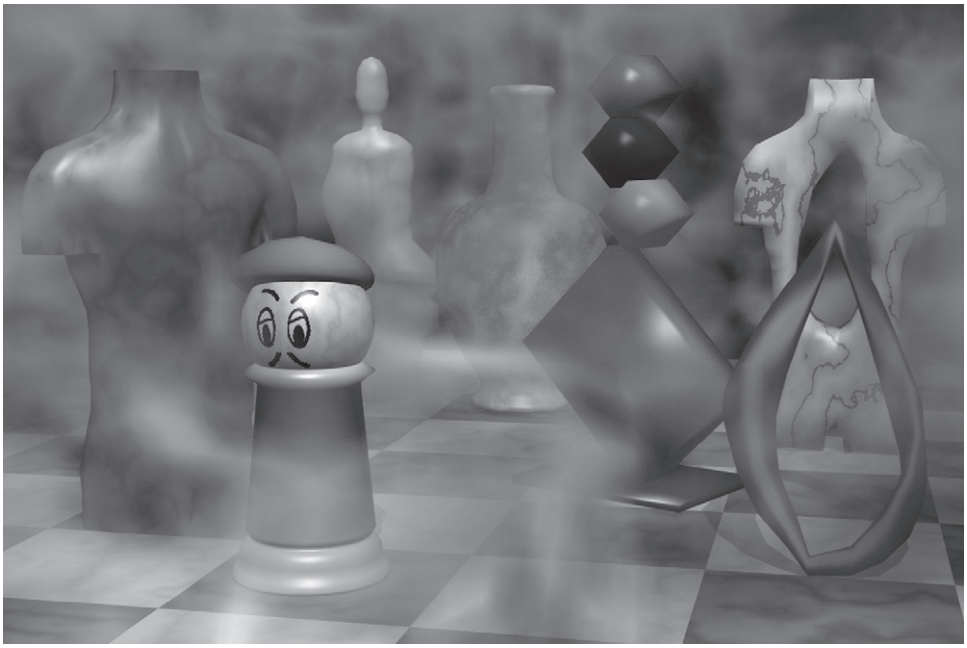

FIGURE 10.4

Texture-mapped facial animation from *Getting into Art.* (©1990 David S. Ebert [10].)

FIGURE 10.5

Simple facial model using rigid components for animation.

(Image courtesy of John Parent.)

The first problem confronting an animator in facial animation is creating the geometry of the facial model to make it suitable for animation. This in itself can be very difficult. Facial animation models vary widely from simple geometry to anatomy based. Generally, the complexity is dictated by the intended use. When deciding on the construction of the model, important factors are geometry data acquisition method, motion control and corresponding data acquisition method, rendering quality, and motion quality. The first factor concerns the method by which the actual geometry of the head of the subject or character is obtained. The second factor concerns the method by which the data describing changes to the geometry are obtained. The quality of the rendered image with respect to smoothness and surface attributes is the third concern. The final concern is the corresponding quality of the computed motion.

The model can be discussed in terms of its static properties and its dynamic properties. The statics deal with the geometry of the model in its neutral form, while the dynamics deal with the deformation of the geometry of the model during animation. Facial models are either polygonal or higher order. Polygonal models are used most often for their simplicity (e.g., [24] [25] [32] [33]); splines are chosen when a smooth surface is desired.

Polygonal models are relatively easy to create and can be deformed easily. However, the smoothness of the surface is directly related to the complexity of the model, and polygonal models are visually inferior to other methods of modeling the facial surface. Currently, data acquisition methods only sample the surface, producing discrete data, and surface fitting techniques are subsequently applied.

Spline models typically use bicubic, quadrilateral surface patches, such as Bezier or B-spline, to represent the face. While surface patches offer the advantage of low data complexity in comparison to polygonal techniques when generating smooth surfaces, they have several disadvantages when it comes to modeling an object such as the human face. With standard surface patch technology, a rectangular grid of control points is used to model the entire object. As a result it is difficult to maintain low data complexity while incorporating small details and sharp localized features, because entire rows and/or entire columns of control information must be added. Thus, a small addition to one local area of the surface to better represent a facial feature means that information has to be added across the entire surface.

Hierarchical B-splines, introduced by Forsey and Bartels [16], are a mechanism by which local detail can be added to a B-spline surface while avoiding the global modifications required by standard B-splines. Finer resolution control points are carefully laid over the coarser surface while continuity is carefully maintained. In this way, local detail can be added to a surface. The organization is hierarchical, so finer and finer detail can be added. The detail is defined relative to the coarser surface so that editing can take place at any level.

10.2.1 Creating a continuous surface model

Creating a model of a human head from scratch is not easy. Not only must the correct shape be generated, but when facial animation is the objective, the geometric elements (vertices, edges) must be placed appropriately so that the motion of the surface can be controlled precisely. If the model is dense in the number of geometric elements used, then the placement becomes less of a concern, but in relatively low resolution models it can be an issue. Of course, one approach is to use an interactive system and let the user construct the model. This is useful when the model to be constructed is a fanciful creature or a caricature or must meet some aesthetic design criteria. While this approach gives an artist the most freedom in creating a model, it requires the most skill. There are three approaches used to construct a model: refining from a low-resolution model, modifying a high-resolution simple shape, and designing the surface out from high-resolution areas.

Subdivision surfaces (e.g., [9]) use a polygonal control mesh that is refined, in the limit, to a smooth surface. The refinement can be terminated at an intermediate resolution, rendered as a polygonal mesh, vertices adjusted, and then refined again. In this manner, the designer can make changes to the general shape of the head at relatively low resolution. Figure 10.6 shows the initial stages of a subdivision-based facial design [27]. Subdivision surfaces have the advantage of being able to create local complexity without global complexity. They provide an easy-to-use, intuitive interface for developing new models, and provisions for discontinuity of arbitrary order can be accommodated [9]. However, they are difficult to interpolate to a specific dataset, which makes modeling a specific face problematic.

Alternatively, the designer may start with a relatively high-resolution simple shape such as a sphere. The designer pushes and pulls on the surface to form the general features of the face and refines areas as necessary to form the face [17] (see Figure 10.7).

Another alternative is to build the surface out from the expressive regions of the face—the mouth and eyes. This approach can be used to ensure that these regions are flexible enough to support the intended animation. See Figure 10.8 for an example of an initial surface design [29].

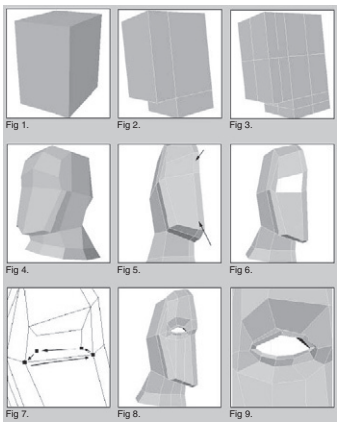

FIGURE 10.6

Early stages in facial modeling using subdivision surfaces [27].

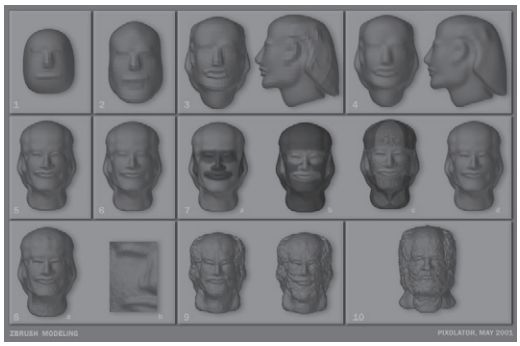

FIGURE 10.7

Facial design by pushing and pulling surface of high-resolution sphere-like shape [17].

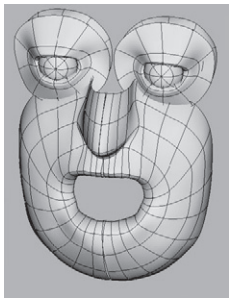

FIGURE 10.8

Facial design by initially defining the eyes and mouth regions [29].

Besides the interactive design approach, there are two main methods for creating facial models: digitization using some physical reference and modification of an existing model. The former is useful when the model of a particular person is desired; the latter is useful when animation control is already built into a generic model.

As with any model, a physical sculpture of the desired object can be generated with clay, wood, or plaster and then digitized, most often using a mechanical or magnetic digitizing device. A two-dimensional, surface-based coordinate grid can be drawn on the physical model, and the polygons can be digitized on a polygon-by-polygon basis. Post-processing can identify unique vertices, and a polygonal mesh can be easily generated. The digitization process can be fairly labor intensive when large numbers of polygons are involved, which makes using an actual person's face a bit problematic. If a model is used, this approach still requires some artistic talent to generate the physical model, but it is easy to implement at a relatively low cost if small mechanical digitizers are used.

Laser scanners use a laser to calculate distance to a model surface and can create very accurate models. They have the advantage of being able to directly digitize a person's face. The scanners sample the surface at regular intervals to create an unorganized set of surface points. The facial model can be constructed in a variety of ways. Polygonal models are most commonly generated. Scanners have the

added advantage of being able to capture color information that can be used to generate a texture map. This is particularly important with facial animation: a texture map can often cover flaws in the model and motion. Laser scanners also have drawbacks; they are expensive, bulky, and require a physical model.

Muraki [22] presents a method for fitting a blobby model (implicitly defined surface formed by summed, spherical density functions) to range data by minimizing an energy function that measures the difference between the isosurface and the range data. By splitting primitives and modifying parameters, the user can refine the isosurface to improve the fit.

Models can also be generated from photographs. This has the advantage of not requiring the presence of the physical model once the photograph has been taken, and it has applications for video conferencing and compression. While most of the photographic approaches modify an existing model by locating feature points, a common method of generating a model from scratch is to take front and side images of a face on which grid lines have been drawn (Figure 10.9). Point correspondences can be established between the two images either interactively or by locating common features automatically, and the grid in three-space can be reconstructed. Symmetry is usually assumed for the face, so only one side view is needed and only half of the front view is considered.

Modifying an existing model is a popular technique for generating a face model. Of course, someone had to first generate a generic model. But once this is done, if it is created as a parameterized model and the parameters were well designed, the model can be used to try to match a particular face, to design a face, or to generate a family of faces. In addition, the animation controls can be built into the model so that they require little or no modification of the generic model for particular instances.

One of the most often used approaches to facial animation employs a parameterized model originally created by Parke [23] [24]. The parameters for his model of the human face are divided into two categories: *conformational* and *expressive*. The conformational parameters are those that distinguish one individual's head and face from another's. The expressive parameters are those concerned with animation of an individual's face; these are discussed later. Symmetry between the sides of the face is assumed. Conformal parameters control the shape of the forehead, cheekbone, cheek hollow,

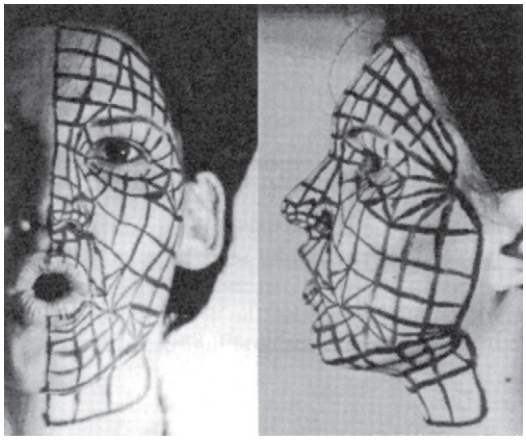

FIGURE 10.9

Photographs from which a face may be digitized [25].

chin, and neck. There are several scale distances between facial features:[1] head x, y, z; chin to mouth and chin to eye; eye to forehead; eye x and y; and widths of the jaw, cheeks, nose bridge, and nostril. Other conformal parameters translate features of the face: chin in x and z; end of nose x and z; eyebrow z. Even these are not enough to generate all possible faces, although they can be used to generate a wide variety.

Parke's model was not developed based on any anatomical principles but from the intuitions from artistic renderings of the human face. Facial anthropometric statistics and proportions can be used to constrain the facial surface to generate realistic geometries of a human head [8]. Variational techniques can then be used to create realistic facial geometry from a deformed prototype that fits the constraints. This approach is useful for generating heads for a crowd scene or a background character. It may also be useful as a possible starting point for some other character; however, the result will be influenced heavily by the prototype used.

The MPEG-4 standard proposes tools for efficient encoding of multimedia scenes. It includes a set of *facial definition parameters* (FDPs) [15] that are devoted mainly to facial animation for purposes of video teleconferencing. Figure 10.10 shows the feature points defined by the standard. Once the model is defined in this way, it can be animated by an associated set of *facial animation parameters* (FAPs) [14], also defined in the MPEG-4 standard. MPEG-4 defines 68 FAPs. The FAPs control rigid rotation of the head, eyeballs, eyelids, and mandible. Other low-level parameters indicate the translation of a corresponding feature point, with respect to its position in the neutral face, along one of the coordinate axes [7].

One other interesting approach to generating a model of a face from a generic model is fitting it to images in a video sequence [8]. While not a technique developed for animation applications, it is useful for generating a model of a face of a specific individual. A parameterized model of a face is set up in a three-dimensional viewing configuration closely matching that of the camera that produced the video images. Feature points are located on the image of the face in the video and are also located on the three-dimensional synthetic model. Camera parameters and face model parameters are then modified to more closely match the video by using the pseudoinverse of the Jacobian. (The Jacobian is the matrix of partial derivatives that relates changes in parameters to changes in measurements.) By computing the difference in the measurements between the feature points in the image and the projected feature points from the synthetic setup, the pseudoinverse of the Jacobian indicates how to change the parametric values to reduce the measurement differences.

10.2.2 Textures

Texture maps are very important in facial animation. Most objects created by computer graphics techniques have a plastic or metallic look, which, in the case of facial animation, seriously detracts from the believability of the image. Texture maps can give a facial model a much more organic look and can give the observer more visual cues when interacting with the images. The texture map can be taken directly from a person's head; however, it must be registered with the geometry. The lighting situation during digitization of the texture must also be considered.

[1] In Parke's model, the z-axis is up, the x-axis is oriented from the back of the head toward the front, and the y-axis is from the middle of the head out to the left side.

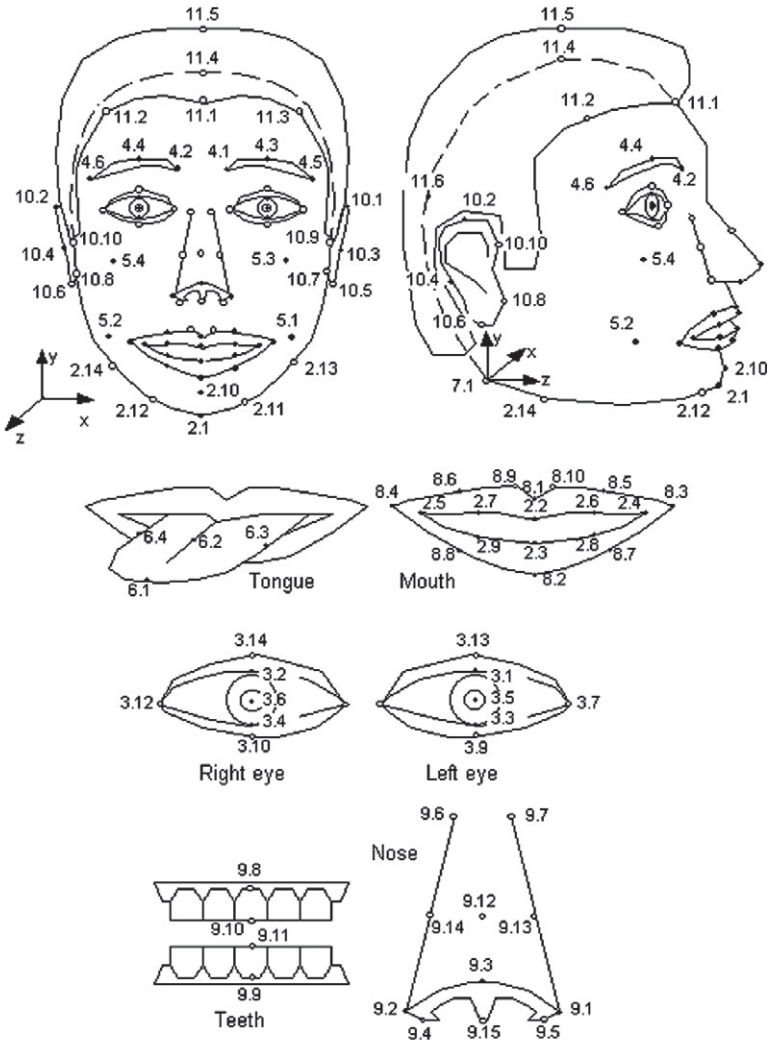

FIGURE 10.10

Feature points corresponding to the MPEG-4 FDPs [15].

Laser scanners are capable of collecting information on intensity as well as depth, resulting in a high-resolution surface with a matching high-resolution texture. However, once the face deforms, the texture no longer matches exactly. Since the scanner revolves around the model, the texture resolution is evenly spread over the head. However, places are missed where the head is self-occluding (at the ears and maybe the chin) and at the top of the head.

Texture maps can also be created from photographs by simply combining top and side views using pixel blending where the textures overlap [1]. Lighting effects must be taken into consideration, and

because the model is not captured in the same process as the texture map, registration with a model is an issue. Using a sequence of images from a video can improve the process.

10.3 Animating the face

Attempts to animate the face raise the questions: What are the primitive motions of the face? And how many degrees of freedom are there in the face?

10.3.1 Parameterized models

As introduced in the discussion of FACS, parameterizing the facial model according to primitive actions and then controlling the values of the parameters over time is one of the most common ways to implement facial animation. Abstractly, any possible or imaginable facial contortion can be considered as a point in an *n*-dimensional space of all possible facial poses. Any parameterization of a space should have complete coverage and be easy to use. Complete coverage means that the space reachable by (linear) combinations of the parameters includes all (at least most) of the interesting points in that space. Of course, the definition of the word *interesting* may vary from application to application, so a generally useful parameterization includes as much of the space as possible. For a parameterization to be easy to use, the set of parameters should be as small as possible, the effect of each parameter should be independent of the effect of any other parameter, and the effect of each parameter should be intuitive. Of course, in something as complex as facial animation, attaining all of these objectives is probably not possible, so determining appropriate trade-offs is an important activity in designing a parameterization. Animation brings an additional requirement to the table: the animator should be able to generate common, important, or interesting motions through the space by manipulating one or just a few parameters.

The most popular parameterized facial model is credited to Parke [23] [24] [25] and has already been discussed in terms of creating facial models based on the so-called conformational parameters of a generic facial model. In addition to the conformational parameters, there are *expression parameters*. Examples of expression parameters are upper-lip position, eye gaze, jaw rotation, and eyebrow separation. Figure 10.11 shows a diagram of the parameter set with the (interpolated) expression parameters identified. Most of the parameters are concerned with the eyes and the mouth, where most facial expression takes place. With something as complex as the face, it is usually not possible to animate interesting expressions with a single parameter. Experience with the parameter set is necessary for understanding the relationship between a parameter and the facial model. Higher level abstractions can be used to aid in animating common motions.

10.3.2 Blend shapes

The simplest approach to facial animation is to define a set of key poses, also called *blend shapes*. For a set of example blend shapes, see Figure 10.12 (Color Plate 10). Facial animation is produced by selecting two of the key poses and interpolating between the positions of their corresponding vertices in the two poses.

This restricts the available motions to be the interpolation from one key pose to another. To generalize this a bit more, a weighted sum of two or more key poses can be used in which the weights sum

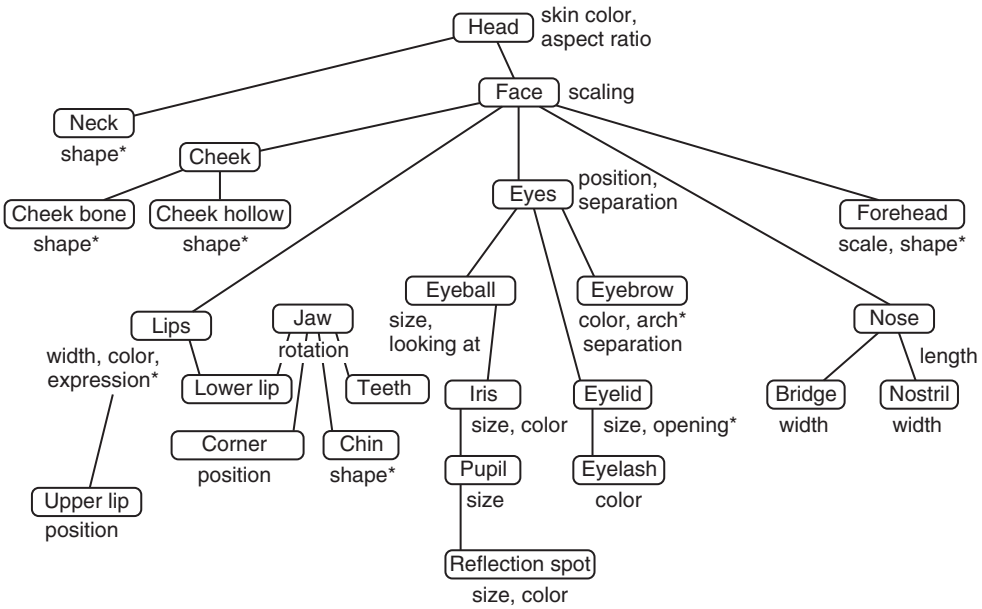

FIGURE 10.11

Parke model. *Indicates interpolated parameters [25].

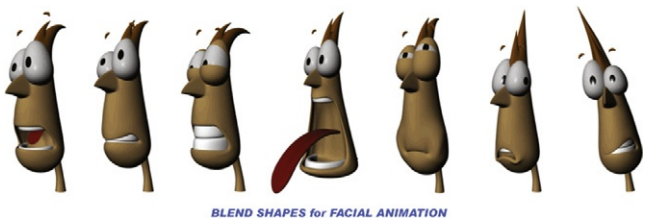

FIGURE 10.12

Blend shapes of a character. (Character design by Chris Oatley; rig design by Cara Christeson; character modeling and posing by Daniel Guinn.)

to one. Each vertex position is then computed as a linear combination of its corresponding position in each of the poses whose weight is non-zero. This can be used to produce facial poses not directly represented by the keys. However, this is still fairly restrictive because the various parts of the facial model are not individually controllable by the animator. The animation is still restricted to those poses represented as a linear combination of the keys. If the animator allows for a wide variety of facial motions, the key poses quickly increase to an unmanageable number.

10.3.3 Muscle models

Parametric models encode geometric displacement of the skin in terms of an arbitrary parametric value. Muscle-based models (e.g., Figure 10.13, Color Plate 11) are more sophisticated, although there is wide variation in the reliance on a physical basis for the models. There are typically three types of muscles that need to be modeled for the face: linear, sheet, and sphincter. The *linear muscle* is a muscle that contracts and pulls one point (the *point of insertion*) toward another (the *point of attachment*). The *sheet muscle* acts as a parallel array of muscles and has a line of attachment at each of its two ends rather than a single point of attachment as in the linear model. The *sphincter muscle* contracts a loop of muscle. It can be thought of as contracting radially toward an imaginary center. The user, either directly or indirectly, specifies muscle activity to which the facial model reacts. Three aspects differentiate one muscle-based model from another: the geometry of the muscle-skin arrangement, the skin model used, and the muscle model used.

The main distinguishing feature in the geometric arrangement of the muscles is whether they are modeled on the surface of the face or whether they are attached to a structural layer beneath the skin (e.g., bone or tissue [27],[33]). The former case is simpler in that only the surface model of the face is needed for the animation system (Figure 10.14). The latter case is more anatomically correct and thus promises more accurate results, but it requires much more geometric structure in the model and is therefore much more difficult to construct (Figure 10.15).

The model used for the skin will dictate how the area around the point of insertion of a (linear) muscle reacts when that muscle is activated; the point of insertion will move an amount determined

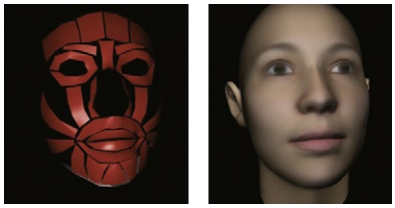

FIGURE 10.13

Example muscles for facial animation [30].

(Image courtesy of Arun Somasundaram.)

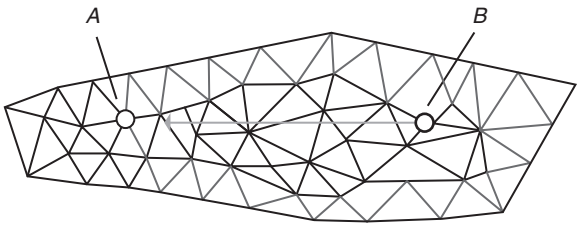

FIGURE 10.14

Part of the surface geometry of the face showing the point of attachment (A) and the point of insertion (B) of a linear muscle; point B is pulled toward point A.

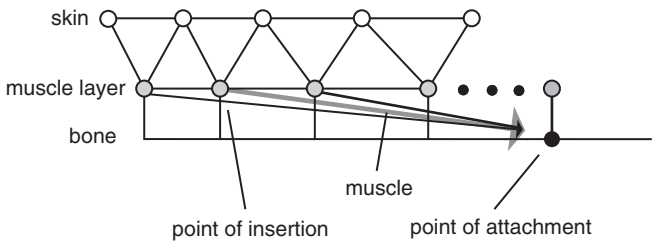

FIGURE 10.15

Cross section of the tri-layer muscle as presented by Parke and Waters [27]; the muscle only directly affects nodes in the muscle layer.

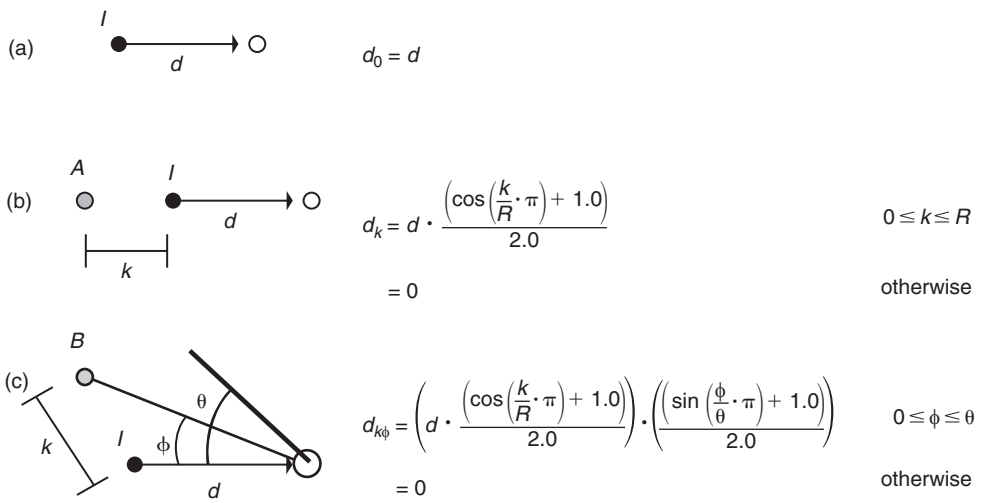

FIGURE 10.16

Sample attenuation: (a) insertion point *I* is moved *d* by muscle; (b) point *A* is moved d_k based on linear distance from the insertion point; and (c) point *B* is moved $d_k\phi$ based on the linear distance and the deviation from the insertion point displacement vector.

by the muscle. How the deformation propagates along the skin as a result of that muscle determines how rubbery or how plastic the surface will appear. The simplest model to use is based on geometric distance from the point and deviation from the muscle vector. For example, the effect of the muscle may attenuate based on the distance a given point is from the point of insertion and on the angle of deviation from the displacement vector of the insertion point. See Figure 10.16 for sample calculations. A slightly more sophisticated skin model might model each edge of the skin geometry as a spring and control the propagation of the deformation based on spring constants. The insertion point is moved by the action of the muscle, and this displacement creates restoring forces in the springs attached to the insertion point, which moves the adjacent vertices, which in turn moves the vertices attached to them, and so on (see Figure 10.17). The more complicated Voight model treats the skin as a viscoelastic

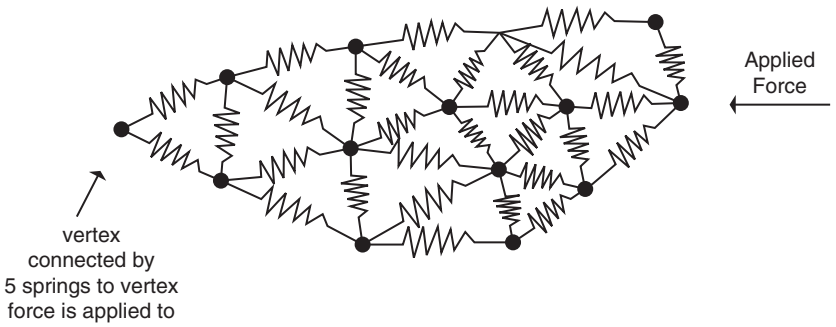

FIGURE 10.17

Spring mesh as skin model; the displacement of the insertion point propagates through the mesh according to the forces imparted by the springs.

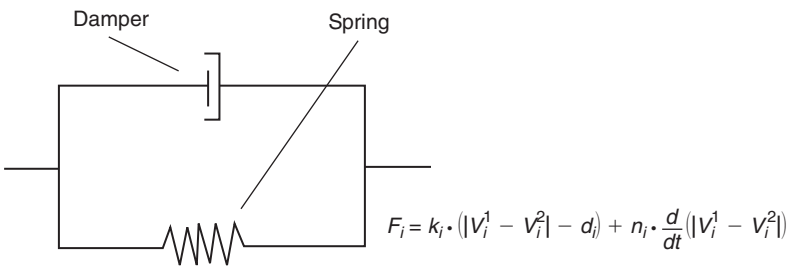

FIGURE 10.18

Voight viscoelastic model; the motion induced by the spring forces is damped. Variables k and n are spring and damper constants, respectively; and d_i is the rest length for the spring.

element by combining a spring and a damper in parallel (Figure 10.18). The movement induced by the spring is damped as a function of the change in length of the edge.

The muscle model determines the function used to compute the contraction of the muscle. The alternatives for the muscle model are similar to those for the skin, with the distinction that the muscles are active elements, whereas the skin is composed of passive elements. Using a linear muscle as an example, the displacement of the insertion point is produced as a result of muscle activation. Simple models for the muscle will simply specify a displacement of the insertion point based on activation amount. Because, in the case of the orbicularis oris, muscles are attached to other muscles, care must be taken in determining the skin displacement when both muscles are activated.

More physically accurate muscle models will compute the effect of muscular forces. The simplest dynamic model uses a spring to represent the muscle. Activating the muscle results in a change of its rest length so as to induce a force at the point of insertion. More sophisticated muscle models include damping effects. A muscle model developed by clinical observation is shown in Figure 10.19. However, spring-based facial muscles often result in a computationally expensive approach, and jiggling of the skin can be difficult to control.

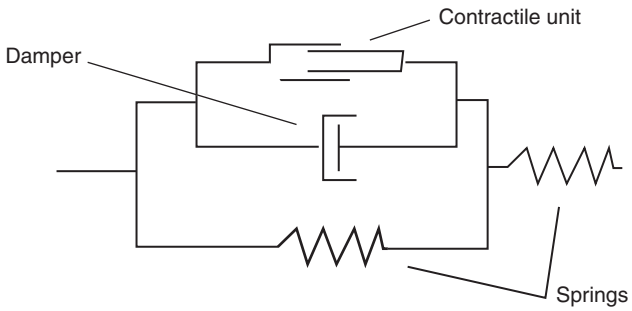

FIGURE 10.19

Hill's model for the muscle.

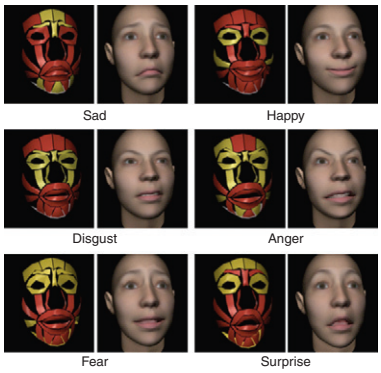

FIGURE 10.20

A basic set of facial expressions [30].

(Image courtesy of Arun Somasundaram.)

10.3.4 Expressions

Facial expressions are a powerful means of communication and are the basis for most facial animation. Any facial animation system will provide for a basic set of expressions. A commonly used set is: happy, angry, sad, fear, disgust, and surprise (Figure 10.20, Color Plate 12).

10.3.5 Summary

A wide range of approaches can be used to model and animate the face. Which approach to use depends greatly on how realistic the result is meant to be and what kind of control the animator is provided. Results vary from cartoon faces to parameterized surface models to skull-muscle-skin simulations. Realistic facial animation remains one of the interesting challenges in computer animation.

10.4 Lip-sync animation

10.4.1 Articulators of speech

Speech is a complex process and not even completely understood by linguists. Involved in the production of speech are the lungs, the vocal folds, the velum, lips, teeth, and tongue. These articulators of speech constitute the *vocal tract* (see Figure 10.21). The lungs produce the primary air flow necessary for sound production. Sound is a sensation produced by vibrations transmitted through air. The various sounds of speech are produced primarily by controlling the frequency, amplitude, and duration of these vibrations. Vibrations are generated by either the vocal folds in the throat or by the tongue in the oral cavity, or by lips. When vibration is produced by the vocal folds, the sound is called *voiced*. For example, the "th" sound in "then" is voiced whereas the "th" of "thin" is not voiced. The sound travels up into either nasal cavity (e.g., "n") or oral cavity depending on the configuration of the velum. In the oral cavity, the sound is modified by the configuration of the lips, jaw, and tongue with certain frequencies resonating in the cavity. In addition, the tongue and lips can vibrate by either partially obstructing the air flow or by completely stopping and then releasing it. If the air flow is relatively unrestricted, then the sound is considered a vowel, otherwise it is a consonant. If the air flow is completely stopped and then released (e.g., "t," "d") then it is referred to as a *stop*. If the air flow is partially restricted creating a vibration (e.g., "f," "th") then it is referred to as a *fricative*.

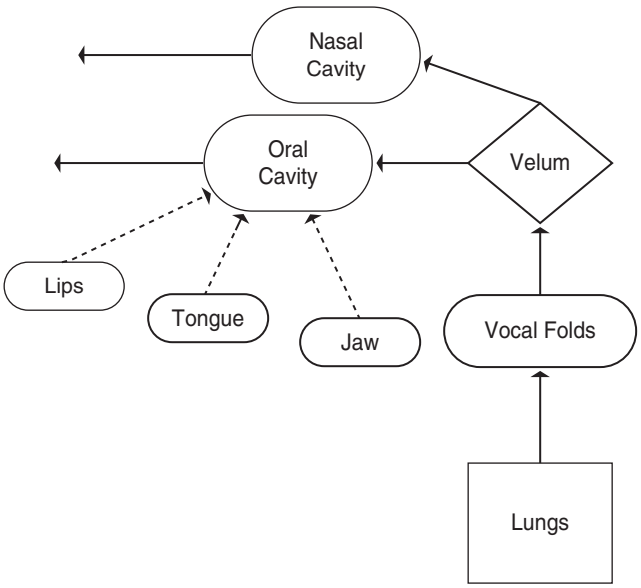

FIGURE 10.21

Schematic of air flow through the vocal tract. Lungs produce air flow. Vocal folds may vibrate. Velum deflects air flow to either nasal cavity or oral cavity. Harmonics of vibrations initiated by vocal folds may be reinforced in oral cavity depending on its configuration; vibrations may be initiated in oral cavity by lips or tongue. Dashed lines show agents that modify oral cavity configuration.

The fundamental frequency of a sound is referred to as F0 and is present in voiced sounds [21]. Frequencies induced by the speech cavities, called *formants*, are referred to as F1, F2, ... in order of their amplitude. The fundamental frequency and formants are arguably the most important concepts in processing speech.

While most of this activity is interior and therefore not directly observable, it can produce motion in the skin that may be important in some animation. Certainly it is important to correctly animate the lips and the surrounding area of the face. Animating the tongue is also usually of some importance.

Some information can be gleaned from a time-amplitude graph of the sound, but more informative is a time-frequency graph with amplitude encoded using color. Using these *spectrographs*, trained professionals can determine the basic sounds of speech.

10.4.2 Phonemes

In trying to understand speech and how a person produces it, a common approach is to break it down into a simple set of constituent, atomic sound segments. The most commonly used segments are called *phonemes*. Although the specific number of phonemes varies from source to source, there are generally considered to be around 42 phonemes.

The corresponding facial poses that produce these sounds are referred to as *visemes*. Visemes that are similar enough can be combined into a single unique viseme and the resulting set of facial poses can be used (Figure 10.22), for example, as blend shapes for a simple type of lip-sync animation.

However, the sounds and associated lip movements are much more complex than can be represented by simply interpolating between static poses. Within the context of speech, a phoneme is

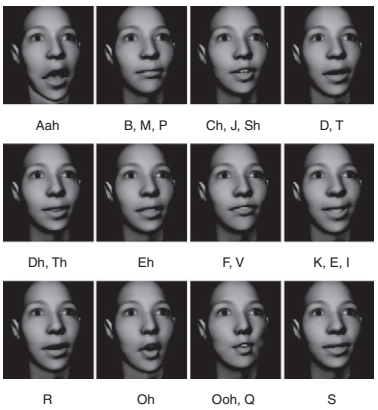

FIGURE 10.22

Viseme set [30].

manifested as variations of the basic sound. These variations of a phoneme are called *allophones* and the mechanism producing the variations is referred to as *coarticulation*. While the precise mechanism of coarticulation is a subject for debate, the basic reason for it is that the ideal sounds are slurred together as phonemes are strung one after the other. Similarly, the visemes of visual speech are modified by the context of the speech in which they occur. The speed at which speech is produced and the physical limitations of the speech articulators result in an inability to attain ideal pronunciation. This slurs the visemes together and, in turn, slurs the phonemes. The computation of this slurring is the subject of research.

10.4.3 Coarticulation

One of the complicating factors in automatically producing realistic (both audial and visual) lip-sync animation is the effect that one phoneme has on adjacent phonemes. Adjacent phonemes affect the motion of the speech articulators as they form the sound for a phoneme. The resulting subtle change in sound of the phoneme produces what is referred to as an allophone of the phoneme. This effect is known as *coarticulation*. Lack of coarticulation is one of the main reasons that lip-sync animation using blend shapes appears unrealistic. While there have been various strategies proposed in the literature to compute the effects of coarticulation, Cohen and Massaro [6] have used weighting functions, called *dominance functions*, to perform a priority-based blend of adjacent phonemes. King and Parent [18] have modified and extended the idea to animation song. Pelachaud et al. [26] cluster phonemes based on deformability and use a look-ahead procedure that applies forward and backward coarticulation rules. Other approaches include the use of constraints [12], physics [2] [30], rules [5], and syllables [19]. None have proven to be a completely satisfying solution for automatically producing realistic audiovisual speech.

10.4.4 Prosody

Another complicating factor to realistic lip-sync animation is changing neutral speech to reflect emotional stress. Such stress is referred to as *prosody*. Affects of prosody include changing the duration, pitch, and amplitude of words or phrases of an utterance. This is an active area of research (e.g., [3] [4] [5] [11] [20]).

10.5 Chapter summary

Facial animation presents interesting challenges. As opposed to most other areas of computer animation, the foundational science is largely incomplete. This, coupled with the complexity inherent in the facial structure of muscles, bones, fatty tissue, and other anatomic elements, makes facial animation one area that has not been conquered by the computer.

References

[1] Akimoto T, Suenaga Y. Three-Dimensional Facial Model Creation Using Generic Model and Front and Side View of Face. IEICE Transactions on Information and Systems March 1992;E75-D(2):191–7.

[2] Albrecht I, Haber J, Seidel H-P. Speech Synchronization for Physics-Based Facial Animation. In: Proceedings of WSCG. Feb 2002. p. 9–16.

[3] Byun M, Badler N. Facemote: Qualitative Parametric Modifiers for Facial Animations. In: Proceedings of the ACM SIGGRAPH Symposium on Computer Animation. 2002. p. 65–71.

[4] Cao Y, Faloutsos P, Pighin F. Unsupervised Learning for Speech Motion Editing. Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation. 2003. p. 225–31.

[5] Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket W, et al. Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. In: Proceedings of ACM SIGGRAPH 1994. p. 413–20.

[6] Cohen M, Massaro D. Modeling Coarticulation in Synthetic Visual Speech. In: Models and Techniques in Computer Animation. Tokyo: Springer-Verlag; 1993.

[7] COVEN. http://coven.lancs.ac.uk/mpeg4/index.html**; January 2001.**

[8] DeCarlo D, Metaxas D, Stone M. An Anthropometric Face Model Using Variational Techniques. In: Cohen M, editor. Computer Graphics. Proceedings of SIGGRAPH 98, Annual Conference Series. Orlando, Fla. Addison-Wesley; July 1998. p. 67–74. ISBN 0-89791-999-8.

[9] DeRose T, Kass M, Truong T. Subdivision Surfaces for Character Animation. In: Cohen M, editor. Computer Graphics. Proceedings of SIGGRAPH 98, Annual Conference Series. Orlando, Fla. Addison-Wesley; July 1998. p. 85–94. ISBN 0-89791-999-8.

[10] Ebert D, Ebert J, Boyer K. "Getting into Art" (animation). CIS Department, Ohio State University; May 1990.

[11] Edge J, Maddock S. Expressive Visual Speech Using Geometric Muscle Functions. In: Proc. of Eurographics UK Chapter Annual Conference (EGUK). 2001. p. 11–8.

[12] Edge J, Maddock S. Constraint-Based Synthesis of Visual Speech. In: Conference Abstracts and Applications of SIGGRAPH. 2004.

[13] Ekman P, Friesen W. Facial Action Coding System. Palo Alto, California: Consulting Psychologists Press; 1978.

[14] FAP Specifications. http://www-dsp.com.dist.unige.it/~pok/RESEARCH/MPEGfapspec.htm**; January 2001.**

[15] FDP Specifications. http://www-dsp.com.dist.unige.it/~pok/RESEARCH/MPEGfdpspec.htm**; January 2001.**

[16] Forsey D, Bartels R. Hierarchical B-Spline Refinement. In: Dill J, editor. Computer Graphics. Proceedings of SIGGRAPH 88 vol. 22(4). Atlanta, Ga.; August 1988. p. 205–12.

[17] International Computer. Three-Dimensional Tutorials: Pixologic ZBrush: Head Modeling Part 1, www.3dlinks.com/oldsite/tutorials_ZOldHead.cfm**; August 2006.**

[18] King S, Parent R. Animating Song. Computer Animation and Virtual Worlds March 2004;15(1):53–61.

[19] Kshirsagar S, Magnenat-Thalmann N. Visyllable Based Speech Animation. Proceedings of Eurographics 2003;22(3):631–9.

[20] Kshirsagar S, Molet T, Magnenat-Thalmann N. Principle Components of Expressive Speech Animation. In: Computer Graphics International. IEEE Computer Society; 2001. p. 38–44.

[21] Lemmetty S. Review of Speech Synthesis Technology. M.S. Thesis, Helsinki University; 1999. http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/**.**

[22] Muraki S. Volumetric Shape Description of Range Data Using Blobby Model. In: Sederberg TW, editor. Computer Graphics. Proceedings of SIGGRAPH 91 vol. 25(4). Las Vegas, Nev.; July 1991. p. 227–35. ISBN 0-201-56291-X.

[23] Parke F. Computer-Generated Animation of Faces. In: Proceedings of the ACM Annual Conference; August 1972.

[24] Parke F. A Parametric Model for Human Faces. Ph.D. dissertation, University of Utah; 1974.

[25] Parke F, Waters K. Computer Facial Animation. Wellesley, Massachusetts: A. K. Peters; 1996. ISBN 1-56881-014-8.
[26] Pelachaud C, Badler N, Steedman M. Generating Facial Expressions for Speech. Cognitive Science 1996;20 (1):1–46.
[27] Ratner P. Subdivision Modeling of a Human, www.highend3d.com/maya/tutorials/-modeling/polygon/189.html**; August 2006.**
[28] Rydfalk M. CANDIDE: A Parameterized Face. Technical Report LiTH-ISY-I-0866, Sweden: Linköping University; 1987.
[29] Silas F. FrankSilas.com Face Tutorial, www.franksilas.com/FaceTutorial.htm**; August 2006.**
[30] Somasundaram A. A Facial Animation Model for Expressive Audio-Visual Speech. Ph.D. Dissertation, Ohio State University; 2006.
[31] Waters K. A Muscle Model for Animating Three-Dimensional Facial Expressions. In: Proceedings of SIGGRAPH, vol. 21(4); 1987. p. 17–24.
[32] Waters K. A Physical Model of Facial Tissue and Muscle Articulation Derived from Computer Tomography Data. SPIE Visualization in Biomedical Computing 1992;1808:574–83.
[33] Waters K, Terzopoulos D. Modeling and Animating Faces Using Scanned Data. Journal of Visualization and Computer Animation 1991;2(4):123–8.