1. INTRODUCTION

Interest in facial animation began with a PhD thesis by Fred Parke in 1974. He was the first to utilise the power of 3-D computer graphics in a way that brought a higher level of control into the act of animating a face. Predicted applications, at that stage, included computer generated actors and a valuable research tool for people studying the area of human facial expression and non-verbal communication. Interest in facial animation has grown rapidly over the last ten years. Now, almost 20 years after the first system was developed many new applications have been touted. Among the most feasable are communications, user interfaces and aids to predict the effects of surgery.

2. BACKGROUND INFORMATION

This thesis brings together information from various disciplines to produce an animation/speech synchronisation system. To give an accurate representation of the human face, the anatomy of the face must be taken into account. This brings in information from the field of medicine. Theories on human expression and methods for recording facial movements have been researched by psychologists. To automate the synchronisation process by analysing sound, ideas from physics and digital signal processing are needed. The implementation of speech movements can draw information from speech and hearing researchers and animators. In an effort to give their actors more character, most facial animation researchers will seek artistic advice from experts in traditional animation. 3D data for creating the original facial model is recorded using photogrammetric techniques and advice from people in the surveying and cartography area. An area undergoing a lot of research over recent years is user interfaces. One of the main motivators behind the work done in facial animation is the development of user interfaces. Anyone developing computer systems should have an interest in improving their user interface. I will outline what is applicable to my thesis in these fields of research as background to facial animation and speech synchronisation.

2.1. Background - Anatomical

To get a true grasp on what is involved in creating a facial animation system, a basic knowledge of human anatomy is required. The main components of the face are: bone (skull), skin and muscle. The skull consists of fourteen major bones, of which the mandible (jaw) is the only free-moving part. (see Fig. 2.1) The size and shape of the bones of the skull varies greatly from person to person. Even so, it is the bone that remains fairly constant throughout a person's lifetime. The muscle and soft tissues can change radically, but the skull's structure determines the shape of the face we recognise (Waters, 1990 p111).

The skin is comprised of two layers, the dermis covered by the epidermis. The epidermis is a layer of dead skin cells that protects the dermis from the elements. Under the skin is a layer of subcutaneous fat, and under the fat is the fascia. The fascia is a fibrous tissue that is connected to the muscle and cartilage of the face. The skin contains collagen (72%) and elastin (4%) fibres which are responsible for its elasticity (Terzopoulos & Waters, 1990). Under low forces, the skin stretches easily. Once the stress passes a certain point, the fully extended fibres

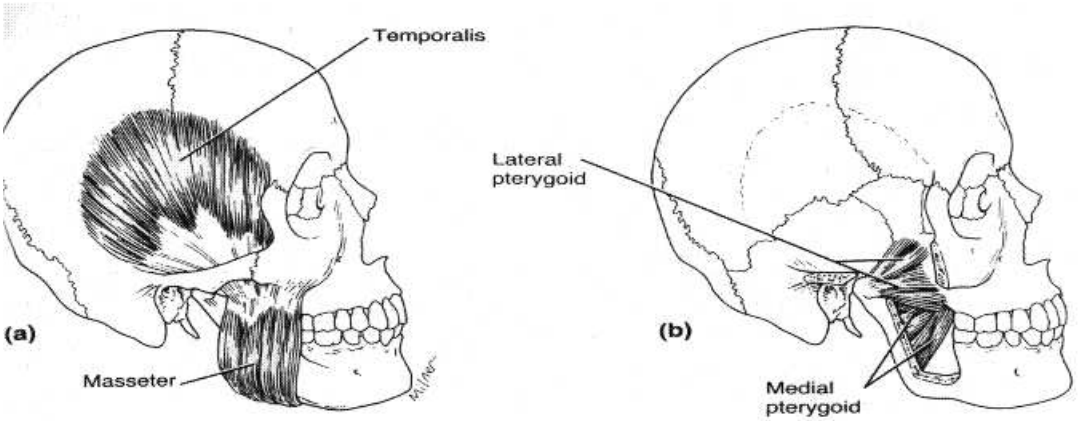

Fig. 2.1 : Major bones of the Skull, Spence and Mason (1983)

are much more resistant to force.

Skin behaves similarly to a rubber sheet, deforming itself around the underlying structure of the face. This can lead to creasing or wrinkling of the skin. These wrinkles become more noticable with age as the skin loses elasticity and fatty tissue. The skin is also affected by gravitational forces, pulling it downwards. The most influential factors affecting the position of a point on the skin at a given time are:

The tensile strength of the muscle and skin.
The proximity of the skin to muscle attachments.
The depth of the underlying tissue and the closeness to facial bones.
The elasticity of the tissue.
The interaction of nearby muscles. (Waters, 1990)

The muscles of the face (see Fig. 2.2) are responsible for facial expression. Most of muscles in the rest of

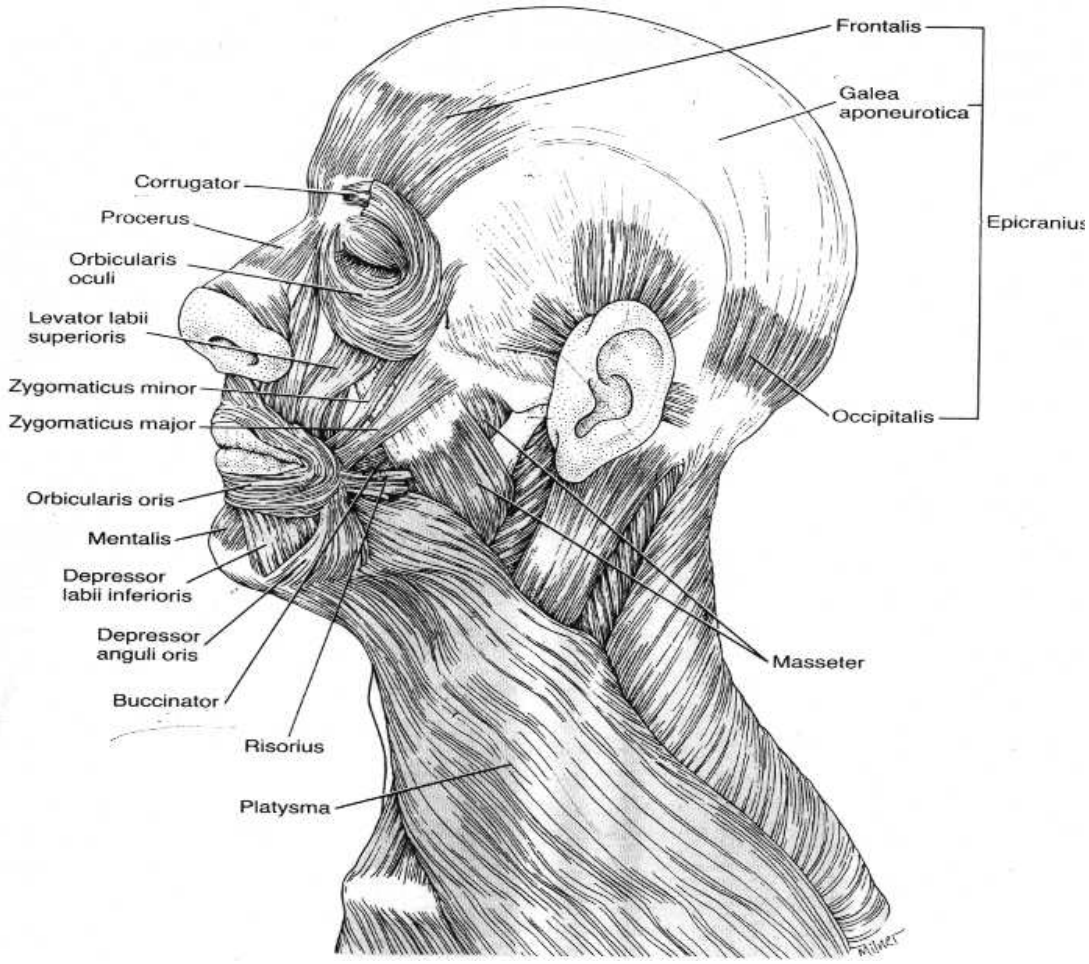

Fig. 2.2 : Facial Muscles, Spence and Mason (1983)

the body have both ends are attached to bone. The muscles of the face are more likely to be connected to bone at one end and to the fascia at the other. There are three types of facial muscle: linear, sheet and sphincter (Terzopoulos & Waters, 1990). A mixture of these muscle types co-ordinate to produce each facial expression.

The following information on muscles comes from an anatomy text, Spence and Mason (1983). Linear, or parallel, muscles are the most common. An example of a linear muscle is the zygomaticus major which pulls up the corner of the mouth. Linear muscles have two attachment points. As they contract, they pull the fascia towards the attachment point on the bone, producing a facial movement. A slightly different type of muscle is the sheet muscle. As with linear muscles, sheet muscles work in one direction. The difference is that they are flatter and wider, and their attachment area is broader. An example of this is the epicranius. This muscle has two parts: the anterior frontalis and the posterior occipitalis. The two muscular sheets are connected by a broad, flat tendon. Contraction of one or the other of the epicranial muscles pulls the scalp backward or forward. Sphincter muscles usually surround body openings. They are ring shaped and will enlarge or reduce the opening by relaxing or contracting. The orbicularis oris is the sphincter muscle that closes the mouth and purses the lips.

Current facial animation systems use varying levels of anatomical realism. Some work by looking at what is externally visible, with no attempt to model the complexity of the facial layers (for example, Parke 1974). Other systems, such as Waters' tri-layer, physically based model (1990), simulate every layer of the face. There needs to be a trade-off between the complexity of the system, and the speed with which it operates.

2.2. Background - Facial Movement and Expression

Analysis of facial expression has been a research area for psychologists for decades. The psychologists' interest in this area is to find out how humans transmit and receive information through the face. They also look at how mental and physical disabilities affect facial movement. The expressiveness of the faces of patients with depression will often go through a series of stages related to their illness. It is hoped that breakthroughs in treatment of patients and the monitoring of illnesses will come from this research into facial movement. Another area of study is facial deception; how to tell if a person is lying to you. When a person lies, you can usually see some conflict in the expression on their face. This perception tends to come naturally. Studying which movements do and don't go with each other will give insights into how we interpret facial movement.

Two psychologists, Ekman and Friesen, have developed a system for anatomical analysis of facial movements. Their Facial Action Coding System, FACS (1978), has proven itself to be the most successful method available for facial expression evaluation. They set about finding a method for detecting and recording facial movements in an objective way. Their research involved taking photographs of their own faces as they selectively fired different muscles. Through this they found out which muscles created which movements. In some cases they found groups of muscles that produced very similar effects. The groups of muscles and single muscles that are responsible for each distinguishable facial movement were named Action Units. (see Fig. 2.3) Ekman and Friesen found 46 Action Units and give full descriptions of each one.

Once the Action Units had been defined, Ekman and Friesen set about finding out how the units combine to create complex expressions. Their manual includes information on how to identify which Action Units are involved in both simple and complex movements. They looked at over 55,000 photos of different facial expressions. From these photos, they decided on six primary expressions. The expressions chosen communicate anger, fear, disgust, sadness, happiness and surprise. (see Fig. 2.4) Most facial animation systems use the information in the FACS as a basis for designing and testing their systems.

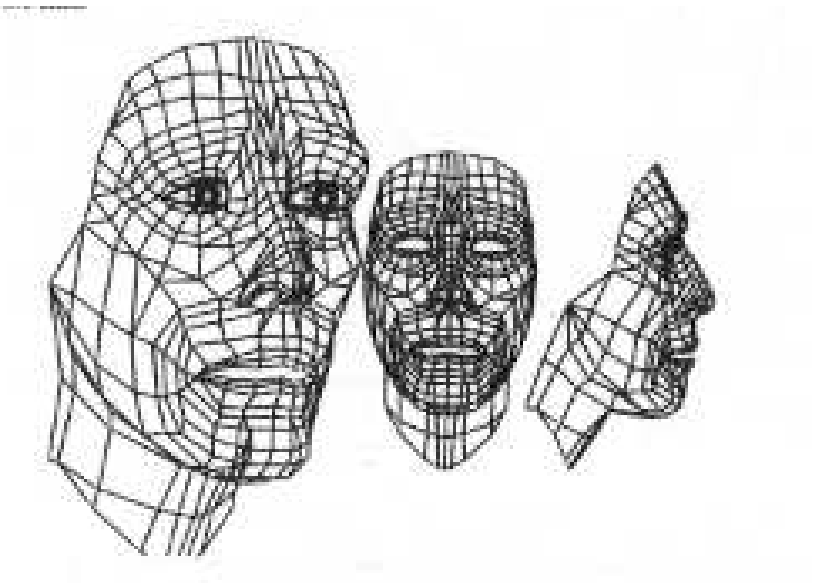

Fig. 2.3 : Eyebrow Action Unit, Waters' (1987)

Fig. 2.4 : The six basic facial expressions, Ekman and Friesen (1975)

2.3. Background - Speech

Tony White (1986) outlines the traditional method of synchronising animation with a speech track. This system involves recording the speech on tape, then noting down the phonetic interpretation of the words frame by frame. It is essential that the timing be done correctly as mistakes can mean having to do an expensive redraw of whole sections of the animation. This method of speech synchronisation is workable, as long as you have a script and you have plenty of time. To make speech synthesis viable in real time, a different approach must be taken. As with parameterisation of the face, the most attractive solution is to take a more abstract view. To do this, we have to look at the structure of language.

The phonetic alphabet is based on the basic elements of speech - sounds. This notation for the spoken word takes us away from letters and into a representation of the sounds that they can make. Mitchell (1964) explains that, "The relationship between letter and sound in the spelling of English is hopelessly confused and inconsistent". He goes on to outline a perfect system of representing sound by visual symbols:-

(a) the same symbol would always represent the same sound.

(b) the same sound would always have the same symbol.

(c) a single sound would be represented by one symbol.

(d) there would never be a symbol in the spelling that did not correspond to a sound.

The statements in (a) to (c) are easily shown to be untrue of English, eg bow (curtsey), bough and bow (tie). The best examples of statement (d) are the silent letters in words like "gnome". Obviously, English is not a good place to start when creating a speech synthesis system.

Within the phonetic alphabet are phonemes. Phonemes are groups of sounds that are variants of a single sound. For example, the letter "t" can be said using a variety of mouth positions. These variations of the letter "t" are said to be in the neighbourhood of each other as their interchange does not affect their meaning. This grouping of closely related sounds lets us use the same symbol in different sound contexts knowing that its pronunciation will not vary by much. A bonus in working with phonemes is their language independence. As long as the full phonetic alphabet is supported, the system should be able to speak in any language.

The parts of the face involved in speaking are: the tongue, epiglottis, food passage, hard palate, lips, pharynx, soft palate, teeth, teeth ridge, uvula, position of the vocal chords and the windpipe. Luckily for us, the only elements we have to worry about in animation are the lips and the teeth. Masden (1969) indicates that lip animation requires the following capabilities:

(a) open lips for the open vowels a, e and i

(b) closed lips for the accent consonants b, m and p

(c) an oval mouth for u, o and w

(d) and the lower lip tucked under the upper front teeth for f and v.

The remaining sounds are formed mainly by the tongue and do not require precise animation (Parke, 1975). So for an animation system, unseen movements can be ignore. The visible movements made by the face to pronounce each phoneme will still need to be worked out, as a speech synchronised system will have to be capable of them all. (see Fig. 2.5) There would need to be an appropriate action to match each phoneme output by the synthesiser.

2.4. Background - Animation

In their 1987 paper, Lasseter and Rafael investigated how traditional animation techniques can be applied to 3-D computer animation. Early computer animation was very similar to that of places like the Disney studios. Techniques such as storyboarding, keyframing and inbetweening were implemented using computers which made animation easier, but didn’t bring anything new to the field. With 3-D computer animation, objects are three dimensional, as in true life. The animator can work with the characters, rather than having to draw them frame by frame. The characters can be controlled like puppets, moving and talking, making better use of the power of computers.

Bergeron (1985) created the computer animated short "Tony De Peltrie" using a 3-D animation package. Tony was made up of an hierarchical skeleton that was manipulated through the TAARNA interactive animation system. Using a hierarchy automates the positioning of body parts as the system knows about the connections and dependencies within the body. Facial expressions were mapped from data recorded by digitising a real face. Thus a mixture of keyframing and 3-D animation came together to produce a very believable piece of animation.

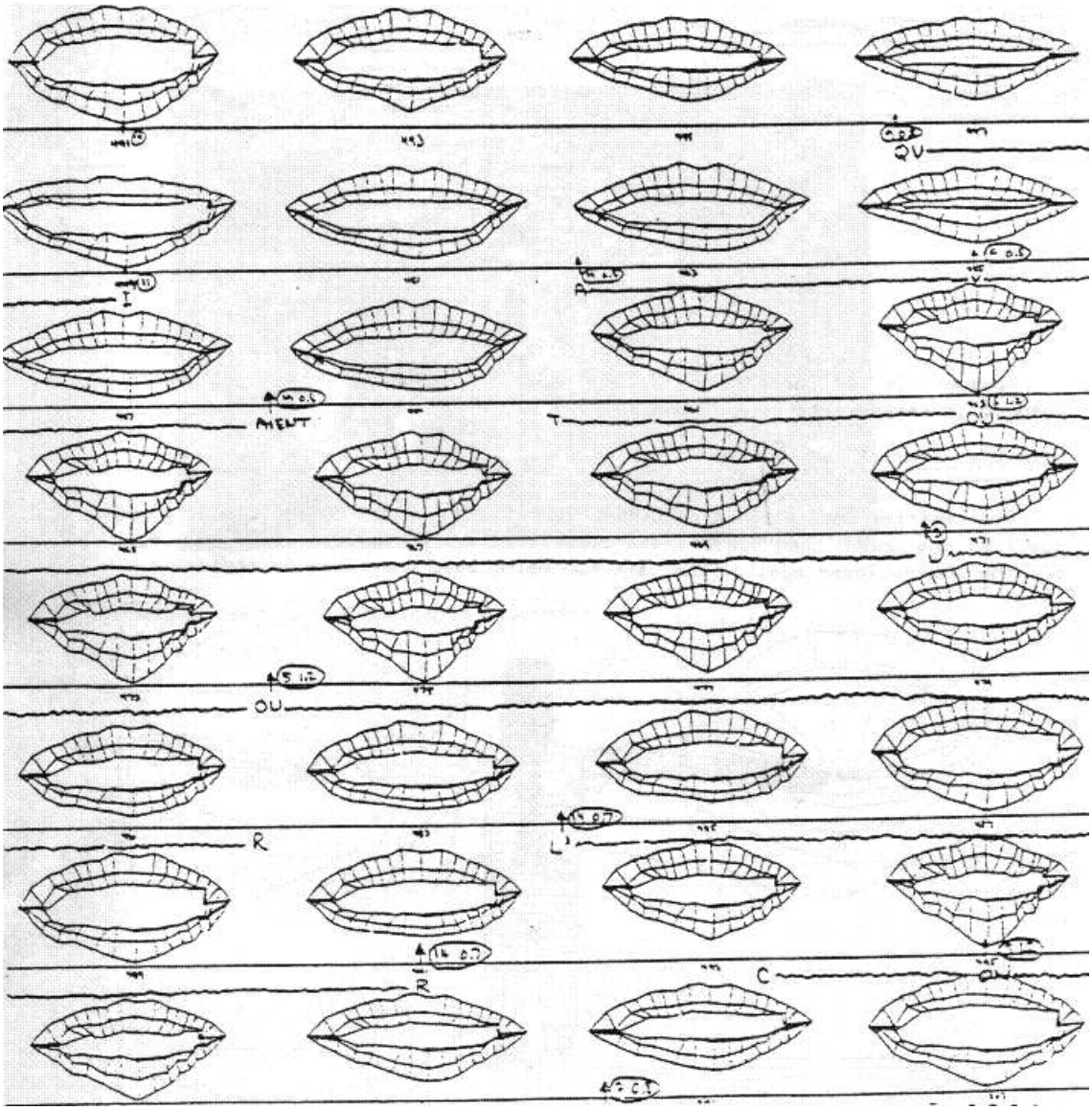

Fig. 2.5 : An array of lip positions, Bergeron (1985)

Todays computer animators can learn more about the craft of animation from its tradition. No matter how much the animation process is automated, it will still require artistic talent to make it work. Lasseter and Raphael (1987) give pointers on traditional tricks of the trade that the new breed of animators should bear in mind.

The principles of animation were created by studying the way things move in reality and working out how to get the same effect in an animated sequence. The most important principle is squash and stretch. When objects move in real life, they retain their volume, but can change their shape. The most obvious example is a bouncing ball. In an animated sequence, it will squash as it hits the ground, then stretch as it bounces away again. Different materials and shapes have their own way of squashing. In facial animation, the squash and stretch of the facial parts as they move in relation to each other is very important. Squash and stretch are also

used to combat strobing between frames.

Timing of movements is also important as it can convey the speed and the size of objects. Thus a heavy object will take more time to get moving and get up to a speed, whereas a small object, like a mouse, will take no time at all. Anticipation of what is to come is another tool that animators can used to get their ideas across. For example, when a character is going to run, they often run on the spot for a few frames before they move. This prepares the audience for what is going to come next. Staging is similar to anticipation in that it helps the viewer to know where to look. Staging gives a focus to the scene, so that the eye of the viewer is fixed on the correct part of the screen to see the next piece of action.

To give a realistic imitation of something stopping, animators use follow through and overlapping action. An example is the follow through after a ball is thrown. There are two main methods of traditional animation: straight ahead and pose to pose. In straight ahead action, the animator knows what they're doing from the top, and goes from the first frame, all the way through. With pose to pose, the technique is much more like the key-framing that most computer animators use. They pick the positions each character will be in during each scene and then do the inbetween frames by interpolating between the two.

Slow in and slow out deal with the spacing of objects during movement. For example, a bouncing ball will be going very fast as it hits the ground, but will slow down as it reaches the top of the bounce. To give a more interesting movement, animators will often move objects in an arc rather than a straight line. They will also exaggerate the movements of a character to add life. Secondary actions, and the appeal of each character also heighten the effect of the animation.

The most important part of character animation is to create a personality for each character. If we want to use a facial animation system to create believable characters, we have to learn the art of animation. What we have is a new set of tools to use for animation, mastery of these tools will give an entertaining result.

2.5. Background - Data Recording

Before any animation can be done, a 3-D model of the face/head is required. There are two components to this 3-D model: the data points, and the topology. The number of data points recorded will vary depending on the accuracy required by the animator. To make the digitising process easier, the model can have markings on the face. The topology of the model is the indicator of the connectivity of the data points. Topological information is often a triangulation of the data. Triangles are chosen as they are easy to render and are always planar. If polygons have four or more sides, it is harder to ensure planarity and validity of polygons.

There are many techniques for recording 3-D spatial information. Good sources of information in this area are surveyers and cartographers. Even though most of their work involves very large objects, they do know a lot about techniques for close range digitising. A manual technique for recording data is to scan two photographs taken from different (known) angles into a computer. Points on each image can be selected and their 3-D position can be calculated. A similar method is to use two slides and an analytical stereo digitiser. The digitiser projects one image onto each eye to give a 3-D effect. A floating mark can be guided over the image, recording each data point as it goes. A description of both of these techniques is in Doak et al (1991). Parke also describes 3-D digitisation in one of his earliest papers (1975b). For any of these methods, the taking of the photographs must be done with great care as it is crucial to the final result. Calculation of each point in 3-space can be very complicated. A new method for solving the equations needed to work out 3-D positions of the data points is described in Naftel (1991).

Once the 3-D data has been recorded, the topological information needs to be defined. (see Fig. 2.6) This can be done manually by selecting the points in each triangle. Order is important when choosing the vertices so

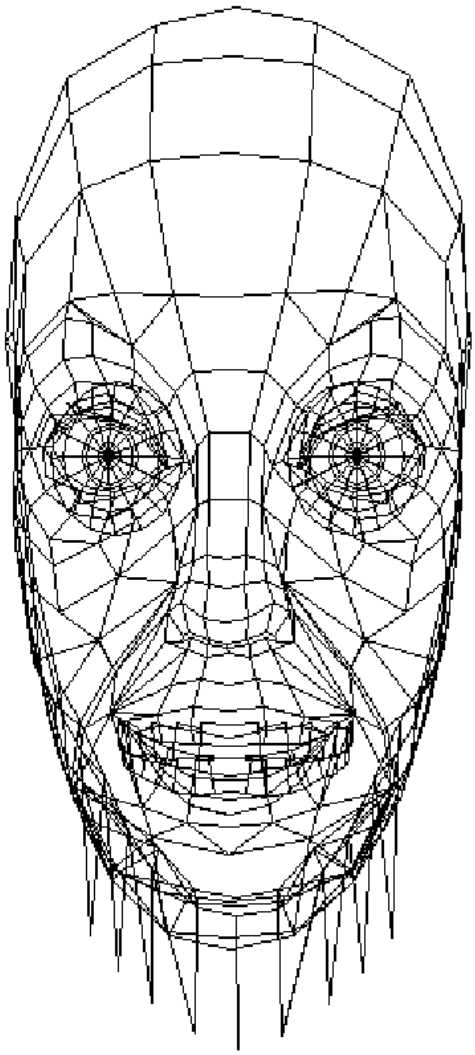

Fig. 2.6 : Wireframe display of the face.

that the triangles will be facing the same direction. Errors made at this stage become obvious when the image is rendered. Another method is to put the points through a triangulation program. There are many algorithms for triangulation, eg. Delaunay triangulation. It is important to choose one that will give the correct connectivity. Quite often, the irregular distribution of the points of the face cause mistakes in connections made during triangulation.

Variations on methods of recording data are very common. The use of 3-D digitisers takes a lot of the manual work out of recording data. Anderson (1990) used a Cyberware 3-D video laser to record facial data for The Abyss. Other techniques, such as light striping using a scanning laser (Yau, 1988), automate the collection of 3-D data. Triangulation of the data points still needs to be done. A more complicated method of recording and storing data uses surface patches rather than triangles. I didn't find any references that cared to explain how

they’d implement surface patches. Quite a few put it in their "improvements" section.

The trade-off between accuracy, speed and cost of machinery must be evaluated before work commences. Stereo digitisers, scanning lasers and 3-D digitisers don’t come cheap. They are quicker and more accurate than quick and dirty methods, however, so a decision must be made. Each method has its pros and cons, the choice really depends on the application.

2.6. Background - User Interfaces

Gasper (1988) states that user interfaces (UI’s) have gone through a series of generations in a similar way to computer languages. The first generation used switches and lights to convey information. Next came keyboards and character output followed by the third generation which introduced graphics and pointers. This is the current phase for user interfaces. We are working towards reaching the fourth generation, voice synthesis and recognition, where the interface will be via speech rather than typed characters and mouse clicks. The fifth generation, if we ever get there, will use synthesised actors and talking faces.

All of these developments are working towards making computers friendlier for users. This seems logical when you consider who it is we are creating the systems for. There are many users who are scared of computers. The most successful interface for data transmission is the human face. It has been shown that only 10% of the information we receive during conversation comes from what is said. The other 90% comes from facial and bodily gestures (Pease, 1982). From this viewpoint, it is obvious that our current interfaces can not be making full use of what we know about human perception. If we can create systems that utilise all we know about how humans communicate, we will be able to transfer more information in a shorter period of time.

One example of how our knowledge of human perception can be used to advantage is the Chernoff face (Marriott, 1990). Chernoff faces are a means of displaying complex, multi-variable data. Each feature on a Chernoff face depends on a different variable. As the values for the variables change, the "expression" on the face changes. Using faces and other familiar objects for displaying data values makes it easier to see what is happening to the data. The alternative would be to produce huge lists of numbers or to graph the data. It is easy to be boggled by pages and pages of numbers, and graphs have their limitations too. Users are more likely to have a good response when data is presented in a manner that is quickly and easily understood.

Facial animation systems will help to make user interfaces friendlier in a similar way to Chernoff faces. The facial model could go beyond simply speaking to the user, it could give non-verbal messages as well. If the user does something wrong, the face could become angry. When the terminal is sitting idle, the face might start to look bored and start whistling to itself. These features are over and above what is expected to happen with facial animation systems being incorporated into user interfaces. Most researchers are working towards building a more natural user interface. This type of interface would have the user talking head to head with the computer in a very natural way (Morishima, 1991). Such systems are already in existence, but they need to become more sophisticated before they can be put into general use. The two main failings of these systems are the quality of the synthesised speech coming from the computer and the ability of the computer to recognise speech.

Creating a more natural interface is one way in which we can encourage new users to make better use of computers. Many people are scared of command line interfaces, and often using a mouse and windows can be a daunting task. By simulating an interface that all people are already familiar with, learning how to use computers will be much easier for the novice. The regular user will find it a lot more comfortable to work with.

3. APPLICATIONS

There is a wide range of applications for facial animation systems. Some of them are purely recreational (film-making), some make life easier for us (user interfaces) and others have a specific practical use (videophone technology). The choice of animation system depends highly on the sort of application it is being developed for. (see Fig. 3.1)

3.1. Applications - Film Making

The most well known application, as far as the general public is concerned, is facial animation for film and video. Animation is widely used as a special effect in movies to create sequences that would often be impossible

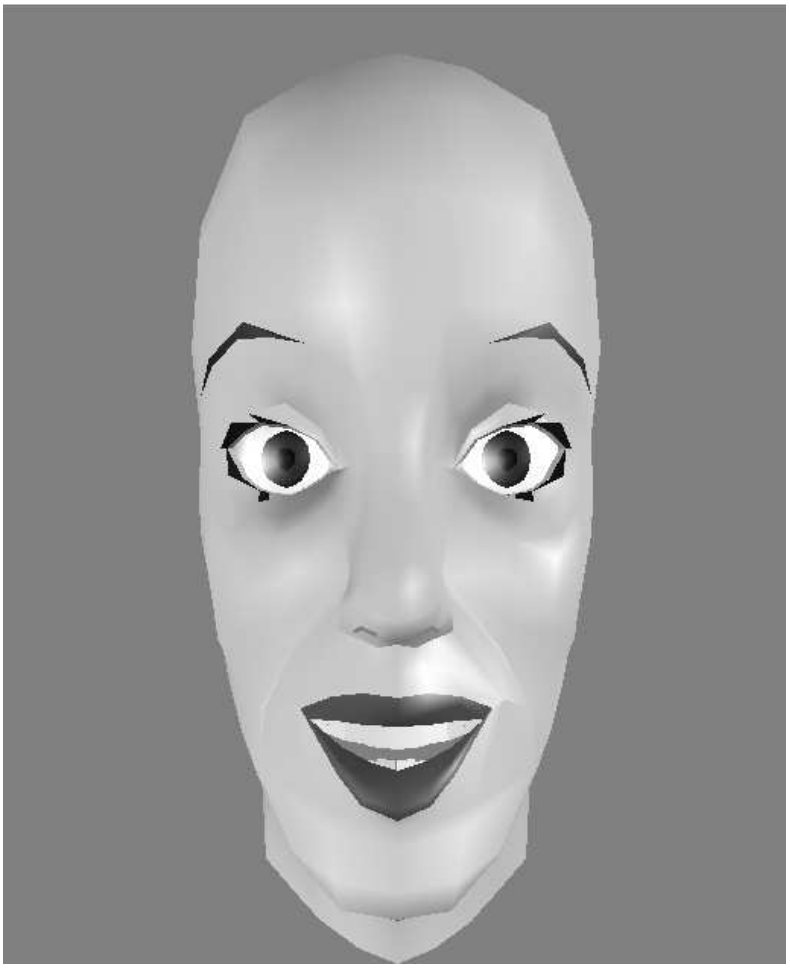

Fig. 3.1 : A smiling face

to do any other way. This is the least restrictive of applications that use facial animation systems as there is no need for a real-time system. Most facial animation that we see on film is done using key-frame animation. This is the method used for the pseudo-pod in The Abyss (Anderson, 1990). Bergeron (1985, 1990) uses packages for his animated shorts, but is still using a variation of key-framing for the actual animation.

Many of the latest movies make use of facial animation. Most of them, like The Abyss and Terminator 2, use full digitisation of each key-frame as a base for their animated sequences. The face of Kane in Robocop 2 is an updated version of Mike the Talking Head (Robertson, 1988). Mike was one of the first widely known results of facial animation, he even has his own manager. That serves as an indication of what can happen when a character is animated well enough to get a following.

DiPaola (1991) is working on extending the capabilities of facial animation systems. Animators are starting to demand greater flexibility from their animation systems. Many are using ideas from research papers to modify and extend their current systems to bring in parametric and anatomical models. Reeves (1990) uses a combination of a hierarchically defined skeleton and a muscle-based face for his animation. As more animators see the benefits of parametric models for facial animation, it is certain that they will begin to demand and make full use of such systems.

3.2. Applications - User Interfaces

Improving user interfaces is one application of facial animation that is very likely to gain acceptance. It would be hard to find a computer user who is fully satisfied with the interface that they use. The problem with human-computer interfaces is that it takes a long time to learn how to use each new system. Most interfaces are completely alien to the novice user. Presenting users with a familiar interface, the human face, will create a more comfortable environment for users to work in. Coupling this up with a speech recognition/synthesis system would take away the embarrassment users feel when faced with a keyboard. Welsh (1990) and Morishima (1991) give the development of user interfaces as one of the major projected uses for their facial animation systems. Hardware and software for speech recognition will have to be developed further before this type of user interface can become commonplace. Hopefully we wont have to wait too long.

3.3. Applications - Medical Research

The main uses for facial animation in medicine will be in the surgical and psychological areas. Parke (1982) predicts that parameterised facial models may become aids for previewing the effects of corrective surgery or dental procedures on patients. This type of application would need a very accurate anatomical model of the patient's face and a means of indicating what changes will take place. Waters' (1987) view on pre-operative techniques is that, "Surgical reconstruction of faces uses a number of techniques to collect 3-D data: Moire patterning, lofting of CAT or EMR scans and lasers. The resultant data can vary enormously from one face to another, and so any resultant parameterisation would, at best, be tedious to implement." This is not to say that it won't happen, Waters is admitting the difficulty of such work.

The use of newly developed facial animation systems by psychologists for researching facial movement and expression is a logical move. Since around 1982, most facial animation systems have used Ekman and Friesen's FACS (1978) as an anatomical guide when constructing the facial model. Research by computing people has thus given the psychologists a plethora of graphical implementations of their theories. Now they have the opportunity to supplement their research with computer models of facial movement rather than having to use photographs or train people to fire muscles at will.

3.4. Applications - Teaching and Speech Aids

One application that has already been tested is the use of a facial animation system as a teaching tool. There are many people in the community who could benefit from a different method of teaching, especially in the language and speech area. Teaching people the correct way to pronounce words is a tedious and repetitive process. This process is often made much more difficult when the student has a speech or hearing disability. Instead of having labour intensive, one-on-one tutoring, the student could work at their own pace with a computer simulated teacher. The student's pronounciation could be tested and feedback given as to how they can improve their speech.

Teaching people with hearing disorders to lip-read could be made easier with an appropriate facial animation system. Not just lip-reading, but teaching the deaf to speak is a noble and highly likely application. Mouth positions could be copied from the computer model and feedback on how well they're speaking would make the learning task a lot easier. People with disabilities would most likely welcome the opportunity to be able to teach themselves communication skills. The satisfaction of being able to teach themselves along with the the skills learnt through an instructional system would make such developments very worthwhile.

One of the applications of HyperAnimation (Gasper, 1992) is Talking Tiles. This program aims to be an aid in teaching language skills. It is phoneme based and is thus language independent. Talking Tiles is mainly for younger people to give them an interesting tutorial tool to teach them how to put sounds together. The player aims to sort out anagrams of words. Words are represented by a series of tiles that can be swapped around. The phonemes can be heard tile by tile and then the final word can be sounded out as a blended combination of the component tiles. By making the learning process seem like a game, the student's attention is held and they can learn more than they would using traditional methods. Lessons in foreign languages would be a matter of adding an extension to the vocabulary.

Teaching correct pronounciation is a time-consuming task. It is usually boring, both for the student and the teacher. Hiding the lessons within a game makes learning more enjoyable. Using a computer as a teacher gives more freedom to the teacher to do less mundane work, and makes it possible for more than one student to learn at one time. Most importantly, the student can learn at their own pace and will not be embarrassed about redoing an exercise they feel needs more work.

3.5. Applications - Criminal Identification

Improving methods of identifying people could help with criminal investigations world-wide. The FACE system developed by Vision Control Australia and the Victorian Police (Eaves et al, 1990) is not really an animation system. It uses a lot of the knowledge from research done on facial animation along with the experience that police have had with the Identikit and Photofit identification systems.

The FACE system uses a series of overlays, similar to those in the Identikit, to create a facial image. A database of facial components from various ethnic groups is available and can be added to. Once the face is put together, different parts of it can be selected and altered to get the best possible fit. Parke (1982) described a possible identification system where a 3-D facial model could be manipulated to match the witness' description. The main advantage of this would be that a 3-D image would result in a more accurate description of the criminal. This is especially true in cases where the witness didn't get a front-on view of the offender. The possibilities are there, whether such a system is viable remains to be seen.

3.6. Applications - Communications

The application getting the most support from industry is the use of facial animation to reduce bandwidth when transmitting facial images. British Telecom, DEC and Sony are some of the companies doing research in the area. By transmitting parameter data related to movements rather than complete images, reductions can be made on the amount of data being sent through communication lines (Parke, 1982). "Low bandwidth teleconferencing ... requires the real-time extraction of facial control parameters from live video at the transmission site and the reconstruction of a dynamic facsimile of the subject's face at the remote receiver" (Terzopoulos & Waters, 1990). So either facial movement analysis or speech recognition systems need to be developed further so that the communications field can reap the benefits of facial animation. Teleconferencing and videophones are the two main applications being looked at by people in communications.

4. TECHNIQUES OF FACIAL ANIMATION

Before implementing a facial animation system, the projected uses must be thought through. For this thesis, a flexible, real time system is required. It needs to give realistic output and be able to take input from text files as well as speech tracks. To make it easier to set up and test the system, an anatomical base would be most suitable.

For all animation methods, the most important features of the face are the eyes and mouth. This is not just because they tend to move the most. Studies have shown that when we look at other people's faces, we spend most of our time looking at the eyes and mouth (Morris, 1982). The results of tracking a person's eyes while looking at a picture of a human face illustrate this point. (see Fig. 4.1) Even though setting up a system for speech synchronisation concentrates on the models mouth, we should not forget to animate the eyes as well. Setting up the system to do occaisional eye movements would probably be enough. The timing for these eye movements would have to be thought out. It is very discomforting to talk to someone who blinks too often, or not often enough. Incorporating slight head movements would be another way of making the system more realistic. These are some of the features that should be available in a facial animation system, if it is to give a realistic result.

There are three main approaches to facial animation: key-framing, parameterised models (Parke, 1974) and physically based models (Waters, 1990). Each method has good and bad features, depending on how fast and how flexible you want the system to be and what you intend to use it for.

4.1. Key Frame Animation

Key frame animation is used extensively in conventional computer animation systems for simpler types of animation such as character animation (Lasseter 1987). The method used in key frame computer animation is to completely define a model by its position and rotation for specified key frames. (see Fig. 4.2) Key frames are

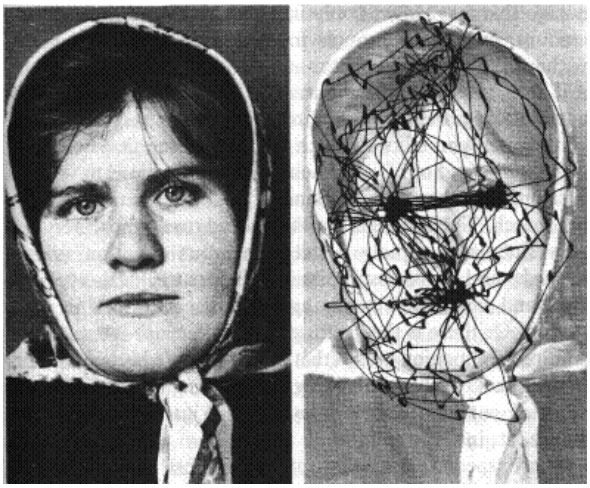

Fig. 4.1 : Tracking the eye movement looking at a face

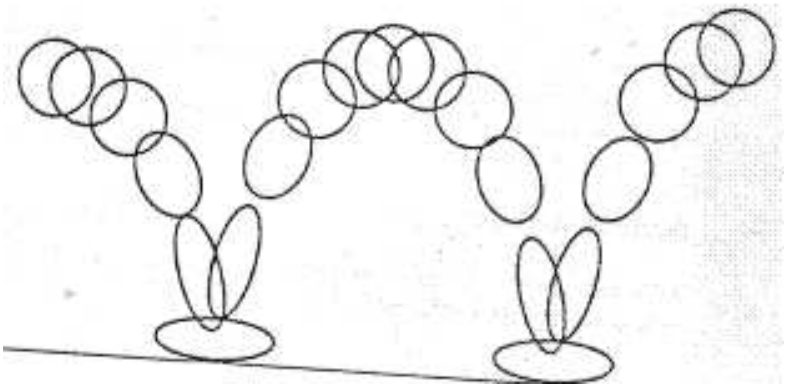

Fig. 4.2 : Inbetweening, Lasseter and Rafael (1987)

separate time instants that the animation system uses to produce
motion of the model. The model's definition between key frames is generated by applying some interpolating algorithm to the key frames, giving the complete animation of that model.

Although key frame animation has been used successfully in 2-D animation systems it is too inefficient for 3-D animation (Waters 1987, Parke 1982). For each key frame a complete specification of the model is required and each change in the model, no matter how small, requires every element's position to be specified for the whole model. For a complex 3-D model with a lot of frames this becomes too costly.

4.2. Parameterisation

Parameterisation is the main technique used in 3-D facial animation. The parameterisation concept takes the individual parts of a model and combines them together in different groups having common criterion or parameters (Parke 1982). Each member of a group can be described by some variation on those parameters. As an example take the set of possible facial expressions. For the set to cover all expressions it would have to have enough parameters to successfully describe any expression you desired. The parameters set could include: pupil size, mouth width, mouth height, etc. depending on the level of realism you required. This set could be broken up into smaller, more manageable sets as the facial expressions become more complex. There is no end to the possible groupings of muscles. With so many possible groupings of parameters, it would be impossible for us to develop a complete parameter set for facial expressions.

The advantage of parameterisation over key framing is that parameterisation allows the animation of a model to be performed as manipulations of specific groups. Small movements can be treated as the re-specification of small groups rather than the whole model. Although more economical than key framing the method used in facial animation to date is not general. One parameterization model can't be used to describe a totally different facial topology (Waters 1987). This is due to the unbounded characteristics of the set of possible faces and their expressions.

Most facial parameterisation models are specific to the current model being animated. These use mainly expressive types of parameters with little, if any emphasis on the facial structure, and are therefore fairly simple and efficient. Parke (1982) works on a more general facial parameterisation method by using conformation

(structure) parameters as well as expression parameters to describe sets of facial objects.

4.3. Physically Based Models

There are varying levels of complexity possible for anatomically based facial animation. Waters created a muscle model in 1987. Another anatomical method used for animating the face is the dynamic simulation of facial skin tissue. Terzopoulos and Waters (1990) have produced a model that simulates the motion of facial skin tissue using Waters' muscle model (Waters 1987) as a base. The physical basis of the model gives an added benefit of automatic error checking. The only movements possible within the model are those that are possible in real life.

The facial skin tissue model of Terzopoulos and Waters (1990) produces improved simulation of the deformable properties of skin tissue under the forces produced by the muscles. (see Fig. 4.3) The biggest advantage of this model is that it can run at interactive speeds while producing more realistic images. This is a direct result of the small parameter set required to describe a wide range of facial expressions.

Terzopoulos and Waters (1990) describe the model as a six level hierarchy of decreasing data abstraction. The expression level executes expression commands in terms base expressions, time intervals and emphasis. The next level is the control level which converts expressions into coordinated movement of the facial muscles. The third level and this level describes the properties of the different facial muscles. Below the muscle is the physics level, containing the physically based facial tissue model which is acted upon by the activated muscles. The fifth level, the geometry level, is the geometric representation of the model which is acted upon by muscle activation and the resulting skin deformation. The last level is responsible for the images. This level uses various graphics techniques to build and render the facial image using dedicated graphics hardware. This allows continuous facial representation at interactive rates.

The main difference between this model and others is its physical base. Terzopoulos and Waters (1990) discuss the structure and properties of real facial skin tissue derived from medical research. They then introduce a mathematical model that uses spring dynamics to simulate the properties of facial tissue. The structure used to

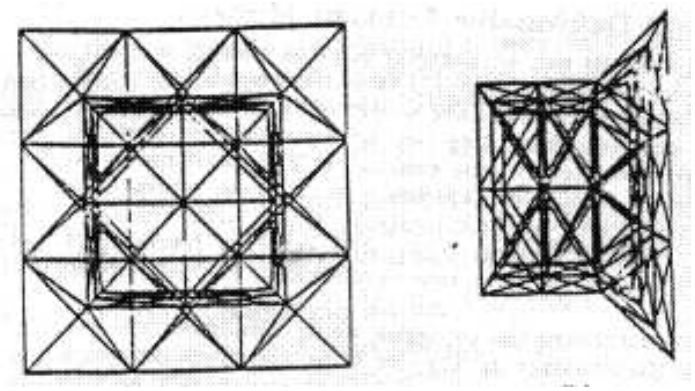

Fig. 4.3 : The Lattice Skin Tissue Model, Terzopoulos and Waters (1990)

physically represent the facial skin model is a tri-layered deformable lattice of point masses connected by springs. Each layer represents the corresponding layer of tissue in a real face. Stiffness of the skin and other properties of each layer are taken into account as spring variables. Each layer is connected by the springs as individual nodes in a layer are also thus producing a stable interconnected lattice. Terzopoulos and Waters (1990) have experimented using this model with real time video, producing promising results.

5. EXAMPLES OF EXISTING FACIAL ANIMATION SYSTEMS

A wide range of facial animation systems have been developed over the last twenty years. Most of them have built on the initial research done by Fred Parke in his 1974 thesis. (see Fig. 5.1) Keith Waters has done a lot of work in the area, publishing information about his physically based system in 1987 and improving on it in the time since then.

Waters and Parke are far from being the only ones making headway in facial animation. DiPaola (1991) is creating a more general facial animation tool to give more freedom to the animator. There is a growing trend

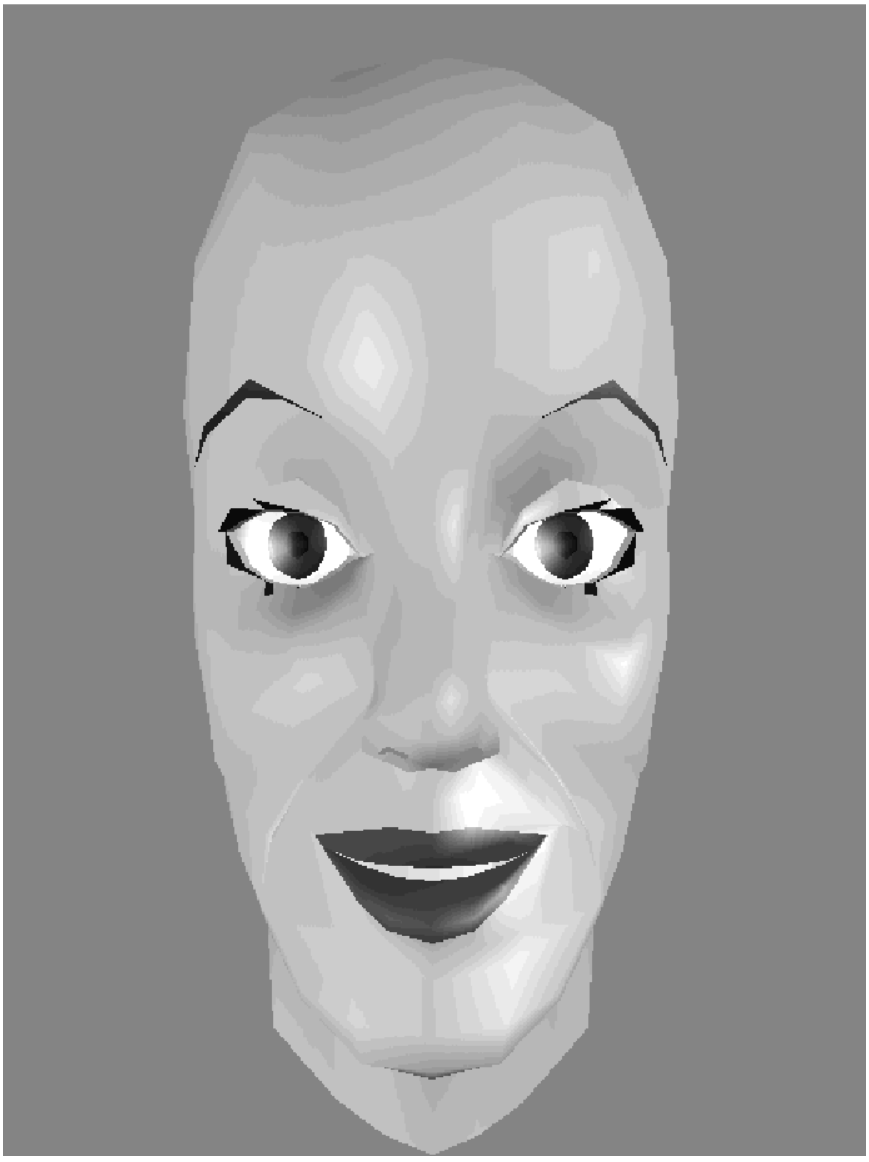

Fig. 5.1 : Neutral face

towards systems that use texture mapping to give a more true to life finish to the output images. Researchers looking into this area include: Yau (1988), Morishima (1991) and Waters (***). Automatic animation is another area undergoing a lot of research. Williams (1990) uses facial movement to drive his system. Speech driven systems have been developed by Welsh (1990), Morishima (1991) and Lewis (1991). In Japan they are working on integrated systems with facial animation being a fundamental part of the user interface (Gross, 1991).

5.1. Parke's Research

Fred Parke has been working in the area of facial animation for almost 20 years. His published works include a PhD thesis and numerous papers.

5.1.1. PhD Thesis

This is the research that sparked the initial interest in facial animation. Parke set about finding a simpler, more flexible model for facial animation. He aimed to develop a system where the user could manipulate a face by inputting a set of parameters, rather than having to define each vertex in the image (key-framing).

The basic idea that makes this work possible is interpolation. By using a coefficient or parameter, the programmer can determine a point between two extremes with the simple expression: x =a(p1) + (1-a)(p2). This idea can easily be expanded into three dimensions. This idea can be used to morph between objects, given that their topologies are the same. Thus, if the topology of a face is fixed, then interpolating between facial positions is a matter of evaluating the mathematical expressions associated with each vertex. The author recorded data from a real face and manipulated it to check that his theory was correct. From his experiments, he found that, indeed, it was possible to use one topology for a moving face.

When developing the parametric model, Parke divided the parameters into two main categories: those controlling facial expression, and those altering the basic shape of the face. The face itself is symmetric in this model. The manipulation capabilities are implemented using parameters to control the interpolation, translation, rotation, and scaling of the facial features. The expressions in the model are mainly a result of the movement of the eyes and mouth, as is the case with real faces. As well as being able to open and close the eyes, the user is able to define the direction in which they are looking. This is crucial in making them a believable imitation of the real thing. For the mouth, teeth were made to give a bit more realism when the mouth was opened. Conformation parameters simulate the differences between faces, for example, nose shape, rather than the change in expression on a single face.

Once the model is set up, the next step is to work out how the parameters should vary over time. The theory behind this is taken from traditional animation techniques. From them, it is possible to find out what movements are involved in each facial expression. Medical books are also a good source of information about facial movements. Speech synchronisation involves matching the movements of the mouth with a recorded speech track. Many levels of animation were attempted. It was found that when six parameters were involved (these included eyes and eyebrows) the result was at least on a par with most conventional speech animation.

Parke concludes that the most useful parameters are the ones involved in mouth, eye and jaw movement. He states the symmetry of the model, although easing the complexity of the model, is a deficiency as it cuts out a lot of expressions and reduces the realism of the image. He refers to the parameterised system as an instrument we do not yet know how to play.

5.1.2. Later Work

Parke went on to refine his animation system by giving it a more anatomical basis. To create a parameterised facial model, the designer must first create the parameter sets. These parameters define the adjustable parts of the face. (movements as well as colour, size, distance and viewpoint) Once this is done, the synthesis model is developed to produce images based on the parameter values. It has two main parts: the parametric model (the data, algorithms and functions for image definition) and the graphics routines to give a visual interpretation of the data.

Again, Parke uses two broad categories of parameters: conformation or structural parameters, and expression parameters. Many of the ideas for expression parameters come from Ekman and Friesen's FACS manual. Conformation parameters let the animator change the shape and size of parts of the face. (see Fig. 5.2) In the process of developing this animation system, the writer found that, the more realistic his model got, the pickier the people he tried it out on became.

The topology of the facial model does not change, just the positions of the vertices. The parameters can be entered into the system interactively or, for animated sequences, through command files. Five types of operations determine the vertex positions for the image. They are: procedural construction (for the eyes); interpolation (to get the position of a vertex between two extremes, depending on the parameter value); rotation (for jaw movement); scaling (for changing the size of specific features) and position offset (to move regions of points, as in the corners of the mouth). The final image is shown with skin coloured Phong shading. Some examples of output images are given. The author believes that the main benefit of parameterised facial systems is that they abstract the animation process and make it simpler for the animator.

5.2. Water's Research

Keith Waters has made a name for himself in the field of facial animation. He takes an anatomical approach to facial models, first with his muscle model, and secondly with the tri-layer tissue model.

5.2.1. The Muscle Model

To develop the parameter sets for the face, Waters worked with the Facial Action Coding System (FACS). FACS provides a notation-based environment with a set of Action Units (AU) to represent muscle groups which work together to produce a single movement. The combination of movements that can be produced by the Action Units create the expressions we know and love. The goal of this research is to model the basic facial expressions (anger, sadness, happiness, fear, surprise and disgust) and test them using the FACS system to validate the results.

To give a realistic model of the face, the anatomy of the human head had to be studied. Firstly, the bone involved in the face. The only moving part is the jaw, which rotates about an axis. The muscles work above the bone, and are often attached to the bone at one end. Waters divides the face into upper and lower sections, stating that the most complex part of the face to model is the mouth. He models two types of muscles: linear/parallel and sphincter. Each node on the facial mesh is affected by one or more muscles, so its position at any point in time is defined as a function of the parameters relating to those muscles. The intricacies of the skin and facial tissue were not taken into account in this model. For example, the effects of ageing on the elasticity of the skin and the differences in the amounts of fatty tissue. To get an idea of a typical layout of the muscles and their

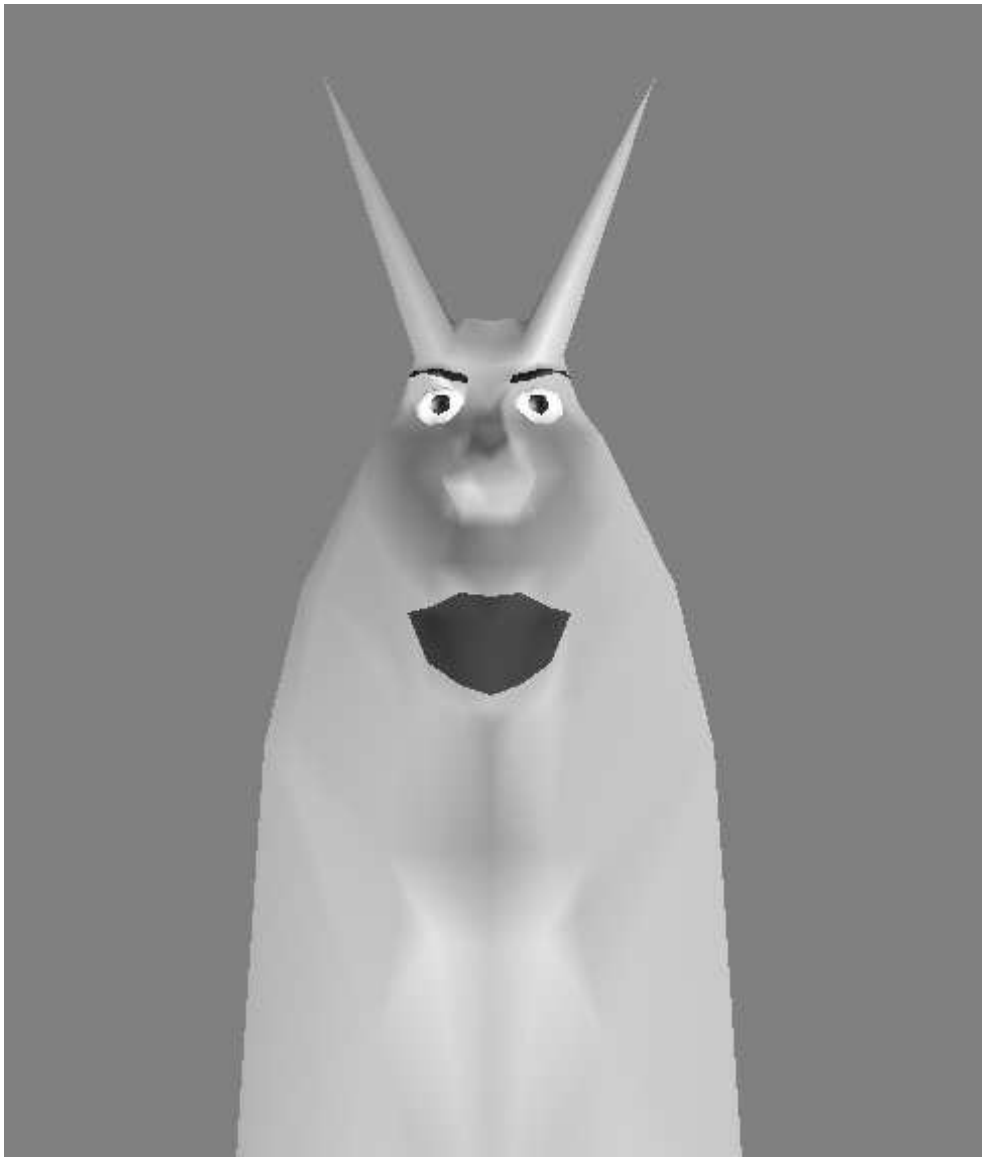

Fig. 5.2 : The effect of the growth factor parameter.

attachment points on the face, accurate measurements were taken of several people. This gave Waters an idea of what differences were likely between people and what the "average" face would look like.

To model the individual muscles, each one was represented as a vector whose magnitude is the pull that the muscle is exerting. For each muscle, a fall-off function, a zone of influence and a maximum movement is defined. The model uses only ten of the muscles of the face to produce its output. By inputting parameters, the user can make the model go through any humanly possible facial expression. The result of this research is a system capable of making believable expressions using parameter input.

5.2.2. The Tri-layer Model

The facial model is a hierarchical system which lets the user control parameters for facial movements at six different levels. These levels are: expression, control (of muscle groups), muscle (individual muscles), physics, geometry and images (light sources and colour choices). Thus, using higher control levels, the user can work with the model at an abstract level without needing to know about the complexity of the underlying system.

The facial tissue model is based on the real thing, with the effects of the epidermis, dermis, subcutaneous fat and then the muscle being taken into account. The geometric model consists of three layers of tetrahedrons. The top layer is the skin, the second is the dermis and fat and the third is muscle. The result is a visual image that is not seeing the muscle movements directly, instead it is seeing them after they have been "filtered" or propagated through the layers of facial tissue.

Facial muscle control is based on the action units outlined in FACS. The muscles themselves can be of three main types: sheet, sphincter and linear. Examples of each include: sheet - the muscle that raises the eyebrows, sphincter - the muscle that pouts the lips, and linear - the zygomaticus major which raises the corner of the mouth. In this model the muscles work through the third layer of the mesh. For each facial movement, all of the effects of all the muscles have to be computed. The model is created with the epidermis as the start point. From the epidermis, the structure for the two lower layers is created using the normals for each polygon on the epidermis.

To track the facial movements, a real model is videoed and this video is fed into an image processing system. To aid in the visibility of the facial features, lines are drawn on the face in strategic places. These lines make it possible to track: the head position - using a line along the hairline; the movement of the zygomaticus major - using the movements of the endpoints of the mouth; nasal movements - using the curve of the nostril; eyebrow movements; and jaw rotation - using the line of the lower edge of the chin. These positions are computed relative to the hairline. To start the dynamic image system, the model's face is processed while in a relaxed state to give an idea of the resting lengths of all the muscles. Using the neutral face as a reference, the facial movements can be approximated on the computer generated image. This has worked in practise to generate a very effective result.

5.3. More Recent Research

Research into developing better techniques for facial animation is being done all over the world. Ideas in the pipeline include: extending the range of facial types available; texture mapping for greater realism; analysis of facial movement and speech to automate animation; and integrated systems where facial animation is a component of a graphical user interface.

5.3.1. Extending the Range of Facial Types

Steve DiPaola (1991) takes a different approach to the area of facial animation. He looks into the creation of an animation tool which lets the animator alter the structure and the rendering of a facial model to a much greater extent than previous models. He aims to create a tool that is more general than current systems. His work is based on Parke's animation model.

Once the author had implemented the natural movements the face is capable of, he went on to expand the system to include more unnatural possibilities. DiPaola thought in terms of what animators would like to be able to do with an animation system. He gives the animator full control of texture and colour, as well as more freedom with the facial movements. Using techniques from traditional animation, he improved the system by adding extra movements and transformations, some of which would be impossible in reality. An example is the ability to scale up facial features; an eye scaling up to half the size of the head. These added movements affected the surrounding area in different ways to the standard facial movements. The techniques for implementing these extra movements were taken from traditional animation and observation and research into animal physiology. Another parameter the author added is used for warping the whole or parts of the face. The warping function can have a subtle effect, as in a warp to produce a facial crease, or a blatant effect as in turning the head into a corkscrew. Stochastic noise deformation is also available to let the animator randomly alter an input face, to create a variety of new faces. It can also be used to distort the face, giving similar effects to warping.

Future areas of research for this system will include the use of patches, developing a hair model that can incorporate a variety of hair styles, and techniques to easily modify and animate a large range of facial wrinkles, furrows and bulges.

5.3.2. Texture Mapping

Yau (1988) outlines a technique for animating the face which gives a more realistic texture than most facial animation systems. Their technique involves taking images of a real face and projecting them onto the surface of a 3-D object. This method is aimed at overcoming one of the problems with facial animation systems: cartoon-like characters that don't look very realistic.

Two 3-D models are used in Yau's animation system. One of them is dynamic and is used to find out the positioning and movements of the face. It is used just before output to the screen to set up the transformation matrix. The second model is static and has the texture mapped onto it. This method speeds up the mapping process before the image is transformed and the lighting calculations take place. The problem with this method is that an open mouth can be mapped onto a closed mouth, which can look pretty silly. Yau gets around this by having some basic facial movements included with the (semi) static model. These are used in cases when the mouth and eyes go through major movements.

Similarly, Morishima (1990) uses a 3-D wire frame model and maps a 2-D texture onto it. Points of importance on the 3-D face are matched up to corresponding points on the texture map using an affine-transformation. The authors set up 17 phoneme positions for the face. The model includes teeth, and the movements of the teeth follow directly from the jaw movements.

Other people looking into texture mapped animation systems include Williams (1990) and Waters (1991). Waters has simplified his tri-layer tissue model to make it possible to run the texture mapped system in real time. Texture mapped systems are quite slow and restricted, but can give some very realistic results. More flexibility will come as more movements can be made in the static model.

5.3.3. Movement Driven Systems

Williams (1990) set out to create a system for animating a face using video input. It extends upon the work done by Parke and Waters by attempting to map texture and expression with continuous motion as input. Using current technologies, both human features and human performance can, in Williams opinion, be extracted, edited, and abstracted with sufficient detail and precision to serve dramatic purposes.

To create the model, a real head was sculpted in plaster and photos were taken from different angles. The scanned data, along with the photographic information, was used to create a warping rule for texture mapping. The result was a cylindrical texture map that could be wrapped around the 3-D facial image. The final model can be stretched in an unrealistic way, resembling a latex mask in some respects.

To do the animation, small dots were put onto the model's skin and then tracked as she went through some facial expressions. Using this as input, the computer generated face copied each change in the facial expression. The reference points on the model were duplicated on the computerised face and the calculations for each movement of a reference point resulted in the appropriate alterations to the facial mesh. This work is a proof of the concept being valid, and will be continued and expanded on in the future.

5.3.4. Speech Driven Systems

One of the most difficult problems in facial animation is speech synchronisation. Sub-standard synchronisation can make an otherwise perfect piece of animation look ridiculous (remember the spaghetti westerns). Traditionally, this problem has been handled using two methods: rotoscoping, where a live model is recorded on video and the animators copy their movements frame by frame; and canonical mapping, where the mouth shapes for each phoneme and/or expression are taken from an animation handbook and then they are formed on the animated face (Lewis, 1991). Both of these methods are highly time consuming as they have to be done manually. Current research is trying to solve the problem of how to automatically obtain mouth movements from a recorded soundtrack.

One technique for speech analysis is the source-filter speech model (Lewis, 1991). In a source-filter model, the speech track can be separated into it's components: periodic harmonics, which are constant and come from the vocal chords; and vocal tract filters which create formants within the sound. Formants alter the speech spectrogram plot, and can thus be identified through a plot. Each formant corresponds to a certain combination of mouth and vocal tract movements which produce each phoneme. From this information, each phoneme can be produced on an animated model. An important feature of this model is that it separates the phonetic information from the intonation. Thus the loudness and softness of the sounds do not affect the formant information. A system that can output a script of phoneme information is a suitable starting point for automated lip-synch.

The most simple technique for automating mouth movements is "loudness equals jaw rotation". As the loudness of the sound increases, the mouth opens wider. This is not really satisfactory as there are sounds that can be made with the mouth shut and the animation tends to look robotic. Another technique is spectrum matching. This involves putting the soundtrack through filters and matching the resulting spectra with reference sounds. There are problems with getting a good match if the pitch of the speech varies. A different approach is speech synthesis. In this approach, the animation and synthesis systems are coupled together. They accept a script of text and each one responds appropriately. The problem here is in the speech synthesisers. Their output isn't very realistic as far as intonation and flow of speech go. The feasibility of this type of system will improve as more and more work is done in the speech synthesis area.

The method favoured by Lewis (1991) is the linear prediction approach. It is a special case of Wiener filtering and involves separating the sound source and vocal tract components of speech. Supersampling the speech is advised for best results. This means that the speech should be analysed more times per second than the frames per second required for the animation. Supersampling helps to keep the facial movements smooth by reducing aliasing. Equations, proofs and references for implementing this method are given in Lewis's 1991 paper. For speech synchronisation, the phonemes have to be within an error bound of a set of reference sounds. Lewis advises concentrating on vowels as they are easier to identify on the speech spectrum. They are usually longer than consonants, so they take up more of the time during speech. Most consonants have very similar

spectra and mouth positions, and for others, there is no set mouth position (Lewis, 1991).

Papers by Welsh (1990) and by Morishima (1991) outline their systems for speech driven facial animation. In Welsh's system (1990), the mouth shape is parameterised using height and width. For each image frame there are two speech frames, which smooths the facial movements. The mouth has 16 possible positions, and each transitional movement from one position to all of the others is given a probability. This is an aid in determining which mouth shape is going to come next by giving a low probability to those that rarely occur. Welsh hopes that they will be able to produce a speaker independent system in the near future. Morishima (1991) uses two methods of voice to image conversion. The first is vector quantisation, and the other is synthesis by neural network. The output of each of these converters becomes the input to the image synthesis system.

5.3.5. Integrated Systems

In Japan, they are working at building better graphical user interfaces (GUI's) (Gross, 1991). The future for GUI's is to progress from the current electronic desktop to the virtual office. As part of the research into producing better GUI's, programmers at Sony are working on System G, a real time video animation and texture wrap system. The most striking demo for the system is a Kabuki mask which can be animated and fully rendered in real time. The mask has a huge repertoire of facial movements. It even has a tongue, unlike most other facial animation systems. The demo proceeds by taking input from an organ and the face sings along with the music with a one frame delay. This facial animation research will serve as a basis for work into the recognition of facial expressions, spoken commands and body language.

In the Visual Perception Laboratory (VPL) within Nippon Telegraph & Telephone (NTT) they have three related research projects in process: computer recognition of people from their faces; facial expression recognition and lipreading and a group working to find interesting uses for optical character recognition and image processing technology. The combination of these projects has already produced a very effective, Max Headroom style interface which can hold a conversation. The technology is also being used to recognise number plates and some basic facial movements.

In all the Japanese are taking a long term approach to bringing virtual reality to the computing industry. They are planning their research, and adopting what they learn into their current systems, easing the task of conversion which is bound to come later. The products that the Japanese are producing on their way to mastering virtual reality are being worked into current systems for long term benefits, rather than producing spin-off solutions that can be applied to immediate problems. For the Japanese, facial animation is a part of the big picture for GUI development.

6. THE PILOT PROJECT AND BEYOND

This thesis follows on from a group graphics project carried out in 1991. The aim of the group project was to do preliminary research into facial animation. Starting with a few landmark articles, the research base was expanded by tracing through the references given in each article. The project also tested out different methods of recording 3-D facial data.

The project report summarised the information gained through reading literature on facial animation. The anatomy of the face along with basic theories on facial expressions are explained. The report then looks at the history of facial animation, outlining different methods and their benefits. Most of the information is based on work by Fred Parke and Keith Waters.

Two methods of digitising facial data were used. One method used two right angled photographs. The photographs were scanned into a Personal Iris and displayed on the screen. A program, facesave, was used to record the 3-d data point positions and the facial topology. (see Fig. 6.1) This method gave quick and dirty, but still satisfactory results. The second method used photos taken from slightly different positions. Slides of these photos were put into an analytical stereo digitiser to record the facial data. Once these points were recorded, a program was written convert the data into a suitable form for triangulation. A few more adjustments were made to the data before it could be used. A full description of both methods and copies of the program are included in the project report.

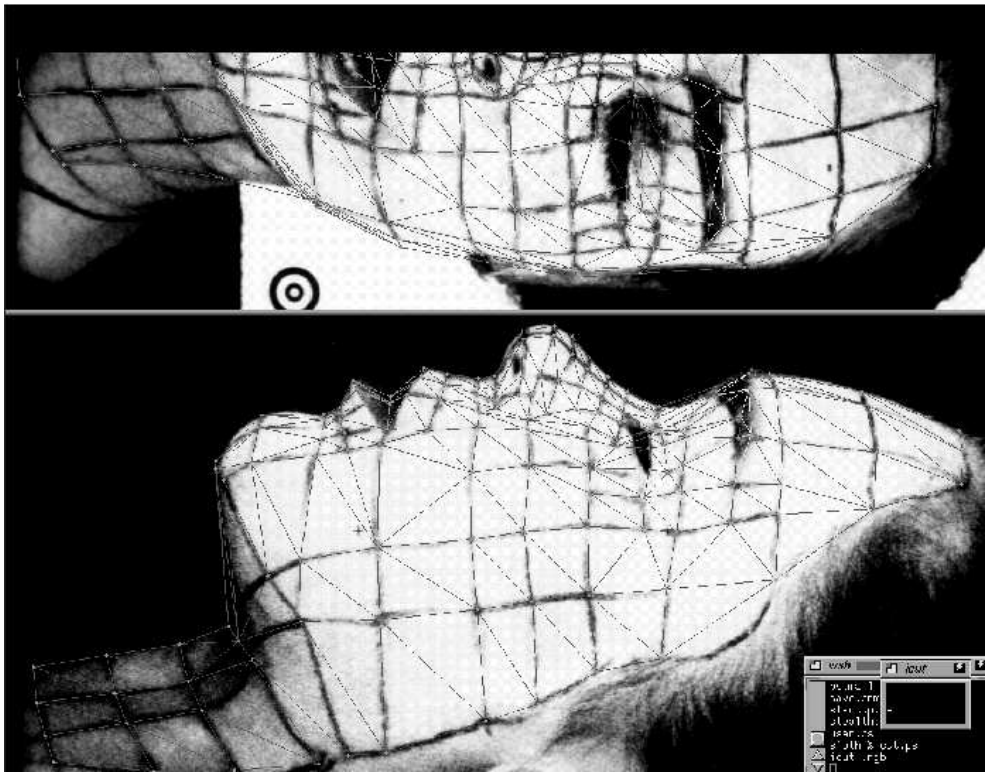

Fig. 6.1 : The screen display for FACESAVE

The project was a success in that it gave insights into what is involved in developing a system for facial animation. Our supervisor, Andrew Marriott, made contact with Fred Parke, who kindly sent a copy of his fascia model to us. Andrew wrote a program to manipulate the fascia data, as well as our facial data. This has been made available for anonymous ftp through the Curtin University computer network.

Since the pilot project, work has continued on facial animation at Curtin. The fascia program has been improved upon, it can move each side of the face independently and has a more elaborate interface. Andrew Marriott has been using fascia to record an animated introduction for the Artificial Intelligence and Simulation conference 1992 which is being held in Perth. He is using a video of a person going through the required movements as a guide for positioning the face for each frame.

I have had some contact with Waters over the last few months. He has given some advice and references to help my thesis along. Other people have been very helpful. Steve Franks has given information about what is available in facial animation and who else is working in the area. His versions of Waters' anatomical facial model are available for anonymous ftp. Email addresses and ftp sites are in appendix ****.

As an aid to my research I will be attending SIGGRAPH 1992 in Chicago. I hope to be able to meet up with other people working in facial animation while there. There will be a special panel on animating human figures that I expect will be highly informative.

7. RESEARCH OUTLINE

The final product from this thesis will be a facial animation system that will be speech driven using text files or real time audio input. At this point, it is difficult to judge how sophisticated the system will be. The animation programs that my thesis will use will undergo some modifications, but, on the whole, they are already well suited to this application. (see Fig. 7.1) The handling of audio input will have to be approached from scratch. There are many different methods for carrying out speech recognition, many of them requiring specialised hardware. Thus, the available resources will be a constraint on the finished product.

This thesis will serve as a basis for further research into facial animation at Curtin. One of the proposed applications for the system is to improve the user interface for communications between computer users. Instead of reading a text message, the user will see and hear the message as it is relayed by a talking head. This is a new area of research with many possible applications. Thus, any new ideas coming from the research will help in the world-wide search for more efficient and robust means of automating speech synchronisation.

7.1. Proposed Research

My work will involve adding new features to existing programs. The animation systems developed by Fred Parke and Keith Waters are available for public use, and will provide a base for my research. Both of these systems are stand-alone animation systems, and will need to be altered to suit my thesis. There are several speech driven facial animation systems that have already been developed in research centres around the world. These systems were built by people under the wing of major companies, British Telecom and DEC, so, although it is possible to build a very sophisticated system, the resources I have are not comparable with those of other

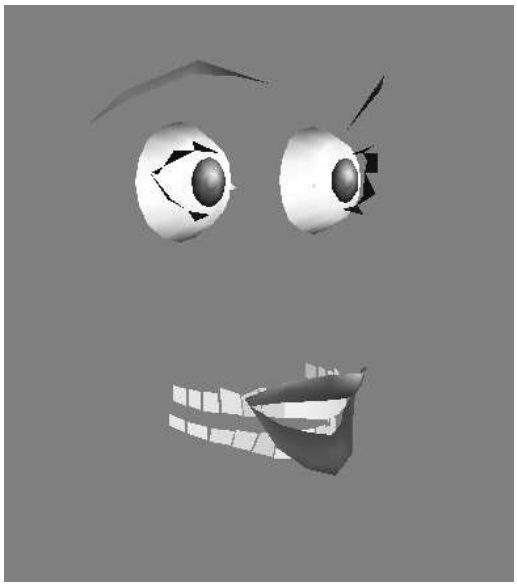

Fig. 7.1 : Parke's facial model minus the skin

researchers in this area. My system will be a simplified version of these elaborate systems. I will be using the information that the developers of these systems have given in conference papers as a guide on how to create a similar system.

I plan to carry out my research in four stages. Through all of the stages of my research, I will be be making changes to the animation systems that I use. These alterations will aim to make the system more efficient, adapt the user interface to my needs and give debug/trace information. The proposed stages are detailed below:

Stage 1: Find information about phonemes and the basic facial expressions. Implement this information as a series of "macro-movements" to give a higher (more abstract) level of control over facial movements.

The macro's must be able to work concurrently with each other so complex expressions can be made. For example, a smile macro could go for a full sequence of animation while other macro's are used to implement movements for speech. A macro for blinking will produce blinking at random times to give a more natural feel to the model. The time period for each macro will have to be included in the macro call.

Stage 2: Set the system up so that the macro-movements can be accessed via a text file of phonemes.

For each macro, there will be a corresponding phoneme or action. I will try to find a fairly "standard" notation for phonemes to use in the text file. Meaningful names (smile, frown, wink) will be used for the other macros.

Stage 3: Link the animation system up with a speech synthesiser. Both systems will be taking textual input.

Hopefully the phoneme notation used in Stage 2 for text file input will be fairly similar to that used by available speech synthesisers. In any case, the conversion should be fairly trivial. An interactive method of input for facial expressions to allow manipulation of the face during speech.

Stage 4: Investigate and implement audio input for the facial animation system. The sophistication of the system's phoneme recognition is dependent on the resources available.

The base level for direct audio input would be to use volume as the only variable and simply open the mouth wider as the input becomes louder. As I continue to research this area, I hope to find more information on how speech recognition is done, and how it can be implemented using the resources I have available.

The work will be done on a Silicon Graphics Indigo, making use of the audio and graphics libraries it has available. Theoretical input and advice will come from conference proceedings as well as email contact with other researchers. Personal interviews with experts in the areas of speech synthesis and recognition, digital signal processing (DSP) and graphics areas will also be undertaken as is necessary.

8. REFERENCES

Anderson S.E. (1990) *Making a pseudopod: an application of computer graphics imagery.* Proc. Ausgraph '90: 303-311.

Bergeron P. (1988) *Artificial intelligence and computer animation.* Ausgraph '88: 105-106.

Bergeron P. (1990) *3-D character animation on the symbolics system*. 3-D Character animation by computer (course notes), AUSGRAPH 1990, Melbourne: Australia.

An informal description of the author's method of character animation. He uses two graphics systems: S-Geometry for space-related work (modelling) and S-Dynamics for time-related work (animation). This is a good article to aid in understanding what's involved in character animation, and is an example of a methodical approach to the problem.

Bergeron P. (1985) *Controlling facial expressions and body movements in the computer-generated animated short "Tony de Peltrie"*, tutorial, SIGGRAPH 1985.

Outlines a method of character animation. The body is animated by getting data from a clay model, and then manipulating the resulting hierarchical skeleton using the TAARNA 3-D graphics system. The facial animation was done by mapping data from expressions made by a human model onto the character's face. All of the speech phonemes were photographed and transferred onto the character.

DiPaola S. (1991) *Extending the range of facial types.* The Journal of Visualization and Computer Animation 2 (4): 129-131.

Doak R., F. Fleming, V. Hall and H.Hillyer (1991) *Facial animation and speech synthesis.* CGI 351 Report, School of Computing Science, Curtin University of Technology: Perth.

Eaves J. and A. Paterson (1990) *FACE - Facial automated composition and editing.* Ausgraph '90: 329-333.

Ekman P. and W.E. Friesen (1978) *Investigators Guide for the Facial Action Coding System.* Consulting Psychologist Press, Palo Alto: California.

Ekman P. and W.E. Friesen (1975) *Unmasking the Face.* Prentice-Hall Inc., Englewood Cliffs: New Jersey.

Aimed at helping people to recognise facial expressions in others. Gives lots of photographs of the six basic expressions: surprise, fear, disgust, anger, happiness and sadness. Points out the facial movements that make up each of the expressions. Also shows what happens when the movements are conflicting, which indicates deceit.

Ekman P. and W.E. Friesen (1977) *Manual for the Facial Action Coding System.* Consulting Psychologist Press, Palo Alto: California.

We have been unable to find this reference. This system is said, by Parke, to be the best basis for complete expression models. Is referenced in just about every article. It catalogues around 55,000 distinguishable facial expressions, with six primary expressions being named. (see "Unmasking the Face", Ekman and Friesen, 1975) These expressions are the result of the changes made to 66 Action Units. Most systems only

use about 50 of these units in animating the face.

Ekman P. and H. Oster (1979) *Facial expressions of emotions.* In: Annual Review of Psychology, 30, pp. 527-554.

Goes into the theoretical side of facial expressions. Refers to studies on the cross-cultural aspect of expressions; ie. interpretation of expressions is independent of culture. Also goes into the learning of expressions by children. Looks at ways of measuring the face and its movements, and their accuracy. The article has five pages of references for psychology papers on facial expressions.

Gasper E. (1988) *Getting a head with hyperanimation.* Dr Dobb's Journal of Software Tools 13 (7): 18.

Gross D. (1991) *Man and Machine.* In: Computer Graphics World, May 1991, pp47-50.

Kendall F.P. and E.K. McCreary (1983) *Muscles Testing and Function* 3rd edition. Williams and Wilkins, Baltimore: USA.

Goes through each area of the human anatomy, illustrating the normal functions of all the muscles in the body. There are diagrams and descriptions for each muscle group, showing what movements are controlled by each muscle. Of primary interest is chapter seven which covers the muscles of the face, eyes and neck. There are a lot of pictures of models' faces to illustrate what effect each muscle has on the face.

Lasseter J. and S. Rafael (1987) *Principles of traditional animation applied to 3D computer animation.* In: Proceedings of SIGGRAPH July 1987 , pp. 35-44.

Describes the basic principles of traditional 2D hand-drawn animation and their application to 3D computer animation. Describes the evolution of the traditional methods and how they effect 2D animation. Goes on to say how these methods can improve the quality of the current 3D computer animation.

Lewis J (1991) *Automated lip-synch: background and techniques.* The Journal of Visualization and Computer Animation 2 (4): 118-122.

Marriott A. (1991) *Personal interviews.* Lecturer: Computing Science, Curtin University, Perth: Australia.

A slightly prickly fellow with a weakness for chocolate cake. Gave us invaluable advice about the project before threatening us with garden gnomes. Wrote the conversion programs: Facesave, Face2mog, Face2fascia. Will do "Anything, Anywhere, Anytime" - for a price.

Marriott A. (1990) *Computer Graphics 252 - Course Notes.* Curtin University, Perth: Australia.

Masden R. (1969) *Animated Film: Concepts, Methods, Uses.* Interland, New York: USA.

Cited as a reference by Parke in his speech synchronisation paper. Aimed more at traditional animation, but can give helpful hints to computer animators.

Mitchell A.G. (1964) *Spoken English.* McMillan and Co. Ltd. London: England.

An Australian text about phonetics. Gives the symbols and sounds used in pronunciation. Shows the movements involved in speech and outlines the more complex attributes of human speech. An easy to read text book on the spoken word.

Montgomery B. (1991) *Personal interviews & help with photography & photogrammetry techniques.* Lecturer: Surveying and Cartography, Curtin University, Perth: Australia.

A man with a passion for photography and an ultimate desire to digitise a woman's body. Was involved in the creation of the big ram on the tourist bureau in Wagin and is the man to see about Photogrammetry.

Morishima S. and H. Harashima (1991) *A natural human-machine interface with model-based image synthesis scheme.* In: Proc. Picture Coding Symposium 1991, Tokyo: Japan, pp 319-322.

Morris D. (1982) *Manwatching,* Triad Paperbacks, London: Great Britain.

A comprehensive guide to human behaviour. Gives analysis of countless observations in many different countries of the way people act. Has plenty of pictures and virtual encyclopaedia of information on the human race.

Naftel A.J. and J.C. Boot (1991) *An iterative linear transformation algorithm for solution of the collinearity equations.* In: Photogrammetric Engineering & Remote Sensing, 57(7), July 1991, pp. 913-919.

Describes a means of solving collinearity equations which gives accurate results while using less computer resources than traditional methods. Gives all the equations required to use the methods outlined in the article, as well as comparing the root-mean-square error when working with subject matter of differing complexity. Concludes that the iterative linear transformation procedure gives fewer errors than the direct linear transformation method.

Parke F.I. (1975a) *A model for human faces that allows speech synchronized animation.* In: Computer and Graphics, 1, 1975, pp. 3-4.

Describes a parametric model for the face which is capable of lip movement and expression animation. The areas affected by the parameters are the eyes, eyelids, eyebrows, lips and the jaw. This is not a muscle model; it uses interpolation between resting and final positions, but it does allow for more than one parameter to affect an area at a time. The model is synchronized in accordance with a timed speech sequence.

Parke F.I. (1975b) *Measuring three-dimensional surfaces with a two-dimensional tablet.* In: Computer and Graphics, 1, 1975, pp.
5-7.

Describes a method for measuring a 3-D object using two photos and a digitising tablet. The corresponding points on each photograph are digitised and matched up. There needs to be at least six points of known position to get the solution of a system of equations which will give the 3-D co-ordinates of each point on the surface.

Parke F.I. (1982) *Parameterized models for facial animation.* In: IEEE Computer Graphics and Applications, 2(9), Nov 1982, pp. 61-68.

A more advanced version of his 1975 model. Gives a good outline of things to think about when developing and parameterising a model. Still uses interpolation of expression parameters between extreme positions for each part of the face. Thus it is not a muscle model like Waters'. Conformation parameters allow the actual structure of the face to change, eg. to shorten the nose.

Parke F.I. (1974) *A Parametric model for Human Faces.* PhD dissertation, order number 75 -8697, University of Utah.

Pease A. (1981) *Body Language: how to read others' thoughts by their gestures.* Camel Publishing: North Sydney.

Porter S. (1990) *Made for the stage: synthetic actors are getting better.* Computer Graphics World 13 (8): 60.

Press L. (1990) SIGGRAPH '89 - tomorrow's PC today. Communications of the ACM 33 (3): 274.

Reeves W.T. (1990) *Simple and complex facial animation: case studies.* AUSGRAPH 1990, Melbourne: Australia.

Gives examples of facial animation using different methods. Is concerned with traditional animation, as well as simulation of the facial structure. The image was recorded using a 3-D digitiser on a clay model. The animation of the body was done using a hierarchical skeleton. The facial animation was done using the techniques outlined by Waters in his 1987 article.

Robertson B. (1988) *Mike the talking head.* Computer Graphics World 11 (7): 57.

Spence A.P. and E.B. Mason (1983) *Human Anatomy and Physiology.* P Benjamin/Cummings Publishing Co., Menlo Park: California.

Contains diagrams and pictures of human anatomy. Has good pictures of skeletal and muscular systems, and goes through a description of the actions of all the muscles.

Terzopoulos, D. and K. Waters (1990) *Analysis of facial images using physical and anatomical models.* IEEE Proceedings of ICCV conference, Osaka: Japan, pp. 727-732.

Creates a tri-layered tissue model to simulate the properties of the muscles and skin in the face. Uses the FACS theory as a base to a hierarchical system of controlling facial movements. Also shows how this system can be used to imitate the movements of a real face, in real time, by tracking deformable contours on the face.

Wasser J.A. (1985) *English to Phoneme Translation, Public Domain Software.* Littleton: USA.

A public domain speech synthesiser. Good to get an idea of how they work, but doesn't give a very realistic output.

Waters K. and D. Terzopoulos (1991) *Modeling and animating faces using scanned data.* The Journal of Visualization and Computer Animation 2 (4): 123-128.

Waters K. (1990) *SIGGRAPH tutorial notes.* SIGGRAPH 1990. (47 pages)

Waters K. (1987) *A muscle model for animating three-dimensional facial expression.* In: Proceedings of SIGGRAPH, July 1987, pp. 17-24.

Outlines a muscle model for the animation of the face. States that this gives a much wider vocabulary of expressions to the model than traditional methods which hard-code the expressions which are available. Gives a parameterised model based on Action Units that control groups of muscles. Aims to produce a model which can produce the six basic expressions and be tested by seeing how well it matches up to Ekman and Friesen's FACS system.

Welsh W.J. et al (1990) *Synthetic face generation for enhancing a user interface.* In: Proceedings of Image Com 1990, 1st International Conference dedicated to professional image chains, Bordeaux:France, pp 177-182.

White T. (1986) *The Animators Workbook.* Watson-Guptill Publications, New York: USA.

An informative text on the basic principles of animation. Good pointers on how to make things look realistic. Describes how to synchronise animation with a soundtrack and what to emphasise when animating the face.

Williams L. (1990) *Performance-driven facial animation.* In: Proceedings of SIGGRAPH Aug 1990, pp. 235-242.

Gives a method of putting control points on an actor's face and then tracking and recording the movements to use in facial animation. Tracking is done in 2-D and then projected to the model. Removes the need for computer inbetweening by actually recording the facial movements when the expression changes. Touches on a means of setting up mirrors to give two views of the face which can be recorded by one camera. This image could then be transformed into 3-D co-ordinates.

Wild (1987) *Wild Aniolyst BC2 Instruction Manual,* Release 5.10.

User guide for the BC2 stereoscopic digitiser.

Yau D. (1988) *A texture mapping approach to 3-D facial image synthesis.* In: Computer Graphics Forum 7(2): 129-134.

9. EQUIPMENT USED

The following equipment was used in the pilot program and in the work done on facial animation since then.

Photographing

(1) Canon T70 50mm camera (2) (Dept of Surveying and Cartography)
(2) Tripods and bar (Dept of Surveying and Cartography)
(3) Theodolite - to work out where to aim the cameras. (Dept of Surveying and Cartography)
(4) Plumb bob - to aid accuracy of equipment placement. (Dept of Surveying and Cartography)
(5) Measuring tape - to align camera angles.
(6) Stool - for person being photographed to sit on.
(7) Marking pen - for positioning of equipment.

Digitising

(8) Wild Aniolyst BC2 analytical digitiser. (Dept of Surveying & Cartography)
(9) Black and white slides - same view, slightly different eye position. (processed by education centre)

Scanner

(10) Epson Scanner - 200 dots per inch. (Computing Centre)
(11) Apple Macintosh - to access scanner and transfer to our system. (Computing Centre)
(12) 8"x10" enlargements - views at right angles. (processed by education centre)

Graphic Display

(13) Silicon Graphics 4D70GT workstation - 96 bitplanes, res: 1280x1024. (School of Computing Science)
(14) Silicon Graphics Indigo workstation. (School of Computing Science)
(15) Apple LaserWriter II (School of Computing Science)

Software

Photogrammetric Method

(16) softsurv - calculates the control points for the BC2.

(17) lens.exe - to work out the distortion of the photos for the 3-D digitiser, written by Bruce Montgomery.

(18) convert - to transform the output data from the digitiser written by Russell Doak.

(19) TGM - to triangulate the data from convert by Michael Evans.

(20) tin2mog - to convert from TGM output to MOG input written by Russell Doak.

(21) tin2fascia - Converts from TGM output to Fascia input written by Russell Doak

Screen/Mouse Method

(22) Facesave - to record data points written by Andrew Marriott.

(23) face2mog - to convert data to mod format written by Andrew Marriott.

(24) face2fascia - to convert data to fascia format written by Andrew Marriott.

Others

(25) fascia - facial animation program by Frederick Parke.

(26) mog - 3-D data display system by Phil Dench

(27) phoneme - speech synthesis program by J. Wasser.

(28) troff - text editor.

(29) library routines - standard and non-standard on the Iris workstation.

LIST OF FIGURES

TABLE OF CONTENTS

APPENDIX A

Email Contacts.

These wonderful people have given me invaluable help with my research.

amber@jaguar.esd.sgi.com - Amber Denker
bamberg@yoda.eecs.wsu.edu - Robert Bamberger
daniel@unmvax.cs.unm.edu - Tommie Daniel
gt8479a@prism.gatech.edu - Ben Watson
jason@monet.UWaterloo.ca - Jason Fischl
kaminski@netcom.netcom.com - Peter Kaminski
mnetor!lsuc!array!colin@uunet.uu.net - Colin Plumb
pieper@MEDIA-LAB.MEDIA.MIT.EDU - Steve Pieper
platt@cs.swarthmore.edu - Steve Platt
raghavac@cis.ohio-state.edu - Saty Raghavachary
stevef@csl.sony.co.jp - Steve Franks
syau%aludra.usc.edu@usc.edu - Scott Shu-Jye Syau
thinman@netcom.netcom.com - Lance Norskog
waters@crl.dec.com - Keith Waters
welsh_w_j@bt-web.british-telecom.co.uk - Bill Welsh

APPENDIX B

The following is an annotated bibliography of 30 facial animation articles.

FACIAL ANIMATION AND SPEECH SYNTHESIS

Thesis Proposal

by

Valerie Hall
875001H

June 1992

Supervisor

Andrew Marriott
Lecturer, School of Computing Science

ABSTRACT

The aim of this thesis is to develop a facial animation system. This system will be capable of automatically generating an animated sequence when given a speech track. Currently, synchronising speech and animation is a time-consuming, manual procedure. Once speech synchronisation is automated, facial animation systems will be able to be used in many practical situations. Possible applications include: video-phones, where they will reduce bandwidth on transmissions; learning aids for people with speech and/or hearing disorders; and, a friendly user interface which will be instantly familiar to new users.

ACKNOWLEDGEMENTS

I'd like to thank the people who have given us advice and assistance over the course of this project. First my supervisor, Andrew Marriott, of the Curtin Computing Science department, for his support in this project and his help with the programs for digitising and displaying the face data. We'd also like to thank Bruce Montgomery, from the Curtin Surveying and Cartography department, for taking the photos and helping with the photogrammetry side of things. Phil Dench and Mike Evans have been kind enough to let us use their MOG and triangulation programs. Lastly, and most importantly, I'd like to thank all the people who have given me information and programs through email and the net.