SPEECH AND EXPRESSION : A COMPUTER SOLUTION TO FACE ANIMATION

Andrew Pearce, Brian Wyvill, Geoff Wyvill, David Hill

Department of Computer Science, University of Calgary.
2500 University Drive N.W.
Calgary, Alberta, Canada, T2N 1N4

Abstract

Animation which uses three dimensional computer graphics relies heavily on geometric transformations over time for the motions of camera and objects. To make a figure walk or make a liquid bubble requires sophisticated motion control not usually available in commercial animation systems.

This paper describes a way to animate a model of the human face. The animator can build a sequence of facial movements, including speech, by manipulating a small set of keywords. Synchronized speech is possible because the same encoding of the phonetic elements (segments) is used to drive both the animation and a speech generator. Any facial expression, or string of segments, may be given a name and used as a key element. The final animated sequence is then generated automatically by expansion (if necessary) followed by interpolation between the resulting key frames.

We present two alternative modelling techniques for constructing the face: a polygon mesh and a functional description using a technique called *soft objects*.

Résumé

L'animation par ordinateur en 3-dimensions compte fortement sur des transformations géométriques sur temps pour les mouvements de la caméra et des objets. Pour faire marcher une silhouette ou bouillir une liquide, il faut un contrôle de mouvement sophistiqué, qui n'est pas généralement disponible aux systèmes d'animation commerciaux.

Cet article fera la description d'une façon d'animer un modèle d'un visage humain. L'animateur peut construire une séquence de mouvements facials, y compris le discours, en manipulant une petite série de mots clés. Le discours est automatiquement synchronisé, parce que les mêmes éléments phonétiques contrôlent l'animation et la restitution vocale, N'importe quelle expression facialle, ou série de segments, peut être nommée, et utilisée comme élément clé. En dernier lieu, la séquence d'animation finale est créée automatiquement par l'expansion (si nécessaire) suivi par l'interpolation entre les images clées.

Nous présenterons deux techniques alternatives de construire le visage: à polygonale maille, et par une description fonctionelle utilisant des *objets moux*.

Introduction

Three dimensional animation using computer graphics often suffers from a lack of sophisticated motion. Current modelling techniques can produce realistic looking images, but are not suited to representing objects in motion. Nor do we have established ways to describe complex motion to the computer system.

The human face is a prime example of an object which moves in a very complicated way, that cannot be easily and convincingly controlled by simple geometric transformations in time, unless constraints are placed on the possible motion. The major work in this area was done by Fred Parke at the University of Utah [Parke 74] and later developed at New York Institute of Technology [Parke 82]. Parke uses a face built from polygons and identifies groups of polygons which can be changed according to a set of parameters to control facial expressions and features. A second approach [Platt 81] is to use a structure based model where the muscles to be moved are described. While simulating the underlying facial muscles allows for exact representations of wrinkles and face motions, an adequate facial model has not been fully developed using this representation. This is due both to the difficulty in encoding all of the facial muscles and the complexity of its motion due to the number of degrees of freedom allowed the animator.

Although much work has been done on the modelling of the face, synchronized speech animation is still effected through rotoscoping or related techniques.

The Graphicsland Animation System

The *Graphicsland* project group [Wyvill, B. 85a] at the University of Calgary has developed an organised collection of software tools for producing animations from models in three dimensions. The system allows the combination of several different kinds of modelling primitive [Wyvill et al 85b]. Thus polygon based models can be mixed freely with fractals [Mandelbrot 83, Fournier 82] and particles [Reeve 83] in a scene. Motions and camera paths can be described, and animations generated. Note that we do not include the use of a two dimensional "paint" system. Our objective is always to construct views of a full three dimensional model.

Our objective in this work was to introduce better techniques for motion control than commonly available and integrate them into *Graphicsland*.

Interfacing to the Parke Model

The face-representation we used has been developed from Fred Parke's work at the University of Utah [PARKE 74]. Parke models the face as a collection of polygons which may be manipulated through a set of 50 numerical parameters. These control such things as length of nose, jaw rotation, shape of chin, and other similar facial features, and allow movement of these features by interpolating the parameters.

To describe motion directly using these parameters is clumsy and difficult. Motions were described as a pair of numeric tuples which identified; the initial frame, final frame, and interpolation type, followed by the parameter, the parameter's initial value, and the final value. In order to aid the animator, a keyword based interface was developed. The interface makes it possible to build up libraries of partial expressions (smile and blink would be two partial expressions) and place them anywhere within an animated sequence. It also has the ability to detect conflict should two simultaneous partial expressions attempt to manipulate a facial feature in opposing directions at any point in the animation.

Expression

We specify each partial expression by means of a set of keywords. We must describe:

1) the part of the face to be moved (eyes, mouth, cheeks...),
2) the type of movement (open, arch, raise...),
3) the initial and final frame number
4) the parameter value (normalized) at the final frame,
5) and optionally, the type of interpolation (default is linear)

For example, to open the mouth the dialogue might be:

open mouth frame 12 25 value 0.8

This would cause the mouth to open with 80% (0.8) of the maximum jaw rotation, begining at frame 12 and ending at frame 25 of the animation. Alternatively, motions may be grouped together into a key element, for example, a blink expression might be specified:

```
animate blink
        close eyes frame 1 2 value 0
        open eyes frame 2 3 value 0.9
end
```

Once an expression has been specified, the animator may place that motion at several places in the animated sequence:

```
add blink frame 25
add blink frame 36
```

Figures 1a,1b and 1c show three frames from the blink sequence. Figure 2,3 and 4 show various expressions. Figure 4 also has had hair grown on the head using a number of particle generators distributed on the polygons which define the scalp.

Speech

In all work so far, the animation of a talking sequence for one of these facial models has been done using a technique similar to rotoscoping. A human actor is filmed, reciting the required script, and the facial model is constrained to follow the sequence of lip and jaw positions needed for each frame of the animation.

This process is tedious and expensive. Various alternative approaches exist, all based on some knowledge of the relationships between speech sounds and the configuration of the articulators. Articulation models giving information about jaw position, lip spreading or rounding, tongue position, and their dynamic relationships, for various speech sounds, can be built up from the literature on lip reading and acoustic phonetics [Walther 1982, Levitan 1977]. Subjective evaluation of the adequacy of such models, followed by correction and re-evaluation ensures that the models are good functional representations. An animator can use such model data to set up key frames corresponding to successive segmental articulations. The computer can interpolate the key frames according to more or less simple rules, and the resulting product can be dubbed in the usual way. Alternatively, the articulatory parameters may be controlled by recognition of the sounds produced by an actor, which ensures natural rhythm for the resulting speech. However, speech recognition is still less than perfect, and such systems tend to rely on sound classification that is both crude and error prone. Our approach is to synthesise both the speech and the sequence of facial expressions by rules, based on the articulatory model for the facial movements, and based on rules for acoustic synthesis for the speech. Thus phonetic script incorporating both segmental and suprasegmental information drives both aspects of the speech animation sequence. Synthetic speech is still somewhat unnatural, but such speech animation is both intelligible and well synchronised. Large quantities of speech animation can be generated at virtually no more cost than the graphical animation that forms part of it. Furthermore, script changes can be incoporated without having to rely on the availability of a particular real speaker. The synthesis is based on a long-standing research project that includes a new model for speech rhythm based on a generalisation of real speech rhythm data [Hill 1978a, Hill 1978b].

Normal speaking rates vary a good deal. Typical segment (speech sound) durations for normal speech vary between 50 and 250 millseconds. Each articulation changes in a basically continuous fashion into the next one. For the acoustic synthesis, piecewise linear interpolation of the acoustic parameters, from one target to the next, according to relatively simple rules a few time divisions, has been found adequate for high quality synthesis. A typical sampling rate would be one sample every 10 milliseconds, but linear changes over periods of 20 to 80 milliseconds occur. A similar approach is being adopted for the interpolation needed for the changing facial expression dictated by the moving articulators. The rate of interpolation varies, just as for the acoustic synthesis (and in synchrony with it), but the rules are few and simple. At 24 frames per second, the average frame rate for a movie is approximately one frame every 40 milliseconds. This is well matched to the sampling rate needed for a fairly accurate representation of the synthetic speech, as might be expected from observations of real speech on film.

As an example of speech, the phrase "Hi there" could be achieved as follows:

```
sentence greet
        h
        ah
        i 170 0.3
        th
        e
        r
end
```

A sentence is specified and denoted by the name "greet" in order to allow the user to refer to the sentence again for modification, deletion or placement. The "end" command marks the end of the sentence, and allows the system to calculate the number of frames required for the speech. The duration and enunciation parameters available for each segment are being used with the diphthong "AH I"; it's duration is now 170 ms. and it's enunciation in terms of mouth position is 0.3 of it's full possible range. If these parameters

are not supplied, the default duration for each element of the phonetic script is used.

Once the animator is satisfied with the placement of the partial expressions, the sequence may be examined for motion conflicts with the command "check". Should a conflict be discovered, the partial expressions which clash are displayed and may be edited.

Figures 5a,b,c,d and e show selected frames produced from the sentence "greet".

Soft Objects

The term "*soft* object" is used to refer to the particular class of objects whose shape varies constantly because of forces imposed on it by its surroundings.

We have been experimenting with a general model for *soft* objects which represents an object or collection of objects by a scalar field. That is a mathematical function defined over a volume of space. The object is considered to occupy the space over which the function has a value greater than some threshold so the surface of the object is an iso-surface of the field function. That is a surface of constant function value within the space considered. The idea of using such surfaces for 3D modelling was first put forward by Jim Blinn [Blinn 82] and refined in the *Graphicsland* system [Wyvill 85c]. Using the field function developed in [Wyvill 85c], such surfaces can be finely controlled by varying the radius of influence and field value due to each key point. Our initial experiments suggest that fewer key points are needed to control the facial movements than with other techniques, and the process is computationally less expensive than using B-splines or polygon meshes.

Although a polygon mesh is a useful representation for the face model, it forms only a crude approximation to the smooth curves of a face, and suffers from the problem that current shading techniques smooth the centre the mesh leaving an un-smoothed polygon silhouette edge. B-spline patches [Huitric 85] have been used to define a smooth surface and fewer control points are needed to define the face than with polygons. However a set of soft object key points share these advantages and have several more. Soft object control points may have different colours associated with them. The colour of a control point affects the colour of a local region of the face, and this colour will be smoothly blended into the colours of the surrounding regions. A face may contain various areas of different colour, for example, rosy cheeks, red lips, a dark chin and pale forehead. The colours of control points can be made to vary with time, causing smooth changes of colour in selected parts of the face, as in a blush.

Conclusion

We have presented some experimental work with a face model. Different modelling techniques for facial animation have been described along with a method of using the model to produce sychronized animation directly from a speech synthesizer. The user interface to the speech and expression program is particularly simple and effective.

Acknowledgements

We would like to acknowledge the help and advice of our colleague Fred Parke, who supplied us with the original face data and has been extremely helpful to our work on *Graphicsland*. We also thank Milan Novacek for his help with the particle hair and Yung-Hsien Yen for his work on the user interface. The work is partially supported by the Natural Science and Engineering Research Council of Canada.

References

Blinn, J. (1982) "A Generalization of Algebraic Surface Drawing" *ACM Transactions on Graphics, 1,* 235.

Fournier, A., Fussel, D., and Carpenter, L. (June 1982) "Computer Rendering of Stochastic Models" *Commun. ACM, 25, 6,* 371-384.

Hill, D.R., Witten, I.H., and Jassem, W. (1978a) "Some results from a preliminary study of British English speech rhythm." *Research Report 78/26/5, Dept. of CS, University of Calgary presented to the 94th. Meeting of the Acoust. Soc. Amer. Miami Dec 1977.*

Hill, D.R. (1978b) "A program structure for event-based speech synthesis by rules within a flexible segmental framework" *Int. Journal of Man-Machine Studies (1978), 10* (3) 285-294.

Huitric, H. and Nahas, M. (March 1985) "B-Spline Surfaces: A tool for computer Painting" *IEEE Computer Graphics & Applications, 5* (3) 39-47.

Levitan, E.L. (1977) "Electronic Imaging Techniques" *Van Nostrand Reinhold Co., New York, N.Y..*

Mandelbrot, B. (1983) "The Fractal Geometry of Nature." *W.H. Freeman and Company. (First Edition 1977).*

O'Neill, J.J. and Oyer, H.J. (1961) "Visual Communication for the Hard of Hearing" *Prentice-Hall Inc., Englewood Cliffs, N.J..*

Parke, F.I. (Dec. 1974) "A Parametric Model for Human Faces" *PhD. dissertation, University of Utah..*

Parke, F.I. (Nov. 1982) "Parameterized Models for Facial Animation" *IEEE Computer Graphics and Applications, 2* (9) 61-68.

Platt, S.M. and Badler, N.I. (Aug. 1981) "Animating Facial Expressions" *ACM Computer Graphics (SIGGRAPH '81), 15* (3) 245-252.

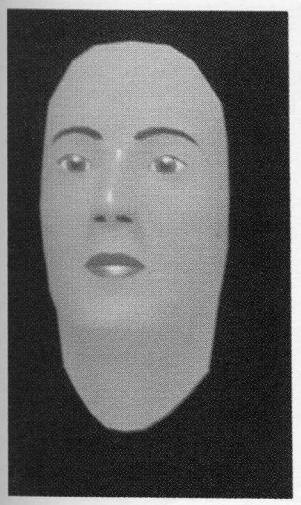
Reeves, W. (Apr 1983) "Particle Systems - A Technique for Modeling a Class of Fuzzy Objects" *ACM Transactions on Graphics, 2,* 91-108.

Walther, E.F. (1982) "Lipreading" *Nelson-Hall Inc., Chicago, Illinois..*

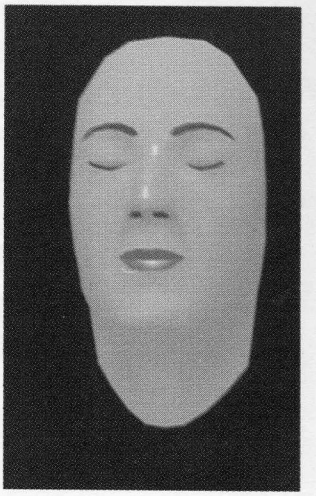
Wyvill, B.L.M., McPheeters, C., and Garbutt, R. (July 1985a) "A Practical 3D Computer Animation System" *The BKSTS Journal (British Kinematograph Sound and Television Society), 67* (6) 328-332.

Wyvill, B.L.M., McPheeters, C., and Novacek, M. (June 1985b) "Specifying Stochastic Objects in a Hierarchical Graphics System" *Proc. Graphics Interface 85, Monteal.*

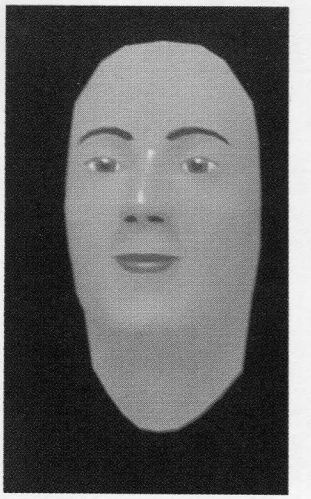
Wyvill, G., Wyvill, B., and McPheeters, C. (October 1985c) "Soft Objects" *Research Report No. 85/215/28,* University of Calgary, Department of Computer Science.



a)



b)



c)

Figure 1. **Blink.**

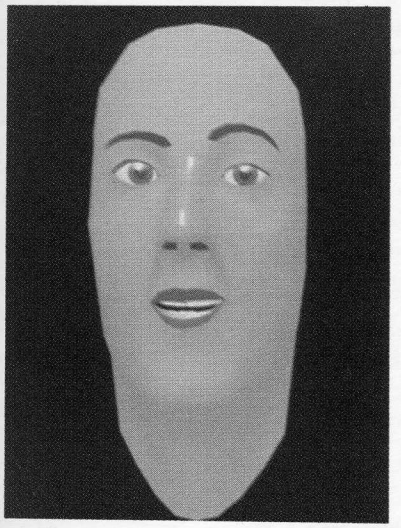


Figure 2. **Smile.**

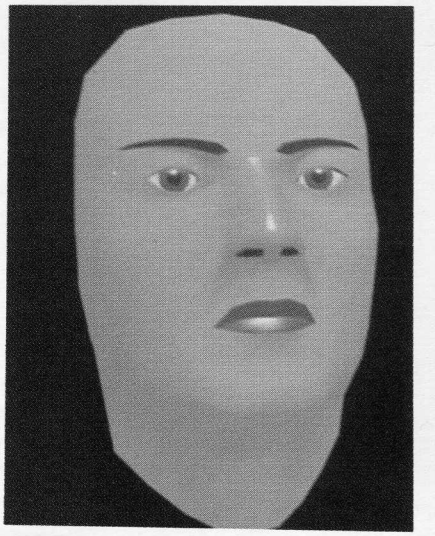


Figure 3. **Frown.**

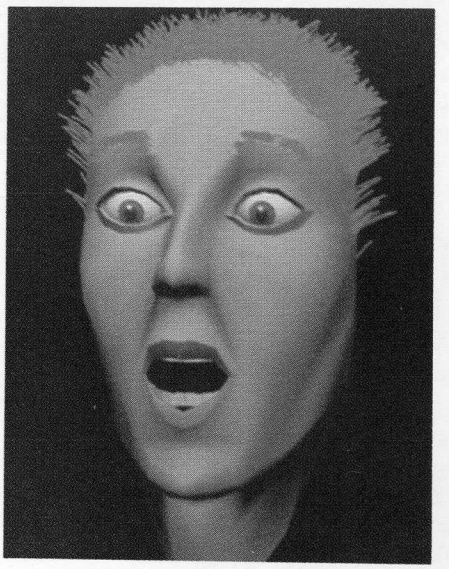
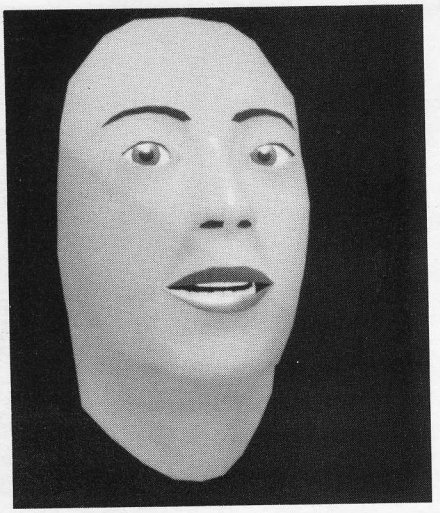
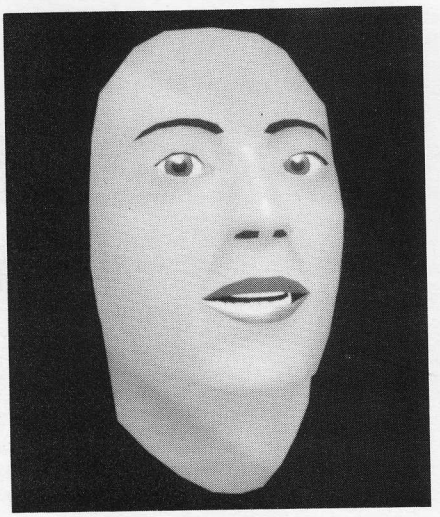


Figure 4. **Scream.**

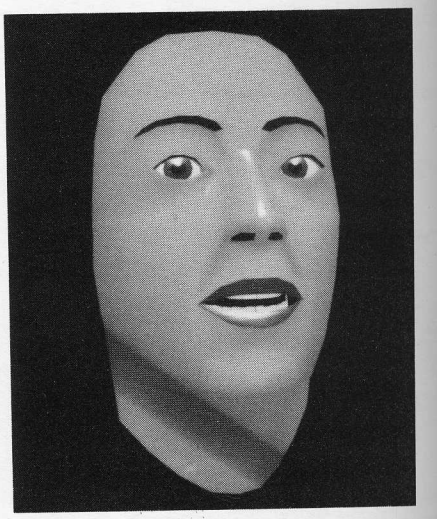


a) 'h'

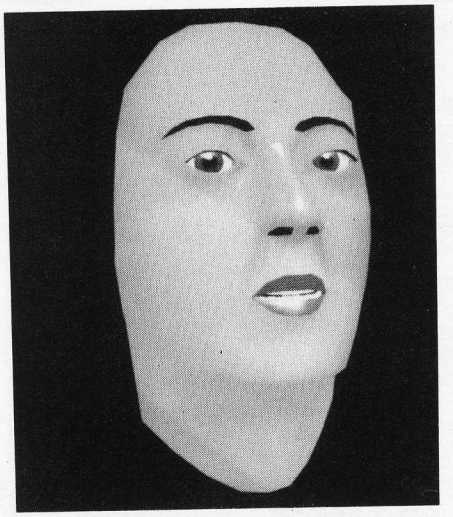
b) 'i'



c) 'th'



d) 'e'



e) 'r'

Figure 5. **Hi there.**