Expressive Visual Speech using Geometric Muscle Functions

James D. Edge and Steve Maddock

Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
Sheffield S1 4DP
<j.edge, s.maddock>@dcs.shef.ac.uk

Abstract

This paper describes a system for the creation of expressive visual-speech animation. Geometric Muscle Functions [Waters 1987, 1988] are used for the control of both facial expression and speech lip postures, allowing the easier integration of these two factors. This addresses the common problem of attempting to combine two separate domain-specific control techniques. The use of muscle functions also allows the control mechanism to be abstracted from the mesh representation, and so the described system is applicable to any reasonable facial model.

Twenty-five simulated muscles, plus jaw rotation, are used to produce both the six universally-accepted emotions (sadness, fear, contempt, surprise, happiness, and anger) and fifty-six identified static speech postures (visemes). The underlying muscular influences of these two factors are then combined together using weighted-blending techniques to create expressive speech postures.

Pre-captured speech and facial expression data is used as input to the system. By varying the relevant muscle influences over time, speech synchronous animation is then created. Example results include spoken digit sequences and simple sentences. Informal user testing suggests that the addition of detailed internal mouth structures, such as the tongue, would improve recognition rates for certain classes of speech gesture.

Keywords: Facial Animation, Visual Speech, Viseme, Expression

1 Introduction

It has been found in the field of Computer Facial Animation that the closer models come to real-life faces, the greater the complexity required to fulfil audience expectations. Cartoon-like figures are readily accepted, yet we have difficulty in simulating the subtlety that is observed in real human faces. The problems of facial animation are increased when we consider Visual Speech - the simulation of talking heads. Here we must not only take into account the problem of reproducing speech postures, but also the audio-synchronicity and realistic facial movement within the resultant animation.

A solution to computer facial animation is eagerly anticipated across a wide section of the graphics industry. Applications can be seen in situations as diverse as entertainment, human computer interaction, and speech therapy. In fact facial animation techniques have recently been commercially demonstrated in Pixar's 'Toy Story' series of feature films, and also Ananova [2000] the Internet newsreader. However, the success of such enterprises to some extent relies upon the audience's suspension of disbelief rather than the realism of the underlying facial model.

The problems experienced with current systems can be categorised into two main areas: rendering/representation, and control/animation. The first set of difficulties arise with the efficient representation of the complex facial mask with simple polygonal or patch constructs, and its subsequent rendering given the multi-layer nature of human skin (see [Parke 1996] for an in-depth discussion of rendering issues). This paper deals with the second set of difficulties, those pertaining to the control and animation of a facial mask representation. The following sections describe in detail the application of a muscle-level facial parameterisation for the control of expression and visual-speech characteristics. The muscle functions implemented are extended versions of those demonstrated by Waters [1987], allowing for facial discontinuity and sphincter muscle protrusion. The use of such muscle-level constructs to produce speech synchronous animation is also demonstrated. It is considered that models which use a combined control mechanism for the creation of both visual speech postures and facial expression will offer benefits over systems that attempt to combine different control techniques. The methods described within this paper are aimed at character-based animation, where virtual actors are required to not only look realistic but also convey emotion to the viewer.

1.1 Background

One of the most important developments within three-dimensional facial animation has been the progression from direct vertex manipulation towards control methods based upon a more manageable set of parameters. Most facial parameterisations inherit from Parke's [1974] original work. These systems directly manipulate vertices from a polygonal mask representation to reproduce expressions. Each parameter within the model will control the movement of a group of vertices, and the combined action of a number of these parameters will allow the creation of facial expression. Parke's approach offers benefits over direct animator control, yet suffers from the close relationship between the structure of the representation and its control mechanism. This limitation indicates that each parameterisation is only applicable to a single mask topology.

Keith Waters's [1987, 1988] work on geometric muscle models builds upon Parke's idea of parameterisation, with the use of muscle functions that are separate from the facial mask structure. Each geometric function mimics the action of an individual muscle upon the skin. Vertices within the function's area of effect are pulled from the muscle's insertion into the skin towards its static attachment to bone. Whereas models based upon Parke's work require a fixed structure, Waters's geometric muscle functions allow the generalisation of techniques across facial topologies. These muscle functions have also been demonstrated with different structural models, for example the Langwidere system [Wang and Forsey 1994] which uses hierarchical B-splines to define the surface representation. This gives some indication of the applicability of muscle-based methods across different surface structures and topologies.

Models of visual speech have coevolved with the methods used to control facial animation, yet most systems are specialised to only deal with the areas of the face used in the production of speech (typically the jaw and lips, as in Guiard-Marigny *et al.* [1994]). These systems need to provide a means of representing individual speech gestures, and driving this representation synchronously with an audio soundtrack. Visual speech also requires that systems take into account coarticulation, which is the effect of articulatory context upon each individual speech posture. Coarticulation is an effect caused by the limitations of the physical system of muscles used in the production of speech. Within natural speech, articulatory context prevents most idealised speech postures from ever being reached.

Baldi [Cohen and Massaro 1993] is perhaps the best-known example of a directly parameterised system driven by a model of speech production. The Baldi model is built upon the basis of Parke's work, adding extra parameters in order to control speech-specific properties of the lips and also including a model of the tongue. These parameters are driven by a set of dominance functions, which are blended across the animation to reproduce the effect of coarticulation. Several other parameterisations have also been demonstrated in the production of visual speech. Examples are FACS action units (described by Ekman and Friesen [1978]) implemented by Pelachaud [1991], FFDs [Kalra 1993], and lip dimension parameters [Guiard-Marigny et al. 1994].

Alternative techniques for visual speech synthesis include methods based upon interpolation or morphing between pre-captured lip positions. Waters's [1993] DECface system is an example of such a technique. DECface uses an interpolation algorithm based upon an approximation of Hooke's law to move between fifty-six predefined lip shapes. The basic DECface system is two-dimensional in nature, yet extensions to three-dimensions are straightforward. DECface does not include an accurate model for predicting coarticulation, which is a problem with most visual-speech techniques that use pre-captured speech postures as key positions within an animation.

Although visual speech research has taken a number of different forms, relatively little work has been conducted into the use of muscle-based control units (see King et al. [2000] for an alternative anatomically-based model for visual speech). Section 2 describes in detail the implementation of such a model for expressive-visual speech production. Sections 3, 4 and 5 evaluate the implemented system and propose future revisions to the model.

2 Implementation

The system produced uses geometric muscle functions to control both the expressive and visual speech features of a facial mask. The advantages to this approach are exhibited in the relatively low amount of work required in converting the control parameterisation to a new mesh structure, and the ease in combining facial expressions with speech postures. The following sections describe in detail the issues involved in implementing muscle functions, and their use in the production of both static facial postures and finally speech-synchronous animation.

2.1 Controlling Facial Expression

The facial mask model used consists of 878 three-dimensional polygons. Individual vertices within the mask are displaced by a combination of twenty-five geometric muscle functions, plus jaw rotation. Two muscle types are simulated within the system: linear, and sphincter [Waters 1987]. Linear muscles are the most common structures used in the control of facial expression; twenty-four of the implemented muscles simulate linear muscle contractions. The model produced does not assume symmetry within the control structures of the face, and so the twenty-four linear muscles consist of twelve left-right pairs. This allows the system to create asymmetrical expressions, for example raising an individual eyebrow. The final muscle is a sphincter muscle simulating the Obicularis Oris, a muscle which surrounds the lips. The contraction of this particular muscle is well documented within speech literature [Bell-Berti and Harris 1981] to correspond to the rounding of the lips as observed in utterances such as ‘book’ or ‘bought’.

2.1.1 Linear Muscle Functions

Linear muscle contractions are characterised by the movement of skin from the muscle’s insertion into the skin towards its stationary point of attachment to bone. This contraction can be described using a vector from the point of insertion to the point of attachment, about which vertices are displaced towards the static attachment to bone. Maximum displacement will occur at the muscle’s insertion into the skin, and no displacement will occur at the edges of the muscle’s area of effect. The area of effect for a linear muscle is defined by an angle about its centre point. Due to the abstract nature of linear muscle functions a surface is used to trim muscular effects that cross the lip boundary. This allows the independent control of the upper and lower lip surfaces, important for the production of visual speech postures (Section 2.2). Previously Waters’s muscle functions could not part the lips, limiting their use in visual speech systems. Figure 1 demonstrates the effect of linear muscle contraction upon a mesh. Equation 1 is used to determine the displacement of an individual vertex due to a linear muscle contraction (k controls the scale of deformation, other symbols refer to Figure 1).

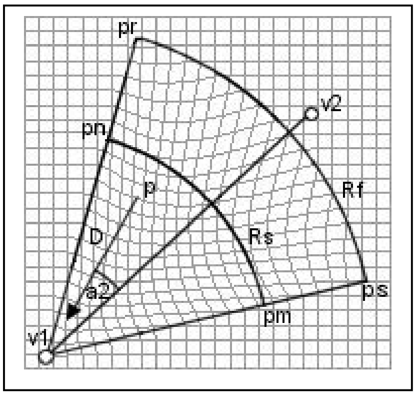

Figure 1 Linear Muscle Contraction

[1]

$$p' = p + akr\frac{\overrightarrow{pv_1}}{\left|\overrightarrow{pv_1}\right|}$$

$$a = \cos(a2)$$

$$D = \left|\overrightarrow{pv_1}\right|$$

$$r = \begin{cases} \cos\left(\dfrac{1-D}{R_s}\right) \text{for p inside sector } (v_1 p_n p_m) \\ \cos\left(\dfrac{D-R_s}{R_f-R_s}\right) \text{for p inside sector } (p_n p_r p_s p_m) \end{cases}$$

2.1.2 Sphincter Muscle Functions

Sphincter muscles consist of a ring of parallel muscle fibres which contract in a manner similar to the closing of a drawstring bag. This form of contraction causes tissue to collect about the centroid of the muscle creating a characteristic bulging. This is simulated within the model using a parametric ellipsoid. Vertices within the ellipsoid are displaced towards its central point (fig. 2). The magnitude of displacement is dependant upon the proximity of each vertex to the centre of the ellipse. Equation 2 is used to control the displacement of an individual vertex due to a sphincter muscle contraction (k controls the scale of deformation, other symbols refer to Figure 2).

[2]

$$p' = p + kd\frac{\overrightarrow{pc}}{\left|\overrightarrow{pc}\right|}$$

$$d = \begin{cases} fg \text{ for } \left|\overrightarrow{\mathrm{pc}}\right| > ly \\ 0 \text{ for } \left|\overrightarrow{\mathrm{pc}}\right| \leq ly \end{cases}$$

$$f = 1 - \frac{\sqrt{ly^2 p_x^2 + lx^2 p_y^2}}{lxly}$$

$$g = \frac{\left|\overrightarrow{pc}\right|}{lx}$$

The central area of the ellipse is also shielded from muscular contractions (for an area with a radius equal to the semiminor axis of the ellipse) to prevent vertices from collecting at the centre point of the muscle. In real lips protrusion increases

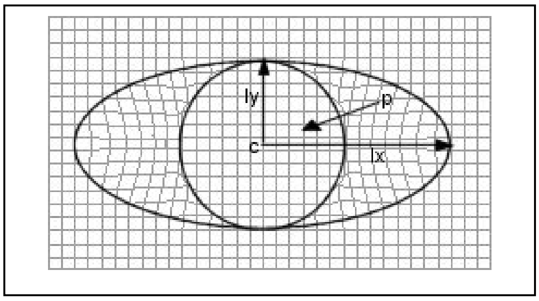

Figure 2 Sphincter Muscle Contraction

towards the centre of the muscle, and so the inverse of the displacement coefficient (d in equation 2) is used to push vertices forward creating a puckering effect. Waters's original sphincter muscle is two-dimensional in nature and so must be modified, as described, to allow for the puckering of the lips.

2.1.3 Muscular Interaction

Combinations of muscle function contractions are used to produce individual expressions (both emotions and visemes) on the facial mask. Where muscle influences overlap, the combined influence is calculated as the simple summation of each of the individual displacements. This captures the antagonistic nature of facial muscle. However, the disadvantage to this approach is that it can result in a disruption of the mask structure. In order to compensate for this the magnitude of each muscle contraction is restricted, preventing combined muscular influence from disfiguring the facial mask.

2.1.4 Jaw Rotation

Jaw rotation is caused indirectly by the contraction of muscles attached to the mandible (jaw bone), which in turn rotates about the mandibular joint causing the skin of the facial mask to stretch and the mouth to open. This is simulated using a simple function which varies the rotation of the skin vertices within the jaw according to their proximity to the centre of the face. The function creates a smoothly curved open mouth shape, with maximum rotation produced at the centre of the lower lip. In order to integrate jaw rotation with the effect of muscular contractions it's effect is also treated as muscular and is integrated with other muscle contractions as described in section 2.1.3.

2.2 Viseme Capture

For this research, visemes are used as the basic units of visual speech. For our purposes, these units are the extreme lip positions used in the instantiation of basic auditory speech units (in this case English phonemes). The same speech classification is used as in the DECFace system [Waters 1993]. Fifty-six visemes (visual speech postures, see fig. 5) are captured including silence and pause speech postures. Each viseme is the direct visual equivalent of a phoneme, yet not every phoneme is visually distinct. The best example of this is the /p, b, m/ cluster for which the captured visemes are visually indistinguishable. Each viseme is captured by hand from observations of real lips using the control structures described in section 2.1. All produced visemes are the result of the interaction between twelve of the implemented muscle functions (including jaw rotation). Table 1 shows the correlation between muscle functions and the important lip dimensions which vary during speech production. It is this correlation of muscle-level units to lip dimensions which allows the relevant functions to be identified for speech. Figure 7 shows the positioning of these muscle functions on the facial mask.

Table 1 Correlation between implemented geometric functions and lip dimensions

Function	Dimension(s)
Obicularis Oris (sphincter)	increases lip rounding, reduces lip width, increases lip protrusion
(left\right) Risorius (linear)	increases lip width
(left\right) Depressor Labii (linear)	lowers the bottom lip
(left\right) Levator Labii Superioris (linear)	raises the upper lip
(left\right) Zygomatic Major (linear)	raises the lip corner
(left\right) Triangularis (linear)	lowers the lip corner
Jaw Rotation	mouth openness

2.2.1 Expression-Viseme Interaction

The muscle functions used for viseme capture are not exclusive to the production of speech. Several of the muscles identified are also used to create emotional facial expressions. Six emotions are demonstrated by the system (fig. 4): sadness, fear, contempt, surprise, happiness, and anger. These have been identified as the six universal emotions [Ekman 1973] as they have been observed to cross cultural boundaries. The system described is capable of producing more than six expressions, yet this set of six common emotions is enough to demonstrate expression-viseme interaction. In order to produce expressive speech the emotion and viseme data must interact. Interaction is produced by parameter blending between the speech and expression data sets (fig. 6). The relative dominance of each factor is then varied by weighting the respective data set. This allows a significant reduction in the data storage required for expressive speech postures. Many previous systems would require storage for 6x56 speech postures to demonstrate the same capability for expressive-visual speech. The current system only requires to store six expressions and fifty-six visemes. The limitations of this approach lie in the assumption of a constant linear relationship between the speech and emotional aspects of human facial expression. This is not necessarily the case for all expression/viseme combinations. However, certain combinations do produce good results (for example, the combination of happiness and vowel lip shape shown in Figure 6). More consistent results may be produced by integrating knowledge of speech production into the method of expression-viseme interaction.

2.3 Speech Synchronous Animation

The system thus far described produces discrete speech postures which directly correlate to the emotional and phonetic units within facial communication. Animation is produced by interpolating between key speech and expression targets, with each frame within the sequence being the result of interaction between the two factors (Section 2.2.1). The production of speech-synchronous results relies upon an available transcription of the speech sample. Each transcription was produced using tools from the MAD demonstrations [Wrigley 1999] within MATLAB. Figure 8 shows an example trajectory from the utterance “One Five Zero Zero Six”.

3 Evaluation

In order to evaluate the quality of the speech postures produced, some informal perceptual testing was performed. Initially phoneme utterances were categorised into eight visually distinct classes (table 2). An example utterance was constructed for each class, consisting of a silence-viseme-silence sequence. Twelve test subjects were subsequently asked to match visual classifications to animations. The subjects were all native English speakers with no formal training in lip-reading or similar audio-visual techniques. Formal experimental conditions were not implemented, and so the following results can only be considered a guide to the quality of visual speech postures produced by the system.

Table 2 Phonetic classification used for perceptual testing

Category	Phoneme(s)	Example
a)	/p,b,m/	pat
b)	/w/	won
c)	/f,v/	vet
d)	/ae,eh,ih/	bat
e)	/ow/	boat
f)	/r,rx/	red
g)	/ao/	bought
h)	/sh,zh,z/	zoo

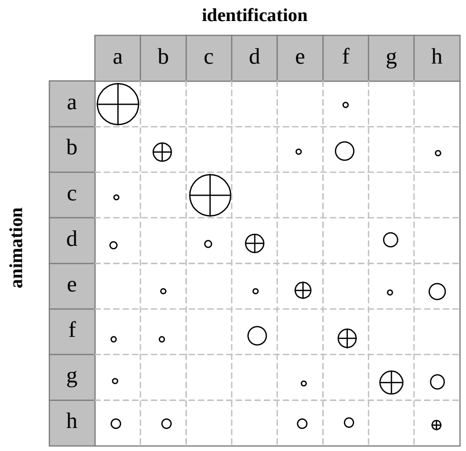

Figure 3 Confusion Matrix for the identification of viseme utterances (a-h refer to the categorisation found in table 2). The area of each circle shows the percentage of identifications. Correct identifications are marked with a cross.

3.1 Results

Figure 3 shows a confusion matrix [Massaro 1998] of the results of the perceptual testing. Test subjects demonstrated varying competence at the perceptual task, yet the results show a trend in the quality of each viseme utterance. The classes /p,b,m/ and /f,v/ are most easily identified, correlating to speech postures with little or no internal mouth structures visible. The classes /ow/, /r,rx/, and /sh,zh,z/ were poorly identified. These highlight two problems with the current system. Firstly, the internal mouth structures are either of poor quality (teeth) or nonexistent (tongue), and secondly, direct interpolation of viseme targets does not sufficiently model the complexity of speech production. It is to be expected that interpolation will produce poor quality animation because it does not take into account the properties of muscle tissue. However, given the limitations of the animation system, the results demonstrate that high quality visual speech postures are produced by the muscle-model. All animations were recognised correctly some of the time, with a maximum recognition rate of 95% (/f,v/ and /p,b,m/) and a minimum of 20% (/sh,zh,z/). These recognition rates are promising for future development of the system. Comparisons with current systems (such as Baldi [Cohen and Massaro 1993]) appear initially favourable. It is believed that with improved modelling of the internal mouth structures and a more detailed speech production mechanism superior identification rates could be achieved.

4 Conclusions

This paper has demonstrated the adaptation of geometric muscle functions [Waters 1987, 1988] for the application of expressive visual-speech. The advantage to such an approach is in the unification of techniques for the creation of expression and visual speech, and thus the simplification of expression-speech interaction. The system is seen as an early step in the evolution of models for expressive facial communication. Where previous methods have taken speech to be an orthogonal property to facial expression we have described an intuitive method for interaction between these two factors. Further to this we have described a system that is conceptually linked to the biological basis of facial control within its use of muscle level control parameters. It is proposed that the distinctions between the control mechanisms for visual-speech and expression will become more closely tied in future systems and perhaps closer in function to the real muscles that are observed in human facial control. These systems will deal with the overall problem of human facial communication, and so avoid the problems inherent in layering different facial control techniques.

5 Future Work

A number of problems are either untackled or insufficiently solved within the current system. Chiefly amongst these is the lack of any model for coarticulation. As previously described coarticulation is the result of articulatory context upon speech, effectively the speech equivalent of joined up handwriting. Given the nature of the model it would appear intuitive to encode the physical attributes of muscles at the geometric function level. This would allow coarticulation to be tackled from a bottom-up perspective, in contrast to many current approaches which work at either the viseme or speech-gesture level. It has also been previously mentioned that improved internal mouth structures may produce superior recognition rates. This is true only for certain classes of viseme, yet it can be seen from the results (fig 3) that it is these open-mouthed speech postures which produced inferior recognition rates. Future revisions of this work will include a model of the tongue to improve the quality of certain open-mouthed visemes.

Further to these points already mentioned there are several elements missing, which belong in a full visual speech system:

- Automatic synchronisation to recognised speech
- Text-to-visual-speech (TTVS) capability
- Inclusion of eye movements and actions (such as blinks or winks) in animation
- A model of the facial substructure (skull and muscular tissues)
- Geometric muscle functions for sheet and triangular muscles (such as the triangularis), which are currently simulated using linear muscles.

6 Acknowledgements

The authors would like to thank Dr. Ian Badcoe of Infogrames Ltd. for his involvement in this research. James Edge would also like to thank Dr Alan Watt, Scott King, Mark Eastlick, and Michael Meredith for their help and comments in producing this report.

7 References

Ananova Ltd. (2000) http://www.ananova.com/

Bell-Berti F. and Harris K.S. (1981) Temporal patterns of coarticulation: Lip rounding, in *Papers in Speech Communication: Speech Production*, Kent R.D., Atal B.S., and Miller J.L. (eds), Acoustical Society of America, 599-604

Cohen M. M. and Massaro, D. W. (1993) Modeling coarticulation in synthetic visual speech, in *Models and Techniques in Computer Animation*, N. M. Thalmann & D. Thalmann (eds), Springer-Verlag, Tokyo, Japan

Ekman P. (1973) Darwin and Facial Expressions, Academic Press, New York, NY

Ekman P. and Friesen W. (1978) Facial Action Coding System, Consulting Psychologists Press inc., Palo Alto, CA

Guiard-Marigny T, Adjoudani A, Benoît C (1994) 3D Models of the lips and jaw for visual speech syntheis, *2nd International Conference on Speech Synthesis*, Newark, NY

Kalra P. (1993) An Interactive Multimodal Facial Animation System, PhD Thesis, EPFL, Lausanne, Switzerland

King S.A., Parent R.E., and Olafsky B.L. (2000) An anatomically-based 3D parametric lip model to support facial animation and synchronized speech, *DEFORM' 2000 Conference*, Geneva, Switzerland, November 2000

Massaro D. (1998) Perceiving Talking Faces, MIT Press, London, 391-413

Parke F.I. (1974) A Parametric Model for Human Faces, PhD Thesis, University of Utah, Salt Lake City, UT

Parke F.I. and Waters K. (1996) Computer Facial Animation, A.K. Peters Ltd., Wellesley, MA

Pelachaud C. (1991) Communication and Coarticulation in Facial Animation, PhD Thesis, University of Pennsylvania, Philadelphia

Wang C. and Forsey D.R. (1994) Langwidere: A New Facial Animation System, University of Calgary and University of British Columbia

Waters K. (1987) A muscle model for animating three-dimensional facial expressions, ACM SIGGRAPH 1987, July 1987

Waters K. (1988) The Computer Synthesis of Expressive Three Dimensional Facial Character Animation, PhD Thesis, Middlesex Polytechnic

Waters K. and Levergood T.M. (1993) DECface: An Automatic Lip-Synchronisation Algorithm for Synthetic Faces, Digital Equipment Corporation

Wrigley S.N., Cooke M., Brown G.J. (1999) Interactive Learning in Speech and Hearing, *ESCA/SOCRATES MATISSE*, London, 16-17 April 1999, 21-24

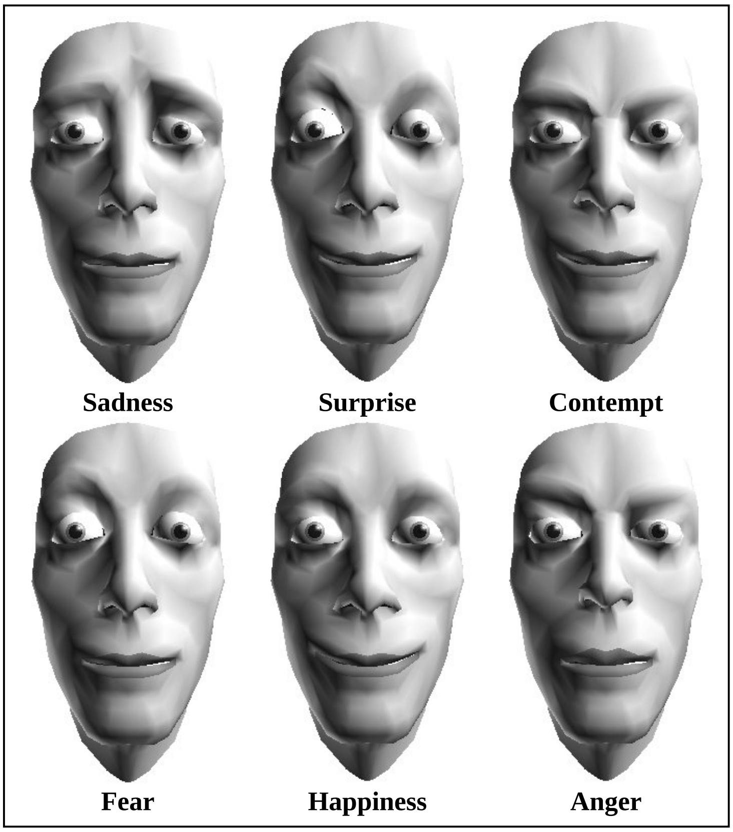

Figure 4 The six universal emotions, as identified in [Ekman 1973]

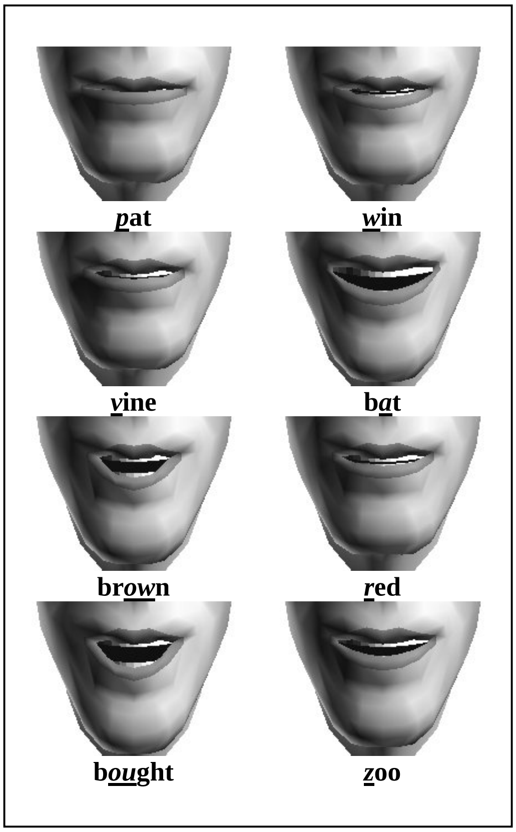

Figure 5 Examples of captured visemes

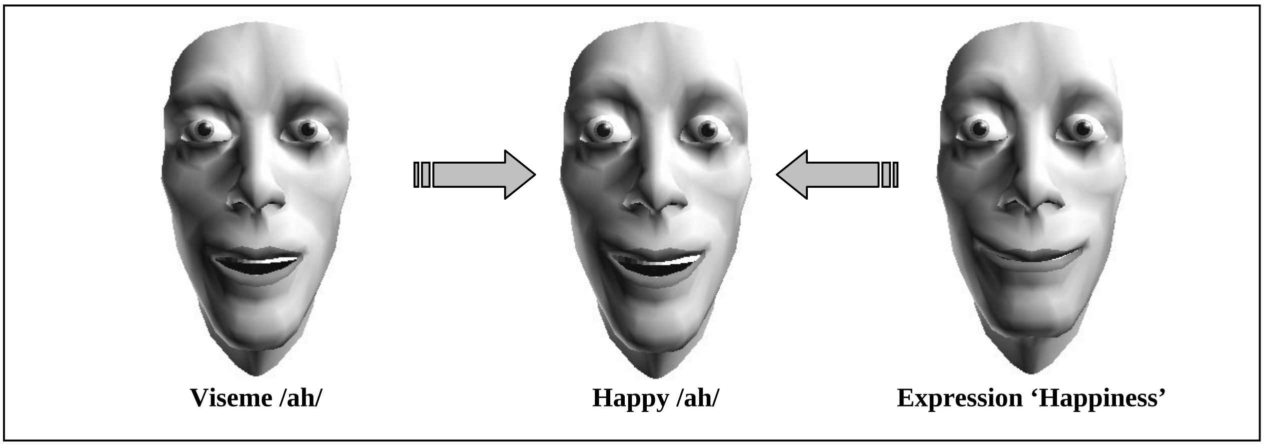

Figure 6 Example of expression-viseme interaction, weighted parameter blending is used to combine the emotion 'happy' with the viseme /ah/

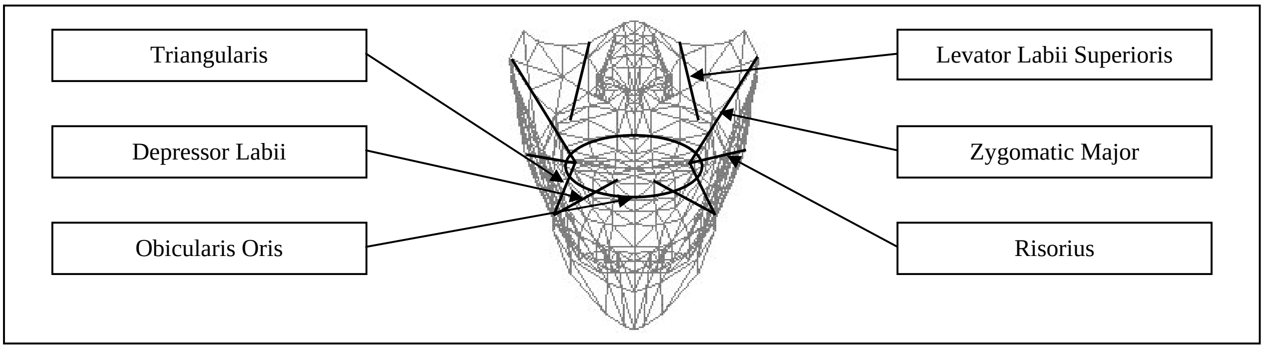

Figure 7 Positioning of muscle functions used in the production of visual speech postures

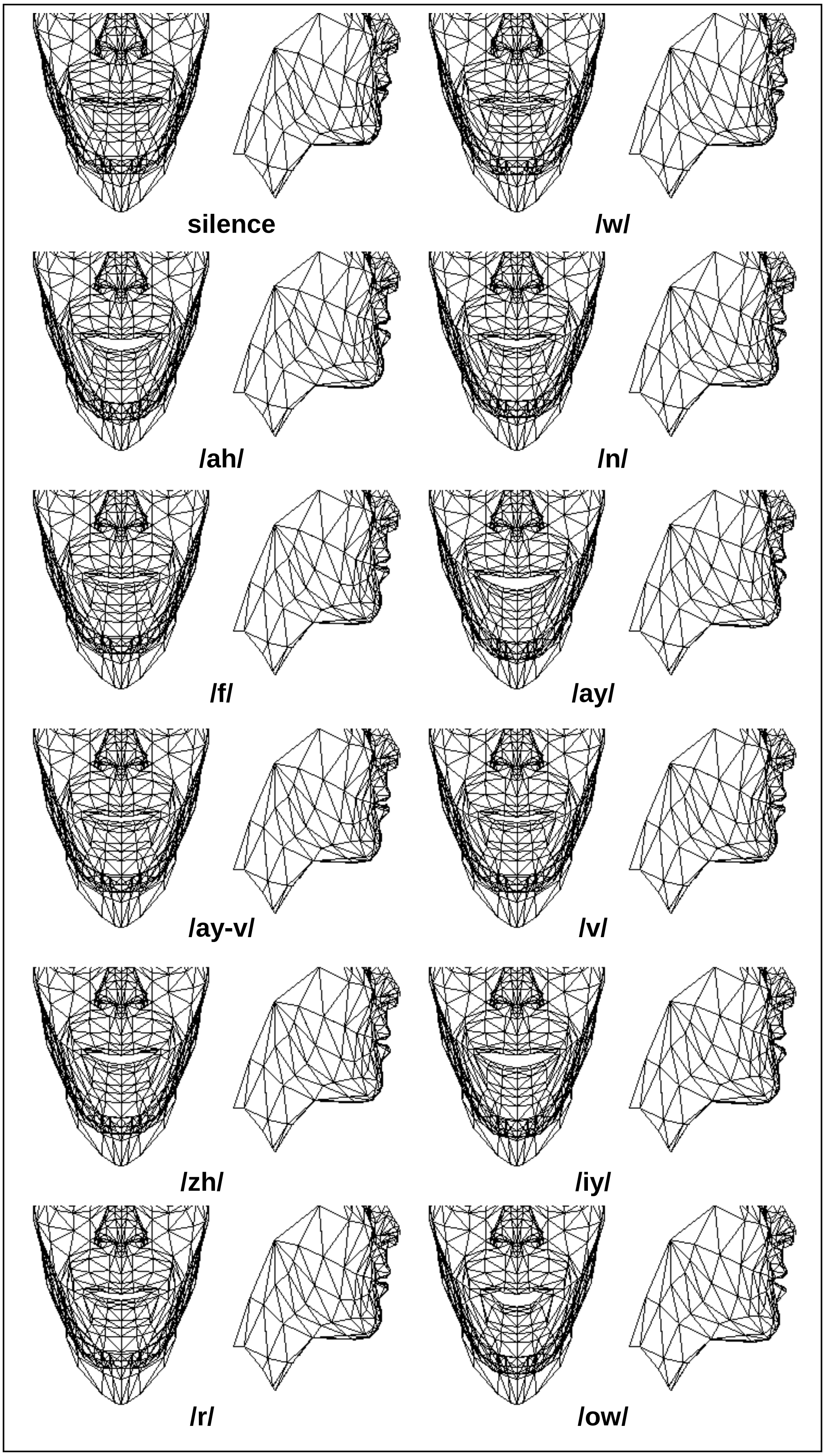

Figure 8 Example lip trajectory for the sentence ‘one five zero…’