

An Investigation into the Generation of Mouth Shapes for a Talking Head

Dr. A. P. Breen, Ms. E. Bowers & Dr. W. Welsh

BTL

ABSTRACT

BT is currently developing a low computation, real time, talking head as an adjunct to the Laureate text-to-speech system[1]. Research into the development of a talking head may be divided into two components; image generation, and face and head movement control. This paper concentrates on the last of the two.

A significant aspect of this work is research into methods of generating convincing mouth shapes when the head is talking. The paper describes a real time method of visual speech generation, which takes into consideration major coarticulation effects. It provides a detailed description of the generation process and compares this with a method of visual speech generation proposed by Cohen and Massaro[2].

1. INTRODUCTION

The BT head generation system[3] (MAXIS) uses one of the original methods of generating computer animation. A so-called wireframe model is used, which is really a data structure contained in the computer's memory. This data structure consists of a set of three-dimensional co-ordinates of vertices which lie on the surface of the object to be modeled e.g. a face. The vertices are joined by edges so that the whole structure represents a set of conjoined triangles. In the MAXIS system, texture mapping is used to project an actual colour image of a person's face onto a 3-D wireframe approximating the shape of the person's head (shown in figure 1). By moving the texture-mapped wireframe, an effect is obtained which has the appearance of seeing the actual person's head in motion.

The wireframe is controlled through a set of action units (AUs). An action unit is an index into a pre-defined set of vectors which describe how a set of vertices are to be modified when the action unit is applied. These action units may be sent singly or in combination to produce an overall effect. Associated with each action unit is a number between zero and one, representing a linearly increasing degree of application. Where a value of zero represents the minimum amount of modification, and a value of one represents the maximum amount of modification. Any number of action units may be defined and groups of action units can be ascribed to specific head gestures. For example, action units are defined to control blinking, eye brow movement and head rotation. In addition, groups of action units may be used to provide emotional cues such as happiness, sadness and amusement etc. The most significant group of action units, with regard to the work described in this paper, are those used to control the tongue, lips and jaw. However, as mentioned above,

action units may be applied in combination, that is, AUs used to generate emotion may be used in combination with those needed to generate effective mouth shapes.

From the definition provided above, it is apparent that an action unit can be very simple in concept, representing a basic facial movement, such as jaw drop, or very complicated representing all the movements needed to produce a viseme. Clearly, the number and variety of action units used to control a head, greatly affects the method of visual speech production.

The production of individual frames of a texture mapped image can often look very convincing, when the model begins to move however, it is a different matter. The problem can be split into two independent parts: the first part is the generation of the moving images and the second is the development of a model to drive the first part. The generation of motion can itself be divided into global motion such as rotations of the head and local motion due to face expression changes and mouth movement corresponding to speech. The generation of arbitrary global motion is easy: the set of vertices corresponding to the head part of the model is globally rotated or otherwise transformed; generation of face expression changes or mouth movements can be done in a similar way by moving sets of vertices according to the action unit definitions discussed above. However, in the second case it is realised that merely deforming the image in this way does not entirely lead to a realistic effect.

2. VISEME SET GENERATION

Speech has evolved over the course of thousands of years, and in the process changed human physiology, such that today, humans are capable of producing a bewildering array of complex sounds. These sounds are produced using the speech articulators. That is, the jaw, lips, teeth, tongue, velum and larynx. Both the nasal cavity and oral cavity are used to produce speech sounds. When we watch someone speaking, we are seeing an intricate interplay of these articulators.

A moment's introspection will reveal that a sound such as 't' in 'tea' differs in lip shape to the 't' in 'two'. These, sometimes dramatic, differences in facial movements due to context, are the visual correlate of the speech effect known as coarticulation. Any method of the mouth shape generation attempting to produce realistic mouth shapes, must have the ability to appropriately model the complex effects of coarticulation. Coarticulation and its affect on the design and collection of a set of visemes is considered below.

A language may be viewed as being composed of a discrete set of abstract symbols, the phonemes of that language. A

phoneme is an abstract representation of a sound, and the set of phonemes in a language is defined as the minimum number of symbols needed to describe every possible word in that language. In standard British English there are approximately forty two such phonemes.

A phoneme is not a sound, it may be viewed as the label for a set of sounds which when spoken as part of a word do not change the meaning of that word. There are an unimaginably large number of sounds used in a language. Due in part to the effects of coarticulation.

The problem is simply this, we as humans, do not speak in discrete units, speech is produced as a continuous flow of articulatory movements which upon introspection can be 'written' as a discrete set of symbols. This flow from one articulated sound to another is *coarticulation*. In other words, it is a term used to describe the effect of local articulation on a given sound. For expediency, coarticulation is often described as being the effect of context on the production of a phoneme.

The set of visemes in a language is often defined as the number of visibly different phonemes in that language, which from the discussion above is clearly non-sensical. The simplest sensible approximation to this definition is to say that a viseme set can be generated from a set of archetypal sounds in a language based on the phonemes of that language.

Table 1 shows how for the purposes of this procedure, the phonemes of standard British English have been group to form a set of visemes. The viseme set is inevitably a compromise between the desire to reduce the number of images needed and the desire to ensure that sufficient phonetic distinctions are retained.

Viseme Group	Phonemes ¹
Consonant 1	p, b, m
Consonant 2	f, v
Consonant 3	D, T
Consonant 4	s, z
Consonant 5	S, Z
Consonant 6	t, d, n, l, r
“Both”	w, U, u, O
Vowel 1	Q, V, A
Vowel 2	3, i, j
Vowel 3	@, E, I, {

¹ SAMP-PA phonetic alphabet

Table 1.

The need for a viseme set is a direct consequence of using complex action units in the generation process. Researchers such as Cohen and Massaro[2] do not use visemes explicitly. The method proposed in their paper attempts to generate realistic movements through an algorithm which combines basic action units to produce and overall effect. The combination process, associates a set of feature based dominance functions with each phoneme in the language. Where a phoneme will have been decomposed into a set of basic attributes such as lip rounding. In contrast, the work described here attempts to pre-generate a database of complex action units. Each action unit describing a specific di-viseme transition.

3. DI-VISEME DATABASE GENERATION

The aim of the work described in this paper is to develop a low computation mouth shape generation algorithm, which takes into account many of the know coarticulation affects observed in continuous speech. The procedure is as follows: The viseme set shown in table 1 is used as a base point for all the visually distinct sounds in southern British English. However, generating a set of action units based solely on these visemes will not produced accurate mouths shapes due to the affects of coarticulation mentioned above. Nor is it possible to devise a set of complex action units which are not subject to some form of continuous speech affect. The method proposed here is to build a database of visemes in context. If all the immediate context affects were considered, approximately 800 tri-viseme² video recordings would be needed. Once recorded, each tri-viseme would be analysed and a complex action unit generate. These action units, when applied to the pre-store image, would produce a close approximation to the desired tri-viseme image. Clearly attempting to analyse approximately 800 tri-visemes is unrealistic. An obvious compromise is to store di-visemes³ rather than tri-visemes. This reduces the number of video images needed to 128. In addition, if the further approximation is made that the coarticulation due to vowel production greatly outweighs the effects produced by consonants, the number of required di-visemes reduces to 48.

To minimise the amount of video data to be stored, the di-visemes were recorded as sets of nonsense syllables. Table 2 shows some of the nonsense syllables used in the production of the database.

² A tri-viseme records the imediate left and right context effects on a centre viseme.

³ A di-viseme records the change in articulation produced when moving from one viseme to another.

SAM-PA Symbols	Orthographic interpretation
pip	peep
fif	feef
sis	sees
SiS	sheesh
TiT	theeth
kik	keek
tit	teet
wiw	weew

Table 2

During the analysis phase, it was observed, that a number of the proposed di-visemes were almost identical or varied only slightly in the amount of mouth shape variation. As a result, a new set of action units were generated which relied on the ability of the MAXIS head image generation code to modify the degree of action unit application. This reduced the number of action units required to 13. Table 3 contains an example of the association of action units and degree to certain viseme transitions

Action Unit	description	Transitions	AU Degree
1	Rounded mouth, protruding lips, teeth close. Mouth shape get more rounded with degree.	Vowel group 2 to consonant group 5.	0.51
		vowel group 3 to consonant group 5.	0.49

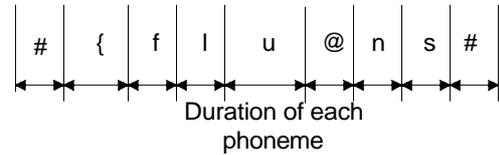
Table 3

4. MOUTH SHAPE GENERATION

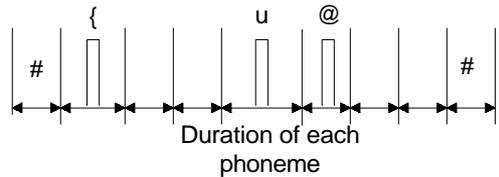
When an utterance is to be synthesised, text is presented to the Laureate text to speech system. Laureate produces synthetic speech and a data structure which contains knowledge gained during the synthesis of the utterance. In particular, this data structure contains phoneme type and duration information. This information forms the input the mouth shape generation component of the MAXIS head.

The generation process is based around the identification of vowels within an utterance. Vowels are used as articulatory targets and hence represent specific points of articulation within the production of an utterance. The current implementation

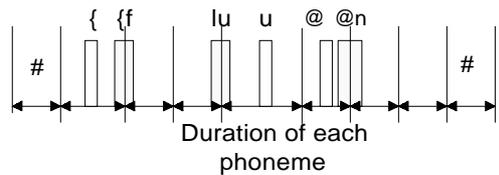
assumes that the mouth shape for a vowel will be at its implementation maximum by the mid duration point. The process is best illustrated through an example. Consider the synthesis of the word “affluence”.



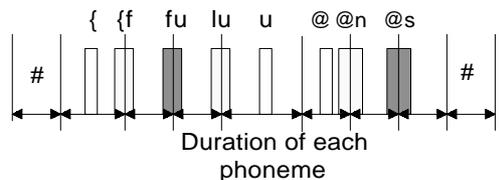
As stated above, the first phase is to determine the appropriate action units for the vowel centres.



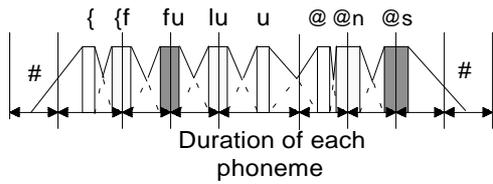
Once the vowels have been found then the appropriate di-viseme action unit for the transitions consonant-vowel and vowel-consonant are applied.



Due to the compromises discussed above, there are no action units to describe the /fl/ transition. Instead, the process selects an action unit based on the consonant-vowel transition /fu/. A similar problem occurs between at the transition /ns/. In this case, as there is no following vowel, the vowel-consonant transition /@s/ is used.



The action units so far assigned, act as visual land marks across the utterance. The final stage uses a linear interpolation function, to specify how the degree of application of each action unit should vary, as the visual image moves from one land mark to the other.



Once calculated, each action unit and its associated degree, is sent to the head player.

5. REFERENCES

1. Page, J. H. and Breen, A. P., 'The Laureate text-to-speech system - architecture and applications', BT Technical Journal, vol. 14, No. 1, January 1996.
2. Cohen, M. N. and Massaro, D. W. (1993), 'Modeling coarticulation in synthetic visual speech', in Thalmann N. M. and Thalmann, D. (Eds.), *Models and Techniques in Computer Animation*, '92, Tokyo: Springer-Verlag.
3. Welsh W. J., 'Model-based coding of moving images at very low bit-rates', Picture Coding Symposium (PCS 87), Stockholm, Sweden, June 1987.

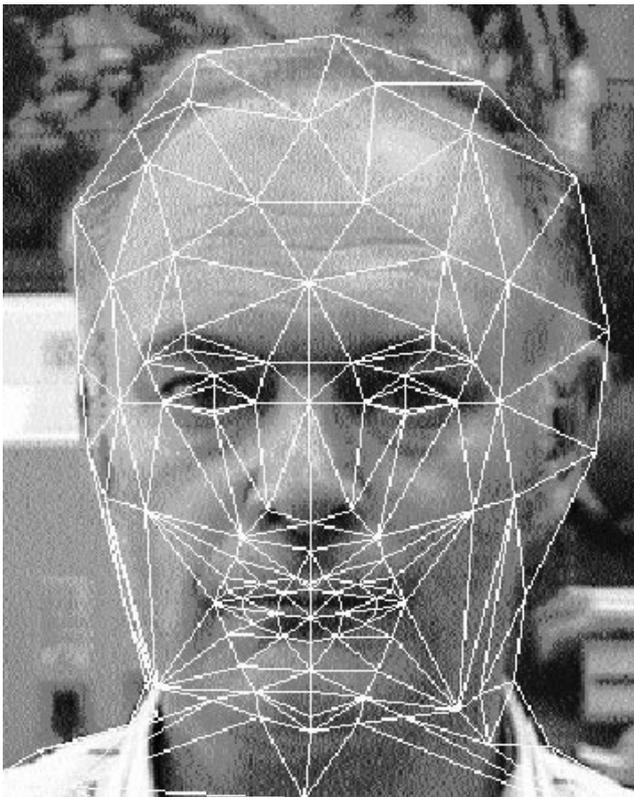


Figure 1
(MAXIS Wireframe superimposed on top of an image)