Introduction

Speech animation

The characters in animated movies are often required to speak. This requires the animator to vary the expressions on the characters' faces, in close synchronisation with the soundtracks provided by real actors, with particular attention to lip and jaw movements. The latter must plausibly represent the real articulations that would be required in speaking the words involved. As computer animation makes increasing inroads into the territory formerly occupied exclusively by human animators, there is an increasing need to synchronise computer generated facial expressions to speech. Manual methods are slow, and relatively costly, so that there is a rising interest in automating speech animation.

In some of the earliest work on the computer modelling and animation of human faces Parke (1974) included a study of what he called "speech animated sequences", using the traditional animator's approach of recording a soundtrack from a human actor, analysing the speech for key sounds, and generating the sequence of expressions needed to mimic the necessary articulations. In this case, he was not copying the expressions, but rather estimating the parameters needed to drive his facial model by self-observation, and programming the resulting data into the key frames for what was, therefore, semiautomatic animation by computer. The process was one of trial and error, as the initial parametric model was too simple (for example, it did not include movement of the upper lip), and algorithms for facial expression selection and specification did not (and still do not) exist. Eventually, he produced a segment he described as "quite convincing" that not only used an extra parameter to control upper lip movement, but also paid much more attention to the eyes, eyebrows and mouth expression in order to get some emotion into the animation. He noted that the lack of head movement (the neck was fixed) and the absence of a tongue were deficiencies.

Even now, when computer-animating a human face that is required to speak, the same basic manual process is still usually used. It is labour intensive, requires specialised skills and equipment, and adds greatly to the effort and cost of the completed product. At the SIGGRAPH 87 Animation Tutorial, the consensus seemed to be that, if speech animation is required, the usual procedures take all the "fun" out of animation.

Only a few studies of problems involved in computer animated speech have been reported since Parke's original studies, including an early paper on the work being reported here (Weil 1982; Bergeron and Lachapelle 1985; Pearce, Wyvill, Wyvill and Hill 1986; Lewis and Parke 1987). The latter work is particularly interesting since, although it involves recording a real actor, it mitigates the problems of the animation process by automatically recognising key sounds in the spoken script and using these to specify key frames in the animation process, thereby automating the lip synchronisation between the computer generated imagery and the real speech. Our approach is based on generating synthetic speech by rules, with suitable minor extensions to control the parameters needed to control facial expression. This not only automates lip synchronisation, but also avoids the need for actors, recordings and the like. However, the approach is not without its problems. Hill (1980) provides a review of speech recognition and synthesis by machine.

Modelling faces

A number of reports have appeared on the closely related problem of modelling faces (Platt 1980; Platt and Badler 1981; Parke 1982; Waters 1987). The trend has been away from the arbitrary structures for facial representation (polygonal nets)

on which Parke's original work was based towards a modelling system that embodies the underlying constraints of facial structure. Moreover, the parameter systems used to control the facial expression are based on a descriptive system devised Ekman and Friesen (1975; 1977) for their research on non-verbal communication -- FACS (Facial Action Coding System). This trend at once tends to simplify the process of controlling facial expression whilst simultaneously making the results more natural. Of course, the model embodies much more knowledge about facial structure. Thus real muscles and their effects are modelled and the next step is to emulate the flow of flesh over the underlying bony structures. It should be noted that Parke carefully distinguishes "animation" from "simulation" on the basis that the latter is intended as an exact model whilst the goal of the former is to communicate (an idea, story, or whatever), a process that allows, may even demand, considerable artistic licence. We are therefore using the term "emulation" which is intended to imply achieving a desired effect, using mechanisms that realistically capture the main real-world facts and structures, without necessarily being an exact simulation.

The rendered images accompanying Waters' paper are strikingly effective in capturing naturalistic expressions of fear, surprise anger, happiness and disgust. In the conference presentation of the work, some very convincing and natural speech animation was shown, based on speech from a real actor with manual synchronisation. Exactly how the parameter values to drive the expressions were derived is not specified in the paper, and the author comments that: "It remains very difficult to extract the parameters from real faces ... it is still open to question as to whether there are the techniques to extract the necessary facial parameters ...", referring specifically to the parameters for his model.

Problems with scripts and inbetweening

A number of problem areas remain, as a result of these kinds of difficulties, the expense of manual methods, and the related lack of flexibility inherent in the use of pre-recorded scripts. The process is not so easy to automate, and considerable problems arise if, after initial animation, someone wishes to change the script slightly. Essentially, a whole chunk is likely to need to be re-recorded and re-animated. Even if patching is reasonable, the original actor would have to record the patch, matching the voice and recording perfectly, and the animator would have to blend the patch in somehow, using ill defined rules. Facial expression is generally not a series of fixed postures. The face is continually mobile, even when not speaking and, since it is a major channel for non-verbal communication, the ebb and flow of expression must remain plausible at all times. The changes from one configuration to the next may not be linear, and different components may change at different rates and in different ways. The process of inbetweening is not straightforward, especially if, as in Lewis and Parke's approach (1987), only strong vowels are recognised. In addition, the extremes of Waters' animated emotions are only rarely achieved, and the subtle nuances of expression that normally play on the face are hard to catalogue and reproduce.

The contribution of speechreading

Thus, speech animation, as Parke (1982) stated, requires more than stylised lip and jaw movements to be convincing. This is why many professionals involved in helping deaf people to use visual cues, in understanding a speaker, prefer the term "speechreading" to "lipreading". Jeffers and Barley (1971, p.4) define speechreading as the art of understanding a speaker's thought by watching the movements of his or her mouth and facial expression, noting that the process is not exact, and that a speechreader must fill in a great deal of missing information using experience, context, general knowledge and the like, since many words (called

homophenes) are indistinguishable on the basis of visual information alone. This is a much worse problem than the auditory problem of words that *sound* alike (*homophones*) because, whilst all 40 or so speech sounds are distinguishable, their visual correlates (lip rounding, lip protrusion, jaw opening, and so on) allow the corresponding facial expressions to be separated into only a few (10 or so) categories (Jeffers and Barley 1971, pp 72-75). People cope with homophones quite well, as they occur relatively infrequently. Homophenes occur constantly. Despite the visual confusion, note that the auditory signal is more comprehensible when accompanied by the sight of the speaker's face (Massaro and Cohen 1983; Massaro, Thompson, Barron and Laren 1986). Nishida (1986) is attempting to supplement and improve acoustically based speech recognition with information gathered about the mouth using an image processing system. Even body movements are important. These facts indicate part of the problem of speaking to people over the telephone and explain why people like to *view* the speaker at public lectures. Sumby and Pollack (1954) showed the enhancing effect of visual cues on the intelligibility of speech heard in noise. If visual cues contribute to intelligibility of speech, it seems clear that they must be correctly produced in detail. Who has not been frustrated by a film, badly dubbed from its original language. Computer-animation of speech certainly needs good lip synchronisation: ultimately it probably demands that attention be paid to the non-verbal channels inherent in movements of the entire body, as well as facial expression.

Speechreading (lipreading) books and methods can provide one source of information for speech animation (Hazard 1971; Jeffers and Barley 1971; Jeffers and Barley 1979; Ordman and Ralli 1976; Walther 1982), though many approaches have derived the required information to specify expression from images of actors -- another tedious and error-prone process. Fromkin (1964) comments on the problems she found in measuring lip positions for vowels. It might be thought that modern measuring techniques (based on laser marking and the like) would have solved the problems. However, there are still difficulties for the researcher wishing to specify shapes, even apart from residual physical stability problems. For one thing, speech is continuous and, just as it is difficult to segment the acoustic waveform into segments corresponding to speech sounds, or to characterise these sounds after the attempt, so it is difficult to segment and characterise the continuously varying facial expression (including lip and jaw movement). Since there are variations between the lip and jaw positions due to context effects (co-articulation) as well as simple variability, recognising the articulations from visual data presents problems anyway.

Then there is the problem as far as lipreading recognition is concerned (previously noted) that many sounds present virtually identical visual cues, and even when they differ (as different vowels differ), the distinctions are of degree, rather than kind, and may be hard to make. Indeed, the variability in vowel appearance will normally exceed the nominal differences, especially with different speakers, and precise versus casual speech. Also many sounds which people think of as simple vowels (the "vowels" in "made", "fight", "toy", "coat", British RP English "fire" and so on) actually consist of two or even three vowel sounds in succession. Lipreading manuals normally group consonant sounds into a mere four categories, with one pair of categories distinguished solely on the basis of tongue movement which is difficult to see, if it is visible at all. All these factors tend to make the recognition and synchronisation of speech based on the analysis of lip and jaw movement more difficult than might be expected, outside of the laboratory.

Synchronised body movements

Condon and Ogston (1971), after a ten year study based on hours of videotapes,

characterised all body movements as controlled by muscle action units. Their work parallels that of Ekman and Friesen (1975; 1977) which led to the FACS system for coding facial expressions. Interestingly, an important unexpected finding in Condon and Ogston's work was that the elementary movements based on action units, besides exhibiting varying group synchronisation amongst themselves, for body movements, were also synchronised with major speech events (phonetically important markers resulting from articulatory processes), not only for the speaker's movements, but also for those of listeners, and even when the listeners were newly born infants, apparently unable to speak or understand. This finding suggests that, quite apart from the obvious need for lip synchronisation and facial expression synchronisation, for the speaker, a mechanism for correlating the body movements *of all those present* with the any concurrent speech is of fundamental importance in computer animation, if the overall intent is to emulate real situations. Artistic licence may even require an exaggeration or caricature of the effect.

Managing expression, body language, and naturalness

The problem of categorising and synthesising body language is a research question at present. Indeed, it is not even known how to control intonation and rhythm in an appropriate manner for speech synthesised from text, and the appropriate categorisation and control of facial expression and other body movements is at least as difficult, if less obtrusive. As already noted, even the information needed to control lip, jaw and tongue movements for computer animated speaking faces is hard to come by, and not entirely obvious in its application. However, the movement towards parametrically controlled faces, based on real muscle structures, should be of considerable help in allowing a small number of meaningful parameters to produce the required complex visual effects. Thus, for example, an older model may allow lips to be decribed in terms of opening height, opening width, and protrusion, but the visual effect may be quite unnatural because the shape may be wrong, or may go through unacceptable transitions during movement. A model which emulates the effect of the *orbicularis oris* -- the sphincter muscle that rings the mouth, the *levator labii superioris alaeque nasi* -- which connects to the upper lip, and so on, should mean that however the controlling parameters are manipulated, the resulting facial expressions will be within the range of natural human expressions. Of course, that still leaves the problem of defining and selecting appropriate expressions and expression sequences for particular purposes such as speaking. Acquiring these from real people speaking the required utterances, as is done, is not and entirely satisfactory solution. Apart from general difficulties, including difficulties of measuring parameters, already noted, there remain some thorny fundamental problems that aggravate the logistical and practical problems of collecting suitable *expressions* for use in speech animation. For example, no two versions of the same utterance are ever spoken in quite the same way, and interactions between successive expressions undoubtedly occur, just as do interactions between successive sounds (co-articulation). Inability to deal with variation, co-articulation, rhythm, and intonation, adequately, is a primary cause of unnaturalness in speech synthesised by rules (as opposed to being re-synthesised from a compressed real speech original). It has been shown (e.g. Holmes 1961; Holmes 1979) that lack of naturalness in parametrically generated synthetic speech is *not* due primarily to synthesis model limitations, but rather to our lack of knowledge about the details of how to drive the model. Analogous problems arise in animating repeated sequences of facial expression and other body movements. For example, Ekman (Ekman and Friesen 1975; 1977) tells us that there are 268 muscles responsible for human facial expression. He learned (with considerable effort) to control these individually and catalogued around 55,000 distinguishable facial expressions. That a subset of 30 of these involves recognisable semantic

distinctions is not as much help as it might seem. There are only 40 or so basic sounds in English (so-called *phonemes*), but putting these together, by rule, into natural sounding synthetic speech still presents serious research problems. This is because phoneme categories are perceptual categories, not sounds *per se*.

An analogy from speech synthesis

By basing a face model upon the real constraints of faces, a great deal of the structural complexity could be wrapped up in the model, and, if the model could be made good enough, and could be tied appropriately to simple, natural, controlling parameters, much of the detail arising from underlying structure could be taken care of automatically. An analogous example arises in the synthesis of speech. One method of synthesising speech (used in the seminal work to elucidate speech cues at the Haskins Laboratories in the '50s and '60s (e.g. Liberman 1957)) controlled the required output by specifying the amount of energy present at all relevant frequencies and times -- a spectrographic synthesis approach. Almost none of the constraints of the human vocal apparatus were present in this "acoustic analog" and, with suitable input, it could just as well make music as speech. The input data required was very detailed, and had to account for the constraints of vocal apparatus explicitly and continuously, as well as controlling the variation in output needed to mimic the individual speech sounds. Another approach used a "resonance model" of the human vocal apparatus that embodied far more of the constraints of the human vocal tract, in particular constraining the output to that which could be produced based on the physical form of the vocal apparatus. Even when the parameters were manipulated in a totally random and violent fashion, the result, whilst not speech, was a very natural sounding "belch" (first author's previously unpublished result). The input data were much simpler as a result. Flanagan (1972) and Witten (1982) provide a thorough treatment of the theory and practice of all aspects of speech analysis, synthesis and recognition by machine.

A simple method for rule based animation

Speech synthesis and facial expression synthesis

In early work on the synthesis of *speech*, problems arose, and solutions were adopted that, with hindsight, are highly reminiscent of those involved in animating facial expressions. Initially, the aim was to produce the same auditory effect as real speech, based on an abstracted (parameterised) model of the sound pattern, rather than on a direct recording of the speech waveform. The advantages of success were seen to be twofold. First, a compressed representation might be possible (in those days engineers wanted compression in order to reduce the transmission bandwidth for long-distance telephone traffic); and secondly, insight into the nature of speech might be gained, with implications for the recognition and generation of speech by machine. Incidentally, the compressed form of speech was quickly seen as providing a basis for encryption of speech for secure telephone conversations. The vocoder, which was an automated form of the acoustic analog mentioned above, parameterised speech in terms of the energy present in a number of contiguous frequency bands, together with parameters specifying whether sounds were voiced or not, and (if voiced) the pitch frequency. By encoding the parameters digitally, and encrypting them, the wartime "scrambler" telephone was made possible. There are parallels with image generation and transmission in general.

Some of the equipment designed for speech synthesis (the Haskins Laboratory "Pattern Playback", for example) lent itself to psychoacoustic experimentation to determine the auditory cues needed to perceive speech, and hence opened the door to modern speech recognition and and rule-based synthesis by machine. The input to the Pattern Playback, was analogous to a polygon mesh representation of the face in the sense that every point in the crude quantised frequency-time

spectrogram needed to be specified, and the Playback machine then "rendered" the spectrogram, converting it into sound with auditory deficiencies equivalent to the visual deficiencies of rendered polygon mesh faces. Figure 1 shows a very detailed spectrogram, derived from the real speech utterance "zero", but with an upper frequency limit of around 5500 Hz. It can be seen that certain features are prominent: vertical striations (representing synchronous excitation of the damped resonances of the vocal tract by successive vocal fold vibrations, seen at many analysis frequencies); slowly varying dark bands (formants) representing the varying resonant frequencies of the changing shapes in the vocal tract; high frequency noise at the beginning, marking the rush of air between tongue and teeth for the inital "zed" (phonetically /z/); and so on. A resonance analogue synthesiser would model these effects explicitly, and use input parameters that controlled them, directly. The Pattern Playback used a simplified form of the spectrogram where these features were caricatured in a simplified artificial spectrogram, used to control the individual amounts of energy in a time-frequency-energy model. This latter approach is like animating faces by controlling the mesh points in a polygon mesh, rather than by controlling key features like lip height and lip width. A face animation system equivalent to the resonance model would have elements that modelled the appearance of mouths, eyes, cheeks, and so on, in terms of important visual features, and would use parameters specifying these features to control them.

We do not know of a facial animator that works just like this although many of the features used to control whole face animation are obviously closely related. Thus Parke's face model (now extended to a whole head model (Parke 1974; Parke 1982; Lewis and Parke 1987)), with controlling software to co-ordinate the movement of mesh points under the control of a few visually obvious parameters, is close.

Waters' approach based on muscles and the FACS parameterisation is not really analagous to a resonance model. It is much more like a third approach to speech synthesis -- namely articulatory synthesis (Coker 1976) in which the behaviour of the vocal tract muscles are modelled, airflow is simulated, and speech results. Some so-called articulatory models control vocal tract areas rather than modelling the muscles *per se*. It is dangerous to push these kinds of analogies too far, but one of the big problems of articulatory synthesis, whether based on physiological or vocal tract area models, is that the underlying parameters, whilst real, are very hard to measure and use as a basis for synthesis. Thus, for facial expression synthesis, if we model the *orbicularis oris* as a basis for manipulating a mouth correctly, and we assume that our model is good enough to "work", how do we know what tensions to apply to achieve the many facial expressions we must use. Also, we have to worry about the categorisation, abstraction and selection of all possible facial expressions for use in particular animation situations (and that raises a horrendous combinatorial problem, given enough parameters).

The perception of speech is truly categorical. That is, for a native speaker of a particular language, one phonetic category does not change continuously into the next category, but there is a relatively sharp perceptual boundary. A sound will be perceived as belonging to one category or the other with essentially no ambiguity. It is unlikely that this is true of facial expression. In a sense, this difference should make facial expression animation easier, because mistakes will lead to untypical expressions rather than wrong expressions. However, it also makes it harder, because psychovisual and psycholinguistic experiments, to structure the cue space, and evaluate success in using the cues, will be much harder to design and much less definitive. That is, it will be much harder to determine what *are* the important visual cues and prototypes for face expression (and other body language), and much harder to find out how to vary these cues appropriately.

At the same time, many of the hoary old problems found in speech synthesis and recognition will apply to expression synthesis and recognition (problems of sequential interaction, individual "accents", and segmentation -- knowing where one expression ceases and a new one begins). It seems likely that there will be characteristic sets of expressions for different "language" groups.

A method for speech animation by rules

Because of the perceived similarities between speech synthesis and facial expression animation, and because it seems a worthwhile approach to investigate in its own right, we decided to try to automate the animation of speaking faces by extending our existing speech rule-based synthesis program (Hill 1978).

We happen to be using a resonance analogue speech synthesiser at present, but this is not essential. Individual speech postures are represented as rows in a table of parameter values. Each parameter controls a perceptually relevant acoustic cue derived from the resonance model on which the synthesiser is based. The method of generating the parameters for a particular utterance is based upon the kind of knowledge gained from work such as that of the Haskins Laboratories, including some in-house research, and there is a close relationship between the parameter variations generated and the necessary perceptual cues (some quasi-static, some dynamic) for the speech postures. Thus speech is created "by rules", without any reference to any "real" speech. Each speech posture may be thought of as an abstract of a phonetic sound category for English, an archetype. We resist the temptation to use the classic term "phoneme" because, although it is accurate, is is greatly misunderstood by non-phoneticians. An utterance is specified as a sequence of posture specifications and the synthesis program uses the table entries, together with knowledge about interactions, intonation, rhythm, and the like to "render" the table values into a continuous set of parameter specifications, suitably interpolated between postures, to drive the synthesiser hardware and thereby create the dynamic acoustic cues. The interpolation from posture to posture is a piecewise-linear approximation to the basic dynamics of articulation which may be thought of crudely as sinusoidal. The hardware could, of course, be simulated by suitable software, although a stage of digital-to-analog conversion would still be required to generate accurate soundwaves.

A year or two ago, courtesy of the author, we obtained a copy of the polygon mesh face model that Fred Parke developed (Parke 1974). The nature of the parametric control for this face, as already noted, is quite analogous to that used for our speech synthesis, although the parameters hook onto fairly crude polygon mesh point adjustment, rather than representing fundamental structural properties of the face. Nine parameters control the jaw and lips as detailed in Table 1.

Parameter	Value Range
Jaw rotation	$0 \leq$ value ≤ 20
Mouth width	$0.5 \leq$ value ≤ 1.5
Mouth expression	$0 \leq$ value ≤ 1
Lip protrusion	$0 \leq$ value ≤ 30
X, Y & Z of mouth corner	$0 \leq$ value ≤ 25
/f/ and /v/ lip tuck	$-20 \leq$ value ≤ 0
Upper lip position	$0 \leq$ value ≤ 20

Table 1: Nine parameters controlling jaw and lips for Parke's face model

To extend the speech synthesis program to handle face animation we simply extended the rows of the rule table for speech postures to include the additional parameter specifications needed for appropriate facial postures (expressions) that included lip and jaw movement. The generated parameter tracks were then sampled at the required frame rate and used to drive the renderings needed to animate the face in exact synchronism with the voice . The face parameter values were derived by measurements taken from photographs illustrating Walther's book (Walther 1982) which, unfortunately, does not include side views. Side view values were estimated. The absolute measurements were scaled appropriately for the computer face. Figures 2 through 5 show some of the illustrations used. A short animated sequence of the face saying "Hi there, how are you" was demonstrated at Graphics Interface 86 (Pearce, Wyvill, Wyvill and Hill 1986). Other aspects of facial expression and body movements generally could be synthesised using the same technique, if it were known what motions were required. The synchronism required by Condon and Ogston's findings would be built in, due to the basis for the computations, assuming the timing requirements were known. Obviously, any reasonable set of parameters could be used, but those clearly related to the appropriate visual cues are easier to work with at this stage of the work, all other things being equal.

Problems and further work

Poor speech quality

There are many obvious problems in our research so far. A major problem, that has concerned us in our speech synthesis research and was noted by Parke (Parke 1987) is the poor quality of speech synthesised by rules. Reasons for this have been suggested, and we have made progress in the area of rhythm (Hill, Witten and Jassem 1978; Jassem, Hill and Witten 1984). A great deal of the synthetic speech that is heard by the average person is *not* synthesised "by rules" but is reconstituted from a compressed version of original real speech. This kind of synthetic speech, usually based on Linear Predictive Coding (LPC) compression and resynthesis schemes these days, is potentially completely natural, but relatively inflexible. It is the form used in some speaking toys (e.g. Texas Instruments' "Speak and Spell" spelling trainer), and many automatic telephone intercept systems. If speech quality is less than perfect, it is only because economies have been made at some stage in the process, often to save storage in the final system (as in the spelling toy). Others (e.g. Parke 1987) have suggested that some mechanical quality of the voice is actually desirable, in order to match the synthetic nature of the animated image. However, even the best speech synthesised "by rules" leaves something to be desired, especially as far as intonation is concerned. However, the low quality may be acceptable for some applications since, unlike reconstituted speech, speech synthesis by rules allows the animation of unrestricted speech from text, without actors, recordings or manual and semi-automatic procedures requiring skill and time. It would be interesting to find out, under these particular circumstances, the extent (if any) to which intelligibility and acceptability are enhanced by the provision of visual cues, effectively repeating Sumby and Pollack's (1954) work in the new conditions.

Lack of detailed knowledge

The next problem is concerned with the details of the animation and parameter manipulation. Part of the problem of good speech synthesis by rules lies in the details, as opposed to the general method of interpolating the parameters; for example, adjusting them for interactions. The same is certainly true in synthesising lip and jaw positions. Indeed, some of the acoustic variation in speech that must be taken into account in improving the details of the acoustic parameter control arises from anticipatory articulations, such as adjusting the lips in anticipation of

the following vowel when articulating a consonant (such as /t/). The /t/ in "tea" looks different to the /t/ in "two". Clearly, this argument runs counter to earlier remarks about the indistinguishability of sounds, but what can be distinguished from point of view of identification is not the same as what looks natural in context. Another detail, already noted, is that the whole face, even the whole body and the bodies of others, need to be animated in some degree of synchronism with the speech. Given synthetic speech, the synchronism *per se* is not a problem, but exactly what movements and expressions need to be synchronised, and with what relative timing, cannot be stated for now.

Realism and "style"

Then there are problems of realism and individual "style". Figures 6 through 10 show the rest position, two consonants (/p/ and /t/), and two vowels (/a/ and /u/, as in "hard" and "who'd" respectively), animated according to the method outlined above in the phrase "Speak to me now, bad kangaroo". Figures 11 through 15 show the same postures articulated by one of the authors. These may be compared with the illustration from Walther's book (figures 2 to 5) for some of the same basic prototypes (the effects of articulatory context are left unconsidered). It will be noted that there are all kinds of discrepancies and differences. It is often assumed that /p/ and similar sounds articulated with lips together "look like" the rest position. This is not so, as may be seen in figures 11 and 12, especially in the side views. There is an additional degree of muscle tension (that varies according to the stress on the carrier word) for the /p/.

The animated face has its own character, and the translation from Walther's specifications are only approximate, to put it charitably. The degree of control in our model is limited, and the quantification involved subjective factors in the conversion. This may not be a problem since there is just as much variation between the two real faces, as may be seen. Which kind of variation (if any) matters, and which is insignificant, is an open question, but it *is* worth reflecting on the fact that both real mouths are making the "same" sounds, and are both controlled by the "same" underlying muscles. Some consistency of style for one animated character, and some distinction between styles for different animated characters, certainly require attention. In this connection, it is also worth noting that there is a considerable difference between carefully enunciated, deliberate (or "citation") speech, and normal conversational speech. In the former, articulations are much more precise and extreme than in the latter. Informally, the difference is like the difference between the elocutionist's "How do you do" and the cocktail party acquaintence's "Harjedoo".

Better face modelling

We should very much like to adapt our method to use a face model based on real muscle and bone structures, using the Facial Action Coding System. In principle this may be straightforward, if we can obtain, or write, the necessary software, and the resulting image should then be much more natural. However, it will not solve the problems already summarised, and the relationship between muscle tensions and visible effects is not obvious. It may well be necessary to delve deeply into the perception and discrimination of facial expression, using computer generated images based on the FACS approach, to get a feel for the cues involved, the method of control, and its application. This would be a major research effort. Work by Boston (1973), De Soete (1987 and Dewdney (1986) seems highly relevant, and Massaro, previously cited for other work, is known to be planning research in the area (private communication).

Aliasing

A technical problem arises from the interaction between parameter construction and

the frame sampling rate. In our synthesis, the parameter variations associated with speech articulation are assumed to occur at a maximum rate of around 50 Hz, and the reconstruction is properly interpolated, based on a sample period of 10 msecs. The result is then sampled at the frame rate for film or video (24 or 30 frames per second), which is a good deal lower than the minimum 100 samples per second Nyquist rate required to avoid aliasing problems. In any case, we should almost certainly increase the maximum parameter frequency to 100 Hz, or even 500 Hz, for a variety of reasons to do with speech quality and ease of computation which could aggravate the situation. To improve the quality of the animation, and avoid temporal aliasing effects, we now apply motion blur to the facial parameter values by applying a Gaussian filter to neighbouring sample values. The same effect occurs in recording real actors movements of all kinds, due to the finite response time of the recording media. It is interesting that the visual perception system actually perceives a better image by observing blurred images in succession than one might expect by examining each individual image , as may easily be seen by observing the difference between an action sports shot, and a single frozen frame from the same sequence, in instant television replays.

Conclusions

The paper has reviewed the background and problems of animating speech and has presented a novel approach to automating the process based on a simple extension of computer speech synthesis by rules. The extra parameters needed to control lips, jaw and facial expression are simply added into the table of parameters needed to control the speech itself. The technique could be applied to other body movements. Two animated sequences: "Hi there, how are you"; and "Speak to me now, bad kangaroo" have been synthesised, and some illustrations of the results appear above as a basis for comparison with real faces. We have neglected to synthesise more material because of resource constraints for rendering, *not* because of any appreciable constraints associated with speech animation. A number of problems require further work, including: speech quality; the details of the face expressions; realism and style of the animation; better face modelling and parameterisation; other body movements; and antialiasing in connection with the framerate sampling of the faster speech events. However, as a first step in low-cost facial expression animation for talking characters, the work has produced satisfactory initial results, and shows promise for the future.

Acknowledgements

We would like to acknowledge with gratitude the support of the Natural Sciences and Engineering Research Council for this work, and to thank Fred Parke, who supplied us with the original face data, for his help and advice with the Graphicsland work. We would also like to thank Trevor Paquette for his skill and willingness in driving the Graphicsland software to produce the examples with which the paper is illustrated, as well as Jim Paterson and Usher Fleising for helpful discussions on facial expression. Richard Esau also deserves thanks as he was the programmer for the current implementation of the first author's speech synthesis by rules algorithm, and has contributed time and thought to matters directly related to the work reported.

References

Boston DW (1973) Synthetic facial communication. Brit J Audiology 7 (1): 95-101

Coker CH (1976) A model of articulatory dynamics and control. Proc IEEE 64 (4): 452-460

Condon WS, Ogston WD (1971) Speech and body motion synchrony of the speaker-hearer. in "The Perception of Language" (DL Horton & JJ Jenkins eds),

Merrill, Columbus, Ohio, pp 150-184
De Soete G (1987) A perceptual study of the Flury-Riedwyl faces for graphically displaying multivariate data. Int J Man-Machine Studies 25 (3): -
Dewdney AK (1986) The compleat computer caricaturist and a whimsical tour of face space. Sci Am 255 (4): 20-28
Ekman P, Friesen W (1975) Unmasking the human face. Consulting Psychologist Press, Palo Alto, California
Ekman P, Friesen W (1977) Manual for the facial action coding system. Consulting Psychologist Press, Palo Alto, California
Flanagan JL (1972) Speech analysis, synthesis and perception. Springer-Verlag, Berlin, Heidelberg, New York
Fromkin V (1964) Lip positions in American English vowels. Language and Speech 7 (3): 215-225
Hazard E (1971) Lipreading for the oral deaf and hard-of-hearing person. Charles C Thomas, Springfield, Illinois
Hill DR (1978) A program structure for event-based speech synthesis by rules within a flexible segmental framework. Int J Man-Machine Studies 10 (3): 285-299
Hill DR (1980) Spoken language generation and understanding by machine: a problems and applications oriented overview. in Spoken Language Generation and Understanding (Simon JC, ed), NATO ASI Series, D. Riedel, Dordrecht: 3-38
Hill DR, Witten IH, Jassem W (1977) Some results from a preliminary study of British English speech rhythm. 94th. Meeting Acoust Soc Amer, Miami, Dec 12-16 (available as Dept of Comp Sci, U of Calgary Report Number 78/26/5)
Holmes JN (1961) Notes on synthesis work. Speech Transmission Laboratory Quarterly Progress Report, Stockholm, April.
Holmes JN (1979) Synthesis of natural-sounding speech using a formant synthesiser. in Frontiers of Speech Communication Research (Lindblom B, Ohman S, eds) Academic Press, London, pp 275-285
Jassem W, Hill DR, Witten IH (1984) Isochrony in English speech: its statistical validity and linguistic relevance. in Pattern, Process and Function in Discourse Phonology (Gibbon D, ed), de Gruyter, pp 203-225
Jeffers J, Barley M (1971) Speechreading (lipreading). Charles C Thomas, Springfield, Illinois
Jeffers J, Barley, M (1979) Look, now hear this. Charles C Thomas: Springfield, Illinois
Lewis JP, Parke FI (1987) Automated lip-synch and speech synthesis for character animation. Proc. Human Factors in Computing Systems & Graphics Interface (CHI+GI 87), April 5-9, Toronto, Carroll JM, Tanner P (eds.): 143-147
Liberman AM (1957) Some results of research on speech perception. J Acoust Soc Amer 29 (1), 117-123
Massaro DW, Cohen MM (1983) Evaluation and integration of visual and auditory information in speech perception. J Exp Psychology 9 (5): 753-771
Massaro DW, Thompson LA, Barron B, Laren E (1986) Development changes in visual and auditory contributions to speech perception. J Exp Child Psychology 41 (1) 93-113
Nishida S (1986) Speech recognition enhancement by lip information. Human Factors in Computer Systems (Proc. CHI 86) (Mantei M, Orbeton P, eds), Boston April 13-17, Assoc for Computing Machinery, New York, pp 198-204
Ordman KA, Ralli MP (1976) What people say. Alexander Graham Bell Association for the Deaf, Washington, DC
Parke FI (1974) A parametric model for human faces. Ph.D. Thesis, Dept. of Computer Science, University of Utah, December
Parke FI (1982) Parameterized models for facial animation. IEEE Computer Graphics and Applications 2 (9): 61-68

Pearce A, Wyvill B, Wyvill G, Hill DR (1986) Speech and expression: a computer solution to face animation. Proc Graphics Interface 86 Conf, Vancouver, May 26-30, Canadian Information Proc Soc, Toronto, Ontario, pp 136-140

Platt SM (1980) A system for computer simulation of the human face. M.Sc. Thesis, The Moore School, University of Pennsylvania, August

Platt SM, Badler NI (1981) Animating facial expressions. Computer Graphics 15 (3): 245-252

Platt SM (1986) Structure-based animation of the human face. Internal Report, Engineering Dept., Swarthmore College

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Amer 26 (2), 212-215

Walther EF (1982) Lipreading. Nelson-Hall, Chicago

Waters K (1987) A muscle model for animating three-dimensional facial expression. Computer Graphics 21 (4): 17-24

Witten IH (1982) Principles of computer speech. Academic Press, London, New York, Paris, San Diego, San Francisco, Sao Paulo, Sydney, Tokyo, Toronto

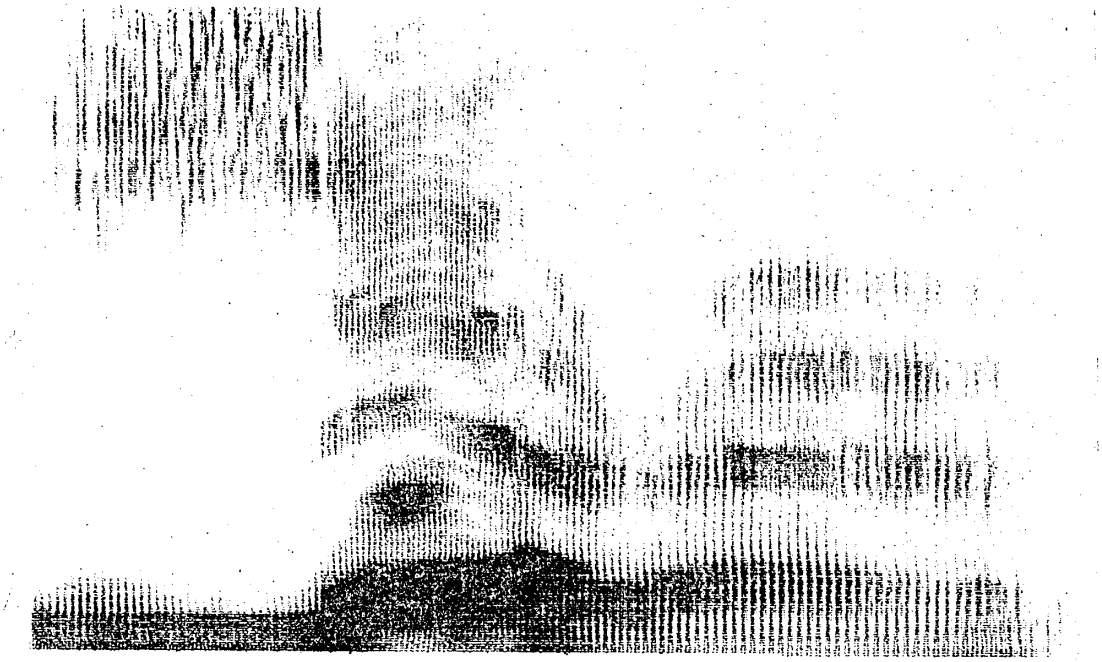


Figure 1: Detailed spectrogram of "zero"

Figure 2: Rest position (Walther)

Figure 3: Articulation of /t/, /s/, ... (Walther)

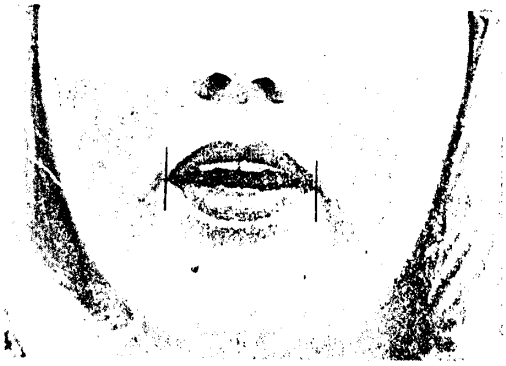


Figure 4: The vowel /a/ (as in "hard") (Walther)

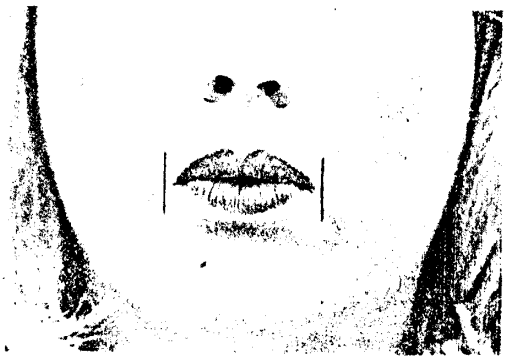
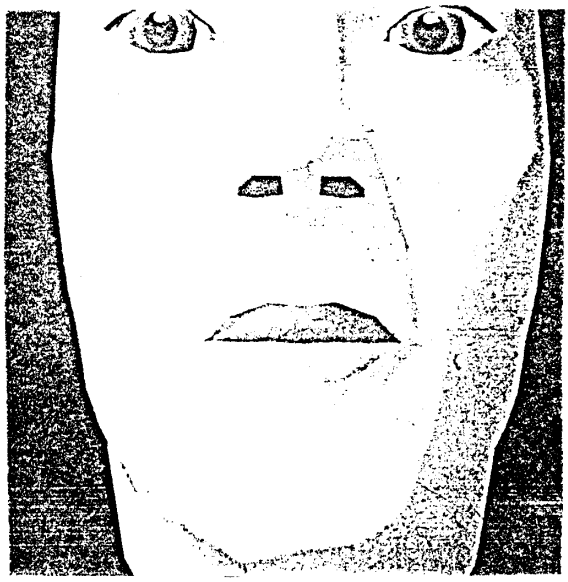


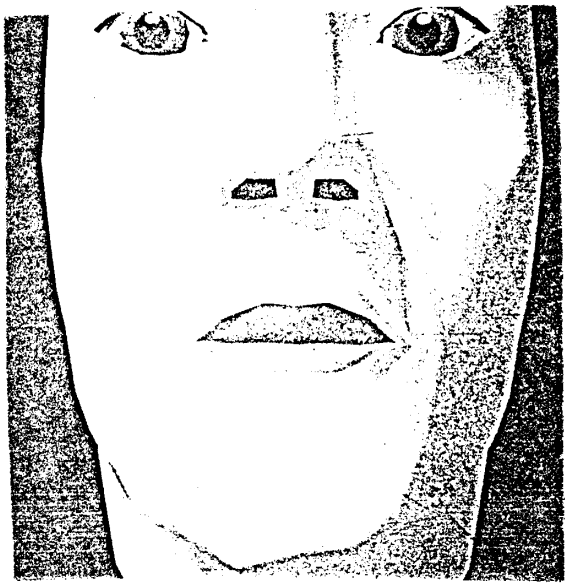
Figure 5: The vowel /u/ (as in "who'd") (Walther)



6(a): Rest position (front)

6(b): Rest position (side)

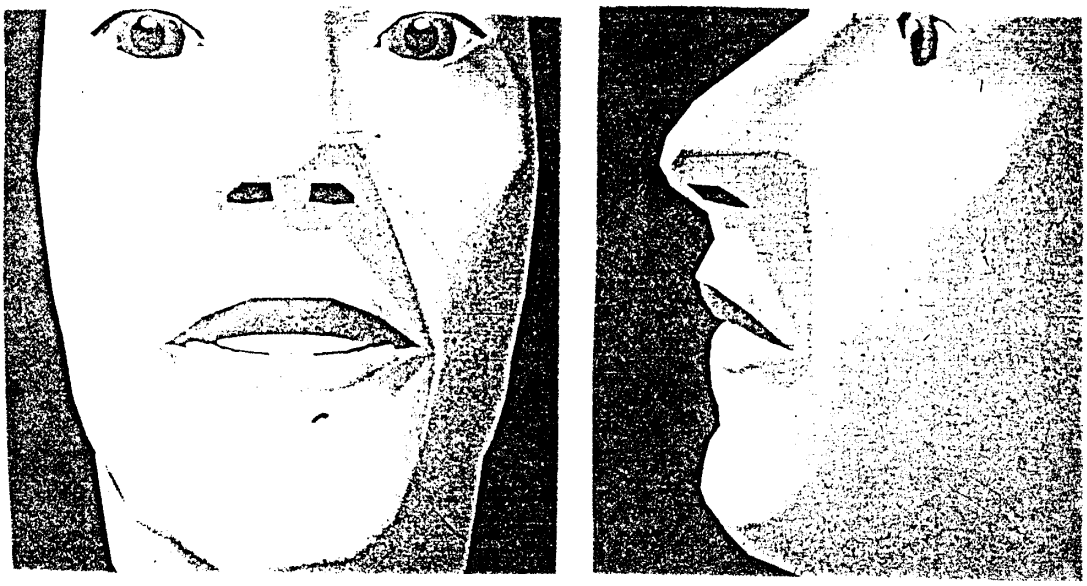
Figure 6: Front and side views of animated face rest position



7(a): Articulation of /p/ (front)

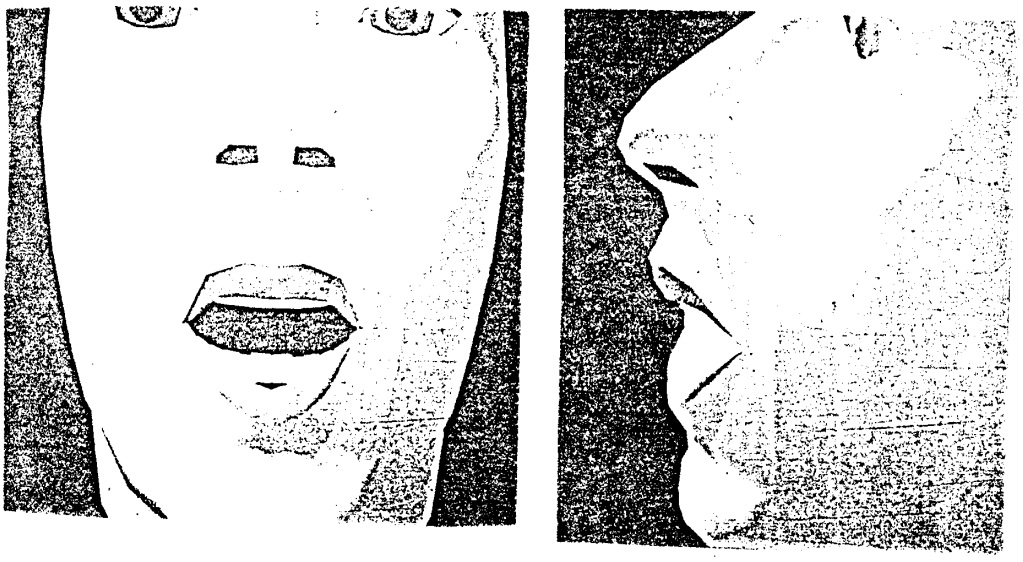
7(b): Articulation of /p/ (side)

Figure 7: Front and side views of animated face /p/ articulation



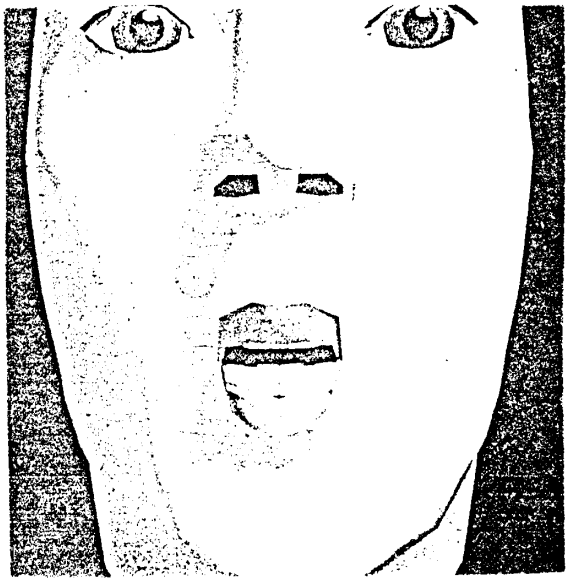
8(a): Articulation of /t/ (front) 8(b): Articulation of /t/ (side)

Figure 8: Front and side views of animated face /t/ articulation



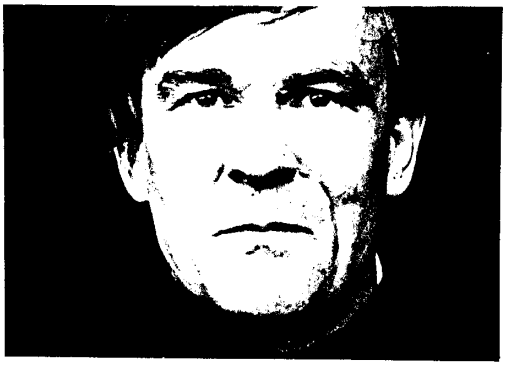
9(a): Articulation of /a/ (front) 9(b): Articulation of /a/ (side)

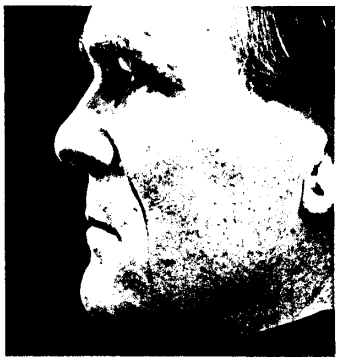
Figure 9: Front and side views of animated face /a/ articulation



10(a): Articulation of /u/ (front) 10(b): Articulation of /u/ (side)

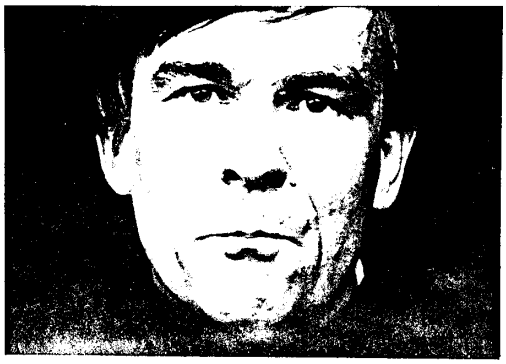
Figure 10: Front and side views of animated face /u/ articulation

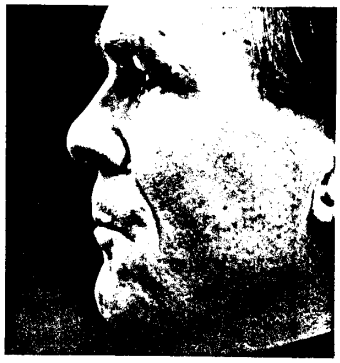




11(a): Rest position (front) 11(b): Rest position (side)

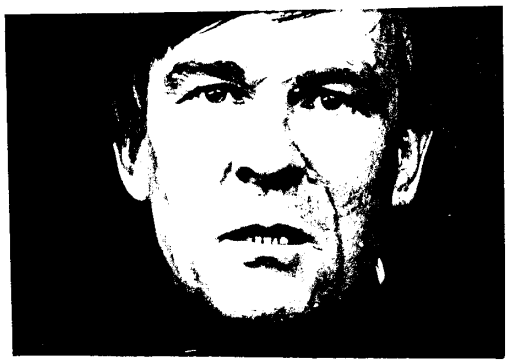
Figure 11: Front and side views of author's rest position

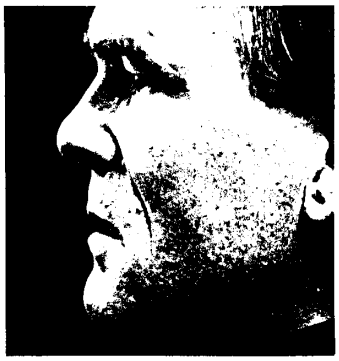




12(a): Articulation of /p/ (front) 12(b): Articulation of /p/ (side)

Figure 12: Front and side views of author's /p/ articulation

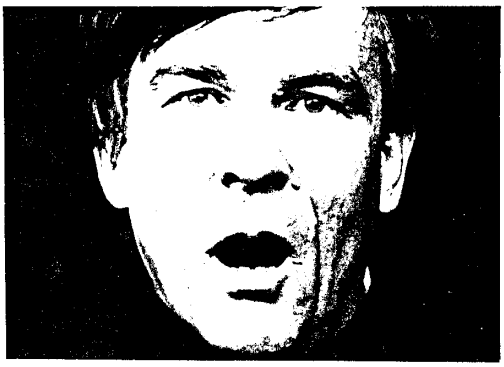




13(a): Articulation of /t/ (front)

13(b): Articulation of /t/ (side)

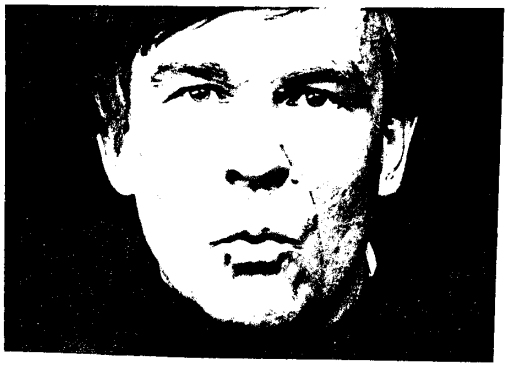
Figure 13: Front and side views of author's /t/ articulation



14(a): Articulation of /a/ (front)

13(b): Articulation of /a/ (side)

Figure 14: Front and side views of author's /a/ articulation



15(a): Articulation of /u/ (front)

15(b): Articulation of /u/ (side)

Figure 15: Front and side views of author's /u/ articulation