Automated Lip-Sync: Background and Techniques

John Lewis*
Computer Graphics Laboratory
New York Institute of Technology

SUMMARY

The problem of creating mouth animation synchronized to recorded speech is discussed. Review of a model of speech sound generation indicates that the automatic derivation of mouth movement from a speech soundtrack is a tractable problem. Several automatic lip-sync techniques are compared, and one method is described in detail. In this method a common speech synthesis method, linear prediction, is adapted to provide simple and accurate phoneme recognition. The recognized phonemes are associated with mouth positions to provide keyframes for computer animation of speech. Experience with this technique indicates that automatic lip-sync can produce useful results.

KEY WORDS: facial animation speech synchronization

INTRODUCTION

Movement of the lips and tongue during speech is an important component of facial animation. Mouth movement during speech is ongoing and relatively rapid, and the movement encompasses a number of visually distinct positions. The movement also must be synchronized to the speech.

Adequate performance on this *lip-sync* problem is not well defined. For example, how accurate must the mouth movement and timing be in order to be satisfying, and how accurate must it be to pass a reality test? While most people cannot read lips (i.e., identify speech from the mouth movement alone [1]), viewers do have a passive notion of correct mouth movement during speech—we know good and bad lip-sync when we see it.

The lip-sync problem has traditionally been handled in several ways. In animations where realistic movement is desired, mouth motion and general character movement may both be obtained by rotoscoping [2]. In this technique, live-action footage of actors performing the desired motion is obtained, and the frames of this footage provide a guide for the corresponding frames of an animation.

A second approach, commonly used in cartoons, is to adopt a canonical mapping from a subset of speech sounds onto corresponding mouth positions. Animation handbooks often have tables illustrating the mouth positions corresponding to a small number of key sounds [3]. The animator must approximately segment the soundtrack into these key sounds. For example, the word "happy" might be segmented as a sequence of two vowels, "aah" and "ee". This approach often neglects non-vowel sounds because vowel sounds correspond to visually distinctive mouth positions and are typically of greater duration than non-vowel sounds. The lip-sync produced using this approach is often satisfactory but is generally not realistic.

Several viable computer face models have been developed, including [4,5,6,7]. Ideally, we might like to control these face models with a high-level animation script, and have an intelligent front end to the face model automatically translate the script into an appropriate sequence of facial expressions and movements. This paper considers the more limited problem of automatically obtaining mouth movement from a recorded soundtrack.

In the following section we describe speech production and the reasons why automatic lip-sync is feasible. Subsequent sections review several approaches to automatic lip-sync. The paper concludes with a discussion of the important but poorly defined problem of matching the realism (or lack of realism) of the facial model with that of the lip-sync motion and speech sounds.

SOURCE-FILTER SPEECH MODEL

Several excellent textbooks on speech principles are available [8,9]. Some relevant points will be mentioned here.

Fig. 1 shows the envelope of the waveform of the phrase "Come quietly or there will be...trouble". It is difficult to visually segment the waveform into words. For ex-

*zilla@computer.org

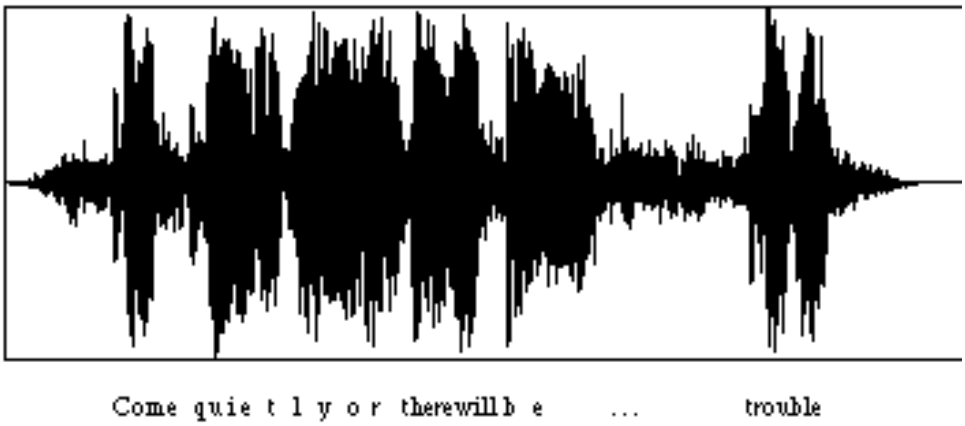

Figure 1: Annotated waveform envelope for the phrase “Come quietly or there will be...trouble”.

ample, there is a gap following the “t” in “quietly”, but there is no gap between the “ly” of “quietly” and the following “or”.

Speech sound generation may be modeled as a broadband sound source passed through a filter. The sound source is vibrations of the vocal cords in the case of voiced sounds and air turbulence in the case of whispered sounds. In the case of voiced sounds the vocal cords in the larynx collide periodically producing a pitched sound with a slowly decaying spectrum of harmonics (Fig. 3a).

Sound produced in the larynx passes through the vocal tract, which consists of the throat, mouth, tongue, lips, and optionally the nasal cavity. The effect of the vocal tract is to filter the sound, introducing resonances (peaks) in the spectrum called formants. Vowel sounds can be characterized by the frequencies of the first two formants [10,9]. The locations of the formants are varied by moving the jaw, tongue, and lips to change the shape of the vocal tract. Formants appear as dark bands in a speech spectrogram plot (Fig. 2). The formant trajectories curve slowly during vowels and change rapidly or disappear in consonants and vowel/consonant transitions.

This source-filter description of speech sound generation is diagrammed in Fig. 3. The plots in this figure are energy spectra, with frequency increasing from zero at the left of each plot. Fig. 3a (source) shows the harmonics of the periodic, roughly triangular pulse produced by the vocal cords. Fig. 3b (filter) shows a vocal tract filter transfer function containing two formants. Fig. 3c (output) shows the spectrum of the resulting speech. The formants are superimposed on the harmonic spectrum of the vocal cords. Note that the formant peak frequencies are independent of the harmonic frequencies.

An important feature of the source-filter model is that it separates intonation from phonetic information. Intonation characteristics, including pitch, amplitude, and the voiced/whispered quality, are features of the sound source, while vocal tract filtering determines the phoneme (“phoneme” is being used somewhat loosely as a term for an “atomic perceptual unit of speech sound”). Human speech production and perception likewise separate intonation from phonetic information. This can be demonstrated by sounding a fixed vowel while varying the pitch or voiced/whispered quality, or conversely by maintaining a constant pitch while sounding different vowels: the mouth position and vowel are both entirely independent of pitch. It should be emphasized that there are various qualifications and details of the source-filter model which are not described here, however, these qualifications do not invalidate the separation of intonation from phonetic information.

In order for automatic lip-sync to be feasible, the position of the lips and tongue must be related in some identifiable way to characteristics of the speech sound. The source-filter model indicates that the lip and tongue positions are functions of the phoneme and are independent of intonation characteristics of the speech sound [9]. A procedure which results in a representation of speech as a timed sequence of phonemes (*phonetic script*) is therefore a suitable starting point for an automated lip-sync approach.

AUTOMATED LIP-SYNC TECHNIQUES

Loudness is jaw rotation

The naive approach to automatic lip-sync is to open the mouth in proportion to the loudness of the sound. It is

Figure 2: A smoothed speech spectrogram (pitch harmonics have been removed). The plot shows energy at frequencies from zero (bottom) to 5000 Hz. and time from zero (at left) to one second. The three primary vowel formants are visible as dark bands.

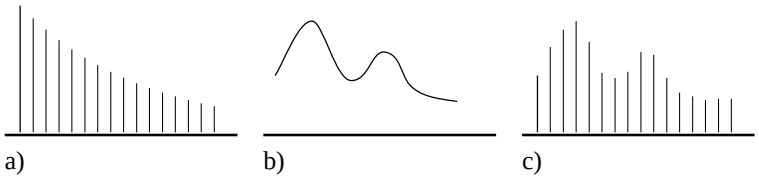

Figure 3: Diagram of the source-filter speech generation model in the frequency domain: a) The vocal cords generate a periodic sound with many harmonics. b) The vocal tract acts as a filter, introducing resonances in the spectrum. c) The resulting speech sound has the resonant peaks (formants) superimposed on the harmonic spectrum generated by the vocal cords.

evident that this is a poor approach: a nasal "m" can be loud although the mouth is closed. Also, the mouth assumes a variety of visually distinct positions during speech; it does not simply open and close. Facial animation produced using this approach has a robotic quality.

Spectrum matching

A more sophisticated approach is to pass the speech signal through a bank of filters, sample the spectra output by the filters at the desired animation frame rate, and then compare these spectra to the spectra of a set of reference sounds (using a least squares match for example). This approach was used in the Transmission of Presence low bandwidth teleconferencing experiments at MIT in the early 1980s [11,12].

This approach can produce acceptable lip-sync, but it is not accurate enough to produce fully realistic lip motion. One problem is that the formant frequencies are quantized to the available filter frequencies. A more significant difficulty with this approach is that the spectrum describes both the vocal tract formants and pitch (in the case of voiced speech), whereas the lip and tongue positions are related only to the formants and are independent of pitch. The pitch in natural voiced speech varies throughout an utterance, so it is unlikely that the pitch of a particular portion of an utterance will match the pitch of the reference sounds. This mismatch degrades the accuracy of the reference sound matching.

Pitch contamination can be reduced by designing the filter bank to smooth the pitch harmonics. There is a tradeoff, however, between smoothing the spectrum and accurately localizing the formant peaks. The best results are obtainable if the filter bank approach is extended to a N-point Fourier transform, where N is sufficient to resolve the pitch harmonics (e.g. two frequency samples per 100 Hz.). The magnitude of this high resolution transform can then be smoothed with a more sophisticated technique such as a smoothing spline.

Speech synthesis

A different approach to the lip-sync problem involves using computer synthesized speech rather than starting from recorded speech. In this approach a phonetic script is either specified directly by the animator or is generated by a text-to-phoneme synthesizer. The phonetic script drives a phoneme-to-speech synthesizer, and it is also read to generate lip motion, resulting in lip-synchronized speech.

This approach has been used successfully in several facial animation systems [13,14,15,6]. An advantage of this approach is that it generates accurate lip-sync, since the speech and the lip motion are both specified by the same script. It is also appropriate when the desired speech is specified textually rather than as a recording, or when the speech content is informative and intonation is a secondary consideration (as is the case in a computerized voice information system).

A drawback of this approach is that it is difficult to achieve natural rhythm and articulation using synthetic speech. Current speech synthesis algorithms produces speech having a slightly robotic quality, while some older systems produce speech which is sometimes unintelligible. Typically the intonation can be improved by adding information such as pitch and loudness indicators to the text or by refining the phonetic script. This requires some additional work, although it is less work than would be required to animate the mouth directly.

LINEAR PREDICTION APPROACH TO LIP-SYNC

Reference [16] described a lip-sync approach based on linear prediction, which is a special case of Wiener filtering [17]. In this approach speech is effectively deconvolved into sound source and vocal tract filtering components. The filtering component is the phonetic script required for lip-sync; no further processing is required to remove pitch harmonics. The algorithm is efficient and maps well onto available matrix algorithms and hardware. This section will describe the linear prediction lip-sync algorithm and several implementation considerations.

Linear prediction speech model

Linear prediction [18] models a speech signal s_t as a broadband excitation signal αx_t input to a linear autoregressive filter (a weighted sum of the input and past output of the filter):

$$s_t = \alpha x_t + \sum_{k=1}^{P} a_k s_{t-k} \qquad (1)$$

This is one realization of the source-filter model of speech production described previously.

The excitation signal x_t is approximated as either a pulse train, resulting in pitched vowel sounds, or an uncorrelated noise, resulting in either consonants or whispered vowels depending on the filter. The filter coefficients a_k vary over time but are constant during a short

interval (analysis frame) in which the vocal tract shape is assumed constant. The analysis frame time should be fast enough to track perceptible speech events but somewhat longer than the voice pitch period to permit deconvolution of the pitch information. An analysis frame time of about 15-20 milliseconds satisfies these conditions. This corresponds to 50-65 frames/second, suggesting that sampling the mouth movement at a standard animation rate (24 or 30 frames/second) may not be fast enough for some speech events (c.f. Fig. 2).

For the purpose of lip-synchronized animation it is convenient to choose the analysis frame rate as twice the film or video frame playback rate. In this case the speech analysis frames can be reduced to the desired animation frame rate with a simple low-pass filter. An alternative is to generate the animation at the higher frame rate (e.g. 60 frames/second) and apply the filter across frames in the generated animation rather than across analysis frames. This supersampling approach reduces the temporal aliasing resulting from quantizing mouth movement keyframes to the animation frame rate, which has been a source of difficulty in previous work [19,14].

Algorithm

Given a frame of digitized speech, the coefficients a_k are determined by minimizing the squared error between the actual and predicted speech over some number of samples. There are a number of formulations of least-squares linear prediction; a simple derivation which results in the autocorrelation method [18] of linear prediction is given here. This derivation views the speech signal as a random process which has stationary statistics over the analysis frame time. The expected squared estimation error

$$E = \mathbf{E}\left\{ s_t - \left[\alpha x_t + \sum_{k=1}^{P} a_k s_{t-k} \right] \right\}^2 \qquad (2)$$

is minimized by setting

$$\frac{\partial E}{\partial a_k} = 0$$

(one proof that this does determine a minimum involves rewriting (2) as a quadratic form), obtaining

$$\mathbf{E}\left\{ s_t s_{t-j} - (\alpha x_t s_{t-j} + \sum_{k=1}^{P} a_k s_{t-k} s_{t-j}) \right\} = 0$$

for $1 \le j \le P$. Since the excitation at time t is uncorrelated with the previous speech signal, the expectation of the product $\alpha x_t s_{t-j}$ is zero. Also, the expectation of terms $s_{t-j}s_{t-k}$ is the $(j-k)$th value of the autocorrelation function. These substitutions result in a system

$$\sum_{k=1}^{P} a_k R(j-k) = R(j) \qquad (3)$$

(in matrix form)

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(P-1) \\ R(1) & R(0) & \cdots & R(P-2) \\ \cdots & \cdots & \cdots & \cdots \\ R(P-1) & R(P-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \cdots \\ R(P) \end{bmatrix}$$

which can be solved for a_k given the analysis frame autocorrelation function R. The latter can be estimated directly from the speech signal using [8]

$$R(\tau) \approx \frac{1}{L} \sum_{t=0}^{L-\tau-1} s_t s_{t+\tau} \quad \text{for } 0 \le \tau \le P$$

where L is the length of the analysis frame in samples. Since the autocorrelation of a stationary process is an even function, $R(j-k)$ is a symmetric Toeplitz matrix (having equal elements along the diagonals). This permits the use of efficient inversion algorithms such as the Levinson recursion [20].

There are a number of other formulations of linear prediction, and the choice of a particular approach depends largely on one's mathematical preferences. The references [8,9] provide speech-oriented overviews of the autocorrelation and another (covariance) formulation, while [18] is an exhaustive (and interesting) treatment of the subject. Many solution algorithms for (3) have also been published. A Fortran implementation of the Levinson algorithm is given in [18] and a version of this routine (**auto**) is included in the IEEE Signal Processing Library [21]. The most efficient solution is obtained with the Durbin algorithm, which makes use of the fact that the right-hand vector in (3) is composed of the same data as the matrix. This algorithm is described in [8] and is presented as a Pascal algorithm in [9]. Alternatively, (3) can be solved by a standard symmetric or general matrix inversion routine at some extra computational cost.

Synchronized speech

The coefficients a_k resulting from the linear prediction analysis describe the short term speech spectrum with the pitch information convolved out.

Analyzed speech is converted to a phonetic script by classifying each speech frame according to the minimum Euclidean distance of its short-term spectrum from

the spectra of a set of reference phonemes. The spectrum is obtained by evaluating the magnitude of

$$H(z) = \frac{\alpha}{1 - \sum_{k=1}^{P} a_k z^{-k}} \qquad (4)$$

(the z-transform of (1)) at N points on the complex z-plane half unit circle with $z = e^{-j\pi k/N}$. In this case the denominator in (4) is effectively a discrete Fourier transform of the negated, zero-extended coefficient sequence $1, -a_1, -a_2, \ldots, -a_P, 0, 0, \ldots$, of length $2N$, permitting implementation by FFT. A resolution of $N = 32$ appears to be sufficient since the linear prediction spectra are smooth. Although a more direct identification approach would be to compare the coefficients a_k to the coefficients of the reference phonemes, least-squares matching on the coefficients performs poorly and it appears that some other norm is required [18].

The selection of the reference phonemes involves a compromise between robust identification and phonetic and visual resolution. Various 'How to Read Lips' books and books on animation [1,3] describe visually distinctive mouth positions and the corresponding sounds (Fig. 4). Previous synchronized speech animation has typically used approximately 10-15 distinct mouth keyframes [11,12,5] (although synthetic speech approaches [15,14] have used many more distinct mouth positions). Our current reference phoneme set consists of the vowels in the words *hate,hat,hot,heed,head,hit,hoe,hug,hoot* (as pronounced in American English), together with the consonants *m,s,f*.

While there are more than thirty phonemes in spoken English [10] (not counting combination sounds such as diphthongs) this reference set includes most of the vowels. Our approach to lip-sync profits from the fact that vowels are easily identified with a linear prediction speech model, since visually distinctive mouth positions correspond to vowels in most cases (Fig. 4), and consonants are also generally shorter than vowels. Also, it is not necessary to have a distinct mouth position for each phoneme, since some consonants such as *d,t* and *f,v* are distinguished by voicing rather than by lip or tongue position. In fact, only a few key sounds and mouth positions are required to represent consonant sounds—the consonants *g,k,s,t* have fairly similar spectra and mouth positions, as do *m,n* (the mouth is closed for *m* and only slightly open for *n*).

We have found that very accurate vowel identification is possible using the linear prediction identification approach with twelve reference phonemes. Currently we are using a 20kHz audio sampling rate with $P = 24$ in (1). The number of coefficients was chosen using the rule of thumb [18] of one pole (conjugate zero pair of the denominator polynomial of (4)) per kHz, plus several extra coefficients to model the overall spectrum shape. Almost all of the semantically important information in speech lies below $4000 - 5000$ Hz, as demonstrated by the intelligibility of telephones, so an audio sample rate of 10kHz is sufficient for analysis applications such as lip-sync. The higher sample rate allows the speech data to be manipulated and resynthesized for a reasonably high quality sound track.

Consonant transitions are an area of theoretical difficulty. In some cases, for example in pronouncing a stop consonant such as "t" at the end of a word, the mouth can remain open following aspiration during a period of silence leading into the next word. Any purely acoustically based lip-sync technique will incorrectly cause the mouth to be closed during this period.

Fig. 5 shows the raw output of the linear prediction lip-sync procedure applied to a phrase which begins "Greetings media consumers..." The columns are (from left to right) the time, the excitation volume, a voiced/unvoiced indicator, the best reference phoneme match (in the starred column) and its associated error, and the second best match and its error. This example is also annotated with the corresponding speech in the right hand column. From the annotation it can be seen that vowels are plausibly identified while consonants are mapped onto other consonants. For example, the "t" sound in the word "greetings" is matched with the "s" reference sound (labeled *es*). The "e" sound in "greetings" is matched with the vowel in the word *hit* rather than with the vowel in *heed* due to pronunciation. The reference sound *eeng* is a variation of the vowel sound in the word *heed*.

Parametric face model

We used the parametric human face model developed by Parke [19,4] in our lip-sync experiments. This model has been extended to several full-head versions by DiPaola and McDermott [22]. The parametric modeling approach allows the face to be directly and intuitively manipulated with a limited and fairly natural set of parameters, bypassing the effort involved in modeling or digitizing keyframes in a keyframe-based approach.

The face model parameters relevant to mouth positioning and lip-sync include those controlling jaw rotation, mouth opening at several points, the lower lip 'tuck' for the f/v sound, and movement of the corners of the mouth. Since the parametric model allows expressive parameters to be manipulated and animated independently of geometric features, an animation script including lip-sync and other expressive parameters can be applied to any available character (geometric database).

Figure 4: Portion of a lip reading chart. Top row, from left to right, the vowels in the words *hat,hot* and the *f/v* sound. Bottom row: the vowels in the words *head,hit,hoot*.

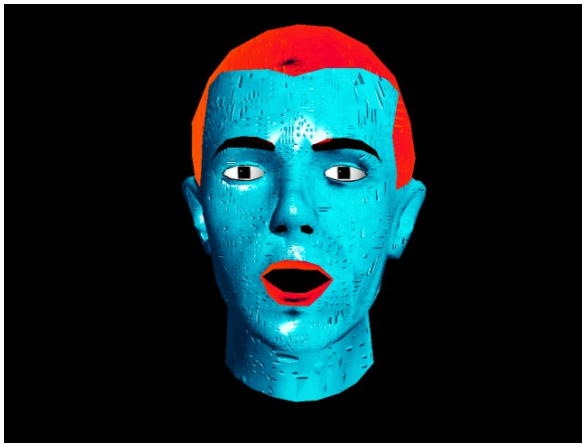

Figure 6: Computer face model positioned for the vowel in the word “hot”.

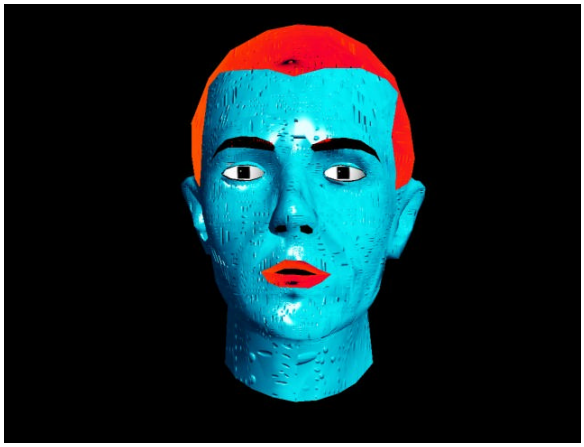

Figure 7: Computer face model positioned for the vowel in the word “hoot”.

```
                          ***
(3.33 gain 0.007 err 0.249 sil1 1.595 sil2 1.901)       ; [silence]
(3.37 gain 0.009 err 0.366 sil2 2.472 sil1 2.965)       ;
(3.40 gain 0.146 err 0.416 em4 5.429 eeng 5.907)        ; GR
(3.43 gain 0.216 err 0.545 hit1 5.985 hit3 6.837)       ; E
(3.47 gain 0.159 err 0.545 hit4 0.000 hit2 4.732)       ; E
(3.50 gain 0.208 err 0.521 hit4 2.914 hit2 4.672)       ;
(3.53 gain 0.053 err 0.585 es2 3.804 sil2 3.872)        ; T
(3.57 gain 0.117 err 0.574 es2 3.854 es1 3.883)         ;
(3.60 gain 0.358 err 0.588 heed2 3.874 heed1 4.995)     ; I
(3.63 gain 0.191 err 0.425 heed2 5.597 es3 5.688)       ;
(3.67 gain 0.244 err 0.475 heed2 5.324 heed3 5.619)     ;
(3.70 gain 0.121 err 0.605 eeng 3.749 eeng 3.749)       ; NG
(3.73 gain 0.066 err 0.401 eeng 4.784 eeng 4.784)       ;
(3.77 gain 0.051 err 0.393 eeng 4.089 eeng 4.089)       ;
(3.80 gain 0.076 err 0.787 em4 4.281 eeng 4.678)        ; [error]
(3.83 gain 0.067 err 0.688 es3 2.991 es2 3.039)         ; S
(3.87 gain 0.065 err 0.515 es2 2.169 es3 3.629)         ;
(3.90 gain 0.007 err 0.253 sil2 1.684 sil1 1.792)       ; [silence]
(3.93 gain 0.027 err 0.488 em2 0. em4 3.829)            ; M
(3.97 gain 0.037 err 0.401 em2 2.629 em4 4.487)         ;
(4.00 gain 0.202 err 0.565 hit4 5.595 heed3 6.360)      ; E
(4.03 gain 0.225 err 0.558 es1 4.623 heed3 5.123)       ; D
(4.07 gain 0.130 err 0.380 es1 6.324 es2 6.911)         ;
(4.10 gain 0.075 err 0.416 es1 5.586 heed3 5.694)       ;
(4.13 gain 0.189 err 0.405 es1 4.732 hit4 5.325)        ;
(4.17 gain 0.211 err 0.463 hit4 4.345 heed3 5.575)      ; I
(4.20 gain 0.250 err 0.669 hit4 5.735 es1 5.917)        ;
(4.23 gain 0.259 err 0.654 head3 6.151 head1 6.203)     ; A
(4.27 gain 0.257 err 0.691 head3 5.967 head1 6.280)     ;
(4.30 gain 0.055 err 0.632 her2 3.671 her1 3.911)       ;
(4.33 gain 0.012 err 0.403 sil2 6.010 sil1 6.019)       ; [silence]
```

Figure 5: Annotated output of the linear prediction lip-sync procedure for the words “Greetings media...”.

Fig. 6 and Fig. 7 show one such character positioned for the vowels in the words *hot* and *hoot*).

Although the tongue motion can be automatically derived from a phonetic script in the same manner as the lips, we are not using this capability since the Parke face model does not currently include a tongue.

Parameter smoothing

The mouth can move rapidly in vowel/consonant transitions, but vowel/vowel transitions are generally smooth (as can be seen from the formant trajectories in Fig. 2). Automated lip-sync in effect performs a vector quantization from a high-dimensional acoustic space onto a one-dimensional, discrete space of phonemes. This quantization results in abrupt transitions between phonemes. It is therefore necessary to smooth the mouth motion somehow.

Since the phoneme space is discrete it is not possible to smooth the phoneme sequence directly. The approach we have used to date is to convert the phonetic script into a set of parameter tracks for the face model, and then smooth these tracks. A fairly sophisticated smoothing technique is needed. A finite impulse response filter did not provide suitable smoothing, since it blurred rapid vowel/consonant transitions and attenuated extremes of the parameter movement. A smoothing spline [23] is currently implemented and provides somewhat better results. Examination of formant trajectories suggests the need for a smoothing technique that preserves large discontinuities.

Linear prediction speech resynthesis

The linear prediction software, once implemented, can also be used to resynthesize the original speech. This enables several manipulations which may be useful for animation. In the most faithful synthesis approach, the difference signal (residual) between the original speech and the output of the linear prediction filter is used as the synthesis excitation signal:

$$x_t = s_t - \sum_{k=1}^{P} a_k s_{t-k}$$

The residual signal approximates an uncorrelated noise for consonants and whispered vowels, and approximates a pulse train for voiced vowels. The linear prediction analysis and the residual together encode most of the information in the original speech. The synthesized speech is highly intelligible and retains the original inflection and rhythm, yet it has a subtle synthetic quality which may be appropriate for computer animation. Variations of this form of synthesis are commonly used for speech compression and the reader has no doubt heard examples of it produced by dedicated linear prediction chips.

Vocoder quality or 'robot' speech is obtained if the excitation signal is a synthetically generated signal, which may be either a pulse train or a random sequence. The Levinson and Durbin algorithms return a per-frame prediction error magnitude which is compared with a threshold to determine which form of excitation to use; normalized errors greater than about 0.3 typically reflect consonants or whispered voice. An important manipulation which is easily possible in the case of synthetic excitation is to speed up or slow down the speech. This is accomplished simply by accessing the linear prediction analysis frames at a faster or slower rate. Since the voice pitch is controlled by the excitation, the speech rate can be changed without producing a ("Mickey Mouse") effect. The linear prediction software has been implemented under a general purpose Lisp-based computer music system [24], so additional sonic manipulations such as reverberation, gender/age change (spectrum shifting), etc. are directly obtainable.

EVALUATION

The linear prediction lip-sync approach described in the previous section produces mouth motion which is tightly synchronized to the speech. The quality of the lip-sync falls short of full realism, but it has been characterized as being better than the lip-sync obtained with the 'lazy rotoscoping' approach employed in [19], in which film footage guided the creation of mouth keyframes every few frames [25]. An animator trained in traditional animation techniques characterized the linear prediction lip-sync method as producing "too much data". This characterization is consistent with the recommendations of animation handbooks, which generally suggest that only lengthy stressed syllables be animated.

Gestalt and specificity

The animator who uses a computer face model faces a strong but poorly defined perceptual phenomenon. Fig. 8 is an attempt to elucidate this phenomenon. This drawing is easily recognized as a face, and we can even infer some "character", despite the fact that the drawing specifies far less (geometric) information than existing computer face models. Information which is clearly omitted from this figure is perceptually ignored or completed. In contrast, while three-dimensional shaded ren-

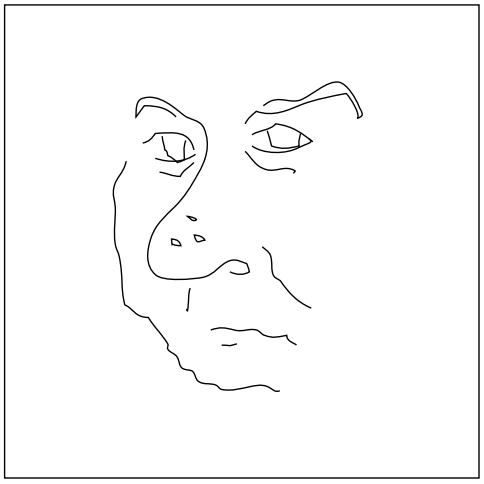

Figure 8: This face sketch specifies much less geometric information than a computer face model.

derings of objects such as cars are often extremely realistic, comparable renderings of computer face models often appear mechanical. It seems that as the face model becomes more detailed and specific, any inaccuracies in the specified information become perceptually prominent.

One view of this problem is that it results from the fact that computer models generally specify unknown information. For example, a set of vertices or control points in a geometric model may be the only "known" detail, and a surface constructed using these points may be one of many plausible surfaces. A shaded rendering of the model can realize only one of these surfaces, however. In the case of a computer face model, the surface interpolation required for computer rendering asserts that the face is quite smooth, whereas the rendering in Fig. 8 does not rule out the possibility of skin imperfections at unspecified locations.

This phenomenon may also affect the use of automated lip-sync in computerized character animation. Lip-sync motion derived from a recorded soundtrack is quite specific but not fully realistic. We can speculate on whether the animation might be more successful if the motion were to be filtered or subsampled to make it less detailed, thereby reducing our perceptual expectations.

Similar considerations can be applied to the soundtrack. The animator should consider whether viewers would be more likely to accept a slightly mechanical face if the speech were also slightly mechanical, as is the case with lip-sync approaches using synthetic speech. If so, recorded speech may be resynthesized by linear prediction in order to achieve a slight synthetic quality while preserving intelligibility and intonation. On the other hand, the successful use of real voices in traditional animation would seem to invalidate a principle that the realism of the soundtrack should match that of the images.

While the preceding comments are philosophical rather than scientific, the successful application of facial animation will require an understanding of these and similar issues [26].

Future directions

Facial animation generated using automated lip-sync looks unnatural if the head and eyes are not also moving. Although head movement during speech is probably quite ideosyncratic, it would seem possible to generate stereotypical head and eye movement automatically from the soundtrack. This would further reduce the animator's work load, and it would enable automated "talking head" presentations of audio monologues [12].

We have not explored possible variations in lip movement for a given utterance. While correct pronunciation considerably constrains possible deviations from 'standard' lip movement, one obvious effect is that increased volume often corresponds to greater mouth opening. The possible effect of emotional expression on mouth movement during speech also has not been considered. This may be an important effect, since mouth position is one of the primary indicators of emotion. A related problem would be to attempt to derive emotional state directly from the speech soundtrack.

ACKNOWLEDGEMENTS

Sean Curran provided an evaluation of the automatic lip-sync technique from an animator's viewpoint.

References

[1] E. Walther, *Lipreading*, Nelson-Hall, Chicago, 1982.

[2] T. McGovern, *The Use of Live-Action Footage as a Tool for the Animator*, SIGGRAPH 87 Tutorial Notes on 3-D Character Animation by Computer, ACM, New York, 1987.

[3] P. Blair, *Animation: Learn How to Draw Animated Cartoons*, Foster, Laguna Beach, California, 1949.

[4] F. Parke, 'Parameterized models for facial animation', *IEEE Computer Graphics and Applications*, 2, (9), 61-68 (Nov. 1982).

[5] P. Bergeron and P. Lachapelle, *Controlling facial expressions and body movements in the computer-generated animated short "Tony de Peltrie"*, SIGGRAPH 85 Tutorial Notes, ACM, New York, 1985.

[6] N. Magnenat-Thalmann and D. Thalmann, *Synthetic Actors in Computer Generated Three-Dimensional Films*, Springer Verlag, Tokyo, 1990.

[7] K. Waters, 'A muscle model for animating three-dimensional facial expression', *Computer Graphics*, 21, (4), 17-24 (July 1987).

[8] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, N.J., 1979.

[9] I. Witten, *Principles of Computer Speech*, Academic Press, London, 1982.

[10] J. Flanagan, *Speech Analysis, Synthesis, and Perception*, Springer-Verlag, New York, 1965.

[11] P. Weil, *About Face: Computergraphic Synthesis and Manipulation of Facial Imagery*, M.S. Thesis, Massachusetts Institute of Technology, 1982.

[12] J. Lewis and P. Purcell, 'Soft Machine: a personable interface' In *Proceedings of Graphics Interface 84*, Ottawa, 223-226 (May 1984).

[13] A. Pearce, B. Wyvill, G. Wyvill and D. Hill, 'Speech and expression: a computer solution to face animation', *Proceedings of Graphics Interface 86*, 136-140 (1986).

[14] D. Hill, A. Pearce and B. Wyvill, 'Animating speech: an automated approach using speech synthesis by rules', *The Visual Computer*, 3, 277-289 (1988).

[15] N. Magnenat-Thalmann, E. Primeau and D. Thalmann, 'Abstract muscle action procedures for human face animation', *The Visual Computer*, 3, 290-297 (1988).

[16] J. Lewis and F. Parke, 'Automated lip-synch and speech synthesis for character animation', In *Proceedings of CHI87*, ACM, New York, 143-147 (Toronto, 1987).

[17] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Wiley, New York, 1949.

[18] J. Markel and A. Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[19] F. Parke, *A Parametric Model for Human Faces*, Ph.D. Dissertation, U. of Utah, 1974.

[20] N. Levinson, 'The Wiener RMS (root mean square) error criterion in filter design and prediction', *Journal of Mathematical Physics*, 25, 261-278 (1947).

[21] *Programs for Digital Signal Processing*, IEEE Press, 1979.

[22] S. DiPaola, *Implementation and Use of a 3d Parameterized Facial Modeling and Animation System*, SIGGRAPH 89 Course Notes on State of the Art in Facial Animation, ACM, New York, 1989.

[23] C. de Boor, *A Practical Guide to Splines*, Springer Verlag, New York, 1978.

[24] J. Lewis, *LispScore Manual, Squonk Manual*, NYIT internal documentation, 1984,1986.

[25] F. Parke, *Personal communication*.

[26] B. Kroyer, *Critical reality in computer animation*, SIGGRAPH 87 Tutorial Notes on 3-D Character Animation by Computer, ACM, New York, 1987.