Automated Lip-Synch and Speech Synthesis for Character Animation

J.P. Lewis and F.I. Parke

Computer Graphics Laboratory
New York Institute of Technology

ABSTRACT

An automated method of synchronizing facial animation to recorded speech is described. In this method, a common speech synthesis method (linear prediction) is adapted to provide simple and accurate phoneme recognition. The recognized phonemes are then associated with mouth positions to provide keyframes for computer animation of speech using a parametric model of the human face.

The linear prediction software, once implemented, can also be used for speech resynthesis. The synthesis retains intelligibility and natural speech rhythm while achieving a "synthetic realism" consistent with computer animation. Speech synthesis also enables certain useful manipulations for the purpose of computer character animation.

KEYWORDS: facial animation, speech synthesis.

RESUME

Cette présentation décrit une méthode automatique pour la synchronisation d'animation faciales avec un texte pré-enregistré. Dans cette approche, on a adapté une méthode courante en synthèse de la parole (prédiction linéaire) à la reconnaissance des phonèmes de facon simple et précise. Les phonèmes identifiés sont ensuite associés à des positions de bouche qui seront elles même utilisées comme images-clefs dans le processus d'animation par ordinateur d'un modèle de visage humain paramétré.

Le logiciel de prédiction lineaire, une fois implementé, peut aussi être utilisé pour la resynthèse de la parole. La synthèse conserve l'intelligibilité, et le rythme "naturel" du langage ainsi qu'un réalisme synthétique adapté aux techniques d'animation par ordinateur. La synthèse de la parole permet par ailleur des manipulations efficaces dans le contexte de l'animation de personnages par ordinateur.

Introduction

While several viable computer face models have been developed [9][8][2], animations employing these models to date have relied in part on non-automated techniques such as rotoscoping. Ideally, an "artificially expressive" front end to the computer face model would intelligently translate an animation script into a sequence of facial movements and expressions which enact the script, while giving the animator a conceptual model with which to control the acting process.

This paper considers a more limited problem: a technique is described for automatically identifying mouth positions corresponding to a given speech segment ('lip-synch'). The approach is to obtain a representation of the speech as a timed phoneme sequence (phonetic script), and then to establish a phoneme to mouth position correspondence in order to drive a parametric face model.

The facial animation system described by Pearce et. al. [10] includes one approach to synchronized speech. In their approach, the phonetic script is specified directly by the animator. The phonetic script is also input to a phoneme-to-speech synthesizer, achieving synchronized speech. This approach is appropriate when the desired speech is specified in a textual rather than auditory form, and the quality of rule-based synthetic speech is acceptable to the purpose. The drawback of this approach is that it is difficult to achieve natural rhythm and articulation when the speech timing and pitch is defined in a script or derived by a rule-based text-to-speech synthesizer. Typically the prosody quality can be improved somewhat by adding information such as pitch and loudness indications to the script.

An alternative approach is to obtain the desired phoneme sequence by analyzing digitized speech. As Pearce et. al. note, current speech recognition technology is somewhat error prone, and in fact most systems are capable only of identifying isolated words. Our problem is considerably simpler than that of recognizing speech, however. Speech recognition involves transforming the speech into a representation in which the speech formant frequencies are emphasized and the pitch information is largely removed, and then parsing this representation to identify words. It is the latter task which is difficult; the former (acoustic preprocessing) step is generally quite

effective and is all that is required for deriving a phonetic script.

The analysis approach used here adopts linear prediction, a parametric speech synthesis model, to obtain speech parameters which can be used to identify phonemes from a limited set corresponding to visually distinctive mouth positions.

Linear Prediction Speech Model

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

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

This is an abstracted but fairly accurate model of speech production, in which the filter models the vocal tract (mouth, tongue, and lip positions) and the excitation signal approximates the acoustic signal produced by the vocal cords. It is also a useful model, since both human speech production and perception likewise separate pitch (determined by the vocal cord tension) from phonetic information (determined by the vocal tract filtering). This can be illustrated by sounding a fixed vowel while varying the pitch or vice versa: the mouth position and vowel are both entirely independent of pitch.

In speech resynthesis 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 msec. 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. 1).

For the purpose of synchronized speech animation it is convenient to choose the analysis frame rate as twice the film or video frame playback rate. In this case the frame rate can be reduced to the desired animation 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 facial movement keyframes to the animation frame rate, which has been a source of difficulty in previous work [9].

Solution 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 [7] 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

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

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

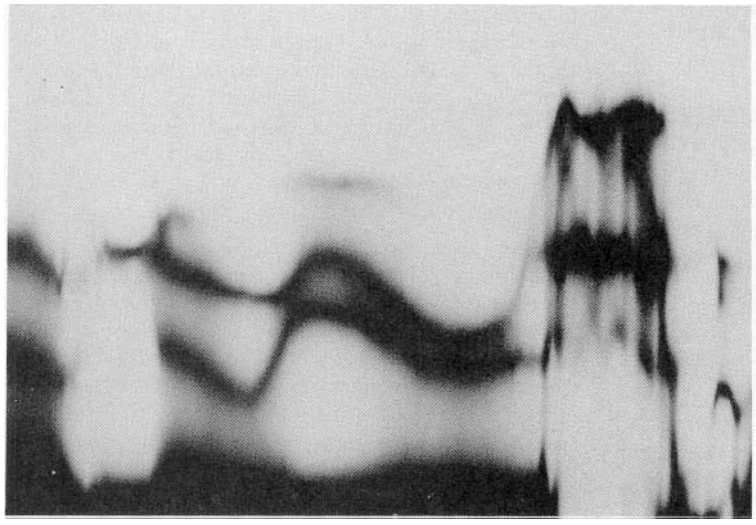

Fig. 1: Speech spectrogram over one second with the pitch information removed. The three primary vowel formants are visible as dark bands. While vowels typically extend over a number of animation frames, transition areas are difficult to track using a standard animation frame rate.

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

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 [11]

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

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), permitting the use of efficient algorithms available for the inversion of these matrices such as the Levinson recursion [4].

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 [11][13] provide speech-oriented overviews of the autocorrelation and another (covariance) formulation, while [7] 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 [7] and a version of this routine (**auto**) is included in the IEEE Signal Processing Library [1]. 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 [11] and is presented as a Pascal algorithm in [13]. Alternatively, (3) can be solved by a standard symmetric or general matrix inversion routine at some extra computational cost. We note in passing that linear prediction is a special case of Wiener filtering, which has other computer graphics applications such as stochastic synthesis [6].

Synchronized Speech

The coefficients a_k resulting from the linear prediction analysis describe the short term speech spectrum with the pitch information convolved out. An analyzed speech frame is classified using the Euclidean distance of its short-term spectrum from the spectra of the 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 e.g. $N=32$ 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 identification on the coefficients performs poorly and it appears that some other norm is required [7].

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 cartooning identify visually distinctive mouth positions and the corresponding sounds (Fig. 2). Previous synchronized speech animation has typically used between approximately 10-15 distinct mouth keyframes [12][5][2]. Our current reference phoneme set con-

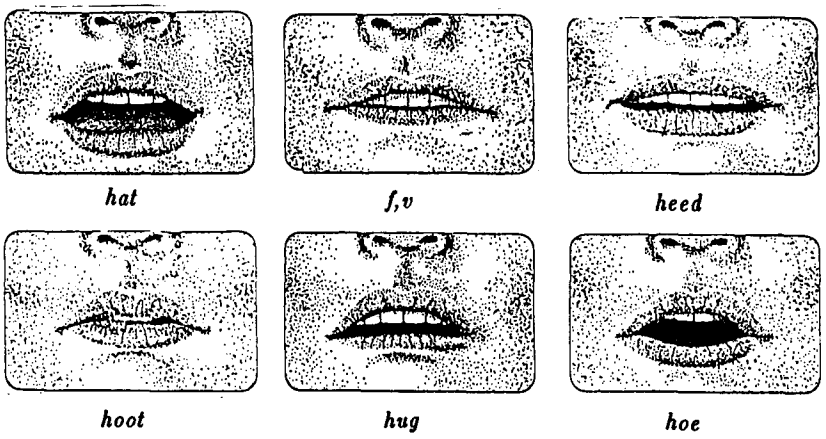

Fig. 2: Portion of a lipsynch chart with vowels indicated by sample words, obtained [12] from a popular 'How to Read Lips' book.

sists 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 [3] (not counting combination sounds such as diphthongs) this reference set includes most of the vowels. Our approach to speech synchronization 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. 2), and consonants are also generally shorter than vowels.

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 [7] 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 AM radio, so an audio sample rate of 10kHz is sufficient for analysis applications such as lipsynch. 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 lipsynch technique will incorrectly cause the mouth to be closed during this period. Another difficulty is the nasal 'm', which presents the inverse situation where the mouth is closed during sound production.

The viability of an automated lipsynch method depends on the nature and purpose of the facial representation. A follow-on to the "Transmission of Presence" low-bandwidth teleconferencing project at the MIT Architecture Machine Group [12][5] included synchronized speech using a simple filter-bank sound categorization method. While this method was able to reliably distinguish only about five sounds, it was sufficient to create tolerable low-resolution facial animation. Informal experiments indicate that people (at least those who do not read lips) do not easily ascertain the sound corresponding to given mouth position (as can be demonstrated by attempting to guess the sounds represented in Fig. 2). We believe that accurate rhythm in the mouth movement is fundamental for lipsynch, while accuracy of mouth positioning becomes necessary in close-up views of the face.

Face Model

The lipsynch system we are developing employs the parametric human face model described in [9][8]. This model has recently been extended to several full-head versions. The parametric modeling approach allows the face to be directly and intuitively manipulated via 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 lipsynch include those controlling jaw rotation, lip opening, raising the upper lip, the lower lip 'tuck' for the f/v sound, and movement of the corners of the mouth. Since the parametric model allows expressive and structural parameters to be manipulated and animated independently, a computed script including lipsynch and other expressive parameters can be applied to any available character employing the model. Fig. 3 shows the face model from the character "T-square", while Figs. 4, 5 show the 'F' and 'hoe' (vowel) positions for this face.

Linear Prediction Speech Synthesis

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 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.

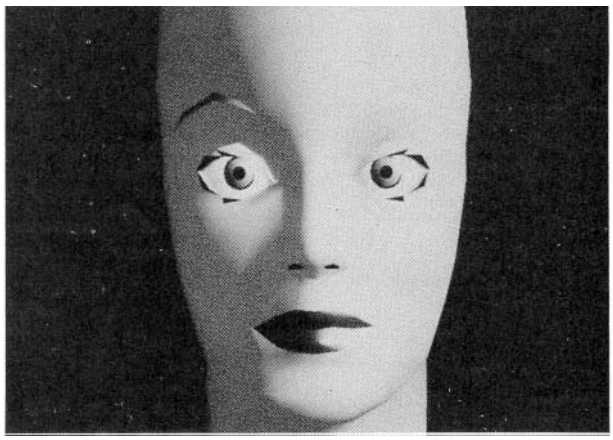

Fig. 3: Face model for the "T-square" character.

Vocoder quality or 'robot' speech is obtained if the excitation signal is a synthetically generated pulse train or 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, e.g. 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 coefficient 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 [14], so additional sonic manipulations such as reverberation, gender/age change (spectrum shifting), etc. are directly obtainable.

Evaluation

Expressive character animation promises to provide the human impact which is generally absent in computer animation. Automated lipsynch and speech manipulation techniques are steps towards this goal. While automated lipsynch is necessarily inferior to rotoscoping, it may be adequate for many purposes, and is probably preferable to the non-rotoscoped manual lipsynch used in cartoon animation. It also provides a first pass at the desired movement in those cases where manual improvement is required.

The computer aided character animator is fighting a strong perceptual effect, however: as the character model becomes more realistic, any remaining flaws become prominent and sometimes even disturbing. Fully expressive semi-automated character animation will require not only lipsynch but also a means of automatically generating characteristic head movements (prosodic nodding, eye movements, expressions, etc.) during speech as well. This problem may be easier than that of speech synchronization in one respect, in that head movement is more likely to be atypical than simply wrong, but in other respects it appears to be quite difficult.

Acknowledgements

This work draws on experience obtained in the MIT Transmission of Presence projects. We would like to thank the reviewers for their helpful comments.

References

1. *Programs for Digital Signal Processing.* IEEE Press, 1979.
2. Bergeron, P., and Lachapelle, P. Controlling facial expressions and body movements in the computer-generated animated short "Tony de Peltrie". *Siggraph Tutorial notes.* 1985.
3. Flanagan, J. *Speech Analysis, Synthesis, and Perception.* Springer-Verlag, New York, 1965.
4. Levinson, N. The Wiener RMS (root mean square) error criterion in filter design and prediction. *J. Math. Phys.* **25**, 1947, 261-278.
5. Lewis, J.P. Methods for stochastic spectral synthesis. *Proceedings, Graphics Interface.* 1986.
6. Lewis, J.P. and Purcell, P. Soft Machine: a personable interface. *Proceedings, Graphics Interface.* May 1984, 223-226.
7. Markel, J. and Gray, A. *Linear Prediction of Speech.* Springer-Verlag, New York, 1976.
8. Parke, F. *A Parametric Model for Human Faces.* PhD. Dissertation, U. of Utah, 1974.
9. Parke, F. Parameterized models for facial animation. *IEEE CG&A.* 9, **2**, Nov. 1982, 61-68.
10. Pearce, A., Wyvill, B., Wyvill, G., Hill, D. Speech and expression: a computer solution to face animation. *Proceedings, Graphics Interface.* 1986, 136-140.
11. Rabiner, L. and Schafer, R. *Digital Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, N.J., 1979.
12. Weil, P. *About Face: Computergraphic Synthesis and Manipulation of Facial Imagery.* M.S. Thesis, Massachusetts Institute of Technology, 1982.
13. Witten, I. *Principles of Computer Speech.* Academic Press, London, 1982.
14. Lewis, J.P. *LispScore Manual, Squonk Manual.* NYIT internal documentation, 1984,86.

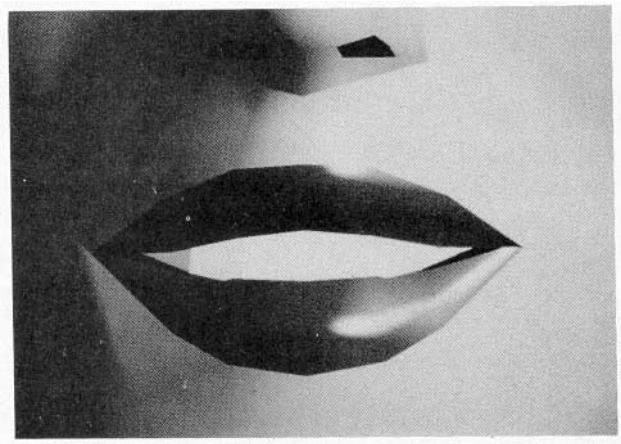

Fig. 4: The face model positioned for the f/v sound. The lower lip is positioned under the teeth (not clearly visible in this view).

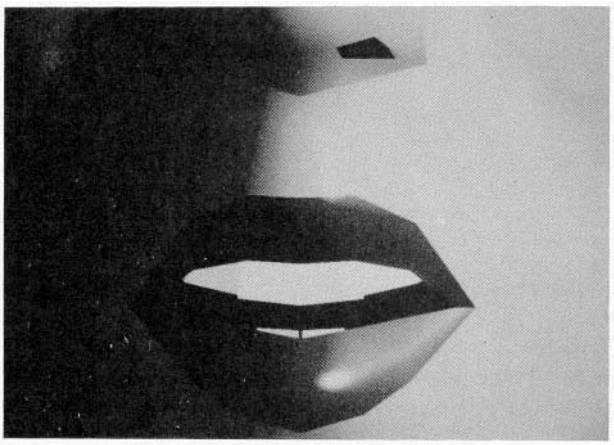

Fig. 5: The face model positioned for the vowel in 'hoe'.