

Predictor Surfaces for Lip Synchronization Animation of Voiced Input

Chandrika Krothapalli, David F. McAllister, Robert D. Rodman

Donald Bitzer, Meng Wang and Jamie Taylor

Voice I/O group

Department of Computer Science

North Carolina State University

Raleigh NC 27606

ABSTRACT

The authors are producing two-dimensional animated human faces with lip motion synchronized to a given speech sound file. The process is called lip synchronization. We assume the only input is the sound file; no text is used to disambiguate mouth shapes. We restrict our discussion to voiced input. (Figures and the Table are at the end of the text.)

Keywords: Lip synchronization, animation, viseme, predictor surface, Delaunay triangulation

1. INTRODUCTION

Lip Synchronization is the process of producing animation of the mouth that is synchronized with speech input. If animation is of a human mouth, an actual portrayal of how the lips, tongue and jaw of the speaker move during the utterance of speech can be deduced from the speech signal [11]. Lip-synching with the help of text is called speech/text directed lip-sync. Such systems segment the text into a set of known *phonemes* for which there is a database of matching *visemes* or mouth shapes [17]. The visemes are played sequentially using the speech signal for synchronization. On the other hand text-free or speech directed lip-sync systems do not attempt to use the phonemic content of speech. They use only the acoustic properties of the recorded sound to identify the mouth position during the speech production. This discussion centers on speech directed lip-synching.

Lip synchronization can be achieved using Hidden Markov Models, Neural Networks, Gaussian mixture models, etc. [8]. We use spectral and shape analysis of the speech signal. Our research is based on the observation that the basic shape of the transform for a given speaker can be reliably correlated with the mouth shape of that sound. The discrete Fourier transform (DFT) is converted to a discrete probability density

function by normalization, and standard statistical shape measures called *moments* are used [13]. Our discussion here will be centered on voiced speech: in particular, vowels, and the voiced fricatives /z/ and /zh/ (middle consonants of *miser* and *measure*).

A pitch tracker is used to determine the period of the glottal pulse (GP). The absolute value of the discrete Fourier transform (ADFT) of each GP is computed. Each ADFT is normalized to have unit area and the mean and variance are found. These values are our measures of the shape of the transform and are combined to produce a 2D curve called a *track*. Figure 1 shows a sound track composed of the first and second central moments of the word 'owie'.

The differences in the mouth shape, size, and voice of different speakers lead to variations in the tracks of different individuals for the same word. To calculate mouth positions associated with a sound track, bivariate surfaces over the moment plane called *predictor surfaces* are constructed. We use four external parameters to define the shape of the mouth (Figure 2): *Jaw* is the distance between the upper and lower teeth; *Edges* is the width of the opening between the upper and lower lip; *Corners* is the distance between the connecting points of the upper and lower lips; and *Flare* is the distance between the bottom of the upper lip and the top of the lower lip at the center of the mouth [12]. These four parameters are sufficient to define the mouth position for most voiced utterances in rapid speech. However there are some sounds whose mouth shape cannot be described by these four parameters alone. For example, tongue plays a major role for the sounds /l/ and /n/.

2. SURFACE GENERATION TECHNIQUES

The four mouth parameters are determined by evaluating two-variable *predictor surfaces* that are computed for a given speaker for each mouth parameter. Several techniques for generating the predictor surfaces have been studied. The training sounds we collect to produce predictor surfaces for a given speaker yield arbitrarily spaced data in the plane. Our goal was to generate a continuous non-negative surface that accurately represented the training data, did not require conversion to a grid or suffered from numerical instabilities. Several interpolation and approximation methods were compared including polynomial and rational least squares approximation, *thin-plate splines* [18], *inverse distance weighting* [15], *Shepards method* [p] and several others. The details can be found in [14]. The *Delaunay triangulation interpolation* (DTI) method [9] was chosen since it meets the criteria described above and most of the others did not. An additional advantage of the method is that it is available in Mathematica, the software we are using for our signal processing and animation frame production.

3. DELAUNAY TRIANGULATION

A triangulation is a subdivision of an area into triangles. There are several ways by which to triangulate any given set of points. One of the most common and useful such triangulation is the Delaunay triangulation [13]. The DT set is a collection of edges satisfying an “empty circumcircle” property. Given a triangle $T(P_i, P_j, P_k)$ belonging to a DT of a set of points P , no other point of P is internal to the circle defined by P_i, P_j, P_k [1]. From all the possible triangulations of a given set of points, DT is the triangulation that gives the largest minimum angle for all the triangular elements [5]. (Other applications are found in [2], [3], and [4].)

Figure 3 (a) shows a set of 20 points generated randomly. Figure 3 (b) represents the DT of the set of points shown in 3 (a). Figure 4 shows the DTI of a set of 120 points generated randomly in 3D.

4. PREDICTOR SURFACES

We use the sounds /aa/, /ee/, /oo/, /ey/, /ae/, /o/, /z/ and /zh/ as in the words ‘calm’, ‘team’, ‘zoom’, ‘name’, ‘jam’, ‘foam’, ‘easy’ and ‘pleasure’ respectively for training the system.

Normalized values of mouth parameters for each of these training sounds are set as shown in the Table 1. These values are calculated using the shape of silent mouth as reference. As shown in the table above, the parameter values for silence

are zero. The training sounds are processed and their first and second moments are captured. The set of points consisting of the mean, variance and mouth parameter value of the training sounds as the X, Y and Z coordinates is used to construct the DTI predictor surface for that parameter. Four surfaces that model the four mouth parameters over the moment space have to be generated. First moment values of the speech for the speakers tested lie in the range $I1 = [1.2, 2.2]$ and the second moment values lie in the range $I2 = [0.9, 1.6]$. We call the Cartesian product of these two intervals, $R = I1 \times I2$, the *moment space*. To handle tracks that may lie outside the DT we create points on the boundary of R and assign height values equal to the value of the closest DT point. This simple algorithm may create discontinuities on the boundary so we smooth the boundary values. Figure 5 shows the predictor surfaces of a speaker with parameter values marked on the DTI surface from the sound track of ‘owie’ (Mathematica automatically converts the DTI to a surface defined over a grid to speed evaluation).

In order to produce predictor surfaces that represent the basic training sounds, it is necessary to collect a cluster of points for each training sound, avoid wide gaps of uncovered regions in the moment space and minimize the overlapping between regions of different sounds.

Predictor surfaces vary over different speakers. This is due to the inter-speaker variations in voice, size and shape of mouth, etc. This research was carried out in parallel to research that investigated the existence of a transformation to map individual predictor surfaces to a set of universal predictor surfaces that could be used to calculate correct parameter values for all speakers. The results of that study appear in [16]. Since there is slight overlap in the moment space of training sound data, the predictor surfaces have some noise. Thus, some smoothing of mouth parameter estimates is required to eliminate jitter in mouth animation. We are examining ways to eliminate the training stage since training may not be possible in all cases.

5. TESTING THE SURFACES

To conduct a thorough test of the surfaces, a set of sounds containing combinations of the vowels and fricatives were recorded. A system was developed to animate the lip movement according to the recorded sound. The points on the sound track of the test utterances were resampled according to the required frame rate, and a set of

points spaced equidistant in time was obtained. For each point in the set, a picture of the appropriate mouth shape was generated. These frames are synchronized with the sound in a QuickTime movie. We use Mathematica to generate animation frames. Since it is not possible to play sound and animation frames simultaneously and control the frame rate in Mathematica, we export animation frames in GIF format and then import them into MovieWorks 4.6 to generate the animation synchronized with the speech recording.

Hermite cubic polynomials are used to represent the edges of the lips and the jaw. The mouth is represented by four curves viz., the upper and lower outer boundary and the upper and lower inner boundary. The jaw is represented by another cubic polynomial. As a first step, a picture of the silent mouth is developed. This shape is used as a reference position and every other mouth position is defined as a displacement from this position. Figure 6 shows the mouth shapes of silence, /aa/, /ee/ and /oo/.

Phonemes sound differently when preceded by different phonemes. This is termed as *co-articulation*. In rapid speech, people do not close their lips after every word of speech. Therefore, we use a neutral or thinking position of the mouth to start and end each utterance. The effect of co-articulation can be produced in the animation by adding extra picture frames at the beginning and end of the original frame sequence. The extra frames represent the movement of lips from the thinking position to the start/end position of the sound.

6. PROCESSING SEQUENCES OF VOICED UTTERANCES

Our system also treats silence. If a silent segment is shorter than 0.5 seconds, the system uses the co-articulation effect to move the mouth from the end position of one segment of speech to the start position of the succeeding speech segment. This is accomplished by using simple S-curve interpolation in Table 1 for each parameter. A cubic Hermite polynomial with zero derivatives at the end points will produce the desired motion.

If the silence separating utterances is longer than 0.5 seconds, the mouth is forced to the closed position between the two utterances. To open the mouth to begin speech from the position of silence, the parameters of the mouth are calculated and linear interpolation is used to

morph from closed to open. In some cases, animation frames must be added at the beginning and end of speech to ensure that mouth motion appears natural. The number of frames depends on the animation frame rate. See ([10]). Similarly for motion from the end of speech to the closed position.

With animator input, the system can also handle the closed mouth sounds /b/, /m/ etc. present in the input signal. We are currently studying the problem of identifying the plosives and the nasals to handle the closed lip cases automatically.

The existing system enables animation of silence and voiced speech utterances. We are implementing a segmentor to divide speech into silence, voiced and unvoiced speech. We have solved the major part of the unvoiced fricative identification problem (see Fu, et al. [7]) and we are developing a complete system to train and build predictor surfaces automatically, accept arbitrary utterances, and produce sound synchronized lip animation automatically.

8. References

- [1] Baker, T.J. *Three-dimensional Mesh Generation by Triangulation of Arbitrary point sets*. AIAA 8th Computational Fluid Dynamics Conference. 1987.
- [2] Belward, John. A. *Surface Fitting with Application to Plant Architectures*. Department of Mathematics, The University of Queensland, Brisbane, Australia. 1999.
- [3] Chandler, Graeme. *Computing in Mechanical Engineering, chapter 4*. Department of Mechanical Engineering, The University of Queensland. 2000.
- [4] Chen, Fang. *Elastic Vector Splines Interpolation*. Research Project, Department of Electrical and Computer Systems Engineering, Monash University, Clayton. 1997.
- [5] Choi, Taek J. *Generating Optimal Computational Grids: Overview and Review*. Department of Mechanical Engineering, Carnegie Mellon University. 1997.
- [6] Fisher, N. I., T. Lewis, and Embleton B. J. J. *Statistical Analysis of Spherical Data*. Cambridge University Press. 1987.
- [7] Fu, H., Rodman, R., McAllister, D., Bitzer, D. and Xu, B. Classification of Voiceless Fricatives through Spectral Moments. *Proceedings of the 5th International Conference on Information Systems Analysis and Synthesis (ISAS'99)*. Skokie,

IL:International Institute of Informatics and Systemics, pp 307-311, 1999.

[8] Huang, Fu Jei. *Real-time Lip-synch Face Animation Driven by Human Voice*. Department of Electrical and Computer Engineering, Carnegie Mellon university, Pittsburgh, PA. 1998.

[9] Krämer, Jörg. *Delaunay Triangulation in Two and Three Dimensions*. Master's thesis. Universität Tübingen, Institut für Informatik, Tübingen, Germany, 1995.

[10] Krothapalli, Chandrika. Developing Predictor Surfaces for Vowels and Voiced Fricatives for Lip Synchronization. MS Thesis. Department of Computer Science, North Carolina State University. 2000.

[11] McAllister, D., Rodman, R., Bitzer, D. and Freeman, A. Speaker Independence in Automated Lip-Sync for Audio-Video Communication. *Computer Networks and ISDN Systems*, V 30, No 21-22, pp 1975-1980, 1998.

[12] Parker, Liam and Jack. M.A. *Utilizing Human Audio Visual Responses for Lip Synchronization in Virtual Environments*. Center for Communication Interface Research. Department of Electrical Engineering, University of Edinburgh. 1996.

[13] Peterson, Samuel. *Computing Constrained Delaunay Triangulation in the Plane*. University of Minnesota. 1998.

[14] Renka J. Robert. *Chapter 8: Curve and Surface Fitting*. The Numerical Algorithms Group LTD, Oxford, UK. 1999.

[15] Shepard, D. *A Two-dimensional Interpolation Function For Irregularly Spaced Data*. Pro. 23rd National Conference ACM. 1968.

[16] Taylor, J., Bitzer, D., Rodman, R., McAllister, D., and Wang, M. Speaker Independence in Lip Synchronization of Vowels.

Proceedings of the SCI 2001 / ISAS 2001 Conference (this conference). Skokie, IL: International Institute of Informatics and Systemics. 2001

[17] Truax, Barry. *Hand Book for Acoustic Ecology*. Second Edition, published by the World Soundscape Project, Simon Fraser University, ARC Publications. 1999.

[18] Wahba, Grace. *Spline models for observational data*. CNMS-NSF Regional Conference series in applied mathematics, 59, SIAM, Philadelphia, Pennsylvania. 1990.

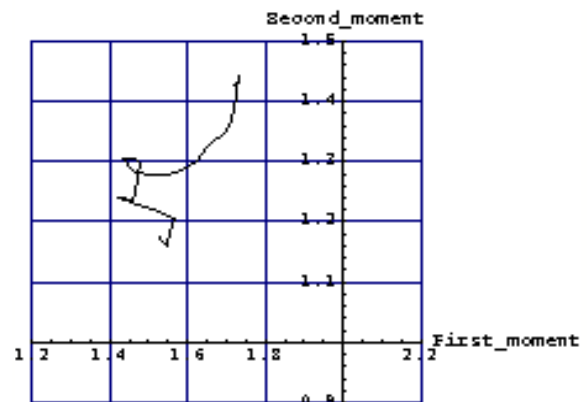


FIGURE 1. A track in moment space

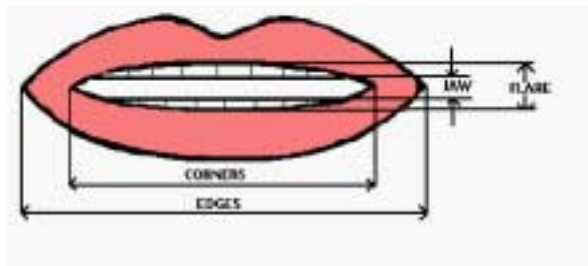


FIGURE 2. Mouth Parameters

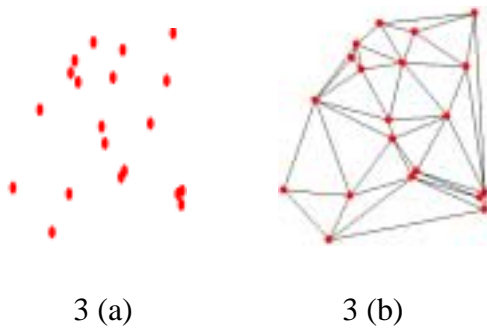


FIGURE 3: Delaunay Triangulation of a set of 20 points generated randomly.

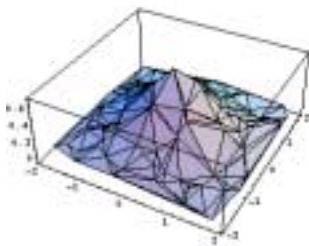


FIGURE 4: DTI of points in 3D

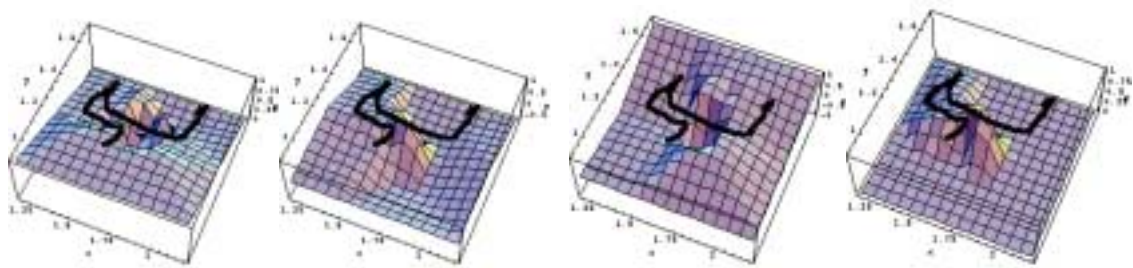


FIGURE 5: Sound track of 'owie' on jaw, flare, corners and edges predictor surfaces.



FIGURE 6a: Mouth shape for silence.



FIGURE 6b: Mouth shapes for sounds /aa/, /ee/ and /oo/ respectively.

Mouth Parameters					
Input	Word	Jaw	Flare	Edges	Corner
/aa/	calm	1	0.15	0	0
/ee/	team	0	-0.75	1	0
/oo/	zoom	0.25	0.75	-1	1
/ey/	name	0.5	-0.75	0.75	0
/ae/	jam	1	-0.75	1	0
/o/	foam	0.5	0.4	-0.25	0.5
/z/	easy	0	-0.1	0	0
/zh/	pleasure	0	-0.75	-0.1	0.15
Silence	_____	0	0	0	0

TABLE 1: Mouth parameter values for training sounds.