

LIP SYNCHRONIZATION OF SPEECH

David F. McAllister, Robert D. Rodman,
Donald L. Bitzer, Andrew S. Freeman

Multimedia Laboratory, Department of Computer Science,
North Carolina State University, Raleigh, NC 27695-8206 USA
Tel. +1 919 515 7480 E-mail: rodman@adm.csc.ncsu.edu

ABSTRACT

Lip synchronization is the determination of the motion of the mouth and tongue during speech. It can be deduced from the speech signal without phonemic analysis, and irrespective of the content of the speech.

Our method is based on the observation that the position of the mouth over a short interval of time can be correlated with the basic shape of the spectrum of the speech over that same interval. The spectrum is obtained from a Fast Fourier Transform (FFT) and treated like a discrete probability density function. Statistical measures called *moments* are used to describe the shape.

For several canonical utterances, video measurements of a speaker's mouth are combined with the corresponding moments to produce continuous predictor surfaces for each of three mouth parameters: jaw position, horizontal opening between the lips and vertical opening between the lips. The method involves smoothing so it is independent of the local behavior of the spectrum.

1. INTRODUCTION

The motion of the lips, tongue, mouth and jaw of a speaker can be deduced from the speech signal without the necessity of speech recognition or previous knowledge of the speech. This speech derivative is called **lip synchronization**, **lip synching** or **lip sync**.

Our emphasis is currently on English vowels and the choice of canonical utterances and transformation techniques that will produce predictor surfaces which generate accurate estimates of mouth motion.

Applications of this work for animation are discussed in [McAl 97a] and for the hearing impaired and other handicaps in [McAl 97b].

We have developed a three-dimensional model of a mouth using bivariate Bezier surface patches to test the results of our methods. The parameters controlling the mouth shape are predicted from an analysis of speech using our techniques.

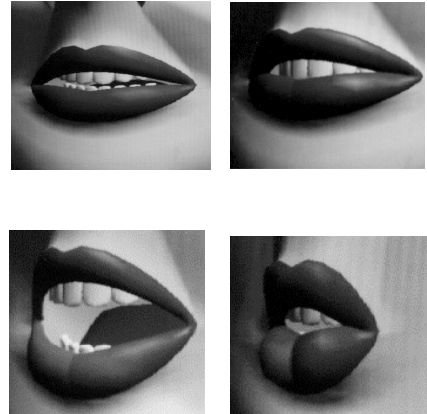


Figure 1: Examples of the Mouth Model

2. METHOD OVERVIEW

We extract the sequence of glottal pulses (GP) of a speech signal using the technique described in [McAl 97b]. We compute the minimum of the sum of the first four odd harmonics of a sequence of FFTs over a window known to contain twice the glottal pulse (GP). We then compute the magnitude of the FFT of the GP, clip the FFT above 4000 Hz, scale it to the interval [0, 22050] and compute shape descriptors of the FFT called moments. We combine moments with measured mouth shape parameters to produce multivariate predictor surfaces for each mouth parameter for a given speaker.

Mouth parameters are measured using video at 30 frames per second. The mouth parameters are JAW, FLARE, and CORNERS. JAW is jaw position and is measured as the distance between the teeth; FLARE is the height of the maximum vertical opening between the lips; and CORNERS is the horizontal opening between the lips (as opposed to a parameter we call EDGES, the distance between the join points of the upper and lower lips). All three are measured and scaled in accordance with the interocular distance of the speaker. Mouth parameter values for all GPs are interpolated from the video values for constructing prediction surfaces.

3. MOMENT CALCULATION

Moments involve an inherent smoothing process which is a global property of a function. If we let f_i denote the i th harmonic of the FFT (where f_0 is the DC term which we set to zero) and n be the number of samples in the GP then we define the k th moment of the FFT to be

$$m_k = \frac{\sum_{i=0}^n i^k f_i}{\sum_{i=0}^n f_i}$$

Central moments are defined as moments about the mean.

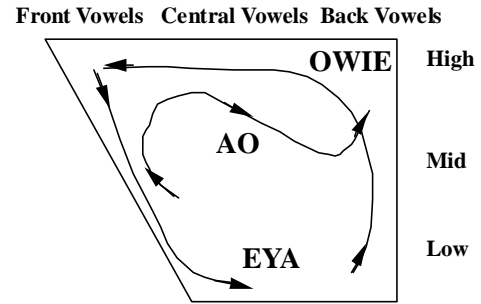
To reduce variation we calculate 10 FTs over 2*GP samples, each shifted by 10% of the GP, and average each harmonic before calculating a moment. This is equivalent to Koster's *dithering* of the FT [Kost 95]. Moving averages are applied to moments and predicted mouth parameters.

The moment for an entire utterance was first divided by the smallest moment to scale them. Moments of utterances for mouth shape prediction were scaled to lie between the minimum and maximum of the corresponding predictor moment.

4. PREDICTOR SURFACES

We have tried many techniques for constructing predicting surfaces. We originally attempted to use the 11 monotone vowels of Koster [Kost 94] which span the vowel chart [McAl 96b] and global least squares approximation assuming a 4 dimensional surface with the first, second and third moments as independent variables. This approach did not provide sufficient accuracy, even using higher order linear surfaces. Rational least squares was no better. We concluded that there were not enough points for adequate surface definition to treat speech transitions.

Our next approach was to produce values for the independent variables using three vowel transitions whose paths "covered" the vowel chart. We refer to them as OWIE, AO and EYA. Phonetically, they are [awi], [eyo^u], and [iya]. Plotting these utterances on a 2D vowel chart for English vowels gives the figure:



Vowel Chart - Tongue Position
Figure 2: Vowel Chart

The number of points used to define the prediction surfaces was the number of GPs collected from these utterances. Three of the moments, the second central moment (the variance), the third central moment (which measures kurtosis) and a nonlinear combination of the two to promote point spread were then used as independent variables for predictor surfaces and parameter estimation. Global least squares approximation failed to have sufficient accuracy. We decided to move to scattered data interpolation methods.

We first tried the Multiquadric surface of Hardy [Hard 71]. It is a global multivariate interpolation technique which is a linear form and predicts the correct relative mouth shape but consistently underestimates the amplitude and sometimes produces negative estimates. It is desirable that our surface be globally nonnegative when given nonnegative interpolation values.

We are now studying the natural extension to three variables of the thin plate spline of Duchon [Duch 77] which is also a linear form and a "minimum bending energy" interpolating surface in the two variable case. In both of the above scattered data methods one need only solve a simple linear system to produce the necessary coefficients.

5. EXAMPLES

We report on a single speaker although similar results have been obtained for other speakers. The test utterance, EIEIO [i yai i yai o^u], from the song *Old MacDonald Had A Farm*, was sung and the mouth movement was predicted and smoothed using both the Multiquadric and the thin plate spline. Video measurements are included to provide a basis of comparison.

The curves containing noise are the predicted values. The thin plate spline appears to overcome the underestimation problem of the Multiquadric. However, it sometimes also predicts negative values due to "crinkles" in the surface. Smoothing often tends to eliminate the problem.

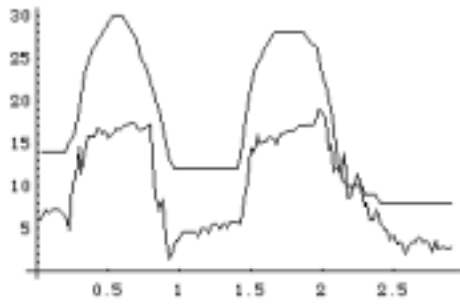


Figure 3: FLARE measured and FLARE predicted for EIEIO using the Multiquadric; x axis is time, y axis in millimeters.

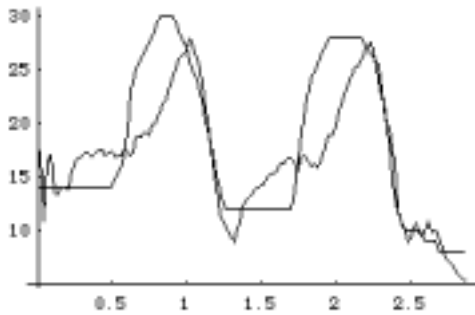


Figure 4: FLARE measured and FLARE predicted for EIEIO using the Thin Plate Spline; x axis is time, y axis in millimeters.

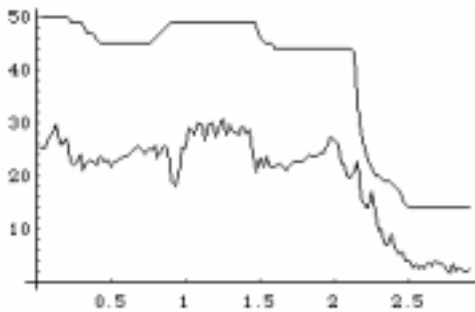


Figure 5: CORNERS measured and CORNERS predicted for EIEIO Multiquadric; x axis is time, y axis in millimeters.

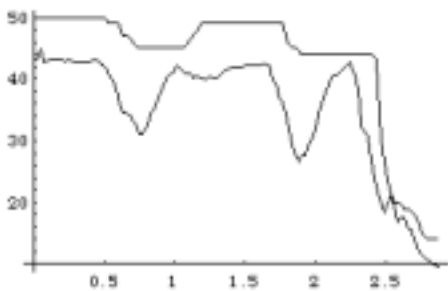


Figure 6: CORNERS measured and CORNERS predicted for EIEIO Thin Plate Spline; x axis is time, y axis in millimeters.

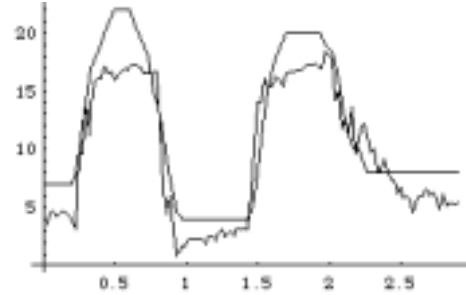


Figure 7: JAW measured and JAW predicted for EIEIO Multiquadric; x axis is time, y axis in millimeters.

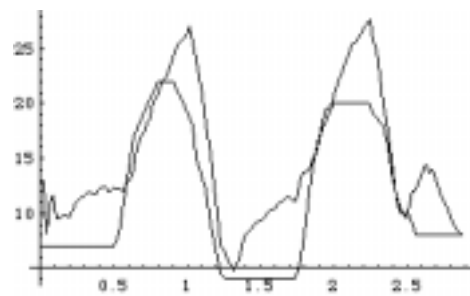


Figure 8: JAW measured and JAW predicted for EIEIO. Thin Plate Spline: x axis is time, y axis in millimeters.

6. CONCLUSIONS

Using moments of the Fast Fourier Transform we have described how to synchronize mouth shape to spoken English vowels for a given speaker. We seek to better understand the relationship between the moments, vowel transitions and mouth motion to produce more accurate predictions.

Our next efforts will be to attempt to produce techniques which are speaker independent and to add consonants in our mouth shape prediction.

7. REFERENCES

- [Duch 77] Duchon, J., "Splines minimizing rotation-invariant semi-norms in Sobolev spaces." In *Constructive Theory of Functions of Several Variables*, (Edited by W. Schempp and K. Zeller), 85-100, Lecture Notes in Mathematics 571, Springer, New York, 1977.
- [Hard 71] Hardy, R. L., "Multiquadric equations of topography and other irregular surfaces," *J. Geophys. Res.* 76, 1905-1915 (1971).

- [Kost 94] Koster, Barrett E., Rodman, Robert D., and Bitzer, Donald L. "Automated Lip-Sync: Direct Translation of Speech-Sound to Mouth-Shape". *Proceedings of the 28th Annual Asilomar Conference on Signals, Systems and Computers*. IEEE publication. 1994.
- [Kost 95] Koster, Barrett E.. Automatic Lip-Sync: Direct Translation of Speech-Sound to Mouth-Animation. Ph.D. Dissertation. Department of Computer Science, North Carolina State University, Raleigh, NC 27695. 1995.
- [McAl 97a] McAllister, David F., Rodman, Robert D., Bitzer, Donald L., and Freeman, Andrew S., "Lip Synchronization for Animation," *Computer Graphics, SIGGRAPH 97*, Los Angeles, CA, August, 1997 (to appear).
- [McAl 97b] McAllister, David F., Rodman, Robert D., Bitzer, Donald L., and Freeman, Andrew S., "Lip synchronization as an aid to the hearing impaired," *Proc. AVIOS 97*, San Jose, CA, September 1997 (to appear).