

LIP SYNCHRONIZATION AS AN AID TO THE HEARING DISABLED

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

**Department of Computer Science
Box 8206
North Carolina State University
Raleigh, NC 27695-8206
(919) 515-7480
rodman@csc.ncsu.edu**

Lip synchronization as an aid to the hearing impaired

1. Introduction

Many properties of speech are elusive and fleeting and require careful analysis to be useful. Still, taken in their entirety they can provide a wealth of information about the speaker that can be applied to manifold useful enterprises.

From speech alone good guesses can be made as to whether the speaker is male or female, adult or child. When previously collected samples of speech from known persons are available, it can be determined probabilistically whether an unknown speaker is one of that group. If an unknown speaker claims to be a certain person, the claim can be verified based on the comparison of speech signals.

Indications of a person's mood, emotional state and attitude may be found in speech. Anger, fear, belligerence, sadness, indignation, reluctance, elation may all be detectable in the speech signal.

Evidence for a person's nationality, region of upbringing, social standing, education level may be found in that person's speech. Whether that person is speaking formally or informally, to intimates or to strangers, to persons of higher social rank or lower social rank, to children or to adults, to foreigners or to nationals may leave a trace in the speech signal.

An accurate portrayal of how the lips, tongue, mouth and jaw of the speaker appear to move during the utterance of the speech may be deduced from the speech signal *without the necessity of speech recognition or previous knowledge of the speech*. This particular speech derivative is variously called **lip synchronization**, **lip synching** or **lip sync**. It is the focus of this paper.

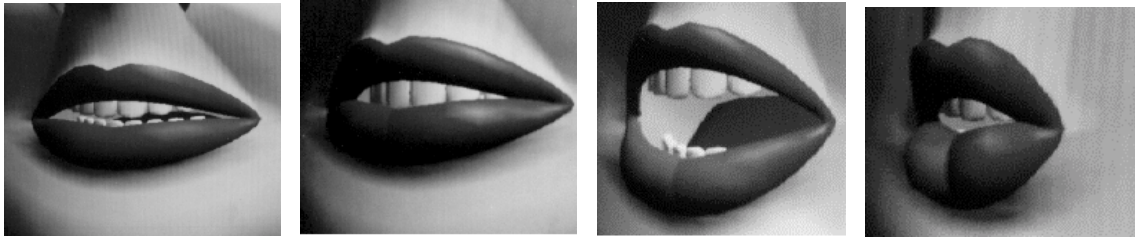
Where possible, everyone uses lip reading to some extent to help disambiguate spoken language [Gree 84]. The hearing impaired depend most heavily on lip reading. A goal of our research is to make it possible to include mouth motion on a graphics screen for the hearing impaired in telephone conversations. Another potential application is use in cockpits or industrial environments where there is heavy background noise [Slag 92]. Other applications are discussed in [McAl 97].

It also plausible that the research would be useful in speech pathology. Without the need for invasive devices, the position of the tongue, lips and jaw can be displayed and used by a speech pathologist to suggest treatment, and by the patient as a means of bio-feedback.

The solution techniques for lip sync would seem to be very much like those for speech recognition, but there are several important differences. Speech recognition

demands distinctions unnecessary for lip sync. For example, /b/, /m/ and /p/ produce similar mouth shapes and need not be differentiated by lip sync. At the same time, lip sync requires timing and other information not needed by speech recognition. For example, *he* and *hoe* have very different shapes for /h/. Speech recognition must take into account accent, language, etc., but these things are of no concern for lip sync. We exploit these differences to get beyond some of the difficulties of speech recognition.

To test the results of our research we have developed a three-dimensional model of a mouth, illustrated below, using bivariate Bezier surface patches. The parameters controlling the mouth shape are predicted from an analysis of speech using our techniques.



We are able to predict the position of the mouth for a speaker by analyzing the behavior of the discrete Fourier transform (FT) of the speaker's voice. Our method is based on the observation that the basic *shape* of the transform for a given sound from a given speaker is relatively static and independent of pitch, and can be reliably correlated with the mouth position of that sound. Hence, rather than approach the problem in the time domain, we analyze the behavior in the frequency domain using simple *shape analysis*. We convert the FT to a discrete probability density function by normalization and use standard statistical shape measures called *moments*.

The goal of our lip sync research is animation suitable for lip reading derived from a speech signal *not known in advance*. If we are successful we intend to design a chip for real time lip syncing independent of the speaker, the text and the language spoken.

The paper is organized as follows:

Section 2: Mouth shape parameters.

Section 3: A new algorithm for locating the glottal pulse.

Section 4: Moments and other parameters used in our predicting equations.

Section 5: A method for reducing the effect of transients on computing mouth parameters.

Section 6: Problems involving pitch variation.

Section 7: Prediction equations.

Section 8: Results and Examples.

Section 9: Curb Cuts

Section 10: Summary and future research.

2. Mouth Parameters

The research described here has been restricted to mouth shape parameters which are externally measurable. We have used actual speaker mouth measurements versus a standard set for all speakers [Kost 95]. These parameters include the horizontal and vertical openings between the lips which we label CORNERS and FLARE respectively. The parameter JAW is a function of tongue height during articulation, and is reflected as the distance between the upper and lower teeth [Park 91, Park 94] (see Figure 1).

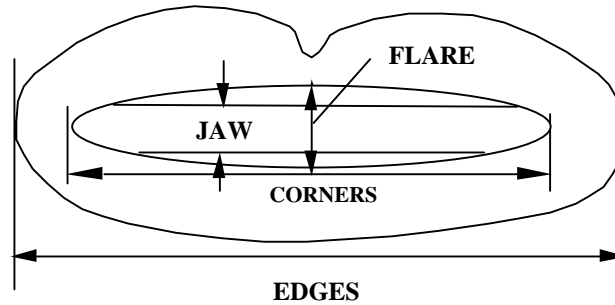


Figure 1: Mouth Parameters

There are actually two parameters which characterize the horizontal movement of the lips: the lip EDGES or join points, where the upper and lower lip connect, and the horizontal opening CORNERS, where the upper and lower lips actually touch during speech. In the experiments reported here we have chosen to report on CORNERS. We have not attempted to predict EDGES independently, but rather allow them to be determined by the other measurements.

Our computational processes are essentially smoothing methods such as computing moments and taking moving averages. By eliminating noise and reducing the effect of transients, we expect sufficient accuracy to be able to exploit continuity in mouth movement to predict mouth position for voiceless utterances. Our discussion here will be centered on voiced utterances.

Video and speech analysis was done using Adobe Premier and Mathematica on a Macintosh 7600. All audio sampling was at 22 KHz. at 16 bit quantization. The mouth measurements are in millimeters. We report on a single speaker although similar results have been obtained for other speakers. However, we have not yet achieved the ability to predict vowel transitions independent of speaker, which remains a goal of our research.

3. Locating the glottal pulse (GP)

The vocal cords vibrate to produce sound which travels through the vocal and/or nasal tract, and is shaped by the tongue, jaw, lips and teeth before emerging from the mouth and/or nose as speech. This opening and closing of the vocal cords create a glottal pulse (GP) that approximates a square wave similar to that produced when rapping on the nose or top of the head to introduce vibrations in the mouth cavity.

While what appears to be the same sound can be made by several different mouth shapes, i.e., the map is many to one, there are differences in the spectrum produced by variations in the mouth parameters which can be identified using the Fourier transform. 2D spectral intensity graphs or spectrograms have been used to identify speech patterns for decades [Lade 93]. Because of transient vibrations which lag behind the termination of a glottal pulse, it is difficult to damp the effect of transients and extract accurate parameters which identify mouth shape if the boundaries of the GP have not been accurately located. Early attempts to quantify mouth parameters used formants which depend on the ability to identify peaks in the FT. Even parameters extracted by smoothing contain sufficient noise as to make the process intractable.

Koster [Kost 95] used the autocorrelation function multiplied by a 7 harmonic function from [Webe 87] and [Stein 86]. We use a different technique. In a region or *window* [L, R] (which depends on pitch) surrounding an estimate of twice the period of the GP, we compute FTs and sum the first four odd harmonics, the reasoning being that all odd harmonics should be zero at $2*GP$ or the sum should produce a relative minimum within [L, R]. In Figure 2 we plot this sum for a monotone (fixed pitch) [ε]. We began searching at L samples, and stop at R samples. We compute the FT for L samples, L+1 samples, L+2 samples, ..., R samples and take the minimum of the sum of the first four odd harmonics over all transforms. This is our estimate of the frequency of $2*GP$. After identifying the first GP, we track using a moving window surrounding twice the old GP of width +/- 20 samples.

For the speaker analyzed in this presentation we set the initial values of $L = 150$ and $R = 300$. Hence, the difference or the length of the abscissa scale in Figure 2 is $R-L = 150$; the origin is $L = 150$. The minimum occurs at 64 and hence the estimate of the GP is half of $150 + 64$ or 107 samples. The speaker is producing a tone at the fundamental frequency of $22050/107 = 206$ Hz.

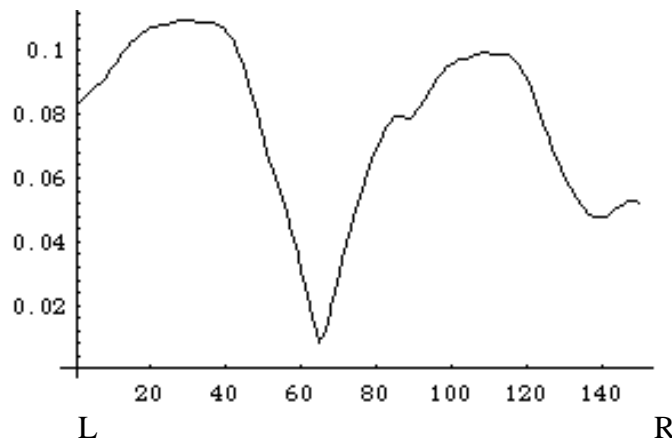


Figure 2: Sum of first 4 odd harmonics for [ε]
x-axis: glottal pulse (GP) number
y-axis: total magnitude

This technique provides for a very stable estimate for the GP during voiced utterances. As an example, Figure 3 shows a graph of the GP's computed for the speaker singing the phrase "EIEIO" from the children's song *Old MacDonald Had a Farm*. No smoothing has been applied. One can identify pitch changes easily. (NB: A higher y-axis value reflects more samples per glottal pulse, which means a greater wave length, hence a lower pitch.)

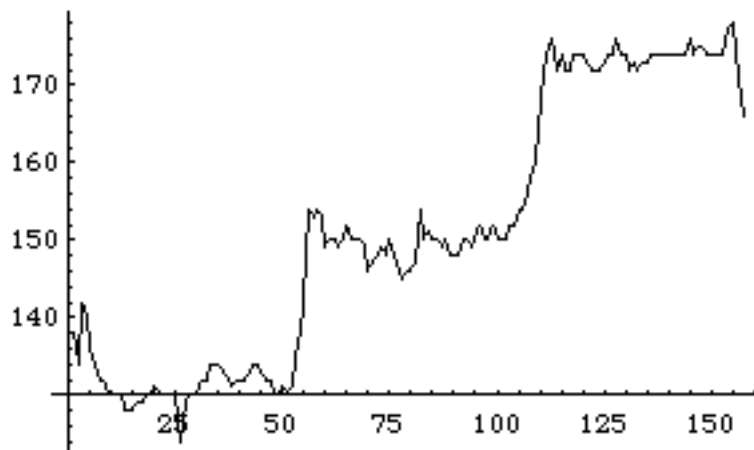


Figure 3: Plot of GP's for EIEIO
x-axis: GP number
y-axis: number of samples in GP

4. FT Shape and Moments

The shapes of the FT for a given speaker for a given mouth shape over various pitches are very similar. The relative location of the harmonics is shifted according to the period. Figures 4 and 5 show examples which illustrate this behavior. (The FFT is symmetric about the midpoint and hence only the first half is shown. The harmonics are interpolated by a linear spline to clarify the shape).

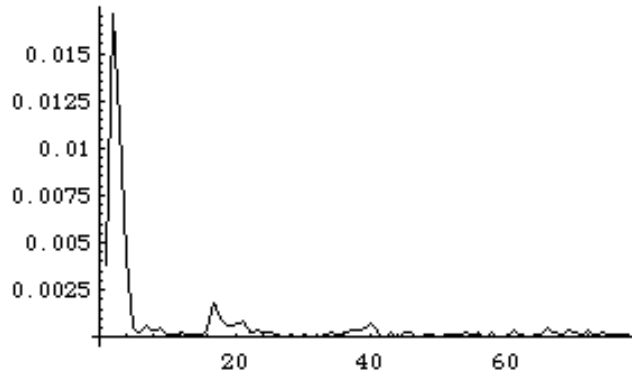


Figure 4: FT for GP=156, [i]
x-axis: harmonic number
y-axis: FT magnitude

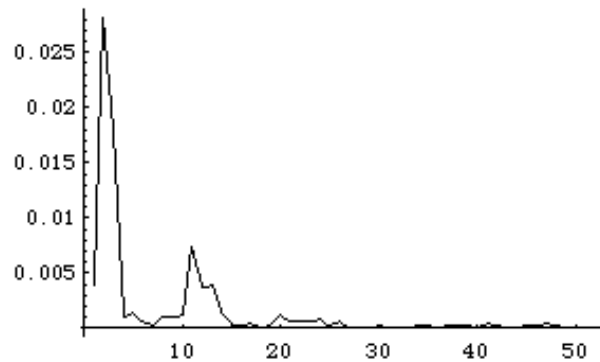


Figure 5: FT for GP=103, [i]
x-axis: harmonic number
y-axis: FT magnitude

The *moments* of a probability density function have long been used as a simple method for describing of the shape of the distribution. Moments involve an inherent smoothing process which is a *global* property of the function vs. a *local* property as would be, for example, the position of local optima. If we let f_i denote the i th harmonic of the FT and n be the number of samples in the GP then we define the k th *moment* of the FT to be:

$$m_k = \frac{\sum_{i=0}^n i^k f_i}{\sum_{i=0}^n f_i} \quad (1)$$

The values of i in equation (1) should be scaled by 22050/GP to ensure that all transforms are over the same frequency interval [0, 22050].

The k th moment about the mean, or the k th central moment, $k \geq 2$, is

$$\overline{m}_k = \frac{\sum_{i=0}^n (i - m_1)^k f_i}{\sum_{i=0}^n f_i} \quad (2)$$

Thus

$$\begin{aligned} \overline{m}_2 &= m_2 - m_1^2 \\ \overline{m}_3 &= m_3 - 3m_1 m_2 + 2m_1^3 \end{aligned} \quad (3)$$

Koster [Kost 94, Kost 95] found various nonlinear combinations of these moments to be useful as single parameter predictors. Some we have found to be effective in constructing training surfaces are defined below:

$$\begin{aligned} m_{20} &= \frac{m_2}{m_1} - m_1 \\ m_{23} &= \frac{\overline{m}_3}{m_2} \\ m_{24} &= \frac{\overline{m}_{23}}{\sqrt{m_2}} \end{aligned} \quad (4)$$

5. Transients in the GP - Harmonic Smoothing

Calculating a set of moments from a single FT for each GP produced results which contained relatively large variation over several successive GPs for a given monotone utterance. We sought to reduce the variation by smoothing the harmonics. See figure 6.

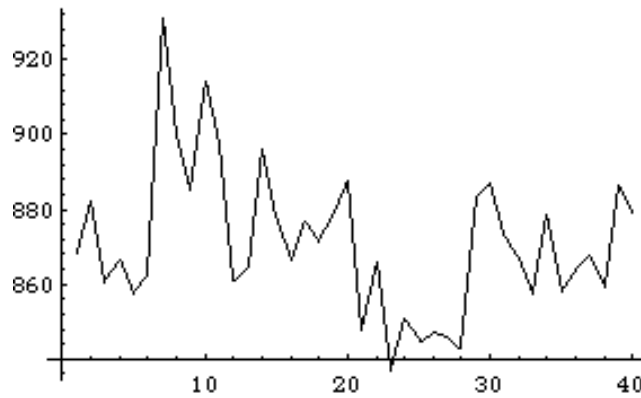


Figure 6: Plot of first moment for [i] over 40 GPs.
x-axis: GP number
y-axis: m_1

The variation could be due to transients in the mouth cavity which had not damped sufficiently during the processing of the following GP. This damping is a very complicated process and may vary for different harmonics for different pitches. We will investigate this in future research.

To reduce variation we chose to calculate 10 FTs over a GP, 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]. We found that computing 100 values of a harmonic by shifting 100 times by a single sample and then averaging did not produce less variation and hence it appears that the 10% shift is sufficient for reducing variation (by harmonic smoothing.) For fixed mouth position and fixed pitch sounds, the maximum relative variation, $(\max - \min)/\min$, was rarely more than 90% as opposed to as much as 500% without the harmonic smoothing.

Additional smoothing was accomplished by applying moving averages for moments and predicted mouth parameters.

6. Pitch Variation - Clipping the FT

After analyzing the FT for many sounds, it is clear that most of the information is contained in several of the lower harmonics (Figures 4 and 5). The remaining harmonics are mostly noise or of very low amplitude. To ensure comparing like-with-like we restrict the calculation of moments to a fixed percentage of the harmonics.

We have chosen the number of harmonics so that the bandwidth is the same over all GPs. Voiced sound frequencies, in general, lie below 4000 Hz (this is not true for voiceless sounds). To equalize bandwidth over all utterances, we have clipped the FT to the first m harmonics where

$$m = (4000/22050)*GP$$

This is equivalent to setting the harmonics to zero above 4000 Hz. It is effective in equalizing moment size over the normal pitch variation which occurs during speech. There appears to be a linear trend when given a wide variation in pitch for a given set of mouth parameters. This trend, however, is not always in the same direction for different sounds for a given speaker. We still do not know the exact reason for this. It could be caused by vocal cord tenseness alterations during pitch changes or biases in our approach. It is a subject for future research. See Figure 7:

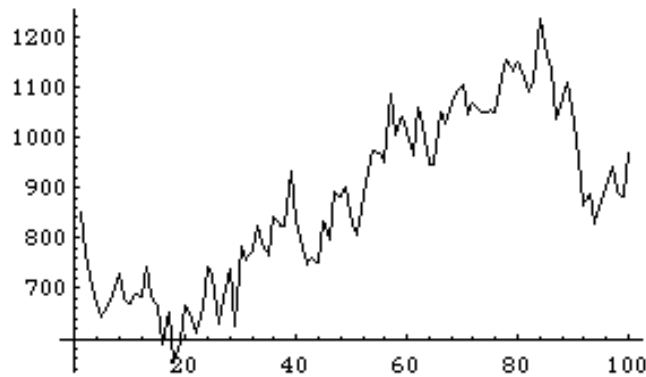


Figure 7: Linear trend in first moment for [o]
 for pitch variation using FT clipping.
 Pitch varies from low to high.
 x-axis: GP number
 y-axis: m_1

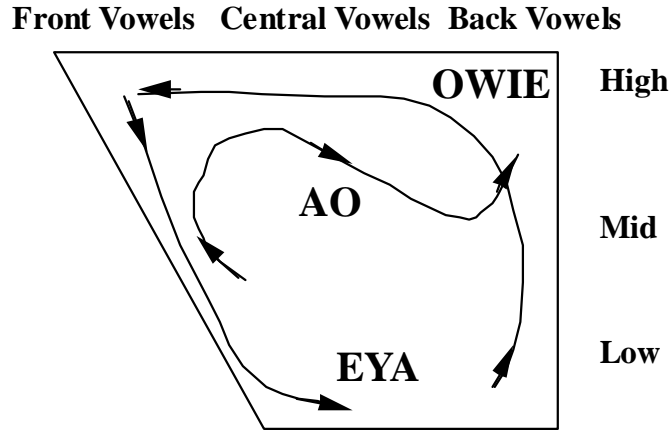
7. Mouth Motion Prediction

We originally attempted to train the system on the following 16 “base sounds,” given in IPA notation approved for American English unless otherwise indicated. By “training” we mean “calculating the predictor equations” for each speaker. Part of the training set was the following 12 vowels, shown in their approximate relative positions in a vowel chart:

i				u
I				U
e	ɜ		o	
ɛ		ɔ]
æ		a		

We found that this approach did not provide sufficient data for establishing an accurate predictor surface.

We decided instead to train the system on three vowel transitions whose paths covered the vowel chart for English vowels. 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 figure 8



Vowel Chart - Tongue Position

Figure 8. Vowel Chart for English with training sounds

The speaker was video taped as the each of the above sounds were uttered. Markers were placed on the nose and lower chin of the speaker to help determine jaw position. When both the upper and lower teeth were visible as in [i], the jaw position could be determined and correlated with the distance between the markers on the nose and chin. Except for variations in the possible motion of the chin marker independent of the jaw opening, this produced an accurate and consistent jaw measurement.

The video was captured at 30 frames per second and the mouth parameters of CORNERS, FLARE, EDGES and JAW were measured and scaled in accordance with the distance between the center of the eyes or the *interocular distance*. This enabled analysis of results which were independent of screen size and video taping session.

Each GP was computed and then deleted from the beginning of each signal and the process repeated for each vowel transition using the technique described in section 3 in constructing the data points for the predictor surfaces. For test utterances, after a GP was computed, 300-500 samples were deleted from the signal before processing the next GP. Moments were calculated for each GP using the FT shifting procedure described in section 4. The set of values for a given moment was always divided by the smallest value in the set. Moment values for test utterances were always linearly transformed so that the maximum value in the set was identical to the maximum value of the corresponding moment used in the predictor surfaces. Three of the moments, the second central moment (the variance), the third central moment and the nonlinear combination m24 described in equations (4) above were used as independent variables for a predictor surface.

Note that the third “independent” variable is a function of the first two and hence the data points lie on a surface in three dimensions. This suggests that it may be possible to get reasonable results from assuming that the problem requires only two independent variables. Preliminary tests show this to be the case using second and third central moments as the independent variables although using the third variable appears to reduce noise in parameter estimates.

First order global least squares surfaces did not provide sufficient accuracy for prediction. We then tried the Multiquadric interpolatory surface of Hardy [Hard 71] which is considerably better in predicting the correct mouth shape but consistently underestimates the amplitude and sometimes produces negative estimates. It is desirable that our surface be globally nonnegative when given nonnegative dependent variable interpolation values.

We are studying other multivariate scattered data interpolation surfaces including gridding methods, and the thin plate splines of Duchon [Duch 77] which are “minimum bending energy” interpolating surfaces.

8. Results

The speaker was video taped for the three training utterances described above and the predictor equations were then calculated. The test utterance, EIEIO [i yai i yai o^u], was sung and the mouth movement was predicted and smoothed using the predictor surfaces. Physical measurements were also taken, as with the training sets, to provide a basis of comparison.

Some of the plots of the predicted vs. measured parameter values are given below. The curves containing noise are the predicted parameter values. The Multiquadric produces excellent agreement in relative mouth movement but underestimates the ranges. Preliminary tests show the thin plate spline appears to overcome the underestimation problem. We will report on these results in a later paper.

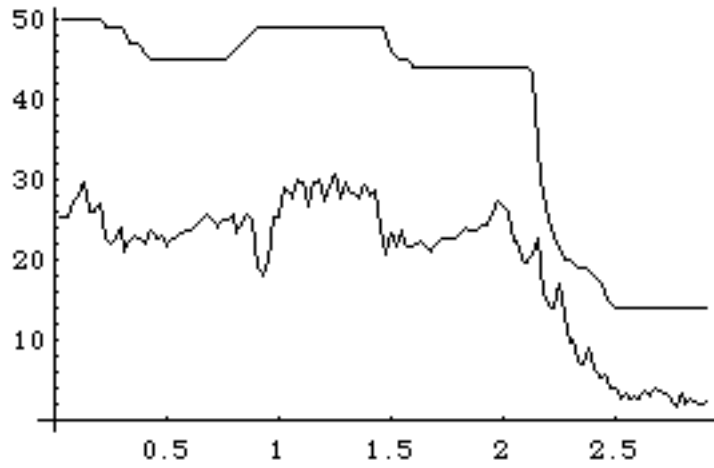


Figure 9: CORNERS measured and CORNERS predicted for EIEIO
x axis is time
y axis in millimeters

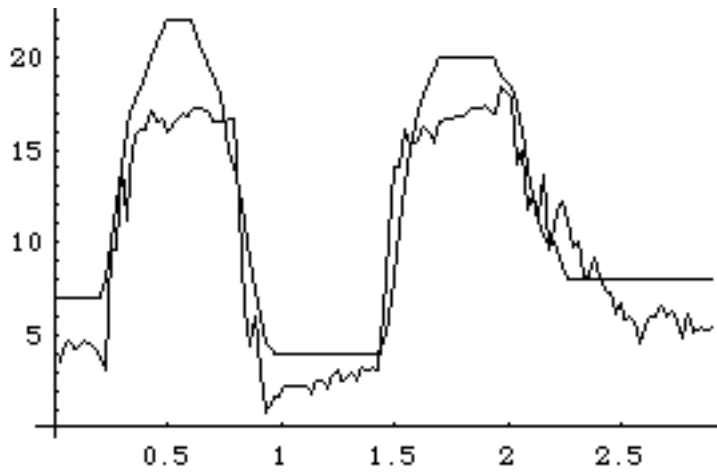


Figure 10: JAW measured and JAW predicted for EIEIO

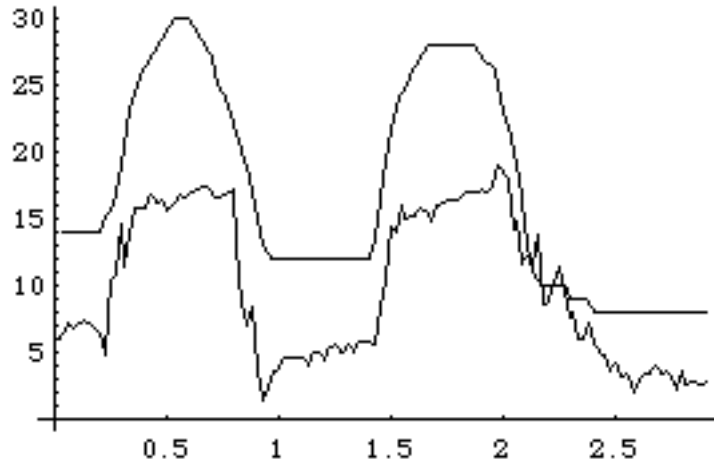


Figure 11: FLARE measured and FLARE predicted for EIEIO

9. Curb Cuts

“Curb cuts” are symbolic of applying technology to improve the quality of life of individuals who, through disease, genetics, accident or age, have a diminished capability, such as a seeing or hearing impairment.

The research described in this report, while preliminary, holds out the hope that accurate mouth movements of spontaneous speech can be calculated and displayed in real time, therefore providing an opportunity for lip reading for the hearing impaired for conversations in which the talking party is remote, as over the telephone. Ideally, the actual face of the talking party would be seen — the video phone concept. Unfortunately, there is not enough bandwidth in current telephone technology to transmit the dozens of video images per second needed for lip reading. Rather, we propose, a graphical talking head be attached to the telephone of a hearing impaired individual. The speech from the other end uses the methods described above to animate the mouth movements based on the voice of the talker.

We have also done some preliminary work in combining computer generated “cued speech” symbols with lip synching. Cued speech, when used by a human talker, are hand gestures that disambiguates sounds that are indistinguishable through lip reading, such as /p/, /b/, and /m/.

We have also begun to search for spectral parameters that can be correlated with tongue movement [Kost 94]. Success in that enterprise would provide a way for speech pathologists to analyze the speech of verbally impaired individuals, such as dysarthrics, without resorting to invasive devices. For persons who have speech defects, a display of tongue position could be the basis of bio-feedback therapy. The subject could observe the tongue position of utterances and experiment with ways of adjusting it toward normal.

10. Summary and Continuing Work

Using moments of the Fourier Transform we have demonstrated the ability to synchronize mouth shape to spoken English vowels for a given speaker. We intend to investigate more sophisticated approximation methods. We expect that improvements in the approximations will yield more accurate predictions and we hope to be able to produce techniques which are speaker independent.

We will then move on to include consonants in our mouth shape prediction. We have found that fricatives and nasals produce discontinuities which cannot be treated using the approach above. Simply recording and analyzing a sound and adding it to the database for a speaker does not produce acceptable results using our current approximation methods. We will probably have to exploit continuity of mouth movement to help estimate mouth positions. Other alternatives include using one-sided moments (e.g. computing the normalized sum of harmonics to the left of the mean) to further clarify shape, and Markovian methods to predict the most likely mouth position.

11. 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
- [Gree 84]. Greenwald, Audrey B., *Lip-reading Made Easy*. Alexander Graham Bell Association for the Deaf. 3417 Volta Place NW, Wash., DC 20007. 1984.
- [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. "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.
- [Lade 93]. Ladefoged, Peter. *A Course in Phonetics, Third Edition*.. Atlanta:Harcourt Brace Jovanovich, Inc. 1993.
- [McAl 97] McAllister, D. F., Rodman, Robert D., Bitzer, Donald L. & Freeman, Andrew S., "Lip Sync for Animation," SIGGRAPH 97 Visual Proceedings, Los Angeles, CA, 1997, (to appear)

[Park 91]. Parke, F. I. "Control Parameterization for Facial Animation," in N. M. Thalmann and D. Thalmann (Eds.) *Computer Animation '91*, Tokyo: Springer-Verlag, 1991.

[Park 94]. Parke, F. I. & Waters, K. (1994) *Computer Facial Animation*, AK Peters. ISBN 1-56881-014-8.

[Slag 92]. Slager, Robert P.. "Device could help hearing impaired with telephone talk". *Western News*. Western Michigan University. V 18, N 18. Pp 1-2. Jan 30, 1992.

[Stein 86]. Steigerwald, Silvi K. *Development Tools for a Speech Mimicking System*. MS thesis for EE at U of IL at Urbana-Champaign. 1986.

[Weber87]. Weber, Richard Joseph. *A Digitalk Speech Processor With Variable Window Sizes*. MS thesis for EE at U of IL at Urbana-Champaign. 1987.