We report on research into the lip synchronization of speech that has not been preprocessed into phonetic units. It is based on the notion that given any spoken words as input, a computer can accurately portray the mouth movements made during the pronunciation of those words, and can do so without the necessity of speech recognition or electromechanical devices attached to the jaw.

Our method is based on correlating parameters of lip and jaw movement with parameters drawn from the spectrum of the speech signal. We found a minimum of three parameters - horizontal lip opening, vertical lip opening and jaw height - sufficient to allow the animation of a model of a generic human mouth. The speech parameters are derived by treating the magnitude of the discrete Fourier Transform as a probability density function and using well known statistical measures called moments to characterize the shape of the distribution.

To establish mouth shape predictor or training functions, which are described mathematically as a surface in multi-dimensional space, we video tape and record a speaker uttering a set of basic training sound transitions that cover the ranges of jaw position, lip rounding and lip spreading. Using a very accurate glottal pulse estimation technique, the first, second and third moments, second and third central moments and various linear and nonlinear combinations of these moments are used to produce a set of independent variables. The mouth measurements obtained from the video are then used as the dependent variable and a multivariate training surface, a thin plate spline, for each mouth parameter is derived from the training utterances.

We have been successful in predicting mouth movements associated with English vowels. The presentation will include graphs which compare actual measured mouth motion and predicted mouth motion. We will also present a short video tape comparing actual speaker utterances and a computer graphics animation of a three-dimensional model of the lips and jaw where the motion of the model is driven by our prediction technique. The model is a collection of bivariate Bezier surfaces in which the control points are modified each glottal pulse in accordance with the mouth positions determined by the training surfaces.

When the speaker who was videotaped to train the system is used to test the system, the accuracy of the predictions is extremely high, well above the threshold necessary to animate the mouth in concert with the speech. We are currently working on construction of training surfaces which would provide adequate predicted mouth motion, independent of the speaker. We are also attempting to incorporate vowels which occur in languages other than English. We will then move to include voiced consonants such as /m/, /n/, /r/, and /z/.

We are augmenting the technique to include additional parameters such as tongue position and a parameter to help differentiate between the edges of the lips and the horizontal
opening of the lips during speech. These parameters will help introduce more realism into
the model. Modeling of actual human faces will be attempted later.

Applications include animation, training in lip reading, speech therapy, video conferencing,
man-machine interfaces, voice compression, training and education, linguistic research and
help in speech recognition.