TOWARD SPEAKER INDEPENDENCE IN AUTOMATED LIP-SYNC

David F. McALLISTER, Robert D. RODMAN,
Donald L. BITZER and Andrew S. FREEMAN
Voice I/O Group, Multimedia Lab
Department of Computer Science
North Carolina State University
Raleigh, NC USA 27695-7534
dfm@csc.ncsu.edu
http://www.multimedia.ncsu.edu/research/voiceio/

ABSTRACT

By analyzing the absolute value of the Fourier transform of a speaker’s voice signal we can predict the position of the mouth for English vowel sounds. This is without the use of text, speech recognition or mechanical or other sensing devices attached to the speaker’s mouth. This capability can reduce the time required for mouth animation considerably. We expect it to be competitive eventually with the speech/text driven solutions [Mori 93] which are becoming popular. Our technique would require much less interaction from the user and no knowledge of phonetic spelling. We discuss the problems of producing an algorithm that is speaker independent. The goal is to avoid having to measure mouth movements off video for each speaker’s training sounds. We have discovered that eliminating variation due to pitch yields moments which are mouth shape dependent but not speaker dependent. This implies that careful construction of predictor surfaces can produce speaker independent prediction of mouth motion for English vowels.

Key words: facial animation, lip synchronization, and speech processing.

INTRODUCTION

We have presented our results for speech driven lip synching of English vowels for a single speaker in [Koster 95], [McAllister et al. 97a], [McAllister et al. 97b] and [McAllister et al. 97c]. We show by computing shape evaluators of the magnitude of the Fourier transform of sections of the digitized input sound that predictor equations can be derived to predict mouth parameters which are used to describe the motion of the mouth corresponding to the input sound. This is without the use of text, speech recognition, or mechanical or other sensing devices attached to the speaker’s mouth. This capability can reduce the time required for mouth animation considerably. Our method runs very fast, requires only simple techniques, and is potentially speaker independent.

We determine the fundamental frequency of the input sound at successive intervals, or, equivalently, the length of the glottal pulse (GP), by optimizing a linear combination of the first eight harmonics of a sequence of FFTs derived from the input signal. Once the GP has been detected and is being tracked accurately, the FFTs of successive intervals are scaled, clipped, smoothed and normalized to
produce a probability density function. The appropriate moments are then computed and scaled and used as the independent variables input to a bivariate predictor function for each visible mouth parameter, which are the dependent variables. These parameters are: Flare, or the vertical distance between the upper and lower lip; Jaw, or the distance between the teeth; Corners, or the horizontal opening between the lips; and Edges, the distance between the join points of the upper and lower lips. Predictor functions are computed using moment sequences from several training sounds.

Our method works well because it does not depend on local behavior of the input signal or its transform. We avoid trying to detect such phenomena as the location of formants, for example. Neither do we use the computationally intensive method of hidden Markov models which is common in speech recognition. Since smoothing is applied at several steps of the process, our method is based on global behavior of the GP and its transform.

The method makes successful predictions of mouth movements for cases in which the mouth measurements and the training sounds are from the same speaker, and the sounds are restricted to English vowels. When we compare the predicted mouth parameters to the actual values measured from video for non-training sound sequences, the accuracy has been good. Indeed, in some cases the technique has been sufficiently accurate to enable us to detect mouth movements missed by the measurement process.

**SPEAKER INDEPENDENCE**

We describe our attempt to extend the results to treat multiple speakers. The goal is to avoid the labor intensive problem of measuring mouth movements from video of everyone who uses the system. There have been recent attempts to automate the process of observing and measuring the parameters of mouth motion [Basu-Pentland97, Reveret et al. 97, Vogt97] but nothing has appeared that produces the accuracy needed for our application.

We seek mouth movement which can be used for applications such as lip reading and cartooning as opposed to highly accurate prediction of a specific speaker’s mouth. Hence, we seek to use as much information as possible from a “standard” speaker’s prediction surfaces. There are three possible approaches to the problem which depend on the independence of moments used to predict mouth motion.

The most desirable approach is to use the same prediction surfaces for every speaker. This assumes that the moments produced from different speakers for the same sound are identical (up to low amplitude noise and transients produced by different configurations of the mouth cavity and minor variations in mouth movement). This technique requires no training of the system for a new user. This technique assumes low intra- and interspeaker variation in moments used to predict mouth positions.

The next best alternative is to use the measurements of a generic speaker and recompute predictor surfaces by matching training sounds from a new user to the training sounds of the generic speaker in some way. This method is required if there is low intraspeaker variation but high interspeaker variation in moments. Training the system is required for each speaker but if training utterances are
matched accurately then no additional mouth measurements are required unless accurate individual
speaker mouth motion is desired, a requirement not usually necessary in animation.

If there is both high inter- and intraspeaker variation, then both training sounds and the associated
mouth measurements from each new user must be made to produce individual predictor surfaces.
This makes the system less convenient to use, but not useless. Users of the system would have to
submit to a video taping session, based on which their own personalized predictor surfaces would be
created. Afterward, their use of the system would be simple and straightforward.

The results below are for three speakers: Speaker A and B are males and Speaker C is female.

INTRASPEAKER VARIATION

Linguistics researchers and experts in voice processing have known for years that shape
characteristics of a given sound are relatively static for a given speaker. For example the relative
locations of the first and second formats in the high front vowel /i/ are relatively constant over pitch
variation for a given speaker. However, locating formants automatically is difficult and error prone
because they represent local behavior of the FFT. We have chosen to use moments, which are
common statistical descriptors for density functions, as the basis of our descriptions of the spectrum.
We have discovered, however, that our initial successes [McAllister et al. 97a, McAllister et al. 97b,
McAllister et al. 97c] were due to the fact that our test speaker spoke in a relatively monotone voice.
When trying to duplicate the experiments for other speakers who spoke with a wider variation in
pitch, the results were not reproducible. Moments tended to vary over 100% with a change in pitch
for a given sound, so that sounds which appeared in one training utterance had no correlation with
similar sounds in other training utterances. The first step toward speaker independence was to
eliminate the effect of pitch on the moments.

In previous attempts we had shifted the FFT by 10% of a GP to help reduce the variation, but the
reduction was not sufficient. We discovered that shifting the FFT by a single sample k times where k
is the number of samples in the GP, averaging harmonics and then computing the cube root,
minimized the effect of the large variation in the amplitudes of the first and second formants and
reduced the variation of the mean and the variance to less than 5%, a considerable improvement. We
will refer to these operations on the FFT as producing the modified FFT.

A similar reduction of variation for the third variable we had used in previous research, the third
central moment, was not forthcoming. Hence, we elected to eliminate the use of this variable and
began a search for another candidate which did not have such large variation, thinking that at least
three variables were necessary on which to base the predictor function for mouth motions. Before
beginning the search we tried using only the mean (i.e., the first moment) and variance (i.e., the
second central moment) to see how well they predicted sounds and discovered that, at least to
predict the mouth shapes for the vowels, those two variables are sufficient. In establishing this fact,
we first plotted points corresponding to a given sound with a static mouth position but including a
pitch change. Figure 1 shows the points for /i/ versus /u/ for speaker B, for example, while Figure 2
shows the track for the utterance OWIE /awi/ for the same speaker. Note how the track passes
through the high back /u/ mouth position and moves on to the high front mouth position of /i/, a
result which demonstrates the consistency of mouth prediction for a given speaker and low intraspeaker variation.

![Figure 1: A comparison of /i/ and /u/ with pitch change using the modified FFT for speaker B.]

**INTERSPEAKER VARIATION**

The variation reduction procedure described in the previous section not only had a positive result for predicting mouth motion for a single speaker, but it also produced results which were similar over several speakers. Tracks of moments versus time produced by our three speakers for the same sound were almost identical except for timing considerations and variations in mouth position. As examples, Figures 3 and 4 below show the tracks for the sound OWIE for speakers A and C respectively. Compare them with the OWIE track for speaker B in Figure 2.

![Figure 2: A plot of OWIE (/awi/) for speaker B.]

It appears that the method produces low interspeaker variation which can be exploited by developing single speaker independent predictor surfaces for each mouth parameter as described in the following section. Our current results suggest the same surfaces can be used for any speaker whose native language is English, which we will investigate in future research.

**PROCEDURE**

We used the same training sounds and mouth measurements of speaker A as in our previous papers, but we now use only the mean and the variance of the modified FFT as independent variables.

We have constructed biquadratic least squares surfaces where we have added points along the boundary of the rectangle $R = [1.3, 2] \times [1.1, 6]$ containing the tracks of training sounds, which both reduces ill conditioning of the problem and ensures that occurrences of (mean, variance) pairs which lie too far from the training tracks do not produce estimates of mouth position which are negative or otherwise outside reasonable bounds. This was accomplished by taking equally spaced points along the boundary of $R$ and using the measurement of the closest training point in each case. We realize this may produce discontinuities of the surface on the boundary and we will investigate other approaches in the future. Simple smoothing along the boundary may, however, eliminate any problems. As an example Figure 5 shows the surface produced by our technique for predicting Jaw motion.

![Figure 3: A plot of the OWIE track for speaker A.](image-url)
CONCLUSIONS AND FUTURE RESEARCH

The above results suggest that we may now have speaker independence which can be described by two independent variables for predicting English vowels. We can now spend our energy on the development of universally effective predictor surfaces. We are eager to move to sounds other than vowels and glides, for example nasals and liquids, to see if similar approaches can be used. With similar success in the vowel-like (resonating) consonants, we would then move on to fricatives, stops and affricates, both voiced and unvoiced, so that ultimately we could lip-sync any user speaking any English phrases. Our results will be reported in a future paper.

Although we are working in English there is nothing about the research that is inherently dependent on English. In the more distant future we anticipate the investigation of lip-synching for all human languages, and, ultimately, the development of linguistically universal predictor surfaces.

REFERENCES

Koster, B. Rodman, R. and Bitzer, D. 1994

Koster, Barrett E. 1995

Lewis, J.P. 1991

[McAllister, David F., Rodman, Robert D., Bitzer, Donald L. and Freeman, Andrew S., 1997

McAllister, David F., Rodman, Robert D., Bitzer, Donald L. and Freeman, Andrew S., 1997

McAllister, David F., Rodman, Robert D., Bitzer, Donald L. and Freeman, Andrew S., 1997

Morishima, S. and Harashima, H. 1993

Parke, F.I., and Waters, K. 1996

Reveret, L., Garcia, F., Benoit, C. and Vatikiotis-Bateson, E. 1997

Vogt, M. 1997