

# Speaker Independence in Lip Synchronization of Vowels and Distinguishing between /m/ and /n/

Jamie TAYLOR, Donald BITZER, Robert RODMAN, David McALLISTER, and Meng WANG  
Voice I/O Group

Department of Computer Science, North Carolina State University  
Raleigh, NC 27606, USA

## ABSTRACT

This paper presents a method for achieving speaker independence in the lip synchronization of vowels by performing an affine transformation on the moments of the spectrum of the speech. A speech sample is separated into samples by glottal pulse period. A track in moment space is made by taking the first and second principle moments of the spectrum of each sample. For a given speaker, a transformation can be created to map the speaker's moment space to a "standard" moment space from which the mouth shape predictor surfaces are created.

Also, we describe an approach to distinguishing between the /m/ and /n/ sounds by adjusting based on the velocity of the track moving into or out of the /m/ and /n/ region in moment space. Results are presented for distinguishing "ahm" and "ahn".

Keywords: Lip Synchronization, Speaker Independence, Viseme, Animation, Predictor Surface

## 1. INTRODUCTION

Lip Synchronization is an area of applied speech processing that is attracting many speech developers. Our Lip Synch system accepts freely spoken, text-free speech as input, and produces parameters to a graphics program that animates the mouth and face of a talking head as if it were speaking the speech. Such programs are very much in demand by various factions in the animation industry, and has other applications as well, such as an aid to the hearing impaired [1].

Today there are competent lip synching systems that accept text, with or without speech, and produce decent animation. The challenge in this area is to achieve lip synching without the benefit of text. One attempt at this is made by running a speech recognition front end, in effect, providing speech-to-text. Errors in such texts – the result of inaccuracies in the difficult process of context-free, continuous speech recognition may result in incorrect animations.

The method of lip synching considered in this paper is to process the speech signal directly, and derive graphics

parameters without text or an intervening level of speech recognition. [1] Mouth shape parameters are computed based on the spectral moments of the input speech signal.

Up until now, the lip synching has been speaker dependent. That is, a modest amount of training is required to establish the speech patterns of an individual, and to construct personal predictor surfaces that define the parameters necessary for animation. [2] This paper describes how to achieve speaker independence for lip synching vowels by creating a rational matrix transformation to map the spectral moments of any speaker onto those of a "standard" speaker.

The /m/ and /n/ sounds share the same region in moment space, and thus are extremely difficult to differentiate. In fact, they are acoustically similar and often confused by human listeners. This paper describes an approach for distinguishing between the /m/ and /n/ sounds by using information about the transition into or out of the /m/ and /n/ region.

## 2. BACKGROUND

In linguistics, a single unit of sound is called a phoneme. The corresponding concept in lip synchronization, the position of the tongue, mouth, and jaw, is called a *viseme* (visual phoneme). Several phonemes may map to the same viseme. For example, /p/, /b/, and /m/ form a single viseme. Lip synchronization is the process of deriving visemes from sounds in a voice data stream, whereas speech recognition must produce phonemes from the data stream. Lip synchronization is easier than speech recognition because multiple acoustically similar phonemes may share a single viseme and not require differentiation. [5-8]

Speech sounds can be divided into two kinds, voiced and unvoiced. Voiced sounds are made by vibrating the vocal chords, whereas unvoiced sounds are made without vibrating the vocal chords. The vibration of the vocal chords produces glottal pulses, where one glottal pulse (GP) is the sound produced by a single vibration of the vocal chords. The spectrum of voiced speech is the product of the glottal pulse (the driving signal) and the mouth shape (the filter). To prevent the driving signal from interfering in our quest to characterize the filter, we

process over the period of only one glottal pulse at a time. To divide the sample into glottal pulses, the glottal pulse periods (GPPs) are computed using discrete Fourier transforms by the algorithm described in our previous work. [1]

Our previous work [1] has also shown that *moments* of the spectrum of a voice sample can be used to derive its associated viseme. After dividing the voice sample into glottal pulses, the *moments* are computed for the spectrum of the sample during each glottal pulse period. The first moment,  $m1$ , is the mean of the spectrum. The second principal moment about the mean,  $m2$ , is the variance. The spectrum of each glottal pulse period is distilled into a single point in  $m1$ - $m2$  space (also referred to as moment space). The points for a voice sample can be interpolated to produce a *track* in  $m1$ - $m2$  space.

The track in  $m1$ - $m2$  space may be mapped onto a continuous predictor surface for each of three mouth shape parameters: jaw position, horizontal lip opening, and vertical lip opening. (There are actually two parameters for horizontal mouth opening.) There may be several different predictor surfaces for the different *types* of speech, e.g. vowels, fricatives, and nasals [3]. This paper focuses only the vowel predictor surface and that of the nasals /m/ and /n/, but one would expect the same technique to work on the other surfaces as well.

The tracks of different speakers uttering the same sound are similar in overall shape and size, but differ significantly in absolute position. This means that the mapping from a track in  $m1$ - $m2$  space to a mouth shape predictor surface is speaker-dependent. This paper shows how well a simple transformation can map the moments of different speakers onto a standardized  $m1$ - $m2$  space used to compute the correct viseme for vowels. This work may be considered an extension of work in "vowel normalization," which has a history dating back to 1890. That history, and a description of the various attempts to normalize vowels and vowel space is discussed in [4].

The lip synchronization system can be constructed to translate from the standardized  $m1$ - $m2$  space to visemes. Once a mapping from an arbitrary speaker to the standard speaker has been found, the speaker can immediately use the lip synchronization system without having to retrain it.

The /m/ and /n/ sounds share the same region in moment space, but have different visemes. In fact, most humans have difficulty distinguishing between the two sounds without the surrounding context. We are lead to believe that the proper way to distinguish between /m/ and /n/ is not to view the position of a few samples in moment space, but rather to examine *how* the track moves into the m-n region in moment space.

### 3. METHODS

#### Vowels

Recordings were made of 10 people saying 12 utterances 5 times each. The 12 utterances were chosen to include the transitions from /ee/ and /oo/ to the consonants /m/, /r/, and /z/, as well as to the vowels /ee/, /ah/, and /oo/. Our previous work [1,2] had indicated that /ee/, /oo/, and /ah/ are fairly discrete (i.e., the bounding boxes for each vowel for the same speaker do not overlap). They also represent extreme mouth positions, thus making them good choices for created a mapping from an arbitrary speaker to a reference  $m1$ - $m2$  space.

The utterances were separated by hand into samples containing one word each. The samples were analyzed by the technique described in our previous work to find the track of the voice in  $m1$ - $m2$  space yielding charts such as Figure 1.

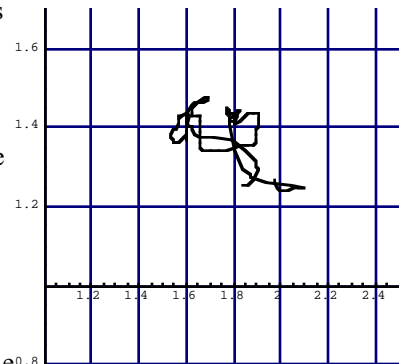
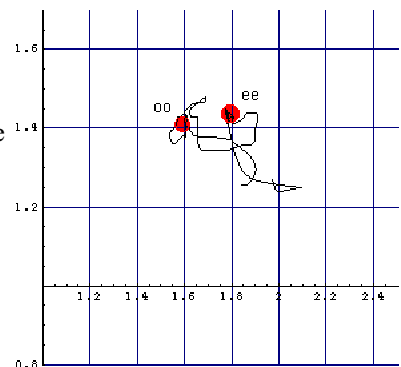


Figure 1: The track in  $m1$ - $m2$  space of a female speaking "oowee".

The positions of the vowels /ee/, /oo/, and /ah/ were marked by hand on the charts when the transition could be identified (see figure 2). Vowels are distinguished by periods of relatively little motion in the  $m1$ - $m2$  track, whereas transitions are usually periods of quick motion without doubling back.



Vowels can also be identified by finding a relatively "flat" segment in the moment vs. GPP chart for either one of the moments. This is illustrated in figure 3.

The mean positions of each vowel for each person were computed and charted, along with the minimum and maximum values as a *bounding box* (see figure 4 below). The bounding box is a rectangle formed by the points  $\{(min1, min2), (max1, max2)\}$ .

The bounding boxes are necessary because each vowel was used in different transitions. Each different transition causes slight variations in the sound and mouth shape, so it is to be expected that the corresponding points in  $m1$ - $m2$  space do not coincide

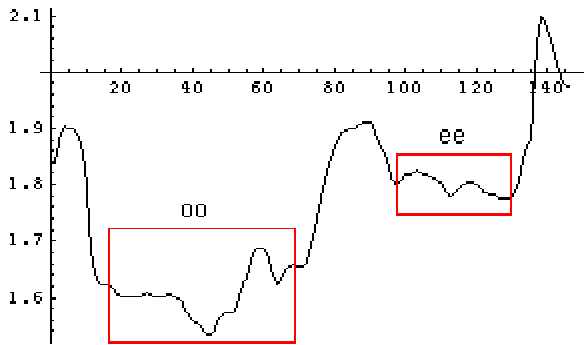


Figure 3: The chart of  $m1$  vs. GPP for the same voice sample as in figures 1 and 2 marked by hand.

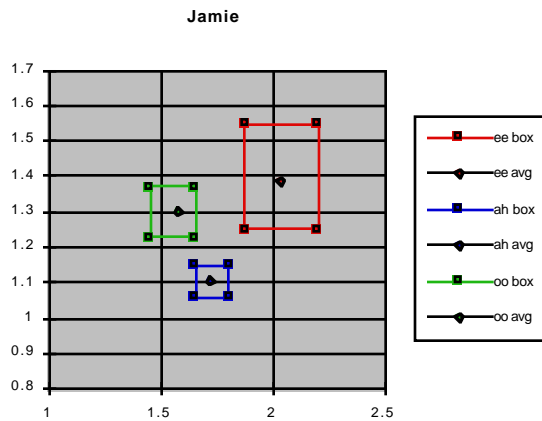


Figure 4: A chart showing the mean positions of /ee/, /ah/, and /oo/, and the bounding boxes for each.

exactly. Specifically, the points in  $m1$ - $m2$  space for the two vowels in a transition are closer together than if the vowel sounds appeared alone. This phenomenon, known as coarticulation, occurs because the actual mouth shapes for the two vowels are closer together than the mouth shapes for the pure vowel sounds. Given the set of samples of a vowel used in the exact same transition, the points in  $m1$ - $m2$  space are much closer together, although they still may not coincide exactly. There are several reasons for this, including tiny variations in the actual sound and mouth shape and small errors that could be introduced when identifying the vowel position by hand.

It can be shown that there exists a unique affine transformation that maps one triangle to another. We wish to find the transformation  $T$  that maps the vertices of the triangle U with mean points  $p_1$  for /oo/,  $p_2$  for /ee/, and  $p_3$  for /ah/ to the triangle V with vertices (1,2), (-1,2) and (0,0). Using homogeneous coordinates, we wish to find the affine transformation

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ 0 & 0 & 1 \end{bmatrix} \text{ such that } Tp_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, Tp_2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}, Tp_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Let

$$p_1 = [a, b]^T, p_2 = [c, d]^T, p_3 = [e, f]^T$$

$$A = \begin{bmatrix} a & b & 1 \\ c & d & 1 \\ e & f & 1 \end{bmatrix}, \delta = \text{Det}(A) = ad + be + cf - bc - de - af$$

$$A^{-1} = \frac{1}{\delta} \begin{bmatrix} d-f & f-b & b-d \\ e-c & a-e & c-a \\ cf-de & be-af & ad-bc \end{bmatrix}$$

Then

$$[t_{11} \ t_{12} \ t_{13}]^T = A^{-1} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \frac{1}{\delta} \begin{bmatrix} b+d-2f \\ a+c-2e \\ be+de-(a+c)f \end{bmatrix}$$

$$[t_{21} \ t_{22} \ t_{23}]^T = A^{-1} \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix} = \frac{1}{\delta} \begin{bmatrix} 2(d-b) \\ 2(a-c) \\ 2e(b-d)+f(c-a) \end{bmatrix}$$

The mean and bounding box were subjected to the matrix transformation and charted (see figure 5). Subjecting the vertices of a polygon (e.g. the bounding box) to a matrix transformation and then drawing the connecting lines will produce a polygon identical to the one that would be formed if every point on the border of the original polygon was subjected to the transformation. The translated chart shows that the matrix transformation produced the points (1,2), (-1,2), and (0,0) as expected, and change the shape of the bounding box.

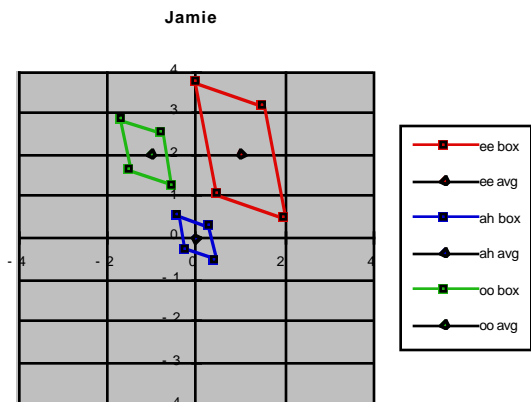


Figure 5: The chart from figure 4 after being subjected to the matrix transformation.

### The /m/ and /n/ Region

In order to predict the appropriate mouth opening for /m/ and /n/, we adjust the value based on its velocity.

Let the original track in moment space be T. Create a new track, T', by adjusting each value of T by a constant factor of the derivative of T. (The constant factor is not necessarily the same for the x component and the y

component.)

Let  $M$  be a function on  $T$  whose values are the mouth opening parameters constructed as follows: For each value in  $T$ ,  $M$  is the value read from the mouth opening predictor surface for the point in  $T$ , plus a fraction of a correction factor. The correction factor is the difference between the mouth openings read from the predictor surface at the current points in  $T$  and  $T'$ , plus a fraction of the correction factor from the previous iteration. The correction factor thus provides an exponentially decaying adjustment to mouth opening based on the velocity-adjusted track through moment space.

#### 4. RESULTS

##### Vowels

The points in  $m1-m2$  space for /ee/, /ah/, and /oo/ are well separated in all of the subjects. This indicates that there is some hope of creating a completely automated method for identifying and discriminating between these vowels in a stream of uninterrupted speech. Furthermore, the relative positioning of the vowels appears to be roughly the same between subjects, implying that the mapping onto a standardized  $m1-m2$  space will not produce any grotesque distortions in the original  $m1-m2$  space. The overlapping color regions in a plot of all points for all speakers (see figure 6) shows the need for a mapping onto a standardized  $m1-m2$  space because a single point in near the center of the chart could represent a viseme formed from any of the vowels.

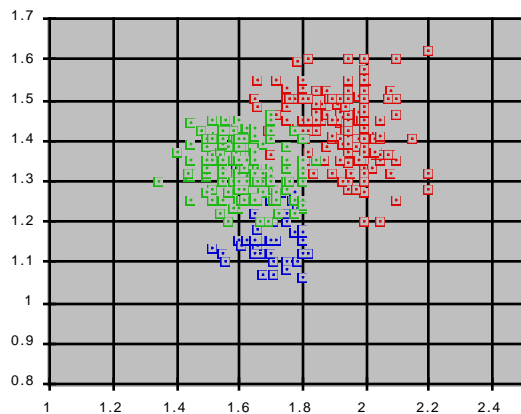


Figure 6: A plot of all points for all speakers before transformation

An overlay of all of the original charts (i.e. before applying the transformation matrix) was created (see figure 7). The bounding boxes were filled in red (/ee/), green (/oo/), and blue (/ah/), the same colors that the outlines of the bounding boxes used. The many overlapping regions of different colors indicate that a viseme predictor surface formed from the  $m1-m2$  space is speaker dependent.

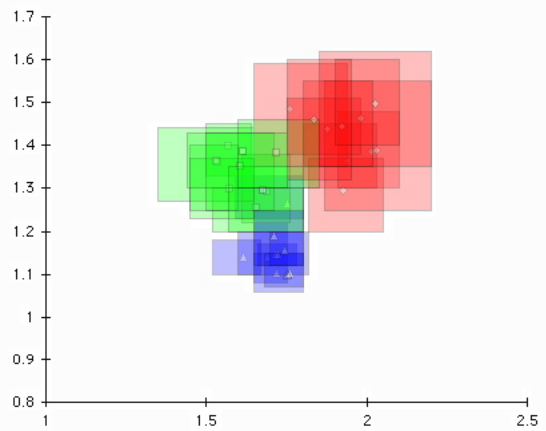


Figure 7: An overlay of all of the original charts before transformation.

Another overlay was created from all of the transformed charts (see figure 8). In this overlay there are areas where different colors overlap, but these areas are much smaller than in the first overlay, and they appear only in the areas between the mean points for the vowels. Also note that the bounding boxes' corner points are not necessarily in the set of data points which the box is bounding, i.e., the bounding box is not necessarily the convex hull of the data points. A viseme predictor surface formed from this  $m1-m2$  space should provide results that, while not perfect for all speakers, are within an acceptable margin of error (i.e. the errors will not be noticed by a casual observer).

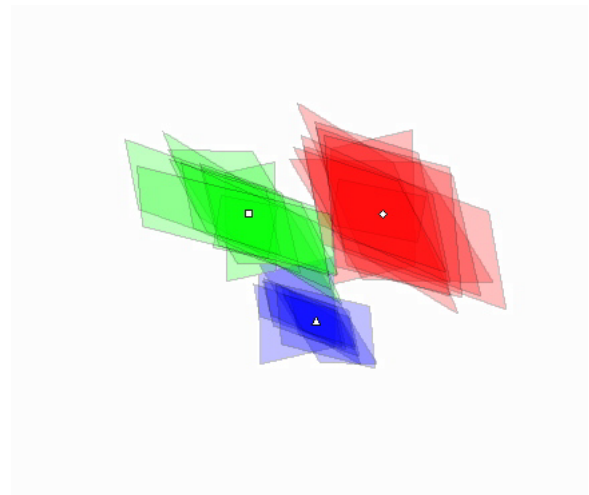


Figure 8: An overlay of all of the transformed bounding box charts.

Figure 9 shows a plot of all points for all speakers after transformation. The two /ah/ points (in blue) which encroach on the /oo/ territory (in green) are from "oowah" from two different speakers. The mouth shapes for those two points are expected to be very close to those of the surrounding /oo/ points because of coarticulation. The

large areas covered by each vowel reflect the large variation in mouth position over the samples representing a variety of transitions.

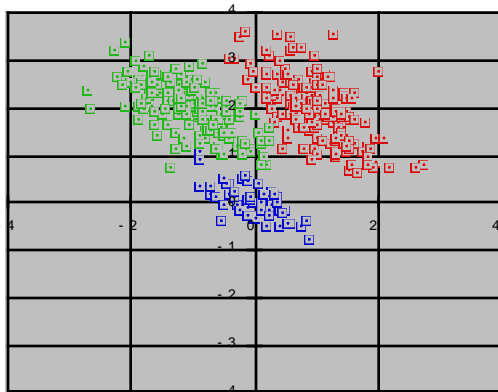


Figure 9: All points for all speakers after transformation.

One quantitative measure of the success of the transformation is a measure of the separation between the three vowel regions. The data points in a vowel region are dispersed throughout the region even for a single speaker. There are several contributing factors to this dispersion, including input signal noise, different vowel shapes due to coarticulation, and variations of mouth shape for the exact same vowel. The transformation cannot improve separation beyond that exhibited by a single speaker.

One measure of the separation between vowel regions is the ratio of the variance of the data points in one region to the variance of the data points to the other two regions. For three regions, there are six ratios. The six ratios were computed before transformation, after transformation, and using each speaker's center points for their own vowel regions (as opposed to the global average). The ratios were summed to produce one number for each of the three calculations. Comparing the square root of the sum of the ratios before and after transformation reveals that the transformation increased the separation between the three vowel regions by approximately 12%. The square root was taken to convert from a measure which grows with the square of the distance to one that is related linearly to distance. Comparing the ratios taken before transformation and using each speaker's own center points shows that the maximum possible improvement in separation between the three vowel regions is approximately 47%. The separation produced by the transformation is approximately 26% of the best possible improvement.

The next step in the lip synchronization process is to create the viseme predictor surfaces using the results in

figure 9. A general outline of the process will be given. Predictor surfaces are constructed for each mouth parameter. Details can be found in [3]. For example, the horizontal lip opening surface would be high in the /ee/ area (red), moderate in the /ah/ area (green), and low in the /oo/ area (blue), with smooth slopes between the areas.

### The /m/ and /n/ Region

In the figures below, the original moment track (T) appears in blue, the velocity-adjusted moment track (T') appears in red, and the mouth opening function (M) appears in green. Note that T and T' are in moment space (with values as indicated on the axes), while M is mouth position as a function of time with smaller y-values indicating mouth closure (scaled so that it appears on the same axes as T and T').

In figure 10 we see the mouth opening close rapidly to /m/, whereas in figure 11 the mouth opening closes less (and less rapidly) to /n/. We expect that this will hold for all transitions to and from /m/ and /n/. Further experiments to show this are in progress.

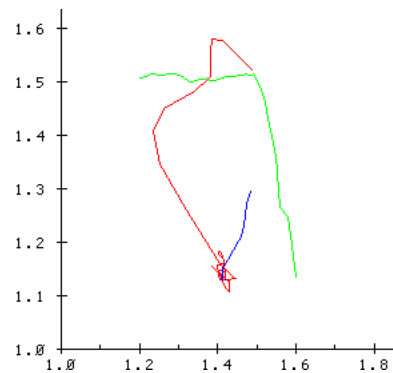


Figure 10: The original moment track (T, blue), velocity-adjusted moment track (T', red), and mouth opening function (M, green) for "ahm".

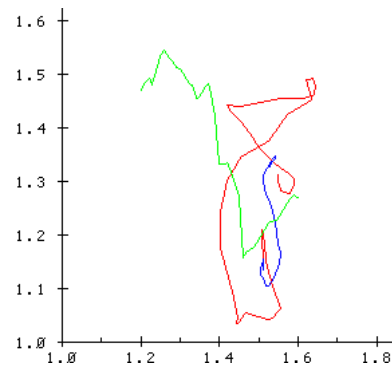


Figure 11: The original moment track (T, blue), velocity-adjusted moment track (T', red), and mouth opening function (M, green) for "ahn".

## 5. CONCLUSIONS

The results show that an affine transformation appears sufficient to map an arbitrary speaker's points for the vowels /ee/, /ah/, and /oo/ in  $m1-m2$  space onto a standardized  $m1-m2$  space so that a viseme predictor surface constructed from the standardized  $m1-m2$  space will produce the correct viseme for different speakers, at least for the three vowels tested. We are in the process of testing the ability of the affine transformation created using the three vowels to correctly transform the complete set of training sounds (/aa/, /ee/, /oo/, /ey/, /ae/, /o/, /z/, and /zh/) for each speaker. Also, constructing the matrix transformation is simple once the average position of the speaker's points in  $m1-m2$  space for each of the vowels /ee/, /ah/, and /oo/ are known. While it is ultimately desirable to create a completely automated method to compute these values from a stream of continuous speech, it is sufficient to obtain samples of the speaker saying each vowel in isolation.

The figures of mouth opening for "ahm" and "ahn" show that /m/ and /n/ can be distinguished in at least some cases by the technique described. Experiments on a wider range of "words" and speakers are ongoing at the time of this writing.

For an "original" version of this paper with color, please contact jltaylor@eos.ncsu.edu

## 6. REFERENCES

- [1] D. McAllister, R. Rodman, and D. Bitzer, "Lip Synchronization as an Aid to the Hearing Impaired", *Proceedings of the American Voice Input/Output Society*, 1997, pp 233-248.
- [2] D. McAllister, R. Rodman, D. Bitzer, and A. Freeman, "Toward Speaker Independence in Automated Lip-Sync", *Proceedings of Compugraphics '97*, 1997, pp 10-15.
- [3] C. Krothapalli, D. McAllister, R. Rodman, D. Bitzer, M. Wang, and J. Taylor. Predictor Surfaces for Lip Synchronization Animation of Voiced Input. *Proceedings of the 5<sup>th</sup> World Multiconference on Systemics, Cybernetics and Information (SCI 2001) and the 7<sup>th</sup> International Conference on Information Systems Analysis and Synthesis (ISAS 2001)*. Skokie, IL: International Institute of Informatics and Systemics, 2001. (This reference)
- [4] J. Harrington and S. Cassidy, *Techniques in Speech Acoustics*. Dordrecht: Kluwer Academic Publishers. 1999.
- [5] R. Rodman, D. McAllister, D. Bitzer, H. Fu, B. Xu, "A Pitch Tracker for Identifying Voiced Consonants". *Proceedings of the 10<sup>th</sup> International Conference on Signal Processing Applications and*

*Technology* (ICSPAT'99). November, 1999.

- [6] Jamie Taylor, *Speaker Independence in Lip Synchronization of Vowels*. M.S. Thesis. Department of Computer Science, North Carolina State University, Raleigh, NC 27695. 2000.
- [7] J. Taylor, D. Bitzer, R. Rodman, D. McAllister, "Achieving Speaker Independence in Automatic Lip Synchronization", *Proceedings of the American Voice Input/Output Society*, 2001, pp 163-171.
- [8] Bowei Xu, *Segmentation and identification of Voiced Fricatives and nasals through Spectral Moments*. M.S. Thesis. Department of Computer Science, North Carolina State University, Raleigh, NC 27695. 1999