

# Automated lip-sync animation as a telecommunications aid for the hearing impaired

David McAllister Robert Rodman Donald Bitzer Andrew Freeman

Dept. of Computer Science  
North Carolina State University  
Raleigh, NC 27695-8206 USA  
rodman@csc.ncsu.edu

## ABSTRACT

Vocal communication is most effective when the listener is able to observe the mouth of the speaker. This is especially true for the hearing impaired, and dramatically true for the deaf, who rely on lip-reading for comprehending speech. Communication over telephone lines is particularly onerous for the hearing impaired as visual information is unavailable. Our research addresses that problem by providing a computational means of taking speech as input and producing an animated mouth as output that moves precisely as if it were articulating the speech.

In this paper we continue reporting on our progress in using moments of spectra — a measure of spectral shapes — to provide a direct mapping from the speech signal to parameters controlling the shape of the lips and position of the jaw during the articulation of the speech. The method requires no text nor does it rely on any form of speech recognition. We report in particular on the progress we have made in distinguishing the visemes — the visible phonemes — corresponding to /m/ and /n/.

## I. INTRODUCTION

Vocal communication is most effective when the listener is able to observe the mouth of the speaker. This is especially true for the hearing impaired, and dramatically true for the deaf, who rely on lip-reading for comprehending speech. Communication over telephone lines is particularly onerous for the hearing impaired as visual information is unavailable. Our research addresses that problem by providing a computational means of taking speech as input and producing an animated mouth as output that moves precisely as if it were articulating the speech.

The use of “video-phones,” should they ever become available, would not obviate the technology we are developing. Video requires tremendous bandwidth, especially at resolutions suitable for lip-reading. Our method requires merely the transmission of facial parameters, which are data easily managed by currently existing telephony. The actual animation of the mouth is created at the receiving end of the phone call.

The methods we use for lip-synching require neither text nor a speech recognition interface. Rather, the speech signal is processed directly and correlated to parameters that control the precise movements of the lips and jaw of an animated mouth. This makes the method suitable for telephone communication, where text is not available, and speech recognition is dubious. Moreover, the method is, in principle, both speaker and language independent. In an actual telephone application, the speech signal would be received and processed, producing a short delay, after which both speech and animated mouth would be presented synchronously to the listener.

The caller would be advised automatically that a lip-synching system was in use, and to expect a short delay in receiving response from the person called. This is a mild inconvenience, but no worse than the delays experienced during the earliest days of international telephone calls transmitted by satellite.

Our method of signal analysis attempts to predict mouth positions directly from the speech signal that produced it without phoneme recognition. The process involves normalizing the signal based on its frequency and then analyzing the shape of the resulting spectrum by treating it as a probability density function and taking statistical moment values for single glottal pulse segments. We have shown that two of these moment functions are sufficient to isolate most voiced sounds to the degree necessary for mouth shape identification. The nasals have proved to be illusive but we will describe some of the techniques we have employed to distinguish them.

Our system requires a training procedure in which the moments identified for each sound segment are associated with a set of lip and jaw parameters needed to form that sound. Taking the moment values as independent variables, and the mouth parameters as dependent variables, we compute a set of surfaces that can later be used to identify mouth positions for new utterances. This process has produced credible facial animations for original utterances based on a limited set of training sounds.

Our success with vowels has been reported earlier [2] - [6]. In this paper we report our progress in the historically difficult problem of separating the nasal consonants, especially distinguishing the bilabial /m/ from the alveolar /n/.

## II. MOMENT CALCULATION

We briefly review our procedure although it has been described elsewhere in more detail [5]. Our sound is captured at 22 KHz, using 8 or 16 bit quantization in WAVE or Mathematica format. A glottal pulse, GP, is identified by minimizing the sum of the first 4 odd harmonics of power spectra computed over increasing sample sizes. This minimum occurs at 2\*GP. We then average the harmonics of power spectra of many sliding samples the same size as the GP to reduce noise. We set the spacing of the harmonics of the average to the fundamental frequency, 22050/GP, clip at 4 KHz, compute the cube root to deflate the influence of the first formant and interpolate with a cubic spline to produce  $S(f)$ . We divide  $S(f)$  by the mass, given as equation (1), to convert  $S$  to a probability density function  $P$  (equation (2)). The two variables we use to identify visemes are the first moment, which is the mean, and given as equation (3); and the second central moment, which is the variance, given as equation (4).

$$\text{mass} = \int_0^{4000} S(f) df \quad (1)$$

$$P(f) = S(f) / \text{mass} \quad (2)$$

$$m_1 = \text{mean} = \int_0^{4000} f * P(f) df \quad (3)$$

$$cm_2 = \text{var} = \int_0^{4000} (f - \text{mean})^2 * P(f) df \quad (4)$$

The resulting moments have very little noise and are pitch independent. (We have tried to introduce the third central moment as an additional variable to help in distinguishing visemes but in the region of interest the three variables are approximately coplanar and hence the third central moment is of little help in providing additional separating power.)

## III. THE NASALS /m/ and /n/

Distinguishing between /m/ and /n/ is a classical problem of acoustic phonetics with a 50 year history of

research well summarized in [1]. Nearly all of the work in this area supports our thesis that the information needed to make the articulation distinction is to be found in the shape of the various spectra of the nasal consonant itself as well as surrounding sounds, including especially the transitions to and from the nasal. Phoneticians have known for years that the location of the anti-formants — frequency bands of lower energy than surrounding frequencies — differ between /m/ and /n/. These differences are reflected in the shape of the various spectra. Phoneticians have also observed that the starting frequency of the second formant provides a cue for distinguishing [m] from [n], again a matter of spectral shape. Further, Harrington claims to have “...shown that a highly coarticulated part of the speech signal, encompassing the transition between the consonant and the vowel, provides some of the most salient cues to the place of distinction in nasal consonants.” [1, p 32] Earlier in the same work Harrington states that “...classifications from combined spectra are most closely related to categorizations from running spectra, in which not only the changing spectrum, but also the actual shape of successive spectra contribute separately useful information to the phonetic identity of the segment.” [1, p 31]

The plots in moment space that we use to determine visemes take into account the “shape of successive spectra.” Our research has focused on two problems. One is parameterizing spectral shapes, for which we are pursuing the use of moments. The second is characterizing the path of successive spectra in whatever moment space we are using. The technique has proven itself for vowels, for approximants such as English /r/ and /l/, and somewhat for voiced English fricatives. (We are still working on distinguishing between /v/ and /dh/ — the medial consonants of *level* and *bother*.) Distinguishing the place of articulation for English nasals *in all contexts* is an important challenge to our approach because /m/ and /n/ are acoustically similar, yet correspond to distinctive visemes.

We find that the nasals as a group are distinguishable from other English sounds. The value pairs for all of the nasals for a given speaker lie in their own region of a two-dimensional moment space in which the x-axis is the mean (first moment) and the y-axis is the variance (second central moment). The nasals themselves, in particular /m/ and /n/, are not distinguishable individually by our standard mean and variance approach.

Our solution to this problem, as suggested above, is to investigate the shape *change* of the normalized spectrum relative to the shape of the adjoining vowel(s) to help identify the behavior of the antiformants. This relative

behavior is examined using the spectrum of the vowel as a *divisor*. For the three speakers we have examined, we are able to distinguish between the nasals /m/ and /n/ in over 90% of the cases, the principal exceptions being when transitioning from the nasal to /i/. In most other cases the path shape together with its initial and/or final position in the mean-variance moment space have been sufficiently different to distinguish the two nasals, although the behavior can be radically different for each speaker. Our approach is as follows:

i) Compute a normalized spectrum for the preceding (or succeeding) vowel and interpolate it using a spline. We have been using the cubic spline provided by Mathematica.

ii) After normalizing the spectrum associated with each glottal pulse as described in (i), and interpolating it, we divide by the normalized vowel spectrum. We note that the DC term in both cases is zero but the limit of the ratio exists from the right assuming that the degree of the polynomial interpolating spline used for the numerator is less than or equal to that used for the denominator.

iii) We compute the moments for the ratio in the same way we compute moments using equations (1) - (4).

The mean and variance of the unit function on the interval [0, 4000] are 2000 and  $1.3333... \times 10^6$ . In all the plots shown in this paper, we have scaled the mean (x-axis) by dividing by  $10^3$  and the variance by dividing by  $10^6$ . Below we give examples of the paths produced by different nasal/vowel and vowel/nasal transitions for three speakers, two males and a female for the transitions /mu/, /nu/ and /um/, /un/, where /u/ is the vowel in English *true*.

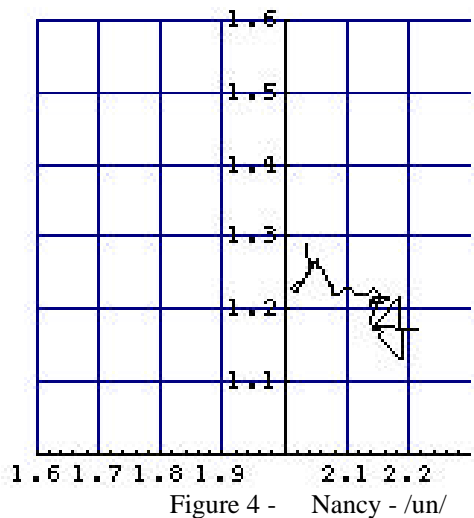
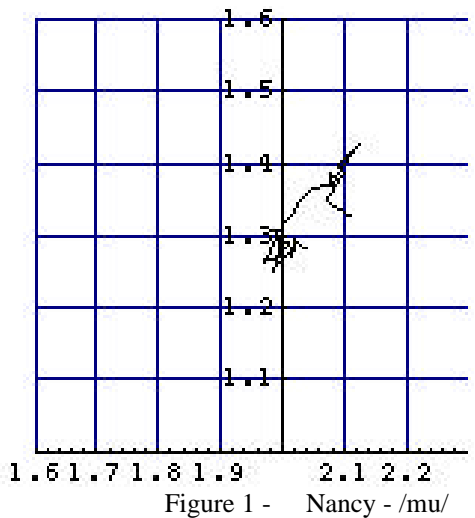
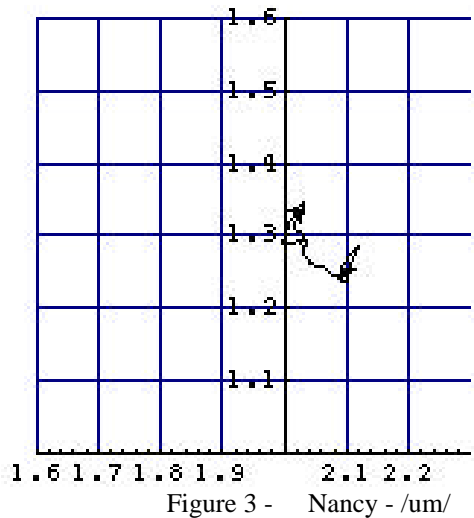
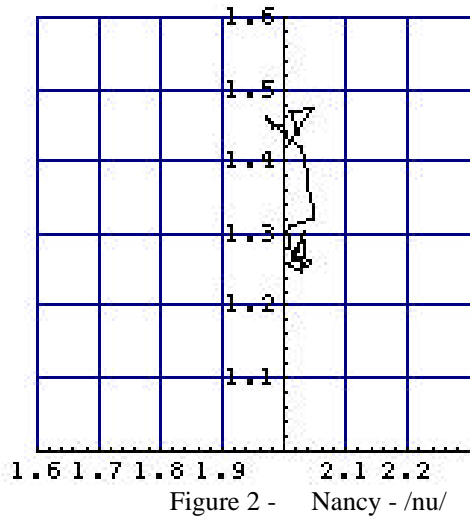




Figure 5 - Dave - /mu/

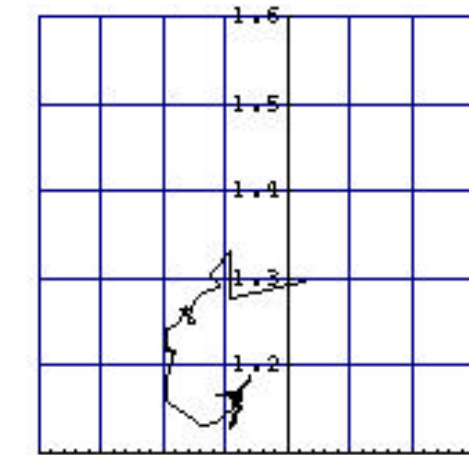


Figure 8 - Dave - /un/

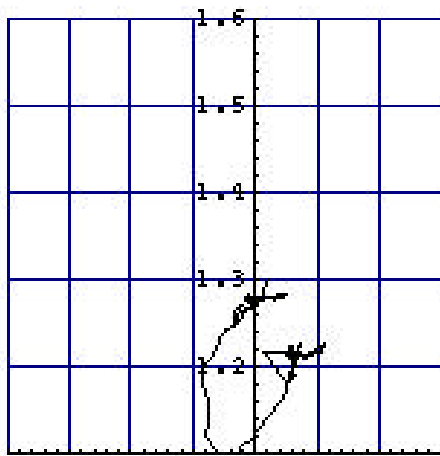


Figure 6 - Dave - /nu/

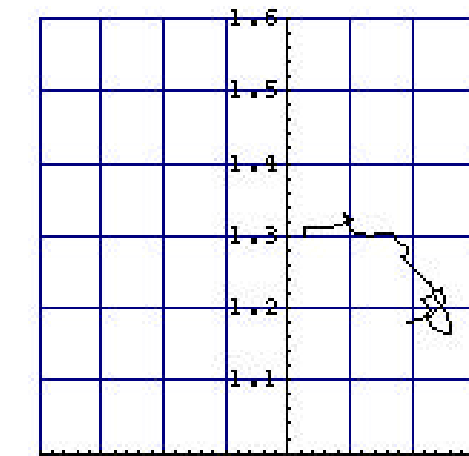


Figure 9 - Robert - /mu/

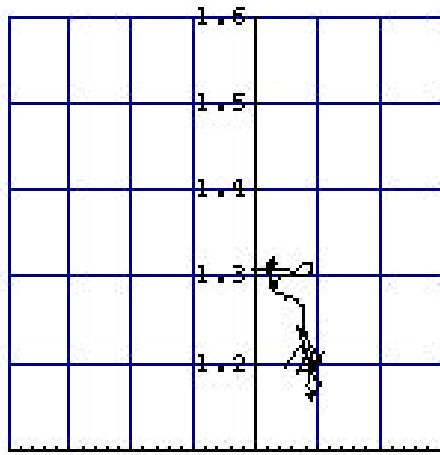


Figure 7 - Dave - /um/

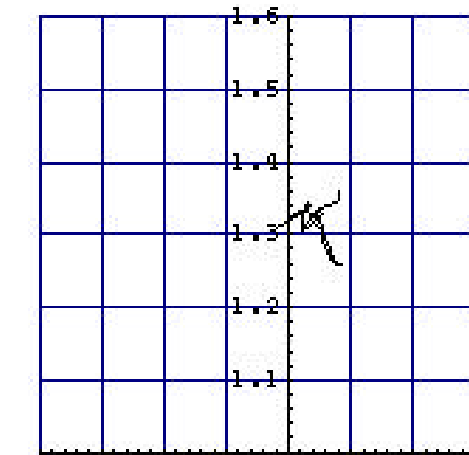


Figure 10 - Robert - /nu/

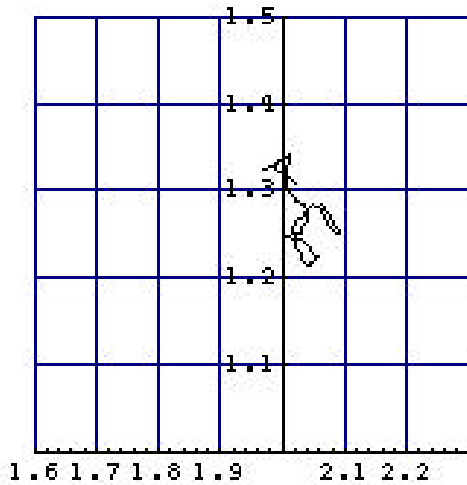


Figure 11 - Robert - /um/

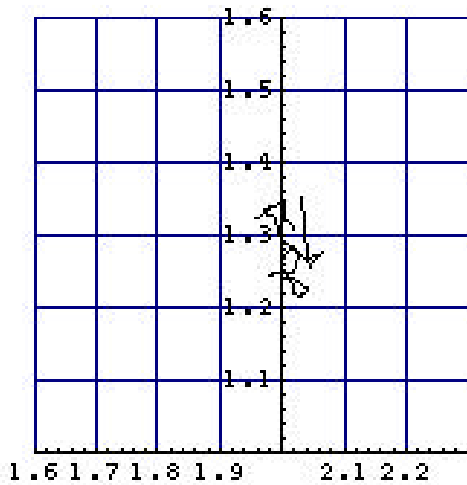


Figure 12 - Robert - /un/

The direction of the path is determined as follows. If the sound begins with a vowel, the path begins very near the point (2.0, 1.3) and usually terminates away from that point. If the sound ends with a vowel, the path terminates in the vicinity of (2.0, 1.3) and usually originates away from that point.

For each speaker there is a significant difference between the path shapes of /m/ and /n/ in the various contexts, with the most significant difference being the degree and direction of curvature. For example in Figures 5 and 6, we note that Dave's /mu/ is approximately linear whereas his /nu/ has considerably more curvature. There are other measurable, significant differences as well. In all other cases the paths are different enough to distinguish between /m/ and /n/ except in Figures 11 - 12. Robert's /mu/, /nu/ paths are too close, and would produce an error in lip synching. We are currently examining methods for comparing the behavior of the ratios in various frequency regions.

Similar results for other transitions such as /ma/, /na/ and /am/, /an/ have been obtained, with the only exception being /mi/, /ni/, where /i/ is the vowel of English *me*. In these cases the high energy of the high, front, tense /i/ appears to drown out the subtle differences in anti-formants that would permit us to distinguish /m/ from /n/. We are currently working to solve this problem. Some relevant plots are given as Figures 13 and 14:

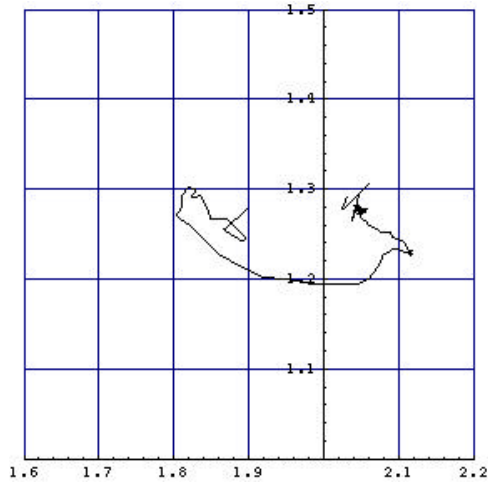


Figure 13 - Nancy - /mi/

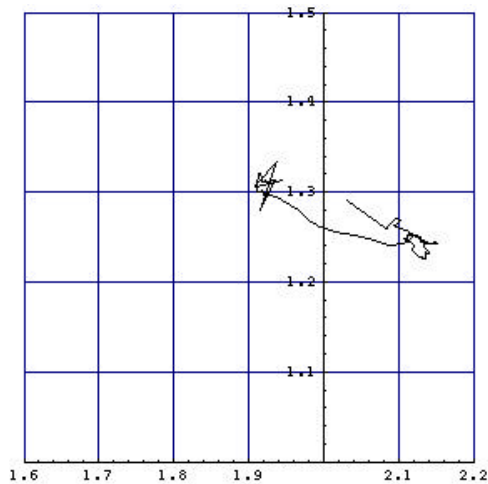


Figure 14 - Nancy - /ni/

There are major differences between speakers in the actual locus of the nasal transition paths. This is to be expected, and is consistent with previous research, which found a high degree of speaker dependence. Fortunately, there is a relatively high degree of speaker consistency, so the system can be trained by having each speaker utter a few key phrases that are used to calibrate the system.

Below we have examples of a second recording of one of the speakers for the /mu/, /nu/ transition cases. We see

that there is a sufficient consistency to allow the system to identify the utterance as a nasal consonant, and to then focus on the transition path to determine place of articulation. Whether such will be the case for the speaking population in general is an unanswered question, and considerably more study must be done before drawing any conclusions.

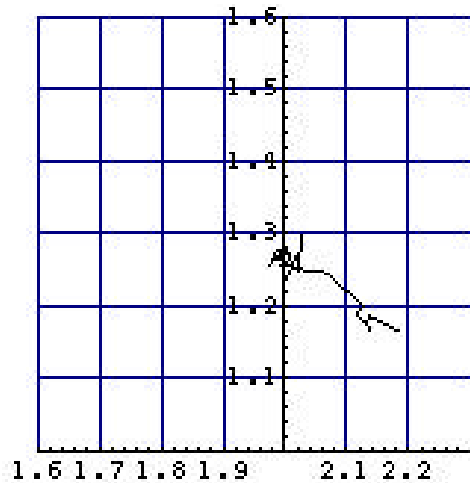


Figure 15 - Dave - /mu/ (2)

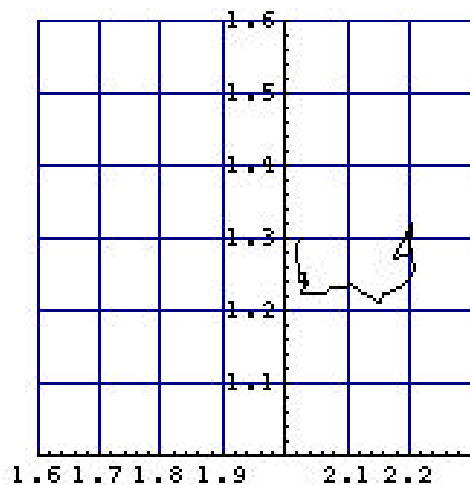


Figure 16 - Dave - /nu/ (2)

## REFERENCES

- [1] Harrington, J. "The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants." *Journal of the Acoustical Society of America*, vol. 96, No. 1, July, 1994, pp 19-32.
- [2] Koster, B. Rodman, R. and Bitzer, D. "Automated Lip-Sync: Direct Translation of Speech-Sound to Mouth Shape," *Proceedings of the 28th Annual Asilomar Conference on Signals, Systems and Computers*, IEEE:1994, pp 36-46.

- [3] 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 27695-8206, 1995.
- [4] McAllister, David F., Rodman, Robert D., Bitzer, Donald L. and Freeman, Andrew S., "Lip Synchronization for Animation," *SIGGRAPH 97 Visual Proceedings*, Los Angeles, CA, August, 1997, p 225.
- [5] 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, pp 233-248.
- [6] McAllister, David F., Rodman, Robert D., Bitzer, Donald L. and Freeman, Andrew S., "Lip Synchronization of Speech," *Proc. AVSP 97*, pp 133-136. Rhodes, Greece, September 1997.

**NB: References 4-6 may be found at URL = <http://www.multimedia.ncsu.edu/research/voiceio/>**