Linguistically Valid Movement Behavior Measured Non-Invasively

Adriano V. Barbosa[1], *Hani C. Yehia*[2], *Eric Vatikiotis-Bateson*[1]

[1]Department of Linguistics, University of British Columbia, Vancouver, Canada
[2]Department of Electronics, Federal University of Minas Gerais, Belo Horizonte, Brazil

adriano.vilela@gmail.com, hani@cefala.org, evb@interchange.ubc.ca

Abstract

We use optical flow to extract reliable kinematics from video for motions of the head, face, torso, and hands during speech and musical performance. Unlike dot- and marker- based measures, these *markerless* measures are non-invasive and require no *a priori* specification of measurement locations. Reliability is compared with marker tracking data and the method's utility is demonstrated for data from Plains Cree, English, and Shona.
Index Terms: optical flow, kinematics, non-invasive measures.

1. Overview

Since the mid-1990's [1], we have been keen to develop video-based tools for measuring spoken communication that would be computationally tractable, reliable, non-invasive, and not restricted to laboratory recording equipment and conditions. At that time, digital image processing was cumbersome and expensive, and everyone thought that video images had to be of the highest resolution possible in order to withstand fine-grained analysis (e.g., [2]). The technology has improved dramatically and we now know that the visible attributes of spoken communication tend to be ubiquitous, simple (e.g., linear), and accessible to perceivers at surprisingly low temporal and spatial resolutions [3, 4, 5]. Given these technical and conceptual advances, the time is ripe for video-based motion analysis tools that can be applied to inexpensively acquired video data.

In this paper, we describe our method for measuring motion from optical flow, test its reliability against 2D flesh point measures [6], and provide sample applications to linguistic performances. The tool, which is part of a larger Matlab toolbox for multimodal data analysis [7], is freely available to the research community and has already proved useful in the analysis of the coordination between speech acoustics, head and face motion, and manual gestures [8].

The text is organized as follows. Section 2 briefly describes the optical flow algorithm and introduces the concept of regions of interest, which are used in order to reduce the high dimensionality of the optical flow signal. The data acquisition and processing procedures are presented in Section 3. Results are discussed in Section 4, where motion measures derived from the optical flow field are presented and compared with those obtained through a video-based marker tracking algorithm for the same data. Lastly, the summary is presented in Section 5.

2. Extracting motion measures from video

The first stage in this process is to get the video data stored in files on the computer. This can be done in one step by recording video directly to the computer via hardware capture, or in two steps by first recording to tape. Once stored on the computer, the movie files can be accessed as image sequences.

The video data acquired in an experiment consist of an image sequence, in which each image (or frame) can be treated as an array of dimension $[M \times N \times 3]$, where M and N are the number of rows and columns in the image, respectively. For example, a *standard definition* (SDTV) frame of NTSC digital video is 480 pixels high by 640 pixels wide ($M = 480$ and $N = 640$). Color images are composed of three channels, though the contents of each individual channel depend on the color space being used. For example, in an RGB color space, the three channels represent the amount of red (R), green (G) and blue (B) in the image, respectively. Other representations use one luminance (with brightness information) and two chrominance (with color information) channels. One advantage of such a representation is that a grayscale version of the image can be easily obtained by simply discarding the two color channels, which results in a $M \times N$ matrix.

Optical flow is a common technique for extracting measures of 2D motion from video. Although there are many algorithms for computing optical flow [9], they all have the same goal of calculating *optical flow fields* corresponding to the projection of the 3D motion of real objects onto the 2D image plane. Roughly speaking, after conversion to grayscale, the algorithm compares consecutive frames of the video sequence and calculates how much and in which direction each pixel in the image moved from one frame to the next. The algorithm then assigns to each pixel a displacement vector corresponding to the difference in the pixel position across the two frames. The array of displacement vectors comprises the optical flow field.

The distance in time between consecutive frames (given by $T = 1/f$, where f is the video frame rate) can be used as a unit of discrete time. Thus, in discrete time, the pixel displacements that comprise the optical flow field can be seen as pixel velocities. Therefore, the optical flow at the point $(x,\ y)$ in the image plane at the discrete time k can be denoted by

$$\overrightarrow{v}(x,\ y,\ k) = [v_x(x,\ y,\ k),\ v_y(x,\ y,\ k)]\,, \qquad (1)$$

where $v_x(x,\ y,\ k)$ and $v_y(x,\ y,\ k)$ are the x and y components of the optical flow vector, respectively.

The optical flow analysis results in a high dimensional signal. For example, in the case of digital NTSC video, there are 640×480 two-dimensional vectors associated with every pair of consecutive frames in the video. There are many ways to reduce the dimensionality of the optical flow analysis that will make the resulting signal more tractable. For example, the image sequence may be filtered on input to reduce the resolution of the analysis, or pixel arrays may be sparsely sampled (e.g., 1 out of every 4 pixels) for analysis. After flow computation, the dimensionality of the result may be reduced by *Principal Component Analysis* (PCA). We have chosen to leave the input sequence alone, but to reduce the dimensionality of

the output to 1 by computing the magnitude of all 640×480 velocity vectors and then summing them for each frame step. This results in a scalar value per optical flow frame that corresponds to the total amount of motion in that frame step. In doing this, the contribution of individual moving components in the scene is, of course, lost. However, as we have shown previously for measuring speaker gestures [8], the temporal variation of this seemingly impoverished measure is surprisingly well-coordinated with time-varying measures made in other domains (for example, the RMS amplitude of the speech acoustics).

A less extreme version of this dimensional reduction is to assess the summed optical flow within specific *regions of interest*. This entails summing the optical flow results within certain regions of interest in the image (obviously, the summed optical flow over the entire frame can still be computed by simply defining a region of interest that encompasses all pixels). The summed optical flow for the n-th region of interest at the discrete time k is computed as

$$v_n(k) = \sum_{x=x_i}^{x_f} \sum_{y=y_i}^{y_f} \|\overrightarrow{v}(x,\, y,\, k)\|\,, \qquad (2)$$

where $\|\cdot\|$ denotes the vector magnitude, and x_i, x_f, y_i, y_f are the initial and final boundary positions of the region of interest in the horizontal and vertical directions, respectively. This procedure computes one scalar per region of interest and, when applied to the entire video sequence, results in one unidimensional signal for each region of interest.

It is also possible to define *regions of disinterest*, which are regions whose contents are zeroed. This is useful for defining regions of the image frame which should be ignored – for example, the incrementing time-code displays. If a region of interest and a region of disinterest overlap, the part that is common to both is ignored. Overlapping regions of interest and disinterest is an effective way of composing non-rectangular regions of interest. Figure 1 shows a frame taken from a video of a speaker telling stories in Shona (a Bantu language spoken in Zimbabwe). Figure 2 shows the associated optical flow frame with the regions of interest (solid line boxes) and region of disinterest delineated (dashed line box).

Figure 1: *A video frame showing a Shona story-teller.*

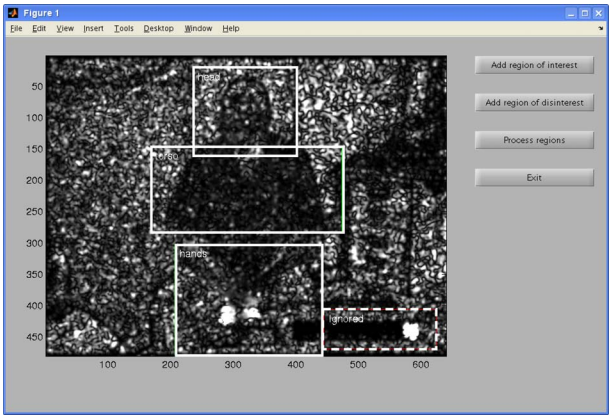

Figure 2: *Defining regions of interest and disinterest in the optical flow domain. Analysis regions for the head, torso and hands are boxed (solid line). Note the boxed (dashed line) region of disinterest enclosing the running clock at the lower right corner of the frame.*

3. Data processing and analysis

We applied the optical flow analysis described in the previous section to video sequences recorded originally for an audio-visual speech production experiment investigating production correlates of Lombard Speech [10]. In the experiment, a female was filmed for recitations of the 100 CID Everyday Sentences in quiet condition and in the presence of noise (for details, see [11]). Blue dots (see Figure 3) were placed on the speaker's face for subsequent 2D position tracking.

Figure 3: *A video frame recorded during sentence recitation in a quiet condition. Two-dimensional face and head motion were measured via tracking of the blue dots shown.*

The videos recorded during the experiment were transferred to a computer and saved as QuickTime movie files. Horn and Schunk's optical flow algorithm [12] was then applied to the video sequences (more specifically, the OpenCV [13] implementation of the algorithm was used). Discarding the color information, optical flow was computed only for the luminance component (brightness) of the video frames. The summed optical flow was computed for the entire frame and also for three regions of interest, namely the speaker's forehead, eyes and lower face (see Figure 4).

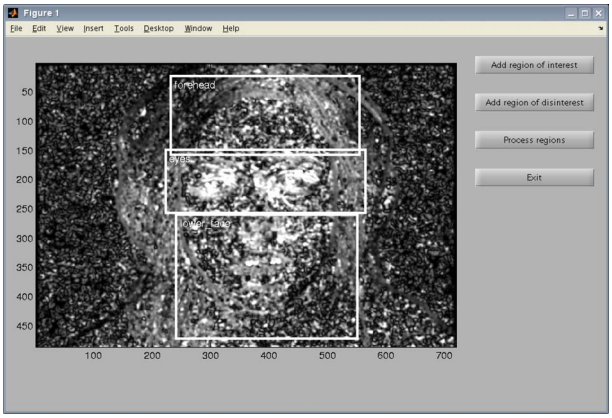

Figure 4: *Regions of interest related to motion of the head (upper box), eyes (middle box), and perioral region (lower box).*

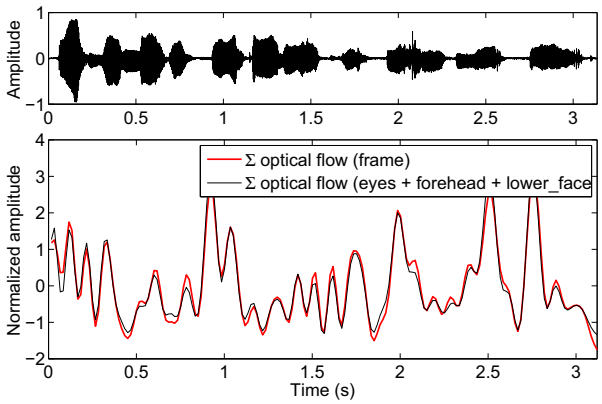

Figure 5: *Comparison between the summed optical flow computed over the entire image frame (thick/red line) and the combined summed optical flow computed over the three regions of interest (thin/black line).*

In addition to the optical flow data, 2D face motion measures were extracted from the video sequences using the video-based marker (blue dot) tracking algorithm developed previously ([6] which also describes the validation of the 2D measures). The recovered marker positions returned by the algorithm were visually inspected for errors in the marker tracking [There is one instance of marker mistracking for a few frames. Since the effect of including the erroneous value is undetectable in the current analyses of the marker tracking data, the sentence has not been excluded.]

Because optical flow corresponds to velocity, not position, comparison (validation) with the measures recovered from the marker tracking data requires that time derivatives be computed for the 2D marker positions, giving marker velocities. Then, since the optical flow measures consist of the sum of the magnitudes of all pixel velocities inside each region of interest, the magnitudes of the velocity vectors for all markers must also be summed. This results in a single value of the *summed marker flow*, or simply *marker flow*, for each frame, which can then be compared frame-by-frame to the summed optical flow.

4. Results and discussion

Here we compare the two measurement techniques. For demonstration purposes, only one sentence (#26) of the 100 CID sentences recorded in the quiet condition is used, *"Those windows are so dirty I can't see anything outside"*. Although this is one of the longest sentences in the data set, with a duration of 3.09 s, utterance duration was not a factor in the analyses discussed in this paper. The tables (discussed later in this section) provide results averaged across all 100 sentences for the *quiet* and *presented-in-noise* conditions.

Before proceeding, we should mention that all comparisons have been made between normalized versions of the signals. The original amplitudes of the summed marker flow and the summed optical flow signals can be very different, since they are obtained by summing different numbers of velocity vectors (which depend on the number of markers, in the case of the summed marker flow, and on the size – and consequently the number of pixels – of the region of interest, in the case of the summed optical flow). Therefore, in order to make comparisons (and also visualizations) easier, all signals are normalized to *zero mean* and *unit variance*.

Figure 5 compares the summed optical flow computed over the entire frame to the summed optical flow computed over the three forehead, eye and lower-face regions defined in Figure 4. The two signals match almost perfectly with a correlation coefficient of 0.98. That is, the three regions of interest account for 96% of the variance recoverable from the entire frame. Similar results were obtained for every sentence in the data set. There are two implications of this that are important to note: 1) the three easy-to-define regions capture so much of the image variance that the size of the analysis space can be reduced with little loss of potential behavioral information; 2) obversely, due to the very small contribution of potential noise to the result, optical flow analysis can be reliably computed for the entire image without requiring the extra effort of defining specific regions of interest.

For the purposes of this paper, we pursue the first implication given above; namely, that we calculate the optical flow only for the regions of interest, without too much fear of lost information, and make reliable comparisons with the marker flow derived from the tracked positons of the blue dots on the speaker's face. Figure 6 shows the time-varying summed marker flow and summed optical flow for Sentence 26. The figure shows

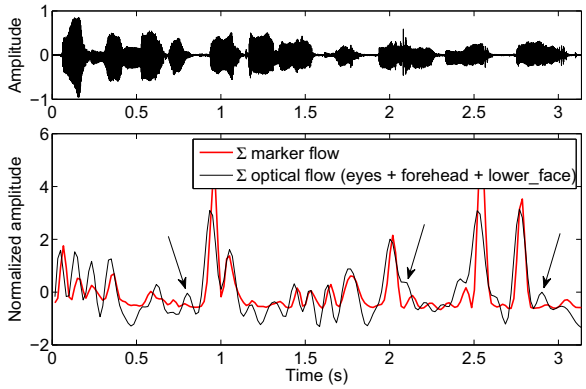

Figure 6: *Comparison between the summed marker flow (thick/red line) and the summed optical flow over the combined regions of interest (thin/black line). The peaks in the summed optical flow pointed by arrows correspond to eye blinks.*

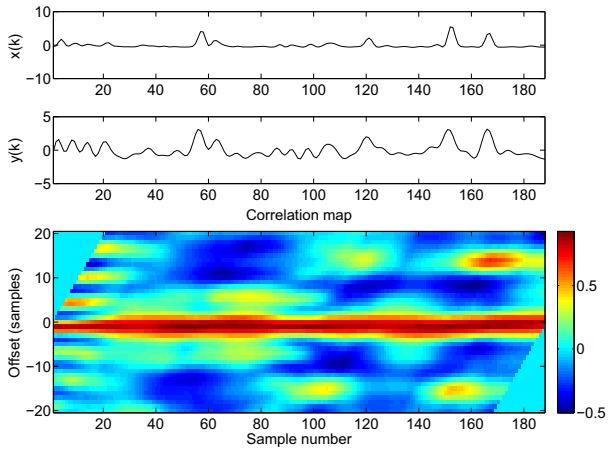

Figure 7: *The instantaneous correlation coefficient between the signals of Figure 6. Note the high correlation along offset zero.*

that for the most part the two signals match up very well, especially in their overall spatial and temporal synchronization. Where the signals do not match so well are indicated by arrows in the bottom panel of the figure. Checking these results against the original movie sequences shows that the extra peaks in the optical flow correspond to eye blinks, which show up in the optical flow signal as regions of high motion activity. As an aside, it is interesting to note how closely eye-blinks coincide with the speech-related elements of the head and face motion. The fact that eye blinks cannot be recovered from marker tracking at all and yet can be selectively included or excluded from the optical flow analysis exemplify a potential advantage of optical flow vs marker analyses. Figure 7 shows the instantaneous correlation [8, 14] between the signals of Figure 6 computed as a function of time and temporal offset. Again, the tight spatial and temporal coordination of the two signals is evident in the high-correlation band at offset zero (bottom panel).

In order to confirm that eye blinks were actually responsible for the signal mismatches in Figure 6, we eliminate the region of interest for the eyes and then redo the comparison of marker

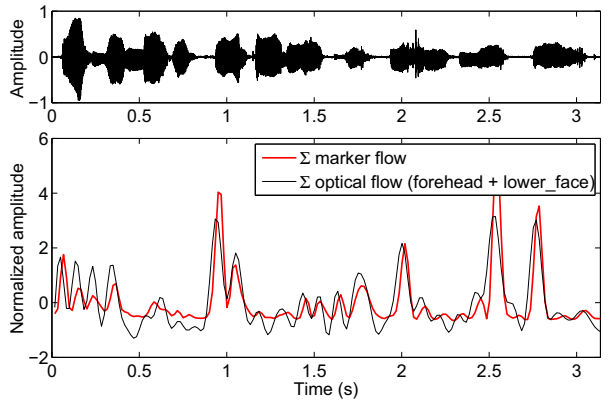

Figure 8: *Comparison between the summed marker velocities (thick/red line, no head correction) and the combined sum of optical flow calculated for the forehead and lower face regions of interest (thin/black line).*

flow and optical flow analyses. Figure 8 shows the summed marker flow and the optical flow summed for the forehead and lower face regions (but not the eyes). Comparing this to Figure 6, we see that the discrepancies (denoted formerly by the arrows) between the marker flow and optical flow signals are no longer present. This confirms that the discrepant peaks in Figure 6 are indeed due to eye blinks.

Finally, we compare how well three different measurement techniques – the PCAs of the 2D marker positions (Markers), the summed marker flow (M-flow), and the summed optical flow (O-flow) within a given region – perform when used in the estimation of cross domain correspondences between visible head and face motion, on the one hand, and acoustic properties such as RMS amplitude and LSP parameters (spectral correlates of vocal tract behavior), on the other.

The results, calculated for the full 100 sentence sets for each performance condition, are shown in Table 1. The correlation coefficient ρ, averaged across the 100 utterances for each production condition, is used to assess the cross-domain estimations shown in the table. In the top half of the table, *Line Spectrum Pairs* [15] extracted from the speech acoustics are mapped onto the three motion measures. Of the three measures tested, 2D marker measures (top three rows) are estimated the best at about $\rho = 0.80$. There is a slight improvement in estimation when head motion has been removed from the marker calculation (row 2 compared with row 1), but note that LSP can also estimate about 52% of the rigid body (3D) head motion (row 3). Comparing the two flow estimates for the summed velocities of the markers and of the image pixels, we see that neither is as good as the 2D marker estimates; however, the optical flow does better than the marker flow. Note that there seems to be no region-specific differences for either motion measure.

While it is clear that the multidimensional (PCA) 2D marker measures are better estimated (65%) from the spectral acoustics than are the uni-dimensional optical flow measures (45%), it is

Table 1: Correlation coefficients ρ between measured and estimated signals. The values shown are the means and the standard deviations for the 100 sentences of the CID data set for the quiet and presented-in-noise (noise) conditions. "M-" and "O-" preceding flow denote *marker* and *optical* flow, respectively.

LSP to motion measure	**ρ (quiet)**	**ρ (noise)**
Markers: Un-Corrected	0.79 (0.10)	0.80 (0.09)
Markers: Head-Corrected	0.82 (0.08)	0.83 (0.08)
Markers: Head (rigid body)	0.72 (0.14)	0.73 (0.14)
M-flow: All	0.61 (0.15)	0.58 (0.15)
M-flow: Head	0.63 (0.16)	0.62 (0.15)
M-flow: Perioral	0.61 (0.15)	0.58 (0.15)
O-flow: Head + Perioral	0.67 (0.15)	0.65 (0.15)
O-flow: Head	0.67 (0.15)	0.65 (0.15)
O-flow: Perioral	0.66 (0.15)	0.65 (0.15)
Motion measure to RMS	**ρ (quiet)**	**ρ (noise)**
Markers: Un-Corrected	0.79 (0.12)	0.82 (0.09)
Markers: Head-Corrected	0.78 (0.13)	0.81 (0.10)
Markers: Head (rigid body)	0.51 (0.14)	0.44 (0.15)
M-flow: All	0.25 (0.16)	0.29 (0.14)
M-flow: Head	0.18 (0.15)	0.19 (0.15)
M-flow: Perioral	0.25 (0.16)	0.29 (0.15)
O-flow: Head + Perioral	0.28 (0.18)	0.32 (0.15)
O-flow: Head	0.23 (0.17)	0.23 (0.15)
O-flow: Perioral	0.28 (0.17)	0.32 (0.15)

impressive how much acoustically relevant information that one motion dimension contains. This is not just a reflection of some high baseline for the estimation function because, as shown in the bottom half of the table, both types of motion flow fail miserably at estimating RMS amplitude, while 2D marker measures can estimate 60-65% of the RMS data.

From these results, it is obvious that there are advantages to using data of greater dimensionality to evaluate audiovisual speech behavior, nor was it ever our aim to argue otherwise. What we have shown though is that this totally non-invasive (and easy!) technique for estimating motion from optical flow does an excellent job of capturing the temporal correspondences and a pretty good job of the spatial ones as well. What is more, the ability to pick and choose which areas of motion to analyze *post hoc*, rather than being limited to marker positions chosen prior to recording, affords a much-needed flexibility even if it will be used only to inform future marker studies. Finally, when combined with the flexible algorithm for calculating and visualizing the spatial and temporal coordination of signals across a range of temporal offsets, the simpler optical flow method may provide the only feasible means for measuring the time-varying coordination of disparate structures such as the hands and body or the motions of additional speaker/performers.

5. Summary

The same analyses presented here have been applied to other performance contexts, including music and speech, with similar or, in the case of musical performance, better results. Examples include multi-camera recordings of Plains Cree narratives and two-person conversations, Shona singing and story-telling, and the incredible display of synchronization between Freddie Mercury and 72,000 British fans at Wembley Stadium in 1985. The ease of data acquisition and analysis associated with this methodology makes it possible to acquire massive amounts of useful, if not dimensionally detailed, data for assessing the time-varying coordination of biological behavior.

6. References

[1] E. Vatikiotis-Bateson, K. G. Munhall, M. Hirayama, Y. Kasahara, and H. C. Yehia, "Physiology-based synthesis of audiovisual speech," in *4th Speech Production Seminar: Models and Data*, P. Perrier, Ed. Autrans, France: ESCA, 1996, pp. 241–244.

[2] C. Kroos, T. Kuratate, and E. Vatikiotis-Bateson, "Video-based face motion measurement," *Journal of Phonetics*, vol. 30, no. 3, pp. 569–590, 2002.

[3] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, no. 3, pp. 555–568, 2002.

[4] K. G. Munhall, G. Jozan, C. Kroos, and E. Vatikiotis-Bateson, "Spatial frequency requirements for audiovisual speech perception," *Perception & Psychophysics*, vol. 66, no. 4, pp. 574–583, 2004.

[5] H. B. de Paula, H. C. Yehia, D. Shiller, G. Jozan, K. G. Munhall, and E. Vatikiotis-Bateson, "Analysis of audiovisual speech intelligibility based on spatial and temporal filtering of visual speech information," in *Speech Production: Models, Phonetic Processes, and Techniques*, J. Harrington and M. Tabain, Eds. London: Psychology Press, 2006, pp. 135–147.

[6] A. V. Barbosa and E. Vatikiotis-Bateson, "Video tracking of 2D face motion during speech," in *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology – ISSPIT'2006*, Vancouver, Canada, August 2006, pp. 791–796.

[7] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, "Matlab toolbox for audiovisual speech processing," in *Proceedings of the international Conference on Auditory-Visual Speech Processing 2007 (AVSP 2007)*, Hilvarenbeek, The Netherlands, 2007.

[8] ——, "Temporal characterization of auditory-visual coupling in speech," in *Proceedings of Meetings on Acoustics*, vol. 1, 2008, pp. 1–14.

[9] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, pp. 43–77, 1994.

[10] E. Lombard, "Le signe de l'lvation de la voix," *Annales des Maladies de l'Oreille, du Larynx, du Nez et du Pharynx*, vol. 37, pp. 101–119, 1911.

[11] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia, "Audiovisual lombard speech: Reconciling production and perception," in *Proceedings of the international Conference on Auditory-Visual Speech Processing 2007 (AVSP 2007)*, Hilvarenbeek, The Netherlands, 2007.

[12] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[13] "Open Source Computer Vision Library," http://www.intel.com/technology/computing/opencv/, accessed in July, 2008.

[14] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, "Algorithm for computing spatiotemporal coordination," in *Proceedings of the International Conference on Auditory-Visual Speech Processing – AVSP 2008*, Tangalooma, Australia, September 2008.

[15] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP," *Speech Communication*, vol. 5, pp. 199–215, 1986.