Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2007) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Suitability of a UV-based video recording system for the analysis of small facial motions during speech

Matthew Craig [a,b,*], Pascal van Lieshout [a,b,c,d], Willy Wong [b,e]

[a] Oral Dynamics Lab (ODL), Department of Speech-Language Pathology, University of Toronto, Rehabilitation Sciences Building, 500 University Avenue, Room 60, Toronto, Ontario, Canada M5G 1V7
[b] Institute for Biomaterials and Biomedical Engineering (IBBME), University of Toronto, Canada
[c] Toronto Rehabilitation Institute (TRI), Canada
[d] Department of Psychology, University of Toronto at Mississauga, Human Communication Lab (HCL), Canada
[e] Department of Electrical and Computer Engineering, University of Toronto, Canada

Received 23 September 2005; received in revised form 4 February 2007; accepted 10 April 2007

**Abstract**

The motion of the face carries great importance in research about speech production and perception. The suitability of a novel UV-based video recording system to track small facial motions during speech is examined. Tests are performed to determine the calibration and system errors, as well as the spatial and temporal resolutions of the setup. Results of the tests are further evaluated through kinematic data of the upper-lip, recorded from human speech, as this articulator typically shows the smallest movements, which would be the strongest test for any movement recording equipment. The results indicate that the current system, with a resolution slightly better than 1 mm, is capable of resolving the relatively small upper-lip motions during the production of normal speech. The system therefore provides an effective, easy-to-use and cost-effective alternative to more expensive commercial systems.


*Keywords:* UV video recording; System resolution; Visual speech; Kinematics

## 1. Introduction

Speech is both an acoustic and visual means of communication. The dynamic motions of a speaker's face can indicate how they physically produce speech, and can influence how that speech is perceived by others, for instance, when the hearing impaired use lip reading to understand speech.

Research involving facial motion during speech or other motor tasks uses a variety of systems to track face kinematics. The act of accurately tracking these motions can be essential in a wide range of studies, from motor speech disorders such as stuttering (Barlow et al., 1983), to speech or oral motor control in general (Clark et al., 2001; Dromey and Benson, 2003; Hasegawa-Johnson, 1998; Shaiman, 2002), or to the development of facial animations with realistic looking kinematics (Yehia et al., 2002).

While the general goal of these tracking systems is to achieve as high a resolution as possible, this aim may not be practical or even necessary when considering the compromise between the system's cost and the required resolution for the study at hand. For instance, sub-millimeter accuracy is generally not necessary when studying motions of the jaw, which are typically in the order of centimeters (Shaiman, 2002). In fact, because jaw motions are relatively large, researchers often do not even address the suitability of the tracking system to resolve them (Tasko and

* Corresponding author. Address: Oral Dynamics Lab (ODL), Department of Speech-Language Pathology, University of Toronto, Rehabilitation Sciences Building, 500 University Avenue, room 60, Toronto, Ontario, Canada M5G 1V7. Tel.: +1 416 946 8552; fax: +1 416 978 1596.
*E-mail addresses:* matt.simon.craig@gmail.com, matt.craig@utoronto.ca (M. Craig), p.vanlieshout@utoronto.ca (P. van Lieshout), willy@eecg.utoronto.ca (W. Wong).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.04.011

McClean, 2004; Recasense, 2002). If, however, the motions of interest are much smaller, such as that of the upper-lip during speech, the lower resolutions of less expensive systems may be inadequate.

The purpose of the current study is to evaluate the suitability of a novel two-camera ultraviolet (UV) system to resolve small motions of the face during speech. The system uses UV black-lights to illuminate glow-in-the-dark stickers on a subject's face, and tracks the motion of these markers using a specialized software suite, APAS, developed by Ariel Dynamics, Inc. (Trabuco Canyon, USA). This setting offers distinct benefits over other measurement systems because it uses unobtrusive passive markers and requires no special camera lenses (i.e., IR lenses) to record motion. This provides a cost-efficient opportunity (compared to other far more expensive commercial systems) to record unbiased facial motions related to speech and other oral motor functions in subjects of all ages. In addition, it can provide easy access to objective assessment of facial motion parameters in various clinical populations where facial expression is limited, as in Parkinson's disease (Simons et al., 2004).

## 2. Existing methods for tracking face kinematics

Strain gauge transducers, articulographs, and infrared video-based measurements are some of the more common techniques to measure face kinematics, all of which typically achieve resolutions of 1 mm or better.

Strain gauge transducers are composed of a head-rest with cantilevers. A bead is thread through each cantilever and attached to midsagittal points on the face. Motion of these points causes a strain on the cantilever, which is transduced and measured as a voltage. This voltage is used to calculate the kinematic signal (Barlow et al., 1983). Such systems are typically calibrated to 1 mm accuracy, and are thus suitable for the study of superior–inferior labio-mandibular motions, which have relatively large amplitudes (Clark et al., 2001; Dromey and Benson, 2003; Shaiman, 2002). The major drawback with this type of system is the restraint on head movements and the use of physical structures (cantilevers) that to some extent can interfere with articulator movements. This makes the system less suitable for use with, for example, very young children of patient populations where head motion is difficult to control.

Articulographs in the 2D version consist of a head mount and three transmitters that create an alternating magnetic field of different frequencies around a subject's head. Small transducer coils are fixed to a subject's face or vocal tract in the midsagittal plane.[1] The motion of articulators within the field induces a current in the transducers, which is inversely proportional to the cube of the distance between the transmitter and transducer (Hasegawa-Johnson, 1998). A time series of articulator motion is then derived from the strength of the induced current in each coil. Because these signals are derived from induced currents, articulographs have access to tongue motions, which cannot be captured by video systems that rely on reflected or emitted light to track such motions. Articulographs can have a relatively high spatial resolution (<.35 mm) (Van Lieshout et al., 2002). In addition, the articulograph hardware is standardized so that the system resolution should not vary significantly from location to location, but values between .35 mm and .7 mm have been reported (Van Lieshout et al., 2002; Zierdt et al., 2000). The main drawback of articulograph systems is the relative invasiveness of the technique. Sensors have to be attached with superglue to the surface of the tongue and gums. In addition, each sensor has a thin wire that to some extent may interfere with speech movements (e.g., Katz et al., 2006). These factors make this technique less suitable for young children, people with strong gag reflexes or patients with limited capability to open their mouth and extend their tongue.

Video-based systems typically consist of two infrared (IR) cameras with IR emitting lights positioned in front of the speaking subject. Circular reflectors are placed on the subject's faces at relevant points of interest. IR equipment provides a useful light source because visible light will not obscure the kinematic measurements. IR systems such as Qualysis, Elite or Optotrack have been shown to have resolutions between .1 and .5 mm (Green et al., 2001; Harris, 2004; Hertrich and Ackermann, 1997; Yehia et al., 2002). An additional benefit of using a video-based system is that, if the analysis requires the corresponding acoustic speech, it is automatically synchronized with the motion data when using an on-camera microphone.

Some distinction should be made between active and passive marker systems. Active markers, such as those used in Optotrack or articulographs, are small three-dimensional pellets fixed to the face or articulators of the vocal tract, with wires attached to transmit the necessary signals. Because of their weight and position, these pellets and wires may provide some intrusion or bias to articulatory motion, and can add unnatural weight when attempting to measure, for example, the motion of the lower-lip with a pellet attached to it. These factors may limit their use with very young children and subjects who resist being 'wired up'.

On the other hand, passive markers are wireless, two-dimensional objects, usually stickers, that act only to reflect light, thereby adding virtually no discomfort through size, weight, or wires. Qualysis is example of such a system that uses small circular IR reflecting stickers (Jiang et al., 2002). Passive markers are more amenable to younger subjects and patient groups with limited tolerance for wired markers. Most commercial video-based systems that use passive markers are fairly expensive and require special (IR sensitive) cameras, which could provide a financial barrier for

[1] 3D articulographs have recently become available, eliminating the requirement for midsaggital placement of the transducers (http://www.articulograph.de/).

Please cite this article in press as: Craig, M. et al., Suitability of a UV-based video recording system for the ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.04.011

their use in labs or clinical settings that have less money to spend on such equipment. In this paper, we wish to introduce a much cheaper solution that allows the use of regular consumer product camcorders with no special lenses and a relatively inexpensive software package (Ariel Performance Analysis System or APAS©) to automatically track the position of facial passive markers in combination with readily available UV light sources.

In the current study, the spatial and temporal resolution, system error, and calibration error of our passive marker, UV system together with the APAS software will be determined. Although similar glow-in-the-dark or UV systems have been used for speech research in other areas, this combination of different existing technologies has never been put to a test for studies involving kinematic motion tracking. If it indeed supplies sufficient resolution to measure small speech motions, our system provides an effective, cost-efficient and easy-to-use alternative to more expensive commercial video-based systems such as Qualysis and Optotrack.

A first system test will determine the average system error and error due to calibration throughout the typical viewing field of the cameras. The system error may further depend on the angle of the markers relative to the UV source. Changes in these angles, which will occur as the face moves, will affect the amount of light reflected off the markers. This is a problem that plagues both IR and UV systems, and will be investigated in a second test. The third and fourth tests will determine the spatial and temporal resolutions of the cameras. Finally, the sufficiency of these resolutions will be assessed through actual kinematic data, extracted from recorded human speech, to examine the suitability of the system to measure such motions.

## 3. Recording setup and equipment

In a typical speech experiment performed in our lab, subjects are seated in a chair in a dark room, illuminated by two 250 W UV black-lights. The UV lights are fixed to the ceiling of the room, pointing towards the subject, at a distance of approximately 2.5 m. Speech motions are tracked with the help of small glow-in-the-dark stickers, which are placed on the subjects' faces on points of interest (see Fig. 1). The stickers are circular with a diameter of approximately $3 \pm 0.5$ mm. In cases where the stickers do not adhere properly to the subjects (e.g., as is caused by short facial hair), a small point of glow-in-the-dark paint of similar size is applied as a substitute. On video, the paint dab is indistinguishable from a sticker, and provides the same results. Subjects are seated less than 1 m in front of the cameras. The two cameras are located just to the left-of-center and right-of-center of the subjects face to provide a three-dimensional view of facial motion. The illuminated stickers in the recorded video are tracked by the APAS software suite, described below.

During the recording, a reference point must remain visible to both cameras at all times which, in our case, is fixed to a plastic arm, held in place by a tripod. Before beginning each recording, the system requires calibration. To this end, we have developed a frame-only cube with 24 markers in pre-defined positions along the edges, which is recorded by both cameras (see Fig. 2). APAS notes that a minimum of six calibration points on the cube are required

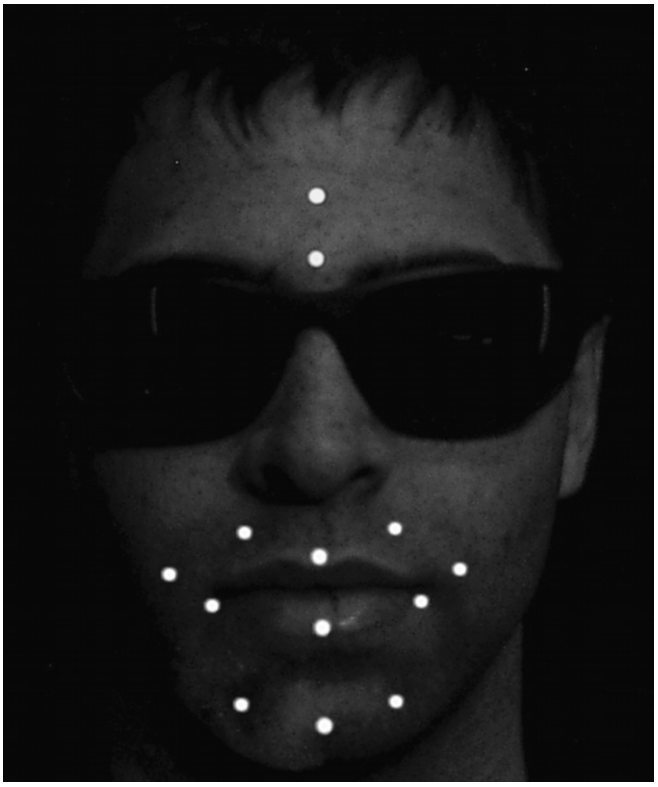

Fig. 1. The placement of passive markers of a subject's face.

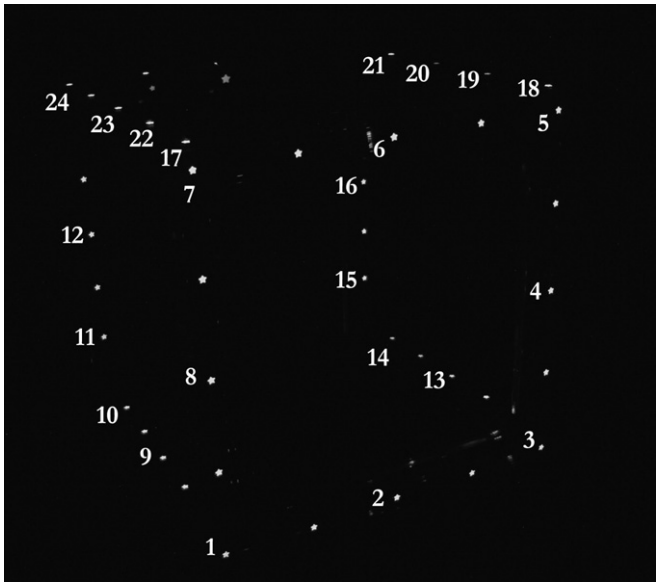


Fig. 2. The calibration cube as seen from the left camera.

Please cite this article in press as: Craig, M. et al., Suitability of a UV-based video recording system for the ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.04.011

for three-dimensional tracking. APAS uses these pre-defined calibration points and the fixed reference point to calculate absolute distances of the motions being tracked.

The two video cameras used in this study are JVC DVL9800U digital cameras with a shutter speed of 30 frames per second and a digital resolution of $720 \times 580$ pixels. Other standard camcorders can also be used, but these were the cameras that came with the APAS system at the time. This resolution compares extremely well to other camcorders, which have a typical vertical resolution between 250 pixels (standard analog camcorder) and 500 pixels (standard digital camcorders). The cameras have a native shutter speed of 30 fps and their standard recording mode uses *interlaced scan* (IS). IS splits each image into even and odd vertical fields and records them alternately, thereby recording 60 fps while halving the vertical resolution to 290 pixels. Every pair of fields (odd field and successive even field, recorded 1/60th of a second later) is combined into a single image, achieving a 30 fps overall rate. If there is movement during that 1/60th of a second time frame, this will show as jitter in the recorded image. The cameras also offer *progressive scan* (PS) mode which records 30 full frames per second, without jitter. Most recordings for this study were performed in IS mode. A comparison between IS and PS mode, in terms of system resolution, is given in Section 5.3. Because horizontal speech motions are much smaller than vertical ones (but not less important) it is preferable to have a higher resolution in the horizontal dimension to capture more subtle movements as shown, for example, by the degree of lip rounding for vowel productions. The current setup of the cameras also allows for lateral motion recordings, which play less of a role in normal speaking subjects, but could be an important diagnostic feature in patient populations.

As mentioned, the cameras are typically positioned 70–100 cm from a subject's face during recordings, and about 50 cm apart from each other. They must remain close to each other (i.e., 60 cm or less) to be able to visualize points on the outer cheeks of subjects.

## 4. Motion tracking

The recorded video from both cameras was processed by the APAS software suite. This suite, designed originally for human gait analysis, extracts the positions of illuminated markers from video streams, and can be readily adapted to record speech motions in the system described above. Three APAS modules are required to extract motion from the recorded videos: "Trimmer", which synchronizes the video streams from both cameras, "Digitize", which tracks the motion of the markers, and "Transform", which uses the system calibration and the output from "Digitize" to calculate the three-dimensional trajectories of the markers. The core module, "Digitize", locates the center of each sticker, based on pixel shading, in each video frame, using a pre-defined search area.

## 5. Equipment testing

The role of spatial and temporal resolution testing is to help understand, (1) whether the system is suitable to perform the measurements that are required for studying oral movements during speech production, and (2) which system configurations are required to achieve optimal results (i.e., camera-to-subject distance, placement of markers, illumination levels, etc.).

The testing apparatus for the tests described in Sections 5.1 and 5.2 consists of two stickers, attached to a ruler. The distance between the stickers was measured with a Vernier calliper to be $49.72 \pm 0.24$ mm apart. Methods similar to those described in tests 1 and 2 were used by Hertrich and Ackermann (1997) to determine the system error of their two-camera IR system.

### 5.1. Calibration error and system error in the viewing field

To determine the error due to calibration and the system error, i.e., the error due to random system noise, the ruler was recorded while being held static on a tripod in 18 locations within the viewing field. The locations, shown in Fig. 3, spanned near the perimeter of the cameras' views and central locations, at distances of approximately 70 cm and 100 cm from the cameras. In a speech experiment, these locations span the spatial extent in which a subject's head can typically be found. At each location, the ruler was held horizontally and recorded for approximately

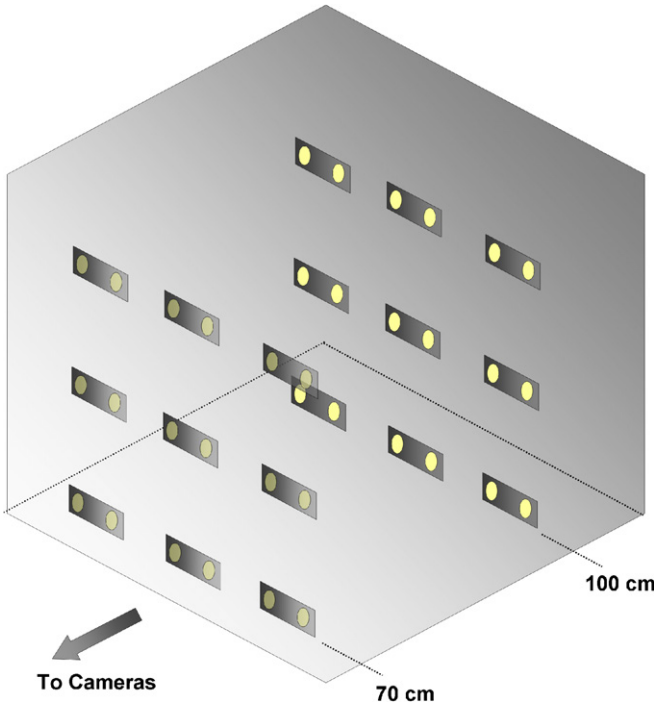


Fig. 3. The 18 recording locations. Angled lines represent the ruler orientation at each location. The placement of the two stickers is indicated on the upper left ruler and would remain constant at all locations. Diagram is not to scale.

Please cite this article in press as: Craig, M. et al., Suitability of a UV-based video recording system for the ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.04.011

3 s. From each 3-s recording, the mean and standard deviation of the distance between the markers was calculated.

Since the actual absolute distance between the markers does not change, the difference between the calliper-measured distance and recorded-digitized distance reflects the calibration of the system related to inaccuracies in the placement of markers on the cube. The mean digitized distance between the markers throughout the 18-s recording was 49.69 mm which compares extremely well to the calliper-measured distance of 49.72 ± .24 mm. This implies that the calibration error falls within the error of the callipers.

The standard deviation across the measured mean distance throughout the recordings can be considered a general index for the variation in calibration error around the viewing field. This variation could reflect a variety of factors, including small inaccuracies in the location of markers on cube, tracking errors in APAS software and slight deviations in positional accuracy for different viewing field positions. The deviation was .67 mm, which equals a relative calibration error of 1.35% (compared to a mean distance of 49.72 mm between marker positions).

The average random system error is determined by averaging the standard deviations of all of the 18 3-s recordings. Because the ruler was static at each location, the local SD value is independent of calibration errors. Average system error amounted to .13 mm. This system error may be taken as a lower limit to the system's resolution.

The average system error was found to be uncorrelated to horizontal or vertical position within the viewing field, but it was lower for the nine points at a 70 cm distance (mean of 0.10 mm) and higher for the remaining points at a 100 cm distance (mean of 0.16 mm).

These results suggest that system error (noise) is worst for points being tracked farther away from both cameras. Moving to the horizontal and vertical extremities of the cameras' views does not depreciate signal quality.

### 5.2. Calibration error and system noise – marker angles

Using a protractor, grooves that were large enough to fit the edge of the ruler were cut into a piece of foam board every 10°. The board was placed in a location central to the cameras views and the ruler was placed horizontally in the first groove, i.e., 0°, which caused it to face centrally towards the two cameras. At this position, the ruler was recorded for 5 s (see Fig. 4). Sequentially, the ruler was fitted to grooves of larger angles and recorded again for 5 s. At each angle, the ruler was rotated 10°, horizontally, away from the left camera and towards the right camera. As a result the ruler is dimmer in the left camera and brighter in the right camera. Beyond a 50° rotation the stickers, as seen by the left camera, were too dim to be tracked.

This procedure was repeated in the vertical dimension by fitting the ruler and foam board vertically and rotating the ruler downwards by 10° steps from a position facing both cameras to facing downwards. In this case, at each successive step, the stickers were dimmer to both cameras.

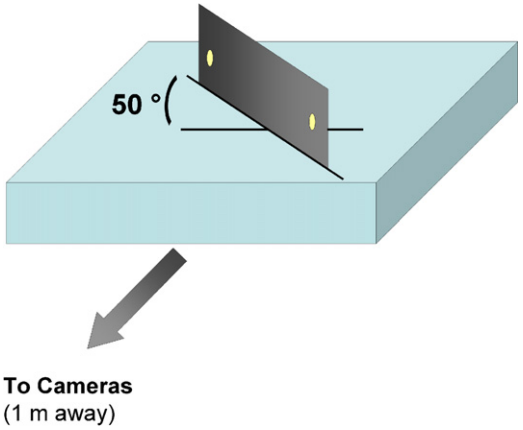


Fig. 4. Fitting of the ruler into a groove angled 50° away from the left camera. The grooves were laid on the base of the calibration cube, indicated by the surrounding lines, which was central to the cameras' views.

Beyond a 30° rotation, the stickers were too dim to be tracked.

At each horizontal and vertical angle, the mean distance between the stickers was calculated along with the standard deviation. Again, the standard deviation across the means calculated at each angle constitutes the calibration error, while the standard deviation at each angle position constitutes the random system error. Throughout the rotations, the calibration error amounted to .1 mm for horizontal rotations and .2 mm for vertical rotations. These values are lower than the value of 0.67 mm reported in test A because the ruler was not relocated throughout the entire viewing field, but rotated in a single spot. System error did not vary systematically through horizontal rotations, but did increase quasi-linearly through the vertical rotation from .1 mm at 0° to .3 mm at 30°.

System error increased steadily only when the angle of rotation was vertical, not horizontal. This result is likely explained by the fact that a change in vertical angle will dim the stickers as seen by both cameras while a change in horizontal angle will dim the stickers as seen by one camera and brighten them as seen by the other, which may compensate for any resolution issues. This encouraging result indicates that changes in horizontal sticker angle, which may occur for non-midsagittal stickers (i.e., on lateral parts of the face), will not affect the digitization results.

### 5.3. System amplitude resolution

The purpose of this test is to determine the minimum amplitude of motion between two stickers that can be considered to be reliably tracked by the system. This is particularly important in speech where the motions being dealt with can be quite small, such as that of the upper-lip. To perform this test, tracking stickers were placed along the edges of large wooden tweezers (see Fig. 5, left). By compressing the ends of the tweezers by hand as shown in Fig. 5, left, a closure of largest amplitude is created at

Please cite this article in press as: Craig, M. et al., Suitability of a UV-based video recording system for the ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.04.011

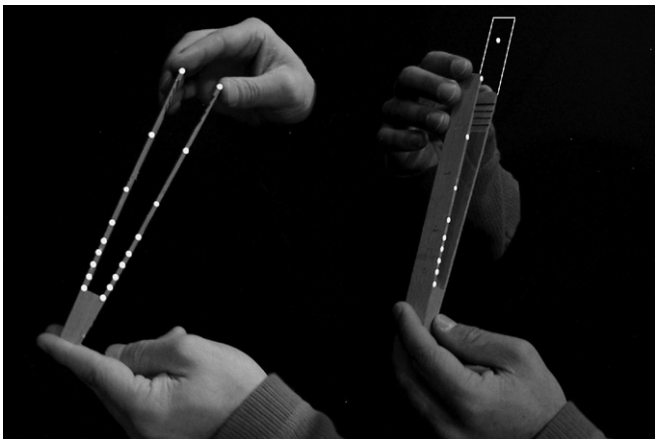

Fig. 5. Left: Hand position for a horizontal closure, with stickers attached to the tweezer edges. Right: Hand position for a depth closure, with stickers attached to the tweezer face and plastic board face. The edges of the transparent board are enhanced for clarity.

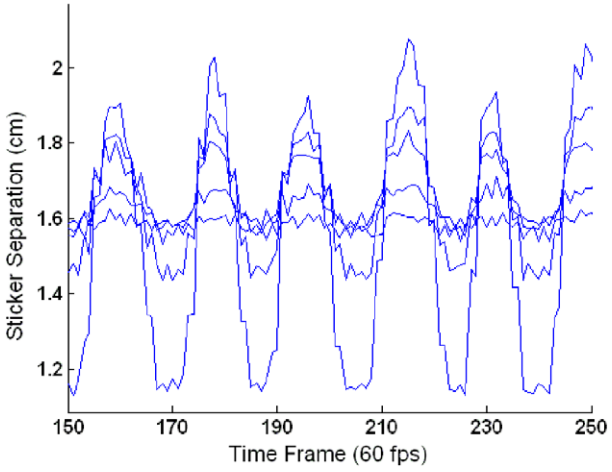


Fig. 6. The signal of the separation of sticker pairs on the tweezers, showing five signals resulting from the same directional input. The largest amplitude is generated by the stickers at the far (open) end of the tweezer.

the ends, and of increasingly smaller amplitude nearer to the vertex of the tweezers, where there will be no closure at all. By opening and closing the tweezers from the ends, the same directional input is applied along the edge of the tweezers, with decreasing amplitude. In an ideal tracking system, the signal that tracks the separation between sticker pairs should be identical (directionally) for all pairs – all pairs will open and close simultaneously – thus the correlation coefficient between the signals for all pairs should be 1. Pair 1, the pair closest to the ends where the manual force is applied by the experimenter, is considered to be the reference signal. Oriented in the vertical ($Y$) and horizontal ($X$) directions, the tweezers were opened and closed 10 times. Pair signals were digitized with APAS. By examining the correlation coefficient of the digitized input signal to the other digitized sticker pair motions of decreasing amplitude, an amplitude threshold limit can be determined. That is, at some threshold amplitude, the correlation coefficient to the input will drop steeply. Thus, this method does not determine the absolute minimal amplitude of a signal that can be measured, but rather (and more importantly) the threshold beyond which smaller amplitudes become unreliable.

Fig. 6 shows an example of the digitized trajectories of five sticker pairs receiving the same directional input, but of different amplitudes. Notice how all signals are closely aligned in time, indicating a comparable linear transfer from finger to tweezer motion at the different locations along the tweezer arms.

To test the resolution in the depth dimension ($Z$), the apparatus was modified slightly. One flat arm of the tweezers was fixed to a rigid, plastic transparent board. Stickers were placed down the side of the other flat arm, facing outwards (see Fig. 5, right). A reference sticker was then placed on the rigid board above the arms of the tweezers. The flat arm then faces both cameras. The signals being compared now are the distances from each sticker on the flat arm to the reference sticker above on the rigid board. Tweezer closures will now simulate motion in the depth dimension, while all stickers are facing the cameras. In a human speech experiment, this setup is similar to that of measuring upper-lip motion relative to a reference sticker on a subject's forehead.

In all three dimensions, the correlation of signals to the input signal fell to between 0.96 and 0.99 for amplitudes of 1 mm (see Fig. 7). The correlation dropped sharply for amplitudes below .5 mm.

Fig. 7 also shows the effect of *interlaced scan* on the accuracy of recordings. Because interlaced scan halves the vertical resolution, the vertical signal ($Y$) decreases in quality more than the other dimensions for any given amplitude below 1 mm. A conservative estimate of the system's resolution threshold is therefore 1 mm.

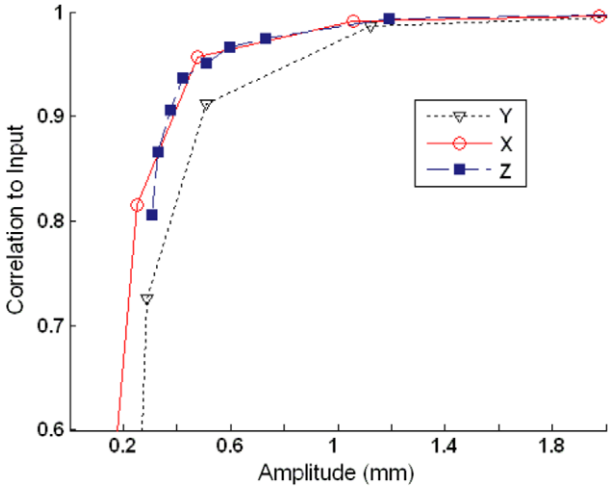


Fig. 7. The correlation of decreasing amplitude signals to the input signal in the vertical ($Y$), horizontal ($X$) and depth ($Z$) dimensions.

Please cite this article in press as: Craig, M. et al., Suitability of a UV-based video recording system for the ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.04.011

### 5.4. Temporal resolution

To test the suitability of the temporal resolution of the cameras for recording speeds of oral articulators that can occur in speech, the following experiment was performed: the experimenter repeated the syllable /pa/ at four different speech rates for a period of approximately 5 s each. During the production, the distance between the upper and lower-lips was digitized. At each speed, the power spectrum of the digitized signal was analyzed to ensure that its peak value corresponded to the desired speech rate. The desired speeds were 2, 4, 6 and 8 Hz, which covers a typical range found for speech repetition rates in both normal speaking subjects and people with speech disorders (Ziegler, 2002). For example, 8 Hz was the fastest speed at which the experimenter could consistently repeat a given syllable. Nyquist's law tells us that the cameras, with a basic shutter speed of 30 fps, will capture information up to 15 Hz. Because speech movements virtually never contain frequency components exceeding this value, the camera shutter speed is sufficient.

From each 5-s recording, the last 3 s were used to derive the power spectrum, to ensure that the experimenter had reached a steady pace.

Fig. 8 shows the combined power spectra of the digitized upper-lip/lower-lip signal at the four speeds, superimposed and normalized by amplitude. The peaks at all four rates are very distinct at their desired values, indicating that the cameras are capable of distinguishing movement repetition rates that fall within the typical range of speech production.

### 5.5. Kinematic testing

In real speech, speakers may use movement undershoot which could lead to smaller movement ranges compared to the repeated syllables used above (Lindblom, 1990). In order to verify our results for real speech, namely, that (1) a 1 mm amplitude threshold is sufficient to record small facial motions; and, (2) that natural speech motions fall well below a 15 Hz limit, a follow-up study was performed. Ten English-speaking, adult subjects, five males and five females, were given the task of reciting a list of 6 groups of 15 phoneme-rich sentences from the TIMIT database (http://www.ldc.upenn.edu/).

Four gestures of motion from each subject were tracked: (1) the motion of the jaw, from a reference point on the forehead; (2) the motion of the upper-lip, from a reference point on the forehead; (3) the separation between the lip corners; and, (4) the separation between the center of the upper and lower-lips. The motion of the upper-lip (motion 2) is the smallest and will therefore provide the strongest test for the resolution of the system. Upper-lip motion is reported in Table 1.

The mean motion of the upper-lip for all subjects was 1.7 mm with a standard deviation of .4 mm. Minimum and maximum mean amplitudes across subjects were 1.1 mm and 2.4 mm. These values suggest that for average speech the recording equipment, with an amplitude resolution threshold of 1 mm, is sufficient to resolve realistic upper-lip motions very well. For certain subjects, there were a number of sentences during which upper-lip motion was below the 1 mm threshold. This was due to the fact that the context of these sentences did not require significant motion of the upper-lip (i.e., they had few bi-labial closures or rounded vowels, as in the sentence "nothing is as offensive as innocence"). Sentences that did require involvement of the upper-lip fell above the threshold. The sentence "help Greg to pick a peck of potatoes" contains four bi-labial closures, a glide and a rounded /o/. For this sentence, the mean upper-lip motion was 1.9 mm for all subjects, which is well above the system's resolution and comparable to upper-lip motion ranges found in other studies using different technologies (e.g., Van Lieshout et al., 2002).

To confirm the suitability of the cameras' temporal resolution for recording speech, the power spectra of the speech motions for all subjects were derived for the four

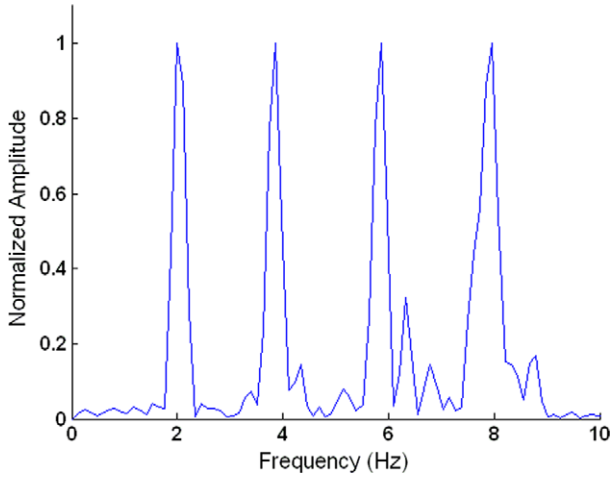


Fig. 8. The four amplitude-normalized power spectra, superimposed, for the separation of the upper and lower-lip during productions of /pa/ at 2, 4, 6 and 8 Hz.

Table 1
Mean upper-lip motion amplitudes and 99% power bandwidth for all 4 gestures for a recording of 90 sentences

| Subject | Mean upper-lip amplitude (mm) | 99% Power bandwidth |
|---|---|---|
| M1 | 1.1 | 6.6 |
| M2 | 1.6 | 5.0 |
| M3 | 2.1 | 5.9 |
| M4 | 1.4 | 6.5 |
| M5 | 1.2 | 7.4 |
| F1 | 1.5 | 5.7 |
| F2 | 2.4 | 5.6 |
| F3 | 1.2 | 6.7 |
| F4 | 2.1 | 6.1 |
| F5 | 1.8 | 6.2 |
| Mean | 1.7 | 6.2 |

articulatory gestures. From these spectra, the bandwidth that carries 99% of the energy was identified. The highest value of this bandwidth for all four gestures for each subject is reported in Table 1. Across all subjects and gestures, the highest 99% bandwidth was 7.4 Hz, which is significantly less than the upper limit of 15 Hz, determined by Nyquist's law. This result is in line with previous investigations in our lab on normal speaking subjects, where it is found that lip movement signals rarely carry meaningful information above 6 Hz (Van Lieshout et al., 2002).

## 6. Discussion

The system error and variation in calibration error for our system, 0.13 mm and 0.67 mm, respectively, are similar to the values reported by Hertrich and Ackermann (1997) for their two-camera IR system, 0.09 mm and 0.54 mm, respectively. Green et al. (2001) using a single camera for two-dimensional measurements reported a resolution better than .1 mm. Their results were determined using a micrometer that was central to the camera's views and with markers that were always facing directly towards the cameras. This is not necessarily the case in measuring subjects who (to an extent) move their head while speaking. Under these circumstances, the current system will show a better performance as well, but we were interested in three-dimensional movement recordings, which complicate matters in terms of camera angles, marker visibility, etc.

In comparison to the results of Green et al. (2001), the measured system error can be considered a lower limit for our system's resolution, at 0.13 mm. However, a more important measure for our purposes is the system amplitude resolution, which tells us the threshold amplitude of a kinematic signal for reliable tracking. This information is important for interpreting the value of data coming from extremely small movements.

The signal quality is very high (near perfect fidelity) at 1 mm amplitude, high (correlation of 0.9 or greater to the reference) above 0.5 mm amplitude, and then drops sharply below this value.

Our findings show that a conservative threshold of 1 mm is sufficient to record even the small upper-lip motion of most adult subjects during the production of normal speech. Since children are found to produce similar magnitudes of facial motion, despite their smaller faces (Riley et al., 2003), our results also support the use of this system for children. Furthermore, the temporal resolution of the cameras is adequate to record typical motions of the face during the production of normal speech. These results signify that out system provides an effective and, for various settings, appealing alternative to more expensive commercial systems, particularly in its use of unobtrusive passive markers. It thus enables a relatively low-cost access to accurate three-dimensional kinematic data of facial motion and expression in populations that normally are more difficult to study due to age (very young children) and disease limitations.

## Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council (NSERC).

## References

Barlow, S.M., Cole, K.J., Abbs, J.H., 1983. A new head-mounted lip–jaw movement transduction system for the study of motor speech disorders. J. Speech Hear. Res. 26, 283–288.

Clark, H., Robin, D., McCullagh, G., Schmidt, R., 2001. Motor control in children and adults during a non-speech oral task. J. Speech Lang. Hear. Res. 44 (5), 1015–1026.

Dromey, C., Benson, A., 2003. Effects of concurrent motor, linguistic, or cognitive tasks on speech motor performance. J. Speech Lang. Hear. Res. 46 (5), 1234–1247.

Green, J.R., Moore, C.A., Higashikawa, M., Steeve, R.W., 2001. The Physiologic development of speech motor control: lip and jaw coordination. J. Speech Lang. Hear. Res. 43 (1), 239–256.

Harris, C., 2004. Exploring smoothness and discontinuities in human motor behaviour with Fourier analysis. Math. Biosci. 188, 99–116.

Hasegawa-Johnson, M., 1998. Electromagnetic exposure safety of the Carstens articulograph AG100. J. Acoust. Soc. Amer. 104, 2529–2532.

Hertrich, I., Ackermann, H., 1997. Accuracy of lip movement analysis. Comparison between electromagnetic articulography and an optical two-camera device. FIPKM 35, 165–170.

Jiang, J., Alwan, A., Keating, P.A., Auer, E.T., Bernstein, L.E., 2002. On the relationship between face movements, tongue movements and speech acoustics. J. Appl. Signal Process. 11, 1174–1188, Special issue of EURASIP.

Katz, W.F., Bharadwaj, S.V., Stettler, M.P., 2006. Influences of electromagnetic articulography sensors on speech produced by healthy adults and individuals with aphasia and apraxia. J. Speech Lang. Hear. Res. 49, 645–659.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 403–439.

Recasense, D., 2002. An EMA study of VCV coarticulatory direction. J. Acoust. Soc. Amer. 111 (6), 2828–2841.

Riley, RR., Smith, A., Smith, A., 2003. Speech movements do not scale by orofacial structure size. J. Appl. Physiol. 94 (6), 2119–2126.

Shaiman, S., 2002. Articulatory control of vowel length for contiguous jaw cycles: the effects of speaking rate and phonetic context. J. Speech Lang. Hear. Res. 45 (4), 663–676.

Simons, G., Smith Pasqualini, M.C., Reddy, V., Wood, J., 2004. Emotional and nonemotional facial expressions in people with Parkinson's disease. J. Int. Neuropsychol. Soc. 10 (4), 521–535.

Tasko, S., McClean, M.D., 2004. Variations in articulatory movement with changes in speech task. J. Speech Lang. Hear. Res. 47 (1), 85–101.

Van Lieshout, P.H.H.M., Rutjens, C.A.W., Spauwen, P.H.M., 2002. The dynamics of interlip coupling in speakers with a repaired unilateral cleft-lip history. J. Speech Lang. Hear. Res. 45 (1), 5–20.

Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. J. Phonetics 30, 555–568.

Ziegler, W., 2002. Task-related factors in oral motor control: speech and oral diadochokinesis in dysarthria and apraxia of speech. Brain Lang. 80 (3), 556–575.

Zierdt, A., Hoole, P., Honda M., Kaburagi T., Tillamn H.G., 2000. Extracting tongues from moving heads. In: Proc. 5th Speech Production Seminar, pp. 313–316.