Published in final edited form as:

J Speech Lang Hear Res. 2014 April 1; 57(2): 426–438. doi:10.1044/2014_JSLHR-S-13-0007.

Accuracy and precision of a custom camera-based system for 2D and 3D motion tracking during speech and nonspeech motor tasks

Yongqiang Feng and

Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

Ludo Max

Department of Speech and Hearing Sciences, University of Washington, Seattle, WA 98105-6246, Haskins Laboratories, New Haven, CT 06511-6695

Abstract

Purpose—Studying normal or disordered motor control requires accurate motion tracking of the effectors (e.g., orofacial structures). The cost of electromagnetic, optoelectronic, and ultrasound systems is prohibitive for many laboratories, and limits clinical applications. For external movements (lips, jaw), video-based systems may be a viable alternative, provided that they offer high temporal resolution and sub-millimeter accuracy.

Method—We examined the accuracy and precision of 2D and 3D data recorded with a system that combines consumer-grade digital cameras capturing 60, 120, or 240 frames per second (fps), retro-reflective markers, commercially-available computer software (APAS, Ariel Dynamics), and a custom calibration device.

Results—Overall mean error (RMSE) across tests was 0.15 mm for static tracking and 0.26 mm for dynamic tracking, with corresponding precision (SD) values of 0.11 and 0.19 mm, respectively. The effect of frame rate varied across conditions, but, generally, accuracy was reduced at 240 fps. The effect of marker size (3 vs. 6 mm diameter) was negligible at all frame rates for both 2D and 3D data.

Conclusion—Motion tracking with consumer-grade digital cameras and the APAS software can achieve sub-millimeter accuracy at frame rates that are appropriate for kinematic analyses of lip/jaw movements for both research and clinical purposes.

Studies of articulatory movements contribute critical information to our understanding of the sensorimotor mechanisms involved in speech production. Positioning and moving the articulators (i.e., shaping the vocal tract) directly determines the final outcome of the speech process – acoustic signals corresponding to speech sounds. Revealing details regarding the nature and organization of movement patterns – and, thus, indirectly, information about the underlying neural control signals – depends on the availability of accurate, reliable, and safe

techniques for transducing the movements of interest. Over the past few decades, a wide variety of motion tracking instruments have been developed and applied to speech production (for an overview, see Earnest & Max, 2003). Examples include x-ray technology, ultrasound, magnetic resonance imaging (MRI), electromagnetic articulography (EMA), and optical or camera-based techniques. Movements of both intra- and extra-oral structures can be tracked using the first four techniques whereas only extra-oral structures can be captured using optical techniques. Sometimes, integration of two or more instrumental techniques allows for better observation of the relevant articulatory movements. For example, the Haskins Optically Corrected Ultrasound System (HOCUS) incorporates an optical system and an ultrasound system for tracking motion of the tongue, lips, and jaw without head constraints (Whalen et al., 2005).

When it is sufficient to track movements of externally visible structures such as the lips and jaw or when non-invasive tracking is required, optical tracking systems can be used. In an active marker system, infrared light emitting diodes (IREDs) are attached as markers of the flesh points that are to be tracked. Each IRED is activated by an electrical signal sent from the control circuitry, and the emitted infrared light is detected by the special cameras' lenses. One widely used example of an active-marker three-dimensional (3D) position tracking system is the Optotrak Certus Motion Capture System (NDI; Northern Digital Inc., Waterloo, Ontario, Canada). In the Optotrak system, every strober (i.e., unit of IRED control circuitry) controls multiple markers and communicates with the main capture unit via wired or wireless connection. The infrared light emission is captured by three lenses. The markers suitable for orofacial structures are as small as 4 mm in diameter, and their positions can be sampled at a rate up to $4600/(n+1.3)$ Hz (where n is the number of markers).

Accuracy and precision are two widely-used performance indices for motion tracking systems. The former is a measurement statistic (often expressed as the root-mean-square error, RMSE) with respect to the true position or displacement of the measured object whereas the latter is a statistic (often expressed as the standard deviation) reflecting the spread of repeated measurements of the position or displacement of an object. For the Optotrak Certus system, the claimed RMSE is 0.1 mm (Northern Digital, 2012a). The system is pre-calibrated at the factory, and a frame with three or more IREDs serves as the reference. Speech researchers typically use a reference frame that is static relative to the skull (e.g., affixed to a head band or goggles) so that all recorded data can be analyzed relative to the head-centric coordinate system. Following similar principles, the Qualisys Oqus systems (Qualisys AB, Gothenburg, Sweden) feature high-speed digital cameras (1, 3+, 4, and 5+ series) with a rate of 180–500 frames per second (fps) at full resolution and 360–1750 fps at reduced resolution. The Qualisys active markers are also available in various sizes.

For speech research, one disadvantage of active-marker systems is that the wires attached to markers on the face may be inconvenient, restrictive, and obtrusive, especially for populations like children, the elderly, and individuals with disabilities. Passive-marker optical systems avoid this problem by using non-wired, reflective markers and a light source. One or more cameras capture the light reflected by the markers for subsequent 2D or 3D movement analyses. Some cameras in this category, such as the Vicon systems (Vicon

Motion Systems Ltd., Los Angeles, CA), function in the range of infrared or near-infrared light. The Vicon Bonita system, for example, uses one or more near-infrared cameras that operate at a rate of 240 or 250 fps. Each Vicon Bonita camera contains a lens and a strobe. The strobe is an array of near-infrared light emitting diodes (LEDs) that surround the lens. The markers reflect the light emitted by the strobe back to the lens. The claimed accuracy is 1 mm for the Bonita 3 system and 0.5 mm for the Bonita 10 system (Vicon, 2012). However, given that the extent of some articulatory movements may be as small as 1–3 mm, sub-millimeter accuracy is needed for speech applications. The BTS Smart D/DX systems (BTS Bioengineering, Milan, Italy) are claimed to offer tracking accuracy in the range appropriate for speech applications: < 0.2 mm for the 100 BTS and 500 BTS models and < 0.1 mm for the 5000 BTS and 7000 BTS models (BTS Bioengineering 2012). In addition, Motion Analysis Corporation (2013) and Qualisys (2013) also provide passive marker systems with sub-millimeter accuracy.

Although such passive-marker systems avoid some of the issues related to subject discomfort and unnatural speaking conditions, and have been used extensively in studies of orofacial motor control (Green, Moore, Higashikawa, & Steeve, 2000; Green, Moore, & Reilly, 2002; Green, Nip, Wilson, Mefferd, & Yunusova, 2010; Wilson & Green, 2009; Wilson, Green, & Weismer, 2012), their high cost remains prohibitive for many research laboratories and makes routine application in the clinical assessment and treatment of speech disorders highly unlikely. A potential solution, if demonstrated to fulfill the relevant technical requirements in terms of sampling rate and spatial accuracy necessary for speech applications, could lie in the use of one (for 2D) or more (for 3D) standard video cameras combined with specialized video analysis computer software (e.g., the APAS system by Ariel Dynamics, Trabuco Canyon, CA, or the MaxTRAQ system by Innovision Systems, Columbiaville, MI). For example, the Ariel Performance Analysis System (APAS, Ariel Dynamics, Trabuco Canyon, CA) is a software package for motion tracking using digital video files recorded with consumer-grade cameras. The pixels of the retro-reflective markers in the video files can be identified, on a frame-by-frame basis, by manual user operation or via automatic tracking algorithms included in the software.¹ Software modules are available for all data extraction and processing steps. Depending on the camera(s), the sampling rate can be 60, 120, or 240 fps (see below Methods section). System calibration requires building a reference frame for 2D tracking or a reference cube for 3D tracking (also described below).

¹The APAS software allows the user to either manually digitize all markers in each frame of one (2D) or two (3D) recorded video streams (in fact, when completing the digitization manually, the user can choose to digitize reliably identifiable anatomical landmarks rather than markers) or to initiate automatic tracking and digitizing of the markers throughout the recordings (note that, as a video-based system, APAS does not allow a real-time display of the acquired data). For automatic digitization, the system uses a proprietary tracking algorithm that is based not only on the brightness and contrast of the markers but also their position, velocity, and acceleration across preceding frames. For such automatic digitization, the markers can be either light or dark relative to the immediate background. Reflective markers, such as used in this work, are recommended but not absolutely necessary as long as the contrast between the marker and its background is sufficiently large (although we have not tested non-reflective markers). For applications on the human face, small reflective markers (here we tested markers with a diameter of 3 or 6 mm) can be placed on discs of black adhesive tape with a diameter that is 2–4 mm larger than the markers themselves. A black background behind each marker also reduces the probability of digitization problems when two markers come close together (when the correct number of markers cannot be identified in a given frame, the APAS digitization process stops, and the user is asked to manually confirm the location of the missing or conflicting markers).

To date, the APAS system has been used in wide variety of limb movement studies (e.g., Ribeiro & Oliveira, 2010; Schlittler, Sanches, Carvalho, & Barela, 2010; Sveistrup et al., 2008), and its accuracy has been verified in the context of large amplitude movements such as human gait analyses (e.g., Klein & De Haven, 1995; Wilson, Smith, & Gibson, 1997). However, verification of this video-based motion tracking system's accuracy for movements as small as the articulatory movements in speech production (i.e., displacements on the order of millimeters), and the small markers (also on the order of millimeters) necessary to track lip and jaw movements, has remained limited (Carter, Pomeroy, & Richards, 2000; Craig, van Lieshout, & Wong, 2007).

Craig et al. (2007) evaluated a custom 3D motion tracking system, built around the APAS software, for potential applications in the area of speech production. The system was tested in the dark, and consisted of two digital video cameras, glow-in-the-dark stickers, ultraviolet black-lights, and the APAS software suite. Craig et al. measured the distance between two static markers on a ruler that was held horizontally for 3 seconds at each of 18 different locations within the view of the cameras – thus, only distance measures along the horizontal x-axis were investigated. The difference between the overall means of inter-marker distance measures obtained with the camera set-up versus hand-held calipers was 0.03 mm (this was considered error due to imperfections in the location of markers used for system calibration). The standard deviation across the 18 camera-based mean inter-marker distances at the various locations in the recording volume was 0.67 mm (this was considered error due to variation in the calibration error across the measured volume). The averaged standard deviation of the measures throughout each 3-second recording was .13 mm (this was considered random system error independent of calibration error). Craig et al. also verified (a) to what extent markers could be angled relative to the camera before tracking was negatively affected, (b) the amplitude resolution of the system, and (c) that the system was able to resolve movements of the upper lip (the articulator that typically shows the smallest displacements) in human subjects during actual speech production. Although the work provided valuable initial information about APAS-based systems for speech research, Craig et al. reported system accuracy (difference between the mean recorded position or distance and the known position or distance) and precision (spread across measurements of a known position or distance) only along the x-axis, only for static markers, only for markers of one size (3 ± 0.5 mm), and only at one sampling rate. The influence of measurement axis (x, y, z), 2D vs. 3D measurements with one vs. two cameras, static versus dynamic markers, sampling rate, and marker size were not investigated. In addition, the feasibility of using a fluorescent light source (rather than UV light in a dark room) remains unknown.

We therefore developed a more comprehensive approach to testing the accuracy and precision of the APAS system under 2D and 3D conditions within the visible light range. Specifically, a set-up combining the APAS software package with two consumer-grade video cameras (we used the exact same cameras as used by Craig et al. (2007) because of their unique ability to operate at 60, 120, or 240 fps) was first calibrated, and the accuracy and precision of this system was then evaluated along all measurement axes in the 2D or 3D recordings of static positions or dynamic displacements of two different sizes of markers that were tracked at all three possible sampling rates. Additionally, the feasibility of using

this complete system with human subjects was evaluated by recording lip and jaw movements during sentence-level utterances produced by an adult male speaker.

Method

System components and overall procedure

The video-based tracking system consisted of a custom-made calibration cube (see below), a fluorescent light source (Caselite 2, Lowel-Light Manufacturing, Hauppauge, NY), circular markers perforated out of a sheet of retro-reflective material (SOLAS Grade Scotchlite Reflective Material, 3M, St. Paul, MN), two digital video cameras (JVC GR-DVL9800, JVC, Yokohama, Japan), a digital video capture card (here Studio DV, Pinnacle Systems, Mountain View, CA, but any digital video capture card can be used), and the APAS software package (Ariel Dynamics, Trabuco Canyon, CA). The JVC GR-DVL9800 cameras were selected for the present testing because they are capable of operating at 60 fps at the maximum spatial resolution (extracting the two individual interlaced fields of each 520×480 video frame), or at 120 fps or 240 fps at reduced spatial resolution (achieved by splitting each field horizontally, vertically, or in four quadrants). These frame rates are similar to those used in many kinematic studies of speech production (e.g., the Carstens AG500 electromagnetic tracking device operates at 200 Hz). It should be noted, however, that the APAS software package can be used with any digital camera(s). In fact, the spatial resolution of most currently available cameras is better than that of the cameras used in this study, and such technological advances can be expected to contribute further improvements in system performance.

The coordinate system and the arrangement of camera(s) and light source with respect to the data acquisition plane or volume are illustrated in Figure 1. The x , y , and z axes corresponded to the medial-lateral, inferior-superior, and posterior-anterior directions, respectively. The fluorescent light was positioned approximately 3 m in front of the capture volume (the regular fluorescent room lights in the laboratory's ceiling were left on). For 2D tracking, one camera was placed approximately 150 cm away from, and oriented perpendicular to, the x - y plane in which the stationary or moving markers were located. For 3D tracking, two cameras were positioned approximately 165 cm from the capture volume and approximately 127 cm apart, oriented toward the center of the data acquisition volume. This positioning was selected to (a) ensure a complete but detailed view of the entire capture volume for each camera (each camera's focus was manually adjusted such that the capture volume was distributed across all available pixels; note that it is the distribution of the image across the available pixels, not the distance between camera and recorded object, that determines the ultimate resolution), and (b) obtaining two images with a sufficiently distinct view of the capture volume (i.e., reducing redundancy between the images) to optimize 3D reconstruction. Given that we recorded from each camera onto separate mini DV tapes, the fluorescent lamp was switched off and on once at the beginning of each 3D recording to serve later as a synchronization signal for the frame sequences from the two videos (note that an alternative approach would be to have two separate video capture cards installed in a computer, and to stream the images directly to the computer's hard drive).

For both 2D and 3D testing, we conducted static (position) as well as dynamic (displacement) tests. During static tests, the tracked markers remained stationary during each data acquisition block, but were relocated to a new position in-between data acquisition blocks. During dynamic tests, the tracked markers were continually moved by a robotic device (Phantom Premium 1.0, SensAble, Wilmington, MA). The potential influence on tracking accuracy of two marker sizes (3 or 6 mm in diameter) and three frame rates (60, 120, or 240 fps) of data acquisition was also evaluated.

When digitizing marker positions in the APAS software, manual identification of the approximate center of each marker in the first frame of the sequence initiated fully automatic, algorithm-based detection of the marker pixels and their center in all subsequent frames.¹ After automatic digitization of the marker coordinates, all data were low-pass filtered with a cut-off frequency of 20 Hz. This cut-off frequency is similar to (e.g., Max, Caruso, & Gracco, 2003) or higher than those typically used in kinematic studies of speech production (Green et al., 2000, 2002; Löfqvist, 2011; Lucero, Munhall, Gracco, & Ramsay, 1997; McClean, Tasko, & Runyan, 2004) and well above the fastest possible rate of articulatory movements (Kent, Kent, & Rosenbek, 1987).

Calibration

The tracking system requires calibrating and establishing a coordinate system defined by locations with known coordinates within the acquisition region. For this purpose, a metal calibration cube (Figure 2.A) was custom built² at a machine shop in the Department of Mechanical Engineering at the University of Connecticut. The cube was open on all sides, had edges 30 cm in length, and was painted black to minimize light reflection. During its construction, 32 shallow indentations for reference markers (16 for 3 mm markers and 16 for 6 mm markers) were drilled at precisely determined coordinates along the four camera-facing edges of each of the two frontal planes (i.e., parallel to the x - y plane). For the testing reported here, 32 retro-reflective markers of 3 mm diameter were placed in the indentations. The cube was mounted on a heavy tripod and placed in the data acquisition region/volume with its top and bottom planes parallel to the transverse plane (i.e., x - z plane).

The APAS system requires identification of a “fixed point” in the calibration images that remains visible in the actual tracking images when the calibration frame has been removed. As shown in Figure 2.B, the fixed point consisted of an extra 3 mm retro-reflective marker, attached to a small wooden extension that was painted black and mounted at the tip of a two-link positioning arm (Variable Friction Magic Arm, Manfrotto, Cassola, Italy).

For 2D calibration, the single camera captured the anterior frontal plane of the cube. Thus, 16 markers were used to define the 2D coordinate system. For 3D calibration, the two cameras captured all 32 markers distributed across the anterior and posterior frontal planes

²When only 2D recordings are required, a very inexpensive calibration device can be constructed by mounting a LEGO® base plate (LEGO Group, Billund, Denmark) on a flat, rigid surface. The molds used in the production of LEGO components have a very high accuracy of 0.01 mm (LEGO Group, 2010), and, thus, the LEGO materials provide a widely-available, low-cost, high-precision alternative to a custom-built calibration device. The studs on such a base plate are 4.8 mm in diameter and evenly spaced at 80 mm intervals. For example, to calibrate 2D recordings, we have previously applied black ink to 16 studs that formed a 320 mm by 320 mm square. These reference locations with known coordinates were then recorded and digitized in the same manner as described for the custom calibration cube.

of the cube, but, due to software limits, only 31 of those markers could be used as calibration points in the APAS software. For each test at a given frame rate, the calibration cube and the fixed point were filmed for approximately 1 s, but only one frame was eventually used for the digitization. Digitizing was accomplished by manually clicking, with a standard computer mouse, the approximate center of each marker in the selected frame image. The final step of the calibration process then involved entering a list of the known Cartesian coordinates for these digitized markers (except for the fixed point).

Note that it is essential that the position of the camera(s) and fixed point remain unchanged from the calibration phase to the subsequent experiment. Therefore, recalibration was required if the camera(s) or fixed point were intentionally or accidentally moved. For the present purposes, a new calibration was completed before each new recording at a different frame rate given that switching to a different frame rate changed the cameras' field of view. The cameras were then also repositioned or reoriented to achieve the best possible view of the cube.

Tests of static tracking

To test the APAS system's accuracy for measuring the position of static markers, one "small" (3 mm diameter) and one "large" (6 mm diameter) marker were attached, approximately 8 mm apart, to the side of the spindle of a depth micrometer with digital read-out (Model 54-225-77, Fowler, Brantford ON, Canada). Both markers were attached to a black background. The position of the spindle could be adjusted with a resolution of 0.001 mm. The micrometer was permanently mounted inside a small cube (with an edge length of 10 cm and with indentations for retro-reflective markers along the edges, similar to the calibration cube) such that the spindle was always parallel to four sides of this small cube and its tip was approximately in the center (Figure 2.C). This micrometer-cube assembly was positioned with the spindle horizontally or vertically for separate tests of tracking accuracy along the x , y , or z axis of the coordinate system (Figure 2.D). Having the spindle mounted inside its own fixed cube allowed us to correct off-line for misalignment between the spindle's axis of movement and the targeted x or y measurement axes (i.e., the axes of the coordinate system defined by the calibration cube). Specifically, the markers on the edges of the 10 cm cube were used to estimate the orientation angles of the spindle within the coordinate system, and to rotate the one-dimensional movement of the micrometer into alignment with the tested x , y , or z axis of the coordinate system.

Each block of testing started by placing the micrometer-cube assembly approximately in the center of the data acquisition region and aligning the spindle with the dimension being tested (i.e., x or y axis for 2D and x , y , or z axis for 3D) by means of a spirit level and visual inspection. Each of the two markers on the spindle was registered as the origin of the micrometer coordinate system for the respective 3 mm and 6 mm tests. The micrometer was then manually adjusted in 1.27 mm increments in the positive direction (0.05 inches, which corresponded to two rotations of the spindle, but actual stop positions were based on the digital display) until the readings indicated 6.35 mm after five steps. Subsequently, starting again from the origin, the spindle was moved toward the negative direction in 1.27 mm increments until the reading reached -6.35 mm. Thus, the excursion process resulted in 12

trials during which each marker was recorded in 11 different positions (2 trials at the origin). For each tested frame rate and axis of movement, this entire process was completed twice. Thus, the x and y axis were tested twice for 2D tracking with a single camera, and the x , y , and z axis were tested twice for 3D tracking with both cameras.

The data acquisition region in which the micrometer-cube assembly and the fixed point were located was recorded for approximately 5 s after the markers were adjusted to each of the described positions. For data analysis, a 1 s sequence of frames was selected from these 5 s videos for each marker position. This selection was arbitrary for 2D tracking but always at a time point after full brightness of the fluorescent light was restored for 3D tracking (as mentioned above, the light was flashed as a signal for synchronization of the two cameras).

Five markers (the small and large target markers, two markers on the small cube, and the fixed point) were necessary to complete the 2D analyses, and six markers (the same five markers plus one extra marker on the small cube) were necessary to complete the 3D analyses. After digitization and filtering (described above), the coordinates of the two target markers were rotated into the micrometer-cube coordinate system and compared with the actual micrometer readings.

Tests of dynamic tracking

To test the APAS system's accuracy for tracking moving targets, two retro-reflective markers (3 and 6 mm in diameter) were secured to a small counterweight attached to the tip of the arm of a robotic device (SensAble Technologies Inc., Phantom Premium, model 1.0A). The robot arm was programmed to move cyclically for 10 s. The effector positions reported by the robot's optical position encoders were sampled at 200 Hz, and these positions were considered as the standard for comparing the APAS system's measurements given that the robot registers positions with a resolution of 0.03 mm. Alignment of the robotic device within the data acquisition region was confirmed with a spirit level and visual examination.

For 2D tracking, the robot arm moved along a straight line parallel to its y axis at a rate of four cycles per second and with a peak-to-peak amplitude of 5 mm. For testing movement along the x axis, the camera – rather than the robotic device – was rotated 90 degrees counterclockwise around the z axis because the robot required its normal upright position to maintain effective force exertion and appropriate alignment of the y axis of its internal coordinate system with that of the calibration volume. For 3D tracking with two cameras, the robot was programmed to move along an elliptical path in the y - z plane at four cycles per second with a peak-to-peak amplitude of 10 mm along the major axis and 5 mm along the minor axis. The angle between the major axis of movement and the y axis was approximately 26 degrees. Although these elliptical trajectories were effectively limited to a 2D plane, they adequately simulated articulatory movements in this sagittal plane and allowed a two-camera examination of movement along the z axis (i.e., the axis not included in the 2D tests with one camera). Principal component analysis (PCA) was applied to both the elliptical trajectories recorded by the robot position encoders and those measured with the APAS system. The peak-to-peak amplitudes along the two axes estimated by the camera tracking system were compared against the readings by the robot position encoders.

Two trials of 2D tracking and four trials of 3D tracking were conducted at each of the three different frame rates (60, 120, 240 fps). Given that the robot was programmed to move for a duration of 10 s, 40 cycles of movement were completed per trial. However, due to inertia of the robot arm, its movements did not reach full stability until four or five cycles had been completed. Consequently, only the peak-to-peak amplitudes of the last 35 or 36 cycles were extracted for analysis.

Data analysis for accuracy and precision testing

In keeping with standard practices and definitions for reporting equipment technical specifications, we calculated both the *accuracy* and *precision* of the APAS measurements that were obtained as described above. To quantify measurement accuracy for tests of static tracking, the root mean square error (RMSE) across the individual video frames from the 1 s recording segment of each trial was calculated along each axis (x , y for 2D tests; x , y , z for 3D tests). To quantify measurement precision, the “compactness” of the cluster of position estimates from the individual video frames of the 1-second recording was calculated as the standard deviation (SD) along each axis. For tests of dynamic tracking, a trial’s accuracy was calculated as the RMSE of the peak-to-peak amplitude estimates along the major and minor axes across the 35–36 individual cycles of the 10 s recording. Precision for these tests of dynamic tracking was calculated as the standard deviation of the 35–36 individual-cycle peak-to-peak amplitude estimates along each of the two measured axes.

All data are presented below as descriptive statistics, illustrated in tabular and graphical format, given that the goal of this work was to provide researchers and clinicians with sufficiently detailed data to support informed decisions regarding the feasibility and appropriateness of using the APAS system for the analysis of orofacial movements during speech and nonspeech motor tasks. That is, the study was not designed to test specific hypotheses regarding differences in accuracy or precision between 2D and 3D tracking set-ups, between markers of different size, or among measurements along the different movement axes. In addition to problems that would be inherent here in defining independently selected experimental units from normally distributed populations, the above described differences (a) in the calculation of RMSE and SD for 2D versus 3D tests and (b) in the number of trials and measurement axes for 2D versus 3D as well as static versus dynamic tests preclude unambiguous inferential statistical comparisons.

Example speech data from a human participant

To verify that the system can be used to track actual articulatory movements during speech production, we conducted a small feasibility test with a human participant. The participant was a 23-year-old male native speaker of American English who produced the utterance “*Buy Bobby a puppy*” at his typical speaking rate.

Data included here were recorded with a single camera operating at 60 fps and placed approximately 2 m from the participant’s face. The dvgrab package for Linux was used to capture the video images directly onto a laptop computer’s hard drive through an IEEE 1394 (“firewire”) interface. The light source was positioned above and approximately 1 m behind the camera. Markers with a 3 mm diameter were first placed on a 6 mm diameter black

background (adhesive tape) and then attached to the participant in the midsagittal plane at the tip of the nose (not used here), at the vermillion border of the upper lip (UL) and the lower lip (LL) and at the distal tip of a wooden dowel pin (length 40 mm, diameter 7.63 mm, mass 1.3 g) that was taped to the undersurface of the chin³ to reflect jaw (J) movement. A small plastic plate (size 35 × 35 mm, mass 2.4 g) with 6 mm reference markers in each corner was attached to the participant's forehead to allow offline corrections for head movement. This set-up of articulatory and reference markers is illustrated in Figure 3.

For offline data analysis, the bottom left reference marker on the forehead plate served as the origin of the coordinate system, and the line connecting the bottom two reference markers served as the *x* axis. Superior-inferior movements of the UL, LL, and J markers were then expressed as time series of positions projected onto the *y* axis. Note that the LL signal reflects contributions of both the lower lip and the jaw given that, for the present purposes, the separate contribution of the jaw (as reflected in the J signal) was not subtracted from the LL signal. Lastly, lip aperture (LA) was calculated as the difference between the *y* coordinates of the UL and LL markers.

Results

Tests of static tracking

Example data from one single trial of 3D, 60 fps tracking of a static 3 mm marker are shown in Figure 4. The left panel shows APAS-based measurements (+ symbols) of the marker's coordinate along the *z* axis (the axis that generally yielded the least accurate results) relative to the actual, micrometer-based coordinates (○ symbols). The APAS data in this figure include all 12 position estimates tested during one trial (1 estimate for each of 5 tested positions in the positive direction of the *z* axis, 1 estimate for each of 5 tested positions in the negative direction of the *z* axis, and 2 estimates at the origin of the *z* axis). Each data point represents the position estimate as obtained by averaging across the 60 individual frames for a 1 s segment of the recording. For this particular 3D test with a 3 mm marker that was tracked at 60 fps, overall accuracy (RMSE) for the *z* axis was 0.09 mm.

For the micrometer steps enclosed by a dotted square in the left panel of Figure 4 (i.e., steps -2 and -3, which yielded relatively poor and relatively good *z* axis accuracy, respectively), full *x*, *y*, and *z* results are shown in detail in the right-side panel of the same figure. Each cluster of APAS data points (+ symbols) includes the 60 single-frame position estimates obtained for the 1 s recording. In this example, measurement precision quantified by the SD along the *x*, *y*, and *z* dimensions was 0.07, 0.09, and 0.14 mm, respectively, for step -2, and 0.10, 0.11, and 0.16 mm, respectively, for step -3.

³One limitation of all optical tracking methods is that movement of a marker attached to the chin would not directly correspond to movement of the mandible itself. Green et al. (2007) reported an average absolute error of 0.81 mm in chin surface markers relative to the rigid mandible, and concluded that it was appropriate to use a chin flesh point to qualitatively describe mandibular behavior but insufficient for precise biomechanical analyses. Possible approaches to reduce or eliminate this particular problem include attaching the marker (a) to a rigid cylindrical object that is adhered to the skin along the undersurface of the jaw or (b) to a splint that is directly attached to the mandibular teeth and that exits from one corner of the mouth. Only the latter method would provide an entirely accurate representation of mandibular movement, albeit at the cost of potential interference with other articulatory movements.

A complete summary of all descriptive accuracy and precision data (2D and 3D tracking; 3 mm and 6 mm markers; all measurement axes) from the static tracking tests is provided in Table 1. The table includes accuracy and precision data that have been obtained by first calculating RMSE and SD across the individual frames of each 1 s recording (as described above) and then averaging the obtained RMSE and SD values across the 24 static tracking trials for a given movement axis. Descriptively, the average accuracy for both marker sizes ranged from 0.03 to 0.11 mm when tracking at a frame rate of 60 fps. Accuracy then slightly worsened to 0.04–0.28 mm for 120 fps, and up to 0.10–0.46 mm for 240 fps. At the highest frame rate, and recording in 3D, measurements along the z axis were associated with the least accurate results (RMSE 0.30 mm for the 6 mm marker and 0.46 mm for the 3 mm marker).

To further illustrate the effect of frame rate on static tracking accuracy of the APAS system, RMSE data collapsed across measurement axis and marker size are shown graphically in Figure 5. It can be seen that for 2D tracking with a single camera, the RMSE mean values at 60 and 120 fps were similar (< 0.1 mm) whereas the RMSE value increased considerably at 240 fps (< 0.2 mm). When tracking in 3D with two cameras, the RMSE mean at 60 fps again remained below 0.1 mm, but at 120 fps, it approached 0.2 mm, and at 240 fps, it exceeded 0.2 mm. In other words, increasing the frame rate from 60 to 120 fps did not decrease the system's accuracy for static tracking in 2D whereas it did have a negative effect on accuracy when tracking in 3D. Nevertheless, even in the worst case of these data averaged across dimensions and marker sizes, the mean RMSE did not exceed 0.25 mm.

In contrast to the above measures of accuracy, precision did not show a similar deterioration with increased frame rate. As can be seen in Table 1 (values in parentheses), SD values remained generally below 0.15 mm, with only the z axis showing relatively worse precision when tracking at 240 fps (SDs of 0.33 and 0.37 mm for 6 and 3 mm markers, respectively).

Tests of dynamic tracking

For illustration purposes, a 5 s portion of one single trial of tracking the 6 mm marker when it was oscillating along the y dimension in the 2D dynamic condition is shown in Figure 6. The marker's trajectory as captured by the APAS system at 60 fps is overlaid on the trajectory registered by the optical encoders of the robotic device that was moving the marker. For quantitative analysis, the peak-to-peak amplitude was determined for all 35–36 stable cycles in the entire 10 s trial (see Method), and the RMSE was calculated by comparing these APAS-estimated peak-to-peak amplitudes with those reported by the robot's optical encoders. For the particular trial that is (partially) shown in Figure 6, the RMSE was 0.16 mm.

A trial of 3D tracking when the 6 mm marker was moved along an elliptical y – z trajectory (i.e., in the sagittal plane) is shown in Figure 7. This trial was tracked by the APAS system at 60 fps. The stable portion of the trajectory that was chosen for analysis included 36 cycles. The ellipses' PCA-derived major and minor axes are indicated by dashed and dotted lines, respectively. For this particular trial, the RMSE of the peak-to-peak amplitudes was 0.08 mm along the major axis and 0.56 mm along the minor axis.

Table 2 includes a complete summary of all descriptive accuracy and precision data from the dynamic tracking tests (2D and 3D tracking; 3 mm and 6 mm markers; all measurement axes). The table includes accuracy and precision data that have been obtained by calculating RMSE and SD across the 35–36 cycles of marker movement in a given trial (as described above) and then averaging the obtained RMSE and SD values across the trials for a given movement axis (2 trials for 2D, 4 trials for 3D). For 2D dynamic tracking, the average accuracy for both marker sizes was 0.12–0.15 mm at a frame rate of 60 fps, and it worsened only slightly at higher frame rates: 0.16–0.20 for 120 fps and 0.07–0.33 mm for 240 fps. In fact, only for the 3 mm marker along the x axis at 240 fps did the RMSE exceed 0.21 mm. For 3D dynamic tracking, on the other hand, accuracy was more negatively affected by frame rate, in particular along the minor axis of the marker's movement ellipse. Although RMSE along the major axis remained limited to 0.38 mm for the 6 mm marker and 0.42 mm for the 3 mm marker, the corresponding values for the minor axis were 0.70 and 0.68 mm.

This effect of frame rate on the dynamic tracking performance of the APAS system is graphically illustrated in Figure 8. The RMSE data in this graph have been collapsed across marker sizes for both the 2D and 3D data and also across measurement axes for the 2D data. The figure clearly illustrates (a) that averaged RMSE never exceeded 0.2 mm when tracking along the x and y axes with a single camera in 2D, but (b) that there was a much more substantial influence of frame rate when tracking along the major movement axis of the elliptical trajectory with two cameras in 3D, and (c) that there was a sudden drop in measurement accuracy when using 240 fps for the minor axis in 3D dynamic tracking.

Table 2 also includes measurement precision data for all dynamic tracking tests (the values reported in parentheses). Quantified by the SD of peak-to-peak amplitude measures across individual movement cycles (and then averaged across trials), precision remained better than 0.25 mm for 20 of the 24 tests, deteriorating to 0.35, 0.39, 0.44, and 0.54 mm in the remaining tests. These four higher values all occurred in 3D tests completed at 240 fps.

Speech data

Figure 9 shows superior-inferior position time series (y axis data relative to the coordinate system defined by the forehead reference markers) for the human participant's UT, LL, and J markers during production of the utterance "*Buy Bobby a puppy*" together with the calculated LA signal. The digitizing module of the APAS system was able to locate all markers in all video images. The bilabial closing and opening gestures associated with the various consonants and vowels/diphthongs are clearly identifiable. In addition, the individual articulator signals as well as the LA signal also accurately reflect known differences in LL and J lowering – and, thus, variation in overall oral opening as reflected by LA – across the vowels and diphthongs of different height.

Discussion

This work evaluated the suitability of a video-based motion tracking set-up for 2D or 3D analyses of small-amplitude movements such as those in the orofacial system during speech or nonspeech motor tasks. For the tests reported here, the system consisted of a custom-built calibration cube, two consumer-grade digital camcorders, retro-reflective markers, a digital

video capture board, and the APAS software package with its modules for digitizing, processing, and analyzing the data. This set-up was tested in 2D (single camera) and 3D (two cameras) configurations with the cameras operating at 60, 120, or 240 fps to track static and dynamic markers of either 3 or 6 mm in diameter.

Results indicate that the overall mean error (RMSE) was 0.15 mm for static tracking and 0.26 mm for dynamic tracking, with corresponding precision (SD) values of 0.11 and 0.19 mm, respectively. These RMS errors compare favorably to those reported by Klein and De Haven (1995) who examined an early version of the APAS system under 3D static conditions for application in general human kinematics. Part of this improvement is likely related to advancements in video camera technology and in the implementation of the automatic tracking algorithms within the APAS software. An additional factor that may have contributed to the excellent spatial resolution may be related to the fact that we custom built a calibration cube that defined a relatively small volume ($30 \times 30 \times 30$ cm) within which all movements of interest took place. That is, by devoting the full set of markers that is used in the calibration images to a relatively small volume in the cameras' field of view, the accuracy of subsequent APAS-based recordings within that calibrated volume may be much improved as compared with a situation in which a larger volume was calibrated by means of the same number of markers.

For applications in speech production research, arguably the most important performance evaluation involves the tests of 3D dynamic tracking with two cameras. Lip and jaw movements for speech occur predominantly in the sagittal plane with minimal lateral movement. Accurately capturing such movements along both the superior-inferior and anterior-posterior axes requires two separate cameras. In the present study, the overall mean error for two-camera dynamic tracking of markers moving in the sagittal plane was 0.35 mm across the three tested frame rates. Especially at the frame rates of 60 and 120 fps, measurement error (generally < 0.3 mm) was comparable to that of the other optical systems reviewed above. In fact, measurement errors of the system evaluated in the present study are essentially equivalent to, or better than, those of speech-specific electromagnetic motion trackers (e.g., Carstens AG501, NDI Wave). For example, Carstens reports the "dynamic positional accuracy" for their model AG501 articulograph to be "0.3 mm RMS for line movements of 100 mm in length" (Carstens Medizinelektronik, 2012), and Yunusora, Green, and Mefferd (2009) reported for the previous generation AG500 a median RMSE under 0.5 mm. Northern Digital reports that their Wave Speech Research System achieves a "static positional accuracy" of 0.6 mm RMS and a "dynamic positional accuracy" of 1.5 mm RMS (Northern Digital, 2012b). Berry (2011) reported a positional error of 0.2 mm for static tracking and an error of < 0.5 mm for dynamic tracking. When we evaluated the APAS system at 60 or 120 fps, errors for both the 3 mm and the 6 mm markers generally (in 34/36 tests) did not exceed 0.3 mm, and this was the case for static as well as dynamic tracking conditions. When we evaluated the APAS system at 240 fps, error was less likely to remain below 0.3 mm (12/18 tests), especially for the minor axis of elliptical movements during 3D tracking. Nevertheless, even in the worst case of our tests (240 fps, 6 mm marker, minor axis of the elliptical trajectory tracked in 3D), error did not exceed 0.70 mm.

The observed trend of increasing measurement error with increasing frame rate probably results from the spatial versus temporal resolution trade-off in the high-speed modes of the JVC cameras. The cameras used for the present work achieve a higher frame rate at the cost of spatial resolution: the full resolution that is available at the rate of 60 fps is reduced at 120 and 240 fps because each video field is split in two halves or four quadrants. Although this trade-off is indeed a limitation of the set-up tested here, it should be noted that the remaining two operating modes (i.e., 60 and 120 fps) on these particular cameras would provide sufficient temporal resolution for most clinical or research applications, and that other cameras are currently commercially available that have not only similar or higher frame rates but also better spatial resolution. It seems reasonable to hypothesize that the use of such cameras will improve the system's performance even further as compared with the accuracy and precision data reported here.

Our results also suggest that the influence of marker size on indices of measurement accuracy and precision was minimal. The ability to use markers as small as 3 mm in diameter has a number of important advantages. First, smaller markers are less likely to interfere with the orofacial structures being tracked (e.g., lips), especially in children. Second, smaller markers allow for more accurate placement on specific anatomical landmarks (e.g., the vermilion border of the upper and lower lips). Third, when many markers are used, a smaller diameter reduces the possibility of failure in the automatic identification algorithm within the digitization software. That is, closer proximity between the edges of relatively large markers could cause the pixels to "merge" in the video images, resulting in measurement error that needs to be manually corrected.

Lastly, it worth noting that factors other than the inherent capabilities of the JVC cameras and APAS software may have contributed error in our measurements. For example, and perhaps most importantly, the custom-built calibration cube may have been less accurate in the anterior-posterior dimension (z measurement axis). Indentations for positioning the calibration markers were precisely drilled in the pre-constructed front- and back-plane squares, but these squares then had to be welded to four parallel front-to-back bars in order to form a complete cube. In addition, markers could be attached only on the front and back planes of the cube but not in intermediate positions. Furthermore, it is likely that there was less than perfect alignment of the coordinate system in which the test markers were positioned/moving with the coordinate system of the recording set-up as determined by the calibration cube. Specifically, for static tracking, imperfect alignment of the micrometer tip with respect to the enclosing small cube would lead to incomplete compensation relative to the calibrated measurement volume. For dynamic tracking, positioning of the robot relative to the calibrated measurement volume was based on visual inspection and a spirit level. Hence, alignment of the robot coordinate system and the APAS coordinate system was also not perfect. Furthermore, the known positions of both markers within the coordinate system of the robot were derived based on the joint angles registered by the device's optical encoders and the manually measured distance of the markers' center to the tip of the robot. Slight inaccuracies in the measurement of that distance would add error in the calculation of the reference positions. Overall, these additional complexities – resulting from the testing set-up rather than being intrinsic to the camera-based tracking system itself – introduced

additional sources of measurement error, and, thus, the accuracy and precision reported here can be safely interpreted as upper limits for the APAS system's RMSE and SD. In other words, the system's actual error can be considered to be "not larger than" the reported values.

In summary, this evaluation of the APAS video-based motion tracking system suggests that the accuracy and precision of the system are appropriate for both clinical applications in the evaluation and treatment of speech disorders (where single-camera tracking of superior-inferior motion may suffice) and more sophisticated research applications (where 3D tracking may be necessary). Moreover, the data recorded from a human participant confirmed that the tested set-up with retro-reflective markers does indeed allow reliable tracking of lip and jaw movements during speech production, including offline corrections for head movement.

Both clinical and research applications require a few precautionary steps to achieve the reported levels of accuracy and precision. First, the system requires user calibration, and for 3D recordings this necessitates the construction of a precisely machined calibration cube (see footnote 2 for an alternative calibration procedure that can be used in the case of 2D recordings). A custom-built calibration device matched in size to the area or volume of interest (as tested in this work) may provide better spatial resolution than the larger calibration device available for purchase with the APAS software (Ariel Dynamics, 2012). Second, for our 3D tracking tests, we manually synchronized in the APAS software the video frame sequences that had been recorded onto separate mini-DV tapes by the two cameras. Our measurements show that this synchronization can be reliably accomplished by simply switching off and on the light source that is used in conjunction with the retro-reflective markers. It is also possible, however, to avoid this manual synchronization by initiating both cameras' recording activity from a single computer with two video capture cards (an option not tested in the present study). Third, given that the system is based on video images, tracking is limited to externally visible structures, and movements inside the oral cavity (e.g., tongue movements) cannot be captured. Similar to the aforementioned HOCUS system (Whalen et al., 2005), however, it is technically feasible to combine optical tracking with an additional system that is capable of registering tongue movements (e.g., ultrasound). Lastly, and this is a limitation of all optical tracking systems, prospective users should be aware that a system like the one tested in this work cannot provide information about the markers' angle of rotation. For studies in which pitch, roll, and/or yaw angles of the sensors need to be known, electromagnetic tracking systems provide a better alternative.

Acknowledgments

This work was supported in part by grant R01DC007603 from the National Institute on Deafness and Other Communication Disorders and grant Y154221431 from the Institute of Acoustics, Chinese Academy of Sciences. The authors thank Thomas J. Mealy for mechanical engineering support.

References

Ariel Dynamics. Product catalog. 2012. Retrieved from <http://www.arielnet.com/cloud/Catalogs/Show/Online>

- Berry JJ. Accuracy of the NDI Wave Speech Research System. *Journal of Speech, Language, and Hearing Research*. 2011; 54:1295–1301.
- BTS Bioengineering. BTS Smart DX: High-precision optoelectronic system for the biomechanical motion analysis. 2012. Retrieved from http://www.btsbioengineering.com/resources/Company/Media-center/Brochure-Library/BTSENG_SMARTDX-1011US-WQ.pdf
- Carstens Medizinelektronik. High precision electromagnetic articulograph AG501. 2012. Retrieved from <http://www.articulograph.de>
- Carter JA, Pomeroy VM, Richards J. The feasibility of a kinematic measure of lip closure during meaningful speech. *Disability and Rehabilitation*. 2000; 22:820–826. [PubMed: 11197518]
- Craig M, van Lieshout P, Wong W. Suitability of a UV-based video recording system for the analysis of small facial motions during speech. *Speech Communication*. 2007; 49:679–686.
- Earnest MM, Max L. En route to the three-dimensional registration and analysis of speech movements: instrumental techniques for the study of articulatory kinematics. *Contemporary Issues in Communication Science and Disorders*. 2003; 30:5–25.
- Green JR, Moore CA, Higashikawa M, Steeve RW. The physiologic development of speech motor control: Lip and jaw coordination. *Journal of Speech, Language, and Hearing Research*. 2000; 43:239–255.
- Green JR, Moore CA, Reilly KJ. The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*. 2002; 45:66–79.
- Green JR, Nip ISB, Wilson EM, Mefferd AS, Yunusova Y. Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*. 2010; 53:1529–1542.
- Green JR, Wilson EM, Wang YT, Moore CA. Estimating mandibular motion based on chin surface targets during speech. *Journal of Speech, Language, and Hearing Research*. 2007; 50:928–939.
- Kent RD, Kent JF, Rosenbek JC. Maximum performance tests of speech production. *Journal of Speech and Hearing Disorders*. 1987; 52:367–387. [PubMed: 3312817]
- Klein PJ, DeHaven JJ. Accuracy of three-dimensional linear and angular estimates obtained with the Ariel Performance Analysis System. *Archives of Physical Medicine and Rehabilitation*. 1995; 76:183–189. [PubMed: 7848077]
- LEGO Group. Company profile: An introduction to the LEGO Group. 2010. Retrieved from <http://cache.lego.com/upload/contentTemplating/AboutUsFactsAndFiguresContent/otherfiles/download98E142631E71927FDD52304C1C0F1685.pdf>
- Löfqvist A. Vowel-related tongue movements in speech: straight or curved paths? *Journal of the Acoustical Society of America*. 2011; 129:1149–1152. [PubMed: 21428476]
- Lucero JC, Munhall KG, Gracco VL, Ramsay JO. On the registration of time and the patterning of speech movements. *Journal of Speech, Language and Hearing Research*. 1997; 40:1111–1117.
- Max L, Caruso AJ, Gracco VL. Kinematic analyses of speech, orofacial nonspeech, and finger movements in stuttering and nonstuttering adults. *Journal of Speech, Language, and Hearing Research*. 2003; 46:215–232.
- McClean MD, Tasko SM, Runyan CM. Orofacial movements associated with fluent speech in persons who stutter. *Journal of Speech, Language and Hearing Research*. 2004; 47:294–303.
- Motion Analysis Corporation. Movement analysis products. 2013. Retrieved from <http://www.motionanalysis.com/html/movement/products.html>
- Northern Digital. Optotrak Certus motion capture system. 2012a. Retrieved from <http://www.ndigital.com/lifesciences/certus-motioncapturesystem.php>
- Northern Digital. The Wave speech research system. 2012b. Retrieved from <http://www.ndigital.com/lifesciences/products-speechresearch.php>
- Qualisys. Oqus camera series. 2013. Retrieved from <http://www.qualisys.com/products/hardware/oqus>
- Ribeiro F, Oliveira J. Effect of physical exercise and age on knee joint position sense. *Archives of Gerontology and Geriatrics*. 2010; 51:64–67. [PubMed: 19716189]
- Schlittler DXC, Sanches MB, Carvalho RP, Barela JA. Ideal treadmill speed to stimulate stepping in infants. *Revista Brasileira de Fisioterapia*. 2010; 14:483–490. [PubMed: 21340242]

- Sveistrup H, Schneiberg S, McKinley PA, McFadyen BJ, Levin MF. Head, arm and trunk coordination during reaching in children. *Experimental Brain Research*. 2008; 188:237–247. [PubMed: 18392615]
- Vicon. Bonita: small and powerful optical capture camera family. 2012. Retrieved from <http://www.vicon.com/products/bonita-features.html>
- Whalen DH, Iskarous K, Tiede MK, Ostry DJ, Lehnert-LeHouillier H, Vatikiotis-Bateson E, Hailey DS. The Haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research*. 2005; 48:543–553.
- Wilson EM, Green JR. The development of jaw motion for mastication. *Early Human Development*. 2009; 85:303–311. [PubMed: 19185434]
- Wilson EM, Green JR, Weismer G. A kinematic description of the temporal characteristics of jaw motion for early chewing: Preliminary findings. *Journal of Speech, Language, and Hearing Research*. 2012; 55:626–638.
- Wilson DJ, Smith BK, Gibson JK. Accuracy of reconstructed angular estimates obtained with the Ariel Performance Analysis System. *Physical Therapy*. 1997; 77:1741–1746. [PubMed: 9413452]
- Yunusova Y, Green JR, Mefferd A. Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech, Language, Hearing Research*. 2009; 52:547–555.

Figure 1.

Data acquisition volume, coordinate system, and placement of the camera(s) and light source for 2D and 3D motion tracking.

Figure 2.

Calibration cube with retro-reflective markers on the camera-facing surfaces of the two frontal planes (A). During calibration (i.e., the calibration cube is filmed without subject), the location of a “fixed point” is digitized together with the markers on the calibration cube; for subsequent movement tracking during a speech experiment (i.e., the subject is filmed without calibration cube), the fixed point needs to remain in the same location in the video images; in this illustration, the fixed point is in front of the subject’s forehead (B). Micrometer-cube assembly used to move retro-reflective markers to positions with known coordinates for all tests of static tracking (C). Illustration of the set-up for static tests along the x axis (D).

Figure 3.

Participant with retro-reflective markers for data collection. Markers are attached midsagittally to the upper lip, the lower lip, and the end of a dowel pin attached under the jaw (the marker on the nose was not used). A rectangular frame with reference markers defining the coordinate system is attached to the forehead (thus allowing corrections for head movement). One of the two markers above the reference frame is used as the “fixed point” that needs to be present in both the calibration images and the data recording images.

Figure 4.

Single trial data from 60 fps 3D tracking of a static 3 mm marker. Left: APAS measurements of the marker's measured z coordinate relative to its actual coordinate. APAS data include 12 position estimates (5 positions in the positive and negative directions each, 2 measurements at the origin) obtained by averaging the position estimates from 60 individual image frames. Right: Full x , y , and z results for micrometer steps -2 and -3. Clusters of APAS position estimates represent the measurements for 60 individual image frames.

Figure 5.

RMSE (in mm) for 2D and 3D static point tracking averaged across trials, measurement axes, and marker sizes. Error bars represent standard errors of the mean.

Figure 6.

Single trial data from 60 fps 2D camera tracking (blue) of a 6 mm marker that moved with a 5 mm peak-to-peak amplitude along the y axis. The camera data are shown together with the output of the optical encoders on the robotic device that moved the marker (red). Only a 5 s portion of the 10 s trial is shown.

Figure 7.

Single trial data from 60 fps 3D tracking of a 6 mm marker that moved along an elliptical trajectory in the y - z sagittal plane (left) together with the output of the optical encoders on the robotic device that moved the markers (right). The estimated major axes are designated by dashed lines and the minor axes by dotted lines. Anterior is to the left.

Figure 8.

RMSE (in mm) for dynamic tracking peak-to-peak amplitude measurements averaged across trials, marker sizes, and measurement axes for the 2D condition and across trials and marker sizes for the 3D condition. Error bars indicate standard errors.

Figure 9.

Position time series of upper lip (UL), lower lip (LL), and jaw (J) markers during an adult male speaker's production of "Buy Bobby a puppy." Data correspond to the markers' y axis position (superior-inferior movements) relative to a coordinate system defined by reference markers on the forehead. The bottom trace represents a lip aperture (LA) signal that was calculated as the difference between the y coordinates of the UL and LL markers.

Table 1

Static marker tracking results for all frame rates and marker sizes. Data (in mm) represent the average accuracy (accuracy for each trial was quantified as RMSE for the position estimates from the individual frames of a 1 s recording) and, in parentheses, the average precision (precision for each trial was quantified as the SD for the position estimates from the individual frames of a 1 s recording) across all 24 trials for a given combination of frame rate, marker size, and movement axis (2 trials at each of 5 positions in the positive direction, 2 trials at each of 5 positions in the negative direction, and 4 trials at the origin).

		Frame rate					
		60		120		240	
		Marker size					
Dimension		6 mm	3 mm	6 mm	3 mm	6 mm	3 mm
2D	x	0.09 (0.05)	0.09 (0.07)	0.09 (0.09)	0.14 (0.10)	0.17 (0.07)	0.22 (0.04)
	y	0.04 (0.06)	0.10 (0.04)	0.04 (0.09)	0.06 (0.11)	0.13 (0.08)	0.24 (0.07)
3D	x	0.04 (0.06)	0.11 (0.05)	0.08 (0.07)	0.07 (0.06)	0.13 (0.09)	0.10 (0.10)
	y	0.06 (0.07)	0.03 (0.08)	0.25 (0.12)	0.28 (0.12)	0.15 (0.12)	0.27 (0.12)
	z	0.08 (0.14)	0.09 (0.15)	0.26 (0.12)	0.22 (0.14)	0.30 (0.33)	0.46 (0.37)

Table 2

Dynamic marker tracking results for all frame rates and marker sizes. Data (in mm) represent the average accuracy (accuracy for each trial was quantified as RMSE of the peak-to-peak displacement estimates for the individual cycles of movement in the trial) and, in parentheses, the average precision (precision for each trial was quantified as SD of the peak-to-peak displacement estimates across the individual cycles of movement in the trial) across trials (two trials for 2D tests and four trials for 3D tests) for a given combination of frame rate, marker size, and movement axis.

		Frame rate					
		60		120		240	
		Marker size					
Dimension		6 mm	3 mm	6 mm	3 mm	6 mm	3 mm
2D	<i>x</i>	0.12 (0.09)	0.15 (0.15)	0.16 (0.12)	0.19 (0.15)	0.21 (0.17)	0.33 (0.21)
	<i>y</i>	0.14 (0.08)	0.14 (0.10)	0.20 (0.18)	0.18 (0.16)	0.07 (0.04)	0.10 (0.07)
3D	<i>major</i>	0.19 (0.08)	0.09 (0.09)	0.23 (0.24)	0.24 (0.23)	0.38 (0.35)	0.42 (0.39)
	<i>minor</i>	0.47 (0.10)	0.18 (0.12)	0.30 (0.21)	0.31 (0.22)	0.70 (0.44)	0.68 (0.54)