The Haskins Optically Corrected Ultrasound System (HOCUS)

TUTORIAL

D. H. Whalen
Khalil Iskarous
Haskins Laboratories,
New Haven, CT

Mark K. Tiede
Haskins Laboratories,
New Haven, CT, and
Massachusetts Institute
of Technology, Cambridge

David J. Ostry
Haskins Laboratories,
New Haven, CT, and
McGill University,
Montreal, Quebec, Canada

Heike Lehnert-LeHouillier
Haskins Laboratories,
New Haven, CT, and
University at Buffalo,
Buffalo, NY

Eric Vatikiotis-Bateson
University of British Columbia,
Vancouver, British Columbia, Canada

Donald S. Hailey
Haskins Laboratories,
New Haven, CT

The tongue is critical in the production of speech, yet its nature has made it difficult to measure. Not only does its ability to attain complex shapes make it difficult to track, it is also largely hidden from view during speech. The present article describes a new combination of optical tracking and ultrasound imaging that allows for a noninvasive, real-time view of most of the tongue surface during running speech. The optical system (Optotrak) tracks the location of external structures in 3-dimensional space using infrared emitting diodes (IREDs). By tracking 3 or more IREDs on the head and a similar number on an ultrasound transceiver, the transduced image of the tongue can be corrected for the motion of both the head and the transceiver and thus be represented relative to the hard structures of the vocal tract. If structural magnetic resonance images of the speaker are available, they may allow the estimation of the location of the rear pharyngeal wall as well. This new technique is contrasted with other currently available options for imaging the tongue. It promises to provide high-quality, relatively low-cost imaging of most of the tongue surface during fairly unconstrained speech.

KEY WORDS: speech production, ultrasound, tongue measurement, kinematics, boundary detection

The tongue is the most important speech organ for forming consonants and vowels, yet it is one of the most difficult of the speech organs to measure. Its usefulness stems in part from its flexibility, but this flexibility adds to the difficulty of measuring tongue shape. A further difficulty is that it is normally hidden from view during speech, making the most convenient and unobtrusive imaging techniques unusable. Even intrusive techniques have difficulty imaging the pharyngeal cavity. Our ability to measure the tongue, therefore, is not commensurate with its importance in speech.

Various ingenious methods have been devised over the years to measure static tongue shapes and dynamic tongue motion (Stone, 1997), each with its strengths and weaknesses. Because the tongue is not normally visible externally, it was natural to study it with X-rays when they became available. Many interesting features could be observed this way (Russell, 1928), especially the critical role of the pharynx in vowel articulation (Carmody, 1941). High dosage levels and a lack of dynamic information limited these early studies, however. Much better information was obtained from later cineradiography (Munhall, Vatikiotis-Bateson, & Tohkura, 1995; Öhman, 1966; Perkell, 1969; Rochette, 1973; Stevens & Öhman, 1963; Wood, 1982), but exposure limitations to ionizing radiation still made this kind of data hard to obtain in large quantities. More recent procedures have allowed for X-ray research protocols of short duration to be viable again (Fitch & Reby, 2001; Stark

1092-4388/05/4803-0543

et al., 1999), but the limited time for acquisition is still a severe limitation on its usefulness. Computed tomography has also been used clinically (Larsson, Mancuso, & Hanafee, 1982; Stutley, Cooke, & Parsons, 1989).

Another approach to tongue measurement in speech was the development of systems that track a small number of points on the surface of the tongue. These depended either on a greatly constrained, and therefore much safer, X-ray microbeam system (Abbs & Nadler, 1987; Kiritani, 1986; Kiritani, Itoh, & Fujimura, 1975) or alternating magnetic fields generated by coils placed outside the head, that is, electromagnetometry (Perkell et al., 1992; Schönle et al., 1987). In both cases, tracking is performed on a foreign object (gold pellet or receiver coil) that has to be glued to the tongue surface. While this interferes minimally with articulation in most cases (Weismer & Bunton, 1999), placement is a lengthy and difficult procedure that is not tolerated by all potential participants. These systems have allowed for extensive tracking of points on the tongue surface in real time. Many areas in phonetics and speech science have benefited from experiments using these devices: speech motor control (Hoole & Nguyen, 1997; Löfqvist & Gracco, 1994; Perkell, Zandipour, Matthies, & Lane, 2002; Westbury, 1994a), phonetic variation (Gick, Iskarous, Whalen, & Goldstein, 2003), speech errors (Pouplier & Goldstein, 2002), and speech disorders (van Lieshout, Alfonso, Hulstijn, & Peters, 1993; Weismer, Yunusova, & Westbury, 2003).

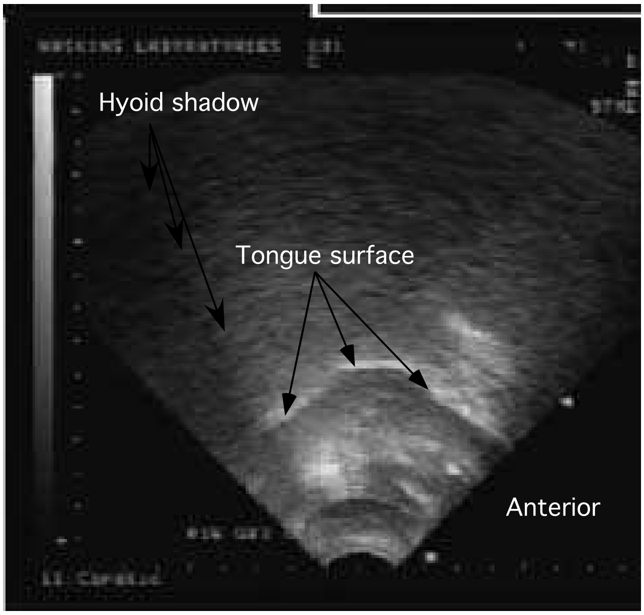
Imaging of the full surface of the tongue, along with the rest of the vocal tract, has been best accomplished to date by magnetic resonance imaging (MRI; Baer, Gore, Boyce, & Nye, 1987; Baer, Gore, Gracco, & Nye, 1991; Lakshminarayanan, Lee, & McCutcheon, 1991; Rokkaku, Hashimoto, Imaizumi, Niimi, & Kiritani, 1986). By manipulation of magnetic gradients, MRI makes possible volumetric imaging of vocal tract structures, with distinct tissue types, including aspects of tongue musculature, distinguished by their differing concentrations of imageable protons. However, various factors make this form of imaging less than ideal, including the supine position of the speaker, the expense of the experimental sessions, the noise the equipment generates, and the relatively slow sampling rate. Its imaging of the velum is unrivaled, however. This is especially true of recent enhancements including "tagged" MRI (Niitsu et al., 1994; Stone et al., 2001), in which an excitation grid is superimposed on the tissue so that deformations of cubes of tongue tissue (e.g., during speech) can be measured (Stone, Epstein, & Iskarous, 2004). Also, a few experimental systems support same-plane sampling rates of up to 20 Hz, making imaging of dynamic speech a possibility (Demolin, Metens, & Soquet, 2000; Narayanan, Nayak, Lee, Sethy, & Byrd, 2004).

The final imaging system for the tongue that we will discuss is ultrasound. The application of ultrasound to the visualization of speech articulation was pioneered by Stone and colleagues (Morrish, Stone, Sonies, Kurtz, & Shawker, 1984; Stone & Davis, 1995; Stone, Faber, Raphael, & Shawker, 1992; Stone & Lundberg, 1996; Stone, Sonies, Shawker, Weiss, & Nadel, 1983), with substantial input from others as well (e.g., Kaburagi & Honda, 1994b). Hardware has now improved to the point where the signals returned by standard settings for ultrasound machines can readily be interpreted as tongue surfaces and are amenable to a wide variety of mathematical descriptions (see Figure 1).

As can be seen in Figure 1, ultrasound captures an almost complete view of the tongue, giving us much more information about the pharyngeal region than is available in point parameterization systems. Imaging speed using modern ultrasound scanners can reach 200 Hz, but systems that require analog video recording of the images force the frame rate to drop to 30 Hz. Availability of digital video capability, however, has removed this obstacle, and our laboratory has begun collecting data at 200 Hz. Spatial resolution is determined by several imaging parameters and is about 1 mm.

If the probe is allowed to move with the jaw, then the position of the tongue is captured only relative to the jaw—the position of the tongue within the vocal tract is not captured. If we are only interested in the shape of the tongue, then measurement in a jaw-centered system

Figure 1. Ultrasound image of the tongue surface during the English vowel /ç/. Note the shadow cast by the hyoid bone at the left side of the image.



is sufficient, but if we are also interested in constrictions, the location of the tongue within the vocal tract must be measured. One way to avoid this problem is to immobilize the head and the probe (Munhall, Ostry, & Parush, 1985; Ostry, Keller, & Parush, 1983; Parush, Ostry, & Munhall, 1983; Stone & Davis, 1995; Wrench & Scobbie, 2003). Because the distance between the head and probe remains constant, the shape of the tongue is captured in a head-centered coordinate system, instead of a jaw-centered one. However, while the jaw is still able to move (especially if an acoustic standoff is used), it is impeded by the stationary probe. It has been established in many experiments on speech motor control that impeding the jaw, even by slight amounts, triggers compensatory mechanisms in the tongue (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984; Lindblom, Lubker, & Gay, 1979). So it is not known whether the data collected using a stationary probe setup reflect the same patterns as speech uttered in a less restrictive setting. Immobilization of the head itself and restricting the general posture of the speaker could also lead to departure from natural speech patterns. Further, some speakers (e.g., children, the elderly, and persons with certain speech disorders) may find it difficult to undergo the required immobilization.

The present article describes a system that takes advantage of ultrasound but that does not require immobilization. The system, the Haskins Optically Corrected Ultrasound System (HOCUS), incorporates both ultrasound imaging of the tongue and optical tracking of the probe relative to the head, and thus tongue surface data can be reoriented to be relative to the head. The head, probe, and jaw are allowed to move, but their motion is tracked and can therefore be used to correct the tongue measurement to a head-based coordinate frame. The probe may either be held to a fixed orientation to collect cross-sectional data during running speech or moved to different orientations during sustained phonation to obtain multiple cross-sections for three-dimensional reconstruction. Optical tracking data can also be obtained for the lips and jaw (and, conceivably, other visible structures) in a way that provides one of the most complete measurements of the vocal tract during running speech that we have seen. The implementation issues that have to be dealt with are (a) collecting the measurements, (b) aligning them in time, (c) putting them into a common coordinate system, and (d) extracting the most relevant speech information. These issues are described below, after which we will conclude.

Instrumentation

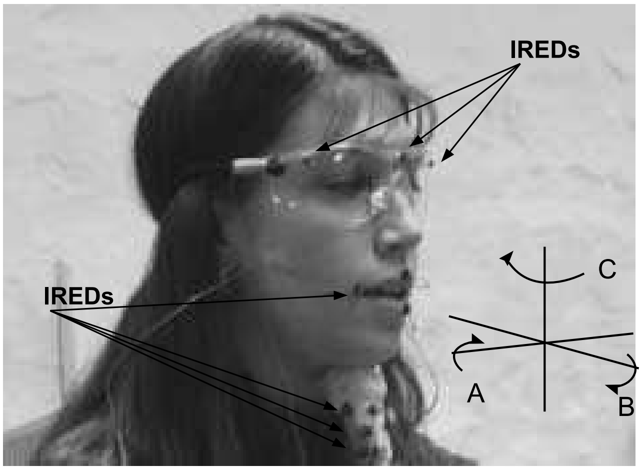
The main two pieces of equipment in the system are an ultrasound device and an optical tracking device. For the ultrasound, we currently use an Aloka SSD-1000 or an SSD-5500, though the system could function with virtually any ultrasound scanner. The data presented here were collected with the SSD-1000, recorded on videotape (with simultaneous recording of the audio signal). For the optical tracking, we use an Optotrak three-dimensional System 3020 from NDI, which also supports concurrent audio recording. In this system, a camera tracks the motion of infrared emitting diodes (IREDs) placed on the probe and head (Ostry, Vatikiotis-Bateson, & Gribble, 1997; Vatikiotis-Bateson & Ostry, 1995). As with the ultrasound, there are other tracking systems, which use either active or passive markers, that could provide the same functionality. What is important is the ability to track multiple points in three dimensions in real time.

The ultrasound transceiver we use is a 3–5-MHz variable-frequency intercostal probe. It has a viewing angle of 90° and an imaging depth of about 17 cm. This probe has a curved surface that makes it ideal for obtaining sagittal images of the tongue. The 90° setting typically allows us as full a view of a speaker's tongue as is possible. The hyoid bone is a limiting factor in the posterior direction, and the shadow of the jaw limits the view in the anterior direction.

Because the ultrasound signal reflects off of various tissues in proportion to the changes in acoustical impedance, we can only image that part of the vocal tract that has a continuous nonair medium between the transceiver and the object of interest; the impedance of the air is too great to allow the signal to pass beyond the air boundary. For the tongue, this means that the tip will not be visible whenever it projects out over an air pocket. The anterior-most point of the tongue detectable in ultrasound, however, is not far from the position of the "tongue tip" markers in point parameterization systems, which is typically 1 cm posterior to the actual tip (e.g., Engwall, 2003; Westbury, 1994b). An experiment comparing point source and ultrasound tongue tracking is currently under way and will address this issue systematically.

Knowing where the ultrasound image is in relation to the rest of the vocal tract is one of the most difficult problems to solve. It is not necessary to solve this problem for every use, as many interesting facts about the tongue shape can be derived from "transceiver-centric" images (Iskarous, Whalen, & Mattingly, 2001), but the most useful information is obtained by aligning the ultrasound image relative to the upper skull so that constrictions can be measured or inferred. In HOCUS, we track the head using at least three IREDs placed on a set of goggles. We do not place the IREDs directly on the head since there are very few flesh-points on the head that do not shift relative to the skull. Therefore, we chose a system that allows for some of

Figure 2. Location of the Optotrak infrared emitting diodes (IREDs) on the goggles, the ultrasound transceiver, and the lips. The intersecting axes, labeled A, B, and C, represent the three directions around which rotation may occur: A, rotation around the lateral dimension (pitch); B, rotation around the anterior/posterior dimension (roll); and C, rotation around the vertical dimension (yaw).



this sliding to occur beneath the goggles, while the main point of stability remains at the bridge of the nose. The goggles are attached to an adjustable elastic band that is tightened so that the goggles cannot move relative to the bridge of the nose. Figure 2 shows the placement of the IREDs on the goggles and on the transceiver.

We have typically kept the transceiver in place by hand for HOCUS, though securing the transceiver to the head with elastic bands has also worked. The handheld approach is not optimal since the hand can shift, but it does have the advantage of allowing greater freedom of the jaw because the probe is held at a fairly constant pressure. However, because the motion of the probe is tracked, it is possible to determine exactly in which frames the probe has rotated or slid out of plane, and those data can be discarded. We describe below a procedure for determining whether the magnitude of probe rotation and translation out of the midsagittal plane is acceptable. The system we are aiming for is one that involves some restraint to hold the alignment of the probe to a single plane and to keep contact with the skin; no more restraint should be needed. The less constrained the speaker, the more likely it is that her speech will resemble her natural speech patterns. Because we are able to correct for the movement of the transceiver relative to the head, our system allows greater freedom of movement and less potential impairment of the speech itself than head restraint systems allow. Furthermore, because the probe's position is tracked, bad data can be identified and error can be quantified.

The ultrasound image is collected at a sampling rate that is a function of the settings for depth of image, line resolution, and scan angle. With the SSD-1000, we often use settings that result in a 57-Hz frame rate. To record continuously in an analog setting, the rate is lowered by the VCR to 30 Hz. For short bursts of data that fit within the ultrasound machine's internal video buffer (24 frames for our current scanner), it is possible to collect the data, then output each frame of ultrasound data to the videotape so that each internal frame is recorded multiple times on the videotape. After the video frames are entered into the computer, it is possible to select a single video frame that represents an ultrasound frame. The synchronization with the sound is lost in this technique, and it is difficult to align the acquisition with a particular utterance, but it is possible to use for special circumstances in which the higher sampling rate is crucial. More sophisticated ultrasound machines support the collection of digital video, which should allow the video frame rate to be the same as the ultrasound machine-internal rate used for continuous recording. We have begun to collect data at 200 Hz using a recently acquired Aloka 5500 scanner, but results presented here are from data collected at 57 Hz, downsampled to 30 Hz by the VCR.

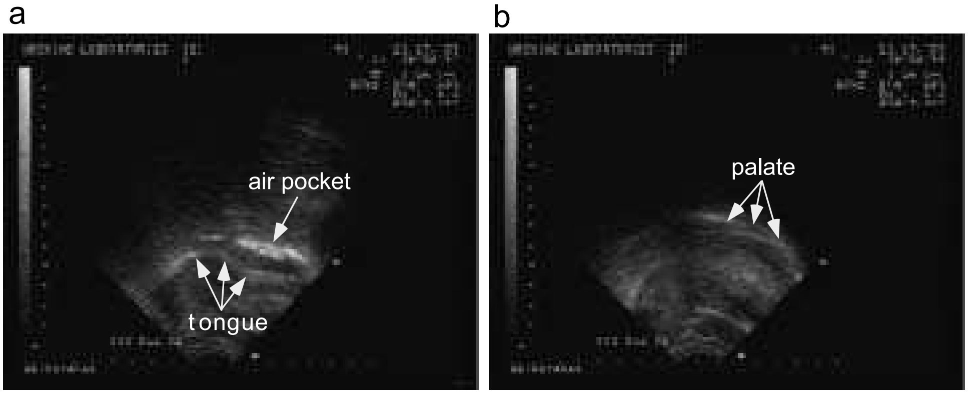
The Optotrak system supports multiple IREDs and concurrent acoustic recording. Individual IREDs are typically tracked at 200 Hz. Our version has an operating range of 2–4 m from the camera. Root-mean-square accuracy at 2 m from the camera is 0.1 mm for x- and y-coordinates and 0.15 mm for the z-coordinate. Lip markers are shown in Figure 2.

Data Collection

HOCUS sessions begin with the signing of informed consent by the participant. A coordinate system is defined with the origin at the upper incisors and a horizontal plane corresponding to the individual's occlusal plane (Westbury, 1994a). The coordinate system is established by simultaneously recording the positions of markers on the goggles and on a triangle, held between the individual's teeth, containing markers at known locations. A reference position for the probe is recorded separately, with the probe held beneath the chin and the teeth clenched. During the experiment we track the motion of the probe as a rigid body relative to the goggles and hence relative to the occlusal plane.

Next, we obtain a trace of the hard palate. This is accomplished by having the participant take a mouthful of water and force it up into contact with the hard palate; this is followed by swallowing the water. This ensures visualization of the palate because the impedance

Figure 3. Ultrasound image taken while a water bolus is being held in the mouth, allowing the imaging of the hard palate. (a) The water bolus is in the mouth, but there is a pocket of air between it and the palate, giving a false impression of where the palate is. (b) In a frame taken just as the speaker swallows, the palate is more clearly outlined.



difference between water and air results in a pattern that is visually distinct from the difference between water and bone. We instruct the participant to force the water up into contact with the hard palate, since otherwise what is visible is the air layer above the water and below the palate (see Figure 3). The palate trace can also be verified from later portions of the run when the speaker swallows (Wrench & Scobbie, 2003). Because the head is tracked during the experiment, the extracted boundary of the palate can be inserted into every frame of the speech trials.

After the palate trace, we apply whatever additional IREDs we need for tracking visible articulators. The setup (including occlusal plane determination and palate trace) generally takes about 20 min. The Optotrak data collection works with individual trials, each starting with a beep. Speech trials are usually 1–2 min long. For calibration, the ultrasound transceiver is positioned manually while the speaker pronounces some nonsense syllables, and the midsagittal position is selected based on visual inspection of the probe and the image. Externally, the probe is in line with the nose; internally, the tongue image tends to be at its most extreme value at the midline (for nongrooved sounds). There are ways of giving the holder of the transceiver more feedback about the desired orientation. For example, a laser pointer can be attached to the probe, and the speaker can keep the laser point within a target on the wall (Gick, 2002). We have not adopted that strategy, because we want the speaker to focus fully on the speech task. We also have means of detecting frames in which the probe orientation was beyond acceptable limits.

Although we have so far described the collection of midsagittal images, it is also possible to use ultrasound in the coronal direction. This can be useful in determining the presence and depth of grooving of the tongue at various individual locations. Multiple scans of a sustained phonation taken at different parasagittal offsets may be combined to produce a three-dimensional static representation of the tongue (Honorof et al., 2003). By using multiple utterances, it is also possible to build up a three-dimensional reconstruction that simulates a video clip by taking a single frame at time points in different utterances that are equivalent to the frames that would have occured in real time (Lundberg & Stone, 1999). Considerable image processing is required for such reconstructions.

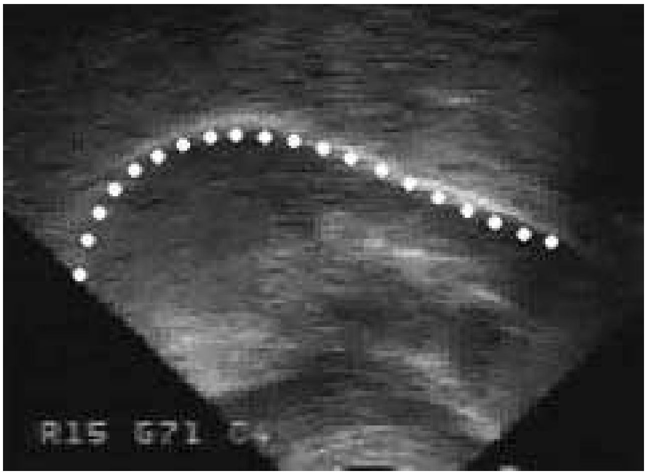
Data Analysis

The video recording of the ultrasound signal has to be digitized so that further processing can take place on the computer. (This process, of course, is avoided with the digital video system.) This portion of the process is the same as that for any video input.

Correction of the video signal assumes that the Optotrak data and ultrasound video are synchronized, which is accomplished by using the acoustic signals that are simultaneously recorded onto the VCR and Optotrak units from the same microphone. The two signals are first resampled to a common rate, then time-aligned based on cross-correlation. The last step in the synchronization is down-sampling the Optotrak signal to a 30-Hz frame rate to match the ultrasound video rate, if necessary.

The tongue surface is extracted by using a “snakes-based” (see below) procedure similar to one described by Iskarous (2005). First, a search window that contains the tongue edge and no other edges is chosen by the user by interactively manipulating the size and shape of the window using five control points. Unless there is contact between the tongue and other structures, it should

Figure 4. Spline fitting of the tongue surface. The white stripe is the air layer above the tongue edge.

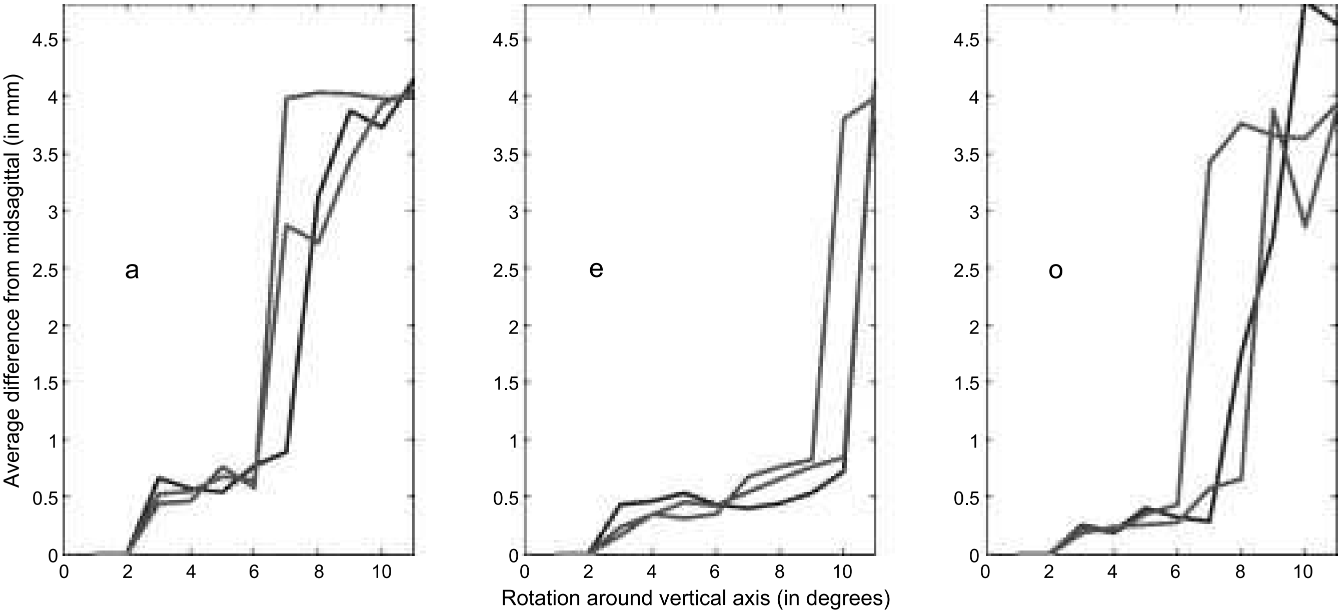


be possible to construct a window containing only the tongue surface. The points within the window with highest luminance gradient are then automatically detected, and B-splines are then fit to the gradient data by using a least squares criterion (Blake & Isard, 1998). Figure 4 shows a sample tongue surface from the ultrasound and the spline that is fitted to it. The procedure is then iteratively applied to the frames in a film.

Rigid body transformation is then performed to locate the ultrasound image of the tongue relative to the rest of the vocal tract. The camera-centric Optotrak data for each speech trial give the location in space of each of the IREDs on the probe and on the head-mounted glasses, which establish head position. To determine the rigid body coordinates (translations and rotations) of the head and probe, a set of MATLAB procedures previously developed for jaw motion detection are used (Ostry et al., 1997). A two-step optimization procedure is used to first correct for head motion relative to the camera and then to specify the motion of the probe in a head-centered coordinate system for that frame.

For each ultrasound frame, the rigid body reconstruction and correction procedures determine six numbers specifying the position and orientation of the probe for that frame. Three of these specify the vertical, lateral, and horizontal position, and the other three specify the pitch, roll, and yaw. Figure 2 shows the coordinate system used. Because the imaged parts of the tongue and the palate change directly with the movement of the probe and the head, the tongue edge and palate splines are then rotated and translated to correct for the motion of the probe and head from frame to frame. However, not all the rigid body coordinates are used for correction. For midsagittal imaging, correction to the edge is based on three of the rigid body coordinates of the probe: 1, vertical displacement; 2, anterior-posterior displacement; and 3, pitch. Variation in these three numbers can be corrected because the image is still in the midsagittal plane. The other three numbers are used to determine whether the midsagittal data are within our tolerance levels (see

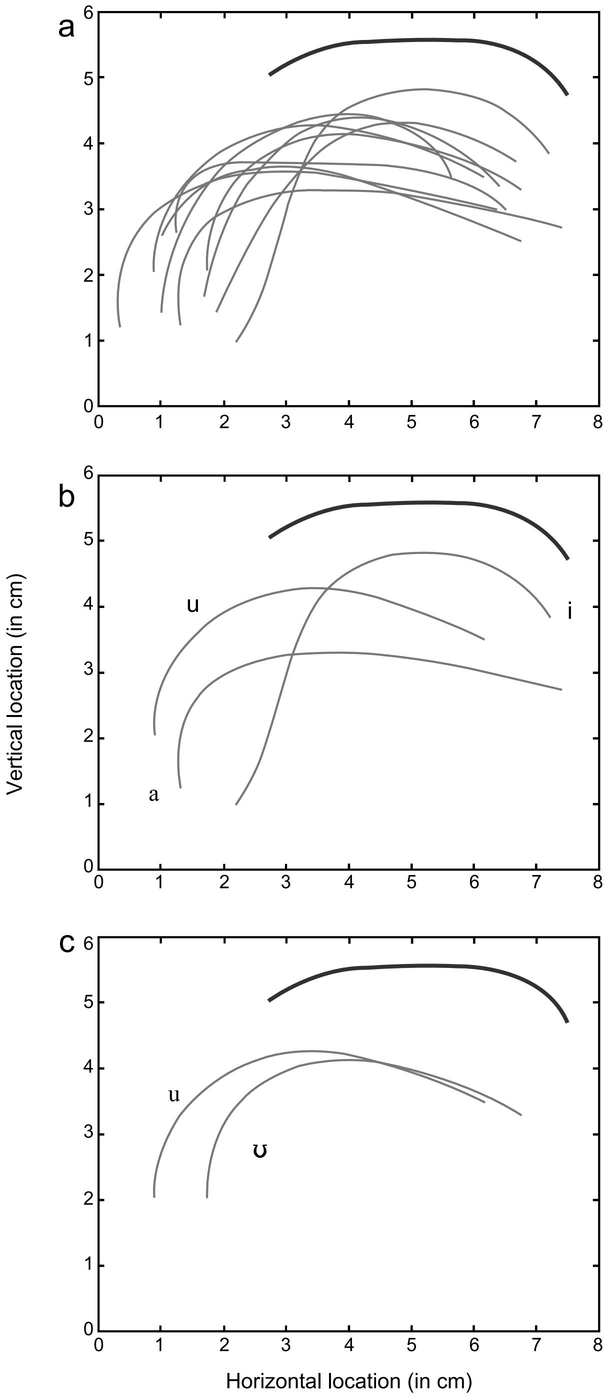
Figure 5. Measurements of average inter-spline error between 0° and the value with varying amounts of rotation around the vertical axis (yaw), for three trials for each of three vowels (/a/, /e/, and /o/).



below). Under our current data collection procedure in which the experimenter holds the probe, departure from the desired plane does occasionally occur, but we are able to exclude segments of video collected when the probe was out of alignment because the remaining three rigid body coordinates specify the lack of alignment. However, if the rotation or translation out of the midsagittal plane is of a small magnitude, the data may still be usable. We have conducted trials in which we have systematically rotated and translated the probe while measuring tongue shape and also measuring the rigid body coordinates of the probe. We then measured the difference in tongue shape of a held articulation as the probe position varied and determined thresholds beyond which the data were unusable. Specifically, by setting threshold values on the lateral position of the transceiver and roll and yaw angles, deviant frames can be discarded. Figure 5 shows an example of threshold determination. The data in this figure were collected by rotating the probe in the vertical plane (yaw) while the participant was saying /a/, /e/, and /o/. What was plotted is the average distance (in millimeters) between the midsagittal slice and every out-of-plane slice. If the probe is rotated less than about 5°, the average error is at most 0.7 mm. As the probe is rotated even more, the shape of the tongue changes greatly and the distance becomes quite large. This data are then used to set a threshold beyond which data should be discarded—in this case, the threshold is about 5°, indicating that 0.7 mm average deviation or less is acceptable. To determine if data of this type are repeatable, we performed the same experiment on the same participant three times. As can be seen from the figure, the data from the three trials are qualitatively similar. In our typical experiments, the speaker performs this task at the end of the experiment, and thresholds are set by averaging the minimal values of deviation for the three vowels. Typically, lateral motion of the transducer is harmful only if it is greater than 2–4 mm. Pitch and roll of less than about 5°–7° provides acceptable data.

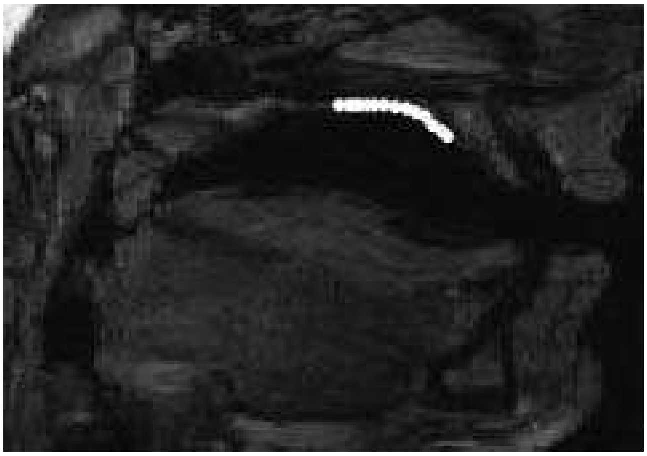
Figure 6 shows a range of tokens of English vowels in (h)Vd words (e.g., *heed, aid*) after the extracted tongue surface has been oriented to the hard palate. While it is not possible to track the individual vowels in Figure 6a, it does show the total vowel space as used by this female talker. Figure 6b shows the point vowels /a/, /i/, and /u/. Figure 6c shows the pair /uʊ/, highlighting the difference in the tongue root. The imaging in the pharyngeal region is extensive and allows us to evaluate the contribution of this part of the vocal tract to speech. If we can assume that the rear pharyngeal wall is fixed, it is even possible to measure changes in pharyngeal depth by noting changes in the location of the anterior pharyngeal wall (i.e., the tongue root).

Figure 6. Tongue shapes for English vowels. Anterior is to the right. The heavy line at the top of each panel is the palate trace. (a) Tongue surfaces for all 11 English vowels from 1 female speaker. (b) Traces for three point vowels. (c) Traces for the tense lax pair u/ʊ.



Distances between the tongue and the hard structures of the vocal tract can be computed with reasonable certainty in the palatal region, because the palate is imaged with the same system as the tongue. We are also

Figure 7. Superposition of a reconstructed palate on a magnetic resonance image of the same speaker (taken separately).



attempting to include the vocal tract obtained from MRI results for the same speakers. An example is given in Figure 7. It appears likely that posture affects the relationship between the posterior pharyngeal wall and the hard palate, so it may be that MRI would need to be done for a variety of possible postures and then matched to the posture adopted by the speaker. If an acceptable fit can be made, it should be possible to predict the posterior pharyngeal wall position from the anterior position, much as we were able to predict anterior pharyngeal tongue position fairly well from fleshpoints on the tongue body (Whalen, Kang, Magen, Fulbright, & Gore, 1999). Because the shape of the supralaryngeal vocal tract is critical for determining the acoustic transfer function, this analysis holds the promise of providing a fairly complete description of the structures involved. This can be tested with articulatory synthesis (Rubin, Baer, & Mermelstein, 1981; Rubin et al., 1996), and that synthesis can in turn be improved with this kind of data (Iskarous, Goldstein, Whalen, Tiede, & Rubin, 2003). However, changes in the position of the anterior pharyngeal wall (i.e., the tongue root) should be

Table 1. Comparison of various speech measurement systems that image the tongue.

	Feature						
System	Tongue imaging	Sampling rate	Imaging tongue root	Imaging velum	Head movement	Special features of speech	Special populations
2-dimensional magnetometry	Fleshpoints (usually 4)	200–500 Hz	No[a]	Yes[b]	Restricted	Receivers affect articulation	Fairly broad
3-dimensional magnetometry	Fleshpoints (usually 4)	Usually 200 Hz	No[a]	Yes[b]	Free	Receivers affect articulation	Fairly broad
Ultrasound	Full-length[c]	30–200 Hz[d]	Yes	No	Free	Probe slightly impinges on jaw	Broad
Ultrasound with head holder	Full-length[c]	30–200 Hz[d]	Yes	No	Restricted	Some effect on jaw	Fairly broad
Ultrasound/Optotrak (HOCUS)	Full-length[c]	30–200 Hz[d]	Yes	No	Free	Probe slightly impinges on jaw	Fairly broad
Static MRI	Full-length	—	Yes	Yes	Restricted	Supine position	Limited
Cine-MRI	Full-length	8–24 Hz	Yes	Yes	Restricted	Supine position	Limited
X-ray microbeam	Fleshpoints (usually 4–5)	40–160 Hz	No[a]	No[b]	Free	Some effects on speech[e]	Fairly broad

Note. Some of the X-ray and computed tomography systems now in use are not described. There are a variety of these systems, with different parameters, making it difficult to include them in this table. HOCUS = Haskins Optically Corrected Ultrasound System; MRI = magnetic resonance imaging.

[a] Only a few speakers with a low gag reflex are able to tolerate a pharyngeal pellet or receiver.

[b] It is possible to glue a receiver or suture a pellet to the underside of the velum, but this is rarely done.

[c] The ultrasound image extends from just above the hyoid bone to near the tip of the tongue. Whenever the signal hits air, it disappears, so any portion of the tongue tip that is over the sublingual cavity will not be imaged. Similarly, retroflex tongue shapes are not well imaged. Note, however, that the coverage of the tip may be similar to that of magnetometry and microbeam, because the receivers for the tip are placed 1 cm posterior to the actual tip; an experiment is under way to see how often this point is imaged by ultrasound.

[d] The internal combinations of settings result in a machine internal sampling rate of anywhere from 30 to 200 Hz. On most machines, the image must be recorded on videotape, which runs at 30 Hz for North American videotape. Machines with digital imaging can record at the true sampling rate.

[e] Weismer and Bunton (1999) found that only a few individuals had noticeable effects of the pellets on their production of one sample sentence. Some other subgroups had tendencies toward perceptible effects.

directly interpretable, assuming only small changes in posture, because the posterior wall has been shown not to move significantly during speech (Magen, Kang, Tiede, & Whalen, 2003).

Comparison Among Systems

The various ways of measuring the tongue have different strengths and weaknesses, as outlined in Table 1. If static imaging is the research goal, then MRI is better than other techniques, because it provides more extensive coverage of the vocal tract, but for running speech, ultrasound or point-tracking devices are superior. (Improvements in acquiring multiple MR images from a single utterance reduce this difference.) If the goal is an imaging situation most similar to natural speech conditions, the HOCUS system is ideal, because head movement is free, there is minimal pressure on the jaw, and the participant is upright. But if the goal is the collection of velocity and acceleration data on particular points of the tongue, point-tracking devices or tagged MRI are appropriate, whereas HOCUS and basic MRI are not as useful.

Resolution is often thought to be a major comparison point between ultrasound and point-tracking systems, but a direct comparison of the accuracy of magnetometer and ultrasound data found that the two agreed on the position of the points of the tongue surface to within 1.16 mm (Kaburagi & Honda, 1994a). Greater temporal resolution is an advantage of X-ray microbeam and electromagnetic articulography systems compared with ultrasound systems with only a video-recorded output, but with the advent of digital video, temporal resolution ceases to be a limiting factor on ultrasound. The main difference is in the object of measurement, tongue shapes for ultrasound and point motions for point-tracking devices. It is certainly the case that the points are easier to quantify and to work with than shapes, but methods for quantifying tongue movement from ultrasound images have shown high reliability and reproducibility (Akgul, Kambhamettu, & Stone, 1999). New methods for global quantification of tongue shape using only a few parameters (Iskarous et al., 2001) also promise to close the quantification gap. Further, there are patterns in articulator movement that cannot be seen in the movement of just a few points on the tongue, such as the pivots between adjacent segments (Iskarous, in press).

Conclusion

Recent advances in ultrasound technology have made it a useful tool for measuring speech articulation, especially for imaging of the tongue in running speech. Our system, HOCUS, which combines ultrasound with optical tracking, allows for relatively unobtrusive measurement of most of the vocal tract. The various techniques complement each other in ways that will allow us in the coming years to measure the speech articulators more completely than has ever been possible. The relative ease of use of HOCUS with normal speakers should allow for more efficient use of laboratory time, and it should make the measurement of special populations—persons with dysarthia, the elderly, and children—much more feasible as well.

Acknowledgments

This work was supported by National Institutes of Health Grant DC-02717 to Haskins Laboratories. We thank the contribution of the ATR Human Information Processing Research Laboratories to the base routines for the rigid body reconstruction. We thank Vince Gracco and two anonymous reviewers for helpful comments.

References

Abbs, J. H., & Nadler, R. D. (1987). *User's manual for the University of Wisconsin X-ray microbeam*. Madison: University of Wisconsin—Madison, Waisman Research Center.

Akgul, Y. S., Kambhamettu, C., & Stone, M. L. (1999). Automatic extraction and tracking of the tongue contours. *IEEE Transactions on Medical Imaging, 18,* 1035–1045.

Baer, T., Gore, J. C., Boyce, S. E., & Nye, P. W. (1987). Application of MRI to the analysis of speech production. *Magnetic Resonance Imaging, 5,* 1–7.

Baer, T., Gore, J. C., Gracco, L. C., & Nye, P. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America, 90,* 799–828.

Blake, A., & Isard, M. (1998). *Active contours: The application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*. London: Springer.

Carmody, F. (1941). An x-ray study of pharyngeal articulation. *University of California Publications in Modern Philology, 21*(5), 377–384.

Demolin, D., Metens, T., & Soquet, A. (2000). Real time MRI and articulatory coordinations in vowels. In P. Hoole (Ed.), *5th Seminar on Speech Production: Models and data* (pp. 93–96). Munich, Germany: Ludwig-Maximilians-Universität.

Engwall, O. (2003). Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication, 41,* 303–329.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society of London Series B, Biological Sciences, 268,* 1669–1675.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association, 32,* 113–121.

Gick, B., Iskarous, K., Whalen, D. H., & Goldstein, L. M. (2003). Constraints on variation in the production of

English /r/. In S. Palethorpe & M. Tabain (Eds.), *Proceedings of the 6th International Seminar on Speech Production* (pp. 73–78). Sydney, Australia: Macquarie University.

Honorof, D. N., Chang, C. Y. C., Iskarous, K., Tiede, M. K., Ostry, D., & Whalen, D. H. (2003, September). *Mandarin /r/ as a grooved approximant: Reconstructed tongue shape data from MRI & OPTOTRAK-ultrasound synchronization.* Paper presented at Societas Linguistica Europea, Lyons, France.

Hoole, P., & Nguyen, N. (1997). Electromagnetic articulography in coarticulation research. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 35,* 177–184.

Iskarous, K. (2005). Detecting the edge of the tongue: A tutorial. *Clinical Linguistics and Phonetics, 19,* 555–565.

Iskarous, K. (in press). Patterns of tongue movement. *Journal of Phonetics*.

Iskarous, K., Goldstein, L. M., Whalen, D. H., Tiede, M. K., & Rubin, P. E. (2003). CASY: The Haskins Configurable Articulatory speech synthesizer. In D. Recasens, M.-J. Solé, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 185–188). Barcelona, Spain: Universitat Autonoma de Barcelona.

Iskarous, K., Whalen, D. H., & Mattingly, I. G. (2001). Modeling tongue shapes with conic arcs. *Journal of the Acoustical Society of America, 110,* 2760.

Kaburagi, T., & Honda, M. (1994a). Determination of sagittal tongue shape from the positions of points on the tongue surface. *Journal of the Acoustical Society of America, 96,* 1356–1366.

Kaburagi, T., & Honda, M. (1994b). An ultrasonic method for monitoring tongue shape and the position of a fixed-point on the tongue surface. *Journal of the Acoustical Society of America, 95,* 2268–2270.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 812–832.

Kiritani, S. (1986). X-ray microbeam method for the measurement of articulatory dynamics: Techniques and results. *Speech Communication, 45,* 119–140.

Kiritani, S., Itoh, K., & Fujimura, O. (1975). Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *Journal of the Acoustical Society of America, 57,* 1516–1520.

Lakshminarayanan, A. V., Lee, S., & McCutcheon, M. J. (1991). MR imaging of the vocal tract during vowel production. *Journal of Magnetic Resonance Imaging, 1,* 71–76.

Larsson, S. G., Mancuso, A., & Hanafee, W. (1982). Computed-tomography of the tongue and floor of the mouth. *Radiology, 143,* 493–500.

Lindblom, B. E., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7,* 147–161.

Löfqvist, A., & Gracco, V. (1994). Tongue body kinematics in velar stop production: Influences of consonant voicing and vowel context. *Phonetica, 51,* 52–67.

Lundberg, A. J., & Stone, M. L. (1999). Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data. *Journal of the Acoustical Society of America, 106,* 2858–2867.

Magen, H. S., Kang, A. M., Tiede, M. K., & Whalen, D. H. (2003). Posterior pharyngeal wall position in the production of speech. *Journal of Speech, Language, and Hearing Research, 46,* 241–251.

Morrish, K. A., Stone, M. L., Sonies, B. C., Kurtz, D., & Shawker, T. (1984). Characterization of tongue shape. *Ultrasonic Imaging, 6,* 37–47.

Munhall, K. G., Ostry, D. J., & Parush, A. (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 457–474.

Munhall, K. G., Vatikiotis-Bateson, E., & Tohkura, Y. (1995). *X-ray film database for speech research* [Videodisc]. Kyoto, Japan: ATR Laboratories.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., & Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America, 115,* 1771–1776.

Niitsu, M., Kumada, M., Campeau, N. G., Niimi, S., Riederer, S. J., & Itai, Y. (1994). Tongue displacement: Visualization with rapid tagged magnetization-prepared MR-imaging. *Radiology, 191,* 578–580.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectographic measurements. *Journal of the Acoustical Society of America, 39,* 151–168.

Ostry, D. J., Keller, E., & Parush, A. (1983). Similarities in the control of speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 622–636.

Ostry, D. J., Vatikiotis-Bateson, E., & Gribble, P. L. (1997). An examination of the degrees of freedom of human jaw motion in speech and mastication. *Journal of Speech, Language, and Hearing Research, 40,* 1341–1351.

Parush, A., Ostry, D. J., & Munhall, K. G. (1983). A kinematic study of lingual coarticulation in VCV sequences. *Journal of the Acoustical Society of America, 74,* 1115–1125.

Perkell, J. S. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies M. L., Garabieta, I., & Jackson, M. T. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America, 92,* 3078–3096.

Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America, 112,* 1627–1641.

Pouplier, M., & Goldstein, L. M. (2002). Asymmetries in speech errors: Production, perception and the question of underspecification. In T. A. Hall, B. Pompino-Marschall,

& M. Rochon (Eds.), *Papers on phonetics and phonology: The articulation, acoustics and perception of consonants* (pp. 73–82). Berlin, Germany: ZAS.

Rochette, C. (1973). *Les groupes de consonnes en français*. Québec, Canada: Les Presses de l'Université Laval.

Rokkaku, M., Hashimoto, K., Imaizumi, S., Niimi, S., & Kiritani, S. (1986). Measurement of the three-dimensional shape of the vocal tract based on the magnetic resonance imaging technique. *Annual Bulletin RILP, 20,* 47–54.

Rubin, P. E., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70,* 321–328.

Rubin, P. E., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., & Browman, C. (1996, May). *CASY and extensions to the task-dynamic model.* Paper presented at the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling and 4th Speech Production Seminar, Autrans, France.

Russell, G. O. (1928). *The vowel: Its physiological mechanism as shown by x-ray*. Columbus: Ohio State University Press.

Schönle, P., Grabe, K., Wenig, P., Hohne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language, 31,* 26–35.

Stark, J., Ericsdotter, C., Branderud, P., Sundberg, J., Lundberg, H.-J., & Lander, J. (1999). The APEX model as a tool in the specification of speaker specific articulatory behavior. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 2279–2282). San Francisco: University of California, Berkeley.

Stevens, K. N., & Öhman, S. E. G. (1963). Cineradiographic studies of speech. *KTH STL-QPSR, 2,* 9–11.

Stone, M. L. (1997). Laboratory techniques for investigating speech articulation. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 11–32). Oxford, England: Blackwell.

Stone, M. L., & Davis, E. P. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *Journal of the Acoustical Society of America, 98,* 3107–3112.

Stone, M. L., Davis, E. P., Douglas, A. S., Aiver, M. N., Gullapalli, R., Levine, W. S., et al. (2001). Modeling tongue surface contours from cine-MRI images. *Journal of Speech, Language, and Hearing Research, 44,* 1026–1040.

Stone, M. L., Epstein, M., & Iskarous, K. (2004). Functional segments in tongue movement. *Clinical Linguistics and Phonetics, 18,* 507–521.

Stone, M. L., Faber, A., Raphael, L. J., & Shawker, T. H. (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [S], and [l]. *Journal of Phonetics, 20,* 253–270.

Stone, M. L., & Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America, 99,* 3728–3737.

Stone, M. L., Sonies, B. C., Shawker, T. H., Weiss, G., & Nadel, L. (1983). Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *Journal of Phonetics, 11,* 207–218.

Stutley, J., Cooke, J., & Parsons, C. (1989). Normal CT anatomy of the tongue, floor of mouth and oropharynx. *Clinical Radiology, 40,* 248–253.

van Lieshout, P. H. H. M., Alfonso, P. J., Hulstijn, W., & Peters, H. F. M. (1993). Electromagnetic articulography (EMA) in stuttering research. *Institut für Phonetik und sprachliche Kommunikation der Universität München — Forschungsberichte, 31,* 215–224.

Vatikiotis-Bateson, E., & Ostry, D. J. (1995). An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics, 23,* 101–117.

Weismer, G., & Bunton, K. (1999). Influences of pellet markers on speech production behavior: Acoustical and perceptual measures. *Journal of the Acoustical Society of America, 105,* 2882–2894.

Weismer, G., Yunusova, Y., & Westbury, J. R. (2003). Interarticulator coordination in dysarthria: An X-ray microbeam study. *Journal of Speech, Language, and Hearing Research, 46,* 1247–1261.

Westbury, J. R. (1994a). On coordinate systems and the representation of articulatory movements. *Journal of the Acoustical Society of America, 95,* 2271–2273.

Westbury, J. R. (1994b). X-ray microbeam speech production database user's handbook [Software manual]. Madison: University of Wisconsin—Madison, Waisman Research Center.

Whalen, D. H., Kang, A. M., Magen, H., Fulbright, R. K., & Gore, J. C. (1999). Predicting pharynx shape from tongue position during vowel production. *Journal of Speech, Language, and Hearing Research, 42,* 592–603.

Wood, S. (1982). *X-ray and model studies of vowel articulation* (Working Paper No. 23). Lund, Sweden: Lund University, Department of Linguistics.

Wrench, A. A., & Scobbie, J. M. (2003). Categorising vocalisation of English /l/ using EPG, EMA and ultrasound. In S. Palethorpe & M. Tabain (Eds.), *Proceedings of the 6th International Seminar on Speech Production* (pp. 314–319). Sydney, Australia: Macquarie University.

Received May 8, 2004

Accepted October 22, 2004

DOI: 10.1044/1092-4388(2005/037)

Contact author: D. H. Whalen, Haskins Laboratories, 300 George Street, New Haven, CT 06511.
E-mail: whalen@haskins.yale.edu