En Route to the Three-Dimensional Registration and Analysis of Speech Movements: Instrumental Techniques for the Study of Articulatory Kinematics

Margaret M. Earnest
University of Connecticut, Storrs

Ludo Max
University of Connecticut, Storrs
Haskins Laboratories, New Haven, CT

Instrumental techniques used to transduce respiratory, phonatory, or articulatory movements have contributed invaluable insights into the physiological processes underlying speech production. Although a number of creative techniques were developed for some of the earliest studies in this area (see R. H. Stetson’s work in Kelso & Munhall, 1988; Kozhevnikov & Chistovich, 1965), most instruments currently in use for this purpose reflect the immense technological advances that have occurred throughout the last three decades. Some examples of the earliest studies in this area (see R. H. Stetson’s work in Kelso & Munhall, 1988; Kozhevnikov & Chistovich, 1965), most instruments currently in use for this purpose reflect the immense technological advances that have occurred throughout the last three decades. Some examples

ABSTRACT: Increasing our understanding of the physiological processes underlying speech production is crucial for both theoretical and clinical reasons. Such gains in theory and clinical application, however, depend on the availability and use of technologically advanced instruments that can accurately reveal the processes of interest. Consequently, it is essential that not only researchers but also clinicians and students in communication sciences and disorders have access to up-to-date information about the most recent developments in the area of instrumental analyses of speech production. Those professionals involved in research continually need to make important decisions regarding the most appropriate instruments to address their specific research questions, whereas those in clinical practice need at least a basic understanding of the same information to interpret the contemporary literature and to decide when a specific instrument is ready to be used for diagnosis or treatment. Therefore, it is unfortunate that, despite numerous rapid and exciting advances in analog and digital technology, instrumental speech analysis procedures receive relatively little attention in most speech-language pathology students’ academic curriculum. Hence, the purpose of this paper is to provide a summary of this important information by presenting an overview of various instrumental techniques that are being used in several laboratories or that are currently in an advanced stage of development. To limit the scope of this work, only instruments suitable for the transduction of articulatory movements are presented. Specific applications, capabilities, and limitations of each of the systems are discussed.

KEY WORDS: speech kinematics, instrumentation, articulation.
of such technological advances include the development of inexpensive microcomputers with fast processors and large data storage capacity, the miniaturization of a wide variety of electronic components and circuits, and the development of an array of noninvasive medical imaging technologies. Most recently, this technological progress has allowed considerable improvements in already established and widely used instruments, as well as the development of several very promising new instruments based on previously unexplored technologies. Consequently, the field of communication sciences and disorders now has at its disposal a relatively wide range of fast and efficient methods for collecting speech movement data with high precision.

Clearly, familiarity with the available instruments is extremely important for those directly involved in research as important decisions need to be made regarding selection of the most appropriate tools for each research endeavor. When making judgments regarding the appropriateness of an instrumental method, it is obviously essential to consider the capabilities and constraints of the instrument in light of both the research question and the target population. It is our position, however, that it is equally important for clinicians and students to also have at least a basic understanding of the capabilities, advantages, and limitations of the available instruments. Indeed, access to the most up-to-date instrumentation-related information is necessary for clinicians and students in order to (a) be able to interpret and critically evaluate the literature on both normal and disordered speech, and (b) pursue innovative applications of such advanced instrumentation in the clinical management of children and adults with various speech disorders. Unfortunately, the rapid nature of current technological developments, the absence of instrumentation courses in many undergraduate and graduate programs in communication sciences and disorders, and often very demanding clinical caseloads appear to prevent the large majority of clinicians and students from familiarizing themselves with this information.

The aim of this paper is to provide both practicing speech-language pathologists and students in communication sciences and disorders—regardless of whether they also consider themselves researchers, prospective future researchers, or research consumers—with an overview of several instrumental techniques that are already well-established or that are currently in an advanced stage of development (and, hopefully, ready for application in the near future). Speech production is an extremely complex process, and instrumental methods are available to assess each of the subsystems involved in this process (i.e., respiration, phonation, articulation). Discussing all available methods for each of those subsystems is impossible within the space constraints of a single paper in a scholarly journal. Therefore, we focus here only on those instruments that are directly applicable to the study of speech articulation (but note that we do not wish to imply that these instruments or this subsystem are of greater clinical or research importance). In particular, we will limit this overview to instruments transducing articulatory movements or articulatory contact, thereby excluding, for example, instruments that transduce the acoustic consequences of those movements or the muscle activation patterns generated to perform the movements. Instruments included in this overview make use of imaging techniques (i.e., providing visual images of the structures of interest), point-tracking techniques (i.e., tracking movements of individual points on the structures of interest), or contact detection techniques (i.e., detecting the location, extent, and time of contact between two structures). For each system, we describe the general design and principles of operation as well as some of the major strengths and limitations that have been identified.

CINERADIOGRAPHY

Cineradiography uses a combination of X-ray and motion picture film techniques to produce dynamic images of all articulators (i.e., lips, jaw, tongue, velum) during speech production. This procedure has been used for approximately four decades to study articulatory movements. Because of safety concerns related to the amount of radiation exposure, its use has decreased recently, and modified X-ray-based techniques have been developed as safer and less invasive alternatives (see below).

X-ray images are produced when a high-frequency electromagnetic field is projected through the head in order to detect various articulatory structures (Figure 1). Detection of those structures is possible because X rays are able to pass through the various tissues, but they are absorbed differently by each type of tissue. The greater the density of the tissue, the more the X rays will be absorbed. This phenomenon makes it possible to identify the different structures (to the extent that they consist of different types of tissue) based on the amount of absorption or “shadow” on a detector surface located on the other side of the head. For example, both illustrations included in Figure 1 show lateral vocal tract images obtained with X-ray technology (a speaker seen from the left in the left panel and a speaker seen from the right in the right panel). Tissues that are relatively dense, including the dentition and mandible, appear very clearly on these images as bright white shapes. Soft tissues, including the lips and velum, appear less clearly as grey shadows. As can be seen in the figure, X-ray imaging is powerful enough to depict most structures, including bone and air-filled cavities, within the vocal tract. In cineradiography, these black and white X-ray images are recorded onto motion picture film, often with a camera speed in the range from 50 to 150 frames per second (Beautemps, Badin, & Bailly, 2001; Daniloff & Moll, 1968; Kent & Moll, 1972). For the study of speech production, lateral images such as those shown in Figure 1 are typically recorded, although frontal imaging also is possible.

Cineradiography has been used in the past to obtain large amounts of data regarding articulatory movements during speech production. It has been used most commonly to study normal speech physiology (Kuehn & Moll, 1976; Perkell, 1969), and it has been applied as a means for testing the accuracy of other techniques (Kuehn, Reich, & Jordan, 1980). In addition, cineradiography has been widely
used to study disorders of speech production, including the speech of individuals with dysarthria, developmental articulation disorders, cleft palate, and hearing impairment (Kent & Netsell, 1975; Kuehn & Tomblin, 1977; Subtelny, Li, Whitehead, & Subtelny, 1989; Tanimoto, Henningsson, Isberg, & Ren, 1994). The main limitation of cineradiography is, of course, that it exposes subjects to relatively large doses of radiation, thus posing a potential health risk. Consequently, the technique is only appropriate when recording time is limited to a few minutes and the radiation intensity is kept sufficiently low to reduce the amount of X-ray exposure (Fujimura, 1980). A second limitation is that articulators consisting of soft tissue (i.e., lips, tongue, and velum) are not imaged as clearly as the teeth and mandible, and that, in fact, they may be obscured by the more dense structures. This limitation is illustrated well by the images shown in Figure 1. In both images, the teeth appear as bright white shapes, whereas the lips are much more difficult to identify. In the panel on the right, the tongue is obscured from view by the dentition. For this reason, some investigators have used radio-opaque pellets (Perkell, 1969) or a contrasting substance (McClean, 1973) to increase soft tissue structures’ visibility on the X-ray images.

**STRAIN GAUGE SYSTEMS**

Abbs and colleagues (Abbs & Gilbert, 1973; Müller & Abbs, 1979) described the use of strain gauges to obtain quantitative data regarding articulatory movements of the lips and jaw during speech tasks. With a number of modifications to the original design, strain gauges have been used extensively in low-cost point-tracking systems for the transduction of articulatory movements of the lips and jaw in one (superior–inferior) or two (superior–inferior and anterior–posterior) dimensions (Barlow, Cole, & Abbs, 1983). The obtained data consist of a time series (or two time series if two-dimensional data are collected) of the position of each flesh point to which a transducer is attached. At the time of this writing, more than three decades after their initial application to the study of articulatory kinematics, strain gauges continue to be used for the same purpose (e.g., Shaiman, 2001).

Strain gauge systems make use of a cantilever beam, consisting of a horizontally oriented, flexible metal strip that is fixed to a stable support structure at one end and connected to a rigid wire at the other end. The free end of the rigid wire is then attached to the articulator. The flexible metal strip is instrumented with strain gauges on both sides (i.e., top and bottom) to register superior–inferior articulatory movements. For two-dimensional transducers, a second metal strip is placed in between the first strip and the rigid wire, with its orientation perpendicular to that of the first strip and with strain gauges again on both sides (i.e., front and back) for the simultaneous registration of anterior–posterior articulatory movements (Müller & Abbs, 1979). Realistic and schematic illustrations of this design and the associated electronics are readily available in several previous publications (see, among others, Abbs & Gilbert, 1973; Baken & Orlikoff, 2000; Barlow et al., 1983).

The strain gauges are resistors that are electrically connected to other resistors located within a physiological amplifier such that this combination of resistors forms a Wheatstone bridge. Movement of the articulator to which the rigid wire is attached causes bending of the metal strip, and this, in turn, causes the strain gauges to be either stretched or compressed, depending on the direction of movement. These stretches and compressions result in time-varying alterations in the strain gauge’s resistance, and, thus, in the output voltage of the Wheatstone bridge. Given that the remaining resistors in the Wheatstone bridge have a fixed resistance, the bridge provides an analog output signal that is proportional to the extent of movement by the articulator.

In the original design, attachment of one end of the cantilever beam to a stable support structure was accomplished by mounting the entire system to a wall, and the subject was required to wear a head restraint (Abbs & Gilbert, 1973). This restraint was uncomfortable, especially for neurologically impaired subjects who exhibit frequent...
involuntary head movements. In addition, it was determined that the head restraint interfered with the naturalness of speech production, because most individuals do show head movements while speaking. Therefore, Barlow et al. (1983) designed a lightweight head-mounted frame that allows accurate data acquisition while permitting the subject’s head to move freely. This frame is adjustable and is positioned on the subject’s head using soft rubber contacts located at five nonmuscular sites of the skull. The three transducers for upper lip, lower lip, and jaw are attached to the frame by means of a support panel and an L-shaped spindle. This setup with a head-mounted frame provides a fixed position of the transducers relative to the head. Again, we refer the reader to the excellent illustrations that have been provided by other authors (Baken & Orlikoff, 2000; Barlow et al., 1983).

Coupling of the transducers to the articulators is usually accomplished by inserting the rigid wire of each transducer through a small bead that is attached to the skin with double-sided adhesive tape. Use of the bead allows almost unrestricted lateral movements of the articulator, and it eliminates difficulties caused by bending of the transducers because of such lateral movements. Lip transducers are attached midsagittally at the vermilion border of the upper and lower lip. The jaw transducer can be attached midsagittally at a point on the undersurface of the chin. This placement can be problematic, however, because the skin of the chin can move independently of the jaw, causing inaccurate transduction of jaw movement (Kuehn et al., 1980). Consequently, great caution must be taken to choose a point on the undersurface of the chin where skin movement is minimal. A tooth-mounted placement has been recommended to eliminate the impact of skin movement (Kuehn et al., 1980). This can be done by placing a dental appliance, with an attached wire, on the lower teeth. The wire is bent to leave the mouth through one of the corners in such a way that its free end is in front of the chin. The strain gauge transducer is attached to the free end of the wire. Although this placement provides accurate jaw movement data, it makes the procedure more cumbersome because a dental appliance needs to be custom-made for each subject. In addition, the wire extending from the dental appliance may interfere with the lower lip movements for bilabial and labiodental consonants.

Strain gauge systems provide a convenient way of studying speech motor control in normal as well as clinical populations. They have been used to obtain valuable information regarding the development of speech motor control in children (Smith & Gartenberg, 1984; Smith & McLean-Muse, 1987) and the kinematic consequences of various conditions affecting speech motor control, including dysarthria, stuttering, and hearing impairment (Caruso, Abbas, & Gracco, 1988; McClean, Levandowski & Cord, 1994; Murdoch, Johnson, & Theodoros, 1997; Tye-Murray & Follkins, 1990). Researchers continue to make use of these systems because they are relatively inexpensive and they provide a noninvasive and minimally obtrusive way of recording kinematic aspects of lip and jaw movements. In addition, they make it possible to collect kinematic data simultaneously with electromyographic or aerodynamic data. Calibration of the transducers is completed statically with a micrometer or dynamically with an electromechanical oscillator (Barlow, Finan, Andreatta, & Paseman, 1997). The major disadvantage of cantilever strain gauge transducers is that they cannot be used for tracking movements inside the oral cavity such as those of the tongue and velum. Additionally, the obtained lip and jaw data are dependent on the orientation of the transducers relative to the axis of movement. Given that the transducers are oriented manually by the researcher based on the visually estimated movement axis, orientation of the transducers can form a source of measurement error. In other words, strain gauge data are not referenced to a standard coordinate system as is recommended when comparing articulatory data among sessions, subjects, and studies (Westbury, 1994a).

### X-RAY MICROBEAM SYSTEM

X-ray microbeam systems track two-dimensional movements of small (2–3 mm diameter) radio-dense (gold) pellets attached to various points on the articulators. The technology was developed as a safer alternative to cineradiography as it uses only a very narrow (.4 mm diameter) beam of X rays aimed at specific points (approximately 6 mm square) on the articulators rather than a broad field of X rays directed at the entire head. An initial prototype X-ray microbeam system was developed at the University of Tokyo in 1966. Subsequently, an additional system was developed at the University of Wisconsin–Madison (Westbury, 1994b). These are the only two X-ray microbeam systems currently in existence. However, an extensive database of X-ray microbeam data is available to the scientific community at no cost. The database, developed and distributed by the research group at the University of Wisconsin–Madison, contains X-ray microbeam data obtained from 57 normal speakers during a wide range of speech and nonspeech tasks.

To track two-dimensional articulatory movements in the midsagittal plane, the computer-controlled system directs a beam of X rays first through a pinhole with a very narrow diameter and then at pellets attached to the articulators as well as reference pellets attached to nonmoving (relative to the skull) locations such as the bridge of the nose and the maxillary incisors. The articulatory pellets are typically placed on the upper lip, lower lip, two locations on the mandible, and four points along the midline of the tongue. The system identifies the location of a pellet based on the point of maximum absorption (i.e., “shadow”) on a detector surface located on the side of the head that is opposite to the pinhole. The coordinates are then determined for the center of the pellet’s shadow. After data processing, movements of the articulatory pellets are re-expressed relative to an anatomically based coordinate system with the x-axis corresponding to the occlusal plane and the origin located at the tip of the upper incisors (Westbury, 1991). This has the important advantage that, because of the fact that all planes and axes are defined based on cranial structures, the same coordinate system can be established for all speakers.
To direct the microbeam when scanning for each pellet, the computer estimates the position of a given pellet based on its previous positions. Differential sampling rates are possible across pellets, with a maximum aggregate sampling rate of approximately 700 samples per second. Typically, pellets are scanned at a rate ranging from 160 samples per second for the tongue tip to 40 samples per second for mandibular and upper lip pellets. However, in data streams that have been further processed (such as those in the aforementioned database), the original pellet time series have been interpolated and resampled to a rate of 160 samples per second for each pellet (Westbury, 1994b).

The data obtained with X-ray microbeam systems consist of x- and y-axis time series of the position of each pellet. Typical ways of displaying these data include plotting the x- and/or y-axis data on an actual time scale (as illustrated in the top panel of Figure 2, which shows the x- and y-axis movements of a tongue tip pellet) or plotting the x- and y-axis data simultaneously in a spatial plot representing the midsagittal plane (as illustrated in the bottom panel of Figure 2, which shows the position of eight pellets at the time point indicated with a vertical cursor in the acoustic waveform). The latter plot can show the position of each pellet at one point in time (as is done in the bottom panel of Figure 2) or, alternatively, the entire trajectory of the pellet.

Westbury (1994b) reported that the average spatial resolution error for the Wisconsin microbeam system during stationary tracking is 0.15 mm. That is, the resolution of the system is such that there is no more than a 0.15 mm discrepancy between the measured and the actual pellet position. The resolution for determining articulatory positions during dynamic tracking has not been determined, but can be expected to be larger. In addition, the accuracy of X-ray microbeam recordings depends on the position of the pellets relative to the midsagittal plane. Movement of the pellet outside of the midsagittal plane constitutes one situation that can result in pellet mistracking. In general, pellet mistracking may occur in two ways. First, the beam may not be able to find the pellet, based on the predicted position, resulting in a loss of data for that pellet. Second, it is possible for the beam to track a different pellet. Because of their close proximity, tracking of the incorrect pellet occurs most frequently for the four tongue pellets. Typically, mistracking of a pellet occurs only for a single pellet and for a small portion of the movement trajectory. Therefore, the proportion of recording time that involves mistracking errors is very small (Westbury, 1994b).

The main advantage of the X-ray microbeam system over other radiographic techniques is the large reduction of subject exposure to radiation. As mentioned, the system employs only a narrow beam of radiation rather than a broad field. This allows for longer data recording sessions per subject without increased health risks. Tracking a given pellet for 1 minute with a sampling rate of 100 samples per second exposes the tissue in a 30–40 mm² area surrounding that pellet for only 60 ms. Nevertheless, because the X-ray microbeam system still involves exposure to potentially harmful radiation, recording duration remains limited. During a typical recording session with a total tracking time of 1200 seconds and a sampling rate of 160 samples per second, subjects are exposed to approximately the same level of radiation as would be expected during routine diagnostic dental procedures such as bitewing X rays (Westbury, 1994b).

Obviously, one disadvantage of the X-ray microbeam technique is the highly complex technology and associated high cost, limiting its availability to only two systems worldwide. Nevertheless, it has been used for a large number of kinematic studies. Many investigators have made use of the Wisconsin database to study various aspects of normal speech physiology (Adams, Weismer, & Kent, 1993; Ostry, Flanagan, Feldman, & Munhall, 1992; Ostry & Munhall, 1994). Also interesting is the work of Stone (1990), who used X-ray microbeam data in conjunction with imaging techniques to develop a three-dimensional model of tongue movement during isolated speech tasks. A relatively small number of X-ray microbeam studies have been conducted on populations with communication disorders, including dysarthria, apraxia, stuttering, and hearing impairment (Alfonso, Watson, & Baer, 1987; Hirose, Kiritani, Ushijima, & Sawashima, 1978; Itoh, Sasanuma, Hirose, Yoshioka, & Ushijima, 1980; Tye-Murray, 1991).

In addition to the limited availability of the system, a second potential limitation is related to the fact that, in order to minimize possible effects on speech naturalness, the subject’s head typically is not restrained during data collection, but head movement may have an unwanted influence on the registered data. However, using mathematical methods to track three-dimensional movements of the head with the two-dimensional microbeam system, Westbury (1991) determined that maximum relative errors resulting from head movements or improper head positioning were never greater than 5%. A third factor that may be considered a potential limitation of X-ray microbeam systems is that placement of the gold pellets on the articulators, and particularly the tongue, may cause distorted productions of certain speech sounds. Acoustic and perceptual analyses of the recorded audio signals, however, have indicated that significant distortions occurred only on a relatively small proportion of the data (Weismer & Bunton, 1999).

### ULTRASOUND

Ultrasound is an imaging technique that has been employed as a safe, noninvasive method of studying tongue position, shape, and movement during speech production. Because exposure to ultrasound involves no known risks to humans, the procedure is especially appropriate when it is necessary to collect large amounts of data. With ultrasound, the tongue and its movements can be investigated using instrumentation that is positioned completely outside the vocal tract and that, therefore, does not interfere with the tongue movements (although it may interfere to some extent with mandible movement). In studies of speech articulation, ultrasound has been used primarily to yield information regarding the shape of the tongue in static
Figure 2. Typical data displays for X-ray microbeam (and electromagnetic articulography) data for a production of the word “school.” The top panel shows, from top to bottom, the acoustic waveform, displacement along the x-axis of a pellet attached to the tongue tip, displacement along the y-axis of the same pellet, and velocity for movement along the y-axis for the same pellet. The bottom panel shows the acoustic waveform and the position of eight pellets at the time point indicated with a vertical cursor in the acoustic waveform (UL = upper lip, LL = lower lip, MI = mandibular incisor, MM = mandibular molar, T1/2/3/4 = tongue).The x-axis is aligned to the occlusal plane; the y-axis is positioned at the tip of the upper incisors.
positions (e.g., Stone & Lundberg, 1996), the shape of the tongue at a single time point within a movement (e.g., Stone, Shawker, Talbot, & Rich, 1988), or the kinematics of tongue dorsum movements (e.g., Keller & Ostry, 1983).

Ultrasound techniques make use of a “transceiver” that contains an array of crystals serving as both transmitters and receivers. The transmitters generate sound waves whose frequencies are above the threshold of human hearing. These sound waves are reflected (echoed) when they encounter a change in acoustic impedance. Changes in acoustic impedance are caused by changes in the density of the medium through which the waves are traveling. The reflected waves are then captured by the array of receivers. The distance between the transceiver and the point of reflection can be computed based on the duration of the interval between signal emission and reception.

When applied to monitor tongue position or movement, the acoustic signal is presented by a transceiver positioned below the mandible and passed upward through the skin and the soft tissue of the tongue. Therefore, a conducting gel is used, and the transceiver may be positioned on a spring-loaded support structure (Stone et al., 1988). As mentioned, the sound waves are reflected back to the receivers when they encounter a change in acoustic impedance. In the oral cavity, the largest change in density is formed by the tissue–air interface at the surface of the tongue because air is a very poor conductor of ultrasound waves. In other words, the border between bodily tissues (high density) and air (low density) represents an area characterized by a sudden and large change in conducting capacity, and the ultrasound waves will be reflected at this location. The latency of the echo can be converted mathematically into a signal representing the position of the point of reflection (in this case, the surface of the tongue).

Ultrasound images can be obtained in different modes. The two most common modes are B mode and M mode. B mode (brightness mode) is the visualization method that is best known from obstetrics and cardiology: The echoes are represented visually in the form of a two-dimensional grey-scale image. In this format, all transmitters in the array emit repeated pulses into the tissue, and all receivers in the array detect echoes. As a result, this mode allows imaging of the entire section (a sagittal or coronal slice) of the tongue that falls within the ultrasound field. On the image, the locations that show a change in acoustic impedance (e.g., upper surface of the tongue) appear as light areas. In particular, the tongue surface is visible as the lower edge of a relatively bright white line. The left panel in Figure 3 provides an example of B-mode imaging of a speaker’s tongue during production of the sound /r/ (the tip of the tongue is toward the right). The shape of the tongue

![Figure 3. B-mode (left) and M-mode (right) ultrasound images of tongue position during /r/ articulation. The B-mode image shows the shape of the upper surface of the tongue as a bright white line. The M-mode image was obtained at the spatial location indicated by the dashed vertical cursor in the B-mode image, and shows the time-varying distance between tongue surface and palate at that point in the oral cavity.](image)

Earnest • Max: Analysis of Articulatory Movements 11
surface can be seen clearly as a white line. This line is the result of a detected change in acoustic impedance (i.e., change in tissue density) as the ultrasound signal encountered air at the tongue surface. Mathematical techniques of edge detection and curve fitting may be applied to the images to estimate the exact location and shape of the tongue surface. If multiple images are taken sequentially, a two-dimensional moving image of the tongue surface movement is obtained. This image can be stored on motion film or videotape or digitized directly into a computer for later analysis. Measurement error using B-mode ultrasound has been reported to be less than 0.7 mm (Stone, 1996).

The vertical dashed line in the B-mode image represents a location at which an M-mode image was taken. The data obtained with M-mode ultrasound are shown in the right-side panel of Figure 3. Although M-mode (transmission mode) also allows one to track the displacement of the tongue over time, the obtained data, again a bright white line, represent a one-dimensional time series instead of an image taken at one point in time. In M mode, information is obtained not about an entire tongue slice, but about the time-varying position of one single location within the oral cavity. Importantly, this single location does not represent a specific point on the tongue surface. That is, if a given point on the tongue surface moves anteriorly or posteriorly, the tracking technique does not follow this specific point. Rather, it tracks changes in position of the tongue surface at one location in the oral cavity regardless of which point along the surface of the tongue is present at that location at that time. As such, M-mode ultrasound presents a time series of the superior–inferior movement of the tongue surface at that point in the oral cavity (see the bright white line in the right-side panel of Figure 3). Given that this movement occurs toward and away from the nonmoving palate, inferences can be made regarding changes in the proximity of the tongue with relation to the palate. Although M-mode ultrasound does not track a specific flesh point on the surface of the tongue, it provides data regarding the varying degree of constriction at a specific location in the vocal tract. Accordingly, those data are closely related to the cross-sectional area of the vocal tract at that location—an important variable in determining the acoustic characteristics of the produced speech sounds.

To further improve the value of ultrasound, Stone and colleagues have developed innovative procedures that allow three-dimensional representations of tongue shape. For example, Stone and Lundberg (1996) described the reconstruction of such three-dimensional tongue shapes based on images obtained with a curvilinear array of 128 crystals. Moving the transceiver by means of a motorized pivot, 60 coronal slices were recorded over a total interval of approximately 10 seconds. Computer software then reconstructed the slices into a three-dimensional tongue surface. Later, Lundberg and Stone (1999) were able to reconstruct three-dimensional tongue surfaces using as few as six coronal slices. The location of the six coronal slices was determined for each subject separately on the basis of a series of midsagittal slices.

Ultrasound machines are designed for various clinical applications. Typically, the transceiver is hand-held during such clinical use. However, even a small movement of the hand, or of the subject’s head, causes rotation or translation of the transceiver and, therefore, off-plane imaging. In addition, stable manual positioning of the transceiver while the jaw moves for speech production is difficult to accomplish. Hence, the validity and reliability of ultrasound measures is clearly at risk when positioning the transceiver by hand, and this prevents quantitative measures to be compared across subjects or across recording sessions within subjects. Ideally, the same position of the transducer relative to the articulator should be maintained. Therefore, a head harness is sometimes employed to maintain the transducer position (Keller & Ostry, 1983). Stone and Davis (1995) designed a head and transducer support system that immobilizes the head and supports the transducer in a stable position with known relationship to the head.

Because it is a safe and relatively accessible imaging method for transduction of tongue activity, ultrasound can be applied with a variety of populations. It has been used to assess the effects of age, anatomy, speaking task, and speaking rate on the kinematics of tongue movement (Ostry, Feltham, & Munhall, 1984; Ostry & Munhall, 1985). Ultrasound techniques also have been used extensively in the study of tongue movement during swallowing (e.g., Litvan, Sastry, & Sonies, 1997; Stone & Shawker, 1986). Importantly, because of its real-time imaging capabilities, ultrasound can be used clinically for biofeedback in speech therapy (Shawker & Sonies, 1985), although its potential for application in the clinic has not been fully explored at this time.

Despite its many advantages, ultrasound is limited to studies focusing on tongue position, shape, or movement. Additionally, because bony structures are not revealed in ultrasound images, the exact distance between the tongue and the mandible or hard palate cannot be measured. Furthermore, ultrasound images provide little or no information about the most anterior portions of the tongue. Although the availability of information about an entire structure rather than a single flesh point is an obvious advantage, it also poses a potential problem in that a selected set of meaningful parameters for data extraction needs to be determined.

## OPTICAL TRACKING SYSTEMS

Similar to strain gauge systems, optical tracking systems allow one to track specific flesh points on extraoral structures (i.e., lips and jaw). Optical tracking systems can be broadly classified into two main categories: those that use passive markers and those that use active markers.

Passive marker systems employ small reflective markers that are secured to the lips, jaw, and possibly other locations on the face. In order to be reliably detected for movement tracking, the markers need to be brighter than the background to which they are attached. Therefore, the markers are highlighted by a light source. Infrared lighting (i.e., wavelengths longer than, and thus frequencies lower than, visible light) is sometimes used instead of a visible
light source to avoid distraction of the subject and additional heating of the room. During movement, the marker positions are tracked by one or more analog or digital cameras. The position of each marker is determined based on its spatial relationship to a set of reference markers that are placed on nonmoving (relative to the skull) facial locations such as the nose and forehead. The marker positions as detected by each of the cameras are converted to Cartesian coordinates and then mapped onto a two- or three-dimensional coordinate system (Klein & DeHaven, 1995).

Techniques are now available to route the signal from a video camera directly into a computer through a video capture system (frame grabber). Movement traces are then generated by specialized software that either automatically determines the spatial coordinates of the marker positions or requires manual identification of the marker locations in each frame. Most standard (i.e., consumer-level) video cameras are capable of recording only 60 images per second. However, special recording techniques and capture hardware and software exist that make it possible to obtain 240 images per second with some standard digital cameras (APAS system, Ariel Dynamics, Trabuco Canyon, CA). Depending on the position and number of cameras used, such video-based systems can record information regarding the superior–inferior, medial–lateral, and anterior–posterior movements of the extraoral articulators. Thus, although only two-dimensional data can be obtained with a single camera, movement tracking in three dimensions is possible if more cameras are used to collect data at different angles (Caruso, Stanhope, & McGuire, 1989).

Passive marker systems require manual calibration. For two-dimensional measurements, this requires the construction of a reference frame. For three-dimensional measurements, a reference cube must be constructed (Trotman, Stohler, & Johnston, 1997). The video equipment is calibrated using the known distances between the markers on the reference frame or cube. Caruso et al. (1989) reported an average calibration error of 0.5 mm using three cameras to obtain three-dimensional data.

An example of a passive marker system (in addition to the aforementioned APAS system) is the ELITEplus system (BTS, Milan, Italy). This system detects passive markers by means of pattern recognition by a hardware-based processor even when the markers are not the brightest objects in the field of view. Markers can be as small as 1 mm or less, and infrared light is used for illumination. The manufacturer of this system indicates that the center of the marker is located with a very high resolution: Accounting also for distortion corrections, final accuracy is reported to be 1/2800 of the field of view. Specifically, with a 280 mm field of view and a 0.8 mm diameter marker, the accuracy is reported to be 0.1 mm (BTS Bioengineering Technology & Systems, n.d.).

Rather than using reflective markers, active marker systems track the movements of infrared light-emitting diodes (IREDs). IREDs contain a semiconductor chip and transmit infrared light. They are attached to various points on the speaker’s face. A set of sensors that contain a lens and signal-processing circuitry are able to detect and process the infrared signals emitted by the diodes and to determine their position in three dimensions (x, y, z).

One active marker system used for speech research is the Optotrak system (Northern Digital, Inc., Waterloo, Ontario, Canada). Optotrak uses a module with three sensors (shown in the left panel of Figure 4) that are able to track the position of multiple markers as small as 4 mm in diameter (shown in the right panel of Figure 4). The sensors sample the position of the various markers up to 200 times per second. During data collection, the subject’s head is allowed to move freely. Consequently, reference markers need to be positioned at stable points on the speaker’s head, and the position of the articulators is tracked relative to the reference markers. Strips of markers are placed on a flexible headband and attached to the speaker’s forehead (Ostry, Vatikiotis-Bateson, & Gribble, 1997). This way, corrections can be made for the subject’s head movements without requiring a head restraint.

The Optotrak is precalibrated by the manufacturer, but the user can define the reference plane and coordinate system. The system is reported to have an average measurement error of 0.1 mm (Northern Digital Inc., 2000). Using Optotrak data and graphical computer software, a

Figure 4. Three-sensor module (left) and infrared emitting diode active markers (right) of the Optotrak system.
method also has been developed that allows the three-dimensional point-tracking data to be constructed into a three-dimensional image of the jaw. As such, it is possible to reconstruct and display jaw movements recorded during speech production (Guiard-Marigny & Ostry, 1997).

Both passive and active optical movement tracking systems are appropriate for use with a variety of populations, including young children. For example, a passive system has been used recently to acquire data regarding the development of lip and jaw movements during babbling and early speech (Green, Moore, Higashikawa, & Steeve, 2000; Green, Moore, & Reilly, 2002). Obviously, the possibility to acquire kinematic data in infants and young children is a major advantage of optical tracking systems. Such information has been previously unavailable, yet it is of great importance to our understanding of both normal and disordered speech development. Optoelectronic systems also have been used in the study of normal speech physiology in both adults and older children (e.g., Ostry, Gribble, & Gracco, 1999; Smith & Goffman, 1998). One advantage of video-based (passive) systems is that they are relatively inexpensive, especially in comparison with optoelectronic (active) systems. A number of disadvantages can be noted with respect to both passive and active systems. First, intraoral markers cannot be detected by the cameras or the position sensors, and, thus, tongue and velum movements cannot be tracked. Second, in order to transduce jaw movement, markers must be placed externally on the chin. As discussed previously, this may be problematic because of excessive movement of the skin on the chin as a result of lip rather than jaw movement. If the marker is placed underneath the chin in order to minimize skin movement artifacts (such as is done for strain gauge transducers), it is not visible to the camera. Therefore, alternative methods have been devised to more accurately transduce jaw movement using optical movement analysis systems. For example, some researchers have attached a dental appliance to the mandibular teeth (Guiard-Marigny & Ostry, 1997). This appliance contains metal rods that are bent upward and outward through the corners of the mouth. These rods hold the markers externally. The dental appliance method has been used mainly in studies focusing exclusively on jaw movements. In such cases, interference of the rods with lip movements may be acceptable or irrelevant.

An additional disadvantage of video-based systems with passive markers is that there may be image distortion resulting from the shape of the video camera lens. Therefore, it is necessary to keep the subject’s orofacial area in the center of the field of view and to use a sufficiently large zoom factor or wide-angle camera lens. Distortion also can occur as a result of subject movement along axes that cannot be represented adequately on the two-dimensional video image. Furthermore, measurement error can be introduced when the orientation of the marker changes relative to the lens of the camera. During lip protrusion, for example, markers at the corners of the mouth may change their orientation (moving further out of the coronal plane), thereby changing their shape on the recorded video image (Borghese, Ferrigno, Redolfi, & Pedotti, 1997). This problem can be reduced somewhat by finding the best possible (i.e., least affected) marker positions during carefully selected oral movements before the actual data collection.

**PALATOGRAPHY**

**Electropalatography**

Electropalatography (EPG; also referred to as palatometry) visualizes the complex patterns of tongue–palate contact during speech production by means of small electrodes mounted in an acrylic artificial palate (Fletcher, McCutcheon, & Wolf, 1975). Given that EPG reveals changes in tongue–palate contact over time, it can be used to display information about the production of lingual consonants and those vowels that involve tongue–palate contact in the lateral regions. EPG was developed several years ago and has been used extensively since then in research related to normal speech physiology (e.g., Butcher & Weiher, 1976; Byrd, 1994). Additionally, EPG has been used to quantify articulatory contact patterns in individuals with articulation/phonological disorders, motor speech disorders, stuttering, and cleft palate (Fletcher, 1985; Gibbon, 1999; Hardcastle & Edwards, 1992; Hardcastle, Gibbon, & Jones, 1991; Wood, 1995). Moreover, because the procedure offers a real-time display of tongue–palate contact patterns, it also is used clinically to provide child as well as adult clients with on-line feedback concerning the accuracy of their productions (Friel, 1998; Howard & Varley, 1995; Kelly, Main, Manley, & McLean, 2000). In fact, several papers have been published reporting the efficacy of EPG biofeedback for establishing improved articulation in clients with acquired or developmental articulation/phonological disorders (Dent, 1995; Gibbon, 1999; Gibbon & Crampin, 2001; Whitehill, Stokes, & Yonnie, 1996) and in speakers of English as a second language (Schmidt & Beamer, 1998). Both the Kay Elemetrics EPG system and the Linguagraph system (see below) are designed to accept input from two simultaneous users so that the clinician can model the correct tongue–palate contact pattern in one window on the display and the client can then attempt to follow the model in a separate window. Repeated attempts can be superimposed or they can be compared with a previously stored target pattern.

For an EPG recording, the subject or client is custom-fitted with an artificial acrylic palate. The palate contains up to 96 electrodes that are embedded along the palatal and dental surfaces. The electrodes are attached to thin wires that pass behind the back molars and exit through the corners of the mouth. During data collection, the patient holds or wears a small electrode that generates a small current. When the tongue touches any of the electrodes embedded in the artificial palate, the electrical circuit is closed, and the location of the electrodes that were contacted can be recorded and displayed differently (e.g., squares for contacted electrodes and dots for uncontacted electrodes) on a two-dimensional graphic display of all
electrodes. Figure 5 contains an illustration of a display obtained using the Kay Elemetrics system. Panel A (upper left) shows the acoustic waveform of a produced utterance, panel B (bottom left) shows the phonetic transcription, and panel C (right) shows the actual EPG display. Square symbols in the EPG display indicate locations of tongue–palate contact at the time point indicated with a vertical cursor in the other two windows. Thus, locations of contact can be seen as dark squares whereas small diamonds indicate uncontacted electrodes.

Using EPG, the pattern of contacted electrodes is digitized into a computer typically at a rate of 100 samples per second. The recordings can be displayed in real time—which makes the system very useful as an online feedback tool for clinical purposes—as well as stored for subsequent detailed analyses. The two most commonly used systems are one developed at the Reading University in the United Kingdom (available from Millgrant Wells Ltd., Rugby, England) and one developed by Kay Elemetrics (Lincoln Park, NJ), which use artificial palates containing 62 and 96 electrodes, respectively. Although both systems reveal similar information regarding contact patterns, the Kay Elemetrics system may hold a slight advantage over the Reading EPG in terms of overall precision (Fougeron, Meynadier, & Demolin, 2000).

Recently, a new electropalatography system, known as the Linguagraph, has been presented as a low-cost, portable system for clinical use (Kelly et al., 2000). The Linguagraph system makes use of the Reading palate with 62 electrodes.

In addition to the typical two-dimensional EPG displays, data also can be shown as contact patterns on a three-dimensional representation of the palate. Jones and Hardcastle (1995) described a procedure for obtaining a three-dimensional palatal image and loading it into the software for the Reading EPG system. The image is created by manually determining the x, y, and z coordinates of the electrodes relative to the occlusal plane while the artificial palate is in a plaster cast. The three-dimensional image is displayed as a “skeleton” or wire-frame image of the palate constructed by graphically connecting all represented electrodes. A similar procedure for manual determination of each electrode’s coordinates has been described for the Kay Elemetrics EPG system (Saw, 1993). A high-tech solution that has become available more recently consists of obtaining the electrodes’ coordinates automatically with a noncontact three-dimensional digitizer (Wakumoto, Masaki, Honda, Kusakawa, & Ohue, 1999).

EPG techniques continue to be improved and refined. For example, Wakumo et al. (1999) have presented work integrating three-dimensional EPG and pressure-sensitive palatography. Pressure-sensitive palatography provides information about the time-varying tongue–palate contact pressure. Thin (0.1 mm) sensor array sheets are placed on the oral surface of a custom-made acrylic palate (0.5 mm). Each sensor array contains from one to four sensors that consist of a pair of small electrodes (3 mm diameter) separated by “pressure-sensitive ink.” The underlying principle is that changes in the distance between the two electrodes can be detected by measuring changes in the electrical resistance of the pressure-sensitive ink between the electrodes. In the absence of any force, the electrical resistance between the electrodes is high and constant. Changes in resistance are induced by alternations in the pressure exerted on the sensors during tongue contact because this pressure changes the distance between the electrodes. Thus, pressure-sensitive palatography continuously monitors these changes in resistance in order to obtain a measure of tongue–palate contact pressure. The total number of sensors and their location can vary based on the subject’s palate size and shape as well as on the expected locations of tongue–palate contact for the targeted speech sounds. Wakumoto and colleagues described placement of 9 to 16 sensors and a sampling rate of 126 samples per second. Their preliminary work was able to demonstrate contact pressure differences between Japanese voiceless and voiced alveolar stop consonants with almost identical contact areas.

Figure 5. Electropalatography display (labeled “palatogram”; Kay Elemetrics System with 96 electrodes) next to waveform and phonetic transcription windows. Square symbols in the electropalatography display indicate locations of tongue–palate contact at the time point indicated with a vertical cursor in the other two windows.
EPG also has been used in conjunction with magnetic resonance imaging (MRI) and electromagnetic articulography (EMA, see below) to develop three-dimensional models of articulatory contact during the production of specific consonants (Alwan, Narayanan & Haker, 1995; Narayanan, Alwan, & Haker, 1997). These studies used EPG to address specifically the issue of inter- and in-subject variability in the production of lingual consonants.

A major advantage of EPG is that it is a relatively unobtrusive method that causes little discomfort, yet it provides valuable information about patterns of lingual-palatal contact—information that presently is not available from other methods. Because of its unobtrusive nature, cost effectiveness, and capability to provide on-line feedback regarding tongue–palate contact patterns, EPG is one of the more practical instrumental methods for diagnosis and intervention in the clinical setting. A disadvantage of EPG is that the artificial palates with embedded electrodes need to be custom-made for each research subject or each client, which is a costly process.

### Optopalatography

Optical systems to transduce the distance between tongue and palate were developed in the late 1970s and the 1980s. Known as glossometry, these systems relied on light emitters and photosensors mounted in an acrylic palate (Chuang & Wang, 1978; Fletcher, McCutcheon, Smith, & Wilson, 1989). Because the intensity of the light detected by the photosensor depends on the distance between the light-emitting source and photosensor on the one hand and the light-reflecting surface on the other hand, the signal detected by the photosensor can be used to extract information about the palate-to-tongue-surface distance. Widespread use of glossometry, however, was prevented by a number of limitations. The sensors were rather thick, and a palate as thick as 5 mm was needed. In addition, the sensors were located only along the midline of the artificial palate.

A more recent technique known as optopalatography (Wrench, McIntosh, & Hardcastle, 1996) uses technology similar to that employed by glossometry, but avoids its major limitations. Optopalatography measures not only the intensity of light reflected from the surface of the tongue and, thus, tongue–palate distance, but also contact location and contact pressure. Small diameter (0.25 to 0.5 mm) plastic optical fibers embedded in an artificial palate emit light from the palate and sense its reflection from the surface of the tongue. The optical fibers are connected to a remote unit that provides the circuitry for light emission and sensing. This design eliminates the need for the light source and sensors to be present inside the oral cavity, it allows the use of a thin artificial palate (0.5–2 mm), and several measurement points both on and off the midline can be used (up to 70). The optical fibers enter and exit the oral cavity in two bundles with a diameter of approximately 2.5 mm, positioned in the corners of the mouth.

After calibration of the system (e.g., using EMA, see below), the distance between the tongue and palate is derived from measurements of the intensity of light reflected from the surface of the tongue. The locations of actual tongue–palate contact can be determined as those locations where the distance decreased to zero. Moreover, contact pressure can be measured based on changes in intensity of the reflected light after the tongue has already contacted the palate. Because of the translucent nature of the tongue, it is indeed possible to detect continuing changes in light intensity after the initial tongue–palate contact has occurred.

The greatest intensity of the reflected light can be collected, and the most precise distance measurements made, when the light emitted from the transmitter fibers shines perpendicular to the tongue surface. Because of the complex movements of the tongue, however, the orientation of the tongue surface relative to the palate changes continually. As a result, the angle of inclination at specific tongue points can induce some error in the resulting distance measurements. Measurement inaccuracies also can be caused by “beam spread” of the light exiting the optical fiber. Wrench et al. (1996) reported that the accuracy of their prototype design was 10% for angles of inclination up to 20 degrees.

### EMA (MAGNETOMETRY)

EMA is capable of providing data similar to those obtained with X-ray microbeam systems—that is, flesh-point tracking of all articulators including those inside the oral cavity (typical ways of visualizing the data are shown in Figure 2)—but using a technique that is believed to be safer and considerably less expensive than that used by the microbeam system. These two advantages make EMA accessible to many individual research laboratories. After initial developments described by Hixon (1971) and Van der Giet (1977), the technique as currently used was developed primarily at the Massachusetts Institute of Technology (MIT, Perkell et al., 1992). The MIT-designed articulograph also is available at Haskins Laboratories (New Haven, CT), whereas other users have commercially available systems (see below).

EMA is based on high-frequency alternating magnetic fields generated by transmitter coils. Each transmitter coil generates an alternating magnetic field with a specific frequency in the range of 10–20 KHz. The magnetic fields induce alternating currents in small receiver coils (as small as 1.5 × 1.5 × 1.0 mm in the MIT design; Haskins Laboratories, n.d.) that are affixed to various points on the articulators and that are connected, by means of very thin lead wires, to the system electronics. The signal induced in a receiver coil by a given transmitter coil is inversely proportional to the cube of the distance between the receiver coil and the transmitter coil. Because of the different frequencies used to drive the three transmitter coils, the distance between a receiver coil and each transmitter coil can be calculated in software. As such, the exact spatial coordinates of the receiver coil can be computed.

Accurate tracking of the sensor coils requires a secure attachment to the articulators. Most users first prepare the
receiver coils by dipping them into a rubber-based compound that provides a protective and smooth coating. The coils are then affixed to the articulators using different types of adhesive, depending on the location of the coil. Extraoral placement, such as for the upper lip, lower lip, and nasal reference coil, is typically accomplished with a skin adhesive (e.g., skin cement). Intraoral placement, such as for the tongue, lower incisors (i.e., jaw), velum, and upper incisor reference coil, is typically accomplished with a dental adhesive (cyanoacrylate). Secure placement of the intraoral coils for the length of the data recording session usually requires careful drying of the articulator surface before placement of the coil. This can be accomplished with a gauze pad or a cool blow dryer. The total amount of time necessary to attach the coils to the articulators is dependent on the number of sensor coils that are needed, but can range from 30 to 60 minutes for 10 coils.

2D EMA

All EMA systems currently in use are two-dimensional systems: they register two-dimensional movements of the receiver coils in the midsagittal plane only. The transmitter coils are positioned parallel to each other, perpendicular to the midsagittal plane, and with their midline lined up to the midsagittal plane. Most systems make use of three transmitter coils (i.e., MIT design as well as the commercially available AG100 design by Carstens Medizinelektronik, Gottingen, Germany), although a few laboratories have used a system with only two transmitter coils (i.e., Movetrack from Botronic, Hagersten, Sweden—at this time, production of the latter system has been discontinued). The transmitter coils of the three-transmitter systems are held in place by a “helmet” in such a way that they form an equilateral triangle, with one transmitter above and in front of the forehead, one at the neck, and one below and in front of the chin close to the chest. The Carstens AG100 is used with either a 32-cm helmet (designed at the University of Nijmegen in The Netherlands; see Alfonso et al., 1993) or a larger 64-cm helmet. The helmet is attached to a pulley system suspended from the ceiling so that its weight does not need to be supported by the subject. When using the Nijmegen helmet, an adjustable headmount is placed on the head and a plexiglass structure with the transmitter coils is then placed over, and affixed to, the headmount.

EMA receiver coils need to be positioned with their axis parallel to the axis of the transmitter coils (i.e., perpendicular to the midsagittal plane), and accurate measurements are obtained only when the orientation of the receivers relative to the transmitters is maintained during data acquisition. One source of error during the registration of speech movements is misalignment of the receiver coils as a result of lateral or rotational movements and soft tissue deformation. In particular, three types of misalignment commonly occur: twist, tilt, and movement of the coil away from the midline (Perkell et al., 1992). These types of misalignment are especially common for tongue movements. Misalignment affects the recorded distance between the transmitter and receiver coils, thus reducing the accuracy of the data. The advantage of three-transmitter coil systems over two-transmitter coil systems is that the former allow corrections to be computed for instances of rotational misalignment. The additional transmitter provides information that can be used by software algorithms for the detection and correction of misalignment errors (Hoole & Nguyen, 1999). However, accurate correction for rotational misalignment can be calculated only if the receiver coils remain within the midsagittal measurement plane (Gracco, 1995).

Commercially available EMA systems, such as the AG100, are capable of sampling rates of up to 500 per second and have been reported to have a spatial resolution between 0.1 and 0.2 mm (Alfonso et al., 1993). Careful calibration is essential when using EMA. Calibration can be performed either manually or by means of an automatic calibration unit available from the manufacturer.

It also is noteworthy that a database (known as the MOCHA database) of two-dimensional EMA data obtained from speakers of British English is currently under development (Wrench & Hardcastle, 2000). In fact, this database will include EPG as well as electroglossography (i.e., registration of vocal fold vibrations) data in addition to seven channels of EMA data.

3D or 5D EMA

As mentioned above, current EMA systems require that all receiver coils are attached midsagittally, they track coil movement in the midsagittal plane in two dimensions, and inaccurate measurements may occur when the coils move off the midline or are misaligned relative to the transmitter coils. Recently, the Carstens company, in collaboration with a number of speech research laboratories, has developed the first units of a system that does not have the midsagittal plane restriction; that is able to track movements in three, rather than two, dimensions; and that is not negatively affected by receiver coil orientation. In fact, this system (a) allows receiver coils to be placed anywhere on any of the articulators without being limited to the midsagittal plane and (b) determines not only the position in three dimensions, but also two angles of orientation for each coil (Zierdt, Hoole, & Tillman, 1999; Zierdt, Hoole, Honda, Kaburagi, & Tillman, 2000). Hence, the system can be considered a five-dimensional system.

This new five-dimensional articulograph, the AG500, uses six transmitter coils that are attached to a supporting cube surrounding the subject’s head rather than to a head-mounted structure. As can be seen in Figure 6, this cube consists of a plexiglass frame that leaves plenty of room for the subject. Note also that the three white circular shapes on the front panel of the cube are the points of attachment of three of the six transmitter coils (the remaining three transmitter coils are located behind the subject). Because receiver coil placement is no longer limited to the midsagittal plane, and because movements can be recovered for coils on the articulators by re-expressing their positions relative to three reference coils on nonmovable parts of the skull, the subject’s head is able to move freely within the measurement area.
The AG500 is capable of sampling rates of 200 samples per second for each channel, with a spatial resolution of 0.5 mm and a rotation detection of 1 degree (Carstens Medizinelektronik, n.d.). The manufacturer reports that the number of channels is unlimited. At the time of this writing, the manufacturer is in the final stages of developing the calibration hardware and procedures.

EMA systems have provided a fast and accurate method of measuring articulatory movements. They have been used to provide insight into several aspects of speech motor control, including coordination of the articulators during normal speech production (e.g., Hertrich & Ackermann, 2000). These systems also have been used to quantify articulatory movement parameters in individuals who stutter, individuals who have dysarthria, individuals with a cleft lip, and users of cochlear implants (Goozee, Murdoch, Theodoros, & Stokes, 2000; Matthies, Svirsky, Perkell, & Lane, 1996; McClean & Runyan, 2000; Rutjens, Spauwen, & van Lieshout, 2001). Although the availability of EMA systems has been limited to research laboratories, at least one study has explored the efficacy of using EMA as a tool to provide visual feedback about tongue position in an adult with apraxia of speech (Katz, Bharadwaj, & Carstens, 1999). The results of that case study suggest that EMA systems may have value as a clinical tool in the treatment of some speech disorders.

One of the main advantages of EMA is related to the capability to transduce intraoral movements, including tongue and velum movements. Furthermore, although no definitive answers are available regarding the possible health risks associated with electromagnetic fields, EMA systems are considered to be safe for use with human subjects. The strength of the electromagnetic fields generated by the MIT system is below all local exposure standards included in a safety evaluation by Hasegawa-Johnson (1998), whereas the Carstens AG100 meets most, but not all, of those standards. Clearly, continued research into the health risks associated with this and similar procedures will be necessary as the technology continues to develop.

Disadvantages of two-dimensional EMA include difficulties with midsagittal placement, measurement inaccuracies resulting from rotational misalignment, and availability of data in only two dimensions. Also, the helmet becomes uncomfortable during lengthy data collection sessions. These limitations and concerns should be largely eliminated with the three-dimensional AG500.

**Radar Systems**

Recently, attempts have been made to use high-frequency (2.3 GHz), low-power (fractions of milliWatts) electromagnetic radar sensors for the study of speech movements (Burnett, Holzrichter, Gable, & Ng, 1999; Holzrichter, Burnett, Ng, & Lea, 1998). Radar systems function by emitting electromagnetic waves from a transmitter antenna and capturing their reflections from conductive discontinuities such as an air–tissue interface. The reflections are captured with a separate receiver antenna. For a type of sensor known as homodyne field disturbance sensors, the distance from the sensor to the point of reflection can be determined by comparing the phase of a reflected wave with that of a local reference wave. When the surface that reflected the wave moves further away from, or closer to, the transmitter antenna, the phase of the reflected wave will change accordingly.

In the case of speech movement transduction, the waves would be reflected by the speech organs located within the path of the transmitted waves. Importantly, the sensors would be positioned external to the vocal tract, and, as such, electromagnetic radar systems would provide a completely noninvasive and unobtrusive way of transducing articulatory movements. In the initial work by Holzrichter and colleagues (Burnett et al., 1999; Holzrichter et al., 1998), measurements of jaw and tongue movements were presented in addition to glottal (or tracheal wall) vibrations. To transduce the jaw and tongue movements, the radar transmitted electromagnetic waves in an upward direction from under the jaw.

Based on the preliminary work by Holzrichter and colleagues (Burnett et al., 1999; Holzrichter et al., 1998), radar sensors appear promising for at least some applications of articulatory measurement. Radar-based systems could be safe (the waves are radiated at intensities reported to be well below the standards for maximum permissible exposure in continuous use), completely noninvasive and unobtrusive (the sensors are located outside the vocal tract and not in contact
with the subject), and inexpensive. However, further research is needed to confirm the validity and reliability of measurements obtained for the articulators.

Although research (Holzrichter et al., 1998; Titze et al., 2000) has provided initial confirmation of the accuracy of radar sensors when used to extract fundamental frequency and vocal fold contact information (but even then the exact contribution of tracheal wall vs. vocal fold vibration remains unclear), no information is available regarding their accuracy when transducing articulatory movements. Therefore, these systems have not been applied yet in actual studies of articulatory movement, nor have attempts been made to use them clinically. Comparisons with data simultaneously collected with systems based on a different technology (e.g., EMA) seem warranted before any conclusions can be drawn regarding the usefulness of these radar sensors for the transduction of articulatory movements. An additional limitation follows from the fact that the current sensors measure only signal changes within a characteristic frequency band associated with a specific structure (e.g., for phonation, around 100–200 Hz). The articulators, however, are at many different locations, and the different articulators move at comparable speeds. Therefore, in order to know which articulator reflected the waveforms, the current sensors require “the speaker to pronounce only those phoneme sequences that involve a known single articulator motion” (Burnett et al., 1999, p. 2247).

MRI AND TAGGED CINE-MAGNETIC RESONANCE IMAGING

MRI is a technique that is best known for its application in imaging the brain. However, it also has been used to provide information regarding the shape and position of the soft tissue articulators during static positions for speech production (Alwan et al., 1995; Baer, Gore, Boyce, & Nye, 1987; Baer, Gore, Gracco, & Nye, 1991; Narayanan et al., 1997; Story, Titze, & Hoffman, 1996, 1998). Given the focus of the present paper on instrumental methods for the transduction of speech movements, we will discuss the basic processes of MRI only to the extent that they are necessary to understand some more recent modifications that have made it possible to obtain dynamic movement information.

In essence, the principles underlying MRI are based on the magnetic properties of the single proton in hydrogen atoms. Hydrogen atoms are abundant in water, which is present in varying amounts in different soft tissues. When not externally perturbed, the hydrogen protons spin around an axis with random orientation. When the MRI scanner’s transmitting electromagnet surrounds the subject’s head with a strong magnetic field, the spinning axes of the hydrogen protons become aligned. They stay aligned until a brief radio frequency pulse causes each proton to wobble out of alignment—a process called precession. However, given the combination of decay of the precession and the continuing presence of the static electromagnetic field, the proton axes become realigned. When realigning themselves, the protons emit a signal with the same radio frequency as that of the pulse that previously tipped them out of alignment. These signals are detected by a receiver coil within the MRI scanner. When the received signals are assembled by computer, they create a two-dimensional grey-scale image of the soft tissues in the recording field. Thus, different soft tissue structures can be differentiated based on their hydrogen content. For example, in a twodimensional midsagittal MRI image, the bony structures and air appear black because they do not contain hydrogen. The tongue, which is high in hydrogen content, appears in a lighter color. By obtaining multiple parallel slices, reconstruction of three-dimensional images of the entire vocal tract is possible.

Construction of a standard MRI image is relatively time-consuming. Each image can take from several seconds to as much as a few minutes to construct. During this entire time period, the speaker must maintain a static articulatory position. Hence, standard MRI technology is useful for imaging the vocal tract during sustained articulatory positions but not during dynamic movements. It should be noted that some MRI techniques are capable of operating at higher rates, such as the ultra-fast turbo-spin technique used by Demolin, Metens, and Soquet (2000) to achieve four or five images per second. The trade-off is that some high-speed MRI procedures are associated with reduced spatial resolution and increased signal noise from the scanner (Stone, 1997). Recently, however, two new MRI-based techniques that provide dynamic articulatory information have been developed. These techniques are known as tagging snapshot MRI or tagged cine-MRI (tMRI; Kumada, Niitsu, Niimi, & Hirose, 1992; Niitsu, Kumada, Niimi, & Itai, 1992; Stone et al., 2001) and stroboscopic MRI articulography (Mathiak et al., 2000, 2000b).

**tMRI**

**tMRI** is a technique whereby measurements are made of the motion and deformation of a grid imposed on a cine-display of the vocal tract. The grid is formed by applying radio frequency pulses and gradients to bands of tissue such that they do not produce a coherent signal and are displayed in black on the MRI image. Horizontal lines (i.e., parallel to the image’s x-axis) are created first, followed by vertical lines (i.e., parallel to the image’s y-axis) to form a complete grid. These bands of tissue then appear as intersecting vertical and horizontal tag lines on the image (Figure 7). In a recent study by Stone et al. (2001), the tag lines were separated by 11 mm. Once the grid has been created, it can be maintained on the display until the spinning protons have recovered a significant portion of their magnetization. During that time window, the motion and deformation of the tag lines can be tracked by collecting a sequence of images. In the aforementioned work by Stone and colleagues, data were acquired for intervals with a duration of 392 ms during which seven images were acquired (referred to as seven time phases). After the imaging sequence, the tag lines must be recreated. Given that seven images were taken over a 392-ms time period, images were acquired with a sampling rate of 18 samples per second.
Because the tag lines are composed of bands of tissue, they move and deform along with the articulator (in work to date, always the tongue) in the cine image. Because of the relatively slow sampling rate, the final image sequence of an articulatory movement is reconstructed from multiple repetitions of the same utterance. The subject in Stone et al. (2001) produced 32 repetitions of a consonant–vowel (/k/ to /A/) tongue movement for each of three slices (left, middle, right), for a total of 96 repetitions. The signal emitted by the protons in each of the seven time phases was summed across multiple repetitions so that a single averaged image of the seven time phases was obtained. For Stone et al.’s subject, the tongue movement from /k/ to /A/ was completed over the first four time phases.

A major advantage of tMRI is that it allows assessment of the deformation of the entire tongue during speech articulation. On the other hand, a number of important disadvantages remain. First, the equipment is large, costly, and therefore not easily accessible. Second, the subject must be in the supine position, and this different environment in terms of gravitational effects on the articulators may influence articulatory movements (Shiller, Ostry, & Gribble, 1999; Stone, 1996; Tiede, Masaki, & Vatikiotis-Bateson, 2000). Third, the temporal (approximately 50 ms) and spatial resolution (approximately 2 mm) are limited. Stone et al. (2001) reported the total error resulting from the sum of (a) temporal and spatial resolution limitations, (b) measurement error, and (c) speaker variation in the form of articulatory precision and head movement to be between 0.5 and 1.8 mm, depending on the direction of movement (x- vs. y-axis) and the type of sound being produced (vowel vs. consonant). Lastly, because of the width of the scanned section (7 mm), there is some spatial distortion in the resulting image. This distortion results from the MRI display of the three-dimensional structures of the vocal tract in two dimensions. Structures that are within the width of the slice but on a different plane will appear to be in the same plane. It should be noted, however, that work is in progress to expand the capabilities of tMRI to track and represent tongue deformation and movement in three dimensions (Stone, Dick, Douglas, Davis, & Ozturk, 2000).

Stroboscopic MRI Articulography

A different method of applying MRI technology to the study of dynamic aspects of speech articulation is stroboscopic MRI articulography. This method relies on a stroboscopy-like procedure. A sequence of MRI images is obtained during repetitive movements, but rather than the sequence being an average across the repetitions, the composite image is obtained by combining distinct data from each repetition. As described by Mathiak and colleagues (Mathiak et al., 2000a, 2000b), a single line in the image was scanned 64 times (repetition time = 8.4 ms) during the interval in which the subject produced a CVC syllable. Approximately 840 ms after beginning the scans for that line (64 × 8.4 ms plus a delay of 300 ms), the next line in the image was scanned while the subject repeated the same syllable. In this manner, the entire image was scanned line by line across a train of 170 repetitions of the same syllable. Given that the scan for each line was repeated with a repetition time of 8.4 ms, the combined images (i.e., all lines combined) formed a reconstructed movie with a sample rate of 120 samples per second.

The major advantage of stroboscopic MRI articulography is similar to that of tMRI: All articulators, including the entire tongue, can be imaged. An obvious disadvantage is that the obtained video image of the production of one syllable is not actually obtained during the production of one syllable. Rather, the video image is reconstructed from multiple productions of a given utterance. This procedure assumes that the speaker’s productions are consistent across the repetitions, and prevents any measures of trial-to-trial variability.

FUTURE DIRECTIONS AND CLINICAL APPLICATION

Information obtained with various types of physiological instrumentation has made it possible to document and understand important sensorimotor aspects of both normal and disordered speech production. This is indeed an exciting era in speech research given the rapid and very promising developments in several technologies that are used to transduce articulatory movements. Examples that we have discussed in some detail are recent improvements in ultrasound and MRI imaging (extending their use from static articulatory postures to dynamic movements) and EMA point-tracking (extending it from a two-dimensional to a five-dimensional system).

Of course, even these important advances will not immediately resolve certain limitations that are inherent in each of the methods. One way in which some of those limitations may be addressed in the future is through an increase in the simultaneous use of two or more different
methods. For example, a number of laboratories are already combining EMA and EPG procedures to obtain information about tongue movement in combination with information about the locations on the palate contacted by the tongue. Although there is a trade-off in terms of speech naturalness because the subjects have both EMA receiver coils and an EPG artificial palate in their mouth (and the wires for both systems), this combination of techniques provides information that is currently unavailable from any single instrument by itself. In light of their increased feasibility for use during dynamic movements, ultrasound and other technologies discussed above may provide similar opportunities for application as part of a combined instrumentation array.

An additional trend that we expect, and hope, to see in the future is an increased use of these instrumental techniques of speech analysis for the purpose of improving speech-language pathologists’ services in the diagnosis and treatment of children and adults with various speech disorders. Admittedly, for some techniques, this will require a rather long process of refinement and further testing as they are still in the early stages of development. For other techniques, however, some possibilities have already been demonstrated at the time of this writing. One technique that can be used to illustrate this point is EPG, which is currently being used rather extensively in the treatment of individuals with articulation disorders and in articulation improvement for individuals who speak English as a second language (e.g., Friel, 1998; Howard & Varley, 1995; Michi, Suzuki, Yamashita, & Imai, 1986; Michi, Yamashita, Imai, Suzuki, & Yoshida, 1993; Schmidt & Beamer, 1998).

Although EMA technology is relatively new as compared to EPG, and its cost may be prohibitive for large-scale clinical application, at least one study has documented positive results when using EMA to provide visual movement-related feedback in the treatment of an adult with aphasia and apraxia of speech (Katz et al., 1999). Also noteworthy in this regard is an ongoing project in The Netherlands that was initiated to develop an instrumental speech motor test for clinical use with individuals who stutter (Peters, Hulstijn, & van Lieshout, 1995; van Lieshout, Peters, & Bakker, 1997). Although this particular test currently involves acoustic, respiratory, and electromyographic data, similar procedures could be developed using instruments that transduce articulatory movements. Such attempts are to be encouraged as they hold great promise to contribute significantly to the continuous improvement and refinement of the clinical services offered by speech-language pathologists.

Indeed, the ultimate goals of the technological advances described in the present paper are not limited to the resulting gains in our theoretical knowledge of the speech production process. At least equally important is how much these technologies will contribute to improving the day-to-day quality of life for individuals whose communication skills and social interactions are negatively affected by a speech disorder. However, only when clinicians will increase the application of these instruments in the clinic while carefully documenting the potential effects on treatment efficacy (treatment outcome as well as treatment duration) will the field be able to evaluate the cost–benefit ratio of each instrument for various clinical populations and treatment approaches. Encouraging such attempts to integrate instrumental techniques in clinical evaluation and treatment and to investigate their possible benefits for clinical service delivery requires that those in the clinical field are first familiarized with the technologies, methods of use, and advantages/disadvantages of the currently available instruments. It is our hope that the present review article may contribute to this goal.

ACKNOWLEDGMENT

We gratefully acknowledge Khalil Iskarous (Haskins Laboratories) for the ultrasound images. Saibal Chakraburtty (Northern Digital, Inc.) for the Optotrak photographs, Steve Crump (Kay Elemetrics) for the EPG data display, Brigitta Carstens (Carstens Medizinelektronik) for the AG500 photograph, and Maureen Stone (University of Maryland School of Medicine) for the tMRI image.

REFERENCES


Beautems, D., Badin, P., & Bailly, G. (2001). Linear degrees of freedom in speech production: Analysis of cineradio-


Contact author: Ludo Max, PhD, University of Connecticut, Department of Communication Sciences, 850 Bolton Road Unit 1085, Storrs, CT 06269-1085. E-mail: ludo.max@uconn.edu