Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels

T. Baer
Department of Experimental Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom

J. C. Gore
Department of Diagnostic Imaging, Yale School of Medicine, New Haven, Connecticut 06501

L. C. Gracco and P. W. Nye
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

(Received 31 July 1990; accepted for publication 22 March 1991)

Magnetic resonance imaging (MRI) techniques were used to gather basic data to apply in computational models of speech articulation. Two experiments were performed. In experiment 1, voice recordings from two male subjects were obtained simultaneously with axial, coronal, or midsagittal MR images of their vocal tracts while they produced the four point vowels. Area functions describing the individual tract shapes were obtained by measurements performed on the MR images. Digital filters derived from these functions were then used to resynthesize the vowel sounds which were compared, both perceptually and acoustically, with the subjects' original recordings. In experiment 2, axial images of the pharyngeal cavity were collected during the production of an ensemble of nine vowels. Plots of cross-sectional area versus the midsagittal width of the tract at different locations within the pharynx and for different vowel productions were used to derive a functional relationship between the two variables. Data from experiment 1 relating midsagittal width to cross-sectional area within the oral cavity were also examined.

PACS numbers: 43.70.Aj, 43.70.Jt, 43.72.Ja

INTRODUCTION

Data on vocal tract shape and dimensions are essential to a full understanding of the articulatory and acoustical processes involved in speech production. The acoustical theory of speech production (Fant, 1960) requires us to view the vocal tract as an acoustical tube with a varying cross-sectional area. To support the early testing of this theory, data on vocal tract shape were collected, largely from radiographic sources (Chiba and Kajiyama, 1941; Fant, 1960; Heinz and Stevens, 1964). These same data were then used extensively in subsequent work on the acoustics of speech production and also served as the basis for early analog models of articulation (e.g., Stevens and House, 1955). Since these pioneering studies, and with the advent of computers, there have been significant advances in the modeling of articulatory and acoustic processes. Articulatory synthesizers have gained importance as research tools (Mermelstein, 1973; Coker, 1976; Maeda, 1982; Rubin *et al.*, 1981; Browman and Goldstein, 1987) and have grown in computational complexity. Many are able to deal with quite detailed area functions and to account for the effects of yielding cavity walls and other mechanisms of energy loss (e.g., Flanagan *et al.*, 1975; Liljencrants, 1985). Recently, there has been a renewed interest in aeroacoustic phenomena in the vocal tract, especially those associated with the acoustic sources of voicing and frication (McGowan, 1988; Shadle, 1985), and even the aeroacoustics of vowel production has received some re-evaluation (Teager and Teager, 1983). Furthermore, models of articulation have addressed more difficult problems such as the three-dimensional (3-D) shape of the tongue (Kakita and Fujimura, 1977). However, few new data on vocal tract dimensions have been added in recent years, and it is now widely recognized that continued progress with both the development of theory and its implementation in computer models requires more detailed data on vocal tract shape and area functions.

Especially considering the importance of vocal tract shape and the area function to speech research, it is indeed surprising that so few new studies have been undertaken. The available data are drawn from relatively few subjects with different language backgrounds, and relate to a limited number of speech sounds. Thus the sources of the samples differ from one study to another and it is difficult to make generalizations. Furthermore, the type of data that is most urgently needed is 3-D in form, whereas most of the available data are two dimensional. In a majority of the early studies, measurements were made from lateral radiographic images (e.g., Chiba and Kajiyama, 1941; Fant, 1960; Abramson and Cooper, 1963; Heinz and Stevens, 1964; Perkell, 1969; Sundberg, 1969), and transverse areas of the airway were obtained by applying transformations to the widths measured from these lateral projections. However, the data on which such transformations from width to area have been based are also limited in quantity. Some data have been derived from casts of the vocal tract (Ladefoged *et al.*, 1971; Sundberg *et al.*, 1987), from measurements of cadavers (Heinz and Stevens, 1964), from measurements made with calipers as subjects tried to hold the positions associated with static sounds (Anthony, 1964), and from less quantitative techniques such as monoscopic fiberoptic endoscopy (Gauffin and Sundberg, 1978). For the pharynx, Heinz and Stevens

(1964) based their transformation on assumptions about shapes and lateral dimensions. There have been a few tomographic studies of vocal tract shape using radiography. Fant (1965) used a conventional radio-tomograph to visualize cross sections of the pharynx, and subsequent studies have used computed tomography (CT) (Kiritani *et al.*, 1977; Johansson *et al.*, 1983; Boe *et al.*, 1988).

The problem of converting midsagittal width to cross-sectional area is perhaps most difficult in the pharyngeal region and, consequently, that is where reliable data are in shortest supply. The results of the studies noted above have led to a number of quite different conclusions about the relationship between the midsagittal width and cross-sectional area of the pharynx at different heights above the larynx during vowel production. Heinz and Stevens (1964) assumed an elliptical cross section whose transverse axis adopts a different fixed magnitude at each position along the tract. Thus their assumption implies that there is a linear relationship between width and area that varies with position. Ladefoged *et al.* (1971) also summarized their results as a graphical relationship between midsagittal width and cross-sectional area that is roughly linear at each position. Sundberg (1969), using data obtained by Fant, proposed the relationship $A = K \cdot S^{1.5}$, where A is the area, K a constant, and S is the midsagittal width that is raised to the 1.5 power, and, in a recent preliminary report of results from x-ray CT, Boe *et al.* (1988) have also used a 1.5 power relationship to fit their data. On the other hand, Sundberg *et al.* (1987) used x-ray CT and derived a square law relationship ($A = K \cdot S^2$). Clearly, more data are needed to clarify this relationship.

At first glance, the x-ray CT method might appear to be the best available to collect the necessary data. It is certainly the case that, over a substantial proportion of the total length of the vocal tract, x-ray CT can, in principle, supply the kind of 3-D information that is needed about tract shape. However, it has been used to image cross sections of the vocal tract at only a few positions for only a few sounds. The reason for this tentative beginning lies in the two major drawbacks of x-ray CT. The first is that, given current knowledge of the damaging effects of even quite low x-ray dosages, the nonmedical use of x-rays on research subjects must be strictly limited. Consequently, it is not possible to obtain even one set of serial CT sections at intervals on the order of 5 mm along the vocal tract without exceeding by several times what many regard as an ethically acceptable risk. The second reason is the limited maneuverability of the subject with respect to the x-ray system. Conventional x-ray CT systems have only a transaxial imaging capability and a table that can be tilted through an angle of up to 45 deg with the subject securely attached. These systems are best suited to imaging the pharyngeal airway of a supine subject because there is only sufficient adjustment to obtain images in planes that are orthogonal to the pharyngeal axis. Repositioning the subject in an effort to maneuver the imaging plane around the bend in the vocal tract and into the upper tract (e.g., by tilting the head with respect to the body) fails because the change in head posture invariably induces significant changes in tract shape.

Alternative methods can be used for obtaining shape information, but all have limitations. Point-tracking methods, such as the x-ray microbeam (Fujimura *et al.*, 1973) and magnetometers (e.g., Schonle *et al.*, 1987), can supply dynamic information about the movement of structures in the oral cavity, but they are designed to obtain measurements only in the midsagittal plane and they cannot supply detailed information about shape. Scanning ultrasound has been used to generate dynamic images of the tongue surface, either in the midsagittal or transverse planes (Shawker *et al.*, 1984; Shawker and Sonies, 1984; Stone, 1990), but it has been useful over only a limited part of the tongue, since for imaging purposes ultrasound does not propagate across tissue boundaries with air and very weakly across boundaries with bone (Minifie *et al.*, 1971). The transmission properties of ultrasound have confined its use to mapping portions of the anterior surfaces of the airway. And, lastly, a stereo fiberoptic endoscope can, in principle, supply quantitative information from the pharynx, but, at present, this instrument has not achieved its potential (Fujimura *et al.*, 1979).

Magnetic resonance imaging, however, is free from many of the disadvantages associated with the methods we have mentioned, although it does have some drawbacks of its own. Among its advantages is the fact that MR technology produces tomographic images that appear similar to those produced by CT but without using ionizing radiation and in three orthogonal planes without tilting the table or repositioning the subject. While in a CT image the pixel values represent the x-ray transmittance of a given volume element of tissue, in an MR image the signal intensity depends on the density of hydrogen nuclei in a tissue element and their magnetic relaxation times $T1$ and $T2$—variables that are determined by the atomic environment of the protons within each molecule. MR techniques involve the use of strong magnetic fields to align the magnetic moments of the hydrogen nuclei and the use of high radio-frequency impulses to set them into resonance. These techniques have no known harmful side effects. Furthermore, images can be obtained with acceptable resolution (about 100 pixels/cm^2) from slices that pass through any point of interest in the vocal tract.

One of the major disadvantages of the MR technique is the time required to perform the imaging process, a period ranging from several tens of seconds to several minutes using the most commonly available equipment and techniques. Therefore, if speech sounds are to be studied, they must be sustainable for the duration of the acquisition process and may, as a consequence, be subject to fatigue effects. (In contrast, CT and ultrasound images can be obtained in under 3 s although the signal/noise ratio of the latter is not as high as that in MR images.) Additional difficulties stem from the fact that calcified structures (bone and teeth) contain little mobile hydrogen and are indistinguishable from the airway in many images. Consequently, measurements of airway dimensions may incorrectly include space occupied by the teeth. A further drawback stems from the fact that the resolution of air–tissue boundaries can, in part, depend on the thickness of the tissue section that generates the MR signal and, to obtain images of sufficient quality for measurement purposes, that thickness can need to be 0.3–0.8 cm. (In comparison, the thickness of CT and ultrasound sections can be

as low as 2 mm.) Thus there is a tendency for the imaging process to reduce 3-D information to two dimensions, although this tendency is offset to some degree by a slice profile (weighting function) that emphasizes the contribution of the MR signal received from the central plane of a slice relative to that received from its two outer faces. Furthermore, the fact that successive slices are usually spaced at intervals approximately equal to their thickness determines the frequency at which the cross-sectional areas are sampled along the tract and, ultimately, the degree to which area functions can specify articulators such as the tongue whose precise shape and place of constriction has been shown to be a sensitive determinant of formant frequency values (Lin, 1990). Next, it must be pointed out that there is often a great deal of acoustic noise associated with the MR imaging process because magnetostrictive effects induced by the rapid switching of magnetic fields produces loud sound impulses that seriously interfere with attempts to record the subjects' productions. And, finally, it should be said that, unlike lateral x-ray and ultrasonic imaging techniques and like x-ray CT, MR requires that the speech sounds be produced by subjects who must lie in a position not commonly used for speaking. Therefore, there is a valid, and as yet unanswered, question whether prolonged periods of exposure to a posteriorly directed gravity vector does measurably affect the performance of the speech organs. However, notwithstanding all these caveats and limitations, and because MR imaging is a rapidly evolving field, there is reason to be optimistic that many of the most serious problems will eventually be overcome. Even now, MR appears to offer the best opportunity to obtain data on the vocal tract shapes associated with sustainable speech sounds such as vowels, laterals, and fricatives.

In early work on this study (Baer *et al.*, 1987), we selected two male subjects and collected two series of MR images representing the entire lengths of their vocal tracts as they produced the vowels /ɑ/ and /i/. In subsequent imaging work, reported here, we have expanded that repertoire to include the vowels /æ/ and /u/, thus completing the set of point vowels, and have progressed further to obtain images of the pharyngeal region for a larger family of nine vowels. The MR images, up to 0.8 cm in thickness, were obtained in the midsagittal plane and the two orthogonal planes at intervals of 5 mm. In this report, we begin with a description of the MR methods and what they can reveal about the vocal tract. We then proceed to two quantitative studies. In experiment 1, we use MRI data to derive vocal tract area functions. The subjects' original vowel utterances are analyzed and their formant frequencies compared with those of waveforms synthesized on the basis of the measured area functions. In experiment 2, we study the relationship between midsagittal width and cross-sectional area at different points in the vocal tract.

I. METHODS

A. Image acquisition: Experiments 1 and 2

The data reported here were collected in two experiments performed over a period of 4 years using two different MR machines. Both experiments were performed on the same two male subjects, TB and PN, who weighed approximately 65–70 kg and were 183 and 175 cm in height, respectively. They were native speakers of English; TB's dialect came from Nassau County, New York, and PN's dialect retained features of Somerset county, in the southwest of England. Neither subject had received any formal voice or phonetic training.

1. EMR Machine (experiment 1)

The first experiment employed an experimental MR (EMR) machine—a whole body imaging system, developed at Yale University in collaboration with General Electric and based on a resistive magnet (Oxford Instruments) which generated a field strength of 0.15 *T*. This machine permitted the introduction of apparatus for head stabilization, sound recording, and reproduction. It was used to collect data representing the vocal tract configurations of two male subjects covering the full length of the tract from the larynx to the lips while they produced the four point vowels /ɑ æ i u/.

Each subject lay in the supine position on a horizontal patient couch with his head inside a saddle-shaped radio-frequency (rf) coil for receiving the resonance signal (Baer *et al.*, 1987). Since it took about 3 to 4 h to collect all the images needed to specify a given vocal tract configuration, and it was necessary to complete the work in two or more sessions, a cephalostat was used to ensure that the subject's head could be returned to its original position. This device consisted of a custom-made head mold attached to a rigid base and, hinged from that base, a plastic locating wand which was designed to make contact with the subject's nose. This device provided sufficient tactile sensation at the nose to inform the subject of the need to minimize residual head movements. The base itself was attached to the couch by means of a positive tongue and slot mechanism, thus enabling the head to be restored to the same position at the beginning of each experimental session. In an attempt to retain the same vocal tract shape over long periods and reduce any tendency to vary vowel quality, the subjects were equipped with an ear insert and plastic tube through which they heard a canonical production of the target vowel. The source of the vowel was a prerecorded tape. Recordings were made of the subjects' vocal output with an electret microphone placed inside the magnet a few cm from the subjects' lips. A screened cable of length 4 m connected the microphone to a battery-powered amplifier which, in turn, was connected to a tape recorder located outside the electromagnetic shield enclosure. The sound recordings were later digitally analyzed both to gain evidence of any deviations from the vowel targets during image acquisition—deviations that might undermine the accuracy of the images—and to obtain the formant frequencies to be compared with the calculated frequencies of resonance using area functions derived by a digital analysis of the images.

The subject's couch carried a scale graduated in cm and was designed to translate linearly under rack and pinion control along its major (z) axis into the bore of the magnet. A 9.5-cm displacement of the couch permitted a segment of a subject's vocal tract, from the superior margin of the verte-

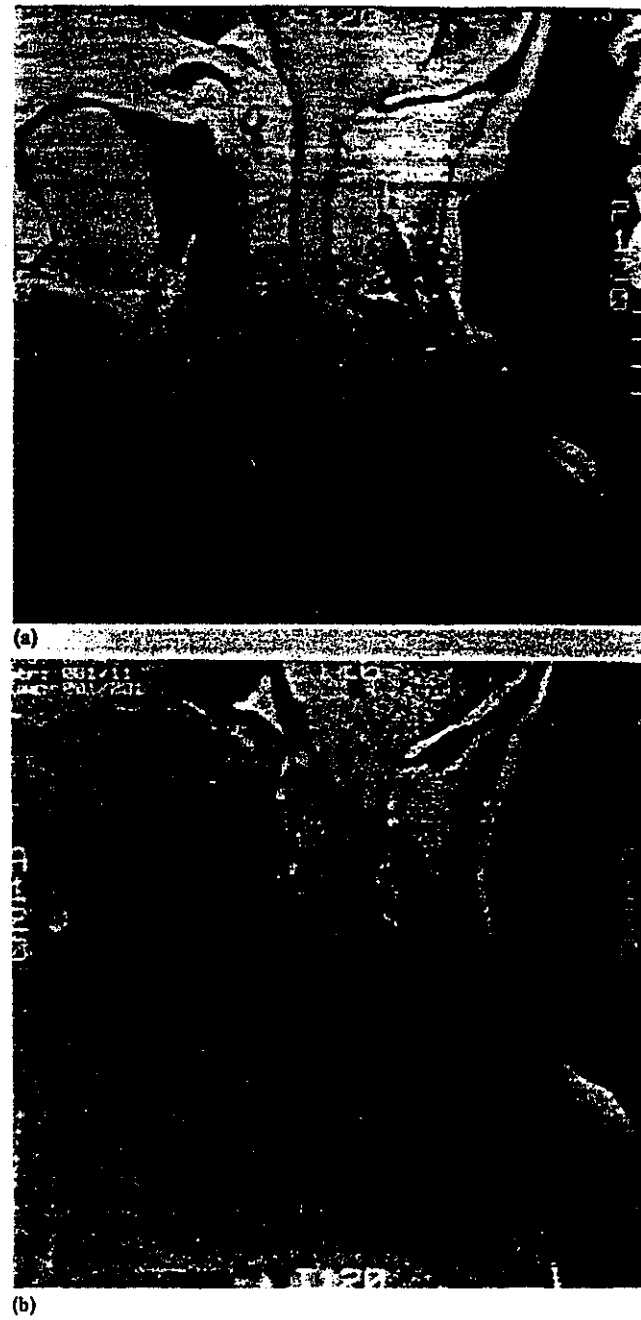
bral axis to the inferior margin of the sixth cervical vertebra, to be examined. This represented the region from the hard palate to the glottis, or the space immediately beneath, depending on larynx height. The procedure involved the acquisition of multiple parallel contiguous axial slices over the 9.5-cm range using a partial saturation spin-echo imaging sequence with a repetition time TR, of 200 ms and echo delay time TE of 11 ms to achieve optimum soft tissue contrast and good signal-to-noise ratio (Bradley *et al.*, 1983). This combination of sequence parameters gave the image a desirable $T1$ weighting. Each image was acquired as 128 phase encoded projections and reconstructed by a two-dimensional Fourier transform to represent an 0.8-cm-thick slice. Slices were spaced at 0.5-cm intervals. Each projection was the average of eight single echoes (four phase cycled pairs) that were each sampled 128 times. A total acquisition time of 3.4 min was required for each 128×128-pixel image which was interpolated to 256×256 pixels for display. Contiguous coronal slices of the same thickness and displaced along the y axis were also obtained at intervals of 0.5 cm by offsetting the rf pulse frequency. These images nominally spanned the region of the vocal tract from the posterior wall of the pharynx to the lips. The subjects were instructed to produce each selected vowel in a sustained monotone for about 15 s between brief inspirations, and to continue doing so throughout the 3.4 min required for image acquisition. Up to 19 axial images and 18 coronal images were required to specify the entire length of the vocal tract; thus the subjects repeated each point vowel over many image acquisition cycles. Not all of the axial and coronal images were equally clear. On several occasions the EMR machine produced poorly resolved images in the region anterior to the alveolar ridge and in the neighborhood of the glottis. These images were not wholely adequate for measurement purposes and may have been a source of error.

2. SIGNA machine (experiment 2)

The second experiment was carried out on a General Electric SIGNA machine specifically designed and primarily used for diagnostic purposes. The competing demands on this machine placed limits on the experimenters' freedom to install equipment that would provide some normally desirable safeguards against experimental error and also limited the available experimental time. For example, circumstances did not permit the use of a cephalostat or sound recording and reproducing facilities and, in deference to time limitations, the majority of the image data selected for collection was confined to axial images of the pharyngeal cavity.

However, some of these disadvantages were offset to a considerable degree by the fact that the SIGNA machine employs a superconducting magnet that develops a higher flux density (1.5 T). This feature makes possible a higher imaging speed than that provided by the EMR system and did much to offset the need for rigorous head stabilization. A full set of images for a given vocal configuration could be obtained in less than 30 min. Axial images were obtained during productions of the four point vowels and five other intermediate vowels /ɪ ɛ ɔ ʊ ʌ/. For one vowel /i/ produced by both subjects, a complete set of coronal and midsagittal images was obtained in addition to the axial images; however, only the axial images of /i/ were analyzed.

The supine position was again adopted for both subjects. Each subject placed his head in a padded universal cranial molding with a built-in rf coil serving both transmitting and receiving functions. To restore the head to a predetermined position, laser reference beams were employed to define the volume of tissue to be imaged, and a digitally controlled mo-



(a)

(b)

FIG. 1. Midsagittal images of the vocal tracts of both subjects, (a) TB and (b) PN, obtained during production of the vowel /i/. Soft-tissue organs of speech are seen more clearly than in a lateral x ray. Calcified structures such as teeth, bone, and some cartilage are outlined by soft tissue boundaries such as fat, gum tissue, or the tongue. The posterior projection of the epiglottis of subject TB, as compared to PN, was a source of differences in midsagittal width between the two subjects. The images of nose and lips on the right edge, and cortex at the bottom of each frame are aliasing artifacts.

tor conveyed the couch and subject the required distance into the bore of the magnet. The subjects employed the same vowel production technique used in the earlier study.

Multislice $T1$-weighted spin echo images were obtained using the two-dimensional Fourier transform method. Each image was acquired with TR = 800 ms, TE = 20 ms, using an image matrix of 256×256 over a field of view of 20 cm. Usually a single acquisition (NEX = 1) was used for each of the 256 phase encoded projections. At the higher field strength and consequently higher magnetic resonance frequency (64 MHz), the images with TR = 800 ms demonstrated similar contrast to the low field images at TR = 200 ms because $T1$ is longer at higher frequencies. The longer TR also permitted the simultaneous acquisition of up to 17 parallel slices. In multislice mode, after each echo is acquired from one location, the rf pulse excitation frequency is incremented and selective pulses are transmitted to produce an echo from an adjacent location. This process is repeated rapidly and permits the acquisition of one projection from each of up to 17 separate slices in the interval TR but takes no longer than the time required to produce one image (3.4 min). However, the procedure also produces a continuous noise generated by magnetostrictive reaction to the gradient switching, which occurs at a frequency of about 50 Hz. This noise precluded the use of a microphone to record the subjects' vowel productions during image acquisition and would have made problematical the task of presenting canonical productions, had there been sufficient time available to attempt it. Therefore, to tell the subject which vowel to produce next, an experimenter would confirm instructions already conveyed over the intercom system by briefly entering the magnet room before image acquisition and repeating the vowel to ensure that it had been heard correctly. Figures 1–5 show examples of images generated by the SIGNA machine.[1]

B. Measurement of cross-sectional areas: Experiments 1 and 2

With minor exceptions, the same methods of image acquisition and dimensional measurement were used in both experiments 1 and 2. In this section, we describe the procedures used to obtain the dimensional measurements.

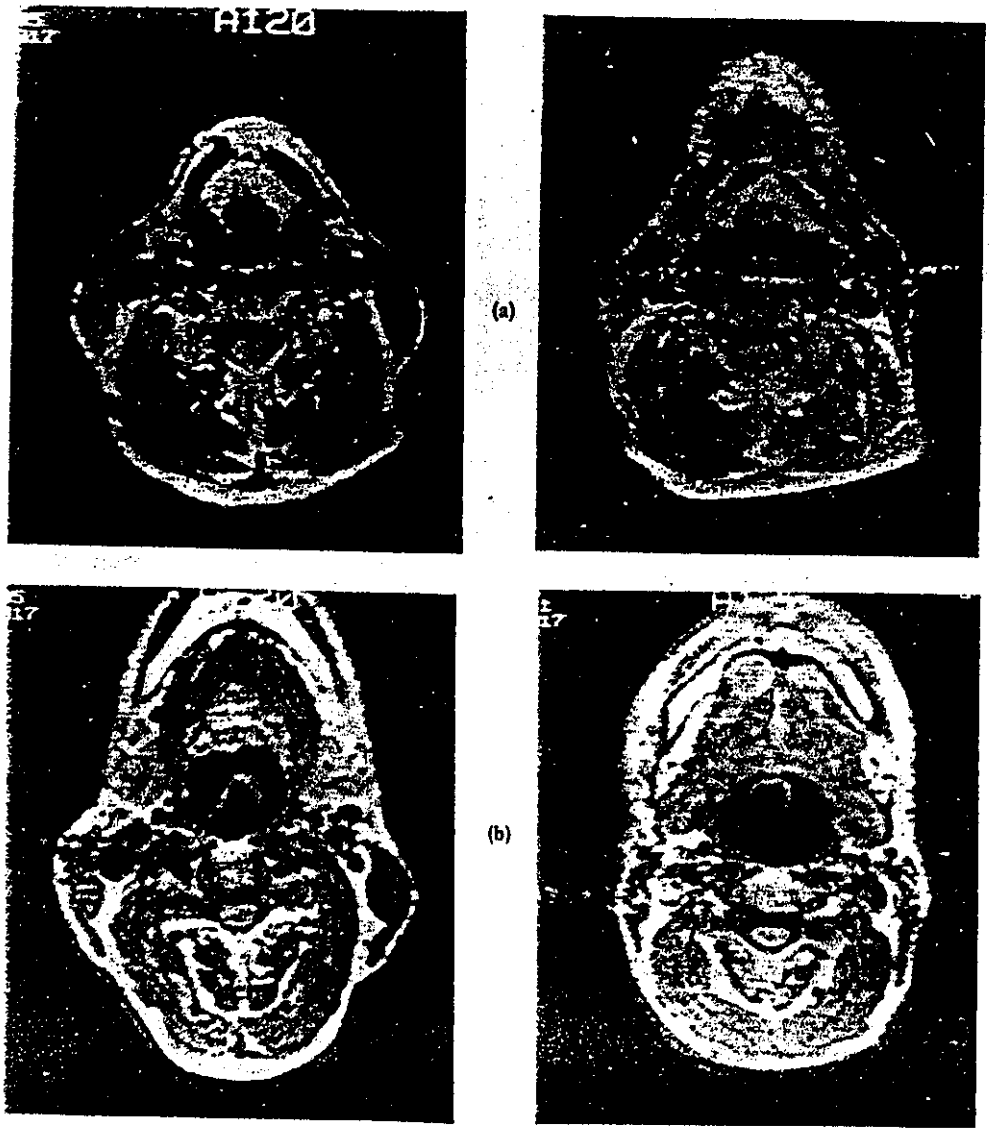


FIG. 2. (a) Examples of axial images near the base of the epiglottis. The aryepiglottic folds separate the central larynx tube from the piriform sinuses on either side. White structures anterior to the airway are the pre-epiglottic fat pad and cartilaginous tissue at the base of the epiglottis. The image plane includes the lower edge of the chin for subject PN (right) but not for TB. (b) Axial transections of the pharynx and epiglottis at a level 2.5 cm higher than the preceding images. The body of the epiglottis appears as a crescent-shaped figure in the vocal tract airway. The hyoepiglottic ligament, which extends from the epiglottis to the hyoid bone and the base of the tongue, can be seen (left) in the midline anterior to the epiglottis.

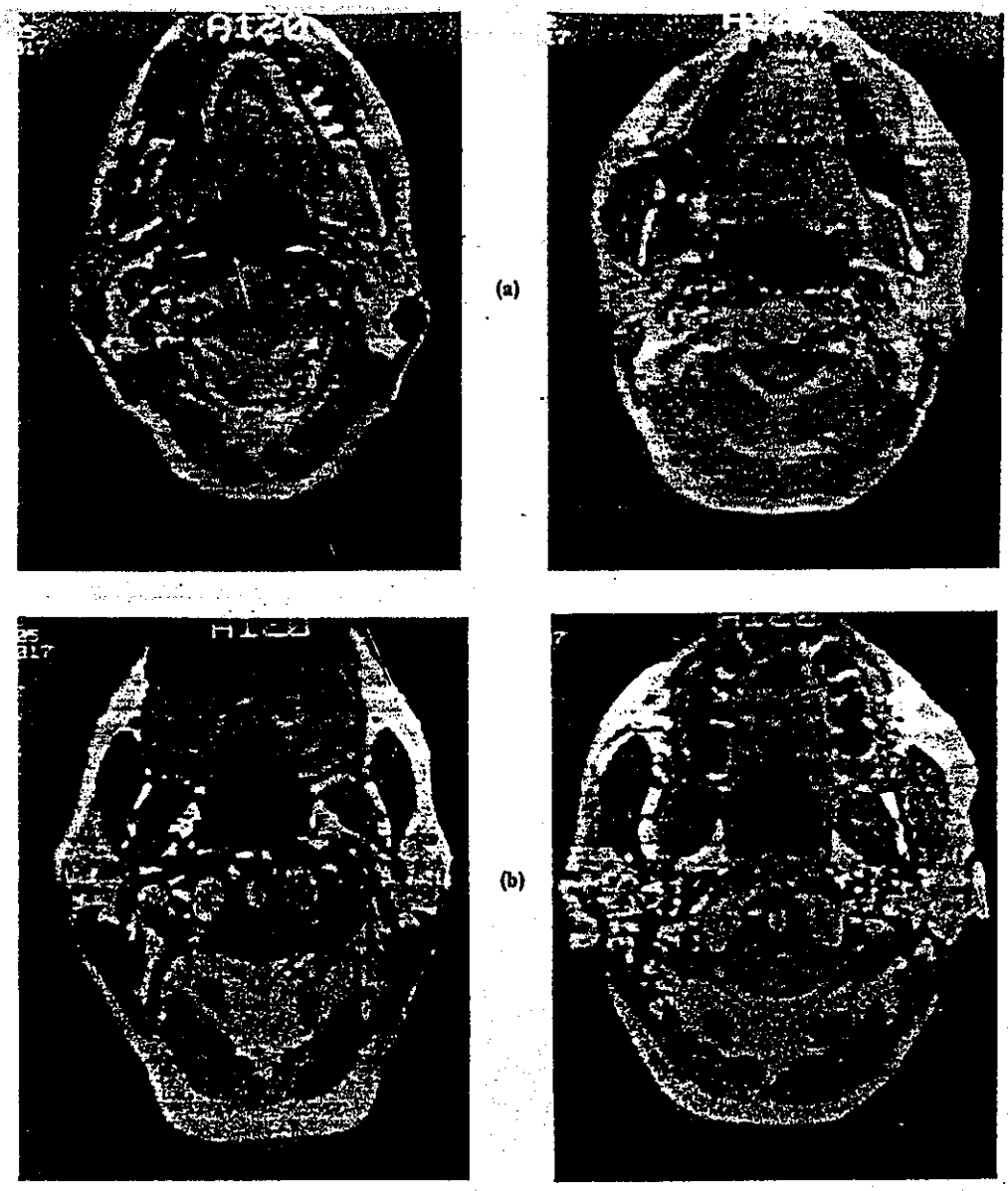


FIG. 3. (a) Axial transections of the pharynx at a level 4.5 cm above the level shown in Fig. 2(a). An angularly shaped airway boundary can be seen that includes a V-shaped tongue groove. The groove is associated with genioglossus muscle activity that typically accompanies the production of high front vowels. The lower teeth appear in contrast to soft gum tissue. (b) Axial images that show the vocal tract shape at the level of the atlas (C1), a distance of 6.5 cm above the images shown in Fig. 2(a). The lateral masses of the C1 are apparent, as well as the centrally placed odontoid process (dens), an extension of the axis (C2). Illustrated in the left image (TB), the uvula is seen adjacent to the posterior pharyngeal wall. The right-hand image (PN), from a slightly higher level, transects the lower region of the velum that separates the vocal tract from the nasal cavity. In both images, the lateral and medial pterygoid muscles are seen lateral to the vocal tract.

1. Boundary tracing

The vocal tract airway was identified in each 256×256 pixel image and the profile of the image density gradient was plotted along vertical and horizontal lines passing through its center. Then, using a trackball or graph pen, the contour defined by the 50% level of the density profile was traced by hand, under ×3 magnification to minimize tracing errors, thus yielding a set of x, y pixel coordinates that followed the perimeter of the airway. In the case of SIGNA images acquired in experiment 2, closed air–tissue boundaries such as typically occurred in the pharyngeal region, were traced by a computer algorithm, which automatically followed the 50% density contour and obtained the sets of x,y coordinates (Martelli, 1976). The boundaries of structures found within the airspace such as the uvula and the epiglottis were traced separately. Along the upper vocal tract, where the image plane intersected the teeth, the lack of a detectable air–tooth density gradient prevented use of the automatic tracing procedure. In these circumstances, we traced the boundaries by hand, under magnification, using data on tooth size and location obtained from x-ray images and visual estimates. Figures 6 and 7 show examples of boundary tracings obtained from EMR images of tract configurations for the vowels /i/ and /ɑ/.

2. Calibration of images

As a precautionary test of the EMR machine, a calibration procedure was devised. This procedure employed a 16.5-cm-long wedge made from 6.5-mm-thick Plexiglas and filled with mineral oil. Axial, midsagittal, and coronal images of this structure were acquired and the perimeters of the strong resonance signal emitted by the oil were measured and compared with the known interior dimensions of the wedge. From this comparison, calibration factors were derived that established the number of pixels per cm in each of the three planes. Similar calibration tests were not undertaken on the SIGNA machine because reliable calibration data were already available and imaging time was in short supply.

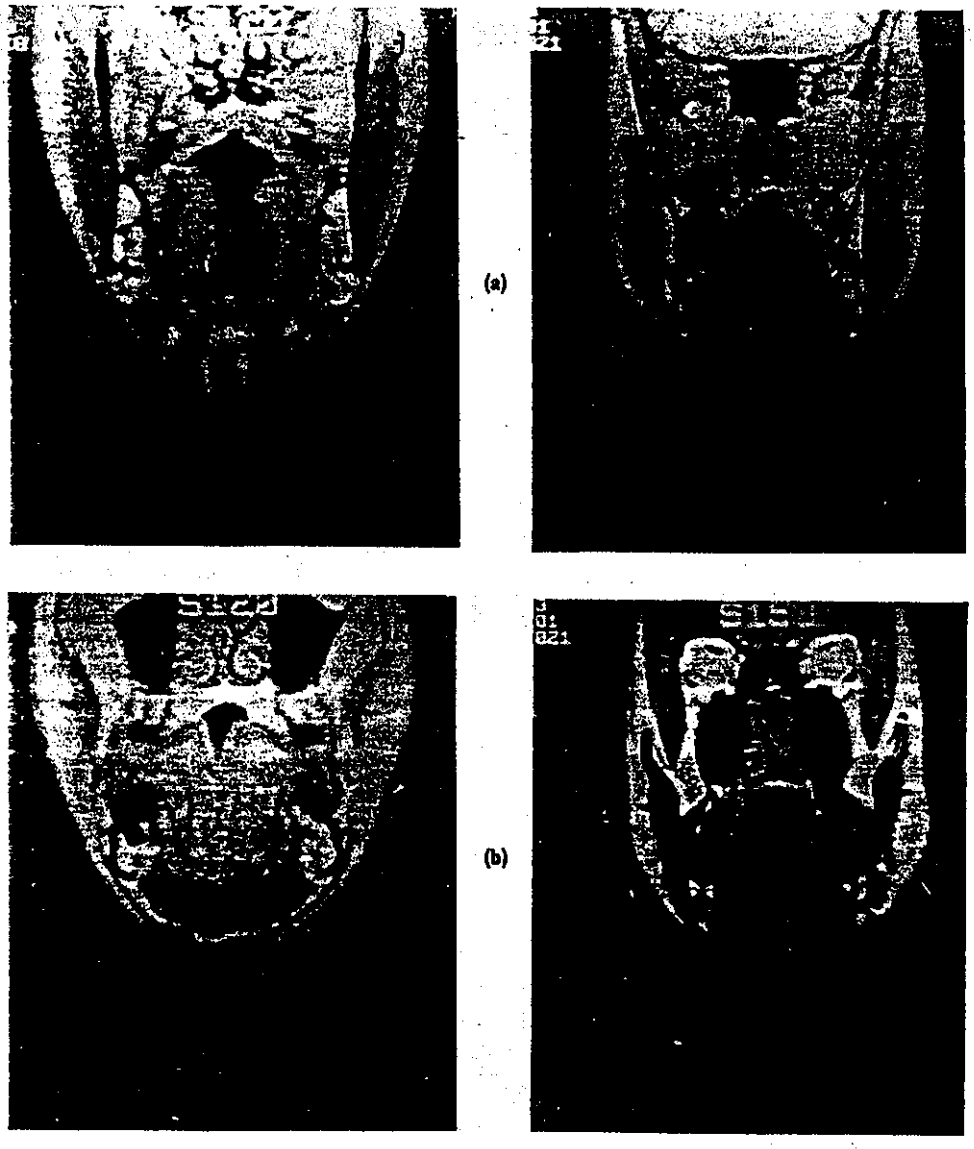


FIG. 4. (a) Coronal sections transecting the anterior region of the pharynx. The image from TB (left) originated from a slightly more anterior plane than the image from PN (right). The vocal tract airways in both images are bound laterally by tongue muscle. Superior to the palate, the two halves of the nasal tract are separated by the nasal septum (TB), and the conchae are seen abutting its lateral walls. The sphenoid sinus appears as a dark space above the nasal cavity (PN). At the inferior margins of the airspaces are the connections between lingual and laryngeal structures. For TB, the fatty tissue at the base of the epiglottis and the ventricular folds appears above the darker boundary of the laryngeal ventrical. For PN, the outline of the epiglottis appears above the superior margin of the thyroid cartilage. (b) Coronal sections through planes located 1.5 cm anterior to those shown in the preceding images. The airway has become a small channel between the domed hard palate and the grooved surface of the tongue. For TB, the groove is sharp and deep because the the image plane transects the tongue surface obliquely. Lateral to the tongue dark areas represent the jaw bone and lower teeth. Above the palate, the nasal airway passes through the turbinates, and is bounded by the maxillary sinuses.

3. Effects of tract motion

Blood flow and respiratory motion introduced artifacts that tended to blur air–tissue boundaries. These artifacts were easily identified and the errors that might have been generated by the boundary tracing algorithm were minimized by substituting hand tracing. Another form of motion arose from the subjects' inability to reproduce exactly the same vocal tract configuration in each of up to 20 consecutive 3.4-min image acquisition cycles attempted during the course of a typical 2-h session in the EMR imager. Thus even two consecutive images of the same location in the tract did not yield precisely the same area measure. Analyses of sample groups of six repeated images made hours or days apart showed the standard deviation of the variation to be within 6% of the mean area and extreme variations in pharyngeal area to approach as much as 15%. In light of the necessarily long session lengths involved, particularly in experiment 1, fatigue probably caused variability in the vocal tract shape and consequent variability in the dimensional and acoustic measurements. It may be assumed that fatigue was a less disruptive factor in the pharyngeal data collected by the SIGNA machine because of its shorter image acquisition time. However, no repeated images were collected during experiment 2 to verify this assumption.

C. Determination of area functions: Experiment 1

In this section we describe the steps involved in deriving area functions from the sets of airway boundary tracings sampled in Figs. 6 and 7.

1. Airway volume reconstruction

The sets of axial and coronal boundary coordinates were first placed in their correct relative positions in a 3-D matrix whose cell density was the same as the pixel density of the original image arrays. Next, the cells located within each image boundary were labeled with nonzero values to distinguish the region occupied by air from the surrounding tissue. Cells representing the uvula and epiglottis were set to zero to prevent the areas of these structures from being included in the airway. Then, the 0.5-cm unoccupied spaces between successive contours were filled by a process of duplication that effectively thickened each axial or coronal section until

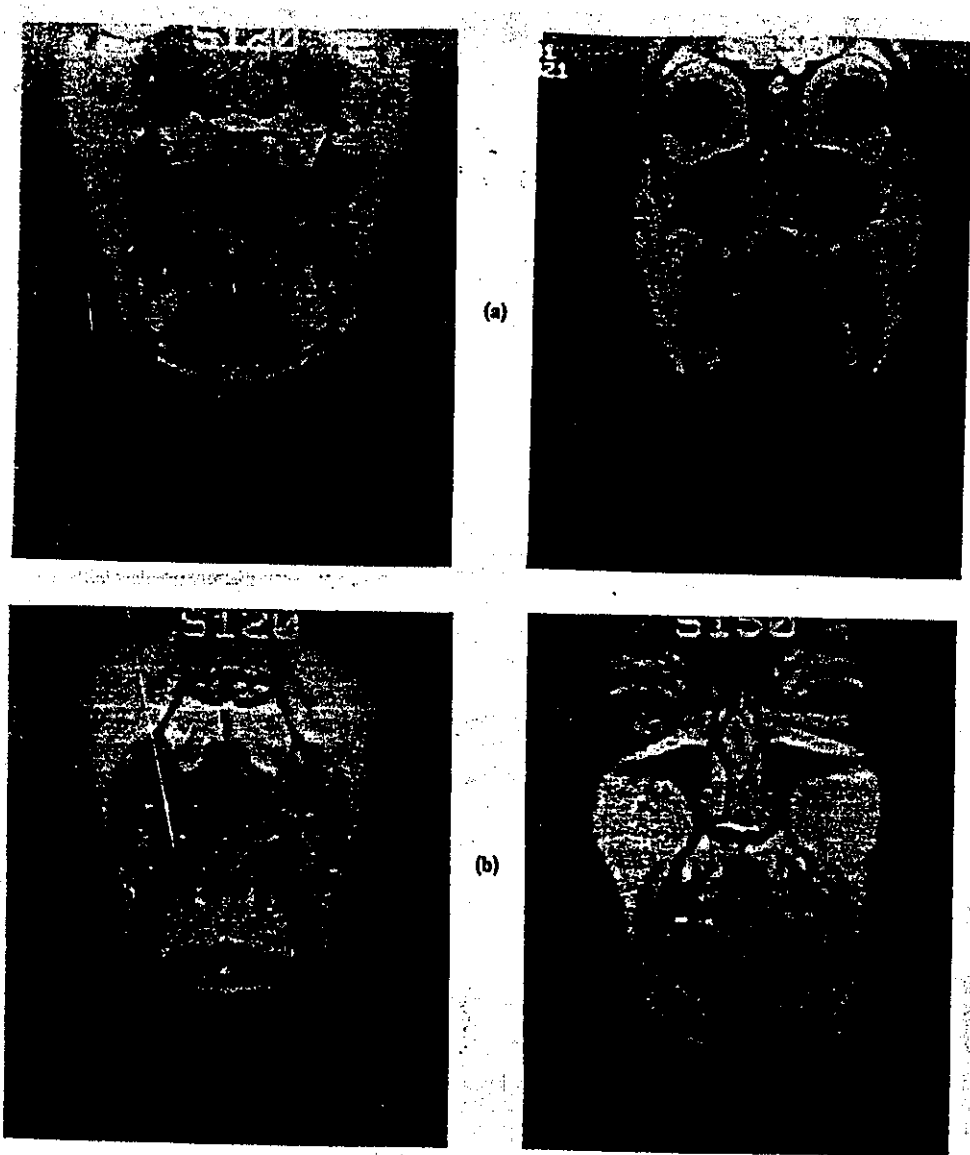


FIG. 5. (a) Coronal sections acquired in planes of intersection lying close to the narrowest oral constriction formed during production of the vowel /i/, a distance of 3.0 cm from the planes shown in Fig. 4(a). The tongue fills most of the palatal dome, leaving a small channel at the midline. For TB (left), the tongue also fills the space between the upper and lower molars which, in contrast, appear as dark spaces. For PN (right), upper and lower teeth also appear to be parted and the nasal airway and maxillary sinuses are still visible. (b) Coronal images that transect the constricted front cavity of the vocal tract at a distance 4.5 cm anterior to Fig. 4(a). For TB (left), the tongue again fills the space between the upper and lower incisors which are seen imbedded in the mandible. The live tissue of the mandible surrounds a round calcified structure, the mental protuberance, located on the midline.

it bridged half of the gaps between itself and adjacent sections. In this manner, a 3-D digital volume of labeled cells was formed. Figures 8 and 9 show the surfaces of two of these volumes, representing the airways of two vowels, drawn in a 3-D format. The images are plotted from a number of different perspectives by a subroutine selected from the NCAR graphics software (Henderson and Clare, 1979).

Variations among repeated productions of the same vowel are the principal cause of the occasional, and otherwise inexplicably abrupt, changes in area between adjacent cross sections of the point vowels shown in Figs. 8 and 9. When such discontinuities were noticed early enough, and the opportunity to reacquire some images remained available, new image data were collected. However, because the acquisition of necessary data processing equipment and software lagged behind data collection activities, not all of the noticeably abrupt changes in area were caught early enough to permit image reacquisition.

The issue of whether a discontinuous datum should be retained, omitted or replaced because of its failure to be consistent with its neighbors (or with the experimenters' assumptions about the underlying anatomy) was resolved in the following way. There were two stages in the data processing at which this issue came to the fore. During the airway boundary-tracing stage, if the shape or area of the boundary appeared inconsistent with neighboring sections and a better conforming boundary existed (e.g., the nonconforming boundary may have been derived from one of the last two or three image-collecting cycles in a given session that were customarily repeated at the beginning of the following session), the better fitting or more plausible boundary was selected. If no substitute boundary trace existed, then the existing observation was preserved intact. Among the potential causes of unusual irregularity were image distortion due to changes in the intake temperature of the water supply used to cool the magnet of the EMR system and the presence of increased pixel noise due to fringe effects at the ends of the receiver coil. Occasionally, as a result of applying the above data-processing rules, the boundaries of minor structures within some images would be omitted. For example, a sinus cavity that made its appearance in one image would fail to show up above the threshold (determined by the mid point

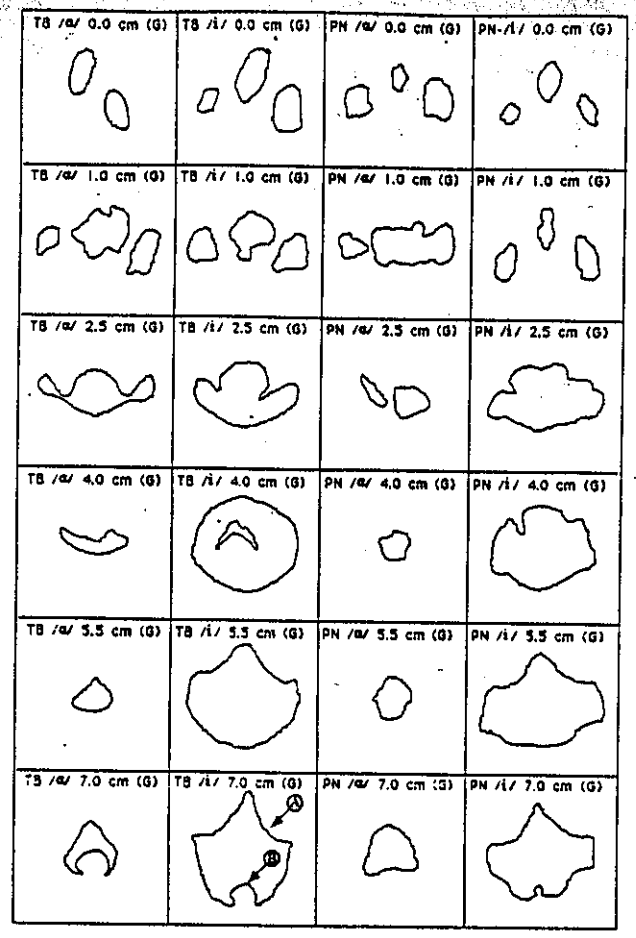


FIG. 6. Digital tracings of axial images of the vocal tract from the EMR machine taken during the production of two point vowels (col. 1, TB/a/: col. 2, TB/i/: col. 3, PN/a/: col. 4, PN/i/). All tracings are oriented with the anterior region of the pharynx at the top of each cell. Cell labels indicate approximate distance in cm from the glottis (G). One of the piriform sinuses in col. 1 is absent because it did not meet the threshold criterion. Arrow (A) indicates the tongue groove and arrow (B) the uvula.

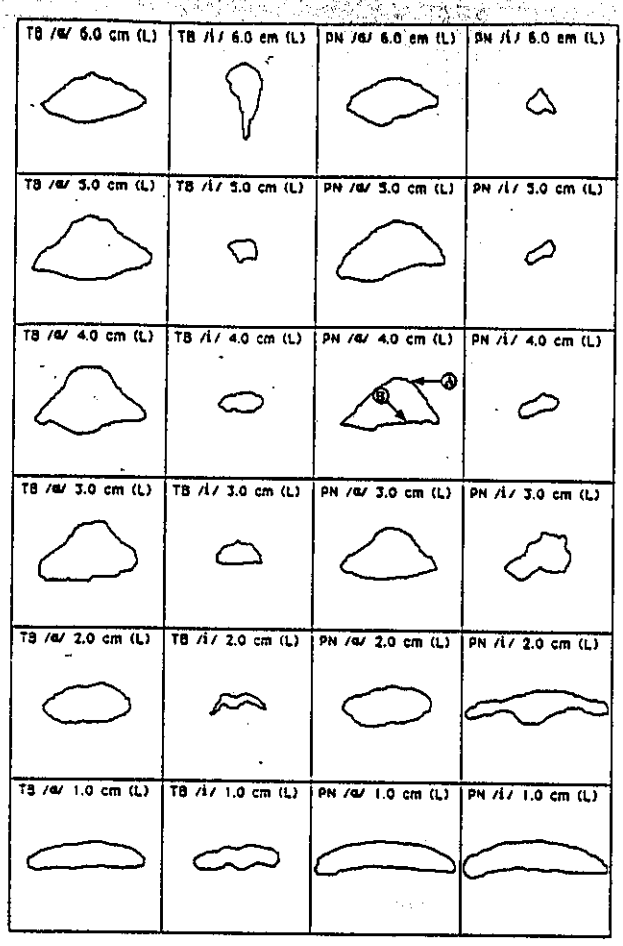


FIG. 7. Digital tracings of coronal images from the EMR machine. The tracings are oriented with the palate vault uppermost. Distances from the lips (L) measured in cm are indicated in each cell. This figure, and the figure that precedes it, illustrate the striking differences in both pharyngeal and upper vocal tract area that are characteristic of the vowels /a/ and /i/. Arrow (A) indicates the dome of the palate and arrow (B) the surface of the tongue.

of the gradient of the airway boundary) in a succeeding image, despite being visible to the eye. Under these circumstances, the structure would be ignored despite the fact that it might reappear again in a third image. This practice was responsible for the isolated islets that appear, as if suspended in space, in Figs. 8 and 9.

Thus, to summarize the data-processing procedure, we endeavored to avoid *post facto* data selection and manipulation and to rely solely upon computer-evaluated objective criteria based on the properties of image gradients to determine the paths of the boundary contours and their inscribed areas.

2. Calculation of area functions

An area function was derived from the digital representation of each airway volume described above. The areas were measured from a series of planes spaced at intervals of about 0.5 cm along the approximate midline of the tract. The reference frame or grid system used to locate those planes is shown in Fig. 10. Close antecedents of this grid are the systems used by Heinz and Stevens (1964), Ladefoged *et al.* (1971), and Mermelstein (1973) although, from this group, the system we adopted resembled most closely that of Mermelstein. In particular, the horizontal gridplanes are coincident with the axial sections and are spaced at 0.5-cm intervals starting in the region of the larynx. The vertical gridplanes coincide with the coronal sections and have the same spacing in the anterior part of the oral cavity. The transition between these regions is bridged by a 90-deg arc with a radius of 4.2 cm—selected because, with suitable vowel-by-vowel adjustments of the origin, a passable fit to the midlines of all four point vowel tract configurations could be found by eye. Radial gridplanes span the arc at 7.5-deg intervals and, thus, intersect the tract midline every 0.55 cm.

The areas of the airway (including all cavities with connections to the main airway) intercepted by the grid planes were calculated by Simpsons Rule and converted to cm^2 with the aid of calibration factors. Adjustments of the area values

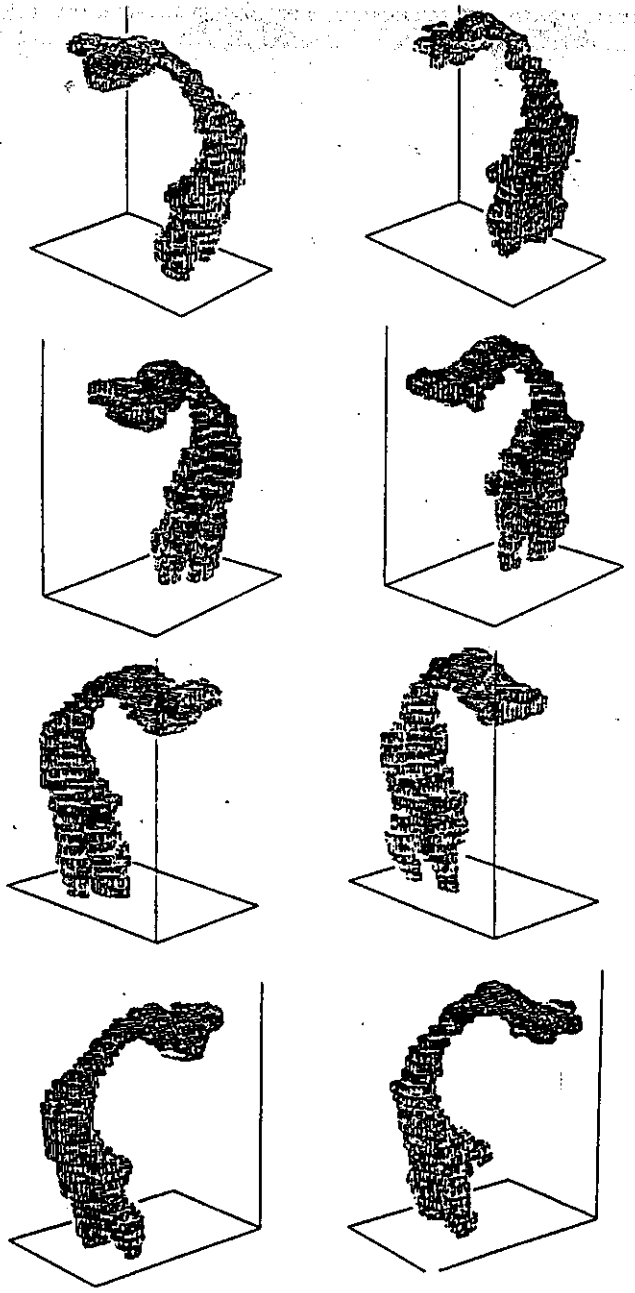


FIG. 8. Computer-generated views of the airway shape adopted for the vowel /u/ and shown for subjects TB (left column) and PN (right column) from four different perspectives. The original images from which these plots evolved were acquired by the EMR machine.

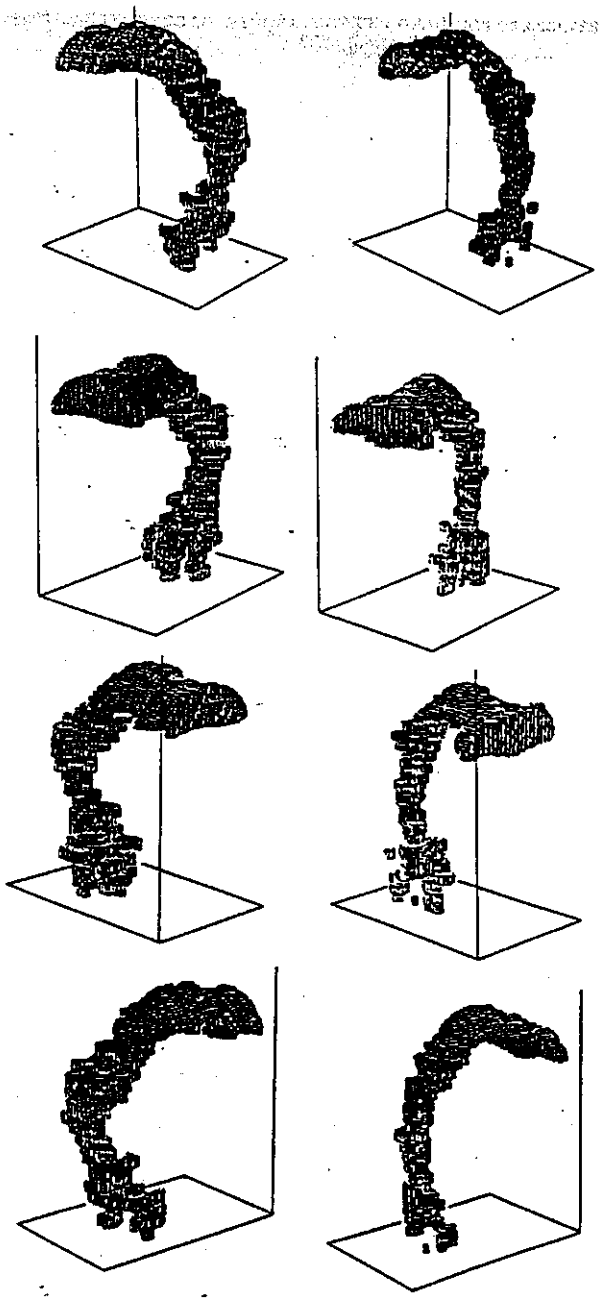


FIG. 9. Computer-generated views of the airway shape adopted for the vowel /æ/. This figure and its predecessor show examples of larger than usual discontinuities between adjacent sections due to the subjects' inability to precisely repeat vocal tract configurations.

were made by linear interpolation to correct for the slight undersampling of the arc, and the adjusted values were then tabulated as a function of their distance from the larynx. Repeating this procedure for each point vowel led to the generation of four area functions for tract lengths computed to the nearest 0.5 cm. Next, for the purposes of comparison, four additional area functions were generated, each omitting the areas of the sinuses. Finally, from each of these data sets, area values were linearly interpolated at intervals of 0.875 cm and employed in waveform calculations using the Kelly and Lochbaum (1962) method, as implemented by Rubin *et al.* (1981).

D. Convergent evidence from other sources: Experiment 1

Measurements relating to the cross-sectional area of the oral cavity, the overall length of the vocal tract and the acoustics of the subjects' original vowel productions were obtained. Their use in refining and assessing the accuracy of the area functions is reported in this section.

1. Oral-dental impressions

With a view to avoiding the error of including areas occupied by teeth as contributing to the airspace, the cross-

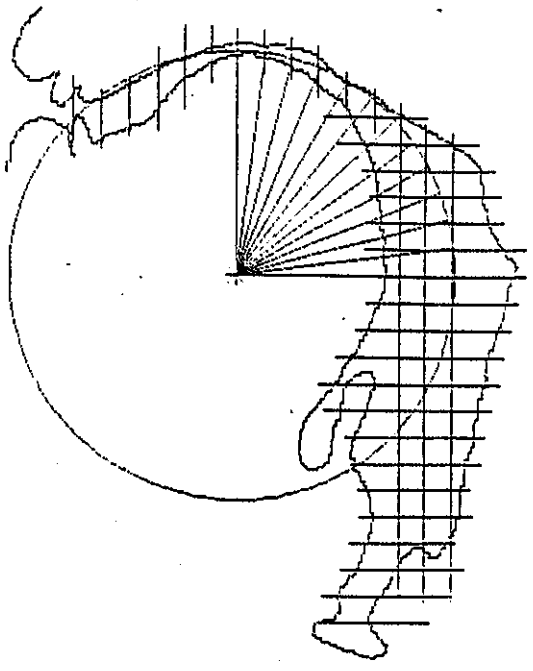


FIG. 10. Gridplane system used to define points of transection of the central axes of 3-D digital models of the vocal tract at intervals of approximately 0.5 cm. The grid is shown superimposed on a digitized tracing of the midsagittal section of a tract shaped for production of the vowel /i/.

sectional areas derived from coronal images of the oral cavity were checked against estimates obtained from dental impressions. Vinyl polysiloxane (3M) impression molds were made of each subject's palate and dentition and repeated 0.5-cm sections were then cut with a rotary slicer in the coronal plane. The outlines of the sections were hand-traced in digital form by means of a graphics tablet and superimposed on corresponding coronal images identified by means of palate height, dental root structure, and related anatomical features. Each of the existing hand-drawn boundary lines was then compared with the new digitized sections obtained from the dental molds and, where necessary, redrawn to reflect the presence of dental constrictions.

2. Comparison of MR tract lengths with x-ray data

The combined uncertainty arising from the 0.8-cm thickness of the MR sections and the lower signal-to-noise levels at the lips and larynx owing to their proximity to the extreme ends of the EMR receiver coil, led the experimenters to the conclusion that the length of the tract could not be determined to better than about ± 0.6 cm at each end. Therefore, as an independent means of verifying the tract length derived from the MR data, sagittal Xerographic x-ray images were obtained from both subjects while they repeated each of the four point vowels in a supine position closely resembling that adopted in the MR machines. The importance of duplicating the MR posture had been brought to our attention by the earlier study of Baer *et al.* (1987). A barium paste was used to enhance the visibility of tract boundaries and, with the aid of a map measure and ruler, the tract center line was traced by hand from each Xerograph. Finally, using as a scale of reference the image of a phantom of known dimensions fixed to each subject's midplane during the x-ray exposure, the measurements of tract length were calibrated and the distances along the midline of the tract from the glottis to the lips were determined to within about 1 to 2 mm.

3. Acoustic analyses of original utterances

The acoustic resonances of tubes having the derived area functions were calculated by the version of the Kelly-Lochbaum algorithm used in Mermelstein's articulatory model (Mermelstein, 1971; Rubin *et al.*, 1981). To compare these resonances with those of the original productions, the recordings of vowel productions made during image acquisition by the EMR machine were analyzed acoustically. The procedure used involved the extraction of the first three formants, by means of a 14-pole LPC analysis of successive 25.6-ms-long frames, the formation of independent histograms for the frequencies of each of the three formants, and the identification of the mode of each distribution. The purpose of this procedure was to achieve the automatic rejection of all portions of the original vowel signal that were acoustically contaminated by the magnetostrictive impulses, which constituted about 17% of the total signal, since the exponential decay time of each pulse was 40 ms in duration.

4. Perceptual analysis of computed vowel waveforms

To ascertain the perceptual acceptability of the vowels that would result from sound excitation of the area functions, four groups of vowel waveforms were computed: Both subjects provided two sets of area functions, one that included the sinuses and another that did not. Thus 16 vowel stimuli were generated. Each stimulus was repeated 20 times, and the entire ensemble of 320 sounds was randomized and recorded on tape in 16 blocks of 20 stimuli each. A pause of 3 s after each stimulus provided enough time for the listener's response and a 10-s interval marked the end of each block. A panel of 20 naive listeners was provided with earphones and response forms. They were asked to identify each stimulus in turn by making a check mark in a column below the most appropriate of the four point vowel symbols appearing on the form. The listeners were also instructed to respond within the 3-s interval provided and not to attempt to enter or correct responses retroactively. This experiment was followed by an identically organized test using 350-ms samples of the subjects' original vowel productions (which had been hand edited to remove the magnetostrictive noise) and a, less formal, open response test performed on a small group of phonetically trained listeners.

E. Measurement of midsagittal dimensions: Experiment 2

The methods used to obtain the midsagittal width and area data from the sequences of pharyngeal images collected in experiment 2 were virtually identical to those described

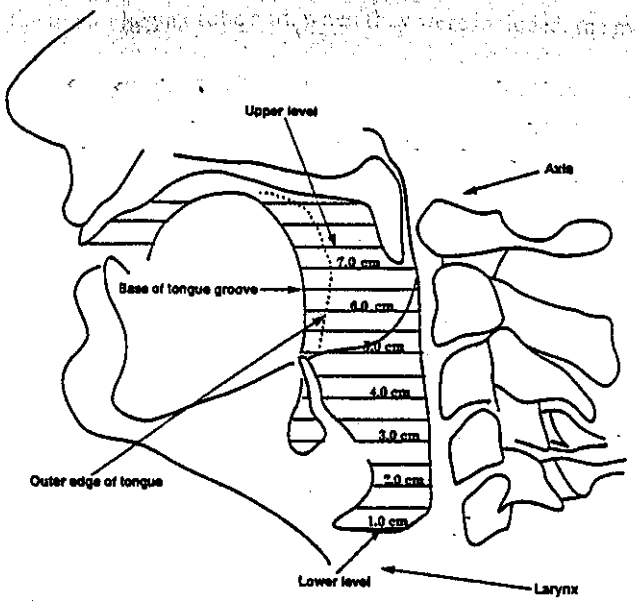


FIG. 11. Positions of the axial images acquired by the SIGNA machine for a study of the relationship between midsagittal width and cross-sectional area in the pharynx. Images from a total of 14 locations (covering a pharyngeal distance of 7 cm) were obtained for each of nine vowels from subject TB. A corresponding set of images covering 12 locations and the same vowel repertoire was also acquired from subject PN.

earlier in experiment 1 and do not need repeating. This section describes the analytical procedures unique to experiment 2.

1. Acquisition of pharyngeal data

Pharyngeal images of 0.5-cm-thick slices spaced at 0.5-cm intervals and spanning an 8.5-cm length of the tract were obtained from the SIGNA machine while each subject produced each of the nine vowels in turn. The approximate orientation of these images with respect to the spine is shown in Fig. 11. However, because the subjects' heads were not identically positioned within the machine, the pharyngeal volume specified by the 17-image arrays was not precisely the same. To confine the field of view to the pharynx and avoid entering the oral cavity, an upper endpoint at the level of the inferior margin of the second cervical vertebra was selected for both subjects. The images that fell below the chosen endpoint formed subsets of the 17-image arrays. For subjects TB and PN the subsets of retained images numbered 14 and 12, respectively. These images were converted into sets of x, y coordinates by the boundary-tracing algorithm.

2. Measurement of pharyngeal dimensions

For each image, the pharyngeal area was calculated from the appropriate set of x, y coordinates. The distance between the most anterior point on the margin of the tongue (or epiglottis) and the most posterior point on the wall of the pharynx was defined as the midsagittal width. In practice, when measuring midsagittal width, the point selected as the posteriormost point of the pharyngeal wall would vary in accordance with whether the piriform sinuses were to be included or excluded from the measurement. When they were excluded, the posterior point lay on the rear wall of the larynx tube and, when they were included, the most posterior point of the boundary of the rearmost piriform sinus was selected. The behavior of the epiglottis complicated the application of our measurement criteria for determining midsagittal width. In all cases, when the cross-sectional area of the pharynx at a height of 3.0 to 4.0 cm above the glottis shrank to about 1.0 to 1.5 cm^2, the epiglottis made intimate contact with the tongue surface and, therefore, for practical measurement purposes, the epiglottis then constituted the surface of the anterior pharyngeal wall. When larger cross sections occurred, however, the epiglottis could lose contact with the tongue root, make a posterior projection into the pharyngeal cavity, and thus transfer the role of defining the anterior pharyngeal wall to the tongue surface. This tendency was considerably more prevalent in case of TB than PN, the tip of whose epiglottis visibly lost contact with the tongue root only during the larger pharyngeal configurations. If, as a result of a posterior projection of the epiglottis, the valecular sinuses became visible, then, for the purposes of measuring midsagittal width without sinuses, the areas occupied by the valecular sinuses would be ignored and, the most anterior point on the posterior surface of the epiglottis would then serve as the anterior boundary of the pharynx.

The distances between the extreme left- and right-most boundaries of the airway (the lateral width) were also measured. In Sec. III C below, these data are used to derive the relationship between midsagittal and lateral dimensions and the relationship between midsagittal width and area. These derivations are then compared to those of Sundberg *et al.* (1987).

II. RESULTS

A. Qualitative observations

1. Image quality

Both MR machines produced images composed of an array of 256×256 two-byte pixels. Figures 1–5 show a few of the more than 600 images that were obtained during the course of this study. Figure 1 contains a pair of midsagittal images of the tract and Figs. 2–5 contain transaxial and coronal images. All of these images were generated by the SIGNA machine during production of the vowel /i/ (see footnote 1). Images obtained from both subjects are shown in pairs that depict roughly corresponding planes through the vocal tract.

In the images, the whitest areas indicate regions with the highest concentrations of hydrogen, such as fatty tissue and bone marrow. Muscle and connective tissue appear in varying shades of grey. The darkest regions show airspaces (the vocal tract and sinuses), calcified structures such as bone and teeth, and blood vessels. The blood within these vessels is, of course, hydrogen-rich, but it does not image because it is flowing rapidly and moves out of the image plane in the time between rf excitation and echo acquisition. Some motion artifact, primarily due to pulsating blood flow in the carotid artery, but also due to respiration, is apparent in some axial images (Fig. 2, subject PN). Much of this artifact was later suppressed by lengthening TE (Fig. 2, subject TB) but, since area measurement was not compromised by the

presence of the artifact, the acquisition procedure was not repeated and these images were retained. Individual cervical vertebrae, which are distinguishable in the images, provided the principal anatomical landmarks used to identify corresponding axial planes in the two subjects' image arrays. Other features such as dental roots, which contrast well with gum tissue, the height of the palate dome, and, the shape of the nasal sinuses served a similar role in the identification of corresponding coronal planes.

2. Vocal tract boundaries

Tracings of the air—soft-tissue boundaries of the vocal tract—are presented in Figs. 6 and 7. The tracings were made from axial and coronal EMR images obtained during the production of two vowels, /i/ and /ɑ/. These vowels have extreme tongue articulations, characterized as high front and low back, respectively. The tracings in Fig. 6 represent a sample of six axial cross sections of the pharynx covering a length of 7 cm, while in Fig. 7 they represent a sample of six coronal cross sections through the upper vocal tract distributed over a distance of 5 cm. The axial planes are identified by the symbol G and their distance in cm from a point located just above the glottis, and the coronal planes are identified by the symbol L and their distance from the lips.

Figure 6 at 0 cm illustrates the appearance of the piriform sinuses just above the glottis. The larynx tube is centrally located, while the piriform sinuses are located posteriolaterally and take the form of two pockets, each approximately 3 cc in volume partitioned from the larynx tube by the aryepiglottic folds (cf., Sundberg, 1974). At a point 1 cm higher, the piriform sinuses remain in view and appear somewhat larger. In tracings made at 2.5 cm above the glottis, the aryepiglottic folds become thinner and reduce the separation of the piriform sinuses from the laryngeal airway. It is at this point that the area of the principal airway begins to change dramatically as a function of vowel identity. The tracings for /ɑ/ assume a more closed posture consistent with a low-back tongue position, while those for /i/ maintain an open configuration consistent with an elevated front tongue position. Meanwhile, the base of the epiglottis during /ɑ/ production projects in a posterior direction, effectively reducing the area of the airway. This phenomenon is particularly evident in the tracings from subject PN to such an extent that the right hand piriform sinus appears to be entirely closed. In contrast, for the vowel /i/ produced by both subjects, the epiglottis adopts a more anterior posture causing the cross-sectional area of the tract to increase.

At the 4.0-cm level in Fig. 6, the vowel differences noted above have become even more marked. In the tracings of /i/ from subject TB, the rim of the epiglottis appears as a crescent-shaped island surrounded by air (the hyoepiglottic ligament and valecular sinuses have been omitted). Such isolation is not apparent in the corresponding tracing from subject PN whose epiglottis consistently adopted a more frontal position that made it appear (as shown here) contiguous with the anterior pharyngeal wall. The presence of tongue grooving, caused by the extreme elevation required for /i/ production, first appears at the 5.5-cm level and extends to the 7.0-cm level. At this point, the uvula makes a bold appearance in both cross sections produced by subject TB. Its much less prominent appearance in the /i/ production of subject PN suggests either that its image was blurred by vibratory motion or that only its tip actually entered the image plane.

Figure 7 contains sample tracings from the upper vocal tract that continue to show the contrast between /i/ and /ɑ/ productions. In the first, most posterior, plane located 6 cm from the lips, the tracing of /i/ from subject TB just captures the remaining portion of the tongue groove. Subsequent planes, located at 5.0 and 6.0 cm from the lips, cover the

TABLE I. Area functions are given for both subjects. Figures not in parentheses represent cross-sectional areas that include the areas of the sinus cavities. Figures in parentheses omit the areas of the sinuses. All cross-sectional areas are calculated at intervals of 0.875 cm starting at the larynx.

Vocal tract area functions for the four point vowels			
Cross-sectional areas (cm^2)			
TB/ɑ/	TB/æ/	TB/i/	TB/u/
1.56 (0.83)	1.91 (0.90)	1.20 (0.81)	3.11 (1.60)
3.10 (1.76)	2.69 (1.22)	2.73 (1.10)	5.18 (2.35)
3.74 (2.96)	4.53	4.11 (3.21)	6.44
2.48	2.05	4.63	6.05
1.28	1.33	6.05	5.76
0.60	1.44	7.62	6.20
0.73	2.82	7.64	5.19
1.28	4.54	7.99	3.72
1.39	4.04	7.09	3.06
1.31	3.24	4.55	2.31
1.43	2.82	2.63	1.05
1.90	3.20	1.81	1.00
3.22	4.32	1.10	0.67
4.44	5.13	0.69	0.92
4.83	4.17	0.85	1.57
3.89	4.98	0.80	2.30
4.72	6.31	0.50	3.78
2.03	5.65	1.10	4.24
2.49	7.03	1.65	3.89
2.82			2.13
			0.66
PN/ɑ/	PN/æ/	PN/i/	PN/u/
2.02 (0.30)	0.48	1.33 (0.71)	0.55
3.34 (2.38)	1.86 (0.28)	1.87 (0.51)	1.92 (0.63)
2.56 (2.04)	1.98 (1.41)	4.31 (3.07)	4.52 (1.41)
1.18 (0.90)	1.92 (1.80)	5.38	6.87
0.76	0.83 (0.74)	6.36	6.87
0.67	0.85	8.11	7.12
0.81	1.68	7.73	6.08
0.80	1.56	6.77	5.09
1.47	1.64	5.68	4.48
2.48	2.34	4.35	2.80
2.85	2.66	2.93	2.11
2.76	2.50	1.64	1.53
3.29	1.77	1.01	0.74
3.60	2.19	0.55	1.20
3.10	2.28	0.54	0.79
2.90	2.35	0.56	1.30
2.53	4.82	1.62	2.03
4.23	7.89	2.37	2.79
		3.33	3.36
			2.44
			1.07

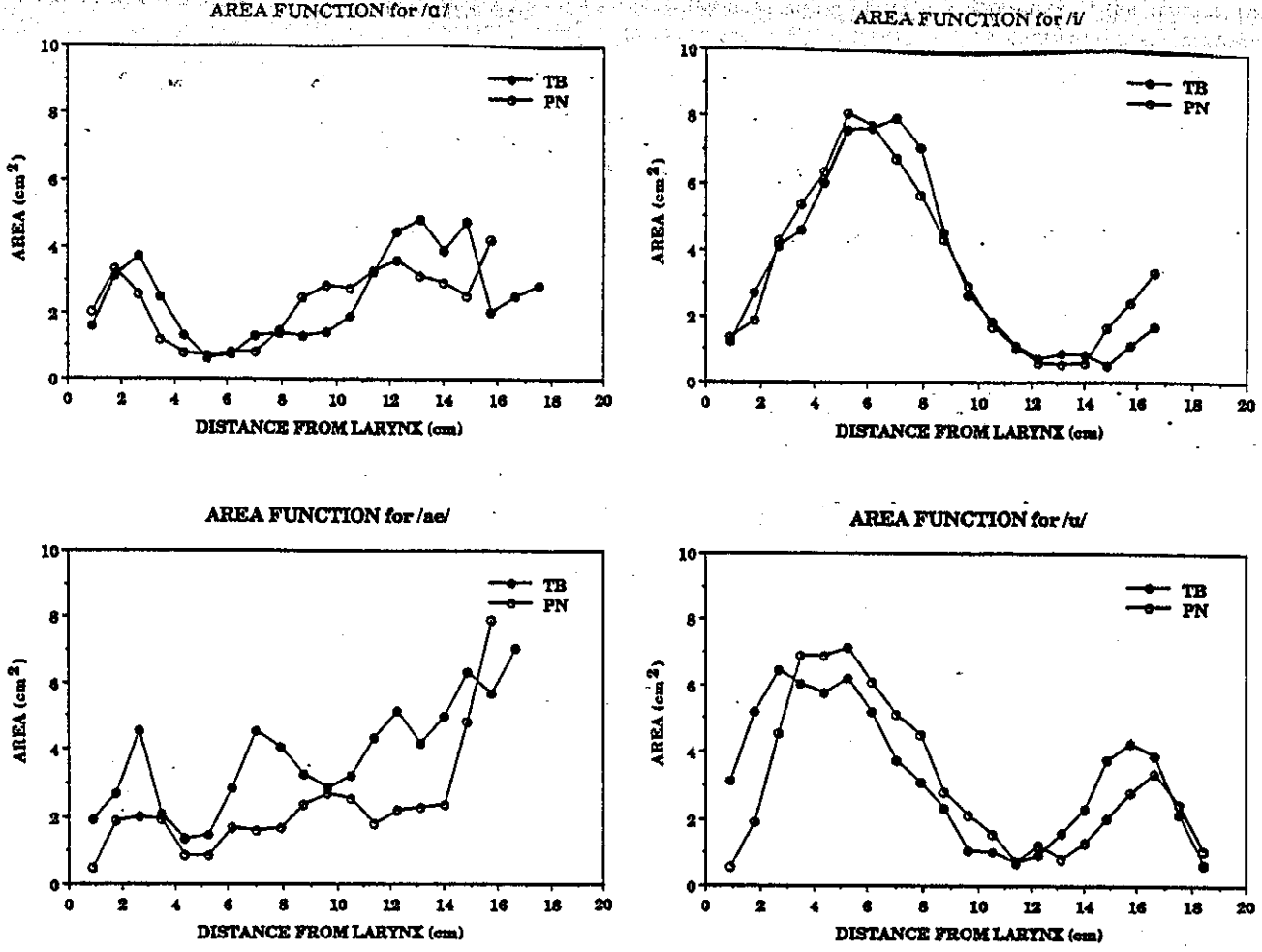


FIG. 12. Graphs of vocal tract area including sinuses for each of the four point vowels as a function of distance from the larynx. Keys to plotting symbols identify subject and vowel. The data plotted in these graphs are available in Table I.

transition from the soft to hard palate, which becomes increasingly arched as the lips are approached. This feature can be seen most clearly in the tracings of the open /ɑ/ configurations from both subjects. In contrast, during /i/ productions, the tongue is thrust upward against the hard palate to form a narrow, often asymmetrically shaped, constriction. The last three planes cover a 2-cm length of the oral cavity from a location lying just anterior to the alveolar ridge to the mid point of the lips. It is in this region that the greatest degree of uncertainty arises due to signal/noise limitations, and the absence of data identifying the location of the incisors and the lateral dimensions of the lips. Thus the image data had to be augmented by data from other sources described in the preceding methods section.

B. Results from experiment 1

1. Area functions

Table I lists the area functions for the four point vowels, /ɑ æ i u/, for both subjects, both including and excluding the sinuses, derived by the methods described above. The data that include the sinus areas are plotted in Fig. 12. Across vowels, the area functions show broadly the expected patterns: narrow pharynx and wide oral cavity for /ɑ/, less constricted pharynx for /æ/, wide pharynx and narrow oral cavity for /i/, and wide oral cavity and pharynx with constrictions in the velar regions and at the lips for /u/. Across subjects, the greatest difference occurs for the /æ/ vowel, for which the TB area function is generally wider and shows more peaks and dips than that for PN. This variability probably reflects the difficulty, which both subjects experienced, with maintaining, for long periods, the vocal tract posture required to produce an /æ/. For both /ɑ/ and /æ/, the calculated lengths differ across subjects. For /i/ and /u/, the area functions for the two subjects have identical lengths and similar shapes. For /u/, the function for TB seems shifted to the left with respect to that for PN, suggesting that these functions may have systematically different pharyngeal starting points. In all cases, the opening at the lips is somewhat larger for PN than it is for TB.

2. Vocal tract lengths

The results of comparing tract lengths calculated from area functions with those measured from the lateral Xerographic x-rays are given in Table II. For five of the eight comparisons, the differences are less than the 0.875-cm resolution of the samples interpolated from the MRI data. Except for one case in which the difference in calculated lengths is negligible, the lengths calculated from the Xerograms are

TABLE II. The vocal tract lengths of the two subjects are given for each of the four vowels as measured from MR and x-ray sources.

Comparison of vocal tract lengths measured from x-ray and MR images				
Subj	Vowel	X-ray (cm)	MRI (cm)	Difference (cm)
TB	/ɑ/	17.6	17.50	0.10
TB	/æ/	16.9	16.62	0.28
TB	/i/	18.1	16.62	1.48
TB	/u/	18.3	18.37	0.07
PN	/ɑ/	16.8	15.75	1.05
PN	/æ/	16.4	15.75	0.65
PN	/i/	17.7	16.62	1.08
PN	/u/	18.7	18.37	0.33

greater. Thus this evidence suggests that our methods of processing MRI data tended to underestimate vocal tract length.

3. Comparison of acoustic analyses

Table III lists the average formant frequencies obtained from the 14-pole LPC analysis of three, 5-s-long natural vowel samples drawn randomly from points close to the start, middle, and end of many image acquisitions. Also present are the formant frequencies computed from the eight area functions shown in Table I (four vowels including the areas of the piriform and valecular sinuses and four without). The formant frequencies of the subjects' utterances differ significantly from the computed resonances of the area functions (with or without sinuses). In the discussion section, we speculate on the possible causes of the formant frequency deviations, and, later in this results section, we show that, despite these differences, the vowel waveforms computed from the area functions in Table I are nevertheless per-

TABLE III. Results of a formant frequency analysis of the subjects' original vowel utterances and vowels synthesized from their area functions. Frequencies are expressed in Hz.

Comparison of natural and computed formants									
			Analyzed			Synthesized			
Subj	Vowel	Sinus	*F1*	*F2*	*F3*	*F1*	*F2*	*F3*	*F4*
TB	/ɑ/	incl.	595	1006	2400	535	1057	2430	3837
TB	/ɑ/	excl.				579	1091	2461	3769
TB	/æ/	incl.	687	1318	2270	624	1416	2259	3737
TB	/æ/	excl.				659	1503	2304	3749
TB	/i/	incl.	246	1917	2608	273	2162	2836	3976
TB	/i/	excl.				278	2263	2947	4066
TB	/u/	incl.	256	777	2146	303	1061	2339	3500
TB	/u/	excl.				312	1068	2462	3592
PN	/ɑ/	incl.	564	910	3006	620	1242	2905	4059
PN	/ɑ/	excl.				763	1383	2935	4127
PN	/æ/	incl.	745	1420	2485	700	1585	3024	3984
PN	/æ/	excl.				769	1727	3083	4441
PN	/i/	incl.	244	2228	2899	284	2270	3674	4761
PN	/i/	excl.				300	2389	3790	4531
PN	/u/	incl.	259	841	2410	311	1442	2426	3545
PN	/u/	excl.				316	1434	2544	3622

TABLE IV. Results of open and closed perceptual tests performed on the synthesized vowels and on the original utterances.

Perceptual data on computed vowels: Closed test

With sinuses
Responses (TB)

	/ɑ/	/æ/	/i/	/u/	
/ɑ/	277	10	0	113	400
/æ/	216	173	3	8	400
/i/	3	1	393	3	400
/u/	3	0	5	392	400
	499	184	401	516	

With sinuses
Responses (PN)

	/ɑ/	/æ/	/i/	/u/	
/ɑ/	368	23	2	7	400
/æ/	67	329	4	0	400
/i/	1	1	396	2	400
/u/	2	2	71	325	400
	438	355	473	334	

Without sinuses
Responses (TB)

	/ɑ/	/æ/	/i/	/u/	
/ɑ/	302	14	2	82	400
/æ/	110	286	1	3	400
/i/	1	1	397	1	400
/u/	3	2	10	385	400
	416	303	410	471	

Without sinuses
Responses (PN)

	/ɑ/	/æ/	/i/	/u/	
/ɑ/	301	97	1	1	400
/æ/	23	373	4	0	400
/i/	0	0	397	3	400
/u/	2	2	121	275	400
	326	472	523	279	

Perceptual data on computed vowels: Open test

With sinuses
Responses (TB)

	/ɑ/	/æ/	/i/	/u/	/ʊ/	/ʌ/	/ɔ/	/ɛ/	
/ɑ/	0	0	0	0	0	42	38	0	80
/æ/	7	15	0	0	0	57	1	0	80
/i/	0	0	74	0	5	0	0	1	80
/u/	0	0	0	46	32	0	0	2	80
	7	15	74	46	37	99	39	3	

With sinuses
Responses (TB)

	/ɑ/	/æ/	/i/	/u/	/ʊ/	/ʌ/	/ɔ/	/ɛ/	
/ɑ/	3	0	0	0	0	46	31	0	80
/æ/	9	54	0	0	0	16	1	0	80
/i/	0	0	78	0	0	0	0	2	80
/u/	0	0	0	34	45	0	0	1	80
	12	54	78	34	45	62	32	3	

With sinuses
Responses (PN)

	/ɑ/	/æ/	/i/	/u/	/ʊ/	/ʌ/	/ɔ/	/ɛ/	
/ɑ/	9	0	0	0	0	66	5	0	80
/æ/	0	72	0	0	0	8	0	0	80
/i/	1	0	76	0	0	0	0	3	80
/u/	0	0	1	0	79	0	0	0	80
	10	72	77	0	79	74	5	3	

Without sinuses
Responses (PN)

	/ɑ/	/æ/	/i/	/u/	/ʊ/	/ʌ/	/ɔ/	/ɛ/	
/ɑ/	75	5	0	0	0	0	0	0	80
/æ/	0	80	0	0	0	0	0	0	80
/i/	0	0	62	0	0	0	0	18	80
/u/	0	0	1	0	79	0	0	0	80
	75	85	63	0	79	0	0	18	

Perceptual data on natural vowels: Closed test

Responses (TB)

	/ɑ/	/æ/	/i/	/u/	
/ɑ/	278	113	0	9	400
/æ/	97	302	1	0	400
/i/	0	1	397	2	400
/u/	1	3	3	393	400
	376	419	401	404	

Responses (PN)

	/ɑ/	/æ/	/i/	/u/	
/ɑ/	267	94	0	39	400
/æ/	120	277	1	2	400
/i/	0	0	399	1	400
/u/	1	1	3	395	400
	388	372	403	437	

ceived as the original utterances by a majority of both skilled and unskilled listeners.

4. Comparison of perceptual analyses

Results of the perceptual analyses are given in Table IV. The results for the first *closed* response test, using vowel sounds synthesized from the area functions, are shown in Table IV(top). Two of the vowels based on data from subject TB were frequently misclassified. A substantial number of the vowels computed from the /æ/ area functions, both including and excluding the sinuses, were misclassified as /ɑ/. There was also a substantial, although smaller, number of cases of /ɑ/ being confused with /u/ among vowels based on data from subject TB. The principal perceptual confusions among the PN stimuli arose when the vowel /u/ was presented. On about 30% of all occurrences, /u/ was classified as /i/. There were also confusions among the /ɑ/ and /æ/ stimuli. However, the overall consequences of these confusions lacked any identifiable pattern. This is illustrated by a comparison of the with- and without-sinus conditions. The results of an information analysis (McGill, 1954), based on the response data contained in Table IV(top), reveal that, of the two bits of information presented, the information transmitted is higher in the with-sinuses condition for subject PN (1.47 bits vs 1.38 bits) and higher in the without-sinuses condition for subject TB (1.42 bits vs 1.28 bits). Only in the case of subject TB does an analysis of variance show the difference between the two conditions to be significant ($F = 42.05, p < 0.0005$). Thus the overall conclusion drawn from the results of the closed response test is that the classification errors show no common pattern across subjects or systematic relation to whether the areas of the sinuses were included or excluded from the synthesis calculations.

Results presented in Table IV(middle) from an additional *open* response test reveal further details of the perceptual shortcomings of the synthesized vowels. Four phonetically trained listeners were instructed to classify the stimuli as American–English vowels. Their responses reveal that most of the /ɑ/ vowels synthesized from TB's vocal tract data (with or without sinuses) were perceived as /ʌ/ or /ɔ/ and that the same confusion occurred with PN's /ɑ/ vowel, but in the with-sinuses condition only. A substantial number of /æ/ vowels synthesized from the TB data were also perceived as /ʌ/ with the greater proportion of that number appearing in the with-sinuses condition. Between 77% and 97% of the /i/ vowels were classified correctly, while the last vowel /u/ was perceived as either /u/ or /ʊ/ with roughly equal frequency in the case of subject TB but, in subject PN's case, it was consistently perceived as /ʊ/.

Lastly, the results of the *closed* response test that employed the subjects' original utterances appear in Table IV (bottom). These data show that the largest proportion of perceptual confusions occurred between the vowels /ɑ/ and /æ/. Consequently, it is clear that the root of a substantial number of the /ɑ/–/æ/ confusions made among the synthesized vowels must lie in the articulatory and acoustic characteristics of the subjects' original vowel productions. However, despite the fact that in terms of the total number of stimuli correctly identified, the overall accuracy achieved with the natural speech stimuli is slightly better than that achieved with the synthesized stimuli, analyses of variance of the natural speech and each set of synthetic speech responses for both subjects show the vowel–condition interaction effects, in all but one instance, to be highly significant ($F(3,57) > 6.28; p < 0.001$) while the overall differences in numerical scores are, in the best case, statistically not significant ($F(1,19) = 2.56; p = 0.13$).

C. Results from experiment 2

1. Relation between midsagittal and lateral pharyngeal width

Measures of both midsagittal and lateral width were made, as described above, from the tracings of the pharyngeal cross sections for all nine vowels at 12 (PN) and 14 (TB) different heights above the glottis. Figure 13 shows, for both subjects, the results of plotting the lateral against the midsagittal measures. The different heights spaced at intervals of 0.5 cm are identified by different symbols. Larger-sized symbols are used to indicate the four uppermost levels.

An identical treatment by Sundberg *et al.* (1987) of a much smaller body of data led these authors to seek evidence

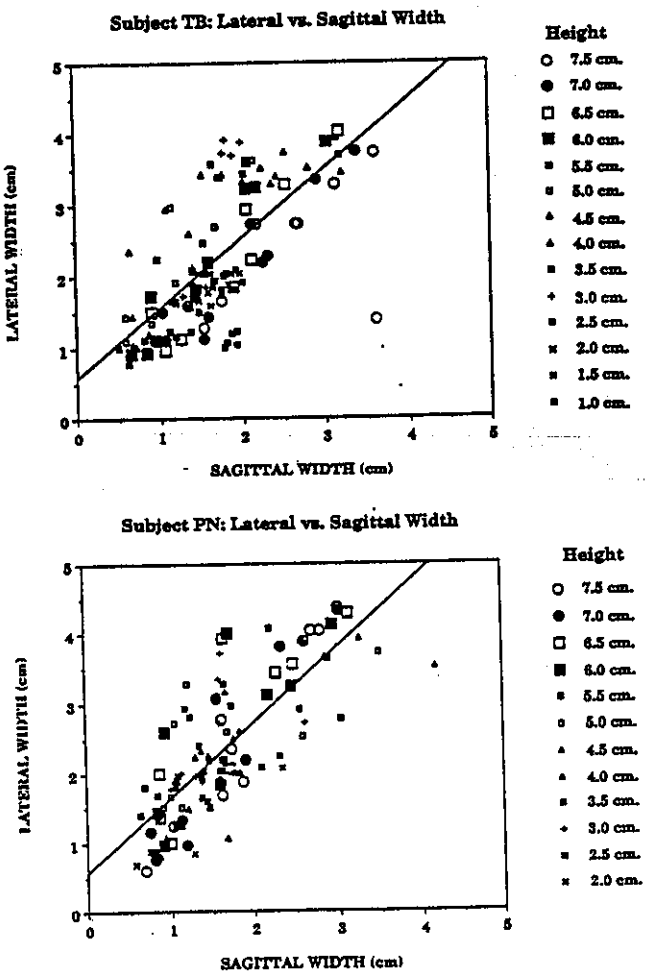


FIG. 13. Graphs of lateral versus midsagittal width of the pharynx. Each plotting symbol represents a measurement made at a different height above an origin located close to the glottis. See text for further details.

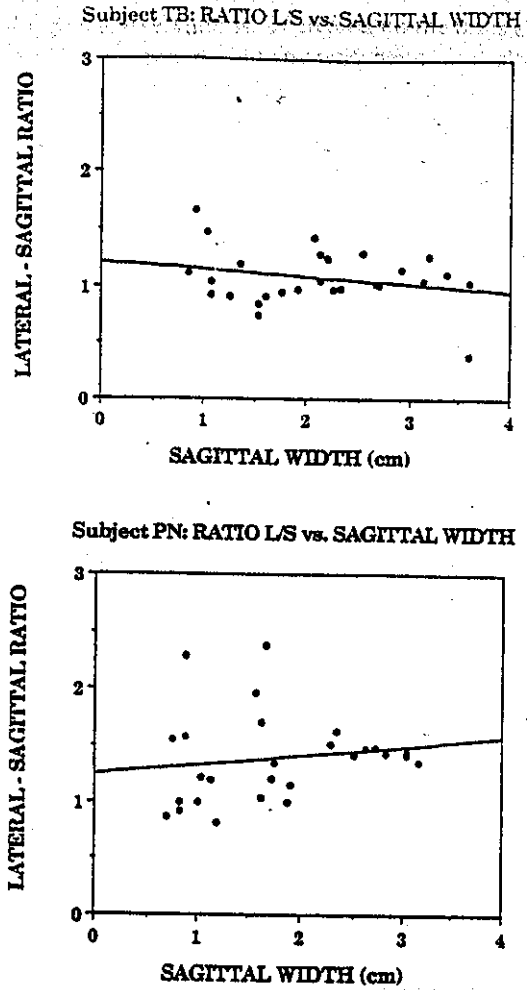


FIG. 14. Graphs of the ratio of lateral to midsagittal width versus midsagittal width plotted with data obtained from the pharynx at points estimated to be 6.0, 6.5, 7.0, and 7.5 cm above the glottis.

of a linear relationship, which they found in the case of a female subject. Results from a male subject were more complex and equivocal, however. Linear regression lines fitted to the data of our male subjects have slopes of 0.95 and 1.06, in close agreement with Sundberg, and intercepts on the *lateral* axis of 0.57 and 0.56 cm, roughly half the value that emerged from Sundberg's study. Correlation coefficients (R) of 0.69 and 0.78 ($n = 126$ and $n = 108$) yield, in both cases, levels of significance that substantially exceed $p = 0.001$. Thus, in contrast to Sundberg *et al.*, we conclude that the relation between midsagittal and lateral width in the pharynx can be plausibly represented by a straight line.

Furthermore, the tendency observed by Fant (1960), Sundberg (1969), and Sundberg *et al.* (1987) for the lateral dimension of the upper pharynx to expand at a slower rate when large midsagittal extensions occur is not fully supported by data from the four uppermost levels shown in Fig. 14. From plots of the ratio of lateral and midsagittal width versus midsagittal width, we would expect to see a downward trend in the ratio as the midsagittal width increases. In fact, a regression line fitted to the data of subject PN has a positive slope that suggests contrary behavior. A similar regression line applied to the ratio data of subject TB, on the other hand, has a negative slope and thus supports the earlier observations. Therefore, on the basis of this evidence, it would appear that any tendency for the slope of a plot such as that shown in Fig. 13 to decrease at extreme midsagittal widths may not represent a universal feature but a behavior of only some individuals.

2. Midsagittal width versus area: The pharyngeal data

We began our analysis of the width versus area data by examining the applicability of the square law hypothesis ($A = K \cdot S^2$) proposed by Sundberg *et al.* (1987) to the data obtained at all heights. Therefore, scatter plots were made of the airway cross-sectional areas versus their squared midsagittal widths. Figure 15 presents these plots, which appear in pairs, representing data sets that exclude (a) and (b) and include (c) and (d), the areas of the sinuses. The data are derived from all nine vowels at each of 12 levels spaced at 0.5-cm intervals above the larynx for subject PN and 14 levels for subject TB. Regression lines having correlation coefficients (R) of 0.88, 0.94, 0.81, and 0.90 fit these data with a statistically high degree of significance ($p < 0.001$, $n > 100$) in all cases. A group of four outlying points (circled) in plot TB(b) have been omitted from that regression calculation with the result that the K and intercept parameters are brought into close agreement with the other data. These regression parameters are summarized in Table V (first five columns) and compared with those of the male subject studied by Sundberg *et al.* This comparison shows that our values for both K and the intercept coefficient are smaller than those calculated from the data of Sundberg *et al.*

In all four plots in Fig. 15, the variance is not constant and appears to vary with the midsagittal width. This sug-

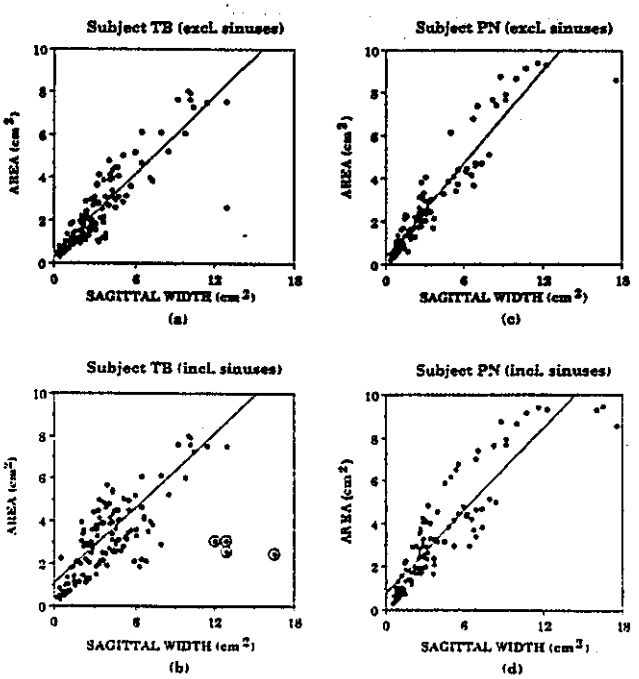


FIG. 15. Graphs showing the relationship between pharyngeal area and the square of midsagittal width. The plots for subjects TB (above) and PN (below) appear in pairs, one member of which excludes and the other includes the areas of the sinuses. Circled points were omitted from the regression calculation.

TABLE V. Coefficients K, exponents r, intercepts, and correlation coefficients of regression lines computed from the entire body of pharyngeal area versus midsagittal width data. Results contrast the square law and power law hypotheses.

		Coefficients based on pooled data from entire pharynx					
		Square law			Power law		
Subject	Sinuses	Coeff K	Intercept	Coeff R	Coeff K	Expo r	Coeff R
TB	incl.	0.57	0.92	0.81	1.09	1.50	0.87
TB	excl.	0.61	0.45	0.88	0.89	1.67	0.91
PN	incl.	0.64	0.75	0.90	0.90	1.89	0.93
PN	excl.	0.72	0.33	0.94	0.84	1.95	0.95
Sundberg	incl.	0.74	1.26	0.91	1.50	1.62	0.95

gests that the assumption of independent variance on which regression analysis is based is not strictly met by the data in their present form. Moreover, acceptance of the square law hypothesis at this stage begs the question of whether some other, perhaps closely related, power law of the type $A = K \cdot S^r$ would achieve a better description of the data. Both of these issues are addressed by logarithmically transforming the area (A) and midsagittal width (S) variables. Regression lines computed from the transformed data, shown plotted in Fig. 16, yield higher correlation coefficients (R) and more nearly equal distributions of the residuals about the lines—thus both graphs indicate closer conformity with the regression assumptions. The gradients of the regression lines now provide estimates of the exponent r, while the intercepts provide corresponding values of K that can be directly compared with the K coefficients derived from the square law analysis. The results of this power law analysis are summarized in the last three columns of Table V. They show that the average value of the power law exponent is approximately 1.75, ranging from about 1.5–2.0. Meanwhile, based on the power law analysis, the average value of K is 0.93 as compared with an average of 0.64 based on the square law.

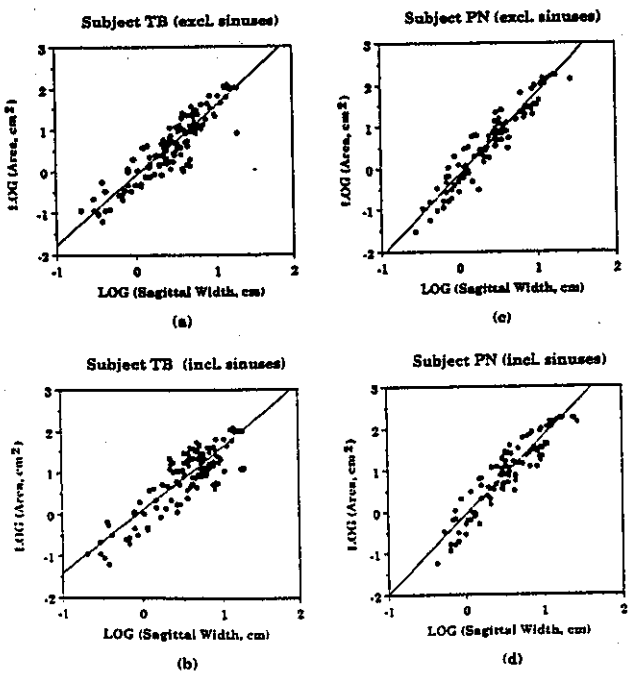


FIG. 16. Graphs of the log transformed pharyngeal area versus midsagittal width data from both subjects, TB (upper) and PN (lower). Data from all measurement locations within the pharynx have been plotted. Left-hand plots (a) and (b) contain data that exclude the areas of the sinuses, whereas right-hand plots (c) and (d) include the sinus areas. See Table V for the parameters of the regression lines.

3. Midsagittal width versus area: The relation to height above the larynx

To this point we have treated the pharynx as if its mode of expansion and contraction were uniform throughout its length. We now examine the behavior of the pharynx at different heights above the glottis, using the area data that exclude the sinuses. Figures 17–19 show plots of the logarithmically transformed area and midsagittal width data at different heights. In many of these plots, it can be seen that the linear regression line provides a remarkably good fit to the data. Table VI lists the correlation and other coefficients of linear regression lines computed on the basis of both the square law and power law at different heights above the larynx. These coefficients have been plotted in Fig. 20, whose left-hand panel compares the values of coeff K derived from the two hypotheses. The right-hand panel, meanwhile, shows values of the exponent r derived from the power law analysis. Several differences in behavior are apparent between the two subjects. In the case of subject TB, the values of coeff K and the exponent tend to be higher in the region between 3.0 and 4.0 cm above the larynx (roughly in the neighborhood of the epiglottis) than elsewhere, and higher than is apparent in the case of subject PN in the same region. The results of averaging the two subjects' K coefficients and exponent values leads to the third pair of graphs in Fig. 20, which, for approximate modeling purposes, might be represented by the parabolic functions appearing there.

Basing their analysis on the square law hypothesis, Sundberg *et al.* (1987) compared the K coefficients, intercept constants and correlation coefficients (R) which they obtained with those of three earlier studies. A comparison of the present results with those assembled by Sundberg *et al.* is shown in Table VII.

4. Midsagittal width versus area: The upper vocal tract

The only complete sets of images of the upper vocal tract were obtained by the EMR machine during the course of experiment 1. These coronal images were obtained only for

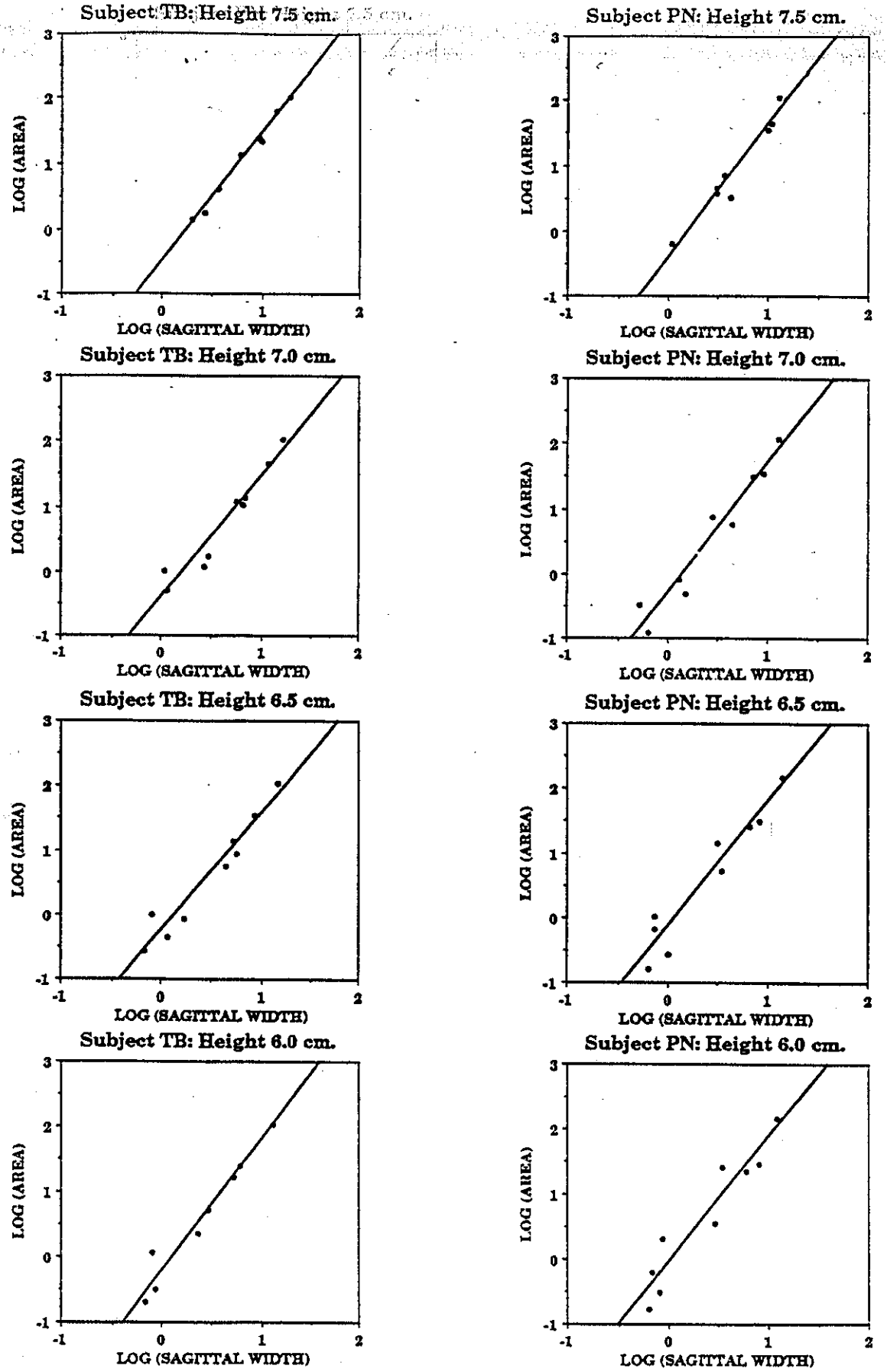


FIG. 17. Graphs of log-transformed pharyngeal area versus midsagittal width data obtained from both subjects, TB (left) and PN (right), at 0.5-cm intervals commencing (top) at the approximate height of 7.5 cm above the glottis. The exponents and constants of the linear regression lines shown here are available in Table VI.

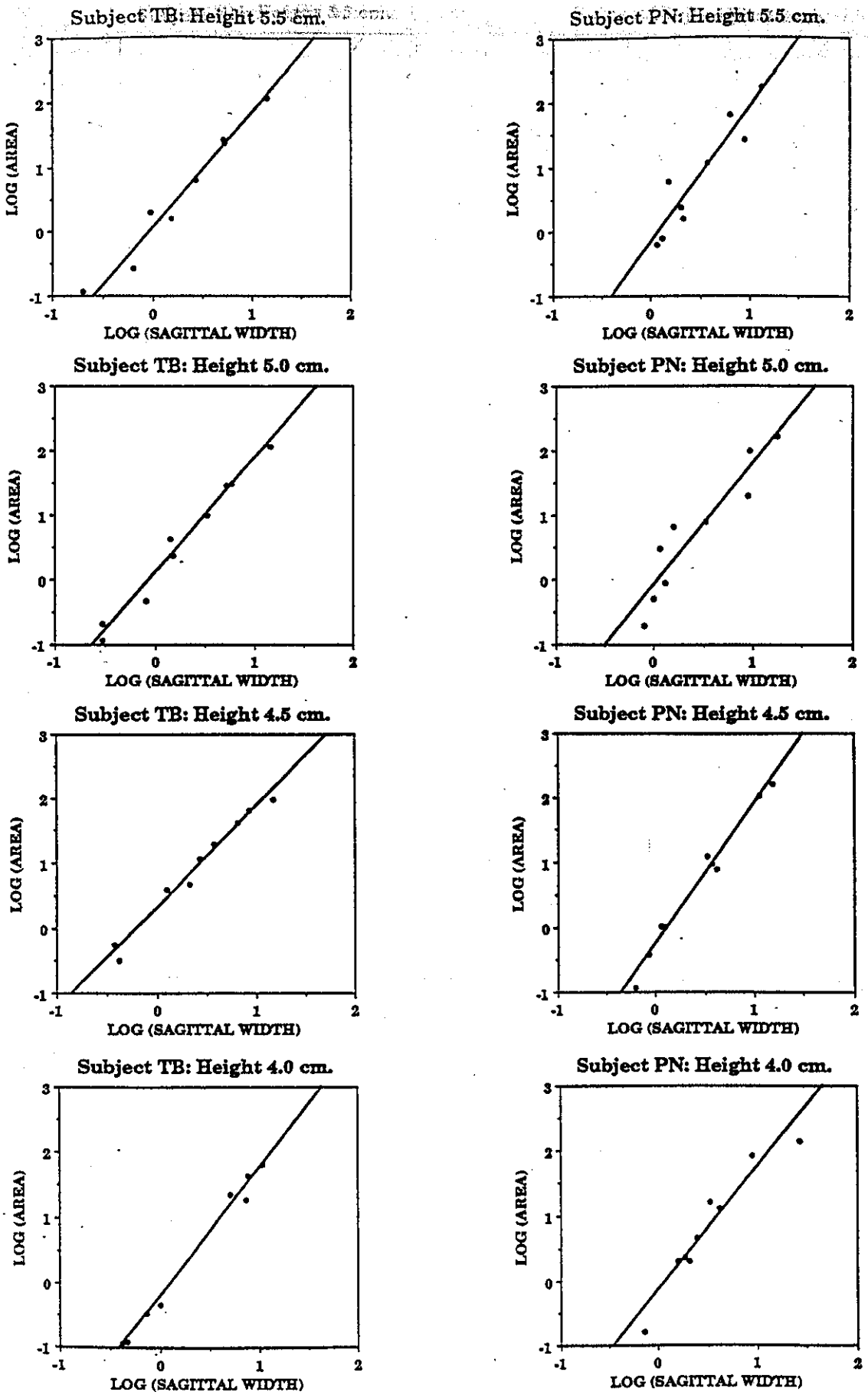


FIG. 18. Graphs of log-transformed pharyngeal area versus midsagittal width data continue from a point 5.5 cm above the glottis. See Table VI for regression parameters.

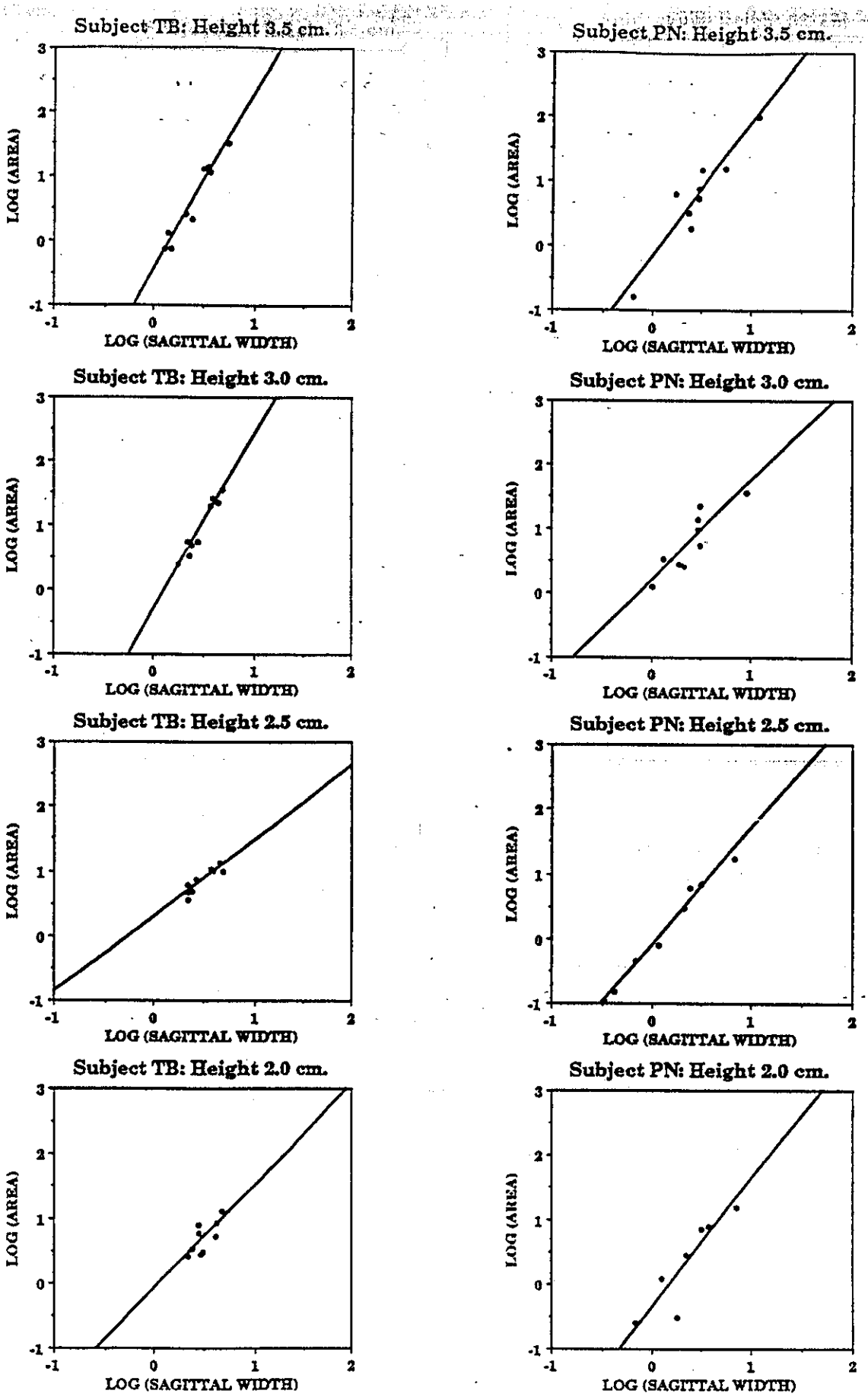


FIG. 19. Graphs of log-transformed pharyngeal area versus midsagittal width data continue from a point 3.5 cm. above the glottis. Regression parameters are available in Table VI.

TABLE VI. Coefficients K, exponents r, intercepts, and correlation coefficients for regression lines computed at each individual level above the larynx. Results contrast the square law and power law hypotheses. Data exclude the piriform and valecular sinuses. All calculations are based on $n = 9$ samples except where otherwise indicated.

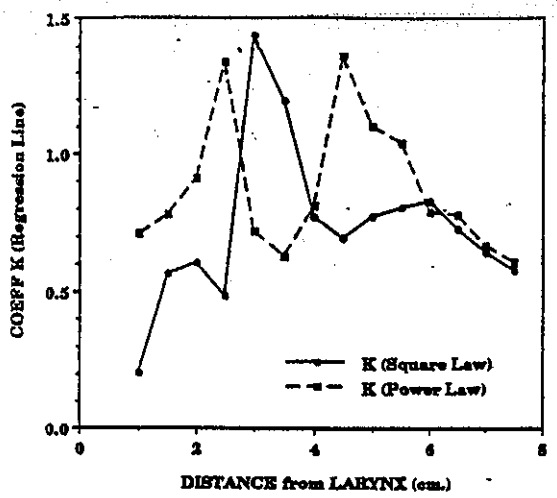
		Coefficients computed at different heights above larynx					
		Square law			Power law		
Height (cm)	Subj	Coeff K	Intercept	Coeff R	Coeff K	Expo r	Coeff R
7.5	TB	0.48	0.01	0.99	0.61	1.96	0.99($n = 8$)
7.5	PN	0.74	−0.20	0.97	0.66	2.02	0.98
7.0	TB	0.65	−0.16	0.99	0.67	1.86	0.97
7.0	PN	0.81	−0.14	0.98	0.74	1.99	0.97
6.5	TB	0.73	−0.11	0.99	0.78	1.82	0.97
6.5	PN	0.82	0.06	0.98	0.88	1.91	0.96
6.0	TB	0.83	−0.03	1.00	0.79	2.03	0.98
6.0	PN	0.89	0.08	0.96	0.95	1.91	0.95
5.5	TB	0.80	0.28	0.99	1.04	1.80	0.98
5.5	PN	0.95	−0.08	0.94	0.83	2.10	0.94
5.0	TB	0.78	0.44	0.99	1.10	1.77	0.99
5.0	PN	0.74	0.35	0.95	0.91	1.88	0.94
4.5	TB	0.70	0.91	0.97	1.36	1.56	0.99
4.5	PN	0.88	0.07	0.99	0.76	2.16	0.99
4.0	TB	0.77	0.05	0.98	0.81	1.95	0.99
4.0	PN	0.48	1.08	0.92	0.86	1.89	0.96
3.5	TB	1.20	−0.70	0.98	0.63	2.73	0.97
3.5	PN	0.85	0.12	0.97	0.85	2.06	0.94
3.0	TB	1.44	−1.02	0.97	0.72	2.72	0.97
3.0	PN	0.59	0.94	0.85	1.20	1.54	0.88
2.5	TB	0.48	1.06	0.89	1.34	1.17	0.90
2.5	PN	0.65	0.33	0.96	0.91	1.79	0.99
2.0	TB	0.61	0.38	0.76	0.91	1.58	0.74
2.0	PN	0.63	0.13	0.95	0.69	1.98	0.96
1.5	TB	0.56	0.23	0.82	0.78	1.59	0.83
1.5	PN	(no data available)					
1.0	TB	0.20	0.66	0.35	0.71	1.02	0.44
1.0	PN	(no data available)					

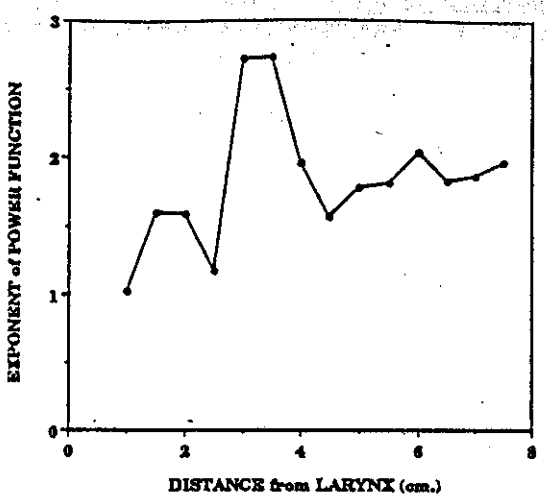
the four point vowels and, consequently, the data that they supplied on the oral cavity were less plentiful than those covering the pharynx. For the purposes of measurement in the coronal plane, the midsagittal width was redefined as the distance between the most superior point on the surface of the palate and the most inferior point on the surface of the tongue.[2] Using the alveolar ridge as a point of alignment, corresponding coronal images from the two subjects were obtained and the dimensional data from those images were combined and plotted in the panels of Figs. 21 and 22.

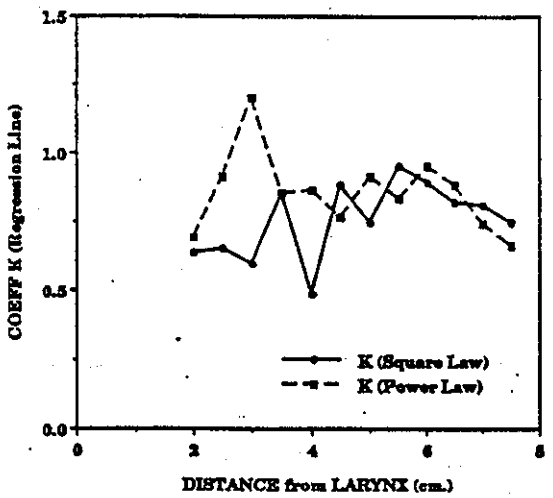
Heretofore, measurements of upper vocal tract dimensions during speech sound production have been possible only with calipers or similar instruments. Boe *et al.* (1988), using axial x-ray tomography, attempted to negotiate the bend in the tract and approach the palatal region by inclining the subject's head. Other studies (Ladefoged *et al.*, 1971; Sundberg *et al.*, 1987) employed casts of the vocal tract that were formed while the subject maintained a vowel-like configuration. Because our image planes are parallel and differ from the planes of section studied by earlier investigators, we cannot directly compare our measurements with those published previously. Sundberg *et al.* (1987) adopted the Heinz and Stevens (1964) polar coordinate system when they measured their casts. They determined that, for male subjects, their data for the mouth cavity as a whole obeyed a relation of the type $A = K \cdot S^r$, where A and S represent the cross-sectional area and midsagittal width respectively, K is a constant between 2.07 and 2.63 and r, the midsagittal exponent, lies between 1.33 and 1.47.

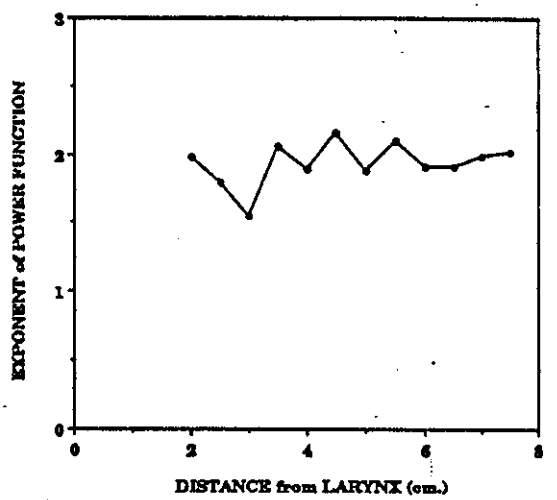
Pooled data from both subjects incorporating all the parallel planes between a point starting just behind the upper incisors and a point just short of where the fronted tongue begins to make a steep descent into the pharynx (a distance of about 6 cm) yield $r = 1.40$ and $K = 1.27$ at a level of significance of $p < 0.001$ ($R = 0.85$, $n > 100$). However, to determine whether the use of pooled data might mask systematic variation in the exponent r and coefficient K as a function of cavity position (as previously noted by Sundberg *et al.*), these parameters were also computed for each plane individually and the results assembled in Table VIII. The series of r exponents from Table VIII, seen plotted in the upper half of Fig. 23, appears to show evidence of a modest upward trend toward the posterior direction, as the image plane number increases. However, a regression line fails to fit the exponent data at a sufficiently high level of significance to lay claim to such a systematic trend. Therefore, we must conclude that for a substantial portion of the upper vocal tract, our data indicate that r is approximately constant with a mean value of 1.97 and standard deviation of 0.41.

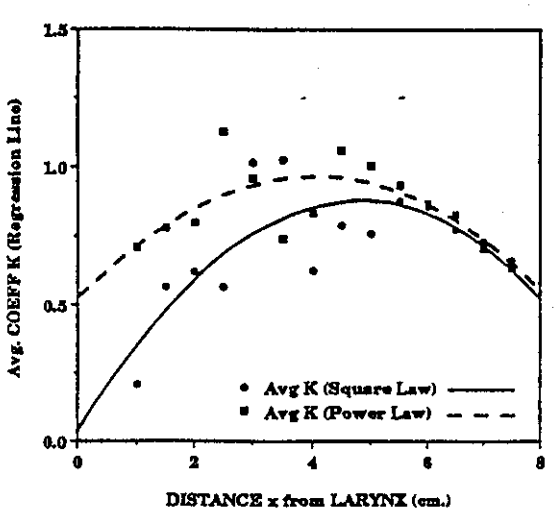
A linear regression line computed on the series of K coefficients, on the other hand, does fit the data points at a











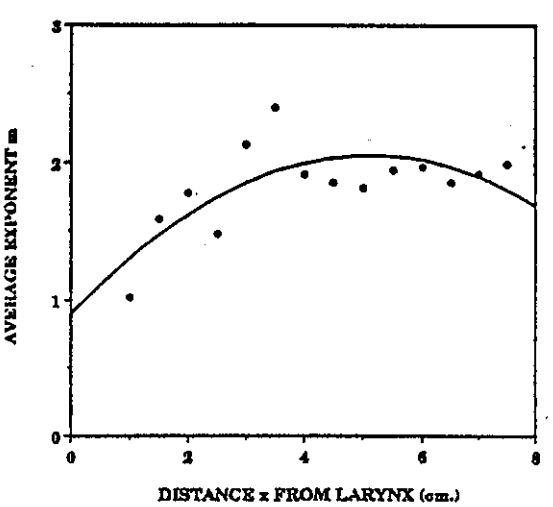


FIG. 20. Left panel: The K coefficients obtained from both square law and power law approximations to the relationship between the cross-sectional area and midsagittal width of the pharynx are plotted here as a function of height above the glottis. Right panel: The exponent of the power law approximation is also plotted as a function of height above the glottis. The two lowest graphs show averages of the two subjects' data and the results of computing parabolic approximations of the data by regression.

TABLE VII. Results from the present study are compared with data (male subjects only) obtained from Table III of Sundberg *et al.* (1987) containing results based on a square law approximation. Because the grid systems used in earlier studies differ from the one used here, corresponding pharyngeal heights between earlier and present data are necessarily approximate. The asterisks indicate that there are no data available.

Study		Comparison with earlier measurements Height (cm)	Coeff K	Intercept (cm²)	Coeff R
Present study	(TB)	7.5	0.48	0.01	0.99
	(PN)	7.5	0.74	−0.20	0.97
Sundberg *et al.* (1987)		7.6	0.61	1.90	0.99
Ladefoged *et al.* (1971)		*	1.75	2.00	*
Heinz & Stevens (1964)		*	2.30	0.00	*
Present study	(TB)	5.5	0.80	0.28	0.99
	(PN)	5.5	0.95	−0.08	0.94
Sundberg *et al.* (1987)		5.7	0.70	0.60	0.99
Ladefoged *et al.* (1971)		*	1.58	4.50	*
Heinz & Stevens (1964)		*	2.30	0.00	*
Present study	(TB)	4.0	0.77	0.05	0.98
	(PN)	4.0	0.48	1.08	0.92
Sundberg *et al.* (1987)		3.8	0.99	0.30	1.00
Ladefoged *et al.* (1971)		*	1.58	4.50	*
Heinz & Stevens (1964)		*	2.00	0.00	*
Present study	(TB)	2.0	0.61	0.38	0.76
	(PN)	2.0	0.63	0.13	0.95
Sundberg *et al.* (1987)		1.9	0.97	1.60	0.91
Ladefoged *et al.* (1971)		*	1.30	4.00	*
Heinz & Stevens (1964)		*	1.60	0.00	*

level of significance of $p < 0.001$ ($R = 0.85$, $n = 13$), sufficiently high to suggest that some evidence exists for systematic variation in K as a function of oral location. The constant K, plotted in the lower half of Fig. 23, has a value of approximately 2.0 at the plane closest to the lips and descends throughout the region of the hard palate reaching about 0.6 at about the 13th plane. This is roughly the point of entry into the pharynx, where the 0.5-cm sections begin to obliquely intersect the anterior pharyngeal wall resulting in broad air–tissue image density gradients and greater uncertainty as to the precise position of the boundary and the area it inscribes. It is also the point where the coronal sections increasingly fail to make an orthogonal intersection with the tract centerline. Consequently, the measurements in this region have limited value for modeling purposes.

III. DISCUSSION

A. Experiment 1

1. Vocal tract lengths

Comparisons of vocal tract length calculated from MRI and x-ray data in Table II generally showed that the MRI derived lengths were shorter. The three vowels in Table II for which the difference in length was greater than the section length (0.875 cm) included /i/ for both subjects and /ɑ/ for subject PN. It is, of course, possible that this disagreement arises from genuine differences in the vowels produced on the two occasions. However, the fact that the /i/ vowels from both subjects showed the largest differences implies systematic error. One possible source of error is evident in Fig. 10, which shows the grid system superimposed on an /i/ vowel configuration. In regions where the grid lines are parallel and intersect the tract centerline orthogonally, each grid line samples a 0.5-cm increment in tract length. However, at the base of the pharynx, and particularly in the front half of the oral cavity, the actual centerline of the vocal tract diverges as much as 30 deg from a perpendicular intersection with the grid lines. Thus, specifically in these regions, the tract length would be underestimated by about 13% using our procedures. Moreover, the cumulative effect of this error can be expected to be greater for an /i/, because the high-front tongue position for that vowel would tend to increase the lack of perpendicularity between the vocal tract centerline and the grid lines in the oral cavity. Such underestimates of tract length would be expected to result in the tendency for synthesized first formant frequencies to be higher than the corresponding analyzed formants. The fact that in Table III, averaged across both subjects, the vowel /i/ (accompanied by /u/) exhibits larger first-formant elevations than the other two vowels, indicates that errors in tract length may be implicated in the lack of agreement between the analyzed and synthesized formants.

2. Role of the piriform sinuses

When analyzing the area data and comparing synthesized vowels with the original utterances, we did not compute the effects of the sinuses (piriform or valecular) as cavity shunts branching off the main airway. Instead we chose to (a) "include" them as if they were a part of the airway or (b) "exclude" them as if they had no acoustic consequences whatsoever. This seemingly cavalier approach was adopted for several reasons. First, as noted earlier, the image quality decreased sharply approaching the larynx due to its close proximity to one end of the receiver coil and caused many gaps in the piriform data. Thus the volumetric data on the sinuses were not complete. Furthermore, even if the dimensions of the cavities could have been measured accurately, the fact remains that the lack of available knowledge about

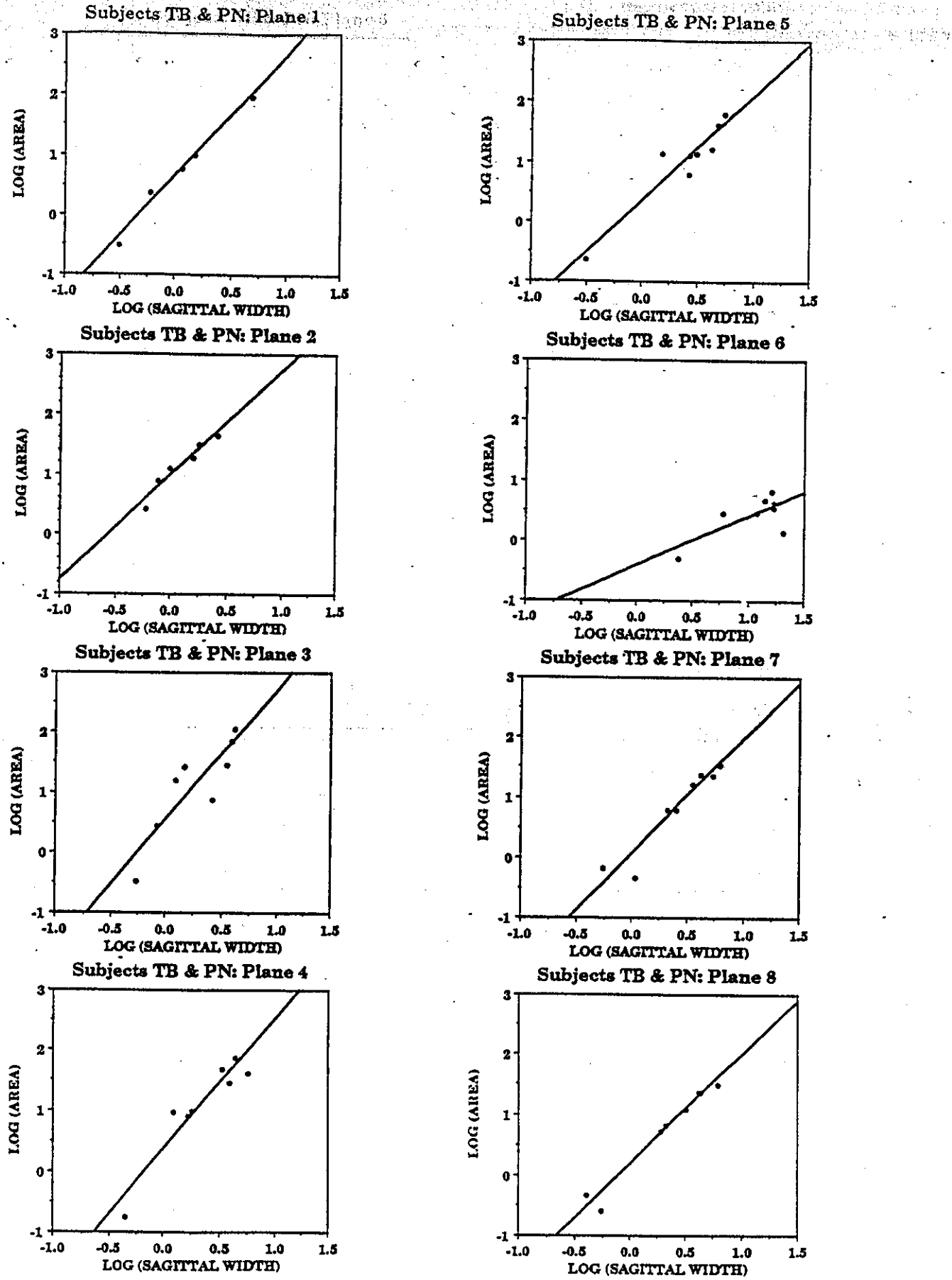


FIG. 21. Graphs of the log-transformed cross-sectional area versus midsagittal width data from the upper vocal tracts of both subjects TB and PN. The coronal planes of intersection with the tract are spaced at intervals of 0.5 cm and numbered 1–8. Plane 1 is located at a point just behind the upper incisors. Linear regression lines are included.

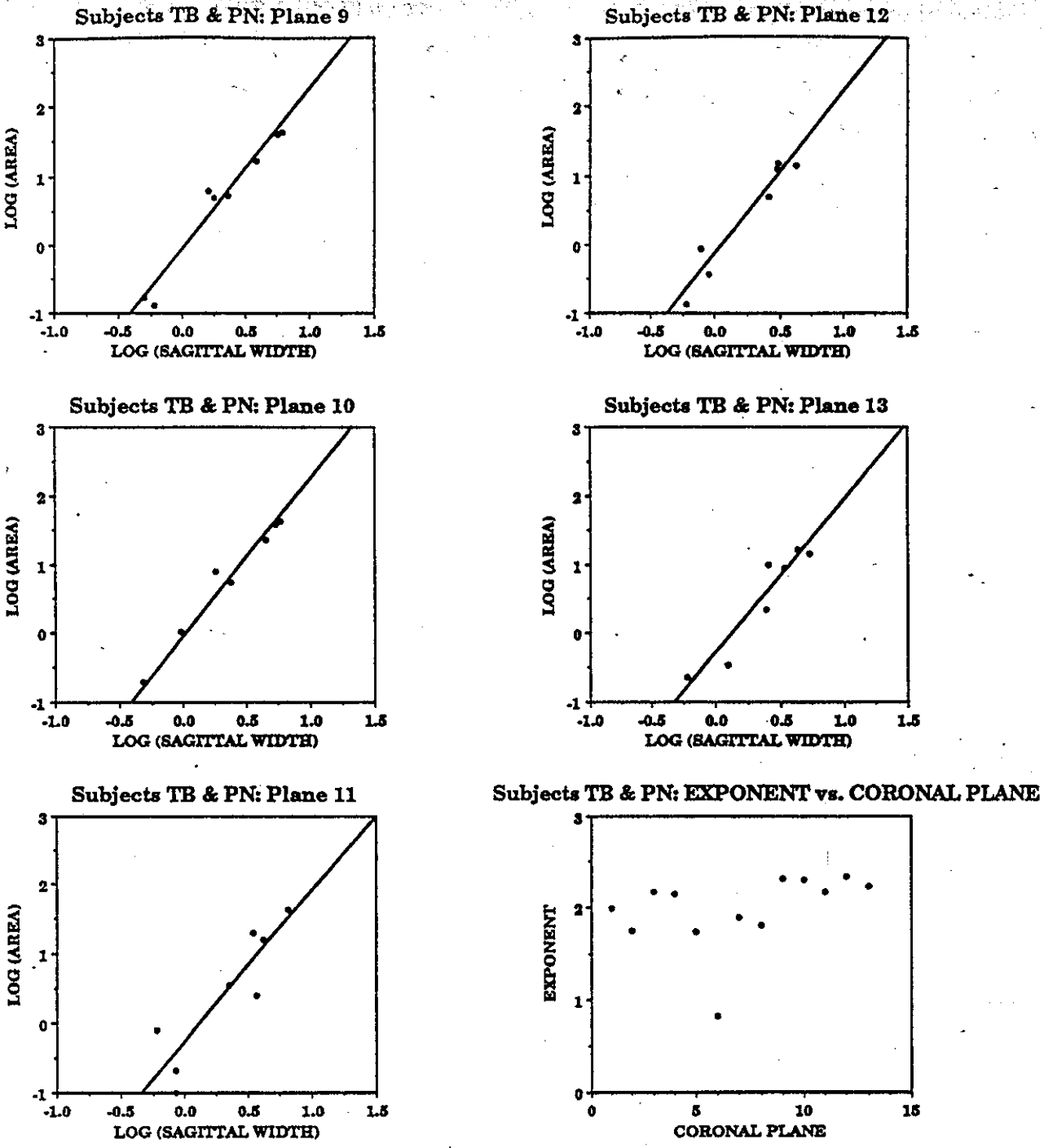


FIG. 22. A continuation of the series of graphs begun in Fig. 21 showing log area versus log midsagittal width for coronal planes 9–13. The sixth graph (bottom right) shows the value of the exponent of the power law approximation plotted as a function of the coronal plane number.

the mechanical properties of the piriform membranes and the degree of mechanoacoustic coupling that exists between the larynx tube and the adjacent sinuses makes it difficult to examine theoretically their acoustic effects with any conviction. It is, perhaps, for this reason that the sinuses have been ignored in most dynamic models of articulation, including the Mermelstein (1973) model used here. Fant (1960), however, did include shunting cavities of fixed size to represent the piriform sinuses in his model and reported that their effect is to insert a zero in the spectrum above 5 kHz and appreciably sharpen the spectral cutoff at that frequency. Lin (1990) further showed in a modeling study that there is a tendency for piriform sinuses of fixed volume and coupling area to reduce formant frequencies to an extent that depends on the place of articulation. The effects are small, however, except in the case of back vowels whose constriction location causes the $F1$ and $F2$ frequency reductions to become more prominent. But, as Fant (1980) acknowledges, and in many instances our data appear to confirm, the piriform cavities are not of fixed size. They can alter their volume during

TABLE VIII. Coefficients K, exponents r, and correlation coefficients of regression lines computed at each of 13 coronal planes in the upper vocal tract. Data from subjects TB and PN have been combined.

Coefficients based on pooled data from upper vocal tract			
Plane	Expo r	Coeff K	Coeff R
1	1.99	1.89	0.99
2	1.74	2.62	0.96
3	2.16	1.68	0.87
4	2.14	1.41	0.94
5	1.74	1.40	0.94
6	0.82	0.71	0.71
7	1.89	1.05	0.95
8	1.80	1.18	0.98
9	2.32	0.92	0.98
10	2.30	0.92	0.98
11	2.17	0.75	0.89
12	2.34	0.85	0.97
13	2.23	0.73	0.94

articulation and often become smaller during the production of back vowels. At present, owing to a lack of anatomical data that would establish the relations between sinus volume, larynx height and the location of tongue constriction,

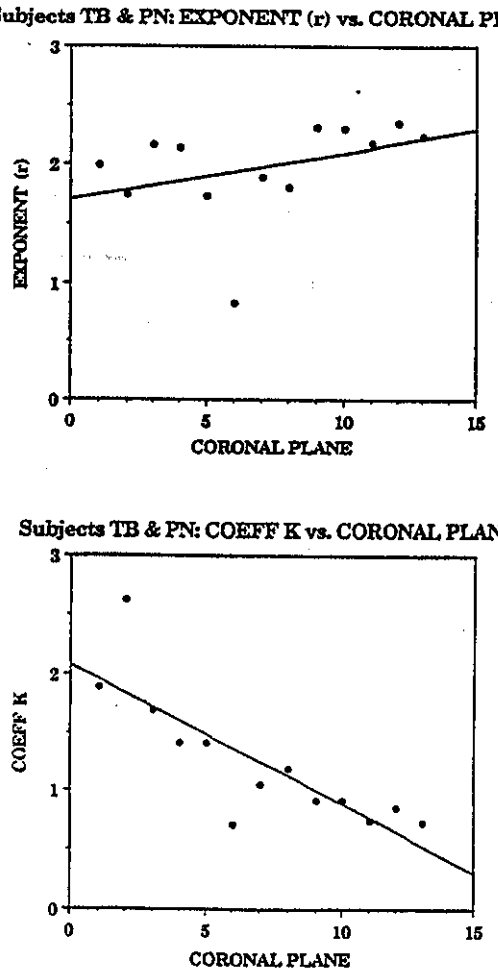


FIG. 23. Plots of the values of the exponent r (upper graph) and coefficient K (lower graph) for each plane in a sequence of 13 coronal planes intersecting the upper vocal tract. The calculations of r and K are based on pooled cross-sectional area data obtained from subjects TB and PN at each roughly corresponding plane.

questions about the role that the piriform sinuses play in the acoustics of normal speech have received little theoretical attention. At least one attempt has been made, however, to explore the issues experimentally by filling the piriform sinuses with cotton to eliminate their resonance entirely (Flach and Schwickardie, 1966). But, as Mermelstein (1967) has pointed out, the assumption that such resonances can be eliminated quite so easily is flawed. Nevertheless, despite the difficulties, the prospects of collecting much of the needed data on sinus dimensions are good since, if coils designed to custom fit the subject's neck are constructed, there is no intrinsic reason why, in future studies, images of superior quality cannot be obtained in the supralaryngeal area.

3. Replicating vowel acoustics by synthesis from tract dimensions

The differences between the formant frequencies of the subjects' original utterances (measured by LPC analysis) and formants computed from the synthesized waveforms were large in several instances. That differences appeared at all should not have been surprising, however, because many simplifying assumptions were made. Among these assumptions is the notion that the vocal tract can be approximated by two straight sections of pipe connected together by a third section bent into a 90-deg arc. That assumption justified use of the simple grid plane system shown in Fig. 10 as a way to resection the 3-D digital model of the airway in planes that were orthogonal to its central axis. However, as we have already pointed out with respect to Fig. 10, there are regions of the vocal tract that fail to fit the pipe model. In these regions the estimated cross-sectional areas will be increased by a factor of sec θ, where θ is the angular deviation of a plane of intersection from true orthogonality with the airway axis. Consequently, the same deviations from orthogonality that were earlier identified as bearing some of the responsibility for underestimates in vocal tract length probably also contributed overestimates of some of the cross sections contained in the area functions.

Another assumption, inherent in the use of LPC analysis, is that the speech signals contained no spectral zeros. However, if nasalization was present in the original vowel productions, as indeed may have been the case based on the velopharyngeal evidence of MR images from subject PN in particular, the analysis procedure, which assumed no nasalization, would have given only approximate resonance values. However, the greater precision required to inquire into this matter was precluded by the less than ideal recording conditions. For example, in addition to the magnetostrictive noise, there was evidence that the natural resonance of the bore of the magnet may have been introducing its own spectral shaping. Recordings of the subjects' productions made in more favorable conditions did, on analysis, exhibit a different pattern of divergence from the synthesized samples.

In the final analysis, however, we have to say that we do not yet know why the synthesized vowels failed to match the formant frequencies of the original utterances. Having explored many avenues to the limits of accuracy of our data, we are now convinced that, in order to get to the bottom of the matter, it will be necessary to repeat the experiment. In the

course of that replication, special efforts will have to be made to minimize several identified sources of error. Among the precautions that should be considered are: (i) confining all future studies to persons who have received phonetic training and are likely to more reliably reproduce the required vocal configurations; (ii) seeking equipment capable of faster imaging times and thinner image planes; (iii) developing special receiver coils that custom-fit each subject's head and neck to enhance the signal/noise ratio and achieve higher image resolution over the entire length of the tract, and using this equipment to obtain better data on the dimensions of the piriform and valecular sinuses; (iv) repeating image acquisitions and examining images more analytically during acquisition to insure that both the equipment and the subject are operating in a stable fashion; (v) adopting a grid system that conforms more closely to the anatomy of the vocal tract and methods of calculation that will produce area estimates for planes oriented more nearly at right angles to the center line of the tract; (vi) upgrading the synthesis procedure, using a time-domain finite-difference algorithm that will exploit the full resolution of the area data, including side cavities, with a view to achieving more accurate resynthesis of the original utterances; and, lastly (vii) upgrading the voice recording equipment used inside the MR magnet and determining the natural modes of acoustic resonance of its bore when a subject is present.

B. Experiment 2

1. Relations between midsagittal and lateral widths

A plot of midsagittal width against lateral width in the pharynx across vowels can be represented with a high degree of significance by a straight line. The failure of our data to replicate the reduced rate of lateral versus midsagittal expansion for the larger dimensions in the upper pharynx is somewhat surprising in light of the fact that the observation has been reported on three previous occasions (Fant, 1960; Sundberg, 1969; Sundberg *et al.*, 1987). A possible explanation for this discrepancy may be related to inherent differences between both the subjects and procedures employed. Our subjects were native speakers of a different language (English) from those employed in the earlier studies. Moreover, the procedures adopted in the present experiment required the subjects to vocalize for longer periods, and it may have been the case that our subjects avoided the strain normally associated with efforts to continuously hold extreme articulations and used less extreme vocal postures that made fewer demands on the ability to expand in either of the two dimensions.

2. Relations between midsagittal widths and areas

It was shown that, in almost all cases, the graphs of log midsagittal width versus the log of the cross-sectional area for different vowels could be adequately represented by a straight line no matter at what height above the level of the larynx the data originated. The exceptions to this rule appear most prominently in the data of subject TB at levels within 2.5 cm of the glottis. One reason for these exceptions may stem from the fact, already noted, that this is a region where image resolution rapidly declines. We believe that in future studies, use of a custom-made receiver coil designed to operate in closer proximity to the larynx will improve image quality. An alternative reason may simply be that little dimensional variation occurs in the larynx tube of this subject and that the midsagittal and area measurements are all clustered about fixed values.

The abrupt increase of both K and r apparent in the graphs of subject TB in the region between 3.0 and 4.0 cm above the larynx (see Fig. 20) is probably due to the already noted tendency for the tip of the epiglottis to draw away from the tongue surface and to project into the pharyngeal airway as the pharyngeal cavity becomes larger. Evidence of similar epiglottal behavior on the part of subject PN appeared at only the more extreme modes of pharyngeal expansion. Thus the midsagittal width measures of subject PN recorded, in most cases, the distance between the surface of the epiglottis and the posterior pharyngeal wall whereas, for subject TB, the recorded measurements were more evenly divided between those in which the tongue and those in which the epiglottis formed the anterior boundary. In consequence, the plotted data of subject TB tend to adopt a bimodal distribution due to the increase in S at the transition point between the two boundary criteria and an increase in the gradient of the regression line results in an elevation of both K and r.

The average of the r exponents at all distances from the larynx is 1.97, which may be regarded as a general-purpose approximation for r that can be applied throughout the pharynx. However, closer inspection of the individual exponents measured at different distances from the larynx show that the values of r are approximately constant over 4 cm of the upper half of the pharynx. Below that level a steep decrease occurs, possibly due to the presence of the epiglottis. At the lowest level, where the data come only from subject TB, the exponent reaches a value near unity, suggesting that the lateral width of the airway remains constant at this level. This interpretation seems intuitively reasonable, since this level is still within the larynx tube, and a closer look at Fig. 13 suggests that it is consistent with the data. Whether, for synthesis purposes, the acoustic results obtained might warrant use of either the parabolic approximation given in Fig. 20 or the individual exponent values computed at each level rather than the single value of 1.97 is not known and will depend on the results of experimentation with particular synthesis algorithms.

The relationship between midsagittal width and cross-sectional area that we have derived for the pharyngeal and oral regions may be used to calculate areas from widths measured from lateral radiographs. However, three caveats are in order. First, the reported data were all measured in either horizontal or vertical planes, so care must be taken in applying them to regions where the centerline of the vocal tract makes a very oblique angle with these planes. Second, in regions where the tongue surface is grooved, the tract width that appears in lateral x rays may not always correspond to the width we measured. More specifically, we used as our reference the medial apex of the groove whereas measures from lateral x rays are likely to represent the midsagittal

distance from a point on the tongue that corresponds more closely to its lateral edge, especially if midline markers are not used. Finally, it must be remembered that our data are derived from static vocal tract shapes, and we can only guess at how well the particular shapes we have studied will relate to the shapes that occur during dynamic speech production.

IV. CONCLUSIONS

The MR equipment that we used, particularly in the initial stages of the study (experiment 1), required long imaging times for single-slice images. This was followed by further periods of extensive data manipulation, software development and analysis. As our experience and the equipment available to us grew more refined, the images could be obtained more rapidly while still sustaining the quality shown in Figs. 1–5. The long imaging times were particularly troublesome because they plagued efforts to maintain or exactly repeat desired vocal tract shapes. Because the EMR machine allowed only single-slice imaging, several hours of imaging time, often spanning more than one session, was necessary to complete a full vocal tract shape. In consequence, the reliability of our data was not as high as we would have wished or now think that we can achieve. The reason for this optimism lies in the more powerful and faster equipment that is currently available.

The experimental difficulties just noted impose certain limitations on the quality of the data. Nevertheless, we have obtained estimates of the vocal tract area functions of two male speakers during productions of the four point vowels and, in a further experiment, have also examined the relations between cross-sectional areas and midsagittal widths in the pharynx over a range of nine different vowel configurations. Although this is not the first vocal tract study to employ MR technology (see, for example, Rokkaku *et al.*, 1986; Lakshminarayanan, Lee, and McCutcheon, 1991), it is the first systematic attempt at measurement and has assembled what probably constitutes the most comprehensive body of dimensional data on the vocal tract presently available. With the ability to obtain images approaching the resolution of x-ray CT but without the use of ionizing radiation and with the prospect of shorter image acquisition times in the future, it is evident that MR techniques have considerable promise and will be used more frequently to collect data on an ever widening repertoire of speech activity.

ACKNOWLEDGMENTS

This research was supported by Grants DC-00121 (formerly NS-13617), DC-00125 (formerly NS-13870), and RR-05596 from the National Institutes of Health and by the Esther A. and Joseph Klingenstein Fund, General Electric Company and the Yale School of Medicine. The authors wish to thank Robin Greene, Donald Hailey, Nianqi Ren, Richard Sharkany, Edward Wiley, and Kenneth Wilkins for providing essential technical assistance, Dr. Ignatius Mattingly for comments on an early draft of the manuscript, Dr. Susan Boyce, Dr. David Garrett, Dr. Hiroshi Muta, and Dr. Kouichi Tsunoda for valuable effort and expertise at various stages in the process of data gathering and data analysis, Dr. Richard McGowan for acoustical advice, and the reviewers for helpful comments.

[1] We have included only SIGNA images because, during the course of manuscript preparation, we discovered that the film-printed versions of all the EMR images, which had been collected for use as illustrations, had their widths shortened by 12% with respect to their heights. The source of this error arose exclusively in the printer interface and did not affect the original digital image data on which our measurements were made. However, the conversion steps that we would have been obliged to take to restore the data to their original format and obtain new high-quality prints were judged to be too time-consuming in relation to their importance for publishing purposes to merit the expenditure of effort. Thus, in order to illustrate this paper, we chose to reproduce in Figs. 1–5, a selection of images generated exclusively by the SIGNA machine. However, no measurements of the coronal images shown in Figs. 4 and 5 were made. All coronal measurements were based on images obtained from the EMR system.

[2] This criterion had the merit of consistency with respect to the pharyngeal study, but as in that study, in instances where tongue grooving occurred, tended to make the accuracy of S (the midsagittal width measure as defined) highly dependent on whether the true depth of the groove was actually resolved. This groove was frequently so narrow that some minor mistuning of the transmitter–receiver assembly could have a large impact on the magnitude of S in the oral region. The true depth of this feature was likely to be more accurately measured at locations where the base of the cleft and the centerline of the tract ran at a right angle to the image plane. The depths of grooves located at the rear of oral cavity, where the planes of intersection are least orthogonal to the tongue surface, appeared exaggerated and, thus, were poorly resolved.

Abramson, A. S., and Cooper, F. S. (1963). "Slow motion x-ray pictures with stretched speech as a research tool," J. Acoust. Soc. Am. 35, 1888–1889 (abs); also ACLS Newsletter 14, 7, 1963.

Anthony, J. (1964). "Replica of the vocal tract," Working Papers in Phonetics, University of California, Los Angeles, 1 (10), 10–14.

Baer, T., Gore, J. C., Boyce, S., and Nye, P. W. (1987). "Application of MRI to the analysis of speech production," Mag. Reson. Imag. 5, 1–7.

Boe, L-J., Perrier, P., and Sock, R. (1988). "Exploitation du modèle articulatoire du CNET. Mise en place d'une base de données articulatoire," Rapport intermédiare. Institut de la Communication Parlée, Unité Associée au C.N.R.S. No. 368, Institut de Phonetique de Grenoble, Juillet 1988.

Bradley, W. G., Newton, T. H., and Crooks, L. E. (1983). "Physical principles of nuclear magnetic resonance," in *Modern Neuroradiology: Advanced Imaging Techniques*, edited by T. H. Newton and D. G. Potts (Clavadel, San Francisco), Vol. 2, pp. 15–62.

Browman, C. P., and Goldstein, L. (1987). "Tiers in articulatory phonology, with some implications for casual speech," Haskins Labs. Stat. Rep. Speech Res SR-92, 1–30; also in *Papers in Laboratory Phonology I: Between the Grammar and The Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge U. P., Cambridge, England, 1990), pp. 341–376.

Chiba, T., and Kajiyama, M. (1941). *The Vowel, Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo).

Coker, C. H. (1976). "A model of articulatory dynamics and control," Proc. IEEE 64, 452–460.

Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).

Fant, G. (1965). "Formants and cavities," in *Proceedings of the Fifth International Congress of Phonetic Sciences*, edited by E. Zwirner and W. Bethge (Karger, Basel), pp. 120–140.

Fant, G. (1980). "The relation between area functions and the acoustic signal," Phonetica 37, 55–86.

Flach, M., and Schwickardie, H. (1966). "Die Recessus Piriformes unter phoniatrischer Sicht," Folia Phoniatr. 18, 153–167.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1975). "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell Syst. Tech. J. 544, 485–506.

Fujimura, O., Kiritani, S., and Ishida, H. (1973). "Computer controlled radiography for observation of movements of articulatory and other human organs," Comput. Biol. Med. 3, 371–384.

Fujimura, O., Baer, T., and Niimi, S. (1979). "A stereo fiberscope with interlens bridge for laryngeal observation," J. Acoust. Soc. Am. 65, 478–480.

Gaufin, J., and Sundberg, J. (1978). "Pharyngeal constrictions," Phonetica 35, 157–168.
Heinz, J. M., and Stevens, K. N. (1964). "On the derivation of area functions and acoustic spectra from cineradiographic films of speech," J. Acoust. Soc. Am. 36, 1037 (abs).
Henderson, L., and Clare, F. (1979). NCAR Graphics Software. Atmospheric Technology Division, National Center for Atmospheric Research, Boulder, CO.
Johansson, C., Sundberg, J., Wilbrand, H., and Ytterbergh, C. (1983). "From sagittal distance to area: A study of transverse, cross-sectional area in the pharynx by means of computer tomography," R. Inst. Technol. STL-QPSR 4/1983, 39–49.
Kakita, Y., and Fujimura, O. (1977). "A computational model of the tongue: A revised version," J. Acoust. Soc. Am. Suppl. 1 62, S15.
Kelly, J. L., and Lochbaum, C. C. (1962). "Speech synthesis," A paper delivered at the IVth Int. Congr. Acoust. (Stockholm), published in *Proceedings of the Speech Communication Seminar*, Stockholm, Speech Transmission Laboratory, Vol. II, 1963, Paper F7.
Kiritani, S., Tateno, Y., Iinuma, T., and Sawashima, M. (1977). "Computed tomography of the vocal tract," in *Dynamic Aspects of Speech Production*, edited by M. Sawashima and F. Cooper (Univ. of Tokyo Press, Tokyo), pp. 203–206.
Ladefoged, P., Anthony, J. F. K., and Riley, C. (1971). "Direct measurement of the vocal tract," UCLA Working Papers in Phonetics 19, 4–13; also in J. Acoust. Soc. Am. 49, 104 (abs) (1971).
Lakshminarayanan, A. V., Lee, S., and McCutcheon, M. J. (1991). "MR imaging of the vocal tract during vowel production," J. Magn. Reson. Imag. 1, 71–76.
Liljencrants, J. (1985). "Speech synthesis with a reflection-type analog," Ph.D. dissertation, Royal Inst. of Technology, Stockholm.
Lin, Q. (1990). "Speech production theory and articulatory speech synthesis," Ph.D. dissertation, Royal Inst. of Technology, Stockholm.
Maeda, S. (1982). "A digital simulation of the vocal-tract system," Speech Comm. 1, 199–229.
Martelli, A. (1976). "An application of heuristic search methods to edge and contour detection," Comm. ACM 19, 73–83.
McGill, W. J. (1954). "Multivariate information transmission," Psychometrika 19, 97–116.
McGowan, R. (1988). "An aeroacoustic approach to phonation," J. Acoust. Soc. Am. 83, 696–704.
Mermelstein, P. (1967). "On the Piriform Recessus and their acoustic effects," Folia Phoniatr. 19, 388–389.
Mermelstein, P. (1971). "Calculation of the vocal-tract transfer function for speech synthesis applications," in *Proceedings of the VIIth International Congress on Acoustics* (Akadémiai Kiadó, Budapest), Vol. 3, pp. 173–176.
Mermelstein, P. (1973). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. 53, 1070–1082.
Minifie, F. D., Kelsey, C. A., and Zagzebski, J. A. (1971). "Ultrasonic investigation of tongue shape," J. Acoust. Soc. Am. 54, 1857–1860.
Perkell, J. S., *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (MIT, Cambridge, MA, 1969).
Rokkaku, M., Hashimoto, K., Imaizumi, S., Niimi, S., and Kiritani, S. (1986). "Measurements of the three dimensional shape of the vocal tract based on the magnetic resonance imaging technique," Ann. Bull. Res. Inst. Logopedics and Phoniatrics 20, 47–54.
Rubin, P., Baer, T., and Mermelstein, P. (1981). "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. 70, 321–328.
Schonle, P. W., Grabe, K., Wenig, P., Hohne, J., Schrader, J., and Conrad, B. (1987). "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain Lang. 31, 26–35.
Shadle, C. H. (1985). "The acoustics of fricative constants," Ph.D. dissertation, MIT, Cambridge, MA.
Shawker, T. H., and Sonies, B. C. (1984). "Tongue movement during speech: a real-time ultrasound evaluation," J. Clin. Ultrasound 12, 125–133.
Shawker, T. H., Sonies, B. C., and Stone, M. (1984). "Soft tissue anatomy of the tongue and floor of the mouth: an ultrasound demonstration," Brain Language 21, 335–350.
Stevens, K. N., and House, A. S. (1955). "Development of a quantitative description of vowel articulation," J. Acoust. Soc. Am. 27, 484–493.
Stone, M. (1990). "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," J. Acoust. Soc. Am. 87, 2207–2217.
Sundberg, J. (1969). "On the problem of obtaining area functions from lateral x-ray pictures of the vocal tract," Royal Inst. Technol. STL-QPSR 1/1969, 43–45.
Sundberg, J. (1974). "Articulatory interpretation of the singing formant," J. Acoust. Soc. Am. 55, 838–844.
Sundberg, J., Johansson, C., Wilbrand, H., and Ytterbergh, C. (1987). "From sagittal distance to area: a study of transverse, vocal tract cross-sectional area," Phonetica 44, 76–90.
Teager, H., and Teager, S. (1983). "The effects of separated air flow on vocalization," in *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, edited by I. R. Titze and R. C. Scherer (College-Hill, San Diego), pp. 124–143.