# Vocal tract area functions from magnetic resonance imaging

Brad H. Story and Ingo R. Titze
*Department of Speech Pathology and Audiology, National Center for Voice and Speech, University of Iowa, Iowa City, Iowa 52242*

Eric A. Hoffman
*Division of Physiologic Imaging, Department of Radiology, University of Iowa College Medicine, Iowa City, Iowa 52242*

(Received 30 October 1995; accepted for publication 20 February 1996)

There have been considerable research efforts in the area of vocal tract modeling but there is still a small body of information regarding direct 3-D measurements of the vocal tract shape. The purpose of this study was to acquire, using magnetic resonance imaging (MRI), an inventory of speaker-specific, three-dimensional, vocal tract air space shapes that correspond to a particular set of vowels and consonants. A set of 18 shapes was obtained for one male subject who vocalized while being scanned for 12 vowels, 3 nasals, and 3 plosives. The 3-D shapes were analyzed to find the cross-sectional areas evaluated within planes always chosen to be perpendicular to the centerline extending from the glottis to the mouth to produce an ''area function.'' This paper provides a speaker-specific catalogue of area functions for 18 vocal tract shapes. Comparisons of formant locations extracted from the natural (recorded) speech of the imaged subject and from simulations using the newly acquired area functions show reasonable similarity but suggest that the imaged vocal tract shapes may be somewhat centralized. Additionally, comparisons of the area functions reported in this study are compared with those from four previous studies and demonstrate general similarities in shape but also obvious differences that can be attributed to differences in imaging techniques, image processing methods, and anatomical differences of the imaged subjects. © *1996 Acoustical Society of America.*

PACS numbers: 43.70.Aj, 43.70.Jt [AL]

## INTRODUCTION

Speech simulation algorithms have been developed to provide a sophisticated representation of acoustic wave propagation through the vocal tract [Ishizaka and Flanagan (1972), Strube (1982), Maeda (1982), Sondhi and Schroeter (1987) and Liljencrants (1985)]. Most speech simulation models are based on the assumption of one-dimensional wave propagation. This means that the tubular vocal tract shape can be approximated as a finite number of cylindrical elements that are ''stacked'' consecutively from the larynx to the mouth. A particular vocal tract shape can be imposed on a model by specifying the cross-sectional area of each cylindrical element as a function of the distance from the glottis. For modeling purposes, any vocal tract shape can be defined by its unique ''area function.'' Hence, a necessary component for the simulation of natural sounding speech is an inventory of vocal tract area functions that correspond to the vowels and consonants used to produce human speech. The success of speech simulators has been limited, in part, by the lack of a body of morphological information about the vocal tract shape on which to base these area functions. Shape information regarding other airways such as the nasal tract and trachea is also needed. The one nearly complete set of area functions that has been used extensively as input to speech simulation models, is that published in 1960 by Fant. These area functions were indirectly acquired with sagittal x-ray projection images and plaster casts of the oral cavity and have proven to be an invaluable resource for researchers in the field of speech synthesis and simulation.

Modern imaging techniques are now being used to acquire three-dimensional shape information about the vocal tract volume and associated airways. Volumetric imaging relies on the ability of an imaging technique to acquire a series of image slices, in one or more anatomical planes, through a desired volume of the human body. This process can be performed with either magnetic resonance imaging (MRI) or electron beam computed tomography (EBCT). Postprocessing of the images typically includes segmentation of the air space from the surrounding tissue and a 3-D reconstruction of the airway shape. Once the 3-D structure has been identified, it can be measured and analyzed. During speech production, a speaker will traverse many different vocal tract shapes in a short period of time, often overlapping one shape into another (i.e., coarticulation). Unfortunately, neither MRI or EBCT can acquire a volume image set fast enough to capture the dynamically changing vocal tract shape. Thus the use of present, commercially available, imaging techniques can be used only to study static vocal tract shapes.

MRI is an attractive imaging technique primarily because it poses no known danger to the human subject being imaged. Since human speech comprises a large number of vowel and consonant shapes it is often desirable to acquire a large number of image sets. Thus the human subject may need to be exposed to hours of imaging, but fortunately, with no apparent risk. However, MRI has a number of distinct disadvantages. The scanning time required for acquisition of a full image set is on the order of several minutes. When this is coupled with the pauses required for subject respiration, a complete image set may require double or triple the actual scan time. Thus the image set and subsequent reconstruction of the airway is based on a large number of repetitions of a particular shape. Also, because of various imaging artifacts associated with air/water interfaces, the air–tissue interface can be poorly defined when imaged via MRI techniques. Additionally, teeth and bone are poorly imaged by MRI because of the low hydrogen (water) concentration, making these structures appear to be the same gray scale value as air.

Baer *et al.* (1991) demonstrated the use of MRI to directly measure the vocal tract shape for the four point vowels with two adult male subjects. In two separate experiments, they collected combinations of image sets in three planes from which they created the first demonstration of 3-D reconstructions of the vocal tract shape using MR imaging techniques. They also reported the corresponding area functions for each vowel that were discretized along the vocal tract centerline at intervals of 0.875 cm and proposed a midsagittal width-to-area transformation based on their data. Moore (1992) performed an MRI study in which sagittal and coronal image sets of the vocal tract were obtained from five adult male subjects for three vowels and two continuants. The study investigated correlations between cavity volumes and resonance frequencies of the vocal tract, as well as other variables such as lip length and lip area. However, it did not include three-dimensional reconstructions of the vocal tract shapes. Another MRI study of the vocal tract was performed by Sulter *et al.* (1992). They used one male subject who was a trained singer. Their primary interest was in correlating the vocal tract length measured for an /ɛ/ vowel to resonance frequencies (formants) predicted by theory. They also imaged the vowels /i/ and /ɑ/ from which they measured cavity volumes. Greenwood *et al.* (1992) imaged five static vowels using one subject. They acquired an image in the midsagittal plane, and a set of images in the axial and coronal planes. Areas functions were extracted from the image slices in close accordance with the vocal tract model proposed by Mermelstein (1973). Dang *et al.* (1994) used MRI to acquire contiguous coronal image sets throughout the nasal tract volume and sagittal image sets of the vocal tract for nasal consonants /m/ and /n/. Using these data they produced 3-D reconstructions of the nasal tract passages and sinus cavities. The nasal tract morphology along with the vocal tract area functions were subsequently used to model the acoustic characteristics of the nasal system and production of nasal consonants. Reconstruction of 3-D vocal tract shapes for five vowels was reported for an adult male, adult female, and an 11-year-old male by Yang and Kasuya (1994). Coronal and axial image sets were acquired in order to define the oral and pharyngeal cavities, respectively. Area functions that were extracted from the vocal tract reconstructions for each subject were implemented in a frequency domain model of acoustic wave propagation (Sondhi and Schroeter, 1987). Formant frequency comparisons between the model and natural recorded speech were quite close, with many cases exhibiting less than a 5% error. Narayanan *et al.* (1995) and Narayanan (1995) have reported the acquisition of coronal and axial image sets from four adult subjects (2 male and 2 female) representing the fricative consonsants, /s,ʃ,f,θ,z,ʒ,v,ð/. 3-D reconstructions of both the vocal tract shape and the tongue were created from these images and a cross-sectional area analysis was performed. This study provides the most accurate information to date of the constrictions and air channels that produce the turbulence generated sound, characteristic of fricative consonants. Lakshminarayanan *et al.* (1991) have focused on using MRI to acquire midsagittal sections of the vocal tract from which they measured the tract widths and converted them to cross-sectional areas using an empirically derived formula from Rubin *et al.* (1981). Perrier *et al.* (1992) and Beautemps *et al.* (1995) have both used x-ray techniques in an attempt to further develop transformations from the midsagittal width to area.

The intent of this experiment was to obtain a collection of vocal tract shapes (and consequent area functions) corresponding to a large number of speech sounds for *one* specific speaker. Such a collection should allow for a unique simulation of speech (i.e., of the imaged subject) in which acoustic pressures and flows generated by a speech production model can be directly compared to that of the subject. Additionally, methods of interpolating or convolving the area functions in time to form ‘‘running speech’’ might be more easily studied since a direct comparison can be made between simulation and recorded natural speech of the subject. Thus a set of area functions from one speaker will allow a more rigorous evaluation of speech modeling algorithms since the simulation can be directly compared to the real human system being modeled. Twenty-two complete vocal tract shapes, trachea, piriform sinuses, and two states of the nasal tract were scanned with MRI for this study. However, only a subset containing 12 vowels, 3 nasals, and 3 plosives will be presented in this paper.

In addition to requiring that one subject produce all of the desired vocal tract shapes it was desired that the image set representing any one particular shape be completely acquired within one imaging session; i.e., the subject would be positioned in the scanner only once for any particular vowel or consonant. This requirement was intended to avoid the potential errors associated with repositioning the subject in the scanner for separate sessions as well as the possibility that the subject might produce a particular vowel or consonant with a slightly different vocal tract shape from session to session. The volumetric imaging studies reviewed above (Baer *et al.*, 1991; Moore, 1992; Sulter, 1992; Greenwood *et al.*, 1992; Dang *et al.*, 1994; Narayanan *et al.*, 1995) all acquired multiple plane image sets. In these studies it has typically been regarded that some combination of two imaging planes is necessary for measurement of the vocal tract shape. But because of the physical demands required of a single subject to produce different vocal tract shapes for 22 full volume scans (with phonation during scanning) plus two volume scans of the nasal tract and one of the trachea, it was concluded that each image set could reasonably include only one imaging plane. Even with the reduction to one imaging plane, the subject spent between 7 and 8 h over the course of three separate sessions lying in the MRI scanner. The axial plane was chosen because it would most accurately represent the small epilaryngeal region and the often constricted pharynx as well allowing a direct comparison to a limited number

TABLE I. MR parameters for vocal tract image acquisition.

Mode=fast spin-echo, using an anterior neck coil
TE=13 ms (echo delay time)
TR=4000 ms (repetition time)
ETL=16 ms (echo train length)
FOV=24 cm (field of view)
NEX=2 (number of excitations)
image matrix=256×256 pixels
resolution=0.938 mm/pixel
slice thickness=5 mm

of image sets acquired with electron beam computed tomography (EBCT) that were part of another experiment (Story, 1995; Story *et al.*, 1996); EBCT allows only the acquisition of axial image sets. Using only axial slices may compromise the measurement accuracy in the oral cavity, particularly for front vowels when there is little air space between the upper surface of the tongue and the hard plate. In such a case, the thickness of the MR slice may be larger than the airway passage. Ideally, multiplane scanning would be performed for every vowel and consonant shape produced by the subject, but subject endurance as well as monetary and time constraints associated with obtaining scanning time prevented this ideal case.

This paper is primarily intended to report area function data for static vocal tract shapes. However, comparisons of formant locations produced by 12 tract shapes for both the natural speech of the subject and a simulation based on the new set of area functions will be presented.

## I. IMAGE ACQUISITION AND ANALYSIS

### A. MRI scanning parameters and image acquisition protocol

The MR images were acquired using a General Electric Signa 1.5 Tesla scanner. MRI produces a slightly blurred boundary between tissue and air (Baer *et al.*, 1991). However, the acquisition mode can be chosen and the pulse sequence parameters adjusted to provide acceptable air–tissue interfaces. The parameters shown in Table I were empirically found to provide acceptable images of the airway.

Volumetric imaging of the vocal tract using MRI for a single subject was completed for 22 different phoneme configurations and also for the nasal tract and trachea (but as stated previously, only the analysis of 18 tract configurations will be discussed in this paper). The subject (BS) was a 29-year-old male with no history of speech or voice disorders and is a native of the midwestern United States. The subject is 5 ft 7 in. tall and weighs approximately 145 pounds. His head circumference was measured to be 57 cm and for the neck, 34 cm, measured just above the prominence of the thyroid cartilage. Height, weight, and head and neck sizes have not typically been reported in previous imaging studies, but such information may be useful to the reader when comparing the data presented here to other data sets. While the body dimensions given certainly do not define a vocal tract size, they may provide a clearer picture of the subject's body structure. The subject is also the first author of this study.

Using the parameters given in Table I, a 26 slice series of 5-mm-thick contiguous, parallel, axial sections was gathered in an interleaved acquisition. This image set, which extended from just above the hard plate down to the first tracheal ring, could be acquired with 4 min and 16 s of scan time. However, because the subject was required to phonate during image acquisition, the actual amount of time required to image one shape was approximately 10 min, allowing for respiration during pauses in scanning.

Prior to the imaging session, the subject spent a significant amount of time practicing phonation while holding the vocal tract shape as steady as possible. The protocol for image acquisition was as follows. The subject was first given ear plugs to attenuate the intense sound of the MR machine. The subject was then positioned in a comfortable supine position on the patient table in the MRI examination room. The subject's head was placed directly on the table (no foam or cushion) and positioned so that the table was perpendicular to the Frankfort plane. Cloth adhesive tape was then used to secure the head to the table in this position. With this approach, very little head movement was possible. An anterior neck coil was brought into position so that the desired portion of the head and neck were within its field of view. With this preparation completed, the patient table was moved such that the subject's head was in the center of the magnet. Prior to each image acquisition session, a sagittal localizer was performed to allow for identification of the appropriate field of view and scan location.

During protocol development a foot signaling method was tested to indicate to the MR technician when a breath was needed. However, the foot movement introduced motion artifacts and blurring. The subject reported that this method interrupted concentration. In the final study the technician started the image acquisition when the subject began phonation (as heard over the intercom) and then ''paused'' the machine after eight seconds of scanning so the subject could take a breath. As soon as the subject began phonating again, the image acquisition was continued. This method allowed the subject to be in control of the time between acquisitions and also keep the vocal tract stable. The subject was given any pertinent instructions over an intercom system. Throughout all image acquisitions a speech scientist, experienced in phonetics, was present in the control room to listen to each speech sound and halt the acquisition if the subject strayed from the desired target.

The phonemes used to create static vocal tract shapes during imaging are shown in Table II with the example word that was given to the subject prior to the image acquisition. For the consonants, the subject imagined that the preceding and following vowel shape was a neutral or schwa vowel (note the rather strange example for /ŋ/).

### B. Image analysis

All image analysis operations were performed with a general Unix-based image display and quantitation package called VIDA™ (volumetric image display and analysis) which has been developed (and continues to be enhanced) by reseachers in the Division of Physiologic Imaging at the University of Iowa (Hoffman *et al.*, 1992). Additional informa-

TABLE II. List of imaged phonemes using phonetic symbols, symbols, and example words.

| Phonetic symbol | Example |
|---|---|
| /i/ | heed |
| /ɪ/ | hid |
| /ɛ/ | head |
| /æ/ | had |
| /ʌ/ | ton |
| /ɑ/ | hod |
| /ɔ/ | paw |
| /o/ | hoe |
| /ʊ/ | hood |
| /u/ | who |
| /ɝ/ | earth |
| /l/ | lump |
| /m/ | mum |
| /n/ | numb |
| /ŋ/ | ung a |
| /p/ | puck |
| /t/ | tuck |
| /k/ | cut |

tion regarding VIDA can be found on the Internet by using a web browser and logging onto http://everest.radiology.uiowa.edu/.

The goal of this experiment was to determine area functions for each of the phonemes listed in Table II. To achieve this goal the image analysis process included three main steps: (1) segmentation of the airway from the surrounding tissue, (2) three-dimensional reconstruction of the airway by shape-based interpolation, and (3) determination of an airway centerline and subsequent extraction of cross-sectional areas assessed from oblique sections calculated to be locally perpendicular to the airway centerline.

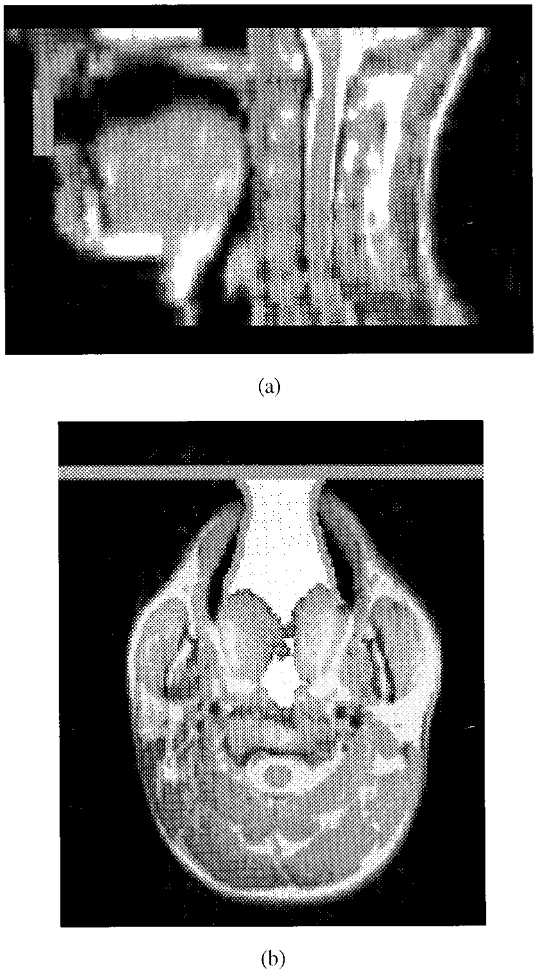
### C. Airway segmentation

The image data sets were transferred from the GE Signa scanner to a Unix-based workstation (via magnetic tape) and translated into a file format recognized by the analysis and quantitation software (VIDA). When files are in this format they can be read into a shared memory structure, and for convenience, images were converted from 16 to 8 bit per pixel gray scale resolution. The upper gray level cutoff was determined by computing a pixel histogram on several slices sampled throughout the data set and choosing the cutoff point so that all pixel values present in the image set were below the cutoff (our version of the GE scanner downloaded the MR data sets as 8 bits of gray scale information stored in 16 bit pixel values).

The airway was segmented from the surrounding tissue by setting all vowels considered to be in the airway to a unique gray scale value. The first step in this process was to attenuate all gray scale values in an image slice by 5%. This ensured that no voxel in the image volume had a value of 255 (the brightest value for an 8-bit gray-scale range). Next a gray scale value that represented the threshold dividing air and tissue was determined by choosing the voxel intensity that was halfway between the darkest part of the airway and the brightest part of the tissue surrounding the airway as determined by a region of interest analysis. Validation of this approach using the image data sets for a phantom study is discussed in the Appendix.

Once the threshold value had been determined the actual segmentation process could be performed. A seeded region growing algorithm was used that changed the gray scale values of all pixels below the threshold value to the brightest possible gray level (Hoffman *et al.*, 1983; Udupa, 1991). This effectively set the airway to a single color that, because of the previous attenuation process, was unique with respect to the rest of the image volume. The seeded region growing process required the airway region to be noncontinuous with air outside the body. For the axial slices in the pharyngeal section this was not a problem, but in the oral region where some slices have an open mouth condition, the algorithm would ''leak'' the brightest pixel value out of airway and into the ''blackness'' (air) outside the image. To correct the problem, an artificial boundary that defined the mouth termination of the vocal tract was constructed. First the image set was viewed in a reformatted midsagittal plane (extracted from the volumetric image data set) where the outline of the lips are easily seen. A mouth termination plane was then defined using the technique of Mermelstein (1973). Figure 1 shows the boundary in the midsagittal plane and the containment of the seeded region growth in the axial plane.

A potential problem that arises in segmentation of the oral cavity is that the 5-mm slice thickness may cut through both tissue and air. For example, the upper 2.5 mm of a slice may contain a portion of the hard palate while the lower 2.5 mm contains air. The result would be a voxel density with some average value of air and tissue which could be above or below the set threshold; i.e., tissue could be included as air and vice versa. The most superior slice in a given image set normally included the lower portion of the nasal tract and also some of the velopharyngeal air space for the nasal consonants. Usually, this slice was ignored except in the case of the nasals. The next slice inferior to the first, typically contained the hard palate and extended down to include the midline raphe on the roof of the mouth. Because of the uneven surface created by the midline raphe on the roof of the mouth, this slice contained some voxels with densities made up of contributions from both air and tissue; their values often would lie below the threshold determined for segmentation. During the segmentation process, this particular slice was always segmented with the threshold-based seeded region growing method. However, the subsequent 3-D reconstruction of the tract shape using methods described in the next section, were performed with and without this particular slice. It was often the case that, when the slice was included, the 3-D reconstruction possessed some rather unnatural ''fins'' on its most superior portion that were most likely the result of including something other than air. Thus in such cases, this particular slice was ignored and the interpolation, surface rendering, and area function analysis were performed without it. A similar error could occur in the inferior part of the oral cavity where portions of the tongue and adjacent airspace might be included within a slice. However, the effect may be less severe in this case because the variety of


(a)

(b)

FIG. 1. Method of terminating the open mouth to contain the seeded region growing, (a) ''painted'' rectangle on the midsagittal slice, (b) containment ''fence'' in the axial plane.

shapes assumed by the tongue in the production of different speech sounds may ensure that one slice would not usually contain the entire upper surface of the tongue. Thus no attempt was made to alter the threshold-based segmentation in this region. With the approach taken, it is possible that a very small volume of air was excluded from the airway which would have the most significant effect on vowels with a front constriction such as /i/, /ɪ/, and /ɛ/.

### D. Volume reconstruction

Interpolation to produce an isotropic (cubic voxel) image set is often used when an image set has been collected in only one image plane and one dimension has a spatial resolution equal to the thickness of an image slice. Voxel densities can be interpolated between consecutive image slices to generate a uniformly sampled image set. The interpolation is typically performed with nearest neighbor, linear, or trilinear techniques (Udupa, 1991). If interaction by the user is required to identify and segment a region or structure of interest, the cubic voxel image set greatly increases the workload. For example, the MR image sets collected for this experiment were 26 slice series of 5-mm-thick slices. An interpolation to a cubic voxel image set would generate 138 slices.

Instead of the more general practice of interpolating voxel densities, a technique called shape-based interpolation (Raya and Udupa, 1990) was used to reconstruct the three-dimensional shape of the vocal tract airway. The shape-based algorithm interpolates segmented image data to form an isotropic (cubic voxel) data set. Thus user interaction can be limited to structure or region identification in the original image set. Since the image segmentation process assigned a gray-scale value to each voxel within the air space that was unique to the rest of the image (in particular the value 255), the segmented image set can be considered to represent a *binary* scene; i.e., each slice can be separated into a ''patch'' of air (gray level=255) or *not* air (gray-level<255). The technique of shape-based interpolation (Raya and Udupa, 1990) utilizes this binary representation of a structure (in this case the airway) within an image set to interpolate slices between the originals using information related to the structure's surface location above and below the new slice location to be generated. Within a given image slice, the shortest distance between each voxel and the boundary of the airway is computed and stored in a 2-D array with the same dimensions as the image slice. For example, if the image consists of a 256×256 pixel matrix and a shortest distance to the airway boundary is measured for the pixel located at a position of (123,50), then this measured distance is stored in the (123,50) location within the 2-D array. If the voxel was in the airway, the measured distance is considered to be positive valued, otherwise the distance is set to be negative. This process is performed for each image slice so that, when finished, a new set of ''slices'' (2-D arrays) containing distance measures has been created that maintains the correct spatial ordering of the original segmented image set. The interpolation between slices is performed on the shortest distance arrays, rather than actual image slices. When the interpolation is complete, the new set of shortest distance slices (2-D arrays) can be mapped back into a binary scene by assigning every positive distance value to a voxel in an image slice with a gray-level value of 255. Every negative distance value is assigned to a voxel with a gray-level value of 0. Thus the shape-based approach interpolates only the gray scale value of the airway while all other colors are ignored, effectively ''stripping'' all tissue away from the airway while creating the same voxel resolution in the axial direction as the other directions; i.e., isotropic resolution.

The interpolated airway produced by shape-based interpolation was subsequently used in both a surface rendering application and a cross-sectional area analysis. The surface rendering consisted of first extracting the edges of the interpolated airway image and then creating a shaded surface display. The result was digital ''cast'' of the vocal tract. Surface rendering applications used effectively in many cardiac and pulmonary applications are discussed in Hoffman (1991). High quality three-dimensional representations of the airway can be rotated and magnified to show many different perspectives. It is only a qualitative tool, but nevertheless an important step in the image analysis because it shows the quality of the segmentation process and provides three-dimensional views of each vocal tract shape.

The area analysis takes advantage of the resultant isotropic, binary representation of the vocal tract to compute oblique cross-sectional areas perpendicular to the local long axis. The details of this analysis are given in the next section.

### E. Area function analysis

To extract the area function from the interpolated vocal tract shape, an algorithm was used that was originally developed to analyze upper airway geometry and volume with regard to sleep disorders (Hoffman *et al.*, 1992; Hoffman and Gefter, 1990). It computes cross-sectional areas from oblique sections calculated to be perpendicular to the local airway long axis.

A segmented and interpolated image is used from which a three-dimensional centerline is computed through the airway. An iterative bisection algorithm is used to compute the centerline. A voxel is chosen by the user at the beginning and the end of the segmented image; i.e., the beginning voxel is chosen near the mouth termination and the end voxel is chosen near the glottis. A line is then drawn through three-dimensional space from the beginning voxel to the end voxel. The plane perpendicular to this line and halfway between the two specified voxels is determined, and all voxels in this plane that have the gray scale value of the airway are identified and used to calculate an average voxel location (i.e., the centroid). The new voxel location along with the original end points are now used to create two new line segments, two new planes perpendicular to the line segments, and consequently two new centroid points within the airway. The process can be repeated for any number of iterations specified by the user. When the iterations are finished the result is a sampled version of the three-dimensional centerline through the airway. Oblique sections were calculated perpendicular to the long axis of the airway at each centerline point. A voxel counting algorithm (summing all voxels of the airway gray scale value) was applied to each oblique section yielding the cross-sectional area, anterior–posterior length, and lateral length as a function of slice location along the airway length. The distance between consecutive cross sections can be calculated with the three-dimensional Pythagorean theorem,

$$L_1=\sqrt{[(x_1-x_0)^2+(y_1-y_0)^2+(z_1-z_0)^2]}, \tag{1}$$

where $(x_0,y_0,z_0)$ and $(x_1,y_1,z_1)$ are the spatial coordinates of the airway centroid locations from consecutive cross sections. However, because of the irregular shape of each cross section, the centroid location often varied within the plane of the section. This will impose small displacements of the centerline out of the axis of wave propagation. Figure 2(a) shows the lateral $(x)$ coordinate for the vowel /ɑ/ as a function of section number. The first section is at a point just above the glottis and the last section is at the lips. The figure shows that there is an overall left and right displacement as well as many small, sharp variations. While the variation from point to point is typically a millimeter or less, the summation of the variations can accumulate and add a centimeter or more to the total vocal tract length. For the frequency range of speech sounds, these variations are a tiny fraction of a given wavelength and it is, therefore, unlikely that a pressure wave would alter its direction of propagation in response to small off-axis variations; i.e., a propagating acoustic wave would not ''see'' these small variations as significant changes in the axis of propagation. To suppress their influence, an averaging filter was applied to the anterior–posterior $(y)$ and inferior–superior $(z)$ coordinates to smooth out sharp variations in the midsagittal plane. The lateral $(x)$ coordinate was not used in the length calculation, hence the tract length was based on the midsagittal vocal tract profile while the true centerline was used for oblique plane and cross-sectional area calculations. Both raw and smoothed tract profiles for the vowel /ɑ/ are shown in Fig. 2(b).

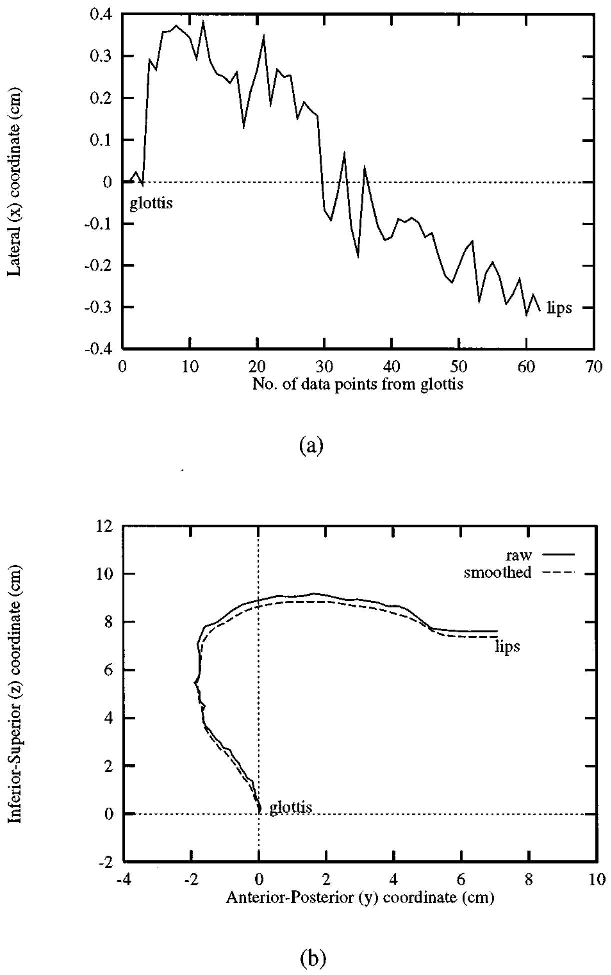


FIG. 2. Off-axis variations of the vocal tract centerline, (a) lateral $(x)$ coordinate variations as a function of section number; first section is at a point just above the glottis and the last point is at the lips, (b) raw and smoothed vocal tract profiles in the midsagittal plane.

It was difficult to locate the seed voxels at exactly the mouth termination plane and the plane just above the glottis. Since the cross-sectional areas of the these planes are important end points for each area function, a region of interest analysis was used to measure the area of the mouth termination by viewing the interpolated airway coronally and the

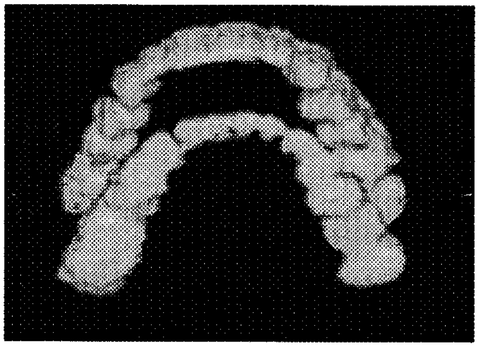


FIG. 3. A ‘‘digital cast’’ of the teeth using the EBCT image set for the vowel /ɑ/ (inferior, frontal view; i.e., looking up the teeth from below).

area just above the glottis was measured from an axial slice. Additionally, for each of these cross sections, the centroid coordinates were determined and added to the data set generated by the iterative bisection algorithm. The region of interest analysis was also used to determine the cross-sectional areas of the lateral pathways for the /l/ by measuring coronal slices through them. The iterative bisection method used for this study did not perform well for branching airways.

Each area function was generated as the set of $x$-$y$ pairs that include the length coordinate ($L_n$) and the corresponding area ($A_n$),

$$\left( L_n = \sum_{k=0}^{n} L_k , \quad A_n \right) . \tag{2}$$

The area function is assumed to begin at the glottal end of the vocal tract (i.e., $L=0$ is just above the glottis) and terminate at the lips.

## F. Exclusion of the teeth

A small concentration of hydrogen causes the teeth to be poorly imaged with MRI making these structures appear to be the same gray scale value as air. To avoid including the teeth during the segmentation process, an estimation of their extent into the oral cavity was required. Story (1995) and Story *et al.* (1996) reported an imaging experiment that used electron beam computed tomography (EBCT) to acquire axial image sets for two vowels, /i/ and /ɑ/ for the same subject (BS) imaged in the present experiment. Since EBCT uses x-ray radiation, the teeth are well defined in the image slices in which they are present. Thus the EBCT image sets provide a convenient means of measuring the dimensions of the teeth, in terms of their extent into the airway, which can then be used as an aid in segmenting the air from tissue and teeth in the MR image sets.

Using the EBCT image set for the vowel /ɑ/, the teeth were removed by segmenting them from all other tissue and air and then applying the shape-based interpolation algorithm described previously to extract only the teeth from the image set. A shaded surface rendering of the teeth extracted from the /ɑ/ vowel is shown in Fig. 3. The dimensions of this ‘‘cast’’ of the teeth can now be measured with a general region of interest analysis program (Hoffman *et al.,* 1992). The space occupied by the teeth within each axial MRI slice in the region of the oral cavity was estimated from these measurements and the airway segmentation was corrected by ‘‘painting out’’ a space approximately equivalent to the size of the teeth. In Fig. 1(b), the spatial contribution of the teeth has been excluded using this method; i.e., observe the black regions lateral to the airway that would have otherwise been included as air space.

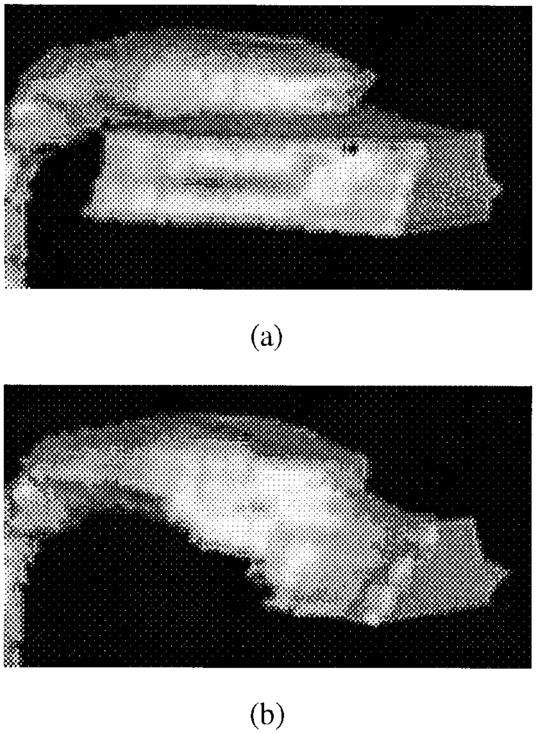


FIG. 4. Shaded surface displays of the vowel /ɑ/ where the flat surface in the upper right portion of the picture is the mouth termination plane: (a) perspective view *including* the teeth-the ‘‘fins’’ that are visible on right side of the oral region (they also exist on the left side) are due to the included presence of the teeth, (b) perspective view *excluding* the teeth.

A demonstration of this method of removing the contribution of teeth is shown in Fig. 4 in which two perspective views of the 3-D reconstruction of the /ɑ/ vowel oral cavity using the MRI data are given. The upper pharynx is seen at the left side of the figure which bends into the oral cavity, and terminates at the lips as represented by the flat plane. Figure 4(a) was reconstructed assuming that the teeth did not exist (i.e., all of the space occupied by the teeth was segmented as air) while Fig. 4(b) represents the same vowel shape but after using the dimensions of the teeth obtained from the EBCT images to estimate their extent and consequently eliminate their contribution to the MR derived air space image. Note that since the dimensions of the teeth do not change, their dimensions measured from the EBCT data can also be used in segmenting any other vowel shape acquired with MRI. Thus one EBCT image set provides the necessary information for estimating the location and size of the teeth in any MR image set when relating the teeth dimensions to their soft tissue surroundings.

## II. SPEECH SIMULATION AND ACOUSTIC RECORDING

### A. Simulation model

A wave-reflection analog vocal tract model (Story, 1995) was used to simulate the vowel sounds based on the area functions measured from the MR image sets. Energy losses due to the yielding properties of the vocal tract walls, fluid viscosity, and radiation from the mouth have been incorporated into the model. An acoustic side branch representing the piriform sinuses was also implemented. The model was sampled at a frequency of 44 100 Hz and each finite section of the area function represented a tube length of 0.396 cm.

The simulation of each vowel sound was performed by injecting a parametrized glottal flow waveform (Titze *et al.*, 1994) into the glottal end of the vocal tract to serve as the voice source. A parametrized source allows precise control of fundamental frequency (pitch) and spectral content (glottal waveshape) so that all of the simulations can be produced with exactly the same voice.

### B. Acoustic recordings of natural speech

In order to compare simulations to the natural speech of the subject, a high quality audio recording was made in which the subject produced speech sounds that corresponded to the static shapes that were acquired with MRI. An attempt was made to simulate, as closely as possible, the conditions experienced during the MRI sessions. The subject was set in a supine position on the suspended floor of an anechoic chamber. The Frankfort plane was perpendicular to the floor. No attempt was made to replicate the acoustic signal produced by the MR scanner but the subject did use ear plugs to create similar acoustic feedback conditions in terms of perceived vocal intensity. The speech sounds were recorded three separate times.

### C. Analysis by linear predictive coding (LPC)

The natural and simulated speech samples were compared in terms of linear prediction spectra. The LPC algorithm was a 50-pole autocorrelation method (Markel and Gray, 1976). For each recorded vowel sound, a sample approximately 0.10 s long was extracted from near the beginning of the total recorded production. An LPC analysis was then performed on this sample. Once the LPC spectrum had been computed, a peak picking algorithm based on a parabolic interpolation method (Titze *et al.*, 1987) was used to find the first three resonance frequencies of the vocal tract. The simulated sounds were subjected to the same analysis as the natural speech.

## III. RESULTS AND DISCUSSION

### A. Vowels

Sagittal views of the surface rendered airways, along with their ''raw'' area functions (i.e., points are not necessarily spaced in equal increments), are shown in Figs. 5, 6, 7, and 8 for the vowels produced by subject BS. Since the imaging was performed on static vocal tract shapes, the /ɝ/ and /l/ are included in the group of ''vowels;'' i.e., any open tract shape was considered to be a vowel. A rotated and tilted view for the /l/ is shown so that the lateral pathways are visible. Additionally, five coronal cross sections spaced at intervals of 0.188 cm show the shape of the lateral pathways. For each surface rendered airway, the most inferior point of the 3-D shape begins with the uppermost section of the trachea. Above the trachea, the airway becomes small in the region of the glottis and then widens, more or less depending on the vowel or consonant, into the lower pharyngeal section. The fingerlike extensions that hang down below the pharynx are the piriform sinuses. As the vocal tract bends into the horizontally oriented oral cavity the airway becomes narrow or wide and terminates with the mouth opening. For all area functions, the glottis is considered to be at the 0-cm point. The figures are grouped to show a progression of

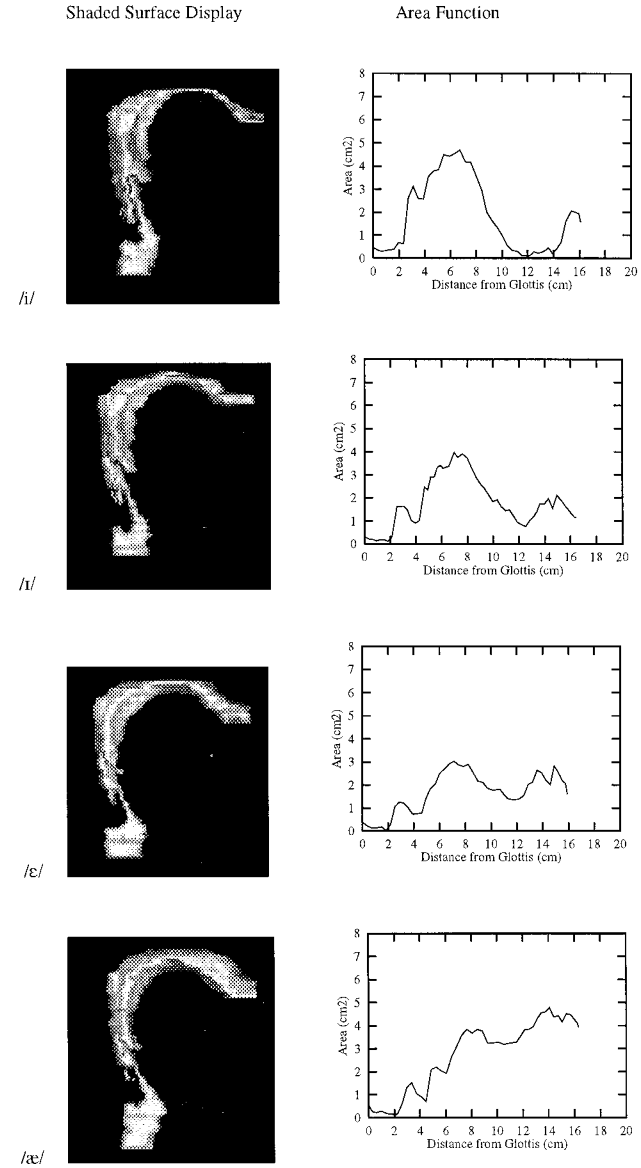


FIG. 5. Surface rendered airways and ''raw'' area functions for /i/, /ɪ/, /ɛ/, and /æ/ (subject BS).

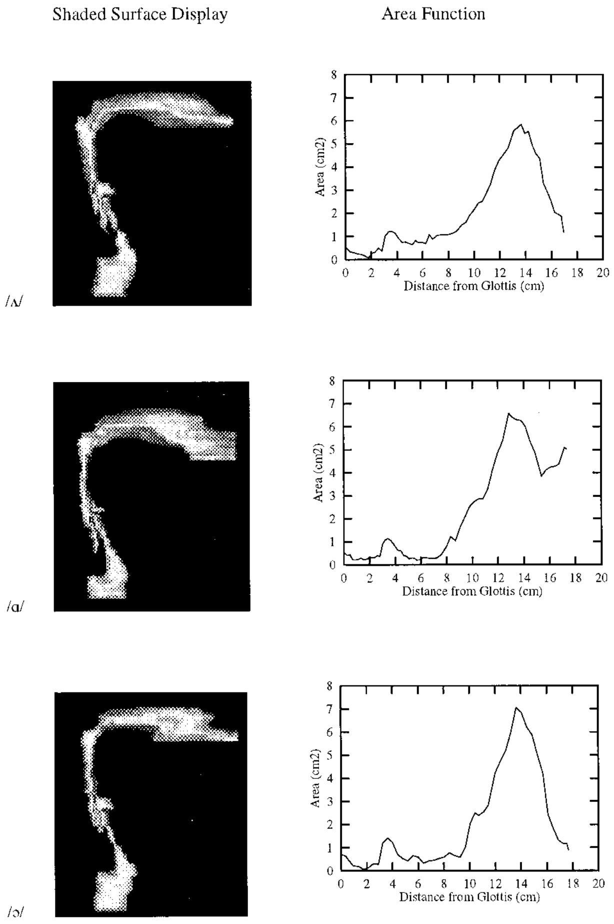


FIG. 6. Surface rendered airways and ''raw'' area functions for /ʌ/, /ɑ/, and /ɔ/ (subject BS).

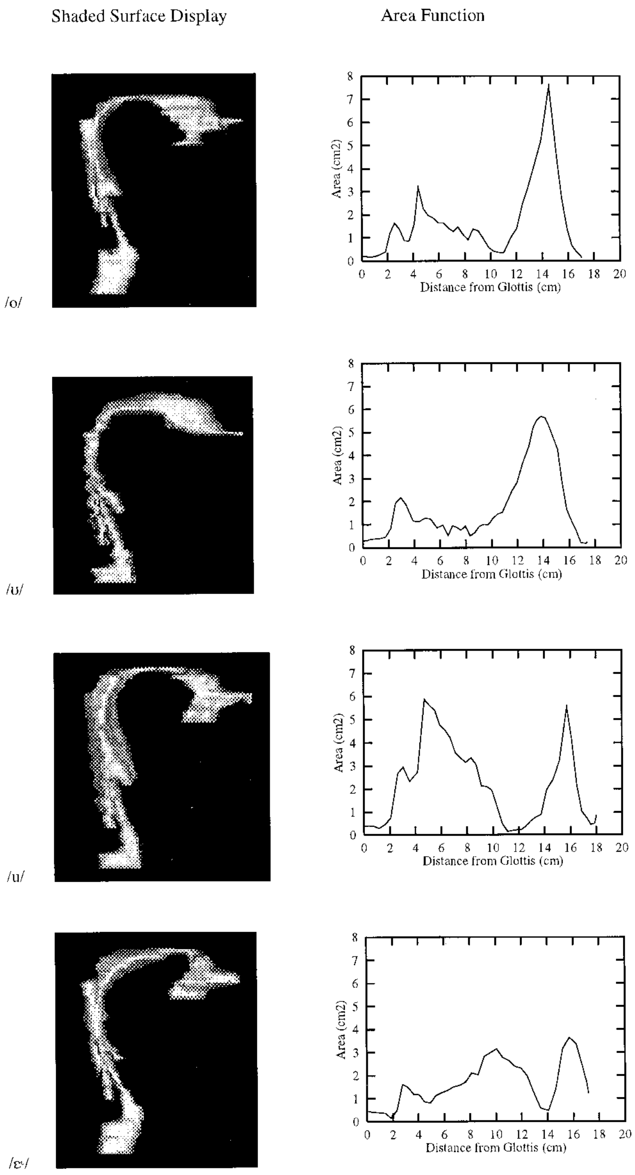


FIG. 7. Surface rendered airways and ''raw'' area functions for /o/, /ʊ/, /u/, and /ɝ/ (subject BS).

shapes that first exhibit a constricted oral section, then a constricted pharyngeal region, and finally a constricted mid-section of the tract.

A general observation with regard to all of the vocal tract shapes is that they show a widening of the tract above the glottis that starts at 2 to 3 cm and narrows again at approximately 4 to 5 cm. This is primarily due to the piriform sinuses merging with the main vocal tract tube. The point above the glottis at which the widening begins is nearly the same for all vowels, suggesting that the analysis of each vowel was consistent in terms of defining the glottal termination.

The first four airway surface renderings and area functions in Fig. 5 (/i/, /ɪ/, /ɛ/, and /æ/), show that the region just above the glottis is nearly constant up to 2 cm above the glottis. Beyond this point there is an abrupt increase in area, with the /i/ vowel achieving the greatest area of 4.7 cm$^2$. The other two ''front'' vowels, /ɪ/ and /ɛ/ have successively lower areas in this region (peak areas of 3.9 cm$^2$ and 3.0 cm$^2$, respectively). In the front half of the area function the cross-sectional area of the /i/ vowel drops to values on the order of 0.2 cm$^2$, far below the /ɪ/ and /ɛ/ areas which are in the range of 1.5 cm$^2$. Thus the /i/ vowel defines the extreme areas in both the front and back vocal tract cavities. The vowel /ɪ/ reaches less extreme areas than the /i/ but more extreme than /ɛ/. The area function for the /æ/ vowel is similar to /i/, /ɪ/, and /ɛ/ in the pharyngeal region and to /ʌ/, /ɑ/, and /ɔ/ (shown in Fig. 6) in the oral cavity. It appears to be a transition vowel between the front and back categories.

Surface renderings of the airways for /ʌ/, /ɑ/, and /ɔ/ (Fig. 6) all show a similar constricted pharynx and widened oral cavity, except that the /ɑ/ clearly has a larger mouth opening. The area functions for these ''back'' vowels are all very similar in the region between the glottis and about 4.5 cm above the glottis at which point the /ʌ/ remains at approximately 0.8 cm$^2$ and the /ɑ/ and /ɔ/ both drop to about 0.25 cm$^2$. All three vowels steadily increase in area from about the 7-cm point up to the 13.5-cm point with /ɔ/ achieving the greatest area of 7 cm$^2$. From the point of peak area out to the mouth termination (at approximately 17.5 cm) the area functions for /ʌ/ and /ɔ/ steadily decrease to a final mouth opening area of about 1.0 cm$^2$. The area function for

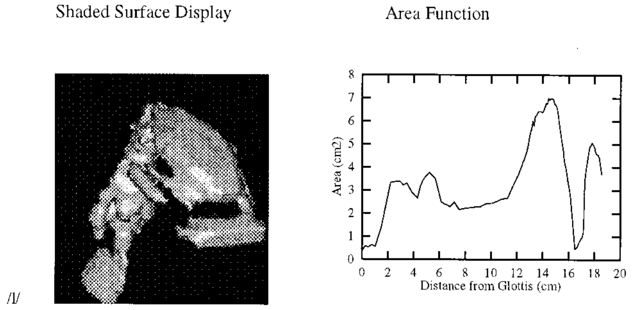


FIG. 8. Surface rendered airway and ''raw'' area function for /l/ (subject BS).

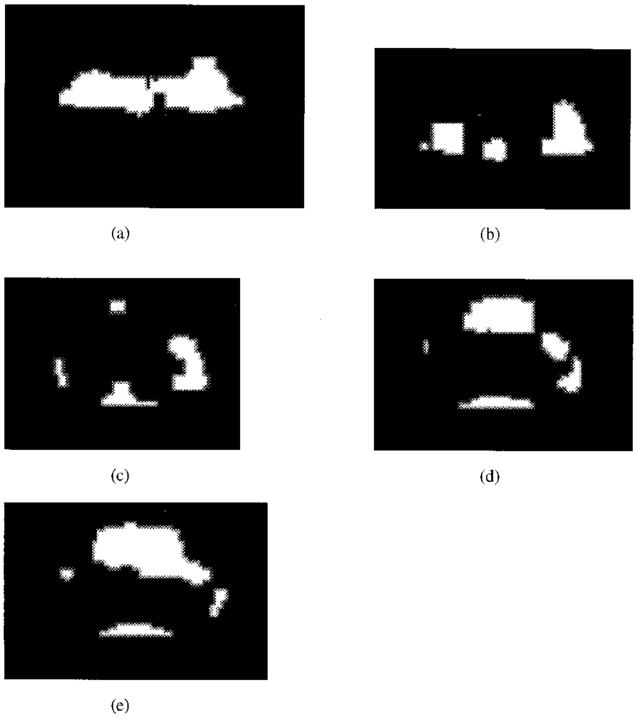


FIG. 9. A series of five coronal slices through a portion of the oral region of the interpolated vocal tract shape for /l/. The slices show the shape of the lateral air space. The interval between slices is 0.188 cm and the slices are shown from a view looking into the vocal tract from the mouth: (a) slice located approximately 1-cm posterior to the mouth termination plane, (b) airway is split into three regions; the right and left region define the lateral air spaces while the region in the lower center represents the small air space below the tongue tip, (c) a fourth region appears which is part of the main vocal tract airway, (d) pathway on the right side (subject's left) splits into two regions, (e) right pathway merges with the main airway—the lower part of the pathway on the right side of the figure, the entire pathway on the left side, and the air space under the tongue never reconnect with the main airway.

/ɑ/ initially exhibits a similar decrease, but at 15.5 cm the area begins to rise and continues out to a mouth termination of 5.0 $\mathrm{cm}^2$.

The airway images for /o/, /ʊ/, /u/ and /ɝ/ in Fig. 7 clearly show the profile of the tongue and especially the tongue tip. The area functions demonstrate how each of these vowels are more or less divided into two distinct chambers by a tight constriction. The /o/ and /ʊ/ both have large cavities in the front of the tract and smaller cavities in the back. However, the constriction in the /o/ is 2 cm closer to the mouth and more well-defined than for the /ʊ/. The front and back cavities of /u/ both have cross-sectional areas that reach about 6.0 $\mathrm{cm}^2$, providing a more equal front–back distribution of cavity size than the other three vowels. The /ɝ/ shape also exhibits a nearly equal front–back cavity distribution, but the largest cross-sectional areas are on the order of 3.5 $\mathrm{cm}^2$. In addition, the point of constriction occurs only 3.0-cm posterior to the mouth, which is closer than for any of the three vowels /o/, /ʊ/, and /u/.

The /l/ is shown in Fig. 8 from a rotated and tilted perspective to give an indication of the lateral air flow (and acoustic wave) path around the tongue. The gap between the oral section and the lip section is due to the presence of the tongue contacting the hard palate. Ideally a lateral path would exist on both sides of the oral cavity but the subject imaged in this experiment closed off the right pathway (right refers to the subject's right). The area function for the /l/ shows a generally uniform area in the range of 2–3 $\mathrm{cm}^2$, throughout the pharyngeal region, but a rapid area increase up to 7 $\mathrm{cm}^2$ just prior to the lateral pathway constriction. Immediately following the constriction, the area again increases rapidly with a final mouth termination area of 3.7 $\mathrm{cm}^2$.

To gain a better view of the lateral pathways, a series of five coronal slices from the interpolated vocal tract shape is shown in Fig. 9. The slices are spaced at intervals of 0.188 cm. The series begins with a section located approximately 1-cm posterior to the mouth termination plane [Fig. 9(a)] and the view is from the perspective of looking into the vocal tract from the anterior end (mouth); i.e., the right side of each figure is the subject's left. At this point the airway is a contiguous region. Figure 9(b) shows the airway split into three regions; the right and left region define the lateral air spaces while the region in the lower center represents the small air space below the tongue tip. A fourth region appears in the superior portion of Fig. 9(c) which is part of the main vocal tract airway. Figure 9(d) shows the pathway on the right side (subject's left) split into two regions. Figure 9(e) shows the merging of the right pathway with the main airway. Note that the lower part of the pathway on the right side of the figure, the entire pathway on the left side, and the air space under the tongue never reconnect with the main airway. Thus the constricted portion of the /l/ area function at approximately 17 cm from the glottis was based on measurements of only the lateral pathway that did reconnect with the main airway.

Two of the area functions shown in this section, /ɛ/ and /ɔ/, contain an extremely small area, less than 0.1 $\mathrm{cm}^2$, located at about 2 cm from the glottis (see Table III for numerical values). It is interesting that the location of this small area is almost exactly the same for the two vowels. A close check of the original image slices indicated that the airway was, in fact, very constricted in this region of the vocal tract. However, for an area as small as these, only a few voxels define the airway; e.g., an area of 0.06 $\mathrm{cm}^2$ would be comprised of only 7 voxels [$0.06\ \mathrm{cm}^2/(0.0938\ \mathrm{cm})^2 \approx 7$]. A slight increase in the threshold value for the seeded region growing

TABLE III. Equal interval (0.396825 cm) area functions for 18 vocal tract shapes. Section 1 is the glottal end of the vocal tract and ''n.c.'' denotes the nasal coupling.

| Sect. no. | /i/ | /ɪ/ | /ɛ/ | /æ/ | /ʌ/ | /ɑ/ | /ɔ/ | /o/ | /ʊ/ | /u/ | /ɝ/ | /l/ | /m/ | /n/ | /ŋ/ | /p/ | /t/ | /k/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.33 | 0.20 | 0.21 | 0.22 | 0.33 | 0.45 | 0.61 | 0.18 | 0.32 | 0.40 | 0.41 | 0.55 | 0.57 | 0.26 | 0.51 | 0.31 | 0.38 | 0.34 |
| 2 | 0.30 | 0.17 | 0.13 | 0.26 | 0.28 | 0.20 | 0.28 | 0.17 | 0.39 | 0.38 | 0.38 | 0.63 | 0.57 | 0.24 | 0.47 | 0.39 | 0.50 | 0.35 |
| 3 | 0.36 | 0.18 | 0.16 | 0.26 | 0.23 | 0.26 | 0.19 | 0.23 | 0.39 | 0.28 | 0.40 | 0.75 | 0.21 | 0.17 | 0.45 | 0.42 | 0.40 | 0.49 |
| 4 | 0.34 | 0.18 | 0.14 | 0.16 | 0.15 | 0.21 | 0.10 | 0.28 | 0.43 | 0.43 | 0.29 | 1.80 | 0.58 | 0.21 | 0.30 | 0.71 | 1.07 | 0.78 |
| 5 | 0.68 | 0.10 | 0.06 | 0.13 | 0.17 | 0.32 | 0.07 | 0.59 | 0.56 | 0.55 | 0.13 | 2.98 | 2.18 | 0.15 | 0.48 | 1.28 | 1.38 | 1.31 |
| 6 | 0.50 | 1.08 | 0.78 | 0.21 | 0.33 | 0.30 | 0.30 | 1.46 | 1.46 | 1.72 | 0.53 | 3.56 | 3.15 | 0.36 | 0.67 | 1.80 | 1.65 | 1.34 |
| 7 | 2.43 | 1.66 | 1.25 | 0.83 | 0.39 | 0.33 | 0.18 | 1.60 | 2.20 | 2.91 | 1.58 | 3.45 | 2.96 | 1.37 | 0.83 | 1.70 | 1.29 | 1.19 |
| 8 | 3.15 | 1.64 | 1.24 | 1.50 | 1.02 | 1.05 | 1.13 | 1.11 | 2.06 | 2.88 | 1.56 | 3.22 | 2.89 | 1.66 | 0.96 | 1.43 | 1.01 | 0.94 |
| 9 | 2.66 | 1.19 | 0.99 | 1.35 | 1.22 | 1.12 | 1.42 | 0.82 | 1.58 | 2.37 | 1.22 | 3.20 | 3.70 | 1.35 | 1.43 | 1.25 | 0.92 | 0.69 |
| 10 | 2.49 | 0.92 | 0.72 | 0.99 | 1.14 | 0.85 | 1.21 | 1.01 | 1.11 | 2.10 | 1.19 | 2.67 | 4.21 | 0.90 | 1.14 | 0.90 | 0.86 | 0.92 |
| 11 | 3.39 | 1.13 | 0.73 | 0.69 | 0.82 | 0.63 | 0.69 | 2.72 | 1.11 | 3.63 | 1.00 | 3.02 | 3.57 | 0.71 | 0.84 | 2.06 | 1.03 | 1.45 |
| 12 | 3.80 | 2.48 | 1.06 | 1.35 | 0.76 | 0.39 | 0.51 | 2.71 | 1.26 | 5.86 | 0.77 | 3.55 | 3.59 | 0.93 | 0.69 | 2.77 | 1.60 | 1.73 |
| 13 | 3.78 | 2.76 | 1.77 | 2.32 | 0.66 | 0.26 | 0.43 | 1.96 | 1.30 | 5.63 | 0.92 | 3.76 | 2.97 | 1.41 | 0.82 | 2.19 | 2.46 | 1.67 |
| 14 | 4.35 | 2.97 | 1.97 | 2.13 | 0.80 | 0.28 | 0.66 | 1.92 | 0.98 | 5.43 | 1.19 | 3.53 | 3.17 | 2.07 | 0.86 | 2.35 | 2.24 | 2.13 |
| 15 | 4.50 | 3.43 | 2.46 | 1.94 | 0.72 | 0.23 | 0.57 | 1.70 | 0.93 | 4.80 | 1.27 | 2.62 | 3.25 | 2.12 | 0.57 | 2.67 | 2.47 | 1.61 |
| 16 | 4.43 | 3.32 | 2.70 | 2.17 | 0.66 | 0.32 | 0.32 | 1.66 | 0.83 | 4.56 | 1.35 | 2.40 | 2.58 | 2.04 | 0.81 | 2.17 | 2.86 | 1.56 |
| 17 | 4.68 | 3.48 | 2.92 | 2.85 | 1.08 | 0.29 | 0.43 | 1.52 | 0.61 | 4.29 | 1.48 | 2.32 | 2.74 | 2.16 | 1.00 | 1.77 | 2.74 | 1.54 |
| 18 | 4.52 | 3.96 | 3.03 | 3.26 | 0.91 | 0.28 | 0.45 | 1.28 | 0.97 | 3.63 | 1.56 | 2.43 | 2.77 | 2.36 | 0.66 | 2.09 | 3.32 | 1.18 |
| 19 | 4.15 | 3.79 | 2.84 | 3.73 | 1.09 | 0.40 | 0.53 | 1.44 | 0.75 | 3.37 | 1.61 | 2.13 | 2.49 | 2.52 | 0.80 | 2.16 | 3.83 | 1.44 |
| 20 | 4.09 | 3.88 | 2.84 | 3.80 | 1.06 | 0.66 | 0.60 | 1.28 | 0.93 | 3.16 | 1.87 | 2.27 | 2.93 | 2.88 | 0.97 | 2.26 | 3.97 | 1.12 |
| 21 | 3.51 | 3.47 | 2.83 | 3.69 | 1.09 | 1.20 | 0.77 | 0.89 | 0.53 | 3.31 | 2.10 | 2.28 | 3.33 | 2.30 | 0.78 | 2.26 | 4.16 | 0.76 |
| 22 | 2.95 | 2.98 | 2.36 | 3.87 | 1.17 | 1.05 | 0.65 | 1.25 | 0.65 | 3.22 | 2.01 | 2.26 | 2.27 | 1.93 | 0.58 | 2.29 | 4.41 | 0.96 |
| 23 | 2.03 | 2.62 | 2.14 | 3.68 | 1.39 | 1.62 | 0.58 | 1.38 | 0.95 | 2.33 | 2.62 | 2.33 | 2.57 | 1.77 | 0.46 | 2.17 | 4.11 | 1.09 |
| 24 | 1.66 | 2.37 | 2.00 | 3.20 | 1.55 | 2.09 | 0.94 | 1.09 | 0.99 | 2.07 | 2.96 | 2.43 | 2.17 | 0.96 | 0.44 | 2.13 | 3.95 | 0.79 |
| 25 | 1.38 | 1.99 | 1.78 | 3.26 | 1.89 | 2.56 | 2.02 | 0.71 | 1.07 | 2.07 | 3.07 | 2.44 | 1.84 | 0.89 | 0.47 | 2.64 | 3.64 | 0.25 |
| 26 | 1.05 | 1.90 | 1.81 | 3.29 | 2.17 | 2.78 | 2.50 | 0.46 | 1.39 | 1.52 | 3.11 | 2.54 | 1.98 | 1.22 | 0.41 | 2.65 | 3.37 | 0.00 |
| 27 | 0.60 | 1.70 | 1.79 | 3.19 | 2.46 | 2.86 | 2.41 | 0.39 | 1.47 | 0.74 | 2.77 | 2.64 | 1.73 | 1.30 | 0.11 | 2.30 | 2.89 | 0.06 |
| 28 | 0.35 | 1.44 | 1.50 | 3.23 | 2.65 | 3.02 | 2.62 | 0.32 | 1.79 | 0.23 | 2.67 | 2.67 | 1.43 | 1.30 | 0.00 | 2.12 | 2.61 | 0.03 |
| 29 | 0.32 | 1.45 | 1.37 | 3.23 | 3.13 | 3.75 | 3.29 | 0.57 | 2.34 | 0.15 | 2.47 | 3.16 | 1.73 | 1.14 | 0.00 | 1.67 | 2.69 | 0.09 |
| 30 | 0.12 | 1.06 | 1.36 | 3.40 | 3.81 | 4.60 | 4.34 | 1.06 | 2.68 | 0.22 | 2.34 | 3.68 | 2.08 | 0.77 | 0.00 | 1.44 | 2.32 | 0.10 |
| 31 | 0.10 | 0.87 | 1.43 | 3.78 | 4.30 | 5.09 | 4.78 | 1.38 | 3.36 | 0.22 | 2.25 | 4.30 | 2.32 | 0.34 | 0.00 | 1.16 | 2.04 | 0.06 |
| 32 | 0.16 | 0.75 | 1.83 | 3.84 | 4.57 | 6.02 | 5.24 | 2.29 | 3.98 | 0.37 | 1.90 | 5.14 | 2.84 | 0.15 | 0.00 | 1.51 | 1.64 | 0.03 |
| 33 | 0.25 | 1.06 | 2.08 | 3.98 | 4.94 | 6.55 | 6.07 | 2.99 | 4.74 | 0.60 | 1.32 | 5.83 | 3.51 | 0.22 | 0.00 | 1.76 | 1.39 | 0.48 |
| 34 | 0.24 | 1.29 | 2.59 | 4.41 | 5.58 | 6.29 | 7.08 | 3.74 | 5.48 | 0.76 | 0.76 | 6.44 | 4.25 | 0.21 | 0.00 | 1.93 | 1.26 | 1.27 |
| 35 | 0.38 | 1.78 | 2.54 | 4.56 | 5.79 | 6.27 | 6.81 | 4.39 | 5.69 | 0.86 | 0.44 | 6.54 | 4.79 | 0.00 | 2.18 | 1.98 | 0.87 | 2.28 |
| 36 | 0.28 | 1.83 | 2.11 | 4.79 | 5.51 | 5.94 | 6.20 | 5.38 | 5.57 | 1.82 | 0.45 | 6.91 | 4.61 | 0.00 | 4.72 | 2.21 | 0.60 | 2.35 |
| 37 | 0.36 | 1.70 | 2.34 | 4.39 | 5.49 | 5.28 | 5.89 | 7.25 | 4.99 | 2.35 | 0.92 | 6.72 | 4.07 | 0.00 | 6.86 | 2.35 | 0.10 | 2.40 |
| 38 | 0.65 | 1.97 | 2.74 | 4.42 | 4.69 | 4.70 | 5.04 | 7.00 | 4.48 | 2.55 | 2.05 | 5.61 | 3.64 | 0.00 | 8.58 | 2.45 | 0.00 | 2.41 |
| 39 | 1.58 | 1.92 | 2.19 | 4.23 | 4.50 | 3.87 | 4.29 | 4.57 | 3.07 | 3.73 | 3.25 | 4.08 | 2.84 | 0.00 | 8.76 | 2.37 | 0.00 | 4.21 |
| 40 | 2.05 | 1.62 | 1.60 | 4.56 | 3.21 | 4.13 | 2.49 | 2.75 | 1.67 | 5.47 | 3.63 | 2.73 | 1.42 | 0.00 | 7.21 | 2.47 | 0.13 | 3.37 |
| 41 | 2.01 | 1.36 |  | 4.31 | 2.79 | 4.25 | 1.84 | 1.48 | 1.13 | 4.46 | 3.59 | 0.45 | 0.29 | 1.59 | 4.92 | 1.75 | 0.18 | 2.46 |
| 42 | 1.58 | 1.18 |  | 3.94 | 2.11 | 4.27 | 1.33 | 0.68 | 0.64 | 2.39 | 3.07 | 0.90 | 0.00 | 1.60 | 3.21 | 1.09 | 1.48 | 2.46 |
| 43 |  |  |  |  | 1.98 | 4.69 | 1.19 | 0.39 | 0.15 | 1.10 | 2.25 | 3.92 | 0.00 | 1.69 | 2.17 | 0.70 | 1.60 | 2.14 |
| 44 |  |  |  |  | 1.17 | 5.03 | 0.88 | 0.14 | 0.22 | 0.77 | 1.20 | 4.99 | 0.00 | 1.17 | 1.41 | 0.00 | 1.43 | 1.50 |
| 45 |  |  |  |  |  |  |  |  |  | 0.41 |  | 4.57 |  |  |  |  |  |  |
| 46 |  |  |  |  |  |  |  |  |  | 0.86 |  | 3.70 |  |  |  |  |  |  |
| n.c. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.04 | 1.26 | 1.09 | 0.00 | 0.00 | 0.00 |
| VT length | 16.67 | 16.67 | 15.88 | 16.67 | 17.46 | 17.46 | 17.46 | 17.46 | 17.46 | 18.25 | 17.46 | 18.25 | 17.46 | 17.46 | 17.46 | 17.46 | 17.46 | 17.46 |

algorithm (see Sec. I C) could have easily included a few more voxels, which would increase the measured cross-sectional area. However, it was decided not to adjust the threshold value for segmentation of these small areas, but maintain a consistent threshold throughout the vocal tract. A potential problem with imaging these vowels is that they are typically not sustained for long periods of time in normal speech (i.e., short vowels), and as a result may have been more susceptible to movement artifact and blurring than the point vowels /i, æ, ɑ, u/ which are often sustained. While such small areas may be due to measurement error or movement artifact, it is shown later in Sec. III D that these two vowels produced very close acoustic matches to natural speech in terms of locations for formants $F1$ and $F2$.

### B. Nasals and plosives

Nasal and plosive tract shapes are included in this section because they both possess occlusions of the vocal tract. In fact, the occlusions should be analogous for /m/ and /p/, /n/ and /t/ and /ŋ/ and /k/. Since the imaging protocol allows only static shapes to be acquired, no distinction between

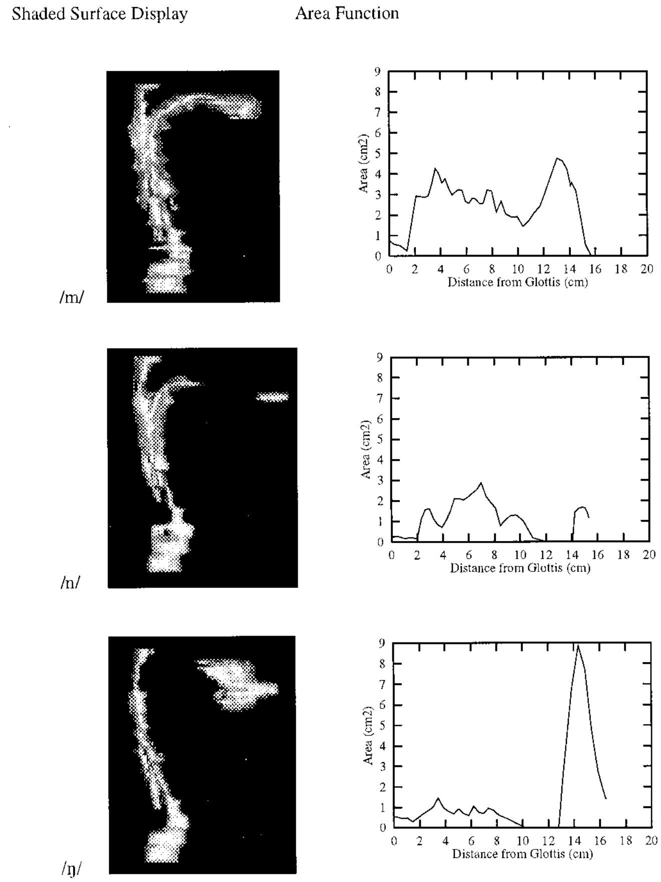


FIG. 10. Surface rendered shapes and area functions for nasals (subject BS).

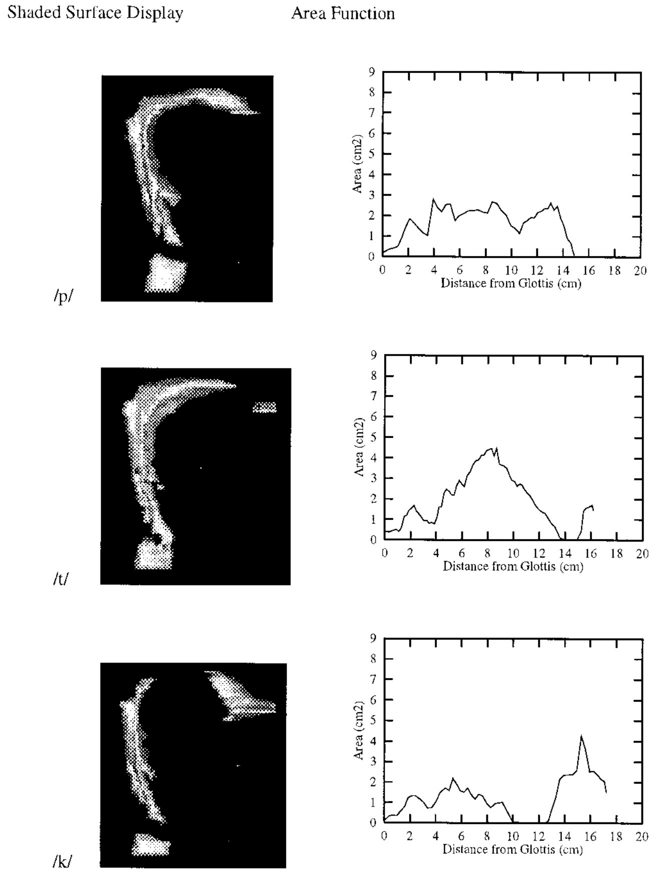


FIG. 11. Surface rendered shapes and area functions for plosives (subject BS).

voiced and unvoiced plosives can be made; i.e., the shape for a /p/ is considered to be equivalently useable as a /b/, /t/ as /d/, etc. The airway surface renderings and area functions for each of the three nasal and plosive consonants are shown in Figs. 10 and 11, respectively. The significant feature in each of these shapes is the location of the main vocal tract occlusion, and in the case of the nasals, the presence of the open velopharyngeal port through which air flow and acoustic waves are diverted. The tract occlusion for the /m/ and /p/ occurs at the lips so that a ‘‘break’’ in the vocal tract shape is not observed other than an apparent shortening relative to the previously shown vowel shapes. Since the mouth is closed, the only outlet for sound during production of /m/ is through the nasal cavity, which means that the entire oral cavity becomes a side branch resonator. Area functions for the /m/ and /p/ demonstrate a generally uniform vocal tract shape with cross-sectional areas in the range of 2–3 $\text{cm}^2$ in the pharyngeal region, followed by a slight narrowing to about 1.5 $\text{cm}^2$ and subsequent expansion in the oral section to 4.5 $\text{cm}^2$ for /m/ and 2.5 $\text{cm}^2$ for /p/ before closing down to zero area. This particular shape is probably due to the requirement that the subject produce the /m/ while imagining that the preceding and succeeding vowels would be a schwa (neutral) vowel. Had an /ɑ/ or /i/ been used as the imagined ‘‘target,’’ the area function might have shown a closer resemblance to those vowels. Simulation of dynamic speech segments may require different area functions for the /m/ and /p/, depending upon the surrounding vowel shapes. The most important information to take from the /m/ images is the area and location of the nasal coupling port. The port area was measured to be 1.04 $\text{cm}^2$ and is located approximately 8 cm from the glottis.

The /n/ and /t/ shapes exhibit a break in the vocal tract where the tongue contacts the hard palate and buccal walls. For the /n/, a small side branch cavity is created by the airspace between the location of the nasal coupling (8 cm from the glottis) and the constriction. The cavity of air near the mouth termination was included only to demonstrate the ‘‘break’’ in the tract. Acoustically speaking, this volume of air is inactive for this particular shape. Area functions for the /n/ and /t/ both show a widened upper pharyngeal region with an abrupt closure in the oral cavity, with the /t/ requiring more volume than the /n/. The vocal tract is also about a centimeter longer for the /t/ than for the /n/, which places the point of constriction at a more anterior location. The nasal coupling area was determined to be 1.09 $\text{cm}^2$ and located 8 cm from the glottis.

Surface rendered airways for the /ŋ/ and /k/ are nearly identical except for the open velopharyngeal passage in the /ŋ/. The occlusion of the tract for /ŋ/ occurs at a point that is posterior enough that there is just a very small side branch cavity created. The 3-D image shows an almost straight tract from the glottis up into the velopharyngeal tube. The cou-

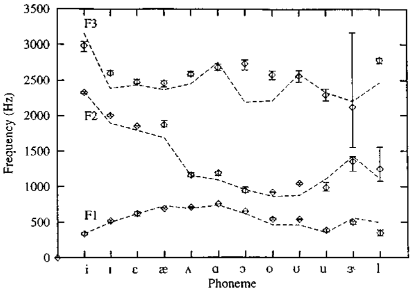


FIG. 12. Comparison of formant locations for $F1$, $F2$, and $F3$ extracted from LPC spectra for both natural and simulated speech. The mean value for three separate recordings sessions of natural speech is depicted by the diamond symbols and the error bars indicate the range. The dashed line represents the simulation.

pling area was found to be 1.26 cm$^2$ and again located 8 cm from the glottis. Like the airway renderings, the area functions show the similarity of the two consonants. The cavities in front of the occlusion are quite different, but since this section of the tract is acoustically inactive, it is not a significant feature.

### C. Numerical area functions

The linear distance between consecutive oblique cross sections determined by Eq. (1) is not necessarily constant throughout the vocal tract. The area functions produced by Eq. (2) contained 50 to 80 data points, depending on the number of iterations that were performed during the analysis and were spaced at intervals that ranged from 0.2 to 0.4 cm. The wave-reflection algorithm discussed in Sec. II A requires that each area function be discretized at equal length intervals and the length of the final area function must be an even integer multiple of the length interval. To transform the ''raw'' area functions determined by Eq. (2) into a usable form for the speech simulation, they were first normalized to a discrete length that was close to their measured length. A wave-reflection-type model dictates that the length of each finite section of the area function be equal to the speed of sound divided by two times the sampling frequency. For this study, the length interval was chosen to be 0.396825 cm which results from using a sampling frequency of 44.1 kHz and a speed of sound equal to 350 m/s. If the measured length of the tract was 17.2 cm, a 44 section area function would be chosen to represent it since 17.46 cm is the closest discrete length. The raw area function would first be normalized to 17.46 cm and then transformed to an equal interval function by fitting it with a cubic spline and then sampling the resulting curve at equally spaced intervals. If desired, all of the raw area functions could be normalized to one particular length so that all of the discretized functions would have, for example, 44 sections. The effect would be a slight compression of the longer area functions and an expansion of the shorter area functions.

Equal interval (0.396825 cm) area functions are shown in Table III and should be read by assuming that the glottal end of the tract is represented by section 1 while the last section for each phoneme represents the mouth termination. The second row from the bottom of the table, labeled ''n.c.'', is the nasal coupling area, which is zero for all shapes except the nasal consonants. The last row indicates the vocal tract length for each area function.

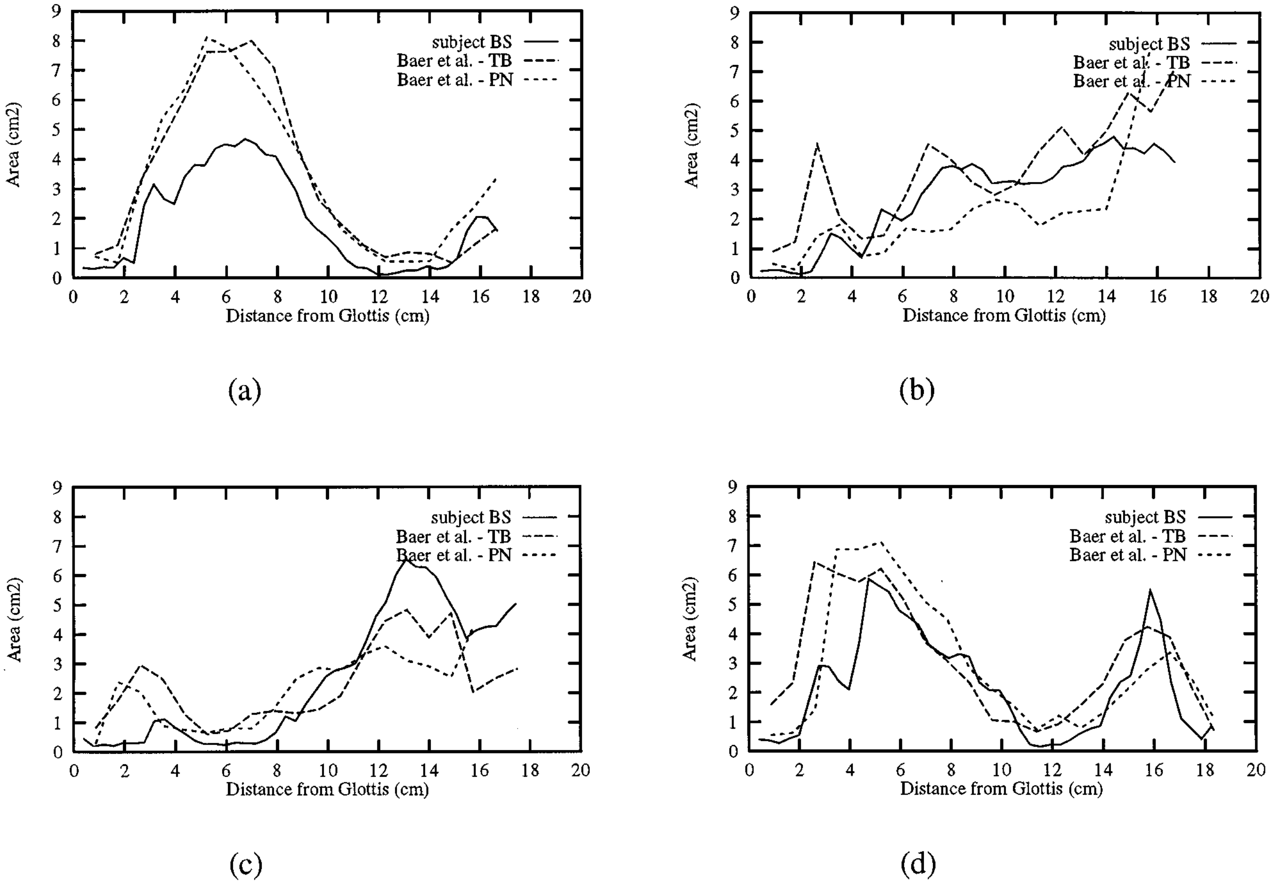
### D. Comparison of natural and simulated vowel sounds

A comparison of the first three formant locations extracted from LPC spectra for the natural and simulated versions of each vowel is shown in Fig. 12. The natural speech sounds were recorded three separate times, so that each diamond shaped symbol represents the mean formant values of the three sets while the error bars indicate the range. The dashed line passes through the simulated values.

The first formants extracted from the simulations show, eight out of the twelve vowels falling within the range of the natural speech, while the remaining four vowels were positioned outside the range. For five of the vowels, the second formants from the simulations fell within the natural speech range while the other seven fell outside the range. Similarly, the third formants show five vowels positioned within the range of the natural speech, but significant deviations from natural speech are observed for /ɔ/, /o/, /ɝ/, and /l/. The large range shown for the third formant of /ɝ/ was an artifact caused by the second and third formants merging in one of

TABLE IV. First three formants from natural recorded speech and simulated speech based on the area functions given in Table III. The superscript ''N'' denotes the natural speech and ''S'' the simulated version. The Δ's represent the percent error of the formants from simulated speech relative to the mean value of the natural speech formants.

| | i | ɪ | ɛ | æ | ʌ | ɑ | ɔ | o | ʊ | u | ɝ | l |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $F1^N$ | 333 | 518 | 624 | 692 | 707 | 754 | 654 | 540 | 541 | 389 | 500 | 348 |
| $F2^N$ | 2332 | 2004 | 1853 | 1873 | 1161 | 1195 | 944 | 922 | 1045 | 987 | 1357 | 1250 |
| $F3^N$ | 2986 | 2605 | 2475 | 2463 | 2591 | 2685 | 2739 | 2584 | 2568 | 2299 | 2124 | 2785 |
| $F1^S$ | 337 | 499 | 621 | 732 | 689 | 738 | 618 | 461 | 461 | 356 | 559 | 500 |
| $F2^S$ | 2340 | 1894 | 1795 | 1689 | 1159 | 1093 | 958 | 861 | 877 | 1108 | 1431 | 1127 |
| $F3^S$ | 3158 | 2388 | 2436 | 2370 | 2454 | 2757 | 2195 | 2217 | 2596 | 2334 | 2206 | 2574 |
| Δ1 | 1.3 | 3.7 | 0.4 | 5.8 | 2.5 | 2.1 | 5.5 | 14.6 | 14.7 | 8.5 | 11.7 | 43.6 |
| Δ2 | 0.4 | 5.5 | 3.1 | 9.8 | 0.2 | 8.5 | 1.4 | 6.6 | 16.2 | 12.3 | 5.5 | 9.8 |
| Δ3 | 5.76 | 8.3 | 1.6 | 3.8 | 5.3 | 2.7 | 19.9 | 14.2 | 1.1 | 1.5 | 3.9 | 7.6 |



(a)

(b)

(c)

(d)

FIG. 13. Comparison of the area functions for subject BS (solid) with Baer *et al.* (1991) area functions for subjects TB (long dashed) and PN (short dashed): (a) /i/, (b) /æ/, (c) /ɑ/, (d) /u/.

the recorded cases so that formant peak picking algorithm skipped to the next formant and reported it as the third. Simply by chance, the mean value of the third formant for /l/ deviated from the simulated version by only 11%. The true deviation in this case was much greater.

The data presented in Fig. 12 is tabulated in Table IV, along with the percentage error for the simulations relative to the mean formant values of the natural cases. Across all vowels and formants, percentage errors range from 0.20% to 43.6% with the majority below 10%. The largest error of 43.6% occurred in the first formant for /l/. This is not surprising considering that the oral cavity for /l/, with the lateral air spaces, was far more complex than any of the other shapes, thus more susceptible to measurement error. In fact, one could argue that an /l/ cannot be adequately described by a simple area function implemented in a one-dimensional acoustic model. However, the second and third formants both were in error by less than ten percent so that the measured area function should not be entirely disregarded. The simulation of the two vowels, /ɛ/ and /ɔ/, both produced formant locations for $F1$ and $F2$ which deviated from the natural speech by less that 6%. The presence of a very small cross-sectional area in these two area functions at approximately 2 cm from the glottis was discussed in Sec. III B.

A study of Fig. 12 reveals that, for many of the simulated vowels, the second and third formants tend to be displaced from the natural speech in the direction of a neutral (schwa) vowel formant structure (i.e., $F1$, $F2$, $F3 \approx 500$, 1500, 2500 Hz). For example, the second formant for the simulated /æ/ is displaced from the natural speech mean value of 1873 Hz down to 1689 Hz; movement toward a more neutral 1500 Hz. In addition, the third formants for /ɪ/, /ɛ/, /æ/, /ʌ/, /ɔ/, /o/, and /l/ are all displaced from natural speech in the direction of 2500 Hz. This suggests that the subject (BS) tended to centralize the production of vowels during the MR imaging sessions. The vowel centralization could be due to fatigue of the articulatory musculature as well as listening fatigue of the aural system. Fatigue effects are quite possible, since the MRI protocol required the subject to produce many repetitions of a given vowel (approx. 30 repetitions for each vocal tract shape). However, another observation with respect to Fig. 12 is that neither the simulated /i/ or /ɑ/ vowels show centralizing effects. In fact, the third formant of /i/ and the second formant of /ɑ/ are displaced in the opposite direction of a neutral vowel. It is possible that the intermediate (short) vowels may require more precise muscular control to maintain the appropriate formant structure than the extreme /i/ and /ɑ/ shapes; /i/ and /ɑ/ are ''asymptotic'' positions that have physiologically imposed constraints or boundaries. In general, the MRI based area functions may be somewhat centralized and should be considered to be an average shape over many productions of the same vowel.

### E. Comparisons of area functions with previous imaging studies

At this point it is of interest to compare the area functions reported in the present study with those given in several previous publications. In the following discussion, whenever reference is made to ''subject BS'' the reader should assume this to mean the area function(s) from the present study.

Figure 13 shows a graphical representation of the discretized area functions (given in Table III) for the vowels /i, æ, ɑ, u/ superimposed with the area functions reported by Baer *et al.* (1991) for the same vowels (two subjects). To be analogous to the area functions in the present study, the Baer *et al.* area functions have been plotted *without* the contributions of piriform sinuses as indicated in Table I of their paper (p. 811). The area functions for each vowel show the same general shapes but demonstrate obvious individual differences. The /i/ vowel in Fig. 13(a) for subject BS is smaller than the Baer *et al.* area functions in every region of the vocal tract except close to mouth termination where it is slightly larger than that for subject TB. This characteristic might suggest that the areas in the present study were underestimated. But the /æ/ area function [Fig. 13(b)] is larger throughout nearly all of the tract than that for subject PN and mostly smaller than the same vowel for subject TB. In the /ɑ/ area functions [Fig. 13(c)] the BS version is smaller than both TB and PN in the pharyngeal region but larger than both in the oral cavity and at the mouth termination. The /u/ [Fig. 13(d)] is similar for all three subjects except that the BS version has a tighter constriction in the middle part of the tract. In general, across all four vowels, the constricted parts of vocal tract are smaller in the BS area functions. In particular, the region just above the larynx (0 to 3 cm) seems to always have smaller cross-sectional area than the Baer *et al.* area functions. The reasons for this could be an underestimation of the areas due to the image processing techniques that were used or it could simply be an individual anatomical difference. However, it should be noted that, based on the height and weight information given in Sec. I A, the subject BS is not a large person. Thus small cross-sectional areas in the vocal tract may not be entirely unexpected.

Figure 14 shows the discretized area functions for the vowels /i, ɑ, u/ along with the area functions given in Yang and Kasuya (1994) for an adult male subject. Again the general shape for each vowel is similar, as would be expected, but individual differences are apparent. The glottal end (0 cm) of each of the Yang and Kasuya area functions has an area 0.5 to 1.0 cm$^2$ larger than the corresponding BS version. The major constrictions are again smaller for the BS area functions but interestingly the front and back chambers for the /u/ vowel are on the order of 2 cm$^2$ larger than for the Yang and Kasuya version. This was not the case when comparisons were made with the Baer *et al.* area functions. One interesting aspect of the Yang and Kasuya data is that the /i/ vowel has a longer vocal tract length than either the /ɑ/ or the /u/. In the present study and the Baer *et al.* paper, the /u/ vowel has the longest vocal tract length.

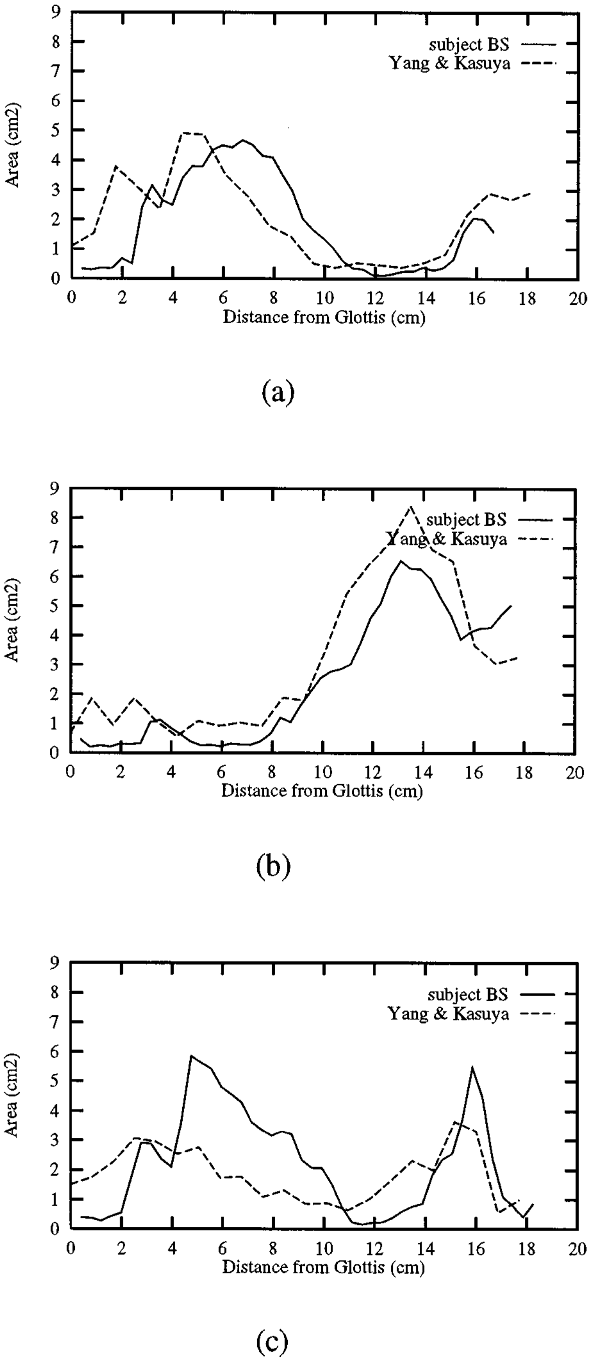
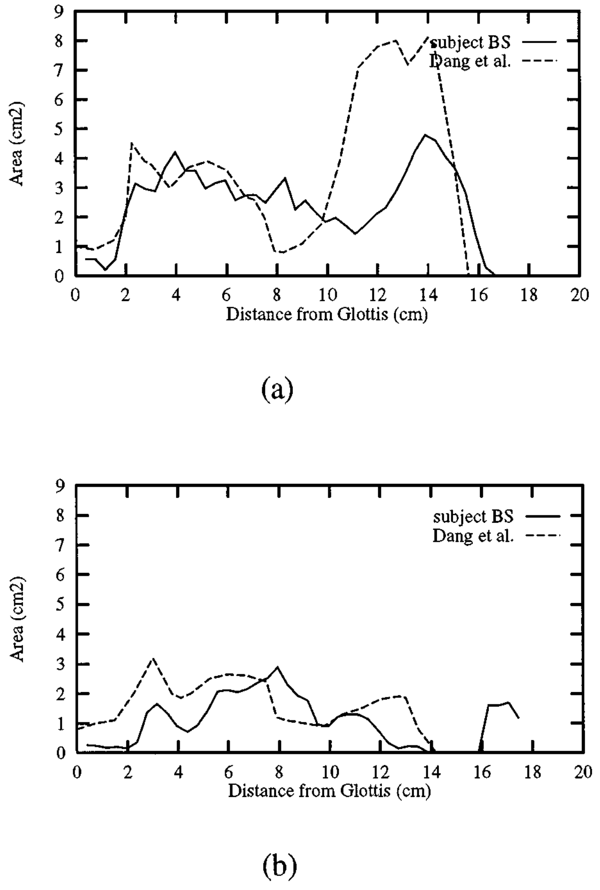


FIG. 14. Comparison of the area functions for subject BS (solid) with Yang and Kasuya (1994) (dashed): (a) /i/, (b) /ɑ/, (c) /u/.

In addition to an extensive analysis of the nasal tract, Dang *et al.*, 1994 also presented vocal tract area functions for the nasal consonants /m/ and /n/. The numerical values for these area functions have been estimated from their Fig. 13 (p. 2097) and are shown with the /m/ and /n/ area functions for subject BS in Fig. 15. The /m/ [Fig. 15(a)] for both subject BS and the Dang *et al.* study exhibit a back chamber in the pharynx with a cross-sectional area of about 3.5 cm$^2$. This region extends from a point approximately 2 cm from the glottis out to the 8-cm point in the Dang *et al.* area function while the same region would be roughly defined to be



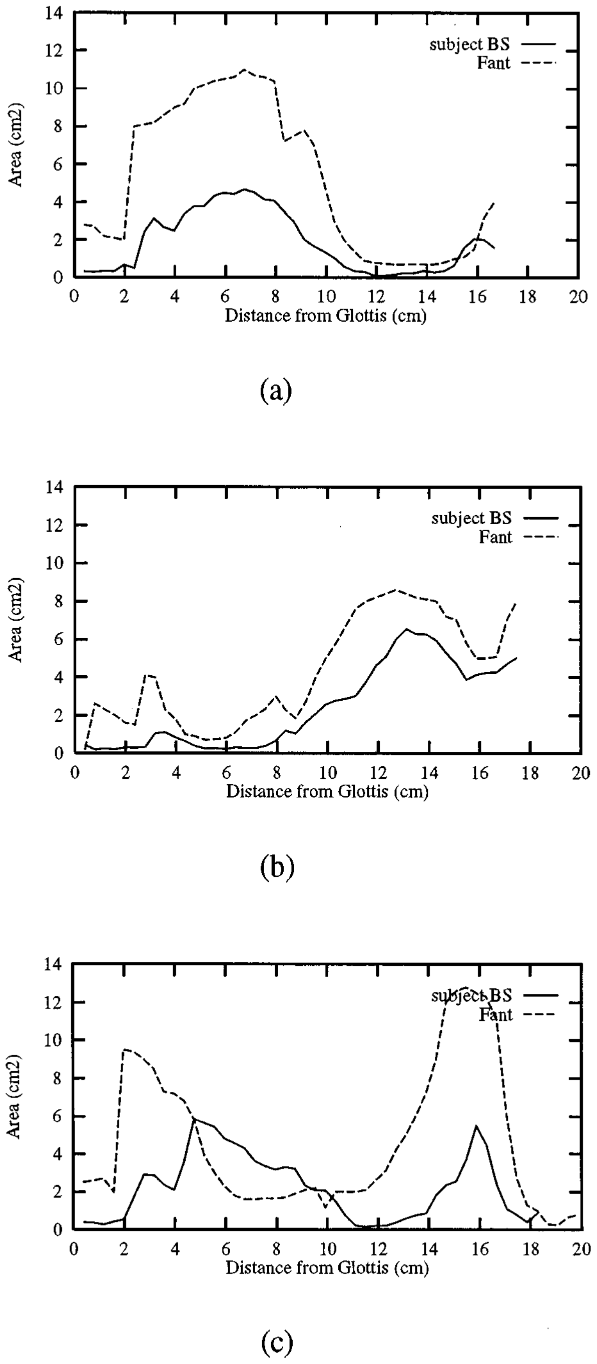
(a)

(b)

FIG. 15. Comparison of the area functions for subject BS (solid) with Dang *et al.* (1994) (dashed): (a) /m/, (b) /n/.

between 2 and 10.5 cm from the glottis for subject BS. Both area functions also show an expanded oral cavity which reaches a cross-sectional area of 8 cm$^2$ in the Dang *et al.* version but only 4.8 cm$^2$ for subject BS. The point of tract closure occurs 1 cm farther from the glottis for subject BS than for the Dang *et al.* area function. In Fig. 15(b), the overall shapes of the /n/ area functions are quite similar. The areas vary between about 1 and 3 cm$^2$ in the region from 3 to 11 cm above the glottis. The point of closure is approximately 0.2 cm farther from the glottis for the Dang *et al.* area function than for subject BS.

Finally, comparisons of the vowels /i, ɑ, u/ from Fant (1960) with those of subject BS are shown in Fig. 16. For all of the vowels, the areas given by Fant seem to be quite large in comparison to those for subject BS and are also large when compared to the Baer *et al.* and Yang and Kasuya area functions. The areas in the pharyngeal region for the vowel /i/ are approximately 5 cm$^2$ larger for the Fant version than for the previous study. The Fant area function for /ɑ/ is almost uniformly larger across the entire vocal tract while the /u/ shows exceptionally large front and back chambers that are on the order of 3 to 7 cm$^2$ larger than the BS version. The Fant /u/ is also more than a centimeter longer than that for subject BS.



(a)

(b)

(c)

FIG. 16. Comparison of the area functions for subject BS (solid) with Fant (1960) (dashed): (a) /i/, (b) /ɑ/, (c) /u/.

## IV. CONCLUSION

MRI has been used to volumetrically image the vocal tract airway of one male subject for 12 vowels, 3 nasals, and 3 plosives. The 3-D image sets were segmented to extract the airway which in turn was analyzed to find the cross-sectional area as a function of the distance from the glottis (along the long axis of the vocal tract). This experiment has provided qualitative and quantitative information about the vocal tract shape. The 3-D surface renderings of the various vowel configurations are a helpful aid in visualizing the effect of articulator positioning on resultant vocal tract shape. The numerical area functions provide a speaker-specific inventory of vocal tract configurations that can be used as input for a speech simulation system. However, since the acquisition of the image set representing each vocal tract shape required many repetitions (approx. 30 repetitions), the area functions need to be considered as an ''average'' shape for a particular vowel or consonant. Due to the large number of repetitions, some fatigue of the articulatory musculature would be expected and as a result, the tract shapes may be somewhat centralized; i.e., fatigue effects may tend to slightly move the vocal tract toward a more neutral or schwalike shape.

The extracted area functions were used as input to a computer model of one-dimensional acoustic wave propagation in the vocal tract. The simulated speech sounds were compared, in terms of formant locations, to recorded natural speech of the subject was imaged. Results indicated that the formant locations were reasonably well represented but some of centralizing effects mentioned above were observed.

The area functions for several vowels and nasal consonants obtained in this study were compared to those given in four previous imaging studies of vocal tract shape. The variability in the area functions observed across these studies are likely due to differences in imaging techniques and procedures, image processing and analysis, and, maybe most importantly, anatomical and physiological differences of the subjects who were imaged. The general shape of each vowel is similar while the details maintain uniqueness. This means, for example, that the /ɑ/ vowel area functions compared in Sec. III E could all be used to generate an /ɑ/ sound but each would possess a unique vowel quality or formant structure.

Measurements of the nasal tract, trachea, and fricative consonants will be presented in the future to augment the data set presented in this paper. A more extensive presentation of speech modeling using these area functions is also planned. Additional work is needed to collect vocal tract shape inventories for more subjects. In particular, there exists little morphological information regarding the female vocal tract and some future studies should be directed specifically at acquiring an area function inventory for female subjects.

## ACKNOWLEDGMENTS

The authors would like to thank Carol Gracco for providing the initial scanning parameters from which we developed those shown in Table I and three anonymous reviewers for their helpful suggestions and comments. This study was supported in part by Grant No. P60 DC00976 from the National Institutes on Deafness and Other Communication Disorders.

## APPENDIX: PHANTOM STUDY

To test the accuracy of the image acquisition and analysis process, a tubular phantom of known dimensions was imaged in the MR scanner. The phantom consisted of a three-section system of hollow (air-filled) plastic tubes connected in a stair-step fashion. The tube system was sealed at both ends and mounted in a water-filled plastic chamber with a 6-in. o.d. (Fig. A1). The middle section of the tube system had a cross-sectional area of 1.14 cm$^2$ if only the air was considered and 1.98 cm$^2$ if both the tube walls and air were included. Both of the end pieces had a 2.45-cm$^2$ cross section for air only and a 3.88-cm$^2$ cross section for both tube and air. Because of the phantom geometry and its positioning in the scanner, accurate cross sectional area could not be obtained from area measurements in a single slice. True area could only be obtained if the analysis software adequately extracted appropriate oblique sections from the volumetric image data set.

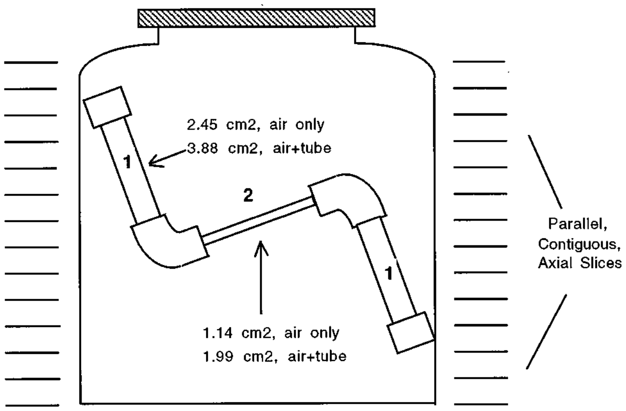


FIG. A1. Schematic diagram of tubular phantom.

The scans consisted of a series of contiguous, parallel, axial slices that included the entire phantom volume. An axial slice refers to the image plane perpendicular to the axis of the cylindrical phantom chamber. The slice thickness was 5 mm.

### 1. Phantom measurements

The image sets for the phantom were segmented, interpolated, and analyzed using the methods described in Sec. I. The set of images does not show the plastic tubing or the air, since hydrogen is not present in either substance.

Table AI shows the known cross sections of the large and small tubes (tube 1 and tube 2, respectively) that make up the phantom along with the cross sections that were measured using the image analysis methods. The measurement overestimated the cross-sectional area of the larger tube by 1.5%. However, measurement of cross-sectional area of the small tube was underestimated by nearly 10%. Since the boundary of a region of interest cannot be determined any closer than ±1/2 a voxel around its perimeter, the resolution of the MR image sets would produce a larger uncertainty in measurement of the region area than would a finer voxel resolution. For a large tube, the voxels on a region boundary would contain a small fraction of the total region area, contributing a small percentage error in area measurement. However, the error in measuring the area would be expected

TABLE AI. Known and measured cross-sectional areas of phantom sections.

| MRI | Tube 1 (cm$^2$) | Tube 2 (cm$^2$) |
|---|---|---|
| Air & tube wall, known | 3.88 | 1.99 |
| Measured MRI image | 3.94 | 1.8 |
| % error | 1.5% | 9.5% |

to grow progressively larger as the cross-sectional area becomes smaller; i.e., the voxels on the boundary of a region of interest will contain a significant fraction of the total area, contributing increasing uncertainty in the total area.

A rough estimate of the expected error for the area measurement of an arbitrarily shaped cross section can be computed by representing the cross-section as an equivalent circle. Using the area, the circumference of an equivalent circle can be computed, which can then be divided by the voxel dimension. This gives an estimate of the number of voxels on the boundary of the cross section (this process can also be performed by considering an equivalent area square; similar answers will result). Since the uncertainty for the voxels on the boundary is $\pm 1/2$ voxel, the uncertainty in the measured area will be the total number of voxels consumed by the cross-section plus or minus one-half of the number of edge voxels. For example, the error produced in the measurement of a 1.99-cm$^2$ cross section using MR images would be calculated as follows. The circumference of the equivalent circle would be computed to be 5.0 cm, which, when divided by a voxel dimension of 0.0938 cm, is equivalent to about 53 edge voxels. The total area consumes 226 voxels [i.e., $1.99/(0.0938)^2$], and the error associated with it should be $\pm 26.5$ voxels (i.e., 53/2). This produces a percentage error of 11.7%, which is similar to the measured error for the small diameter tube given in Table AI.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). ''Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels,'' J. Acoust. Soc. Am. **90**, 799–828.

Beautemps, D., Badin, P., and Laboissiere, R. (**1995**). ''Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data,'' Speech Commun. **16**, 27–47.

Dang, J., Honda, K., and Suzuki, H. (**1994**). ''Morphological and acoustical analysis of the nasal and the paranasal cavities,'' J. Acoust. Soc. Am. **96**, 2088–2100.

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

Greenwood, A. R., Goodyear, C. C., and Martin, P. A. (**1992**). ''Measurements of vocal tract shapes using magnetic resonance imaging,'' IEEE Proc.-I **139**(6), 553–560.

Hoffman, E. A. (**1991**). ''An historic perspective of heart and lung imaging,'' in *3D Imaging in Medicine*, edited by J. K. Udupa and G. T. Herman (CRC, Boca Raton, FL), pp. 285–311.

Hoffman, E. A., and Gefter, W. B. (**1990**). ''Multimodality imaging of the upper airway: MRI, MR spectroscopy, and ultrafast x-ray CT,'' *Sleep and Respiration*, E. F. G. Issa, P. M. Suratt, and J. E. Remmers (Wiley–Liss, New York), pp. 291–301.

Hoffman, E. A., Sinak, L. J., Robb, R. A., and Ritman, E. L. (**1983**). ''Noninvasive quantitative imaging of shape and volume of lungs,'' Am. Physiol. Soc. 1414–1421.

Hoffman, E. A., Gnanaprakasam, D., Gupta, K. B., Hoford, J. D., Kugelmass, S. D., and Kulawiec, R. S. (**1992**). ''VIDA: An environment for multidimensional image display and analysis,'' SPIE Proc. Biomed. Image Proc. and 3-D Microscopy, 1660, San Jose, CA, 10–13 Feb.

Ishizaka, K., and Flanagan, J. L. (**1972**). ''Synthesis of voiced sounds from a two-mass model of the vocal cords,'' Bell Syst. Tech. J. **51**, 1233–1268.

Lakshminarayanan, A. V., Lee, S., and McCutcheon, M. J. (**1991**). ''MR imaging of the vocal tract during vowel production,'' J. Mag. Res. Imag. **1**(1), 71–76.

Liljencrants, J. (**1985**). ''Speech Synthesis with a Reflection-Type Line Analog,'' DS Dissertation, Dept. of Speech Comm. and Music Acoust., Royal Inst. of Tech., Stockholm, Sweden.

Maeda, S. (**1982**). ''A digital simulation method of the vocal-tract system,'' Speech Commun. **1**, 199–229.

Markel, J. D., and Gray, A. H. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York).

Mermelstein, P. (**1973**). ''Articulatory model for the study of speech production,'' J. Acoust. Soc. Am. **53**, 1070–1082.

Moore, C. A. (**1992**). ''The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images,'' J. Speech Hear. Res. **35**, 1009–1023.

Narayanan, S. S. (**1995**). ''Fricative consonants: An articulatory, acoustic, and systems study,'' Ph. D. thesis, UCLA, Dept. of Electrical Engineering, Los Angeles, CA.

Narayanan, S. S., Alwan, A. A., and Haker, K. (**1995**). ''An articulatory study of fricative consonants using magnetic resonance imaging,'' J. Acoust. Soc. Am. **98**, 1325–1347.

Perrier, P., Boe, L-J., and Sock, R. (**1992**). ''Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients,'' J. Speech Hear. Res. **35**, 53–67.

Raya, S. P., and Udupa, J. K. (**1990**). ''Shape-based interpolation of multidimensional objects,'' IEEE Trans. Med. Imag. **9**, 32–42.

Rubin P., Baer, T., and Mermelstein, P. (**1981**). ''An articulatory synthesizer for perceptual research,'' J. Acoust. Soc. Am. **50**, 1180–1192.

Strube, H. W. (**1982**). ''Time-varying wave digital filters for modeling analog systems,'' IEEE Trans. Acoust. Speech Signal Process. **ASSP-30**(6), 864–868.

Sondhi, M. M., and Schroeter, J. (**1987**). ''A hybrid time-frequency domain articulatory speech synthesizer,'' IEEE Trans. Acoust. Speech Signal Process. **ASSP-35**(7).

Story, B. H. (**1995**). ''Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract,'' Ph. D. dissertation, University of Iowa.

Story, B. H., Hoffman, E. A., and Titze, I. R. (**1996**). ''Vocal tract imaging: A comparison of MRI and EBCT,'' SPIE Proc. Physiology and Function from Multidimensional Images, 2709, Newport Beach, CA, 10–15 Feb.

Sulter, A. M., Miller, D. G., Wolf, R. F., Schutte, H. K., Wit, H. P., and Mooyaart, E. L. (**1992**). ''On the relation between the dimensions and resonance characteristics of the vocal tract: A study with MRI,'' Mag. Res. Imag. **10**, 365–373.

Titze, I. R., Horii, Y., and Scherer, R. C. (**1987**). ''Some technical considerations in voice perturbation measurements,'' J. Speech Hear. Res. **30**, 252–260 (1987).

Titze, I. R., Mapes, S., and Story, B. (**1994**). ''Acoustics of the tenor high voice,'' J. Acoust. Soc. Am. **95**, 1133–1142.

Udupa, J. K. (**1991**). ''Computer aspects of 3D imaging in medicine: A tutorial,'' in *3D Imaging in Medicine*, edited by J. K. Udupa and G. T. Herman (CRC, Boca Raton, FL), pp. 1–64.

Yang, C-S, and Kasuya, H. (**1994**). ''Accurate measurement of vocal tract shapes from magnetic resonance images of child, female, and male subjects,'' Proc. ICSLP **94**, 623–626, Yokohama, Japan.