PHONOSCOPE

THREE DIMENSIONAL VOCAL TRACT SHAPES IN /r/ AND /l/: A Study of MRI, Ultrasound, Electropalatography, and Acoustics

Darryl Ong, M.S.

Department of Biomedical Engineering
The Johns Hopkins University
Baltimore, Maryland

Maureen Stone, Ph.D.

Division of Otolaryngology, Head and Neck Surgery
University of Maryland School of Medicine
Baltimore, Maryland

The relative inaccessibility of the vocal tract during articulation creates difficulties in the study of speech production. Magnetic resonance imaging (MRI) allows exceptional imaging of the soft tissue of the vocal tract in multiple planes. This study attempted to develop a procedure for reconstructing three-dimensional (3D) airway volumes from two-dimensional (2D) MRI or computed tomography (CT) slices in multiple planes to characterize the airway shapes for /r/ and /l/ and to study these shapes in relation to acoustic, ultrasound, and electropalatography (EPG) data. /r/ and /l/ were studied because they appear acoustically similar but are physiologically distinct. Edge detection and reconstruction algorithms produced an airway reconstruction sufficient to characterize both sounds and reveal their underlying similarities and differences. Posteriorly, the two sounds had very similar airway shapes until 8 cm above the glottis. Anteriorly, the /r/ was produced with a small area function in the palatal vault region. The /l/ had a larger palatal area function. Physiological differences were corroborated with EPG and ultrasound data. Formant patterns for both sounds were generated from the 3D shapes and midsagittal distances using the "Vtcalcs" program. Results were consistent with known spectra for both sounds and suggested that anterior tongue positions are related to differences in the third formant (F3).

PHONOSCOPE
Volume 1, Number 1, pp. 1–13

A problem that frequently arises in the study of speech production is the relative inaccessibility of the vocal tract during articulation. Data on vocal tract dimensions and shapes are important in numerous investigations such as understanding the speech production mechanism, improving the speech of patients with dysarthria, or designing modern speech synthesizers. Prior to the advent of modern techniques, direct measurement of the vocal tract in imaging was very difficult. Indeed, even indirect estimations of the dimensions and shape of the vocal tract from x-ray projections were challenging. However, recently developed imaging techniques have solved some of the problems associated with visualizing the vocal tract.

Magnetic Resonance Imaging (MRI) is of great interest to speech researchers. Because of its exceptional imaging of the soft tissue of the vocal tract, MRI has several advantages over other imaging techniques, for example, the ability to image multiple planes in any orientation and to observe tissue deformation via tagging. Its drawbacks include the inability to image calcified structures such as teeth and bone, limited slice thickness, and currently rather poor temporal resolution, permitting only static speech conditions to be investigated. For example, Baer et al[1,2] have obtained pharyngeal dimensions of the vocal tract for steady state vowels to calculate vocal tract area functions. Moore,[3] Sulter et al,[4] and Story et al[5] have independently studied the relation between vocal tract resonance characteristics and dimensions of the vocal tract derived from MR images. The main intent of the current study was to characterize the airway shapes for /r/ and /l/, study these shapes in relation to acoustic, ultrasound, and EPG data, and develop a procedure for reconstructing three-dimensional (3D) airway volumes from two-dimensional (2D) MRI or computed tomography (CT) slices in multiple planes. Multiple, 2D MR images were acquired in the coronal, midsagittal, oblique, and axial planes. The resulting images were processed using an active contour algorithm to extract the cross-sectional area of the vocal tract airway in each frame and converted into a 3D volume.

There have been some studies of the perception of /r/ and /l/,[6] but much less work has been done on the production of these semivowels. The American /r/ has been classified as a voiced linguapalatal glide.[7] It is characterized mainly by a change of resonance produced by gliding the tongue past the palate and briefly making lateral contact. Native American English speakers produce /r/ in two ways: the first, known as a retroflex /r/, is described as being made with the "curling back" of the anterior portion of the tongue. The second, or bunched /r/, is made with the tongue arched toward the palate.[8]

The American English /l/ has been classified as a voiced lingua-alveolar lateral,[7] and occasionally as a voiced lingua-alveolar glide, and its standard target position is described as tongue tip against the alveolar ridge, with the tongue adjusted such that its edges do not touch the teeth and palate at the sides. Stone and colleagues[9-13] have shown, using coronal ultrasound scans, x-ray microbeam, and EPG data, that /l/ is produced with a compressed region just behind the tongue tip and a convex posterior portion.

We chose to investigate the production of /r/ and /l/ primarily because of their acoustic similarity; /r/ has a lower third formant than /l/, but the two sounds are virtually identical in their first and second formants. We were interested in studying the vocal tract to observe how this acoustic similarity occurs, despite the apparent differences in production.

METHODS

In this section, we will discuss how the images were obtained, processed, and reconstructed. The first section provides a brief primer on MR physics and deals primarily with image acquisition. The second section details the image processing, which consists of (1) the edge detection using active contour models and (2) the coordinate reconstruction. The third section discusses some of the theory behind the α-shape program used to reconstruct the airway and the relationship between airway shape and airway resonance. Section four discusses the other measurements that were made: ultrasound, electropalatography, and acoustic recordings of the /l/, bunched /r/, and retroflex /r/.

Magnetic Resonance Imaging

MRI exploits the magnetic properties of the hydrogen nuclei (the protons), which are abundant in the water and fat molecules present in biological tissue. Each hydrogen proton spins on an axis and is called a "spin." In the absence of an external field, the spins are randomly oriented, resulting in zero net magnetization. However, on the application of a large external magnetic field, the spins align parallel to the longitudinal axis of the external field and a net magnetization (or polarization) is observed. Once this occurs, a radiofrequency pulse is applied, which causes a transition from the low energy state, in which the hydrogen proton is aligned parallel to the longitudinal axis, to the higher energy state where it is aligned anti-parallel to that axis. As the dipole reverts to its low energy state, it emits a photon, which can be detected by a receiver coil tuned to that frequency.

This MR signal decays exponentially in a process known as the Free Induction Decay (FID). The processes that account for the exponential decay of the MR signal are longitudinal and transverse relaxation, which involve a loss of synchrony of precession of the protons. The first type of relaxation, T_1, is also called longitudinal, or spin-lattice, relaxation. It is the loss of energy associated with the return of the protons (spins) to their original longitudinal orientation. T_2 relaxation, also called transverse or spin-spin relaxation, results from the interaction of spins with their neighbors, causing them to dephase and eventually revert to a state of maximum entropy. As the protons revert to their initial, random phase, the bulk signal decreases because the signals from the neighboring protons tend to cancel each other. This process is characterized by its time constant, T2. The second type of relaxation is longitudinal or spin-lattice relaxation. It is the interaction of the protons (spins) with their surroundings (the lattice of other spins) and is characterized by its time constant, T1. Both T1 and T2 relaxation occur simultaneously, although T2 is always shorter than the T1 process. That is, the spins dephase faster than they move back into alignment with the external magnetic field. Typical biological materials have T1 on the order of several hundred milliseconds, whereas T2 is approximately between 10 and 100 milliseconds. Tissue contrast in MRI is achieved by maximizing the signal differences caused by differences between T1 and T2.

Image Acquisition

The method of acquisition was based on a sequence developed by McVeigh and Atalar[14] and initially designed for dynamic imaging of the human heart on a standard Signa MRI scanner (General Electric Medical Systems, Milwaukee, WI). It is essentially a variation of the Gradient Recalled Acquisition in the Steady State (GRASS) technique with an asymmetric RF pulse and a partial echo readout in order to shorten TR (repetition time) and TE (time-to-echo). Final imaging parameters used were TRÅ6.5ms, TEÅ2ms, 2NEX (2 averages per phase encode), 32 phase encodes/pass with a 28 cm Field of View (FOV) and 5 mm slice thickness. The images obtained from the MR scanner were high-resolution, 256 x 256 images.

A neck coil was placed over the lower part of the subject's face and throat. The neck coil defined the area to be imaged and held the head and neck region securely, minimizing any registration errors caused by inadvertent movement by the subject. Data were collected for /l/ and retroflex /r/, but not for bunched /r/. In isolation, the subject naturally produced retroflex /r/. He produced a bunched /r/ in certain vowel contexts, but could not reliably produce it in the isolated, multiple repetition task required for the MRI data. The following images were acquired, one sequence for /r/ and one for /l/: 13 contiguous coronal slices, beginning immediately behind the front teeth; 8 contiguous axial slices beginning at the level of the hard palate to the second cervical vertebra (C2); 7 axial slices spaced 5 mm apart beginning at the level of C3 to C6; 5 oblique slices 9° apart, beginning around the mid-palatal vault, and ending around the upper pharyngeal area; and one midsagittal image. A subset of the slices used in the reconstructions is marked in Fig 1. The subject was allowed to rest every 32 phonations, to avoid fatigue. Each image required 10 repetitions of the semivowel, for a total of 880 repetitions. Acquisition took a total of approximately 4 hours. The first author served as subject.

The subject, age 22, speaks with a Canadian accent. He was born in Montreal, Canada and learned English as his first language. At the age of 5, he began learning Mandarin Chinese. At 7 years of age, the subject began attending an English-speaking school in Singapore and English was spoken at home. English is the principal language of Singapore. The subject returned to Canada at age 17 and has lived in Baltimore, Maryland since the age of 18.

One major assumption of this study was the repeatability of vocalization. The MR images were collected a few slices at a time, thus there were different utterances for the coronal /r/s than for the axial /r/s, and each slice was the average of four repetitions. We know that human speech has considerable variability, even for the same speaker. However, in this study it was assumed that the short, single-sound utterances were essentially consistent between repetitions. A second assumption concerned acoustics. In an ideal situation, the recording of the speech wave would be made simultaneously with the MR image. However, due to the high magnetic field, we were not able to include a recording device within the gantry of the MR machine. Moreover, the loud noise in the MR scanner caused by the rapid switching of the gradients would have obscured most of the speech signal. Thus the acoustic data were collected later, simultaneously with the electropalatography (EPG) data.

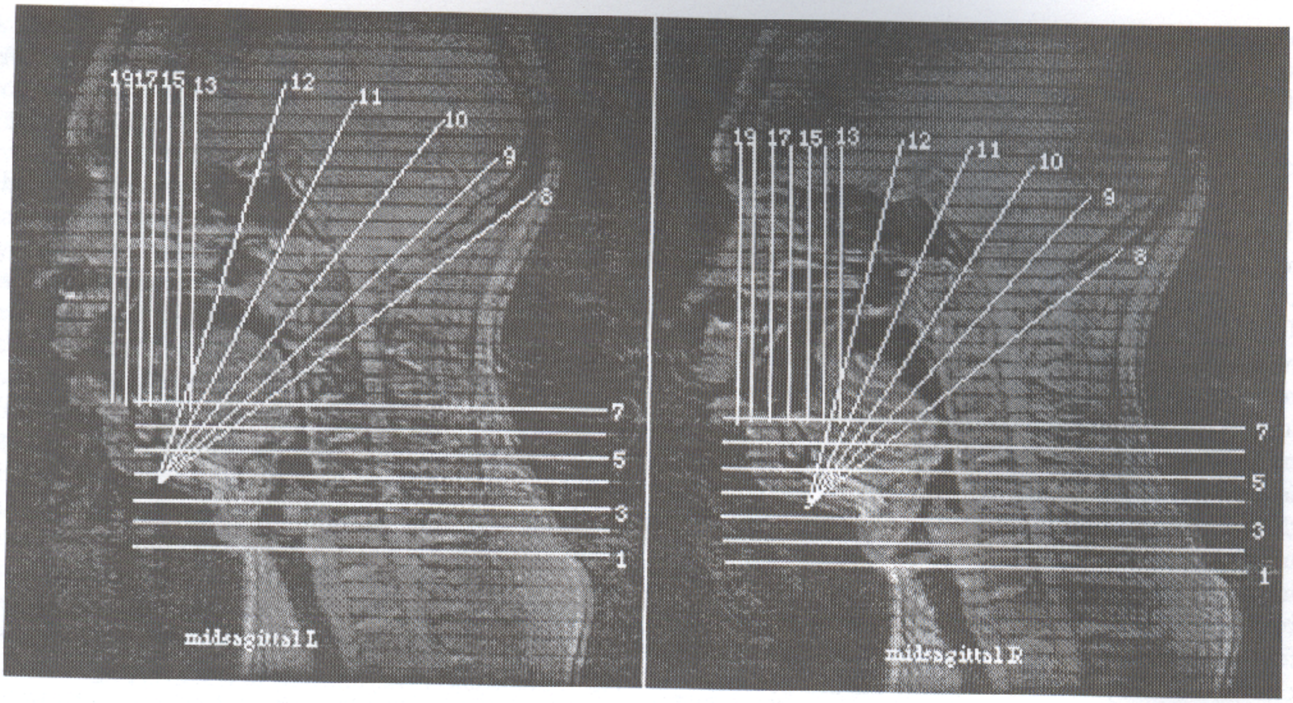


Fig 1. MRI image with white lines indicating the location of the coronal, axial, and oblique slices used in the VTcalcs program.

Processing

File Format Conversion

The GE Signa MRI images were imported into NIH Image (Rasband, NIH) by stripping out their header information. The images were saved as raw data files.

Edge Detection

An edge detection algorithm was developed based on a recent technique known as active contour models, or "snakes." First proposed by Kass, Witkin, and Terzopoulos,[15] an active contour is an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it toward nearby edges, thereby localizing them accurately. This active contour model is defined by an energy functional, and a solution is found based on Euler equations, a technique of variational calculus. Some inherent problems with this technique include numerical instability and a tendency for points to bunch up on strong parts of an edge contour.

Amini et al[16] presented an algorithm for active contour modeling using dynamic programming. This method, although more stable and enabling the inclusion of hard constraints in addition to the soft constraints inherent in the formulation of the functional, is very slow, being on the order of $O(nm^3)$ where n is the number of points in the contour and m is the size of the neighborhood in which a point can move during a single iteration. More recently, Williams and Shah[17] have developed a "greedy" algorithm (one that never undoes what it did earlier) for active contours and curvature estimation. This algorithm allows the inclusion of hard constraints, and is much faster than the Amini algorithm, being on the order of O(nm). We experimented both with the Kass et al algorithm, and the Williams and Shah algorithm on various MR images. Implementing both algorithms on Matlab (The MathWorks, Natick, MA), we found that, for the same initial contour, the Williams and Shah algorithm not only converged faster, but also resulted in a more accurate localization of the edge of the vocal tract, as quantified by a comparison of the mean squared error computed on a known phantom. The phantom was created in NIH Image (Rasband, NIH), and consisted of a circle, a crescent and an ellipse, all set within a larger ellipse. The mean squared error between the detected edge and the actual edge was less than 1 mm in the worst case and approximately 0.35 mm on average. The Kass et al algorithm performed slightly worse, giving an average of slightly over 0.47 mm. These numbers are an average of the mean squared error computed on the circle, the ellipse, and the crescent in the phantom.

In general, the number of points the user needed to define in order for the algorithm to produce a satisfactory contour ranged from 16 to 45. The main reason for the large range of values was due to (1) large variation in the size of the airway areas, (2) some images that had two or more airway areas due to bending of the vocal tract, and (3) some images that had particularly complicated airway areas, in particular the axial slices right around the vocal folds. The Williams and Shah algorithm was implemented on all images except those that had vocal tract shapes with sharp corners, or irregular shapes, where the algorithm failed to give a satisfactory contour and those axial slices that included the buccal cavity, where the calcified structures (teeth and jaw-bone) were not included in the airway. Those edges were delineated manually. Once all of the contours were delineated, the x–y coordinates were written to a file for transformation into 3D coordinates.

Image Reconstruction

The 2D x–y coordinates of the vocal tract shapes were transformed into 3D x–y–z coordinates relative to a single coordinate axis. The first coronal slice was picked arbitrarily to define the z = 0 plane of the coordinate axis, and the rest of the coronal images were transformed simply by adding 5, 10, 15 mm, etc. to their z component. The oblique and sagittal images were transformed using the transformation: $x' = x$, $y' = y \cos \Theta$, $z' = z \sin \Theta$, where Θ is defined as the angle of rotation and then displaced by the appropriate distance. A simple C program was written to realign the 2D coordinate points of the delineated airway edges back into 3D space and, at the same time, to rescale the axes from pixels to millimeters.

Three-dimensional images were reconstructed from the aligned coordinate points using a program named Alvis, which is based on the theory of alpha shapes, Delaunay triangulations, and simulated perturbation. The program is freely distributed by its authors, Edelsbrunner and Mucke, both of the University of Illinois at Urbana-Champaign. The theory behind the program is beyond the scope of this paper; the readers are referred to Edelsbrunner and Mucke[18] for further explanation.

Electropalatography, Ultrasound, and Acoustic Measures

The subject was able to produce both a retroflexed and a bunched /r/, although his preferred /r/ was retroflexed. Electropalatography (EPG) data were collected while the subject produced the two /r/s in isolation and then /iri/ and /ara/ using the Electropalatography 6300 system from Kay Elemetrics (Lincoln Park, NJ). Speech acoustic data were collected simultaneously using Computer Speech Lab (Kay Elemetrics, Lincoln Park, NJ). At a later date, Ultrasound data (Acoustic Technology Labs, Seattle, WA) were collected of the midsagittal tongue contour while the subject produced /r/s intervocalically for each of the 11 vowels of English. The procedures for both these data collections are well documented in other papers.[11,19]

RESULTS

Image Reconstruction

The 3D reconstructions of /r/ and /l/ are presented in Fig 2. In the left image, the 3D airway is rotated about the vertical axis so that the anterior oral cavity is protruding slightly out of the plane of the paper. In the center image, the airway is tilted slightly back to show the tongue shaping the bottom of the airway. In the right image, the airway is viewed from the subject's left side. The results are quite interesting. Although there are differences in the /l/ and /r/ airway, they are nonetheless remarkably similar. Both have a narrow pharyngeal region (occurring higher for /l/) and a wide oral region. This can also be seen in Fig 3. The lower edge of the reconstruction is the surface of the tongue. The retroflex tip seen for /r/ in the middle and right images of Fig 2, B contrasts with the level surface seen for /l/ (Fig 2, A).

Area Functions

Area functions were calculated for both sounds from the 3D vocal tract volumes. A total of 18 slices was used in calculating the area functions: 6 axials beginning with the slice through the glottis and the next 5 superior slices, the 5 oblique slices, and the 7 most anterior coronal slices. The distance between the centers of the cross-sectional vocal tract areas in these slices was approximately 1 cm apart, with the exception of the coronal slices, which were 0.5 cm apart. The slices are shown in Fig 1. These slices were chosen so that we could use VTcalcs, a program developed by Shinji Maeda to compute the formants of a tube based on cross-sectional areas of a tube. VTcalcs creates a multisegmented, tubular model of the vocal tract, and calculates the resonances of the complex tube. The segment areas are based on

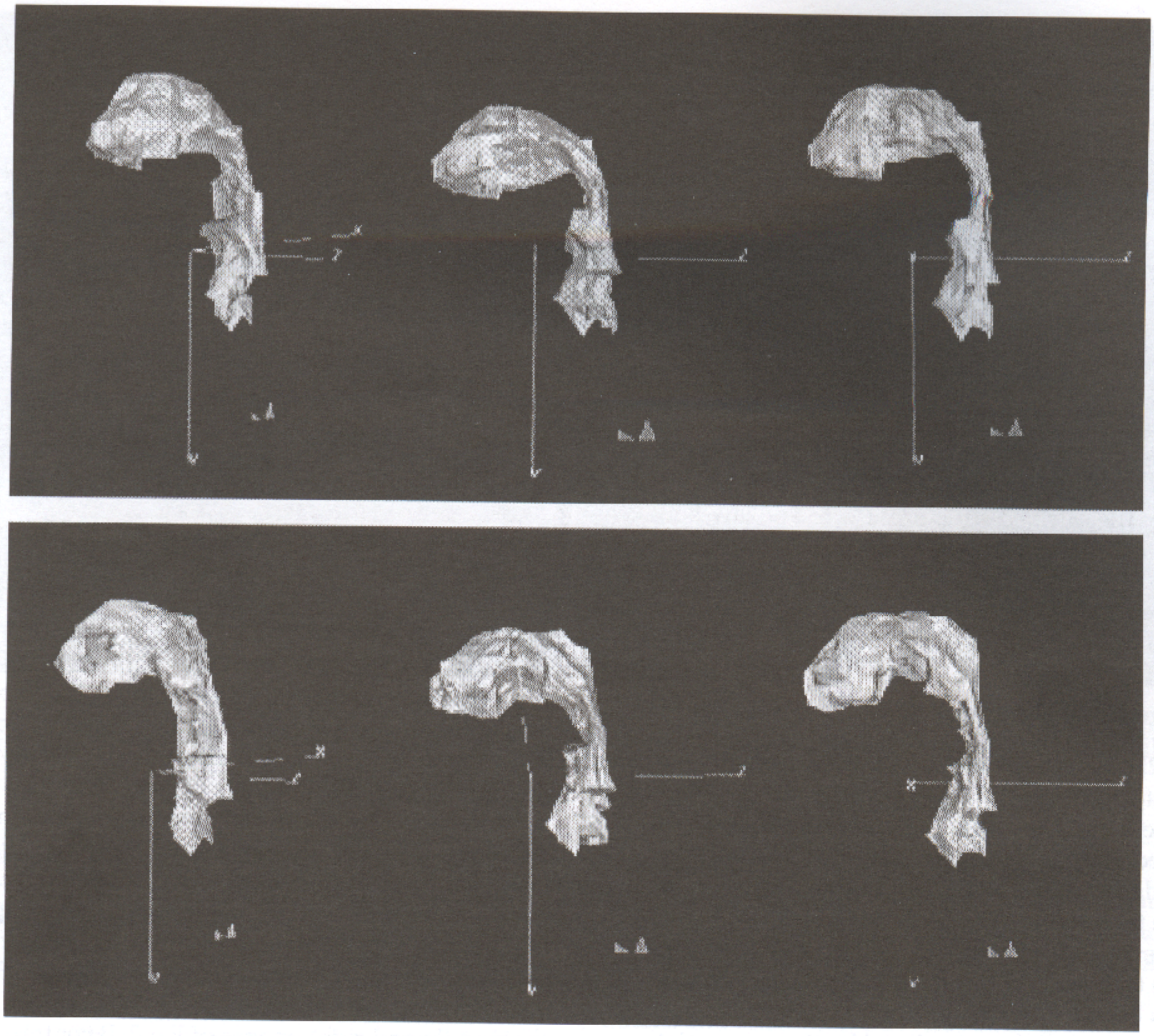
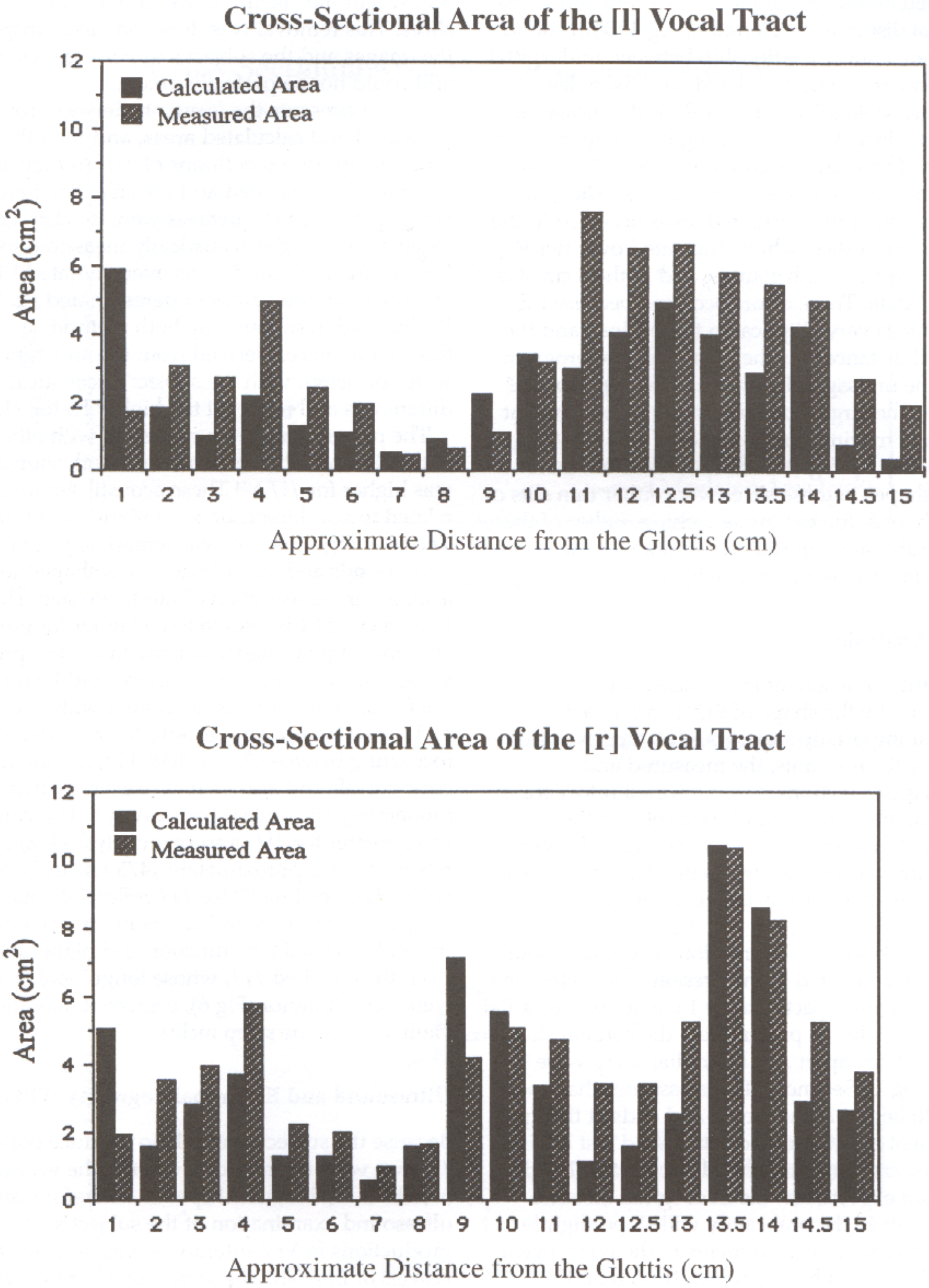


Fig 2. Three-dimensional reconstructions of /l/ (top) and /r/ (bottom). The images show rotation about the vertical axis (left), the horizontal axis (center), and a traditional lateral view (right).

midsagittal, cross-sectional distance measures. Since VTcalcs is based on X-ray tracings, not 3D data, it is only able to estimate the actual volume of the vocal tract, based on the formula $A = \alpha x^{\beta}$, where x is the midsagittal dimension of the vocal tract, and α and β are parameters which were determined based on the position of x. Nonetheless, Maeda's program has shown rather accurate results in comparison to actual speech wave spectra, including the present data.

The area of the vocal tract was measured on these slices using Matlab (The Mathworks, Inc), by means of a program available in the Mathwork's m-file library. This m-file, denoted area.m, calculates areas of polygons using Green's Theorem. Another program was written to extract the midsagittal dimension of the slice based on these area measurements. The midsagittal dimensions were then processed using Maeda's midsagittal distance to area conversion formula. Both the measured areas, and the areas calculated from midsagittal dimensions were plotted as a function of distance from the glottis for both the /r/ and /l/ vocal tracts (Fig 3). The major distinctions between the area functions for the two sounds were a very narrow palatal vault region and a large alveolar region for /r/, but not for /l/. These differences can also be seen in Story[5] (Figs 5-8 and 5-9), Narayanan et al,[20] and Alwan et al.[21] The different areas for the palatal vault region are very important in distinguishing between /r/ and /l/. Pickett[22] notes that a narrowed palatal region for /g/ causes a drop in third formant (F3) frequency for that sound. This



Cross-Sectional Area of the [l] Vocal Tract

Cross-Sectional Area of the [r] Vocal Tract

Fig 3. Area functions for /l/ (top) and /r/ (bottom). Black bars are areas calculated from midsagittal distances using VTcalcs. Striped bars are actual area measurements.

region, then, appears to be critical in creating the local spectral differences between /r/ and /l/.

Measured and calculated areas as a function of the midsagittal distance are plotted in Fig 4. Both sound sets showed a linear relationship between midsagittal distance and area. The calculated areas were less variable and looked like a linear fit to the measured areas (open circles), but with a slight underestimation of the area. These underestimations, as well as some overestimations, are also seen in Fig 3. The largest overestimation of the calculated areas occurred in the first supraglottal slice, which calculated over twice the correct area for both sounds, and in the 8 cm slice for the /r/ data. These errors occurred because the vocal tract was very elliptical in these slices, and the midsagittal distance was the long axis. The program assumes the midsagittal distance is the short axis of the ellipse. The largest underestimations occurred at the region of maximal cross-sectional area for /l/ and at the retroflex tip region for /r/. The actual cross-sectional shapes in these slices were either thin slits or crescent shaped due to tongue arching. Either of these shapes would cause an underestimation of the area in the program, as it assumes an ellipse.

Acoustic Analysis

The formants, or resonant frequencies, are characterized by the shape of the tract and are the single most important way of modulating the voice.[23] To calculate the formants, the measured and calculated area functions were imported into VTcalcs. In addition, the recorded audio samples of the bunched /r/, and the retroflexed /r/ and /l/ were analyzed and the formant frequencies extracted using an Autocorrelation (or Linear Predictor Coefficient [LPC]) Analysis.

VTcalcs makes three assumptions about the input data that were violated in the present study. First, the parallel slices are expected to be 1 cm apart and radial slices 11° apart. In the present data the coronal slices were only 0.5 cm apart and the radial slices were 9° apart (see Fig 1). Second, VTcalcs assumes that vocal tract length begins at the glottis and ends at the lips. In the present data the oral cavity ended just behind the incisors, creating a truncated vocal tract. Third, the program expects that the midsagittal distance measurements in the oral cavity will slope slightly backward and that measurements in the pharyngeal cavity will slope slightly downward, so that the normals to these measurements form an acute angle. In the present data, the oral slices were vertical and the pharyngeal slices were horizontal, thus perpendicular to each other. These differences created a distortion of the oral cavity that probably affected the calculation of the formants. Moreover, the measured areas may have been distorted slightly by removal of the "teeth" space from the oral cavity slices. This removal was done via visual inspection of the images and the subject's upper dental stone cast and could not be perfectly precise.

Table 1 presents the formants derived from the measured and calculated areas, and the LPC extracted formants for two repetitions of /l/ and retroflex /r/. For both the measured and the calculated areas, the derived formant frequencies were in fairly good agreement with the acoustically measured formants. For /r/ the measured areas overestimated F1 and F3, and the calculated areas underestimated F2. For /l/ F3 was underestimated by both methods. However, the F3 differences derived from the area functions were consistent with the subject's acoustical differences and reflected the higher F3 for /l/.

The present data were consistent with other studies, in that F1 and F2 were similar for both sounds, and F3 was higher for /l/.[6,24] These acoustic features can be related to vocal tract shape similarities and differences. The pharyngeal region was remarkably similar for both sounds and, in each case, was shaped to contain a secondary constriction of moderate size. The frequency of F1 is known to be higher for greater amounts of pharyngeal constriction.[22] The present F1 values (400 to 450 Hz) were in the middle region of the F1 range for vowels, consistent with mid vowels and a moderately constricted pharynx. The F2 values (occurring between 750 to 1000 Hz) ran the range of back vowels and also were consistent with a moderately constricted back cavity.[22] The values of F3 were greater for /l/ (approximately 3 kHz) than for retroflex /r/ (approximately 1475 Hz) or bunched /r/ (1597 Hz). The low F3 for /r/ reflects the narrow midpalatal constriction[22] and is in direct contrast with the wider palatal area function and higher F3 for /l/. Even the bunched /r/, whose tongue-palate contacts were fairly anterior (Fig 6), contacted only the palatal vault and not the steep incline.

Ultrasound and Electropalatography (EPG)

Because the subject was able to produce both types of /r/, we were interested in whether he ever used the bunched /r/ in natural speech. Therefore, an ultrasound examination of the subject's /r/ productions in VrV utterances was done for the 11 vowels of English. The same vowel preceded and followed the /r/. The results indicated that the subject used the bunched /r/ with the front vowels /i/, /ɪ/, /e/, /ɛ/, /æ/, and the retroflex /r/ with the others (Fig 5). The bunched /r/s are very similar to each other, but with a slight elevation of the tip in

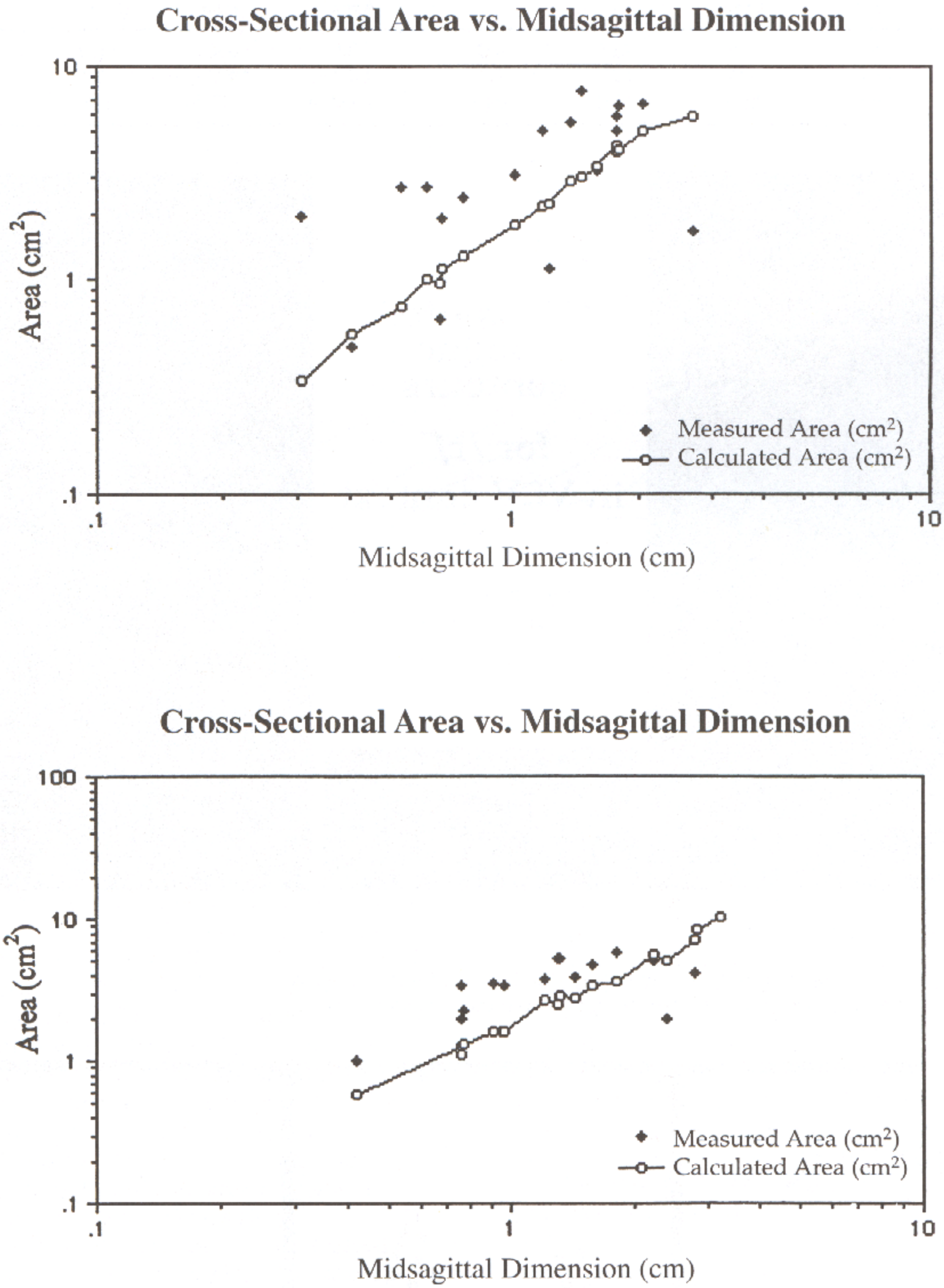


Fig 4. Plot of the measured and calculated areas by the midsagittal distance.

Table 1. Formants as computed by VTCALCS and CSL.

	/l/				/r/			
Formant	Measured Area	Calculated Area	Acoustic 1	Acoustic 2	Measured Area	Calculated Area	Acoustic 1	Acoustic 2
F1	480	376	413	411	538	480	452	442
F2	864	824	826	771	1104	992	1105	1116
F3	2695	2258	3031	2908	2032	1509	1484	1463

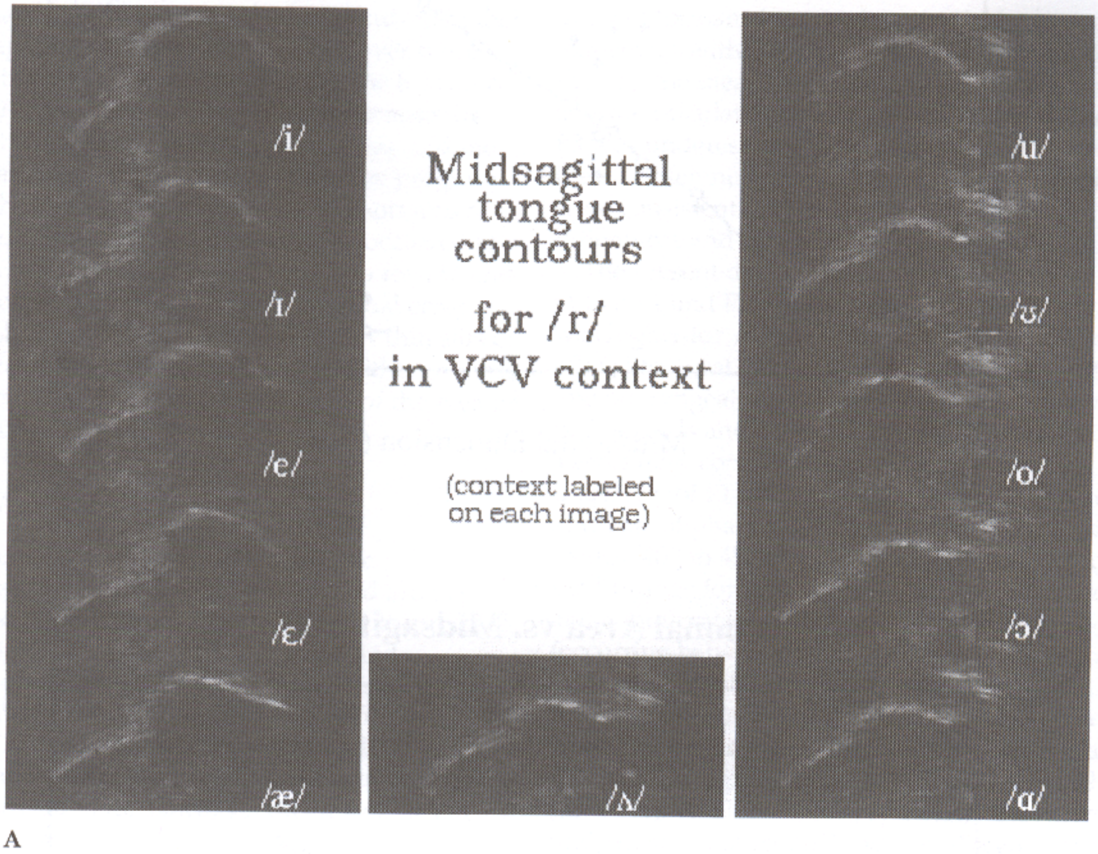

/i/
/ɪ/
/e/
/ɛ/
/æ/
/ʌ/
Midsagittal tongue contours for /r/ in VCV context
(context labeled on each image)
/u/
/ʊ/
/o/
/ɔ/
/ɑ/
A

Fig 5. A, Ultrasound images of midsagittal tongue contours for /r/ in VCV context (context labeled on each image). B, Contours traced from the ultrasound images.

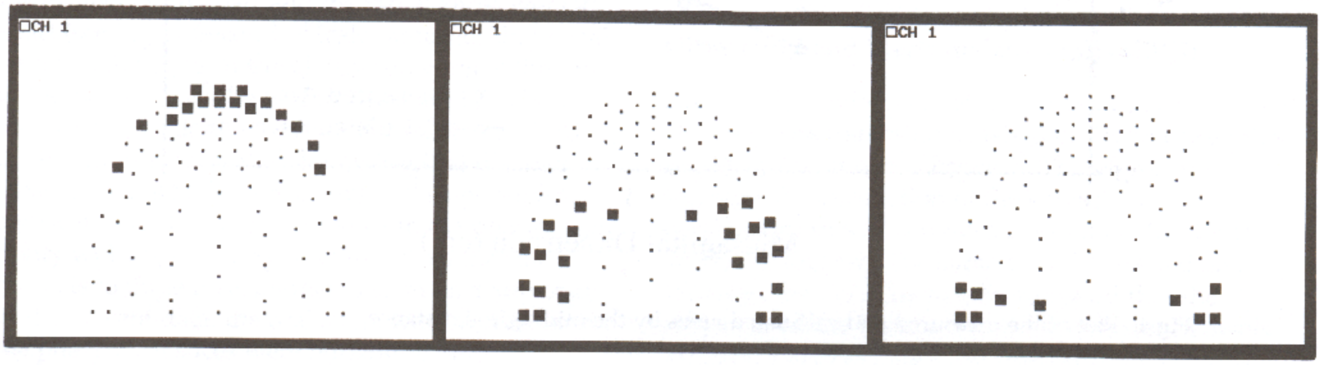
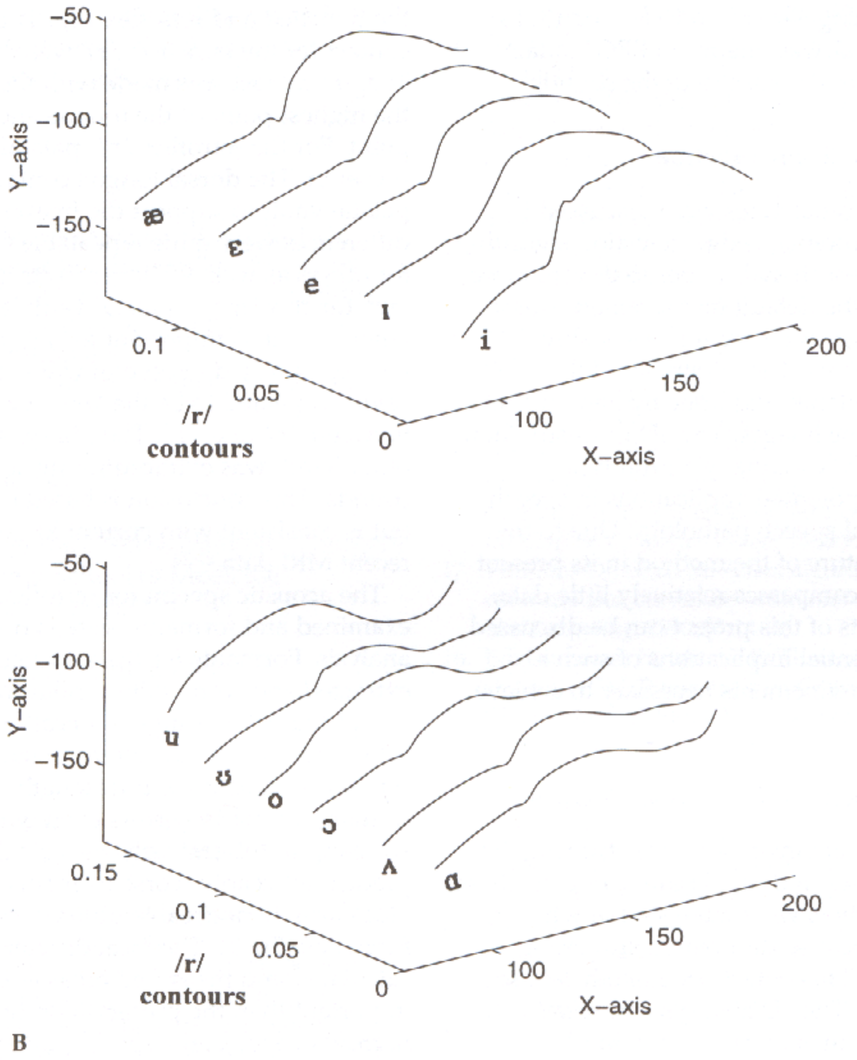


Fig 6. EPG patterns for /l/ (left), bunched /r/ (center), and retroflex /r/ (right).



/æ/ context. The retroflexed /r/s are more varied, but they all have a slight bunching in the middle region of the tongue, very likely at the location of palatal contact. A previous 3D ultrasound study[13] found that the front, high vowels and the bunched /r/ had quite similar tongue shapes, ie, front raising.[25] Therefore, it was quite reasonable for the bunched /r/ to be used in these contexts and for the subject to revert to his preferred /r/ with the back vowels. Moreover, both the bunched and retroflexed /r/ created a vastly reduced area function in the palatal vault, critical for F3 lowering.

The EPG patterns indicated anterior crosswise tongue-palate contact for the /l/, bilateral contact for the bunched /r/, and a shorter, more posterior, bilateral contact for the retroflex /r/ (see Fig 6). During the bunched /r/, palatal contact occurred at the point of maximum constriction and also behind it, but always in the palatal vault, not the more anterior palatal slope. For the retroflexed /r/, contact occurred in the dorsal tongue region and posterior palatal vault, providing stabilization for the posterior tongue body, and no contact at the retroflexed tip. Figure 6 shows steady-state utterances, consistent with the MRI data. The /r/ pattern in /iri/ was identical to the steady bunched /r/, and the /r/ in /ara/ was identical to the steady retroflex /r/. This supported the ultrasound data indicating a bunched /r/ for high front vowels. It also suggested this subject used a coarticulatory strategy of gesture substitution rather than shape/position modification to accommodate front versus back vowels.

DISCUSSION

This study had three components: (1) to develop a computerized procedure for reconstructing the vocal tract airway from MRI slices, (2) to use the computerized procedure to characterize the 3D airway shapes for English /r/ and /l/, and (3) to relate the resulting airway shapes to EPG contact patterns and the acoustic spectra of the sounds.

Vocal Tract Airway Reconstruction

The reconstruction procedures were discussed in detail above. The automatic edge detection algorithm was better for cross-sectional shapes that were more circular. However, the default measurement was a hand-measured edge, allowing for reasonably accurate measurements. The alignment of the different scanning planes was done by hand, using landmarks within the images. The 3D reconstruction of the vocal tract was essentially a prototype technique that has potential applications in speech research and clinical speech pathology. Due to the time-consuming nature of the method in its present form, this study encompasses relatively little data. However, the results of this project can be discussed in terms of the potential implications of such a procedure and the refinements necessary to achieve future goals.

Characterization of /r/ and /l/ Airway Shapes

The second component involved reconstructing 3D airway shapes for /r/ and /l/. Figure 2 depicts these shapes. The researchers were interested in why /r/ and /l/, which appear so different in studies of tongue shape and EPG contact, are nonetheless so similar acoustically. The 3D airways and area functions provided an answer. The tongue configurations, although different, produced great similarities in the resulting vocal tract tube area function. The largest similarities were the global vocal tract shapes, particularly to about 8 cm from the glottis. These shapes revealed small pharyngeal volumes and large oral volumes (Fig 3). The pharyngeal similarity is undoubtedly related to the similarity of F1 and F2 for these two sounds. In the palatal region, slices 10 to 13 showed a much smaller area function (ie, a higher tongue) for the /r/ than the /l/ and a large cavity just anterior to the /r/'s retroflex tip (slices 14–15) due to the absence of any tongue mass.

EPG and Acoustic Correlates

The third objective of the study was to relate the airway shapes to the EPG patterns and acoustic spectra. The EPG data confirmed the differences in tongue-palate contact pattern already well known for /l/ and /r/ (see Fig 6). It also distinguished between the bunched and retroflex /r/. The region of palatal contact for the bunched /r/ was the vault area. Tongue contact was made with the lateral margins of the highest point of the tongue and posterior to this point. For the retroflex /r/, palatal contact was more posterior. The dorsal tongue contacted the posterior palatal vault to support the freestanding tip. The differences were quite repeatable for this subject and for others as well.[26] These results imply that a small area function in the palatal vault, and a large cavity anterior to it, is important for /r/ acoustics, and can be accomplished in several ways. The anterior oral cavity appeared to be the key area that distinguished between /r/ and /l/. The pharyngeal region, on the other hand, was extraordinarily similar for the two sounds. This conclusion is based on only one subject, but is consistent with current knowledge and other recent MRI data.[27,28]

The acoustic spectra for retroflex /r/ and /l/ were examined and formants were extracted using LPC analysis. Formants for the bunched /r/ also were extracted and were quite similar to the retroflex /r/ except that F3 was higher. Recall that the present data set violated the assumptions of slice angle, slice separation, and vocal tract length, resulting in formant estimation errors even when the exact areas were input. Interestingly, the calculated areas did not predict noticeably worse formants than the measured areas despite noticeable differences in those areas' functions (Fig 3). The local differences between the calculated and measured areas apparently were less important than the preservation of a moderately narrow pharynx and, for /r/, a narrow palatal constriction. These larger features resulted in the preservation of similar F1 and F2 values for the two sounds (narrow pharynx), and differing F3s (small palatal area for /r/ only).

CONCLUSION

Multi-slice MRI and 3D reconstructions provided fairly complete information on the shape of the vocal tract airway. This procedure allowed a thorough examination of one subject's /r/ and /l/ productions. Similar pharyngeal area functions resulted in similar F1/F2 values for the two sounds. A large oral cavity difference (small palatal vault area and large alveolar area for /r/) created a focal spectral difference (lower F3). Finally, both bunched and retroflexed /r/ caused a small palatal area resulting in similar F3 lowering for the two /r/s. From the 3D image, a clearer idea has emerged of the physiological similarities and differences of /r/ and /l/. These data, when combined with other physiological data, provided new insight into the links between the acoustic similarities and differences inherent in these two sounds.

Acknowledgments

Supported (in part) by research grant No. R01 DC 01758 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. The authors would like to thank Shinji Maeda for his cooperation and assistance in adapting VTcalcs for this application.

Address correspondence to

Maureen Stone, Ph.D.
Division of Otolaryngology—Head and Neck Surgery
University of Maryland School of Medicine
16 South Eutaw Street, Suite 500
Baltimore, MD 21201
E-mail: mstone@surgery2.ab.umd.edu

References

1. Baer T, Gore J, Boyce S, Nye P. Application of MRI to the analysis of speech production. *Magn Reson Imag.* 1987;5:1–7.
2. Baer T, Gore J, Gracco C, Nye P. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *J Acoust Soc Am.* 1991;90:799–828.
3. Moore CA. The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images. *J Speech Hear Res.* 1992;35:1009–1023.
4. Sulter AM, Miller DG, Wolf RF, Schutte HK, Wit HP, Mooyaart EL. On the relation between the dimensions and resonance characteristics of the vocal tract: a study with MRI. *Magn Reson Imag.* 1992;10:365–373.
5. Story BH, Titze IR, Hoffman E. Vocal tract area functions from magnetic resonance imaging. *J Acoust Soc Am.* 1996;100:537–554.
6. Miyawaki K, Strange W, Verbrugge R, Liberman AM, Jenkins JJ, Fujimura O. An effect of linguistic experience: the discrimination of /r/ and /l/ by native speakers of Japanese and English. *Percept Psychophys.* 1975;18:331–340.
7. Tiffany WR, Carrell J. *Phonetics: Theory and Application.* New York: McGraw-Hill; 1977.
8. Delattre P, Freeman DC. A dialect study of American r's by x-ray motion picture. *Linguistics.* 1968;44:29–68.
9. Stone M. A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *J Acoust Soc Am.* 1990;87:2207–2217.
10. Stone M. Toward a three-dimensional model of tongue movement. *J Phonetics.* 1991;19:309–320.
11. Stone M, Faber A, Raphael L, Shawker T. Cross-sectional tongue shape and lingua-palatal contact patterns in [s], [sh] and [l]. *J Phonetics.* 1992;20:253–270.
12. Stone M. How the tongue takes advantage of the palate during speech. In: *Producing Speech: Current Issues: a Festschrift for Katherine Safford Harris.* New York: American Institute of Physics; 1995:143–154.
13. Stone M, Lundberg A. Three-dimensional tongue surface shapes of English consonants and vowels. *J Acoust Soc Am.* 1996;99:3782–3790.
14. McVeigh ER, Atalar E. Cardiac tagging with breath-hold cine MRI. *Magn Reson Imag.* 1992;28:318–327.
15. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. *Int J Comp Vision.* 1991;1:71–76.
16. Amini AA, Tehrani S, Weymouth TE. Using dynamic programming for minimizing the energy of active contours in the presence of hard constraints. *Proc IEEE Int Conf on Computer Vision.* 1988:95–99.
17. Williams DJ, Shah M. A fast algorithm for active contours and curvature estimation. *CVGIP: Image Understanding.* 1992;55:14–26.
18. Edelsbrunner H, Mucke E. Three-dimensional alpha shapes. *ACM Transactions on Graphics.* 1994;13:43–72.
19. Chi-Fishman G, Stone M. A new application for electropalatography: swallowing. *Dysphagia.* 1996;11:239–247.
20. Narayanan SS, Alwan AA, Haker K. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. *J Acous Soc Am.*1997;101(2):1064–1077.
21. Alwan A, Narayanan S, Haker K. Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data. Part II: The rhotics. *J Acoust Soc Am.*1997;101(2):1078–1089.
22. Pickett JM. *The Sounds of Speech Communication.* Baltimore, Md: University Park Press; 1980.
23. Parsons T. *Voice and Speech Processing.* New York, NY: McGraw-Hill; 1987.
24. Tarnoczy T. Resonance data concerning nasals, laterals and trills. In: Lehiste I, ed. *Readings in Acoustic Phonetics.* Cambridge, Mass: MIT Press; 1996:111–117.
25. Harshman R, Ladefoged P, Goldstein L. Factor analysis of tongue shapes. *J Acoust Soc Am.* 1976;62:693–707.
26. de Jong KJ. The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. *J Acoust Soc Am.* 1995;97:491–504.
27. Patterson DK, Pepperberg IM. A comparative study of human and parrot phonation: acoustic and articulatory correlates of vowels. *J Acoust Soc Am.* 1994;96:634–648.
28. Tanimoto K, Henningsson G, Isberg A, Ren YF. Comparison of tongue position during speech before and after pharyngeal flap surgery in hypernasal speakers [see comments]. *Cleft Palate-Craniofac J.*