TR－H－025

Reconstructing the Vocal Tract During Vowel Production using Magnetic Resonance Images

Barton F. Lane
Eric Vatikiotis-Bateson

1993. 9. 9

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2　☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

RECONSTRUCTING THE VOCAL TRACT DURING VOWEL PRODUCTION USING MAGNETIC RESONANCE IMAGES

Barton F. Lane and Eric Vatikiotis-Bateson
ATR Human Information Processing Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

The goal of this study was to reconstruct and accurately quantify a three-dimensional (3D) vocal tract during the pronunciation of different vowels. This was done using magnetic resonance image (MRI) data recorded during sustained pronunciation of five vowels. Since MRI cannot distinguish between bone and air, it has been difficult to reconstruct reliable volumes for the oral cavity since the teeth do not produce an air-tissue boundary. Therefore, a major task in this study was to find a means to identify the teeth in the images and subsequently to subtract them from the reconstructed oral cavity volume. Of the five vowels examined, two had cavity shapes suitable for tooth subtraction using threshold-detection techniques. Cross-validation of the subtracted volumes with a more direct, physical measure of tooth volume suggest an accuracy of more than 90 percent. Using configurations in which accurate subtraction is possible to derive constant values for the teeth it is then possible to apply these values and to estimate reasonable volumes for all configurations produced by the same speaker.

INTRODUCTION

This study had two purposes. First was to address an inherent problem of MRI, namely to find a way around the problem of estimating realistic volumes for the oral cavity in which the teeth have indistinguishable air-tissue boundaries. Second, and somewhat independent of the first, was to provide volumetric estimates for a particular speaker's vocal tract structures, which could be used to parametrize the dynamical model developed for that speaker's muscle EMG and articulatory motion data collected by flesh-point measurement techniques (e.g., magnetometer, OPTOTRAK; see Hirayama et al., 1992; Vatikiotis-Bateson et al., 1993). Ideally, if the contribution of the teeth could be subtracted from the total oral cavity volume, accurate cross-sectional areas could be computed and used to adjust our acoustic synthesis model derived from the mapping between articulator kinematics and PARCOR coefficients. In this paper, we concentrate on describing the methodology and results pertaining to volumetric reconstruction and subtraction of estimated tooth volumes.

METHODOLOGY

Image acquisition and processing

MRI provides a means to visualize the physical configuration of the entire vocal tract during a single (prolonged) speech production event. In this experiment, MRI images were obtained for the vocal tract during pronunciation by an English speaker (EVB) of the five vowels /a, i, u, e, o/. In addition, image sequences were obtained for several static postures — e.g., the protruded tongue was clamped between the teeth to determine the orientation and shape of the teeth, particularly the teeth anterior to the molars (see Figure 1). Images were acquired using a Shimadzu 1.0 Tesla field strength machine; the overall dimension of each acquired image was 300 x 300 mm. For each vowel, 14 axial slices (orthogonal to the longitudinal axis of the body) and 10 coronal slices (parallel to the face plane) were taken, as shown in Figure 1. The axial slices were used to measure the pharyngeal area; the first slice began just below the larynx. The coronal slices were used to measure the oral cavity; the first slice began just behind the pharynx. The slices were 10mm thick and contiguous (i.e., no gaps). The entire set of 24 axial and coronal images for a single vowel were recorded over a period of approximately 30 seconds, therefore the subject was required to maintain the vocal tract conformation for the vowel for that period of time.

After acquisition, images were transferred to a Silicon Graphics workstation (IRIS). There they were transformed into 8-bit gray-scale images in VoxelView format and calibrated in software-specific units (Voxels). Images were tagged by the program for volumetric reconstruction and manipulation. Subsequently, these gray-scale images were moved to an Apple Macintosh as PICT files. On the Macintosh, Adobe Photoshop was used for further image analysis and area measurement.

Figure 1. Mid-sagittal head during the bite trial, showing the distribution and orientation of coronal and axial scans. Vertical lines indicate the boundaries of the rightmost and leftmost coronal slices (slice nos. 1 and 10, respectively). Horizontal lines indicate the lowest (no. 1) and highest (no. 14) axial scans.

Vocal tract area measurements

Originally, we had hoped to reconstruct the volume for the entire head and then, using air-tissue boundary detection, subtract the entire volume of the vocal tract from that of the head. The dimensions of the resulting vocal tract volume could then be analyzed in the coordinate orientation of the head. Although this method is feasible in principle, the image quality must be very high and the slice thickness much thinner (less than 2-3mm). Therefore, in the current study, we were relegated to an analyzing individual axial and coronal slices and reconstructing the volume from pieces.

Estimation of vocal tract area entailed several steps. First was to determine which areas of the vocal tract were imaged better (i.e., with less distortion) by the axial and coronal slices. Once the set of image slices was selected, air-tissue boundaries were marked with an automatic detection algorithm, edited when necessary, and converted to physiological measures of the vocal tract cross-sectional area. An important task here was identification and subtraction of the teeth from the cavity estimation. The net vocal tract areas for each slice were then summed to give volume estimates for the different vowels.

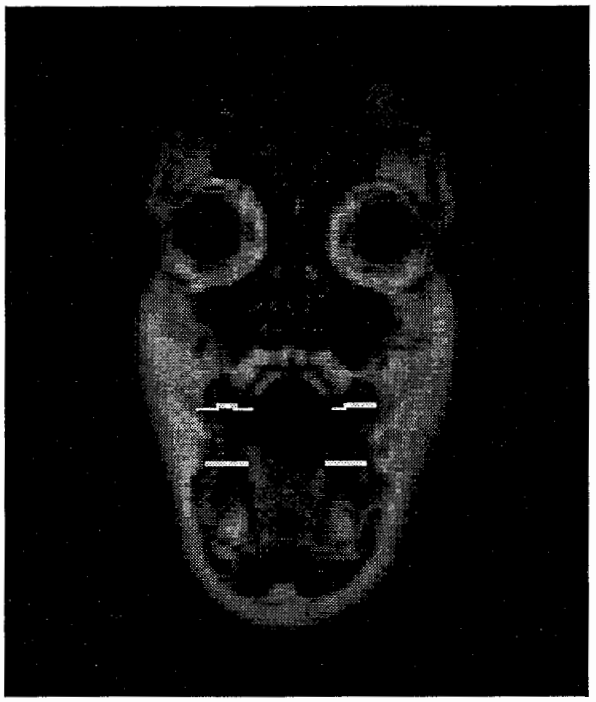

Figure 2 Coronal slice showing how the teeth were eliminated from the vocal tract area detection. The entire blacked out vocal region was originally selected; the white lines in the figure show where the cuts were made eliminating the lower teeth and the upper teeth above the gumline. The upper teeth themselves remain in the selected area.

Vocal tract outline. Determining the vocal tract outline (air-tissue boundary) was carried out as follows. Each image was opened on the Macintosh at 256 X 256 pixel resolution. Image brightness and contrast were adjusted to provide optimal definition of the vocal tract cavity. Then, an edge detection and area selection algorithm (Adobe Photoshop's 'magic wand tool') was used to select the vocal area automatically. In cases where the cavity was defined for two or more disconnected areas, such as occurred for the axially imaged airspace anterior to the epiglottis, multiple selections were made and grouped. Area measures were recorded (in pixel values) and converted to physiological measures (mm). The selected area was isolated and saved as a binary raw data file.

Although modification of the automatic edge detection algorithm was avoided as much as possible, some modifications were necessary to generate complete sequences of images within each orientation. These were restricted to sequences of scans used in visual comparison of the axial and coronal orientations for the different vowels (see Figures 3a-b), but were not used for the volume calculations: the superior axial slices, which could not be used for volume reconstruction because they gave a distorted rendering of the oral cavity, and the posterior coronal slices, which gave a distorted rendering of the pharyngeal area. This distortion was the result of the relatively thick slices not being oriented orthogaonally to the vocal tract in these areas.

Tooth detection and subtraction. Modification of the automatic edge detection algorithm also was necessary when making oral cavity measures from the coronal slices. As there was no way to discern the tooth-air boundary, entire teeth including the roots were typically included in the airspace detection. However, it was possible to visualize the gumline, so the selected area could be easily modified to exclude the portions of the teeth lying beneath the gumline. This editing of the detected cavity outline, depicted in Figure 2, resulted in a smoother extracted area. It also reduced the tooth subtraction problem to just those geometrically more tractable portions of the teeth protruding above the gumline.

It was possible to eliminate the lower teeth (on the mandible) entirely from the coronal slice by way of the tongue-tooth boundary; this was possible for every vowel except in the most anterior slices of the /u/ vowel (see Figure 3b). To correct for the inclusion of the teeth in some of the coronal area measurements, the tooth area was obtained from a separate bite scan containing 10 coronal images acquired while the subject held his tongue between his teeth (Figure 1). The purpose of the bite scan was to make visible the tongue-tooth boundary. The tooth area from these images was selected in the same way as was the vocal area. Similarly, the area selections were modified to remove the area below the gumline, so that only the volume of the teeth above the gumline remained. These values are shown as the separate upper and lower tooth volumes in Table 1b.

Table 1a

Area measurements of the pharyngeal portion of the vocal tract, taken from *axial* slices. Height is measured in mm from slice #1, just below the larynx. The volume (mm^3) and total area (ΣArea in mm^2) of /i/ was calculated using one more slice than the other four vowels.

Height	a	i	u	e	o
0	265	255	298	243	254
10	168	139	233	78	115
20	63	152	137	41	71
30	44	185	119	36	89
40	69	353	151	168	218
50	170	349	255	199	148
60	30	442	220	147	93
70	29	466	104	183	44
80	29	505	117	233	29
90	54	456	91	272	33
100	92	468	76	255	71
110		214			
ΣArea	1013	3984	1801	1855	1165
Volume	**10121**	**39853**	**18018**	**18553**	**11673**

Removing the volume of the teeth from the vowels /a/ and /o/ involved simply subtracting the entire volume of the upper teeth, since the coronal vocal area measurements for these two vowels included all of the upper teeth. The procedure for the other three vowels was more complicated. As can be seen in Table 1b, only some slices contain the tooth areas for these vowels. The problem was determining which slice from the bite scan to use to subtract from the total volume, since corresponding slices from the vowel and the bite scans do not necessarily correspond to the same area of the vocal tract. From an anatomical standpoint, as the upper teeth are fixed in relation to the vocal tract, a reconstructed tooth volume should correspond to the same location for every vowel. However, slight changes in head position between subsequent vowel scans had the effect of shifting the overall orientation of the vocal tract for each vowel. For example, it can be seen in Table 1b that the area of the upper teeth only at the 80mm slice scan (second column from right) is greater than the area of the teeth and vocal tract at the 80mm slice from the /i/ vowel. Furthermore, where the last slice of the bite volume clearly showed some tooth area, some of the vowel images showed no tooth area in the final slice. Therefore, when subtracting tooth volumes for these three vowels, the tooth volume slice

Table 1b

Area measurements of the mouth portion of the vocal tract, taken from *coronal* slices. Height is measured from slice #1 (moving forwards). A 't' indicates that the teeth (u=upper, l=lower) are included in the area. For vowels /a/ and /o/, the volume without teeth was determined by subtracting the total tooth volume of the upper teeth. For /e/ and /i/, the 80mm upper tooth slices were subtracted. For /u/, the 70 and 80mm tooth slices from both upper and lower teeth were subtracted. * indicates the tooth volume obtained from a vacuum mold of the teeth.

height	a	i	u	e	o	ut	lt
10	228				114		
20	93	338	30	91	187		
30	154	45	19	29	368		
40	352 ut	16	21	27	663 ut	19	
50	573 ut	4	63	22	869 ut	108	
60	618 ut	12	163	43	920 ut	114	
70	562 ut	41	817 ult	49	694 ut	74	117
80	354 ut	163 ut	389 ult	216 ut	404 ut	196	30
90	240 ut	108	23	211 ut	87	80	
S Area	475	564	319	261	756	591	147
Volume:							
total	**31737**	**7292**	**15257**	**6880**	**43053**	**5919***	
no teeth	**25818**	**6492**	**11027**	**6080**	**37153**		

which appeared to match the tooth shape in the vowel slice was used, whether or not it corresponded in terms of slice number (see Table 1b).

Once pixel values had been obtained for the areas from every image, they were converted into physiological values (mm). The resulting area measurements are shown for each vowel, and for the teeth, in Tables 1a-b. In addition, a total volume calculation is given for each vowel. The volume was calculated by taking the original slice thickness to be 10mm; the images were then stacked to give a volume measure. As shown in Table 1, not all of the slices were used from each set in the volume calculations. As mentioned before, those slices which gave distorted views (i.e., not orthogonal to the vocal tract) were not used. Generally, for the axial slices, slices up to and including the uvula were used. For the coronal plane, slices from the lip region moving back to where the pharynx comes to dominate the area were used.

As a confirmation for determining the tooth volume from images, the volume of the teeth was obtained from high quality vacuum molds made from dental casts of the subject's upper and lower teeth. For example, the mold for the upper teeth was filled with water from a graduated syringe to the level of the gumline and the volume was thus calculated; this value is given in Table 1b. As can be seen, the tooth volume obtained from the MRI images is an overestimate. The volume from the mold itself may be a slight overestimate as the mold does not make a perfectly tight fit with the original tooth model, though this might be counteracted by slight shrinkage of the dental impression.

VOLUME RECONSTRUCTION

In order to reconstruct calibrated vocal tract volumes, the vocal areas derived from Photoshop were transferred back to the IRIS workstation and the total vocal tract was reconstructed for each vowel. These volumes can be seen in Figures 3a-b. The procedure for reconstructing the volume involved an interpolation procedure. Interpolating the slices "fills in" areas between the slices to create smoother transitions than in a simple stack of the slices. This then increases the overall volume of the vocal tract compared to that of summing the areas of each slice. For the axial reconstruction, the first 13 slices were used; for the coronals all 10 slices were used. The volume was constructed with each slice being interpolated 9 times, to create 121 axial images and 91 coronal images. The voxel scaling values were then adjusted to give real world measurements of 300 x 300 x 130 mm for the axial reconstructed volume, and 300 x 300 x 100 mm for the coronal reconstructed volume. Because the interpolation procedure added area between the slices, the resulting volume's dimensions were larger than those calculated using the procedures described above and shown in Table 1. The volumes obtained from the reconstruction are shown in Table 2; the volumes have been taken through the same distances as with the slices in Table 1.

A better estimation of volume would be to merge the appropriate subsets of axial and coronal slices to create a comprehensive volume of the entire vocal tract. This was not

Table 2a

Vocal tract volumes (mm^3) of the pharynx for the five vowels, as computed by theVoxelView volumetric reconstruction. The regions (mm) were measured from slice #1, just below the larynx.

Vowel	Volume	region
a	13203	0-108
i	49794	0-118
u	22886	0-108
e	22192	0-108
o	14509	0-108

Table 2b

Vocal tract volumes of the mouth area for the 5 vowels, using VoxelView. For the vowels /a/ and /o/, the total upper tooth volume was subtracted; it was not possible to determine a suitable tooth volume to subtract from the other volumes (see text). The region measured is the distance from slice #1, which was located at the back of the pharyngeal area, except * where the area is measured from the first slice (posterior) where the teeth are visible.

Vowel	Volume	no teeth	region
a	42368	31799	11-100
i	10117		22-100
u	26587		22-100
e	8570		22-100
o	56707	46138	11-100
Upper teeth	10569		0-59*

done due to the uncertainty surrounding the accuracy of the reconstructed volumes. The thickness of the slices precluded accurate measurements of areas of the vocal tract not sampled cross-sectionally, such as the transition between the oral cavity and the pharynx in the vicinity of the velum.

In order to correct for the teeth in the volume reconstruction, the tooth area images derived earlier were also interpolated to create a tooth volume (not shown); the total volume is shown in Table 2b. Because of the interpolation procedure, the tooth volume measured is approximately twice that calculated from the original images. To extract the teeth from the volumes proved to be impossible except for the /a/ and /o/ vowels, where

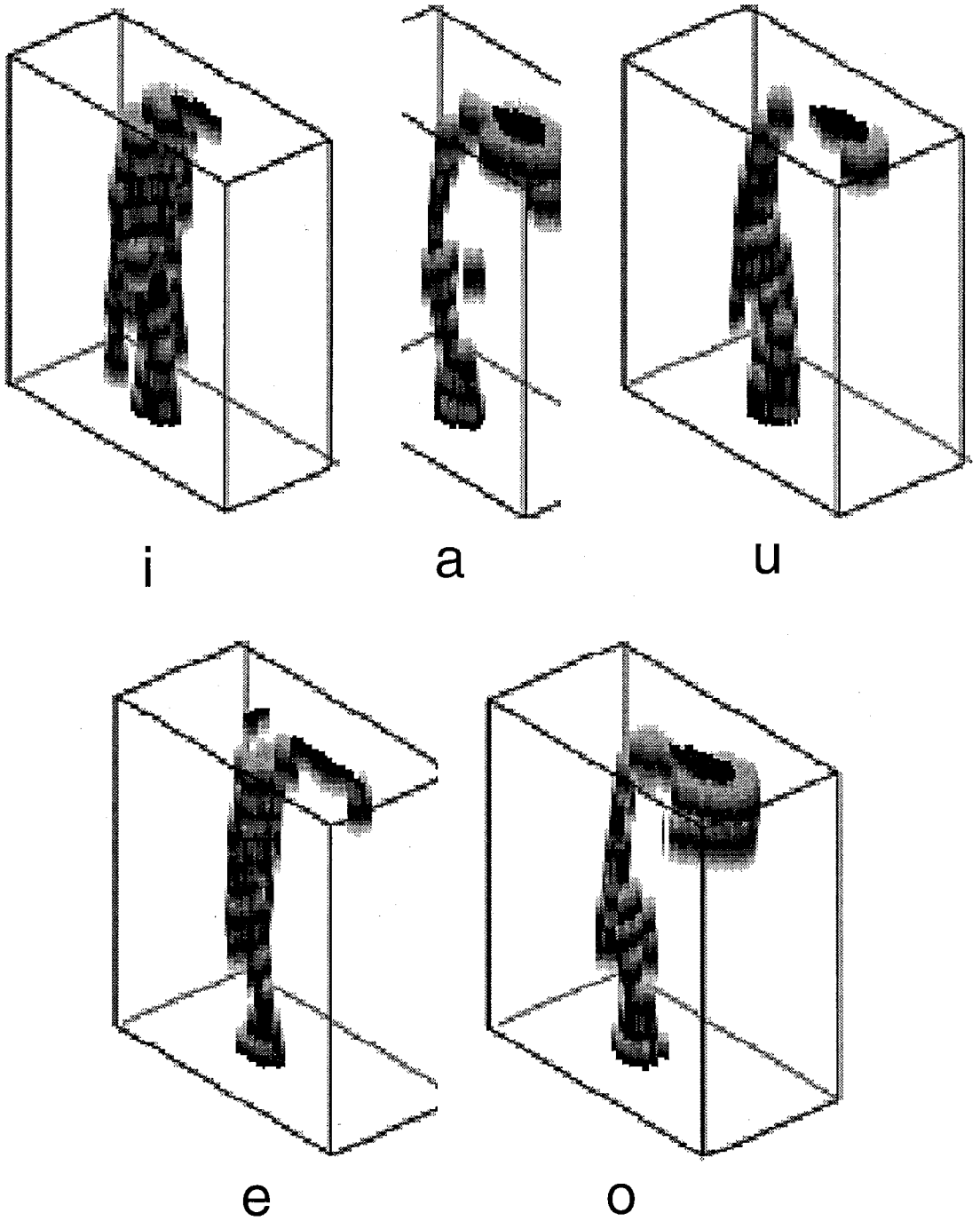

Figure 3a: Volumes reconstructed from axial slices using voxel view. The gray scaling is a result of the interpolation procedure: the original slices take on black values; the interpolated slices are shaded lighter as their distance from the original slice increases.

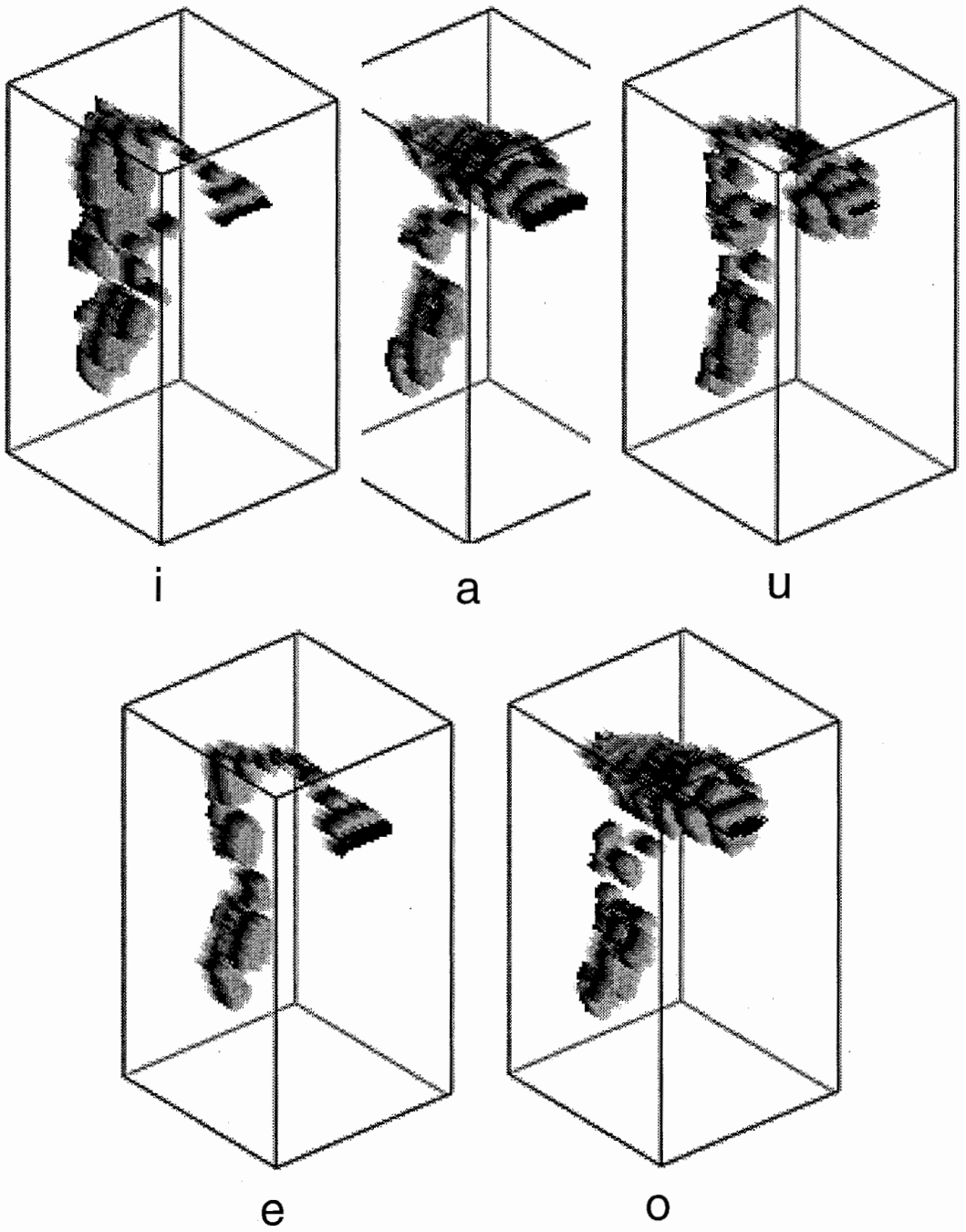

Figure 3b: Volumes reconstructed from coronal slices using VoxelView. The gray scaling is a result of the interpolation procedure: the original slices take on black values; the interpolated slices are shaded lighter as their distance from the original slice increases.

the entire volume could be subtracted. It was not possible to determine with any degree of certainty where to slice the reconstructed tooth volume to isolate the correct tooth values for the other three vowels. Since the interpolation procedure expands the volume differentially for each set of slices, it would be invalid to subtract the tooth volumes obtained for /a/ and /o/ from the other three vowels.

DISCUSSION

The reconstructed volumes (Figure 3) show clear differences in vocal tract shape for the five vowels. Although the axial slices show the best results, since there are no teeth to interfere with visualization, the differences can clearly be distinguished in the coronal slices as well. Whether or not these volumes can actually be used for further speech analysis depends upon whether or not the measurements determined in this experiment can be verified.

The major obstacle to determining accurate area measurements was the inclusion of the teeth in the coronal images. It was decided that the best way to eliminate the teeth was to try and obtain tooth measurements using the same procedure used to measure the vocal tract area; the errors were thus minimized. However, due to the small size of the teeth in relation to the thickness (10mm) of the images, potentially large overestimates could not be avoided.

There was also a problem relating to the interpolation procedure. In this experiment, the differences in values between the interpolated volume and the volume calculated from the original slices are significant. Because of the thickness of the slices, the vocal tract area was under sampled, and therefore the interpolated slices dominate the volume reconstruction. As the acquired images were contiguous slices, interpolating and thus adding slices creates additional volume. The question remains then which method of volume measurement should be used for the final analysis of the vocal tract. For the purposes of this experiment, the interpolated volumes are useful as visualizations of the vocal tract, providing a smoother image than simply stacking the original slices. A midsagittal image acquired for each vowel would have helped to confirm the volume reconstruction, and would have provided an immediate cross-reference index of the axial and coronal slice orientations. This would be a good idea for future experiments.

Despite these difficulties relating to under sampling of data, the procedures to quantify the volume of the teeth were successful. The volume calculated for the teeth was confirmed by comparison with the vacuum mold impression. Furthermore, Figure 3b shows that the teeth could successfully be eliminated from the volume reconstructions, to produce striking visualizations of the vocal tract in mid-speech. Given higher resolution MRI data, it should be possible to more accurately calculate the volume of the teeth, using techniques such as the tongue-bite procedure. Similarly, if specific tooth landmarks could be established, then a single reconstructed tooth volume could be used to subtract the teeth from any orientation of the vocal tract for a single speaker, despite the inherent variability of the position of the tongue and mandible.

CONCLUSIONS

The research methods used in this experiment were valid in their approach to the problem. In the future such methods could be used with better data to reconstruct accurate vocal tract volumes. Better sampling of the vocal tract using thinner image slices would likely provide more accurate results, especially where the teeth are concerned. The axial volumes rendered in this experiment showed that the thickness of the acquired images, at 10mm/slice, was too great to show adequate detail. There were substantial variations in the images from adjacent slices.

This experiment was also successful in showing relative configurations of the vocal tract for different vowels. In the axial volumes especially, but also in the coronal volumes, the overall shape and orientation of the vocal tract was clearly distinguished. Also, the data provided good 3D reconstructions of the pharynx, which otherwise cannot be measured, and further demonstrates the value of MRI imaging for extracting measurable vocal tract parameters during speech production.

ACKNOWLEGDMENT

Kevin Munhall provided helpful criticism of an earlier draft.

REFERENCES

Baer, T., Gore, J.C., Gracco, L.C., & Nye, P.W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, *J. Acoust. Soc. Am.*, **90**, 799-828.

Chiba, T., and Kajiyama, M. (1941). *The vowel, its nature and structure* . Tokyo: Tokyo-Kaiseikan.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Honda, K. (1992). Neural network modeling of speech motor control. In *The International Conference on Spoken Language Processing-1992*, **2** (pp. 883-886). Banff, Canada.

Vatikiotis-Bateson, E., Hirayama, M., Wada, Y., & Kawato, M. (1993). Generating articulator motion from muscle activity using artificial neural networks. *Annual Bulletin of RILP* (Univ. of Tokyo), **27**, 67-77.