A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/

Xinhui Zhou[a] and Carol Y. Espy-Wilson[b]
Speech Communication Laboratory, Institute of Systems Research, and Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742

Suzanne Boyce[c]
Department of Communication Sciences and Disorders, University of Cincinnati, Mail Location 0394, Cincinnati, Ohio 45267

Mark Tiede[d]
Haskins Laboratories, 300 George Street Suite 900, New Haven, Connecticut 06511

Christy Holland[e]
Department of Biomedical Engineering, Medical Science Building 6167, Mail Location 0586, University of Cincinnati, Cincinnati, Ohio 45267

Ann Choe[f]
Department of Radiology, University Hospital G087C, Mail Location 0761, University of Cincinnati Medical School, Cincinnati, Ohio 45267

(Received 16 April 2007; revised 26 February 2008; accepted 29 February 2008)

Speakers of rhotic dialects of North American English show a range of different tongue configurations for /r/. These variants produce acoustic profiles that are indistinguishable for the first three formants [Delattre, P., and Freeman, D. C., (1968). "A dialect study of American English r's by x-ray motion picture," Linguistics **44**, 28–69; Westbury, J. R. *et al.* (1998), "Differences among speakers in lingual articulation for American English /r/," Speech Commun. **26**, 203–206]. It is puzzling why this should be so, given the very different vocal tract configurations involved. In this paper, two subjects whose productions of "retroflex" /r/ and "bunched" /r/ show similar patterns of F1–F3 but very different spacing between F4 and F5 are contrasted. Using finite element analysis and area functions based on magnetic resonance images of the vocal tract for sustained productions, the results of computer vocal tract models are compared to actual speech recordings. In particular, formant-cavity affiliations are explored using formant sensitivity functions and vocal tract simple-tube models. The difference in F4/F5 patterns between the subjects is confirmed for several additional subjects with retroflex and bunched vocal tract configurations. The results suggest that the F4/F5 differences between the variants can be largely explained by differences in whether the long cavity behind the palatal constriction acts as a half- or a quarter-wavelength resonator.
 [DOI: 10.1121/1.2902168]

PACS number(s): 43.70.Bk, 43.70.Fq, 43.70.Aj [BHS] Pages: 4466–4481

I. INTRODUCTION

It is well known that different speakers may use very different tongue configurations for producing the rhotic /r/ sound of American English (Delattre and Freeman, 1968; Hagiwara, 1995; Alwan *et al.*, 1997; Westbury *et al.*, 1998; Espy-Wilson *et al.*, 2000; Tiede *et al.*, 2004). While the picture of variability in tongue shape is complex, it is generally agreed that two shapes, in particular, exhibit the greatest degree of contrast: "retroflex" /r/ (produced with a raised tongue tip and a lowered tongue dorsum) and "bunched" /r/ (produced with a lowered tongue tip and a raised tongue dorsum). Figure 1 shows examples of these shapes drawn from our own studies of two different speakers producing their natural sustained /r/ (as in "pour"). Similar examples of this contrast may be found from Delattre and Freeman (1968) and Shriberg and Kent (1982). These examples are typical in showing three supraglottal constrictions along the vocal tract: a narrowing in the pharynx, a constriction along the palatal vault, and a constriction at the lips. However, the locations of constrictions and the degrees and lengths of constriction significantly differ, especially along the palate. At first glance, the degree of difference between the two configuration types for /r/ appears to be similar to that between, say, /s/ and /ʃ/ or /i/ and the unrounded central vowel /ɨ/. Thus, it might be expected that the two types of /r/ would show clear acoustic and perceptual differences. However, the question of an acoustic correlation between formant frequencies and tongue shape was investigated by Delattre and Freeman (1968) and, more recently, by Westbury *et al.* (1998).

[a]Electronic mail: zxinhui@glue.umd.edu
[b]Electronic mail: espy@glue.umd.edu
[c]Electronic mail: boycese@uc.edu
[d]Electronic mail: tiede@haskins.yale.edu
[e]Electronic mail: Christy.Holland@uc.edu
[f]Electronic mail: Ann.Choe@uc.edu

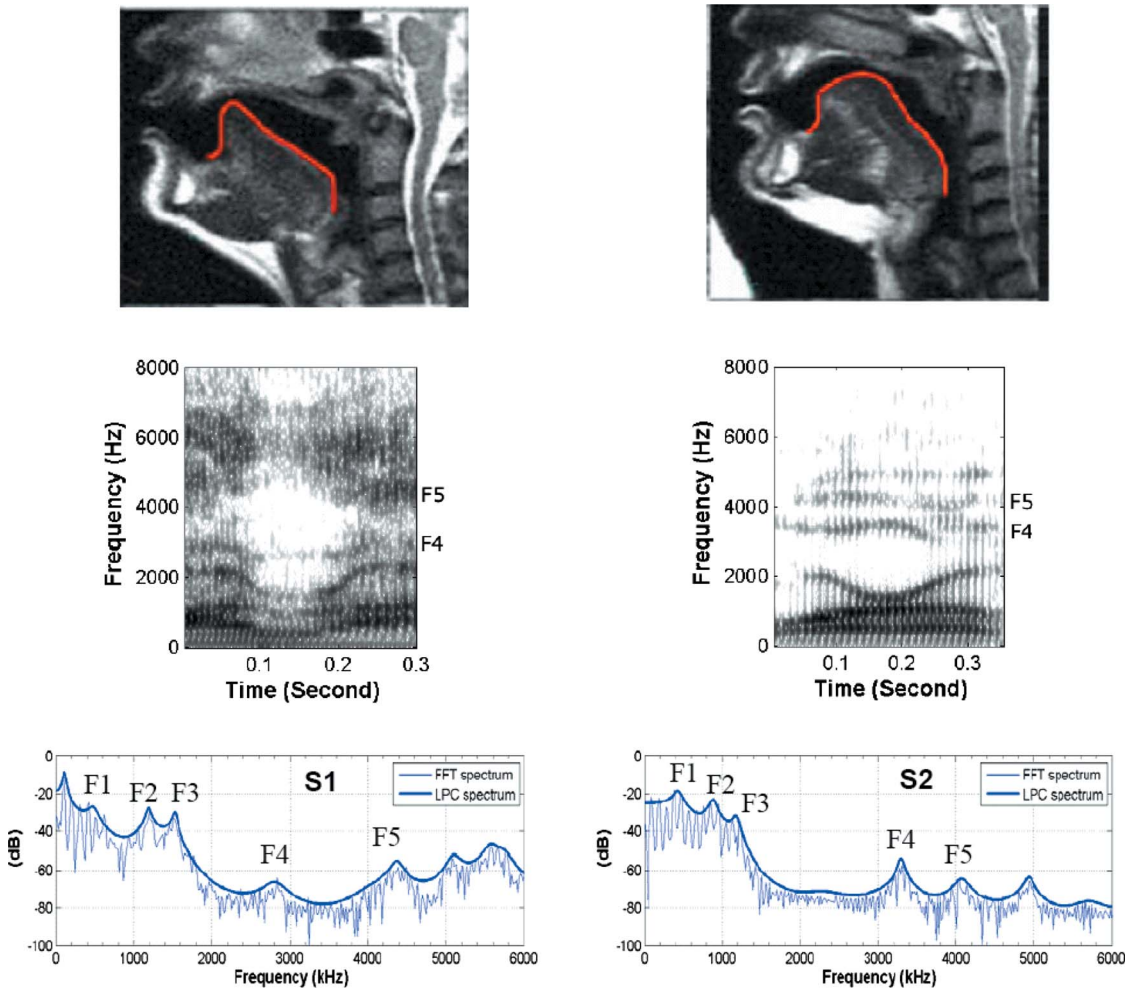

FIG. 1. (Color online) Top panel: Midsagittal MR images of two tongue configurations for American English /r/. Middle panel: Spectrograms for nonsense word "warav." Lower panel: Spectra of sustained /r/ utterance. The left side is for S1 and the right side is for S2.

Interestingly, no consistent pattern was found. In a recent perceptual study, Twist *et al.* (2007) found that listeners also appear to be insensitive to the difference between retroflex and bunched /r/.

American English /r/ is characterized by a lowered third formant frequency (F3) sitting in the region between 60% and 80% of average vowel F3 (Hagiwara, 1995) and often approaching F2 (see Lehiste, 1964; Dalston, 1975; Espy-Wilson, 1987). This low F3 is the most salient aspect of the acoustic profile of /r/ (Lisker, 1957; O'Connor *et al.*, 1957). F1 and F2 typically cluster in the central range of a particular speaker's vowel space, consistent with the common symbol of hooked schwa (or schwar) for /r/ when it acts as a syllabic nucleus.

As noted above, previous attempts have failed to find a correlation between formant frequency values and tongue shapes for /r/. However, these previous studies focused on the first three formants, F1–F3. In recent years, Espy-Wilson *et al.* have suggested that the higher formants may contain cues to tongue configuration and vocal tract dimensions (Espy-Wilson and Boyce, 1999; Espy-Wilson, 2004). Typically, researchers have not looked at higher formants such as F4 and F5 because their lower amplitude in the spectrum can make them difficult to identify and measure. In addition, the process of speech perception appears to largely depend on the pattern of the first three formants. However, higher formants are particularly responsive to smaller cavities in the vocal tract (e.g., piriform sinuses, sublingual spaces, the laryngeal cavity), and thus may give more detailed information regarding the vocal tract shape. Such knowledge may contribute to human speech perception and speaker identification to some extent. In addition, detailed knowledge of the vocal tract shape from acoustics is desirable for automatic speech and speaker recognition purposes.

In this paper, we investigate a case of two subjects with similar vocal tract anatomy who produce very different bunched and retroflex tongue shapes for /r/. These are the subjects shown in Fig. 1. As the middle panel of the figure shows, the subjects' acoustic profiles resemble those discussed by Delattre and Freeman (1968) and Westbury *et al.* (1998) in that their F1–F3 values are similar. However, the two subjects also show very different patterns for F4 and F5. In particular, the distance between F4 and F5 for the retroflex /r/ is double that of the bunched /r/. The lower panel of Fig. 1 shows examples of the same F4/F5 pattern drawn from running speech, this time from production of the nonsense word /warav/. In this paper, we investigate the question of whether different patterns of the higher formants are a con-

sistent feature of bunched versus retroflex tongue shape. If so, this difference in acoustic signatures may be useful for a number of purposes that involve the mapping between articulation and acoustics, i.e., speaker recognition, articulatory training, speech synthesis, etc. Alternatively, the different patterns of F4 and F5 may derive from structures independent of tongue shape, for instance, additional cavities in the vocal tract such as the laryngeal vestibule (Kitamura *et al.*, 2006; Takemoto *et al.*, 2006a) or the piriform sinuses (Dang and Honda, 1997). The key piece of evidence is whether such structures differ in such a way as to explain the F4/F5 patterns across /r/ types.

In this paper, we approach the task of understanding this difference in formant pattern in the following way. First, magnetic resonance imaging (MRI) is used to acquire a detailed three-dimensional (3D) geometric reconstruction of the vocal tract. Second, we used the finite element method (FEM) to simulate the acoustic response of the 3D vocal tract and to study wave propagation properties at different frequencies. Third, we derive area function models from the FEM analysis of our 3D geometry. The resulting simulated acoustic response is verified against the 3D acoustic response. The area function models are then used to isolate the effects of formant-cavity affiliations. The results of the simulation are compared to actual formant values from the subjects.

II. MATERIALS AND METHODOLOGIES

A. Subjects

The data discussed in this paper were obtained as part of a larger study on the variety of tongue shapes in productions of American English /r/ and /l/. For the purposes of this paper, we concentrate on /r/ data from two native speakers of American English, referred to here as S1 and S2.[1] As Fig. 1 shows, S1 produces a retroflex /r/ and S2 produces a bunched /r/. Both subjects are male. S1 was 48 years old and S2 was 51 at the time the data were collected. S1 had lived in California, Minnesota, and Connecticut and S2 had lived in Texas, Massachusetts, and Southwestern Ohio. Both spoke a rhotic dialect of American English.[2] The subjects were similar in palate length, palate volume, overall stature, and vocal tract length (see Table I).[3] We also compare the data from S1 and S2 to that from other subjects with similar retroflex or bunched tongue shapes for /r/ collected in the larger study. These subjects are referred to as S3–S6. The articulatory data collected for all subjects include MRI scans of the vocal tract for sustained natural /r/, dental cast measurements, computed tomography (CT) scans of the dental casts, and acoustic recordings made at various points in time.

B. Image acquisitions

MR imaging was performed on a 1.5 T G.E. Echospeed MR scanner with a standard phased array neurovascular coil at the University Hospital of the University of Cincinnati, OH. Subjects were positioned in supine posture, with their heads supported by foam padding to minimize movement. The subjects were instructed to remain motionless to the extent possible during and between scans. For hearing protection and comfort, subjects wore earplugs during the entire session. In addition, the subjects' ears were covered by padded earphones.

TABLE I. Dimension sizes of S1 and S2 in overall height, and volume, length, depth, and width of the palate. The measurements of the palate are based on the dental casts of the subjects. The width of the palate is the distance between edges of the gum between the second premolar and the first molar on both sides of the upper jaw. The length of the palate is the distance of the edges of the gum between the upper middle two incisors and the cross section of the posterior edge of the back teeth. The depth of the palate is the distance from the floor of the mouth to the cross section with the lateral plane. The volume of the palate is the space surrounded by the margin between the teeth and gums, the posterior edge of the back teeth, and the lateral plane. We used several techniques to calculate the volume, all of which gave the same answer within a certain range, and the average volume as a matter of displacement in water is reported here. That measure was done three times.

	S1	S2
Height of subject	188 cm	188 cm
Length of palate	35.8 mm	33.6 mm
Depth of palate	16.1 mm	13.2 mm
Width of palate	25.5 mm	25.0 mm
Av. volume of palate	29.1 mm^3	29.1 mm^3
Maxillary teeth volume	3.4 mm^3	3.3 mm^3

Localization scans were performed in multiple planes to determine the optimal obliquities for orthogonal imaging. A midsagittal plane was identified from the brain morphology. Axial and coronal planes were then oriented to this midsagittal plane. During each subsequent scan, the subject was instructed to produce sustained /r/ as in "pour" for a defined period of time (between 5 and 25 s depending on the sequence). T2 weighted 5 mm single shot fast spin echo images were obtained in the midline sagittal plane with two parasagittal slices. T1 weighted fast multiplanar spoiled gradient echo images (repetition time (TR) of 100–120 ms, echo delay time (TE) of 4.2 ms, 75° flip angle) were obtained in the coronal and axial planes with a 5 mm slice thickness. There was no gap between adjacent slices. The scanning regions for the coronal and axial planes include the region from the surface of the vocal folds to the velopharyngeal port and the region from the rear wall of the velopharynx to the outside edge of the lips. Depending on the dimensions of the subjects' vocal tract, the data set comprised 24–33 images in the axial and coronal planes. For all images, the field of view was 240×240 mm^2 with an imaging matrix of 256×256 to yield an in-plane resolution of 0.938 mm per pixel.

The MR imaging technique we used does not distinguish between bony structures such as teeth and air due to the low levels of imageable hydrogen. Thus, to avoid overestimation of oral tract air space, CT scans of each subject's dental cast were acquired on a GE Lightspeed Ultra multidetector scanner with a slice thickness of 1.25 mm, subsequently superimposed on the volumes derived from MRI as described below. Images were resampled to 1.25 mm at 0.625 mm intervals to optimize 3D modeling. The field of view was 120 mm with an imaging matrix of 512×512 to yield an in-plane image resolution of 0.234 mm per pixel.

C. Acoustic signal recording

During the MRI sessions, the subject's phonation in the supine position was recorded using a custom-designed microphone system (Resonance Technology Inc.) and continuously monitored by a trained phonetician to ensure that the production of /r/ remained consistent over the course of the experiment. Subjects were instructed to begin phonation before the onset of scanning and to continue to phonate for a period after scanning was complete. A full audio record of the session was preserved using a portable DAT tape recorder (SONY TD-800). Due to the noise emitted by the scanner during the scans, the only portions of the subject's productions of /r/ that can be reliably analyzed occur in 500 ms after phonation began, before the scanner noise commenced, and in 500 ms after the scanner noise ceased while the subject continued to phonate. The recordings are still quite noisy, but it was possible to measure F1–F3 with reasonable accuracy during most scans.

Subjects were also recorded acoustically in separate sessions in a sound-treated room by using a Sennheiser headset microphone and a portable DAT tape recorder (SONY TD-800). Subjects recorded a set of utterances encompassing sustained productions of /r/ plus a number of real and nonsense words containing /r/. As in the MR condition, subjects were instructed to produce /r/ as in "pour." In addition, they recorded sustained /r/ as in "right," "reed," and "role." For the sustained productions, subjects were recorded in both upright and supine postures. The nonsense words were "warav," "wadrav," "wavrav," and "wagrav," repeated with stress either on the first syllable or the second syllable. The real words included /r/ in word-initial, word-final, and intervocalic positions. For the real and nonsense words, subjects were recorded in the upright posture. Acoustic data recorded in the sound-proofed room are referred to as sound booth acoustic data. Recording conditions were such that, in addition to F1–F3, F4, and F5 could be reliably measured.

D. Image processing and 3D vocal tract reconstruction

We used the software package MIMICS (Materialise, 2007) to obtain a 3D reconstruction of the vocal tract. This software has been widely employed in the medical imaging field for processing MRI and CT images, for rapid prototyping, and for 3D reconstruction in surgery.

Our reconstruction proceeded in four steps. Step (1) involved segmentation between the tissue of the vocal tract and the air space inside the vocal tract for each MR image slice in the coronal and axial sets. Because the cross section of the oral cavity is best represented by the coronal slice images, and the cross section of the pharyngeal and laryngeal cavities are best represented by the axial slices, we used the following procedure to weight them. First, the segmented axial slices were transformed into a 3D model. Then, the coronal slices were overlapped with the axial-derived model. As in the study by Takemoto *et al.* (2006b), we extended the cross-sectional area of the last lip slice with a closed boundary halfway to the last slice in which the upper and lower lips are still visible. The coronal slice segmentation in the pharyngeal and laryngeal cavities was then corrected by reference to the axial slice 3D model.

Step (2) involved compensation for the volume of the teeth using the CT scans, which were made in the coronal plane. The CT images were segmented to provide a 3D reconstruction of the mandible and the maxillae with the teeth. (This process was considerably easier than for the MR slices described above, given the straightforward nature of the air/tissue boundary in that imaging modality.) The 3D reconstruction of the dental cast was then overlapped with the MRI coronal slices. The reconstruction of the maxilla cast was positioned on the MR images by following the curvature of the palate. The reconstruction of the mandible cast was positioned with reference to the boundary provided by the lips. In step (3), the final segmentation was translated into a surface model in stereolithography (STL) format (Lee, 1999). Finally, the 3D geometry surface was smoothed using the MAGICS software package (Materialise, 2007). The validity of the reconstructed 3D vocal tract geometry was evaluated by comparing midsagittal slices created from the reconstructed 3D geometry to the original midsagittal MR images. We also used this method to check for the possibility that subjects had changed their vocal tract configuration for /r/ across scans. The data sets of all the subjects in this study show very good consistency, and overall boundary continuity between the tissue and the airway was successfully achieved.

As noted above, the difference in the F4/F5 formant pattern between S1 and S2 must be derived from a difference in vocal tract dimensions, either in small structures such as the piriform sinuses and laryngeal vestibule (Dang and Honda, 1997; Kitamura *et al.*, 2006; Takemoto *et al.*, 2006a) or in tongue shape differences. The laryngeal vestibule cavities were included in the 3D model, but given the resolution of the MR data, the representation is relatively crude. The dimensions of the piriform sinuses were measured and found to be similar to the range in length of 16–20 mm and in volume of 2–3 cm^3 reported by Dang and Honda (1997).[4] Because no significant differences were found between the subjects for either structure, we conclude that the tongue shape differences between S1's retroflex and S2's bunched /r/ are likely the major factor determining their differences in the F4/F5 pattern. Possibly, these cavities at the glottal end of the vocal tract are less influential for /r/ than for vowels due to the greater number, length, and narrowness of constrictions involved.

E. 3D finite element analysis

The FEM analysis was used in this study to obtain the acoustic response of the 3D vocal tract and to obtain the wave propagation at different frequencies. The pressure isosurfaces at low frequency were used to extract area functions. The governing equation for this harmonic analysis is the Helmholtz equation,

$$\nabla \cdot \left(\frac{1}{\rho} \nabla p \right) + \frac{\omega^2 p}{\rho c^2} = 0, \quad (1)$$

where p is the acoustic pressure, ρ (1.14 kg/m^3) is the density of air at body temperature, c (350 m/s) is the speed of sound, and ω is the angular frequency ($\omega = 2\pi f$, where f is the vibration frequency in hertz and the highest frequency in our harmonic analysis is 8000 Hz). The boundary conditions for the 3D finite element analysis are as follows: for the glottis, a normal velocity profile as sinusoidal signal at various frequencies; for the wall, rigid; for the lips, the radiation impedance Z of an ideal piston in an infinitely flat baffle (Morse and Ingard, 1968),

$$Z = \rho c(1 - J_1(2ka)/(ka) + jK_1(2ka)/(2ka)), \quad (2)$$

where $k = 2\pi f/c$, $a = \sqrt{A_1/\pi}$ (A_1 is the area of the lips opening), J_1 is the Bessel function of order 1, and K_1 is the Struve function of order 1. The volume velocity at the lips is measured by velocity integration over the cross section at the lips, and the acoustic response of the vocal tract is defined as the volume velocity at the lips divided by the volume velocity at the glottis. Note that for the purpose at hand, the ideal piston model has been shown to be computationally equivalent to a 3D radiation model at the lips (Matsuzaki *et al.*, 1996).

The finite element (FEM) analysis was performed using the COMSOL MULTIPHYSICS package (Comsol, 2007). The mesh for FEM was created using tetrahedral elements as in the STL format.

F. Area function extraction

Area functions are generated by treating the vocal tract as a series of uniform tubes with varying areas and lengths. The extraction of area functions from imaging data is typically an empirical process. Baer *et al.* (1991), Narayanan *et al.* (1997), and Ong and Stone (1998) based their area function extractions on a semipolar grid (Heinz and Stevens, 1964). In contrast, Chiba and Kajiyama (1941), Story *et al.* (1996), and Takemoto *et al.* (2006b) extracted area functions by computing a centerline in air space and then evaluating the cross-sectional areas within planes chosen to be perpendicular to the centerline extending from the glottis to the mouth.

In general, because our area functions were derived from the 3D FEM, it might be expected that the area function simulation and the simulated acoustic response from the 3D model should be the same. However, it should be noted that area function extraction, by transforming the bent 3D geometry of the vocal tract into a straight tube with varying cross-sectional areas (Chiba and Kajiyama, 1941; Fant, 1970), necessarily involves considerable simplification. An additional and related problem is that it assumes planar wave propagation, and thus tends to neglect cross-mode wave propagation and potential antiresonances or zeros. Thus, we expect some small differences between the simulation results using area function analysis and planar wave propagation from simulation results directly obtained from the corresponding 3D geometry (Sondhi, 1986).

In this study, we used the low-frequency wave propagation properties resulting from the 3D finite element analysis to guide the area function extraction from the reconstructed 3D geometry. This approach is quite similar to the centerline approach. The logic of this procedure was as follows. As noted above, area-function-based vocal tract models assume planar wave propagation. Finite element analysis at low frequencies such as 400 Hz (around F1 for /r/) produces pressure isosurfaces that indicate approximate planar acoustic wave propagation. Thus, a tube model derived from area functions whose cutting plane follows these pressure isosurfaces should constitute a reasonable one-dimensional model for the 3D vocal tract. In this study, as the curvature of the vocal tract changes, the cutting orientation in our method was adjusted to be approximately parallel to the pressure isosurface at 400 Hz. This procedure was performed by recording the coordinates of the isosurfaces. Those coordinates are then used to determine the cutting planes. The distance between two sampling planes was set to be the distance between their centroids. The vocal tract length was estimated as the cumulative sum of the distance between the centroids. The cutting plane gap was about 3 mm. Since this method was based on the 3D reconstructed geometry instead of sets of MR images, pixel counting and other manipulations such as reslicing of images were not needed. The area calculation was based on the geometric coordinates of the reconstructed vocal tract.

As noted above, the reduction of a vocal tract 3D model to area functions requires considerable simplification. To assess the degree to which our area function extraction preserved essential aspects of the vocal tract response, we compared the simulation output from the 3D FEM to the acoustic response of Vocal Tract Acoustic Response (VTAR), a frequency-domain computational vocal tract model (Zhou *et al.*, 2004) which takes area functions of the vocal tract as input parameters and includes terms to account for energy losses due to the yielding wall property of the vocal tract, the viscosity and the heat conduction of the air, and the radiation from the lips. The vocal tract response from the 3D model and from VTAR were, in turn, evaluated by comparison with formant measurements from real speech produced by the subjects, as described below.

G. Formant measurement of /r/ acoustic data

Formants from both sound booth and MR acoustic recordings were measured by an automatic procedure that computed 24th order LPC (Linear Prediction Coding) spectrum over a 50 ms window from a stable section of the sustained production. The 50 ms window for the MR acoustic data was taken from the least noisy segment of the approximately 500 ms production preceding the onset of MR scanning noise. Only F1–F3 were measured in the MR acoustic recording because the noise in the high-frequency region masked the higher formants very effectively. Both sets of measurements are shown in Tables III and IV. To maximize the comparability of the MR and sound booth acoustic measures, the latter were measured from productions recorded when the subjects were in supine posture. The formant val-

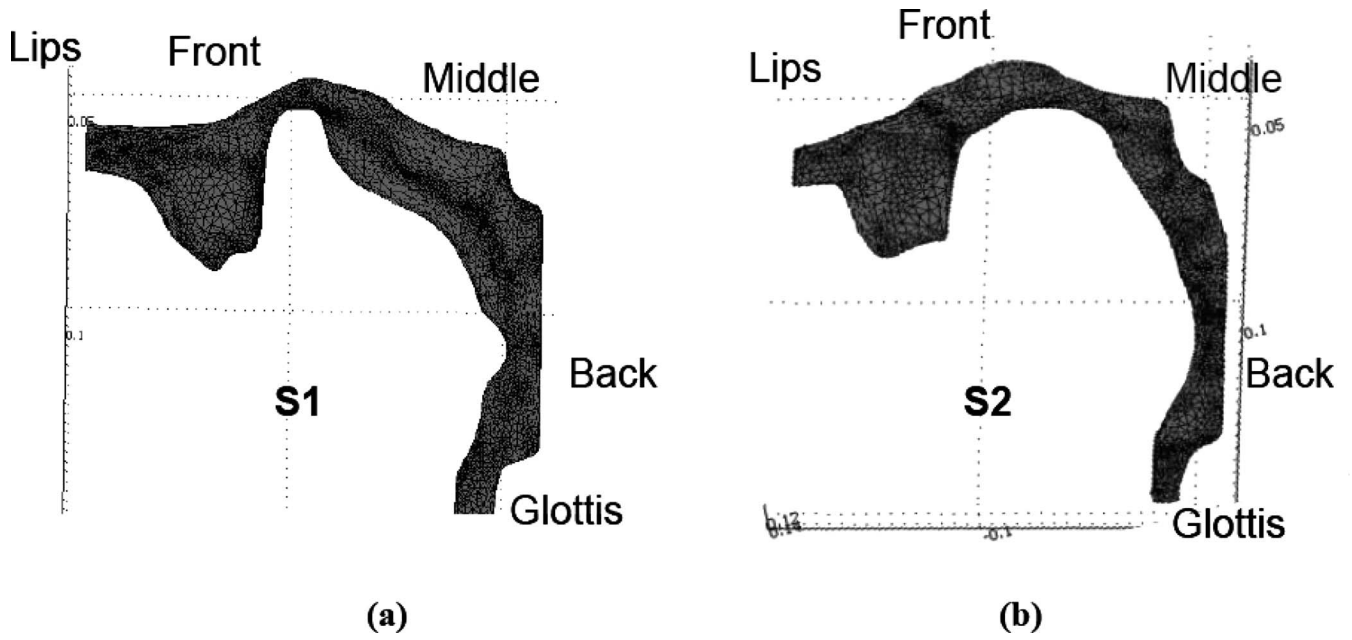

FIG. 2. FEM mesh of the reconstructed 3D vocal tract. (a) The retroflex tongue shape. (b) The bunched tongue shape.

ues of the sustained /r/ in MRI sessions are the average of the measurements from all the scans including midsagittal, axial, and coronal scans.

The difference in F4/F5 pattern between the subjects previously alluded to is clearly shown in Tables III and IV for the sound booth recording of sustained /r/. As Fig. 1 shows, the pattern in question was even more strongly evident in the more dynamic real word condition. We concluded from these data that the patterns shown in sustained /r/ are representative of patterns shown in running speech.

H. Reconstructed 3D vocal tract geometries

The reconstructed 3D vocal tract shapes for the retroflex /r/ of S1 and the bunched /r/ of S2 are shown in Fig. 2. The two shapes are significantly different in several dimensions that are likely to cause differences in cavity affiliations. First, S1's retroflex /r/ has a shorter and more forward palatal constriction, leading to a slightly smaller front cavity. At the same time, the lowered tongue dorsum of the retroflex /r/ leads to a particularly large volume of the midcavity between the palatal and pharyngeal constrictions. Further, the transition between the front and midcavities is sharper for the retroflex /r/. This difference makes it more likely that the front and midcavities are decoupled for the retroflex /r/ of S1 than for the bunched /r/ of S2. Unlike the speakers analyzed by Alwan *et al.* (1997) and Espy-Wilson *et al.* (2000), neither S1 nor S2 shows a sublingual space whose geometry is clearly a side branch to the front cavity. However, the two subjects' overall vocal tract dimensions from the 3D model are very similar. These dimensions are shown in Table II.

I. FEM-based acoustic analysis

In previous work, FEM analysis has been used to study the acoustics of the vocal tract for open vocal tract sounds, i.e., vowels (Thomas, 1986; Miki *et al.*, 1996; Matsuzaki *et al.*, 2000; Motoki, 2002). Zhang *et al.* (2005) applied this approach to a two-dimensional vocal tract for a schematized geometry based on a single subject producing /r/. In this study, we extend the work of Zhang *et al.* (2005) by computing the pressure isosurfaces at various frequencies to 3D vocal tract shapes based on S1's retroflex and S2's bunched /r/. As Fig. 3 shows, the retroflex and bunched /r/ shapes have similar wave propagation. For both, as expected, the wave propagation is almost planar up to about 1000 Hz. Between 1500 and 3500 Hz, a second wave propagates almost vertically to the bottom of the front cavity. Above 4500 Hz, the isosurface becomes more complex and part of the acoustic wave propagates to the two sides of the front cavity. The results show that the wave propagation property should be kept in mind when assuming planar wave propagation along the vocal tract, particularly for antiresonances. Note that for both subjects, F4 and F5 occur in the transition region below 4500 Hz. This will be discussed later. The cutting orientations for the area functions based on the pressure isosurfaces are shown in the upper panel of Fig. 4 as grid lines. The area functions themselves are shown in the lower panel of Fig. 4.

III. RESULTS

The purpose of this study was to determine if the F4/F5 difference in pattern between bunched and retroflex /r/ occurs as a result of tongue shape differences. The approach involves comparing the results of calculations to acoustic

TABLE II. Measurements on the reconstructed 3D vocal tract in surface model (STL file format).

	S1	S2
X dimension	51 mm	46 mm
Y dimension	106 mm	107 mm
Z dimension	106 mm	100 mm
Volume	62 909 mm^3	48 337 mm^3
Surface area	14 394 mm^2	12 243 mm^2

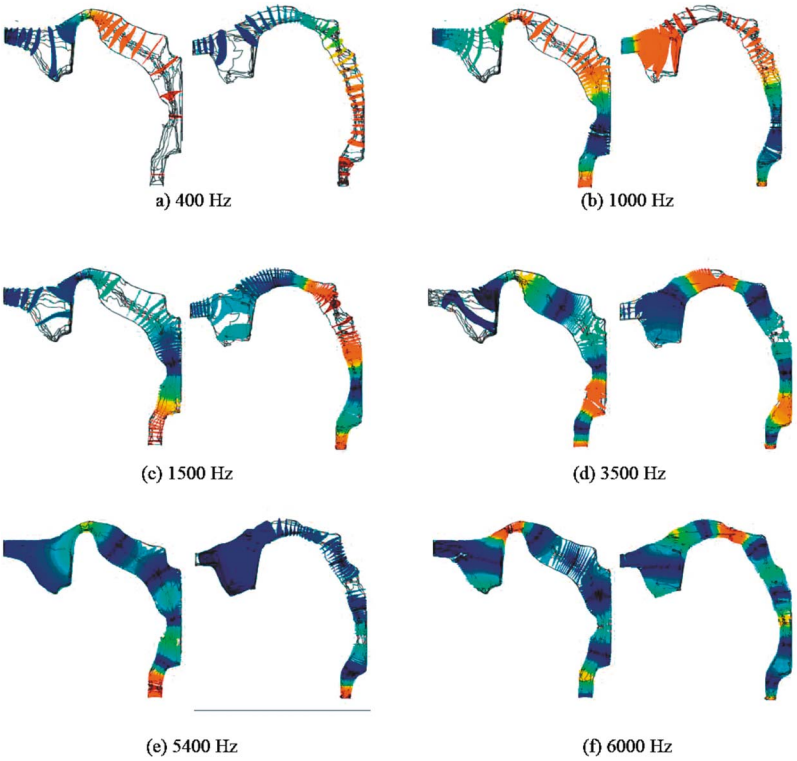

FIG. 3. (Color online) Pressure isosurface plots of wave propagation inside the vocal tracts of the retroflex /r/ (S1 on the right side) and the bunched /r/ (S2 on the right side) at different frequencies. (Pressure isosurfaces are coded by color: the red color stands for high amplitude and the blue color stands for low amplitude.) (a) 400 Hz, (b) 1000 Hz, (c) 1500 Hz, (d) 3500 Hz, (e) 5400 Hz, and (f) 6000 Hz.

spectra from actual productions by the subjects during (a) MR and (b) sound booth acoustic sessions, respectively. The calculated results include (c) generating an acoustic response from the FEM analysis based on the 3D model, (d) generating an acoustic response from the VTAR computational model using FEM-derived area functions, (e) generating sensitivity functions for better understanding of formant-cavity affiliations and manipulating the VTAR computation model to isolate the effects of particular cavities and constrictions, and (f) generating simple-tube models to understand the types of resonators that produce the formants. The FEM analysis makes no assumptions regarding planar wave propagation, whereas the area functions are derived from cutting planes determined by the FEM at low frequency. The isolation of cavity/constriction influences is done by using VTAR to synthesize changes in the dimensions of a particular cavity/constriction while holding the rest of the vocal tract constant. In effect, we compare the acoustic responses from the 3D FEM and the area functions with the subjects' actual production.

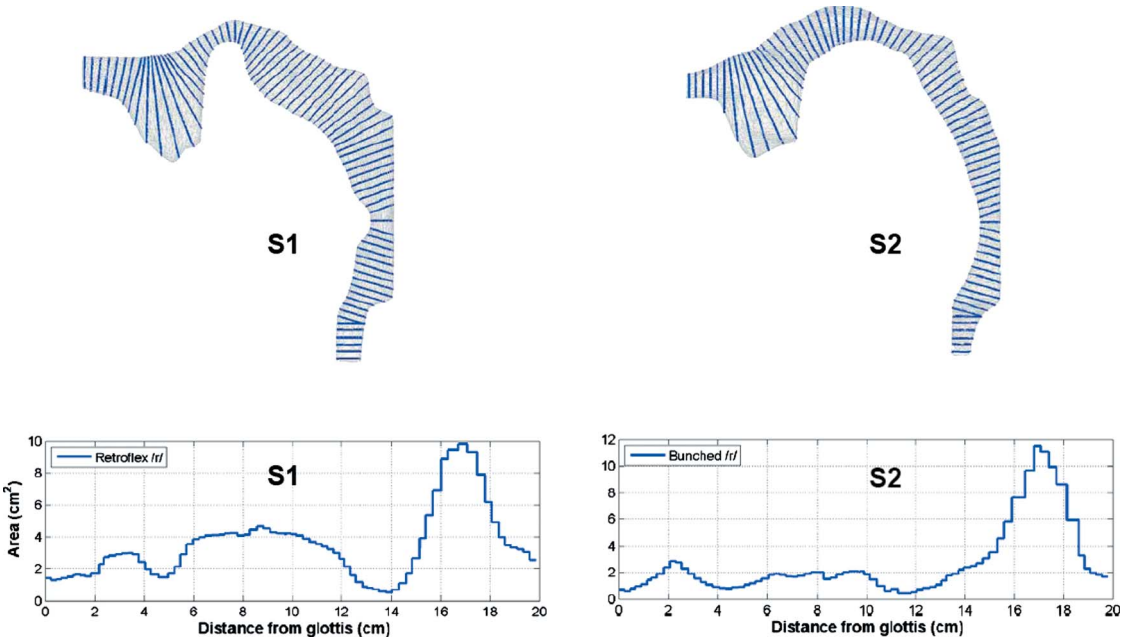

FIG. 4. (Color online) Top panel: Grid lines for area function extraction inside the vocal tract. Lower panel: Area function based on the grid lines. (In each panel, the left side is for S1 and the right side is for S2.)

TABLE III. Formants measured from S1's retroflex /r/ compared with calculated values from the 3D FEM, tube model with area function model, and simple-tube model, respectively (Unit: Hz). The percentage difference between the FEM formant values and the actual subject formant values from MR (Δ1) and sound acoustic (Δ2) sessions are also given. Note that due to background noise, only F1–F3 could be consistently measured from the MRI acoustic data.

	Retroflex /r/ (S1)							
	MRI acoustic data	Second both supine position	Sound booth upright position	3D FEM			Area function tube model	Simple tube model
				Formant	Δ1 (%)	Δ2 (%)		
F1	522	391	438	380	27.2	2.81	383	418
F2	1075	1234	1188	1160	7.91	6.0	1209	1262
F3	1534	1547	1563	1580	3.0	2.13	1609	1660
F4		2797	2828	2940		5.11	3002	2936
F5		4328	4234	4280		1.11	4366	4233
F5-F4		1531	1406	1340			1364	1297

MR versus sound booth acoustic data. Because the FEM analysis and area functions are both based on MR data, the F4/F5 patterns would ideally have been extracted from the simultaneously recorded acoustic signal ("MR acoustic data"). As noted previously, however, F4 and F5 are masked in the MRI condition by the noise of the scanner. Hence, acoustic data recorded in a sound booth (from the supine posture) were used for comparisons with the calculated acoustic response results. Comparison between the MR and sound booth acoustic data for the first three formants show that the subjects' productions are, for the most part, highly similar, as shown in Tables III and IV. There are notable deviations in the F1 and F2 produced by S1 and in the F3 produced by S2. While these differences probably indicate a slight difference in articulatory configuration for sustained /r/, this same alternation between formant values can also be seen in their running speech for both real and nonsense words.[5] In all cases, the characteristic F4/F5 pattern is maintained.

The difference in F4/F5 patterns between the retroflex configuration of S1 and the bunched configuration of S2 is also observed when subjects produce /r/ in the upright posture. This is shown for running speech in Fig. 1. In addition, the formant values from sound booth acoustic sustained productions recorded in the upright posture are reported in Tables III and IV, for comparison to the values recorded in supine posture.

Comparison of actual formants to acoustic response from FEM and area function. In Fig. 5, spectra from the subjects' actual productions are shown along with acoustic responses from the models for S1 and S2. As shown in Figs. 5(a) and 5(c) (in addition to Tables III and IV), the FEM provides formant values for F1–F3 similar to those measured from actual productions in MRI sessions by each speaker. The percentage differences (between modeled and measured acoustics) are also given in Tables III and IV. As Fig. 5(b) and Tables III and IV also show, the spacing between F4 and F5 in the sound booth data for actual speaker production is much larger for the retroflex /r/ than for the bunched /r/ (a difference of 1531 Hz versus 796 Hz for the supine position, and 1469 Hz versus 651 Hz for the upright position). Notably, the FEM also replicates this pattern of different spacing between F4 and F5. A similar difference in spacing is also predicted by the VTAR computer model using the extracted area functions (see Tables III and IV). Thus, these results support our methods for deriving a 3D model. They also

TABLE IV. Formants measured from S2's bunched /r/ compared with calculated values from the 3D FEM, area function model, and simple-tube model, respectively (Unit: Hz). The percentage difference between the FEM formant values and the actual subject formant values from MR (Δ1) and sound acoustic (Δ2) sessions are also given. Note that due to background noise, only F1–F3 could be consistently measured from the MRI acoustic data.

	Bunched /r/ (S2)							
	MRI acoustic data	Second both supine position	Sound booth upright position	3D FEM			Area function tube model	Simple tube model
				Formant	Δ1 (%)	Δ2 (%)		
F1	445	453	391	480	7.87	5.96	457	472
F2	1008	906	891	1040	3.17	14.79	998	1047
F3	1469	1203	1219	1660	13.0	37.99	1626	1680
F4		3313	3281	3260		1.60	3330	3190
F5		4109	4016	4000		2.65	3912	3841
F5-F4		796	735	740			582	651

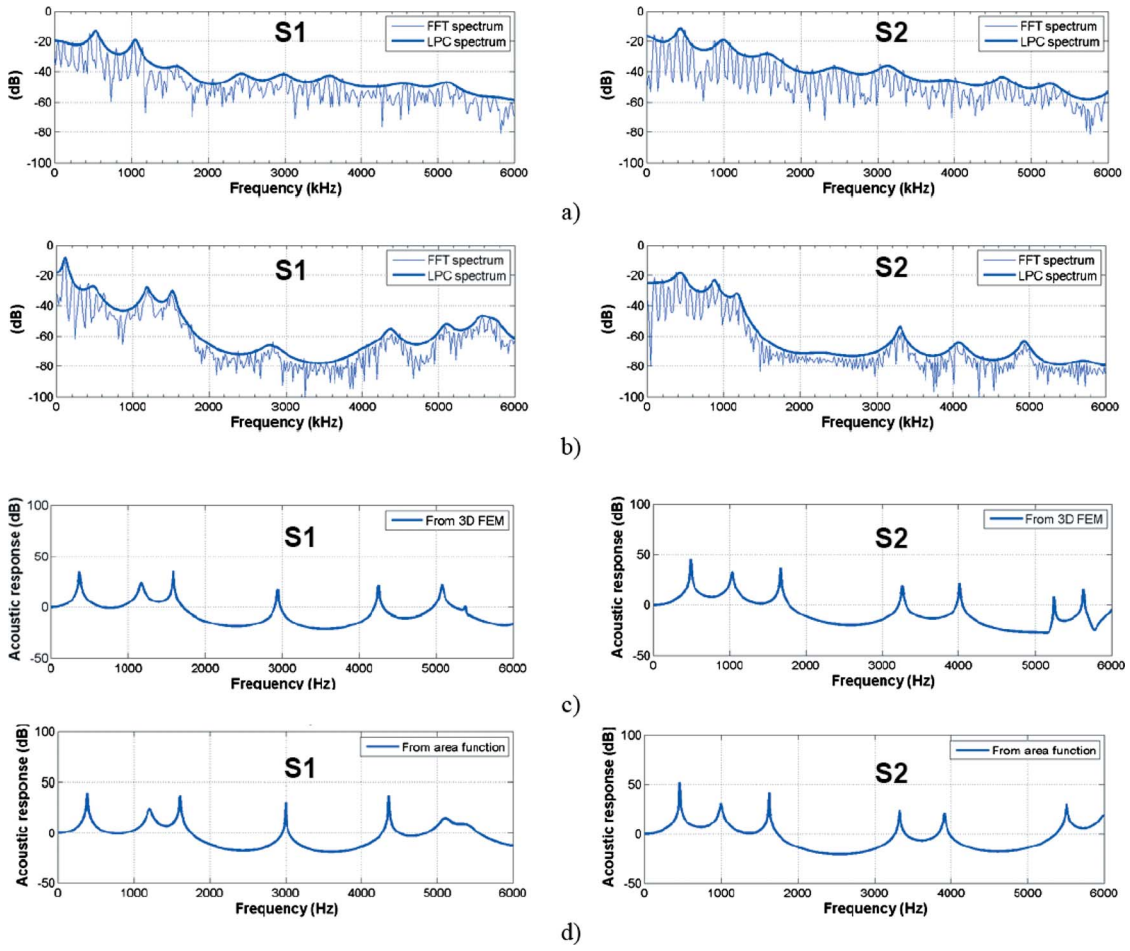

FIG. 5. (Color online) For S1 and S2: (a) Spectrum of sustained /r/ utterance in MRI session, (b) spectrum of sustained /r/ utterance in the sound booth acoustic data, (c) the acoustic response based on 3D FEM, and (d) the acoustic response based on the area function.

suggest that the source of the differences in the F4/F5 pattern between the bunched and retroflex /r/ follows from their respective differences in overall tongue shape.

A. FEM-derived area functions

Spectra generated from 3D FEM and area function sources are shown in Figs. 5(c) and 5(d). Formant values generated are shown in Tables III and IV. Both comparisons show that the results from the two methods match within 5% of each other. Note, however, that although the FEM produces zeros above 5000 Hz, they are not produced by the area function vocal tract model because it does not contain side branches and is based on only plane wave propagation.

B. Sensitivity functions and simple-tube modeling based on FEM-derived area functions

To gain insight into formant-cavity affiliations, the area function models were used to obtain sensitivity functions for F1–F5. Additionally, the area function models were simplified to arrive at models consisting of 3–8 sections (as opposed to about 70 sections) in order to gain insight into the types of resonators from which the formants originate and the effects of area perturbations of these resonators. These will be referred to as simple-tube models.

1. Sensitivity functions for F1–F5

The sensitivity functions of the formants are calculated as the difference between the kinetic energy and potential energy at the formant frequency as a function of distance starting from the glottis, divided by the total energy of kinetic and potential energies in the system (Fant and Pauli, 1974; Story, 2006). The relative change of the formant that corresponds to the change in the area function can be described as

$$\frac{\Delta F_n}{F_n} = \sum_{i=1}^{N} S_n(i)\frac{\Delta A_i}{A_i}, \tag{3}$$

where F_n is the nth formant, ΔF_n is the change of the nth formant, S_n is the sensitivity of the nth formant, A_i is the area of the ith section, and ΔA_i is the area change of the ith section. Section I is the first section starting from the glottis, and N is the last section number at the lips.

The calculated sensitivity functions are shown in Fig. 6 (the left panel is for S1 and the right panel is for S2). At a point where a curve for a given formant passes through zero, a perturbation in the cross-sectional area will cause no shift in the formant frequency. Otherwise, the curve shows how the formant will change if the area is increased at that point. If S_n is positive at a certain point, increasing the area at that point will increase the value of the nth formant. If S_n is

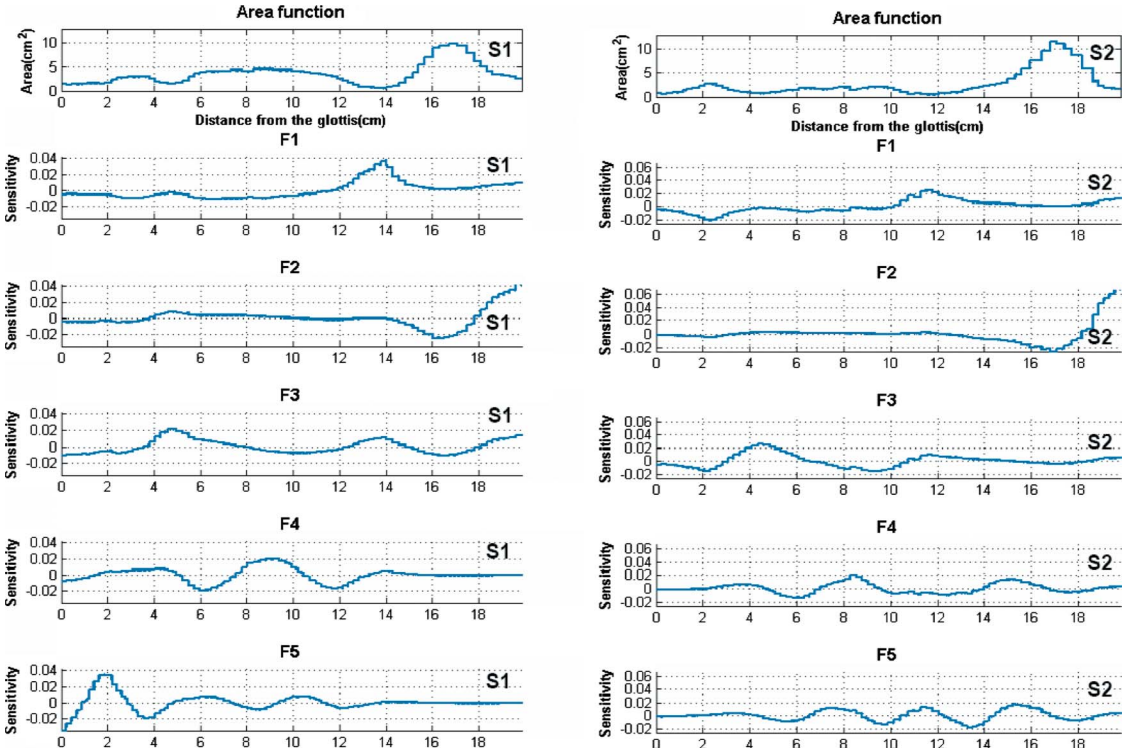

FIG. 6. (Color online) Acoustic sensitivity functions of F1–F5 for the retroflex /r/ of S1 and S2.

negative at a certain point, increasing the area at that point will decrease the value of the *n*th formant. The number of such zero crossings on a curve is equal to $2N-1$ (1, 3, 5, 7, and 9 for F1–F5, respectively) (Mrayati *et al.*, 1988), where N is the formant number for that curve.

As shown in Fig. 6, the sensitivity functions for F1–F3 have some similarities in their patterns for both the retroflex /r/ and the bunched /r/. In both cases, F2 is mainly affected by the front cavity where the lip constriction with small area plus the large posterior volume between the lip constriction and the palatal constriction act as a Helmholtz resonator. The frequency of a Helmholtz resonator is given by

$$F_H = \frac{c}{2\pi}\sqrt{\frac{A_1}{l_1 A_2 l_2}},$$

where A_1 and l_1 are the area and length of the lip constriction and A_2 and l_2 are the area and length of the large volume behind the lip constriction. From this equation, F_H will increase if the area of the lip constriction increases or if the area of the large volume behind the lip constriction decreases. The sensitivity functions for F2 show this behavior since it is significantly positive during the portion of the tube that corresponds to the lip constriction and, conversely, significantly negative during the portion of the tube that corresponds to the large volume.

This conclusion is supported by the spectra in Figs. 7 and 8. Figures 7 and 8 compare the spectra from the full vocal tract model with the spectra from the shortened vocal tract that includes only the front cavity as highlighted (acoustic responses were calculated with radiation at the lips) and the spectra from the shortened vocal tract that includes only the back cavity as highlighted (pressure on the front side is assumed to be zero). As can be seen, the first resonance of the front cavity is F2 from the full vocal tract for both subjects.

Based on the area function data of S1, Fig. 9 shows how the F2/F3 cavity affiliations switch when the front cavity volume is changed by varying its length. When the front cavity volume exceeds about 17 cm^3, there is a switch in formant-cavity affiliation between F2 and F3. The front cavity resonance is so low that it becomes F2 and the resonance of the cavity posterior to the palatal constriction becomes F3. It seems that the front cavity resonance may be F2 or F3 depending on the size of the volume of the Helmholtz resonator. This conclusion is supported by the findings from two different subjects showing bunched configurations discussed by Espy-Wilson *et al.* (2000). In that study, F3 was clearly derived from the Helmholtz front cavity resonance. However, the subjects in that study had much smaller front cavity volumes (of 5 and 8 cm^3) relative to those of the current subjects S1 and S2 (of 24 and 27 cm^3), respectively.

Due to coupling between cavities along the vocal tract, F1 and F3 of both retroflex and bunched /r/ can be affected by area perturbation along much of the vocal tract. However, there are differences. The F1 sensitivity function for S1's retroflex /r/ shows a prominent peak in the region of the palatal constriction (between 12.6 and 14.6 cm), whereas the F1 sensitivity function for S2's bunched /r/ shows a prominent peak and large positive value in the region of the palatal constriction (between 10.7 and 12.3 cm) and also a prominent peak dip in the region posterior to the pharyngeal constriction (between 1.6 and 2.8 cm). This difference in the F1 sensitivity functions of the retroflex and bunched /r/ is due to the differences in the area functions posterior to the front

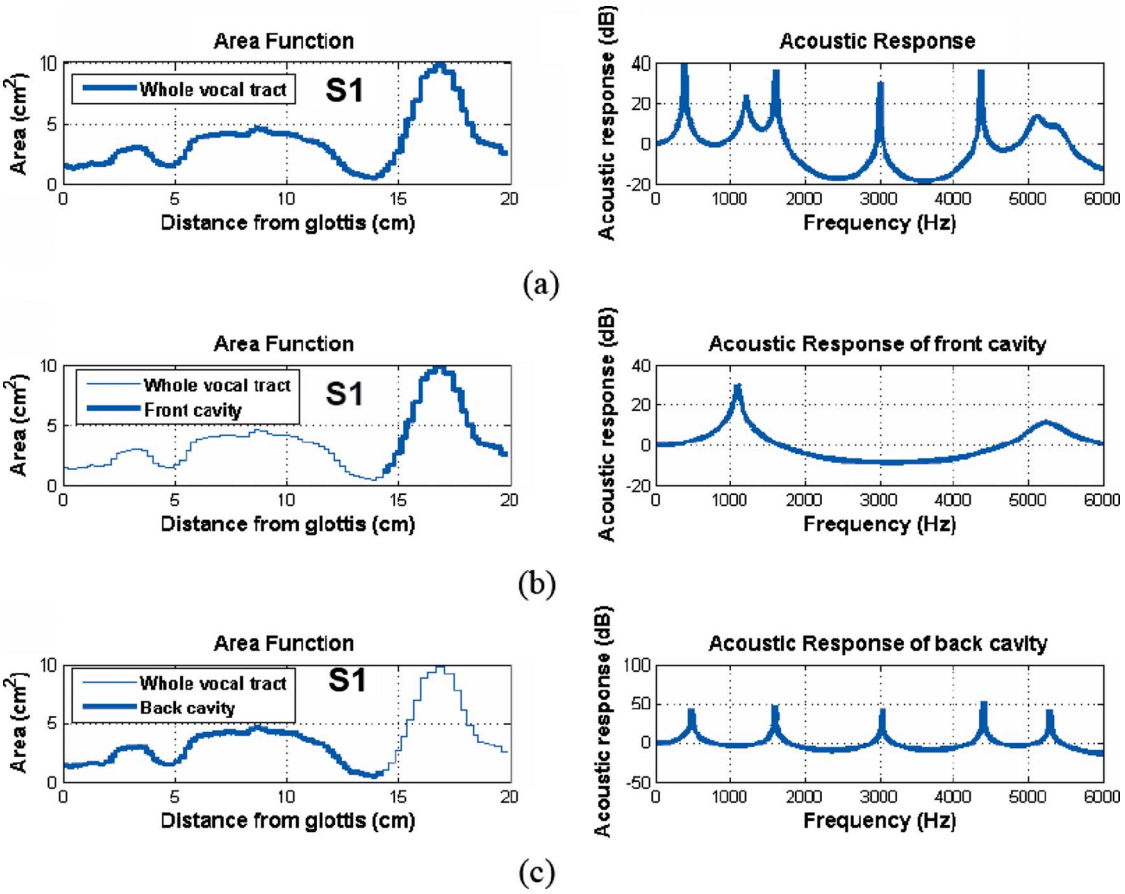

FIG. 7. (Color online) Acoustic response of S1's retroflex /r/ area function with front and back cavities separately modeled. (The left side is the area function and the right side is the corresponding acoustic response). (a) Area function of the whole vocal tract and its corresponding acoustic response. (b) Area function of the front cavity and its corresponding acoustic response. (c) Area function of the back cavity and its corresponding acoustic response.

cavity. In the retroflex /r/, the areas of the palatal constriction are much smaller than the areas of the back cavity posterior to the palatal constriction. This shape is more like a Helmholtz resonator for F1. In the bunched /r/, the overall shape of the area function posterior to the front cavity is similar to that of the retroflex /r/. However, the areas are more uniform so that F1 is the first resonance of a uniform tube (see discussion of simple-tube modeling below).

As the sensitivity functions indicate, F3 can be decreased by narrowing at each of the three constriction locations along the vocal tract. Note, however, that in both of these cases, F3 is most sensitive to the perturbation of the pharyngeal constriction. It is relatively much less sensitive to the palatal constriction and even less to the lip constriction. This result confirms the finding of Delattre and Freeman (1968) that the percept of /r/ depends strongly on the existence of a constriction in the pharynx.

Sensitivity functions for F4 and F5 have very different patterns for the retroflex /r/ and the bunched /r/. In the retroflex /r/, F4 and F5 are only minimally affected by the area perturbation of the front cavity, starting at the location about 14.8 cm from the glottis, which means that they are resonances of the cavities posterior to the palatal constriction. This conclusion is supported by the spectra in Fig. 7 which shows that the first four resonances of that part of the vocal tract behind the palatal constriction are close to F1–F5. In the bunched /r/, F4 and F5 are not sensitive to the area perturbation of the cavity posterior to the pharyngeal constriction and they are affected to some extent by the front cavity. Again, this sensitivity to the front cavity is probably due to a higher degree of coupling between the back and front cavities for the bunched /r/ relative to the retroflex /r/. Given the more gradual transition between the back and front parts of the vocal tract for the bunched /r/, Fig. 8 shows two possible divisions. In one case, the front cavity is assumed to start at 11.8 cm from the glottis. In the other case, it starts 2.9 cm further forward, at 14.7 cm from the glottis. In both cases, the first resonance (a Helmholtz resonance formed by the lip constriction and the large volume behind it) of the front cavity is around 1000 Hz, the frequency of F2 in the spectrum derived from the full vocal tract. However, this choice of a division point has a significant effect on the location of the second resonance (a half-wavelength resonance of the large volume between the lip constriction and the palatal constriction) from the front cavity. If the front cavity starts at 11.8 cm, the second resonance is around 3300 Hz, the region of F4 from the full vocal tract spectrum. If the front cavity starts around 14.7 cm, the second resonance of the front cavity is around 5500 Hz, which corresponds to the region around F6 in the spectrum derived from the full vocal tract.

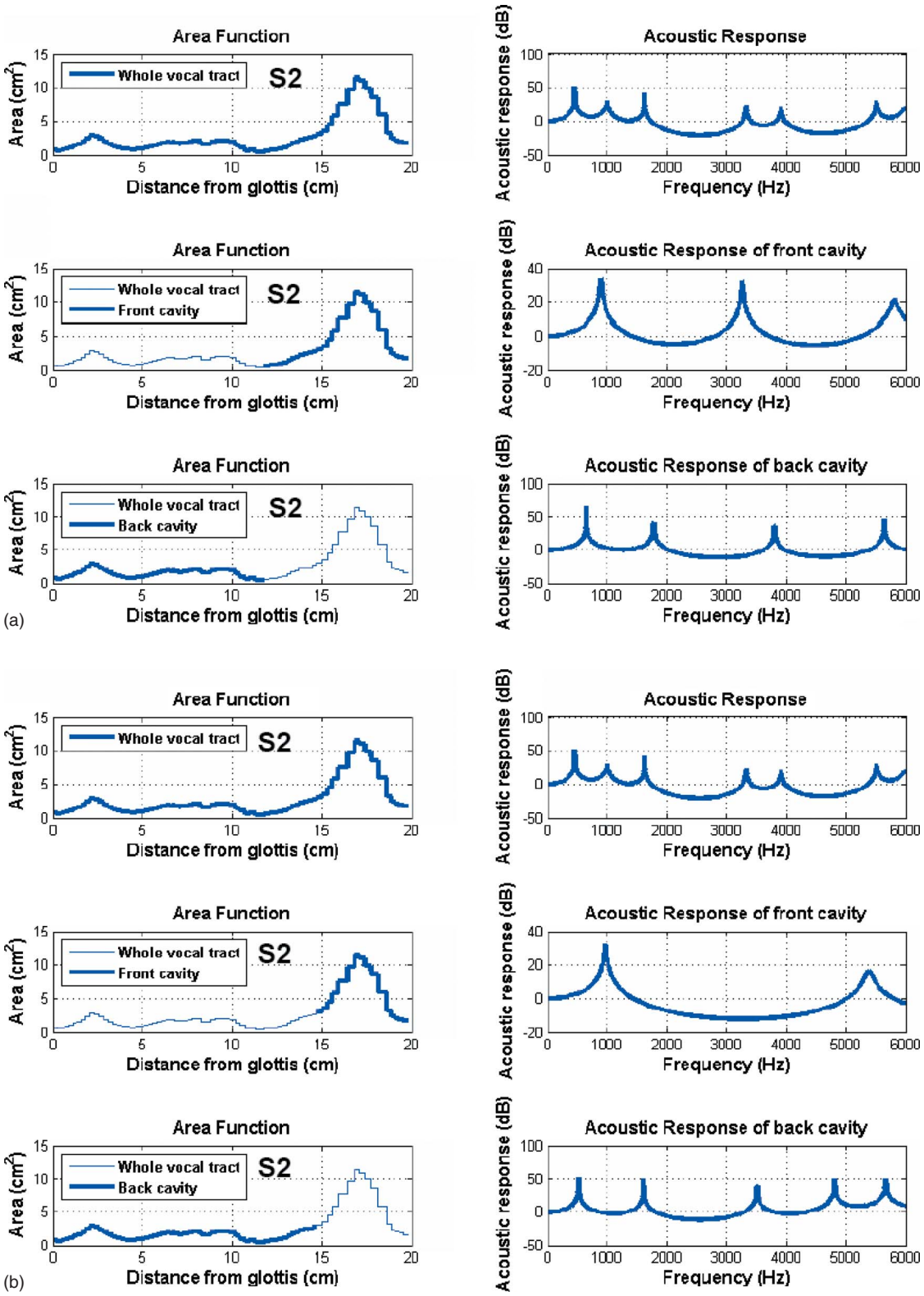

FIG. 8. (Color online) Acoustic response of S2's bunched /r/ area function with front and back cavities separately modeled. (The left side is the area function and the right side is the corresponding acoustic response). (a) The dividing point between the front cavity and the back cavity at about 12 cm. (b) The dividing point between the front cavity and the back cavity at about 15 cm.

2. Simple-tube models based on FEM-derived area functions

Figure 10 shows simple-tube models for the retroflex and bunched /r/ along with the original area functions and the corresponding acoustic responses. In the first case of the retroflex /r/, as shown in Fig. 10(a), the simple model consists of four tubes: a lip constriction, a large volume behind the lip constriction, a palatal constriction, and a long tube posterior to the palatal constriction [see Fig. 10(a)]. Henceforth, the area forward of the palatal constriction will be

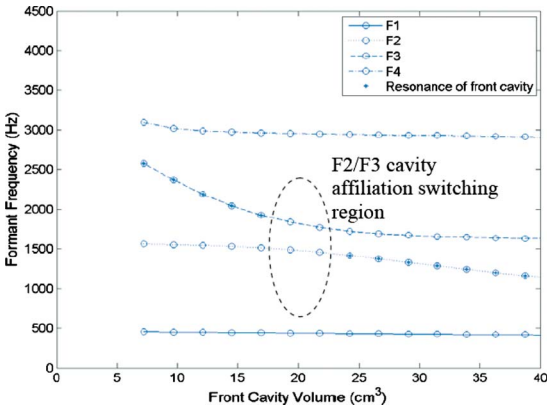

FIG. 9. (Color online) F2/F3 cavity affiliation switching with the change of the front cavity volume by varying its length (based on the area function data of S1).

referred to as the front cavity, while the area from the palatal constriction backward to the glottis will be referred to as the long back cavity. As we saw from the sensitivity functions, F2 comes from the front cavity, acting like a Helmholtz resonator at low frequencies. F1 comes from the long back cavity plus the palatal constriction, which together act as a Helmholtz resonator at low frequencies. F3–F5 are half-wavelength resonances of the long back cavity. The fact that the three formants are fairly evenly spaced [see Figs. 10(a) and 10(b)] is thus explained. Refinement of the simple tube, by allowing additional discrete sections as in Fig. 10(b), indicates that if we include the pharyngeal narrowing in our model, F3 is further lowered in frequency. In addition, if we include the narrowing in the laryngeal region above the glottis, F4 and F5 rise in frequency. The net results from these perturbations can be seen in Fig. 10(b). These formant-cavity affiliations agree well with our understanding from the sensitivity functions. Further, Tables III and IV show that there is close agreement between the formant frequencies measured from the actual acoustic data and those predicted both by the FEM-derived area functions and the simple-tube model.

In the case of the bunched /r/, the long back cavity has a wide constriction in the pharynx and is more uniform overall, so that we model it initially as a quarter-wavelength tube [see Fig. 10(c)]. If we then account for the pharyngeal narrowing, F3 is lowered and F5 is raised. If we include the palatal constriction itself, F4 is raised and F5 is lowered. Finally, including the laryngeal narrowing in the model raises F4 and (to a lesser extent) F5. The net results of these manipulations are shown in Fig. 10(d). Again, Tables III and IV show that there is close agreement between the formant frequencies predicted by both the FEM-derived area functions and the simple-tube model and measured from the actual acoustic data.

C. Formants in acoustic data of sustained /r/ and nonsense word "warav"

At this point, it appears plausible that the F4/F5 pattern shown by S1 and S2 is a function of their retroflex and bunched tongue shapes. As a partial confirmation of this hypothesis, we investigated acoustic data from sustained /r/ data for two subjects (S3 and S4) who have retroflex /r/ tongue shapes similar to S1 and two subjects (S5 and S6) who have bunched /r/ tongue shapes similar to S2. The averaged spectra (from a 300 ms segment of sound booth acoustic recordings) of the sustained /r/ sounds produced by the six subjects in the upright position are shown in Fig. 11. As can be seen, the retroflex /r/ has a larger difference in F4 and F5 than the bunched /r/. The differences between F4 and F5 for S3 and S4 are about 1900 and 2000 Hz, respectively, while the differences between F4 and F5 for S5 and S6 are about 500 and 600 Hz, respectively. These results are consistent with the results obtained from S1 and S2 in that the spacing between F4 and F5 is larger for the retroflex /r/ than for the bunched /r/.

In addition, the formant trajectories of the nonsense word "warav" for all the six subjects are shown in Fig. 12 (note that the spectrograms of Fig. 1 are repeated here for comparison). The differences between F4 and F5 of /r/ at the lowest point of F3 for S1, S3, and S4 are about 2100, 1500, and 1600 Hz, respectively, while the differences between F4 and F5 of /r/ at the lowest point of F3 for S2, S5, and S6 are about 700, 900, and 600 Hz, respectively. These results indicate that, for these subjects, the difference between F4 and F5 for the retroflex /r/ in dynamic speech is relatively larger than that in the bunched /r/ and provides additional support for the simulation result from the 3D FEM and computer vocal tract models based on the area functions.

IV. DISCUSSION

In this paper, we investigate the relationship between acoustic patterns in F4 and F5 and articulatory differences in tongue shape between subjects. The primary data come from S1 and S2, who produce sharply different bunched and retroflex variants of /r/ associated with different patterns of F4 and F5. S1 and S2 are particularly comparable because they resemble each other in terms of vocal tract length and oral tract dimensions. The results suggest that bunched and retroflex tongue shapes differ in the frequency spacing between F4 and F5. Further, the F4/F5 patterns produced by S1 and S2 can be derived from a very simple aspect of the difference between the two vocal tract shapes. For both S1's retroflex /r/ and S2's bunched /r/, F4 and F5 (along with F3) come from the long back cavity. However, for S1, these formants are half-wavelength resonances, while for S2, these formants are quarter-wavelength resonances of the cavity. Additionally, the finding of an F4/F5 difference in pattern is replicated in the acoustic data from an additional set of four subjects, two with bunched and two with retroflex tongue shapes for /r/. These results suggest that acoustic cues based on F4-F5 spacing may be robust and reliable indicators of tongue shape, at least for the classic (tongue tip down) bunched and (tongue dorsum down) retroflex shapes discussed here.

It appears that this spacing between F4 and F5 is due to the difference in long back cavity dimension/shape. In the case of the retroflex /r/, there is one long back cavity posterior to the palatal constriction. Our simple-tube modeling and the sensitivity functions show that F4 and F5 are half-

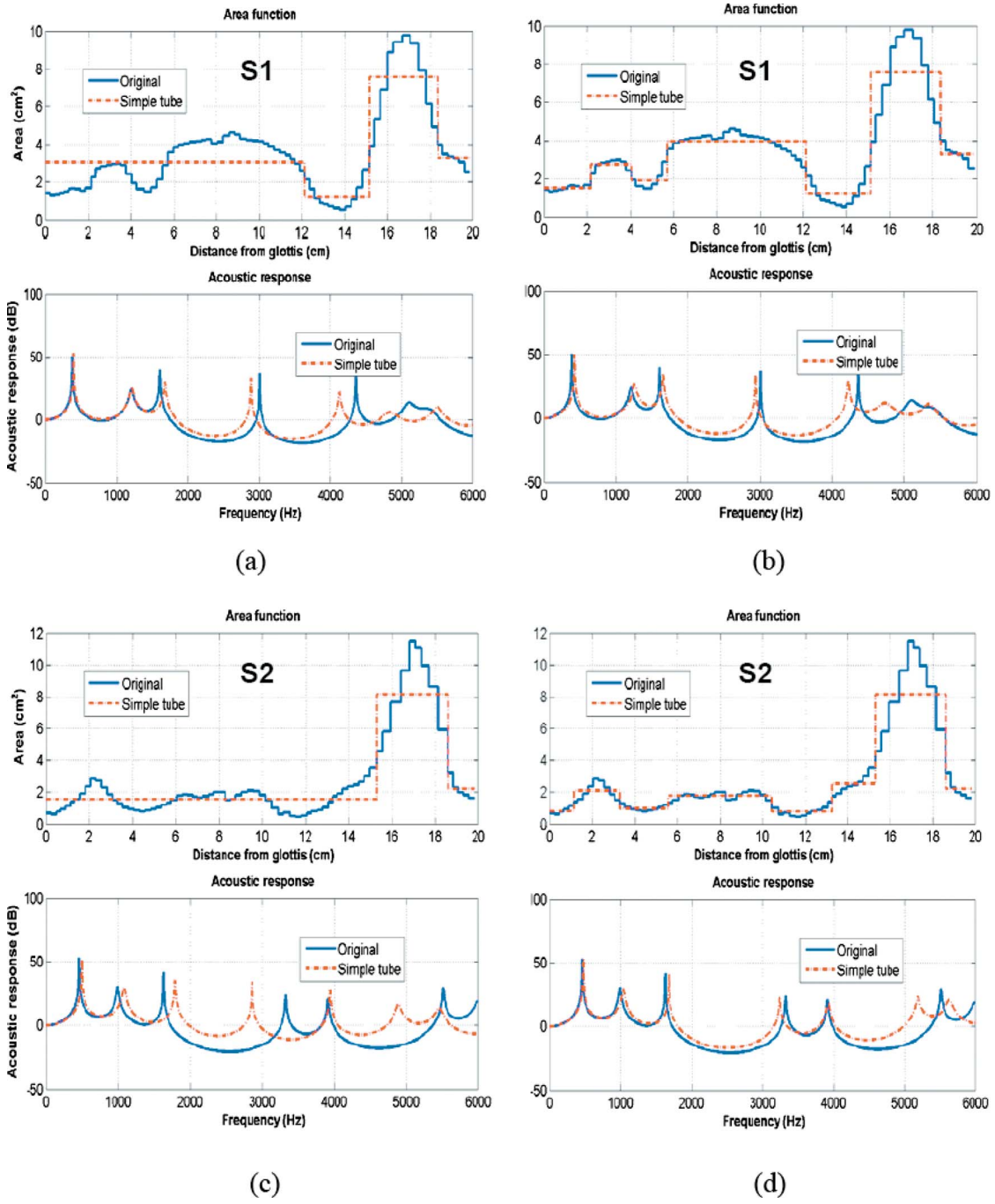

FIG. 10. (Color online) Simple-tube models overlaid on FEM-derived area functions (top panel) and corresponding acoustic responses (bottom panel). (a) Four element simple-tube model of the retroflex /r/ of S1. (b) Seven element simple-tube model of the retroflex /r/ of S1. (c) Three element simple-tube model of the bunched /r/ of S2. (d) Eight element simple-tube model of the bunched /r/ of S2.

wavelength resonances of the back cavity. In fact, F4 and F5 are the second and third resonances of the back cavity (F3 is the first resonance of this cavity). For S1, this half-wavelength cavity is about 12 cm long which gives a spacing between the resonances of about 1460 Hz. The narrowing in the laryngeal regions shifts F4 and F5 upward by different amounts so that the spacing changes to about 1300 Hz. This spacing agrees well with the 1469–1531 Hz measured from S1's sustained /r/. For the bunched /r/, the back cavity can be modeled as a quarter-wavelength tube. Our simple-tube modeling shows that F4 and F5 are the third and fourth resonances of this cavity. The sensitivity functions, on the other hand, show that F4 and F5 are influenced by the front cavity. This is probably due to the higher degree of coupling between the front and back cavities for the bunched /r/ of S2. The length of the back cavity for S2 is about 15 cm. Thus, the spacing between F4 and F5 for the bunched /r/ should be about 1150 Hz. However, the narrowing in the laryngeal, pharyngeal, and palatal regions decreases this difference to about 650 Hz, as seen in Fig. 10(d). This formant difference agrees well with the value of 651–796 Hz measured from S2's sustained /r/. As a point of interest, the spacing between F4 and F5 in the spectrograms of Fig. 12 is generally greater across all of the consonants and vowels for the speakers who

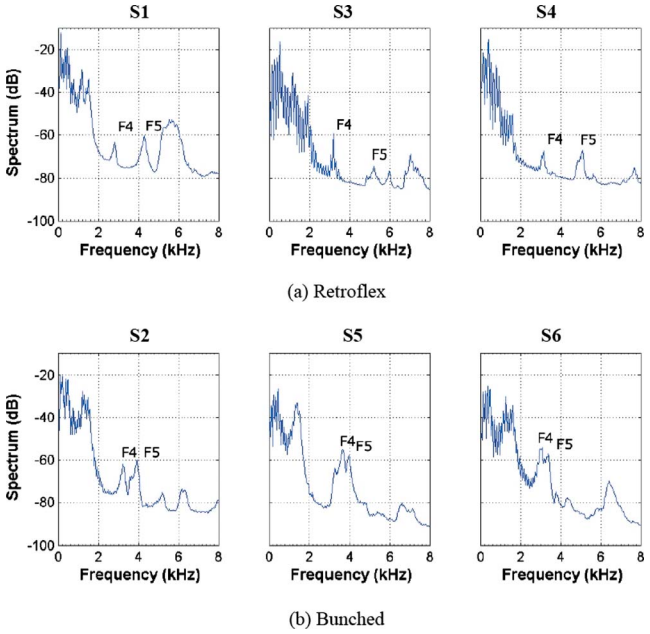

FIG. 11. (Color online) Spectra of sustained /r/ utterances from six speakers (three retroflex /r/ and three bunched /r/). (a) Retroflex /r/ (left: S1; middle: S3; right: S4.) (b) Bunched /r/ (left: S2; middle: S5; right: S6).

produce the retroflex tongue shape for /r/ than it is in the spectrograms for the speakers who produce the bunched tongue shape for /r/. However, the difference does appear to be considerably enhanced during the /r/ sounds with the lowering of F4 and the slight rising of F5 during the retroflex /r/, and the rising of F4 for S2 during the bunched /r/.

The relationship of tongue shapes for /r/ to specific acoustic properties as found in this study may be useful for the development of speech technologies such as speaker and speech recognition. For example, knowledge-based approaches to speech recognition heavily rely on acoustic information to infer articulatory behavior (Hasegawa-Johnson *et al.*, 2005; Kinga *et al.*, 2006; Juneja and Espy-Wilson, 2008). In addition, speakers appear to use tongue shapes in very consistent ways (Guenther *et al.*, 1999). Thus, the use of a particular tongue shape for /r/ may produce acoustic characteristics that are indicative of a speaker's identity, even if these characteristics are not relevant to the phonetic content.

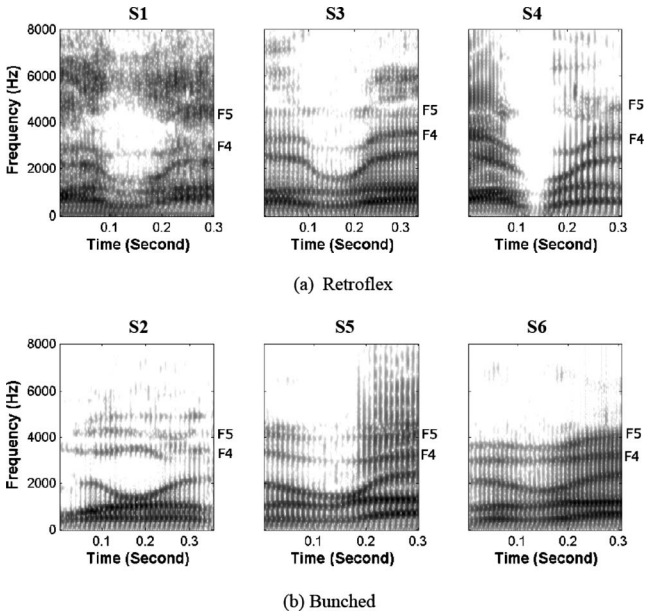

FIG. 12. Spectrograms for nonsense word "warav" from six speakers (three retroflex /r/ and three bunched /r/; only portions of spectrograms are shown in the figure with /r/ in the middle). (a) Retroflex /r/ (left: S1; middle: S3; right: S4). (b) Bunched /r/ (left: S2; middle: S5; right: S6).

ACKNOWLEDGMENT

This work was supported by NIH Grant No. 1-R01-DC05250-01.

[1]In the larger study (see Tiede *et al.*, 2004), subjects S1, S2, S3, S4, S5, and S6 are coded as subjects 22, 5, 1, 20, 17, and 19, respectively.

[2]Linguists distinguish between rhotic dialects, in which /r/ is fully pronounced in all word conditions, and nonrhotic dialects, in which some postvocalic /r/s are replaced by a schwalike vowel. Nonrhotic dialects are typically found throughout the southern states and in coastal New England.

[3]Ideally, productions of both a retroflex and bunched /r/ from a single speaker would be compared. Some speakers do indeed change their productions between true retroflex and bunched shapes in different phonetic contexts (Guenther *et al.*, 1999). However, this behavior appears to be a reaction to coarticulatory pressures in dynamic speaking conditions and is not easily elicited or trained in a sustained context. We in fact trained S2 to produce /r/ with his tongue tip up, and we collected a full set of MRI data for this production, in addition to the set with his natural /r/ production. However, even with training, S2 was not able to produce /r/ without a raised tongue dorsum as well as a raised tongue tip; thus, we were not able to compare 3D models of both a bunched configuration and a true retroflex tongue shape. While S1 was able to produce bunched /r/ in context, he was not able to sustain it consistently. At the same time, all of our speakers produced the same tongue shape consistently when asked to produce their natural sustained /r/. Thus, we contrast sustained bunched and retroflex /r/ as produced by subjects whose age and vocal tract dimensions are as similar as possible.

[4]Measured piriform dimensions: for S1, 18 mm in length and 2.3 cm^3 in volume; for S2, 12 mm in length and 2 cm^3 in volume.

[5]For subject S2, we collected a separate session of sound booth acoustic data in which his tongue shape for /r/ was monitored via ultrasound (Aloka SD-1000, 3.5 MHz probe held under the jaw). In all cases (upright running speech, supine and upright sustained /r/), S2 used a bunched tongue configuration with the tongue tip down when producing his natural /r/.

Alwan, A., Narayanan, S., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," J. Acoust. Soc. Am. **101**, 1078–1089.

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.

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

Comsol (**2007**). COMSOL MULTIPHYSICS (http://www.comsol.com, accessed 12/20/2007).

Dalston, R. M. (**1975**). "Acoustic characteristics of English /w,r,l/ spoken correctly by young children and adults," J. Acoust. Soc. Am. **57**, 462–469.

Dang, J. W., and Honda, K. (**1997**). "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101**, 456–465.

Delattre, P., and Freeman, D. C. (**1968**). "A dialect study of American English r's by x-ray motion picture," Linguistics **44**, 28–69.

Espy-Wilson, C. Y. (**1987**). Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Espy-Wilson, C. Y. (**2004**). "Articulatory strategies, speech acoustics and variability," Proceedings of Sound to Sense: Fifty+ Years of Discoveries in Speech Communication, pp. B62–B76.

Espy-Wilson, C. Y., and Boyce, S. E. (**1999**). "The relevance of F4 in distinguishing between different articulatory configurations of American English /r/," J. Acoust. Soc. Am. **105**, 1400.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., and Alwan,

A. (**2000**). "Acoustic modeling of American English /r/," J. Acoust. Soc. Am. **108**, 343–356.
Fant, G. (**1970**). *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations* (Mouton, The Hague).
Fant, G., and Pauli, S. (**1974**). "Spatial characteristics of vocal tract resonance modes," Proceedings of the Speech Communication Seminar, pp. 121–132.
Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., and Perkell, J. S. (**1999**). "Articulatory tradeoffs reduce acoustic variability during American English /r/ production," J. Acoust. Soc. Am. **105**, 2854–2865.
Hagiwara, R. (**1995**). "Acoustic realizations of American /r/ as produced by women and men," UCLA Working Papers in Phonetics, Vol. 90, pp. 1–187.
Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., and Tianyu, W. (**2005**). "Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 213–216.
Heinz, J. M., and Stevens, K. N. (**1964**). "On the derivation of area functions and acoustic spectra from cinéradiographic films of speech," J. Acoust. Soc. Am. **36**, 1037–1038.
Juneja, A., and Espy-Wilson, C. Y. (**2008**). "Probabilistic landmark detection for automatic speech recognition using acoustic-phonetic information," J. Acoust. Soc. Am. **123**, 1154–1168.
Kinga, S., Frankel, J., Livesc, K., McDermott, E., Richmon, K., and Wester, M. (**2006**). "Speech production knowledge in automatic speech recognition," J. Acoust. Soc. Am. **121**, 723–742.
Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., and Honda, K. (**2006**). "Cyclicity of laryngeal cavity resonance due to vocal fold vibration," J. Acoust. Soc. Am. **120**, 2239–2249.
Lee, K. (**1999**). *Principles of CAD/CAM/CAE Systems* (Addison-Wesley, Reading, MA).
Lehiste, I. (**1964**). *Acoustical Characteristics of Selected English Consonants* (Indiana University, Bloomington).
Lisker, L. (**1957**). "Minimal cues for separating /w,r,l,y/ in intervocalic position," Word **13**, 256–267.
Materialise (**2007**). Trial versions of Mimics and Magics (http://www.materialise.com, accessed 12/20/2007).
Matsuzaki, H., Miki, N., and Ogawa, Y. (**2000**). "3D finite element analysis of Japanese vowels in elliptic sound tube model," Electron. Commun. Eng. **83**, 43–51.
Matsuzaki, H., Miki, N., Ogawa, Y., Matsuzaki, H., Miki, N., and Ogawa, Y. (**1996**). "FEM analysis of sound wave propagation in the vocal tract with 3D radiational model," J. Acoust. Soc. Jpn. (E) **17**, 163-166.
Miki, N., Mastuzaki, H., Aoyama, K., and Ogawa, Y. (**1996**). "Transfer function of 3-D vocal tract model with higher mode," Proceedings of the Fourth Speech Production Seminar (Autrans), pp. 211–214.
Morse, P. M., and Ingard, K. U. (**1968**). *Theoretical Acoustics* McGraw-Hill, New York.
Motoki, K. (**2002**). "Three-dimensional acoustic field in vocal-tract," Acoust. Sci. & Tech. **23**, 207–212.
Mrayati, M., Carré, R., and Guérin, B. (**1988**). "Distinctive regions and modes—a new theory of speech production," Speech Commun. **7**, 257–286.
Narayanan, S. S., Alwan, A. A., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," J. Acoust. Soc. Am. **101**, 1064–1077.
O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1957**). "Acoustic cues for the perception of initial /w,j, r,l/ in English," Word **13**, 24–43.
Ong, D., and Stone, M. (**1998**). "Three dimensional vocal tract shapes in [r] and [l]: A study of MRI, ultrasound, electropalatography, and acoustics," Phonoscope **1**, 1–13.
Shriberg, L. D., and Kent, R. D. (**1982**). *Clinical Phonetics* (Macmillan, New York).
Sondhi, M. M. (**1986**). "Resonances of a bent vocal tract," J. Acoust. Soc. Am. **79**, 1113–1116.
Story, B. H. (**2006**). "Technique for 'tuning' vocal tract area functions based on acoustic sensitivity functions (L)," J. Acoust. Soc. Am. **119**, 715–718.
Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**, 537–554.
Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., and Honda, K. (**2006a**). "Acoustic roles of the laryngeal cavity in vocal tract resonance," J. Acoust. Soc. Am. **120**, 2228–2238.
Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (**2006b**). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," J. Acoust. Soc. Am. **119**, 1037–1049.
Thomas, T. J. (**1986**). "A finite element model of fluid flow in the vocal tract," Comput. Speech Lang. **1**, 131–151.
Tiede, M., Boyce, S. E., Holland, C., and Chou, A. (**2004**). "A new taxonomy of American English /r/ using MRI and ultrasound," J. Acoust. Soc. Am. **115**, 2633–2634.
Twist, A., Baker, A., Mielke, J., and Archangeli, D. (**2007**). "Are 'covert' /r/ allophones really indistinguishable?," Selected Papers from NWAV 35, University of Pennsylvania working Papers in Linguistics **13**(2).
Westbury, J. R., Hashi, M., and Lindstrom, M. J. (**1998**). "Differences among speakers in lingual articulation for American English /r/," Speech Commun. **26**, 203–226.
Zhang, Z., Espy-Wilson, C., Boyce, S., and Tiede, M. (**2005**). "Modeling of the front cavity and sublingual space in American English rhotic sounds," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 893–896.
Zhou, X. H., Zhang, Z. Y., and Espy-Wilson, C. Y. (**2004**). "VTAR: A MATLAB-based computer program for vocal tract acoustic modeling," J. Acoust. Soc. Am. **115**, 2543.