Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002

Brad H. Story[a)]
Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

(Received 7 March 2007; revised 12 October 2007; accepted 16 October 2007)

A new set of area functions for vowels has been obtained with magnetic resonance imaging from the same speaker as that previously reported in 1996 [Story *et al.*, J. Acoust. Soc. Am. **100**, 537–554 (1996)]. The new area functions were derived from image data collected in 2002, whereas the previously reported area functions were based on magnetic resonance images obtained in 1994. When compared, the new area function sets indicated a tendency toward a constricted pharyngeal region and expanded oral cavity relative to the previous set. Based on calculated formant frequencies and sensitivity functions, these morphological differences were shown to have the primary acoustic effect of systematically shifting the second formant (F2) downward in frequency. Multiple instances of target vocal tract shapes from a specific speaker provide additional sampling of the possible area functions that may be produced during speech production. This may be of benefit for understanding intraspeaker variability in vowel production and for further development of speech synthesizers and speech models that utilize area function information.
 [DOI: 10.1121/1.2805683]

PACS number(s): 43.70.Bk, 43.70.Aj [CS] Pages: 327–335

I. INTRODUCTION

Collections of vocal tract area functions are useful for the development and testing of many types of speech production models and speech synthesizers. Since the early 1990s, magnetic resonance imaging (MRI) has been used extensively to acquire volumetric image sets of the head and neck from which area functions can be directly measured (e.g., Lakshminarayanan *et al.*, 1991; Baer *et al.*, 1991; Yang and Kasuya, 1994; Dang *et al.*, 1994; Dang and Honda, 1997; Story *et al.*, 1996, 1998; Narayanan *et al.*, 1995, 1997; Alwan *et al.*, 1997; Narayanan *et al.*, 1997; Story, 2005b). These area functions are typically obtained for static vocal tract shapes and are assumed to be representative of a particular speaker's "normal" production of specific vowels or consonants.

It has been shown, however, that a particular speaker can produce a range of different vocal tract shapes for the same target vowel. Story *et al.* (2001) reported area functions obtained from one male and one female speaker who were each asked to deliberately produce the vowels [i ɑ æ u] with three different qualities: normal, twang, and yawny. Production of each quality was hypothesized to require a different overall "setting" of the vocal tract shape (Laver, 1980). The results for each speaker indicated quite different area functions for each vowel across the three quality conditions, but production of each quality involved fairly systematic changes to the vocal tract shape across the vowels. Specifically, the yawny quality was characterized by a general widening of the oral cavity and lengthening of the vocal tract, whereas the twang was produced with a widened lip opening, constricted oral cavity, and shortened tract length. Although these speakers were trained vocal performers and the twang and yawny qualities imposed rather extreme changes on the vocal tract shape, the results suggest that a typical speaker would have the ability to generate a variety of tract shapes for a given target vowel depending on speaking conditions (e.g., Lindblom, 1996).

Takemoto *et al.* (2006) has also demonstrated differences in area functions of the same target vowel obtained from the same person, albeit one for dynamic speech and the other for static. They used MRI techniques that allow a three-dimensional (3D) volume of the head and neck to be acquired over the time course of an utterance (Takemoto *et al.*, 2006). Postacquisition image analysis was then used to generate a time-varying area function. Area functions taken from specific points in time within the vowel sequence [ɑeuio] were compared with those, of the same five vowels, obtained from the same speaker with a more conventional static imaging procedure. The area functions for each vowel were quite similar in overall shape, but there were localized differences in the magnitude of the cross-sectional area in both the anterior and posterior portions of the area function, depending on the particular vowel.

Information concerning multiple instances of target vocal tract shapes of a specific speaker is useful from the point of view that it provides a better sampling of the possible area functions that may be produced during speech production. This may be of particular benefit for understanding intraspeaker variability as well as for further development of vocal tract models based on statistical analyses of area functions

[a)]Electronic mail: bstory@u.arizona.edu

(e.g., Meyer *et al.*, 1989; Yehia *et al.*, 1996; Story and Titze, 1998; Story, 2005a, b; Mokhtari *et al.*, 2006).

The purpose of this article is to report a new set of vocal tract area functions for vowels obtained from the same speaker as that reported in Story *et al.* (1996). Between 2001 and 2003, image sets were collected from six speakers whose area functions were reported in Story (2005b). During this time, the opportunity existed to also obtain image sets of vowels from the speaker in Story *et al.* (1996). The aim in collecting these images was simply for purposes of comparison to the previous set, and to supply an additional instance of each vowel produced by the same person. Reporting similar data collected from the same speaker twice is perhaps a bit unusual, and, if such data were identical, would not be particularly interesting. The results, however, indicate that the new set of area functions support a downward shift in the second formant (F2), relative to the Story *et al.* (1996) set, over almost the entire [F1, F2] vowel space. This suggests that a different vocal tract setting, as described previously, was used for production of these vowels. The specific aim of this article is to report the area functions in numerical form, compare them graphically and acoustically with those from Story *et al.* (1996), and provide some explanation of how the differences observed between the two area function sets support a downward shift in F2.

II. METHOD

A. Speaker

The speaker was the same person who was scanned with MRI for the Story *et al.* (1996) study (henceforth referred to as STH96). It is acknowledged that he was the first author of STH96 and is the sole author of the present study. All new image sets were collected on 22 May 2002 at the University of Arizona Medical Center in Tucson, AZ. At this time the speaker was 37 years old. The previous data (reported in 1996) had been collected in June 1994. In the intervening years the speaker moved from Iowa City, IA to Denver, CO where he lived until 2000; he then relocated to Tucson, AZ. The speaker's height and weight were identical to those stated in STH96 (ht.=5 ft 7 in. wt.=145 lbs).

B. Image collection

MRI was used to obtain volumetric image sets of vocal tract shapes that corresponded to the speaker's production of the American English vowels [i ɪ e ɛ æ ʌ ɑ ɔ o ʊ u]. The MR images were acquired with a General Electric Sigma 1.5 Tesla scanner. The data acquisition mode was fast spin echo and the scanning parameters were set to TE=13 ms, TR=4000 ms, ETL=16 ms, and NEX=2. During the speaker's production of a particular vocal tract shape, a 28 slice series was collected with an interleaved acquisition sequence. Each image set consisted of contiguous, parallel, axial sections (slices) extending from a location just superior of the hard palate to an inferior location near the first tracheal ring. The field of view (slice dimensions) for each slice was 24 cm×24 cm which, with a 256×256 pixel matrix, provided an in-plane spatial resolution of 0.938 mm/pixel. Although the STH96 images were obtained with a rigid anterior/posterior neck coil, at the time these new images were collected only a flexible anterior neck coil was available. The scanning parameters were set to allow an image slice thickness of 4 mm for all image sets. In STH96 the slice thickness was 5 mm.

The procedure for acquiring the image sets was nearly identical to that described in Story (2005b). Each vowel was produced as if it were to be spoken in an hVd syllable, but was instead sustained. Thus, any vowels typically spoken as diphthongs (e.g., [e] and [o]) would be represented in the image set as the onset vowel of that diphthong. Shortly after the speaker began phonation for a particular target vowel, the MR technologist initiated the scan. After 8 s the scan was paused to allow time for the speaker to breathe. The scanning was continued when the speaker resumed phonation. The scanning time required for each image set (i.e., for one complete tract shape) was 4 min and 32 s as compared to 4 min and 16 s in STH96. This required approximately 30 repetitions of each target vowel. With pauses for respiration between repetitions, each image set was completed in about 10–15 min. The speaker's goal was a "normal" production of each vowel, with a strong focus on maintaining a consistent shape.

C. Image analysis and area functions

The image analysis was the same as that described in Story (2005b) and Story *et al.* (1996, 1998, 2001). In brief, for each vocal tract shape the procedure included segmentation of the airspace from the surrounding tissue, shape-based interpolation to generate a 3D reconstruction of the airspace, and cross-sectional area analysis of the airspace. The collection of areas obtained, which extends from just above the glottis to the lips, along with the distance of each cross section from the glottis, comprise the area function. Each area function was subsequently resampled with a cubic spline from which 44 area sections were obtained at equal length increments. A smoothing filter was subsequently applied to remove small discontinuities assumed to be imaging artifacts (see Story *et al.*, 2001, p. 1653). The piriform sinuses were segmented in each image set but were not included in the cross-sectional area analysis. Hence, information about them will not be reported here.

D. Audio recording and formant analysis

On the day immediately following image collection, the speaker's production of all 11 vowel sounds was recorded. For this session the speaker inserted earplugs and lay supine on a cushioned table in a sound-treated booth. At least three repetitions of each vowel were produced as long sustained productions with approximate durations of 4–8 s, similar to the duration of a single vowel repetition in the MR scanner. The speaker attempted to produce the vowels as similarly as possible, in both quality and loudness (self-perception), to those produced the day before in the MR scanner. The audio signal was transduced with an AKG CK92 microphone positioned 30 cm from the speaker and off-axis at 45°. The signal was recorded on digital audio tape at a sampling frequency of 44.1 kHz and later transferred to separate digital

computer files for each vowel. In addition, audio recordings obtained in 1994 that coincide with STH96 were transferred to digital files so that they could be analyzed with the same methods used for the 2002 recordings. These were also long sustained vowels recorded in the supine position with earplugs.

For the 2002 recordings formant frequencies were estimated over the time course of three repetitions of each vowel with PRAAT's formant analysis module (Boersma and Weenink, 2007). Depending on the particular vowel, formant analysis parameters were manually adjusted so that the formant contours of F1, F2, and F3 were aligned with the centers of their respective formant bands in a simultaneously displayed wide-band spectrogram. All time-dependent formant values for each vowel were transferred to MATLAB (Mathworks, 2006) where means and standard deviations were computed. Depending on the vowel, the three repetitions provided 15–25 s of recorded signal over which the analysis was performed. The 1994 recordings were analyzed in exactly the same manner but for some vowels there were only two repetitions available. Nonetheless, this still provided 10 or more seconds of recorded signal for analysis.

E. Calculation of formant frequencies

Frequency response functions were calculated for each newly obtained area function as well as for the ten vowels of STH96. This was performed with a frequency-domain technique (Sondhi and Schroeter, 1987; but specifically as presented in Story *et al.*, 2000) that included energy losses due to yielding walls, viscosity, heat conduction, and radiation. Prior to the calculations, the STH96 area functions were smoothed with the same process applied to those of the present study (see Sec. II C). Formant frequencies were determined by finding the peaks in the frequency response functions with parabolic interpolation (Titze *et al.*, 1987).

Calculation of the formant frequencies reported in STH96 was performed differently than in the present study. These differences do have an effect on the actual formant values, hence, they will be briefly summarized. For STH96 a wave-reflection algorithm (Liljencrants, 1985; Story, 1995) was used to generate a time-domain simulation of a particular vowel sound. The algorithm used a glottal flow pulse signal as the voice source and included losses due to yielding walls, viscosity, heat conduction, and radiation. This particular simulation required that each section (tubelet) of the area function have a length of $c/(2F_s)$, where c is the speed of sound and F_s is the sampling frequency. Accordingly, the area functions were reported as a variable number of sections, each with the same section length (STH96, p. 547). In the present study, all area functions contain 44 sections but the section length can vary across vowels. Another important difference is that a side-branch representing the piriform sinuses was coupled to the main vocal tract at a point 2.4 cm from the glottis for all vowels. The cross-sectional areas for the piriform sinuses were reported in Story (1995). Each simulated vowel was then subjected to exactly the same LPC analysis that was used to measured formants from recorded speech. This involved an initial preemphasis prior to the LPC analysis.

Each of these differences can potentially contribute to a slightly different calculation of the formants than would be given by the frequency-domain approach used in the present study. As can be discerned from Sec. III, the formants calculated for the STH96 area functions with the present frequency-domain approach are not the same as those originally reported. A piriform sinus branch, however, was added to the frequency-domain model and used to recalculate the formants of the STH96 area functions. In this case, the formants were well-matched to the originals, suggesting that the inclusion of the piriform sinus is the largest contributor to the difference in the calculated formants. Although the effects of these sinuses on vowel formants are an important area for further study (e.g., Dang and Honda, 1997), they were considered to be outside the scope of the present report. Hence, the calculated formants for both sets of area functions are reported in Sec. III for the condition *without* a piriform sinus branch.

III. RESULTS

The area functions are presented numerically in Table I. Each column contains 44 cross-sectional areas that extend from the glottis (section 1) to the lips (section 44). The bottom two rows contain the section length and total tract length of each area function, respectively. The measured and calculated acoustic characteristics are shown in Table II. In the top row are the measured fundamental frequencies (F0) that range from 149 Hz for [e] to 159 Hz for [i]. These are somewhat higher than the speaker's typical F0 but were thought, by the speaker, to be representative of those produced during MR scanning. Measured and calculated formant frequencies are shown in the middle portions of Table II. The lower three rows indicate the percent error of the computed formants relative to the mean value of the natural speech formants. These range from a low of 1.6% for the second formant of [e] to high of 35.7% for the second formant of [ʌ].

Each new area function is plotted along with its STH96 counterpart in Fig. 1. To accurately depict the time period in which the respective image sets were acquired, the legends refer to the area functions from the present study as "2002" and those from STH96 as "1994." An [e] vowel was not reported in STH96, hence, the subplot for it contains only the 2002 version. There are both similarities and differences between the two area function sets. For example, the variation in the cross-sectional area along the initial 2–4 cm of VT length is approximately the same for both versions of the vowels [i ɪ ɛ æ o ʊ u]. For [ʌ], [ɑ], and [ɔ], however, an increase in area occurs closer to the glottis in the new versions than in the old. This is apparently a consequence of lengthening the pharyngeal section, perhaps by larynx lowering, which extends from about 3.5 to 9 cm above the glottis in the 1994 area functions and from 2.5 to 9 cm in the new ones. The new versions of these same three vowels, along with [o] and [ʊ], also exhibit larger areas within the oral cavity but have nearly the same lip termination area as

TABLE I. Area vectors for each vocal tract shape. Each original area function has been resampled to consist of 44 area sections given in cm^2; the length of each section is given by Δ in cm. The glottal end of each area vector is at section 1 and the lip end at section 44. The total vocal tract length (VTL) in cm is computed as 44Δ.

Section	i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u
1	0.51	0.28	0.29	0.37	0.31	0.23	0.56	0.27	0.38	0.37	0.54
2	0.59	0.21	0.26	0.30	0.21	0.34	0.62	0.43	0.45	0.38	0.61
3	0.62	0.21	0.30	0.24	0.18	0.47	0.66	0.54	0.57	0.49	0.66
4	0.72	0.30	0.40	0.23	0.23	0.60	0.78	0.67	0.77	0.62	0.75
5	1.24	0.47	0.55	0.29	0.33	0.77	0.97	0.83	1.31	0.85	1.13
6	2.30	0.71	0.74	0.41	0.50	1.06	1.16	0.92	1.92	1.28	1.99
7	3.30	1.12	0.99	0.58	0.78	1.26	1.12	0.89	1.74	1.62	2.83
8	3.59	1.48	1.09	0.82	0.96	1.09	0.82	0.73	1.11	1.47	2.90
9	3.22	1.35	0.90	0.97	0.85	0.80	0.55	0.55	0.75	1.04	2.52
10	2.86	1.05	0.69	0.82	0.63	0.65	0.45	0.44	0.59	0.81	2.40
11	3.00	0.92	0.77	0.62	0.46	0.56	0.37	0.37	0.57	1.03	2.83
12	3.61	0.92	1.31	0.54	0.36	0.47	0.29	0.28	0.68	1.44	3.56
13	4.39	1.19	2.13	0.54	0.33	0.37	0.21	0.18	0.73	1.49	3.99
14	4.95	1.94	2.74	0.68	0.46	0.24	0.15	0.14	0.67	1.28	3.89
15	5.17	2.83	3.03	1.09	0.73	0.17	0.16	0.15	0.58	1.06	3.50
16	5.16	3.31	3.23	1.62	1.00	0.18	0.25	0.14	0.49	0.85	3.04
17	5.18	3.48	3.33	2.03	1.30	0.23	0.34	0.13	0.44	0.69	2.64
18	5.26	3.60	3.27	2.35	1.66	0.27	0.43	0.14	0.42	0.56	2.44
19	5.20	3.64	3.09	2.51	1.97	0.28	0.54	0.18	0.49	0.42	2.31
20	5.02	3.49	2.84	2.39	2.06	0.29	0.61	0.20	0.53	0.27	2.07
21	4.71	3.20	2.66	2.22	2.03	0.33	0.67	0.23	0.38	0.27	1.80
22	4.13	2.90	2.46	2.13	2.01	0.52	0.98	0.49	0.30	0.41	1.52
23	3.43	2.59	2.14	2.00	1.89	0.97	1.76	1.03	0.45	0.51	1.14
24	2.83	2.21	1.79	1.78	1.66	1.50	2.75	1.58	0.61	0.49	0.74
25	2.32	1.87	1.44	1.58	1.49	1.91	3.52	2.06	0.71	0.47	0.42
26	1.83	1.54	1.17	1.43	1.42	2.23	4.08	2.61	0.79	0.50	0.22
27	1.46	1.20	1.00	1.31	1.37	2.65	4.74	3.35	0.86	0.58	0.14
28	1.23	0.92	0.88	1.23	1.34	3.29	5.61	4.34	1.01	0.80	0.20
29	1.08	0.74	0.80	1.24	1.41	4.13	6.60	5.51	1.41	1.19	0.47
30	0.94	0.59	0.81	1.38	1.58	5.00	7.61	6.70	2.09	1.62	0.89
31	0.80	0.52	0.85	1.61	1.82	5.77	8.48	7.75	3.00	2.27	1.15
32	0.67	0.54	0.84	1.82	2.19	6.33	9.06	8.63	4.10	3.24	1.42
33	0.55	0.59	0.86	1.96	2.63	6.61	9.29	9.29	5.16	4.16	2.17
34	0.46	0.65	0.96	2.01	2.97	6.63	9.26	9.59	6.22	5.00	3.04
35	0.40	0.71	1.18	2.00	3.17	6.45	9.06	9.42	7.34	5.70	3.69
36	0.36	0.67	1.35	1.95	3.40	6.04	8.64	8.78	8.15	6.11	4.70
37	0.35	0.61	1.48	1.77	3.56	5.39	7.91	7.82	8.61	6.21	5.74
38	0.35	0.57	1.62	1.48	3.57	4.42	6.98	6.50	8.37	6.29	5.41
39	0.38	0.50	1.49	1.30	3.58	3.29	6.02	4.95	6.76	6.24	3.82
40	0.51	0.48	1.29	1.21	3.44	2.37	5.13	3.47	4.37	4.91	2.34
41	0.74	0.54	1.24	1.10	3.15	1.74	4.55	2.15	2.30	2.61	1.35
42	0.92	0.73	1.17	0.99	3.38	1.36	4.52	1.38	1.06	1.09	0.65
43	0.96	0.93	1.04	0.91	3.99	1.17	4.71	1.11	0.58	0.63	0.29
44	0.91	0.82	0.95	0.78	4.17	0.99	4.72	0.90	0.47	0.59	0.16
Δ (cm)	0.384	0.376	0.386	0.393	0.366	0.390	0.388	0.395	0.417	0.440	0.445
VTL (cm)	16.90	16.55	16.98	17.30	16.11	17.14	17.09	17.40	18.33	19.34	19.59

they did previously. For [ʌ ɔ o ʊ], the speaker maintained a more constricted pharyngeal region in the new versus the old area functions that, with the expansion of the oral cavity, would suggest they were produced with an increased degree of the "back" dimension. The overall shape for [ɪ], [ɛ], and [æ] is similar across the 1994 and 2002 sets even though the magnitude of the areas is different. This is especially prominent for [æ] where peaks occur near 8 and 14 cm from the glottis, respectively, in both versions, but the more recent area function is on the order of 1 to 2 cm^2 smaller within this region.

For nearly all vowels, the vocal tract length in the new set of area functions was greater or equal to those from 1994. The exception is the [ɑ] vowel which is slightly shorter. The [ɛ], [o], [ʊ], and [u] were longer by more than 1 cm. In the case of [ɛ], the entire length axis of the area function appears

TABLE II. Fundamental frequencies F_0, and measured and calculated formants for the 11 vowels produced by the speaker. Each measured formant (denoted by superscript "N") is the mean across several seconds of recording and s.d. is the standard deviation. The calculated formant values are denoted by "C." The Δ's represent the percent error of the computed formants relative to the mean value of the natural speech formants (e.g., $\Delta 1 = 100(F1^C - F1^N)/F1^N$). All values have units of Hertz except the Δ's which are percentages.

	i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u
F_0	159	154	149	154	153	155	154	154	155	156	155
$F1^N$	295	478	509	659	765	752	715	665	563	518	430
s.d.	±17	±9	±10	±15	±15	±13	±9	±19	±9	±8	±5
$F2^N$	1923	1965	1917	1740	1724	1300	1180	932	835	947	917
s.d.	±28	±11	±10	±26	±22	±23	±31	±12	±30	±13	±14
$F3^N$	2797	2688	2617	2485	2501	2698	2691	2814	2637	2355	2087
s.d.	±60	±27	±42	±69	±59	±33	±21	±54	±26	±30	±14
$F1^C$	325	413	484	565	810	593	694	557	499	477	314
$F2^C$	2139	2103	1887	1595	1655	836	942	709	768	789	702
$F3^C$	2976	2626	2404	2200	2310	3099	2948	3176	2421	2499	2298
$\Delta 1$	10.0	−13.6	−4.9	−14.3	5.8	−21.1	−3.0	−16.2	−11.3	−7.9	−27.1
$\Delta 2$	11.2	7.0	−1.6	−8.3	−4.0	−35.7	−20.2	−23.9	−8.0	−16.7	−23.5
$\Delta 3$	6.4	−2.3	−8.1	−11.5	−7.6	14.9	9.6	12.9	−8.2	6.1	10.1

to be stretched, whereas, for the latter three vowels, the length increase could be attributed to more extreme lip rounding.

The F1 and F2 formant frequencies calculated for both the 1994 and 2002 versions of each vowel are plotted against each other in Fig. 2(a). A data point for [e] is only present for the 2002 vowels. Relative to the 1994 vowels, a predominant feature is that the entire 2002 vowel space is shifted downward along the F2 dimension, except for [ɪ]. Such a global change in F2 suggests a systematic, rather than random, difference in the vocal tract shapes of the 2002 vowels. To investigate this difference, acoustic sensitivity functions (Schroeder, 1967; Fant and Pauli, 1975) corresponding to F2 were calculated for each of the 2002 vowels using the method described in Story (2006, p. 715). For each vowel, regions along the vocal tract length were identified from the F2 sensitivity function where an increase or decrease in cross-sectional area would increase the frequency of F2. That is, what change in the area function would move F2 toward the value calculated for the corresponding 1994 vowel. Two of these regions are indicated by the bold portion of each 2002 area function shown in Fig. 1; the solid dots and vertical lines denote the division of the regions. The arrows above the area function, within each region, show the direction of cross-sectional area change that would perturb F2 upward in frequency. In all cases, a synergistic expansion and constriction of the marked posterior and anterior regions, respectively, would increase the frequency of F2. For all of the vowels, these prescribed changes in area (to increase F2), in at least one of the regions, would account for differences relative to the 1994 vowels. For instance, a decrease in the area of the 2002 [i] vowel between 8.5 and 14 cm from the glottis would bring it more in line with the cross-sectional area of the 1994 [i], and consequently increase F2. Increases in area for at least some portion of the posterior regions of the vowels [ɛ æ ʌ ɑ ɔ o ʊ u] and decreases in the area of the anterior regions of [ʌ ɑ ɔ o ʊ u], all of which would increase F2, are also in the direction needed to more closely match the 1994 versions of these same vowels.

Certainly other characteristics of the area function, such as tract length and cross-sectional area near the lips, may contribute to the differences in F2. But the dominant structural difference between the two sets of area functions that corresponds to their acoustic differences is a tendency toward a constricted pharynx and expanded oral cavity in the 2002 vowels. This is exemplified by the mean area functions calculated for both sets, that are plotted in the lower right panel of Fig. 1. Relative to the 1994 version, the 2002 mean area function is similar in the initial 2–4 cm of VT length, slightly constricted in the pharyngeal region (between 4 and 11 cm from glottis), slightly expanded in the oral cavity (between 11 and 16 cm from glottis), and again similar near the lip termination. These differences result in an F2 that is about 200 Hz lower than that calculated for the 1994 mean.

The F1 and F2 formant frequencies measured from the recorded speech are shown in Fig. 2(b). The center of each black ellipse represents the mean [F1,F2] value over the time course of the 15–25 s duration of each of the 2002 vowels. The horizontal and vertical extent of each ellipse indicates ±1 s.d. The 1994 formants are similarly plotted with gray ellipses. The standard deviations are small enough that there is no overlap of the corresponding vowels in the two sets except for and [ɔ] and nearly for [ɪ] (the ellipses for the 1994 [ʌ] and the 2002 [ɑ] do, however, overlap). There is a downward trend in the second formant (F2) of some of the 2002 vowels relative to those of 1994, although less prominent than for the calculated formant values. Specifically, the second formants of [i], [æ], and [ɔ] decreased by 115 Hz or more, whereas, [ɛ], [ʊ], and [u] also showed decreases in F2 but to a lesser degree. The remaining three vowels [ɪ], [ʌ], and [ɔ], had nearly identical F2 values as those of the 1994 vowels (their F1 values were different enough to prevent

them from overlapping in the [F1,F2] space). There is also a notable upward shift in the first formant (F1) of [æ], similar to that observed for the calculated formants.

IV. DISCUSSION

A new set of area functions for 11 vowels has been obtained from the same speaker who provided those reported in STH96. Eight years separated the collection of these two data sets. Comparing them, differences were observed in the cross-sectional area variation along the vocal tract axis as well as differences in vocal tract length. In a general sense, the 2002 vowels tended to be slightly more constricted in the pharyngeal region and slightly more expanded in the oral cavity than their 1994 counterparts, although there are exceptions for a few vowels. Based on calculated sensitivity functions, these morphological differences were shown to have a primary acoustic effect of systematically moving F2 downward in frequency. Although the results have indicated how the observed acoustic differences are related to the area function differences, two questions remain unanswered. First, why are the 1994 and 2002 area functions sets different? Second, why are the differences in calculated formant frequencies for each of the two sets, especially F2, greater in magnitude than the differences observed from the measured formants?

One obvious potential source of area function differences is the 8 year separation between collection of the two image sets. During this time the speaker aged from 29 to 37 years, and it is known that structural changes to the craniofacial complex and pharynx can occur in adulthood (Kollias and Krogstad, 1999; West and McNamara, 1999). Anatomical changes seem somewhat unlikely, however, because of the way in which many of the peaks and valleys in the two sets of area functions are fairly well aligned, even though cross-sectional areas are different. This was noted in particular for the [æ] but can also be observed for most of the other vowels. A detailed anatomical analysis based on the original MR image sets could potentially reveal whether structural changes did occur, but this was not carried out for the present study.

A second possible source of change is the speaker's relocation during the 8 years between image collections. As stated in Sec. II, he first moved from Iowa City, IA to Denver, CO, and then to Tucson, AZ. There are some dialectal differences relative to these three cities. Particularly notable is the merging of /ɔ/ with /ɑ/ in the western states (Labov, 1996). There is little evidence of these two vowels merging in the 2002 area function set, however, as the tract shapes of [ɔ] and [ɑ] are quite different near the lip termination and their corresponding [F1, F2] values are well separated in the vowel space plot. This does not rule out the possibility that other dialectal changes could have affected the speaker's vowel production, but does show that a well-known vowel change relative to these three cities was not observed in the present data.

A third possible source of area function change could be slight differences in the imaging procedure. The new image sets were collected using a flexible neck coil and 4 mm axial slices thickness, whereas a rigid anterior/posterior neck coil and 5 mm axial slices were used for acquiring the 1994 image data. Although different coils could affect image quality and slice thickness could affect the accuracy of the 3D vocal tract reconstruction, especially in the oral cavity because of exclusive use of axial images, one would expect such differences to be rigidly systematic with respect to particular regions of the vocal tract. For example, if slice thickness were the cause of the area function differences, the oral cavity portion might be expected to be consistently under-or overestimated in area across all the vowels. But this is not what was observed: Six of the new vowels have larger areas in the oral cavity than the 1994 versions and three are clearly smaller.

A more likely reason for the differences in the two area function sets is that the speaker had, over time, acquired a slightly different habitual "setting" of the vocal tract shape. Laver (1980) referred to such settings as "tendencies to maintain a particular constrictive (or expansive) effect" within some region of the vocal tract that biases the resulting formant frequency patterns toward a particular type of global timbre. The downward shift in F2 supported by the 2002 area functions indicates such a bias. In addition, the overall differences between area functions of the 1994 and 2002 data sets coincides with Laver's (1980) description of a "pharyngealized" quality where a constrictive effect is imposed on the middle pharynx which also results in some expansion of the oral cavity. Area functions and formant frequencies reported for a "yawny" voice quality (Story *et al.*, 2001) also coincide somewhat with the overall shape changes and the downward shift in F2 observed in the present study. That a change in habitual setting occurred during the 8 years between the data collection is, perhaps, not surprising. During this time, in addition to relocating, the speaker became familiar with many voice therapy techniques as well as methods for coaching and enhancing the professional voice. He also studied various voice qualities that involved subtle changes to the vocal tract shape (e.g., Story *et al.*, 2001; Story and Titze, 2002). Through these investigations he became well-practiced in producing many of the various qualities described by Laver (1980). Taken together these experiences likely had an effect on his typical speech production pattern that is perhaps exemplified in the differences observed between the two sets of area functions.

Also supporting the notion that a vocal tract "setting" could account for differences in the area functions is the fact that the formant frequencies of the recorded vowels from 1994 and 2002 were in closer proximity than the calculated formants based on the two area function sets. This suggests that the speaker was not obligated by anatomical, dialectal, or even habitual influences to produce the vowels with the same degree of decrease in F2 as generated by calculation using the 2002 area functions. For whatever reason, the speaker appears to have reduced the degree of a pharyngeal/yawny-type setting during the audio recording on the day following image collection. Alternatively, it might be concluded that the speaker cannot actually produce vowels with formant frequencies like those resulting from the calculations, and that the area functions are simply in error.

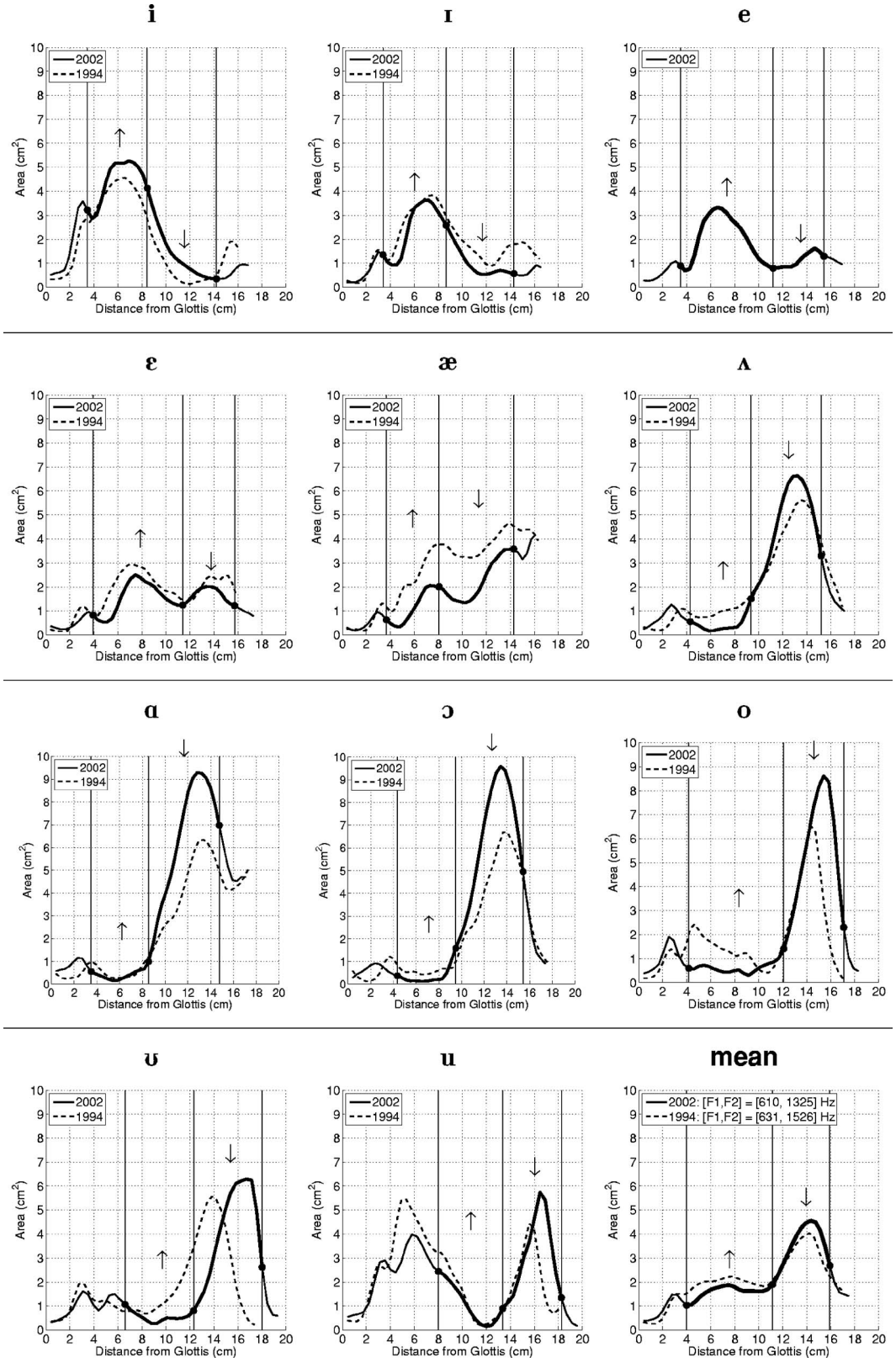
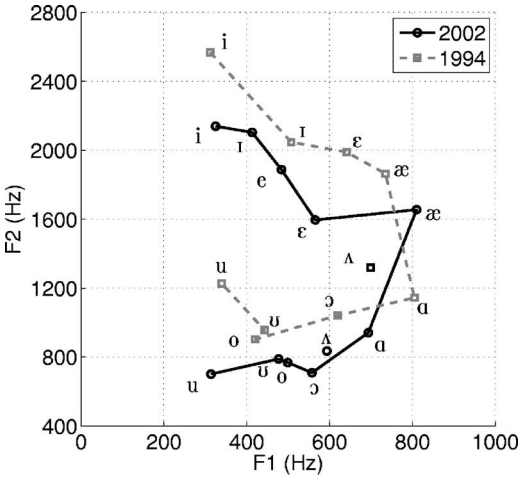
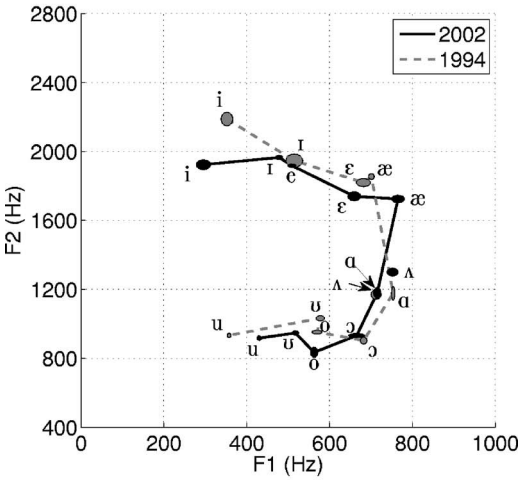


FIG. 1. Two sets of vocal tract area functions obtained from the same speaker. The solid lines represent area functions from the present study and the dashed lines are those area functions reported in STH96. The lower right subplot contains the mean area functions from each set. The legend on each subplot designates each set according to the year (1994 or 2002) in which the images were collected with MRI. The solid dots on the 2002 area functions along with the vertical lines denote regions within which a change of cross-sectional area, in the direction of the arrows, would increase F2.

The alternative conclusion begs the question of whether the speaker can actually produce vowels with the same downward shift in F2 as produced by the calculations. To test this, an audio recording was recently (September 2007) made of the same speaker producing two different series of 11 hVd syllables (a vowel embedded between an initial /h/ and final



(a) Calculated formant frequencies



(b) Measured formant frequencies

FIG. 2. Vowel space plot of F1 and F2 frequencies calculated and measured from the 1994 and 2002 versions of each vowel (the [e] exists only for the new set). The data points are connected by solid or dashed lines to clarify the set to which they belong and to provide a rough outline of the possible vowel space. Because of their positions in the F2 vs F1 plane, the two [ʌ] vowels are not connected to the other vowels within their respective sets. (a) Calculated formant frequencies. (b) Measured formant frequencies. Here the data points are indicated by ellipses in which the horizontal and vertical extents denote ±1 s.d. in the F1 and F2 dimensions, respectively; the mean value is located in the center of each ellipse.

/d/) containing the same 11 vowels as represented by the 2002 area function set. In the first series, the syllables were spoken in a pharyngealized/yawny quality. This was done based on the speaker's knowledge that the pharynx should tend toward a constrictive configuration to the degree allowed by a given vowel. For contrast, a second series of hVds was produced by releasing the constrictive effect in the pharynx and imposing somewhat of a constrictive tendency in the palatal region. The speaker, who was in the supine position with ear plugs inserted, was recorded while producing three repetitions of each hVd. The audio signals were saved in digital form directly to a computer disk. The speaker received no feedback concerning formant frequency locations during the recording. Formant frequencies were extracted from the vowel portion of each hVd (three repetitions of each) with an LPC technique programmed in MATLAB

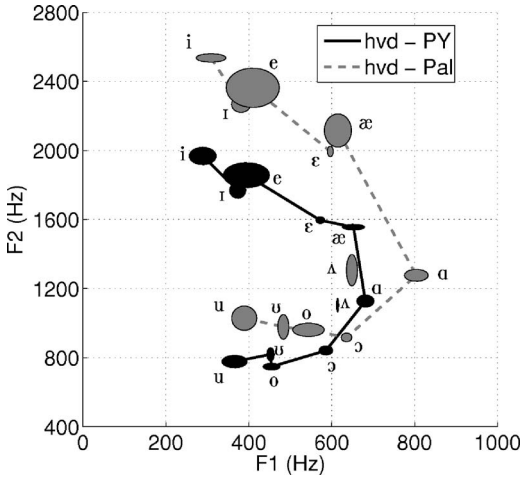


FIG. 3. Vowel space plot of F1 and F2 frequencies measured for the vowels in two series of hVd syllables. The data points are indicated by ellipses as they were in Fig. 2. The vowel formants denoted by solid black points are for a pharyngealized/yawny quality (PY), whereas the gray points indicate formants for the palatalized quality (Pal).

(these were verified to be essentially the same as would have been given by the PRAAT analysis algorithm). The results are shown in Fig. 3. The vowels of the pharyngealized/yawny quality, indicated by the black ellipses (horizontal and vertical extent represents ±1 s.d.), have a range of F2 values similar to those produced by the 2002 area functions and are shifted downward relative to the platalized vowels (gray ellipses). The standard deviations are larger than for the sustained vowels shown in Fig. 2(b) because of the nearly continuous movement of the vocal tract required to produce an hVd syllable.

That the speaker was indeed capable of producing a set of vowels with second formant frequencies as low as those calculated from the 2002 area functions, and at the same time could also produce a set of vowels with much higher values of F2 (more like those of the 1994 set), suggests that the observed differences in area function shapes could have resulted from different vocal tract "settings" rather than anatomical or dialectal change, or from imaging method differences. The reason that differences in F2 were less extreme between the measured formants from 1994 and 2002 [Fig. 2(b)] compared to the F2 differences of the corresponding calculated formants is apparently because the speaker utilized a more extreme version of the pharyngealized/yawny setting during image collection than in the subsequent audio recording. Perhaps this setting allows for production of vocal tract shapes that are easier to maintain over many repetitions in the noisy environment of a MR scanner.

The second set of area functions reported in the present study provides additional instances of target vocal tract shapes produced by one specific speaker. These data show that the vocal tract shape may be highly variable for the same target vowel depending on the particular setting used by the speaker. Such multiple instances of target vocal tract shapes may be useful for understanding intraspeaker variability and for purposes of speech synthesis and speech production modeling.

ACKNOWLEDGMENTS

The author would like to thank Ted Trouard for consulting on image acquisition and Jennifer Johnson for operating the MR scanner. This research was supported by NIH Grant No. R01-DC04789.

Alwan, A. A., Narayanan, S. S., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. 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.

Boersma, P., and Weenink, D. (**2007**). "PRAAT, Version 4.6.09," www.praat.org. (last viewed on 8 August, 2007).

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

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

Fant, G., and Pauli, S. (**1975**). "Spatial characteristics of vocal tract resonance modes," in Proceedings of the Speech Communications Seminar, Vol. **74**, Stockholm, Sweden, 1–3 August, pp. 121–132.

Kollias, I., and Krogstad, O. (**1999**). "Adult craniocervical and pharyngeal changes—a longitudinal cephalometric study between 22 and 42 years of age. I. Morphological craniocervical and hyoid bone changes," Eur. J. Orthod. **21**, 333–344.

Labov, W. (**1996**). "The organization of dialectic diversity in North America," Presented at the Fourth International Conference on Spoken Language Proceedings, Philadelphia, 6 October. Available online at www.ling.upenn.edu/phono_atlas/ICSLP4.html. (last viewed on 6 August 2007).

Lakshminarayanan, A. V., Lee, S., and McCutcheon, M. J. (**1991**). "MR imaging of the vocal tract during vowel production," J. Magn. Reson Imaging **1**, 71–76.

Laver, J. (**1980**). "The Phonetic Description of Voice Quality," Cambridge University Press, Cambridge, UK.

Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type line analog," DS dissertation, Dept. of Speech Communication and Music Acoustica, Royal Institute of Technology, Stockholm, Sweden.

Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**, 1683–1692.

The Mathworks (**2007**). "MATLAB, Version 7.4.0.287," R2007a.

Meyer, P., Wilhelms, R., and Strube, H. W. (**1989**). "A quasiarticulatory speech synthesizer for German language running in real time," J. Acoust. Soc. Am. **86**, 523–539.

Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. (**2006**). "Principal components of vocal tract area functions and inversion of vowels by linear regression of cepstrum coefficients," J. Phonetics **35**, 20–39.

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

Narayanan, S. S., Alwan, A. A., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. I. The laterals," J. Acoust. Soc. Am. **101**, 1064–1077.

Narayanan, S. S., Alwan, A. A., and Song, Y. (**1997**). "New results in vowel production: MRI, EPG, and acoustic data," Proceedings of the 1997 European Speech Proceedings Conference, Rhodes, Greece, Vol. **2**, pp. 1007–1009.

Schroeder, M. R. (**1967**). "Determination of the geometry of the human vocal tract by acoustic measurements," J. Acoust. Soc. Am. **41**, 1002–1010.

Sondhi, M. M., and Schroeter, J. (**1987**). "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**, 955–967.

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

Story, B. H. (**2005a**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am., **117**, 3231–3254.

Story, B. H. (**2005b**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.

Story, B. H. (**2006**).

Story, B. H., Laukkanen, A. M., and Titze, I. R. (**2000**). "Acoustic impedance of an artificially lengthened and constricted vocal tract," J. Voice **14**, 455–469.

Story, B. H., and Titze, I. R. (**1998**). "Parameterization of vocal tract area functions by empirical orthogonal modes," J. Phonetics **26**, 223–260.

Story, B. H., and Titze, I. R. (**2002**). "A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function," J. Phonetics **30**, 485–509.

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.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1998**). "Vocal tract area functions for an adult female speaker based on volumetric imaging," J. Acoust. Soc. Am. **104**, 471–487.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**2001**). "The relationship of vocal tract shape to three voice qualities," J. Acoust. Soc. Am. **109**, 1651–1667.

Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (**2006**). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," J. Acoust. Soc. Am. **119**, 1037–1049.

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

West, K. S., and McNamara, J. A. (**1999**). "Changes in the craniofacial complex from adolescence to midadulthood: A cephalometric study," Am. J. Orthod. Dentofacial Orthop. **115**, 521–532.

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

Yehia, H. C., Takeda, K., and Itakura, F. (**1996**). "An acoustically oriented vocal-tract model," IEICE Trans. Inf. Syst. **E79-D**, 1198–1208.