Acoustic modeling of American English /r/

Carol Y. Espy-Wilson[a)]
Electrical and Computer Engineering Department, Boston University, Boston, Massachusetts 02215 and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Suzanne E. Boyce
Department of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, Ohio 45267 and Electrical and Computer Engineering Department, Boston University, Boston, Massachusetts 02215

Michel Jackson
Electrical and Computer Engineering Department, Boston University, Boston, Massachusetts 02215

Shrikanth Narayanan
AT&T Labs, Florham Park, New Jersey 07932

Abeer Alwan
Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, California 90024

(Received 29 December 1998; accepted for publication 22 March 2000)

Recent advances in physiological data collection methods have made it possible to test the accuracy of predictions against speaker-specific vocal tracts and acoustic patterns. Vocal tract dimensions for /r/ derived via magnetic-resonance imaging (MRI) for two speakers of American English [Alwan, Narayanan, and Haker, J. Acoust. Soc. Am. **101**, 1078–1089 (1997)] were used to construct models of the acoustics of /r/. Because previous models have not sufficiently accounted for the very low $F3$ characteristic of /r/, the aim was to match formant frequencies predicted by the models to the full range of formant frequency values produced by the speakers in recordings of real words containing /r/. In one set of experiments, area functions derived from MRI data were used to argue that the Perturbation Theory of tube acoustics cannot adequately account for /r/, primarily because predicted locations did not match speakers' actual constriction locations. Different models of the acoustics of /r/ were tested using the Maeda computer simulation program [Maeda, Speech Commun. **1**, 199–299 (1982)]; the supralingual vocal-tract dimensions reported in Alwan *et al.* were found to be adequate at predicting only the highest of attested $F3$ values. By using (1) a recently developed adaptation of the Maeda model that incorporates the sublingual space as a side branch from the front cavity, and by including (2) the sublingual space as an increment to the dimensions of the front cavity, the mid-to-low values of the speakers' $F3$ range were matched. Finally, a simple tube model with dimensions derived from MRI data was developed to account for cavity affiliations. This confirmed $F3$ as a front cavity resonance, and variations in $F1$, $F2$, and $F4$ as arising from mid- and back-cavity geometries. Possible trading relations for $F3$ lowering based on different acoustic mechanisms for extending the front cavity are also proposed. [S0001-4966(00)00407-0]

PACS numbers: 43.70.Fq [AL]

INTRODUCTION

Historically, models for the more articulatorily complex liquids have been less well developed than models for vowels and obstruent consonants (Chiba and Kajiyama, 1941; Fant, 1960, 1980; Stevens, 1999; Rubin, Baer, and Mermelstein, 1981; Harshman, Ladefoged, and Goldstein, 1977; Maeda, 1982). Recently, however, several researchers have proposed models of the acoustics of American English /r/ (Stevens, 1999; Alwan *et al.*, 1997; Hagiwara, 1995; Veatch, 1990; Ohala, 1985) and similar rhotic sonorants (McGowan, 1994; Narayanan *et al.*, 1999). Several aspects of /r/ make this a complicated and interesting task. First, /r/ is characterized by a particularly stable acoustic pattern of $F3$ lowering close to the value of $F2$ (Boyce and Espy-Wilson, 1997; Guenther *et al.*, 1999). However, the acoustic means by which $F3$ is lowered has not been clear. Second, speakers of ''rhotic'' varieties of American English use a multitude of articulatory configurations (Westbury *et al.*, 1999; Delattre and Freeman, 1968; Zawadaski and Kuehn, 1980; Alwan *et al.*, 1997; Ong and Stone, 1998). These configurations may involve substantially different tongue shapes and different parts of the tongue as primary articulators, but researchers so far have failed to link patterns of acoustic variability in formant values with the different articulatory configurations. For instance, Delattre and Freeman's (1968) ground-

a)Electronic mail: espy@bu.edu

breaking x-ray study of articulatory configuration types across speakers failed to identify consistent acoustic differences associated with different types. Similarly, studies by Westbury *et al.* (1999) and Guenther *et al.* (1999) have found that different types of articulatory configurations could not easily be correlated with specific patterns of formant values. The conclusion has been that different configurations produce essentially equivalent acoustical profiles (Delattre and Freeman, 1968; Westbury *et al.*, 1999). As such, American English /r/ is often cited as an example of a many-to-one articulatory–acoustic relationship. Third, acoustic models of /r/ must deal with three constrictions along the vocal tract, and the dimensions of the cavities thus formed were not known from physiological data until recently (Moore, 1992; Alwan *et al.*, 1997; Ong and Stone, 1998). This is particularly true of the sublingual space, which is not easily estimated using surface point-tracking systems such as x-ray microbeam or palatography (but see Sundberg *et al.*, 1992, for an alternative method).

Because of the difficulty in obtaining precise physiological data for vocal tract dimensions during a segment of interest, investigators developing acoustic models have for the most part been forced to make plausibility the criterion for assumptions about physical vocal tract features, such as constriction size, cavity volume, etc. In addition, the criteria for a successful match between predicted and actual acoustic patterns have been in terms of ability to capture general patterns of phonemic difference across multiple speakers. Recently, however, advances in physiological data collection methods have made it possible to test the accuracy of predictions against speaker-specific vocal tracts and acoustic patterns. This advance enables model makers to use a more demanding set of criteria; namely, a successful match between predicted and actual patterns for a given speaker's vocal tract. For instance, Story *et al.* (1996) used MRI-derived vocal-tract area functions from a single speaker to obtain simulated formant frequencies for vowels that were almost always within 10% of those measured from natural speech. In a similar study, Yang and Kasuya (1994) obtained vowel formant frequencies from a model using MRI-derived area functions that were within 5% of those measured from natural speech.

In this paper, we use recently available magnetic resonance imaging (MRI) data for American English /r/ (Alwan *et al.*, 1997) to examine various current theoretical models of the acoustics of /r/ in greater detail and with narrower criteria than has previously been possible. These data, and modifications thereof, are used to supply vocal-tract dimensions for VTCALCS, Maeda's vocal-tract computer modeling software (1982). Our major aim is to understand the acoustic mechanism responsible for the distinctive characteristics of /r/, in particular, its unusually low $F3$.

A. Acoustics of /r/

American English /r/ occurs both as a syllable nucleus and in consonantal position, where it is classified as a sonorant liquid.[1] The characteristic formant pattern for both involves an $F1-F2$ pattern similar to that of a canonical central, rounded vowel (Espy-Wilson, 1992), together with a very low $F3$. A typical example of intervocalic /r/, illustrating the severe dip in $F3$, is shown spectrographically in Fig. 1. As in this case, $F3$ is often low enough to approach and/or merge with $F2$ (Stevens, 1999). $F1$ and $F2$ values are predictable from the general articulatory shape of /r/, and accordingly overlap with those of vowels with similar place and height features. The $F3$ of such vowels is around 2500 Hz or above for most speakers (cf. Hagiwara, 1995). Consequently, the major problem in modeling the acoustics of American English /r/ is accounting for the very low third formant. Although formant values for /r/ vary somewhat from speaker to speaker and across prosodic conditions such as initial vs final word position, syllabic vs consonantal function, etc., $F3$ remains low relative to other segments. In a study of male–female and subject-to-subject variability in formant frequencies for /r/, Hagiwara (1995) used $F3$ in the neutral vowel to normalize between-speaker formant frequencies. He then determined that $F3$ for any one subject falls between 60% and 80% below the $F3$ value in the neutral vowel. For most speakers, this puts $F3$ below 2000 Hz. The range of actual values reported in the literature (across sex, syllabic vs consonantal prosodic position, etc.) is approximately 250–550 Hz for $F1$, 900–1500 Hz for $F2$, and 1300–1950 Hz for $F3$ (Delattre and Freeman, 1968; Lehiste, 1962; Zawadaski and Kuehn, 1980; Nolan, 1983; Espy-Wilson, 1992; Westbury *et al.*, 1999; but see Hagiwara, 1995, for outliers).

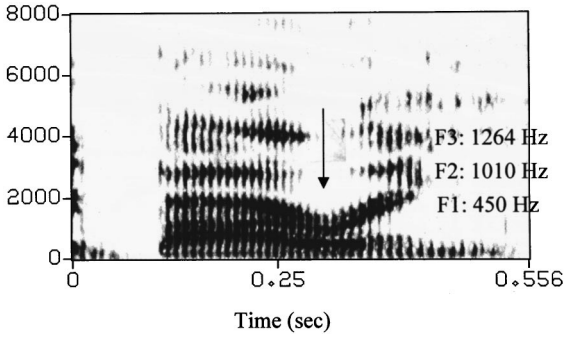

FIG. 1. Spectrogram of the word ''barring'' spoken by a male speaker.

B. Articulation of /r/

Overall, articulatory configurations for /r/ involve three constrictions: in the pharynx, along the palatal vault, and at the lips. The configurations differ most by what happens in the palatal region, i.e., by whether the effective constriction occurs (1) at the alveolar ridge and is made solely by the tongue tip, (2) in the palatovelar region and is made solely by the tongue dorsum with a lowered tongue tip, or (3) in both alveolar and palatovelar regions, and is made by the simultaneous raising of the tongue tip and tongue dorsum. Traditionally, these configurations have been divided into contrasting categories of ''retroflex'' (in which the tongue tip is raised and the tongue dorsum is lowered) versus ''bunched'' (in which the tongue dorsum is raised and the tongue tip lowered) (Delattre and Freeman, 1968; Shriberg and Kent, 1982; Kent, 1998). However, as a number of re-

searchers have pointed out, these two categories are only the extremes in a continuum that includes many incremental variants (Zawadaski and Kuehn, 1980; Delattre and Freeman, 1968; Westbury *et al.*, 1999; Alwan *et al.*, 1997; Guenther *et al.*, 1999; Ong and Stone, 1998). Further, the variant in which both tongue tip and tongue dorsum are raised does not easily fit into the traditional dichotomy of retroflexed versus bunched. In this paper, therefore, the major types of configurations will be categorized as (1) tip-up retroflex /r/, (2) tip-up bunched /r/, and (3) tip-down bunched /r/. It is worth noting that these different configurations occur both within and across speakers; that is, while some speakers may use one type of configuration exclusively, other speakers switch between two or three different types of configurations for /r/ in different phonetic contexts (Delattre and Freeman, 1968; Guenther *et al.*, 1999), and according to prosodic variables such as word position, syllabic versus consonantal function, etc. (Delattre and Freeman, 1968; Zawadaski and Kuehn, 1980).

C. Modeling /ɹ/: Acoustic sources of *F*3

Proposed models of the acoustics of American English /r/ divide into two types: (1) the Perturbation Theory account, which has been promulgated primarily through teaching and laboratory demonstrations,[2] but is described in Johnson (1997) and Ohala (1985), and (2) decoupling accounts, exemplified by the model of Stevens (1999) and modified versions by Alwan *et al.* (1997) and Narayanan *et al.* (1999). The Perturbation Theory account of /r/ is based on a general principle of tube acoustics; namely, that for a relatively open tube, constrictions at points where standing waves have maximum volume velocity have the effect of lowering the natural resonances of the resonating tube (Chiba and Kajiyama, 1941; Heinz, 1967; Schroeder, 1967). Such points of sensitivity to constriction are commonly invoked to explain the formant-lowering effect of lip-aperture narrowing and to predict the effect of certain articulatory changes on vowel acoustics (Stevens, 1999, p. 284; also Mrayati, Carré, and Guérin, 1988). It happens that when the vocal tract is modeled as a quarter-wavelength tube, maximum volume velocity points for *F*3 occur in the pharyngeal, palatal, and labial regions. Because the common denominator across various types of /r/ is the presence of constrictions in these three regions, some investigators have suggested that the lowering effects of all three combine to produce the low *F*3 typical of /r/ (Hagiwara, 1995; Veatch, 1990; Ohala, 1985). Indeed, modeling trials where realistic constrictions are inserted at points of maximum velocity produce *F*3 values in the appropriate range for /r/ (Espy-Wilson *et al.*, 1997). The usefulness of this approach for modeling speech sounds in general is controversial (Boe and Perrier, 1990); with regard to /r/ this is primarily because the perturbation effect applies only to cases where constrictions are mild.[3] Another prediction of the Perturbation Theory account is that speakers place their actual constrictions for /r/ at extremely specific points along the vocal tract, i.e., the points of maximum volume velocity. Thus, a finding that speakers place constrictions at other points along the vocal tract would be evidence against this model of /r/ acoustics.

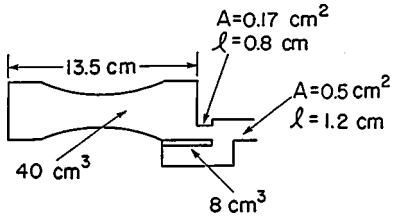

FIG. 2. Stevens' simple tube model for a tip-up retroflex /r/. The symbol ''*A*'' stands for area and the symbol ''*l*'' stands for length. The orientation of this model is such that the glottis is at the left edge and the lips are at the right edge. (Reprinted with permission from K. Stevens, *Acoustic Phonetics*, MIT Press.)

In contrast to the Perturbation Theory account, decoupling accounts, such as Fant (1960), Stevens (1999), Alwan *et al.* (1997), and Narayanan *et al.* (1999), assume that the vocal tract is divided into several different tubes, of differing areas and lengths, and that different formants have their origin as resonating frequencies of different tubes. Thus, a major issue for decoupling accounts is the question of boundary conditions for the particular shapes, lengths, and proportions of tubes in the vocal tract. If a constriction at one end of a cavity is narrow enough, for instance, the cavity is modeled as a tube with a closed end. If a constriction is wide enough, cavities on either side of the constriction are no longer decoupled. If both ends of a cavity have narrow constrictions, and the cavity-to-constriction ratio is high enough, the cavity and the constriction anterior to it can be modeled as a Helmholtz resonator. If, on the other hand, the constriction posterior to the cavity is narrow and the constriction anterior to the cavity is wide, the cavity and the anterior constriction can be modeled as a quarter-wavelength tube. The choice of such boundary conditions determines the appropriate equations for estimating formant frequencies and, conversely, determines which of many possible combinations can match a particular set of formant frequencies. Until recently, the dimensions of cavities and constrictions in the vocal tract had to be estimated, based on what seemed anatomically likely and what worked to produce formant frequencies approximately in the correct range. As we show in this paper (see Sec. II E, simulation experiment 5), availability of directly observed, segment-specific physiological data allows the boundary conditions for modeling to be determined substantially more accurately.

The most detailed model of the acoustics of /r/ can be found in Stevens (1999), which was primarily designed for articulatory configurations found in tip-up /r/'s where the tongue dorsum is lowered. A sketch of this model is shown in Fig. 2. Note that this model assumes a sizable difference between the area of the back cavity (about 3 cm^2) and the area of the front cavity (0.5 cm^2). Also, Stevens included a substantial sublingual space with a volume of 8 cm^3. Similar models are given in Fant (1960), Alwan *et al.* (1997), and Narayanan *et al.* (1999). In Stevens' model, the palatal constriction is assumed to be narrow enough, and the lip constriction narrow enough, that the cavity behind the lips (i.e.,

the ''front'' cavity) and the lip cavity can be modeled together as a quarter-wavelength tube that is closed at the palatal end and open at the lip end. $F3$ is assumed to be the lowest resonance of this quarter-wavelength tube.[4] Given this assumption, the front cavity must be 5 cm long to produce a resonance at 1750 Hz—a value in the middle of the typical $F3$ range for English speakers. It must be correspondingly longer in order to produce a resonance in the lower portion of the typical range, e.g., 1300–1600 Hz. This model may also be modified in such a way that, if the lip constriction is narrow enough, the front part of the vocal tract may resemble a Helmholtz resonator. That is, it may consist of a relatively large cavity with an anterior constriction (due to either lip rounding or natural tapering by the teeth and lips). In this case, a high volume-to-constriction degree ratio is required in order to produce an $F3$ within the typical range.

I. METHOD

A. Magnetic resonance imaging data

The data used for this study were collected by Alwan *et al.* (1997). They used magnetic resonance imaging (MRI) to collect data from four phonetically trained native American English speakers who produced sustained /r/'s in two different production conditions, as described below. Subjects sustained each sound for 13–16 s in a supine position, enabling four or five image slices to be recorded in that time period (about 3.2 s per image). Recordings were made in sagittal, axial, and coronal planes. Subjects repeated each sound six to nine times, with a pause of 3 to 10 s between repetitions, to enable the entire vocal tract to be scanned. Further details of the MRI recording, data acquisition, and analysis methodologies are provided in Alwan *et al.* (1997) and Narayanan *et al.* (1997).

Production conditions for Subject PK consisted of instructions to produce a retroflex and a bunched /r/. Tracings of the midsagittal profiles for these productions are shown in the left and right lower panels of Fig. 3. Note that the production labeled as tip-up bunched /r/ in Fig. 3 came from the retroflex condition, and the production labeled as tip-down bunched /r/ came from the bunched condition.[5] In other words, for the retroflex condition, the speaker produced an /r/ with both dorsum and tongue tip raised. This was somewhat of a surprise, as the traditional image of a retroflex /r/ involves a lowered tongue dorsum (Delattre and Freeman, 1968; Kent, 1998). Production conditions for Subject MI consisted of intentional production of /r/'s as they would occur in word-initial position and syllabically. These are again shown in the left and right upper panels of Fig. 3. Notably, Subject MI also produced /r/'s with raised, or bunched tongue dorsum; his tongue tip is slightly raised in his ''word-initial'' /r/. Altogether, we consider four sets of vocal-tract dimensions from two speakers, corresponding to the four vocal-tract profiles in Fig. 3. All fit the category of bunched, i.e., the tongue dorsum is up.

MRI data consisted of cross-sectional images, from which vocal-tract area functions were calculated. Area functions were measured separately for the sublingual space

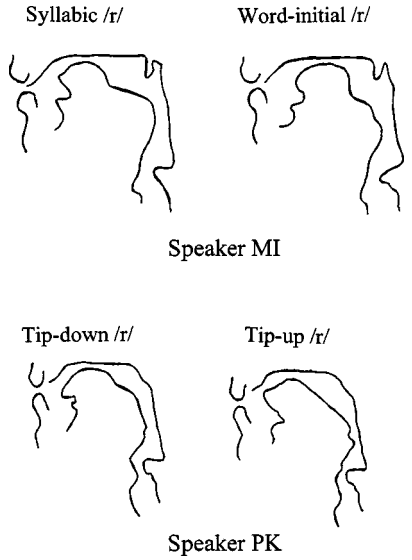

FIG. 3. Vocal-tract profiles for /r/ in midsagittal plane (adapted from Alwan *et al.*, 1997).

(when appropriate) and for the continuous supralingual space running from the lips to the glottis. Figure 4 shows the area functions for the supralingual space for each of the four data sets. The data are positioned such that the glottis is to the left, at 0 cm, while the lip opening is on the right. Larger areas under the curve indicate vocal-tract cavities; constrictions occur where the distance between the data curve and the abscissa are small. These area functions served as input for the Maeda (1982) vocal-tract model VTCALCS, a computer program using standard acoustical tube assumptions to predict formant frequencies from vocal-tract dimensions. We used a MATLAB version of the original VTCALCS program developed by Dr. Ronan Scaife (personal communication). For Subject PK this involved a model vocal tract of 15.3 cm, divided into 51 sections of 0.3-cm length. For Subject MI, this involved a model vocal tract of 18 cm, divided into 60 sections of 0.3-cm length.

The cross-sectional areas of the sublingual space measured from MRI data for each subject and production condition are listed in Table I. Note that sub- and supralingual space was measured from coronal sections in the superior–inferior plane; thus, the sublingual space is defined as any space bounded vertically by the tongue and buccal floor. For PK, the sublingual space was measured as being three sections (i.e., 0.9 cm) long for both her tip-up and tip-down production conditions. Sublingual space lengths were measured as four sections (i.e., 1.5 cm) long in the case of MI's word-initial /r/ and four sections (1.2 cm) in the case of MI's syllabic /r/. The sublingual cavity started at 3 cm from the lips for both of MI's productions, and at 2.1 cm from the lips in both of PK's production conditions.[6]

B. Audio recording

Due to noise in the experimental chamber, it was not possible to record speakers' acoustic output during the MRI sessions. However, each subject recorded four repetitions of nine (PK), or ten (MI) real words containing consonantal /r/ or syllabic /r/. A list of the words is given in the Appendix.

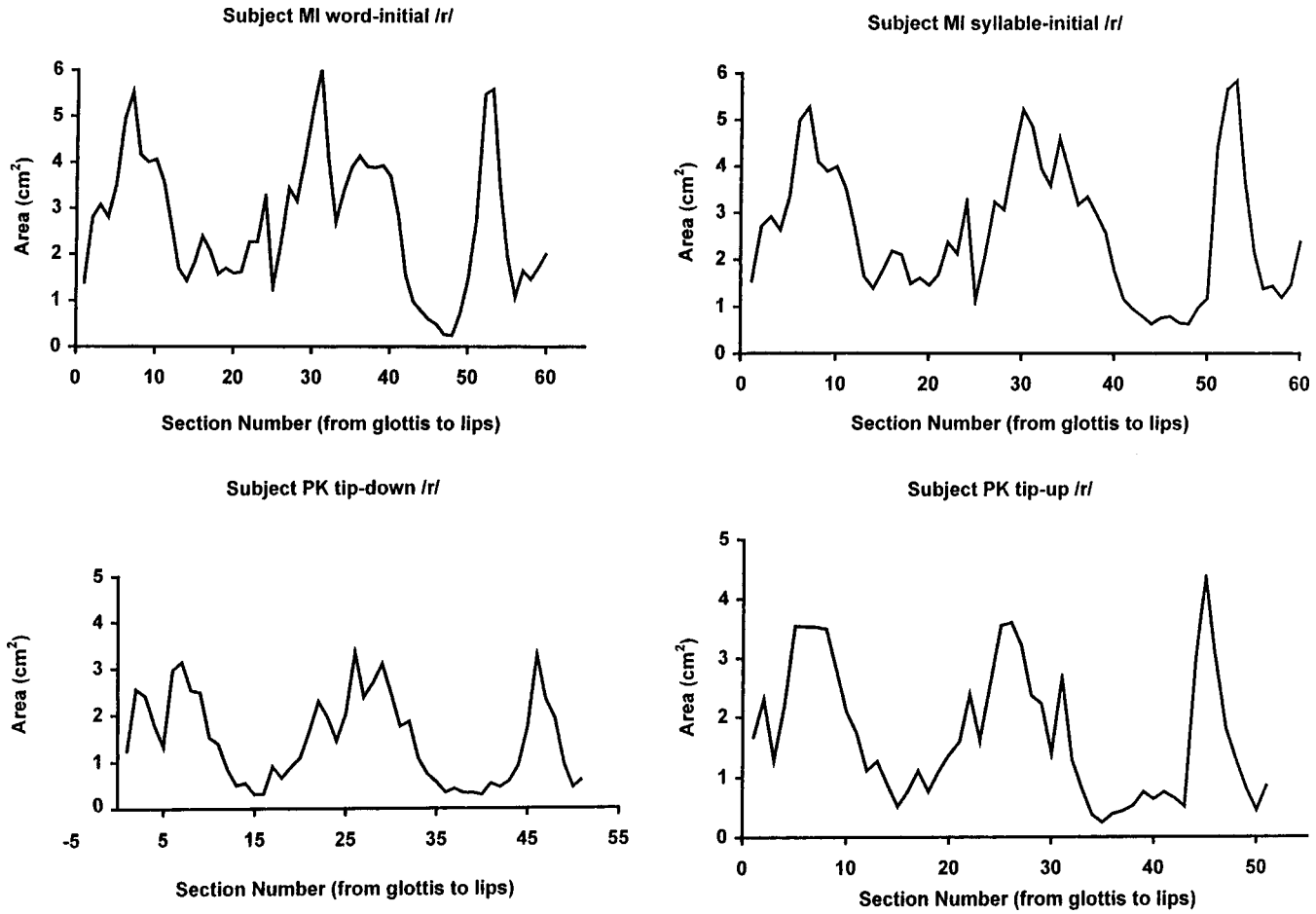

FIG. 4. MRI-derived supralingual area functions for the /r/'s produced by PK and MI.

These words were embedded in the carrier phrase ''Say ___ again.'' The speech data were recorded in a soundproof facility at 44.1 kHz directly onto a Sun workstation and were later downsampled to 11.025 kHz. An omnidirectional microphone (Beyerdynamic M101) with a flat frequency response (within 4 dB) between 40–20 000 Hz was placed at approximately 22 cm from the subject's mouth at about a 15° angle off the midline. The Entropic Signal Processing Systems (ESPS) Waves environment was used to display spectrograms of each utterance and automatically track the first four formants. Measurement of the formants was made at the lowest point of $F3$. In a few utterances (one for PK and three for MI), it was not possible to get the frequency of $F4$ when $F3$ was at its lowest point. The energy above $F3$ was very weak, so $F4$ was not visible in the spectrogram and was not properly tracked. Thus, these values of $F4$ were not considered in our analysis.

TABLE I. MRI-derived areas of sublingual space (in cm^2). Each section is of length 0.3 cm, resulting in a sublingual cavity of 0.9 cm for PK, 1.5 cm for MI's word-initial /r/ and 1.2 cm for MI's syllabic /r/. The sublingual cavity starts 3 cm from the lips in the case of MI, and 2.1 cm from the lips in the case of PK.

PK tip up	2.707	0.897	0.594		
PK tip down	1.025	0.984	0.465		
MI initial	1.833	0.865	0.542	0.419	0.156
MI syllabic	2.88	1.251	0.659	0.306	

C. Vocal-tract acoustic model

Vocal-tract modeling in this study is based on the VTCALCS program developed by Maeda (1982). This time-domain simulation of the vocal tract assumes one-dimensional wave propagation and includes acoustic losses due to yielding walls, fluid viscosity, and radiation effects from the mouth. VTCALCS as originally written allows for a side branch corresponding to the nasal cavity in the velar region, but this option is not suitable for modeling the sublingual space. To model the sublingual space, we modified the original VTCALCS program (in its MATLAB version) to allow for a parallel side branch in the palatal region (Jackson *et al.*, submitted). This modeling procedure is discussed in detail in Sec. III D.

II. SIMULATION EXPERIMENTS

An ideal model of the acoustics of /r/ will account for the full range of variability in formant values exhibited by all speakers of American English. However, since the MRI data from our speakers may not match all speakers' vocal tracts, our aim was to match the range of variability in formant values for our particular speakers, as shown in their real word productions. All experimental simulations in this paper used as their base the vocal-tract area functions reported in Alwan *et al.* (1997) and here in Fig. 4 and Table I. Table II contains the $F1-F4$ data measured from the subjects' pro-

TABLE II. Real word formant frequencies (in Hz). For Subject PK, $N=36$; for Subject MI, $N=40$.

	Subject PK			Subject MI		
	Mean	Range	s.d.	Mean	Range	s.d.
$F1$	349.5	202–517	98.0	388.0	234–489	78.3
$F2$	1355.4	989–1698	147.7	1383.8	989–1586	153.7
$F3$	1833.8	1479–2157	137.7	1664.9	1400–1946	127.4
$F4$	4110.8	3898–4483	134.8	3113.7	2742–3483	139.3

duction of real words containing /r/. These data, including the full range of $F3$ values produced by the speakers, are the standard against which we measure the success or failure of a particular modeling schema. Given the difficulty of accounting for very low $F3$ in current models, we particularly focus on accounting for the full range of $F3$.

In experiment 1, we use these data to estimate whether the location of actual constrictions seen in MRI data match the locations predicted by the Perturbation Theory. In experiment 2, we present estimates of formant values produced when the supralingual vocal-tract area functions of Fig. 4 are input to the VTCALCS program, and compare the results to those from real words. In experiments 3 and 4, we present estimates of the formant values when the sublingual area functions are taken into account. Finally, in experiment 5, we develop a simple tube model that accurately predicts $F1$ through $F4$ while accounting for formant cavity affiliations across different articulatory configurations for /r/.

A. Experiment 1: Locations predicted by perturbation theory

As noted above, the Perturbation Theory approach to /r/ assumes that subjects' constriction locations will coincide with the points of maximum volume velocity predicted by the Perturbation Theory.

To determine the predicted locations for our speakers, we calculated points of maximum volume velocity (where constriction decreases $F3$), and maximal pressure (where constriction increases $F3$), for the vocal-tract lengths 15.3 cm (for PK) and 18 cm (for MI). Table III shows how these predicted constriction locations compare with the constriction locations found in the MRI data. Because the MRI data constriction locations extended over several sections, these are expressed as ranges across which the constriction was maximal. The criterion for constriction beginning and end was set at area$=1.0$ cm^2 for PK and at area$=2.4$ cm^2 for MI. Wide ranges indicate stretches for which constriction met criterion.

TABLE III. Real versus predicted constriction locations.

	Actual palate	Predicted palate	Actual pharynx	Predicted pharynx
PK tip up	10.2–13.3	9.6	4.3–6.2	3.1
PK tip down	10.5–13.6	9.6	4.3–5.9	3.1
MI initial	12.6–15.0	10.8	3.9–6.9	3.6
MI syllabic	12.0–15.0	10.8	3.9–6.3	3.6

TABLE IV. $F1-F4$ values predicted by VTCALCS from supralingual MRI data.

	PK tip up	PK tip down	MI initial	MI syllabic
$F1$ (Hz)	367.5	373.8	310.9	340.2
$F2$ (Hz)	1336.4	1175.7	1173.4	1210.2
$F3$ (Hz)	1938.9	2167.5	1921.0	1883.5
$F4$ (Hz)	4369.2	4132.2	3068.0	3275.9

It is clear that for both speakers, real palatal constrictions are long, and appear to cover an area considerably forward of the location predicted by the Perturbation Theory to have the maximal lowering effect on $F3$ (see Table III). Indeed, for both subjects, palatal constrictions center over areas predicted by the Perturbation Theory to correspond with maximal pressure, making constrictions in these areas more likely to raise $F3$ than to lower it. For PK, the pharyngeal constriction ranges forward of the predicted point. For MI, the pharyngeal constriction apparently covers an area that may be conducive to $F3$ lowering, but is longer than necessary. Thus, neither the pharyngeal nor the palatal constriction is located as would be predicted by the Perturbation Theory. In particular, the palatal constrictions here cover areas that should affect $F3$ in the wrong direction. This is true regardless of the type of /r/; for instance, PK's ''tip-down'' /r/ and ''tip-up'' /r/ have slightly different constriction lengths but similarly forward constriction locations and similar constriction degrees. We conclude that subjects are not taking advantage of points of maximum volume velocity along the vocal tract to lower $F3$ in any obvious way.

B. Experiment 2: Formant frequencies from MRI-derived supralingual area functions

As a first approximation to modeling the acoustics of /r/ from MRI data, the supralingual area functions from the four MRI data sets were used as input to VTCALCS. The resulting estimates of $F1-F4$ values are shown in Table IV.

These estimated formant values are within the general range of $F1$, $F2$, and $F3$ reported across different studies using different speakers of American English (see the Introduction, Sec. A above). Further, the estimates for $F1$ and $F2$ compare favorably with the subjects' real word data, being squarely within their respective ranges and not far from the average frequency values. In contrast, the estimates of $F3$, at 1938.9 and 2167.5 for PK and 1921 and 1883.5 for MI, are in the very high portion of the reported range across studies, and 2–3 standard deviations higher than the average $F3$ formant frequency values reported for speakers of American English. Further, these $F3$ values match only the highest values in each speaker's range. Thus, the shape of the supralingual vocal tract for these subjects (as measured by MRI) is apparently enough to specify $F1$ and $F2$, but is not enough to specify the most notable mid-to-lower values of $F3$.

There are several possible explanations for this result. First, because the MRI data were collected with the subject in a supine position, differences in the effect of gravity might account for the high $F3$ values. However, previous studies contrasting vowel formant frequencies produced in supine

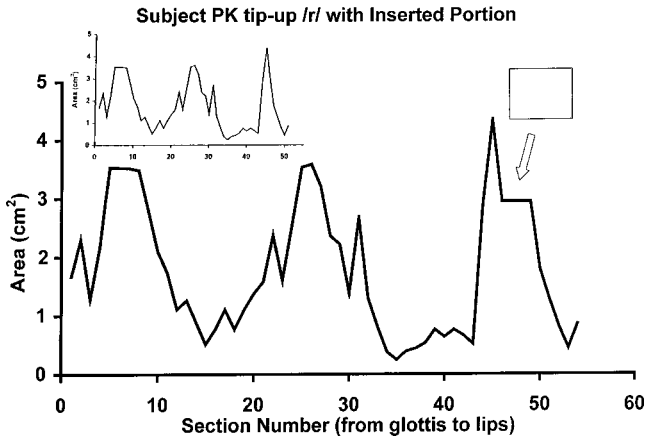

FIG. 5. MRI-derived area function for PK's tip-up /r/ with a uniform tube (the length of PK's sublingual space with area equal to the maximum cross-sectional area of the sublingual space) inserted into the front cavity at the measured location of the opening into the sublingual space.

versus upright position have found very minor effects (Shiller, Ostry, and Gribble, 1999; Tiede *et al.*, 1997; Tiede, personal communication). Second, because the MRI and audio data were collected at different times, it is possible that during MRI data collection the subject was producing the very high $F3$ values we see here. A more powerful explanation, however, comes from looking at the dimensions of the supralingual MRI vocal-tract data. According to Fant (1960), $F3$ is a front cavity resonance. As noted above, when modeled as a quarter-wavelength tube, the front cavity must be around 5 cm long to produce a resonance in the midrange of $F3$. The length of the front cavity in our supralingual MRI data is less than 2 cm. Thus, given the proportions of the MRI supralingual vocal tract, additional length, or alternative acoustic models, are needed to account for the mid-to-low $F3$ values produced by these speakers.

C. Experiment 3: Formant frequency estimates from extending the front cavity

Stevens (1999) and Alwan *et al.* (1997) suggest that the sublingual space adds to the volume of the front cavity and thereby lowers the front cavity resonance ($F3$). Thus, in our first attempt to model the sublingual space, we increased the volume of the front cavity by inserting a uniform tube into the front cavity portion of the MRI-derived area function plots, at the measured location of the opening into the sublingual space. The cross-sectional area of the uniform tube corresponded to the largest area of the sublingual space. The length corresponded to the measured length of the sublingual space. An example of this type of modification is shown in Fig. 5.

Table V shows the $F1-F4$ values predicted from the extended front cavity. A comparison of these data with those of Table IV shows that the added volume to the front cavity lowers $F3$ by 200 Hz to 300 Hz.

Alternatively, the volume of the front cavity can be increased by adding the area of the sublingual space.

TABLE V. $F1-F4$ values predicted by VTCALCS with extended front cavity.

		PK tip up	PK tip down	MI initial	MI syllabic
Length of front cavity increased	$F1$ (Hz)	360.8	360.6	295.7	337.7
	$F2$ (Hz)	1284.6	1162.8	1156.2	1180.0
	$F3$ (Hz)	1650.2	1949.9	1637.4	1616.3
	$F4$ (Hz)	4287.5	4079.6	3051.8	3255.1
Area of front cavity increased	$F1$ (Hz)	368.2	375.1	313.6	349.9
	$F2$ (Hz)	1317.8	1173.2	1170.8	1201.9
	$F3$ (Hz)	1766.2	1996.1	1793.7	1751.0
	$F4$ (Hz)	4357.3	4126.5	30568.2	3272.5

D. Experiment 4: Formant frequency estimates using side branch

In this part of the study, we modeled the sublingual cavity as a side branch. To do so, the MATLAB version of Maeda's VTCALCS program was revised to allow a side branch, as shown in Fig. 6. Sublingual MRI-derived area functions, measured as noted above, were used as input to the model. (See Table I for the data.) Each section of the side branch is modeled in the same way as the sections in the main tube, and the side branch terminates in an open circuit to model the effects of a hard wall (infinite impedance). Since the volume element of this side branch and the volume element of the front cavity are in parallel, they effectively add for the purposes of calculating the front cavity resonances. In addition, the model assumes that the sublingual space, acting as a side branch, generates an antiresonance that is proportional to its length. Given the 3–5-section length of the measured sublingual space, this antiresonance can be expected to fall in the range of 5–6 kHz. Details of the side-branch model are reported in Jackson *et al.* (submitted).

The formant values obtained are given in Table VI. As

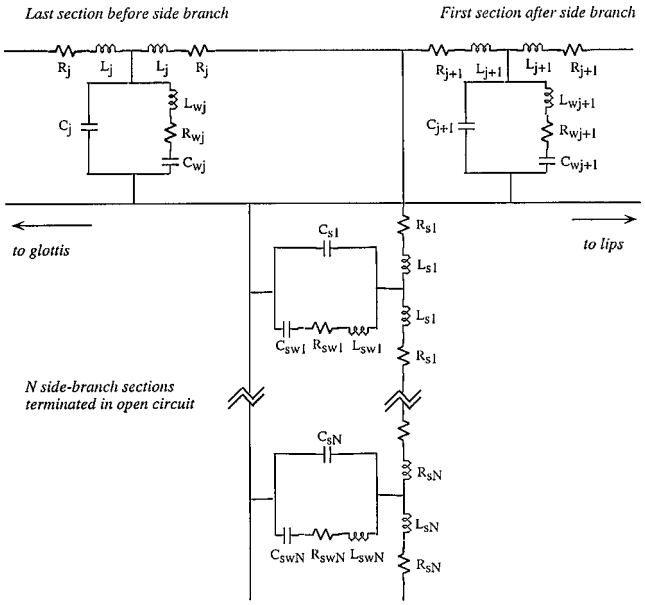

FIG. 6. A diagram representing how Maeda's vocal-tract model was modified to include a side branch. The side branch is terminated in an open circuit to model the effects of a hard wall (infinite impedance). R_w, L_w, and C_w model the impedance of the vocal-tract walls.

TABLE VI. $F1-F4$ values predicted by VTCALCS with sublingual cavity modeled from MRI data as a side branch.

	PK tip up	PK tip down	MI initial	MI syllabic
$F1$ (Hz)	366.6	373.3	310.5	348.4
$F2$ (Hz)	1302.4	1170.4	1167.2	1194.0
$F3$ (Hz)	1692.5	1947.6	1719.8	1682.5
$F4$ (Hz)	4334.9	4124.7	3064.1	3269.9

can be seen, taking the sublingual space into account does not change the frequencies of $F1$ and $F2$ very much, but it does lower $F3$ by about 200 Hz. These results are very close to those obtained for the manipulation of experiment 3 (see Table V), where the sublingual space was accounted for by increasing the volume of the front cavity by a simple extension. For the purpose of producing $F3$ values, these two different ways of accounting for the sublingual space are essentially equivalent. These results suggest that the sublingual space acts primarily to increase the volume of the front cavity, at least for the purpose of producing appropriate formant values. (Note that these different methods of accounting for the sublingual space make different predictions regarding other variables such as existence of an antiresonance, changes in formant bandwidth, etc. that are outside the scope of this paper.) Note also that the methods of experiments 3 and 4 may not necessarily be equivalent for articulatory configurations involving longer sublingual space dimensions, such as might be found in true tip-up retroflex /r/'s. However, both sets of results show a definite positive advance on previous modeling attempts, as we have achieved $F3$ values that match (approximately) the average $F3$ produced by our speakers (see Table II, real data).

E. Experiment 5: Formant frequency calculations using a simple tube model for /ɹ/

1. Simple tube models

In this section, we consider simple tube models for /r/, to determine which cavities contribute to which formants. We consider first the sources of formants $F1$, $F2$, and $F4$ as being the easier to predict, and then consider the source of $F3$. Overall, we attempt to build a single model with the potential to account for data-derived variations in constriction type, constriction degree, and cavity dimensions. For this analysis, we assume plane wave propagation and no acoustic losses. To be consistent with Maeda's vocal-tract modeling program, the value used for the speed of sound, c, in all calculations was 35 000 cm/s.

Stevens (1999) provides a simple acoustic tube model, complete with approximate dimensions, for the tip-up retroflex type of /r/ configuration as shown in Fig. 2. To account for the fact that our MRI-derived area functions show slightly different cavity and constriction dimensions and describe bunched /r/'s (some with the tongue tip up), we adapted Stevens' simple tube model to the generic version shown in Fig. 7. Differences between our model and that of Stevens include the following: (1) Stevens model assumed a relatively small supralingual front cavity with dimensions equal to the area of the lip opening. Because the MRI data show both a considerably larger oral cavity and tapering of

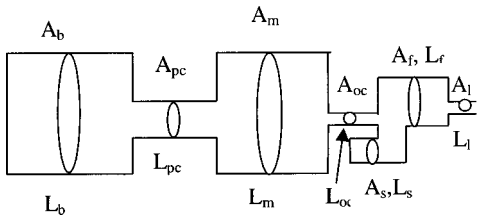

FIG. 7. Our simple tube model for the bunched /r/'s produced by PK and MI. A_b and L_b correspond to the area and length of the back cavity; A_{pc} and L_{pc} correspond to the area and length of the pharyngeal constriction; A_m and L_m correspond to the area and length of the midcavity between the pharyngeal constriction and the oral constriction; A_{oc} and L_{oc} correspond to the area and length of the oral (palatal) constriction; A_f and L_f correspond to the area and length of the front cavity between the oral constriction and the lip constriction; and A_l and L_l correspond to the area and length of the lip constriction.

the front part of the mouth around the teeth and lips, our model consists of two tubes, a larger one representing the front cavity and a narrower one representing the lip constriction. (2) Because the physiological data represents bunched /r/'s (both tip up and tip down), the palatal constriction in our model is considerably longer than that proposed by Stevens. As a result, the total length of each cavity between the palatal constriction and the glottis is shortened. (3) As Stevens did, we take into account the sublingual space in the acoustic modeling of /r/. However, Stevens assumes a considerably longer sublingual space than we found to exist in the MRI-derived data. Accordingly, in our model the sublingual space is shorter. (4) Finally, in Stevens' model the pharyngeal constriction is modeled as a slight narrowing of the tube rather than as a tight constriction. In our data, there was a qualitative difference between subjects' data in degree of pharyngeal constriction. To account for this, we allow for a variant of the simple tube model, in which the pharyngeal constriction is modeled as a separate tube rather than as a perturbation of a larger tube. For purposes of precision in terminology, we refer to the cavity between the lips and the palatal constriction as the ''front'' cavity, the cavity between the palatal and pharyngeal constrictions as the ''mid'' cavity, and the cavity behind the pharyngeal constriction as the ''back'' cavity. Based on data presented here, and elsewhere (Espy-Wilson *et al.*, 1997; Stevens, 1999; Narayanan *et al.*, 1999), we assume that $F3$ is a front cavity resonance.

Our simple tube model, although based on MRI-derived dimensions, is considerably simplified compared to the real dimensions. In what follows, we describe the series of tests we ran to determine how well our estimates of formant frequency values based on simple tube dimensions match estimates based on actual dimensions. To simplify comparisons, we compared formant frequency patterns calculated from the supralingual portion of our model (i.e., without reference to sublingual space effects) to formant frequencies estimated by the VTCALCS program from MRI supralingual dimensions. The supralingual portion of the model includes all the dimensions listed in Table VII, except those related to the augmented front cavity (A_f' and L_f') and the augmented back cavity (A_b' and L_b'), which refer to sublingual dimensions.

A sketch of the simple tube model, modified to reflect only supralingual data, is given in Fig. 8. Speaker-specific

TABLE VII. Dimensions for simple tube model estimated from the MRI-derived area functions in Fig. 4. A_l and L_l refer, respectively, to the area and length of the lip constriction, A_f and L_f refer to the area and the length of the cavity between the lip constriction and the palatal constriction, A_f' and L_f' refer to the area and length computed by combining the sublingual cavity with the cavity between the lip constriction and the palatal constriction, A_{oc} and L_{oc} refer to the area and length of the palatal (oral) constriction, A_m and L_m refer to the area and length of the cavity between the oral constriction and the pharyngeal constriction, A_{pc} and L_{pc} refer to the area and length of the pharyngeal constriction, A_b and L_b refer to the area and length of the cavity posterior to the pharyngeal constriction, and A_b' and L_b' refer to the area and length of the cavity between the palatal constriction and the glottis.

	PK tip up	PK tip down	MI initial	MI syllabic
A_l (cm^2)	0.71	0.67	1.58	1.62
L_l (cm)	0.90	0.90	1.50	1.50
A_f (cm^2)	2.66	2.03	3.84	4.47
L_f (cm)	1.5	1.5	1.5	1.5
A_f' (cm^2)	2.19	1.58	2.57	3.05
L_f' (cm)	2.40	2.40	3.00	2.70
A_{oc} (cm^2)	0.53	0.45	0.78	0.87
L_{oc} (cm)	3.00	3.00	2.70	3.00
A_m (cm^2)	2.18	2.07	3.88	3.70
L_m (cm)	4.20	4.20	4.50	4.20
A_{pc} (cm^2)	0.94	0.62	1.94	1.93
L_{pc} (cm)	2.40	2.40	4.20	4.20
A_b (cm^2)	2.56	2.12	3.53	3.56
L_b (cm)	3.30	3.30	3.60	3.60
A_b' (cm^2)	2.29	1.74	3.12	3.25
L_b' (cm)	9.9	9.9	12.3	12

dimensions of the different tubes were derived by averaging the areas of the sections appropriate to a particular tube, and summing the corresponding section lengths. In the case of PK, areas less than 1 cm^2 were taken as part of a constriction and areas larger than 1.5 cm^2 were taken as part of one of the larger cavities. The inflection point (i.e., the largest first difference) in the transition region was taken as the boundary between the constriction and the larger cavity. In the case of MI, areas less than 1.5 cm^2 in the lip area and in the palatal region were taken as part of the constriction. Areas larger than 2.0 cm^2 were taken as part of a larger cavity. The region behind the palatal constriction was divided into several cavities where the areas greater than 2.6 cm^2 formed the larger mid- and back cavities and areas less than this formed the smaller cavity in the pharyngeal region. Table VII shows the

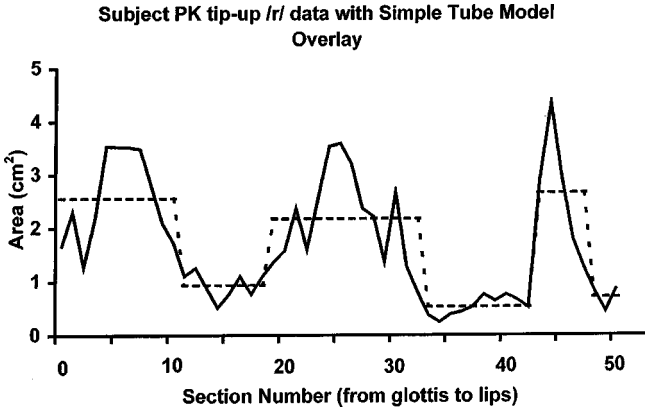

FIG. 8. MRI-derived area function for PK's tip-up bunched /r/ with the corresponding simple tube model from Table VIII superimposed.

TABLE VIII. Formant frequencies estimated from VTCALCS using supralingual simple tube model dimensions as input.

	PK tip up	PK tip down	MI initial	MI syllabic
$F1$	377.1	389.4	349.9	358.3
$F2$	1360.6	1282.5	1207.0	1229.3
$F3$	1892.4	2253.6	2020.1	1923.6
$F4$	3798.9	3905.2	3108.4	3156.8

tube dimensions derived for our model for each speaker. For purposes of comparison, Fig. 8 shows the supralingual simple tube model for PK's tip-up /r/ superimposed on PK's MRI-derived supralingual area function from Fig. 7.

The supralingual vocal-tract dimensions of the simple tube model were put into VTCALCS, and the resulting formant frequencies are given in Table IX. Compare these results with those of Table V, which were obtained via VTCALCS with the actual MRI-derived supralingual dimensions. As can be seen, differences between $F1-F3$ values in Tables V vs IX are no greater than 110 Hz in any case. In fact, for many comparison pairs the correspondence is much closer; for example, the $F1$ values for PK's tip-up /r/ differ by 10.4 Hz while her $F3$ values differ by 46.5 Hz. Given the simplifications inherent in any simple tube model, these result suggest that our supralingual model is equivalent to the actual MRI-derived dimensions, in terms of its accuracy in predicting our speakers' real formant frequencies.

2. F1, F2, and F4 calculation (resonances from behind the palatal constriction)

In this section, we dissect the cavity origins of $F1$, $F2$, and $F4$. Both Stevens (1999) and Alwan *et al.* (1997) assume these formants come from a cavity that either involves the palatal constriction or is posterior to it. Stevens (1999) models $F1$ and $F2$ of a tip-up retroflex /r/ configuration (see Fig. 3) as follows: (1) $F1$ is a Helmholtz resonance formed by the large back cavity (extending from the palatal constriction to the glottis) and the palatal and labial constrictions (see footnote 3), and (2) $F2$ is a half-wavelength resonance of the back cavity behind the palatal constriction. Alwan *et al.* (1997) propose a similar model, with the difference that $F1$ is considered to be a Helmholtz resonance formed by the palatal constriction and the back cavity behind it. In contrast to Stevens' approach, the Alwan *et al.* paper did not take lip constriction into account. They added the observation that $F2$ may be a Helmholtz resonance between the pharyngeal constriction and the cavity posterior to it if the pharyngeal constriction is narrow enough.

In this context, our data are particularly interesting, because the /r/ productions of the two speakers PK and MI show very different degrees of pharyngeal constriction. This can be seen in Table VIII, for instance, by comparing the narrow pharyngeal constriction value (A_{pc}) of 0.94 cm^2 for PK's tip-up /r/ (and similar value for tip-down /r/ against the wider constriction of 1.94 cm^2 for MI's word-initial /r/ (and similar value for syllabic /r/). As Alwan *et al.* (1987) note, if the pharyngeal constriction is narrow enough, separate large resonating cavities are formed in front and behind. This may be the case for speaker PK. On the other hand, if the pha-

TABLE IX. $F1$ calculated as a Helmholtz resonance and $F2$ calculated as a half-wavelength resonance.

	PK tip up	PK tip down	MI initial	MI syllabic
$F1$	418	453	406.5	406.9
$F2$	1767	1767	1422	1458

ryngeal constriction is wide enough, the area behind the palatal constriction may act as a single resonator. We call the latter variant of the simple tube model the long *back cavity* model, and the former variant the *additional cavity* model. To determine which model best predicts the different subjects' data, we ran separate calculations for these two variants of the simple tube model, for each speaker and speaking condition.

In the *long back cavity* variant of the simple tube model, $F1$ is a Helmholtz resonance formed by the palatal constriction and a back cavity posterior to it (extending from the palatal constriction to the glottis, with dimensions A_b' and L_b' in Table VII). Accordingly, $F1$ is calculated by summing the admittances of the back cavity $[-j(A_b'/\rho c)\tan \omega L_b'/c]$ and the palatal constriction $[-j(A_c/\rho c)\cot \omega L_c/c]$ and finding the frequency of the first zero crossing. $F2$ is calculated as a half-wavelength resonance of the back cavity $(c/2l_b')$. $F4$ is calculated as the second half-wavelength resonance of the back cavity $(2{*}c/2l_b')$. The frequencies of $F1$, $F2$, and $F4$ resulting from this method are given in Tables IX and XI. Note that in this simple tube calculation, as in Stevens' model, the pharyngeal constriction is not accounted for. Our data show, however, that it occurs at a point along the tube where a perturbation would be expected to lower the formants by 100–200 Hz.

In the *additional cavity* method, we model the back part of the vocal tract with a separate tube for the pharyngeal constriction. In this case, the palatal constriction and the cavities posterior to it form a double-Helmholtz (coupled) resonator. To estimate formant frequencies from this complicated configuration, we had three options for simplifying the calculation of $F1$ and $F2$ values at this point; (1) by assuming decoupling in a distributed system, (2) by assuming decoupling but using a lumped approximation, and (3) by using a double-Helmholz lumped equation. In option (1), we decoupled the double-Helmholtz resonator into two single-Helmholtz resonators. $F1$ was calculated by summing the admittances of the midcavity $[j(A_m/\rho c)\tan \omega L_m/c]$ and the palatal constriction $[-j(A_c/\rho c)\cot \omega L_c/c]$ and finding the frequency of the first zero crossing. Similarly, $F2$ was calculated by summing the admittances of the back cavity $[j(A_b/\rho c)\tan \omega L_b/c]$ and the pharyngeal constriction $[-j(A_{pc}/\rho c)\cot \omega L_{pc}/c]$ and finding the frequency of the first zero crossing. For option (2) we again assumed decoupled Helmholtz resonators but used lumped approximation to obtain values for $F1$ $[c/2\pi(A_{oc}/L_{oc}A_mL_m)^{1/2}]$ and $F2$ $[c/2\pi(A_{pc}/L_{pc}A_bL_b)^{1/2}]$. Finally, we computed the frequencies for $F1$ and $F2$ by using Fant's double-Helmholz lumped equation (Fant, 1960) to account for the interconnection between the two Helmholtz cavities. These values of $F1$ and $F2$ are given in Table X. Note that one consequence of using the lumped-coupled model is that, relative to the lumped-decoupled values, $F1$ is shifted down by the proportion that $F2$ is shifted up. To calculate $F4$, we decoupled all of the tubes making up the double-Helmholtz resonator. $F4$ was then calculated as the half-wavelength resonance of the longest cavity. For all the subjects and conditions, this meant that $F4$ was the half-wavelength resonance of the midcavity $(c/21_m)$. (Note that in the case of MI's syllabic /r/, the lengths of the tube formed by the pharyngeal constriction and the midcavity are the same.) Again, the $F4$ frequencies by this method are given in Table XI.

TABLE X. $F1$ and $F2$ calculated from a double-Helmholtz model. The decoupled values of $F1$ and $F2$ are computed by disconnecting the two Helmholtz resonators and using the appropriate admittance functions. The decoupled and lumped values are computed again with the Helmholtz resonators disconnected, and with the lumped approximation used. The coupled and lumped values are calculated from the double-Helmholtz lumped formula (Fant, 1960, p. 286).

	PK tip up	PK tip down	MI initial	MI syllabic
$F1$ (decoupled)	715.8	682.1	668.0	705.6
$F1$ (decoupled–lumped)	774.2	732.1	716.9	761.3
$F1$ (coupled–lumped)	528.3	511.9	513.2	525.5
$F2$ (decoupled)	1072.4	976.9	912.2	907.6
$F2$ (decoupled–lumped)	1200.0	1071.0	1062.5	1055.3
$F2$ (coupled–lumped)	1756.6	1530.1	1482.8	1527.4

Whether the *long back cavity* or the *additional cavity* variant of the model works best can be seen for each speaker by comparing the $F1$, $F2$, and $F4$ values in Tables IX and XI. Remember that speaker MI had relatively wide pharyngeal constrictions, while speaker PK had relatively narrow pharyngeal constrictions. A comparison of the values in Tables IX and X shows that modeling the back part of the vocal tract as a single long cavity gives reasonable values for MI's /r/ productions. For instance, the values for MI's $F2$ in Table IX are 1422 and 1458 Hz.[7] These frequencies compare well with those from the real data (see Table II) which range from 989–1556 Hz. However, as Table X shows, the additional cavity model also produces reasonable $F2$ values, from 912.2 and 907.6 Hz for the decoupled calculation to 1482.8 and 1527.4 Hz for the coupled-lumped calculation. For PK, on the other hand, modeling the pharyngeal constriction as an additional cavity produces better $F2$ values. For instance, in the case of PK's (tip-down) bunched /r/, modeling the back part of the vocal tract as a double-Helmholtz resonator results in an $F2$ frequency of 1530 Hz (see Table X, coupled-lumped option). This is much lower than the 1767-Hz value obtained when the *long back cavity model* is used (see Table IX). It should be noted that calculations us-

TABLE XI. $F4$ calculated as the second resonance of a half-wavelength tube model for the back cavity (between the palatal constriction and the glottis) and the lowest resonance of the midcavity (between the palatal and pharyngeal constrictions) from the double-Helmholtz model.

	PK tip up	PK tip down	MI initial	MI syllabic
$F4$ (half-wavelength tube model)	3535.4	3535.4	2845.5	2916.7
$F4$ (double-Helmholtz model)	4166.7	4166.7	3888.9	4166.7

ing a lumped approximation tend to overshoot the true value by 100–200 Hz. A demonstration of this general tendency can be seen in Table X by comparing the lumped and distributed estimates of $F2$ for a decoupled system. Thus, we expect that the true value of $F2$ (relative to the coupled-lumped method) for PK's tip-down /r/ is closer to 1350 Hz, while the likely true value for her tip-up /r/ is closer to 1550 Hz. Similarly, we expect MI's true $F2$ value to be in the region of 1250 Hz. The difference between predictions of the *additional cavity* simple tube model for PK's tip-up and tip-down /r/ are likely due to the larger area ratios (A_m/A_{pc} and A_b/A_{pc}) between cavities (see Table VII).

Looking at the $F4$ calculations, we see that for PK the values of 4166.7 Hz obtained from the additional cavity model (see Table XI) are within range of the 4110.8-Hz average $F4$ frequencies observed in her real data (see Table II). At the same time, the $F4$ frequencies predicted by the *long back cavity* model, at 3535.4 Hz, are too low. On the other hand, the $F4$ values calculated by the *additional cavity* model for MI are too high at 3888.9 and 4166.7 Hz, compared to the average $F4$ value of 3113.7 Hz in his real data (see Table II). In contrast, the values calculated by the *long back cavity* model, at 2845.5 and 2916.7 Hz are appropriately near the real data average of 3113.7 Hz.[8]

3. F3 calculation

Stevens (1999) and Alwan *et al.* (1997) assume that $F3$ results from the front cavity represented by a large volume, including the sublingual cavity, and by a narrowed lip opening. Both assume that for tip-up /r/'s the volume of the front cavity may not be large enough to account for $F3$ unless the sublingual space is included. For tip-down /r/'s, both assume the front cavity is large enough to account for $F3$ as a front cavity resonance. Because the contribution of the sublingual space to the volume of the front cavity is reduced as the tongue tip moves down, Alwan *et al.* note the possibility of a trading relation between sublingual space volume (for tip-up /r/'s) and a relatively more posterior palatal constriction for tip-down /r/'s. In other words, front cavity volume, and consequent low $F3$, may be maintained by either increasing the sublingual space or by moving the palatal constriction back and thereby increasing the length of the front cavity. Although neither paper specifically suggests a mechanism whereby the increase in volume would lower $F3$, there are two possibilities: (1) that the sublingual space acts to increase the length of the front cavity, thereby lowering its full set of resonant frequencies, or (2) that the sublingual space, by increasing the front cavity volume-to-lip constriction ratio, contributes to the formation of a front cavity with the shape of a Helmholtz resonator.

To calculate $F3$, we assume that the front cavity can be decoupled from the palatal constriction. Further, the results of experiment 3 suggest that the sublingual space acts to increase the volume of the front cavity and thereby lower the frequency of $F3$. Thus, to calculate $F3$, we combined the front cavity and the sublingual space into one uniform tube. The lengths of the oral cavity and the sublingual cavity were added and the average of the areas involved was taken as an estimate of the uniform area. $F3$ was calculated by summing the admittance of the combined oral and sublingual cavity $[-j(A'_f/\rho c)\tan \omega L'_f/c]$ and the admittance of the lip constriction $[-j(A_l/\rho c)\cot \omega L_l/c]$ and determining the first zero crossing. This value for $F3$ is given in Table XII. Two further effects cancel each other out: if we take into account the radiation impedance of the mouth opening, then the value of $F3$ is lowered by about 200 Hz, while if we take into account the effect of the acoustic mass that results from the palatal constriction, $F3$ is raised by about 200 Hz (Guenther *et al.*, 1999).

TABLE XII. $F3$ calculated from the cavity anterior to the palatal constriction including the sublingual space.

	PK tip up	PK tip down	MI initial	MI syllabic
$F3$	1874.3	2064.5	1675.9	1698.7

III. DISCUSSION

In this paper, we consider several issues important to the development of an acoustic model for American English /r/. First, we show that the Perturbation Theory does not make appropriate predictions of constriction location for /r/'s of the type produced by our speakers. It is likely that this difficulty extends to /r/'s of similar articulatory configuration reported in Guenther *et al.* (1999), Westbury *et al.* (1999), Zawadaski and Kuehn (1980), Lindau (1985), Kent (1998), Delattre and Freeman (1968), and additional speakers reported in Alwan *et al.* (1997), among others.

In experiments 1, 2, and 3, we use MRI-derived dimensions and the Maeda computer simulation program to present evidence from several sources that $F3$ is a front cavity resonance. In experiment 1, we show that the effects of eliminating the pharyngeal constriction on $F3$ are minimal. In experiments 2 and 3, we further demonstrate that in fact the addition of the sublingual space is crucial for achieving $F3$ values that match the full range of speakers' values. By comparing the results of experiments 3 and 4, we further show that the sublingual space acts to extend the front cavity. Furthermore, for articulatory configurations with limited sublingual area, such as the ones reported in this study, the branch cavity antiresonance is well above the region of $F3$ (>5000 Hz) so that it is not a factor in predicted $F3$ for /r/. [Note that, in a tip-up retroflex /r/ where the tongue dorsum is lowered (Stevens, 1999; Narayanan *et al.*, 1999), the sublingual cavity may be considerably longer so that the antiresonance may occur in a region close to $F3$ or $F4$.] In experiment 5, we developed a simple tube model based on the MRI-derived dimensions, and show that $F3$ can be derived by positing (1) a lip constriction formed by the tapering gradient of the teeth and lips (with or without rounding), and (2) a large-volume cavity behind it and anterior to the palatal constriction. The larger volume in turn results from a combination of the front cavity proper together with the volume of the sublingual space. Narayanan *et al.* (1999) discuss a similar role for the front cavity in Tamil retroflex liquids. For these retroflex liquids, however, a tongue tip-up articulatory configuration with less-pronounced lip constriction and a

long sublingual space allowed application of Stevens' model (1999) of the front cavity as a quarter-wavelength resonator.

An interesting aspect of the difference between the Stevens (1999) model and our own model is the contrast between the role of the lip constriction and the sublingual space. If the lip constriction and tapering produce a front cavity of a Helmholtz shape, adjustments to lower $F3$ must have the effect of adding to the volume of the cavity. As noted above, this effect may be produced either by the addition of a sublingual space, or by more posterior placement of the palatal constriction. We may imagine here a trading relation between the placement of the palatal constriction and the existence of a sublingual space, such that each contributes to increasing the volume of the cavity. On the other hand, if the lip constriction is wide enough that the front cavity is best modeled as a quarter-wavelength resonator, the dimension that matters most for lowering $F3$ is an increase in the length of the front cavity. Again, we may imagine a trading relation between increases to the length of the front cavity and the narrowing of the lip constriction, in essence, a trading relation between a Helmholtz-type and quarter-wavelength model of the front cavity acoustics. These putative trading relations are slightly different, but complementary to, the trading relations between measures of palatal constriction location, degree, and length discussed in Guenther *et al.* (1999). In addition, the formation of a separate lip protrusion channel has a lowering effect on $F3$. Because the MRI-derived configurations of speakers PK and MI describe bunched configurations, it might be tempting to assume that the Helmholtz shape is typical of bunched configurations, while the quarter-wavelength shape characterizes retroflex configurations such as those in Narayanan *et al.* (1999). Please note, however, that lip constriction dimensions are not available for the retroflex /r/ of Ong and Stone (1998), or for any other American English tip-up retroflex /r/. [The dimensions in Fant (1960) are based on a Russian trilled /r/.] Thus, although we can posit a general trading relation, we cannot relate it to the classical contrast between retroflex and bunched /r/'s in any detail.

In this paper, we have discussed the sublingual space only in the context of how it helps to lower $F3$. However, the sublingual space also functions as a side branch (that can itself be modeled as a quarter-wavelength tube), introducing an antiresonance in the spectrum of /r/. For the articulatory configurations considered in this paper, the side branch is no more than 1.5 cm long, producing an antiresonance between 5–6 kHz. Note that configurations exist where the sublingual space is considerably longer, as in the Tamil /r/ described by Narayanan *et al.* (1999) and the retroflex /r/ assumed by Stevens (1999). In such cases, the antiresonance is between 3–4 kHz. An antiresonance in this region of the spectrum may have a considerable effect on the distinctive acoustic profile of /r/.

Another interesting result of the modeling involves the role of the pharyngeal constriction. In the case of PK, the pharyngeal constriction is narrow enough so that the back part of the vocal tract behind the palatal constriction is best modeled as a double-Helmholtz resonator. On the other hand, for MI, the pharyngeal constriction is wider so that this portion of the vocal tract is best modeled as a half-wavelength resonator. These different models result in very different values of $F4$ for PK and MI. In both cases, $F4$ is a half-wavelength resonance. However, in the case of PK, the length of the cavity is only about 4 cm long, resulting in a frequency for $F4$ around 4200 Hz. On the other hand, for MI, the length of the half-wavelength resonator is around 12 cm, resulting in an $F4$ frequency around 2900 Hz.

IV. CONCLUSION

Our primary aim in this paper was to examine and modify existing acoustic models of American English /r/, using recently obtained MRI-derived vocal-tract dimensions. A second aim was to model the full range of formant frequencies, and in particular $F3$, produced by our speakers. We first considered the Perturbation Theory account of /r/ acoustics, and found that the placement of constrictions in the MRI-derived data did not match predicted locations. We then proceeded by using MRI-derived vocal-tract dimensions as input to the Maeda VTCALCS program. Comparing these results to the range of actual formant frequencies produced by our speakers, we found that even in the case of tip-down bunched /r/, the addition of the sublingual space to the front cavity was necessary to achieve $F3$ frequencies that match speakers' mid- and low-range $F3$ values. We then developed a simple tube model whose output formant frequencies match those derived via the Maeda computer simulation model and actual MRI-derived dimensions. As such, it adequately accounts for the cavity affiliations and frequencies of $F1$, $F2$, $F3$, and $F4$. In this model, $F3$ is a front cavity resonance where the front cavity includes a lip constriction formed by the tapering gradient of the teeth and lips (with or without rounding) and a large volume cavity behind it that includes the sublingual space. The sublingual space acts to increase the volume of the cavity and thereby lower $F3$. The results also suggest that $F1$, $F2$, and $F4$ arise from the palatal constriction and/or the cavities posterior to it.

Although to this point we have largely succeeded in modeling the mid-to-low range of $F3$, we have not yet fully accounted for the lowest $F3$ values. This may stem from the fact that the MRI-derived dimensions and the real words were recorded at different time points; in other words, during the MRI experiment the speakers may actually have been producing /r/'s with $F3$'s in their own midrange. Also, the nature of the MRI technique, and the orientation of the magnetic coil, means that peripheral structures such as lips and glottis are imaged with less accuracy. The error in imaging the lip area may be as much as 1 cm; an increase or decrease of this magnitude in lip protrusion or narrowing would have a major effect on $F3$. In addition, other acoustic mechanisms that we have not considered may be at work. Recent work, Story, Titze, and Hoffman (1998) and Dang and Honda (1997), for instance, document the role of the piriform sinuses in lowering formant frequencies. We hope to explore these issues further in future research.

ACKNOWLEDGMENTS

We thank Ken Stevens for his help and comments, Shinji Maeda for the VTCALCS program, and Ronan Scaife for the MATLAB version of VTCALCS. This work was supported by an NIDCD Grant No. IR03-C2576-01 to Suzanne Boyce and NSF Grant No. IRI-9310518 and NIH Grant No. 1 K02 DC00149-01A1 to Carol Espy-Wilson. We also thank Maureen Stone for furnishing her MIT data.

APPENDIX: LIST OF REAL WORDS SPOKEN BY PK AND MI

(1) beaker;
(2) beeper;
(3) beater;
(4) kirk;
(5) perk;
(6) reed;
(7) rod;
(8) rude;
(9) turd;
(10) turk (MI only).

[1]Phoneticians typically classify American English /r/ into a syllabic, or vocalic, allophone and a consonantal allophone /r/. Because modeling issues are the same for all such allophones, we have chosen to treat /r/ in this paper as a single entity. Thus, we have maximized the challenge to our modeling results by including data from all allophonic types of /r/ in our discussions of formant values typical of /r/, including discussion of ranges of formant values.

[2]The currency of the perturbation model for /r/ among teachers of speech acoustics has been confirmed in conversations with, among others, John Ohala, Pat Keating, Louis Goldstein, Mary Beckman, Keith Johnson, and Peter Ladefoged. It is the basis of the discussion of American /r/ acoustics in Johnson (1997).

[3]Conditions under which a constriction is narrow enough that cavities on either side are best modeled as separate resonators, rather than as coupled areas separated by a perturbation, have been a matter of empirical investigation. In general, Mrayati *et al.* (1988) estimate that the cross-over point for ''one-tract-mode'' (OTM) resonance vs ''two-tract-mode'' (TTM) resonance is a 0.8 cm^2 constriction in a 5-cm^2 tube, or a 0.65-cm^2 constriction in a 4-cm^2 tube—essentially a 6-to-1 ratio. On this logic, a vocal tract with less extreme constrictions might plausibly be modeled as resonating in OTM (perturbation) mode. In the case of /r/, modeling trials have established that when tapering palatal, pharyngeal, and lip constrictions of 0.8 cm^2 are inserted in a tube of uniform 4-cm^2 area and 15.3-cm length (essentially a 5-to-1 ratio), $F1$ is at 480 Hz, $F2$ at 1274 Hz, and $F3$ at 1728 Hz. These numbers are securely in the range of attested $F3$ values for /r/. Thus, an idealized vocal tract configured as predicted by the Perturbation Theory does in fact produce acceptable formant values for /r/. [It should be noted that $F3$ decreases further as constriction size decreases at these points. This may be an instance of what Mrayati *et al.* (1988) term ''observations that... the relation predicted by 'small' perturbation theory continues to hold even if 'large' variations are introduced.''] However, the actual constrictions measured from the MRI data for our subjects in Fig. 4 are typically much smaller—with palatal constrictions typically less than 0.65 cm^2 and as low as 0.25 cm^2. These data on constriction size suggest that Perturbation Theory may not be the appropriate model for /r/ as produced by our subjects.

[4]At low frequencies, the impedance of the palatal constriction and the impedance of the lip constriction are in series, since the impedance of the front cavity (which consists of an acoustic compliance) can be ignored. For more details, see Stevens (1999).

[5]In Alwan *et al.* (1997), PK's vocal-tract profiles are labeled according to production condition, i.e., as ''retroflex'' and ''bunched.''

[6]The sublingual areas used in this study differ from those reported in Alwan *et al.* (1997). The sublingual areas given in Table II of Alwan *et al.* (1997) were measured directly from raw coronal images of the oral cavity and included only the airspace areas that appeared directly below the tongue surface. The sublingual areas reported in this paper are based on an improved and more realistic specification of the sublingual cavity: (1) The contiguous raw coronal image scans were used to obtain a 3D reconstruction of the oral cavity that included the sublingual cavity and any airspace along the sides of the tongue. (2) Cross sections of the supra- and sublingual cavities were then obtained from resectioning these cavities along planes perpendicular to their midlines. Specifically, note that the ''improved'' sublingual areas also included some contributions from air space along the sides of the front tongue.

[7]As noted above, if we model the pharyngeal narrowing as a perturbation in MI's *long back cavity* model, its position is such that it would lower the frequencies of $F2$ (1422 and 1458 Hz) by 100–200 Hz.

[8]Again, if we model the pharyngeal narrowing as a perturbation in MI's long back cavity model, we would expect it to raise $F4$ by 100–200 Hz, bringing the values of 2845.5 and 2916.7 Hz even closer to the real data average of 3113.7.

Alwan, A., Narayanan, 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.

Boe, L., and Perrier, P. (**1990**). ''Comments on Distinctive regions and modes: A new theory of speech production,'' Speech Commun. **7**, 217–230.

Boyce, S., and Espy-Wilson, C. (**1997**). ''Coarticulatory stability in American English /r/,'' J. Acoust. Soc. Am. **101**, 3741–3753.

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

Dang, J., 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. (**1968**). ''A dialect study of American Rs by x-ray motion picture,'' Language **44**, 29–68.

Espy-Wilson, C. (**1992**). ''Acoustic measures for linguistic features distinguishing the semivowels in American English,'' J. Acoust. Soc. Am. **92**, 736–757.

Espy-Wilson C. Y., Narayanan, S., Boyce, S. E., and Alwan, A. (**1997**). ''Acoustical modeling of American English /r/,'' Proceedings of Eurospeech '97, September, Patras, Greece.

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

Fant, G. (**1980**). ''The relations between area functions and the acoustic signal,'' Phonetica **37**, 55–86.

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 Phonetics Laboratory Working Papers, 90.

Harshman, R., Ladefoged, P., and Goldstein, L. (**1977**). ''Factor analysis of tongue shapes,'' J. Acoust. Soc. Am. **62**, 693–707.

Heinz, J. M. (**1967**). ''Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues,'' STL-QPSR 1/1967, pp. 1–14.

Jackson, M., Espy-Wilson, C., and Boyce, S. (**1999**). ''Verifying a vocal tract model with a closed side branch,'' J. Acoust. Soc. Am. (submitted).

Johnson, K. (**1997**). *Acoustic and Auditory Phonetics* (Blackwell, Cambridge).

Kent, R. (**1998**). ''Normal aspects of articulation,'' in *Articulation and Phonological Disorders*, edited by J. Bernthal and N. Bankson (Allyn and Bacon, Boston).

Lehiste, I. (**1962**). ''Acoustical characteristics of selected English consonants,'' University of Michigan Communication Sciences Laboratory Report #9.

Lindau, M. (**1985**). ''The Story of /r/,'' in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic, Orlando), pp. 157–168.

Maeda, S. (**1982**). ''Digital simulation method of the vocal tract system,'' Speech Commun. **1**, 199–299.

McGowan, R. S. (**1992**). ''Tongue-tip trills and vocal-tract wall compliance,'' J. Acoust. Soc. Am. **91**, 2903–2910.

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

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., Alwan, 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., Byrd, D., and Kaun, A. (**1999**). ''Geometry, kinematics, and acoustics of Tamil liquid consonants,'' J. Acoust. Soc. Am. **106**, 1993–2007.
Nolan, F. (**1983**). *The Phonetic Bases of Speaker Recognition* (Cambridge University Press, Cambridge, England).
Ohala, J. (**1985**). ''Around flat,'' in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic, Orlando), pp. 223–241.
Ong, D., and Stone, M. (**1998**). ''Reconstruction of vocal tract shape from magnetic resonance images during production of [r] and [l],'' Phonoscope **1**, 1–14.
Rubin, P., Baer, T., and Mermelstein, P. (**1981**). ''An articulatory synthesizer for perceptual research,'' J. Acoust. Soc. Am. **70**, 321–328.
Scaife, R. (**1997**). Personal communication.
Schroeder, M. R. (**1967**). ''Determination of the geometry of the human vocal tract by acoustic measurements,'' J. Acoust. Soc. Am. **41**, 1002–1010.
Shiller, D., Ostry, D., and Gribble, P. (**1999**). ''Effects of gravitational load on jaw movement in speech,'' J. Neurosci. **19**, 9073–9080.
Shriberg, L., and Kent, R. (**1982**). *Clinical Phonetics* (Macmillan, New York).
Stevens, K. N. (**1999**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).
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.
Sundberg, J., Lindblom, B., and Liljencrants, J. (**1992**). ''Formant frequency estimates for abruptly changing area functions: A comparison between calculations and measurements,'' J. Acoust. Soc. Am. **91**, 3478–3482.
Tiede, M. (**1999**). Personal communication.
Tiede, M., Masaki, S., Wakumoto, W., and Vatikiotis-Bateson, E. (**1997**). ''Magnetometer observation of articulation in sitting and supine conditions,'' J. Acoust. Soc. Am. **102**, 3166.
Veatch, T. C. (**1991**). ''English vowels: Their surface phonology and phonetic implementation in vernacular dialects,'' University of Pennsylvania Ph.D. dissertation.
Westbury, J. R., Hashi, M., and Lindstrom, M. J. (**1999**). ''Differences among speakers in lingual articulation of American English /r/,'' Speech Commun. **26**, 203–226.
Yang, C. S., and Kasuya, H. (**1994**). ''Accurate measurement of vocal tract shapes from magnetic resonance images of child, female, and male subjects,'' in Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, pp. 623–626 (unpublished).
Zawadaski, P., and Kuehn, D. (**1980**). ''A cineradiographic study of static and dynamic aspects of American English /r/,'' Phonetica **37**, 253–266.