A modeling investigation of articulatory variability and acoustic stability during American English /r/ production

Alfonso Nieto-Castanon[a) and Frank H. Guenther
*Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street,
Boston, Massachusetts 02215*

Joseph S. Perkell
*Research Laboratory of Electronics, Massachusetts Institute of Technology, 77 Massachusetts Avenue,
Room 36-413, Cambridge, Massachusetts 02139-4307*

Hugh D. Curtin
*Department of Radiology, Massachusetts Eye and Ear Infirmary, 243 Charles Street,
Boston, Massachusetts 02114*

This paper investigates the functional relationship between articulatory variability and stability of acoustic cues during American English /r/ production. The analysis of articulatory movement data on seven subjects shows that the extent of intrasubject articulatory variability along any given articulatory direction is strongly and inversely related to a measure of acoustic stability (the extent of acoustic variation that displacing the articulators in this direction would produce). The presence and direction of this relationship is consistent with a speech motor control mechanism that uses a third formant frequency ($F3$) target; i.e., the final articulatory variability is lower for those articulatory directions most relevant to determining the $F3$ value. In contrast, no consistent relationship across speakers and phonetic contexts was found between hypothesized vocal-tract target variables and articulatory variability. Furthermore, simulations of two speakers' productions using the DIVA model of speech production, in conjunction with a novel speaker-specific vocal-tract model derived from magnetic resonance imaging data, mimic the observed range of articulatory gestures for each subject, while exhibiting the same articulatory/acoustic relations as those observed experimentally. Overall these results provide evidence for a common control scheme that utilizes an acoustic, rather than articulatory, target specification for American English /r/. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1893271]

I. INTRODUCTION

When producing a given phoneme, speakers use a set of articulators (e.g., tongue, jaw, lips) to affect the vocal-tract shape and, ultimately, the characteristics of the resulting acoustic signal. The vocal-tract configuration for the production of a given phoneme is not uniquely defined by phoneme identity. Different speakers will use different articulatory configurations when producing the same phoneme, and often the same speaker will use a range of different articulatory configurations when producing the same phoneme in different contexts. In particular, the American English phoneme /r/ has been associated with a large amount of articulatory variability (Delattre and Freeman, 1968; Westbury *et al.*, 1998; Guenther *et al.*, 1999). While large, the degree of articulatory variability present in natural speech does not seem to hinder phoneme recognition by listeners, and it is often conceptualized as an expression of control mechanisms that make efficient use of a redundant articulatory system. Such efficient use of redundancy in biological motor systems is often referred to as *motor equivalence*.

Current speech movement control theories dealing with the motor equivalence problem can be roughly classified de-

pending on the type of phonemic targets hypothesized (see MacNeilage, 1970, for motivations of a target-based approach to speech motor control theories). The task-dynamic model of Saltzman and Munhall (1989) exemplifies a type of computational model in which phonemic targets are characterized in terms of *tract variables* representing specific aspects of the vocal-tract shape that can be independently controlled by the speech control mechanism (e.g., lip aperture, tongue dorsum constriction location, etc.) In this model, articulatory variability can arise as a consequence of "blending" effects from the context phonemes. For example, when producing a /b/ in a VCV context, a full bilabial closure represents the targeted tract variable. Other aspects of the vocal-tract not affecting the targeted tract variable, such as tongue shape, will vary depending on the shape adopted in the production of the leading vowel, while also being subject to anticipatory movements towards the following vowel configuration. In this way, articulatory variability in different phonetic contexts would reflect the interplay between constraints imposed by current and contextual phonemic targets.

The DIVA model (e.g., Guenther *et al.*, 1998; 2003) exemplifies a second type of computational model of speech motor control in which the phonemic targets are characterized in terms of *acoustic/auditory variables*[1] (for example,

———————————
a)Electronic mail: alfnie@bu.edu

formant frequency descriptors). In this model, the control mechanism moves the articulators in the direction that would bring the formants of the resulting auditory signal closest to the targeted formants, without reference to an explicit vocal-tract shape target. Articulatory variability then arises naturally as a consequence of the many-to-one mapping between the articulatory configurations and the audible acoustic characteristics of the produced sound. In other words, for these models articulatory variability reflects the variety of articulatory configurations that would produce the desired acoustic properties.

Often (e.g., Saltzman and Munhall, 1989; Guenther *et al.*, 1998) the distinction is emphasized between the articulatory configurations (the state of articulatory variables, such as jaw aperture) and the resulting vocal-tract shapes (the state of tract variables, such as tongue dorsum constriction degree). This highlights the redundancy of the speech articulatory system. For example, a particular tongue dorsum constriction degree can be achieved with a relatively low jaw height and a relatively high tongue-body height (relative to the jaw), or a higher jaw height and lower tongue-body height can be used to achieve the same constriction degree. More generally, both articulatory and tract variables represent different coordinate frames that can be used to represent the state of the vocal-tract apparatus (see MacNeilage, 1970, for an introduction to the concept of coordinate systems in speech production). Tract variables represent a more abstract coordinate frame than articulatory variables, since there is a one-to-many relation between tract variables and articulatory variables defined by the geometrical relations among them. In the same way, acoustic or auditory variables (Guenther *et al.*, 1998) can be simply thought of as yet another coordinate frame for the representation of the articulatory state. They also represent a more abstract coordinate frame than articulatory variables, in that there is a one-to-many relation between auditory and articulatory variables. The analysis of variability in articulatory configurations in the production of a given phoneme, similar to the analysis of errors in a pointing task (Carozzo *et al.*, 1999; McIntyre *et al.*, 2000), is a useful approach for uncovering an appropriate coordinate frame for the representation of targets in speech production. We thus believe that the analysis of articulatory variability should serve to direct the definition of motor control models of speech production. Based on this view, the goal of the current paper is twofold: (1) to characterize, in a paradigmatic example of articulatory variability (American English /r/), the extent of articulatory variability in relation to hypothesized target representations (relevant tract and acoustic variables); and (2) to test whether a model of speech motor control based on an acoustic target definition, together with a speaker-specific vocal-tract model, can explain the specificities of the observed articulatory variability in individual speakers. To these ends, we first present new, model-based analyses of electromagnetic midsagittal articulometer (EMMA) data on seven subjects from a previous study (Guenther *et al.*, 1999). These analyses characterize the experimentally observed articulatory variability in relation to hypothesized target variables. We then provide simulation results of an auditory target model controlling the move-
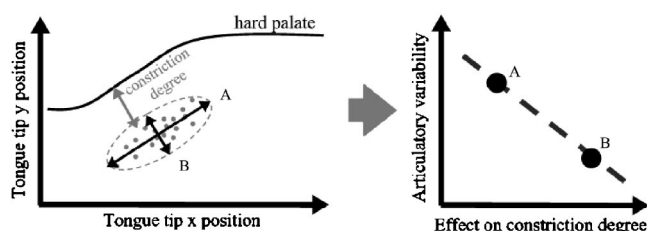

FIG. 1. Schematic example of articulatory variability analysis for a single articulatory measure of interest (tongue tip position). Left: Hypothetical configuration of tongue-tip positions in the production of a phoneme that could be characterized by a tongue tip constriction degree phonemic target. A and B represent the directions of the tongue tip movement resulting from a principal component analysis (PCA) of the tongue-tip articulatory covariance of multiple repetitions. The gray arrow represents the direction of the tongue-tip movement affecting the constriction degree the most. Right: Plot relating the extent of articulatory variability along each of the articulatory directions (A and B) versus the effect that each of these directions has on the hypothetical target variable (constriction degree). The actual analyses performed in this section attempt to provide evidence for several theoretically motivated phonemic target definitions by extending this simple scheme to the case of multiple articulatory measures of interest (indicated by six transducer positions located on the tongue, lips, and jaw of the speakers; see the text for details).

ments of speaker-specific vocal-tract models based on magnetic resonance imaging (MRI) scans of the vocal tracts of two of the seven experimental subjects. The model movements are then compared to those of the modeled speakers. Note that the present study addresses only the production of American English /r/. Several aspects of this paper's methodology (to be described later) are specific to the class of vowel and semivowel productions. The extent to which the presented results generalize to the production of other phoneme classes (in particular, consonants) can only be addressed by further studies.

A. Variability analysis rationale

Previous analyses (Guenther *et al.* 1999) showed that articulatory trade-offs during /r/ production act to reduce $F3$ variability. In this paper we attempt to assess this kind of finding in the context of different speech motor control theories by testing the ability of theoretically motivated phonemic target variables to predict the observed variability in articulatory configurations. Our rationale is exemplified in Fig. 1. Let us only consider the movement of the tongue tip in this example. Imagine, during the production of a hypothetical phoneme, the phonemic target consists of accomplishing a given tongue-tip constriction degree (distance between the tongue tip and the hard palate). The expected array of final configurations of the tongue tip for the production of this phoneme would be expected to take the approximate form shown in Fig. 1 left. The axes labeled A and B represent the directions of articulatory movement resulting from a principal component analysis (PCA) of the final articulatory covariance[2] of a number of productions of the phoneme, and the gray arrow characterizes the direction of articulatory movement affecting the degree of the tongue-tip constriction the most. The right side of Fig. 1 plots for each articulatory direction (A and B) their effect on the hypothesized target variable (*effect on constriction degree*) on the x axis, and their extent of *articulatory variability* on the y axis. This plot

J. Acoust. Soc. Am., Vol. 117, No. 5, May 2005

Nieto-Castanon *et al.*: A modeling investigation of articulatory 3197

schematizes the observation that those articulatory dimensions affecting the target variable the most (B, in this case) would be expected to show a lesser extent of articulatory variability than those dimensions affecting the target variable the least (A). The EMMA analyses in this paper (Sec. II A) extend the simple scheme in this example (with only one transducer reflecting the tongue-tip position) to the case of multiple transducers (six transducers, reflecting tongue, jaw, and lips configurations). The simultaneous analysis of multiple transducers on different articulators allows the articulatory dimensions (12 for each subject) that result from a PCA to characterize complex movements of one or several articulators, such as those described in the literature as trading relations between and within articulators (for example a simultaneous raising of the tongue back and decrease of lip rounding, as in Perkell *et al.*, 1995; or a simultaneous raising of the tongue tip and lowering of the tongue back as in Guenther *et al.*, 1999). As in the example shown here, a functional relationship between the extent of articulatory variability along each of the resulting articulatory dimensions and their associated effect on a hypothesized target variable is taken as indicative of the use of a specific target scheme in the articulatory movement data being analyzed.

In the current article we report the results of analyses of this type performed on the data from each speaker. Subsequent pooling of these results across different speakers allows us to determine whether commonalities exist in the target specification for /r/ across speakers. While we acknowledge that the control strategy for the production of /r/ could be different for different speakers, and the literature has historically emphasized these differences across speakers and phonetic contexts in the articulatory specification of /r/ (e.g., Delattre and Freeman, 1968), our results indicate that commonalities can in fact be found when using an appropriate frame of reference. In particular, we demonstrate that when the articulatory frame of reference is aligned to correspond with important acoustic features, commonalities in the target specification for /r/ are apparent again. These commonalities indicate that a simple control scheme, common across speakers, that utilizes an acoustic production target for /r/ can provide a straightforward and parsimonious explanation for the articulatory variability within and between speakers, whereas control schemes utilizing a common constriction target for /r/ cannot account for the results. To that end the analyses will test both acoustic and tract variables as hypothetical target variables using the methodology outlined above. Note that from these analyses we investigate the possibility of acoustic or tract variables *forming part* of the global target specification for /r/, not whether they fully define it. More complex analyses would be needed to test the possibility of multiple target variables fully defining the target specification for /r/.

B. Modeling and simulations rationale

The analysis of articulatory variability outlined above attempts to identify the nature of the phonemic target for /r/. The results will reveal that there is a great deal more evidence indicative of acoustically defined phonemic targets (in particular one based on $F3$), rather than targets based on

vocal-tract variables. Nevertheless, the previous analyses do not explicitly test whether using a common control strategy based on acoustically defined targets is sufficient to explain the variety of articulatory configurations different speakers use in producing /r/. In order to address this issue, in the current article we explicitly simulate the outcome of a control strategy for /r/ production based on acoustic targets. These simulations are performed using specific models of two of our subjects' vocal tracts, so that the results can be directly compared to these subjects' observed articulatory configurations during the production of /r/.

In order to simulate the effect of a common control strategy based on acoustic targets for different speakers, we must first understand for each speaker the relationship between their articulators and the resulting acoustics. There are several reasons why we cannot use the previously obtained EMMA data and acoustic recordings for each subject in order to characterize this relationship. First, independent data pools for modeling and testing are always preferable, as this offers a generally more valid approach to hypothesis testing. Second and equally important, EMMA data have limited potential to characterize the articulatory–acoustic relationship given the relative scarcity of relevant articulatory information, which is limited by the number of available transducers. Articulatory–acoustic mappings obtained from EMMA data are not only less accurate but also lead to limited interpretability, as the researcher is left to speculate the vocal-tract profile from a limited sampling of interpolating points. MRI data, in contrast, provide a more satisfying characterization of vocal-tract morphology. We thus used simultaneous recording of MRI and acoustic data for two subjects to characterize the relationship between each subject's articulatory configurations and the resulting acoustics (see Sec. II B). Then, we simulated the effect of the hypothesized control strategy on each subject's vocal-tract model during the production of /r/ using different leading phonetic contexts, and the modeled results were compared to each subjects' productions (Sec. II C). While this methodology has the added complexity of combining MRI and EMMA data, it is a more valid and informative approach than one based on EMMA data alone. Furthermore, we believe the analyses in these sections not only add an important modeling examination of /r/ production but also contribute to efforts in speech production modeling that addresses speaker-specific behavior, rather than the behavior of an average or idealized speaker.

II. METHODS

A. EMMA data collection and analysis

An EMMA system (Perkell *et al.*, 1992) was used to track the movement of six transducer coils indicating the tongue shape (tongue back, tongue dorsum, and tongue tip), jaw aperture (transducer located on the lower teeth), and lips (upper and lower lip) in the midsagittal plane during the production of /r/ in five different phonetic contexts ("warav," "wabrav," "wavrav," "wagrav," "wadrav") for seven American English speakers. Each subject repeated each production between two and five times. A directional microphone was used to record the subjects' speech simulta-
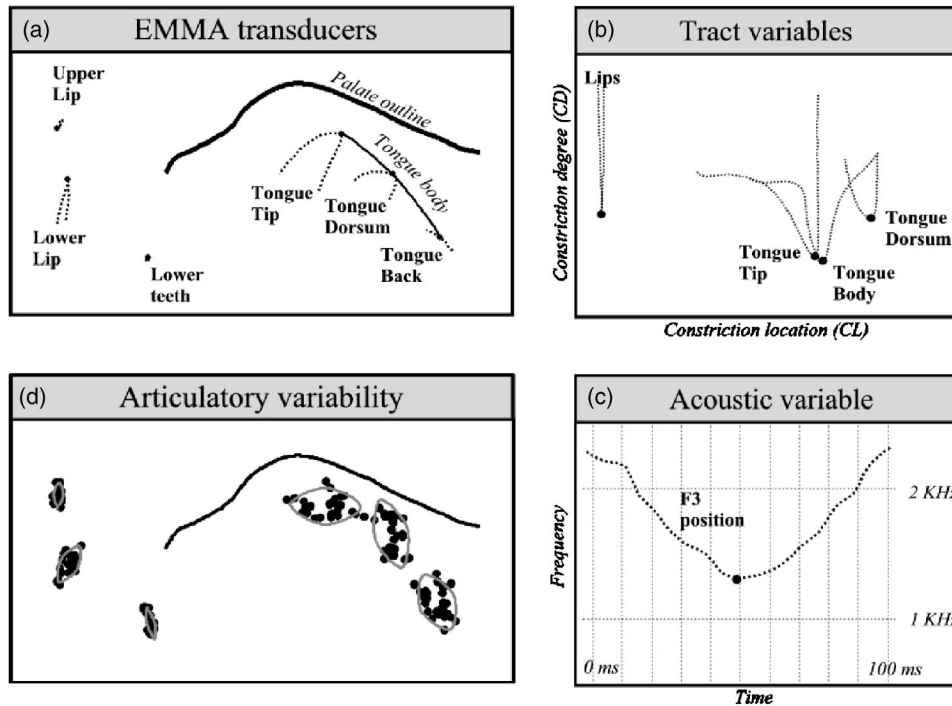
FIG. 2. Main elements in the analysis of EMMA data for each subject. (A). EMMA transducers: Example of the location of the six transducers during production of /r/ in /warav/. Dotted lines represent the trajectories of each transducer. Black dots indicate the center of the /r/ defined from the inflexion point of the tongue-dorsum (see the text for details on alternative definitions of /r/ centers). The line uniting the three tongue transducers was created using a Catmull–Rom spline. (B). Tract variables: Eight variables representing constriction degree and location are derived from the transducer positions to represent four relevant vocal-tract constrictions. Tongue-tip and tongue-dorsum constrictions represent the relative positions of these transducers to the palate outline. A tongue-body constriction was also defined using the relative position of the point on the tongue-body line closest to the palate. The lip constriction represents the relative positions of the two lip transducers (lip aperture and protrusion). (C). Acoustic variable: Trajectory of the third formant ($F3$) around the /r/ center. (D). Articulatory variability: Example of articulatory variability in /r/ production. Ellipsoids represent 95% confidence intervals of each transducer position during the /r/ for a series of /r/ productions in different phonetic contexts. The analyses in this section test the ability of the acoustic variable $F3$ and the eight tract variables defined above to characterize the observed articulatory variability.

neously with the EMMA signals. The details of the methodology are described in Guenther *et al.* (1999). The primary acoustic cue for /r/ is a deep dip in the trajectory of the third formant frequency, or $F3$ (Boyce and Espy-Wilson, 1997; Delattre and Freeman, 1968). The acoustic signal was therefore processed to extract the $F3$ trajectory. An initial definition of the acoustic center of the /r/ was constructed in terms of the time point of the $F3$ minimum. Figure 2 shows the main elements in the analysis of the EMMA data. The plot labeled (A) illustrates the trajectory of the six transducers for a window of 100 ms around the /r/ center during a ''warav'' production, and the plot labeled (C) shows the corresponding $F3$ trajectory.

In addition to the acoustic variable $F3$ [Fig. 2(C)], we defined eight vocal-tract variables reflecting the degree and location of four relevant tongue and lip constrictions [Fig. 2(B)]. *Tongue tip and tongue dorsum constriction degree* were defined as the distance between the hard palate outline and the tongue tip and tongue dorsum transducer positions, respectively. *Tongue tip and tongue dorsum constriction location* were defined as the positions along the hard palate outline of the point closest to each of these transducers. *Lip constriction degree and location* were defined as lip aperture (distance between upper and lower lip transducers) and lip protrusion (average horizontal position of the upper and lower lip transducers), respectively. To accommodate the possibility that the tongue transducers were not optimally

located at places of relevant constrictions, we defined an additional tongue constriction by connecting the three tongue transducer locations using a Catmull–Rom spline [shown in Fig. 2(A) as a solid line], and estimating the degree and location of the constriction formed by the point along the resulting tongue outline closest to the hard palate. We call the resulting measures associated with this additional constriction the *tongue-body constriction degree and location*. No tongue back constriction was defined due to the lack of information regarding the pharyngeal wall position for each subject. Based on these constrictions we constructed corresponding articulatory-based definitions for the /r/ centers. These were manually identified as the inflexion point in the trajectories of the four previously defined constrictions (three tongue constrictions and one lip constriction) within a window of 100 ms around the acoustically defined /r/ center. The /r/ centers are indicated in Fig. 2 by dots [in plots (A) and (D) dots indicate the tongue–dorsum /r/ center; in plots (B) and (C) dots indicate the corresponding constriction- or acoustically defined /r/ center]. The articulatory defined /r/ centers occurred on average 7 ms (95% CI [6, 9] milliseconds; $t_{591} = -11.4$; $p < 0.001$) before the acoustically defined /r/ centers. Among the articulatory defined centers the main difference was for that defined from the lip constriction. While the lip constriction extreme occurred on average 19 ms (95% CI [16, 22] ms, $t_{147} = -13.7$; $p < 0.001$) before the acoustically defined /r/ center, the different tongue con-

strictions were only 4 ms (95% CI [2, 5] ms; $t_{443} = -5.6$; $p < 0.001$) before the $F3$ minimum (and approximately at synchrony among them; ANOVA analysis, 7% intergroup variance, $F_2 = 2.64$; $p = 0.07$).

The articulatory data were analyzed in terms of the articulatory variability of the transducer positions [Fig. 2(D)] at the /r/ centers, using as hypothesized target variables the acoustic and tract variables defined above [Figs. 2(B) and (C)]. The details of this analysis follow. Variables associated with transducer positions were normalized independently and separately for each subject in order to appropriately compare across subjects, and also to reduce possibly confounding effects from the different ranges of operation of each of these variables (e.g., the lower teeth transducer showing a smaller range of movement than the tongue transducers). We computed for each subject the articulatory covariance matrix Ω_0 at the /r/ center. For the analyses involving an acoustic target variable we used the acoustically defined /r/ centers, and for the analyses involving a tract target variable we used the corresponding articulatory defined /r/ centers.

For each subject a principal component analysis of the articulatory covariance led to the definition of a set of 12 vectors or principal articulatory directions $\mathbf{q}_j$ ($j = 1, ..., 12$) defining a base in the articulatory space. Each of these unit vectors $\mathbf{q}_j$ represents a direction of change of the EMMA positions characterizing the observed articulatory variability. For each articulatory direction j the *percentage of articulatory variability* associated with this direction was computed as

$$\sigma_j \equiv \frac{\mathbf{q}_j^t \cdot \Omega_0 \cdot \mathbf{q}_j}{\Sigma_k \mathbf{q}_k^t \cdot \Omega_0 \cdot \mathbf{q}_k}.$$

Nine target variables were then hypothesized, eight corresponding to tract variables (constriction degree and location for each of the previously defined vocal-tract constrictions) and one corresponding to the acoustic variable $F3$. For each combination of an articulatory direction j and a target variable i, the effect of the articulatory dimension on the target variable was estimated as

$$\lambda_{ij} \equiv \frac{|\mathbf{q}_j^t \cdot \mathbf{X}^+ \cdot \mathbf{y}_i|}{\Sigma_k |\mathbf{q}_k^t \cdot \mathbf{X}^+ \cdot \mathbf{y}_i|}.$$

Here, $\mathbf{y}_i$ is a vector representing the time courses of the target variable i for a window of 10 ms around the /r/ center for all contexts and repetitions, and the matrix $\mathbf{X}^+$ represents the pseudoinverse of a matrix $\mathbf{X}$ containing the corresponding time courses of the transducer positions. The numbers λ_{ij} represent the absolute value of the expected change in the ith target variable associated with moving the articulators along the jth articulatory direction (normalized across all articulatory directions). They can be interpreted as a *percentage load* of the target variable on each of the articulatory dimensions.

We performed two set of analyses on these data, one categorical and one continuous. In the categorical analysis the articulatory dimensions were divided, for each target variable independently, into two sets Θ_i^{small} and Θ_i^{large}, corresponding to the *small-* and *large effect on target variable*

dimensions, and each defined as the six dimensions associated with the six lowest or the six largest λ_{ij} values, respectively. We then computed the percentage of articulatory variability associated with small effects on each target variable by combining the variability over of the associated articulatory dimensions: $\sigma_i^{\text{small}} \equiv \Sigma_{j \in \Theta_i^{\text{small}}} \sigma_j$. This leads to a value σ_i^{small} for each subject and for each target variable. Under the null hypothesis (no association between articulatory variability and effect on target variable), the expected percentage of articulatory variability associated with each of these sets would be 50%. We estimated the associated probability level of the observed data using Monte Carlo simulations on randomly defined sets Θ_i^{small}. For collapsing the results across subjects we computed the average of σ_i^{small} for each target variable, and the associated null hypothesis distribution was formed from an equal-weighted mixture of each of the conforming Monte Carlo distributions.

In the continuous analysis we constructed plots relating, for each hypothesized target variable i, the observed articulatory variability along each articulatory direction (σ_j) versus its effect on the target variable (λ_{ij}). The resulting plots were fit using a linear regression on the log variables. R^2 and p values, as well as confidence intervals for the linear fit parameters, are reported in Sec. III.

B. Construction of speaker-specific vocal-tract models

A *speaker-specific vocal-tract model* is a characterization of the range of configurations a speaker's vocal-tract could adopt, together with the acoustic output any configuration would produce under glottal excitation. To estimate the former, a set of 2D MRI midsagittal profiles was acquired for two subjects (the first two subjects in the EMMA experiment) while producing a set of phonemes. To estimate the latter (the associated acoustic outputs), acoustic data were collected at the start of each scan. The following paragraphs describe the data acquisition and the procedure used to interpolate and generalize from the limited available articulatory and acoustic data to other nonobserved configurations. The results provide a simple characterization of the full range of articulatory configurations and acoustic outputs a speaker can produce.

1. Data acquisition

Scans were performed with a 1.5-tesla Siemens scanner using a 14-s TR acquisition, 4-mm midsagittal slice with 256×256 matrix size. Subjects were asked to produce a simple utterance (either a steady-state vowel or a /VC/ sequence) and hold the last phoneme during the 14 s of the image acquisition procedure. Their productions were recorded using a microphone placed in the scanner near the subject's mouth. The MR acquisition started when the subject was holding the last phoneme to allow clear audio recording of their productions prior to the onset of scanner noise. Data for 27 and 15 phoneme productions were acquired for subject 1 and 2, respectively. Productions included several American English vowels (ε i æ Λ uw ɜ ei ow), semivowels (r), fricatives (ʃ s f), nasals (m,n), and stop (p t k b d g) consonant sounds. All utterances were used to

construct the articulatory models. However, since formants could only be reliably extracted for the vowel and semivowel utterances, only these utterances were used to formulate the mapping between articulator configurations and acoustics.

2. Analysis of vocal-tract configurations

Previous approaches to the creation of a parametric description of articulatory movements (e.g. Perrier *et al.*, 1992; Story *et al.*, 1996, 1998) create a grid in the midsagittal plane and obtain the vocal-tract area function from the intersection of this grid with the vocal-tract outline. An articulatory model based directly on a vocal-tract area function representation is, nevertheless, unlikely to produce optimally realistic articulatory movements, given the discontinuity between natural vocal-tract articulator movements and the corresponding area function representation using the grid method. For example, forward movement of the tongue body creates discontinuities in the associated area function changes each time the tongue tip crosses a grid line. These discontinuities are particularly marked when a cavity is formed below the tongue tip, as occurs in some /r/ productions. In this paper we chose to create a parametric definition of the articulator space from a principal component decomposition of the outlines of different vocal-tract segments (tongue, jaw, and lips). In this way the resulting characterization is expected to be both articulatorily meaningful and continuous with respect to movement of the articulators. MR images were inspected visually for movement artifacts, and trials with a large amount of movement were removed from further analyses. In each resulting raw MR image, the region associated with air (vocal cavity and the head exterior) was identified. Pixel intensities were automatically clustered into eight clusters. The idea was to identify the lowest intensity cluster with the regions of air in the midsagittal image. The user then selected a starting point from this air region and a flood-fill algorithm was used to define the air area. Images were manually edited to correct for the cases when the air area comprised multiple disconnected regions (e.g., when the lips were closed). The outline of the resulting air region was then extracted for each image. These vocal-tract outlines were aligned spatially using the hard palate outline to correct for subject movement in the scanner. They were then divided into different segments of interest (tongue body, jaw, lips, hard palate, velum, laryngeal region). Each segment was interpolated by a fixed number of equally spaced 2D points along the identified segment outline. To obtain a simple descriptor of each segment's shape we concatenated both the x- and y coordinates of all the points along a given segment outline. For the present study we concentrated on the effect of tongue, lower lip, and jaw. PCA was applied to each of these shape descriptors to obtain a set of five articulatory components: three for the tongue body, and one each for the jaw and lower lip. The variability in articulatory configurations explained by movements of the jaw was removed prior to the estimation of the tongue and lip principal components in order to remove redundancies in their definition (cf. Maeda, 1990). The resulting set of principal articulatory components was used as a characterization of the range of articulatory configurations the subject could produce. In this way, any articulatory configuration the subject's vocal-tract model could produce was represented by a five-element vector, describing the contribution of each of the five articulatory components to the vocal-tract shape.

3. Analysis of acoustic signals and the articulatory to acoustic mapping

Acoustic recordings of the subject's production of each utterance made while in the MRI scanner (just before the onset of the scanner noise) were analyzed using linear predictive coding (LPC) ($p = 26$, $F_s = 22$ KHz). The acoustic signal was pre-emphasized with a single delay FIR filter (a1 = 0.95) to reduce the effects due to radiation and the glottal pulse (Wakita, 1973). The first three formant values were extracted for each production.

In order to approximate the vocal-tract articulatory/ acoustic mapping, past studies have typically used a transformation from midsagittal cross dimensions to an area function. Then, from acoustic theory the frequency response of a particular vocal-tract shape is computed. In this transformation there are several unknowns that cannot be obtained from simple misagittal MR images, most importantly the midsagittal cross section to area function relationship. Previous models have either fitted these parameters to the subject's acoustic productions (e.g., using a relatively difficult to tune elliptical approximation to the area cross sections; Maeda, 1990) or an elegant but more complex estimation procedure based on multiple 3D volumetric MRI representations of the vocal-tract (Tiede *et al.*, 1996). The collection of 3D volumetric data for multiple phonemes is time-consuming and can suffer from problems in determining the location of the teeth, which do not show up on MR images and thus adversely affect the measured area function. In contrast to this approach, here we use a purely statistical approach characterized by a linear mapping fitting the relationship between the articulatory and formant descriptors for each subject. In this way, the proposed model offers only an approximation to the articulatory–acoustic relationship, but has the advantages of requiring a relatively small amount of MRI and acoustic data for each subject and avoiding the complications derived from the estimation of the area function. The linear mapping best fitting the relationship between articulatory and acoustic components for each subject's data was then estimated using linear regression on the articulatory and acoustic descriptors from vowels and semivowels (nine and six configurations for Subjects 1 and 2, respectively).

The validity of this approach was first estimated by creating a random sample of vocal-tract configurations, and computing the corresponding acoustic outputs using a standard articulatory synthesizer (Maeda, 1990). A random set of 10 000 valid articulatory configurations was created using a normal distribution of the model's articulatory parameters (mean zero, standard deviation one) hard limiting between −3 to 3 standard deviations (the full valid range of articulatory parameters in Maeda's 1990 vocal-tract model). For these data we found a very significant linear relationship ($R^2 = 0.97$) between the articulatory and formant descriptors. Deviations from linearity were most apparent in extreme configurations (close to a closure). For each articulatory con-

figuration **x** we constructed an approximate measure of percentage extent of closure as $100/k$, where the value k is the minimum value such that the articulatory configuration $x_0 + k(x - x_0)$ would result in a closed vocal-tract configuration ($\mathbf{x}_0$ represents a rest configuration). This measure is 0% for a rest configuration, and 100% for a closed configuration. For this measure we observed that the previously estimated articulatory–acoustic fit provided good approximations ($R^2 > 0.9$) for relatively open configurations ($100/k < 80\%$), but this fit was considerably poorer ($R^2 = 0.65$) for configurations near closure ($100/k > 90\%$). For comparison, average articulatory configurations for /r/ production for Subjects 1 and 2 were reasonably open ($100/k \approx 60\%$). These results indicate that a linear mapping between articulatory and acoustic dimensions is reasonable for our present analysis demands, and in general it is appropriate if the vocal-tract is restricted to nonextreme configurations (e.g., vowels and semivowels). In other words, this methodology would not be appropriate for modeling many consonant productions. As a last validation analysis we estimated the effect that a limited amount of available data points (nine and six configurations for Subjects 1 and 2, respectively) would have in our estimation procedure. The average errors in the estimation parameters (linear regressors) using randomly selected sets of nine and six configurations were found to be relatively low (2% and 11%, respectively, for Subjects 1 and 2).

C. Simulations of /ɹ/ production

The DIVA model (Guenther *et al.*, 1998) was used as a controller for the movement of the speaker-specific vocal-tract articulators to produce an acoustic /r/ target in different phonetic contexts. The DIVA model can be characterized as a derivative controller in the acoustic space. The implementation reduces, at each time point, to iteratively moving the articulators in the articulatory direction that brings the current acoustic output closest to the desired acoustic target. In mathematical terms, the model uses a pseudoinverse of the Jacobian matrix relating articulator movements to their acoustic consequences to move in a straight line (in acoustic space) to the target (see Guenther *et al.*, 1998 for details). While in the complete DIVA model this is accomplished by learning this pseudoinverse transformation through experience (e.g., Guenther *et al.*, 1998), in the current implementation we used an explicit calculation of the pseudoinverse of the articulator-to-acoustic mapping. The articulatory space was defined in terms of the PCA components as described above, and the acoustic space was defined in terms of the first three formants of the spectrum (in Hz). The acoustic target in the model was defined from each subject's own /r/ production formants. To compare the results of the DIVA model simulations to the experimentally obtained EMMA data for each subject, the estimated transducer locations were manually identified on a rest configuration of the modeled speaker-specific vocal tract. The approximate location where the tongue transducers were placed was visually identified following the directives of the original EMMA experimental paradigm, as 1, 2.5, and 5 cm back from the tongue tip. The initial vocal-tract configurations of three phonetic contexts (/ar/, /dr/, and /gr/) were manually edited from the original

MRI data to approximate the observed initial transducer configuration (75 ms before $F3$ minimum) in the corresponding contexts for each subject. Simulations of the DIVA model were run starting from these configurations to a "final" configuration at the $F3$ minimum for /r/. The estimated direction of movement (difference between the final and starting transducer positions) was compared to the measured transducer movement in the same contexts (correlation coefficients are reported). Finally, using all available MRI configurations as initial vocal-tract configurations (not just the three used for the preceding analyses), additional simulations were run using the same acoustic target for /r/, and the resulting articulatory variability across the model's /r/ productions was determined. On these data we performed articulatory variability analyses similar to those performed on the original EMMA data.

III. RESULTS

A. Predictive relations between hypothetical target variables and articulatory variability

This section deals with the analysis of articulatory movement data in an attempt to show the ability of different phonemic target hypotheses to account for the observed articulatory variability in the production of /r/. In particular, it was expected that the choice of an "appropriate" phonemic target would provide good separability of those directions of articulatory movement showing large versus small articulatory variability. The main result shows that, among the hypothesized target variables, the acoustic variable $F3$ provides the best predictions of the articulatory variability in /r/ production. In particular, for any direction of articulatory movement, its effect on the acoustic variable $F3$ is strongly related (for each subject and across subjects) to the extent of articulatory variability along this direction. On the other hand, none of the tract variables tested (corresponding to an articulatory phonemic target representation hypothesis) provides as good predictability across subjects of the articulatory variability in the production of /r/. This section presents these comparative results, and provides a series of analyses describing the observed relationship between effect on $F3$ and articulatory variability.

Figure 3 shows, for each subject, and collapsed across all seven subjects, the percentage of articulatory variability associated with dimensions that have small effects on each of the hypothesized target variables (this percentage of articulatory variability is labeled σ_i^{small} in Sec. II, where i represents each of the hypothesized target variables). Under the null hypothesis (no association between articulatory variability and effect on a target variable) these percentages would be 50%. Higher numbers indicate inverse association between effect on a target variable and articulatory variability, and are taken as indicative of a control strategy utilizing the target variable in the definition of the phonemic target. For example, the articulatory variability for Subject 2 (shown for reference in the left-most column of the figure) shows a tongue-tip distribution similar to that schematized in the example of Fig. 1 (indicating a possible tongue-tip constriction degree target), and the corresponding cell in the table indi-
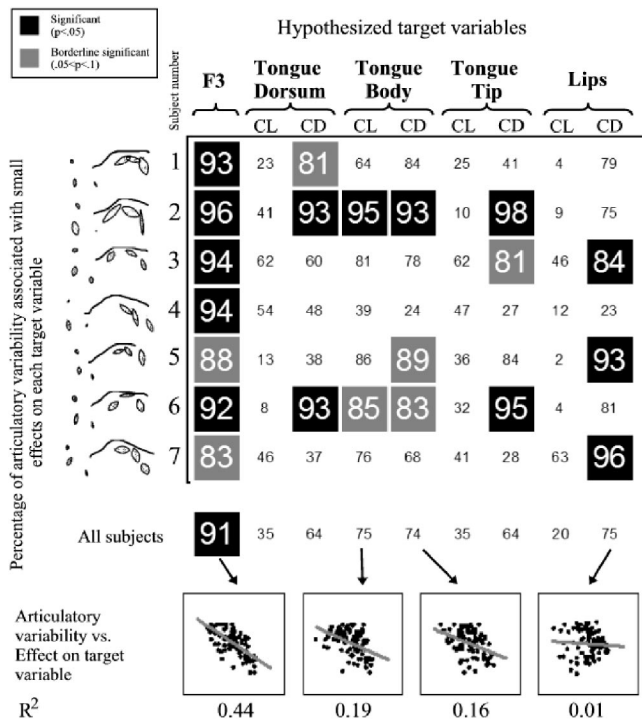
FIG. 3. Relation of hypothetical target variables to articulatory variability during /r/ production. Top: Categorical analyses. Table shows the percentage of articulatory variability associated with small effects on each hypothesized target variable (columns) for each subject (rows), and across all subjects (last row). Statistically significant percentages are highlighted. For reference, plots at the left of the table schematize the shape of the articulatory variability for each subject. Bottom: Continuous analyses. Plots show the relation between each articulatory dimension's variability (abscissa) and its effect on some of the most likely target variables (ordinate). Each dot in the plots represents an articulatory dimension (i.e., a direction of movement of the articulators) for a given subject. Both articulatory variability and effect on target variable are represented in log percentage units.

across all subjects (*All Subjects row*) indicate that $F3$ is the most consistent phonemic target among the hypothesized variables. Small effects on $F3$ are associated on average with a significant amount of articulatory variability (91%, $p = 0.03$). In contrast, small effects on none of the hypothesized tract target variables are found to be significantly associated across subjects ($p > 0.21$) with the extent of articulatory variability. Small effects on tongue-body constriction degree and location and lip constriction degree are among the best competing tract variable hypotheses, each associated with about 75% of the articulatory variability (not significantly greater than 50%, $p = 0.21$). The plots at the bottom of the figure show the associations between effect on each target variable and articulatory variability in a continuous form. Again, the acoustic target $F3$ is best supported by our data, showing the strongest inverse association ($R^2 = 0.44$), as expected from a motor control strategy that utilizes $F3$ as a phonemic target.

These results indicate that, among the target variables tested, $F3$ is the most likely target variable that appears consistently across subjects in the production of /r/. In particular, they show that if, for a given subject, deviating from an average /r/ configuration along a given articulatory direction was found to have a relatively large impact on $F3$ (low $F3$ stability), then that subject tended to show little articulatory variability along this articulatory dimension. Conversely, if deviating along a given articulatory direction was found to have relatively little impact on $F3$ (high $F3$ stability), then the subject tended to show a larger amount of articulatory variability along this articulatory dimension. We will refer to this as a *predictive relationship between acoustic stability and articulatory variability*. Figure 4 highlights the continuous (left) and dichotomous (right) description of this relationship. Each dot in the left plot represents an articulatory dimension for a given subject. Their position represents the relative effect of each articulatory dimension on $F3$ (in percentage load, compared to other dimensions for the same

cates that in fact for this subject a significant amount of articulatory variability (98%) could be associated with this constriction target. While each subject shows indication of one or more possible phonemic targets, the collapsed results
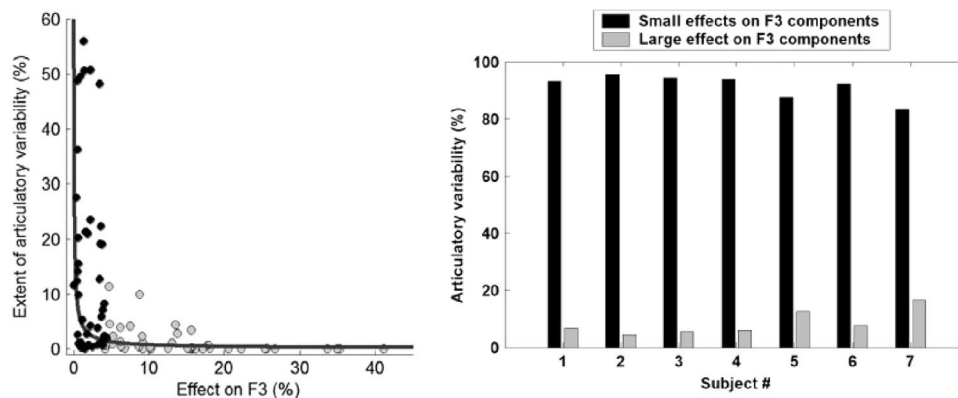

FIG. 4. Predictive relationship between acoustic stability and articulatory variability. Left: The extent of articulatory variability (in percentage of total variability for each subject) vs the effect on third formant frequency (in percentage load for each subject—see the text for details) for all articulatory dimensions for all subjects (each dot represents an articulatory dimension—a direction of movement of the articulators—for a given subject). The thick line represents the inverse relation fit to data (approximating the curve $y = 10/x$). Black/gray points represent the articulatory dimensions that, for each subject, would be categorized as small/large effect on $F3$ components. The inverse relation shown in this plot is identified as a *predictive relation between acoustic stability and articulatory variability* Right: Consistency of found articulatory/acoustic relations across subjects. The percentage of articulatory variance associated with large/small effect on $F3$ components is shown for each subject. Under the null hypothesis (articulatory variability not associated with the effect on $F3$) these percentages would be equal (50% each). A strong bias of the articulatory variability towards those articulatory dimensions that have a small effect on $F3$ is apparent in all the experimental subjects.

subject) versus the extent of articulatory variability found along this articulatory dimension (percentage of total variability for each subject). The solid line represents the linear fit on the log variables, which approximates the curve $y = 10/x^{1.2}$ ($F_{1,82} = 65.2$; $p < 0.001$, 95% confidence intervals [5, 17] and [0.9, 1.5] for the constant in the numerator and the exponent of x, respectively). Dots are colored based on their effect on $F3$, dichotomized to only two equal-sized levels: dark or light dots represent those dimensions that have a small or large effect on $F3$, respectively. The bar plot in Fig. 4 right represents, for each subject, the cumulative variability associated with each of these two levels. We maintain that this relationship is the hallmark of a control mechanism that utilizes an $F3$ target. In computer simulations reported below we validate this claim by simulating a speech control mechanism utilizing an $F3$ target that replicates this relationship.

To assess the statistical significance of the continuous version of the observed predictive relationship between acoustic stability and articulatory variability across subjects, we performed a Monte Carlo test involving replication of all the analysis steps using a series of simulated datasets conforming to a predefined null hypothesis. The null hypothesis represents the case where there is no relation between articulatory variability and acoustic stability. In a worst-case scenario an artifactual relationship could stem solely from measurement noise in the estimation of $F3$. The Monte Carlo dataset consisted of simulated transducer positions at the /r/ center following the same distribution as those observed in our data, and a simulated target variable randomly distributed and independent of the transducer positions. The 95th percentile of the R^2 distribution under this null hypothesis (from 10 000 Monte Carlo simulations) was relatively large ($R^2 = 0.42$), just below the observed R^2 value from our data ($R^2 = 0.44$; $p = 0.03$). Under this test, only the predictive relationship using the acoustic variable $F3$ survives a 0.05 significant level for the pooled data. For the tract variables the significance level of their predictive relationships is always greater than $p = 0.89$. Yet, these are very conservative tests as they do not take into account the observed degree of association between transducer positions and tract variables, which generally indicate a small presence of measurement noise (an average of 89% of the acoustic variable and $>95\%$ of each tract variable was linearly associated with the transducer positions). When this is incorporated into the Monte Carlo simulations (by creating a simulated target variable equal to the average transducer position plus a variable amount of independent random noise), the 95th percentile of R^2 under the null hypothesis drops to a value of $R^2 = 0.03$. Under this more liberal test the predictive relationships using not only $F3$ but also tongue-body constriction location and degree would become statistically significant ($p < 0.05$). While the across-subject results need to be interpreted with care, due to the limited amount of subjects in this study, the consistency of the individual subject results together with the Monte Carlo simulations indicate that the observed relation between acoustic stability and articulatory variability is statistically significant beyond possible artifactual causes.

An important source of contextual variability in the cur-rent experimental setup is the phonetic context preceding the /r/ production. Articulatory/constriction target models often employ context-dependent articulatory targets (e.g., blended targets in the task-dynamic model of Saltzman and Munhall, 1989), as they are believed to explain the source of articulatory variability. According to these models, in our analyses of articulatory variability, context would be acting as a confounding effect. What we mean by this is that the observed relationship between acoustic stability and articulatory variability could simply be addressing how these context-dependent targets are organized, instead of addressing the target space definition in the speech control strategy. To address this concern, we replicated our original analyses but now explicitly treating context as a confounding effect and removing its effect on the observed articulatory variability by analyzing the intracontext variability in transducer positions. Interestingly, the percentage of *intracontext* articulatory variability associated with small effects on $F3$ was 88% across subjects, very similar to the original 91% of *total* articulatory variability associated with small effects on $F3$. This result was still the only one statistically significant ($p = 0.03$) among the tested target variables (next competing tract variable was tongue-body constriction location, 76%; $p = 0.15$). What these results indicate is that the observed relationship between acoustic stability and articulatory variability is not an effect of the phonemic context. Furthermore, they indicate that the evidence for tract-variable targets does not significantly improve when considering the effect of the phonetic context on the articulatory variability (i.e., when allowing a different target for each phonetic context). This supports the interpretation of the observed relationship in terms of a motor control mechanism utilizing acoustic targets, rather than one utilizing context-dependent tract-variable targets.

Overall, the positive results in this section highlight a strong and consistent relationship between the acoustic variable $F3$ and articulatory variability. This result is schematized in Fig. 5 to facilitate interpretation. This relationship is consistent with that expected from a control mechanism using an $F3$ target; i.e., the final articulatory variability is lower for those articulatory directions most relevant to determining the $F3$ value (axis A in the plot). Furthermore, this relationship appears both when looking at the total articulatory variability (dotted black ellipsoid) and when looking at the intracontext articulatory variability (dotted gray ellipsoids; the articulatory variability within each of the phonetic contexts tested). These results suggest that an acoustic target motor control mechanism utilizing the same acoustic target across contexts can account for the observed range of articulatory configurations during /r/ production. The next subsection further investigates this assertion with a specific model that utilizes an acoustic target for /r/.

B. Speaker-specific vocal-tract models

For the first two subjects participating in the previous analyses, we constructed from MRI and acoustic data a simple model characterizing the specificities of their vocal tracts and the range of acoustic signals (limited to the first three formant values) that different configurations would produce. PCA of the articulatory configurations led to a set

Nieto-Castanon *et al.*: A modeling investigation of articulatory
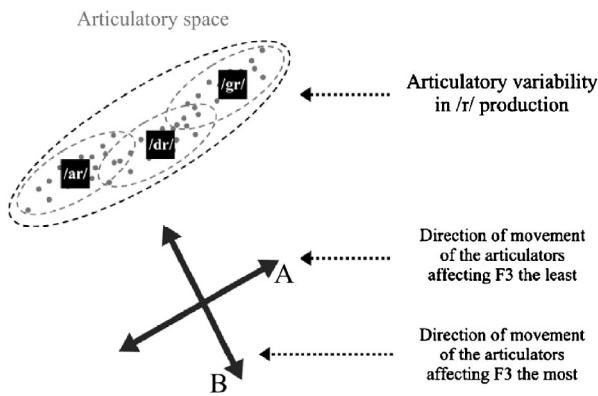
FIG. 5. Diagram summarizing the main results in this section. The plot represents in a schematic way the range of articulatory configurations (dots in the plot) reached in the production of /r/ under different phonetic contexts (black boxes). The main results are (a) An acoustic variable ($F3$) is the best predictor among the phonemic target variables tested for the shape of the articulatory variability in the production of American English /r/. The articulatory variability is maximal along the directions of movement of the articulators associated with small $F3$ changes, and minimal along the directions of movement of the articulators associated with large $F3$ changes. (b) The intracontext articulatory variability (the articulatory variability for each of the phonetic contexts) shows the same association with the effect of $F3$, indicating not the action of a context-dependent target definition, but possibly a common control mechanism utilizing an acoustic phonetic target.

FIG. 6. Characterization of speaker-specific vocal-tract models. Sample movements of the models for Subjects 1 and 2 to change $F1$ (left), $F2$ (center), and $F3$ (right) are shown. For each subject, the deviations from a neutral articulatory configuration necessary to produce an individual change (increase/decrease) in each of the first three formants of the resulting auditory signal is shown in each column (e.g., the first column represents the movements associated with changes in $F1$ while keeping $F2$ and $F3$ constant). The gray area represents the configuration that produces the highest formant value (for the corresponding formant) among the configurations represented.

of five meaningful articulatory components covering 75.4% and 83.7% of the total observed variability in shape for the two subjects, respectively. The jaw component primarily describes the aperture/closure of the mouth, along with the associated lip aperture/closure, and lowering/raising of the tongue body; the three tongue components describe approximately the raising/lowering of the apical and dorsal areas of the tongue and its front/back movement; the lip component describes the frontal extension (protrusion) of the lips (cf. Maeda, 1990; see also Sec. IV). Components derived from other vocal-tract segments (a velum component, describing the opening/closing of the nasal cavity; and a laryngeal component, describing the raising/lowering of the base of the laryngeal region), were estimated but not explicitly used in the simulations presented in this paper (other than any of their movement that was associated with the jaw component). The articulatory to acoustic mapping was then estimated by a linear fit between the articulatory configurations (defined by the positions of each of these five components) and the corresponding acoustic output (defined by the first three formant values measured during the MRI scans). Figure 6 characterizes the resulting mappings by illustrating the movements of the resulting speaker-specific vocal-tract models to achieve changes in $F1$, $F2$, and $F3$. Each column represents for each subject the movement of the articulators, starting from a rest or average configuration, that would be associated with changes in an individual formant. The results are consistent with standard characterizations (Schroeder, 1967; Fant, 1980) of high/low tongue configurations associated with low/high values of $F1$, respectively (left column in Fig. 6), and front/back tongue configurations associated with high/low values of $F2$, respectively (middle column in Fig. 6). At the same time, the resulting vocal-tract models accommodate the specificities of each subject. For example, Subject 2 tended to use lip protrusion more actively to lower $F2$
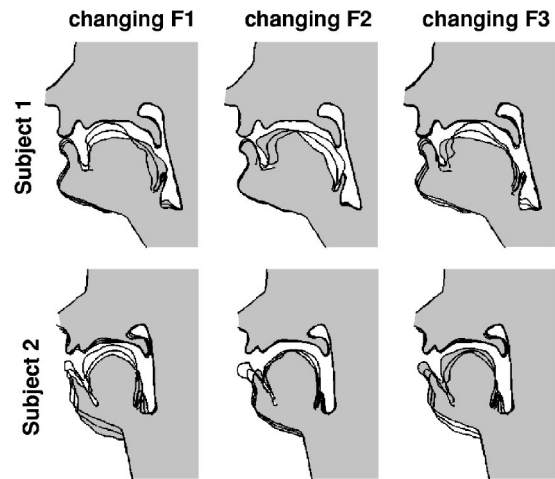
(see for example Perkell *et al.*, 1993, 1995, where trading relations between lip protrusion and tongue-body raising, argued to stem from their motor equivalence in the control of $F2$, were investigated in the context of /u/ production). With respect to the action on $F3$, Subject 1's movement to decrease $F3$ can be interpreted from an acoustic theory analysis as an increase in the front cavity length together with a decrease of the palatal constriction area, both acting to lower the third formant value. Subject 2 appears to decrease $F3$ primarily by increasing the size of the front cavity.

C. Simulations of /ɹ/ production

A simplified version of the DIVA model (Guenther *et al.*, 1998) was used to control movements of the speaker-specific vocal-tract models for Subjects 1 and 2 while performing /r/ productions in different phonetic contexts. An acoustic /r/ target was defined by its first three formants values ([593, 1238, 1709] Hz for Subject 1, and [376, 1476, 1990] Hz for Subject 2), and the simulations were run starting from articulatory configurations representative of the leading context phonemes (see Sec. II C for details). In order to compare the model simulations to the EMMA data, approximate transducer locations were manually identified (see Sec. II) on each subject-specific vocal-tract model. Acoustic and articulator trajectories for the production of /r/ in the contexts /ar/, /dr/, and /gr/ were then obtained using the DIVA model. These contexts were chosen to represent the full range of articulations seen in the experimental data.

Figure 7 compares the experimentally measured EMMA data (first row) to the simulation results (second row) for each subject, in terms of the direction of movement of the tongue transducers. The initial transducer positions in the simulations is fixed to that obtained from the EMMA data 75 ms before the $F3$ minimum (dashed lines). The results indi-

J. Acoust. Soc. Am., Vol. 117, No. 5, May 2005

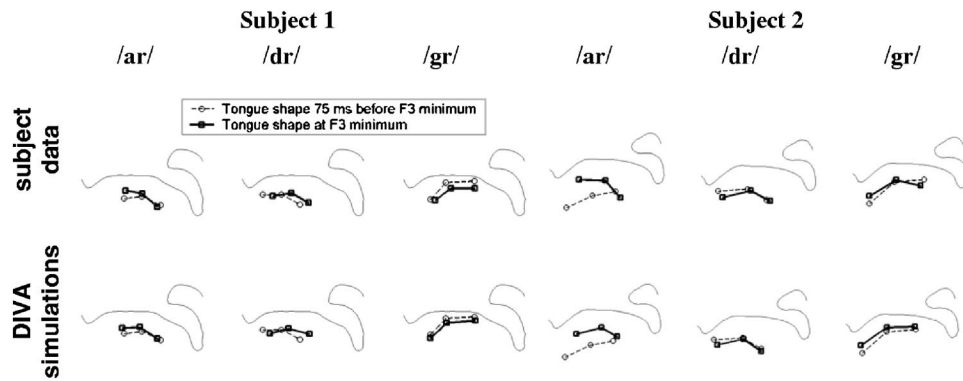Nieto-Castanon *et al.*: A modeling investigation of articulatory   3205

FIG. 7. Simulations of the DIVA model producing /r/ in different leading phonetic contexts. Top row shows the average lingual gestures used by each subject when producing /r/ in the contexts (from left to right) /ar/, /dr/, and /gr/ as measured using electromagnetic midsagittal articulometry (EMMA). Bottom row shows the simulation results using the DIVA model (with a subject-specific acoustically defined target for /r/) in conjunction with each subject-specific vocal-tract model. Dashed lines represent the initial (75 ms before $F3$ minimum) transducer positions, which are fixed to the experimentally observed values in the DIVA simulations. Solid lines represent the final transducer positions (at the $F3$ minimum for /r/). The outline of the hard palate and velum is included for reference. The correlation between the modeled and experimental movement of the tongue (tongue gestures) was $r = +0.86$ and $r = +0.93$ for Subjects 1 and 2, respectively.

cate that the direction of movement estimated using the DIVA model for the three leading phonetic contexts closely approximates the experimentally measured data for both subjects. The correlation between modeled and experimental change in transducer positions (tongue gestures) was $r = +0.86$ and $r = +0.93$ for Subjects 1 and 2, respectively. Qualitatively, the model mimics the range of /r/ configurations used by each subject in the phonetic contexts tested (thick black lines in Fig. 7).

Next, we investigated the ability of an acoustic target speech motor control scheme to predict the emergence of the articulatory/acoustic relationship observed in the experimental data. To that end, we analyzed the /r/ production simulation final articulatory configurations when using a wide range of leading phonetic contexts. All available configurations from the MRI data of each subject were used as starting articulatory positions and the DIVA model was run using the same acoustic /r/ targets as in the preceding simulations. Analysis of the resulting articulatory variability led to the

results shown in Fig. 8. For each subject, the five articulatory dimensions show the expected predictive relations between acoustic stability and articulatory variability (Fig. 8 left; cf. the experimental results in Fig. 4, left). The relation between articulatory variability and effect on $F3$ predicted by the model is close to linear in the log variables ($R^2 = 0.93$), justifying the use of this family of curves when fitting the experimental data. For the simulated data, the linear regression on log variables shows a significant relationship of the form $y = 17/x^{0.8}$ between the tested variables despite the limited data ($F_{1,8} = 99.8$; $p < 0.001$, 95% confidence intervals [10, 28] and [0.6, 1.0] for the constant in the numerator and the exponent of x, respectively). As an additional test, we analyzed the initial articulatory variability (the variability of the contextual articulatory configurations, prior to any movement of the articulators) and confirmed that the articulatory/acoustic relation was not present in the contextual configurations prior to the action of the speech controller ($p > 0.39$). This indicates that the relationship resulted from the
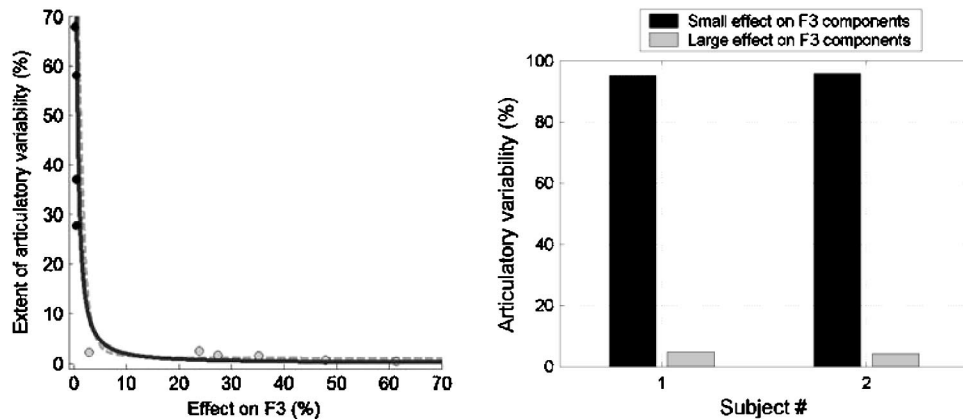

FIG. 8. Simulated articulatory/acoustic relations in /r/ production using the DIVA model (compare to experimental relations in Fig. 4). Left: The extent of final articulatory variability vs the effect on $F3$ for all articulatory dimensions of both subjects' simulations. The solid curve represents the inverse relation fit to these data. The dotted curve represents the expected predictive relation as theoretically derived from the DIVA model (see the Appendix). Black/gray points represent the articulatory dimensions that would be categorized as small/large effect on $F3$ components, respectively. Right: Consistency of simulated articulatory/acoustic relations across subjects. The percentage of articulatory variance in the simulated /r/ productions associated with large/small effect on $F3$ components is shown for each subject. As in the experimental data (see Fig. 4, right), a strong bias of the articulatory variability toward those articulatory dimensions that have a small effect on $F3$ is apparent in both subjects' simulations.

Nieto-Castanon *et al.*: A modeling investigation of articulatory

movements produced by the DIVA model. Furthermore, the simulation results mimic the expected relationship as derived theoretically from the DIVA control equations (dotted line in Fig. 8, left; see the Appendix for this derivation). The nature of the inverse relation predicted by the model ($y \propto x^{-0.8}$) was slightly shallower than the one observed in the EMMA data ($y \propto x^{-1.2}$), but the confidence intervals for the two curve parameters overlap as an approximate $y = 10/x$ relation. For completeness, Fig. 8 (right) illustrates the consistency of articulatory/acoustic relations in the simulations across the two subjects (cf. the experimental results in Fig. 4, right). Overall, these results indicate that an acoustic target controller, such as the one used in the present simulations, predicts the relationship between acoustic stability and articulatory variability observed in the experimental data. Furthermore, the DIVA model produces articulatory movements that closely mimic those of a particular speaker when controlling a speaker-specific vocal-tract model.

IV. DISCUSSION

A. On coordinate frames and articulatory dimensions

In target-based speech motor control models, the question of what coordinate frame is used by each model is usually identified with the proposed target representation. The task-dynamic model of Saltzman and Munhall (1989) exemplifies a type of computational model that uses a vocal-tract shape coordinate frame (vocal-tract targets defined by tract variables). The DIVA model (Guenther *et al.*, 1998) exemplifies a computational model that uses an acoustic coordinate frame (targets defined by acoustic variables). While there are many different coordinate frames one could use to represent the articulatory state, a major question for speech production modelers is, what coordinate frame(s) provides a simpler or more parsimonious characterization of behavioral data? In the same way as physical laws can be more readily unveiled when using an appropriate coordinate frame (e.g., planet orbits from an earth-centered vs a sun-centered coordinate frame), for speech production the use of an appropriate coordinate frame should allow the researcher to more clearly expose functional relations in the data. Finally, the ability of different coordinate frames to characterize the available motor speech production behavioral data could direct and facilitate the modeler's enterprise in proposing specific motor control strategies, and in particular it directly relates to the question of appropriate target definitions in target-based motor control schemes.

The behavioral data dealt with in this study is the articulatory variability present in American English /r/ production. Since articulatory variability is a local property (it characterizes the local departures in articulatory configurations from an average configuration) a linear approximation to the articulatory space geometry is appropriate. The issue of coordinate frames, under a linear approximation, becomes the simpler issue of characterization of vector spaces. Under this framework the articulatory space is a multidimensional vector space, and its characterization reduces to the definition of an appropriate base (a set of independent articulatory dimensions, each describing a direction—or vector—in the articu-

latory space). Different bases would in this way characterize different coordinate frames for the description of the articulatory state. Each of the columns in Fig. 6, for example, describes a different articulatory dimension (i.e., a direction of movement, or vector, in the articulatory space). The three articulatory dimensions in this figure characterize an acoustic coordinate frame (one based on three formant descriptors).

B. Predictive relations between acoustic stability and articulatory variability

A purely empirical approach to describing appropriate coordinate frames for the characterization of articulatory variability in /r/ production could be potentially given by a PCA of the articulatory covariance. This analysis provides the set of independent articulatory dimensions that best (most simply) characterize the observed articulatory variability. Conceptually, these correspond to the articulatory dimensions that offer an optimal separability of the total articulatory variability. In a two-dimensional case, for example, the resulting two articulatory dimensions would correspond to those dimensions associated with the largest and smallest variability, respectively (i.e., there is no one-dimensional subspace comprising more variability than that associated with the first articulatory dimension; equally, there is no one-dimensional subspace comprising less variability than that associated with the second articulatory dimension). A purely empirical approach like this, nevertheless, has potentially limited generalizability; i.e., since articulatory variability is a local property, the characterization resulting from the analysis of /r/ production might not be appropriate for other production examples. Furthermore, the researcher is left to interpret the resulting articulatory dimensions in terms of his/ her theoretical constructs.

In this paper we opted for a mixed empirical/theoretical characterization of the observed articulatory variability. In this way, we tested the ability of theoretically motivated articulatory dimensions to offer good separability of the observed variability in articulatory configurations. We feel that this approach has a better chance to generalize to other cases of speech production data, and that it offers a more useful source of information for the development of motor control models of speech production. We also take the view that an account which involves a common control strategy across speakers is preferable to an account that requires different strategies across speakers as it is the more parsimonious account. From this perspective, the relevance of the results presented in the $F3$ column of Fig. 3 is that they show how an articulatory dimension defined by an acoustic property ($F3$, a salient acoustic cue for /r/ perception), offers a good separability of the observed articulatory variability in /r/ production for all subjects tested. In particular, an average of 91% of the articulatory variability concentrates along articulatory dimensions that have a relatively small effect on the third formant ($F3$) value, while only 9% concentrates along articulatory dimensions which have a relatively large impact on $F3$. This result indicates that an acoustically defined articulatory dimension would be a good candidate to enter an appropriate coordinate frame characterization of the presented speech production behavioral data. Furthermore, fol-

lowing the original motivation for searching appropriate coordinate frame characterizations, we showed (Fig. 4) that using an acoustically defined coordinate frame can also be useful for unveiling functional relations in the behavioral data. In particular, we showed that the degree of articulatory variability associated with any particular articulatory dimension is related to the associated extent of change in $F3$ by a linear relationship in the log variables ($R^2 = 0.44$; $p = 0.03$). This relationship is conceptualized as a predictive relation between acoustic stability and articulatory variability. The form of this relationship is again consistent with that expected from a control mechanism using an $F3$ target; i.e., the final articulatory variability is lower for those articulatory dimensions most relevant to determining the $F3$ value.

The previous results show that an acoustic frame of reference can offer a useful characterization of the observed articulatory variability in American English /r/. In terms of the implications of these results for speech production modeling, the results indicate that, while no tract variable dimension was used consistently across speakers for the specification of /r/, all of the subjects showed evidence of an acoustic specification of /r/. The most parsimonious interpretation of these results points to the use of a common control strategy that utilizes acoustic, rather than articulatory, phonemic targets. Note that the results explicitly address the possibility of common acoustic variables *forming part* of the global target specification for /r/, not whether they fully define it. In this way the results indicate that $F3$ is likely to form part of the target specification for /r/, but we would neither claim nor expect it to be the only component in the target specification for this phoneme.

An important issue regarding the observed articulatory/acoustic relations examines the extent to which they favor acoustic target motor control models in contrast to vocal-tract target models. Several results of the present study build a very strong case for the acoustic target hypothesis. First, the results in Fig. 3 indicate that, while the tested acoustic variable ($F3$) shows a significant relation with the extent of articulatory variability (91%; $p = 0.03$), making it a potential candidate for a useful articulatory coordinate frame definition, the hypothesized vocal-tract variables fail to show such a relation (less than 75%; $p > 0.21$). This negative result addresses mainly the lack of consistency across subjects when hypothesizing tract-variable targets, and also the small evidence for some subjects of any form of tract-variable targets (e.g., Subject 4, although this could be related to the inability of our EMMA data to inform us about possible pharyngeal wall constrictions). Another piece of comparative evidence between acoustic and vocal-tract target hypotheses addresses the possibility of context-dependent effects (context here refers to the phoneme preceding /r/). Our results indicate that the observed articulatory/acoustic relations do not stem solely from the context-dependent articulatory variability, and can be equally observed when focusing on the intracontext articulatory variability (i.e., the articulatory variability resulting from /r/ production in each specific phonetic context). This result again points towards hypotheses that posit the observed trading relations as resulting from the motor control strategy (such as the acoustic target hypothesis),

rather than explanations that rely on context-dependent targets (such as the possibility of different articulatory targets for /r/). Last, the possibility of context-dependent articulatory targets was also directly addressed by trying to show predictive relations between tract variables and intracontext articulatory variability. Our failure to observe such relations indicates that using context-dependent articulatory targets does not seem to significantly improve the predictive ability of hypothesized tract variables on the observed articulatory variability. Overall, the results indicate that, for American English /r/, subjects consistently act as though they attempted to produce stable $F3$ configurations. The articulatory variability is reliably minimal along those articulatory dimensions that are important for determining $F3$. No vocal-tract target variable tested offers this level of generalization across subjects. One might argue that, given the linear nature of our analyses, articulatory targets defined as linear combinations of tract variables are completely equivalent to acoustic targets. From this perspective the results simply indicate that, if articulatory targets are being used, they are probably not defined by simple vocal-tract constriction targets but could possibly be defined by nontrivial linear combinations of these variables. Even more specifically, in order to conform to the functional relationship between articulatory variability and acoustic variability observed in this experiment, they could be defined parsimoniously by those linear combinations that best relate to the effect on relevant acoustic cues, as exemplified by $F3$ in the current /r/ production data. Such targets would be in this case more simply characterized as acoustic.

C. Speaker-specific vocal-tract models

The simulation results shown in this paper also indicate that it is possible to construct simple speaker-specific vocal-tract models approximating the specificities of each subject's speech production apparatus from a limited amount of MRI and acoustic data. We were interested in obtaining a simple characterization of the relationship between articulatory configurations and formant positions for two subjects. The model we used is a purely statistical one defined as a simple linear relation between these variables. Compared to physically based models that estimate the area functions and from this calculate the acoustic characteristics, the linear model presented here provides a purely statistical approximation to the true articulatory–acoustic relationship, and as such it offers only an estimation and description (but not a physical explanation) of the articulatory acoustic relationship. However, it has the advantage of requiring only a relatively small amount of MRI and acoustic data for each subject and not requiring an accurate estimation of the area functions (which poses technical difficulties, e.g., the teeth not being portrayed in MR images). A locally linear approximation between articulatory parameters and formant positions is predicted by perturbation theory (Schroeder, 1967; Fant, 1980). Our preliminary validation analyses (see Sec. II B) suggest that this approximation is appropriate ($R^2 > 0.9$) for a relatively large range of articulatory configurations in our modeled speakers. This implies that approximate speaker-specific vocal-tract models can be estimated using simple linear models with

Nieto-Castanon *et al.*: A modeling investigation of articulatory

minimal demands on the amount of necessary data. A detailed analysis of the general accuracy of these models is beyond the scope of this paper. Nevertheless, for Subject 1, for whom we have redundant degrees of freedom to estimate the level of accuracy of the resulting mapping, a significant linear relation between articulatory and formant descriptors was in fact found (general linear model, $R^2 = 0.90$; $\Gamma_{5,3} = 40.0$; $p < 0.01$; dof$= 4$). The speaker-specific vocal-tract models estimated in this paper are in agreement with standard characterizations of articulatory to acoustic relations (such as the differences between high and low, front and back, tongue configurations) while accommodating the specificities of each subject's vocal tract and their effective articulatory degrees of freedom. We believe the use of subject-specific vocal-tract models, in conjunction with a speaker-independent motor control strategy, is a promising approach to fit the specificities of different subjects' speech movements.

D. Acoustic target model predictions and simulations

Speech motor control models based on acoustic targets posit that the target for production of a phoneme is defined in terms of its acoustic properties, rather than as a specific vocal-tract configuration. In this way the variability in articulator configurations in the production of a given phoneme would reflect the one-to-many relation between the acoustically defined target and the articulatory space (i.e., the range of articulator configurations that are able to produce sounds with equivalent acoustic properties). The DIVA model is an example of such a model. The simulations presented in this paper use this model in conjunction with appropriate speaker-specific vocal-tract models to replicate two of the subjects' articulatory data. Note that while the results of our EMMA study showed evidence of the acoustic specification of /r/ ($F3$ forming part of the production target for /r/), the simulations in this section go beyond that by indicating that an acoustic /r/ target definition (a target defined *only* by acoustic dimensions) can account for the observed data. The simulation results of /r/ production in different leading phonetic contexts (Fig. 7, bottom) mimicked the range of articulatory gestures used by the two subjects being modeled (Fig. 7, top). The correlation between the experimental and modeled tongue gestures was $r = +0.86$ and $r = +0.93$ for Subjects 1 and 2, respectively. Furthermore, the simulated articulatory configurations reached by the DIVA model showed articulatory/acoustic relations (Fig. 8) similar to those found in the experimental data (Fig. 4). In effect, the articulatory variability in the simulations along each articulatory dimension was inversely related to its associated effect on $F3$.

The ability of the DIVA model simulations to fit the specificities of each subject's lingual gestures for the characteristic phonetic contexts tested emphasizes the idea that a relatively wide range of the articulatory variability in /r/ production can be explained by a simple speech motor control scheme using acoustic targets (without the need to appeal to possible multiple articulatory targets). In Fig. 7 top, for example, the tongue tip for each of the subjects moves in different directions for each context, and these directions do not seem to aim at any common lingual configuration. Interest-

ingly, this can be modeled simply as a movement in the articulatory direction that in each case brings the acoustic output closest to a fixed acoustic target. Similarly, as shown by the simulations, the same acoustic target model parsimoniously explains the emergence of predictive relations between acoustic stability and articulatory variability. The expected articulatory/acoustic relation theoretically derived from this model is exemplified in Fig. 8, left (dotted line).

E. Limitations

There are several limitations of this study. First, the study is restricted to the analysis of American English /r/ production. The results presented could only be generalized if the motor control strategy used in speech production, which predicts the emergence of the observed articulatory/acoustic relations, is common across different phonemic targets. Evidence of articulatory trading relations argued to limit acoustic variability in the production of /u/ (Perkell *et al.* 1993) suggests another case where acoustic variables could potentially predict the extent of articulatory variability. It is thus likely that the descriptive ability of the acoustic-target hypothesis generalizes to other vowel and semivowel cases. Whether articulatory- or mixed articulatory/acoustic variables are more instrumental in the description of consonant productions is an issue that could potentially be addressed following a methodology similar to the one presented in this paper. Our expectation would be that the exact nature of the phonemic targets (auditory and/or somatosensory) is learned, and it would depend on the amount of language- and subject-specific allowed variability in these two spaces for that phoneme. Second, the presented articulatory/acoustic relation analyses are restricted to changes in $F3$. While this is an important acoustic cue for /r/ production, it is most probably not the only one. A more complex study showing the form of these relations when multiple acoustic cues are considered could potentially deepen our knowledge on the motor control strategies in speech production. In relation to this issue the simulations presented in this paper use the first three formants as a descriptor of the acoustic /r/ target. The presence of a predictive relationship between $F3$ stability and articulatory variability in the simulations shows that, for these relations to emerge, it is not necessary for the targeted variable to be the sole descriptor of the target coordinate frame. Third, regarding the speaker-specific vocal-tract models, the presented methodology is limited by the linear nature of the analyses involved. The relation between articulatory configurations and the acoustic output is complex. Nevertheless, this relation seems to be well approximated by a linear relation between articulatory and formant descriptors if relatively open configurations (such as vowels and semivowels) are considered. In this way, the validation presented in Sec. II indicates that the appropriateness of the linear model extends for a relatively large proportion of the articulator space (as indicated by the good linear fits between articulatory and acoustic formant descriptors estimated using Maeda's realistic tube model). The proposed speaker-specific vocal-tract models represent a simple first-order approximation to the complexities of the vocal-tract apparatus and the corresponding acous-

J. Acoust. Soc. Am., Vol. 117, No. 5, May 2005

Nieto-Castanon *et al.*: A modeling investigation of articulatory   3209

tic output. This approximation is especially valid for vowels and semivowels. For the production of consonants different strategies should be investigated. Fourth, regarding the DIVA simulations, this paper does not address how the phonemic targets are learned or transferred between subjects, issues still open to further discussion and research. The DIVA simulations for each subject used an acoustically defined target for /r/ based on his/her own productions. In this way we were simply testing the ability of a single acoustic target for each subject to account for the range of articulatory configurations reached in the production of /r/ in different phonetic contexts. It is possible that some sort of speaker normalization allows each speaker to define acoustic targets that are somehow informed of the actual range of acoustic productions that this speaker can produce. This paper does not attempt to address these issues. More detailed analysis of intersubject differences in vocal-tract morphology and its possible relationship with phonemic target specification could provide very relevant information but is beyond the scope of this paper. Finally, the small number of subjects modeled limits our ability to generalize the model's ability to fit the specificities of each subject's articulatory configurations in different phonetic contexts (cf. Westbury *et al.*, 1998, for a large sample analysis of intersubject articulatory variability in /r/). Our expectation would be that the intersubject variability, assuming a speaker-independent motor control strategy, is mainly affected by differences in vocal-tract morphology, and hence could be accounted for by using appropriate speaker-specific vocal-tract models such as the one presented in this paper.

Future studies using speaker-specific vocal-tract models could in this way help better understand the sources of intersubject variability.

V. SUMMARY

The analysis of articulatory movement data on seven subjects during the production of American English /r/ in different phonetic contexts shows a functional relationship between acoustic stability and articulatory variability. This relation indicates that the extent of articulatory variability along any given articulatory dimension is well predicted by the effect that the articulatory dimension has on a relevant acoustic cue ($F3$): most of the articulatory variability present in the production of American English /r/ is concentrated along articulatory dimensions that produce minimal change in $F3$. Both the presence and direction of the observed relationship are consistent with speech motor control mechanisms utilizing an acoustic ($F3$) target representation. In contrast, no significant relationship was found consistently across subjects between hypothesized vocal-tract target representations and articulatory variability. The combined results indicate that if phonemic targets are being used, they do not seem to be simply defined by constriction variables, but as nontrivial linear combinations of them. Such variables are more parsimoniously defined in terms of an acoustic frame of reference.

The second part of this paper investigated the ability of auditory or acoustic target models to explain the specificities of the range of articulatory gestures observed in the produc-

tion of American English /r/. Speaker-specific models capturing the specificities of two subjects' vocal tracts were constructed from a combination of MRI and acoustic data. Simulations of the DIVA model (an example of an acoustic target motor control scheme) controlling each speaker-specific vocal-tract model produced articulatory movements that closely mimic those of each speaker. Furthermore, the articulatory configurations realized by this model exhibit similar articulatory/acoustic relations as those observed in the experimental data. The results demonstrate the ability of motor control speech production models utilizing a purely acoustic target representations to mimic central aspects of the experimental articulatory data on a particular example of speech production.

APPENDIX: DERIVATION OF ARTICULATORY/ACOUSTIC RELATION FROM THE MOTOR CONTROL EQUATIONS OF THE DIVA MODEL

In the DIVA model, the differential equation governing the articulator vector $x(t)$ given an acoustic target vector y takes the form

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{J}^+ \cdot (\mathbf{y} - \mathbf{f}[\mathbf{x}(t)]) - \alpha \cdot \mathbf{\Pi}(\mathbf{J}) \cdot \mathbf{x}(t),$$

where $\mathbf{f}(\mathbf{x})$ represents the articulatory to acoustic mapping, $\mathbf{J}$ represents the Jacobian (the multivariate derivative) of this mapping at each point $\mathbf{x}(t)$, $\mathbf{J}^+$ and $\Pi(\mathbf{J})$ represent its pseudoinverse and its null space projector operator, respectively, and α is a small factor in the model (relaxation factor) controlling the degree of articulatory relaxation toward a neutral configuration (without loss of generality this is assumed to be $\mathbf{x}=0$). Under a linear approximation of the articulatory to acoustic mapping $[\mathbf{f}(\mathbf{x}) = \mathbf{A} \cdot \mathbf{x}]$, and using a regularized form of the pseudoinverse, the explicit form of the previous equation is

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}^t \cdot (\mathbf{A} \cdot \mathbf{A}^t + \mu \cdot \mathbf{I})^{-1} \cdot [\mathbf{y} - \mathbf{A} \cdot \mathbf{x}(t)]$$

$$- \alpha \cdot [\mathbf{I} - \mathbf{A}^t \cdot (\mathbf{A} \cdot \mathbf{A}^t + \mu \cdot \mathbf{I})^{-1} \cdot \mathbf{A}] \cdot \mathbf{x}(t)$$

$$= \mathbf{A}^t \cdot (\mathbf{A} \cdot \mathbf{A}^t + \mu \cdot \mathbf{I})^{-1} \cdot [\mathbf{y} - (1-\alpha) \cdot \mathbf{A} \cdot \mathbf{x}(t)]$$

$$- \alpha \cdot \mathbf{x}(t),$$

where $\mathbf{A}$ is the linear mapping between the articulatory and acoustic spaces, and μ is a small regularization factor of the pseudoinverse. The solution of this differential equation is the articulatory trajectory $\mathbf{x}(t)$

$$\mathbf{x}(t) = \mathbf{x}_0 + (\mathbf{I} - e^{-\mathbf{K} \cdot t}) \cdot (\mathbf{x}_\infty - \mathbf{x}_0),$$

$$\mathbf{K} \equiv (1-\alpha) \cdot \mathbf{A}^t \cdot (\mathbf{A} \cdot \mathbf{A}^t + \mu \cdot \mathbf{I})^{-1} \cdot \mathbf{A} + \alpha \cdot \mathbf{I},$$

where $\mathbf{x}_0$ is the initial articulatory configuration, and $\mathbf{x}_\infty$ is the articulatory configuration that would be reached allowing infinite time ($\mathbf{x}_\infty$ depends on the acoustic target $\mathbf{y}$, and its solution is not relevant to the following discussion). Repeated productions under different initial articulatory configurations will reach, after time T, the articulatory configuration $\mathbf{x}(T)$, following a distribution with average

$$\langle \mathbf{x}(T) \rangle = \mathbf{x}_\infty - e^{-\mathbf{K} \cdot T} \cdot (\mathbf{x}_\infty - \langle \mathbf{x}_0 \rangle),$$

and covariance

$$\mathbf{\Omega}_T = e^{-\mathbf{K} \cdot T} \cdot \mathbf{\Omega}_0 \cdot e^{-\mathbf{K}^t \cdot T},$$

where $\langle \mathbf{x}_0 \rangle$ and $\mathbf{\Omega}_0$ are the average and covariance, respectively, of the initial articulatory configurations. For simplicity, let us assume the distribution of initial articulatory configurations to be normal, with covariance $\sigma_0 \cdot \mathbf{I}$. In this case, the articulatory covariance of the final articulatory configurations takes the form

$$\mathbf{\Omega}_T = \sigma_0 \cdot e^{-2 \cdot \mathbf{K} \cdot T}.$$

Let us, finally, define the vector $\mathbf{q}$ to be any eigenvector of the matrix $\mathbf{\Omega}_T$ (corresponding to one of the articulatory directions resulting from PCA of the final articulatory covariance). The *acoustic effect* of this articulatory direction $\mathbf{q}$ is defined as the associated change in the acoustic vector when moving the articulators along the direction $\mathbf{q}$, and it is computed as $\lambda(\mathbf{q}) \equiv \|\mathbf{A} \cdot \mathbf{q}\|$, and the *articulatory variability* associated with the same articulatory direction $\mathbf{q}$ is computed as $\sigma(\mathbf{q}) \equiv \mathbf{q}^t \cdot \mathbf{\Omega}_T \cdot \mathbf{q}$. Using the definition of the matrices $\mathbf{\Omega}_T$ and $\mathbf{K}$, and noting that their eigenvectors (they are the same for both matrices) will correspond to the right eigenvectors of the matrix $\mathbf{A}$, the articulatory variability $\sigma(\mathbf{q})$ can be expressed, as a function of the acoustic effect $\lambda(\mathbf{q})$, as

$$\sigma(\mathbf{q}) = \sigma_0 \cdot e^{-2 \cdot [(1-\alpha) \cdot [(\lambda^2(\mathbf{q})) / (\lambda^2(\mathbf{q}) + \mu)] + \alpha] \cdot T}.$$

More simply, the articulatory/acoustic relation predicted from the DIVA equations belongs to the class of functions

$$\sigma(\lambda) \propto \varepsilon^{\lambda^2 / (\lambda^2 + \mu)},$$

where ε and μ are two small factors. The dashed line in Fig. 8, left is an example of such a function approximating the simulation results ($\varepsilon = 0.01$; $\varepsilon = 0.001$).

---

[1]The current version of the DIVA model (Guenther, Ghosh, and Nieto-Castanon, 2003) uses a combination of auditory and somatosensory targets. As a result of learning in the model, sounds whose characteristic acoustic signal can be produced with a wide range of articulator shapes end up with primarily auditory targets, while sounds that can only be produced with a consistent somatosensory pattern (e.g., lip tactile information signaling full closure for a bilabial stop) will have both auditory and somatosensory targets. In other words, the model hypothesizes that the exact nature of the target (auditory and/or somatosensory) for a sound will depend on the amount of variability in the two spaces that is allowable for that sound in the infant's native language. In the current article we will deal only with /r/, which we believe to have a primarily auditory target in American English.

[2]The covariance is a multivariate extension of the common univariate concept of variance. Conceptually it characterizes not only the spread or range of each variable but also the level of association between the variables.

Numerically it is defined as a symmetric matrix, and the elements in its diagonal correspond to the variance of each of the individual variables. PCA is a common statistical technique for the characterization of multivariate data. Conceptually it is similar to factor analysis. It offers a decomposition of the data in terms of factors or components that successively comprise most of the data variance, and are, in this sense, most explanatory of the data. If the data are normally distributed, forming a rough ellipsoid in an arbitrary multidimensional space, the resulting principal components correspond to the axes defining this ellipsoid. Numerically it is computed as an eigenvector decomposition of the data covariance matrix. See Mardia *et al.* (1979) for a highly detailed exposition of these and other multivariate concepts.

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

Carrozzo, M., Stratta, F., McIntyre, J., and Lacquaniti, F. (**2002**). "Cognitive allocentric representations of visual space shape pointing errors," Exp. Brain Res. **174**(4), 426–436.

Delattre, P., and Freeman, D. C. (**1968**). "A dialect study of American r's by x-ray motion picture," Linguistics **44**, 29–68.

Fant, G. (**1980**). "The relations between area functions and the acoustic signal," Phonetica **37**(1–2), 55–86.

Guenther, F. H., Ghosh, S.S., and Nieto-Castanon, A. (**2003**). "A neural model of speech production," Proceedings of the 6th International Seminar on Speech Production, Sydney, Australia.

Guenther, F. H., Hampson, M., and Johnson, D. (**1998**). "A theoretical investigation of reference frames for the planning of speech movements," Psychol. Rev. **105**, 611–633.

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**(5), 2854–2865.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modeling*, edited by W.J. Hardcastle and A. Marchal (Kluwer Academic, Boston), pp. 131–149.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (**1979**). *Multivariate Analysis* (Academic, London).

McIntyre, J., Stratta, F., Droulez, J., and Lacquaniti, F. (**2000**). "Analysis of pointing errors reveals properties of data representations and coordinate transformations within the central nervous system," Neural Comput. **2**(12), 2823–2855.

MacNeilage, P. F. (**1970**). "Motor control of serial ordering of speech," Psychol. Rev. **77**(3), 182–196.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M. (**1992**). "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., and Jordan, M. I. (**1993**). "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot 'motor equivalence' study," J. Acoust. Soc. Am. **93**(5), 2948–2961.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., and Jordan, M. I. (**1995**). "Goal-based speech motor control: a theoretical framework and some preliminary data," J. Phonetics **23**, 23–35.

Perrier, P., Boe, L. J., and Sock, R. (**1992**). "Vocal-tract area function estimation from midsagittal dimensions with CT scans and a vocal-tract cast: modeling the transition with two sets of coefficients," J. Speech Hear. Res. **35**(1), 53–67.

Saltzman, C., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological Psychol. **1**, 333–382.

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

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal-tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**(1), 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**(1), 471–487.

Tiede, M., Yehia, H., and Vatikiotis-Bateson, E. (**1996**). "A shape-based approach to vocal-tract area function estimation," 4th Speech Production Seminar/ETRW, 41–44.

Wakita, H. (**1973**). "Direct estimation of the vocal-tract shape by inverse filtering of acoustic speech waveforms," IEEE Trans. Audio Electroacoust. **AU-21**(5), 417–427.

Westbury, J. R., Hashi, M., and Lindstrom, M. J. (**1998**). "Differences among speakers in lingual articulation of American English /r/," Speech Commun. **26**, 203–226.

Nieto-Castanon *et al.*: A modeling investigation of articulatory