1998-01

Articulatory Tradeoffs Reduce Acoustic Variability During American English /r/ Production

Guenther, Frank H.

Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems

http://hdl.handle.net/2144/2336

Boston University
Articulatory Tradeoffs Reduce Acoustic Variability During American English/r/Production

Frank H. Guenther, Carol Y. Espy-Wilson, Suzanne E. Boyce, Melanie L. Mathies, Majid Zandipour, and Joseph S. Perkell

January 1998

Submitted to Journal of the Acoustical Society of America.
Articulatory Tradeoffs Reduce Acoustic Variability During American English /r/ Production

Frank H. Guenther¹,², Carol Y. Espy-Wilson³,², Suzanne E. Boyce³,², Melanie L. Matthies⁴,², Majid Zandipour²,¹, and Joseph S. Perkell²,⁵

Submitted to Journal of the Acoustical Society of America

Address correspondence to:
Prof. Frank H. Guenther
Boston University
Center for Adaptive Systems and
Department of Cognitive and Neural Systems
677 Beacon Street
Boston, MA, 02215
Fax Number: (617) 353-7755
Email: guenther@cns.bu.edu

¹Department of Cognitive and Neural Systems, Boston University.
²Research Laboratory of Electronics, Massachusetts Institute of Technology.
³Department of Electrical and Computer Engineering, Boston University.
⁴Department of Communication Disorders, Boston University.
⁵Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
ABSTRACT

Acoustic and articulatory recordings reveal that speakers utilize systematic articulatory tradeoffs to maintain acoustic stability when producing the phoneme /r/. Distinct articulator configurations used to produce /r/ in various phonetic contexts show systematic tradeoffs between the cross-sectional areas of different vocal tract sections. Analysis of acoustic and articulatory variabilities reveals that these tradeoffs act to reduce acoustic variability, thus allowing large contextual variations in vocal tract shape; these contextual variations in turn apparently reduce the amount of articulatory movement required. These findings contrast with the widely held view that speaking involves a canonical vocal tract shape target for each phoneme.
1. Introduction: The targets of phoneme production

It has long been recognized that the production of a speech sound, or phoneme, involves the generation of an appropriate acoustic signal. In other words, some form of “acoustic target” is either explicitly or implicitly utilized at some level in the speech production process. In the speech production modeling literature, it has also been traditionally assumed that the production of a phoneme involves a target shape of the vocal tract. For example, Henke (1966) and MacNeilage (1970) hypothesized that phoneme production involves the achievement of target spatial positions of the speech articulators. That is, to produce a particular phoneme, the speech motor control system simply moves each articulator toward a target position specific to that phoneme, and when the articulators have all reached their target positions, the vocal tract is in an appropriate shape for producing the phoneme. To explain motor equivalence phonemona such as bite-block speech, Lindblom, Lubker, and Gay (1979) suggested that, instead of the positions of individual articulators, the target for a phoneme is a vocal tract area function that might be achieved with different combinations of the positions of individual articulators. A prominent recent theory of speech movement control, the task-dynamic model (Saltzman and Munhall, 1989), hypothesizes that production involves the achievement of a target set of vocal tract constrictions for each phoneme, and the model of Guenther (1995a) extended this to target regions, rather than points, within a constriction-based reference frame.

According to all of these theories, then, production of a phoneme involves some kind of canonical vocal tract shape target for that phoneme. This viewpoint has also influenced the fields of speech perception and phonology, serving as the basis of the motor theory of speech perception (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman and Mattingly, 1985) and articulatory phonology (Brownman and Goldstein, 1990a,b).

Of course, any viable theory of speech production must account for the fact that some variability occurs in speech and is therefore presumably tolerable for both listener and speaker. Vocal tract shape target theories generally attribute variability in vocal tract shape for a particular speech sound to one of the following sources. First, the motor system may intentionally undershoot movements toward the target shape for a sound, particularly during rapid speech. For vowels, this undershoot is

---

1. This is not to say that the canonical vocal tract shape target necessarily specifies the shape of the entire vocal tract. It is often assumed that only key portions of the vocal tract are actively controlled (e.g., Saltzman and Munhall, 1989; Guenther, 1995a). Furthermore, a phoneme’s target shape may vary with time, e.g. for diphthongs and glides.
often referred to as “vowel reduction”. Second, coarticulation or coproduction (Fowler, 1980; Ohman, 1966, 1967), usually involving portions of the vocal tract that are not actively controlled for the current sound, can deform the vocal tract shape for the current sound away from its canonical target and toward the targets for neighboring sounds. Finally, variability may arise due to unplanned “sloppiness” in the production process. For example, incomplete compensation for inertia during rapid speech may lead to overshoot or undershoot of the desired vocal tract shape.

In contrast to vocal tract shape theories, several recent theoretical considerations of speech motor behaviors have posited that the speech production process may involve only canonical acoustic or auditory targets, without any corresponding vocal tract shape targets\(^2\) (e.g., Guenther, 1995b; Guenther, Hampson, and Johnson, 1997; Perkell, Matthies, Svirsky, and Jordan, 1993, 1995; Perkell et al., 1997; Savariaux, Perrier, and Orliaguet, 1995a; Savariaux, Perrier, and Schwartz, 1995b; see also Johnson, Ladefoged, and Lindau, 1993). A computational approach that explicates how the motor system can produce speech movements utilizing only acoustic targets has been embedded in the DIVA model of speech production (Guenther, 1995b; Guenther et al., 1997). This model utilizes a control scheme that is related to Jacobian pseudoinverse control techniques from the robotics literature (Baillieul, Hollerbach, and Brockett, 1984; Klein and Huang, 1983; Liégeois, 1977). For the current purposes, it suffices to note that a Jacobian pseudoinverse is a matrix that can be used to transform desired velocities in a planning space (e.g., formant velocities) into velocities in the effector space (e.g., articulator velocities) that achieve the desired planning space velocities. A controller that utilizes a particular pseudoinverse, the Moore-Penrose pseudoinverse, will use the least amount of articulatory movement possible to follow a straight line to the target in planning space coordinates (e.g., Klein and Huang, 1983). When moving to the same acoustic target starting from different articulator configurations (such as when the same phoneme is produced in different phonetic contexts), this type of controller will generally end up in different articulator configurations at the point where the acoustic target is reached. In other words, the canonical acoustic target for the phoneme is not associated with any canonical vocal tract shape target; instead, different vocal tract shapes are used.

\(^2\) From this point on, we will use the term “acoustic target theories” to refer to theories positing that the canonical targets of speech production are more closely related to the acoustic signal or its representation in the auditory system than to vocal tract shapes, although some authors have preferred the terms “auditory” or “auditory perceptual” in describing these targets (e.g., Guenther et al., 1997; Savariaux et al., 1995a,b).
shapes will be used to achieve the same acoustic result in different phonetic contexts.

A major difference between constriction and acoustic target theories is that the latter allow for the existence of trading relations between constrictions to achieve stability in the acoustic signal. Assume, for example, that narrowing either of two constrictions at different locations along the vocal tract has the same effect on an important acoustic cue for a phoneme. If the speech motor system utilizes an acoustic target without a corresponding constriction target for that phoneme, then in different phonetic contexts, the motor system could use different combinations of the two constrictions to achieve the same acoustic effect; this would appear as a negative covariance between the sizes of these two constrictions across phonetic contexts. This added flexibility in choosing a vocal tract shape for a sound can allow the motor system to decrease the effort required to move the articulators by utilizing the constriction combination that is most easily achieved in the current phonetic context. As discussed above, pseudoinverse-style controllers that plan movements in acoustic space possess this property. In contrast, constriction theories do not predict systematic tradeoffs between constrictions to reduce acoustic variability.

Recent experiments have investigated the trading relations issue (de Jong, 1997; Perkell et al., 1993; Perkell, Mathies, and Svirsky, 1994; Savariaux et al. 1995a), but the results have been inconclusive: some subjects behave as expected according to acoustic target theories, while others do not. A possible reason for this is that these studies have primarily concentrated on one hypothesized trading relationship, and subjects who do not utilize this trading relation may be using other, unanalyzed trading relations to reduce acoustic variability. For example, Perkell et al. (1993) investigated an hypothesized trading relation between lip rounding and tongue body raising for the vowel /u/. Three of four subjects showed weak trading relations, but the fourth subject showed the opposite pattern. This fourth subject may have been using other trading relations that overrode the effect of the lip rounding/tongue body raising relationship.

In the current study, we employ analysis procedures that allow us to assess the combined effects of multiple articulatory covariances on the variability of the acoustic signal. The American English phoneme /r/ was chosen for this study because it has often been associated with relatively large amounts of articulatory variability (Alwan, Narayanan, and Haker, 1997; Delattre and Freeman, 1968; Espy-Wilson and Boyce, 1994; Hagiwara, 1994, 1995; Ong and Stone, 1997; Westbury, Hashi, and Lindstrom, 1995). An interesting feature of this phoneme is that the endpoints of the articulatory continuum for /r/ can be analyzed as
functionally different articulator configurations that use different primary articulators (tongue tip vs. tongue dorsum). These endpoints have been characterized in the literature as "bunched" (using the tongue dorsum) and "retroflexed" (using the tongue tip). At the same time, the primary acoustic cue is relatively simple and stable: a deep dip in the trajectory of the third spectral energy peak of the acoustic waveform, or third formant frequency (F3) (Boyce and Espy-Wilson, 1997; Delattre and Freeman, 1968), and no consistent acoustic difference between bunched and retroflexed /r/’s has been discovered. The existence of very different articulator configurations for /r/, often within the same subject, is in itself problematic for vocal tract shape target theories, although one might still claim that these different configurations are just rather extreme variations around a single vocal tract shape target. According to this viewpoint, the articulatory variability results from incomplete, sloppy, or blended movements toward a canonical vocal tract shape target, and trading relations that maintain acoustic stability would not be expected.

2. Methods

2.1. Data collection

An electromagnetic midsagittal articulometer (EMMA) system (Perkell et al., 1992) was used to track the movements of six small transducer coils attached to the tongue, lips, and lower incisor. The current study focused on the positions of the three tongue transducers, which are located approximately 1, 2.5, and 5 cm back from the tongue tip with the tongue in a neutral configuration. Each of seven subjects produced 4-6 repetitions of the carrier phrase “Say _____ for me” for each of the five test utterances /warav/, /wagrav/, /wadrav/, /wavrav/, and /wabrav/. Acoustic data were collected simultaneously. The articulatory and acoustic data were time-aligned to allow direct comparisons between the two data types.

2.2. F3 extraction and alignment

The minimum F3 value during /r/ production, corresponding to the acoustic “center” of /r/, served as a landmark for time-alignment of the data across utterances for each speaker. Formant tracks were computed for all utterances using the ESPS/WAVES formant tracker and a 51.2 ms window and 3.2 ms frame rate. The F3 minimum was detected using an automatic procedure that first identified all sonorant regions, then located the point of minimal F3 from the relevant sonorant regions. F3 values and transducer positions within a 150 ms time window centered at the F3 minimum were extracted. Extracted F3 traces for some utterances were corrupted due to technical difficulties in automatically tracking
low amplitude and low frequency values of F3 after stop consonants. Therefore, utterances whose F3 tracks changed by more than 200 Hz in a 3.2 ms time step were eliminated from the study, leaving 12 to 27 analyzed utterances per subject.

2.3. Effects of vocal tract shape parameters on F3

The vocal tract shape for /r/ involves a tongue constriction in the anterior third of the tract. We take the resonance of the front cavity to be the third formant frequency, F3 (Stevens, 1997). Acoustic theory predicts that the frequency of this resonance can be decreased by tongue movements that: (i) Lengthen the front cavity and/or (ii) increase the acoustic mass of the oral constriction behind the front cavity by either decreasing the area of the constriction or lengthening the constriction. The predicted effects of these movements on F3 were confirmed using vocal tract area functions derived from structural MRI scans of a speaker producing /r/.

Two area functions were derived: one representing a “bunched” /r/ configuration, and one representing a “retroflexed” /r/ configuration. Three manipulations were carried out on each area function to test the effects on F3 predicted from acoustic theory: (i) the tongue constriction was extended dorsally by narrowing the vocal tract area immediately behind the constriction, (ii) the front cavity was lengthened by displacing the tongue constriction dorsally, and (iii) the vocal tract area at the tongue constriction was decreased. For all three manipulations, an acoustic signal was synthesized (using S. Maeda’s VTCALCS program; Maeda, 1990) and compared to the signal synthesized from the original area function. Each manipulation resulted in a lower F3 in both the bunched and retroflexed /r/ cases, as expected from the acoustic theory analysis.

Because all three manipulations act to lower F3, subjects could maintain a relatively invariant F3 despite vocal tract shape variations if these variations involved tradeoffs between the different manipulations. That is, the following strategies could be used to maintain a relatively invariant F3 across utterances while allowing variations in vocal tract shape: (i) a tradeoff between constriction length and front cavity length, (ii) a tradeoff between constriction length and constriction area, and (iii) a tradeoff between front cavity length and constriction area.

---

3. Although the direction of the effect on F3 for all of these vocal tract shape manipulations should be the same, the magnitude of the effect on F3 will be different for each manipulation. We account for this by including the magnitudes of the effects when analyzing the combined acoustic effect of these movements; this is done through the $A_i$ terms in Equations 1-3 below.

4. The vocal tract area functions were provided by Abeer Alwan and colleagues from the Electrical Engineering Department at the University of California, Los Angeles.
2.4. Predicted articulatory covariances

For tongue configurations during /r/ production, a backward movement of the tongue tip transducer generally corresponds to a lengthening of the front cavity, an upward movement of the tongue tip transducer generally corresponds to a decrease in the area of the constriction for /r/, and, since the point of maximal constriction for /r/ is typically anterior to the tongue back transducer, an upward movement of the tongue back transducer generally corresponds to a lengthening of the tongue constriction and possibly a decrease in the area of the constriction. This indicates that the three trading relation strategies described above should be evidenced by the following articulatory covariances: (i) a positive covariance between tongue back height and tongue tip horizontal position, (ii) a negative covariance between tongue back height and tongue tip height, and (iii) a positive covariance between tongue tip horizontal position and tongue tip height. Note that the use of all three trading relations by a single subject is unlikely given that they impose competing constraints; i.e., if tongue back height and tongue tip horizontal position are positively correlated as in relation (i), and tongue tip horizontal position and tongue tip height are positively correlated as in relation (iii), it is very likely that tongue back height and tongue tip height will also be positively correlated, thus violating relation (ii).

2.5. Analysis of articulatory and acoustic variances

To quantify the combined effects of articulatory covariances on F3 variability, an analysis was performed using both acoustic and articulatory data to estimate F3 variance as a function of articulatory variances. The relationship between transducer coordinates and F3 during /r/ can be written for each speaker as follows:

\[ F3 = A_0 + \sum_{i=1}^{N} A_i c_i + E \]  \hspace{1cm} (1)

where the \( A_i \) are constants, the \( c_i \) are the transducer coordinates, \( N \) is the number of transducer coordinates considered in the analysis, and \( E \) is a residual term that accounts for the effects on F3 due to all other sources, including articulators not included in the analysis, measurement errors, and nonlinearities in the relationship between F3 and the transducer coordinates. The equation relating F3 variance to articulatory variances at each point in time is then:
\[ \text{Var}(F3) = \sum_i A_i^2 \text{Var}(c_i) + \text{Var}(E) + 2 \sum_{i<j} A_i A_j \text{Cov}(c_i, c_j) + 2 \sum_i A_i \text{Cov}(c_i, E). \quad (2) \]

Note that this equality is exact for the measured F3 and transducer coordinates due to the inclusion of the \( E \) term in the variance calculation. To determine the effects of articulatory covariances on F3 variability, we can compare the variance estimate of Equation 2 to the following variance estimate that excludes the covariances between the analyzed transducer coordinates:

\[ \text{Var}(F3) = \sum_i A_i^2 \text{Var}(c_i) + \text{Var}(E) + 2 \sum_i A_i \text{Cov}(c_i, E). \quad (3) \]

If the F3 variance estimate in the absence of articulatory covariances (Equation 3) is significantly larger than the variance estimate including the articulatory covariances (Equation 2), we conclude that the primary effect of the articulatory covariances is a reduction in the variance of F3.

Strictly speaking, a comparison of the F3 variance estimates in Equations 2 and 3 tells us only about the effects of the covariances of the linear component of each transducer's relation to F3. However, the relationship between F3 and transducer coordinates should be linear near a particular configuration of the vocal tract, since F3 is presumably a continuous nonlinear function of the vocal tract area function, and such functions are locally linear. One would further expect that the relationship is still approximately linear for the relatively limited range of vocal tract configurations utilized by a particular subject for /r/. The linear approximations reported below captured approximately 80% of the variance when using only three pellet coordinates, providing support for the assertion that the primary effect of articulatory covariances on F3 variance can be captured by considering only the linear component of each transducer's relationship to F3. Furthermore, the sign of an articulatory covariance's contribution to F3 variance depends only on the sign of the corresponding \( A_i \) terms, and we are primarily interested in the sign of the combined effects of articulatory covariances on F3 variance. The expected signs of the \( A_i \) for tongue back height, tongue front horizontal position, and tongue front height can be deduced from acoustic theory considerations (Sections 2.3 and 2.4). \( A_i \) values were estimated for each subject using multiple linear regression on the acoustic and articulatory data. All 21 estimated \( A_i \) values (3 values for each of 7 subjects) were of the sign expected from these acoustic theory considerations.
3. Results

3.1. Temporal progression of tongue shapes

Inspection of the temporal progression of tongue shapes during /r/ indicates that contextual shape variations were not simply the result of incomplete (or "blended") attempts to reach the same vocal tract shape target, as illustrated by the sample tongue movements in Figure 1. The tongue position at an instant in time is illustrated by a line connecting the positions of the three tongue transducers at that instant. Each panel shows the tongue shapes at the acoustic center of /r/ (solid line) and 75 ms before and after the acoustic center (dashed lines); dashed arrows indicate the progression of the tongue shape through time. The top half of the figure shows sample movements from Subject 1. This subject used a pronounced upward movement of the tongue tip for /r/ in /warav/; this was the most commonly used articulation for /r/ across subjects and contexts. If this movement was aimed at a canonical tongue tip target for /r/, then one would expect movements toward the same target in different phonetic contexts. For /wagrav/, however, the tongue tip is not raised for /r/ even though it starts out well below the tongue tip height for /r/ in /warav/. Similarly, Subject 2 (bottom) does not raise the tongue tip for /r/ in /wadrav/ even though it is well below the tongue tip height for /r/ in /warav/. Similar patterns were seen in other subjects and contexts; e.g., Subjects 6 and 7 lowered or maintained the height of the tongue tip for /r/ in /wagrav/ even though it started out well below the tongue tip height for /r/ in /warav/, and Subject 4 lowered the tongue tip for /r/ in /wadrav/ even though it started out well below the tongue tip height for /r/ in /wagrav/.

3.2. Tongue shapes at acoustic center of /r/

Figure 2 shows tongue configurations at the F3 minimum of /r/ for each of the seven speakers. For each utterance, the three tongue transducer positions are connected by a straight line. The tongue configurations for all repetitions in all phonetic contexts are superimposed for each speaker. Thus, the fact that different numbers of utterances were analyzed for different subjects and contexts is reflected in this figure. As previously reported elsewhere (e.g., Delattre and Freeman, 1968; Hagiwara, 1994, 1995; Ong and Stone, 1997; Westbury, Hashi, and Lindstrom, 1995), a wide range of tongue shapes is seen both within and across subjects. Also of note is the fact that, although most subjects seem to use an approximate continuum of tongue shapes (e.g., S2, S3, S6, and S7), others show a more bimodal distribution of tongue shapes (e.g., S4, S5). Guenther et al. (1997) describe how both of these patterns can be accounted for by an acoustic target model of phoneme production. Finally, the tongue shapes across subjects appear to
FIGURE 1. Sample tongue movements indicating that a single tongue tip constriction target cannot account for /r/ articulations in different phonetic contexts. The tongue position at an instant in time is illustrated by a line connecting the positions of the three tongue transducers at that instant. Each panel shows the tongue shapes at the acoustic center of /r/ (solid line) and 75 ms before and after the acoustic center (dashed lines); dashed arrows indicate the progression of the tongue shape through time. Subject 1 (top) used a pronounced upward movement of the tongue tip for /r/ in /warav/; this was the most commonly used articulation for /r/ across subjects and contexts. If this movement was aimed at a canonical tongue tip target for /r/, then one would expect movements toward the same target in different phonetic contexts. For /wagrav/, however, the tongue tip is not raised for /r/ even though it starts out well below the tongue tip height for /r/ in /warav/. Similarly, Subject 2 (bottom) does not raise the tongue tip for /r/ in /wadrav/ even though it is well below the tongue tip height for /r/ in /warav/.
form an approximate continuum between a bunched configuration (e.g., S6) and a retroflexed configuration (e.g., S4).

Figure 3 shows the midsagittal palatal outline (thick solid line) and mean tongue shapes at the time of the F3 minimum for /r/ for each of the seven subjects. For each subject, mean configurations from two phonetic contexts (solid and dashed lines) are shown to illustrate the range of tongue shapes used by that subject. Tongue outlines were created by connecting the average positions of the three tongue transducers for a given utterance with a smooth curve. A line was then extended downward from the tongue tip transducer position, then forward to the lower incisor transducer position, to provide a rough estimate of the relative size of the front cavity across contexts. Also shown in the upper left corner of this figure are two superimposed, highly schematic vocal tract outlines that illustrate trading relations for maintaining a relatively stable F3. The effect on F3 of the longer front cavity of the dashed outline is counteracted by the effects of the longer and slightly narrower constriction of the solid outline. Similarly, the vocal tract outlines for all subjects indicate that shorter front cavity lengths are accompanied by a compensating increase in tongue constriction length and/or decrease in the constriction area. Furthermore, the tongue shapes during /wagrav/ (solid lines) are generally much closer in shape to tongue shapes for /g/ than are the /r/ shapes for /wabrav/ or /warav/ (dashed lines), suggesting that subjects utilize /r/ configurations that are reached relatively easily in the current phonetic context.

3.3. Articulatory trading relations

For each subject, Pearson correlation coefficients corresponding to the predicted covariances described in Section 2.4 were estimated across utterances at the point of F3 minimum and are listed in Table 1. All subjects showed a significant positive correlation between tongue back height and tongue tip horizontal position, indicative of a trading relation between constriction length and front cavity length. Six of seven subjects also showed a second strong trading relation, either between constriction length and constriction area (five subjects) or front cavity length and constriction area (one subject). One subject showed only very weak correlations other than the strong trading relation between tongue back height and tongue tip horizontal position.

---

5. The lower incisor location was not available for Subject 2, so the vocal tract outlines for this subject in Figure 3 are based on a lower incisor position estimated from the lower lip position. Furthermore, the palatal trace for this subject was slightly misaligned relative to the tongue transducer data. To correct for this, the palatal trace for this subject in Figure 3 has been raised approximately 3 mm relative to the tongue transducer positions.
FIGURE 2. Tongue configurations at the F3 minimum of /r/ for each of the seven speakers. For each utterance, the three tongue transducer positions are connected by a straight line. The tongue configurations for all repetitions in all phonetic contexts are superimposed for each speaker.
FIGURE 3. Trading relations during /r/ production. The upper left corner shows two superimposed, highly schematized vocal tract outlines (dashed and solid lines) illustrating trading relations between front cavity length and tongue constriction length and area. Also shown are vocal tract outlines that illustrate the range of tongue shapes used by each of the seven subjects to produce /r/ in different phonetic contexts. Thin solid lines correspond to the tongue shapes for /r/ in /wagrav/ (averaged across repetitions), and dashed lines correspond to the /r/ in /wabrav/ or /warav/, depending on the subject. Thick solid lines indicate palatal outlines. Each outline is formed by connecting the three tongue transducer positions with a smooth curve, then projecting downward and forward from the tongue tip transducer to the lower incisor transducer. All seven subjects show tradeoffs between the front cavity length and the constriction length and/or area when producing /r/ in the two different contexts.
Table 1. Articulator correlation coefficients. Significant correlations that are consistent with hypothesized trading relations are shown in boldface. TBY = tongue back height, TTX = tongue tip horizontal position, TTY = tongue tip height.

<table>
<thead>
<tr>
<th>Subject</th>
<th>TBY-TTX</th>
<th>TBY-TTY</th>
<th>TTX-TTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.77*</td>
<td>-0.76*</td>
<td>-0.69*</td>
</tr>
<tr>
<td>2</td>
<td>0.92*</td>
<td>-0.69*</td>
<td>-0.88*</td>
</tr>
<tr>
<td>3</td>
<td>0.77*</td>
<td>-0.74*</td>
<td>-0.46</td>
</tr>
<tr>
<td>4</td>
<td>0.86*</td>
<td>0.64*</td>
<td>0.77*</td>
</tr>
<tr>
<td>5</td>
<td>0.64*</td>
<td>-0.49*</td>
<td>-0.57*</td>
</tr>
<tr>
<td>6</td>
<td>0.55*</td>
<td>-0.81*</td>
<td>-0.60*</td>
</tr>
<tr>
<td>7</td>
<td>0.84*</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>

* Statistically significant (p < 0.01).

3.4. Analysis of acoustic and articulatory variabilities

The results in Section 3.3 indicate that most subjects utilized two of three hypothesized articulatory trading relationships to reduce acoustic variability. Furthermore, as described in Section 2.4, it is unlikely or impossible for a subject to utilize all three trading relations because they counteract one another. However, it is still possible that the significant correlations that violate the trading relations could effectively “override” the beneficial articulatory tradeoffs, potentially nullifying or even reversing the effect of the utilized trading relations on acoustic variability. It is therefore necessary to estimate the net effect of all three articulatory covariances, as outlined in Section 2.5.

F3 variance estimates with and without covariance terms (Equations 2 and 3, respectively) were calculated using the tongue back height, tongue tip horizontal position, and tongue tip height transducer coordinates. The corresponding F3 standard deviations were then averaged across subjects. The $A_i$ values for each speaker were estimated using multiple linear regression across utterances and time bins and are provided in Table 2; the value of $E$ for a particular time bin was simply the residual of the regression in that time bin. $R^2$ values for the F3 fit (without the residual term) ranged from 0.75 to 0.85 for the different subjects, with an average $R^2$ of 0.79. If covariances are high and the actual effect of an articulator’s position on F3 is very low, the regression analysis can possibly result
in estimates of transducer contributions that have the wrong sign, which could in turn cause some articulatory covariances to decrease estimated F3 variability when in reality they increase or have no significant effect on F3 variability. The fact that none of the transducer contribution estimates produced by the regression were of the opposite sign as expected from acoustic theory considerations and the tube model analysis indicates that this potential problem did not affect our results.

Table 2. Regression coefficients indicating the relationship between transducer coordinates and F3. Units are Hz/mm.

<table>
<thead>
<tr>
<th>Subject</th>
<th>A1 (TBY)</th>
<th>A2 (TTX)</th>
<th>A3 (TTY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-27.54</td>
<td>19.38</td>
<td>-33.44</td>
</tr>
<tr>
<td>2</td>
<td>-81.13</td>
<td>92.77</td>
<td>-35.25</td>
</tr>
<tr>
<td>3</td>
<td>-12.28</td>
<td>25.93</td>
<td>-51.50</td>
</tr>
<tr>
<td>4</td>
<td>-11.71</td>
<td>55.14</td>
<td>-56.93</td>
</tr>
<tr>
<td>5</td>
<td>-24.04</td>
<td>21.44</td>
<td>-30.49</td>
</tr>
<tr>
<td>6</td>
<td>-21.68</td>
<td>10.99</td>
<td>-30.95</td>
</tr>
<tr>
<td>7</td>
<td>-46.29</td>
<td>33.71</td>
<td>-31.87</td>
</tr>
</tbody>
</table>

Figure 4 shows the results as a function of time starting at the F3 minimum for /r/, averaged across subjects. (Standard deviations were plotted in place of variances to produce values whose units are Hz.) Also plotted is the standard deviation obtained from measured values of F3 (solid line). When articulatory covariances are included, the F3 standard deviation estimates are equal to the measured F3 standard deviation; this is as expected because of the inclusion of the residual term in the variance estimate calculations. When articulatory covariances are removed from the estimates, however, the estimated F3 standard deviation increases substantially. The dashed line in Figure 4 represents estimated F3 standard deviation without covariances using the three tongue transducer coordinates. According to this estimate, then, F3 standard deviation would be 105% higher if the motor system did not utilize articulatory tradeoffs.

The increase in the F3 variance estimate without covariances is seen at the F3 minimum for all subjects. This suggests that the ambiguous results from previous studies may have been at least partly due to analyzing only one articulatory tradeoff at a time, since in our study no subject used all three hypothesized trading
relations, but all subjects showed a net decrease in acoustic variability due to the combined effects of the articulatory covariances. Assume, for example, that the data listed in each column of Table 1 were the result of an independent research study. Researchers investigating the trading relation in column 2 (in which 5 of 7 subjects used the trading relation) would sharply disagree with researchers investigating the trading relation in column 3 (in which only 1 of 7 subjects used the trading relation) as to whether or not trading relations are reliably used, and both sets of researchers would report ambiguous results since in neither case do all subjects behave in the same way. A much clearer picture has emerged from the current study due to the analysis of the combined effects of the articulatory covariances.

Also evident in Figure 4 is a steady decrease with time of the effects of the covariance terms on F3 as the /r/ transitions into the following /a/. This decrease, evident in six of the seven subjects, is suggestive of a decrease in the use of trading relations as vocal tract shape differences across utterances (due to the different phonemes preceding /r/ in different utterances) diminish.

4. Discussion

Together, the results in this report paint a clear picture of the speech motor control system utilizing systematic tradeoffs between the shapes of different parts of the vocal tract to achieve a stable acoustic end while allowing a large amount of variation in the positions of individual articulators. This "acoustic target" view of the speech production process differs from the traditional view that speech production involves some sort of canonical vocal tract shape target for each phoneme, such as the locations and degrees of key constrictions in the vocal tract. The first piece of evidence against the traditional view is that, although the most common movement for /r/ is an upward and backward movement of the tongue tip, thus suggesting a high tongue tip constriction target for /r/ according to vocal tract shape target theories, the tongue tip movements for /r/ in other phonetic contexts are often inconsistent with a high tongue tip target (Section 3.1). Although unattractive from the viewpoint of parsimony, one might propose that subjects use two different vocal tract shape targets for /r/, choosing between them in different phonetic contexts. However, this is only consistent with the tongue shapes across contexts for a minority of subjects who show a roughly bimodal distribution of tongue shapes; most subjects instead show an approximate continuum of tongue shapes for /r/ across contexts (Figure 2). An acoustic target theory can account for both patterns (Guenther et al., 1997). The next piece of evidence for the acoustic target viewpoint arises from inspection of the tongue shape extremes for each subject, which show clear tradeoffs between the length of the front cavity and the
length and/or area of the tongue constriction (Figure 3). These tradeoffs would be expected to reduce acoustic variability across contexts despite large variations in vocal tract shape and are predicted by acoustic target theories but are not accounted for by vocal tract shape target theories. Analysis of articulatory covariances indicates that most speakers use two of three hypothesized articulatory trading relationships to reduce acoustic variability (Section 3.3). Furthermore, the use of all three trading relationships is very unlikely given that they counteract each other. Finally, analysis of the combined effects of these articulatory covariances indicates that they strongly influence F3 variability across contexts, effectively cutting F3 standard deviation in half compared to what it would have been without the articulatory covariances (Section 3.4; Figure 4).

A likely reason for the use of articulatory tradeoffs is that they can reduce the amount of effort required to move the articulators through a set of acoustic targets.
For example, the tongue shapes for /r/ in /wagrav/ were generally closer to the tongue shapes for /g/ than the tongue shapes for /r/ in other contexts, suggesting that, to a first approximation, subjects move to the closest vocal tract shape that can be used to produce the appropriate sound in the prevailing phonetic context. As discussed in the introduction, a controller that uses an acoustic or auditory planning space and a pseudoinverse-style control scheme, such as the DIVA model of speech production (Guenther, 1995b, Guenther et al., 1997), possesses this property. Another desirable property of pseudoinverse-style controllers is that they are capable of immediate compensation for constraints on the speech articulators such as a bite block. Such compensation occurs automatically; i.e., it does not require any experience or learning with the constraints. This property is treated at length elsewhere (Guenther, 1994, 1995a,b; Guenther et al., 1997).

Unlike earlier trading relations studies that reported mixed results across subjects (e.g., de Jong, 1997; Perkell et al., 1993; Perkell, Matthies, and Svirsky, 1994; Savariaux et al. 1995a), the reduction of F3 variability due to articulatory covariances was seen at the acoustic center of /r/ for all seven subjects in the current study. We believe that the following factors contributed to this difference. First, the current study investigated a phoneme known to exhibit a large amount of articulatory variability across contexts. Such a sound would be expected to exhibit stronger trading relations due to the larger overall articulatory variability. Second, the current study investigated the combined effects of multiple articulatory covariances. Although the combined effect of articulatory covariances was a reduction of F3 variability in all seven subjects according to the analysis of Section 3.4, different subjects used different combinations of the individual articulatory trading relations (Table 1). It is therefore not surprising that in earlier studies, which investigated articulatory covariances individually, some subjects did not use an hypothesized trading relationship. The results of the current report suggest that these subjects may well have used other, unanalyzed trading relations that reduced acoustic variability.

5. Acknowledgements

We thank Ken Stevens and Dan Bullock for their help and comments, Shinji Maeda for the VTCALCS program, and Abeer Alwan and colleagues for the MRI data. Supported by NIDCD grants 1R29-DC02852-02 to Frank Guenther, 5R01-DC01925-04 to Joseph Perkell, and 1R03-C2576-01 to Suzanne Boyce, and NSF grant IRI-9310518 to Carol Espy-Wilson. Frank Guenther is also supported by the Alfred P. Sloan Foundation.
REFERENCES


