Production Allophones of North American English Liquids

Mark Tiede[1], Suzanne Boyce[2], Michael Stern[3], Teja Rebernik[4], Martijn Wieling[4]

[1]*Dept. of Psychiatry, Yale University & Haskins Laboratories, New Haven, USA*
[2]*Dept. of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, USA*
[3]*Dept. of Linguistics, Yale University, New Haven, USA*
[4] *Dept. of Information Science, University of Groningen, Groningen, Netherlands*

mark.tiede@yale.edu, boycese@ucmail.uc.edu, michael.stern@yale.edu, t.rebernik@rug.nl, m.b.wieling@rug.nl

Abstract

The syllabic liquids [ɚ] (as in "purr") and [əɫ] (as in "pull") have well-defined acoustic targets but are produced with a wide range of heterogenous tongue postures. This work surveys midsagittal tongue shapes from a large (N=78) number of speakers producing these sounds, to illustrate their variety, and to determine systematically how this variety can be quantified. In particular we propose that a categorization based on just two parameters—degree of tongue dorsum convexity and tip orientation—is sufficient to classify observed shapes, and superior to defining *ad hoc* prototypes.

Keywords: rhotics, liquids, speech production, MRI, ultrasound

1. Introduction

The North American English (NAE) syllabic liquids /r/ [ɚ] (as in "purr") and velarized (dark) /l/ [əɫ] (as in "pull") form a natural class phonologically and phonetically by traditional acoustic criteria; however, they show a high degree of production variability across speakers (Delattre & Freeman, 1968; Westbury et al., 1998; Mielke et al., 2016). The multiple attested articulatory variants of /r/ in particular converge on a perceptually equivalent acoustic profile with F1 and F2 characteristic of a central vowel and an F3 at 80% or less of the 3rd natural resonating frequency of the vocal tract (Hagiwara, 1995; Espy-Wilson et al., 2000). Laterals are similar but with F3 shifted in the opposite direction.

Broadly speaking, both /r/ and /l/ variants have been grouped into tip-down ('bunched'/laminal) and tip-up ('retroflex'/apical) categories. While some modeling evidence for /r/ suggests F4 differences between these types (Zhou et al., 2008), no perceptual data exist showing that listeners are able to distinguish exemplars of these two production allophones reliably (see e.g. Twist et al. 2007 for a representative null result). Other continuants with production variants typically show consistent acoustics maintained over a smoothly varying range of motor equivalent "trading relations": /u/ for example can be produced with a consistent formant pattern by manipulating the extent of lip protrusion *vs*. laryngeal lowering. /r/ is unusual in that no comparable trading relations exist providing a smooth transition from one postural type to the other, raising questions of how many types exist, how speakers learn their preferred posture, and whether the production goal is driven by an auditory or proprioceptive target. Here we use data scanned using MRI and midsagittal ultrasound from a range of speakers producing NAE syllabic /r/ and /l/, to survey their production variety, and to support a new approach for their categorization.

2. Methods

2.1. Participants

Midsagittal imaging data were collected from two non-overlapping cohorts during production of syllabic /r/ and /l/. The first group was imaged in supine posture using magnetic resonance imaging (MRI) at the University Hospital of the University of Cincinnati. The second group was imaged by stabilized ultrasound in sitting posture using the facilities of the mobile SPRAAKLAB (Wieling et al. 2023). In total 78 speakers (39F) provided the data surveyed here, ranging in age from 16 to 68 (mean 34.8, s.d. 12.6).

2.1.1. MRI

29 native NAE speakers (10F) were scanned with 5 mm slice thickness and 128x128 voxels (1.07 pixel/mm resolution) using midsagittal MRI. Speakers were instructed to produce "purr" or "pull" and to sustain the liquid during the 1.2 s scan duration. Speaker audio recorded immediately prior to and following scanning was used to confirm achievement of the expected acoustic target. All provided informed consent and were compensated for their participation.

2.1.2. Ultrasound

To increase power, an additional 70 Dutch speakers were recorded in SPRAAKLAB producing five repetitions of (English) "purr" and "pull" with midsagittal ultrasound during the 2022 Noorderzon Festival (Groningen) using the UltraFit probe stabilizer (Spreafico et al., 2018), recorded with synchronized audio by AAA software (Articulate Instruments). The imaging frame of 720x540 pixels mapping 4.7pixels/mm was recorded at 82 frames/sec. Speakers provided informed consent but were uncompensated volunteers. Following review by two native English listeners 21 of these participants were excluded for inconsistency across repetitions or productions that did not achieve native formant targets, retaining 49 speakers (29F, 1 Other).

2.2. Analysis

2.2.1. MRI-specific

Midsagittal tongue shapes for /r/ and /l/ were obtained by fitting a thin plate spline to the lingual surface, from the top of the epiglottis to the anterior-most point of the apex. Four landmarks were identified along the distal vocal tract wall (base of the pharynx, anterior apex of the second vertebra, highest visible point of the palatal vault, and base of the alveolar ridge), and used to define a semipolar grid to 'unwrap' the tract (Figure 1). Distance functions sampled along gridlines were parameterized as the sum of the first three

coefficients from a Fourier transform (Liljencrants, 1971). Unsupervised *k*-means clustering using elbow and silhouette heuristics was used to determine optimal group separation, addressing the question of how many distinct classifications of /r/ and /l/ were present in this data sample.

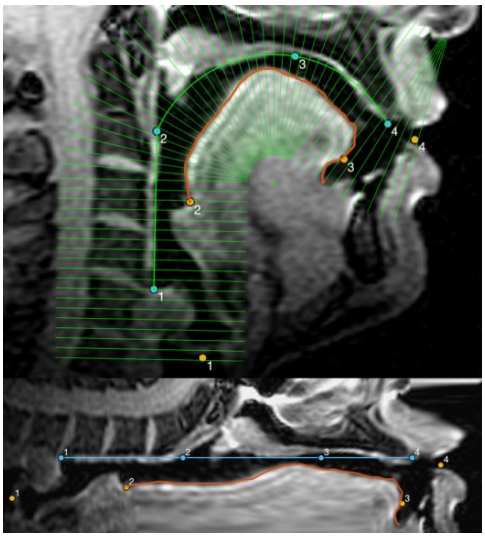

Figure 1: *Semipolar grid (above) used to sample vocal tract distance function along 'unwrapped' tract (below).*

2.2.2. *Ultrasound-specific*

Phone segmentations of the productions of "pull" and "purr" were identified using the Montreal Forced Aligner (McAuliffe et al., 2017). Tongue surface contours at the centers of these acoustically determined liquid intervals were extracted from the ultrasound video using DeepEdge (Chen et al., 2020). Three consecutive frames were averaged for each repetition, and these averages were in turn averaged across repetitions by speaker.

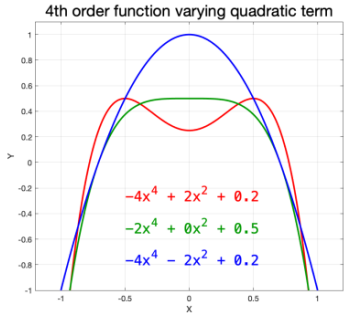

Figure 2: *Illustration of how quadratic coefficient (C2) tracks convexity: >0 concave; ~0 flat; <0 convex.*

2.2.3. *Tongue shape*

The 58 speaker tongue shapes obtained from the MRI and ultrasound cohorts for /r/ and /l/ were normalized as follows:

- resampled to an equal number of mm-based coordinates
- fitted with an ellipse enclosing 95% of all coordinates
- rotated such that the major axis of this ellipse was aligned with the horizontal coordinate axis
- 'curled-under' points at the beginning and end were trimmed (to ensure horizontal monotonicity)
- centered on the midpoint of the ellipse major axis and scaled by its length

This procedure resulted in a tongue shape *y* expressed as a function of *x* for each contour, which was parameterized by a least-squares fit to a 4th order polynomial (higher orders improved the fit but did not significantly affect the quadratic term). In addition, the rotation and scaling factors provide indices of speaker vocal tract morphology. As illustrated in Figure 2, the quadratic coefficient (C2) of this polynomial tracks the degree of convexity of the fit, and as such provides a useful characterization of tongue dorsum shape: concave shapes (bowed down/inward) have positive sign, flat shapes are close to zero, and convex shapes (bowed up/outward) have negative sign.

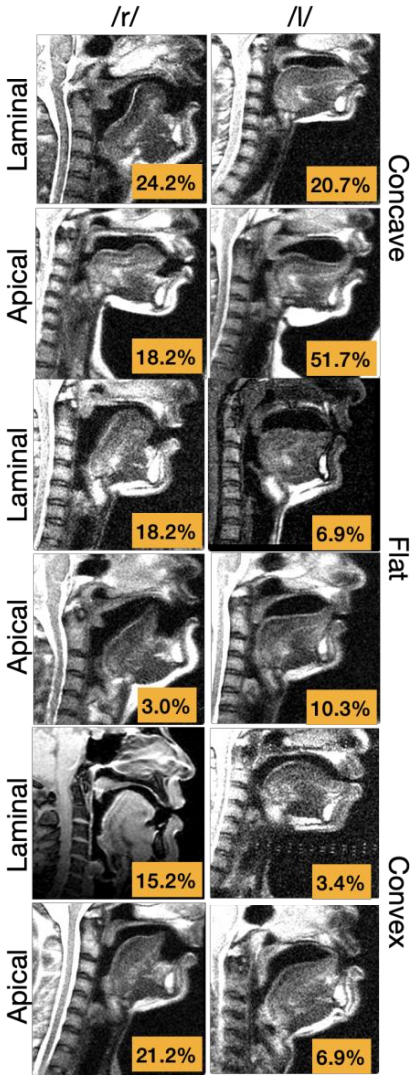

Figure 3: *Representative MRI tongue shapes for syllabic liquids showing apical and laminal tongue tip variants for concave, flat and convex tongue dorsum postures. Insets show percentage of observed speakers with that shape.*

The orientation of the tongue tip was determined with a similar parameterization. The anterior-most 30% of the original extracted tongue shapes were fitted by an enclosing ellipse, rotated, and scaled as above, though retaining non-monotonic points. The average rotation in this case is

approximately 90° CCW. Aligned in this way, the center of gravity (COG) of the polygon determined by the scaled and rotated points has negative sign in *x* for tip-up ('retroflex'/apical), and positive sign for tip-down ('bunched'/laminal) tongue shapes. Note that this secondary analysis is restricted to the 29 speakers of the MRI cohort, as the sublingual cavity and/or mandibular shadow preclude accurate imaging of the tongue tip using ultrasound.

3. Results

As a first approximation observed tongue shapes derived from MRI can be sorted into the six shapes exemplified in Figure 3. These distinguish between concave, flat and convex tongue dorsum shapes, further separated by whether the tongue tip is tilted up (apical) or down (laminal). When these shapes are characterized by Fourier decomposition of their respective distance functions as described in Section 2.2.1 above, a *k*-means classification of their associated coefficients clusters optimally into three groups by both silhouette and elbow heuristics (N=29). Principal component analysis of all speaker tongue shapes (N=78) showed independently that three components accounted for 95% of variance for both /r/ and /l/.

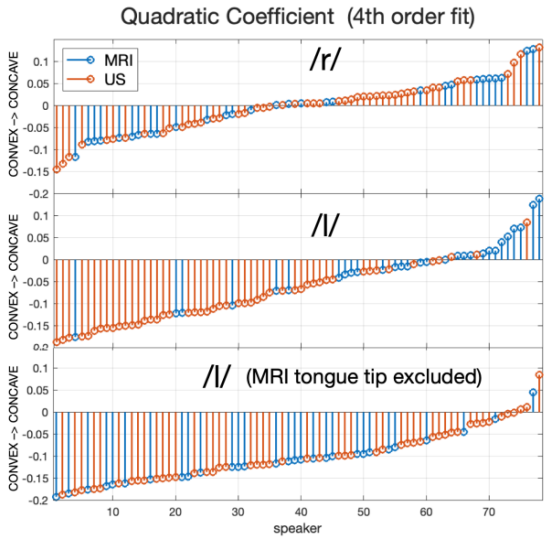

Figure 4: *Distribution of quadratic coefficients for 4th order polynomial fit to normalized tongue dorsum shapes (N=78).*

The results of fitting a 4th order polynomial to the normalized tongue dorsum data from all speakers are shown in Figure 4 (N=78). It can be seen that for /r/, shapes derived from MRI are distributed across the range about the same as those from ultrasound, confirmed by a linear model predicting C2 from data source (t(76) = 0.07 n.s.). For /l/, however, there is a strong bias towards negative (convex) shapes for the ultrasound data not seen in the MRI shapes (t(76)= -5.52 ***), which likely reflects the latter including tongue tip information not available from ultrasound. When fits are computed for /l/ with the anterior-most 30% of the MRI excluded (Fig. 4, bottom panel), this difference is no longer significant (t(76) = 1.41 n.s.), and we therefore conclude that data from both modalities can be successfully combined with this exclusion operative. Normalized tongue dorsum shapes (excluding MRI tongue tips) are shown averaged across concave, flat and convex values of C2 in Figure 5 (threshold for "flat" +/- .02).

When only complete (tongue tip included) MRI shapes are considered, there is a significant correlation between C2 values for /r/ and /l/ (r = 0.39 *). For the tongue tip, we observed that using the rhotic apical (COG<0) *vs*. laminal (COG>0) pattern as a prior predicted the same pattern for the corresponding within-speaker lateral: 83.3% of apical /r/ speakers produced an apical /l/. However, the converse was at chance: 50.0% of apical /l/ speakers produced laminal /r/ (MRI only; N = 29).

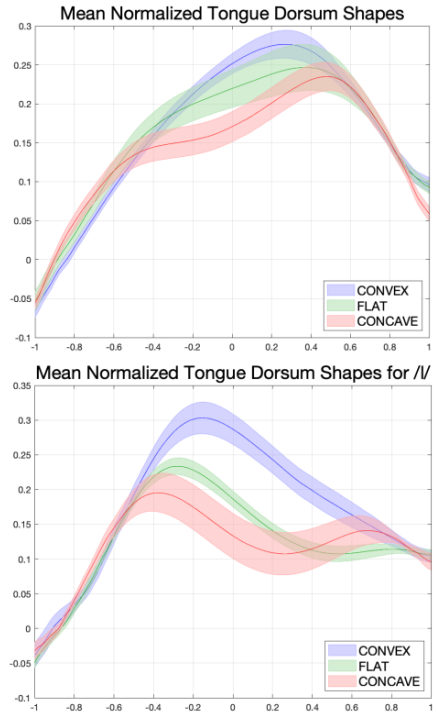

Figure 5: *Normalized tongue dorsum shapes (excluding MRI tongue tips); error bars show SEM.*

Although not directly part of the parameterization, scaling and rotation factors used to normalize tongue dorsum shapes showed an interesting gender-distinct pattern predictable from known differences in morphology (e.g. Vorperian et al., 2005): Rotation angles were consistently smaller for female speakers (t(153) = -3.23 **), likely reflecting shorter pharynx lengths relative to overall vocal tract length and thus less scope for tongue body rotation. Similarly, scale factors were also reliably larger for female speakers (t(153)= 2.71 **), likely reflecting smaller head and tongue sizes.

4. Discussion and conclusion

The extensive variety of observed midsagittal tongue shapes used to produce perceptually equivalent acoustic signatures for /r/ and /l/ likely reflects their interaction with individual differences in speaker palatal morphology. (While misalignment of the sampling plane is also a possibility, MRI shapes were verified against a midsagittal cross-section of coronally-oriented volumes collected during the same session.) Given this variety, how do language learners settle on a preferred shape? Syllabic liquids are notoriously among the last NAE sounds to be acquired, unsurprising given that they require coordination of at least three constrictions (lips, and two or more of the tongue within the vocal tract). One possibility may be that children, given sufficient exploration of articulatory possibilities guided by their own perceptual feedback and reinforcement from their parents and peers

eventually stumble into a configuration that succeeds in producing the appropriate acoustics.

However, a second possibility is that coproduction with other speech targets may expose them to alternative strategies which are close to liquid targets: In two instances participants in this study succeeded in producing separately scanned apical and laminal variants of /r/ with the same acoustics but very different dorsal shapes. Additional scanning of coproduced onset (/Cr/) contexts showed an apical posture during the rhotic for the former and a laminal posture for the latter (Figure 6). Alternative /r/ postures employed by the same speaker have also been found using EMA (Guenther et al., 1999; Tiede et al., 2010) and ultrasound (Mielke et al., 2016). This suggests that fluent NAE speakers have access to more than one production strategy for liquids, selected on least-effort principles during coproduction, but favoring one over others in syllabic contexts as being easier (for them) to produce and sustain.

Figure 6: *Coproduced /Cr/ onset contexts from the same speaker show contrasting apical (left) vs. laminal (right) tongue postures.*

Predicting a given speaker's preferred tongue posture for liquids on the basis of their vocal tract morphology would be useful for guiding possible clinical intervention, but this remains a very challenging problem, with parasagittal shape, tongue size, muscle interdigitation and asymmetry just some of the unknown free variables affecting the observed midsagittal projection. Previous studies of midsagittal shapes of liquids have mostly followed the pioneering efforts of Delattre & Freeman (1968), who categorized the 48 shapes they observed using cineradiography into one of eight prototypes. Because our own survey found shapes that could not be readily accounted for by these prototypes, a useful step towards addressing the prediction problem is a more precise way of quantifying midsagittal shape. The two parameter approach proposed here represents an improvement over prototype classification in that it accurately separates the six basic shapes found in our survey, is arbitrarily extensible to parameterizing any midsagittal shape, and provides quantified values that can correlated with available morphological measures.

5. Acknowledgements

Work supported by NIH grants DC05250 (Boyce) and DC002717 (Whalen). Thanks to Alan Wrench for ultrasound/audio alignment, Defne Abur for assisting with data collection during Noorderzon, and especially all the participants who volunteered their time at the festival.

6. References

Chen, W-R., Tiede, M., & Whalen, D. (2020). DeepEdge: automatic ultrasound tongue contouring combining a deep neural network and an edge detection algorithm. Paper presented at the *12th International Seminar on Speech Production* (ISSP 2020). https://github.com/WeirongChen/DeepEdge.

Delattre, P. & Freeman, D. (1968) A dialect study of American R's by X-ray motion picture. *Linguistics*, 44, 29-68.

Espy-Wilson, C., Boyce, S., Jackson, M., Narayanan, S. & Alwan, A. (2000) Acoustic modeling of American English /r/. *The Journal of the Acoustical Society of America*, 108(1), 343-356.

Guenther, F., Espy-Wilson, C., Boyce, S., Matthies, M., Zandipour, M., & Perkell, J. (1999) Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of the Acoustical Society of America*, 105(5), 2854-2865.

Hagiwara, R. (1995) Acoustic realizations of American English /R/ as produced by women and men. *UCLA Working Papers in Phonetics*, 90, 1-187

Liljencrants, J. (1971). Fourier series description of the tongue profile. *KTH Speech Transmission Laboratory – Quarterly Progress Status Reports*, 12(4), 9-18.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech 2017* (Stockholm), 498–502.

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language*, 101-140.

Spreafico, L., Pucher, M., Matosova, A. (2018). UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science. *Proc. Interspeech 2018* (Hyderabad), 1517-1520.

Tiede, M., Boyce, S., Espy-Wilson, C. & Gracco, V. (2010). Variability of North American English /r/ production in response to palatal perturbation. In Speech Motor Control: New Developments in Basic and Applied Research, 53-67, B. Maassen & P. van Lieshout, Eds. Oxford University Press.

Twist, A., Baker, A., Mielke, J., & Archangeli, D. (2007). Are "covert" /ɹ/ allophones really indistinguishable?. *University of Pennsylvania Working Papers in Linguistics*, 13(2), 207-216.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 117(1), 338-350.

Westbury, J., Hashi, M., & Lindstrom, M. (1998) Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, 26, 203-226.

Wieling, M., Rebernik, T., & Jacobi, J. (2023). SPRAAKLAB: a mobile laboratory for collecting speech production data. In *Proceedings of the 20th International Congress of Phonetic Sciences* (Prague), 2060-2064.

Zhou, X., Espy-Wilson, C., Boyce, S., Tiede, M. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America*, 123, 4466-4481.