Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation

William F. Katz,[a] **Sonya Mehta, and Matthew Wood**
Department of Communication Sciences and Disorders, The University of Texas at Dallas,
800 West Campbell Road, Richardson, Texas 75080-3021, USA
wkatz@utdallas.edu, sxm053100@utdallas.edu, mtw130130@utdallas.edu

Jun Wang
Department of Communication Sciences and Disorders, Department of Bioengineering,
The University of Texas at Dallas, 800 West Campbell Road, Richardson,
Texas 75080-3021, USA
wangjun@utdallas.edu

Abstract: This study examined the contributions of the tongue tip (TT), tongue body (TB), and tongue lateral (TL) sensors in the electromagnetic articulography (EMA) measurement of American English alveolar consonants. Thirteen adults produced /ɹ/, /l/, /z/, and /d/ in /ɑCɑ/ syllables while being recorded with an EMA system. According to statistical analysis of sensor movement and the results of a machine classification experiment, the TT sensor contributed most to consonant differences, followed by TB. The TL sensor played a complementary role, particularly for distinguishing /z/.

[AL]
Date Received: May 3, 2016 **Date Accepted:** December 12, 2016

1. Introduction

An ongoing challenge for speech science is how to best represent tongue movement using standard measurement techniques. Two approaches to measurement are commonly employed: Whole-tongue imaging methods and flesh-point tracking. Whole-tongue imaging provides detailed information concerning tongue shape and position. Historically, techniques such as cineradiography and videoflourography have provided useful data, although ultrasound[1] and magnetic resonance imaging (MRI)[2,3] have become increasingly prominent and informative. These systems provide comprehensive views of tongue shape, movement, and contact, although it is more difficult to use such systems to determine spatial coordinates for quantitative analysis. A common flesh-point tracking technique, electromagnetic articulography (or EMA), involves attaching small sensors to a talker's tongue, lips, and jaw. The movement of these sensors is tracked as the sensors pass through an alternating magnetic field, resulting in positional data for multiple articulatory regions during speech. EMA allows for accurate measurement and recording of articulatory movement with effective time resolution.

This study uses EMA to address manner of articulation differences in speech sounds (i.e., how sounds are produced), focusing on consonants produced at the alveolar ridge. We present a detailed examination of American English talkers' /ɹ/, /l/, /z/, and /d/ productions, with the aim of determining (1) whether EMA data can distinguish these manner distinctions, and (2) the extent to which lateral sensor data (assessing convex/concave shape change) play a role in these distinctions.

Briefly, the literature on American English alveolar consonants suggests that measurement of lateral tongue position may be informative. American English /ɹ/ is the most variable among these consonants, often described as consisting of "bunched" and "retroflex" variants. X- ray motion picture analysis suggested six types of articulations that could be used to achieve an "r" percept in American English.[4] Substantial articulatory variability was also found in /ɹ/-containing word productions analyzed using the x-ray microbeam (flesh-point tracking) system.[5] Studies using MRI[6] and combined ultrasound/electropalatography (EPG)/MRI techniques[7] have confirmed that American /ɹ/ productions are achieved by a variety of tongue configurations, including

[a] Author to whom correspondence should be addressed.

a common pattern of a concave (inwardly drawn) tongue body (TB) and a more convex tongue anterior. Altogether, the literature suggests that talkers produce a range of tongue shapes for American /ɹ/ and that these vary between bunched and retroflex endpoints.

The lateral approximant /l/ involves sound shaped by air flow in small channels formed along the sides of the tongue. Early cineradiography work supported the general impression that /l/ has two allophonic variants, an anterior lingual form ("light /l/") in prevocalic productions and a posterior, velar form ("dark /l/") in post-vocalic productions.[8] Subsequent x-ray microbeam research suggested that these articulatory types may more properly be considered as endpoints on a continuum.[9] Additional findings concerning /l/ production, including convex/concave shape distinctions, have been provided by MRI and EPG imaging. Data for American English /l/ have shown patterns of linguo-alveolar contact and inward lateral compression (causing a convex cross-section) along the tongue posterior body region.[10] Other data indicate the anterior oral cavity is the main region distinguishing American English /l/ and /ɹ/, and the pharyngeal region is similar for the two sounds.[7] In summary, kinematic data provide important details about the variability of /l/ production, including tongue shape differences and how these differences may affect the shape and symmetry of the lateral flow channels.

Early x ray studies traced regularities in fricative vocal tract positioning across different phonetic contexts.[11] These data were chiefly used to determine patterns of consistency along the midsagittal plane, but not for cross-sectional morphology (i.e., coronal plane determinations). Ultrasound, EPG, and MRI studies have supplemented this midsagittal information with additional anatomical detail. For instance, ultrasound data have revealed deep lingual grooving in the case of certain fricatives, such as /s/. MRI data for talkers producing fricatives of American English showed that the anterior TB of /s/ and /z/ had concave cross-sectional shapes, resulting in a notable area function difference behind the constriction, compared to postalveolars[12] (although this pattern may vary somewhat across different vowel contexts).[3]

In summary, whole-tongue imaging findings to date suggest that American English /r/ productions are realized with a convex TB and concave tongue posterior, /l/ productions show a more flattened tongue posterior and convex cross-sections of the tongue posterior, while /z/ is produced with a deep central groove involving tongue retraction toward midline. Given these emerging data addressing tongue shape during consonant production, we investigated whether it would be beneficial to incorporate tongue lateral (TL) sensor data into EMA analyses of speech production. Our aim was to determine whether EMA sensor movement in the coronal plane could improve the description of consonant manner of articulation in American English. Based on our previous work,[13] we hypothesized that the EMA tongue tip (TT) sensor would provide the most useful information for distinguishing these four American English consonants (/ɹ/, /l/, /z/, and /d/), followed by the TB sensor. We further predicted that tongue shape information supplied by a TL sensor would enhance classification for this set of sounds.

2. Methods

2.1 Participants

A total of 13 adult volunteers (ages 18–36 yrs, mean = 24) participated. All were native speakers of English studying at UT Dallas (5 males and 8 females). None reported any history of speech, hearing, or language difficulties.

2.2 Stimuli

Talkers produced /ɹ/, /l/, /z/, and /d/ consonants in /ɑCɑ/ (nonword) utterances. These stimuli were embedded in the carrier phrase "It's a ___ game" (e.g., "*It's a ra game.*"). Each subject repeated the stimulus list a minimum of ten times at their natural volume and speaking rate.

2.3 Tongue motion tracking procedure

An AG501 EMA system (Carstens Medezinelektronic GmbH, Bovenden, Germany) was used to record speech movement concurrently with a synchronized speech signal. Kinematic data were sampled at 250 sample/s; acoustic data at 22 kHz (*.wav format). Speech movement was recorded for tongue sensors (TT, TB, and TL) and lips (upper lip, UL, and lower lip, LL). The TT sensor was placed 0.5 to 1 cm posterior to the apex, the TL sensor approximately 2 cm posterior to the TT sensor (displaced 1.5 cm left of midline), and the TB sensor approximately 4 cm posterior to the TT sensor (see

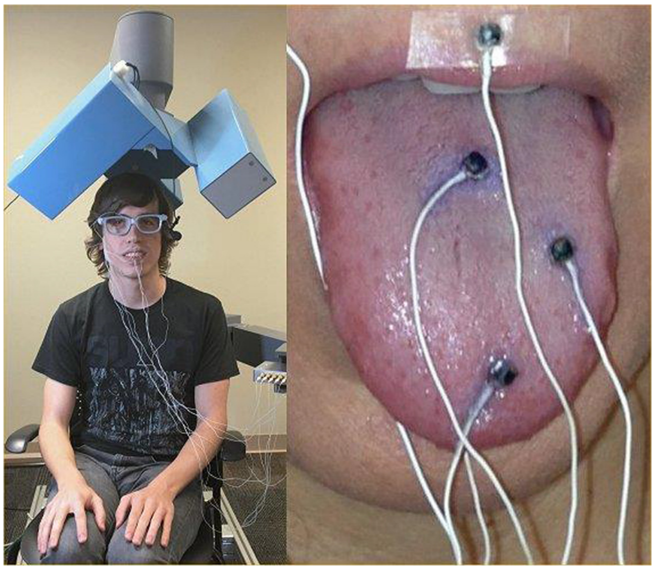

Fig. 1. (Color online) EMA recording system (left) and close-up showing tongue midline and lateral sensors. Sensors were also attached to the UL and LL. Reference sensors attached to a pair of glasses were used to eliminate head movement and to establish a local reference plane.

Fig. 1). Three sensors, attached to a pair of glasses worn by the subject, established a head-reference frame, which was used to provide head-independent movement information.[13] After head-movement correction, a low-pass filter (20 Hz) was applied for removing noise using SMASH,[14] a software program for articulatory data processing, visualization, and analysis.

2.4 Analysis

Talkers' kinematic patterns were visually inspected, and subsequently analyzed both qualitatively and quantitatively. Plots of talkers' utterances were first used to identify qualitative differences in articulator extent and direction, talker-particular patterns, and to infer distinctions such as bunched versus retroflex /ɹ/ patterns. Sample three-dimensional (3D) plots are shown in Fig. 2. The data were next analyzed using two quantitative approaches, including (1) direct comparisons of the positional changes of TT, TB, TL, and TLx relative to TTx, and (2) a time series pattern analysis of lingual sensor movement using Dynamic Time Warping (DTW).

TT, TB, TL, and TLx - TTx sensor position changes. In order to compare vertical (y), horizontal (z), and lateral (x) motion across individual talkers, the sensor movement data were normalized by subtracting a reference position from data within each carrier phrase. The consonant-vowel (CV) consonant offset position was used as a reference point, transforming the data from all talkers into a common, local coordinate system. The 3D coordinates for each of the tongue sensors (TT, TB, and TL)

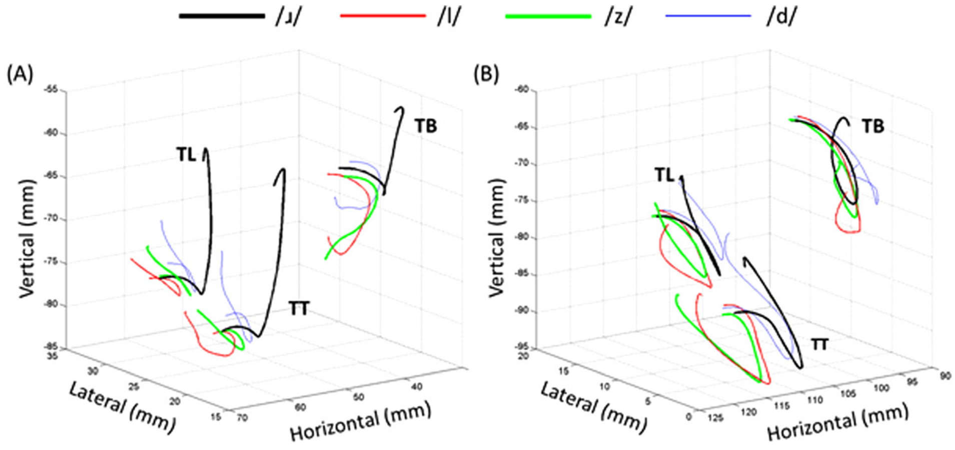

Fig. 2. (Color online) Representative examples of averaged 3D motion paths of the TT, TB, and TL sensors for ten repetitions of *ra*, *la*, *za*, and *da* spoken by two talkers. Distances are in mm. The *ra* patterns (boldest lines) differ qualitatively between the two talkers, suggesting more of a bunched pattern in (a) than in (b).

were obtained at CV onset and offset using customized MATLAB scripts. The CV portion of the utterance was located based on both acoustic and kinematic cues. Consonant onset was defined at (or immediately prior to) the first vertical peak for TT sensor position following the preceding /ə/ in "*It's a CV game.*" Vowel offset was defined at the stop closure for the /g/ of "*game*" (maximum vertical position for the TB sensor). An example is provided in Fig. 3.

Statistical analyses were performed on the positional changes of TT, TB, TL, and TLx relative to TTx. The current analysis examined the vertical (y), horizontal (z), and lateral (x) movement components of the individual sensors between consonant release (onset) and offset. Based on previous work by Wang and colleagues,[13] the TTy component was predicted to be most informative, followed by TTz. We further hypothesize a possible TTx (lateral) contribution.

We also computed a normalized measure of the lateral (x) movement of the TL sensor by subtracting the lateral TT data as a reference. This measure allowed us to better estimate the role of the TL sensor, independent of overall TL movement, during consonant production. The TLx relative to TTx (independent of TT) measure was calculated using this equation:

(1) Lateral positional change of TL relative to change of TT, from time points a to b:

$$\Delta TLx - \Delta TTx = (TLx_b - TLx_a) - (TTx_b - TTx_a)$$

where x_a = first time point (consonant onset)
and x_b = second time point (CV offset)

DTW for analyzing temporal movement patterns. DTW is widely accepted as the best distance measure among time-series signals.[15] DTW has been successfully used in many domains, including speech kinematics.[15] The standard DTW algorithm calculates the summed distances between data points of two time-series signals, after aligning the peaks. Thus, DTW is particularly useful in signals that have temporal

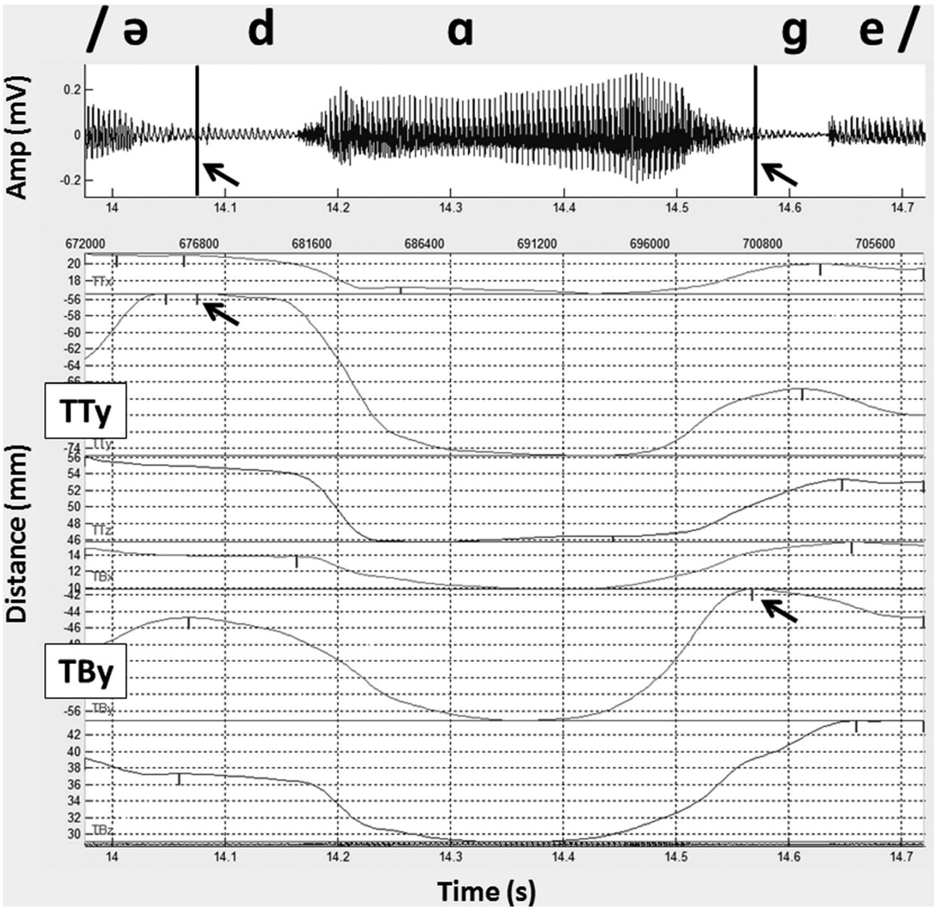

Fig. 3. Example of data selection for CV portions of kinematic signal (based on acoustic consonant onset and offset) and for the /g/ of game. Vertical lines on the acoustic waveform (top panel) indicate onset and offset cutoffs that correspond with TT and TB vertical movement peaks (shown below).

variations (e.g., speech). In this paper, DTW was used as a classifier to distinguish different sounds from articulatory time-series data (i.e., the spatial coordinates of sensors attached on the tongue). The procedure of using DTW for classification includes three steps:[15] First, the training process generates a "representative" or "reference" signal based on multiple training samples for each sound. Second, the distances between the test sample and each reference signal for each sound (here, for each consonant) are calculated. Third, the sounds whose reference signal has the shortest distance with the test sample are recognized (assigned to a consonant). This was performed in a cross-validation manner, where half of the data set was used for training and the other half for testing in one execution and switched in a second. The performances from the two executions were averaged as the final result.

2.5 Results

Statistical analysis of movement patterns. 3D plots revealed differences in movement patterns between the manners of articulation, as well as peculiarities inherent to individual speakers. Figures 2(a) and 2(b) suggest that maximally distinct articulatory variants for /ɹ/ (bunched and retroflex) could be found in the data, as was evident by the higher and backed sensor patterns for the bunched productions [Fig. 2(a)]. Four of the 13 talkers showed productions that appeared to be bunched, based on a raised TB sensor and lowered TT sensor. In general, talkers were quite regular with the direction and extent of their articulatory patterns, although one talker showed an unusual pattern of UL lowering for /ɹ/.

The positional data were quantified as follows: Analysis of variance (ANOVA) examined the change over time of sensor position from the onset of the consonants /ɹ/, /l/, /z/, and /d/ to the beginning of /g/, e.g., in "*ra game*." This was done to capture the fullest extent of the consonantal movement across a syllable. In order to provide an estimate of TL contribution to consonant classification (independent of TT), separate two-way (Consonant × Movement direction) ANOVAs were conducted for the TT, TB, and TL data, as well as for the differences between the TL and TT sensors (TL - TT).

The ANOVA results for TT (dv = position in mm) were significant for the (y) movement dimension [$F(3,36) = 13.83$, $p < 0.001$]. *Bonferroni*-adjusted pairwise comparisons of this main effect revealed that /l/ and /z/ were produced with more raising movement toward the /g/ of *game* than were /d/ and /ɹ/. That is, /d/-/l/, /d/-/z/, /l/-/ɹ/, and /ɹ/-/z/ contrasts were significantly different, while /l/ - /z/ and /d/ - /ɹ/ were not. The horizontal (z) movement data showed a significant difference as a function of consonant [$F(3,36) = 46.48$, $p < 0.001$]. Pairwise comparisons investigating this main effect for horizontal movement indicated all consonants were distinguished from each other, except for /d/ and /z/. The most anterior movement relative to the /g/ of *game* was found for /ɹ/, followed by /d/, /z/, and /l/ (all three of which showed successively more posterior net movement). TT lateral (x) dimension analyses were not significant [$F(3,36) = 0.246$, $p < 0.864$].

The TB results largely parallel the findings for TT. There was a significant main effect for the vertical (y) movement data [$F(3,36) = 33.1$, $p < 0.001$], with pairwise comparisons indicating all consonants were distinct, with the exception of /l/-/z/. In short, all alveolar consonants showed raising towards the /g/ of *game*, ranked in the order of (/l/ = /z/) > /d/ > /r/. A significant main effect was found for horizontal (z) movement [$F(3,36) = 16.8$, $p < 0.001$], with pairwise comparisons indicating significant differences between /ɹ/ and the other three consonants, all of which showed net anterior movement. This indicates /d/ was produced with the most anterior movement toward /g/. The main effect for lateral (x) movement was not significant [$F(3,36) = 0.573$, $p < 0.637$].

For TL, the vertical (y) data differed significantly between consonants, [$F(3,36) = 23.89$, $p < 0.001$]. *Post hoc* comparisons suggested an effective difference for /l/, as it was produced with more height moving toward /g/ than the other three consonants. Also, the horizontal (z) data differed significantly between consonants, [$F(3,36) = 136.97$, $p < 0.001$]. Pairwise comparison showed that /ɹ/ was produced with more anterior movement moving toward the /g/ of *game* than the other three consonants. The lateral (x) TL data were significant, [$F(3,36) = 5.19$, $p = 0.030$]. Pairwise comparison showed the consonant /z/ was produced with more lateral excursion from consonant onset to the /g/ of *game* than was /l/ and /d/.

Recall that the measure of TLx relative to TTx was devised to allow estimation of relatively specific TLx movement [e.g., and not to more general left/right positioning of the entire tongue in the mouth, Eq. (1)]. A one-way ANOVA revealed a main effect of consonant [$F(3,12) = 9.72$, $p < 0.001$]. Pairwise comparisons showed

Table 1. Contribution of TT, TB, and TL sensors for DTW classification of /ɹ/ and /z/.

Sensor	Average accuracy for /ɹ/	Average accuracy for /z/
Tongue tip		
TT*x*	59%	59%
TT*y*	56%	71%
TT*z*	90%	61%
Tongue back		
TB*x*	58%	48%
TT*y*	79%	56%
TT*z*	78%	69%
Tongue lateral		
TL*x*	49%	65%
TL*y*	66%	76%
TL*z*	86%	61%

significant differences for /z/ versus /d/ and for /z/ versus /ɹ/. Here, /z/ showed a greater degree of lateral excursion from onset to the /g/ of *game* than /d/ and /ɹ/.

DTW classification. The DTW results are shown in Table 1. Overall, the procedure yielded significant classification performance ($p < 0.05$).[16] Of the three lingual sensors, TT contributed a slightly higher mean degree of overall consonant classification accuracy [65.58%, standard deviation (SD) = 3.2] than both TL (60.5%, SD = 8.0) and TB (60.1%, SD = 8.5), although this difference did not reach significance. There was a trend for horizontal (z) information to be most informative (66.99%, SD = 13.5), followed by vertical (y) at 64.5% (SD = 8.5), with lateral (x) information significantly less accurate than vertical (54.8%, SD = 7.8).

We next examined whether including the TL sensor helps classify any particular consonant of the series. Table 1 (right column) shows that TL data (67%, average x, y, z accuracy) play a bigger role than TT (63% average) and TB (57% average) in distinguishing /z/ from the other consonants. In contrast, TL does not show a predominant role (compared to TT and TB) in distinguishing the other consonants (/d/, /l/, /ɹ/). These DTW data thus resemble the measured results for sensor positional change over time during the transition from consonant onset to a following consonant.

3. Discussion and conclusions

The current findings from both movement analyses and machine classification results suggest an EMA sensor placed on the tongue's lateral surface provides information that is helpful in characterizing the American English alveolar consonants /ɹ/, /l/, /z/, and /d/. Although the TT sensor provided the highest amount of information useful for distinguishing these consonants (as noted in the DTW results), distinct patterns were noted in the measured positions for the TL sensors. For instance, /z/ was produced with greater TL lateral displacement as it moved from consonant onset to the following /g/, compared with /d/, /l/, and /ɹ/. This suggests a more midline tongue position for /z/ than the other alveolar manners of articulation, agreeing with previous reports of /z/ production based on MRI data.[12] Also, DTW classification results indicated marked accuracy for /z/, particularly for the TL sensor (Table 1). While lateral tongue motion may otherwise be considered insignificant in general speech measurement,[13] the current findings suggest judicious use of a lateral sensor in systems such as EMA may yield useful information concerning tongue shape and sound-related motion. Such information may be useful for a variety of purposes, including a more fine-grained characterization of different speech sounds (e.g., bunched versus retroflex /ɹ/), and developing speech training systems based on the provision of real-time tongue position visual feedback. Future studies should also explore bilateral placement of lateral sensors to investigate tongue symmetry in alveolar consonant production.

Acknowledgments

This project is in part supported by the National Institutes of Health through a R03 grant (Grant No. DC013990) and the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We also thank Amy Berglund, Tristanne Mock, Kelly Fazel, Marcus Jones, Aarthy Sundar, Amanda Clark, Mona Dirbashi, and the participants who volunteered their time for this research.

References and Links

[1] M. Stone, "A guide to analysing tongue motion from ultrasound images," Clin. Ling. Phon. **19**(6–7), 455–502 (2005).

[2] X. H. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of 'retroflex' and 'bunched' American English /r/," J. Acoust. Soc. Am. **123**, 4466–4481 (2008).

[3] C. Shadle, M. I. Proctor, and K. Iskarous, "An MRI study of the effect of vowel context on English fricatives," in *Proceedings of the Joint Meeting of the Acoustical Society of America and European Acoustics Association*, Paris, France, pp. 5099–5104 (2008).

[4] P. Delattre and D. C. Freeman, "A dialect study of American r's by x-ray motion picture," Linguistics **6**, 29–68 (1968).

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

[6] A. Alwan, S. Narayanan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," J. Acoust. Soc. Am. **101**, 1078–1089 (1997).

[7] D. Ong and M. Stone, "Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics," Phonoscope **1**, 1–13 (1998).

[8] S. B. Giles and K. L. Moll, "Cinefluorographic study of selected allophones of English /l/," Phonetica **31**, 206–227 (1975).

[9] R. Sproat and O. Fujimura, "Allophonic variation in English /l/ and its implications for phonetic implementation," J. Phon. **21**, 291–311 (1993).

[10] S. Narayanan, A. Alwan, and K., Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," J. Acoustic. Soc. Am. **101**(2), 1064–1077 (1997).

[11] P. Badin, "Fricative consonants—Acoustic and x-ray measurements," J. Phon. **19**, 397–408 (1991).

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

[13] J. Wang, W. Katz, and T. F. Campbell, "Contribution of tongue lateral to consonant production," Interspeech 174–178 (2014). http://www.isca-speech.org/archive/interspeech_2014/i14_0174.html.

[14] J. Green, J. Wang, and D. Wilson, "SMASH: A tool for articulatory data processing and analysis," Interspeech 1331–1335 (2013). http://www.isca-speech.org/archive/interspeech_2013/i13_1331.html.

[15] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dynamic time warping to the multi-dimensional case requires an adaptive approach," Data Mining Knowledge Discovery 1–31 (2016).

[16] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," J. Neurosci. Methods **250**, 126–136 (2015).