INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

An automatic, simple ultrasound biofeedback parameter for distinguishing accurate and misarticulated rhotic syllables

Sarah R. Li[1], *Colin T. Annand*[2], *Sarah Dugan*[2,3], *Sarah M. Schwab*[2], *Kathryn J. Eary*[1], *Michael Swearengen*[1], *Sarah Stack*[1], *Suzanne Boyce*[3], *Michael A. Riley*[2], *T. Douglas Mast*[1]

[1]Biomedical Engineering, University of Cincinnati, USA
[2]Psychology, University of Cincinnati, USA
[3]Communication Sciences and Disorders, University of Cincinnati, USA

lisr@mail.uc.edu

Abstract

Characterizing accurate vs. misarticulated patterns of tongue movement using ultrasound can be challenging in real time because of the fast, independent movement of tongue regions. The usefulness of ultrasound for biofeedback speech therapy is limited because speakers must mentally track and compare differences between their tongue movement and available models. It is desirable to automate this interpretive task using a single parameter representing deviation from known accurate tongue movements. In this study, displacements recorded automatically by ultrasound image tracking were transformed into a single biofeedback parameter (time-dependent difference between blade and dorsum displacements). Receiver operating characteristic (ROC) curve analysis was used to evaluate this parameter as a predictor of production accuracy over a range of different vowel contexts with initial and final /r/ in American English. Areas under ROC curves were 0.8 or above, indicating that this simple parameter may provide useful real-time biofeedback on /r/ accuracy within a range of rhotic contexts.

Index Terms: speech disorders, ultrasound imaging, speech production, articulation, speech therapy, rhotics

1. Introduction

Speech movement biofeedback makes explicit otherwise opaque parameters related to production success or failure, enabling the speaker to refine the production of a target. For speakers with residual speech sound disorder (RSSD), biofeedback provides information to improve articulation accuracy for sounds in error. One method of providing articulation information is ultrasound imaging of the midsagittal tongue surface, which shows much of the tongue root, dorsum, and blade in real time and is increasingly accessible in clinics [1-3]. Displaying these ultrasound images as biofeedback in speech therapy (ultrasound biofeedback therapy or UBT) generally results in positive outcomes [4].

However, non-responders for UBT remain [5-9]; one cause may be high attentional and cognitive demands required for interpreting ultrasound images and translating those interpretations into accurate speech movement patterns. Distinguishing accurate vs. misarticulated patterns of tongue movement can be challenging in real time because of the fast, independent movement of tongue regions [10-12]. In UBT, speakers must learn to interpret these rapidly changing tongue surface curves from ultrasound images and compare the resulting shapes to model images, which comprise the speaker's best previous attempts, example videos from other children, or demonstrations from the clinician [1, 3, 13]. Comparison to models is challenging because varying tongue shapes are used in typical speech. In particular, the rhotic /r/ of American English can be produced by configurations that are very different, described by a continuum ranging from "bunched" to "retroflex" /r/ [14-16]. Differences in tongue size and rotational position on the image due to anatomy and transducer placement may also interfere with comparisons of tongue shapes. Figure 1 shows two ultrasound images of /r/ (left: a more retroflex shape, right: a more bunched shape) in /ɑr/ productions from typical child speakers in our database, demonstrating how differently ultrasound images may appear for varying tongue shapes.

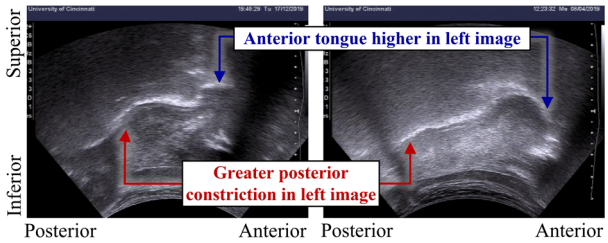

Figure 1: *Example ultrasound images of typical /r/ productions with differing tongue shapes in this study.*

These difficulties for the speaker could be reduced if tongue movement for these different but accurate tongue shapes could be characterized by a single parameter. In this vision, biofeedback could be calibrated to a single target value or range of values (e.g., visualized as a needle-style meter). Thus, speakers could more easily compare their tongue movements to accurate models and potentially improve UBT outcomes for /r/ articulation, without interpreting images of tongue shape. Such a biofeedback parameter would need to be clinically relevant, with the ability to discriminate between perceptually accurate speech and misarticulations. The parameter would be more valuable if usable over a wide range of phonemic contexts, varying but accurate tongue shapes, and individual speaker movement patterns. Computational simplicity would also be desirable to allow for display during speech production.

Toward the goal of simplified UBT, we have developed a custom program, TonguePART (Tongue Profiles with Automatic Rapid Tracking), that tracks the tongue surface in midsagittal ultrasound images and quantifies tongue movement as displacement trajectories of the tongue root, dorsum, and blade [17, 18]. We have used support vector machine (SVM) classifiers to evaluate whether these displacement trajectories distinguish accurate from misarticulated productions of /r/-final syllables [19, 20], showing high classification accuracies.

http://dx.doi.org/10.21437/Interspeech.2021-1749

In this paper, we expand our analysis of displacements from TonguePART by transforming these values into a single biofeedback parameter, the time-dependent blade displacement minus the dorsum displacement. We use receiver operator characteristic (ROC) curves to evaluate the ability of this parameter to predict production accuracy over a range of different vowel contexts with both final and initial /r/.

2. Methods

2.1. Data collection

2.1.1. Participants

Seventeen children (8-17 years old) with typically developing (TD) speech or RSSD were recruited. All children were native speakers of a rhotic American English dialect. The mean age (± standard deviation) of the TD speakers (four female, three male) was 14.2 ± 2.1 years. The mean age (± standard deviation) for the RSSD speakers (seven female, three male) was 11.4 ± 3.2 years.

2.1.2. Ultrasound imaging of speech production

For this study, we selected words with /r/ in three vowel contexts (560 total productions) from a larger dataset of words. Speakers produced 5–10 productions for each word:

- /r/-final words: /ir/ (“ear”), /ɑr/ (“are”), /buər/ (“booer”)
- /r/-initial words: /rip/ (“reap”), /rɑ/ (“raw”), /ru/ (“rue”)

Midsagittal ultrasound images were recorded with procedures described previously [17], using a head stabilizer and a Siemens Acuson X300 Premium Edition diagnostic ultrasound system with a C6-2 curved linear array transducer (90° field of view), an image depth of 8 cm, center frequency of 4.0 MHz, and frame rate of 36 fps. These images were collected as videos at 60 fps, with nearest-neighbor temporal interpolation. Audio concurrent with the ultrasound imaging was recorded at 44.1 kHz with a cardioid condenser USB microphone (Audio-Technica ATR2500-USB). Child speakers used a range of tongue configurations.

2.1.3. Auditory perceptual rating data

Naïve listeners participated in an online perceptual experiment programmed using jsPsych [21] and hosted on JATOS [22]. Listeners were asked to judge accuracy of the speakers’ productions from the larger dataset (section *2.1.2*) using a continuous scale with endpoints labelled “incorrect” and “correct,” mapped to 0 and 10 respectively in this analysis.

During the perceptual rating experiment, the associated text for each word was displayed on the screen, and the sound for each word could be replayed an unlimited number of times. Listeners judged a common subset, presented to all 103 listeners, and a randomly assigned subset. If a listener’s median time for rating each production in a subset was less than 1.5 seconds (including loading the audio file and listening, with an average and standard deviation audio file duration of 0.46 ± 0.18 seconds), their ratings for the subset were excluded. This resulted in ~81% of productions rated by 12–21 listeners and ~19% rated by 102 listeners. Intraclass correlation coefficients for mean ratings, ICC(2,*k*), ranged from 0.73–0.91 for the randomly assigned subsets to 0.99 for the common subset (the higher value due to more listeners), indicating moderate (> 0.5 and < 0.75) to excellent (> 0.9) reliability [23].

The ratings for the specific portion of words analyzed (see *2.1.2*) were then selected. Due to implementation errors, some productions were missing or improperly presented to listeners (e.g., incorrect audio edits) and subsequently excluded (3.4%, 19 out of 560). Listener ratings were then averaged and used to define whether a production was accurate (rating ≥ 5.5 [20]) or misarticulated (rating < 5.5) for the following analysis.

2.2. Automatic biofeedback parameter

Tongue movement in these productions was quantified as tongue part displacement trajectories automatically calculated using TonguePART [17], which performs low-pass filtering on each video frame of ultrasound and then delineates the tongue surface as brightness maxima within vertical search windows. Search window positions are initialized by a user-provided spatial calibration point and updated along the posterior and anterior directions using Taylor series estimates, with ends of the tongue surface determined by user-calibrated brightness threshold values. If necessary, calibration by the user was updated for different words and speakers.

The determined tongue surface contour is then automatically divided into three regions and reference points with equal horizontal spans representing the root, dorsum, and blade. Displacements are calculated as distances between the average reference points for the initial frame and following frames in a production. Positive displacement values in these trajectories represent narrowing of the vocal tract due to the specific tongue part, while negative values represent widening.

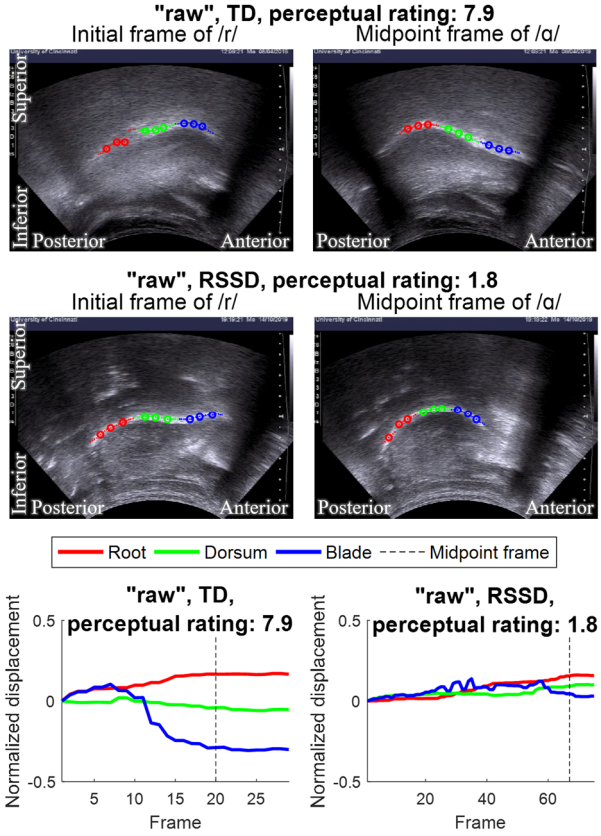

Figure 2: *Example ultrasound images, tracking, and displacement trajectories using TonguePART for /ra/ from a TD (top row, bottom-left graph) and an RSSD speaker (middle row, bottom-right graph).*

Displacement trajectories for these productions were calculated from the start to end of vocalization: displacements were computed relative to the initial frame of the vowel for an /r/-final word or the initial frame of /r/ for an /r/-initial word. Displacements were normalized via dividing by a reference tongue length, the average of distances between the root and blade middle reference points for the initial frames of /ɑr/ productions for each speaker. Tracking errors, associated primarily with ultrasound image artifacts or anomalies, were identified automatically by differential displacements [17] or by visual inspection, resulting in exclusion of 22.9% of the productions with ratings. Thus, 56–90 productions were analyzed for each word. Examples of tracking for productions from a TD speaker and RSSD speaker are shown in Figure 2.

These trajectories were transformed into a candidate biofeedback parameter calculated for each frame by subtracting dorsum displacement from blade displacement (a simple linear combination). Values of this parameter at frame 32 (out of 39 frames after displacement trajectories were interpolated to the same number of frames) were selected for analysis. This frame represents an estimated temporal acoustic midpoint for the /r/ in /r/-final productions [20] or vowel in /r/-initial productions.

2.3. ROC curve analysis

Capability of this biofeedback parameter as a predictor for production accuracy was evaluated using ROC curves for each vowel context with initial or final /r/. The perfcurve function in MATLAB (R2020b, *The MathWorks, Inc.*) was used to calculate ROC curves, and optimal threshold values were then defined with the nearest approach of the ROC curve to the upper left corner of the ROC plot. These optimal values indicate biofeedback parameter targets that balance sensitivity and specificity for distinguishing accurate productions from misarticulations. In general, area under the curve (AUC) values calculated from ROC curves range from 0.5 to 1, with 0.5 indicating prediction equivalent to chance and 1 indicating perfect prediction. For /r/-initial words, the direction of the ROC curve was reversed by inverting the sign of the biofeedback parameter.

3. Results and Discussion

Results for the analyzed rhotic words demonstrate that the proposed biofeedback parameter, i.e., the blade displacement minus the dorsum displacement, can discriminate most of the accurate from misarticulated productions. Because different tongue shapes for accurate productions (retroflex vs. bunched) were included (Figure 1), these results are expected to be generalizable for larger populations. Figure 3 shows ROC curves with associated AUC values for the six rhotic words analyzed. All AUC values were 0.8 or above, significantly greater than the null hypothesis (AUC=0.5). High AUC values also imply that the optimal threshold, with sensitivity and specificity at the circles in Figure 3, can be usefully altered. For example, if having fewer false negatives (accurate speech indicated by the biofeedback as inaccurate) is more desirable than fewer false positives in UBT, the threshold can be reduced to result in higher sensitivity with some decrease in specificity.

A closer investigation of the tongue part displacement trajectories measured by TonguePART explains why the parameter achieves discrimination. The top and middle rows of Figure 4 display time-dependent means and standard deviations for the tongue part displacement trajectories quantified from accurate and misarticulated /ir/ and /buər/ productions. Despite substantial variance in both accurate and misarticulated movement patterns, it is clear that relationships between dorsum and blade movement differ for accurate vs misarticulated productions. For the /i/-context /ir/, this is a difference of whether the dorsum or blade displacement has the greater magnitude, while for the /u/-context /buər/ the dorsum and blade move in opposing directions during accurate productions but move in more parallel directions for misarticulations. These findings support the use of blade and dorsum displacement difference as a biofeedback parameter.

The bottom row of Figure 4 shows the mean and standard deviation of the biofeedback parameter calculated at each timepoint, with the optimal threshold from the ROC analysis plotted as a magenta line. Although the standard deviation of this parameter is relatively large, it is generally lower than the difference needed for discrimination. Similar variation and separation occur for the /r/-initial and /ɑ/-context words (supplemental figures for these analyzed words can be found at https://github.com/SarahRLi/interspeech-2021-submission).

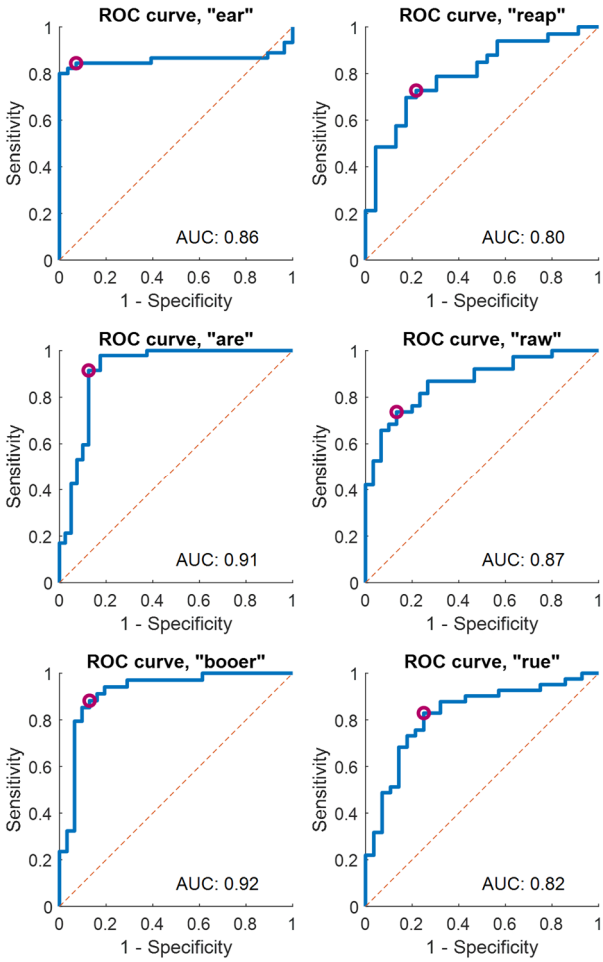

Figure 3: *ROC curves and AUC values, with the magenta circle indicating the optimal threshold.*

Figure 5 shows scatterplots of dorsum and blade displacements at the estimated midpoint, as used to compute the biofeedback parameter for ROC analysis. The distance of each point to the diagonal dotted line (blade=dorsum) is proportional to the biofeedback parameter, i.e., blade minus dorsum displacement. Magenta diagonal lines represent the optimal

thresholds found in the ROC analysis. Points to the upper left of this line are predicted as accurate in these plots (with a reversal in axis directions for /r/-initial words). The fill color for each point indicates the average perceptual rating as shown on the color bar, with the outline color yellow for accurate productions (rating ≥ 5.5) and blue for misarticulations (rating < 5.5). Circular and triangular markers indicate whether the production was from a TD or RSSD speaker, respectively.

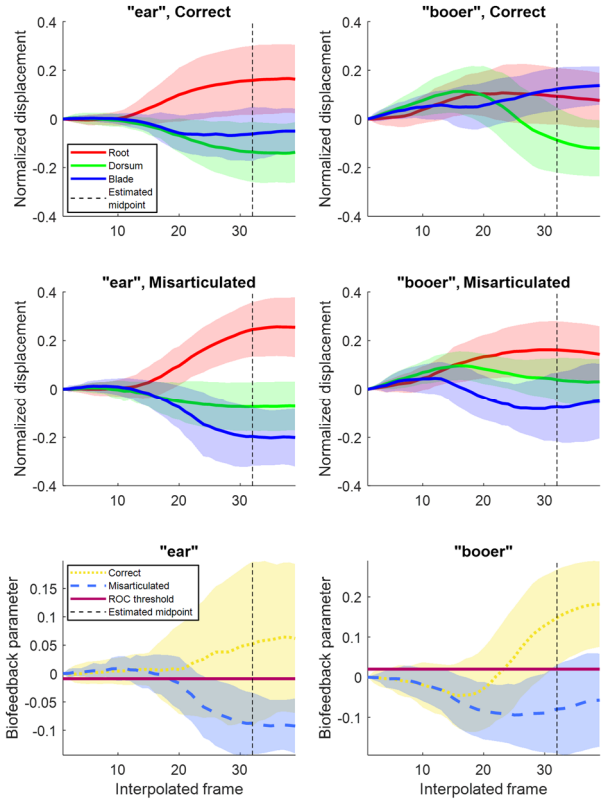

Figure 4: *Example time-dependent mean and standard deviation of tongue part trajectories for accurate (top row) and misarticulated (center row) /ir/ and /buər/ productions, with the associated biofeedback parameter trajectories (bottom row).*

These graphs also provide insight into articulatory reasons for the discriminative ability of the proposed biofeedback parameter. Accurate /r/ shapes require both anterior and posterior constrictions [14, 16, 24]. The positive blade displacements for accurate /ɑr/ and /buər/ (negative for /rɑ/ and /ru/) suggest that blade displacement represents the anterior constriction necessary for accurate /r/. For production of /i/, the tongue partitioned as the blade is in a high position within the mouth and therefore may not require much additional displacement for accurate /r/ productions. The negative dorsum displacement for accurate /r/-final words or positive displacement for accurate /r/-initial words may represent tongue flattening or grooving necessary for the posterior constriction [14]. However, some productions do not follow these patterns, and new insight may be gained by refinement of the tracking, displacement calculation, or biofeedback parameter definition.

Due to demonstrating discriminative ability across varying movement patterns and contexts, the proposed simple biofeedback parameter may be particularly useful for UBT. With this parameter, a single target could be used for each word, with the clinician or future biofeedback software needing to change only the target when the speaker is learning different words. Because the displacements compare movement to an initial frame and are normalized to a reference tongue length for each speaker, the biofeedback parameter is not complicated by differences in tongue positions and shapes that make comparing images of the tongue more difficult. Use of one target for each word may also be advantageous. By not restricting the comparison to one shape, this feature may allow a speaker to "self-organize" movement patterns and explore potential solutions that work better for them [25]. Learning outcomes can be improved when a speaker tries different configurations [7].

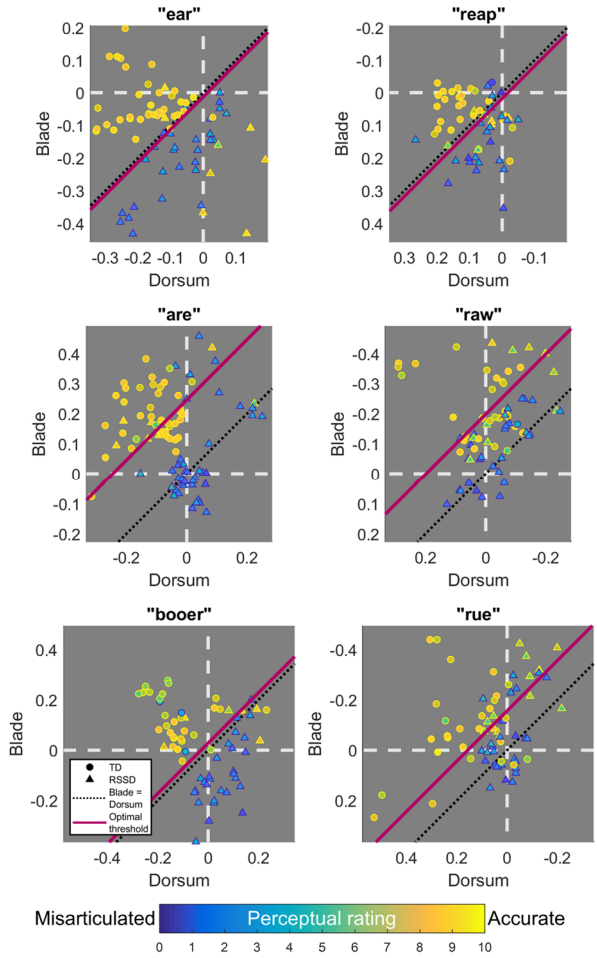

Figure 5: *Scatterplots showing dorsum and blade displacements at estimated midpoint. Axis directions for /r/-initial words are reversed for illustration.*

4. Conclusion

These results demonstrate potential benefits and clinical promise for a proposed simple biofeedback parameter, based on relative displacement of the tongue dorsum and blade, in UBT. Future testing is required to evaluate effectiveness of this approach for clinical biofeedback in speech therapy.

5. Acknowledgements

This work was funded by University of Cincinnati Creating Our Third Century funding support and by NIH/NIDCD grants R01 DC013668 and R01 DC017301. We also thank Siemens Medical Solutions USA, Inc., for making their Acuson ultrasound scanner available for this project.

6. References

[1] J. L. Preston et al., "Ultrasound images of the tongue: a tutorial for assessment and remediation of speech sound errors," *Journal of Visualized Experiments*, no. 119, pp. e55123, Jan. 2017, doi: 10.3791/55123.

[2] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics & Phonetics*, vol. 19, no. 6–7, pp. 455–501, Jan. 2005, doi: 10.1080/02699200500113558..

[3] J. Cleland, A. Wrench, S. Lloyd, and E. Sugden, "ULTRAX2020 : Ultrasound technology for optimising the treatment of speech disorders : clinicians' resource manual," *Strathprints - The University of Strathclyde Institutional Repository*, 2018. doi: 10.15129/63372.

[4] E. Sugden, S. Lloyd, J. Lam, and J. Cleland, "Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders," *International Journal of Language & Communication Disorders*, vol. 54, no. 5, pp. 705–728, Sep. 2019, doi: 10.1111/1460-6984.12478.

[5] G. M. Sjolie, M. C. Leece, and J. L. Preston, "Acquisition, retention, and generalization of rhotics with and without ultrasound visual feedback," *Journal of Communication Disorders*, vol. 64, pp. 62–77, Nov. 2016, doi: 10.1016/j.jcomdis.2016.10.003.

[6] J. L. Preston, M. C. Leece, and E. Maas, "Motor-based treatment with and without ultrasound feedback for residual speech-sound errors: Motor-based treatments for residual speech errors," *International Journal of Language & Communication Disorders*, vol. 52, no. 1, pp. 80–94, Jan. 2017, doi: 10.1111/1460-6984.12259.

[7] T. M. Byun, E. R. Hitchcock, and M. T. Swartz, "Retroflex versus bunched in treatment for rhotic misarticulation: evidence from ultrasound biofeedback intervention," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 6, pp. 2116–2130, Dec. 2014, doi: 10.1044/2014_JSLHR-S-14-0034.

[8] J. L. Preston et al., "Remediating residual rhotic errors with traditional and ultrasound-enhanced treatment: a single-case experimental study," *American Journal of Speech-Language Pathology*, vol. 28, no. 3, pp. 1167–1183, Aug. 2019, doi: 10.1044/2019_AJSLP-18-0261.

[9] Q. Heng, P. McCabe, J. Clarke, and J. L. Preston, "Using ultrasound visual feedback to remediate velar fronting in preschool children: A pilot study," *Clinical Linguistics & Phonetics*, vol. 30, no. 3–5, pp. 382–397, May 2016, doi: 10.3109/02699206.2015.1120345.

[10] J. R. Green and Y.-T. Wang, "Tongue-surface movement patterns during speech and swallowing," *Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2820–2833, May 2003, doi: 10.1121/1.1562646.

[11] M. Stone, M. A. Epstein, and K. Iskarous, "Functional segments in tongue movement," *Clinical Linguistics & Phonetics*, vol. 18, no. 6–8, pp. 507–521, Sep. 2004, doi: 10.1080/02699200410003583. doi: 10.1006/jpho.1996.0006

[12] N. Nguyen, A. Marchal, and A. Content, "Modeling tongue-palate contact patterns in the production of speech," *Journal of Phonetics*, vol. 24, no. 1, pp. 77–97, 1996.

[13] B. Bernhardt, B. Gick, P. Bacsfalvi, and M. Adler-Bock, "Ultrasound in speech therapy with adolescents and adults," *Clinical Linguistics & Phonetics*, vol. 19, no. 6 – 7, pp. 605 – 617, Jan. 2005, doi: 10.1080/02699200500114028.

[14] S. Boyce, "The articulatory phonetics of /r/ for residual speech errors," *Seminars in Speech and Language*, vol. 36, no. 04, pp. 257–270, Oct. 2015, doi: 10.1055/s-0035-1562909.

[15] P. Delattre and D. C. Freeman, "A dialect study of American r's by x-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968, doi: 10.1515/ling.1968.6.44.29.

[16] X. 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/," *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 17, Jul. 2008, doi: 10.1121/1.2902168.

[17] S. Dugan et al., "Tongue part movement trajectories for /r/ using ultrasound," *Perspectives of the ASHA Special Interest Groups*, vol. 4, no. 6, pp. 1644–1652, Dec. 2019, doi: 10.1044/2019_PERS-19-00064.

[18] C. T. Annand *et al.*, "Using ultrasound imaging to create augmented visual biofeedback for articulatory practice," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 974–975.

[19] S. R. Li et al., "Classification of accurate and error tongue movements for /r/ in children using trajectories from ultrasound," *Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1799, Apr. 2019, doi: 10.1121/1.5101588.

[20] S. R. Li et al., "Classification of accurate and misarticulated rhotic syllables for simplified ultrasound biofeedback therapy," *Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2470, Dec. 2020, doi: 10.1121/1.5146836.

[21] J. R. de Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behavior Research Methods*, vol. 47, no. 1, pp. 1–12, Mar. 2015, doi: 10.3758/s13428-014-0458-y.

[22] K. Lange, S. Kühn, and E. Filevich, ""Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies," *PLOS ONE*, vol. 10, no. 6, pp. e0130834, Jun. 2015, doi: 10.1371/journal.pone.0130834.

[23] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2016.02.012.

[24] A. Alwan, S. Narayanan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," *Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1078–1089, Feb. 1997, doi: 10.1121/1.417972.

[25] M. M. Pacheco, C. W. Lafe, and K. M. Newell, "Search strategies in the perceptual-motor workspace and the acquisition of coordination, control, and skill," *Frontiers in Psychology*, vol. 10, pp. 1874, Aug. 2019, doi: 10.3389/fpsyg.2019.01874.