Classification of accurate and error tongue movements for /r/ in children using trajectories from ultrasound

Sarah R. Li,[1] Sarah Dugan, PhD CCC-SLP,[2] Colin Annand,[3] Hannah M. Woeste,[1] Jack A. Masterson,[1] Suzanne Boyce, PhD CCC-SLP,[2] T. Douglas Mast, PhD,[1] & Michael A. Riley, PhD[3]

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

Introduction

- Ultrasound biofeedback therapy (UBT) for children with residual speech sound disorders (RSSD) [1] can be slow, possibly due to difficulty of interpreting ultrasound images [2].
- Our long-term goal is to develop a simplified ultrasound biofeedback system to improve speech therapy outcomes.
- Development first requires quantitative comparisons between normally articulated ("accurate") and misarticulated ("error") tongue movement.
- We focused on children and on American English /r/ (IPA /ɹ/).
- Clinicians refer to two general shapes as strategies for accurate /r/: retroflex and bunched.
  - Definitions for these shapes can vary [3-4] and represent a continuum [5].
  - Different /r/ strategies can produce perceptually equivalent results [4].
- We compared movement as tongue part trajectories [5-6] rather than tongue shapes, but patterns in previously observed shapes may extend to trajectory results.
- We used **TonguePART** (Tongue Profiles with Automatic Rapid Tracking, poster 3aSC1) to provide quantitative tongue root, dorsum, and blade displacement trajectories.

Hypothesis

- From time-dependent displacement of tongue parts, we can
  1) **Identify distinct strategies** for accurate movement
  2) **Classify** accurate vs. error tongue movement

Data Acquisition and Preparation

- Participants: 17 children with typically developing (TD) speech and 23 with RSSD, aged 8-17. All were speakers of a rhotic American English dialect.
- Stimuli: 15-20 productions of /ɑr/ recorded with a Siemens Acuson X300 PE ultrasound system at 36 fps
- Perceptual ratings: Trained listeners rated auditory accuracy of productions on a 10-point continuous scale (10 most accurate; ≥ 8 clinically accurate) [7].
- Data processing:
  - TonguePART (poster 3aSC1) to measure displacement trajectories of tongue root, blade, & dorsum
  - ~10% of trajectories excluded for sharp jumps, indicative of tracking errors
  - Interpolation of trajectories to match production lengths (39 frames)

Acknowledgements

This work was funded by NIH/NIDCD grants 1R01DC013668-01A1 and 1R01DC017301-01 and by University of Cincinnati Creating our Third Century Funding Support. We are also grateful to Siemens Medical Solutions for lending the Acuson X300 scanner.

Methods

- **Identifying articulatory strategies** for accurate movement:
  - Principal component analysis (PCA) to find patterns in fewer dimensions; first four principal components (PCs) used, which sum to explain > 90% variance
  - Hierarchical clustering on PC scores of accurate productions
  - Strategies identified as mean trajectories from each cluster
- **Classification**:`
  - *k*-fold cross-validation – Clusters found as general strategies were evenly divided across the *k*=5 folds. Same folds used on all tests
  - Comparison of three support vector machine (SVM) models with clinical accuracy ("accurate" with perceptual rating ≥ 8 or "error" < 8) as class labels:
    1. Linear SVM on dorsum and blade displacement at midpoint of /r/
    2. Radial basis function (RBF) kernel SVM on correlation distance (1 – Pearson *r*) to each identified strategy
    3. RBF kernel SVM on correlation distance to mean of all accurate trajectories

Results

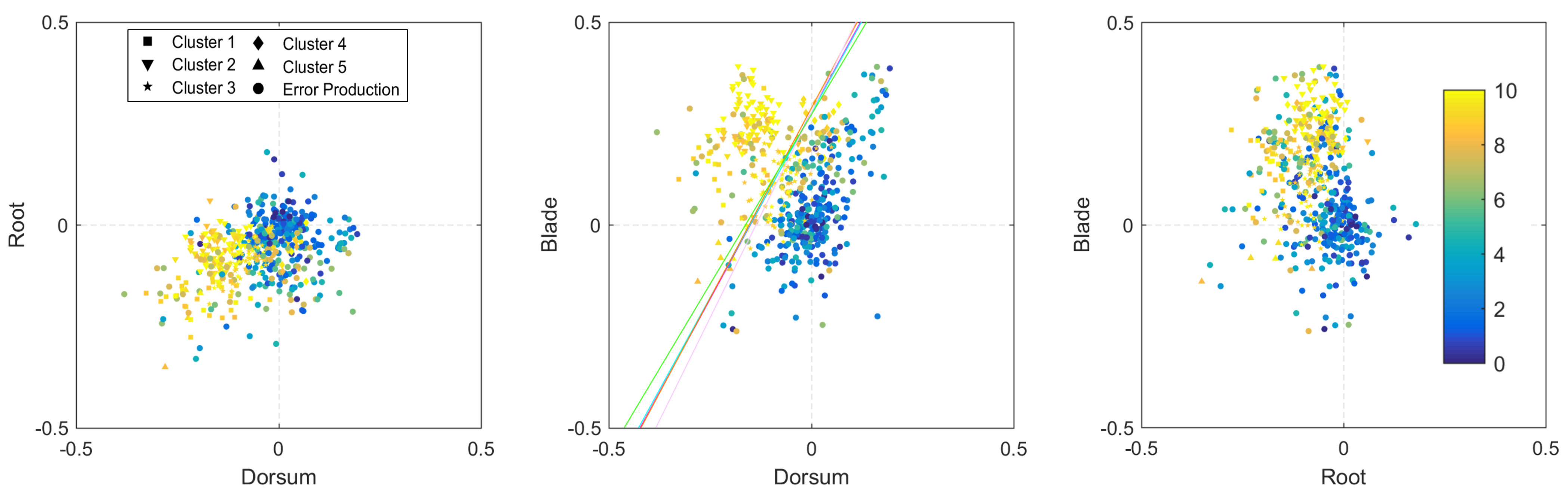


Figure 1: Scatterplots of normalized displacements of tongue parts at frame 32 of 39, representative of the midpoint of /r/. Color illustrates auditory perceptual ratings. Shapes represent cluster identity (accurate production in clusters 1-5 or error production). The colored lines in the center panel represent the hyperplane of the Linear SVM trained for each *k*-fold (model #1). Grid lines illustrate that error productions typically have smaller displacement values. Accurate vs. error productions are roughly separable by displacements of the three tongue parts.

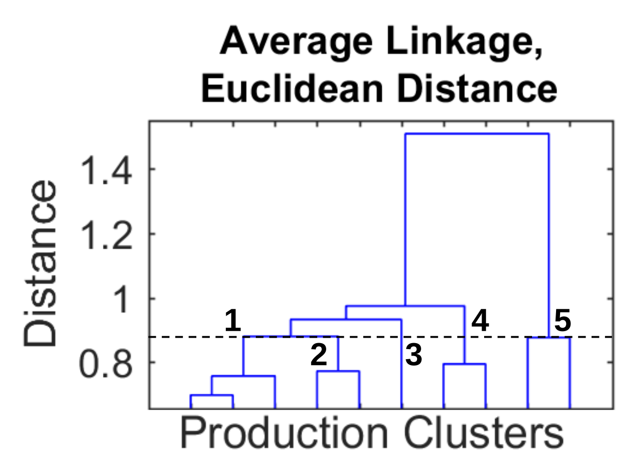


Figure 2: Dendrogram showing clusters found from PCs on trajectories of accurate productions. Cluster 5 is furthest from the other clusters, while clusters 1-3 are more similar.

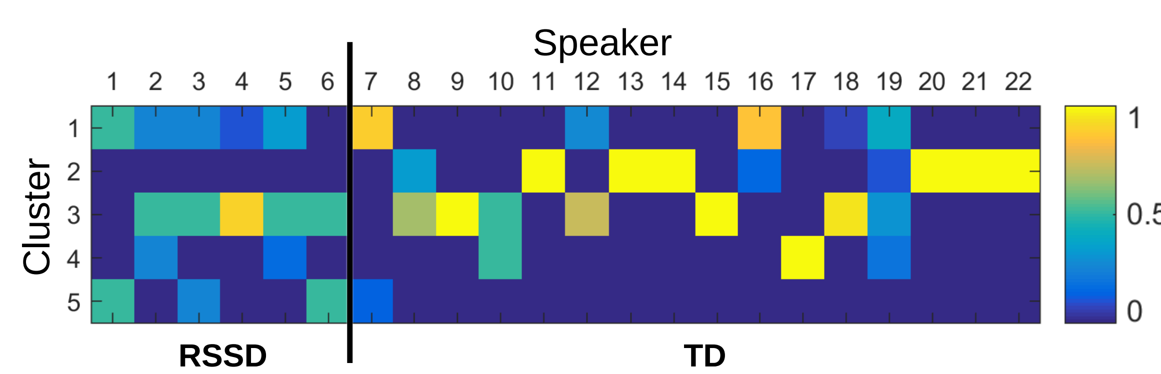
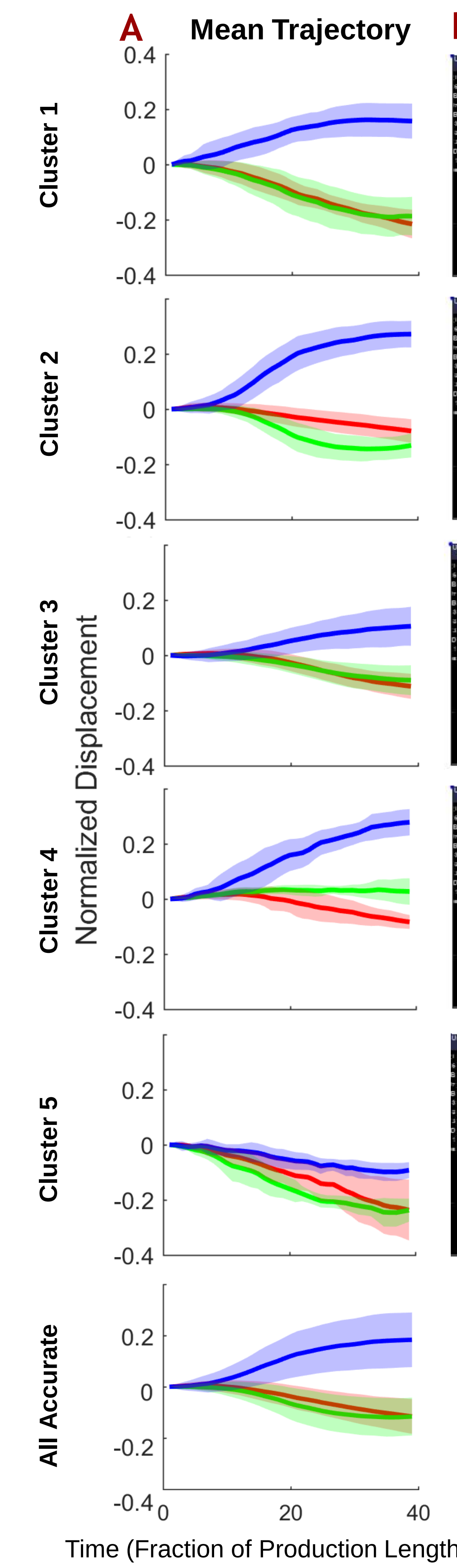


Figure 3: Proportion of accurate productions from each speaker belonging to each identified cluster. Cluster 5 is a pattern mostly used by RSSD speakers. Many TD but no RSSD productions are in cluster 2.



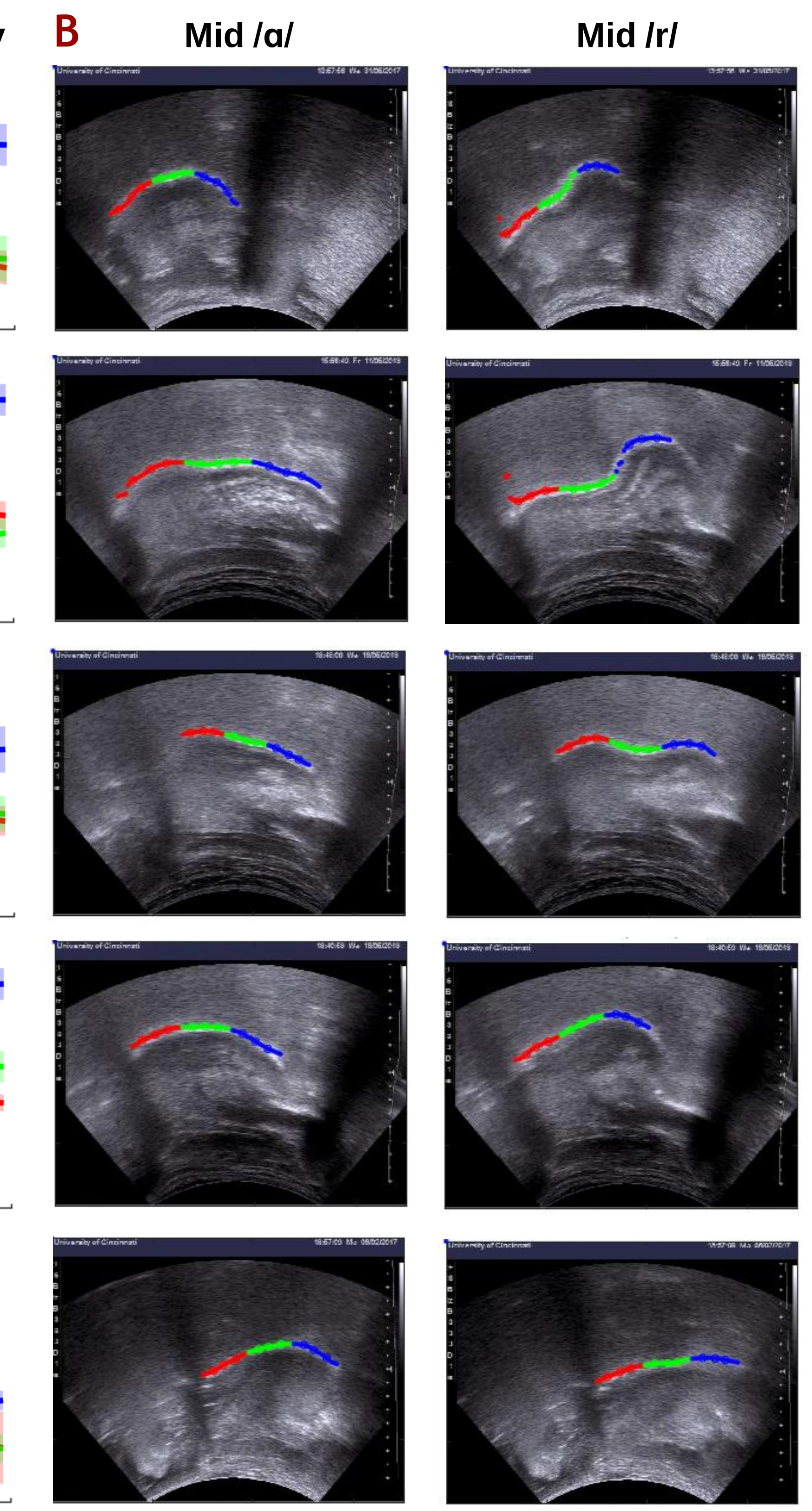


Figure 4: (A) Mean of trajectories within each cluster, banded by one standard deviation. (B) Example B-mode images at the midpoint of /ɑ/ and typical midpoint of /r/ (frame 32) from each cluster, with TonguePART tracking.

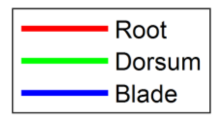


Table 1: Classification accuracy (cross-validation)

Model	Accuracy
1. Linear SVM on overall dorsum, blade displacement	81.39%
2. RBF SVM on correlation distance to strategy	80.87%
3. RBF SVM on correlation distance to mean of accurate	85.39%

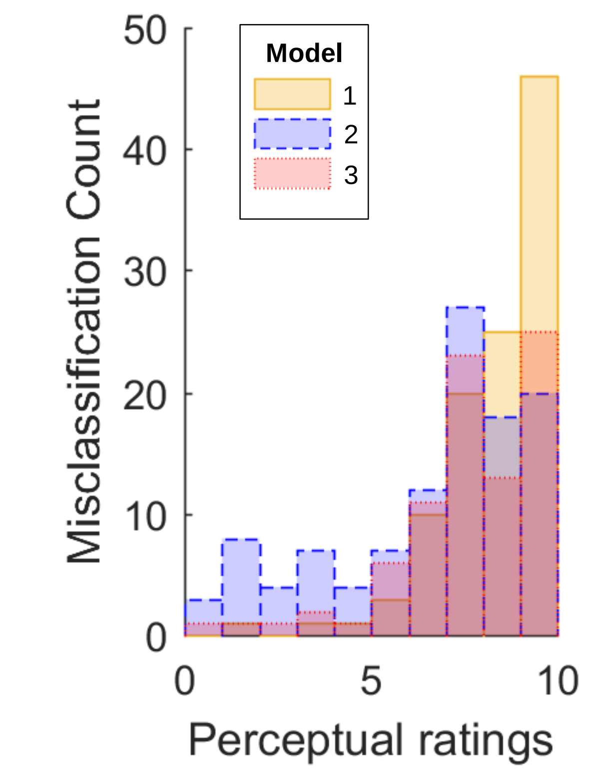


Figure 5: Histogram of misclassification counts and perceptual ratings.

Discussion

- Classification of accurate vs. error with TonguePART trajectories is achievable, supported by classification accuracy rates ≥ 80%.
- The 5 identified clusters have low time-dependent standard deviation of displacement, consistent with distinct strategies for achieving accurate /r/, beyond retroflex vs. bunched.
- Cluster 5 comprised mostly accurate RSSD productions, with trajectories close to some inaccurate RSSD productions. This may explain poorer classification using correlation to identified strategies (model #2).
- Except for model #1, misclassifications occur most commonly with borderline ratings such as 7, consistent with known patterns of listener judgments of accuracy [7].
- All 3 SVM models provided similar classification accuracy of 80%-85%.

Works Cited

[1] J. Preston *et al*., "Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors," *Journal of Visualized Experiments*, no. 119, 2017.
[2] J. Preston, L. Maas, J. Whittle, M. Leece and P. McCabe, "Limited acquisition and generalisation of rhotics with ultrasound visual feedback in childhood apraxia", *Clinical Linguistics & Phonetics*, vol. 30, no. 3-5, pp. 363-381, 2015. [3] L. Magloughlin, "Accounting for variability in North American English /ɹ/: Evidence from children's articulation", *Journal of Phonetics*, vol. 54, pp. 51-67, 2016.
[4] X. Zhou *et al*., "A magnetic resonance imaging-based articulatory and acoustic study of 'retroflex' and 'bunched' American English /r/," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4466–4481, 2008.
[5] J.R. Westbury *et al*., "Differences among speakers in lingual articulation for American English /ɹ/," *Speech Communication*, vol. 26, no. 3, 1998, pp. 203-226.
[6] F. Campbell, B. Gick, I. Wilson and E. Vatikiotis-Bateson, "Spatial and Temporal Properties of Gestures in North American English /r/", *Language and Speech*, vol. 53, no. 1, pp. 49-69, 2010.
[7] S. Dugan, N. Silbert, T. McAllister, J. Preston, C. Sotto and S. Boyce, "Modelling category goodness judgments in children with residual sound errors", *Clinical Linguistics & Phonetics*, vol. 33, no. 4, pp. 295-315, 2018.