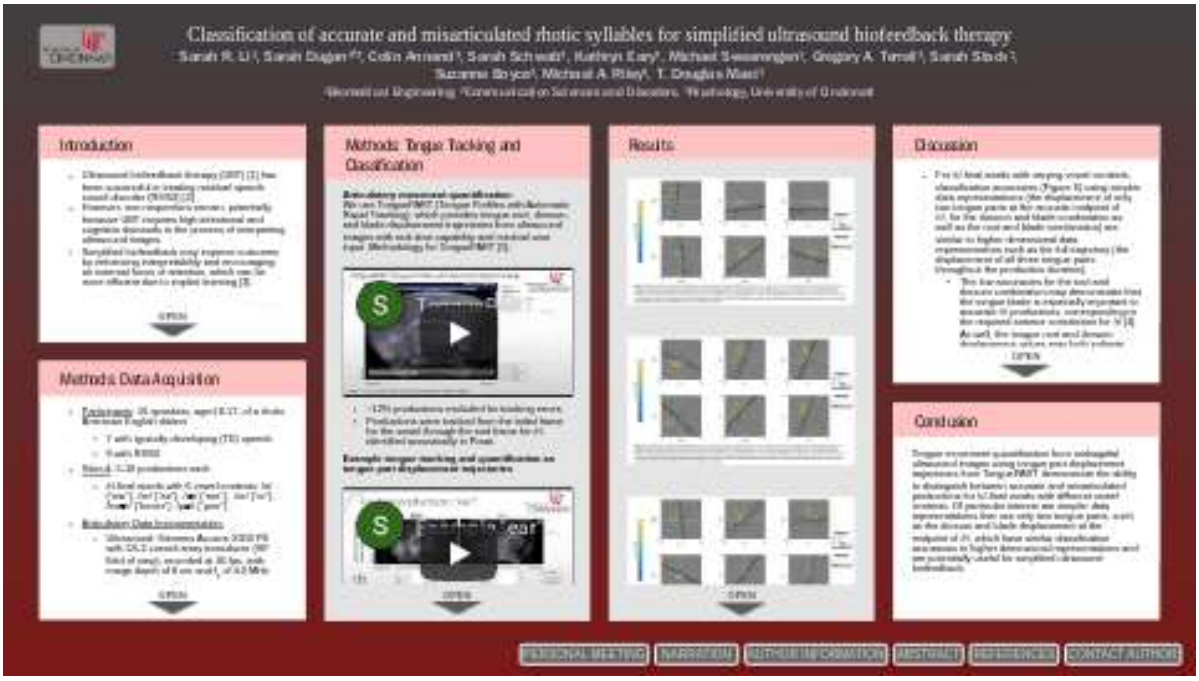
Classification of accurate and misarticulated rhotic syllables for simplified ultrasound biofeedback therapy



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

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

PRESENTED AT:

ACOUSTICS VIRTUALLY EVERYWHERE
ASA ACOUSTICAL SOCIETY OF AMERICA
179th Meeting
7 -11 December 2020

INTRODUCTION

- Ultrasound biofeedback therapy (UBT) [1] has been successful in treating residual speech sound disorder (RSSD) [2].
- However, non-responders remain, potentially because UBT requires high attentional and cognitive demands in the process of interpreting ultrasound images.
- Simplified biofeedback may improve outcomes by enhancing interpretability and encouraging an external focus of attention, which can be more efficient due to implicit learning [3].

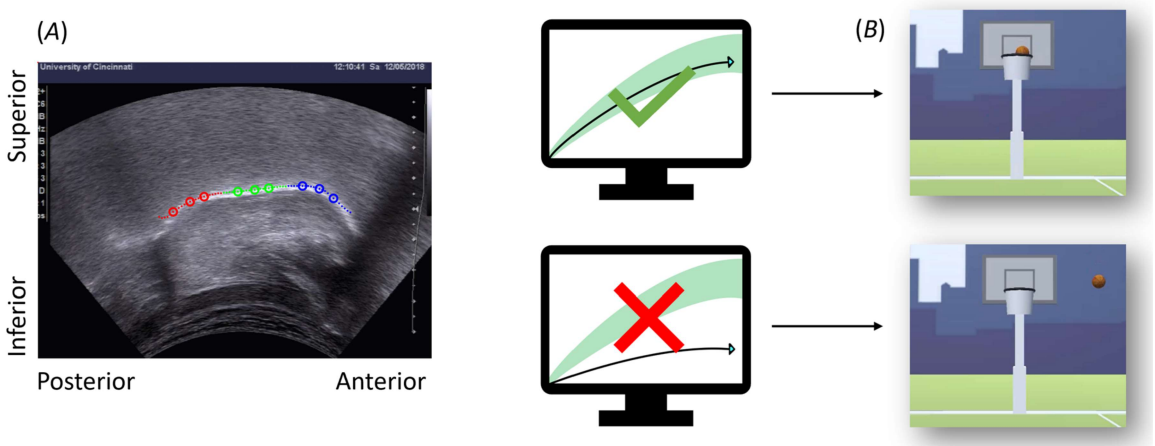


Figure 1: A visual example of simplifying biofeedback. Tongue movement (quantified from ultrasound images as shown in *A*) is instead represented as scoring baskets in a basketball game (*B*). If the movement is considered accurate, the basketball scores a basket, which is easily interpretable as correct. The distance between the basketball and basket would also illustrate how far the tongue movement is from the target accurate movement pattern.

- Simplifying biofeedback requires identifying biofeedback parameters (quantitative representations of tongue movement) that are easily interpretable (Figure 1) and can differentiate between accurate and misarticulated tongue movement for the syllables of interest: /r/-final words.
  - Different accurate configurations for /r/: bunched and retroflex [4]
- We have shown that accurate and misarticulated productions of /ɑr/ from children can be differentiated by using quantification from TonguePART, which quantifies tongue movement from midsagittal ultrasound images in real-time as tongue part displacement trajectories [5, 6]. We are interested in vowel contexts additional to /ɑ/.

Goal: Determine relevant articulatory features of accurate /r/-final syllables with different vowel contexts by comparing classifiers trained on tongue part displacement trajectory quantification. Simple data representations that can be easily interpreted are preferred as possible biofeedback parameters.

METHODS: DATA ACQUISITION

- Participants: 16 speakers, aged 8-17, of a rhotic American English dialect
  - 7 with typically-developing (TD) speech
  - 9 with RSSD
- Stimuli: 5-10 productions each
  - /r/-final words with 6 vowel contexts: /ir/ (“ear”), /er/ (“air”), /ɑr/ (“are”), /or/ (“or”), /buər/ (“booer”), /pɛr/ (“purr”)
- Articulatory Data Instrumentation:
  - Ultrasound: Siemens Acuson X300 PE with C6-2 curved array transducer (90° field of view), recorded at 36 fps, with image depth of 8 cm and f_c of 4.0 MHz
  - Head stabilizer [5]
- Acoustic Data Instrumentation: cardioid condenser USB microphone (Audio-Technica ATR2500-USB), recorded at 44.1 kHz
- Auditory Perceptual Ratings: Naive listeners were recruited and rated productions through an online experiment for a larger dataset; the /r/-final words used in this analysis made up ~17% of this analysis (464 out of 2764 productions). For ~81% of the full dataset, the average of 5-8 raters was used as the perceptual rating for a production. A verification subset (~19% of the full dataset) used the average of 47 raters; for these ratings, the intraclass correlation coefficient was 0.98, showing good interrater agreement (> 0.9).

METHODS: TONGUE TRACKING AND CLASSIFICATION

Articulatory movement quantification
We use TonguePART (Tongue Profiles with Automatic Rapid Tracking), which provides tongue root, dorsum, and blade displacement trajectories from ultrasound images with real-time capability and minimal user input. Methodology for TonguePART [5]:

[VIDEO] https://www.youtube.com/embed/1j29Yryv24c?rel=0&fs=1&modestbranding=1&rel=0&showinfo=0

- ~12% productions excluded for tracking errors
- Productions were tracked from the initial frame for the vowel through the end frame for /r/, identified acoustically in Praat.

Example tongue tracking and quantification as tongue part displacement trajectories

[VIDEO] https://www.youtube.com/embed/a8EFwAHmrcM?rel=0&fs=1&modestbranding=1&rel=0&showinfo=0

[VIDEO] https://www.youtube.com/embed/9MkBBc-QN-M?rel=0&fs=1&modestbranding=1&rel=0&showinfo=0

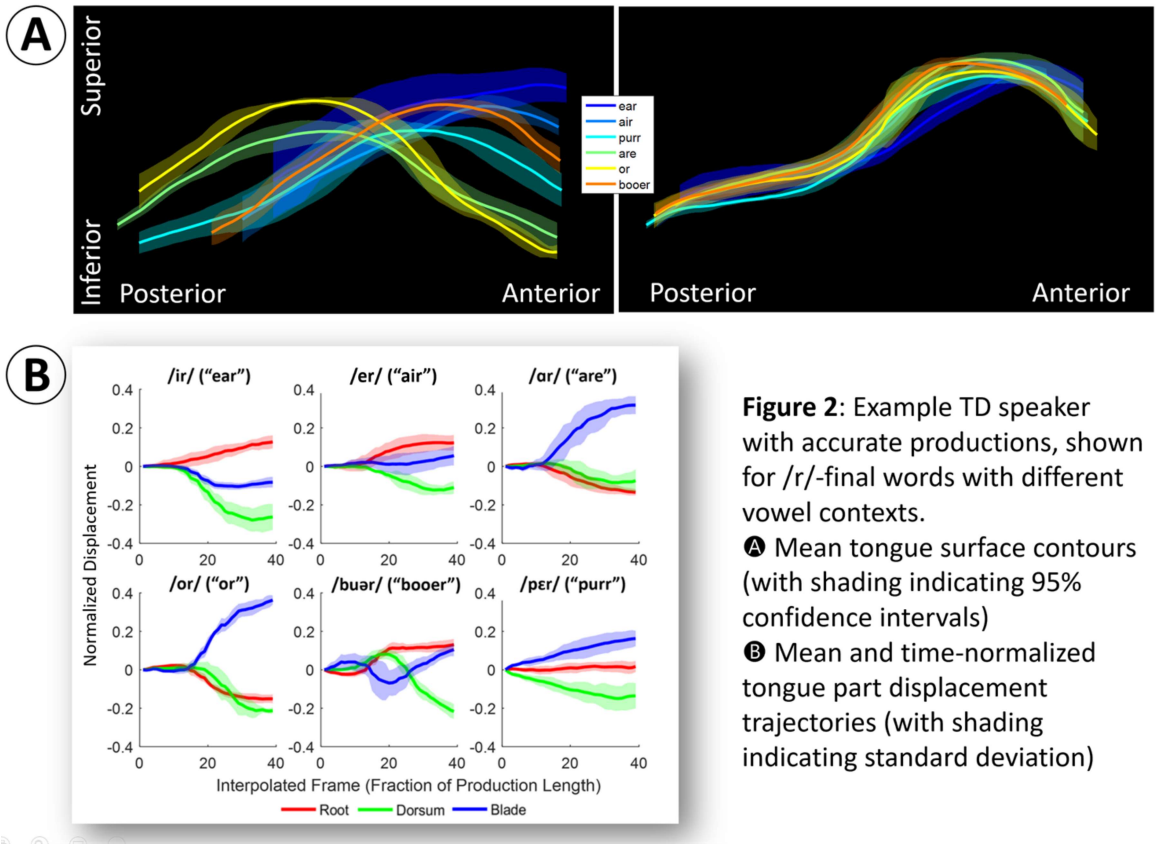


Figure 2: Example TD speaker with accurate productions, shown for /r/-final words with different vowel contexts.
❶ Mean tongue surface contours (with shading indicating 95% confidence intervals)
❷ Mean and time-normalized tongue part displacement trajectories (with shading indicating standard deviation)

Classification

We train linear support vector machine (SVM) classifiers (MATLAB) to predict whether a production is accurate or misarticulated using data representations from tongue part trajectories.

- Accurate production: average auditory perceptual rating > 5.5 (out of 10, selected due to highest classification accuracies on another dataset for "are")
- Data representations from tongue part trajectories (the midpoint of /r/ refers to the frame associated with the acoustic midpoint of /r/, representing the displacement between two frames: the initial frame for the vowel and midpoint frame for /r/):
  - Full trajectory
  - Root, dorsum, and blade displacement at midpoint of /r/
  - Dorsum and root displacement at midpoint of /r/
  - Dorsum and blade displacement at midpoint of /r/

- Root and blade displacement at midpoint of /r/

- 10-fold cross-validation
- The box constraint hyperparameter for SVM classifiers was optimized with Bayesian optimization for most classifiers and manually adjusted for a few classifiers.

RESULTS

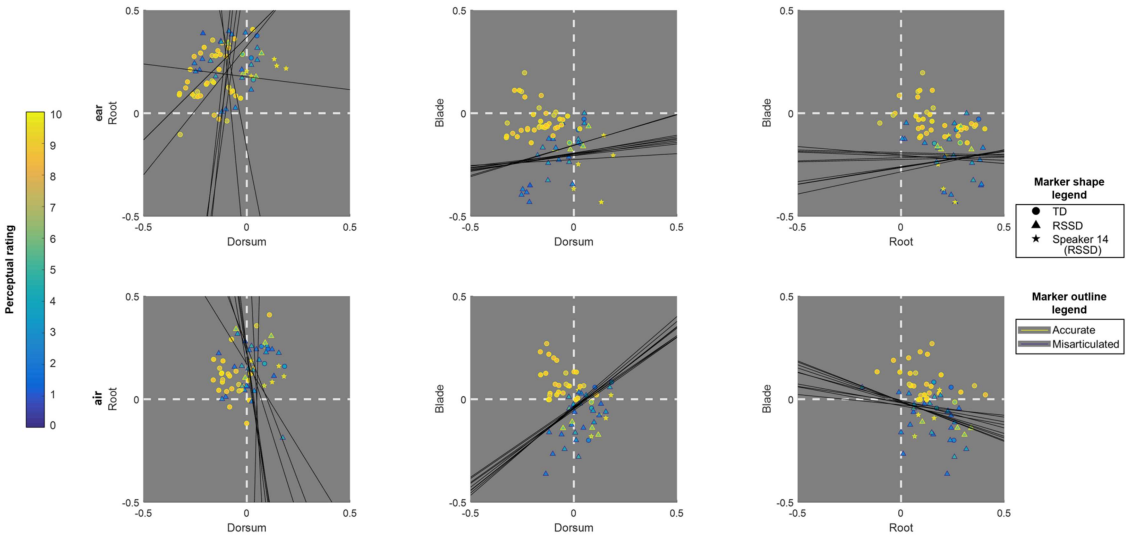


Figure 3: Scatterplots of tongue part displacement values at the acoustic midpoint of /r/ for “ear” (top row) and “air” (bottom row). Shading illustrates the average auditory perceptual rating, with outline color indicating whether the class label for the production was accurate (rating > 5.5) or misarticulated. Marker shape indicates speaker association; Speaker 14 is specified as an example pattern commonly misclassified for certain vowel contexts (“ear” and “air” in this figure). Black lines indicate the decision boundaries of the linear SVM classifiers trained on each fold for the associated tongue parts (e.g., dorsum and root displacement at the midpoint of /r/ for the plots on the left).

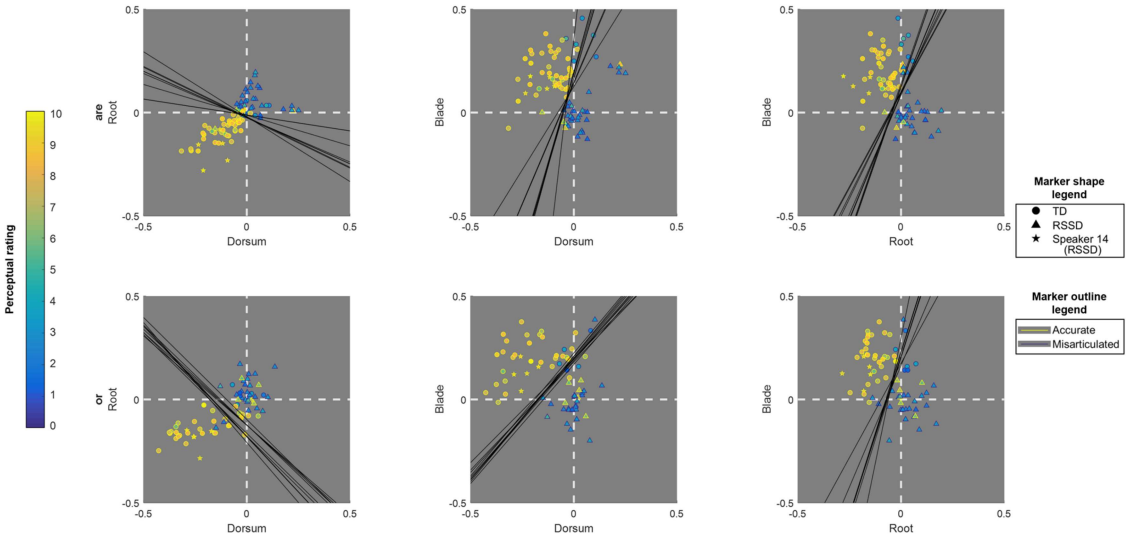


Figure 4: Scatterplots of tongue part displacement values at the acoustic midpoint of /r/ for “are” (top row) and “or” (bottom row). Shading illustrates the average auditory perceptual rating, with outline color indicating whether the class label for the production was accurate (rating > 5.5) or misarticulated. Marker shape indicates speaker association; Speaker 14 is specified as an example pattern commonly misclassified for certain vowel contexts (“ear” and “air” in Figure 3). Black lines indicate the decision boundaries of the linear SVM classifiers trained on each fold for the associated tongue parts (e.g., dorsum and root displacement at the midpoint of /r/ for the plots on the left).

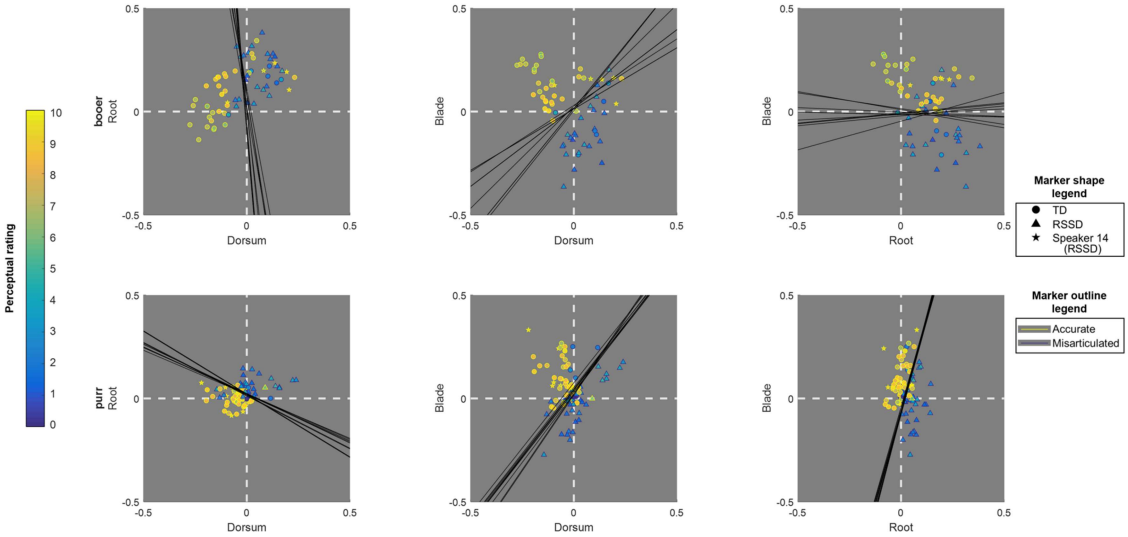


Figure 5: Scatterplots of tongue part displacement values at the acoustic midpoint of /r/ for "booer" (top row) and "purr" (bottom row). Shading illustrates the average auditory perceptual rating, with outline color indicating whether the class label for the production was accurate (rating > 5.5) or misarticulated. Marker shape indicates speaker association; Speaker 14 is specified as an example pattern commonly misclassified for certain vowel contexts ("ear" and "air" in Figure 3). Black lines indicate the decision boundaries of the linear SVM classifiers trained on each fold for the associated tongue parts (e.g., dorsum and root displacement at the midpoint of /r/ for the plots on the left).

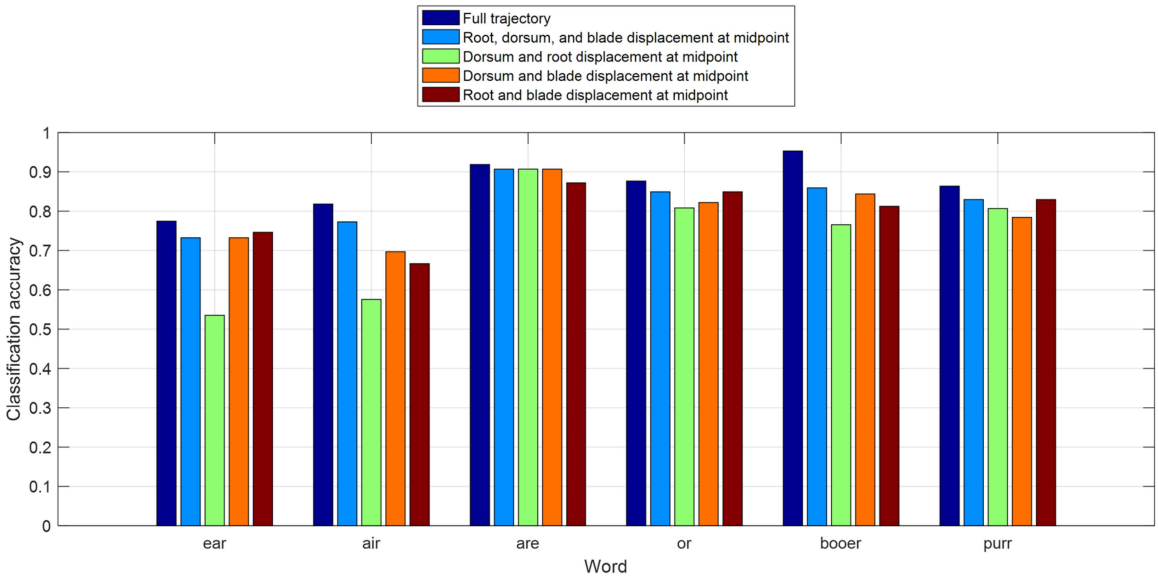


Figure 6: Classification accuracies for the classifiers trained on different data representations from tongue part displacement trajectories, indicated for each word.

DISCUSSION

- For /r/-final words with varying vowel contexts, classification accuracies (Figure 6) using simpler data representations (the displacement of only two tongue parts at the acoustic midpoint of /r/, for the dorsum and blade combination as well as the root and blade combination) are similar to higher dimensional data representations such as the full trajectory (the displacement of all three tongue parts throughout the production duration).
  - The low accuracies for the root and dorsum combination may demonstrate that the tongue blade is especially important to accurate /r/ productions, corresponding to the required anterior constriction for /r/ [4]. As well, the tongue root and dorsum displacement values may both indicate similar information, such as the posterior constriction for /r/.
  - Variability across vowel contexts
- Some potential reasons for misclassifications include:
  - Variation due to using naive raters. Some RSSD productions were rated only slightly above the rating threshold 5.5 (out of 10, considered accurate).
  - Methodology does not account for some tongue movement patterns (e.g., accurate productions from RSSD Speaker 14 for "ear" and "air" in Figure 3)
  - Variation due to imaging quality (while tracking errors were excluded, productions were included if they did not image a portion of the tongue root or tip).
- A range of typical tongue configurations (some more bunched, some more retroflex) are included in these high classification accuracies.
- These results indicate that a simplified biofeedback parameter based on the displacement of two tongue parts can be considered. An example using a linear combination of the dorsum and blade displacement (based on the linear SVM classifier) on recorded data from different speakers:

[VIDEO] https://www.youtube.com/embed/zMsMrBidUxA?rel=0&fs=1&modestbranding=1&rel=0&showinfo=0

CONCLUSION

Tongue movement quantification from midsagittal ultrasound images using tongue part displacement trajectories from TonguePART demonstrate the ability to distinguish between accurate and misarticulated productions for /r/-final words with different vowel contexts. Of particular interest are simpler data representations that use only two tongue parts, such as the dorsum and blade displacement at the midpoint of /r/, which have similar classification accuracies to higher dimensional representations and are potentially useful for simplified ultrasound biofeedback.

AUTHOR INFORMATION

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

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

lisr@mail.uc.edu

ABSTRACT

Ultrasound biofeedback therapy (UBT) provides real-time imaging of tongue movements and has demonstrated positive speech remediation outcomes; however, some individuals have limited or no response. UBT outcomes could be further improved by a simplified biofeedback display to enhance motor learning. Such simplification requires automatic processing of ultrasound images to determine biofeedback parameters and targets. We investigate potential biofeedback parameters using TonguePART, a method that automatically tracks the tongue surface on midsagittal ultrasound images to quantify displacement trajectories of the tongue root, dorsum, and blade. Our focus is rhotic syllables (/i/, /u/, /o/, /e/, /ɛ/, and /ɑ/ with final /r/) from children with residual speech sound disorders and children with typically-developing speech. We train support vector machines on measured tongue part displacement trajectories to distinguish between accurate and misarticulated productions as determined from auditory perceptual ratings. Preliminary data indicate that a linear combination of the tongue dorsum and blade displacements, between the vowel and consonant, can distinguish between accurate and misarticulated productions of rhotic syllables. These results suggest a real-time biofeedback parameter based on projections of real-time dorsum and blade displacements, along with potential target values, different for each vowel, for this parameter in simplified UBT for speech remediation.

REFERENCES

[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, doi: 10.3791/55123.

[2] 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.

[3] E. Maas et al., “Principles of Motor Learning in Treatment of Motor Speech Disorders,” *Am J Speech Lang Pathol*, vol. 17, no. 3, pp. 277–298, Aug. 2008, doi: 10.1044/1058-0360(2008/025).

[4] S. Boyce, “The Articulatory Phonetics of /r/ for Residual Speech Errors,” *Semin Speech Lang*, vol. 36, no. 04, pp. 257–270, Oct. 2015, doi: 10.1055/s-0035-1562909.

[5] 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, 2019, doi: 10.1044/2019_PERS-19-00064.

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