Comparing gamified and image-based ultrasound biofeedback therapy for speech remediation: Preliminary results

Sarah R. Li, PhD,[1] Sarah Dugan, PhD CCC-SLP,[2] Nicholas S. Schoenleb,[1] Maxim Lushpin,[1] Elizabeth Farmer, CCC-SLP, Brittany Fletcher, PhD CCC-SLP,[2] Mehraban Salaryfar,[2] Reneé Seward Nettle, MGD,[3] Michael A. Riley, PhD,[4] Suzanne Boyce, PhD CCC-SLP,[2] and T. Douglas Mast, PhD[1]

[1]Biomedical Engineering; [2]Communication Sciences and Disorders; [3]Design; [4]Rehabilitation Sciences | University of Cincinnati

Introduction

- In ultrasound biofeedback therapy (UBT) [1], ultrasound imaging is used to display the tongue shape and thus facilitate the learning of tongue movement patterns, such as for children with speech sound disorders.
- Despite accumulating evidence of UBT efficacy [2], non-responders remain (from 1 in 6 to 1 in 3 participants [3, 4, 5]).
  - One hypothesized reason for non-response is difficulty interpreting rapid, complex tongue movement from ultrasound images.
- A simplified, gamified presentation of UBT may elicit an external focus of attention on outcome (vs. internal on the tongue), demonstrated in non-speech motor control studies to improve motor learning [6].

Hypotheses

For a crossover (gamified, image-based) UBT experiment:
(***H1***) After the first block, participants receiving gamified UBT in the first block would have improved progress compared to image-based.
(***H2***) After both blocks, efficacy would be comparable to prev. UBT studies.

Methods: Analysis

Measures analyzed

- Perceptual ratings on audio recordings of word probes
  - Binary (misarticulated vs. accurate) from one clinician
  - Automatic rater trained on predicted % (0 to 1) of accurate ratings from previous perceptual experiment [8] and on PERCEPT-R [9] (Table 1)
  - Automatic ratings used, excluding productions with rating disagreement

Automatic rater

- Recurrent neural network, input of MFCCs after cutting audio to vowel + /r/ (identified with Montreal Forced Aligner [10])

Table 1: Performance of auto-rater

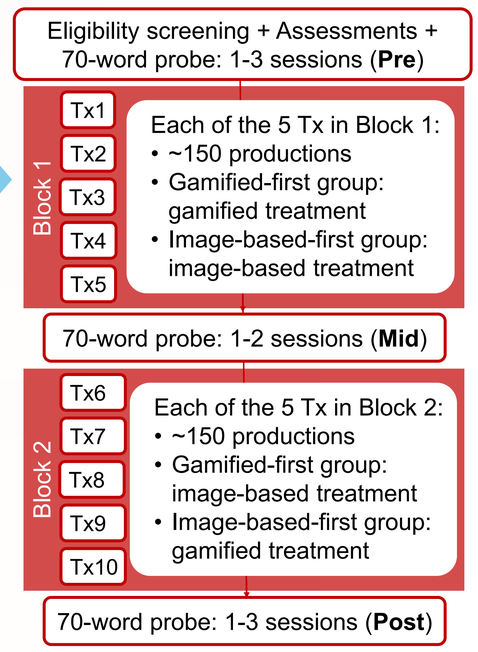
Test set	Acc. (%)	F1-score	ICC (A-1)
Previous experiment [8]	84.1	0.847	0.711
PERCEPT-R (20%)	85.6	0.789	0.794

Analyses performed

- Effect size comparisons (Cohen's *d* / Busk & Serlin's d_2 [11]: mean difference between participant's (***H1***) **Pre vs. Mid** or (***H2***) **Pre vs. Post** % accuracy, divided by pooled standard deviation)
  - Calculated separately for /r/-final and /r/-initial contexts
- Linear mixed models

Methods: Experimental Design

Treatment design



Each treatment (Tx) session:
- Every 10 productions, provide focusing or verbal cue
- Every 20 productions, switch target phonetic context when accuracy is > 80%

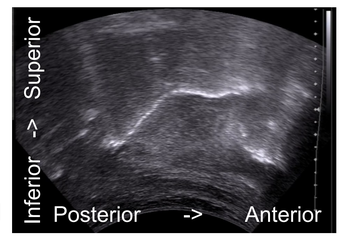
Participants

- 23 child (aged 7-15) speakers of American English with consistent /r/ misarticulations (> 20%)
- Most had previous or ongoing non-ultrasound therapy.

Image-based treatment

- "Traditional" UBT [1]
  - Participant views ultrasound image of tongue.
  - Clinician provides cues about tongue shape.
- Example cue used for attempt of /ir/ ("ear") (Fig. 1): "Leave that top lifted when you get to the back" (referring to keeping the anterior tongue up).

Fig. 1



Gamified treatment

- Tongue movement from ultrasound images translated to a gamified display (Fig. 2):
  1. Tongue surface is semi-automatically tracked in real time [7].
  2. Movement represented as tongue part displacements, with derived δ parameter (blade − dorsum) shown previously to represent production accuracy [7, 8].
  3. Derived parameters from displacements are mapped to the position of a game object (goat) during a game action (jump). Movement outcome is also displayed after production end (target turns green for accurate tongue movement).
     - Vertical goat position: δ parameter
     - Horizontal goat position: blade + dorsum displacement
- To emphasize external focus:
  - Participant does not view the tongue shape, only seeing the gamified display.
  - Clinician provides cues about the movement of the game object.
- Example cues
  - "Keep the goat's nose up in the target."
  - Used for production of "fear" (Fig. 2) : "Sounds good when you draw a fish" (referring to the shape of movement).

SCAN ME
More visuals

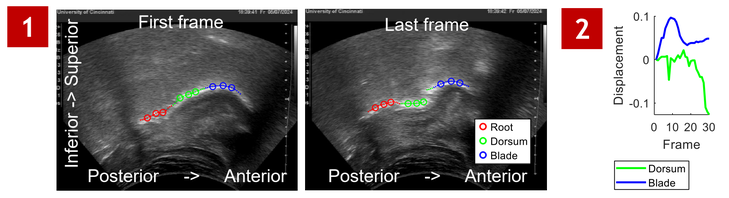


Fig. 2

Results

Overview

Table 2: Overall progress across participants: **Pooled percentage (%) of accurate /r/**

		Auto-rater (rating ≥ 0.5)	Clinician (binary)
Word probe session	Pre	10.0	17.1
	Mid	14.3	34.9
	Post	21.9	49.2

- 25.4% (2390 of 9419) ratings disagreed, excluded from rest of analysis.

Linear mixed models

- Response variable: *auto_rating*
- Fixed effects: *age*, *sex*, *session_time* (Pre vs. Mid or Pre vs. Post), interaction between *treatment_first* (gamified-first vs. image-based-first) and *session_time*
- Random effects: interaction between /r/ *position* (/r/-final vs. /r/-initial contexts) and *speaker*

(***H1***) No significant factors for model with **Pre** and **Mid**.
(***H2***) For the model only including **Pre** and **Post**, *session_time* was significant (β = −1.3 for Pre, p = 1e-6). This model also found the *treatment_first* interaction significant, which was not hypothesized (β = 0.7 for Pre and image-based-first, p = 0.042).

Model (Wilkinson notation): *auto_rating* ~ 1 + *session_time* + *age* + *sex* + *treatment_first*:*session_time* + (1 + *session_time* | *position*:*speaker*)

Effect sizes: After first block (Pre vs. Mid)

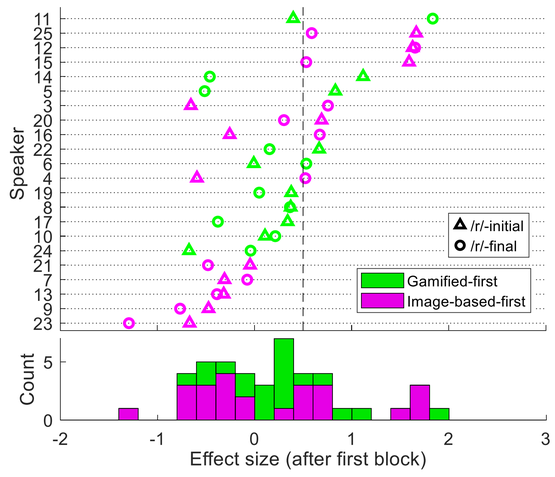


Fig. 3

(***H1***) t-test did not find a significant difference (*M* for gamified-first = 0.26, *M* for image-based-first = 0.18, p = 0.71 > α = 0.05).

Effect sizes: After both blocks (Pre vs. Post)

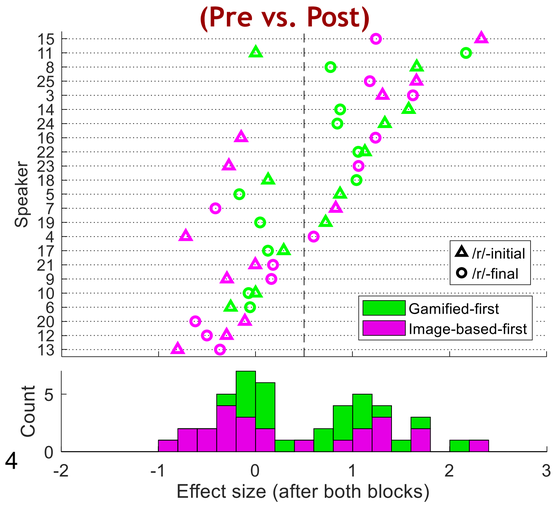


Fig. 4

(***H2***) 15 out of 23 participants had an effect size d_2 > 0.5 (Fig. 4)
- One of these participants only had modest absolute improvement (Speaker 4, Fig. 5).
- While mean effect size for gamified-first was greater (*M* = 0.64 compared to *M* = 0.37), t-test did not find a significant difference (p = 0.26 > α = 0.05).
  - Significant if not excluding disagreements with clinician ratings (*M* = 0.56 compared to *M* = 0.13, p = 0.028)

Details: Individual participant progress

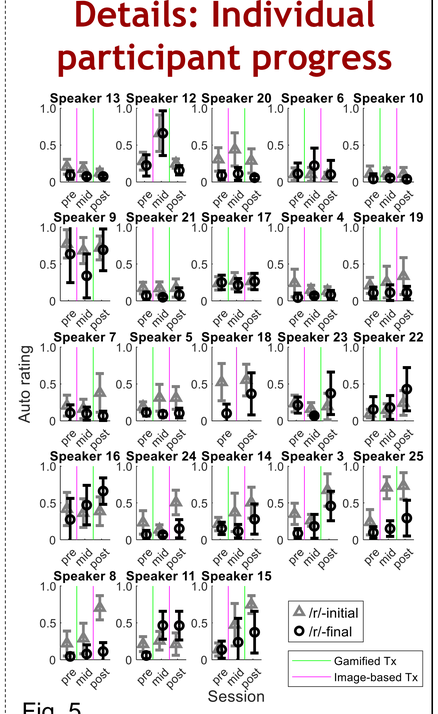


Fig. 5
- Markers and error bars represent mean and standard deviation.

Discussion and Conclusion

- (***H1***) Results do not support hypothesis. No differences were found for gamified vs. image-based treatment types after the first block.
  - Future analysis: differences in progress not generalized to word probes.
- (***H2***) Proportion of responders (~2/3) consistent with prev. UBT experiments.
- (Not hypothesized) Gamified-first may have slightly improved progress.
  - Emphasized with the more critical (fewer accurate ratings) auto-rater.

Overall, results indicate that gamified UBT is a valid method for administering UBT. Because differences between UBT types are slight, benefits to gamified may be more qualitative (e.g., younger age, ease of use or participant mood) or as "priming" for image-based UBT.

References

[1] J. L. Preston et al., "Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors," *J Vis Exp*, no. 119, p. e55123, Jan. 2017.
[2] E. Sugden et al., "Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders," *Int J Lang Commun*, vol. 54, no. 5, pp. 705–728, Sep. 2019.
[3] J. L. Preston et al., "Treatment for residual rhotic errors with high- and low-frequency ultrasound visual feedback: A single-case experimental design," *J Speech Lang Hear Res*, vol. 61, no. 8, pp. 1875–1892, Aug. 2018.
[4] T. McAllister Byun and H. Campbell, "Differential Effects of Visual-Acoustic Biofeedback Intervention for Residual Speech Errors," *Front Hum Neurosci*, vol. 10, Nov. 2016.
[5] N. R. Benway et al., "Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single-case randomization design," *Am J Speech Lang Pathol*, vol. 30, no. 4, pp. 1819–1845, Jul. 2021.
[6] G. Wulf, C. Shea, and R. Lewthwaite, "Motor skill learning and performance: a review of influential factors: Motor skill learning and performance," *Med Educ*, vol. 44, no. 1, pp. 75–84, Jan. 2010.
[7] S. R. Li et al., "Classification of accurate and misarticulated /ɑr/ for ultrasound biofeedback using tongue part displacement trajectories," *Clin Linguist Phon*, pp. 1–27, Mar. 2022.
[8] S. C. Biehl et al., "Optimization of classifying accurate and misarticulated speech sounds for use in a gamified real-time ultrasound biofeedback system," *J Acoust Soc Am*, vol. 155, no. 3_Supplement, pp. A335–A335, Mar. 2024.
[9] N. Benway et al.,"PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/," *Proc Interspeech*, pp. 3648–3652, Sep. 2022.
[10] M. McAuliffe et al., "Montreal forced aligner: Trainable text-speech alignment using kaldi," *Proc Interspeech*, pp. 498-502, Sep. 2017.
[11] P. M. Beeson and R. R. Robey, "Evaluating Single-Subject Treatment Research: Lessons Learned from the Aphasia Literature," *Neuropsychol Rev*, vol. 16, no. 4, pp. 161–169, Dec. 2006.

Acknowledgements

NIH/NIDCD grant R01 DC017301.
Siemens Medical Solutions for lending the Acuson X300 ultrasound scanner.

University of CINCINNATI