PERCEPT-US: A MULTIMODAL AMERICAN ENGLISH CHILD SPEECH CORPUS SPECIALIZED FOR ARTICULATORY FEEDBACK

Amanda Eads[1]; Heather Kabakoff[2]; Nina Benway[3]; Elaine Hitchcock[4]; Jonathan L. Preston[5]; Tara McAllister[1]

INTRODUCTION

- Speech Sound Disorder (SSD) can cause lasting academic and social challenges[1-3].
- Speech-language pathologists' (SLPs) large caseloads make it difficult to fully remediate SSD [4-5].
  - AI tools to extend SLP services could improve outcomes.
- Residual /ɹ/ distortions are common due to articulatory complexity[1].
- Current AI tools, e.g., PERCEPT[6,19], can classify /ɹ/ productions as accurate or inaccurate.
  - Clinically, we also want to provide tongue shape and articulatory cueing support.
- Tongue shapes for American English /ɹ/ sound alike[7] but differ in higher formants (lower for retroflex shapes)[11-13].
- We present the PERCEPT-US corpus—audio and ultrasound of /ɹ/— aiming to infer tongue shapes from acoustics to support clinical cueing.

DATA COLLECTION

Table 1: *Children in the PERCEPT-US corpus*

	Children with RSSD	Children without RSSD
Number of children	46	80
Females, Males	17, 29	41, 39
Mean age (SD)	10;7 (1;5)	12;7 (2;3)
Age range (yrs; mos)	9;0 - 14;8	8;8 - 17;8

Table 2: *Speech production tasks with counts of participants and utterance counts in parentheses. Word-TSC represents the tongue shape complexity task and Word-MP represents the minimal pair task.*

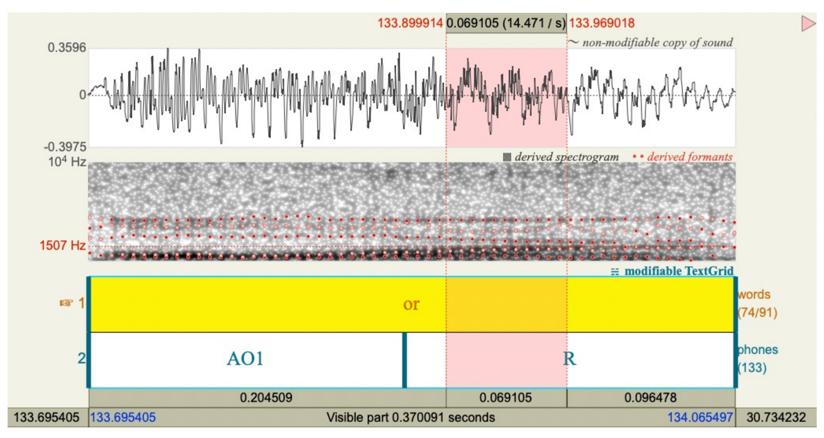
Speech Production Task	Children with RSSD	Children without RSSD
Syllable	39 (3240)	77 (3465)
Word - /ɹ/	46 (4000)	78 (3900)
Word - /s, z/	13 (754)	40 (2320)
Word - TSC	34 (1675)	38 (950)
Word - MP	13 (936)	41 (2952)
Sentence	13 (65)	78 (390)
Total Utterances:	**10,670**	**13,977**

- IRB-approved data collection at Haskins Labs, NYU, Montclair State U, and Syracuse U.
- Audio and midsagittal ultrasound were recorded simultaneously.
- 126 participants (mean age = 11;10, range = 8;8–17;8; 58 female, 68 male).
- Similar in age, American English rhotic dialect, and absence of speech-language-hearing differences other than residual speech sound disorder (RSSD).
  - Children with RSSD participated in our previous biofeedback speech therapy studies.
- Corpus combines four studies that differ in head–transducer alignment techniques.
- Current classification is on perceptually accurate data only (69 speakers, 2,385 utterances)
  - Tongue shape coding for perceptually inaccurate utterances is in progress.
- Total corpus size compares favorably to previous multimodal child speech corpora[10].

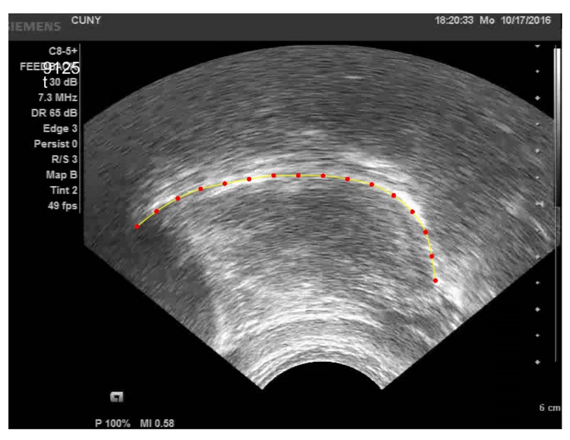
DATA PREPARATION & ANNOTATION

Preparation:

- Original MP4 and MKV screen-recorded files converted to lossless WAV and MP4 (H.264 + MP3).
- Praat[17] TextGrid files created with orthographic transcriptions for all audio.
- Montreal Forced Aligner[18] and PERCEPT pre-trained acoustic model v 3.0[19] used to segment phone boundaries.
  - Hand correction of boundaries where needed.

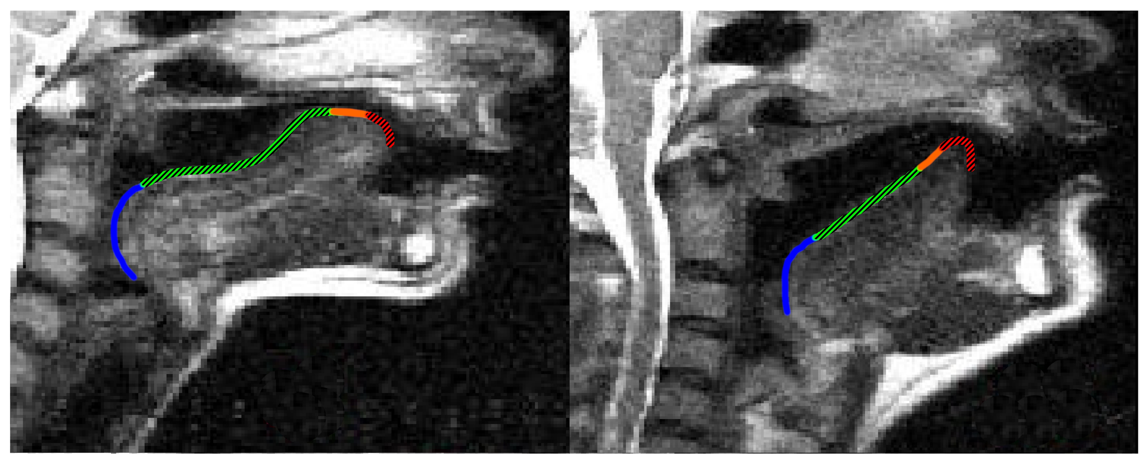


Transcription and Segmentation



Ground Truth Coordinates

Traced frames within /ɹ/ intervals using the MATLAB program *GetContours*[20-21].



Ground Truth Tongue Shape Classification

Flowchart from [24] used to classify tongue shapes for /ɹ/ into five categories, subsequently collapsed to bunched (tip down) versus retroflex (tip up) variants.

Ground Truth Ratings

Perceptual accuracy from expert and/or crowdsourced listeners[22].

CORPUS DEMONSTRATION

Does binary tongue shape variant (bunched-retroflex) predict mean F3-F5 Hz in the subset of the English /ɹ/ syllable task with ground truth labels for both perceptual ratings and tongue shape variant?

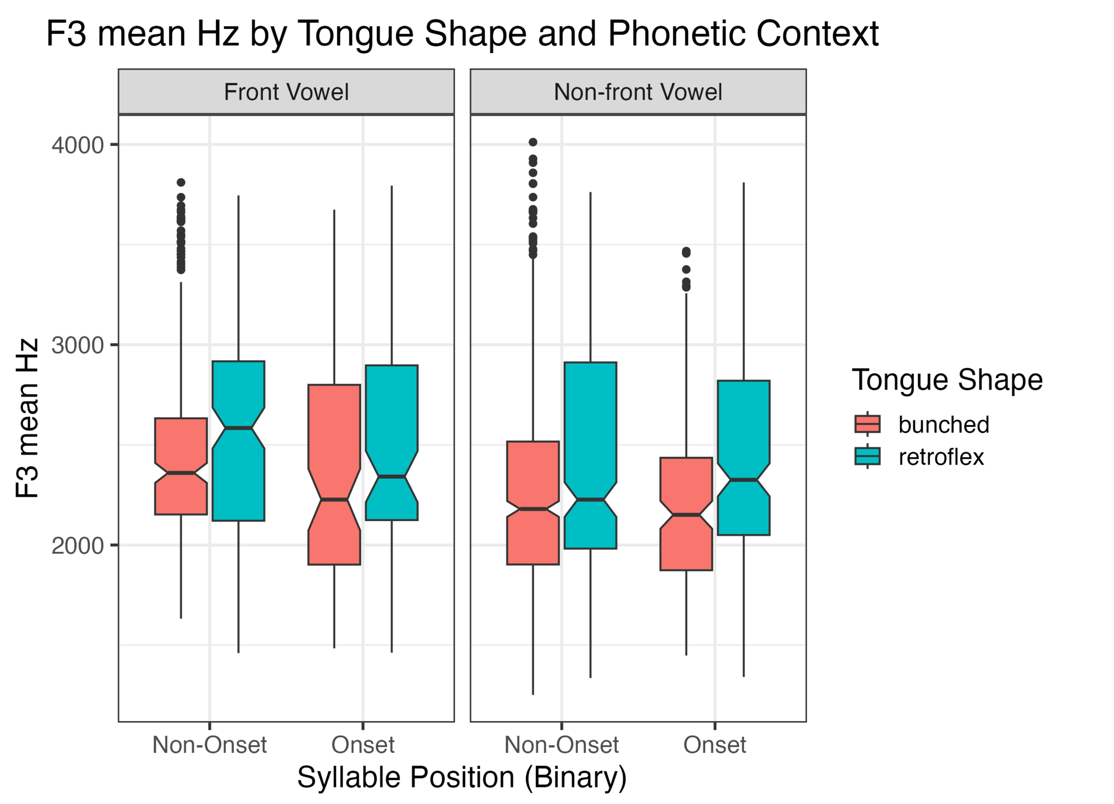
Acoustic Measurement:

- Adapted the Voweltine script[28] to select best formant settings per token/speaker and measure F1-F5
- Extracted mean Hz value from the steadiest 25 milliseconds

Analysis:

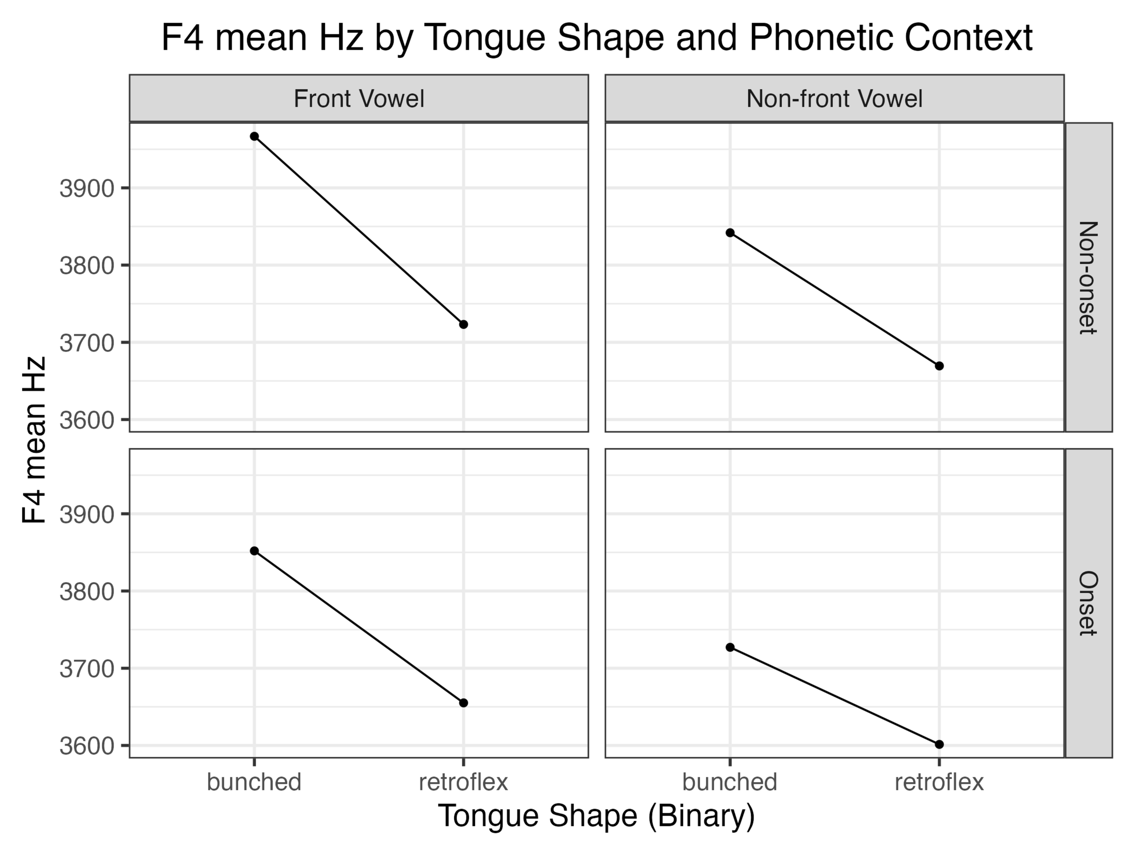
- Outliers excluded (1,966 paired observations)
- 3 Mixed-effects linear regression models [29-34]

lmer(**f3meanhz** ~ tongueshape_binary * (onset + vowel_binary) + Sex * Age + group + (1 + onset + vowel_binary | participant) + (1 | word)



F3 tongue shape differences by vowel context.

- Significant effect of vowel context ($p < 0.001$) with F3 lowering in non-front vowel contexts.
- Significant interaction between tongue shape variant and vowel context ($p < 0.01$) where non-front vowels + retroflex shape resulted in higher F3.



Retroflex tongue shapes have lower F4 Hz values.

- Significant effect of tongue shape ($p<0.02$);
- Significant effect of vowel context ($p= 0.01$).

CONCLUSIONS & FUTURE DIRECTIONS

Conclusions:

- Acoustic differences between tongue shape variants must be considered in relation to syllable position and vowel context!
  - Statistically significant findings for F3 and F4, but not F5.
  - Lower average F4 mean Hz in retroflex shapes aligns with previous work.[11-13]

Future Directions:

- Ground Truth labelling of perceptually inaccurate /ɹ/
- Releasing the corpus
- Developing the acoustic classifier

This research was supported by NIH NIDCD:
- F31DC022514 (PI: Eads)
- F31DC018197 (PI: Kabakoff)
- R01DC013668 (PI: Whalen)
- R01DC017476 (PI: McAllister).

Additional support provided by the American Speech-Language-Hearing Foundation, the Acoustical Society of America, and the Council of Academic Programs in Communication Sciences and Disorders.

If you have any questions and/or feedback, please contact Amanda Eads at are326@nyu.edu