Performance of forced-alignment algorithms on children's speech

Tristan Mahr [1], Kan Kawabata [2], Vikram C. Mathad [2], Visar Berisha [2], Sharon Tang [1], Helen Vradelis [1], Julie Liss [2], Katherine Hustad [1]
[1] University of Wisconsin–Madison. [2] Arizona State University.

Background

- To perform acoustic measurements on speech sounds, recordings must be segmented into separate intervals for individual phones.
- This process is time-consuming, so forced-alignment algorithms can automate this task. These algorithms use a speech sample, transcript, pronunciation dictionary, and statistical model of acoustic patterns to create (*force*) an alignment of phone labels and audio intervals.
- These *aligners* are validated against adult speech corpora.
- However, there are physiological, articulatory, and acoustic differences between adult speech and child speech, so we cannot assume forced alignment will work on child speech.

Current study

- Which of four available aligners performed best, compared to manual alignment, on samples from 3–6-year-old children?

Method

- **Participants.** Speakers were 42 typically developing 3–6-year-olds (39–83 months; 20 boys, 22 girls).
- **Task.** Speech samples were collected in a structured repetition task based on the TOCS+ (Hodge & Daniels, 2007). Prompts included 40 single words and 2–7-word utterances with 10 prompts per utterance length.
- *We omitted plosives from our analyses due to inconsistencies between human raters.*

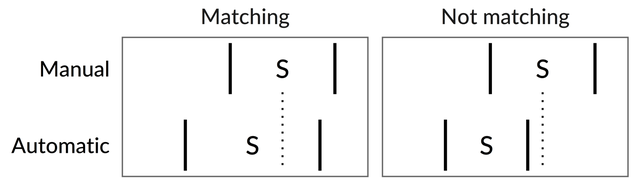
Aligners

All automatic aligners were used off-the-shelf with adult speech acoustic models.

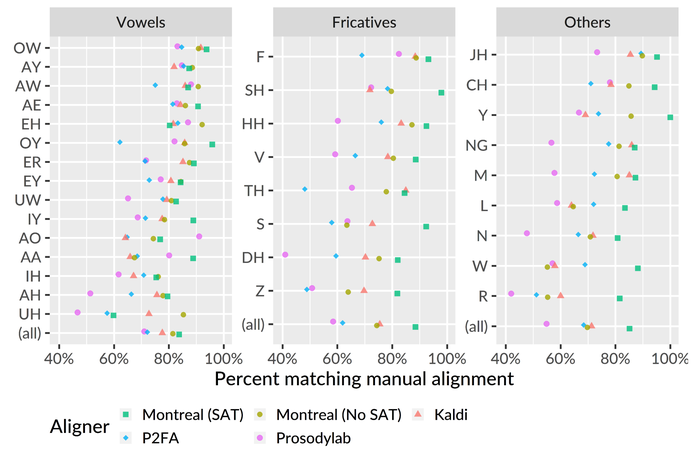
- Manual: Two human aligners corrected alignments from the Prosodylab aligner.
- Montreal Forced Aligner with/without speaker adaptation training (SAT) (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017)
- Kaldi: Alignments from the Kaldi speech recognition system (Povey et al., 2011)
- Prosodylab: Prosodylab Aligner (Gorman, Howell, & Wagner, 2011)
- P2FA: Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008)

Accuracy

- An aligner's interval *matches* the gold standard if the boundaries of an automatic alignment interval contain the midpoint of the manual one (Knowles, Clayards, & Sonderegger, 2018).

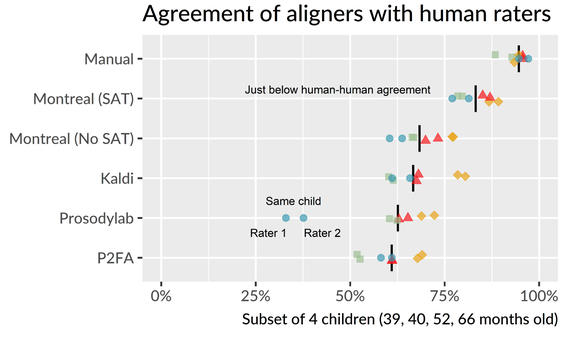


- Montreal with speaker adaptation had the best performance, including some large gains on individual phones (/r/, /s/, /w/).



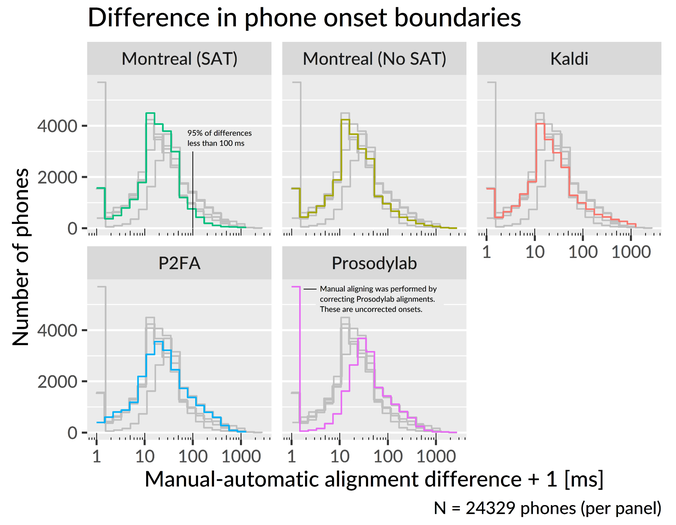
Interrater Agreement

- Two human raters each aligned the same 4 children.
- For each child, how well does Rater 1 match Rater 2 and how well does Rater 2 match Rater 1?
- Now, substitute aligners for each rater.
- Montreal with speaker adaptation approached human-level reliability.



Onset differences

- We also measured absolute difference in phone onset times between automatic and manual alignments.



- Bulk of differences for all aligners is between 10–50 ms.
- This analysis reveals a difference between the Montreal and Kaldi aligners. The right tails of the distributions show Montreal had the fewest differences over 100 ms and that Kaldi had the most differences around 500 ms.
- Although Kaldi and Montreal performed best in general, Kaldi demonstrated more severe alignment errors than Montreal.

Conclusions

- Speech recognition-based aligners (Kaldi, Montreal) performed better on children's speech off-the-shelf than Hidden Markov Model Toolkit (HTK) based ones (P2FA, Prosodylab).
- Speaker adaptation made large improvements to developmentally variable sounds (/s/, /r/, /w/, /ʃ/)
- Montreal-human agreement fell just below human-human agreement. Training the acoustic model on children's speech may narrow the gap further.

Funding provided by NIDCD (R01DC015653, R01DC006859) and NICHD (U54 HD090256)
Conference on Motor Speech
Santa Barbara, CA: February, 2020
Contact: Tristan Mahr (tristan.mahr@wisc.edu)