Viability of forced alignment for segmentation of dysarthric speech

Thea Knowles[1], Jeff Berry[2], Maura Philippone[1]
[1]Communicative Sciences & Disorders, Michigan State University
[2]Speech Pathology & Audiology, Marquette University

BACKGROUND

- **Acoustic-phonetic segmentation of dysarthric speech is very challenging.**
- Segmentation relies on precise identification of phonemic boundaries. Manual segmentation can be time-consuming and create reliability challenges that limit segmental analyses of larger datasets, especially in connected speech.
- **Automatic forced alignment** can be used to automatically segment words and phones in speech by predicting temporal boundaries, given an orthographic transcription and a trained acoustic model.
- Automatic forced alignment is highly effective for non-disordered adult speech and shows viability with highly variable child speech (Knowles et al., 2018; Mahr et al., 2022). However, forced alignment in dysarthric speech has received less attention.

Previous key findings on use of forced alignment on variable speaker populations:

- Training specifically on the speech-to-be-aligned resulted in improved alignment of child speech in one trainable aligner (Knowles et al., 2018; Gorman et al., 2011).
- More target-like phones aligned with better accuracy, e.g., shorter, adult-like /s/ vs. longer /s/were more accurately aligned in child speech (Knowles et al., 2018)
- The Montreal Forced Aligner with Speaker Adaptation (McAuliffe et al., 2017) outperformed other force-alignment technology in child speech (Mahr et al., 2022).

Purpose

- Evaluate efficacy of automatic forced alignment of vowels in a passage read by speakers with and without dysarthria.
- Evaluate an initial set of factors impacting alignment accuracy of dysarthric speech: **Acoustic model, vowel class, vowel duration**

METHODS

Speakers:

- **Speakers with dysarthria**: n = 5, aged 24 – 53 (3f, 2m), Midwest English speakers. Mixed dysarthria secondary to brain injury/stroke as well as mild-mod expressive aphasia/?apraxia of speech. Speech characterized predominantly by *articulatory imprecision, phoneme distortions, and reduced rate.*
- **Controls**: n = 5, aged 20 – 22 (3f, 2m).

Speech stimuli: Caterpillar Passage manually divided into 17 utterances.

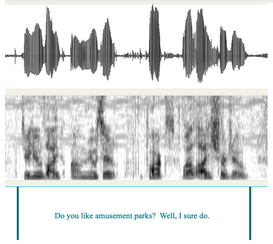
Manual vowel segmentation: All measurable occurrences of corner vowels and diphthong /ai/, resulting in ~30 instances of /ai/ & ~50 instances of corner monophthong vowels measured per speaker by trained research assistants.

Automatic Forced Alignment Details

- **Montreal Forced Aligner** (McAuliffe et al., 2018)
- **Acoustic models**: 1) *Default US English* (ARPA; trained on 982 hours of US English from 2000+ speakers in LibriSpeech corpus) with no speaker adaptation; 2) Speaker-Adapted US English, or 3) Retrained on speech-to-be-aligned.
- **Pronunciation dictionary**: US ARPA
- **Accuracy**: Did the midpoint of the aligned phone "match up" with the manual alignment? (%-Match, Knowles et al., 2018)

Statistical analysis: 2 generalized linear mixed effects regression to model 1) effect of acoustic model, group, and vowel class and 2) group, vowel class, and vowel duration within the best performing alignment.

FORCED ALIGNMENT: INPUT INGREDIENTS



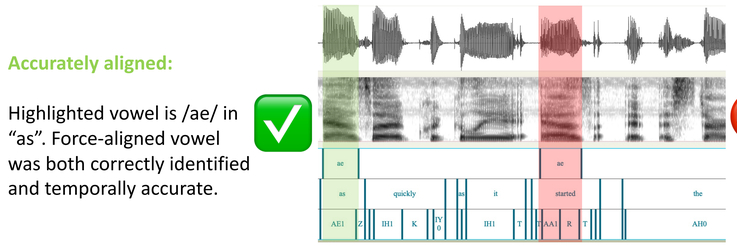
	Speech-to-be-aligned	Segmented into equal utterances & orthographically transcribed (Caterpillar Passage).
	Acoustic model	1. Default, no speaker adapt (SA) 2. Default, with SA 3. Retrained on speech-to-align
	Pronunciation dictionary (Arpabet)	Look-up word-phone key for standard US English pronunciations

RESULTS

FORCE-ALIGNED OUTPUT: Examples

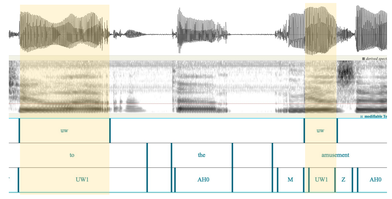
Accurately aligned:

Highlighted vowel is /ae/ in "as". Force-aligned vowel was both correctly identified and temporally accurate.



Inaccurately aligned:

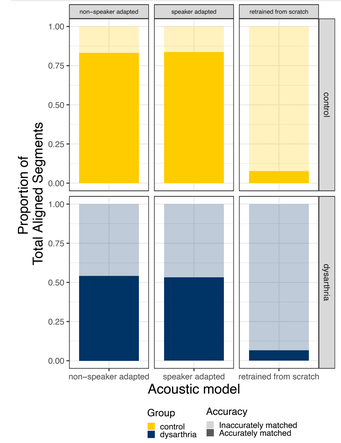
Highlighted vowel is /ae/ in "as", but aligner has gotten tripped up due to schwa-insertion earlier in the phrase and has incorrectly identified this as /ar/.



Vowel accurately aligned but temporally imprecise:

- In both cases /u/ is correctly identified by the aligner.
- In the first case an already long /u/ is force-aligned with a correct onset but an offset that includes disfluencies.
- In the second case the vowel is more or less accurately aligned but starts and ends a bit off from the manual annotated boundaries.

Acoustic Model

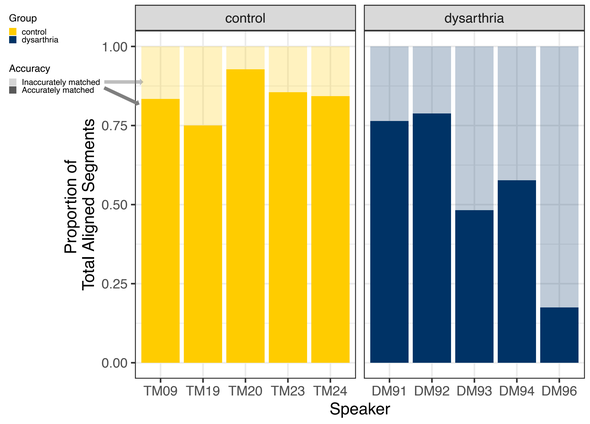


Pretrained > Retrained
Retraining on the speech-to-be-aligned was MUCH worse than either of the pretrained models, likely due to the small size corpus of heterogeneous speech.

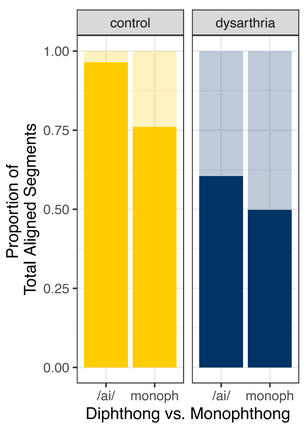
Speaker adaptation?
Speaker adaptation didn't improve (or worsen) alignment accuracy compared to the default un-adapted model, likely again due to small n of utterances.

Accuracy within the speaker-adapted model…

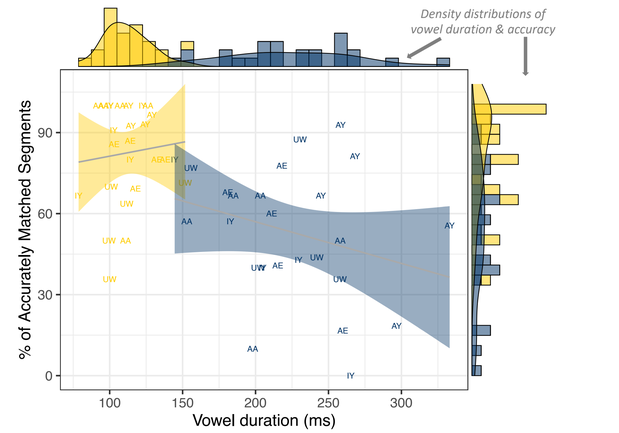
Control group > Dysarthria group



/ai/ > corner monophthongs



Longer vowels more accurately aligned for controls, less accurately for dysarthria



Key Take-Aways and Practical Tips

- **Better forced alignment accuracy as expected for control speakers vs. speakers with dysarthria.**
- **Better accuracy for diphthong /ai/ vs. corner vowel monophthongs for both groups.**
- **Longer vowels aligned with better accuracy for controls, and worse accuracy for dysarthria.**
- Retraining on the speech-to-be aligned → worse accuracy in this small corpus: Inconsistent with Knowles et al., 2018. Likely too little speech (<30 minutes) that was too heterogeneous?
- Speaker adaptation wasn't noticeably different than the un-adapted model. Inconsistent with Mahr et al., 2022. Speaker adaptation may work better for larger corpora (McAuliffe, personal comm).
- Phones that were more "typical" (in duration, at least) were force aligned with better accuracy (consistent with Knowles et al., 2018 for /s/ in child speech).
- Speaker with lowest intelligibility aligned with poorest accuracy (DM96), but variability in others.

Next steps: Trialing different, larger, systematically varied training data/transcription protocols and exploring specific dysarthric speech features and speech stimuli.

Practical tips for using forced alignment dysarthric speech:

- **Use it to facilitate, not replace manual segmentation:** Automatic forced alignment can be used as a tool to *facilitate* segmentation of dysarthric speech, but *cannot yet be relied upon unsupervised*. Use it as a first pass to speed up and facilitate the process.
- **Best used on shorter utterances:** When possible and practical, the aligner tends to do better with shorter versus longer phrases.
- **Transcribe deviations from orthography:** If a speaker deviates from the text, this will have a potentially large impact on accuracy. Revised orthographic transcription will help achieve better accuracy.

www.theaknowles.com
thea@msu.edu
@theaknowles
Email for references