JSLHR

Research Article

Performance of Forced-Alignment Algorithms on Children's Speech

Tristan J. Mahr,[a] Visar Berisha,[b] Kan Kawabata,[b,c] Julie Liss,[b] and Katherine C. Hustad[a,d]

Purpose: Acoustic measurement of speech sounds requires first segmenting the speech signal into relevant units (words, phones, etc.). Manual segmentation is cumbersome and time consuming. Forced-alignment algorithms automate this process by aligning a transcript and a speech sample. We compared the phoneme-level alignment performance of five available forced-alignment algorithms on a corpus of child speech. Our goal was to document aligner performance for child speech researchers.
Method: The child speech sample included 42 children between 3 and 6 years of age. The corpus was force-aligned using the Montreal Forced Aligner with and without speaker adaptive training, triphone alignment from the Kaldi speech recognition engine, the Prosodylab-Aligner, and the Penn Phonetics Lab Forced Aligner. The sample was also manually aligned to create gold-standard alignments. We evaluated alignment algorithms in terms of accuracy (whether the interval covers the midpoint of the manual alignment) and difference in phone-onset times between the automatic and manual intervals.
Results: The Montreal Forced Aligner with speaker adaptive training showed the highest accuracy and smallest timing differences. Vowels were consistently the most accurately aligned class of sounds across all the aligners, and alignment accuracy increased with age for fricative sounds across the aligners too.
Conclusion: The best-performing aligner fell just short of human-level reliability for forced alignment. Researchers can use forced alignment with child speech for certain classes of sounds (vowels, fricatives for older children), especially as part of a semi-automated workflow where alignments are later inspected for gross errors.
Supplemental Material: https://doi.org/10.23641/asha.14167058

Research on children's speech production requires analyses on large corpora due to the developmental and individual variability in production. The traditional workflow for acoustic analysis at the phoneme level—manually annotating recordings for words and phonemes as well as repeatedly playing segments of audio and tweaking boundaries—is time consuming. For example, 1 min of annotation for a 2-s token (a conservative estimate) would represent an annotation to speech duration ratio of 30. As phonetic corpora grow ever larger, manual annotation cannot scale for these larger scale data sets. We posit that manual annotation is a rate-limiting factor in gaining a deep understanding of the ways in which phoneme production develops in children.

In this work, our aim is to evaluate the accuracy of several *forced-alignment* algorithms that automatically map the words and phones onto intervals of speech by comparing the algorithms to trained human aligners. These algorithms use models similar to those used in speech recognition (a pronunciation dictionary of words and a statistical model of acoustic patterns), a speech sample, and a transcript of what was said in the sample to create (*force*) an alignment of phone labels and audio intervals. In acoustic–phonetics, we routinely annotate/segment recordings into meaningful intervals (turns, utterances, words, phones, etc.) and take measurements of those intervals (durations, frequencies, etc.). In this respect, alignment is fundamental for downstream analysis of speech data, both at the segmental level (e.g., measuring the spectrum of a particular fricative) and at the suprasegmental level (e.g., measuring speech rate;

[a]Waisman Center, University of Wisconsin–Madison
[b]Department of Communication Sciences and Disorders, Arizona State University, Tempe
[c]Aural Analytics, Inc., Scottsdale, AZ
[d]Department of Communication Sciences and Disorders, University of Wisconsin–Madison
Correspondence to Tristan J. Mahr: mahr@wisc.edu

Editor-in-Chief: Cara E. Stepp
Editor: Erika S. Levy
Received May 18, 2020
Revision received October 30, 2020
Accepted November 24, 2020
https://doi.org/10.1044/2020_JSLHR-20-00268
Publisher Note: This article is part of the Special Issue: Select Papers From the 2020 Conference on Motor Speech.

Disclosure: *Visar Berisha and Julie Liss are cofounders and have equity in Aural Analytics, Inc. The other authors have declared that no other competing interests existed at the time of publication.*

Shinozaki & Furui, 2003; Tu et al., 2018; Yuan & Liberman, 2011). Therefore, algorithms that accurately align children's speech would enable researchers to answer a variety of questions on a data scale that has not been feasible to date.

Although there are several popular forced alignments in the literature, broadly speaking, they all use the same fundamental underlying statistical machinery: (a) modeling the distribution of the low-level acoustics associated with the speech and (b) modeling the temporal relationship between phonemes (Keshet, 2018). The existing state-of-the-art approaches differ in *how* this modeling is done. For example, some of the algorithms do not directly account for coarticulation and model the phonemes without consideration for context (Gorman et al., 2011; Yuan & Liberman, 2008, 2011); whereas others use *triphone* models that consider the sounds that precede and follow the phoneme of interest (McAuliffe et al., 2017; Povey et al., 2011). More recent approaches allow for adaptation of pretrained acoustic models to account for differences in acoustics between the way the model was trained (e.g., on adult speech) and the way the model is used after training (e.g., on children's speech). In Table 1, we list the algorithms in our consideration set and provide a brief description of each one.

Previous work on the performance of forced-alignment algorithms has largely focused on adult speech (MacKenzie & Turton, 2020). However, child speech is different from adult speech: Children's speech anatomies are still developing into adult proportions, and their articulatory abilities and phonological representations are immature and more variable than adult speech. Thus, automatic speech recognition systems have larger error rates on child speech (see review in Beckman et al., 2017). Forced-alignment algorithms are built on similar acoustic models to those used in automatic speech recognition systems; as such, these aligners will likely be less accurate for child speech than adult speech.

One notable recent test of forced alignment on child speech is the work by Knowles et al. (2018), which evaluated how well the Prosodylab-Aligner performed on children's speech while manipulating different alignment parameters (training data, pronunciation dictionary, corpus, etc.). For the default training set (adult speech), the accuracy of sibilant alignments was less than 50%. For stops and vowels, alignment accuracy was between 60% and 88% for one of the corpora and all under 50% for the other. Alignment accuracy improved with age so that the default Prosodylab acoustic model yielded more accurate alignments on older children. As expected, training the acoustic model on child speech improved alignment accuracy when compared against the default adult-speech acoustic model.

The work from Knowles et al. (2018) leaves open the question of how different alignment approaches (e.g., triphone alignment and alignment based on speaker adaptive triphone models) fare on child speech. In this study, we took a broader view and tested the performance of several publicly available alignment algorithms (see Table 1) on speech from a probe used to evaluate speech in children. We focus on speech samples from children ages 3 to 6 years old and evaluate five forced-alignment algorithms along two related dimensions: accuracy and onset-time differences.

The output of forced-alignment systems can be evaluated in a variety of ways, and the metric of interest is application specific. As a result, we aim to answer several research questions that holistically capture the performance of an aligner. These include:

1. What was the accuracy of the aligners relative to gold-standard manual alignments?
2. Which classes of sounds had the most accurate alignment? How did each aligner perform on each class of sounds?

Table 1. Comparison of the forced-alignment algorithms under consideration.

Algorithm	Engine	Alignment	English training set	Remark
P2FA (Yuan & Liberman, 2008)	HMM-GMM on PLP features. HTK backend.	Monophone	25 hr of U.S. Supreme Court oral arguments	Not trainable.
Prosodylab (Gorman et al., 2011)	HMM-GMM on MFCC features. HTK backend.	Monophone	10 hr laboratory-recorded North American speech	
Kaldi (Povey et al., 2011)	HMM-GMM on MFCC features. Kaldi backend.	Two passes: monophone, triphone	Librispeech (Panayotov et al., 2015): 1,000 hr of adult-read audiobooks	Kaldi is a speech recognition engine but recipes are available for forced alignment.
MFA-No-SAT (McAuliffe et al., 2017)	HMM-GMM on MFCC features. Kaldi backend.	Two passes: monophone, triphone	Librispeech	Automates Kaldi alignment recipes. Developed by same lab as Prosodylab.
MFA-SAT (McAuliffe et al., 2017)	HMM-GMM on MFCC features. Kaldi backend.	Three passes: monophone, triphone, speaker-adapted triphone	Librispeech	

Note. P2FA = Penn Phonetics Lab Forced Aligner; HMM = Hidden Markov model; GMM = Gaussian mixture model; PLP = perceptual linear predictor; HTK = Hidden Markov Model Toolkit (Young et al., 2015); MFCC = Mel-frequency cepstral coefficient; MFA = Montreal Forced Aligner; No-SAT = No speaker adaptive training; SAT = speaker adaptive training.

3. Did children's ages predict aligner performance?
4. How did human–human and human–automatic interrater agreement compare?
5. What was the distribution of phone-onset time differences for the aligners?

This comparative evaluation of performance between the five alignment algorithms will provide immediate value for both clinical speech researchers and technologists. We expect that the results presented here will help inform the analysis decisions that researchers make in other child speech studies. In addition, the results will provide a target against which speech technologists can compare new algorithms they develop for alignment.

Method

Participants

A total of 42 typically developing children (21 girls, 21 boys) contributed speech samples for this study. Children met the following criteria: (a) American English as the primary language in the home, (b) hearing within normal limits as indicated by parent report and passing a pure-tone hearing screening or distortion product otoacoustic emission screening bilaterally, (c) speech within normal limits as indicated by standardized articulation test scores, and (d) language within normal limits as indicated by standardized language test scores.

The subset of children examined in this study was selected based on their chronological age. Children were randomly selected from the following age bands: 36–47 months (n = 10), 48–59 months (n = 10), 60–71 months (n = 12), and 72–83 months (n = 10). Half of the children in each age band were boys and half were girls. Children in this sample represented the local community, which is skewed toward White middle-class and upper middle-class families.

Experimental Task

Children produced a standard set of speech stimuli from the Test of Children's Speech (TOCS+; Hodge & Daniels, 2007), administered by a research speech-language pathologist in a sound-attenuating suite. In an elicitation task involving a recorded model played on an iPad, children produced a series of single words and a series of multiword utterances that were the same for each child. Single-word stimuli were 38 individual words, including all items from the TOCS-30 word probe (Hodge & Daniels, 2007). Multiword stimuli were 60 sentences ranging from two to seven words (10 items of each sentence length). The multiword protocol started with the 10 two-word utterances and advanced to the 10 three-word utterances and so on. Some of the younger children were not able to produce all 10 utterances of a given length, so the elicitation protocol was stopped if a child could not produce at least five of the 10 utterances. For example, if a 3-year-old child only produced four of the five-word utterances, then none of the five-word utterances were included. We accepted all child productions of the words, regardless of whether they correctly articulated the word or not, as our goal was to assess aligner performance on actual child speech. Lexical errors (additions, substitutions, omissions, transposition of whole words) were also accepted; we updated the utterance transcript as needed to match the words that the child said. Recordings of children were made using a digital audio recorder (Marantz PMD 570) at a 44.1-kHz sampling rate (16-bit quantization) and a condenser studio microphone (Audio-Technica AT4040) positioned next to each child using a floor stand. The level of the signal was monitored and adjusted on a mixer (Mackie 1202 VLZ) to obtain optimized recordings and to avoid peak clipping. Individual utterances were extracted from each recording into separate audio files, so that there was one file per TOCS item per child. Audio files ranged in duration from 0.7 to 6.9 s with a mean duration of 2.1 s.

Materials and Procedure

Forced aligners. Table 1 describes the five forced-alignment algorithms under consideration. We selected these aligners because of their prior use in the literature and their availability for public use. However, we note that this selection is not exhaustive. We used each aligner's default acoustic models and configurations.

Manual alignment. Manual alignments of boundaries for all phonemes produced by each child were made by two research assistants, one who was a graduate student in speech-language pathology and one who was a certified speech-language pathologist. We considered these human alignments as determined by either of the two research assistants to be the "gold standard." Both research assistants had specialized training in acoustic–phonetics that was specific to this project, involving extensive experience in evaluating child speech samples using acoustic tools. The two researcher assistants divided the children's samples between themselves, but overlapped on 10% of the data (four children) so that interrater reliability could be assessed. Most alignments took approximately between 1 and 5 min per file.

To make gold-standard alignments, research assistants manually corrected the output from the Prosodylab-Aligner, which involved listening to each speech sample produced by each child and performing manual boundary adjustments on Praat textgrids (Boersma & Weenink, 2015) for each of the automatically generated phoneme boundaries. We chose to correct prepopulated alignments rather than create alignments de novo because of the tremendous time demands of creating versus adjusting alignments. We used the Prosodylab as the starting point because of prior experience with it. Research assistants calibrated their judgments of phoneme boundaries by working on the same sets of speech samples, making separate judgements, which were then compared as part of a training set. This training set included two children (a 6-year-old followed by a 3-year-old). Differences between raters, as well as questions that arose, were discussed with the first author and among the research assistants. Across children and utterances, there were 34,205 manually

aligned phones. Due to pronunciation dictionary differences and transcription errors (i.e., the transcript for alignment not matching the child's production), some phones could not be compared against manual alignments (~2% of phones excluded).

Interrater reliability involved having both judges evaluate and adjust phoneme boundaries independently for 10% of the sample (all utterances produced by four different children). We then compared accuracy and alignment timing differences for the two human raters. We report interrater reliability in the results section, where we use this index of human agreement as a benchmark for comparison of aligner performance.

Outcome Variables

Two outcome variables were of interest for this study. These were alignment accuracy and alignment timing differences. We evaluated aligner *accuracy* using the same criterion as Knowles et al. (2018). Two intervals *match* if the boundaries of an automatic (forced) alignment interval contain the midpoint of the manual interval. Figure 1 shows the alignments for a single token with examples of matching alignments. This criterion provides a gross measure of accuracy: Did the aligner "find" the same sound as the manual alignment? In some cases, one force-aligned phone interval can be trivially accurate by spanning with the width of several phones, but that wide interval causes a mismatch in the other automatic intervals relative to the manual alignments. We measured alignment timing differences based on *absolute difference in phone-onset times* between automatic and manual alignments. Because most onset times were also the offset time of a prior phone—for example, in /bi/, the onset of [i] is the offset of [b]—we considered only the onset times for this comparison. We examined both metrics across all phones and within four classes of sounds: vowels (/i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ə/ʌ, ɚ/ɝ, aɪ, aʊ, ɔɪ/), fricatives (/f, v, θ, ð, s, z, ʃ, h/; i.e., all but /ʒ/), plosives (/p, b, t, d, k, g/), and sounds from other classes (affricates /tʃ, dʒ/, liquids /l, ɹ/, nasals /m, n, ŋ/, and glides /w, j/).

Statistical Analyses

We modeled the accuracy of the aligners with a logistic mixed-effects regression model. The outcome variable was aligner *accuracy*—that is, whether an interval produced by an aligner overlapped with the midpoint of the gold-standard, human-aligned interval. Accuracy is a binary measurement, so we used a logistic regression model. We report accuracy estimates using percentages (rather than proportions). Our baseline model included population-average (fixed) effects for aligner, sound class, and aligner-by-sound class interaction. These effects estimated how accuracy on average changed as a function of aligner and sound class. The model's varying (random) effects included by-child intercepts and by-child-by-aligner intercepts. These intercepts allowed children to vary in their overall alignment "difficulty" (by-child intercepts) and in their relative difficulty for each aligner (by-child-by-aligner). To assess the effect of age, we augmented the model to include age and the interactions of age with other predictors, and we used model comparison to determine if age significantly improved model fit over and above the baseline model. Age was centered at 5 years and scaled in years (i.e., Age 0 corresponded to 5 years old and Age 1 corresponded to 6 years old). Additionally, a secondary analysis was performed where we characterize phone-onset time differences using descriptive statistics.

Analyses were carried out in R (Version 4.0.0; R Core Team, 2020) with model fitting by lme4 (Version 1.1.23; Bates et al., 2015). We report effects with estimated marginal means and adjusted p values calculated by the emmeans package (Version 1.4.6; Lenth, 2020). Supplemental Material S1 provides the analysis code and results.

Results

We report our findings for each research question below.

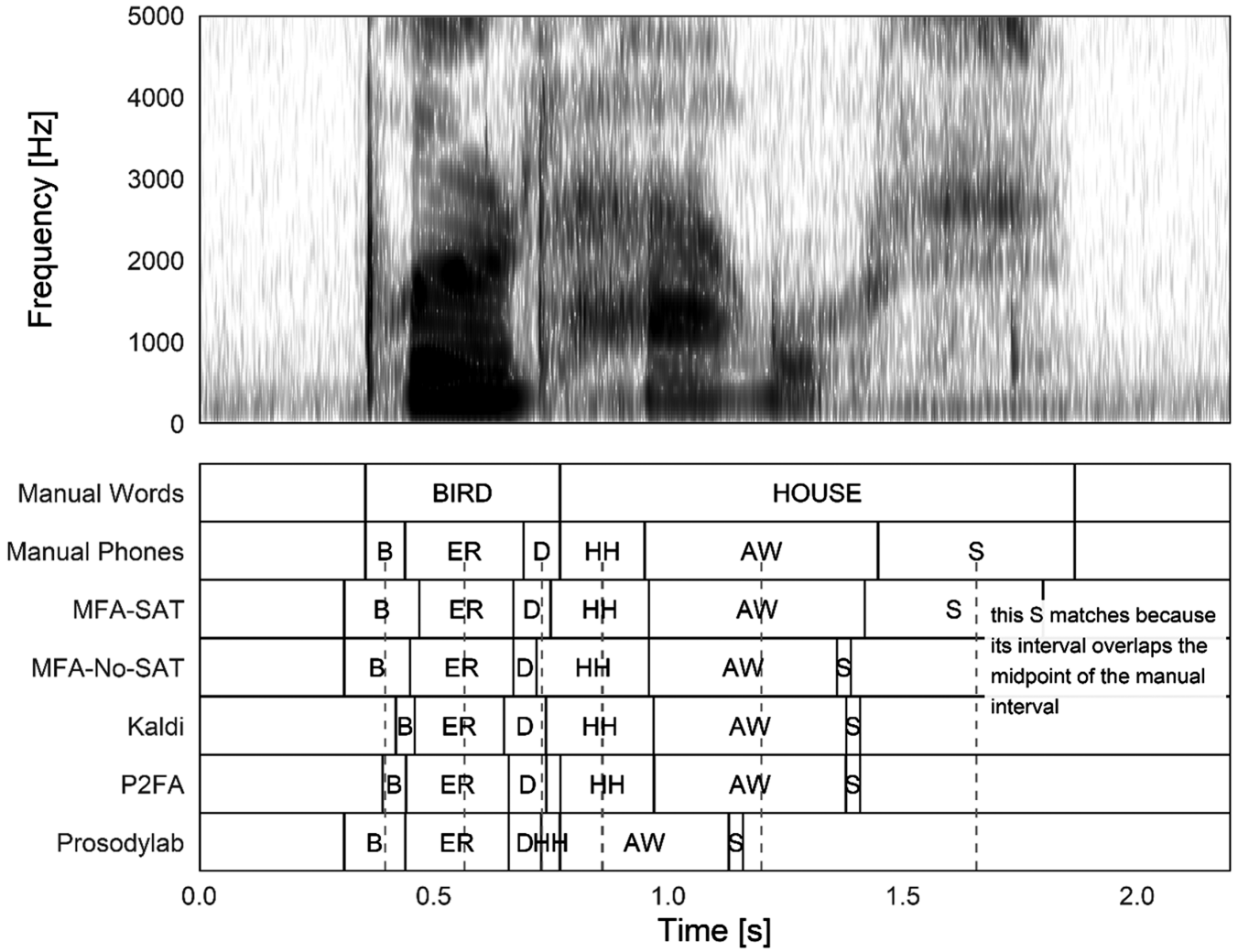
1. What was the accuracy of the aligners relative to gold-standard manual alignments? Averaging over all speech sound classes, the Montreal Forced Aligner with speaker adaptive training (MFA-SAT) was the most accurate aligner, *Average Percent-Accuracy (Acc)* = 86%, 95% CI [84, 87], followed by MFA-No-SAT, *Acc* = 77%, [75, 79]; Kaldi, *Acc* = 76%, [74, 78]; Penn Phonetics Lab Forced Aligner (P2FA), *Acc* = 67%, [65, 69]; and Prosodylab, *Acc* = 61%, [58, 63]. All pairwise log-odds differences between aligners were significant (with Bonferroni-adjusted p values) except for the Kaldi versus MFA-No-SAT contrast, *Odds Ratio* (Kaldi/MFA-No-SAT) = 0.98, $SE = 0.05$, $z = -1.26$, $p = 1.00$. Thus, the aligners all performed differently on average, except for MFA-No-SAT and Kaldi.

2. Which classes of sounds had the most accurate alignment? How did each aligner perform on each class of sounds? Averaging over all alignment algorithms, alignment was more accurate for vowels, *Acc* = 83%, 95% CI [82, 84], compared to other speech sound classes: *Acc*(plosives) = 71%, [69, 72], *Acc*(fricatives) = 72%, [71, 74], *Acc*(others) 69%, [67, 71]. All pairwise log-odds differences between classes were significant (with Bonferroni adjusted p values), but the overlapping confidence intervals on the percentage scale suggest that the key contrast here is between vowels and nonvowels.

Table 2 reports the observed accuracy for each aligner and class of sound. Nearly all of the Aligner × Class contrasts were significant (with false-discovery-rate-adjusted p values) on the log-odds scale under effect (sum-to-one) contrast coding. This scheme compared each cell mean (e.g., Kaldi × Fricative) to the mean of the 20 Aligner × Class cell means. Within each aligner, vowels were the most accurate. For Prosodylab, plosives were the least accurate. For P2FA, plosives and fricatives were the least accurate. For both iterations of MFA and for Kaldi, the other-sounds class was the least accurate.

3. Did children's ages predict aligner performance? We augmented the Aligner × Class baseline model to include age. First, we included age and all two-way interactions

Figure 1. Example spectrogram and textgrid of alignments for the token "bird house." The top two tiers of the grid are the manually aligned words and phones for the utterance, and the lower five tiers are the phone boundaries from the forced aligners. The dashed lines extending from the manual phones are the midpoints of each phone. In this example, the spectrogram shows mid-to-high frequency noise for the final /s/ sound, and the manual interval for /s/ covers this region. Only the MFA-SAT interval also covers this region, and because it includes the midpoint of the manual alignment (the dashed line), this automatic alignment *matches* the manual one. For the vowel in *bird,* all of the automatic alignments match the manual alignment. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; P2FA = Penn Phonetics Lab Forced Aligner.



(Age × Aligner, Aligner × Class), and model comparison showed a significant improvement in model fit, $\chi^2(8) = 308$, $p < .001$. Next, we allowed Age × Aligner × Class three-way interactions, and model comparison showed a significant improvement in model fit, $\chi^2(12) = 59$, $p < .001$.

Averaging over all speech sound classes, Kaldi, P2FA, and MFA-No-SAT showed a significant improvement in accuracy in age. Odds ratios for a 1-year increase in age were as follows: *OR*(MFA-SAT) = 1.03, 95% CI [0.95, 1.12], *OR*(MFA-No-SAT) = 1.19, [1.10, 1.29], *OR*(Kaldi) = 1.26, [1.16, 1.37], *OR*(P2FA) = 1.19, [1.10, 1.29], and *OR* (Prosodylab) = 0.96, [0.89, 1.05]. For the aligner with the largest age-related effect, Kaldi, the estimated expected accuracy was 76% at 5 years old and 80% at 6 years old; hence, on the percent scale, these improvements were on the order of less than five percentage points in accuracy.

Averaging over all the alignment algorithms, the effect of age was greatest for fricatives and for other sounds, odds ratios for a 1-year increase in age: *OR*(fricatives) = 1.29, 95% CI [1.21, 1.39], *OR*(others) = 1.15, [1.07, 1.23]. There was not a statistically significant effect of age for plosives or vowels, *OR*(plosives) = 1.06, [0.99, 1.13], *OR*(vowels) = 1.00, [0.94, 1.07].

Figure 2 shows the estimated alignment accuracy by age and sound class for each aligner and for the marginal means across the five alignment algorithms. For each of the five aligners, the effect of age was greatest for fricatives compared to all other classes. For MFA-SAT, the only class

Table 2. Observed percentages of accurate automatic alignments (i.e., alignment intervals that cover the midpoint of the manual alignment).

Subset	Aligner	Accuracy (%)
All sounds (33,545)	MFA-SAT	87
	MFA-No-SAT	79
	Kaldi	77
	P2FA	69
	Prosodylab	63
Plosives (9,149)	MFA-SAT	85
	MFA-No-SAT	75
	Kaldi	75
	P2FA	62
	Prosodylab	51
Vowels (13,010)	MFA-SAT	90
	MFA-No-SAT	87
	Kaldi	81
	P2FA	76
	Prosodylab	77
Fricatives (6,485)	MFA-SAT	86
	MFA-No-SAT	75
	Kaldi	75
	P2FA	62
	Prosodylab	60
Others (4,901)	MFA-SAT	80
	MFA-No-SAT	70
	Kaldi	72
	P2FA	68
	Prosodylab	55

Note. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; P2FA = Penn Phonetics Lab Forced Aligner.

with a significant positive age effect was the fricatives, *OR* (fricatives, MFA-SAT) = 1.30, [1.17, 1.45].

4. How did human–human and human–automatic interrater agreement compare? We measured agreement between two human aligners on four children. The by-child percentage of matching intervals between the two was 85%–96%. For comparison, we computed by-child agreement between the automatic aligners and the human aligners on the subset of four children: MFA-SAT 70%–89%, MFA-No-SAT 59%–78%, Kaldi 60%–80%, P2FA 50%–71%, Prosodylab 30%–71%. Figure 3 visualizes these agreements by aligner. The only automatic aligner to overlap with human aligners was the MFA-SAT aligner. The upper end of MFA-SAT-to-human raters (89%) overlapped with the lower end of human-to-human agreement (85%).

5. What was the distribution of phone-onset time differences for the aligners? Table 3 shows the absolute differences in phone-onset times between manually aligned intervals and automatically aligned intervals and the percentages of onset-time differences under certain tolerances. Three sets of time differences are included.

First, we examined time differences for all intervals, including those that did not match the manually aligned intervals. These times reflect *average-case* performance. All aligners had median time differences of less than 30 ms, and the distributions of the time differences were right-skewed with the median differences being much smaller than the average differences: For example, Kaldi showed a median of 20 ms and a mean of 63 ms. This discrepancy occurred because phone-onset times errors can have cascading effects on each other: A poor alignment for a word will likely affect the alignment of subsequent words. MFA-SAT, MFA-No-SAT, and Kaldi all performed comparably in terms of median time differences, but Kaldi had more extreme timing differences (those above 100 ms) so its average time difference was much larger than its median.

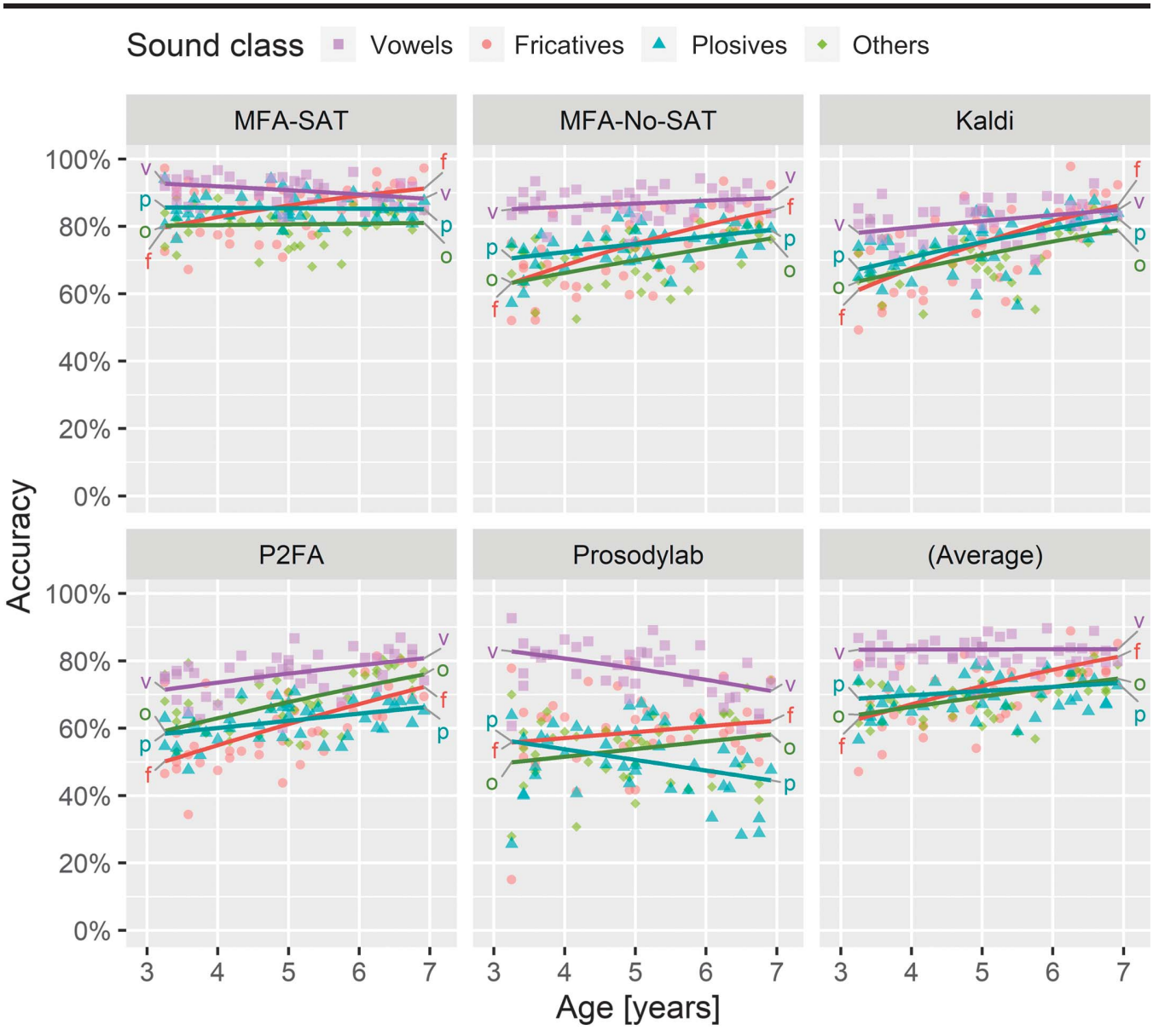
Next, we examined time differences for only the accurate intervals that matched the manually aligned intervals. These times reflect *best-case* performance in which the aligner found the sound in the speech sample. In this set, the aligners all performed comparably: median time differences of 17–20 ms, mean differences of 26–29 ms, 85%–89% of differences smaller than 50 ms. Indeed, the largest difference among the aligners was in the *numbers of intervals* tested: for example, 29,066 for MFA-SAT versus 21,190 for Prosodylab. These results suggest the most important feature for temporal accuracy was how reliably the aligner could find the target sound.

To put these time differences into perspective, we also computed time differences for the interrater agreement subset of speech samples. For this comparison, one of the raters, randomly selected, served as the gold standard for the other rater and for the alignment algorithms. Manual alignment yielded much smaller time differences than the forced-alignment algorithms: A median difference of 10 ms and 72% of differences were smaller than 25 ms. Because the manual alignment process started by correcting boundaries on Prosodylab intervals, any boundary that was not adjusted by both raters would automatically have a time difference of 0 ms. Put differently, both raters start with 0-ms difference on every boundary by default and diverge from each other by correcting boundaries. Therefore, we checked what percentage of differences was 0 ms to evaluate whether human–human interrater agreement was inflated by unadjusted boundaries: 6.6% for manual intervals, 3.2% for MFA-SAT, 3.6% for MFA-No-SAT, 3.4% for Kaldi, 0% for P2FA, and 17.3% for Prosodylab. Unadjusted boundaries only accounted for 6.6% of differences, and when excluding these cases, the median difference for human alignment was 13 ms.

Discussion

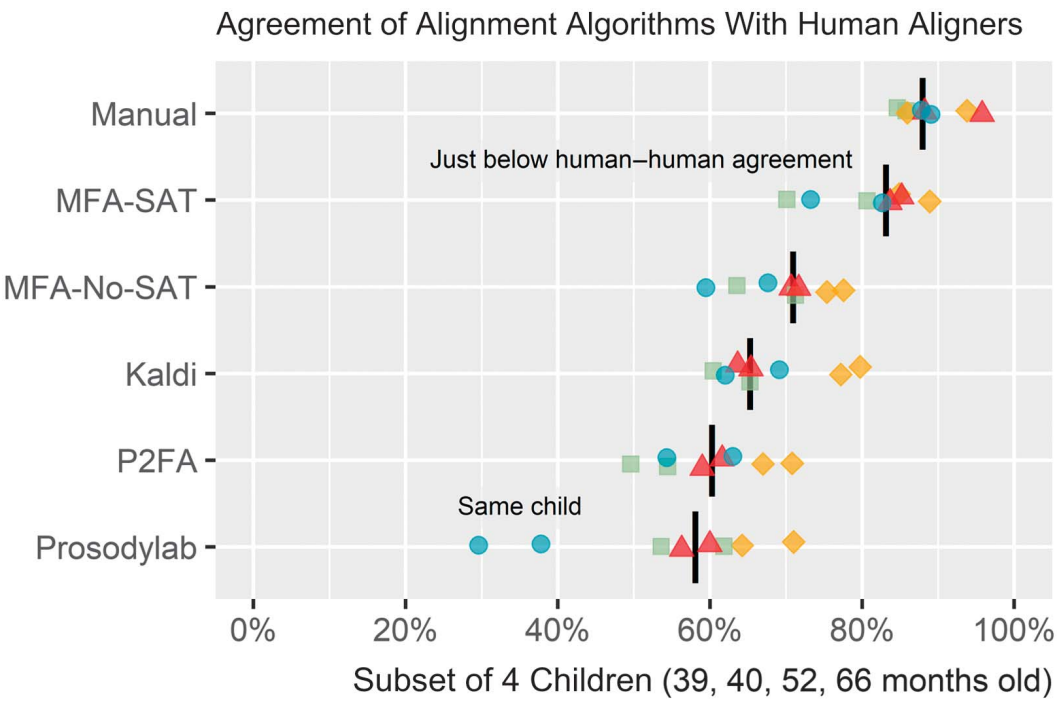
In this study, we performed a "bake-off" (an empirical evaluation of several algorithms) with five different forced-alignment algorithms on speech samples of 3- to 6-year-old children. We assessed the accuracy of these aligners by evaluating whether intervals produced by forced alignment contained the midpoints of intervals produced by manual alignment, and we asked whether speech sound class and child age affected alignment accuracy. We found that the MFA-SAT (McAuliffe et al., 2017) performed the best overall. Vowels were the least difficult class of sound for forced alignment, and age had the largest effect on accuracy for the fricatives. Finally, for accurately aligned sounds, phone-onset time differences were comparable across aligners.

Figure 2. Estimated alignment accuracy by sound class by age for the five alignment algorithms and the marginal average over the five aligners. Lines represent the estimated population-average (fixed-effects) accuracy. Points represent the average accuracy for the classes of sounds for each child, so one point represents one child's accuracy for that sound class. Several key findings are visible here: (a) MFA-SAT was the most accurate overall; all of its age-trend lines are above 75%. (b) Vowels were the most accurately aligned sounds; the topmost age-trend line in each panel is the vowel line. (c) Fricatives were the sound class most affected by age: In every panel, there is a positive slope for the fricative age-trend line. The letters v, f, p, o are included to label the sound class for each line. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; P2FA = Penn Phonetics Lab Forced Aligner.



Why did MFA-SAT perform the best overall? Broadly speaking, we found three different tiers of aligner algorithms in terms of average overall accuracy: MFA-SAT performed best with 86% accuracy, followed by MFA-No-SAT and Kaldi with approximately 77% accuracy, followed by P2FA and Prosodylab with less than 70% accuracy. We can interpret these tier differences in terms of underlying alignment technology. P2FA and Prosodylab perform alignment at the monophone level, and their acoustic models were trained on smaller corpora of adult speech (25 hr or less). Kaldi and both MFA types perform alignment at the triphone level, and their acoustic models were trained on a 1,000-hr corpus of adult-read speech. The combination of contextual variation (with triphone alignment) and a richer acoustic model may make these aligners perform a step above the Prosodylab and P2FA aligners. The large performance gain in MFA-SAT over Kaldi and MFA-No-SAT can be attributed specifically to speaker adaptive training. Although the

Figure 3. Percent agreement between automatic forced-alignment algorithms and human aligners. Each point represents the percentage of agreement for an alignment algorithm and a human aligner for one child. Points are colored to uniquely identify children. There were two human aligners, so each child appears twice per row (two points of same color). Vertical bar marks the median in that row. Ages in months were 39 (blue circles), 40 (green squares), 52 (orange diamonds), and 66 (red triangles). MFA = Montreal Forced Aligner; SAT = speaker adaptive training; P2FA = Penn Phonetics Lab Forced Aligner.



acoustic model for MFA-SAT was trained on adult speech, speaker adaptation appeared to allow the aligner to normalize or adjust for developmental differences in children.

The architectural differences between aligners also help explain differences among aligners for different classes of sounds. The triphone aligners (Kaldi and MFA) can incorporate contextual information into their acoustic models. Plosives show reliable positional variation (e.g., aspiration or unreleased closure), but the lack of contextual information in monophone aligners (P2FA, Prosodylab) would make

Table 3. Percentages of absolute phone-onset differences under various tolerances.

Set	Aligner	No. of intervals	Differences in onset times (ms)			Percent differences		
			Median	IQR	*M*	< 25 ms	< 50 ms	< 100 ms
All intervals	MFA-SAT	33,545	20	31	35	60	85	95
	MFA-No-SAT	33,545	20	32	42	57	81	92
	Kaldi	33,545	20	35	63	55	77	87
	P2FA	33,545	25	44	56	50	73	86
	Prosodylab	33,545	28	44	56	46	74	86
Only accurate intervals	MFA-SAT	29,066	18	27	26	64	89	98
	MFA-No-SAT	26,471	17	25	26	64	89	97
	Kaldi	25,871	17	25	28	64	88	96
	P2FA	23,019	19	25	28	62	87	97
	Prosodylab	21,190	20	38	29	58	85	96
All intervals from interrater reliability subset	Manual	2,215	10	26	26	72	85	94
	MFA-SAT	2,215	20	31	34	56	84	95
	MFA-No-SAT	2,215	24	37	48	51	77	89
	Kaldi	2,215	26	54	91	49	71	81
	P2FA	2,215	28	54	62	46	69	82
	Prosodylab	2,215	30	45	64	42	72	85

Note. MFA = Montreal Forced Aligner; SAT = speaker adaptive training; P2FA = Penn Phonetics Lab Forced Aligner.

these sounds more difficult. Plosive sounds were among the most difficult classes for these two aligners.

In what situations is forced alignment most reliable for child speech? Vowels on average had the most accurate alignments, and fricative accuracy was most strongly affected by age. Intuitively, this finding makes sense as vowels can be parameterized using formants, whereas fricative properties span the spectrum. Said another way, the speaker has additional degrees of freedom when producing fricatives; therefore, their spectral signatures are more variable. It is clear from the results that this variability cannot be readily normalized away with speaker adaptation. Another possible reason for the age-fricative relationship is that children also show considerable anatomical and articulatory development during the age range of this study. Additionally, fricative sounds are also those that are most sensitive to noise during the recording process, although we do not expect that this played a large role in our analyses because the recording conditions were well controlled.

Age of acquisition norms can help explain some of the effects of age on alignment accuracy. Averaging over the aligners, age did not influence alignment accuracy for vowels or plosives. These sounds have simpler motor demands that support earlier acquisition (Kent, 1992). For example, the review of consonant acquisition by McLeod and Crowe (2018) puts the average age of acquisition (75%–85% correctly produced) for plosive sounds /p,b,d,b,k,g/ (i.e., all but /t/) at 2 years of age but fricatives are acquired over the full 2- to 6-year age range. A related consideration, though open to further empirical work, is that these aligners do not know what children sound like. Because they are trained on adult corpora, the acoustic models are not familiar with common child articulatory strategies (stopping fricatives, gliding liquids, etc.). Alignment accuracy for fricatives increased with age because children started to develop more adult-like productions.

Differences in phone-onset times were consistent across the five aligners on accurately aligned intervals. Therefore, the key problem for forced alignment is *accuracy*; that is, finding where the sound occurs in a speech sample. Onset-time differences were smaller for human–human differences (around 10 ms) compared to human–manual time differences (around 20 ms), but this result is expected. The human aligners received laboratory training with check-ins on protocol drift (i.e., meetings to make sure human aligners were using the same rules or heuristics during alignment). Minimizing human–human differences is an ongoing concern in multirater research designs, but this kind of retuning or recalibration is not part of the automatic alignment workflow.

This work provides a target against which speech technologists can evaluate the performance of their child speech alignment algorithms. The current aligners perform well on child speech, but not on par with forced-alignment algorithms on adult speech or with the gold-standard human labels. On adult speech benchmarks, the best performing aligner (MFA-SAT) had 72%–77% of phone boundaries within 25 ms of gold-standard boundaries with median absolute time differences of approximately 11 ms (McAuliffe et al., 2017). For this set of child speech, we found 58% of onset boundaries were within 25 ms of the gold standard with a median time difference of 21 ms for the same aligner. Importantly, this result does not suggest that forced alignment should not be used in research done on child speech. The paper provides an estimate of expected accuracy and timing errors by sound class and age.

We have two suggestions for how to use forced-alignment algorithms. First, forced alignment can be used as part of a semi-automated workflow where intervals are first set automatically and then later manually corrected. For instance, a variation of this workflow was used by Stuart-Smith et al. (2015) where automatic voice onset time measurements were screened as *correct, correctable* (then corrected), or *not usable* (due to large error or noise, etc.). They report an efficiency of 1 min of annotation time per 1 min of speech time. Our phone-class results then set a priority list for correction: Vowels will likely need less correction, but fricatives in younger children will need more attention. Second, researchers using forced alignment for child speech statistically control for accuracy and timing errors in their statistical models to ensure that their findings are not confounded by these variables. This approach might become more important when a data set grows too large for the semi-automated workflow. Both of these workflows, however, would benefit if aligners also generated scores for the confidence of the phone alignments or provided other diagnostics (e.g., Baghai-Ravary et al., 2011) that can indicate whether an alignment interval might need further review (or exclusion) in downstream analyses.

Limitations and Future Directions

There are two key limitations of this study. First, we only used the aligners in their default configuration and with their adult-trained acoustic models. Therefore, these results set a lower bound on alignment performance on child speech. All of the aligners except P2FA support the training of new acoustic models, so training on a child speech corpus or a mixed adult–child speech corpus can improve alignment performance. Indeed, Knowles et al. (2018) found that retraining on child speech provided a substantial increase in alignment accuracy for the Prosodylab-Aligner. Other strategies may improve alignment performance: training separate models for different age ranges or normalizing child speech with preprocessing before alignment. These avenues require further research and experimentation.

The other main limitation of these results is that we tested the aligners on high-quality child speech data. They were recorded in a well-controlled environment as part of a picture-prompted word and sentence repetition task; hence, environmental noise and articulatory–linguistic variability were minimized. We have not tested aligners on longer samples or samples of spontaneous or conversational speech. We would expect MFA-SAT to perform the best in such situations, based on our results, but we are hesitant to extrapolate beyond elicited laboratory speech.

Methodologically, our gold-standard manual alignments were created by correcting alignments produced by

the Prosodylab-Aligner. A purer approach, albeit more time consuming, would have started with randomly placed boundaries. The present workflow, though efficient, biased time differences toward the Prosodylab-Aligner, but only for the smallest differences. Any boundary not adjusted manually had an onset-time difference of 0 ms, so the number of 0-ms differences was inflated compared to the number of 1- to 10-ms differences. For example, 19% of Prosodylab differences and 4% of MFA-SAT differences were equal to 0 ms, but 6% of Prosodylab differences and 21% of MFA-SAT differences were between 1 and 10 ms.

Our results provide a snapshot of the state of the art in forced-alignment algorithms, but with ever-improving technological developments, these results will require updating. It is impressive that the accuracy of the best-performing aligner approached that of human–human agreement on an interrater reliability probe. This result suggests that, as the sizes of publicly available corpora grow and new technology is developed, it will not be long before the state-of-the-art aligner will bridge this gap, at least on simple speech elicitation tasks (e.g., prompting, repetition) or on subsets of phonemes (e.g., vowels).

Acknowledgments

This study was funded by Grants R01 DC015653 (awarded to Hustad) and R01 DC006859 (awarded to Liss/Berisha) from the National Institute on Deafness and Other Communication Disorders. Support was also provided by a core grant to the Waisman Center, U54 HD090256, from the National Institute of Child Health and Human Development. The authors thank the children and their families who participated in this research, and the students and staff at the University of Wisconsin–Madison and Arizona State University who assisted with data collection, data reduction, and analyses.

References

Baghai-Ravary, L., Grau, S., & Kochanski, G. (2011). Detecting gross alignment errors in the Spoken British National Corpus. arXiv. Retrieved January 01, 2011, from https://ui.adsabs.harvard.edu/abs/2011arXiv1101.1682B

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beckman, M. E., Plummer, A. R., Munson, B., & Reidy, P. F. (2017). Methods for eliciting, annotating, and analyzing databases for child speech development. *Computer Speech & Language, 45,* 278–299. https://doi.org/10.1016/j.csl.2017.02.010

Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer* [Computer program]. Retrieved October 1, 2020, from https://www.praat.org

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics, 39*(3), 192–193.

Hodge, M., & Daniels, J. (2007). *TOCS+ Intelligibility Measures* [Computer software]. University of Alberta. http://www.tocs.plus.ualberta.ca/software_Intelligibility.html

Kent, R. D. (1992). The biology of phonological development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 65–90). York Press.

Keshet, J. (2018). Automatic speech recognition: A primer for speech-language pathology researchers. *International Journal of Speech-Language Pathology, 20*(6), 599–609. https://doi.org/10.1080/17549507.2018.1510033

Knowles, T., Clayards, M., & Sonderegger, M. (2018). Examining factors influencing the viability of automatic acoustic analysis of child speech. *Journal of Speech, Language, and Hearing Research, 61*(10), 2487–2501. https://doi.org/10.1044/2018_JSLHR-S-17-0275

Lenth, R. (2020). *emmeans: Estimated Marginal Means, a.k.a. Least-Squares Means.* https://CRAN.R-project.org/package=emmeans

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard, 6*(Suppl. 1). https://doi.org/10.1515/lingvan-2018-0061

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In F. Lacerda (Ed.), *Proceedings of Interspeech 2017* (pp. 498–502). International Speech Communication Association.

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology, 27*(4), 1546–1571. https://doi.org/10.1044/2018_AJSLP-17-0100

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. V. Clarkson, & J. Manton (Eds.), *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE.

R Core Team. (2020). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing, Vienna, Austria.* http://www.R-project.org/

Shinozaki, T., & Furui, S. (2003). Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation. In J. Bilmes, & W. Byrne (Eds.), *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)* (pp. 417–422). IEEE. https://doi.org/10.1109/ASRU.2003.1318477

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology, 6*(3–4), 505–549. https://doi.org/10.1515/lp-2015-0015

Tu, M., Grabek, A., Liss, J. M., & Berisha, V. (2018). *Investigating the role of L1 in automatic pronunciation evaluation of L2 speech.* arXiv. https://doi.org/10.21437/Interspeech.2018-1350

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Ragni, A., Valtchev, V., Woodland, P., & Zhang, C. (2015). *The HTK book for HTK version 3.5*. Cambridge University Engineering Department.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America, 123*(5), 3878. https://doi.org/10.1121/1.2935783

Yuan, J., & Liberman, M. (2011). Automatic detection of "g-dropping" in American English using forced alignment. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 490–493). IEEE. https://doi.org/10.1109/ASRU.2011.6163980