JSLHR

Research Article

Examining Factors Influencing the Viability of Automatic Acoustic Analysis of Child Speech

Thea Knowles,[a,b] Meghan Clayards,[c,d,e] and Morgan Sonderegger[c,e]

Purpose: Heterogeneous child speech was force-aligned to investigate whether (a) manipulating specific parameters could improve alignment accuracy and (b) forced alignment could be used to replicate published results on acoustic characteristics of /s/ production by children.

Method: In Part 1, child speech from 2 corpora was force-aligned with a trainable aligner (Prosodylab-Aligner) under different conditions that systematically manipulated input training data and the type of transcription used. Alignment accuracy was determined by comparing hand and automatic alignments as to how often they overlapped (%-Match) and absolute differences in duration and boundary placements. Using mixed-effects regression, accuracy was modeled as a function of alignment conditions, as well as segment and child age. In Part 2, forced alignments derived from a subset of the alignment conditions in Part 1 were used to extract spectral center of gravity of /s/ productions from young children. These findings were compared to published results that used manual alignments of the same data.

Results: Overall, the results of Part 1 demonstrated that using training data more similar to the data to be aligned as well as phonetic transcription led to improvements in alignment accuracy. Speech from older children was aligned more accurately than younger children. In Part 2, /s/ center of gravity extracted from force-aligned segments was found to diverge in the speech of male and female children, replicating the pattern found in previous work using manually aligned segments. This was true even for the least accurate forced alignment method.

Conclusions: Alignment accuracy of child speech can be improved by using more specific training and transcription. However, poor alignment accuracy was not found to impede acoustic analysis of /s/ produced by even very young children. Thus, forced alignment presents a useful tool for the analysis of child speech.

Supplemental Material: https://doi.org/10.23641/asha.7070105

Acoustic analysis of speech has traditionally required labor-intensive hand annotation of segment boundaries or acoustic events. The time-consuming nature of the process has limited the scale of these studies. There has been a growing interest in very large spoken language corpora in order to facilitate more large-scale research on systematic variation in speech (Beckman, Plummer, Munson, & Reidy, 2017; Coleman, Liberman, Kochanski, Burnard, & Yuan, 2011). Such research depends on the ongoing development of tools for increased automation of the process.

One such tool is *forced alignment,* or the automatic time alignment of a phonetic transcription to an acoustic speech signal using automatic speech recognition (ASR) tools. Forced alignment takes as input an orthographic transcription of the speech signal, the speech signal itself, a pronunciation dictionary, and acoustic models trained to recognize the phones of the pronunciation dictionary. As output, it aligns phone and word-level transcripts to the acoustic signal, producing an automatic phonetic segmentation. In cases where more than one possible pronunciation is listed in the pronunciation dictionary (e.g., "talking" vs. "talkin'" as in Yuan & Liberman, 2011b), the aligner is forced to choose one. These pronunciation choices as well as the segmentation of the aligner can then be used for subsequent analyses (Gorman, Howell, & Wagner, 2011; Milne, 2014; Renwick, Baghai-Ravary, Temple, & Coleman, 2013; Schiel, 2004; Yuan & Liberman, 2008, 2011a). Several degrees of freedom can affect alignment accuracy, including the speech data on which it was trained and on which

[a]School of Communication Sciences & Disorders, Western University, London, Ontario, Canada
[b]Health & Rehabilitation Sciences, Western University, London, Ontario, Canada
[c]Linguistics, McGill University, Montréal, Québec, Canada
[d]School of Communication Sciences and Disorders, McGill University, Montréal, Québec, Canada
[e]Centre for Research on Brain, Language and Music, McGill University, Montréal, Québec, Canada

Correspondence to Thea Knowles: tknowle3@uwo.ca

Editor-in-Chief: Julie Liss
Editor: Megan McAuliffe

Received July 19, 2017
Revision received January 22, 2018
Accepted May 17, 2018
https://doi.org/10.1044/2018_JSLHR-S-17-0275

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

it will be used, as well as the variants in the pronunciation dictionary.

Although many aligners come with pretrained, default, acoustic models for users (Bigi, 2012; Gorman et al., 2011; McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), some forced aligners are also *trainable,* which means that the user may retrain acoustic models using other audio data. Typically, *training* a forced aligner is much like using one, except that the aligner must learn what the best acoustic models are instead of having acoustic models provided. Importantly, it is not given the alignments themselves to learn from, only the orthographic transcription, dictionary, and sound files. Thus, with a trainable aligner, a researcher always has the option of training on their own, unaligned data and then subsequently aligning it. In fact, this is sometimes encouraged as a means of improving alignment accuracy (McAuliffe et al., 2017). However, this may or may not be the best choice for a given data set, and here, we explore some of the factors that might help determine that.

Forced alignment has been successful for automating acoustic analysis of adult productions, for example, related to sibilant spectral center of gravity (CoG; Clayards & Doty, 2011), acoustic reduction (Schuppler, Ernestus, Scharenborg, & Boves, 2011), word- and syllable-final consonant realization (Adda-Decker & Snoeren, 2011; Milne, 2014; Schuppler, van Dommelen, Koreman, & Ernestus, 2012; Yuan & Liberman, 2011a, 2011b), nasal place assimilation (Renwick et al., 2013), and vowel change (Labov, Rosenfelder, & Fruehwald, 2013). The success of these attempts suggests that this is a viable new tool in the speech researcher's toolkit that could find many applications.

Automatic Recognition of Child Speech

One such application would be extending these techniques to other populations such as children. However, ASR technology is known to perform more poorly with highly variable speech, such as with child utterances (see Beckman et al., 2017; Benzeghiba et al., 2007, for reviews), with error rates generally inversely correlated with age. Child speech differs from adult speech in that it is more variable, slower, and systematically different in spectral dimensions (Lee, Potamianos, & Narayanan, 1999). In fact, human listeners often also have more difficulty in understanding very young children's speech (D'Arcy & Russell, 2005). Although most ASR systems are trained only on adult data, the differences between adult and child speech make recognition of children's speech using acoustic models trained on adult speech problematic (Wilpon & Jacobsen, 1996). Warping children's speech using vocal tract normalization so that it more closely matches adult acoustics improves performance (Gerosa, Giuliani, & Brugnara, 2007; Potamianos, Narayanan, & Lee, 1997), as does training acoustic models with child speech (Wilpon & Jacobsen, 1996, though see Gerosa, Giuliani, & Brugnara, 2009), the latter of which may be more successful (Elenius & Blomberg, 2005). Another source of difficulty for automatic systems is that children do not always pronounce words with the same phones as would be found in an adult pronunciation dictionary (Benzeghiba et al., 2007).

In ASR, the system must determine what the words were as well as where the segments are. In forced alignment, however, the transcription is provided, making the task more constrained. As such, forced alignment is a potentially viable tool for analyzing children's speech. For example, Lee et al. (1999) used it to facilitate analysis of acoustic properties of speech of 5- to 11-year-olds. Given that forced alignment is an ASR-based system, however, it is likely that its accuracy is subject to similar pitfalls. Relatively little work has examined factors affecting the accuracy of forced alignment for children's speech. This article does so by exploring how accuracy is affected by parameters that researchers may be able to manipulate.

The first half of this article explores the effects of three alignment parameters on the accuracy of forced alignment in child speech: the type of data used to train acoustic models (whether it includes adult or child speech, including the exact speech to be aligned), the type of transcription used (orthographic or phonetic), and the speech segment to be aligned (vowels, stops, sibilants). We also explore the effects of speaker age and, more qualitatively, speaking conditions (spontaneous vs. elicited). The first half aims to explore the options that would typically be available to speech researchers looking to force-align their data in order to better understand how these options affect alignment performance. The second half asks an important follow-up question: However accurate the alignments are, can they successfully replace hand segmentations for acoustic–phonetic analysis? We explore this question by attempting to replicate the findings of Bang, Clayards, and Goad (2017) on /s/ productions in children using automatic alignment of the same data. If acoustic analysis conducted on force-aligned /s/ data leads to the same conclusions drawn in Bang et al. (2017) where the speech was manually aligned, this would indicate that some analyses of child speech may benefit from this automation technique as has been demonstrated for adult speech. /s/ productions may be a good candidate for forced alignment, given that the spectral properties of frication are relatively stable throughout any particular production (see, however, Iskarous, Shadle, & Proctor, 2008, for evidence of important dynamic patterns). This may mean that acoustic analysis of /s/ is less reliant on highly accurate temporal alignment. If, however, the variability inherent in child /s/ production poses too great a challenge for accurate alignment to reliably capture the relevant acoustic signal, the utility of forced alignment for analyzing child speech is still limited.

Child /s/ Production

In children, /s/ production is highly variable over the course of development (e.g., Nittrouer, 1995; Smit, Hand, Freilinger, Bernthal, & Bird, 1990). Target-like word-initial /s/ production is not typically achieved by most (90%) English-speaking children until after age 7 (Li, Edwards, & Beckman, 2009; Smit et al., 1990); young children instead may produce distortions, phonetic substitutions, or omissions. Such variable developmental acquisition may depend on structural or motoric constraints of sibilant production (Green, Moore, & Reilly, 2002; McAllister Byun, 2011, 2012; Mugitani & Hiroya, 2012; Vorperian et al., 2009, 2011). Children's productions of /s/ tend to have a lower CoG and a smaller spectral slope and are more coarticulated with following vowels compared to adult /s/ (Nissen & Fox, 2005; Nittrouer, 1995; Nittrouer, Studdert-Kennedy, & McGowan, 1989). Sex/gender differences in /s/ production that cannot be explained by anatomical differences alone have been found to occur in the speech of very young children (Bang et al., 2017; Li et al., 2016). For example, Bang et al. (2017) found that all children between the ages of 2 and 5 years old produced /s/ differently from adults, but they also found that male children produced more adult male /s/-like productions and female children produced more adult female /s/-like productions. This difference, measured by differences in spectra analysis of /s/, was apparent even as early as 3 years of age and increased as children got older. In our analysis, we will attempt to replicate this gender difference and its interaction with age.

Purpose

This article explores the use of automatic forced alignment on the heterogeneous speech of young children in order to determine variables that lead to improvements in automatic analysis of highly variable speech. Specifically, we predict that (a) automatic forced alignment, compared to manual alignment, will yield similar but much less accurate boundary predictions of child speech segments; (b) a subset of modifiable parameters for forced alignment will lead to greater alignment accuracy when used with child speech; and (c) that the application of these parameters will lead to more accurate acoustic analysis, comparable to analysis performed using hand-segmented data. In order to test these predictions, this article is divided into two parts. In Part 1, using a trainable forced aligner, we systematically explore the effects of four modifiable variables on the accuracy of force-aligned child speech—pronunciation dictionary, training data, phonetic class to be aligned, and child age—for two different speech corpora. In Part 2, we use a subset of these parameters to automatically align and analyze child /s/ productions, using the same data set and methodology used in Bang et al. (2017), to determine whether forced alignment is a viable tool for automatic acoustic analysis of child speech.

Part 1: Examining the Viability of Forced Alignment on Child Speech

Method

We compared manually aligned and force-aligned child speech data in order to identify mutable alignment parameters that optimized the force-aligned output. Using a trainable forced aligner on two distinct speech corpora, we explored three alignment parameters that represent methodological choices in a standard research setting: pronunciation dictionary, training of acoustic models, and phonetic segment of interest. These parameters are described below and, when referred to as predictors, identified in SMALL CAPS throughout the text. Distinct levels of parameters are identified in *italics*.

Speech Corpora

We analyzed data from two speech corpora available from the Child Language Data Exchange System (MacWhinney, 2000). These corpora were chosen in part to represent different testing environments and paradigms used to elicit child speech. The *Julia* corpus included approximately 2 hours of speech from one female Canadian English–speaking child. Spontaneous speech data were collected longitudinally from ages 1;5 to 3;6 (years;months) in a naturalistic play setting (Goad, 2010). The English version of the *Paidologos* corpus included approximately 5 hours of speech from 81 children (40 girls, 41 boys) from Columbus, OH, ages 2;0 to 5;11 (Edwards & Beckman, 2008). Speech consisted of single-word productions elicited during a picture-prompted word repetition task. Both corpora included orthographic and full or partial phonetic transcriptions. The speech audio files were segmented at the utterance level to prepare for alignment.

Forced Alignment

Automatic segmentation was performed for all data using the ProsodyLab-Aligner (Gorman et al., 2011), which uses the Hidden Markov Model Toolkit (Young et al., 1994). A full description of how forced alignment was applied to these data can be found in Knowles et al. (2015).

Predictors: Alignment Parameters

Table 1 describes the speech corpora and the alignment predictors included in Part 1: PRONUNCIATION DICTIONARY, TRAINING DATA for acoustic models, and phonetic SEGMENT of interest.

Pronunciation dictionaries. Forced alignment requires a phonetic transcription of the audio speech data to be aligned, often provided in the form of a pronunciation dictionary in which orthographic forms are mapped to phonetic transcriptions. We included two PRONUNCIATION DICTIONARIES: *Standard,* composed of target-like English pronunciations, and *Customized,* developed from the phonetic transcriptions of child utterances. The *Standard* dictionary was used for alignment of both corpora and consisted of a standard North American English transcription, the Carnegie Mellon University (CMU) Pronunciation Dictionary (Weide, 1998). The CMU Pronunciation Dictionary is a machine-readable pronunciation dictionary for North American English that provides over 134,000 words and their phonetic transcription in ARPAbet.[1] *Julia* was also

[1] ARPAbet is a standard set of phonetic symbols for speech recognition. See further explanation at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Table 1. Alignment parameters used in Part 1.

Parameter	Levels	Description
Corpus	*Julia*	One child in a naturalistic setting
	Paidologos	Multiple children in a word repetition task
Pronunciation Dictionary	*Standard*	CMU Standard dictionary
	Customized	Composed of actual phonetic realizations
Training of Acoustic Models	*Adult-only*	Default acoustic models trained on adult laboratory speech
	Adult–child	Mix of adult and child speech
	Child-general	Mix of child speech (nonspecific)
	Child-specific	Speech of specific child or children to be aligned
Phonetic Segment	*Voiceless stops*	p, t, k
	Voiceless sibilants	s, ʃ
	Vowels	Various

Note. CMU = Carnegie Mellon University.

aligned using a customized speaker-specific pronunciation dictionary consisting of the phonetic transcription of her utterances (a full phonetic transcription was not available for *Paidologos*). Each utterance was given a unique entry in the pronunciation dictionary. The supplied narrow phonetic transcription was collapsed into a broader set of ARPAbet characters to provide more exemplars for each ARPAbet category. See Table 2 for an example of ARPAbet entries for the two pronunciation dictionaries.

Training of acoustic models. Many widely used forced aligners have been pretrained on a large speech data set, and retraining is either impossible (e.g., Forced Alignment and Vowel Extraction; Rosenfelder, Fruehwald, Evanini, & Yuan, 2011) or difficult (e.g., Munich AUtomatic Segmentation; Schiel, 2004). One advantage of using a trainable aligner is that the data that the researcher wishes to align can be used to train the aligner directly (Gorman et al., 2011; McAuliffe et al., 2017). This has been recommended before as a potential way to improve alignment accuracy (McAuliffe et al., 2017).[2] An open question is whether this is the best way to train an aligner—in particular when one has a small data set—or whether using other data sets that are larger is better—even if they are unlike the data of interest. This is one of the questions we set out to test. It is particularly relevant for small data sets of child speech that are very unlike the large data sets of adult speech normally available, for example, in pretrained systems. The input training data are likely to affect the alignment accuracy because it helps the aligner identify likely acoustic representations of the phones to be aligned (McAuliffe et al., 2017). Alignment using an acoustic model that is trained on speech that is highly dissimilar from the speech to be aligned may be less likely to lead to accurate output. However, if there is too little speech data on which to train, even if it has a high degree of similarity to the speech to be aligned, the acoustic models generated during the training stage may not have enough exemplars to produce consistently reliable boundary predictions. This study included four TRAINING conditions designed to vary (a) in acoustic similarity to the child speech to be aligned and (b) inversely, in the amount of data used for training.

The four TRAINING conditions were as follows: *Adult-only (AO)* training included acoustic models trained on approximately 10 hours of North American English adult laboratory speech, which are the default models distributed with the Prosodylab-Aligner (Gorman et al., 2011). *Adult–child (AC)* training included a combination of adult laboratory data (the same as for *AO*) and a subset of child data from both corpora (approximately 6 hours of audio in total). *Child-general (CG)* training included all child data from the two corpora (approximately 7 hours of audio) and no adult data. *Child-specific (CS)* training included acoustic models trained only on the specific corpus to be aligned (*Julia* or *Paidologos*). That is, training of acoustic models in this final condition was restricted to the exact data that would be aligned.

Phonetic segments. We analyzed voiceless stops and sibilants as well as vowels to determine whether the phonetic class yielded differences in forced alignment accuracy. Consonants of interest occurred word initially for *Paidologos* and in multiple word positions for *Julia*. In the case of the *Julia* corpus, there were two possible sets of segmental transcriptions. Alignments using the *Standard* pronunciation

Table 2. Example of ARPAbet character entries and corresponding International Phonetic Alphabet (IPA) for three productions of the word “dog” in the (a) *Standard* and (b) *Customized* pronunciation dictionaries.

Word	Standard	Customized	IPA
dog	D AO1 G	D AO1 G	d ɔ g
		D AO1	d ɔ
		D AA1	d æ

[2] Note that training on the same data is an acceptable approach for forced alignment, where the goal of the task is not necessarily to generalize the acoustic model results to new speech data but rather to obtain the *best* alignment for the given data set. This is in contrast with the notion of training and testing for ASR and machine learning tasks, in which the goal of training is to develop acoustic models that will perform accurately for new data sets. In this case, researchers will generally avoid testing their models on the same data that were used for training in order to avoid overfitting.

dictionary contained only target-like segmental transcriptions, regardless of whether the actual production was realized as target-like. On the other hand, the *Customized* pronunciation dictionary contained transcriptions of utterances exactly as they had been phonetically transcribed for that child. For example, in the case that a sibilant /s/ was phonetically realized as a /t/, it would be analyzed as a /t/ in alignment conditions utilizing the *Customized* dictionary and analyzed as an /s/ in alignment conditions using the *Standard* dictionary.

Manual Segmentation

Manual segmentations were collected for both corpora for comparison to the automatic segmentations. For *Julia,* manual segmentation of voiceless stops, voiceless sibilants, and vowels was completed by research assistants using conventional criteria in Praat (Boersma & Weenink, 2011). Phoneme boundaries that were too difficult to determine due to background noise or ambiguity in the signal (e.g., two stops with no release between them) were discarded. For *Paidologos,* manual segmentations of word-initial consonants and the following vowels were provided with the corpus.[3]

Comparisons

Manual and automatic segmentations were compared across each of the TRAINING and DICTIONARY conditions for all segments of interest. For *Julia,* four TRAINING conditions by two DICTIONARY conditions led to eight total alignment conditions. *Paidologos* was aligned under four conditions, as alignment did not vary by dictionary. Two broad measures of accuracy were included for analysis: (a) alignment accuracy, designed to capture whether the aligned segments overlapped with the corresponding manual alignment, and (b) temporal accuracy, which captured differences in duration and boundary placements between the overlapping aligned segments and the corresponding manual alignments.

Alignment accuracy was measured by the proportion of force-aligned segments that occurred in *approximately* the correct location. The operational definition of "approximately correct" for this study was as follows: the force-aligned segment overlapped with the midpoint of the corresponding manually aligned phone. Such segments were considered "matched" with the true phone. "Unmatched" force-aligned segments may have overlapped with the beginning or end of the true phone or may not have overlapped at all but crucially did not overlap with the *midpoint* of the true phone. Figure 1 provides examples of matched and unmatched force-aligned segments. In these examples, the segmentations in the second row from the top are manually aligned, and the phones highlighted in yellow are force-aligned. The distinction of "matched" versus "unmatched" was chosen to reflect gross accuracy measures that researchers working with large data sets would be interested in using in order to facilitate automated analysis. Importantly, this metric allowed us to identify whether a phone was *more or less* in the correct position and more accurately identify gross alignment errors, which can be an important component of facilitating semiautomated analysis (Baghai-Ravary, Grau, & Kochanski, 2011).

Measures of temporal accuracy provided closer examination of "matched" segments and included absolute differences of duration, onsets, and offsets between the matched forced and manual alignments. Many speech science researchers may find the gross accuracy measure of %-Match to be of greatest interest, though measures of temporal accuracy are also necessary to evaluate alignment performance in greater detail and to compare to previous work evaluating forced alignment quality, in particular (DiCanio et al., 2012; Gorman et al., 2011; McAuliffe et al., 2017; Milne, 2014; Renwick et al., 2013; Yuan & Liberman, 2011a).

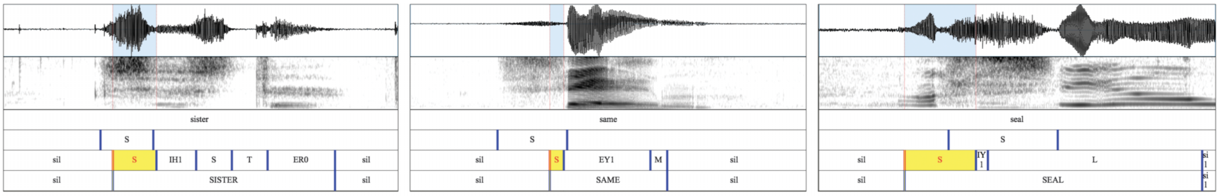
Statistical Models

We modeled alignment accuracy as a function of the parameters described above. We fit one mixed-effects logistic regression of alignment accuracy (matched vs. unmatched segments) for each of the two corpora (*Paidologos* and *Julia*) using the glmer() function from the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2014). We fit one linear mixed-effects regression for each of the three temporal measures (duration, onset, offset, log-transformed after adding 0.001 s) for each corpus (six linear models in total) using the lmer() function from lme4. All categorical variables were coded with contrast schemes such that the intercept was the grand mean and all continuous variables were centered. Therefore, the intercepts of the models reported below may be interpreted as the predicted value of the response (e.g., %-Match) when all predictor variables are held at their average values. Main effect terms may be interpreted as the expected value of the response averaged over other variables given the random effects. Fixed-effect *p* values were calculated using the Satterthwaite approximation as implemented in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2015).

Fixed effects. Fixed-effect predictor variables (identified in SMALL CAPS) in the *Paidologos* model included TRAINING, AGE, and SEGMENT. The model fit for *Julia* included these three predictor variables in addition to PRONUNCIATION DICTIONARY. All possible interaction terms were included in order to examine the potential relationship between variables. A summary of the fixed effects included in Part 1 appears in Table 3. AGE was treated as a continuous variable and *standardized,* that is, centered and divided by 2 *SD*s (Gelman & Hill, 2007). Discrete variables with more than two levels, namely TRAINING (four levels) and SEGMENT (three levels), were coded using Helmert contrasts, which allows the mean of each level to be compared to the overall mean of the subsequent levels. To investigate the effect of the four TRAINING conditions on alignment accuracy

[3] The annotations provided with *Paidologos* were meant to capture an approximation of the segment's boundaries (Beckman, personal communication, 2015). For the analyses in Part 1, this is sufficient to capture the approximate accuracy of forced alignment. For a more detailed analysis involving more precise segment boundaries, annotation was redone on a subset of the data for Part 2.

Figure 1. Alignment examples: Matched and unmatched segments.



(a) Matched (b) Unmatched (c) Unmatched

(*AO, AC, CG, CS*), the Helmert contrast interpretations were as follows: (a) TRAINING1: acoustic models trained exclusively on adult speech versus models trained on some or exclusively child speech (*AO* vs. *AC, CG, CS*), (b) TRAINING2: models trained partially on adult speech versus exclusively on child speech (*AC* vs. *CG, CS*), and (c) TRAINING3: models trained on all children (from both corpora) versus the specific child/children to be aligned (*CG* vs. *CS*). The interpretations of the contrasts for SEGMENT (vowel, sibilant, stop) were as follows: (a) SEGMENT1: vowels versus consonants and (b) SEGMENT2: sibilants versus stops. PRONUNCIATION DICTIONARY was coded using sum contrasts (*Standard* vs. *Customized*). Main effects and two-way interactions of interest are reported here. For greater detail (including two-way interactions not explicitly reported in this text), see the Supplemental Materials.

Random effects. All models for *Paidologos* included by-speaker and by-word random intercepts, as well as all possible by-speaker random slopes (training and segment), in order to account for variability beyond that captured by the alignment parameters. Further random slopes led to problems with model convergence and were therefore omitted, at the risk of anticonservative *p* values (Barr, Levy, Scheepers, & Tily, 2013). The random effects structure for *Julia* included all possible by-utterance random slopes and intercepts (by-speaker random effects were not possible because this corpus contained the longitudinal speech of a single child).

Table 3. Summary of fixed effects included in regression models for Part 1.

Predictor	Subcomparisons	Description
Training	Training1	Adult-only vs. (Adult–child, Child-general, Child-specific)
	Training2	Adult–child vs. (Child-general, Child-specific)
	Training3	Child-general vs. Child-specific
Age	NA	Continuous centered variable (2;0–5;0)
Segment	Segment1	Vowels vs. consonants
	Segment2	Stops vs. sibilants
Dictionary	NA	

Note. NA indicates that there was only a single comparison. Age was treated as a continuous variable, and Dictionary contained a single comparison (Standard vs. Customized).

Results

Alignment Accuracy: %-Match

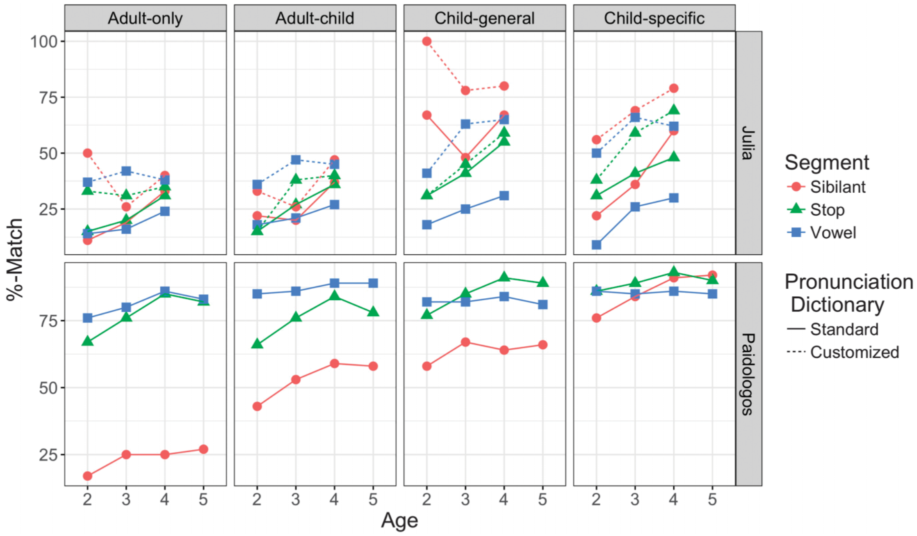
%-Match refers to the percentage of force-aligned segments that overlapped with the midpoint of the correct hand-aligned segment. A "matched" segment thus was force-aligned in approximately the correct location relative to the true (manually aligned) segment. Figure 2 shows how the proportion of matched segments depends on the variables of interest in the empirical data. Fixed effects for the statistical models for *Paidologos* and *Julia* are reported in Supplemental Material S1.

Training. For both corpora, training on adult speech led to poorer accuracy than training on child speech and can be summarized as follows: *Adult-only* < *Adult–child* < *Child speech only*. The distinction between training on adult versus child speech training is captured by TRAINING1 and TRAINING2, which were significant for both corpora (*Paidologos*: TRAINING1: $\hat{\beta} = -1.185$, $p < .001$; TRAINING2: $\hat{\beta} = -0.643$, $p < .001$; *Julia*: TRAINING1: $\hat{\beta} = -1.02$, $p < .001$; TRAINING2: $\hat{\beta} = -1.044$, $p < .001$). With regard to training on child speech, captured by TRAINING3, training exclusively on the speech to be aligned led to better accuracy than training on child speech in general for *Paidologos* ($\hat{\beta} = -0.718$, $p < .001$), but a significant difference was not found for *Julia*. ($\hat{\beta} = 0.067$, $p = .3$).

Age. Alignment accuracy improved with AGE for both corpora (*Paidologos*: $\hat{\beta} = 0.427$, $p < .001$; *Julia*: $\hat{\beta} = 0.655$, $p < .001$). AGE did not interact with TRAINING for any of the comparisons with the exception of TRAINING1 for *Paidologos* ($\hat{\beta} = 0.191$, $p = .02$), indicating that, for the most part, the age of the child did not alter how much of an impact training data improved alignment accuracy (except in the case of training containing exclusively adult speech).

Segment. For clarity and ease of interpretation, only main effects of SEGMENT are reported in the results. For more detail on interactions involving the type of SEGMENT aligned, see the Supplemental Materials. The pattern of most accurately aligned segments was reversed for the two corpora: for *Paidologos,* vowels were aligned with the greatest accuracy, followed by stops and then sibilants, whereas for *Julia,* the order of accuracy was sibilants, stops, vowels. This is captured by a positive main effect of SEGMENT1 and a negative main effect of SEGMENT2 for *Paidologos* (SEGMENT1: $\hat{\beta} = 0.975$, $p < .001$; SEGMENT2: $\hat{\beta} = -1.455$, $p < .001$) and the opposite pattern for *Julia*

Figure 2. Average percentage of matched segments (%-Match) for Part 1 by training, corpus, segment, and pronunciation dictionary (applicable to *Julia* only).



(SEGMENT1: $\hat{\beta} = -0.299$, $p < .001$; SEGMENT2: $\hat{\beta} = 0.361$, $p < .001$).

Dictionary. The difference between the CMU dictionary (*Standard*) and a *Customized* dictionary based on the phonetic transcription was tested for *Julia*. Overall, the *Customized* dictionary led to better alignment accuracy than the *Standard* version ($\hat{\beta} = -0.477$, $p < .001$). All possible interactions with DICTIONARY were found to be significant. A significant interaction with AGE suggests that, although alignment accuracy for both dictionaries improved as the child aged, they became *more similar* as *Julia's* age increased ($\hat{\beta} = 0.206$, $p < .001$). Presumably this reflects the fact that the *Customized* dictionary was tailored to *Julia's* specific utterances at each age, accounting for less room for improvement overall. It could also be due to her productions becoming more adultlike as she aged. Not all training conditions benefited equally from the *Customized* dictionary, accounting for significant interactions between all TRAINING and DICTIONARY comparisons (DICTIONARY, TRAINING1: $\hat{\beta} = 0.163$, $p = .004$; TRAINING2: $\hat{\beta} = 0.271$, $p < .001$; TRAINING3: $\hat{\beta} = 0.179$, $p = .006$). The more customized the training data, the greater benefit the customized pronunciation dictionary provided. For example, *Child-specific* training saw the greatest improvement between the *standard* and *customized* dictionaries (41%–68% matched segments, a difference of 27%), whereas *Adult-only* training benefitted only by 11% (25%–36%).

Temporal Accuracy of Matched Segments

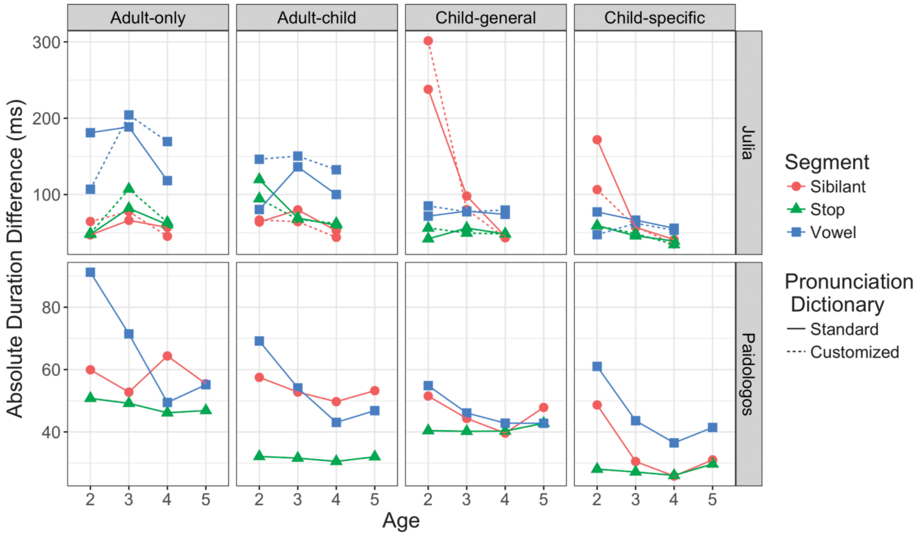
Absolute *duration* differences as a function of AGE, SEGMENT, TRAINING, and DICTIONARY are reported in Figure 3, and absolute *boundary* differences (segment onsets and offsets) are reported in Figure 4. All analyses were done on log-transformed data. Fixed effects for each model of temporal differences between force-aligned and manually aligned segments, including all main effects and interactions of all variables of interest, are reported in full in Supplemental Materials S2 and S3. Because segment duration is often a measure of interest for speech researchers, the absolute differences between force-aligned and manually aligned durations are also important to evaluate. Theoretically, differences between the accuracy of aligning segment onsets and offsets are of interest for researchers interested in phenomena that may occur at phoneme boundaries in child speech.

Absolute duration differences of matched segments. The absolute differences between the durations of force-aligned segments and their corresponding (matched) manual alignments are reported in this section. Training on adult speech led to poorer accuracy (greater durational differences) than training on more specific child speech data. Consistent with %-Matched, the pattern is the same for both corpora: *Adult-only* < *Adult–child* < *Child-general* < *Child-specific*. This is captured by the significant main effects of all TRAINING comparisons (*Paidologos,* TRAINING1: $\hat{\beta} = 0.225$, $p \leq .001$; TRAINING2: $\hat{\beta} = 0.119$, $p \leq .001$; TRAINING3: $\hat{\beta} = 0.417$, $p \leq .001$; *Julia,* TRAINING1: $\hat{\beta} = 0.322$, $p \leq .001$; TRAINING2: $\hat{\beta} = 0.4$, $p \leq .001$; TRAINING3: $\hat{\beta} = 0.213$, $p \leq .001$).

Despite variability shown in Figure 3, overall, durational differences significantly decreased with AGE for *Julia* ($\hat{\beta} = -0.276$, $p \leq .001$), but not for *Paidologos* ($\hat{\beta} = -0.063$, $p = .064$).

Overall, the age effect was not modulated by the type of training, as demonstrated by the absence of significant interactions between AGE and TRAINING. An exception to this is TRAINING3 for *Paidologos,* which contrasts the two child-speech-only TRAINING conditions ($\hat{\beta} = 0.216$, $p \leq .001$). As can be seen in Figure 3, the *Child-general* training condition

Figure 3. Average absolute duration differences for Part 1 by training, corpus, segment, and pronunciation dictionary (dictionary applicable to *Julia* only).



demonstrates an overall flatter rate of improvement (decrease in durational differences) compared to the *Child-specific* training. That is, AGE had a greater effect on alignments with *Child-specific* than *Child-general* training.

Force-aligned vowels demonstrated smaller durational differences than consonants for *Paidologos,* though the opposite was observed for *Julia*. For *Paidologos,* sibilants showed greater durational differences than stops but did not significantly differ for *Julia* (*Paidologos*, SEGMENT1: $\hat{\beta} = -0.157$, $p < .001$; SEGMENT2: $\hat{\beta} = -0.093$, $p = .005$; *Julia,* SEGMENT1: $\hat{\beta} = 0.449$, $p < .001$; SEGMENT2: $\hat{\beta} = -0.055$, $p = .281$). Interactions between SEGMENT1 and AGE for both corpora indicate that vowels and consonants were affected by age differently across the two corpora. Age had a greater effect on consonants than vowels for *Julia* and a smaller effect on consonants than vowels for *Paidologos* (*Paidologos*: $\hat{\beta} = -0.236$, $p < .001$; *Julia*: $\hat{\beta} = 0.24$, $p = .013$).

There was no main effect of DICTIONARY for *Julia,* indicating that the specificity of the transcription provided did not significantly affect the force-aligned phoneme durations. There were no significant interactions between TRAINING or AGE and DICTIONARY, though significant interactions between DICTIONARY and SEGMENT indicated that, overall, the *Customized* dictionary led to less accurate alignment of vowels and stops compared to better alignment of sibilants (SEGMENT1: $\hat{\beta} = -0.098$, $p = .006$; SEGMENT2: $\hat{\beta} = 0.096$, $p = .009$).

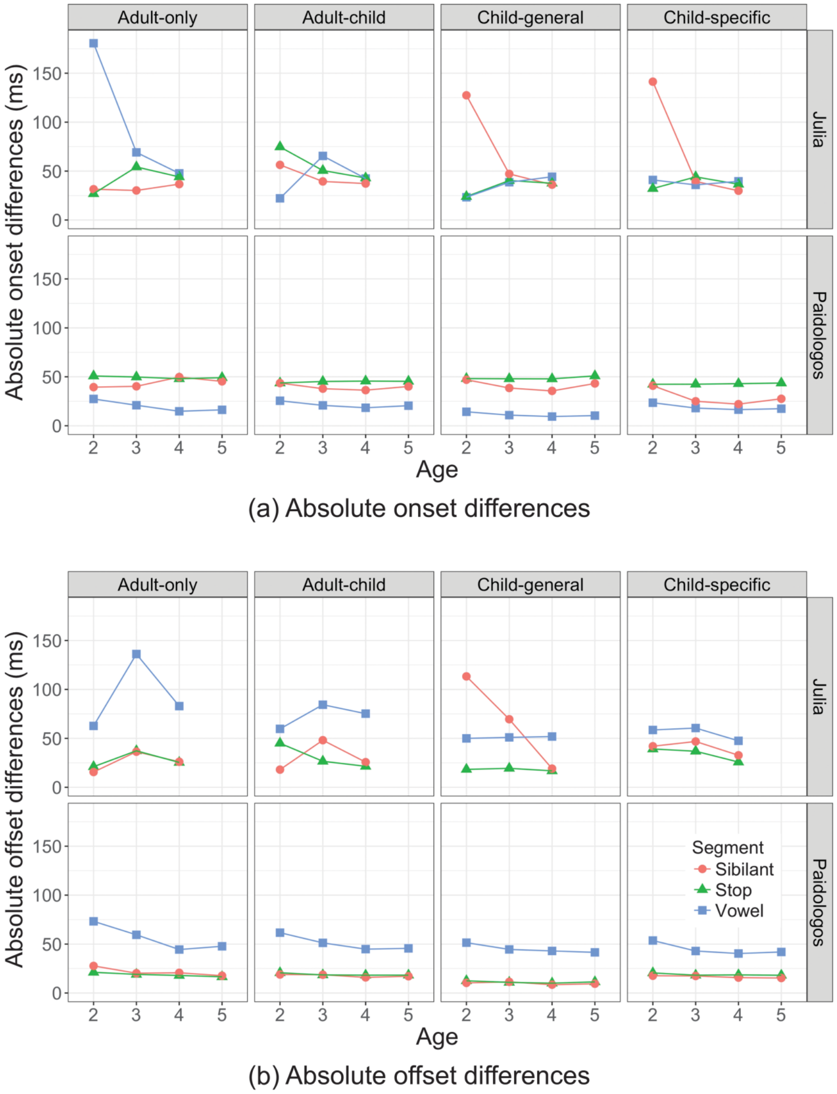
Absolute boundary differences of matched segments. For simplicity, only main effects are reported in this section, and emphasis is placed on comparisons that differed between onsets and offsets. Full coefficient tables for absolute onset and offset differences, including all main effects and interactions of all variables of interest, can be found in Supplemental Materials S3 and S4.

As with %-Match and durational differences, in general, training on adult speech yielded worse outcomes (significantly larger boundary errors). Almost all TRAINING conditions demonstrated a significant main effect on both onset and offset differences for both corpora, with the exception of TRAINING3 onsets and TRAINING1 offsets for *Paidologos*. These findings mirror the global pattern, seen in Figure 4, of lower absolute differences for more specific (i.e., less adult speech) training. As seen previously, the speech of older children was also generally aligned with greater accuracy with regard to boundary differences. However, AGE did not impact the accuracy of consonant onsets for *Paidologos,* which were more poorly identified by the aligner, as seen in Figure 4 (see discussion below).

Across the different segment types, differences emerged between how well onsets and offsets were aligned. Specifically, for *Paidologos,* vowel onsets were more accurately aligned than vowel offsets, whereas consonant onsets were more *poorly* aligned than offsets, as can be seen in the bottom panels of Figure 4. This pattern was not systematic for *Julia*. One possible reason for this discrepancy may have been the elicitation method in *Paidologos*: the consonants studied here all occurred word initially, which may have been a more difficult task for the aligner. It appears that consonant-vowel boundaries in particular, that is, consonant offsets and vowel onsets, may have been easier for the aligner to detect. Finally, the *Customized* dictionary for *Julia* led to improvements in boundary accuracy, consistent with improvements in other accuracy measures presented above.

In summary, overall alignment accuracy as measured by general phone identification (%-Match) and temporal accuracy measures (durational and boundary differences) was better with older children and when using training data that were similar to the speech being aligned (i.e.,

Figure 4. Average absolute differences for segment onset and offsets for Part 1 by training, corpus, segment. Only the standard dictionary is pictured for simplicity.



(a) Absolute onset differences

(b) Absolute offset differences

more child data). There was a wide range of error rates, ranging from < 25% to 100% matched segments across conditions and temporal differences ranging from 0 ms to > 1 s. These errors are of interest as measures of alignment quality, but do they actually affect the conclusions researchers would draw from analyzing these data? We turn to this question in Part 2.

Part 2: Using Forced Alignment to Examine Spectral Properties of Child Sibilant Productions

Bang et al. (2017) examined word-initial /s/ productions from children included in the *Paidologos* corpus and found an increase in spectral CoG in older children, as well as a divergence in CoG in male and female children, with older female children producing higher CoG than younger children and older male children. Using the acoustic measures obtained from the original manual segmentation from Bang et al. (2017) as a comparison, Part 2 of this study sought to determine whether forced alignment, without additional manual adjustment, could be used to replicate the results of the original study.

Method

To replicate the methodology of Bang et al. (2017), we selected two alignment conditions we considered representative of available options in real-world research settings: alignments trained on *Adult-only* speech, representing the out-of-the-box acoustic models available with the Prosodylab-Aligner, and *Child-specific* speech, which generated acoustic models by training the aligner on the exact data set to be aligned. Recall that the results from Part 1 identified that, overall, these conditions also represented the worst (*Adult-only*) and the best (*Child-specific*) alignment accuracy for *Paidologos.* Comparisons were made to the *manual* segmentations provided by Bang et al. (2017), which were slight adjustments made to the segments provided with the *Paidologos* corpus. All /s/ segments of interest, regardless of whether they passed our "matched" accuracy measure, were included. The general pattern demonstrated in Part 1, namely, that alignments performed using acoustic models generated with *Child-specific* training led to more accurate alignments than those with *Adult-only* training, held for the subset of /s/ segments of interest in this section.

Acoustic Analyses: Spectral CoG

The first spectral moment, CoG, was obtained from a discrete Fourier transform spectrum computed by averaging six spectra of 15 ms evenly distributed across the middle 80% across the fricative (Shadle, Koenig, & Preston, 2011), as was done in Bang et al. (2017). This procedure was repeated once for each of the TRAINING conditions as well as for the manual alignment, resulting in three CoG measures per fricative token.

Statistical Models

We used linear mixed-effects regression (refer to Part 1 model details) to model CoG of /s/ as a function of AGE and ALIGNMENT: *Manually aligned, Adult-trained* (force-aligned, trained on adult lab speech), and *Child-trained* (force-aligned, trained on the same child speech to be aligned). In order to replicate the analysis performed in Bang et al. (2017), the model also included fixed effects of speaker SEX and the interaction between AGE and SEX to determine if differences in male and female speakers increased with age. AGE and SEX were standardized as previously described (treated as a continuous variable and standardized).

All models included by-word and by-speaker random intercepts to account for the variability in the acoustic measures beyond the effects of the primary variables of interest. In addition, we included all possible by-word and by-speaker random slopes to account for variability among items and speakers. Correlations between random effects terms were omitted to facilitate model convergence.

Results

CoG

CoG was measured for all /s/ tokens of interest (regardless of alignment accuracy) across three ALIGNMENT conditions (*Manual* alignment, *Adult-trained* forced alignment, *Child-trained* forced alignment). Figure 5 shows how CoG varies by alignment condition and child age and gender. Full model results are reported in Supplemental Material S5. There was no main effect of ALIGNMENT ($p > .5$), nor were there any interactions involving ALIGNMENT ($p > .25$ for all possible interactions). That is, the overall CoG measured using both forced alignments, regardless of alignment accuracy, was similar to the CoG measured from the manual alignments. A significant positive effect of AGE ($\hat{\beta} = 1097.294$, $p = .001$) indicates that, overall, older children produced /s/ with higher CoG. However, the significant interaction between AGE and SEX ($\hat{\beta} = -1797.488$, $p = .006$) reveals that CoG continued to increase with age for girls, but decreased for boys. This can be seen in Figure 5, in which CoG diverges for boys and girls after age 3. These findings replicate the results of Bang et al. (2017).

No significant interactions between ALIGNMENT and AGE or SEX were found; changes in CoG as a function of these speaker variables were captured equally well by all ALIGNMENT conditions. That is, a similar pattern of CoG increasing with age and diverging for boys and girls after the age of 3 was found across all alignment conditions.

Summary: Replication of Bang et al. (2017)

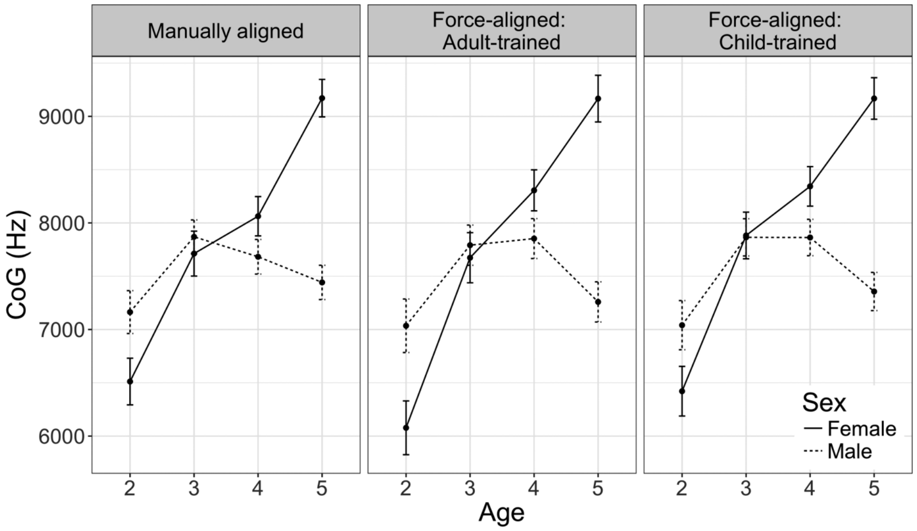
Despite inaccuracies in alignment as seen in Part 1, the use of forced alignment—even when trained only on adult data—did not significantly affect the conclusions made by analyzing the CoG of the /s/ segments of interest. In other words, the same qualitative results for Bang et al. (2017) would have been obtained with either manually aligned or force-aligned data. This is interesting given that we found that, for the alignments trained on adult data only, only a small percentage met the "Match" criterion (< 25%). Thus, it appears that force-aligned segments need not overlap with the midpoint of the true phone (which constitutes a positive "match") in order to capture an accurate representation of the spectral frequency distribution. We explore possible underlying reasons for this in the Discussion.

Discussion

In this study, we explored the consequences of changing specific parameters of forced alignment on alignment accuracy, as well the viability of using forced alignment to facilitate acoustic analysis in child speech. The findings described above demonstrate that modifying inputs to forced alignment do indeed have quantifiable ramifications on the accuracy of the segmentation. However, despite inaccuracies in alignment (and especially for the standard out-of-the-box pretrained alignment), forced alignment allowed replication of the findings of Bang et al. (2017) regarding CoG in /s/ as a function of age and sex in young children.

Overall, increased alignment accuracy (as measured by %-Match) was found with the *Paidologos* corpus (picture-prompted single word repetition) compared to the *Julia* corpus (naturalistic spontaneous speech). Although there are too many uncontrolled differences between these two corpora to draw concrete conclusions for the asymmetry, certain variables are likely to have had an effect. First of all, spontaneous speech is a more challenging task for forced alignment in general, in large part because reductions and substitutions in continuous speech reflect a different acoustic realization than what may be expected from the canonical pronunciation dictionary (Benzeghiba et al., 2007). Single-word utterances are more isolated acoustic

Figure 5. Mean center of gravity (CoG) values for manually aligned and force-aligned /s/ (*Adult-trained* and *Child-trained*). Error bars represent standard error.



events than running speech, and automated methods may more easily be able to determine word and segment boundaries. Second, a naturalistic setting, such as a play-based interaction in a room with toys in which the *Julia* recordings took place, may allow for greater levels of background noise as the child moves around and plays with objects.

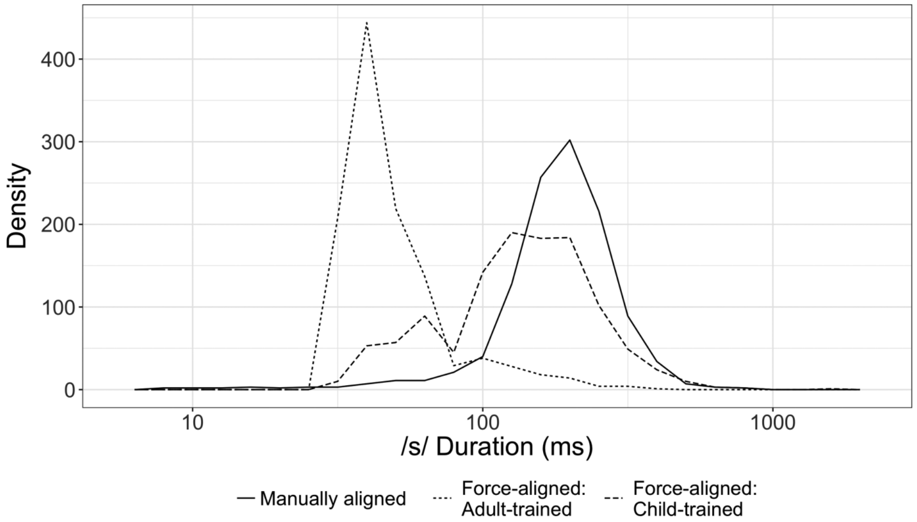
Regarding customizable components of forced alignment, transcription (i.e., dictionary) and training both led to better performance when they were more similar to the audio to be aligned. When a full narrow phonetic transcription of the child speech was available, as was the case with *Julia,* a customized pronunciation dictionary specific to the child's actual utterances led to better performance. This is not surprising, as children may use different phonetic realizations to approximate typical adult speech productions. Substitutions and omissions typical in early speech development may mean that a transcription representing the adult target utterance may not accurately correspond to the child's production. A more specific transcription allows forced alignment to more accurately map the transcription to the acoustic signal, thereby improving its performance. Speech from older children was also aligned with greater accuracy. This too is, in part, related to the reduction in overall variability in the distortions and phonetic realizations of a child as they begin to produce more adultlike speech.

Of all training conditions, input training data containing *entirely* child speech consistently led to better outcomes in alignment accuracy for both corpora. In this study, the similarity of the training audio to the audio to be aligned was of greater importance than other benefits that adult speech might yield, such as greater consistency or clearer targets. That is, training on the specific type of speech to be aligned captured the acoustic properties of child speech that differ from adult speech. These results support similar findings in the literature on understudied populations, that the more similar the phones to be aligned are to the phones on which alignment is modeled, the more accurate the output (e.g., Wilpon & Jacobsen, 1996, for children and the elderly; DiCanio et al., 2012, 2013, for endangered languages).

A notable asymmetry existed between the two corpora for the *Child-specific* training data. *Child-specific* training for *Julia* contained the spontaneous speech of a single child and overall less audio compared to *Child-specific* training for *Paidologos,* which contained the speech of multiple children and more speech overall. This could, in part, explain why *Child-specific* did not provide additional benefit over *Child-general* for *Julia,* despite doing so for *Paidologos*. It is presently unknown at what point the amount and quality of the training data fails to lead to better alignment. McAuliffe et al. (2017) examined this question in adult laboratory speech, concluding that, although further investigation is required, training on more similar data often yields improvement over greater quantities of data. Future research would benefit from examining different training conditions and more precisely controlling for the amount of audio data provided for training.

The finding that more specific training led to improvements in alignment accuracy overall held for the sibilant analysis in Part 2. Curiously, regardless of the finding that the majority of /s/ segments did not overlap with the midpoint of the "true" segment in alignments trained on adult speech (< 25% "match"), both alignment conditions were still able to lead to replication of the CoG measures obtained by Bang et al. (2017). To explore the potential underlying causes of how mediocre automatic alignments could still yield the same results as a study using manual alignments, we pursued a more detailed analysis of our /s/ alignments. Specifically, we examined two aspects of alignment: (a) durational measures and (b) whether or not the segment of interest overlapped *in some way,* but perhaps not in a way that was captured by the %-Match criterion. Figure 6

Figure 6. /s/ Duration (ms).



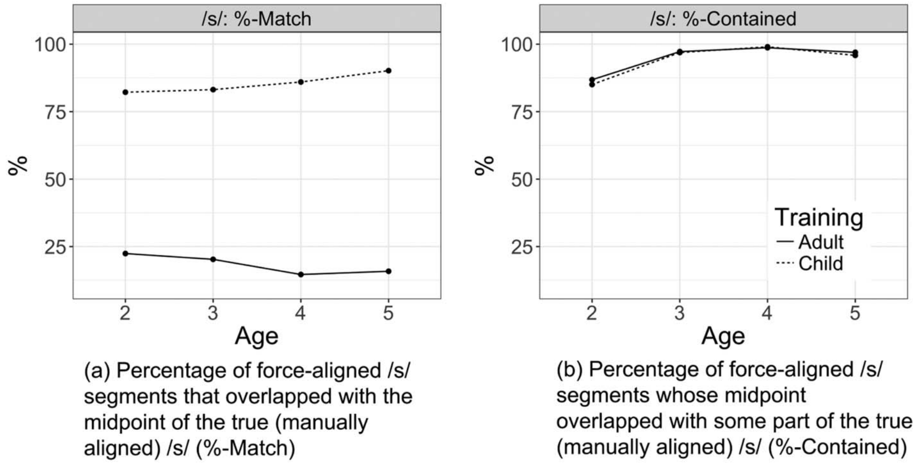
demonstrates an asymmetry in the two alignment conditions with regard to duration: /s/ segments aligned in the Adult-trained condition were much shorter than those aligned in the Child-trained condition. Manually aligned /s/ segments were longest of all. This indicates that, although both force-aligned conditions led to shorter /s/ durations, the Child-trained /s/ were more "childlike" with regard to their duration. This finding thus does not explain why CoG measures, which are calculated over the whole duration of a segment, were so similar across conditions.

We thus next explored whether a different measure of accuracy aside from %-Match would help to explain the CoG findings. Specifically, we looked at whether the midpoint of the *aligned* segment occurred within the boundaries of the true phone. In contrast with the Match criterion, where the midpoint of the true phone overlapped with the aligned segment, this new criterion, herein referred to as %-Contained, provided a less stringent measure of accuracy. An example of this appears in Figure 1b. Figure 7 demonstrates that, although a large difference existed between the two training conditions for %-Match, nearly identical performance was found with %-Contained. That is, the majority of force-aligned /s/ segments, regardless of training, did indeed overlap (at the force-aligned midpoint) with the manual alignment. The finding that both the more and less specific training conditions yielded the same pattern for CoG is likely a consequence of this: Even when the force-aligned /s/ did not land in the middle of the correct phone, it needed only to overlap with at least part of intended signal to reproduce the results of Bang et al. (2017).

Recent work has suggested that fricative productions are not stable and that acoustic variability is present throughout the time course of sibilant production (Iskarous et al., 2008). Nonetheless, the variability did not hinder the acoustic analysis presented in the current study. When CoG was extracted from within the boundaries of the segmentation, both *Adult*- and *Child-trained* alignments yielded the same pattern of results as the more accurate manual segmentations. This may indicate that CoG is a robust spectral measure and perhaps is not as sensitive to dynamic changes across the course of the fricative. We did not explore other acoustic cues or phonemic classes in this study. As such, this finding is not necessarily generalizable to the use of automation in all cases. Given the poor performance regarding %-Match for the *Adult* training in particular, analyses using acoustic measures more sensitive to accurate temporal demarcations may be less likely to be replicated than analyses using CoG. Nevertheless, the task of forced alignment is to identify the part of the acoustic signal corresponding to the segment to be aligned. The replication of the sibilant acoustic analysis affirms that, even in variable speech, forced alignment is mostly successful in this task. That is, it is at least successful enough to yield a correct analysis when averaging over enough tokens.

In all cases of automatic segmentation, there were instances of gross alignment errors such that the aligned segment did not capture the relevant part of the acoustic signal. This is not uncommon with automation of very large speech corpora, especially in the case of background noise, untranscribed or inaccurately transcribed speech. Baghai-Ravary et al. (2011) sought to systematically address gross alignment errors in the Spoken British National Corpus by developing algorithms designed to detect suspicious alignment anomalies and alert the user to alignment failures. Such methods would be of value when integrating the use of forced alignment in very large corpus studies of highly variable speech. Further work is needed to conduct a more detailed exploration of alignment parameters to optimize their performance with child speech. In this study, we did not control for amount of training data or length of speech data to be aligned. Nevertheless, the findings of this study suggest the promise of semiautomation for

Figure 7. More (a) and less (b) stringent measures of accuracy for both training conditions in Part 2.



(a) Percentage of force-aligned /s/ segments that overlapped with the midpoint of the true (manually aligned) /s/ (%-Match)

(b) Percentage of force-aligned /s/ segments whose midpoint overlapped with some part of the true (manually aligned) /s/ (%-Contained)

phonetic analysis of child speech and its viability as a tool for speech researchers. Despite limitations, the parameters identified here may improve the accuracy of forced alignment and allow for the investigation of much larger-scale theoretical questions related to variable speaker populations. Most importantly, training on the data to be aligned was quite successful, even with small amounts of data, and phonetic transcriptions also provided clear gains. However, even when using an out-of-the-box forced aligner with poor alignment performance on child speech, forced alignment was still able to reproduce CoG results found with manual segmentation, underscoring the promise of semiautomation for future investigations of child speech. Currently, forced alignment can be performed with freely available software that can be downloaded on any computer and used without advanced technical skills. As technology advances, aligners will only become easier to use and even more accurate than what we found here.

Acknowledgments

This research was supported by the McGill Collaborative Research Development Fund awarded to Meghan Clayards, Aparna Nadig, Kristine Onishi, Morgan Sonderegger, and Michael Wagner. We thank Aparna Nadig, Kristine Onishi, and Michael Wagner for their contributions to a preliminary version of this work, which was presented at the 170th Meeting on Acoustics and appears in the proceedings (Knowles et al., 2015). We would also like to thank Hye-Young Bang for providing the manual alignments and acoustic analysis scripts for these data. Finally, we would like to thank A. Azzimmaturo, L. Bassford, L. Harrison, J. Kobelski, M. Schwartz, and A. Vong for assisting with data preparation.

References

Adda-Decker, M., & Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, *39*(3), 261–270.

Baghai-Ravary, L., Grau, S., & Kochanski, G. (2011). Detecting gross alignment errors in the Spoken British National Corpus. In *VLSP 2011: New tools and methods for very-large-scale phonetics research* (pp. 103–106). Philadelphia, PA: University of Pennsylvania. Retrieved from https://arxiv.org/pdf/1101.1682.pdf

Bang, H.-Y., Clayards, M., & Goad, H. (2017). Compensatory strategies in the developmental patterns of English /s/: Gender and vowel context effects. *Journal of Speech, Language, and Hearing Research*, *60*(3), 571–591. https://doi.org/10.1044/2016_JSLHR-L-15-0381

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Linear mixed-effects models using Eigen and S4. *Journal of Statistical Software*, *67*(1), 1–48. Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

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

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., . . . Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Bigi, B. (2012). SPPAS: A tool for the phonetic segmentations of speech. *Proceedings of LREC 2012* (pp. 1748–1755). Istanbul, Turkey.

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer [Computer program] (Version 5.3). Retrieved from http://www.fon.hum.uva.nl/praat/

Clayards, M., & Doty, E. (2011). Automatic analysis of sibilant assimilation in English. *Canadian Acoustics*, *39*(3), 194–195.

Coleman, J., Liberman, M., Kochanski, G., Burnard, L., & Yuan, J. (2011). Mining a year of speech. In *VLSP 2011: New tools and methods for very-large-scale phonetics research* (pp. 16–19). Philadelphia, PA: University of Pennsylvania. Retrieved from http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf

D'Arcy, S., & Russell, M. J. (2005). A comparison of human and computer recognition accuracy for children's speech. In

Interspeech (pp. 2197–2200). Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_2197.pdf

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., & García, R. C. (2012). Assessing agreement level between forced alignment models with data from endangered language documentation corpora. In *Interspeech* (pp. 130–133).

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., & Garcıa, R. C. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, *134*(3), 2235–2246.

Edwards, J., & Beckman, M. E. (2008). Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics*, *22*(12), 937–956.

Elenius, D., & Blomberg, M. (2005). Adaptation and normalization experiments in speech recognition for 4- to 8-year-old children. *Proceedings of Interspeech 2005* (pp. 2749–2752). Lisbon, Portugal. Retrieved from https://pdfs.semanticscholar.org/4bef/001b8e82c29f32e550c1e6af7340e7d556a3.pdf

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, *49*(10), 847–860.

Gerosa, M., Giuliani, D., & Brugnara, F. (2009). Towards age-independent acoustic modeling. *Speech Communication*, *51*(6), 499–509.

Goad, H. (2010). English-Goad: Online corpus of phonological development. *PhonBank*. Retrieved from http://phonbank.talkbank.org/access/Eng-NA/Goad.html

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

Green, J. R., Moore, C. A., & Reilly, K. J. (2002). The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research*, *45*(1), 66–79.

Iskarous, K., Shadle, C. H., & Proctor, M. (2008). Evidence for the dynamic nature of fricative production: American English /s/. *Proceedings of the 8th International Seminar on Speech Production* (pp. 405–408). Strasbourg, France.

Knowles, T., Clayards, M., Sonderegger, M., Wagner, M., Nadig, A., & Onishi, K. H. (2015). Automatic forced alignment on child speech: Directions for improvement. *Proceedings of Meetings on Acoustics*, *25*(1), 060001. https://doi.org/10.1121/2.0000125

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models (R package version 2.0-29). Retrieved from https://cran.r-project.org/web/packages/lmerTest/index.html

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30–65.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455–1468.

Li, F., Edwards, J., & Beckman, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics*, *37*(1), 111–124. https://doi.org/10.1016/j.wocn.2008.10.001

Li, F., Rendall, D., Vasey, P. L., Kinsman, M., Ward-Sutherland, A., & Diano, G. (2016). The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. *Journal of Phonetics*, *57,* 59–70.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McAllister Byun, T. (2011). A gestural account of a child-specific neutralisation in strong position. *Phonology*, *28*(3), 371–412.

McAllister Byun, T. (2012). Positional velar fronting: An updated articulatory account. *Journal of Child Language*, *39*(5), 1043–1076. https://doi.org/10.1017/s0305000911000468

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montréal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech 2017* (pp. 498–502). Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-1386

Milne, P. (2014). *The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French* (doctoral dissertation). Université d'Ottawa/University of Ottawa, Canada.

Mugitani, R., & Hiroya, S. (2012). Development of vocal tract and acoustic features in children. *Acoustical Science and Technology*, *33*(4), 215–220.

Nissen, S. L., & Fox, R. A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *The Journal of the Acoustical Society of America*, *118*(4), 2570–2578. https://doi.org/10.1121/1.2010407

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *The Journal of the Acoustical Society of America*, *97*(1), 520–530. https://doi.org/10.1121/1.412278

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, *32*(1), 120–132. https://doi.org/10.1044/jshr.3201.120

Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. *Proceedings of 5th European Conference on Speech Communication and Technology* (pp. 2371–2734). Rhodes, Greece.

Renwick, M. E., Baghai-Ravary, L., Temple, R., & Coleman, J. S. (2013). Assimilation of word-final nasals to following word-initial place of articulation in United Kingdom English. *Proceedings of Meetings on Acoustics*, *19*(1), 060257. https://doi.org/10.1121/1.4800279

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) program suite*. Retrieved from http://fave.ling.upenn.edu

Schiel, F. (2004). MAUS goes iterative. *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1015–1018). Lisbon, Portugal.

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96–109. https://doi.org/10.1016/j.wocn.2010.11.006

Schuppler, B., van Dommelen, W. A., Koreman, J., & Ernestus, M. (2012). How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, *40*(4), 595–607. https://doi.org/10.1016/j.wocn.2012.05.004

Shadle, C. H., Koenig, L. L., & Preston, J. L. (2011). Acoustic characterization of /s/ spectra of adolescents: Moving beyond moments. *Proceedings of Meetings on Acoustics*, *12*(1), 060006. https://doi.org/10.1121/1.4862854

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, *55*(4), 779–798.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., . . . Gentry Lindell, R. (2009). Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *The Journal of the Acoustical Society of America*, *125*(3), 1666–1678. https://doi.org/10.1121/1.3075589

Vorperian, H. K., Wang, S., Schimek, E. M., Durtschi, R. B., Kent, R. D., Gentry Lindell, R., & Chung, M. K. (2011). Developmental sexual dimorphism of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of Speech, Language, and Hearing Research*, *54*(4), 995–1010.

Weide, R. (1998). *The CMU Pronunciation Dictionary* (release 0.7a). Pittsburgh, PA: Carnegie Mellon University.

Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 349–352). Atlanta, GA.

Young, S. J., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Woodland, P. (1994). The HTK book. *Cambridge University Engineering Department*, *3,* 175.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*. Paris, France. Retrieved from http://languagelog.ldc.upenn.edu/myl/ICASSP_final.pdf

Yuan, J., & Liberman, M. (2011a). /l/ variation in American English: A corpus approach. *Journal of Speech Sciences*, *1*(2), 35–46. https://doi.org/10.1002/9780470753460.ch10

Yuan, J., & Liberman, M. (2011b). Automatic detection of 'g-dropping' in American English using forced alignment. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 490–493). Okinawa, Japan. Retrieved from http://www.ling.upenn.edu/~jiahong/publications/cn1.pdf