Simon Gonzalez[a,*], James Grama[a] and Catherine E. Travis[a]

Comparing the performance of forced aligners used in sociophonetic research

https://doi.org/10.1515/lingvan-2019-0058
Received February 20, 2019; accepted October 19, 2019

Abstract: Forced aligners have revolutionized sociophonetics, but while there are several forced aligners available, there are few systematic comparisons of their performance. Here, we consider four major forced aligners used in sociophonetics today: MAUS, FAVE, LaBB-CAT and MFA. Through comparisons with human coders, we find that both aligner and phonological context affect the quality of automated alignments of vowels extracted from English sociolinguistic interview data. MFA and LaBB-CAT produce the highest quality alignments, in some cases not significantly different from human alignment, followed by FAVE, and then MAUS. Aligners are less accurate placing boundaries following a vowel than preceding it, and they vary in accuracy across manner of articulation, particularly for following boundaries. These observations allow us to make specific recommendations for manual correction of forced alignment.

Keywords: forced alignment; accuracy comparison; sociophonetics; vowels; workflow optimization

1 Introduction

A first step in conducting (socio)phonetic analysis is to create reliable segmentations between the acoustic signal and phonemic segments. Typically, this is done by orthographically transcribing speech and then time-aligning phonemes with the waveform and spectrogram. Manual alignment is incredibly time-consuming, having been reported to take approximately 800 times longer than the duration of the speech itself to implement (Schiel et al. 2012: 111). However, the past decade has seen advances in computational methods that have allowed for automated phonemic transcriptions, time-aligned at the segment level, to be derived from orthographic transcriptions, time-aligned at the utterance level. This process, known as forced alignment, tremendously increases the number of tokens that can be analyzed, adding to both the power of analyses and the generalizability of results. It is no wonder then that forced alignment has become a mainstay in sociophonetic research (cf. Stuart-Smith et al. 2017).

While forced alignment is highly effective (e.g., Brognaux et al. 2012), accuracy is nevertheless impacted by various factors, including data type and recording quality (Fromont and Watson 2016), phonological context (Cosi et al. 1991: 4; DiCanio et al. 2012: 132), speech rate (MacKenzie and Turton 2020), as well as the aligner itself (Watson and Evans 2016; McAuliffe et al. 2017). There is, however, little in the way of direct comparisons across the multiple aligners that are widely used. This paper, then, presents a comparison of the four major forced aligners in use in linguistic research today: the Munich Automatic Segmentation System (MAUS, Schiel 1999), the Forced Alignment & Vowel Extraction suite (FAVE, cf. Rosenfelder et al. 2014), the Language, Brain and Behaviour Corpus Analysis Tool (LaBB-CAT, Fromont and Hay 2012), and the Montreal Forced Aligner (MFA, McAuliffe et al. 2017). We first present an overview of each of these aligners (Section 2), and then examine how well they perform on spontaneous speech, utilizing Australian English sociolinguistic

[a]All authors contributed equally to this paper.

***Corresponding author: Simon Gonzalez,** ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, Australia, E-mail: simon.gonzalez@anu.edu.au
James Grama and Catherine E. Travis: ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, Australia, E-mail: james.grama@anu.edu.au (J. Grama); catherine.travis@anu.edu.au (C. E. Travis)

interview data (Section 3). To do this, we compare the segmentation produced by each of the forced aligners with those produced by two trained phoneticians, paying particular attention to the impact of preceding and following phonological environment (Section 4). In closing, we offer an interpretation of the differences in accuracy observed (Section 5), and conclude with some recommendations for working with force-aligned data (Section 6).

2 Overview of four common forced aligners

2.1 Components of a forced aligner

Forced alignment for linguistic research typically takes an orthographic transcription as the basis from which to create the corresponding phonemic segmentation. Transcriptions are made machine-readable through the use of a grapheme-to-phoneme (G2P) dictionary – standardized orthographic transcriptions mapped to the phonemic representation (or representations) of each orthographic word. The phonemes are then matched with corresponding acoustic information in the audio file, based on statistical representations of the distinctive sounds in the acoustic model. The acoustic models are most commonly created from Mel Frequency Cepstral Coefficients (MFCC) (Chodroff 2018). For forced aligners used by sociolinguists, these models are built using either the Hidden Markov Model Toolkit (HTK) (Young et al. 2006) or the Kaldi toolkit (Povey et al. 2011). Both are based on Hidden Markov Models (HMMs), which have been fundamental in the development of models for language processing, with Kaldi having been developed more recently (for comparison, see Gaida 2014).

The acoustic models used by a forced aligner may be either *pre-trained* (built from existing datasets of time-aligned speech corpora and incorporated into the aligner prior to the input of data for forced alignment) or established via a *train/align* process (built on the basis of the data fed to the aligner). Both pre-trained and train/align acoustic models can be developed on a variety of speech types (from controlled speech, such as word list data or isolated sentences read aloud, to spontaneous speech), and on different languages and language varieties.

The forced aligners we consider here differ in the toolkits they use (HTK vs. Kaldi), the way the acoustic model is developed (pre-trained vs. train/align), and the data the acoustic model is based on (more controlled vs. more spontaneous speech, from different varieties of English). To date, the impact of these parameters has been relatively understudied. While we do not directly test each factor here, the comparisons presented allow us to draw some hypotheses about the impact they may have.

To contextualize the analyses that follow, we first provide a brief overview of each of the four aligners.

2.2 Munich Automatic Segmentation System (MAUS)

MAUS is a software package available from the Bavarian Archive for Speech Signals (Schiel 1999). It can be installed locally on the user's machine, and can also be accessed online through WebMAUS (Kisler et al. 2017). MAUS uses HTK and works with pre-trained models. There are pre-trained models available for several languages, including English, German and Icelandic, as well as a pan-Australian model for Aboriginal languages (cf., Stoakes and Schiel 2017). MAUS can also be adapted to work with languages lacking pre-trained acoustic models, as has been done, for example, for Barunga Kriol (a variety of an English-based creole spoken in Roper River, North Australia, Jones et al. 2017) and Bora (a Witotoan language spoken in Peru, Strunk et al. 2014). For the Australian English model, MAUS has been pre-trained on a subset of the AusTalk database (Wagner et al. 2010), comprised of word lists and read sentences from 95 speakers. An acoustic model pre-trained on spontaneous speech is also available for New Zealand English, and we consider the difference in performance of the models built from controlled and spontaneous speech below (Section 4.1).

2.3 Forced Alignment & Vowel Extraction (FAVE)

FAVE was developed in the University of Pennsylvania Linguistics Lab, based on the Penn Phonetics Lab Forced Aligner suite (p2fa, Yuan and Liberman 2008), for the purpose of aligning English sociolinguistic interview data (Rosenfelder et al. 2014). While FAVE is used as a locally installed program, much of its functionality can be accessed online via DARLA (Reddy and Stanford 2015). FAVE uses the HTK toolkit and the Carnegie Mellon University (CMU) Pronouncing Dictionary (Carnegie Mellon University 1993–2016). The acoustic model is pre-trained on the SCOTUS corpus (Yuan and Liberman 2008), consisting of 25 hours of hand-aligned oral arguments from the Supreme Court of the United States, recorded over a 50-year period. As well as forced alignment, FAVE offers a pipeline for measuring and extracting formant values from vowels in the aligned data. It has been used widely in studies of English, has been adapted for Spanish (Wilbanks 2018), and has also been applied to other languages using the English acoustic model, for example, Bribri (a Chibchan language spoken in Costa Rica, Coto-Solano and Flores Solórzano 2017) and Bequia Creole (an English-based creole spoken on the island of Bequia, in St. Vincent and the Grenadines, Walker and Meyerhoff 2020).

2.4 Language, Brain and Behaviour Corpus Analysis Tool (LaBB-CAT)

LaBB-CAT is a corpus building and annotation tool (Fromont and Hay 2012). It was created by the New Zealand Institute of Language, Brain and Behaviour, to facilitate the development of the corpus for the Origins of New Zealand English project (ONZE, Gordon et al. 2007). As with the other aligners discussed here, LaBB-CAT is installed locally, and while it has no generally accessible online interface, it is possible to build a server for remote use. Like MAUS and FAVE, LaBB-CAT employs HTK. It is designed specifically to interface with the CELEX dictionary, based on British English (Baayan et al. 1995), but any G2P dictionary can be wrangled to work with the program, as has been done, for example, with the CMU Dictionary for U.S. English, and with a dictionary of Māori (Keegan et al. 2012). LaBB-CAT uses a default train/align system for speech samples longer than five minutes, reverting to a pre-trained model for shorter segments. As a powerful corpus building and management tool, LaBB-CAT has functionality for data tagging (including tagging with formant values that have been extracted via forced alignment, as well as a range of other linguistic, social, and interactional features).

2.5 Montreal forced aligner (MFA)

MFA was developed at McGill University as an update to Prosodylab-Aligner (McAuliffe et al. 2017). It, too, can function as a locally installed program, and some features are accessible online via DARLA. MFA is unique among the aligners evaluated here in that it employs Kaldi. The acoustic model in MFA is based on the LibriSpeech corpus (Panayotov et al. 2015), which consists of 80 hours of audiobooks, read by volunteers primarily from the U.S. (McAuliffe et al. 2017: 501). Similar to LaBB-CAT, MFA offers both pre-trained and train/align models. Pre-trained models are currently available in 22 languages, including U.S. English (which is what we use for the comparisons here), as well as, for example, Korean, Mandarin, Swahili, Hausa, and Russian. Using train/align, MFA can be applied to languages lacking an acoustic model, and has been found to achieve similar accuracy to that achieved for English data (e.g., for Matukar Panau, an Oceanic language spoken in Papua New Guinea, Gonzalez et al. 2018). For the purpose of comparability, in the analyses here we utilize the pre-trained model.[1]

Table 1 provides a summary of the components of the four aligners examined here, as they were parameterized for the analyses presented.

1 Preliminary testing suggests that MFA performs similarly using train/align or pre-trained models.

Table 1: Components for each aligner as implemented in this study.

Program	Toolkit	Acoustic model	Speech type for training data	
			Controlled	Variety
MAUS	HTK	Pre-trained	Yes	Australian English
			No	New Zealand English
FAVE	HTK	Pre-trained	No	U.S. English
LaBB-CAT	HTK	Train/align	No	Australian English
MFA	Kaldi	Pre-trained	No	U.S. English

3 A methodology for comparing the performance of forced aligners

In order to assess the performance of these forced aligners, we utilize transcriptions of sociolinguistic interviews with four Anglo Australians, two males and two females, one adult and one teenager of each sex. These recordings were made on cassette tape in the late 1970s for the Sydney Social Dialect Survey (Horvath 1985). The findings discussed here are thus applicable to data collected in less-than-ideal recording conditions (as is often the case for legacy data). A total of approximately 1.5 hours of speech, or 16,000 words, were aligned, relatively evenly divided across the four speakers.[2]

The recordings were orthographically transcribed in short utterance lengths (mean = 9.85 syllables/utterance), generally between pauses, to create intervals that would maximize aligner performance. The four files were then separately aligned at the segment level by each of the four forced aligners. We extracted 10% of the transcribed section for manual correction by two trained phoneticians (the first and second authors), who independently made boundary adjustments based on auditory and acoustic cues. To guard against bias by the output of any one aligner, the two human coders corrected output generated by a different aligner for each speaker.

We focus on vowels in stressed syllables, thus avoiding possible confounds of phonological reduction for unstressed vowels, and consistent with standard practice in sociophonetic studies. This process yielded 1079 vowel tokens on which to test the performance of the aligners (ranging from 229 to 328 tokens per speaker).

As the forced aligners we compare use dictionaries based on distinct varieties of English, labels were coerced to represent the aligned variety, Australian English, allowing for reliable comparison of the same segments. For example, in coercing the output from FAVE and MFA, post-vocalic /r/ was merged into the preceding vowel, and unstressed vowels were assigned to /ə/ (consistent with CELEX dictionary representations). In cases where aligners selected distinct variants of the same words (e.g., two- vs. three-syllable variants of *different*), representations were converted to the shortest form.

Alignment outputs were compiled for comparison in Praat textgrids. Figure 1 gives a sample textgrid with tiers corresponding to each comparison. The first tier represents the input orthographic transcription and the second tier corresponds to the human benchmark, produced by the first human coder, which serves as the comparison point for all other tiers. The third tier represents the alignment produced by the second human coder for the human-to-human (H2H) comparison, and the remaining tiers represent each of the four forced aligners. While some boundaries are very close across the aligners (e.g., between /s/ and /ʌ/, in *someone*), others show marked differences (e.g., between /e/ and /l/ in *else*). We will see below that preceding fricatives are generally not problematic for aligners, while following laterals are.

2 The four interviews are: AAF_151, AAM_305, ATF_166, and ATM_128, drawn from the Sydney Speaks corpora (Travis et al. In Progress). There was some difference in the quality of the alignment across these speakers, in particularly for one file (AAF_151), which we attribute to its poorer recording quality.

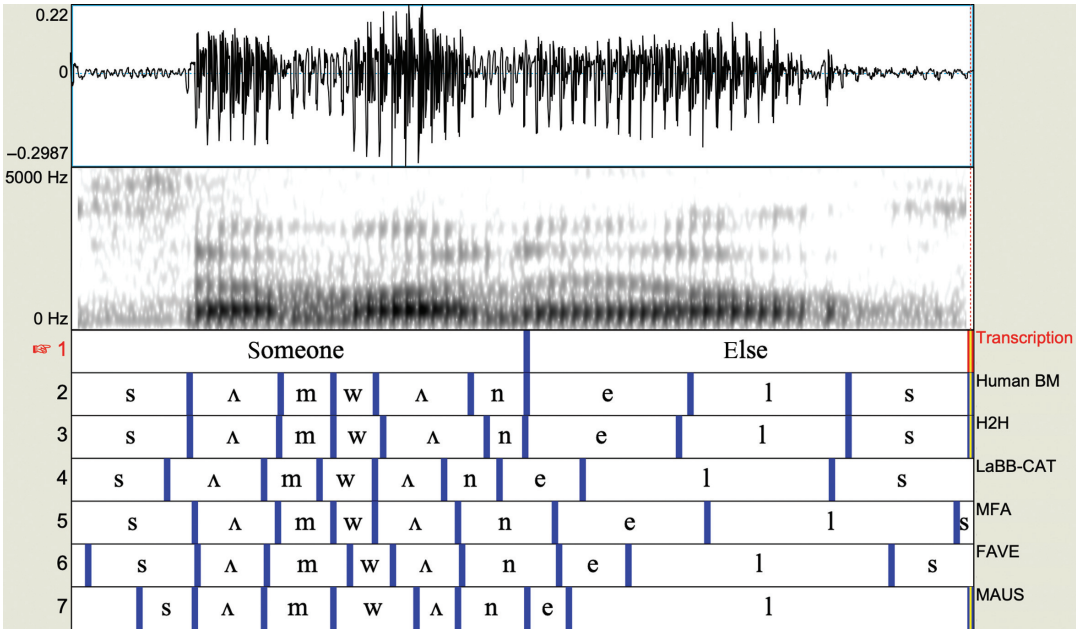

Figure 1: Sample Praat textgrid comparing the human benchmark with the second human coder (H2H) and the four aligners.

4 Results: Measuring accuracy

We employ two measures to evaluate the accuracy of the automated segmentation: *Overlap Rate*, which calculates the proportion of overlap between the intervals established by the human benchmark and the comparison alignments; and *Boundary Displacement*, which calculates the time difference in milliseconds between the boundary placed by the human benchmark and the comparison alignments (cf., Coto-Solano and Flores Solórzano 2017; McAuliffe et al. 2017). We consider the results for each of these in turn.

4.1 Overlap rate

Overlap Rate (OR) was calculated relative to the human benchmark. We applied a formula developed by Paulo and Oliveira (2004: 39), presented in (1) below, in which dur_{SHARED} is the time (in milliseconds) shared between the interval produced by the human benchmark (dur_{HUM}) and the comparison alignment, that of the second human coder or the forced aligners (dur_{COMP}). This formula gives a score from 0 (representing no overlap) to 1 (representing complete overlap). A linear mixed-effects model was fit to OR, with forced aligner identity and scaled vowel duration as fixed effects, and speaker, vowel identity, and phonological context (preceding and following) as random intercepts (see Appendix, Table 2).[3]

$$\mathrm{OR} = \frac{dur_{SHARED}}{dur_{HUM} + dur_{COMP} - dur_{SHARED}} \quad (1)$$

Figure 2 depicts the estimates and confidence intervals from the model; the human-to-human (H2H) comparison serves as the baseline. As can be seen, H2H exhibits the highest estimated OR at nearly 80%, significantly better than MFA at 65%; LaBB-CAT and FAVE are in turn both significantly lower than MFA, at around 57%; and MAUS, at 53%, is significantly lower than all other aligners.

In order to test whether the relatively low OR in MAUS is due to it being trained on more controlled speech samples, we fit another model that included a build of MAUS trained on spontaneous New Zealand

3 Linear mixed-effects models were run in R (R Core Team 2018) using *lme4* (Bates et al. 2010), and *p*-values were calculated using *lmerTest* (Kuznetsova et al. 2017).

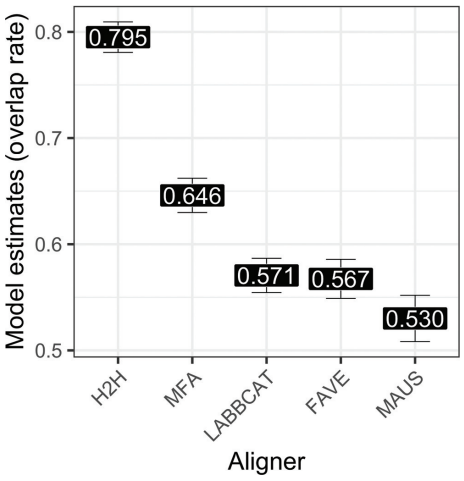

Figure 2: Model estimates for Overlap Rate across aligners (from Table 2).

English speech (see Appendix, Table 3). MAUS trained on spontaneous speech produced an OR of 54%, a non-significant improvement over MAUS trained on controlled speech, and still significantly below the other aligners. We also independently tested the four aligners on a wordlist of 16 words from these same speakers (using the build of MAUS pre-trained on word lists and read sentences), and the relative performance of each was comparable to that for the spontaneous speech data. This indicates that the poorer performance of MAUS is not attributable to the type of training data used.

We also observe that accuracy varies as a function of phonological context. Random intercepts from the model indicate that laterals, nasals, and approximants tend to produce lower ORs than stops and fricatives. To explore the impact of phonological environment, we turn to Boundary Displacement.

4.2 Boundary displacement

Boundary Displacement (BD) provides a measure of the difference between the boundaries set by the human benchmark from that of the comparison alignment. Lower BD represents lesser distance, and thus corresponds to greater accuracy. The formula applied is given in (2), where $bound_{HUM}$ represents the timepoint in milliseconds at which the boundary for the human benchmark was placed, and $bound_{COMP}$ that of the comparison alignment.[4]

$$|\text{BD}| = bound_{HUM} - bound_{COMP} \tag{2}$$

For initial investigation, a linear mixed-effects model was fit to BD with forced aligner identity, scaled vowel duration, and position (preceding vs. following) as fixed effects, and speaker and vowel identity as random intercepts (see Appendix, Table 4). BD was marginally lower for preceding compared with following segment ($p = 0.058$), indicating that higher accuracy is achieved in boundary placement at the onset of a vowel compared with the offset (cf. DiCanio et al. 2012: 133).[5] Again, the human coders produced the most similar results, with an average BD of 10 ms for preceding and 14 ms for following environment.

BD was further significantly impacted by aligner identity. Specifically, MFA and LaBB-CAT performed equally well (with mean BDs of 28 ms and 23 ms, respectively for preceding context and 30 ms and 28 ms

4 Note that this formula returns an absolute value; we leave for future work comparisons of whether the force-aligned boundary sits within the human-marked interval or outside it.

5 Preceding and following segment includes both across and within words. Boundary position (word-initial, word-medial, word-final) when included in the model does not reach significance, suggesting that the effects observed are generalizable across and within the word.

for following), and significantly better than FAVE or MAUS (41 ms and 69 ms for preceding, 45 ms and 73 ms for following, respectively). This differs from the results for OR, in which MFA outperformed all of the other aligners. The difference between the performance of MFA and LaBB-CAT according to these two measures can be partially accounted for in terms of the placement of the force-aligned boundaries relative to the human benchmark. Since OR calculates the overlap between intervals, a higher rate of overlap is attained in cases where one interval encompasses the other – either the benchmark interval falls within the automated interval, or the other way around. While close to half (46%) of the annotated intervals follow this pattern for MFA, just one quarter (26%) of the intervals do so for LaBB-CAT. This contributes to greater OR for MFA, without necessarily resulting in boundary placements that are closer to the human benchmark.

We found no effect of place of articulation for either preceding or following segment; we therefore focus the discussion on manner of articulation. Separate models were fit to BD for preceding and following segment. Each of these models included forced aligner identity, scaled vowel duration, and manner of articulation as fixed effects, with an interaction between forced aligner and manner; speaker and vowel identity were random intercepts.[6]

4.2.1 Preceding phonological segment

Figure 3 presents the estimates from the model fit to preceding BD across aligner and manner of articulation. H2H comparisons uniformly produce the most similar alignments, with no significant differences across the

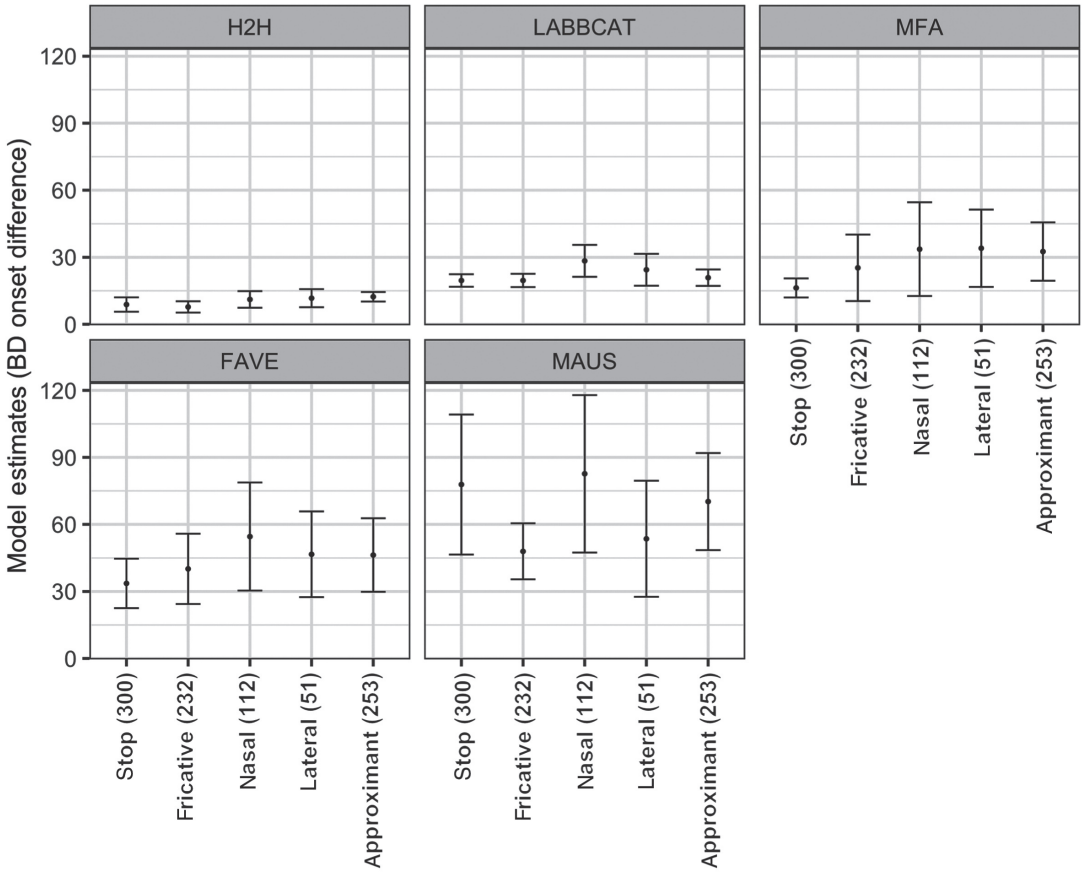

Figure 3: Model estimates for Boundary Displacement across aligners: Preceding manner of articulation (from Table 5).

6 We excluded tokens in unclear speech ($n = 67$ preceding, $n = 67$ following) and environments with low token numbers, including preceding ($n = 23$) or following affricates ($n = 21$) and preceding ($n = 21$) or following vowels ($n = 31$).

distinct manners (see Appendix, Table 5). Notably, while the boundaries placed by the second human coder appear to be closer to those of the human benchmark than LaBB-CAT and MFA, these differences are not significant, indicating that (at least for vowel onsets) these programs produce alignments that are very similar to those of human coders. Both FAVE and MAUS, on the other hand, produce significantly higher BD than the H2H comparison, with MAUS exhibiting by far the highest displacement.[7]

Aligners behave relatively uniformly with respect to manner of articulation; preceding nasals, laterals and approximants generally yield (non-significantly) higher BD (cf., Cosi et al. 1991; DiCanio et al. 2012: 133) and have greater variance than fricatives and stops. MAUS, however, performs significantly better with preceding fricatives than stops; and it returns relatively large confidence intervals for preceding stops and nasals, indicating less internally consistent alignments in these contexts. By contrast, alignments produced by LaBB-CAT exhibit relatively small confidence intervals, indicating more internally consistent alignments than those of other forced aligners.

4.2.2 Following phonological segment

Figure 4 presents the estimates from the model fit to following BD across aligner and manner of articulation (see Appendix, Table 6). As with preceding environment, the second human coder uniformly produces

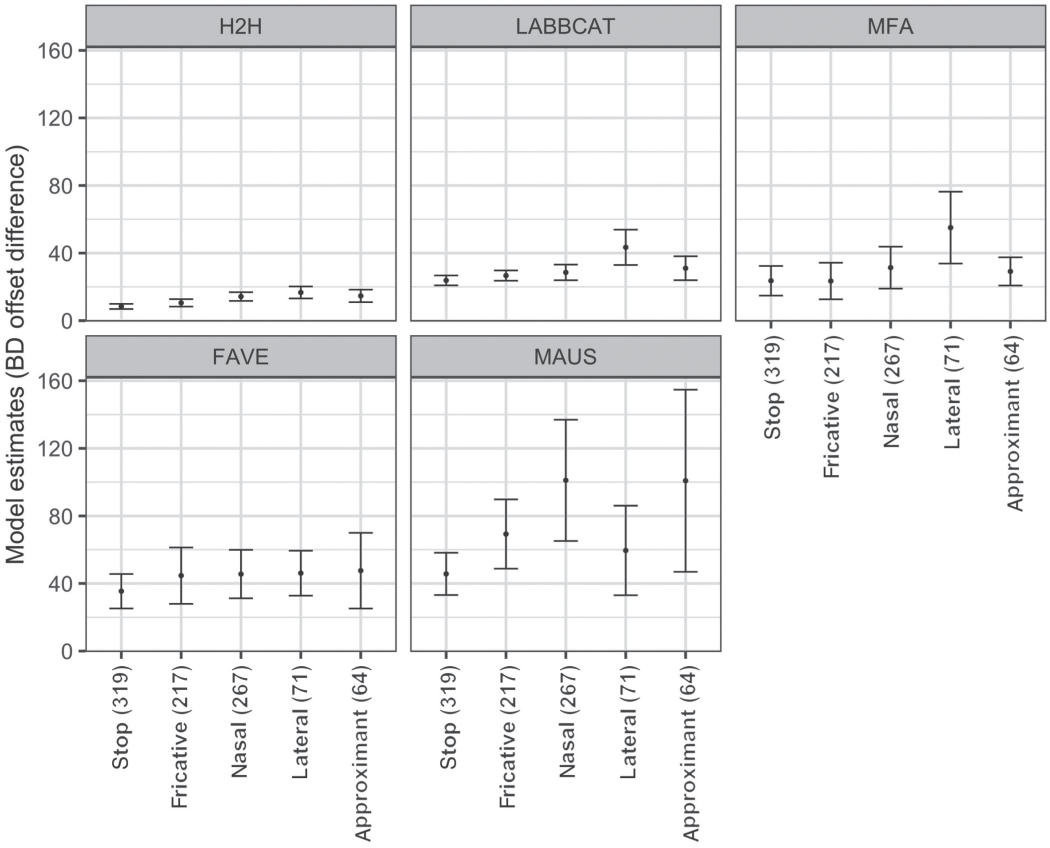

Figure 4: Model estimates for Boundary Displacement across aligners: Following manner of articulation (from Table 6).

7 In a study applying MFA (run through DARLA) and FAVE to British English sociolinguistic interview data, MacKenzie and Turton 2020) similarly find that MFA places onset boundaries very close to those placed by humans, but in contrast to the current study, they find little difference between MFA and FAVE. What might account for these different results is unknown, but two key differences are the comparison method (they compared boundaries placed by a human coder vs. forced alignment, whereas we have compared boundaries placed by two human coders with those placed by a human coder vs. forced alignment), and the object of study (they tested the alignment of all phones, while we have tested that of stressed vowels).

alignments closest to the human benchmark, with little difference across manner; and both MAUS and FAVE are outperformed by MFA and LaBB-CAT, with no significant differences arising between the former or latter two. Unlike what we observe with preceding environments, MFA ($p = 0.07$) and LaBB-CAT ($p = 0.066$) produce higher BD than the H2H comparison, consistent with the general observation that following phonological environments are more problematic than preceding environments.

Also unlike preceding phonological environment, aligners differ substantially in how they behave across manners of articulation. While FAVE and LaBB-CAT exhibit no significant differences across manner of articulation, this is not the case for MAUS and MFA. MAUS returns higher BD in the context of following approximants and nasals than it does with following stops or laterals, and a releveled model indicates that MFA returns significantly higher BD for following laterals as compared with other manners. LaBB-CAT, again, exhibits smaller confidence intervals, indicating greater internal consistency in its boundary placement.

5 Interpreting accuracy differences

As an overall pattern, we first observe that the two human coders produced extremely uniform alignments, seen in the high rate of overlap, and in the small boundary displacement across phonological contexts. Humans significantly outperformed all of the forced aligners in OR, and in BD for following phonological context. But for preceding phonological context, MFA and LaBB-CAT produced boundaries that were similar to human boundaries, a testament to the high quality of the alignment these programs produce.

Among the forced aligners, MFA and LaBB-CAT yielded the most accurate alignments, followed by FAVE, and finally, MAUS. What might account for these differences in performance? First, MFA is the only forced aligner that uses Kaldi, while the other three use HTK. The high performance of MFA may be an indication that Kaldi's use of more advanced ASR techniques results in more accurate segmentations. The high performance of LaBB-CAT may be attributable to its use of train/align, rather than pre-trained, models. Indeed, the fact that LaBB-CAT performs as well as MFA in terms of BD suggests that the train/align protocol may help close the gap in performance between HTK and Kaldi.

MAUS exhibits the poorest performance, despite being trained on Australian English, the variety being aligned. This is not due to the controlled nature of the training data used to construct its acoustic model, as performance did not significantly improve with an acoustic model based on (New Zealand) spontaneous speech. While it has been proposed that higher accuracy is obtained when the training data and the data to be processed are of a similar speech type (Cassidy and Schmidt 2017: 216), this finding supports the contrasting view that training on spontaneous speech yields the best alignment accuracy for both controlled and spontaneous speech (Fromont and Watson 2016: 426). And while MFA was trained on read speech from audio books, we would expect the nature of this genre to render data closer to spontaneous speech than to reading passages collected in more constrained settings.

We find no evidence to suggest that dictionary (G2P mapping) or variety greatly impact performance. MFA and FAVE are produced with U.S. English models and using U.S. English pronouncing dictionaries, but they outperform MAUS, again despite the latter being trained on Australian English. Thus, data type, the process by which acoustic models are generated, and toolkit would all appear to impact alignment accuracy above and beyond the variety that is being aligned (cf. also MacKenzie and Turton 2020). Future studies directly testing of the impact of variety may be able to confirm this prediction.

6 Conclusions

This study lends empirical support to the common wisdom that humans are far more consistent in creating alignments than are forced aligners, indicating that regardless of the aligner used, alignment accuracy will be enhanced by manual correction. Nevertheless, the accuracy differences identified here allow for some general recommendations for prioritizing the correction of vowels. First, with all programs, following contexts yield higher error rates than preceding contexts. For MFA and LaBB-CAT, vowel onset boundaries so closely mirror

those of humans that these should require less manual correction. Beyond that, a human coder appears to be less impacted by phonological environment than the forced aligners are, and thus correction priorities are program specific. With MFA, following laterals may require more checking, while for MAUS, preceding stops and following nasals and approximants warrant more attention. For FAVE and LaBB-CAT, on the other hand, there is little difference across manners of articulation, such that equal effort can be given to checking alignment accuracy.

The results of the comparisons presented here indicate that, assessed on the robustness of their forced alignment, MFA, LaBB-CAT and FAVE are more efficacious choices than MAUS for investigating vowels in English, with MFA and LaBB-CAT providing the highest quality alignments for English sociophonetic work.

Acknowledgements: We gratefully acknowledge support from the ARC Centre of Excellence for the Dynamics of Language, and funding from a Transdisciplinary & Innovation Grant (TIG952018). We thank Robert Fromont, Debbie Loakes, and the anonymous *Linguistics Vanguard* reviewers for valuable feedback on the paper, as well as Miriam Meyerhoff, Jim Stanford, and Hywel Stoakes for help in formulating the ideas presented here.

References

Baayan, Harald, R. Piepenbrock & Lennart Gulikers. 1995. *The CELEX Lexical Database (Release 2, CD-ROM)*. University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2010. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-33.

Brognaux, Sandrine, Sophie Roekhaut, Thomas Drugman, & Richard Beaufort. 2012. Automatic phone alignment: A comparison between speaker- independent models and models trained on the corpus to align. *Proceedings of the 8th International Conference on NLP, JapTAL*, 300–311. Kanazawa, Japan.

Carnegie Mellon University. 1993–2016 CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Cassidy, Steve & Thomas Schmidt. 2017. Tools for multimodal annotation. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 209–227. Netherlands: Springer.

Chodroff, Eleanor. 2018. Corpus Phonetics Tutorial. *ArXiv:* abs/1811.05553.

Cosi, Piero, Daniele Falavigna & Maurizio Omologo. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *EUROSPEECH-91*, 2nd European Conference on Speech Technology, 693–696.

Coto-Solano, Rolando & Sofía Flores Solórzano. 2017. Comparison of two forced alignment aystems for aligning Bribri speech.*CLEI Electronic Journal* 20(1). 2:1–2:13.

DiCanio, Christian, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith & Rey Castillo Garcia. 2012. Assessing agreement level between forced alignment models with data from endangered language documentation corpora. *INTERSPEECH-2012*, Portland Oregon, 130–133.

Fromont, Robert & Jennifer Hay. 2012. LaBB-CAT: An annotation store. *Proceedings of the Australasian Language Technology Workshop*, 113–117.

Fromont, Robert & Kevin Watson. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora* 11(3). 401–431.

Gaida, Christian, Patrick Lange, Rico Petrick, Patrick Proba, Malatawy Ahmed & David Suendermann-Oeft. 2014. Comparing open-source speech recognition toolkits. DHBW Stuttgart Technical Report, Project OASIS. (http://suendermann.com/su/pdf/oasis2014.pdf).

Gonzalez, Simon, Catherine E. Travis, James Grama, Danielle Barth & Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In Julien Epps, Joe Wolfe, John Smith & Caroline Jones (eds.), *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, 145–148.

Gordon, Elizabeth, Margaret Maclagan & Jennifer Hay. 2007. The ONZE corpus. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and digitizing language corpora*, 82–104. London: Palgrave Macmillan.

Horvath, Barbara. 1985. *Variation in Australian English: The sociolects of Sydney*. Cambridge: Cambridge University Press.

Jones, Caroline, Katherine Demuth, Weicong Li & Andre Almeida. 2017. Vowels in the Barunga variety of North Australian Kriol, *INTERSPEECH-2017*, Stockholm Sweden, 219–223.

Keegan, Peter J., Catherine I. Watson, Jeanette King, Margaret Maclagan, & Ray Harlow. 2012. The role of technology in measuring changes in the pronunciation of Māori over generations. In Tania Ka'ai, Muiris Ó Laoire, Nicholas Ostler, Rachael Ka'ai-Mahuta, Dean Mahuta & Tania Smith (eds.), *Language endangerment in the 21st Century: Globalisation, technology and new media, Proceedings of Conference FEL XVI*, 65–71. AUT University, Auckland, New Zealand: Te Ipukarea: The National Māori Language Institute, AUT University/Foundation for Endangered Languages.

Kisler, Thomas, Uwe D. Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26.

MacKenzie, Laurel & Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*. 6(s1).

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *INTERSPEECH-2017*, Stockholm Sweden, 498–502.

Panayotov, Vassil, Guoguo Chen, Daniel Povey & Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books, *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. Brisbane, Australia.

Paulo, Sérgio & Luís C. Oliveira. 2004. Automatic phonetic alignment and its confidence measures. In José Luis Vicedo, Particio Martínez-Barco, Rafael Munoz & Maximiliano Saiz Noeda (eds.), *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004*, 36–44. Berlin/Heidelberg: Springer-Verlag.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer & Karel Vesely. 2011. The Kaldi speech recognition toolkit. Paper presented at the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawaii.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Reddy, Sravana & James Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1(1). 15–28.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2 10.5281/zenodo.22281.

Schiel, Florian. 1999. Automatic phonetic transcription of non-prompted speech, *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)*, 607–610. San Francisco.

Schiel, Florian, Christoph Draxler, Angela Baumann, Tania Elbogen & Alexander Steen. 2012. The Production of Speech Corpora. Ms. (https://www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/).

Stoakes, Hywel & Florian Schiel. 2017. A Pan-Australian Model for MAUS. *Paper presented at the Australian Linguistic Society Annual Conference*, 4–7 December, University of Sydney.

Strunk, Jan, Florian Schiel & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14*, 3940–3947. Reykjavik, Iceland.

Stuart-Smith, Jane, Brian José, Tamara Rathcke, Rachel Macdonald & Eleanor Lawson. 2017. Changing sounds in a changing city: An acoustic phonetic investigation of real-time change over a century of Glaswegian. In Chris Montgomery & Emma Moore (eds.), *Language and a sense of place: Studies in language and region*, 38–64. Cambridge: Cambridge University Press.

Travis, Catherine E., James Grama & Simon Gonzalez. In Progress. *Sydney Speaks Corpora*. Australian Research Council Centre of Excellence for the Dynamics of Language, Australian National University: http://www.dynamicsoflanguage.edu.au/sydney-speaks/.

Wagner, Michael, Dat Tran, Roberto Togneri, Phil Rose, David Powers, Mark Onslow, Debbie Loakes, Trent Lewis, Takaaki Kuratate, Yuko Kinoshita, Nenagh Kemp, Shunichi Ishihara, John Ingram, John Hajek, David Grayden, Roland Göcke, Janet Fletcher, Dominique Estival, Julien Epps, Robert Dale, Anne Cutler, Felicity Cox, Girija Chetty, Steve Cassidy, Andy Butcher, Denis Burnham, Steven Bird, Cathi Best, Mohammed Bennamoun, Joanne Arciuli & Eliathamby Ambikairajah. 2010. *The big Australian speech corpus (The Big ASC) Paper presented at the 13th Australasian International Conference on Speech Science and Technology*. Melbourne, Australia.

Walker, James & Miriam Meyerhoff. 2020. Pivots of the Caribbean? Low-back vowels in Eastern Caribbean English. *Linguistics*.

Watson, Catherine I. & Zoe E. Evans. 2016. Sound change of experimental artifact?: A study on the impact of data preparation on measuring sound change. In Christopher Carignan and Michael D. Tyler (Eds.), *Proceedings of the 16th Australasian International Conference on Speech Science and Technology*, 261–264. Sydney, Australia.

Wilbanks, Erik. 2018. faseAlign (Version 1.1.9) [Computer software]. Retrieved Oct 11, 2018 from https://github.com/EricWilbanks/faseAlign.

Young, Steve, Gunnar Evermann, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Julian Odell, Dave Ollason, Daniel Povey, Valtcvho Valtchev & Phil Woodland. 2006. *The HTK Book (For Version 3.4)*. Cambridge: Cambridge University Engineering Department.

Yuan, Jiahong & Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America* 123. 5687–5690.

Appendix

Table 2: Linear mixed-effects model fit to Overlap Rate – Predictors: aligner identity, scaled vowel duration; Random intercepts: speaker, vowel identity, preceding and following manner of articulation.

Predictors	Estimates	CI	*t*	*p*
(Intercept) = H2H	0.77	0.71–0.82	27.08	**<0.001**
Aligner = FAVE	−0.23	−0.25 to −0.21	−21.02	**<0.001**
Aligner = LABBCAT	−0.22	−0.25 to −0.20	−20.71	**<0.001**
Aligner = MAUS	−0.26	−0.29 to −0.24	−24.45	**<0.001**
Aligner = MFA	−0.15	−0.17 to −0.13	−13.76	**<0.001**
Scale(dur)	0.04	0.04–0.05	10.81	**<0.001**

Bold values denote statistical significance at the $p < 0.05$ level.

Table 3: Linear mixed-effects model fit to Overlap Rate (with MAUS acoustic model from New Zealand English spontaneous speech) – Predictors: aligner identity, scaled vowel duration; Random intercepts: speaker, vowel identity, preceding and following manner of articulation.

Predictors	Estimates	CI	*t*	*p*
(Intercept) = H2H	0.76	0.70–0.82	24.76	**<0.001**
Aligner = FAVE	−0.23	−0.25 to −0.21	−20.55	**<0.001**
Aligner = LABBCAT	−0.22	−0.25 to −0.20	−20.25	**<0.001**
Aligner = MAUS	−0.26	−0.29 to −0.24	−23.91	**<0.001**
Aligner = MAUSNZ	−0.25	−0.28 to −0.23	−22.89	**<0.001**
Aligner = MFA	−0.15	−0.17 to −0.13	−13.45	**<0.001**
Scale(dur)	0.04	0.03–0.05	10.88	**<0.001**

Bold values denote statistical significance at the $p < 0.05$ level.

Table 4: Linear mixed-effects model fit to Boundary Displacement – Predictors: aligner identity, scaled vowel duration, position; Random intercepts: speaker, vowel identity.

Predictors	Estimates	CI	*t*	*p*
(Intercept) = H2H, preceding	8.92	2.46–15.38	2.71	**0.007**
Aligner = FAVE	31.24	24.53–37.95	9.12	**<0.001**
Aligner = LABBCAT	13.46	6.75–20.17	3.93	**<0.001**
Aligner = MAUS	60.44	53.73–67.16	17.65	**<0.001**
Aligner = MFA	17.36	10.65–24.07	5.07	**<0.001**
Position = following	4.11	−0.13–8.36	1.90	0.058
Scale(dur)	12.51	10.21–14.81	10.65	**<0.001**

Bold values denote statistical significance at the $p < 0.05$ level.

Table 5: Linear mixed-effects model fit to preceding Boundary Displacement – Predictors: aligner identity; scaled vowel duration, preceding manner of articulation; interaction between preceding manner and aligner; Random intercepts: speaker, vowel identity.

Predictors	Estimates	CI	*t*	*p*
(Intercept) = H2H, stops	7.40	−5.37–20.16	1.14	0.256
scale(dur)	9.03	5.72–12.34	5.34	**<0.001**
Manner = approximant	5.99	−12.41–24.39	0.64	0.523
Manner = fricative	−0.48	−19.24–18.28	−0.05	0.960
Manner = lateral	4.03	−28.45–36.51	0.24	0.808
Manner = nasal	1.64	−22.08–25.36	0.14	0.892
Aligner = FAVE	24.78	7.33–42.23	2.78	**0.005**
Aligner = LABBCAT	10.77	−6.68–28.22	1.21	0.226
Aligner = MAUS	69.02	51.57–86.47	7.75	**<0.001**
Aligner = MFA	7.44	−10.00–24.89	0.84	0.403

Table 5 (continued)

Predictors	Estimates	CI	*t*	*p*
Manner = approximant:Aligner = FAVE	9.18	−16.61–34.98	0.70	0.485
Manner = fricative:Aligner = FAVE	7.58	−18.84–34.00	0.56	0.574
Manner = lateral:Aligner = FAVE	10.17	−35.60–55.95	0.44	0.663
Manner = nasal:Aligner = FAVE	18.68	−14.79–52.14	1.09	0.274
Manner = approximant:Aligner = LABBCAT	−2.24	−28.03–23.56	−0.17	0.865
Manner = fricative:Aligner = LABBCAT	1.09	−25.33–27.51	0.08	0.935
Manner = lateral:Aligner = LABBCAT	1.95	−43.82–47.72	0.08	0.933
Manner = nasal:Aligner = LABBCAT	6.50	−26.96–39.97	0.38	0.703
Manner = approximant:Aligner = MAUS	−11.12	−36.92–14.67	−0.85	0.398
Manner = fricative:Aligner = MAUS	−28.83	−55.25 to −2.41	−2.14	**0.032**
Manner = lateral:Aligner = MAUS	−27.12	−72.90–18.65	−1.16	0.245
Manner = nasal:Aligner = MAUS	2.52	−30.95–35.98	0.15	0.883
Manner = approximant:Aligner = MFA	12.79	−13.00–38.59	0.97	0.331
Manner = fricative:Aligner = MFA	10.06	−16.36–36.48	0.75	0.455
Manner = lateral:Aligner = MFA	14.92	−30.85–60.69	0.64	0.523
Manner = nasal:Aligner = MFA	15.05	−18.41–48.52	0.88	0.378

Bold values denote statistical significance at the $p < 0.05$ level.

Table 6: Linear mixed-effects model fit to following Boundary Displacement – Predictors: aligner identity, scaled vowel duration, following manner of articulation; interaction between following manner and aligner; Random intercepts: speaker, vowel identity.

Predictors	Estimates	CI	*t*	*p*
(Intercept) = H2H, stops	9.11	−2.77–21.00	1.50	0.133
scale(dur)	11.42	8.22–14.62	7.00	**<0.001**
Manner = approximant	5.77	−22.80–34.35	0.40	0.692
Manner = fricative	−0.87	−19.21–17.47	−0.09	0.926
Manner = lateral	5.56	−21.91–33.03	0.40	0.691
Manner = nasal	5.73	−11.60–23.06	0.65	0.517
Aligner = FAVE	27.03	10.58–43.48	3.22	**0.001**
Aligner = LABBCAT	15.43	−1.02–31.89	1.84	0.066
Aligner = MAUS	37.28	20.83–53.74	4.44	**<0.001**
Aligner = MFA	15.22	−1.24–31.67	1.81	0.070
Manner = approximant:Aligner = FAVE	5.93	−34.32–46.18	0.29	0.773
Manner = fricative:Aligner = FAVE	7.11	−18.75–32.97	0.54	0.590
Manner = lateral:Aligner = FAVE	2.39	−36.17–40.96	0.12	0.903
Manner = nasal:Aligner = FAVE	4.28	−20.10–28.65	0.34	0.731
Manner = approximant:Aligner = LABBCAT	0.95	−39.30–41.20	0.05	0.963
Manner = fricative:Aligner = LABBCAT	0.74	−25.12–26.59	0.06	0.956
Manner = lateral:Aligner = LABBCAT	11.32	−27.24–49.88	0.58	0.565
Manner = nasal:Aligner = LABBCAT	−1.13	−25.51–23.25	−0.09	0.928
Manner = approximant:Aligner = MAUS	48.88	8.63–89.13	2.38	**0.017**
Manner = fricative:Aligner = MAUS	21.49	−4.37–47.35	1.63	0.103
Manner = lateral:Aligner = MAUS	5.58	−32.98–44.14	0.28	0.777
Manner = nasal:Aligner = MAUS	49.54	25.16–73.91	3.98	**<0.001**
Manner = approximant:Aligner = MFA	−0.71	−40.96–39.54	−0.03	0.972
Manner = fricative:Aligner = MFA	−2.27	−28.13–23.59	−0.17	0.864
Manner = lateral:Aligner = MFA	23.19	−15.37–61.75	1.18	0.239
Manner = nasal:Aligner = MFA	1.94	−22.44–26.31	0.16	0.876

Bold values denote statistical significance at the $p < 0.05$ level.