Laurel MacKenzie* and Danielle Turton

Assessing the accuracy of existing forced alignment software on varieties of British English

https://doi.org/10.1515/lingvan-2018-0061
Received February 14, 2019; accepted July 15, 2019

Abstract: This paper presents an analysis of the performance and usability of automatic speech processing tools on six different varieties of English spoken in the British Isles. The tools used in the present study were developed for use with Mainstream American English, but we demonstrate that their forced alignment functionality nonetheless performs extremely well on a range of British varieties, encompassing both careful and casual speech. Where phone boundary placement is concerned, substantial errors in alignment occur infrequently, and although small displacements between aligner-placed and human-placed phone boundaries are found regularly, these will rarely have a significant effect on measurements of interest for the researcher. We demonstrate that gross phone boundary placement errors, when they do arise, are particularly likely to be introduced in fast speech or with varieties that are radically different from Mainstream American English (e.g. Scots). We also observe occasional problems with phonetic transcription. Overall, we advise that, although forced alignment software is highly reliable and improving continuously, human confirmation is needed to correct errors which can displace entire stretches of speech. Nevertheless, sociolinguists can be assured that the output of these tools is generally highly accurate for a wide range of varieties.

Keywords: forced alignment; British English varieties; computational automatic speech recognition tools; sociolinguistics; sociophonetics; dialectology

1 Introduction

Tools for forced alignment of speech data are gaining currency in sociolinguistic research for the analysis of large-scale datasets (e.g. Labov et al. 2013). These tools convert an orthographic transcription into phones, then time-align words and phones to the speech signal. However, this process has generally been trained on a particular language variety, and moreover it relies on a standard dictionary of reference pronunciations. The extent to which these tools may be used effectively with data from dialects that differ phonologically from those reference points has not been systematically investigated, despite the fact that this would be beneficial for many researchers in sociolinguistics. In this paper, we investigate whether these tools perform to an acceptable standard on data from dialects that they were not built to work on. Our goal is to determine whether researchers who work with sociophonetic data that diverges markedly from the standard need bespoke alignment software tailored to the dialect they study, or whether existing software can be used without compromising performance.

In the following sections, we argue that there is a need among researchers and students of sociolinguistics for off-the-shelf forced alignment systems that can perform robustly on a range of different language varieties without training or extensive manipulation. We note that no published study has yet investigated the potential of existing forced alignment systems to fit a range of distinct varieties in this way. We then evaluate the performance of two different forced alignment tools which are well-known and widely used in

***Corresponding author: Laurel MacKenzie,** New York University, New York, NY, USA, E-mail: laurel.mackenzie@nyu.edu.
https://orcid.org/0000-0001-6427-4762
Danielle Turton: Lancaster University, Lancaster, UK, E-mail: d.m.turton@lancaster.ac.uk

sociolinguistic research: the Forced Alignment and Vowel Extraction suite, or FAVE (Rosenfelder et al. 2014), and the Dartmouth Linguistic Automation program, or DARLA (Reddy and Stanford 2015a, Reddy and Stanford 2015b). We test the aligners on six dialects of British English, selected to represent a range of regional and social varieties. We compare the placement of phone boundaries by each aligner for each dialect with that of a human annotator, assessing aligner accuracy in terms of agreement between aligner-placed and human-placed boundaries. We also weigh each aligner's accuracy against its technical requirements and ease of use. We conclude by providing recommendations for researchers in sociolinguistics who wish to use these tools. Overall, we advise that, although forced alignment software is highly reliable and is improving continuously, human confirmation is needed to correct occasional errors which can displace entire stretches of speech. Nevertheless, sociolinguists can be assured that the output of these tools is generally highly accurate for a wide range of varieties.

2 Background

2.1 What is forced alignment?

As detailed above, forced alignment uses automatic speech recognition to convert an orthographic transcript to a time-aligned phonemic transcription. Generally, in order to conduct forced alignment, the researcher begins with an audio file alongside a time-aligned orthographic transcript of the speech for analysis. Behind the scenes, the aligner software matches each orthographic word in the transcript with a phonemic representation in a pronouncing dictionary. Each phone in the pronouncing dictionary, and its surrounding phonetic environment, is matched with its ideal acoustic realization, which is based on a statistical model trained on a range of possible phonetic realizations. This may use pre-trained models (e.g. FAVE and DARLA) or be based on the speech input itself (e.g. LaBB-CAT; Fromont and Hay 2012; Fromont and Watson 2016). The final output is formed by the alignment of each part of the speech signal with the most likely acoustic model given the transcription. This process is visualized in Figure 1, reproduced from Bailey (2016: 12). For a more detailed overview of the inner workings of the acoustic matching process, see Evanini (2009: 52).

The typical output of these systems is a fully time-aligned Praat TextGrid (Boersma and Weenink 2017), where each word and phone identified in the audio is paired with its transcription. This then allows for the efficient and automated search and identification of segments and words of interest to the analyst by serving as input to scripts for coding of acoustic properties, such as the Praat handCoder (Fruehwald 2011), which plays each sound of interest in sequence for the researcher to code auditorily. Information about the phonological environment of a word or sound of interest, e.g. preceding/following segment, position in word, duration of sound, can be taken automatically, vastly increasing the processing speed and eliminating human error.

2.2 Customizing forced alignment

With the right technical knowhow (that is, a fairly proficient level of programming experience) and enough audio data, a researcher can train their own acoustic models on any language or language variety. This,

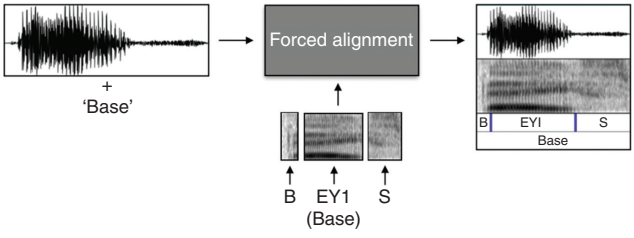

Figure 1: Visualization of the forced-alignment process of the word *base*, from Bailey 2016: 12.

combined with a pronouncing dictionary, is the basis that any researcher needs for a forced alignment system. This approach has allowed researchers to force align languages including French (Milne 2011), Spanish (Wilbanks 2015), Dutch (Schuppler et al. 2011), Mandarin Chinese (Lai et al. 2010), and Matukar Panau (González et al. 2018a).

In principle, researchers wishing to force align varieties of English which existing English-based aligners were not expressly designed for could do the same, i.e. train an acoustic model and construct a pronouncing dictionary that reflects the variety under study. Typically, however, the technical knowledge required to do this presents a barrier for many researchers (although recent advancements in the field are making this easier; see McAuliffe et al. 2017). Moreover, it effectively eliminates the possibility that these models could be used in teaching, with students who are technical neophytes. By contrast, off-the-shelf aligners are student-friendly, and have been used in many undergraduate-level classes and independent research projects (e.g. Warburton 2016; Lee 2017).

In this paper, we thus explore how well existing English-based forced aligners, with no or only very few simple modifications, fare on English dialects that they have not been trained on.

2.3 Testing the accuracy of forced alignment

A growing body of work compares the placement of phone boundaries by different forced alignment systems to the placement of those same boundaries by humans. Because these studies assess the performance of aligners on varieties that they were trained on, they constitute a baseline for the present study. The findings of these studies are summarized in Table 1. We follow customary practice in reporting the percentage of aligned phone boundaries which fall within a certain threshold of the human's placement.

Because the advantage of forced alignment is that it can relieve a human annotator of a tedious burden, it is also instructive to compare these performance measures to human–human agreement on the same task. If aligner–human agreement is comparable to human–human agreement, then there is no disadvantage to using an aligner over having a human hand-place phone boundaries. The last column of Table 1 provides (where available) human–human agreement statistics on the same speech for which human–aligner agreement was assessed.

Human–aligner agreement is generally high. Moreover, when human–human performance is compared to human–aligner performance, humans agree somewhat more, but not by a large margin. This has led many researchers to conclude that forced alignment is an acceptable substitution for human phone annotation, at least when aligners are used on the dialects they have been trained on. Few published studies look specifically at aligner performance on varieties or languages that are unfamiliar to the aligner, as we do here (although

Table 1: Summary of human–aligner and human–human phone boundary placement agreement from published studies.

Reference	Language	Speech style	Threshold	Human–aligner agreement rate	Human–human agreement rate
Hosom (2009)	English	Read	20 ms	80–93%	93%
Goldman (2011)	English	Spontaneous	20 ms	76%[a]	79%
			10 ms	51%	62%
McAuliffe et al. (2017), Raymond et al. (2002)[b]	English	Spontaneous (Buckeye corpus)	25/20 ms[c]	77%	79%
			10 ms	41%	62%
Goldman (2011)	French	Spontaneous	20 ms	81%	81%
			10 ms	51%	57%
Cosi et al. (1991)	Italian	Spontaneous	20 ms	56–64%	
Wilbanks (2015)	Spanish	Spontaneous	20 ms	69%	
			10 ms	45%	

[a]Goldman compares the performance of one aligner to two humans; we average those two human–aligner agreement rates.
[b]The human–aligner statistics are from McAuliffe et al.; the human–human statistics are from Raymond et al.
[c]25 ms is the human–aligner threshold; 20 ms is the human–human threshold.

see Meer and Matute Flores 2018 on Trinidadian English, González et al. 2018b on Australian English, and DiCanio et al. 2015 on Mixtec).[1] However, Knowles et al. (2015) evaluate the performance of forced alignment on a different "non-standard" population of language users: children. They find much poorer results than have previously been found for adult language data: for spontaneous child speech, only 15% of aligner-placed phone boundaries are within 10 ms of a human's placement, and 46% are within 25 ms.

3 The present study

3.1 The aligners

Our study compares two aligners: Forced Alignment and Vowel Extraction suite (FAVE; Rosenfelder et al. 2014) and Dartmouth Linguistic Automation (DARLA; Reddy and Stanford 2015a, Reddy and Stanford 2015b).

FAVE is an adaptation of the Penn Phonetics Lab Forced Aligner (P2FA; Yuan and Liberman 2008) and uses the HTK speech recognition toolkit (Woodland et al. 1995). FAVE's acoustic models are based on 9000 hours of speech by US Supreme Court justices; accordingly, its models are based on what can be called Mainstream American English (Wolfram and Schilling 2015), with 15 unique vowel phonemes. FAVE was developed specifically for sociolinguistic use, and thus has a number of advantages for researchers dealing with interview data and other types of spontaneous speech: it can align multiple talkers and overlapping speech, and can allow for background noise.

FAVE uses the CMU Pronouncing Dictionary v.0.7 (Weide 2008) for its orthographic–phonemic matching. The aligner must be installed locally; instructions and a user manual are provided on GitHub.[2] This local installation allows some customizability: namely, the user can substitute their own pronouncing dictionary (provided that it conforms to FAVE's acoustic models; see Section 5) or add additional words to the built-in dictionary (with the same caveat). FAVE requires the user to provide orthographic transcriptions at the sentence or breath group level and returns TextGrids suitable for use with the sound analysis program Praat.

DARLA is "a system [...] that automates the entire pipeline from transcribing audio to alignment and formant extraction" (Reddy and Stanford 2015a: 17). DARLA provides users with the option to furnish their own orthographic sentence-level transcription of their sound file ("semi-automated alignment"), or users can instead have their file undergo full automatic speech recognition ("completely automated alignment"). DARLA's forced alignment system is the Montreal Forced Aligner (MFA; McAuliffe et al. 2017), which involves several training passes of the data (including monophone, triphone, and speaker-specific). DARLA also uses the CMU Pronouncing Dictionary for phone-level transcriptions. Unlike FAVE, it does not allow the user to customize the pronouncing dictionary; instead, it provides its own best-guess transcription of unknown words using the Sequitur grapheme-to-phoneme converter (Bisani and Ney 2008). Note that local installation of MFA does allow this level of customizability.

Although both FAVE and DARLA provide the ability to automatically measure vowel formants, in this paper, we set this technology to the side and focus only on phone-level transcription and alignment. For details of the accuracy of automatic vowel extraction, see Evanini (2009) and Labov et al. (2013: 35–38).

We selected these two aligners because each has unique advantages for sociolinguists. FAVE easily enjoys the most currency among sociolinguists at present: for instance, of the nineteen abstracts presented at the 2017 New Ways of Analyzing Variation conference that mention using forced alignment on English data, twelve used FAVE, while only four used DARLA. (The remaining three did not specify which system they used). As mentioned earlier, FAVE allows for multiple speakers and speaker overlap, but it requires local installation which can be technically demanding. DARLA, by contrast, does not (as of this writing) accommodate multiple speakers or speaker overlap, but it has a user-friendly online interface which lends itself well

1 There is also Kisler et al. (2017: 333), who reference the existence of an "internal project report" in which an aligner trained on Standard Southern British English is tested on Scots English. The aligner is found to have "an error rate twice that of human experts."

2 https://github.com/JoFrhwld/FAVE/.

to teaching. In selecting these two aligners, we hope to assess whether these trade-offs come with trade-offs in performance, as well.

3.2 The varieties

We selected six varieties of British English for this study. These are enumerated in Table 2, which also contains information on the speaker selected to represent each variety and a source that provides an overview of the variety's phonological system.

These six varieties comprise five non-standard regional varieties (Manchester, Blackburn, Sunderland, Hastings, and Westray, all plotted on the map in Figure 2) plus Received Pronunciation, the English standard. All varieties are represented in the study with data from a sociolinguistic interview (Labov 1984) or similar,

Table 2: Overview of varieties and speakers used in the study.

Variety	Speaker demographics	Overview of phonological system
Received Pronunciation	Male, 70s, recorded in 2000 (interview) and 2006 (reading)	Wells (1982)
Manchester	Female, 20s, interviewed in 2012	Baranowski and Turton (2015)
Blackburn	Female, 80s, interviewed in 2015	Turton (2015)
Sunderland	Female, 20s, interviewed in 2012	Burbano-Elizondo (2008, 2015)
Westray	Female, 80s, interviewed in 2007	Tamminga (2009)
Hastings	Female, 79, interviewed in 2011	Holmes-Elliott (2015)

Figure 2: The locations of the varieties under study. Received Pronunciation is non-regional.

Table 3: Word count and duration of analyzed excerpt, and total number of hand-corrected phone boundaries per excerpt–aligner pair.

	Word count	Duration (sec.)	Number of phone boundaries	
			DARLA	FAVE
RP – spontaneous	334	89	1045	1038
RP – reading	277	147	1086	1125
Manchester	284	85	912	944
Blackburn	317	103	942	1097
Sunderland	308	88	1008	1056
Westray	261	131	931	997
Hastings	340	88	1022	1049

though we also include a single passage of read speech from the Received Pronunciation speaker as a control case, in order to assess performance on speech which is slow and clear. In total, then, we evaluate the performance of our two chosen aligners across seven different recordings.

We selected the varieties that we did in order to represent a diversity of locations and phonological features. We were additionally interested to see how the aligners would cope with widespread British English phonological variants involving the deletion or substitution of consonants (Hughes et al. 2012; MacKenzie 2017): specifically, what effect these missing or lenited sounds would have on the ability of the aligner to correctly choose the boundary for surrounding sounds, and whether they would introduce a significant problematic effect on the overall alignment.

3.3 Methodology

All audio files were orthographically transcribed in ELAN (Wittenburg et al. 2006) by the authors or trained research assistants. Transcription followed the protocol suggested for use with each aligner.

Audio files and their accompanying transcripts were then run through FAVE and DARLA. Each aligner's built-in pronouncing dictionary was used for alignment in all cases; no supplementary dictionary files were used, though out-of-dictionary words were manually transcribed by the authors for FAVE alignment. (DARLA does not provide the user with the opportunity to manually transcribe out-of-dictionary words.)

After alignment, an excerpt from each sound file of approximately 300 words (precise counts in Table 3) was selected for hand-correction. To the extent possible, in order to capture the most vernacular speech, these excerpts were chosen to contain narratives of personal experience (Labov 2006[1966]). Then, the word and segment boundaries produced by FAVE and by DARLA were hand-corrected by the authors – both trained sociophoneticians with considerable experience with several of the varieties under study – in Praat (Boersma and Weenink 2017). Figure 3 provides an example. These human-placed annotations were then compared to the aligners' placement of the same boundaries to measure amount of human–aligner displacement for the onset and offset of every phone. The results of this analysis are presented in Section 4.

As is apparent from Figure 3, in the output of forced alignment, the onset of every phone is the offset of the previous phone, and the offset of every phone is the onset of the next phone (with the exceptions, of course, of the very first and very last phones in a given audio file). It is thus redundant to provide overall statistics on both phone onset displacement and phone offset displacement. We report phone onset displacements in this paper unless otherwise indicated.

Phones were excluded from analysis when the aligner chose a pronunciation variant that was simply not alignable with the audio. There were two words for which we observed this to occur: the preposition *for*, which has a possible (presumably erroneous) transcription "F R ER0" in the CMU Pronouncing Dictionary, and the pronoun *us*, for which the aligner occasionally selected the (unintended) pronunciation "Y UW1 EH1 S" (i.e., *U.S.*). In such cases, all phones associated with the erroneously transcribed word were omitted from analysis.[3]

3 This was rare and resulted in the omission of only 19 phones in total.

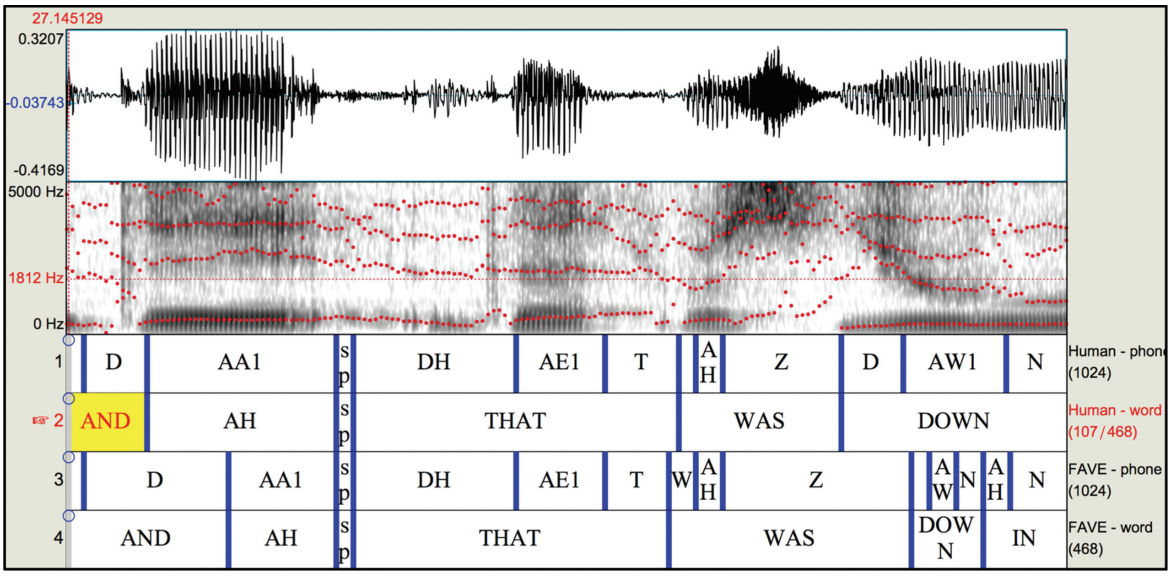

Figure 3: Hand-correction of FAVE output. The bottom two tiers are FAVE output; the top two tiers are manual correction of that output. Variety: Westray.

4 Results

In this section, we present quantitative results on phone boundary placement (Section 4.1), assess aligner performance (Section 4.2), and discuss alignment problems (Section 4.3). The word count and duration of each analyzed excerpt is provided in Table 3. Also provided is the total number of phone boundaries that were hand-corrected for each excerpt–aligner pair. This number differs by aligner even within an excerpt because aligners, given identical word-level transcripts, may nonetheless choose different pronunciations for a given word if multiple possible entries exist in the dictionary.

As is evident from comparing the word count and duration columns on each row, speaking rates vary widely in this sample, from 3.85 words/sec (Hastings) to 1.88 words/sec (RP – reading). We address the implications of this for aligner performance in Section 4.3.

4.1 Quantitative results

Following, e.g., Goldman (2011), Hosom (2009), and Wilbanks (2015), we provide summary statistics on human–aligner phone onset boundary displacement, in Table 4. The raw data that these summary statistics are based on is plotted in Figure 4. Figure 5 shows the percentage of aligner-placed phone onset boundaries that were within 10 ms and 20 ms of the human's placement.

Table 4: Phone onset boundary displacement for two forced alignment methods. All values in milliseconds.

	DARLA			FAVE		
	mean	median	s.d.	mean	median	s.d.
RP – spontaneous	12.9	0	50.7	17.6	0	64.1
RP – reading	6.4	0	16.5	8	0	19
Manchester	10.9	0	26.7	11.7	0	32.8
Blackburn	9.7	0	28.8	7.5	0	27.3
Sunderland	17.5	0	61.5	16.8	0	32.8
Westray	14.6	0	42.3	19.5	0	46.8
Hastings	15.2	0	57.7	19.6	0	64.5

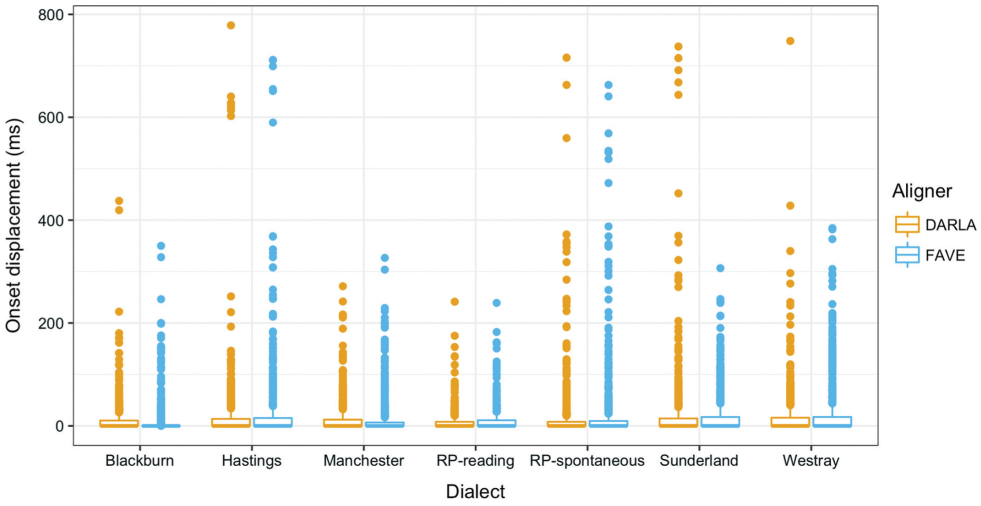

Figure 4: Phone onset boundary displacement for two forced alignment methods.

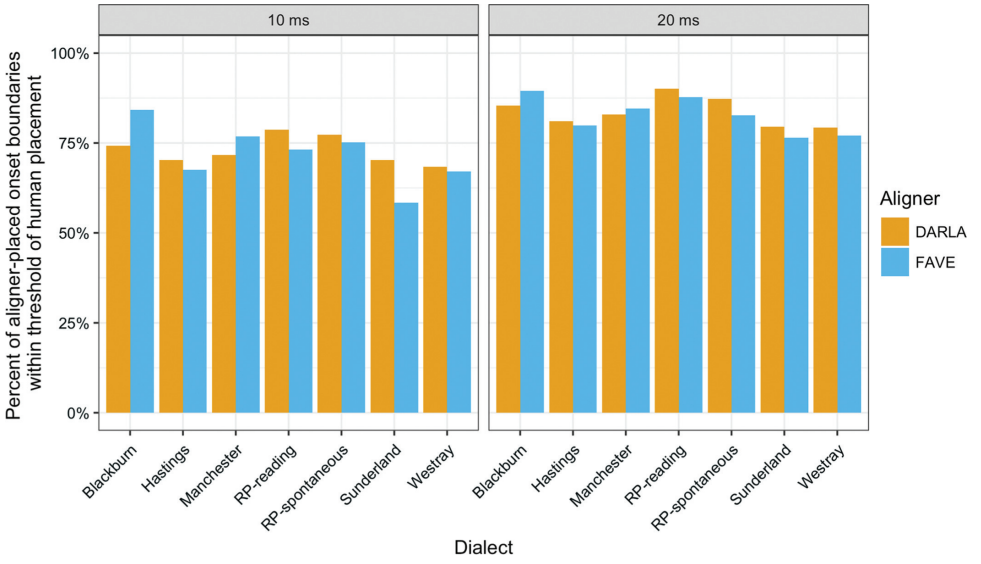

Figure 5: Phone onset boundary displacement for two forced alignment methods.

4.2 Performance assessment

First, we compare the results in Table 4 to the "baseline" results of previous research, in which aligners were tested on the same variety they had been trained on. McAuliffe et al. (2017) evaluated the performance of both MFA (the aligner that DARLA serves as a wrapper for) and FAVE on the Buckeye Corpus. They found an average human–MFA phone boundary displacement of 17 ms. DARLA's performance in our study improves on this for five of the studied dialects, and effectively matches for the sixth (Sunderland). Where FAVE is concerned, McAuliffe et al. found an average human–FAVE phone boundary displacement of 19.3 ms. FAVE's performance here actually improves on this for four of our studied dialects (RP, Manchester, Blackburn, Sunderland) and effectively matches it for the remaining two (Westray and Hastings). Aligner–human agreement in the present study is thus excellent.

In fact, aligner–human agreement in this study often exceeds human–human agreement as found in previous work. Raymond et al. (2002) found a mean boundary placement difference between humans of 17 ms (again using the Buckeye Corpus). Both aligners' performance improves on this in nearly every case. This

should serve as reassurance for researchers who wish to use these aligners on English data that differs phonologically from the data they have been trained on: on the whole, phone boundary placement is nevertheless just as good as what a human would achieve.

When we compare the aligners' absolute performance to each other, DARLA performs significantly better than FAVE via a Wilcoxon rank sum test with continuity correction ($p < 0.001$). But in practical terms, the actual difference between the two is negligible: DARLA achieves an overall mean onset boundary displacement of 12.4 ms and FAVE, 14.3 ms.

Turning to Figure 5, we look at the aligners' performance relative to 10 ms and 20 ms thresholds in light of previous work. Though no previous research has used these particular accuracy metrics with the specific aligners we test here, Goldman (2011) investigates human–aligner agreement with the aligner EasyAlign, finding 51% of aligner-placed boundaries to be within 10 ms of the human's placement, and 76% to be within 20 ms. Both DARLA and FAVE meet or exceed EasyAlign's accuracy rate within both thresholds in all cases. We can also compare the aligners' performance on this metric to attested human–human agreement rates, and again, the aligners perform well. Raymond et al. (2002) found that 62% of phone boundaries placed by humans were within 10 ms of each other; DARLA and FAVE meet or exceed this in thirteen out of fourteen cases (with only FAVE + Sunderland coming up short, at 58%). They also found that 79% of phone boundaries placed by humans were within 20 ms of each other; DARLA and FAVE meet or exceed this in twelve out of fourteen cases. (Coming up short, by only two or three percentage points, are FAVE + Sunderland and FAVE + Westray.)

There is no significant difference between the aligners on either the 10 ms or the 20 ms threshold metric via a chi-square test ($p > 0.05$).

It is clear that both aligners have performed very well. The fact that they have been provided with phonological systems that differ – sometimes rather radically – from the systems they have been trained on has not hindered their performance.

4.3 Alignment problems

In the previous section, we established that aligner performance in terms of phone boundary placement relative to a human's is, on the whole, very good. Still, it is apparent from looking at Figures 4 and 5 that the aligners perform worse on some sound samples than others. Has anything systematic led to these differences?

One contender is speaking rate, which, as noted previously, varied widely across the excerpts selected for study. Table 5 reorders the excerpts from fastest to slowest speaking rate (calculated by dividing the number of words in the excerpt by the duration of the excerpt) and provides the mean displacement measures and 20 ms threshold accuracy measures from Table 4 and Figure 5.

Looking first at the mean boundary displacement measures, neither aligner shows a significant correlation between speaking rate and mean displacement via a Spearman's rank test ($p > 0.05$). But this may be because both aligners perform fairly poorly on Westray, despite the speaker's slow speaking rate.

Table 5: Speaking rate for each excerpt, arranged from fastest to slowest, accompanied by accuracy metrics reproduced from Table 4.

	Speaking rate (words/sec.)	Mean displacement, DARLA (ms.)	Mean displacement, FAVE (ms.)	% within 20 ms, DARLA	% within 20 ms, FAVE
Hastings	3.86	15.2	19.6	81%	80%
RP – spontaneous	3.75	12.9	17.6	87%	83%
Sunderland	3.5	17.5	16.8	80%	76%
Manchester	3.34	10.9	11.7	83%	85%
Blackburn	3.08	9.7	7.5	85%	90%
Westray	1.99	14.6	19.5	79%	77%
RP – reading	1.88	6.4	8	90%	88%

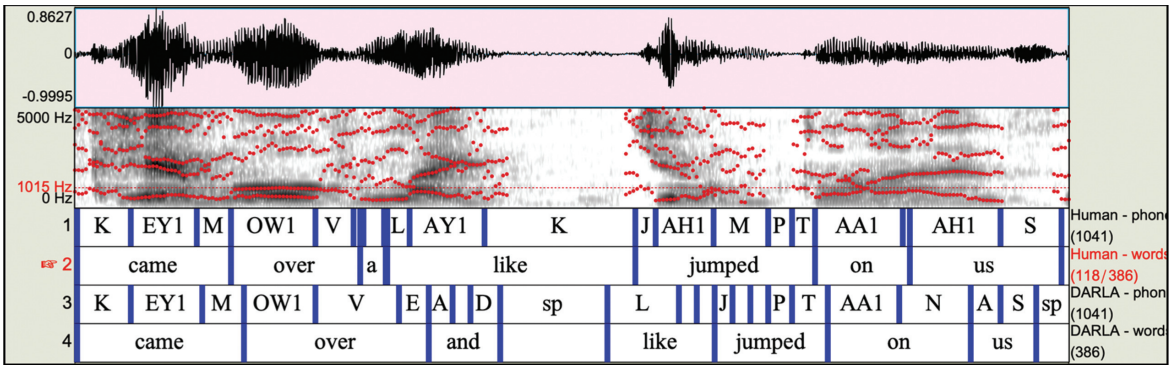

Figure 6: Extreme misalignment of the string *over and like jumped*, induced by massive reduction of *over and*, which the speaker pronounces [ovn̩]. Variety: Sunderland.

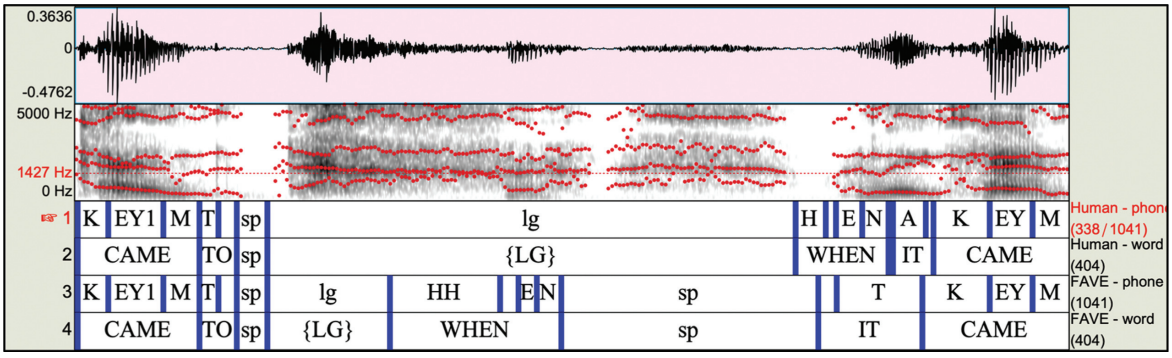

Figure 7: Extreme misalignment of the string *{LG} when*. {LG} marks a period of laughter. Variety: RP-spontaneous.

When we omit the outlier Westray excerpt, FAVE's accuracy does significantly correlate with speaking rate (rho $= 0.942$, $p = 0.017$). DARLA's does not (though $p = 0.058$).

FAVE's sensitivity to speaking rate has been shown before (Bailey 2016). But the Westray results show that slow speaking rate is no guarantee of a successful FAVE performance. The Westray variety's considerable phonological differences from Mainstream American English result in particularly poor FAVE performance even at a slow rate of speech. DARLA, by contrast, is less sensitive to speech rate, and comparatively less sensitive to radical phonological deviation from North American English.

The 20 ms threshold placement accuracy metric, which is not sensitive to outlying values like the mean is, does not correlate with speaking rate for either aligner ($p > 0.05$ with and without Westray). This implies that fast speech induces extreme cases of boundary displacement, rather than minor deviations.

With fast speech comes massive phonetic reduction (Johnson 2004), which is the likely source of speaking rate-induced poor performance on the part of the aligners. Figure 6 gives an example of reduction-induced misalignment. We also noticed that considerable misalignment could occur with laughter, which the aligners often misplaced (Figure 7; {LG} indicates laughter).

5 Discussion

Our analysis has shown impressive performances from both DARLA and FAVE, and we have full confidence in recommending that researchers who work on non-American and non-standard varieties of English use these tools for forced alignment. On the whole, human–aligner boundary placement disagreements are rare, and, as the examples in Figures 6 and 7 demonstrate, even when alignment is thrown off, it quickly recovers. This said, the aligners clearly perform worse on faster, reduced speech, and on the Westray variety of Scots. We would suggest that researchers take care in checking alignment output when working with varieties which

stray so far from the standard they were designed to work with. This warning also stands for rapid speech rates and massively reduced speech.

We have also found little meaningful difference between the aligners in terms of boundary placement accuracy. Given that the two aligners are essentially indistinguishable in terms of performance on this measure, in this section we discuss ways in which they do differ, so researchers can choose which one best suits their needs.

One major difference between the two is that FAVE allows the user to override its default pronouncing dictionary. This can be valuable for sociolinguists, who can modify the dictionary to incorporate possible variants for words that undergo widespread cases of sociolinguistic variation. For instance, *-ing* final words can be transcribed with both alveolar- and velar-final variants, and then the aligner can choose, for any given *-ing* word, which transcription best matches the audio. A number of researchers have found this to be a reliable way of automating the coding of sociolinguistic variation (Milne 2011; Schuppler et al. 2011; Yuan and Liberman 2011a, Yuan and Liberman 2011b; Milne 2012; Schuppler et al. 2012; Milne 2015; Bailey 2016).[4]

In terms of ease of use, the division is clear. DARLA has a user-friendly online interface which not only outputs an aligned transcript, but also provides raw and normalized formants, and a vowel plot. It is ideal for student projects, relies on simpler transcription conventions than FAVE, and its guess-the-pronunciation feature of unknown words is particularly useful for British spellings or non-standard lexical items, the transcription of which can be time-consuming in FAVE.

FAVE, on the other hand, has many benefits to sociolinguists, such as a built-in ability to handle multiple talkers and annotations of contextual style. When installed locally, FAVE is customizable, and allows for transparent error tracking and troubleshooting. That said, FAVE's local installation is non-trivial, and may be off-putting for researchers who are not comfortable with command line tools or the more technical aspects of dealing with FAVE's multiple requirements. This also makes FAVE near-impossible to use in most undergraduate teaching.

6 Conclusion

This study investigated the performance of two off-the-shelf forced alignment systems, which were developed for use with Mainstream American English, on a diverse set of social and regional varieties of British English. Our results, based on comparing aligners' phone boundary placement to humans', indicate that researchers working with such varieties can rest assured that both DARLA and FAVE will generally do an excellent job of aligning their sound files. As ever, researchers need to check the results of such automated tools: errors are particularly likely to be introduced in fast speech or with varieties that are radically different from Mainstream American English (e.g. Scots), and there are occasional problems with phonetic transcription. But, overall, these problems are outweighed by the elimination of tedious manual analysis and the vast speeding-up of the research process, resulting in more ambitious projects and much larger datasets.

One of our original motivations for carrying out this study was to test whether the aligners' phone boundary placement accuracy would be affected by processes of British English consonantal variation, such as *h*-dropping and post-vocalic *r*-vocalization, and to assess whether a user would be better off with a pronouncing dictionary which incorporates these variants (MacKenzie and Turton 2013). While it is true that the aligner will be looking for these sounds, which may not always be present, unless such variation is near-categorical and causing persistent alignment issues (which we did not find it to do in this study), it is probably not worth users modifying a pronouncing dictionary to incorporate these variants.

That said, one of the biggest issues that does arise when using American aligners on British data – not in terms of phone boundary placement, but in terms of transcription – is the differing number of phonemic

4 In principle, researchers could even replace FAVE's dictionary entirely, with, say, a British English one, such as BEEP (Robinson 1997). In practice, however, phonemic mismatches between British and American English make this nearly impossible: Standard British English has one more unique low back vowel than General American English does, and FAVE cannot align vowels it has no acoustic models for, i.e., that do not exist in American English.

distinctions in the vocalic domain. Aligners cannot make phonemic distinctions that they were not trained on, which means that British English speech will be phonetically transcribed by an American English aligner to have vocalic "mergers" that are not actually present. For now, we advise that the user goes ahead with the American model and corrects the transcription of vowel categories post-hoc at the analysis stage. This seems to be the most efficient method until these aligners can be made to handle more phonemic categories.

In terms of the performance of specific aligners, DARLA performs slightly better than FAVE where accuracy is concerned, and also has the upper hand in terms of ease-of-use, accessibility, and teaching purposes. FAVE performs slightly better than DARLA where customizability and the specific needs of sociolinguistic research are concerned, though it struggles slightly more than DARLA on rapid speech.

Overall, if these aligners are used in the manner for which they were designed – as tools, and not as the complete replacement of a dedicated researcher – they have highly positive consequences in terms of improving the speed, accuracy, and efficiency of the data analysis process.

Acknowledgments: Many thanks to Sophie Holmes-Elliott and Meredith Tamminga for sharing their data, Adam Mearns for access to DECTE, James Stanford for help with DARLA, audiences at NWAV 42 for helpful comments and questions, and the reviewers and editors of this special issue for suggestions which have improved the paper.

References

Bailey, George. 2016. Automatic detection of sociolinguistic variation using forced alignment. In Helen Jeoung (ed.), *Penn working papers in linguistics 22.2: Selected papers from NWAV 44*, 11–20. Philadelphia: Penn Graduate Linguistics Society.

Baranowski, Maciej & Danielle Turton. 2015. Manchester English. In Raymond Hickey (ed.), *Researching Northern English*, 283–305. Amsterdam & Philadelphia: John Benjamins.

Bisani, Maximilian & Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication* 50(5). 434–451.

Boersma, Paul & David Weenink. 2017. Praat: Doing phonetics by computer, version 6.0.29. http://www.fon.hum.uva.nl/praat/ (accessed 28 July 2017).

Burbano-Elizondo, Lourdes. 2008. *Language variation and identity in Sunderland*. Sheffield, UK: University of Sheffield dissertation.

Burbano-Elizondo, Lourdes. 2015. Sunderland. In Raymond Hickey (ed.), *Researching Northern English*, 183–204. Amsterdam & Philadelphia: John Benjamins.

Cosi, Piero, Falavigna, Daniele & Omologo, Maurizio, 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In *Proceedings of the Second European Conference on Speech Communication and Technology*, 693–696.

DiCanio, Christian, Hosung Nam, Jonathan D. Amith, Rey Castillo García & Douglas H. Whalen. 2015. Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. *Journal of Phonetics* 48. 45–59.

Evanini, Keelan. 2009. *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*. Philadelphia, PA: University of Pennsylvania dissertation.

Fromont, Robert & Jennifer Hay. 2012. LaBB-CAT: An annotation store. In Paul Cook & Scott Nowson (eds.),*Proceedings of the Australasian Language Technology Association Workshop* 2012, 113–117.

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

Fruehwald, Josef. 2011. handCoder [Praat script]. http://val-systems.blogspot.co.uk/2011/02/handcoder-praat-script.html (accessed 28 July 2017).

Goldman, Jean-Philippe. 2011. Easyalign: An automatic phonetic alignment tool under Praat. In *Proceedings of the 12th Conference of the International Speech Communication Association*, 3233–3236.

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

González, Simón, James Grama & Catherine Travis. 2018b. Comparing the accuracy of forced-aligners for sociolinguistic research. Poster presented at CoEDL Fest, University of Melbourne, 5–8 February.

Holmes-Elliott, Sophie. 2015. *London calling: Assessing the spread of metropolitan features in the southeast*. Glasgow, UK: University of Glasgow dissertation.

Hosom, John-Paul. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication* 51(4). 352–368.

Hughes, Arthur, Peter Trudgill & Dominic Watt. 2012. *English accents and dialects: An introduction to social and regional varieties of English in the British Isles*. London: Hodder Education.

Johnson, Keith. 2004. Massive reduction in conversational American English. In Kiyoko Yoneyama & Kikuo Maekawa (eds.), *Spontaneous speech: Data and analysis: Proceedings of the 1st Session of the 10th International Symposium*, 29–54. Tokyo: The National International Institute for Japanese Language.

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

Knowles, Thea, Meghan Clayards, Morgan Sonderegger, Michael Wagner, Aparna Nadig & Kristine H. Onishi. 2015. Automatic forced alignment on child speech: Directions for improvement. *Proceedings of Meetings on Acoustics* 25. 060001.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh and Joel Sherzer (eds.), *Language in use: Readings in sociolinguistics*, 28–66. Englewood Cliffs, N.J.: Prentice Hall.

Labov, William. 2006 [1966]. *The Social stratification of English in New York City*. New York: Cambridge University Press.

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

Lai, Catherine, Yanyan Sui & Jiahong Yuan. 2010. A corpus study of the prosody of polysyllabic words in Mandarin Chinese. In *Proceedings of Speech Prosody* 2010, 100457:1–4.

Lee, Sarah. 2017. Style-shifting in vlogging: An acoustic analysis of "YouTube Voice". *Lifespans & Styles: Undergraduate Working Papers on Intraspeaker Variation* 3(1). 28–39.

MacKenzie, Laurel. 2017. Frequency effects over the lifespan: A case study of Attenborough's r's. *Linguistics Vanguard* 3(1). 1–12.

MacKenzie, Laurel & Danielle Turton. 2013. Crossing the pond: Extending automatic alignment techniques to British English dialect data. Paper presented at New Ways of Analyzing Variation 42, Pittsburgh, 17–20 October.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*, 498–502.

Meer, Phillip & José Matute Flores. 2018. Making FAVE ready for New Englishes: Applying and modifying FAVE for semi-automatic acoustic analyses of Trinidadian English vowels. Paper presented at New Ways of Analyzing Variation 47, New York University, 18–21 October.

Milne, Peter M. 2011. Finding schwa: Comparing the results of an automatic aligner with human judgments when identifying schwa in a corpus of spoken French. *Canadian Acoustics* 39(3). 190–191.

Milne, Peter M. 2012. The effects of syllable position on allophonic variation in Québec French /r/. In Hilary Prichard (ed.), *Penn working papers in linguistics 18.2: Selected papers from NWAV 40*, 67–76. Philadelphia: Penn Graduate Linguistics Society.

Milne, Peter M. 2015. Improving the accuracy of forced alignment through model selection and dictionary restriction. Ms., McGill University.

Raymond, William D., Mark A. Pitt, Keith Johnson, Elizabeth Hume, Matthew J. Makashay, Robin Dautricourt & Craig Hilts. 2002. An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 1125–1128.

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

Reddy, Sravana & James N. Stanford. 2015b. A web application for automated dialect analysis. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 71– 75.

Robinson, A. J. 1997. British English Example Pronunciation (BEEP). URL http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, accessed 6 June 2017.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2014. FAVE 1.1.3. http://dx.doi.org/10.5281/zenodo.9846 (accessed 28 July 2017).

Schuppler, Barbara, Mirjam Ernestus, Odette Scharenborg & Lou Boves. 2011. Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics* 39(1). 96–109.

Schuppler, Barbara, Wim A. van Dommelen, Jacques Koreman & Mirjam Ernestus. 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.

Tamminga, Meredith. 2009. Insular Scots front vowels in Westray, Orkney. *Scottish Language* 28. 67–87.

Turton, Danielle. 2015. 4,000 /r/s from Blackburn, Lancashire: A (socio)phonological analysis of rhoticity in Northern England. Paper presented at the Manchester Forum in Linguistics, University of Manchester, 6–7 November.

Warburton, Jasmine. 2016. *Phonetic variation in the North East of England: On intra-regional differences in /u/, /ə/ and /l/ realizations*. Newcastle: Newcastle University MA thesis.

Weide, Robert. 2008. The CMU Pronouncing Dictionary. Carnegie Mellon University. http://dx.doi.org/10.5281/zenodo.9846 (accessed 28 July:2017).

Wells, J. C. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Wilbanks, Eric. 2015. The development of FASE: Forced Alignment System for Español and implications for sociolinguistic methodologies. Paper presented at New Ways of Analyzing Variation 44, University of Toronto, 22–25 October.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Wolfram, Walt & Natalie Schilling. 2015. *American English: Dialects and variation*. Malden, MA: John Wiley & Sons, third edition.

Woodland, Philip C., Chris J. Leggetter, J. J. Odell, Valtcho Valtchev & Steve J. Young. 1995. The 1994 HTK large vocabulary speech recognition system. In *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing*, 73–76.

Yuan, Jiahong & Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123. 3878.

Yuan, Jiahong & Mark Liberman. 2011a. Automatic detection of g-dropping in American English using forced alignment. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, 490–493.

Yuan, Jiahong & Mark Liberman. 2011b. /l/ variation in American English: A corpus approach. *Journal of Speech Sciences* 1(2). 35–46.