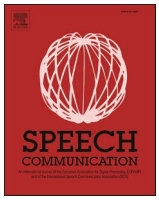
Speech Communication 158 (2024) 103042

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom



Analysis of forced aligner performance on L2 English speech

Samantha Williams *, Paul Foulkes, Vincent Hughes

Department of Language and Linguistic Science, University of York, York, UK

ARTICLE INFO

Keywords:
Automatic methods
L2 english
Forced alignment

ABSTRACT

There is growing interest in how speech technologies perform on L2 speech. Largely omitted from this discussion are tools used in the early data processing steps, such as forced aligners, that can introduce errors and biases. This study adds to the conversation and tests how well a model pre-trained for the alignment of L1 American English speech performs on L2 English speech. We test and discuss the impact of language variety, demographic factors, and segment type on the performance of the forced aligner. We also examine systematic errors encountered.

Forty-five speakers representing nine L2 varieties were selected from the Speech Accent Archive and force aligned using the Montreal Forced Aligner. The phoneme-level boundary placements were manually corrected in order to assess differences between the automatic and manual alignments. Results show marked variation in the performance across language groups and segment types for the two metrics used to assess accuracy: Onset Boundary Displacement, a distance metric between the automatic and manual boundary placements, and Overlap Rate, which indicates to what extent the automatically aligned segment overlaps with the manually aligned segment. The highest accuracy on both measures was obtained for German and French, and lowest accuracy for Russian. The aligner's performance on all varieties was comparable to that on conversational American English and non-standard varieties of English. Furthermore, the percentage of boundary placements within 10 and 20 ms of the corrected boundary was similar to that observed between transcribers. Apart from errors due to variety mismatch, most issues encountered in the alignment were due to issues not exclusive to L2 speech such as inaccurate orthographic transcriptions, hesitations, specific voice qualities, and background noise.

The results of this study can inform the use of automatic aligners on L2 English speech and provide a baseline of potential errors and information to help the development of more robust alignment tools for further development of automatic systems using L2 English.

1. Introduction

Forced aligners provide a semi-automatic method of aligning an acoustic signal with phoneme-level segmentation. Provided with an orthographic transcript they can greatly reduce the amount of manual work by giving an estimate of the time-alignments of words and segments. While forced aligners perform well on languages they were trained on (e.g., McAuliffe et al., 2017), the underlying language models tend to be based on monolingual speakers of standard varieties of majority languages such as English and French. Unless a researcher has adapted an existing model to a new variety or built a new model based on smaller datasets (outlined in McAuliffe, 2021b), many languages and varieties do not have models readily available. This means that the performance of forced aligners on those varieties may suffer. Previous studies on the performance of forced aligners have compared the relative performance of different aligners (e.g., Gonzalez et al., 2020), investigated the effects of various factors on their accuracy (e.g., Fromont and Watson, 2016), and tested variety mismatch of the acoustic model with non-standard varieties of English (e.g., regional British English varieties in Mackenzie and Turton, 2020). Surprisingly few studies, however, have considered the impact of L2 speech. Yet there are estimated to be well over twice as many L2 speakers of English as L1 speakers (1.08 billion L2 speakers in 146 countries compared to 373 million L1 speakers; Eberhard et al., 2022).

The degradation in the performance of speech technologies for speakers of L2 or even non-standard varieties has long been discussed in both research and media (e.g., Chan et al., 2022; Markl, 2022; Koenecke et al., 2020; Wu et al., 2020; Harwell et al., 2018). However, this topic is typically restricted to the user-facing aspects of speech technology such as automatic speech recognition (ASR) or text-to-speech. By comparison, the methodological tools used behind the scenes, such as forced alignment, are largely omitted from the discussion. This is perhaps due to

* Corresponding author.
E-mail address: samantha.williams@york.ac.uk (S. Williams).

https://doi.org/10.1016/j.specom.2024.103042
Received 28 April 2023; Received in revised form 27 November 2023; Accepted 5 February 2024
Available online 16 February 2024

primarily being used in research contexts for large-scale analyses in areas such as (socio)phonetics and forensic speech science. However, as force-aligned data is used as the foundation for training many speech technologies, it stands to reason that issues and biases stemming from phoneme-level models which rely on tools such as forced aligners could be translating to issues further upstream (as discussed in Hutiri and Ding, 2022).

Examining the complications that L2 speech can present for tools such as forced alignment is increasingly relevant due to a recent shift towards segmentally informed models (i.e. phoneme-level segmentation) for many applications (e.g., Brown et al., 2021; Shi et al., 2020; Ferragne et al., 2019). Accurate time-aligned phonemic transcriptions are crucial when building this type of model. Inaccurate alignments indicate that the models may not accurately represent what they are trying to describe, classify, or identify. However, with the need for large amounts of data to train most state-of-the-art systems, manual segmentation becomes an unworkably time-consuming task.

1.1. The present study

This study contributes to better understanding of background processes in speech technology by testing how well a forced aligner trained for the alignment of L1 General American (GA) English speech performs on L2 English speech.

While there are several options for selecting the acoustic model used for forced alignment, in cases where several varieties are being analysed, building or adapting separate models for each is prohibitive. It is therefore useful to explore how accurate an 'off-the-shelf' standard-variety model would be. Additionally, this data will be used for the comparison of L2 varieties as part of an L1 identification system in future research. As the intention is to directly compare the realisations of phonemes, it is thus beneficial to use the same dictionary, phone set, and acoustic model across varieties (Brown, 2014; Huckvale, 2004). Using a single language model with its accompanying grapheme-to-phoneme dictionary therefore allows for straightforward comparison between varieties.

While only a few studies have tested off-the-shelf models on a mismatched variety (a variety the model was not trained on), the performance has been reassuring. The GA English model for the Montreal Forced Aligner (MFA; McAuliffe et al., 2017) has been tested on non-standard English varieties (Mackenzie and Turton, 2020), other languages (Babinski et al., 2019), and New Englishes (institutionalized Englishes such as Indian English; Meer 2020). Results have been comparable to the performance on American English, suggesting that a standard GA English model could be sufficient for L2 English speech, at least in certain contexts (discussed further in Section 2.2).

This paper will first provide some background on forced aligners and the performance of the MFA in prior studies, as well as outline factors that could present difficulties for alignment. We then examine the performance of the aligner with respect to L2 variety, sociophonetic/demographic factors, and segment type, followed by outlining systemic issues and in what contexts errors occur. Finally, we discuss whether an off-the-shelf GA English model is sufficient for use on L2 English speech and provide some recommendations for how to mitigate or address specific error types.

2. Background

2.1. Forced alignment

Forced aligners require three components: an orthographic transcript, an acoustic model, and a pronunciation dictionary. The orthographic transcript can be from a manual transcription, a set text (if recordings are of a read passage), or using speech-to-text software, although this introduces its own errors. The acoustic model and pronunciation dictionary are ideally matched with the target variety. This is because the acoustic model provides information about the expected pronunciation for each phone in the variety it has been trained on, while the pronunciation dictionary provides a list of words with their corresponding phonemic pronunciation(s). In addition, unless grapheme-to-phoneme capability has been added, for example with Pynini (Gorman, 2016) or Sequitur G2P (Bisani and Ney, 2008), the aligner is unable to predict the pronunciation of words outside of those in the dictionary.

The acoustic models for the MFA were built using Kaldi Speech Recognition Toolkit (Povey et al., 2011) and use a gaussian mixture model - hidden markov model (GMM-HMM) architecture. These models contain the expected distribution of acoustic features over time for each phone in the phone set. Contextual information is also taken into account by training on triphones along with monophones (McAuliffe et al., 2017). There are a large number of pretrained acoustic models available (40 languages as of writing). However, these have mostly been trained on L1 speakers of standard varieties or majority dialects such as Standard British English.

For non-standard varieties, forced alignment is typically limited to utilizing existing standard variety acoustic models or those adapted with additional data. If used in their off-the-shelf form, the acoustic model may not match or include all of the phones required to provide an accurate phonemic transcription. For example, Meer (2020) found that the MFA performed worse on vowels that were specific to Trinidadian English (and therefore lacked a matched vowel/acoustic model) than vowels that were present in both the standard and Trinidadian English varieties. Even if an L2 speaker's L1 acoustic model might be more appropriate, a mismatch in the phone set between the model and the dictionary will cause errors.

2.2. Performance of MFA in prior studies

A summary of the performance of the MFA from relevant studies, along with inter-rater agreement ratings from early studies using manual methods, is provided in Table 1. The accuracy of the aligner has been assessed using various thresholds and measures depending on the purpose for using the forced aligner as well as if the MFA was used directly or through another software. For example, DARLA (Dartmouth Linguistic Automation; Reddy and Stanford, 2015), which uses the MFA, extracts formant measurements to produce vowel plots. Babinski et al. (2019) then used the vowel space and subsequent distance of F1 and F2 from the manually measured means as a test of aligner accuracy. However, performance is more commonly measured directly, by the mean boundary displacement (BD), or percentage of BD measurements under a given threshold (see further detail in Section 3.4). While it is difficult to compare the accuracy measures from different studies directly, typical accuracy ranges from 6.4 to 28 ms for mean BD and 77 - 90 % for percentage of boundary placements less than 25 ms.

Gonzalez et al. (2020), MacKenzie and Turton (2020), and Meer (2020) all used a standard pre-trained acoustic model (General American English) to align data from non-standard varieties of English. These include Anglo-Australian English, regional British English varieties, and Trinidadian English, respectively. Findings showed that although variety had minimal impact on the overall performance (cf. target variety matched with acoustic model, McAuliffe (2021a) and McAuliffe et al. (2017); Table 1) the aligner tended to perform worse on varieties and phones that deviated markedly from American English. For example, Westray, a variety of Scots, had 79 % BD $<$20 ms compared to RP (Received Pronunciation had 90 % of BD $<$20 ms (MacKenzie and Turton, 2020). Using adapted pronunciation dictionaries, standard English acoustic models have also been used to align other languages with relative success (e.g., Babinski et al., 2019; Table 1).

2.3. Complications for alignment

There are several factors that can present difficulties for the aligner.

Table 1
Summary of accuracy scores from studies where an English acoustic model was used with the Montreal Forced Aligner (MFA) for the matched and mismatched test-train conditions, along with inter-rater agreement scores. Studies shown in reverse chronological order.

	Paper	Aligner	Acoustic Model	Testing Variety	Number of Speakers	Number of Boundaries	Metric	Value
Matched	McAuliffe (2021a)	MFA	American English (LibriSpeech)	American English (Buckeye Corpus)	40	–	Mean BD	16.3 ms
	McAuliffe et al. (2017)	MFA	American English (LibriSpeech)	American English (Buckeye Corpus)	40	–	% < 10 ms % < 25 ms Mean BD Std Dev BD	41 % 77 % 17 ms 11.2 ms
Mismatched	Meer (2020)	MFA	American English	Trinidadian English	11	1352	% < 10ms % < 20 ms	onset = 47.2 % offset = 50.2 % onset = 89.1 % offset = 77.8 %
	Gonzalez et al. (2020)	MFA	American English (LibriSpeech)	Anglo-Australian (stressed vowels)	4 (2 M, 2F)	2158	Mean Onset BD Mean End BD	28 ms 30 ms
	MacKenzie and Turton (2020)	MFA (via DARLA)	American English (LibriSpeech)	Six British English varieties	6 (5F, 1 M)	~1000 per speaker	% < 20 ms Mean Onset BD	~79–90 % 6.4 – 17.5 ms
	Babinski et al. (2019)	MFA (via DARLA)	American English	Yidiny	2	–	Mean F1 Mean F2	< 6Hz < 20Hz
Inter-rater	Raymond et al. (2002)	Manual	–	American English (Buckeye Corpus)	–	2813	% < 10 ms % < 20 ms	62 % 79 %
	Kvale (1993)	Manual	–	Norwegian	4 (2 M, 2F)	–	% < 20 ms	96.5 %
	Cosi et al. (1991)	Manual	–	Italian (IRST-MAIA IWSDB database)	–	–	% < 20 ms	88–90 %

Sociolinguistic and demographic factors such as gender and regional variety can result in phonetic realisations which deviate from the standard acoustic model, or introduce words not contained in the grapheme-to-phoneme dictionary. This is typically due to underrepresentation of speech from certain groups within the training data. For this reason, the performance of the aligner in our study was tested against the various demographic information provided by the speakers (see Section 3.1 and Appendix).

Another key factor that introduces variation into the realisation of phones is phonological context. Gonzalez et al. (2020) found that there was a significant impact on accuracy due to the phonological context of vowel segments. Surrounding sonorants (laterals, nasals, and approximants) tended to affect the accuracy more negatively than stops and fricatives, because vowel-sonorant boundaries are often acoustically unclear. However, the onset boundaries for vowels closely aligned with manual boundary placement. Meer (2020) also observed a negative impact on performance depending on the segment categories surrounding the boundary. They found reduced performance for vowel-vowel boundaries, pre- and post-pausal position, and – more generally – contexts where there is no clear acoustic boundary. However, these tended to also be boundaries that would also cause problems when manually labelling (Wesenick and Kipp, 1996).

Speaker-specific factors also impact the performance of forced aligners. Faster speech rate, for example, has been shown to negatively impact performance, although less so for the MFA compared with other aligners (MacKenzie and Turton, 2020; Bailey, 2016). For example, MacKenzie and Turton (2020) found that massive phonetic reduction, characteristic of fast speech, caused instances of extreme misalignment when using FAVE (Forced Alignment and Vowel Extraction suite, Rosenfelder et al., 2014).

Besides understanding all the factors that could impact the accuracy of the forced aligner, additional systematic errors due to particularities of the aligner are inevitable. When and why these occur is useful to know, especially if there is expected to be minimal manual intervention post-alignment, as is the case with many automatic systems. Previous studies have cautioned against leaving alignments unchecked for measurements that rely on precise alignments (e.g., Gonzalez et al., 2020; and MacKenzie and Turton, 2020; Babinski et al., 2019).

2.4. Complications arising from L2 varieties

L2 speech introduces sources of variability within- and between-speakers. Factors such as varying degrees of competence, strength of accent, as well as phonetic and phonological transfer from the L1 introduces larger within-variety differences than for non-standard L1 varieties (Lo and Wong, 2024; Davidson, 2011; Little, 1995; Flege and Bohn, 1989). For example, Wade et al. (2007) showed that Spanish speakers of English had approximately 33 % more variability in the acoustic realisation of vowels, in both height and backness, than L1 speakers of English. The L2 speakers in Laturnus' (2020) study showed on average twice as much variability in vowel production. Some L2s varied primarily along a single dimension (e.g., F1 for the Italian speaker of English) and others along both F1 and F2 (e.g., Thai and Russian speakers of English). Furthermore, the vowel space for each L2 variety differed in a unique way from L1 English vowel productions.

Insertions and deletions, resulting in deviations from the standard model, may also be more prevalent in L2 speech, although it is unclear how consistent or predictable they may be. For example, Broselow et al. (1998) discusses data from Wang (1995) testing how Mandarin Chinese speakers of English simplify stops /p, t, k/ and /b, d, g/ in coda position, none of which are permissible in the L1 phonology. Of the 81 % of voiceless and 98 % of voiced stops pronounced incorrectly, speakers most often either omitted the final stop (43–46 %) or added a vowel (36 %). Both strategies lead to CV structure, in line with the general

demands of the L1 phonology. However, in 19 % of cases, the voiced stops were instead devoiced. This would not be an expected simplification based on the L1 grammar. Insertions and deletions result in a mismatch between the acoustic signal and the grapheme-to-phoneme dictionary. This is likely to cause confusion with boundary placement, leaving the aligner looking for a missing phoneme or figuring out how to fit an additional one in. How systematically this is done by the aligner will be explored later in this paper.

While there could be adjustments made to dictionary entries, it is clear they would have to be language-specific due to differences between languages in the treatment of illegal grammatical structures. Hancin-Bhatt and Bhatt (1997) compared the production of complex onsets and codas of monosyllabic words with Japanese and Spanish learners of English. While both vowel epenthesis and consonant deletion were used for complex onsets, in the word-final codas consisting of liquid+obstruent or liquid+nasal, consonant deletion was overwhelmingly preferred over vowel epenthesis for both languages. Notably, however, they differed in what part of the consonant cluster was deleted. The Japanese speakers tended to delete the non-final consonant (the liquid) while the Spanish speakers tended to delete the final consonant (the nasal or obstruent). Davidson (2005, 2006) provides another interesting example of L1 English speakers who showed some difficulty with the articulatory movements required to produce certain onset consonant clusters in pseudo-Czech and Polish words. Not being able to fully overlap the articulation of successive consonants, such as /zg/, led to the impression of vowel epenthesis, although at a much shorter length than would be expected for actual lexical vowel epenthesis.

Variation between L2 speakers is likely to be further influenced by a range of social factors. Flege et al. (1995) identifies specific measurable factors found to impact the perceived strength of accent. In particular, significant factors included: age of learning (AOL), length of residence in an English-speaking country (LOR), speaker sex, and how often speakers use their L2 relative to their L1. AOL, LOR and speaker sex are included in the demographic information collected with the dataset for the current study (see Appendix for the full list of questions). While this information does not provide a comprehensive assessment of the speaker's competency and strength of accent, it does offer additional insight into aligner performance.

The types of inconsistencies presented above all have the potential to cause reliability issues with the performance of the aligner on the current dataset.

2.5. Research questions

The research questions in this study aim to evaluate the performance of the MFA with respect to issues identified in previous research regarding forced aligners more generally and with using a standard pretrained model with non-standard speech.

RQ1 To what extent is aligner performance impacted by L2 variety?

RQ2 To what extent is aligner performance impacted by other sociophonetic/demographic factors?

RQ3 Does the aligner perform better or worse on specific segment types?

RQ4 How does the aligner deal with inconsistencies in transcription and dictionary?

3. Methods and materials

3.1. Data

The recordings used in this study were taken from the speech accent archive (SAA) corpus (Weinberger, 2015). The recordings are each approximately 30 s in length and consist of a speaker reading a 69-word passage (included in the Appendix). This passage contains most of the consonants, vowels, and consonant clusters of standard American English. In addition, the speakers were asked for information about their linguistic background (questions listed in Appendix), including place of birth, other language(s) spoken, and how they learned English.

Five speakers from each of nine L2 English varieties were selected for analysis (Table A.4), with a balance of male and female speakers (24 F, 21 M). The nine L2 varieties as labelled in the corpus were: Arabic (Jiddah,[1] Saudi Arabia), Dutch (Antwerp, Belgium), French (Montreal, Canada), German (various cities, Germany), Italian (Naples, Italy), Korean (Seoul, South Korea), Mandarin (Shanghai, China), Portuguese (Sao Paulo, Brazil), and Russian (Moscow, Russia). The varieties will henceforth be referred to by the L1 of the speakers. For example, 'German' refers to L2 English material spoken by German L1 speakers. Where possible, speakers for a given language group were selected from the same city of birth (although this was not possible for German). Residence listed as 'USA' or 'Canada' was prioritized to ensure a target L2 of GA English. A full list of speakers can be found in the Appendix (Table A.4). While we expect degree of accentedness and fluency to have an impact on the performance of the aligner, this information was not provided with the corpus and therefore was not controlled for in this study. Potential correlations are briefly discussed in Section 5.1 (RQ1).

In total, 45 recordings were force aligned with the passage and then manually corrected for alignment (see Section 3.3).

3.2. Forced aligner

Several aligners were considered for this study, including WebMAUS (Munich Automatic Segmentation system, Schiel 1999), LaBB-CAT (Language, Brain and Behaviour Corpus Analysis Tool; Fromont and Hay 2012), and FAVE (Rosenfelder et al., 2014), each offering their own benefits and drawbacks. The decision to use the MFA (McAuliffe et al., 2017) was made based on recommendations from related studies which have directly compared aligner performance (mainly MacKenzie and Turton, 2020 and Gonzalez et al., 2020) as well as ease of use, accuracy, replicability, and ability to make modifications if necessary.

The CMU English pronunciation dictionary uses ARPABET (stressed) notation, which indicates stress via a number attached following a vowel and eliminates the use of reduced vowel transcriptions in the dictionary (Carnegie Mellon University, 1993). These labels will be referred to as segments and indicated by bolded uppercase letters e.g., **T** or **AH**, to distinguish them from a phone or phoneme which will be indicated using their IPA symbol. Approximate ARPABET-to-IPA list for the phonemes present in this dataset can be found in the Appendix (Tables A.2 and A.3).

3.3. Manual correction

The Praat script provided in the Corpus Phonetics Tutorial by Chodroff (2018) was used to generate a TextGrid for each recording as input to the MFA. Following automatic alignment, a duplicate of the MFA phone transcription in Praat (duplicate TextGrid tier) was created and hand-corrected by the first author to allow for direct comparison of boundary differences.

Notes on each of the speakers were made during corrections such as any significant impressions of voice quality, speaking rate, or large deviations in pronunciation that could impact the accuracy of the aligner, along with any decisions made in adjusting boundaries.

An auditory-acoustic approach was taken to make decisions with respect to boundary placement. This meant listening for sound transitions as well as looking at changes in the spectral characteristics and waveform, mainly following the principles laid out in Turk et al. (2006). Some error in the manual correction is expected; inter-rater agreement

[1] Note that this is also spelled Jeddah but is referred to here as Jiddah to remain consistent with the labelling in the Speech Accent Archive (Weinberger, 2015).

for a 20 ms threshold has ranged from 79 to 96.5 % in previous studies (e.g., Raymond et al., 2002; Kvale, 1993; Cosi et al., 1991; Table 1).

Some difference in phonetics and phonology relative to the target variety was expected since the pronunciation dictionary is based on American English. Therefore, decisions had to be made about how to treat inconsistencies. The segment/phoneme labels did not always match the phone. However, since realisations of phonemes with the same label are going to be compared in future research, as described in Section 1.1, the segment labels were left as they were. For other research questions this may not be the case, and a more accurate phone set/dictionary may be appropriate. Epenthesis and deletion were also expected issues. Inserted sounds were included with the phoneme they most likely represent according to existing literature on the variety. For example, in Spanish initial /sC / clusters are not phonotactically possible, and thus a Spanish L2 speaker of English might break the cluster by inserting a vowel preceding the cluster (e.g. [əskul] for 'school'). In such cases, the two phones would not be represented separately in the aligner transcription; both would be merged under the **S** segment label. This is more difficult to correct when the ungrammatical phoneme cluster occurs between words. In this case, the inserted sound was merged with the word-final segment (e.g., Fig. 15 in Section 4.4.3).

The following additional decisions were made consistently when adjusting the alignment:

- Voicing boundaries and onset/offset of a periodic waveform were used for decisions on vowels and used the 'max' criterion that includes laryngeal activity (Turk et al., 2006:17).
- The initial boundary for stops was left alone when it was unclear when the closure occurred (e.g., for phonemically voiceless stops following a pause).
- Stop releases were included with the stop segment to remain consistent with existing automatic systems (Jurafsky and Martin, 2009: 255).
- For vowel-approximant boundaries, the boundary marker was placed at the approximate midpoint of the transition.
- The **N-D** boundary of "and" was left alone because it was almost always reduced to some variation of /æn/. This variant was later added as a dictionary item.

3.4. Measurements and calculations

The accuracy of the automatic alignment was analysed by comparing the placement of boundaries from the aligner (MFA) with those of the manually corrected boundaries. Two main metrics were used to assess accuracy for individual segments/boundaries: *Boundary Displacement* (1) and *Overlap Rate* (2). Two additional metrics were then calculated to assess the accuracy for a whole recording or group: *Percent <20 ms and 10 ms*, and *Total Overlap Rate (TOR)* (3).

3.4.1. Boundary Displacement (BD)

Further specified as Segment Onset Boundary Displacement (OBD), and Segment End Boundary Displacement (EBD). This measures the absolute displacement between a boundary placed by the automatic aligner and the manually corrected placement of the boundary measured in milliseconds. The lower the boundary displacement, the greater the accuracy.

$$BD = |boundary_{manual} - boundary_{automatic}| \quad (1)$$

An additional measure, Directional BD is calculated using Eq. (1) without taking the absolute value, to provide information on whether there is any bias towards early or late placement of the boundary. The resultant value is positive when the automatic aligner has placed a boundary too early (i.e., need to add 5 ms to the boundary placement), and negative when the boundary was placed too late (i.e., need to subtract 5 ms from the boundary placement).

3.4.2. Overlap Rate (OvR)

This is a duration-independent measure based on the percentage of overlap between the automatically aligned segment and the manually aligned segment (Paulo and Oliveira 2004: 39). Both onset and end boundary displacement are considered in this measurement, therefore providing a better general assessment of the accuracy for a segment than the boundary displacement values individually. A value of 1 indicates complete overlap while a value of 0 indicates no overlap.

$$OvR = \frac{dur_{shared}}{dur_{man} + dur_{auto} - dur_{shared}} \quad (2)$$

Where,

dur_{shared} = duration between the latest time stamp and the earliest time stamp

dur_{man} = duration of manually aligned segment

dur_{auto} = duration of automatically aligned segment

3.4.3. Percent <20 ms; Percent <10 ms

Percentage of boundary displacements within 10 milliseconds and 20 milliseconds of manually corrected boundary placement. For this study, the percentage of OBD was used because all segments were included in the analysis, and it would therefore be redundant to take both the onset and end displacements. This metric is an indicator of the percentage of the boundary corrections that are small. Additionally, Cosi et al. (1991) showed that choosing a threshold of 20 ms was optimal for inter-rater agreement. The consensus following this has been that 20 ms is a good threshold for final evaluation of aligner accuracy (Fromont and Watson, 2016) and has been used as the primary measure for accuracy in most of the related literature. Interpretation of these scores should also take into consideration that the MFA works in 10 ms frames. Therefore, the percent <20 ms can be interpreted as within two frames of analysis and percent <10 ms within one frame.

3.4.4. Total Overlap Rate (TOR)

While the percent <20 ms OBD measurement does provide a good indication of the accuracy of the alignment, and in particular, the magnitude of errors, it can be skewed by length of phones and speaking rate. Another measure of accuracy, using the time independent OvR, involves measuring how much the automatically aligned segments overlap with the manual alignments for the whole language group. This can be measured as in (3) for a given group where n is the number of segments:

$$TOR = \frac{\sum_1^n OvR_i}{n} \quad (3)$$

Eq. (3) calculates how accurate the aligner is across all segments in a recording, and therefore can provide a good estimate of overall accuracy for a speaker or language group. By taking the average, each segment is treated as equal instead of each unit of time, providing a more accurate assessment of accuracy than taking the overall ratio of shared duration and total duration.

3.5. Removal of speaker errors

Before the analysis of accuracy (RQ 1–3), speaker errors were removed. Since all speakers were tasked with reading the same passage, the orthographic transcription used as input to the aligner was the same. Therefore, speakers who deviated from the passage created unexpected input for the aligner; albeit errors may also indicate competency in the language.

Segments were removed from analysis only if they were due to errors unrelated to the accuracy of the aligner and there was no assumption that the aligner could deal with them without manual intervention. Reasons for segment removal included:

- Repetition of a phrase
  'to bring- to bring these things' [Korean 23; 3.0 s]

- False starts
  'into th- into three' [Arabic 44; 22.8 s]
- Incorrect order of words / inserted words
  ***from the** store*
  *'**home from** store' [Russian 6; 4.3 s]*
- Deletion of word(s)
  *big toy **frog** for the kids*
  'big toy for the kids' [Italian 20; 14.9 s]
- Incorrect word used (not a mispronunciation)
  *things **into** three red bags*
  *'things **in** three red bags' [Portuguese 28; 22.4 s]*
- Segments that were condensed due to a duration hard coded in the initial script used to create the TextGrids. When generating the TextGrid files, the script expected 20 ms of silence before the end of the recording leading to occasional early placement of the last word boundary (compare the automatically aligned *Phones-Auto* tier to the corrected *Phones-Hum* tier in the area indicated by the red box in Fig. 1).

Further investigation of these kinds of errors is discussed in the results for RQ4 (Section 4.4). This information was used to help create error-checkers and inform adjustments to scripts for future use of the forced aligner on L2 data.

Data was not discarded where the issue was something the aligner could reasonably be assumed to handle. For example, Fig. 2 shows the forced aligner mistaking background noise for a fricative. Comparing the *Phones-Auto* tier to the *Phones-Hum* tier, we can see that the aligner has included background noise as part of the **F** segment in "for" highlighted by the red box. In total, 9852 observations were retained in the analysis out of 9931 (99.2 % retained).

4. Results

4.1. (RQ1) to what extent is aligner performance impacted by L2 variety?

The Total Overlap Rate (TOR) was calculated for each variety, indicating how much of the aligned segments overlapped with the manually aligned segments to provide a segment length-independent measure of accuracy (Table 2; Fig. 3). These values ranged from 81.4 – 92.1 % overlap with the corrected alignments. Also listed is the standard deviation of TOR and of OvR. The first of these is relative to individual speaker TOR values, and as such, displays the degree of variation between speakers in a language group (1.0 – 4.5 %). The latter shows the variation of OvR for all segments in a language group. These ranged from 14.0 – 21.8 %.

The ranking of TOR is slightly different from that of the standard percent <20 ms ranking (cf. Table 3). For languages whose ranking has improved, this suggests that while there may have been a higher percentage of larger boundary displacements (>20 ms), these were not as significant relative to the overall length of the segment. The standard deviation of the TOR shows there is little difference in accuracy between speakers within a language group for the OvR metric. The standard deviation of the segmental OvR however, ranges from 14.0 – 21.8 % for the tested varieties. This shows there is considerable variation in the OvR between segments.

For direct comparison against previous studies (Table 1), statistics based on OBD were calculated (Table 3; Fig. 4). The MFA showed good performance on the tested varieties with the percentage of OBD <20 ms for the tested varieties ranging from 80.0 - 93.0 % and mean OBD values from 4.6 – 14.1 ms. The aligner performed best on German, Italian, and French (all having over 92 % of tokens with OBD <20 ms), and by far the worst on Russian (80.0 %).

The mean OBD values are all lower than those for aligning American English ($\bar{x} = 16.9$ ms; McAuliffe, 2021a). However, there is considerably more variation in the displacement measurements. The standard deviation of OBD for the languages in Table 2 range from 11.8 - 36.2 ms compared with 11.3 ms for American English (McAuliffe, 2021a). This suggests that while the mean accuracy may be better, the consistency in performance is worse than for American English. It is useful to note that the measurements for American English used in previous studies were based on CVC boundaries in conversational speech (Buckeye Corpus; Pitt et al., 2007) and therefore in some ways presents a more challenging speech signal to align. However, they also ignore pronunciation variation that includes insertions and deletions that are present in the data for this study.

The results here also show comparable performance to the use of the General American English model/dictionary on non-standard varieties of British English ($\bar{x} = 6.7$–17.5 ms, MacKenzie and Turton, 2020). This may present a more comparable measure of accuracy given all segments were included in the analysis as opposed to only CVC boundaries. The percentage of boundaries placed within 20 ms of the corrected boundary (80.0–93.0 %; Table 3) were also relatively similar to the 77–90 % <20 ms reported by MacKenzie and Turton (2020).

The percent <10 ms OBD threshold has a slightly different ranking by language from the 20 ms threshold, as well as lower accuracy scores

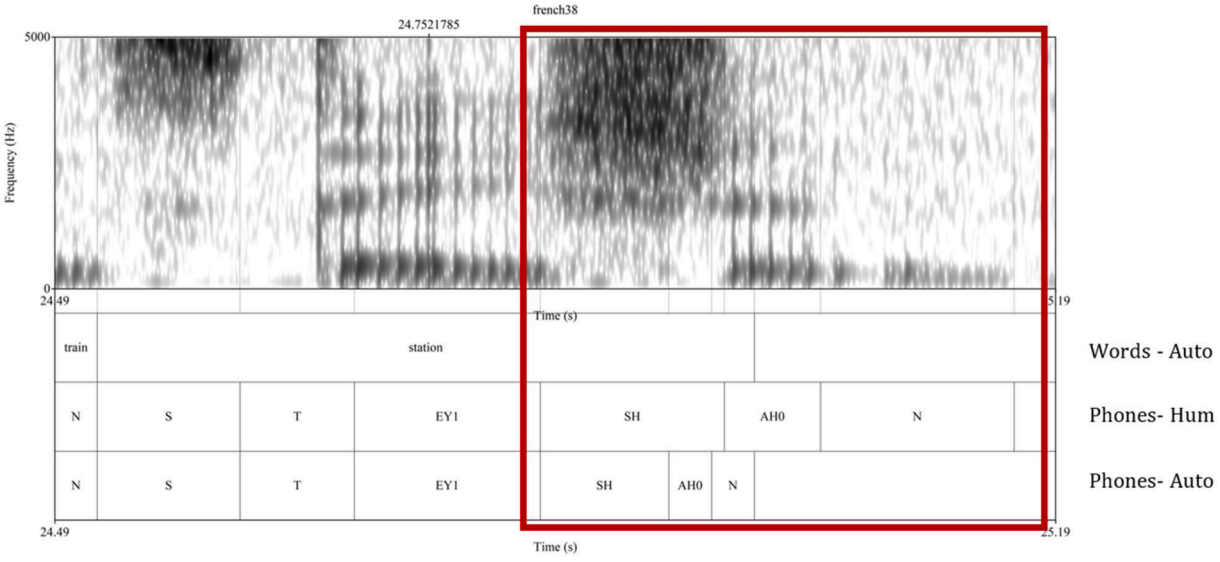


Fig. 1. Speaker French 38 saying "*station*" which finished beyond where the script has placed the last word boundary on the Words-Auto tier. This resulted in the remaining segments being condensed into the remaining space within the word boundaries on the Phones-Auto tier.

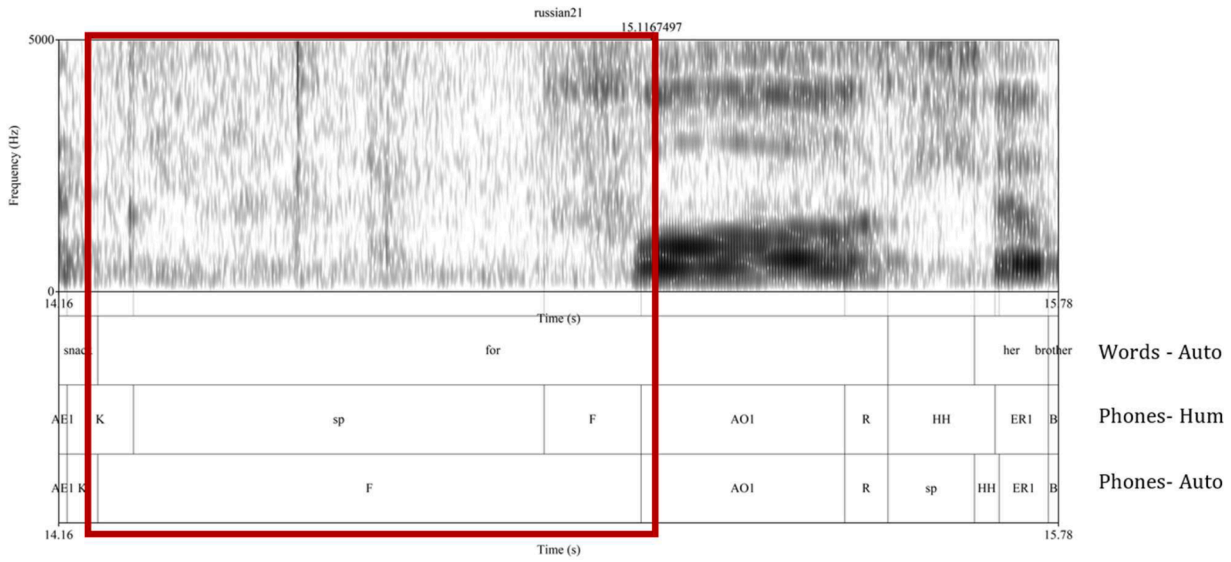


Fig. 2. Speaker Russian 21. Example of the forced aligner mistaking background noise for a fricative. This is considered an error of the aligner and therefore not an observation to be removed from the analysis.

Table 2
Performance of the MFA on each L2 variety as measured by the average percentage of overlap (TOR) between the automatically and manually aligned segments with speaker errors removed (as described in Section 3.5). Arranged from the highest to lowest percentage of TOR. Both standard deviation of OvR and TOR between speakers is provided.

Language	Std. Dev. Of TOR between speakers (%)	Std. Dev. Of OvR (%)	TOR (%)
French	1.0	14.0	92.1
German	1.8	15.1	91.8
Italian	1.9	16.1	90.9
Dutch	2.2	16.9	89.0
Arabic	2.6	15.1	88.1
Korean	1.3	18.0	87.6
Mandarin	2.2	18.0	85.9
Portuguese	1.4	19.3	83.4
Russian	4.5	21.8	81.4

Table 3
Statistics of speaker means for each L2 variety with speaker errors removed (as described in 3.5). Arranged from the highest to lowest percentage of OBD <20 ms. Percentage of OBD <10 ms is also listed although has a slightly different ranking.

Language	Mean Speaking Rate (segments/*sec*)	Mean OBD (ms)	Std. Dev. of OBD (ms)	% OBD <20 ms	% OBD <10 ms
German	11.6	5.0	19.7	93.0	83.2
Italian	11.0	5.4	19.2	92.7	82.4
French	11.2	4.6	11.8	92.3	83.9
Dutch	11.0	6.5	13.3	90.4	78.7
Portuguese	11.6	8.9	15.9	87.1	69.6
Korean	9.4	7.7	17.6	87.2	76.3
Arabic	9.3	9.8	27.9	85.6	74.9
Mandarin	9.9	9.3	18.8	85.4	71.9
Russian	10.0	14.1	36.2	80.0	66.8

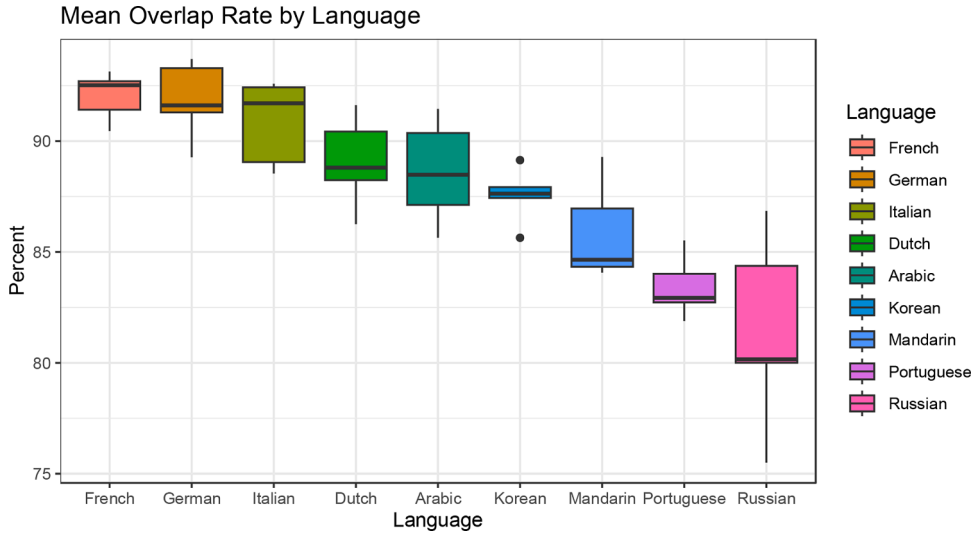


Fig. 3. Distribution of speaker means of TOR ranked by aligner performance from best to worst (cf. Figure 4).

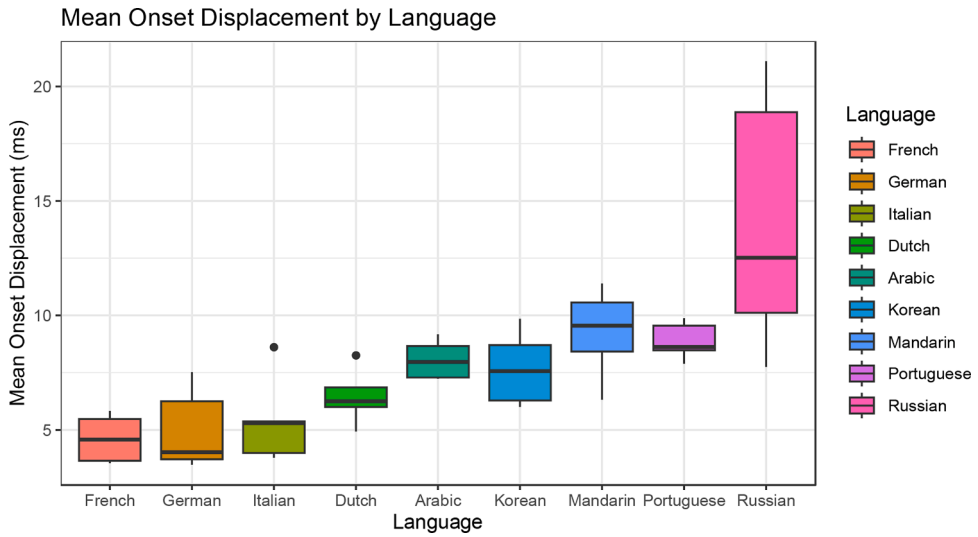


Fig. 4. Distribution of speaker means for Onset Boundary Displacement by language. Listed in order of TOR rankings in Table 2 to maintain consistency between plots (cf. Figure 3).

for all varieties. These values range from 66.8 – 83.9 %. Some of the varieties, such as Portuguese, Mandarin, and Russian, have notably lower accuracy with this reduced threshold. This indicates that fewer corrections made for these varieties were small boundary adjustments.

As a general pattern, it is worth noting that the speakers of the varieties with lower accuracy had slower average speaking rates. It is possible this could be correlated with the fluency of the speakers in the groups, as speaking rate is often used as an indicator of fluency for L2 speakers (Tavakoli and Wright, 2020). Speaking rate is further investigated in Section 4.2.

To test for significant performance difference between varieties, a linear model was fitted to the speaker means OBD data (Table 3) as well as by-speaker TOR data (Table 2) with a fixed effect of language (Table 4). The reference was adjusted to the best performing variety according to the TOR ranking (French). Table 4 shows that there was no significant difference in aligner performance between French and German, or Italian for either OBD or TOR. A significant difference was found with Dutch but only for TOR (Table 4 Right). There was, however, a significant difference ($p < 0.05$) between French and the remaining varieties for both OBD and TOR.

4.2. *(RQ2) to what extent is aligner performance impacted by other sociophonetic/demographic factors?*

A linear mixed model was fit to the segmental OBD and OvR data to determine the impact of speaker demographic factors on overall performance. The model included *AOL, LOR, Number of years knowing English* (an extrapolated value: *(age – (age began learning English))*), along with *Speaking Rate* as fixed effects, and *speaker* and *previous phonetic context* as random effects. The only significant predictor was *Language*, which is illustrated in Section 4.1.

While *Speaking Rate* was not a significant predictor in the mixed model, for comparison with prior studies, Spearman's rank correlation tests were performed against by-speaker mean OBD values as well as the by-speaker percent < 20 ms threshold accuracy metric. The tests showed significant, but relatively weak, correlations indicating better accuracy of the aligner and smaller boundary displacements on average for faster speaking rates (Figs. 5 and 6; OBD: $S = 20{,}628$, $p = 0.01$, rho = −0.36;% < 20 ms: $S = 8398$, $p = 0.00$, rho = 0.45). However, there are two speakers who appear to fall outside the general trend for the OBD metric (the Russian speakers with high OBD and low speaking rate in Fig. 5) that may be skewing the correlation slightly.

As indicated by the mixed model, using the individual segment OBD tokens as opposed to speaker means, a Spearman's rank correlation test shows that the correlation between OBD and speaking rate is not

Table 4
Linear regression model for distribution of speaker mean OBD and TOR within each variety. The reference value (Intercept) is set at the best performing variety in TOR – French. * Indicates statistical significance.

Model Info								
Observations			45				45	
Dependent Variable			OBD				TOR	
Model Fit								
R^2			0.63				0.75	
Adj. R^2			0.55				0.70	
	Est.	**S.E.**	**t val.**	**p**	**Est.**	**S.E.**	**t val.**	**p**
(Intercept) French	4.62	1.05	4.38	0.00	92.04	1.01	91.44	0.00
Arabic	3.45	1.49	2.31	0.03*	−3.43	1.42	−2.41	0.02*
Dutch	1.84	1.49	1.24	0.22	−2.98	1.42	−2.09	0.04*
German	0.38	1.49	0.26	0.80	−0.21	1.42	−0.15	0.88
Italian	0.80	1.49	0.53	0.60	−1.18	1.42	−0.83	0.41
Korean	3.07	1.49	2.06	0.05*	−4.49	1.42	−3.15	0.00*
Mandarin	4.63	1.49	3.11	0.00*	−6.19	1.42	−4.35	0.00*
Portuguese	4.27	1.49	2.86	0.01*	−8.63	1.42	−6.06	0.00*
Russian	9.46	1.49	6.34	0.00*	−10.67	1.42	−7.49	0.00*

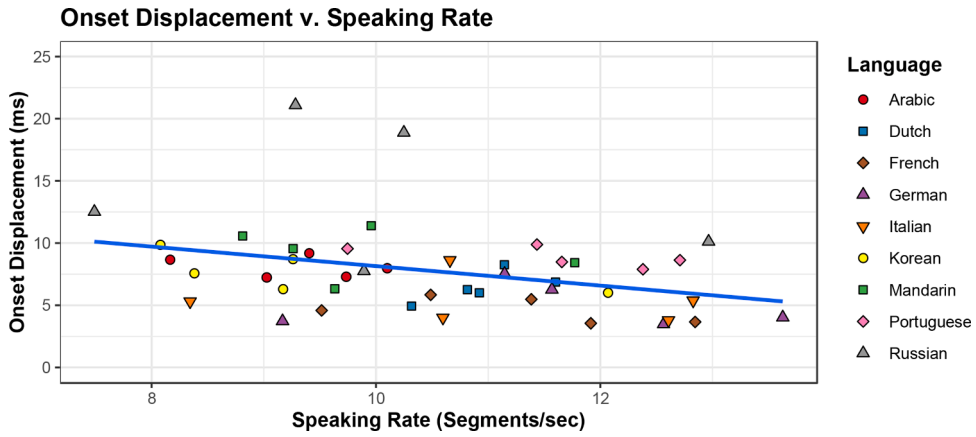


Fig. 5. Patterning of OBD as a function of speaking rate using speaker mean OBD values ($p = 0.01$, rho $= -0.36$).

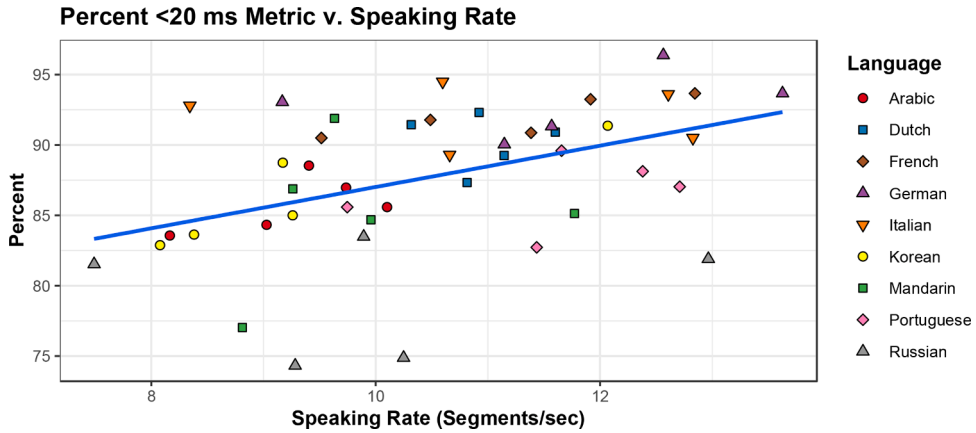


Fig. 6. Patterning of the percent OBD <20 ms as a function of speaking rate ($p = 0.00$, rho $= 0.45$).

significant: ($S = 1.61e{+}11$, p-value $= 0.43$, rho $= -0.01$). Additionally, there was no significant correlation between speaking rate and by-speaker TOR, indicating speaking rate has a stronger impact on the absolute displacement rather than general accuracy.

It is worth noting that there also appear to be language-specific patterns displayed (also seen in Table 3). For example, most Mandarin speakers tended to have slower speaking rates, while most Portuguese speakers tended to have faster speaking rates. This could be an indicator of fluency as mentioned in the previous section. However, there could also be interactions with features of the L1. For example, Coupé et al. (2019) showed that speaking rate can be constrained by the structure of the language being spoken. Of the 17 languages they tested, syllable-timed languages such as Spanish and Italian tended to have faster speaking rates, while tone languages such as Mandarin tended to be slower. If there are language constraints impacting the L2, resulting in consonant cluster simplification for example, it is possible this could have a secondary impact on the speaking rate. Conversely, other confounding factors such as pronunciation competence may affect the performance without affecting speaking rate. If the phone realisation strongly deviates from the phoneme model, the aligner will have difficulty correctly placing boundaries and could result in more, or larger errors (discussed further in 4.4.2).

4.3. *(RQ3) does the aligner perform better or worse on specific segment types?*

A more fine-grained analysis of the performance of the aligner was conducted by categorising segments by type, defined in terms of manner of articulation, and tagging them with new labels following Gonzalez et al. (2020). The classification is summarised in Tables A.1 and A.2. Vowels were additionally tagged to indicate stress according to their dictionary label. It is possible to break down the performance further by segment. However, due to limited data, it is unlikely any strong conclusions can be drawn from this information, and segment by segment data is therefore not shown here.

It is notable that boundary displacements are more than likely due to the interaction between adjoining segments rather than the performance of the aligner on any one segment category alone. Therefore, OvR was used as the main measure for individual segment accuracy since it takes into account both the Onset (OBD) and End BD (EBD), while OBD alone was used for segment clusters. Furthermore, some of the variation in performance may occur due to mismatches in manner of articulation, resulting in accuracy being influenced by both context and acoustic model mismatches.

The presentation of results below begins by discussing segment accuracy as a whole for all languages, followed by a by-language discussion of segment accuracy and a focus on vowels to provide comparison with previous studies.

4.3.1. All segments. Pre-segmental pauses heavily impacted the distribution of the aligner accuracy on different segment types; $\bar{x} = 7.7$ ms, sd $= 20.4$ ms when post-pausal boundaries are included, compared with $\bar{x} = 6.1$ ms, sd $= 14.8$ ms when they are not. The phone was always contained within the post-pausal segment. However, as a buffer, there would often be a period of silence included as well. The adjustments made to segment boundaries following a pause or silence were mostly due to a personal preference to indicate as accurately as possible the start of a phone (further discussed in Section 4.4.4). For the following analysis, segments that followed a pause were therefore removed to

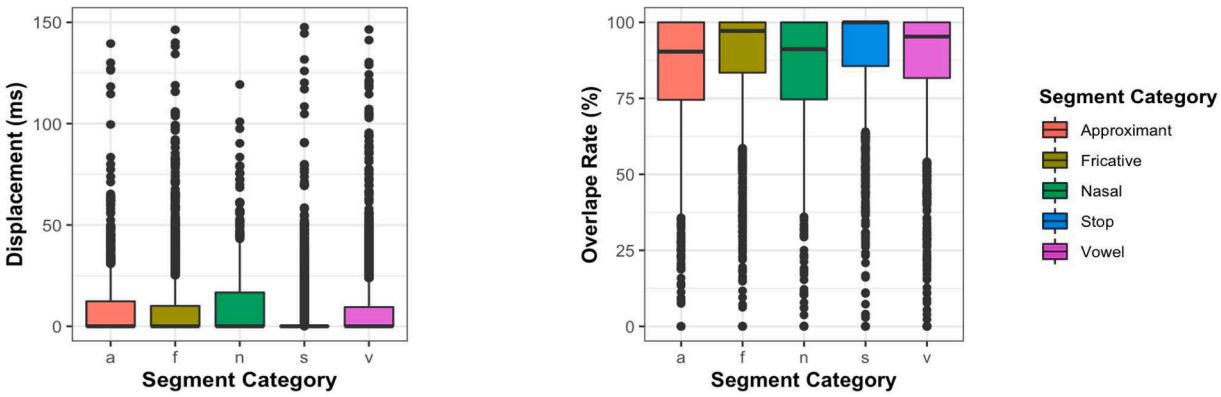


Fig. 7. OBD (7a) and OvR (7b) by segment category for all languages.

allow for better investigation of the aligner's performance on different segment types. A total of 691 of the 9861 observations (7 % of the data) were removed.

Fig. 7 shows the performance using OBD in milliseconds (7a) and OvR (7b) for each segment category. Fig. 7a shows that there is generally higher OBD for nasals than other segments. This indicates that segments with following nasals require the most manual correction and therefore tend to produce the least accurate automatic alignment. It is clear that the aligner is very good at identifying the beginning of stops, as they displayed the lowest OBD.

The OvR values indicate that nasals, approximants, and vowels have the lowest correct overlap percentage (Fig. 7b). Stops and fricatives, on the other hand, had the highest OvR with the least variation, indicating they are more accurate and robust to alignment error.

Vowel-vowel (v-v), fricative-fricative (f-f), and vowel-nasal (v-n) clusters had the highest BD values and were therefore less well aligned than other cluster types for all varieties (Table 5). This is likely due to the acoustic similarities between vowels and nasals, and the difficulty in separating adjacent phones produced in the same manner. For fricative-fricative clusters, depending on the variety, the center of gravity could be either very similar between the two segments, or one could be more similar to the acoustic model than the other, causing issues with alignment. The best alignment was between approximant-stop (a-s), nasal-fricative (n-f), and stop-stop (s-s) clusters. All the best alignments seemed to occur where there was a clear difference in the spectral characteristics between adjacent segments. No skew in direction of the BD was found for any of the segment clusters.

4.3.2. By language. The performance of the aligner on individual segment categories, along with relative performance between segment types, was dependent on the L2 variety (Fig. 8). For most varieties, the aligner had more difficulty correctly aligning nasals and approximants, and displayed similar trends of performance across categories. One notable exception was Arabic, which was the only case in which the aligner performed the best on nasals.

The BD between all segment clusters was analysed for each language (Table 6). On average, the displacements were 10 ms and below, indicating very good performance by the aligner.

While the performance on specific cluster types mostly ranked similarly between varieties, there were some cluster types that patterned differently in certain languages. The coloured cells in Table 6 display a few clusters with this type of pattern. For example, the boundary between stops and nasals (orange cells) is highly accurate for most of the languages except Korean, Mandarin, and Russian, while fricative-fricative boundaries (blue cells) are generally more difficult for the aligner with the exception of French. These patterns could of course be influenced by the limited contexts available in the dataset.

4.3.3. Vowels. The aligner performed slightly better on stressed vowels than unstressed vowels. Nonetheless, the difference in accuracy did not reach statistical significance for any of the varieties. Secondary stressed vowels were ignored due to a comparatively small number of tokens. As a whole, the EBD of vowels followed by a nasal or vowel was higher than for vowels followed by any other category (Fig. 9; see also Table 5). Additionally, boundary placement in vowel-consonant clusters was generally more difficult for the aligner than consonant-vowel clusters ($\bar{x}_{v-c} = 7.5$ ms; $\bar{x}_{c-v} = 5.2$ ms).

There was of course, a reduced phone set due to the limited number of vowel-X combinations, where X is any non-empty unit, in the reading passage (e.g., there are only 4 unique v-v combinations). However, of the approximants, **L** had more of an impact than **R** on the EBD of vowels. Of the nasals, **N** and **NG** were more likely to cause boundary placement problems than **M**. In addition, apart from **S**, all segments that heavily impacted the end boundary displacement of vowels were voiced. Even then, it is possible that the **S** segments were realised as [z] even though they were not labelled as such.

Table 5
Average BD for segment clusters listed from lowest BD (most accurate) in milliseconds to highest BD (least accurate). Where, (a) = approximant, (f) = fricative, (n) = nasal, (s) = stop, and (v) = vowel.

Lowest BD										
a-s	**n-f**	**s-s**	**n-s**	**f-s**	**v-s**	**n-v**	**s-v**	**f-v**	**s-f**	**f-n**
2.5	2.8	3.2	4.2	4.4	4.8	4.8	4.9	5.2	5.3	5.8
s-n	**a-v**	**v-f**	**f-a**	**a-f**	**v-a**	**s-a**	**v-n**	**f-f**	**v-v**	
5.8	6.2	6.2	6.2	7.3	7.7	7.9	11.9	12.6	13.4	
										Highest BD

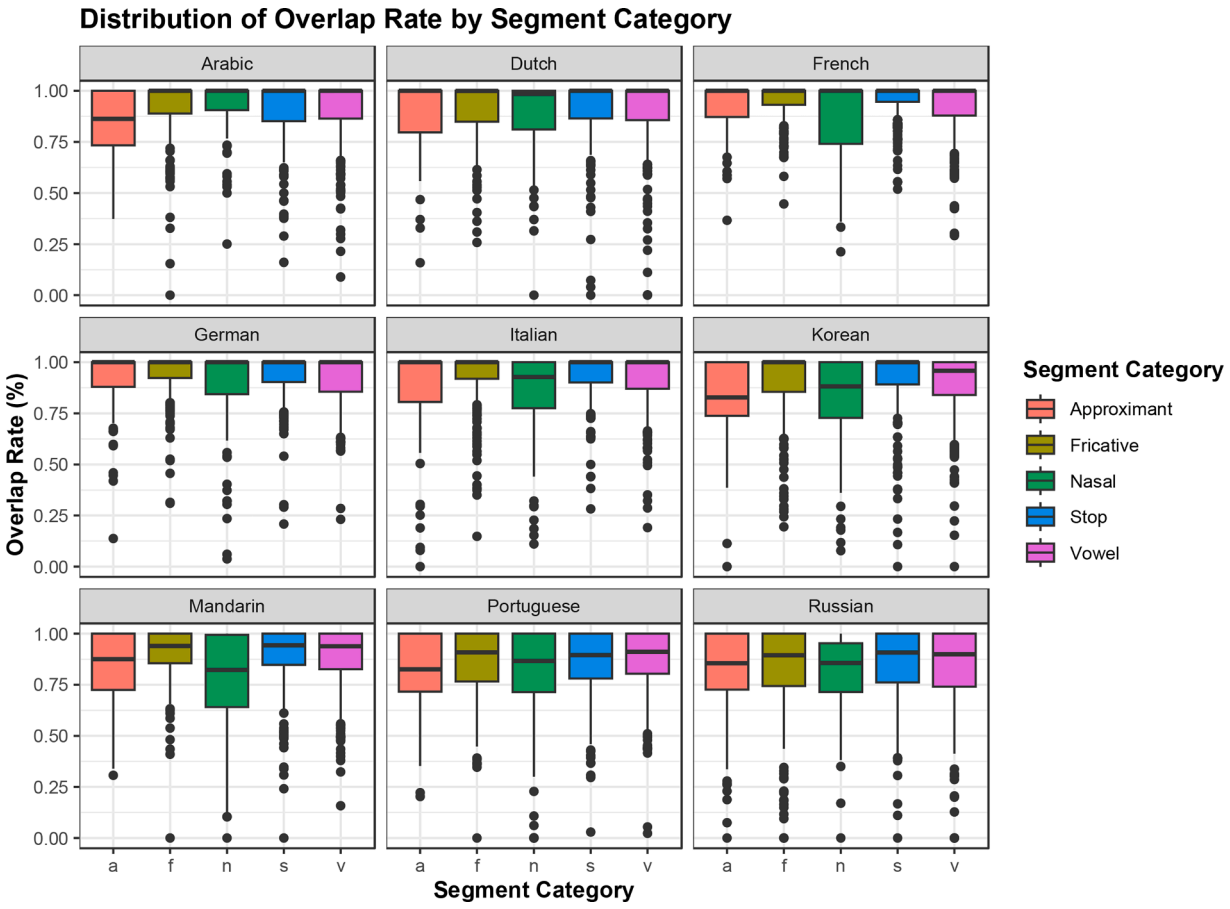


Fig. 8. Distribution of OvR by segment category. Coloured by segment category.

Table 6
Ranking of segment cluster types from lowest BD to highest BD for each language (decreasing in accuracy). Blue and orange cells depict examples of segment clusters that are similarly ranked for most languages but show very different rankings for a few varieties.

Arabic	Dutch	French	German	Italian	Korean	Mandarin	Portuguese	Russian
s-n	s-n	s-n	s-n	a-s	a-s	s-s	a-s	n-s
n-f	f-n	s-s	f-n	s-n	s-s	a-s	n-f	s-s
n-s	v-s	f-s	a-s	n-f	n-f	n-f	f-s	a-s
f-s	n-s	a-s	n-s	s-s	f-s	a-f	s-n	f-v
n-v	f-s	n-f	f-s	n-s	s-f	v-s	n-v	n-v
s-s	a-v	s-f	n-f	f-s	n-v	f-a	a-v	v-a
a-s	n-f	n-s	s-s	v-s	a-f	f-n	v-a	s-v
f-v	s-v	f-f	s-f	a-f	f-n	s-v	v-v	f-a
s-v	n-v	v-s	a-f	s-v	n-s	f-v	s-v	n-f
v-s	v-f	f-a	a-v	v-v	v-f	v-f	v-s	s-f
f-n	f-v	s-v	f-a	f-v	f-v	v-a	a-f	v-s
a-v	v-a	v-a	n-v	s-f	s-v	f-s	v-f	a-v
v-f	s-a	a-f	f-v	f-n	v-s	a-v	s-f	v-f
v-n	a-f	a-v	v-s	v-f	f-a	n-v	n-s	s-n
f-a	s-s	v-f	s-a	n-v	s-a	s-f	f-n	v-n
s-f	s-f	f-n	v-f	a-v	a-v	s-a	s-s	f-s
s-a	f-a	s-a	s-v	v-a	v-v	f-f	f-v	s-a
f-f	v-v	n-v	f-f	s-a	s-n	v-v	s-a	f-n
a-f	v-n	f-v	v-v	f-a	v-a	n-s	f-a	f-f
v-a	a-s	v-v	v-a	f-f	v-n	v-n	f-f	a-f
v-v	f-f	v-n	v-n	v-n	f-f	s-n	v-n	v-v

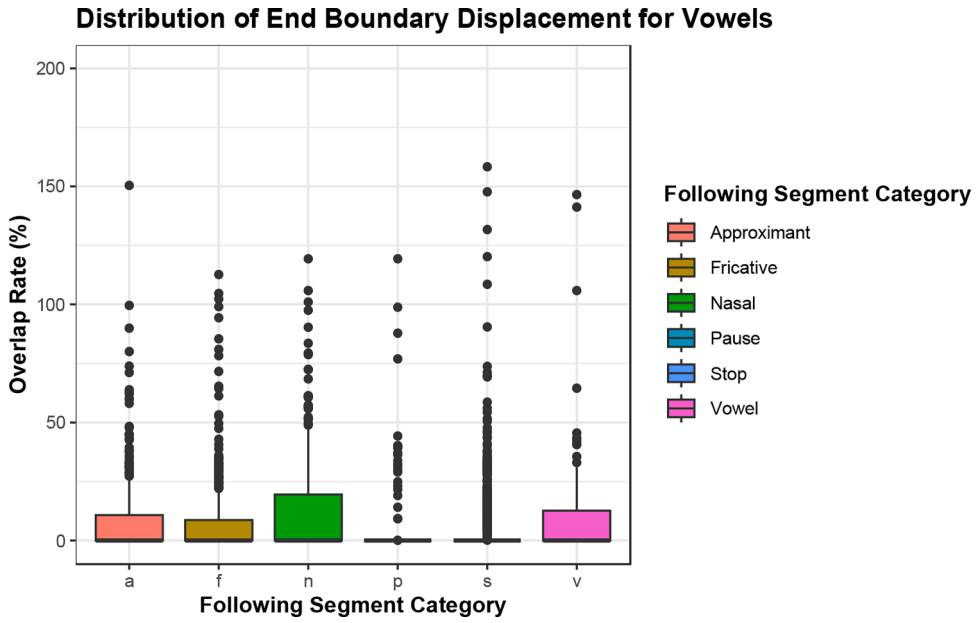


Fig. 9. End Boundary Displacement distribution for vowels according to the following segment category. Segments are coloured by segment category.

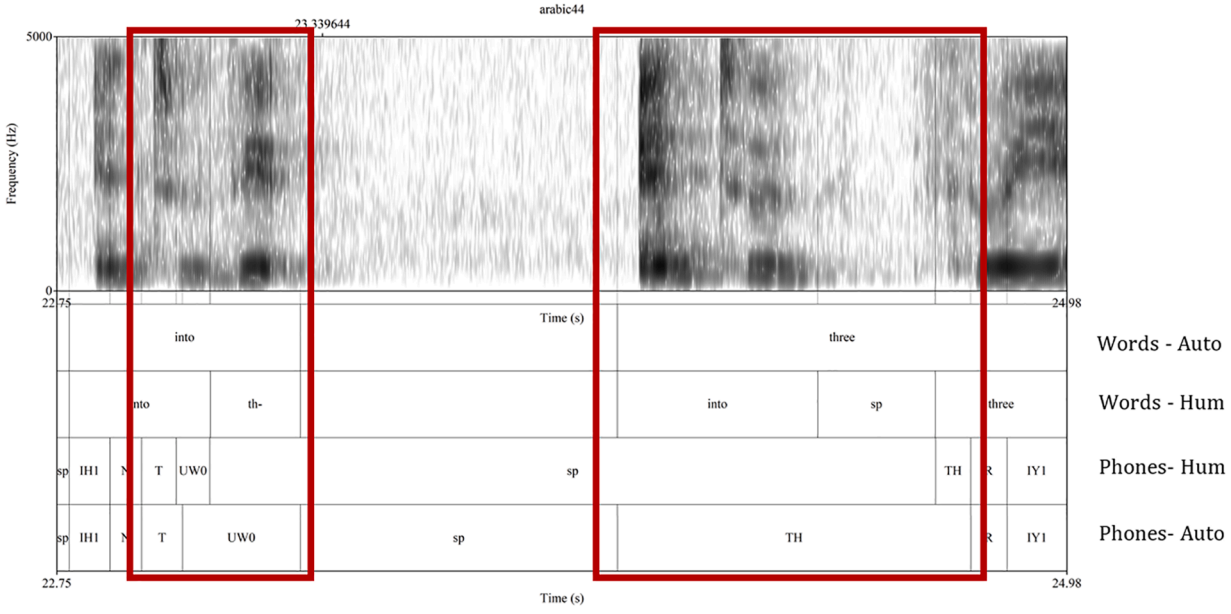


Fig. 10. Example of a hesitation and repetition of a word. "*into* ***th- into*** *three*" by speaker Arabic 44. Only **UW0** /u/ for "into" and the **TH** /θ/ of "three" were extended where there were speaker errors.

4.4. (RQ4) how are inconsistencies with the transcription, dictionary, and acoustic model treated? Can systemic issues inform post-alignment error-checks?

The following section revisits the data that was removed from analysis in the previous research questions (Section 3.5) and discusses how the aligner responds to insertions, deletions, and speaker errors such as hesitations or repeated phrases when there is no verbatim transcription. Additionally, boundary placement errors that were greater than 80 ms are categorised

4.4.1. Speaker errors. Speaker errors that involve additional words or phrases such as with false starts and word/phrase repetition (e.g., Fig. 10 and Fig. 11), are all treated by the aligner in a similar manner. In these instances, only the nearest following segment appears to be affected. The aligner extends the segment to encapsulate the error unless it is following a period of silence. In that case, the initial triggered segment is extended. For example, in Fig. 10, in the red box on the left, on the *Words-Hum* tier, we can see the speaker has a false start for the word "three" and they then repeat the phrase "into three" in the red box on the right. As can be seen on the *Phones-Auto* tier, the vowel portion of "into" (**UW0** /u/) is extended to include the false start until the pause, while the **TH** /θ/ segment has been extended following the pause. Similarly, in Fig. 11, the additional phrase "bring those- ask her to" shown on the *Words-Hum* tier was captured under the **AH0** /ʌ/ segment from "to." On rare occasions, the aligner correctly ignored a word which was not present in the transcript, such as the example in Fig. 12 where the additional words "of the," indicated by the red box, were marked as "sp" (silence/non-speech).

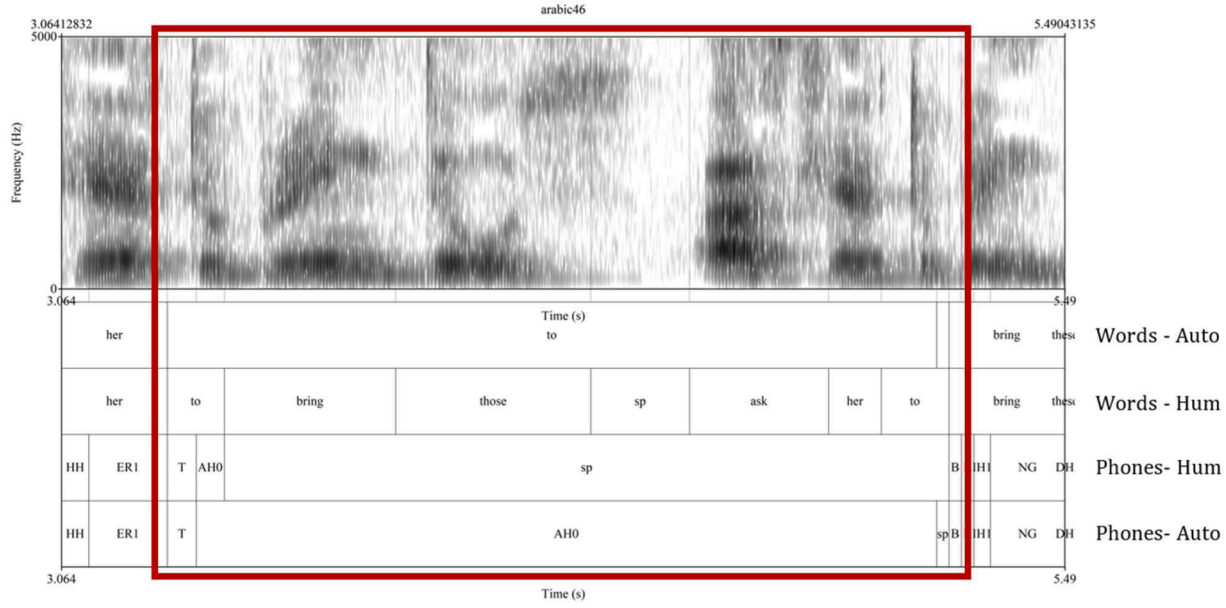


Fig. 11. Example of a speaker repeating a phrase due to self-correction. "*Ask her to bring* ***those- ask her to bring these***" as spoken by speaker Arabic 46.

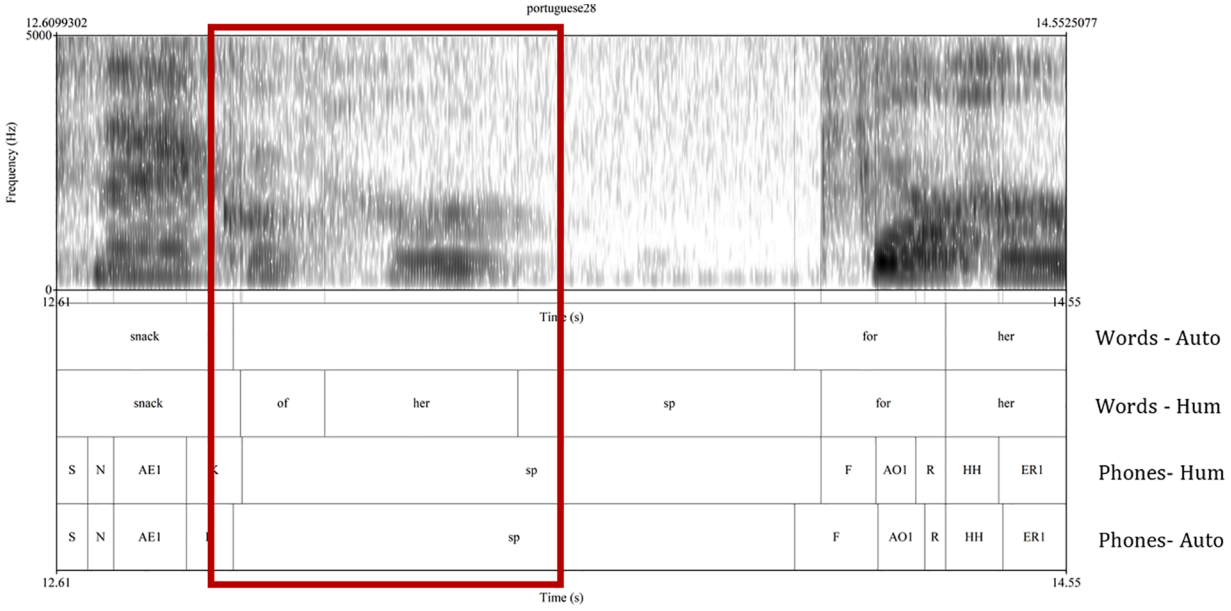


Fig. 12. Example of the aligner correctly ignoring a repetition and incorrect word, marking it as "sp" (silence/non-speech). "*snack* ***of her- for her***" as spoken by speaker Portuguese 28.

When there is an incorrect word spoken (e.g., Arabic 50 "this" instead of "the," shown in Fig. 13), the word-level segmentation is often correct. That is, the boundaries for the word are correct, but the phone alignment within the word is skewed in a manner similar to the type of error it most closely resembles. For example, it may be treated more like phonetic insertion or deletion (Section 4.4.3) depending on the substituted word. For instance, in Fig. 13, the word "the" has been replaced with "this" by the speaker. Consequently, both the number and labels of segments differed from the transcript. Following how the aligner tends to treat insertions, the segment **AH0** /ʌ/ from "the" (on the *Phones-Auto* tier) has been extended to include the additional **S** segment (correctly transcribed on the *Phone-Hum* tier).

4.4.2. Phonetic substitution. Phonetic substitution results in a mismatch with the acoustic model. How the aligner treats phonetic substitution depends both on what a phone is being substituted with, as well as the expected categories of the surrounding segments. This is due to the degree of confusability between consecutive segments. For example, Fig. 14 shows two instances of phonetic substitution for the segments **DH** /ð/ and **TH** /θ/. In this example, both fricatives have been replaced with stops. However, for **DH**, neither of the surrounding segments is a stop, and the realizations of those sounds match the expected realisations resulting in correct alignment of the segment. For **TH**, there is a preceding fricative (**Z** /z/) which the aligner divides in an attempt to correctly align both fricatives. The actual realisation of **TH** by the speaker is treated like an insertion with the 'additional phone' being captured under the **TH** segment, discussed further in the following Section (4.4.3).

4.4.3. Phonetic epenthesis and deletion. Phonetic epenthesis is similarly

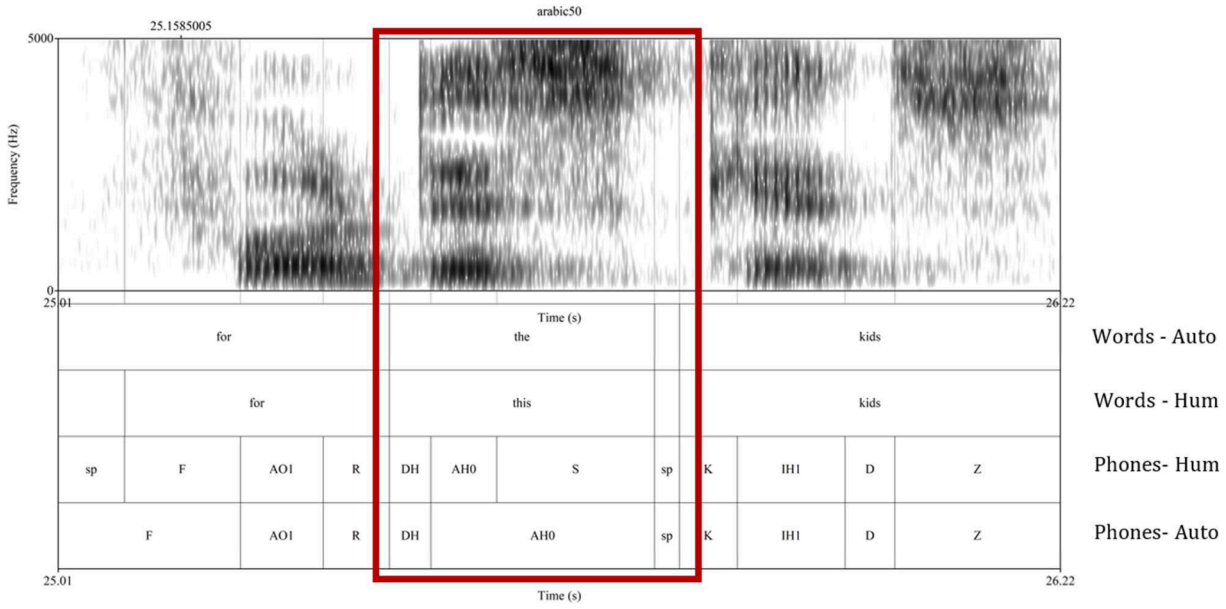


Fig. 13. Example of a speaker using an incorrect word. "*for **this** kids*" instead of "*for **the** kids*" as spoken by speaker Arabic 50.

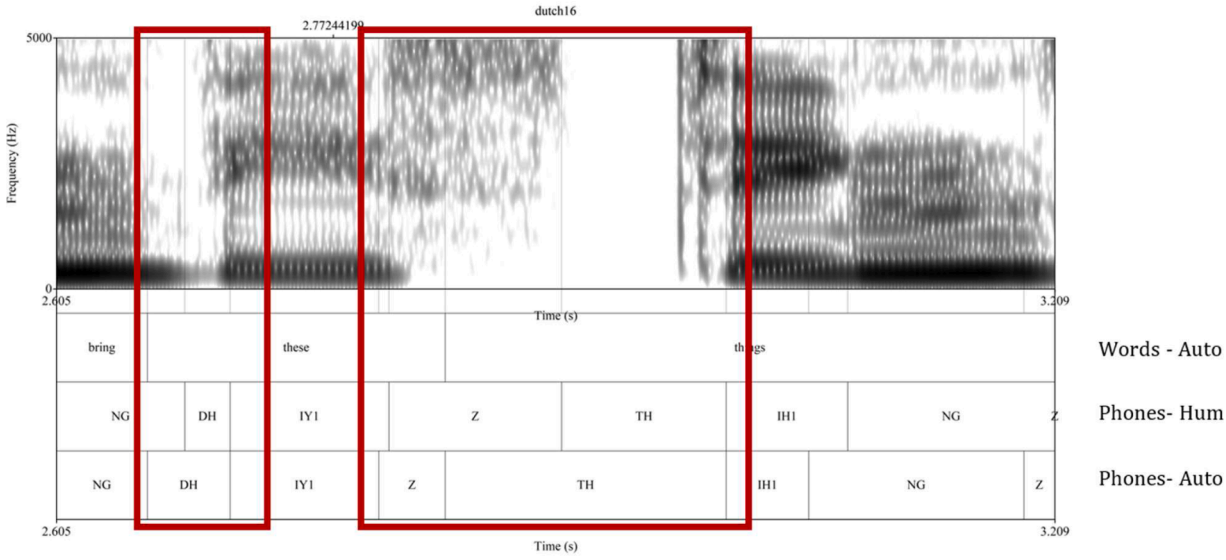


Fig. 14. Two examples of phonetic substitution for the segments **DH** and **TH**. "*bring these things*" as spoken by speaker Dutch 16.

dependent on the categories of the surrounding phones. If, for example, an additional sound is placed at a v-n boundary, and the additional sound has similar spectral characteristics, the aligner will tend to put a boundary at the midpoint of the inserted phone. Alternatively, if the category matches either of the surrounding phone categories, it will be included with the phone it more closely matches. An example of this is seen in Fig. 15. There is vowel epenthesis following the **D** segment in "and" which can be seen in the spectrogram indicated by the red box. In this case, the spectral characteristics of the vowel more closely resemble the expected following vowel segment (**AH0** on the *Phones-Auto* tier) than the preceding stop (**D** on the *Phones-Auto* tier) and is therefore captured under the vowel segment. As can be seen on the corrected *Phones-Hum* tier, the additional vowel was not labelled, but included under the **D** segment (this was a choice of the first author in line with how the data will be used for future work).

When phonetic deletion occurs, the aligner collapses two segments under one label and just catches up with the next correct segment. Therefore, one sound is divided in two to accommodate two segment labels. This happened often with "and" due to reduction; /ænd/ was realised as some variant of /æn/ as described in Section 3.3.

4.4.4. Analysis of boundary displacement errors over 80 ms. Following removal of non-aligner related outliers (as described in Section 3.5), the OBD values were arranged in descending order, and displacements over 80 ms (amounting to 96 tokens) were investigated. This was done to understand the cause of the top 1 % of errors.

The errors fell into four distinct categories, as summarised in Table 7: **(I)** Too much silence following a pause, **(II)** Hesitations and background noise, **(III)** Voice quality, and **(IV)** Variety-specific differences.

In particular, the aligner produced a lot of errors for Russian speakers, providing 32 of the 96 tokens examined (33 % of the tokens).

Occasionally, when there is a pause, the aligner treats vowel-like hesitations or audible breathing as part of the sound. This results in an early boundary placement for word-initial sounds. While the whole

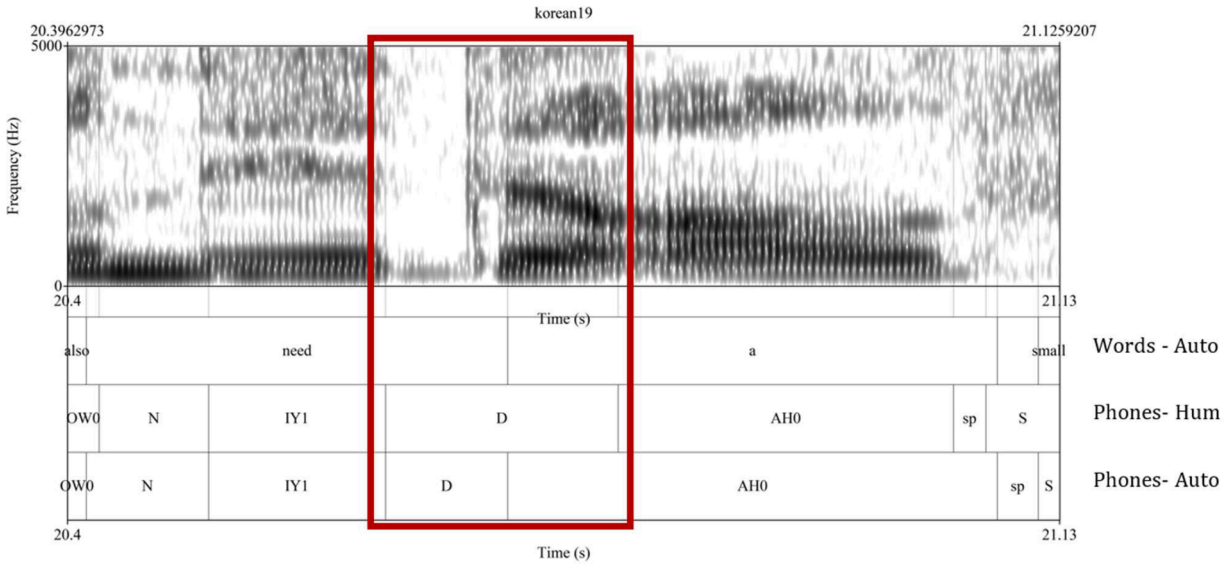


Fig. 15. "*also need a small*" as spoken by speaker Korean 19. Vowel epenthesis following **D** has been combined with the vowel segment **AH0** due to spectral similarity.

Table 7
Classification of errors causing boundary displacements over 80 ms.

	Error Type	Description	Percentage of Errors
(I)	**Too much silence following a pause**	• Tended to be triggered early by background noises	28 %
(II)	**Hesitations and background noise**	• Vowel-like hesitations were sometimes mistaken for segments	14 %
(III)	**Voice quality (speaker-related)**	• Creaky voice and breathy voice tended to cause alignment issues	11 %
(IV)	**Variety-specific differences**	• Vowel was longer than expected or there was a difference in vowel quality • Different realization of sounds e.g., [t] instead of [θ] • Sounds with similar spectral characteristics next to one another (e.g., a dark /l/ next to a back vowel) • Long stop closures	47 % *Almost half from Russian speakers

sound is captured within the labelled segment, the duration is likely to be misrepresented, and thus acoustic measurements of the sound are also potentially unreliable. This is likely caused by the silence model from the LibriSpeech corpus (which the English model is trained on). This model does not include much noise. Therefore, it is expected that noise can heavily impact the accuracy and is more likely to be confused for speech than with a noisier silence model. Silence following a pause accounts for a large percentage of the errors (28 %). Depending on the purpose of the alignment, these could be considered errors or not:

(i) *If location of segments is needed* - early segment boundaries should not be considered an error because the whole sound is captured within the segment.
(ii) *If acoustic measurements of segment are needed* - early segment boundaries could throw off duration measures, typicality measures (if taking the segment as a whole), and mid-point measurements (the midpoint of the segment is not equal to the midpoint of the sound and in some cases may not even include the target sound).

Variety-specific differences account for the largest number of errors (47 %). This type of error is the result of deviation from the acoustic model. This suggests that it will be useful to identify large errors when dealing with a variety that does not match the acoustic model.

5. Discussion

The main purpose of this study was to determine whether the General American English model and dictionary provided with the MFA are sufficient for use with L2 English speech. This was addressed through investigating the aligner accuracy by variety, speaker demographics, segment type and context, and by analysing the types of errors found in the data. A summary of key findings can be found in Table 8.

5.1. Discussion of the results

RQ1 Overall, the results indicate that while the MFA trained on General American English does show variation in performance for both OBD and TOR between varieties, variety does not significantly impact the accuracy. In fact, the aligner performed well across varieties. The results for all measures were comparable to existing studies that have tested an 'off-the-shelf' model on other L1 varieties of English and nativised Englishes, as well as to the L1 English and inter-rater benchmarks. We refer to Table 1 for direct comparison.

The mean BD and OvR in this dataset were lower, and the percentage of boundary placements within 10–20 ms of the manual placement were higher, in comparison to the matched condition (L1 English) in McAuliffe et al. (2017) and McAuliffe (2021a). This could be due to the amount and diversity of boundary data included in our study in addition to a speech style difference. McAuliffe et al. (2017) and McAuliffe (2021a) included only boundaries from 534 CVC words from conversational speech which a) presents a more difficult speech signal than read speech and b) is a mismatch in style to the training data (LibriSpeech; Panayotov et al., 2015) which is based on read speech. The standard deviations of the BD in the present study, however, were generally much higher compared with studies based on matched conditions (L1 English; McAuliffe, 2021a; McAuliffe et al., 2017). Therefore, despite appearing to perform better on average, the aligner performs less consistently on L2 speech. One possible explanation specific to this study could be the lack of verbatim transcription, causing larger displacements due to deviations from the read passage. The slightly better performance than the MacKenzie and Turton (2020) study on non-standard varieties of British English could be due to a closer match in the target English variety (American vs British English) for the L2 speakers despite

Table 8
Summary of findings.

Research Question		Key Findings		
RQ1	Is the performance of the aligner impacted by variety?	*OBD*	$\bar{x}$	4.6 - 14.1 ms
			sd	11.8 - 36.2 ms
			% *<20 ms*	80.0 - 93.0%
			% *<10 ms*	66.8 – 83.9%
		TOR	$\bar{x}$	81.4 - 92.1%
			sd_{TOR}	1.0 - 4.5%
			sd_{OvR}	14.0 - 21.8%
RQ2	Is there any significant effect from other sociophonetic / demographic factors?	*significant*		Speaking rate Age of learning Length of residence Years learning
		not significant		
RQ3	Is the performance of the aligner impacted by manner of articulation?	*segment*		• Stops and fricatives more accurate than nasals and approximants • Similar spectral characteristics next to one another caused issues for boundary placement (e.g., v-a, f-f, v-n) • No skew in direction of displacement
		language		• Performance on segments and segment clusters varied by language (see Fig. 8 and Table 6)
		vowels		• No significant difference in performance between stressed and unstressed vowels • BD *v−consonant* > BD *consonant−v*
RQ4	How are inconsistencies with the transcription, dictionary, and acoustic model treated?			• Many errors are not variety-specific or specific to L2 speech and will be encountered with any difference in the transcription or segmental realisation

expecting similar types of phonetic and phonological deviations.

There was minimal deviation from the group mean TOR by speakers in each language, suggesting little within-group variation. It also suggests that the performance was impacted more by variety differences than speaker differences. This result is not too surprising, as the Kaldi-based acoustic models include speaker-adaptation which normalises the speaker feature space (Chodroff, 2018). Additionally, a review of impressionistic notes and post hoc assessment by the second author and two experienced phoneticians suggests the performance of the aligner may be linked more to fluency and technical quality rather than speaker-specific qualities or features. This is supported by findings from Tu (2018), who asked native speakers of US American English to rate the degree of accentedness for a set of speakers from the same corpus used in this study (30 speakers from each of German, French, Mandarin, and Spanish). The results showed a lack of strong accent ratings for both French and German speakers in the corpus, while the Mandarin speakers were deemed to have the strongest accent overall. Additionally, the French and German speakers were judged to sound more 'native.' These results could partially explain the better performance for German and French speakers and the poorer performance for Mandarin in the present study.

That there was a significant difference in performance between varieties shows that variety does impact how well the aligner performs. Even so, the performance was relatively good on all varieties. The poorer performance on Russian could be explained by the comparatively worse technical quality of the recordings along with more hesitations, word order errors etc. that could in principle be fixed with corrected orthographic transcriptions.

RQ2 Available demographic and sociolinguistic factors including length of residence in an English-speaking country, age the speaker first began to learn English, and speaking rate, were analysed for impact on aligner accuracy. Apart from speaking rate, none of the other features patterned in any meaningful way with aligner accuracy. One explanation could be limited data resulting in not enough speakers for each grouping category. The present study was constrained to prioritize the impact of variety and therefore controlled the data for city of birth as a proxy for regional variety as opposed to any of the other factors that could have been controlled for.

The impact of speaking rate on aligner accuracy contrasts with findings from previous studies on L1 English varieties such as MacKenzie and Turton (2020) and Bailey (2016), which found faster speech rates to negatively impact performance. In this study, while speaking rate was not a significant predictor of accuracy for individual segment boundaries, there was a significant but relatively weak correlation with speaker mean OBD and percent <20 ms metrics which indicated slightly better accuracy and fewer large displacements for faster speaking rates. MacKenzie and Turton (2020) did, however, note one exception with the Westray variety, where FAVE performed relatively poorly despite the slower speaking rate. The variety's extreme phonetic and phonemic deviation from General American English was suggested to be the cause of this phenomenon. For the present study, we suspect the effect of speaking rate is due to a combination of fluency (slower = generally less fluent) (Cucchiarini et al., 2000; Tavakoli and Wright, 2020), less reduction present overall in L2 speech leading to longer duration of unstressed vowels (e.g., Duckinoska, 2021; Laturnus, 2020; Kim and Lee, 2005), and interactions with the L1 (e.g., whether the variety tends to have long or short vowels for example). Taken together this suggests that the aligner performance will decrease for extreme speaking rates in either direction.

RQ3 The performance of the aligner was not only impacted by variety, but also segment category and boundary context within each language. In general, stops and fricatives had the highest OvR with minimal spread, indicating they were not only well-aligned, but consistently so. In contrast, nasals and approximants had less correct overlap and were therefore generally less well aligned, indicating that post hoc corrections of alignment should be focused on these segment types. This corroborates findings from similar studies such as Gonzalez et al. (2020) and MacKenzie and Turton (2020).

The minimal difference in performance on stressed versus unstressed vowels could be due to the nature of L2 speech, i.e., that speakers reduce their vowels less often, or to a lesser degree, (e.g., Duckinoska, 2021) and thereby produce unstressed vowels closer to their full form. This in turn could make it easier for the aligner to identify the correct vowel.

The by-language rankings of BD between segment clusters displayed interesting patterns. While the aligner performed consistently well or badly on some cluster types (e.g., nasal-fricative and vowel-nasal respectively), other clusters ranked differently depending on the language. For example, the aligner consistently performed well on stop-nasal clusters for all varieties except Korean and Mandarin, whereas the aligner consistently had more difficulty with vowel-vowel clusters

except for with Italian and Portuguese (Table 6). It is possible these patterns are caused by how closely matched the realisation of these categories are for a given variety, differences in which segments diverge from the acoustic model, as well as potential impacts from stress patterns in the L1 phonology. This explanation is supported by the findings in Meer (2020), who showed that the vowels specific to Trinidadian English were less well aligned than those shared with the variety the acoustic model was trained on. The inconsistency in performance across languages further demonstrates how variety can impact the performance of the MFA.

Some of the patterns from the overall alignments appear to apply more generally. For example, the findings that vowel-consonant clusters were more difficult for the aligner than consonant-vowel clusters supports findings from McAuliffe et al. (2017) based on American English speech. Specifically for vowel boundary placements, the results from this study support findings from Gonzalez et al. (2020) that vowel-nasal, vowel-vowel, and vowel-approximant clusters are more difficult for the aligner; contexts where boundaries can also be difficult to reliably place when manually transcribing (e.g., Turk and Nakai, 2006; Wesenick and Kipp, 1996).

One solution to combat alignment errors would be to ignore the beginning or end portion of a segment in automatic analyses. We hesitate to recommend favouring one over the other as there was no bias in the direction of the BD (whether boundary placement was consistently early or late) for any of the clusters or varieties. However, removing a portion of the segment prior to acoustic analysis will be less likely to remove important information from vowels, nasals, and fricatives than stops and approximants. Mid-point measures from vowels, nasals and approximants will still have a good chance of being accurate even without correction, except possibly for post pausal position.

RQ4 Many of the types of errors that arise due to variety mismatch such as phonetic substitution, deletion, and epenthesis are treated by the aligner in a systematic way. Knowing this, one can isolate which clusters may be problematic and approach them in a way that best suits the use. This could be by adding dictionary entries, as in Bailey (2016), to ensure more correct labelling, or by making manual corrections to the alignments if duration measurements are important. However, in general the MFA is very quick to catch up with the next correct alignment, especially with speaker errors, therefore not excessively throwing off the alignment of the remaining speech.

The majority of systematic errors leading to large boundary displacements (described in Section 4.4.4) were the result of issues not related to variety mismatch. Therefore, they are to be expected in forced alignments for any variety, potentially explaining why variety mismatch has a relatively low impact on accuracy. These errors include post-pausal excess silence, interference from hesitations and background noise, and issues arising from voice quality. These findings support the conclusions of McAuliffe et al. (2017) and McAuliffe (2021a), who described much larger boundary displacement following a pause than not. Hesitations and background noise being mistaken for speech sounds depends on how silence or non-speech sounds are modelled in the acoustic model. In this case the LibriSpeech silence model is very clean, and therefore will be more likely to confuse background noise as speech than one which includes more background noise and non-speech sounds in their silence model. Creaky voice and breathy voice, while more common in some varieties than others, can happen in almost any speech (Ladefoged and Maddieson, 1996). These particular voice qualities were likely causing issues due to their disruption of the expected spectral characteristics of the phones. Creaky voice, for example, introduces more irregularity in the vibration of the vocal folds, while breathy voice tends to introduce more high-frequency noise into the signal (Ladefoged and Maddieson, 1996).

Based on the types of errors encountered when using a standard General American English model with L2 English speech (Section 4.4), some recommendations have been presented in the following section. This includes recommendations for how to catch or mitigate errors, and when it's best to make adjustments post-hoc.

5.2. Recommendations

5.2.1. Speaker errors. If the forced aligner is going to be used on a large amount of data without manual correction, being able to screen for repetitions or insertions would be useful. Due to the way the aligner treats these types of errors, it follows that a useful measure could be segment duration. In this dataset, there is a mean segment length of 590 ms when there is a speaker error; the smallest duration being 60 ms and the longest 1,790 ms (Table 9). While the smaller errors, often due to hesitations, may be difficult to find, the larger errors would likely be easy to identify.

Based on the average duration for each segment type for each language ('Segment Average' in Table 9; also listed in summary Table A.3), the extended segments are far longer than one would expect for a given segment (other than an extended filled pause, perhaps). This means that duration of segments would be a good indicator of speaker error in the data collection step following automatic alignment. Screening for these errors would hopefully identify repeated phrases or false starts without needing to pre-screen the recordings. However, this might not be sufficient to identify all errors. For example, in one case the segment is only 30 ms below the average segment duration. Ignoring this segment, the shortest duration of a segment with this type of error is 260 ms. In the data set, only 99 segments with corrected alignment are 260 ms or over (1 % of the data).

For vowels and fricatives, 260 ms appears to be a good threshold to check for errors for this dataset, but for the remaining segment categories, the threshold would be better set at a lower duration. Keep in mind some of the values may be longer in duration to account for vowel epenthesis and that some of the segment categories do not match the actual phone, so there will be some variation within a group. The

Table 9
Examples of inserted words or hesitations and the duration of the segment that was affected.

Speaker	Affected Segment		Corrected Duration (ms)	Automatic Duration (ms)	Difference (ms)	Segment Average (ms)	Difference from Segment Average (ms)
	ARPABET	IPA					
Arabic44	**UW0**	/u/	74.6	260.0	185.4	110.0	150
Arabic46	**AH0**	/ʌ/	76.3	1790.0	1713.7	105.4	1684.6
Arabic46	**S**	/s/	114.8	320.0	205.2	115.3	204.7
Korean23	**AH0**	/ʌ/	108.9	600.0	491.1	124.0	476
Korean4	**T**	/t/	29.4	60.0	30.6	89.5	−29.5
Korean4	**AH0**	/ʌ/	104.2	590.0	485.8	124.0	466
Mandarin9	**AH0**	/ʌ/	90.8	360.0	269.2	101.3	258.7
Portuguese28	**ER1**	/ɝ/	204.4	370.0	165.6	115.2	254.8
Portuguese28	**V**	/v/	76.7	650.0	573.3	76.3	573.7
Russian6	**P**	/p/	70.2	280.0	209.8	114.0	166
Russian21	**TH**	/θ/	66.9	510.0	443.1	123.8	386.2

majority of segment categories, however, were correctly assigned.

The exact thresholds set will depend on both speaking rate and average durations for the dataset under investigation. However, the top 1 % of durations, i.e., those approximately >2.3 standard deviations from the mean, for a given segment category should be sufficient to catch any large errors. At this point, the affected segments could either be excluded from analysis or manually corrected. A summary of average segment durations for each of the languages is provided in the Appendix (Table A.3).

5.2.2. Phonetic substitution. If greater/improved accuracy is required in terms of a specific label, or excessive and consistent errors are being caused by substitutions, the only way to mitigate this type of potential error without training a new model is to provide additional dictionary entries. This entails either adding a direct phone match to account for common substitutions or replacing with a phone of the same category of the substitution if there is no equivalent phone in the phone set. Bailey (2016) provides a method for adjusting a large number of dictionary items.

To check for errors post hoc, more familiarity with the variety being studied is required. Potentially problematic sound clusters can be identified based on expected substitutions and their context in the transcript then manually corrected as necessary.

5.2.3. Phonetic epenthesis and deletion. While phonetic epenthesis presents similarly to additional speech (Section 4.4.1), duration will not provide an effective error check. Adjustments for both epenthesis and deletion would be better made at the dictionary stage, or with manual correction following alignment.

5.2.4. Summary. While the overall magnitude of the displacement errors in this study are not large, we would agree with Babinski et al. (2019) and DiCanio et al. (2013) among others, that manual corrections should be made if the required measures need very accurate alignment, such as for duration or stop VOT. Foulkes et al. (2018) further emphasize this message and advise caution when making theoretical claims based on small effects in corpus-based studies. They specifically point to forced alignment as introducing uncertainty into the reliability of acoustic and duration measurements taken from corpora.

It is recommended to have a verbatim transcript prior to phone-level alignment if possible, either manual orthographic transcription or with the help of automatic speech recognition software, as it could resolve many of the errors related to repeated or deleted words/phrases and incorrect order of words, especially if trying to automate analysis as much as possible. However, as Markl (2022) shows, similar issues with L2 speech are also present in the output of automatic speech recognition systems. To account for consistent variation in the speech, a focus on adding dictionary transcriptions would be beneficial, and if there is no matched phone label in the acoustic model, manner of articulation category will likely be sufficient to mitigate many of the expected errors described in Section 4.4.

Knowledge of how the variety under investigation differs from the standard model will help determine which boundary contexts may prove to be particularly difficult for the aligner. This could of course be challenging if there is no existing literature available.

If the alignment is so unreliable that making the above adjustments would not help, adapting the acoustic model to new data or training a new model may be the only solution (McAuliffe et al., 2017).

5.3. Caveats

Our results are based on a small subset of speakers that were selected for each variety, and therefore performance may vary due to additional factors not explored in this study. These include regional accent (many large urban cities were selected; see Table A.4), general articulatory fluency, fluency in English as an L2, and technical quality of the recording. Therefore, the exact ordering or average OBD/TOR should not be taken as a reflection of absolute performance on these varieties. However, given that the averages were either not significantly different, or were very close to the matched training and testing condition, the General American model can reasonably be assumed to perform well on L2 English speech. Controlling for any of the other sociophonetic / demographic factors or determining scores for fluency and strength of accent could provide potential avenues of further study to identify key issues of using a standard acoustic model on L2 speech.

Additionally, we suspect the generalisability of our findings to other forced alignment systems will depend crucially on the phone models (training data) and grapheme-to-phoneme dictionary used.

6. Conclusion

The findings of this study support the conclusion that while the performance of the Montreal Forced Aligner (GMM-HMM architecture) is impacted by variety mismatch, it does not differ drastically from use on a mismatched style (such as conversational speech) or other variety mismatches. The accuracy, while comparable to testing with the standard variety, is less consistent across L2 varieties. However, the majority of systematic errors encountered that led to large boundary displacements were the result of issues not exclusive to variety mismatch such as inaccurate orthographic transcriptions, hesitations, specific voice qualities, and background noise. Whether the specific errors that arise from using a mismatched model are insignificant enough to warrant the use of the standard model depends on the research question being asked and the specific measurements that will be taken. For most, a pretrained, or 'off-the-shelf,' model will work sufficiently with L2 speech, especially if combined with targeted manual corrections.

CRediT authorship contribution statement

Samantha Williams: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft. **Paul Foulkes:** Conceptualization, Supervision, Validation, Writing – review & editing. **Vincent Hughes:** Conceptualization, Formal analysis, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Funding

This work was supported by the University of York YGRS (York Graduate Research School) Overseas Research Scholarship

Appendix A

Passage participants were asked to read for the Speech Accent Archive (Weinberger, 2015):
Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Additional questions speakers were asked to self-report in the Speech Accent Archive:

1. Where were you born?
2. What is your native language?
3. What other languages besides English and your native language do you know?
4. How old are you?
5. How old were you when you first began to study English?
6. How did you learn English? (academically or naturalistically)
7. How long have you lived in an English-speaking country? Which country?

Table A.1
Categorization of ARPABET symbols separated by manner of articulation. The table contains only the sounds that were present in the recordings (according to the 'english_us_arpa' dictionary (Jurafsky and Martin, 2009:251)).

Category	Symbol	Segment Labels (In ARPABET)	IPA Symbol
Plosive	s	**P, B, T, D, K, G**	/p, b, t, d, k, g/
Fricative / Affricate	f	**F, V, TH, DH, CH, S, Z, SH, HH**	/ f, v, θ, ð, ʧ, s, z, ʃ, h /
Nasal	n	**M, N, NG**	/m, n, ŋ /
Approximant (Liquids and Glides)	a	**W, L, R**	/w, l, ɹ /
Vowel (Unstressed)	v (0)	**IY, IH, AH, OW, UW, ER**	See Table A.2
Vowel (Stressed)	v (1)	**IY, IH, EY, EH, AE, AA, AO, AH, OW, UW, AY, OY, ER, (EY2)**	See Table A.2
Pause	p	sp, sil	–

Table A.2
ARPABET symbols for vowels with primary stress along with their approximate IPA equivalent (Jurafsky and Martin, 2009:252).

ARPABET	Approximate IPA Equivalent
IY	/i/
IH	/ɪ/
EY	/eɪ/
EH	/ɛ/
AE	/æ/
AA	/ɑ/
AO	/ɔ/
AH	/ʌ/
OW	/oʊ/
UH	/ʊ/
UW	/u/
AY	/aɪ/
AW	/aʊ/
OY	/ɔɪ/
ER	/ɝ/

Table A.3
Average duration in milliseconds for each segment category for each language. From left to right: Approximant, Fricative, Nasal, Stop, Vowel.

	a	f	n	s	v
Arabic	78.0	121.3	84.2	88.7	129.7
Dutch	63.4	116.7	81.8	85.0	91.0
French	70.0	101.3	70.3	76.9	102.8
German	64.4	102.9	74.3	79.3	94.1
Italian	67.0	110.4	70.9	83.6	102.4

(*continued on next page*)

Table A.3 (*continued*)

	a	f	n	s	v
Korean	79.2	117.7	81.9	88.6	131.5
Mandarin	73.9	120.7	79.7	90.6	113.0
Portuguese	63.0	96.0	70.8	75.3	100.7
Russian	78.1	120.8	90.3	92.0	112.4

Table A.4
Speaker demographic information from Speech Accent Archive (*Weinberger, 2015*).

L1	Birth City, Country	Speaker ID	Age	Sex
Arabic	Jiddah, Saudi Arabia	40	19	M
		44	29	F
		46	28	M
		50	36	M
		52	21	F
Dutch	Antwerp, Belgium	16	23	F
		18	23	M
		23	23	M
		31	23	F
		37	21	F
French	Montreal, Canada	12	19	F
		16	19	F
		26	27	F
		33	62	M
		38	22	M
German	Meissen, Germany	3	19	F
	Halle, Germany	5	47	M
	Berlin, Germany	7	20	M
	Bernburg, Germany	8	25	M
	Elsterwerda, Germany	22	21	F
Italian	Naples, Italy	6	32	F
		7	24	M
		20	32	M
		31	59	F
		36	61	F
Korean	Seoul, South Korea	4	29	F
		7	32	M
		19	21	F
		23	49	F
		24	42	M
Mandarin	Shanghai, China	4	24	F
		9	38	M
		24	21	F
		28	45	M
		66	22	F
Portuguese	Sao Paulo, Brazil	6	44	M
		9	18	M
		28	36	F
		35	22	F
		37	43	F
Russian	Moscow, Russia	6	26	F
		8	66	M
		13	26	M
		21	46	F
		27	21	M

References

Babinski, S., Dockum, R., Craft, J.H., Fergus, A., Goldenberg, D., Bowern, C., 2019. A Robin Hood approach to forced alignment: english-trained algorithms and their use on Australian languages. Proc. Lingu. Soc. Am. 4 (1) https://doi.org/10.3765/plsa.v4i1.4468, 3-1.

Bailey, G., 2016. Automatic detection of sociolinguistic variation using forced alignment. In: University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV 44). York, pp. 10–20. Available at: https://repository.upenn.edu/pwpl/vol22/iss2/3.

Bisani, M., Ney, H., 2008. Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun 50 (5), 434–451. https://doi.org/10.1016/j.specom.2008.01.002.

Brown, G., 2014. Y-ACCDIST: An Automatic Accent Recognition System for Forensic Applications. [MA by research thesis, University of York]. http://etheses.whiterose.ac.uk/7603/.

Brown, G., Franco-Pedroso, J., González-Rodríguez, J., 2021. A segmentally informed solution to automatic accent classification and its advantages to forensic

applications. Int. J. Speech, Lang. Law 28 (2), 201–232. https://doi.org/10.1558/ijsll.20446.
Broselow, E., Chen, S., Wang, C., 1998. The emergence of the unmarked in second language phonology. Stud. Second Lang. Acquis 20 (2), 261–280. https://doi.org/10.1017/S0272263198002071.
Carnegie Mellon University, 1993. CMU Pronouncing Dictionary, 2016. http://www.speech.cs.cmu.edu/cgibin/cmudict.
Chan, M.P.Y., Choe, J., Li, A., Chen, Y., Gao, X., Holliday, N., 2022. Training and typological bias in ASR performance for world Englishes. In: Proceedings of the 23rd Conference of the International Speech Communication Association. https://doi.org/10.21437/Interspeech.2022-10869.
Chodroff, E. (2018). Corpus phonetics tutorial. ArXiv: abs/1811.05553.
Cosi, P., Falavigna, D., Omologo, M., 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In: Proceedings of Eurospeech 1991. Genova, Italy, pp. 693–696, 24–26 September.
Coupé, C., Oh, Y.M., Dediu, D., Pellegrino, F., 2019. Different languages, similar encoding efficiency: comparable information rates across the human communicative niche. Sci. Adv 5 (9), eaaw2594. https://doi.org/10.1126/sciadv.aaw2594.
Cucchiarini, C., Strik, H., Boves, L., 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. J. Acoust. Soc. Am. 107 (2), 989–999. https://doi.org/10.1121/1.428279.
Davidson, L., 2005. Addressing phonological questions with ultrasound. Clin. Linguist. Phon 19 (6–7), 619–633. https://doi.org/10.1080/02699200500114077.
Davidson, L., 2006. Phonology, phonetics, or frequency: influences on the production of non-native sequences. J. Phon 34 (1), 104–137. https://doi.org/10.1016/j.wocn.2005.03.004.
Davidson, L., 2011. Phonetic and phonological factors in the second language production of phonemes and phonotactics. Lang. Linguist Compass 5 (3), 126–139. https://doi.org/10.1111/j.1749-818X.2010.00266.x.
DiCanio, C., Nam, H., Whalen, D.H., Bunnell, H.T., Amith, J.D., García, R.C., 2013. Using automatic alignment to analyze endangered language data: testing the viability of untrained alignment. J. Acoust. Soc. Am. 134 (3), 2235–2246. https://doi.org/10.1121/1.4816491.
Duckinoska, I., 2021. Vowel reduction in english grammatical words by Macedonian EFL learners. Eng. Pronunc. Instruc. 279 https://doi.org/10.1075/aals.19.12duc.
Eberhard, D.M., Simons, G.F., Fennig, C.D. (Eds.), 2022. Ethnologue: Languages of the World. Twenty-fourth edition Dallas. Texas: SIL International. Online version: https://www.ethnologue.com/guides/most-spoken-languages.
Ferragne, E., Gendrot, C., Pellegrini, T., 2019. Towards phonetic interpretability in deep learning applied to voice comparison. In: ICPhS, ISBN-978 halshs.archives-ouvertes.fr.
Flege, J.E., Munro, M.J., MacKay, I.R., 1995. Factors affecting strength of perceived foreign accent in a second language. J. Acoust. Soc. Am. 97 (5), 3125–3134. https://doi.org/10.1121/1.413041.
Flege, J.E., Bohn, O.S., 1989. An instrumental study of vowel reduction and stress placement in Spanish-accented English. Stud. Second Lang. Acquis 11 (1), 35–62. https://doi.org/10.1017/S0272263100007828.
Foulkes, P., Docherty, G., Hufnagel, S.S., Hughes, V., 2018. Three steps forward for predictability. Consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory. Linguis. Vanguard 4 (s2). https://doi.org/10.1515/lingvan-2017-0032.
Fromont, R.A., Hay, J., 2012. LaBB-CAT: an annotation store. In: Proceedings of the Australasian Language Technology Workshop, pp. 113–117. Available at: http://hdl.handle.net/10092/15624.
Fromont, R., Watson, K., 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. Corpora 11 (3), 401–431. https://doi.org/10.3366/cor.2016.0101.
Gonzalez, S., Grama, J., Travis, C.E., 2020. Comparing the performance of forced aligners used in sociophonetic research. Lingu. Vanguard 6 (1). https://doi.org/10.1515/lingvan-2019-0058.
Gorman, K., 2016. Pynini: a Python library for weighted finite-state grammar compilation. In: Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata, pp. 75–80.
Hancin-Bhatt, B., Bhatt, R.M., 1997. Optimal L2 Syllables: interactions of Transfer and Developmental Effects. Stud. Second Lang. Acquis 19 (3), 331–378. https://www.jstor.org/stable/44487972.
Harwell, D., Mayes, B., Walls, M., Hashemi, S., 2018. The accent gap. The Washington Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.
Huckvale, M., 2004. ACCDIST: a metric for comparing speakers' accents. In: Eighth International Conference on Spoken Language Processing. Available at. https://discovery.ucl.ac.uk/id/eprint/12139.
Hutiri, W.T., Ding, A.Y., 2022. Bias in automated speaker recognition. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 230–247. https://doi.org/10.1145/3531146.3533089.
Jurafsky, D., Martin, J.H., 2009. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J., 2nd ed. Pearson Prentice Hall, London.
Kim, J.M., Lee, O.H., 2005. Reduced vowel quality accounts for Korean accent of English. Stud. Engl. Lang. Literature 31, 73–93.
Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S., 2020. Racial disparities in automated speech recognition. Proc. National Acad. Sci. 117 (14), 7684–7689. https://doi.org/10.1073/pnas.1915768117.
Kvale, K. (1993). Segmentation and labelling of speech [PhD thesis, Norwegian institute of technology]. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2368838.
Ladefoged, P., Maddieson, I., 1996. The Sounds of the World's Languages. Blackwell Publishers, Oxford.
Laturnus, R., 2020. Comparative acoustic analyses of L2 english: the search for systematic variation. Phonetica 77 (6), 441–479. https://doi.org/10.1159/000508387.
Little, D., 1995. Learning as dialogue: the dependence of learner autonomy on teacher autonomy. System 23 (2), 175–181. https://doi.org/10.1016/0346-251X(95)00006-6.
Lo, J.H.J., Wong, S.G., 2024. Multilingualism and code-switching. In: Nolan, F., McDougall, K., Hudson, T. (Eds.), The Oxford Handbook of Forensic Phonetics. Oxford University Press.
MacKenzie, L., Turton, D., 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. Linguistics Vanguard 6 (s1). https://doi.org/10.1515/lingvan-2018-0061.
Markl, N., 2022. Language variation and algorithmic bias: understanding algorithmic bias in British english automatic speech recognition. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 521–534. https://doi.org/10.1145/3531146.3533117.
McAuliffe, M., 2021a. Update on montreal forced aligner performance. Memcauliffe. https://memcauliffe.com/update-on-montreal-forced-aligner-performance.html.
McAuliffe, M., 2021b. How much data do you need for a good MFA alignment? Memcauliffe. https://memcauliffe.com/how-much-data-do-you-need-for-a-good-mfa-alignment.html.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M., 2017. Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Proceedings of the 18th Conference of the International Speech Communication Association, pp. 498–502. https://doi.org/10.21437/Interspeech.2017-1386.
Meer, P., 2020. Automatic alignment for new Englishes: applying state-of-the-art aligners to Trinidadian English. J. Acoust. Soc. Am. 147 (4), 2283–2294. https://doi.org/10.1121/10.0001069.
Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964.
Paulo, S., Oliveira, L.C., 2004. Automatic phonetic alignment and its confidence measures. In: International Conference on Natural Language Processing (in Spain). Springer, Berlin, Heidelberg, pp. 36–44. https://doi.org/10.1007/978-3-540-30228-54.
Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E., 2007. Buckeye Corpus of Conversational Speech. Department of Psychology, Ohio State University (Distributor, Columbus, OH, 2nd releaseAvailable at: www.buckeyecorpus.osu.edu.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K., 2011. The Kaldi speech recognition toolkit. IEEE 2011 Workshop On Automatic Speech Recognition and Understanding (No. CONF). IEEE Signal Processing Society.
Raymond, W.D., Pitt, M.A., Johnson, K., Hilts, C., 2002. An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. In: Proceedings of 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002. Denver, Colorado, USA. September 16-20, 2002. Available at: https://www.isca-speech.org/archive_v0/archive_papers/icslp_2002/i02_1125.pdf.
Reddy, S., Stanford, J.N., 2015. Toward completely automated vowel extraction: introducing DARLA. Linguistics Vanguard 1 (1), 15–28. https://doi.org/10.1515/lingvan-2015-0002.
Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., Yuan, J., 2014. FAVE (forced alignment and Vowel extraction) Suite Version 1.1.3. Software. https://doi.org/10.5281/zenodo.9846.
Schiel, F., 1999. Automatic phonetic transcription of non-prompted speech. In: Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS). San Francisco, pp. 607–610.
Shi, X., Yu, F., Lu, Y., Liang, Y., Feng, Q., Wang, D., Qian, Y., Xie, L.. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. https://www.datatang.ai/shujutang/static/file/AESRC2020.pdf.
Tavakoli, P., Wright, C., 2020. Second Language Speech fluency: From research to Practice. Cambridge University Press. https://doi.org/10.1017/9781108589109.
Tu, M., 2018. A Computational Model For Studying L1's Effect on L2 Speech Learning (Doctoral Dissertation). Arizona State University.
Turk, A., Nakai, S., Sugahara, M., 2006. Acoustic segment durations in prosodic research: a practical guide. Methods Empir. Pros. Res. 3, 1–28. https://doi.org/10.1515/9783110914641.
Wade, T., Jongman, A., Sereno, J., 2007. Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. Phonetica 64 (2–3), 122–144. https://doi.org/10.1159/000107913.

Wang, C., 1995. The acquisition of English word-final obstruents by Chinese speakers. (Unpublished doctoral dissertation). State University of New York at Stony, Brook, NY.

Weinberger, S., 2015. Speech Accent Archive. George Mason University. Retrieved from. http://accent.gmu.edu.

Wesenick, M.B., Kipp, A., 1996. Estimating the quality of phonetic transcriptions and segmentations of speech signals. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, 1. IEEE, pp. 129–132. https://doi.org/10.1109/ICSLP.1996.607054.

Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P.R., Clark, L., Cowan, 2020. See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–9. https://doi.org/10.48550/arXiv.2006.06328.

Further reading

Skirgård, H., Roberts, S.G., Yencken, L., 2017. Why are some languages confused for others? Investigating data from the great language game. PLoS ONE 12 (4), e0165934. https://doi.org/10.1371/journal.pone.0165934.