Maximizing accuracy of forced alignment for spontaneous child speech

Robert Fromont,
Lynn Clark,
Joshua Wilson Black,
Margaret Blackwood

University of Canterbury, Christchurch, Aotearoa New Zealand

Abstract: Sociophonetic study of large speech corpora generally requires the use of forced alignment — the automatic process of determining the start and end time of each speech sound within the recording — in order to facilitate large-scale automated extraction of acoustic measurements of targeted vowels or consonants. There is an extensive literature evaluating alignment accuracy of a number of forced alignment tools and procedures, processing speech data from a range of languages and dialects. In general, these evaluations use typical adult speech data, often elicited in a controlled laboratory environment. There is little literature on the effectiveness of forced alignment systems on child speech, and none on speech elicited in field environments. This presents a problem for research at the intersection of language acquisition and sociophonetics as there is no established best practice for automatically aligning child speech. Child speech presents special challenges to automated tools, as it includes more variation in speech sounds and voice quality, and non-standard pronunciation and prosody. We evaluated three commonly used forced aligners, the Montreal Forced Aligner (MFA), the Hidden Markov Model Toolkit (HTK) integration provided by the LaBB-CAT corpus analysis tool, and the Penn Aligner (P2FA), using different configurations to force align non-rhotic child speech elicited in a preschool environment. Against many of our expectations, we found that volume of training data trumps similarity to the speech; MFA, using rhotic acoustic models pre-trained on adult speech, performed best. This paper provides a clear methodology for other researchers in sociophonetics to evaluate the success or otherwise of phonetic alignment.

Keywords: child speech; language acquisition; sociophonetics; speech corpora; forced alignment

Corresponding author: Robert Fromont, New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, NZ. Email: robert.fromont@canterbury.ac.nz.

ORCID IDs: https://orcid.org/0000-0001-5271-5487; https://orcid.org/0000-0003-3282-6555; https://orcid.org/0000-0002-8272-5763

Citation: Fromont, R., Clark, L., Wilson Black, J., & Blackwood, M. (2023). Maximizing accuracy of forced alignment for spontaneous child speech. *Language Development Research, 3*(1), 182–210. https://doi.org/10.34842/shrr-sv10

Introduction

There has been a progressive development in the collection and use of large digitised speech corpora containing hundreds of hours of spontaneous speech in sociophonetic research, e.g. the Origins of New Zealand English (ONZE) corpus containing over 3 million words (Gordon et al. (2007)), or the Spoken BNC2014 corpus over 11 million words (Love et al. (2017)). Such corpora are not amenable to painstaking manual alignment to the phone level, which can take 800 times longer than the duration of the speech (Schiel et al. (2012), Section 8.5.1, p. 111, footnote 11). 'Forced alignment', the automated process of locating the start and end times of speech sounds within speech recordings, has been described as 'transformative' by Coto-Solano (2022) (p. 2) allowing the large-scale extraction and study of segments from such corpora.

Although automatically generated alignments of extracted speech sound tokens can be manually checked and adjusted for accuracy, as the number of tokens extracted increases, the practicality of manually checking each and every one decreases. Developing highly accurate tools and procedures for forced alignment is critical, and there is a decades-long literature evaluating different systems and techniques when applied to adult speech. Current best practice in sociophonetics research on adult talkers combines methods which use the most accurate forced alignment configuration, together with procedures for automatically weeding out erroneous tokens after extraction. This method can result in the loss of incredible amounts of data (e.g. Brand et al. (2021) report losing 80% of their data during the filtering process) and yet it still allows measurement and analysis of hundreds of thousands of tokens.[1] Maximising the accuracy of automatic alignment is crucial to minimising such exclusion of data.

Although the literature is well established for typical adult speech, very little work has been done to establish best practices for accurate alignment of child speech. During language development, speech includes more variation in pronunciation (Lee et al. (1999), Assmann & Katz (2000)), duration (Smith (1992), Lee et al. (1999)), and prosody (Athanasopoulou & Vogel (2016)), which can be a challenge for automatic tools that are calibrated for typical adult speech.

After reviewing the current literature on forced alignment of adult and child speech, we describe our own child spontaneous speech corpus, present experiments we ran to determine the most accurate procedure for force aligning our data with three commonly used forced aligners, and the methods we used to measure accuracy. Finally, we present the results of these experiments, and discuss the implications of those results.

[1] In addition to Brand et al. (2021), see recent work by Stuart-Smith et al. (2019). A comprehensive survey of forced alignment used for sociophonetic research is provided by Coto-Solano (2022) (Section 6).

Forced Alignment Tools and Procedures

Since the 1990's a number of computational techniques have been applied to the problem of forced alignment, including Dynamic Time Warping (DTW; Cosi et al. (1991), Coleman (2005)), Hidden Markov Models (HMMs; Young et al. (2006)), and Deep Neural Networks (DNNs; Hawkins et al. (2017)). Forced alignment procedures have sometimes included post-alignment error correction by modelling errors based on a small number of manual alignments (Toledano & Gómez (2002), Adell et al. (2005)).

Most current forced aligners commonly used for phonetics research use one of two HMM-based Automatic Speech Recognition (ASR) software toolkits: the HMM Tool Kit (HTK; Young et al. (2006)) and Kaldi (Povey et al. (2011)).

Although the ASR toolkits themselves support a wide array of options for preparing, processing, and aligning speech data, the forced aligners that have been developed to simplify and automate parts of this process for phonetics generally employ a two-phase process.

Phase one requires three ingredients:

1. a collection of speech recordings,
2. corresponding orthographic transcripts with start and end times of utterances, and
3. a mapping of orthographic spelling to pronunciation using some set of phoneme symbols (usually a pronunciation dictionary).

Hidden Markov Model Gaussian Mixture Models (HMM-GMMs) are trained using the toolkit, which uses Mel Frequency Cepstral Coefficients (MFCC) computed from the audio signal[2], producing a set of acoustic models, either one for each phoneme symbol (monophone models) or one for each distinct cluster of three phonemes (triphone models).

Phase two requires four ingredients:

1. a collection of recordings,
2. corresponding orthographic transcripts,
3. a mapping of orthographic spelling to pronunciation using the same set of phoneme symbols used during phase one, and
4. the acoustic models trained during phase one.

Phase two involves using acoustic models from phase one, either as-is or adapted for each speaker, to align the word pronunciations with the audio, output being a set of

[2]Gaussian Mixture Models (GMMs), are used to model the distribution of the coefficients

start and end times for the words and corresponding phones found in the recordings.

Aligners that use 'pre-trained models' are those where the recordings and transcripts used in phase one are different from those used in phase two. Conversely aligners that use a 'train/align' procedure are those where the same recordings/transcripts are used in *both* phases.

If the recordings in phase one are all from the same speaker, then the models are *speaker-specific*, otherwise they are *speaker-independent*, although some aligners support adapting speaker-independent models to individual speakers during phase two. We refer to the former as *speaker-adapted* models and the latter as *unadapted*.

HTK and Kaldi

HTK and Kaldi are both toolkits for developing ASR systems. They both use HMMs (although Kaldi supports using DNNs instead) and can both be used for training monophone or triphone models.

HTK, developed from 1989 to 2016 by Cambridge University Engineering Department (CUED), is older than Kaldi. Kaldi has been in development since 2009 at Johns Hopkins University, using more 'modern and flexible code' than HTK[3]. While the source code for both toolkits is available, the HTK license requires users to register. Kaldi is released with the Apache License v2.0 licence, and is fully open source.

Current Forced Alignment Systems

Forced alignment systems currently used in sociophonetic research each use their own combination of toolkits, models, and procedures. Widely used systems include:

- Penn Phonetics Lab Forced Aligner (P2FA; Yuan & Liberman (2008)), which uses monophone HTK models pre-trained on American English speech;
- Munich AUtomatic Segmentation (MAUS; Schiel (1999), Schiel (2015)), an HTK-based system with pre-trained models for a wide variety of languages, also available via BAS Web Services[4](Kisler et al. (2017));
- Prosodylab Aligner (Gorman et al. (2011)), an HTK-based system that allows for training of new acoustic models;
- Montreal Forced Aligner (MFA; McAuliffe et al. (2017)), the successor of Prosody-Lab Aligner[5], built on Kaldi's HMM capabilities, including speaker-adapted models, and supporting both pre-trained triphone models (acoustic models and pro-

[3] https://www.kaldi-asr.org/doc/about.html
[4] https://www.bas.uni-muenchen.de/Bas/BasMAUS.html
[5] McAuliffe et al. (2017) p. 1.

nunciation dictionaries for a wide range of languages and varieties are available) and also a train/align mode of operation;
- LaBB-CAT (Fromont & Hay (2012)), a speech corpus management system that integrates with HTK, P2FA, MFA and BAS Web Services, supporting both pre-trained and train/align procedures; and
- Gentle (Hawkins et al. (2017)), which like MFA is built on Kaldi, but unlike MFA, uses DNNs instead of HMMs[6], and supports English only.

Evaluations on Adult Speech

Over the last three decades, the accuracy of many forced alignment tools and configurations has been evaluated using adult speech.

Factors considered in these evaluations include: which tool set is used (Chen et al. (2004), Adell et al. (2005), Niekerk & Barnard (2009), DiCanio et al. (2013), McAuliffe et al. (2017), Meer (2020)), the amount of data used for training (Toledano & Gómez (2002), Chen et al. (2004), Brognaux et al. (2012), Fromont & Watson (2016)), speech style (e.g. read vs. spontaneous) (Chen et al. (2004), Fromont & Watson (2016)), whether monophone or triphone models are used (Toledano & Gómez (2002), Brognaux et al. (2012), McAuliffe et al. (2017)), using pre-trained models or the train/align procedure (Niekerk & Barnard (2009), Brognaux et al. (2012), Fromont & Watson (2016), McAuliffe et al. (2017), Gonzalez, Grama, et al. (2018)), using speaker-independent or speaker-specific models (Toledano & Gómez (2002), Niekerk & Barnard (2009), Brognaux et al. (2012)), how finely chunked the speech is (Chen et al. (2004)), applying automated post-alignment corrections based on a manual aligned sample (Toledano & Gómez (2002), Adell et al. (2005)), or by force-aligning data recursively, adding more data for each new cycle (Moreno et al. (1998), Gonzalez, Grama, et al. (2018)). The literature includes data from different languages (e.g. Afrikaans, English, French, isiZulu, Matukar Panau, Russian, Setswana, Spanish) and language varieties (e.g. American, Australian, Blackburn, Hastings, Liverpool, Manchester, New Zealand, Sunderland, and Westray English), including cases where models were pre-trained on a different language (Niekerk & Barnard (2009), DiCanio et al. (2013), Babinski et al. (2019), Tang & Bennett (2019)) or variety (Fromont & Watson (2016), MacKenzie & Turton (2020)) from the speech being aligned.

Various metrics have been used for comparing manual alignments with automatic ones, including comparing aggregate acoustic measurements (pitch peak, vowel space, and mean duration) resulting from automatic and manual alignments (Babinski et al. (2019)), error thresholds for absolute differences in boundaries (Cosi et al. (1991), Toledano & Gómez (2002), DiCanio et al. (2013), McAuliffe et al. (2017), Tang & Bennett (2019), Meer (2020), Gonzalez, Grama, et al. (2018), Gnevsheva et al. (2020)) or interval

[6]Early versions of MFA included the possibility of using DNNs, but MFA version 2.0 does not

mid-points (Gonzalez, Travis, et al. (2018)), mean / median differences between boundaries (Chen et al. (2004), Gorman et al. (2011), McAuliffe et al. (2017), Gonzalez, Grama, et al. (2018), Tang & Bennett (2019), Meer (2020), Gonzalez et al. (2020)), and the 'Overlap Rate' — the proportional degree of overlap of intervals (Niekerk & Barnard (2009), Fromont & Watson (2016), Gonzalez, Travis, et al. (2018), Gonzalez et al. (2020)).

General conclusions from the literature are that the finer the data is chunked the better and that speaker-specific models are more accurate than speaker-independent models, as are models trained on more data. A mismatch in speech style between the training and alignment data leads to lower accuracy, and using a sample of manual alignments to model post-alignment corrections also boosts accuracy. There is conflicting evidence about whether monophone or triphone models are more accurate. HMM-based systems represent the current state of the art, with a recent preference towards Kaldi-based MFA rather than older HTK-based ones (Gonzalez, Grama, et al. (2018), Gonzalez et al. (2020)).

Evaluations on Child Speech

Work on forced alignment has skewed towards 'high resource' data, i.e. 'mainstream' languages such as English, and high-status varieties of those languages, such as US English. This skew also has a demographic dimension. Development and evaluation of forced alignment tends to use readily available non-pathological adult speech.

However other types of speech also warrant sociophonetic research; child speech has special challenges not usually present in most adult speech. As children are still in the process of developing their language faculties, they show more variability in their phonology, volume, and articulation. The authors have also found unusual prosodic phenomena such as mid-word pauses in our own data (Fromont et al. (2022)).

Alignment accuracy with child speech has only recently received any attention from researchers. Knowles et al. (2018), Mahr et al. (2021), and Szalay et al. (2022) have performed some evaluations which we now describe. Knowles et al. (2018) investigated the effect of various factors on the accuracy of forced alignment of child speech, using a specific forced alignment tool, ProsodyLab-Aligner. They used two corpora of child speech: one comprising 2 hours of spontaneous speech by a single Canadian English speaking child at different ages (1;5 - 3;6), and another including 5 hours of single-word controlled speech by 40 girls and 41 boys aged between two and six years, speaking US English recorded in a laboratory.

Using the attributes of the corpora themselves, they examined the effects of speech style and speaker age. They also compared alignments produced using different types of training data: adult speech only, adult and child speech, and child speech only, training both speaker-independent models and speaker-specific models. In addition they

compared the use of two different dictionaries: a 'standard' dictionary (the CMU Pronouncing Dictionary, Rudnicky & Weide (2014)), and a dictionary manually customised to match the child's speech.

They concluded that controlled speech had more accurate alignment than spontaneous speech,[7] the speech of older children was more accurately aligned, child-only models performed better, and the customised dictionary, which more closely matched the child's actual speech, performed better than a 'standard' dictionary. Vowels and sibilants were best aligned. The best accuracies produced, using their midpoint overlap metric (see below), were 75%-90%.

Mahr et al. (2021) compared different forced aligners - MFA, Kaldi with triphone models, Prosodylab Aligner, and P2FA - using a corpus of 42 US English speaking children aged between 3 and 6 years, recorded in a laboratory. Unlike Knowles et al. (2018), the utterances were generally sentences (up to 60 per participant) rather than single words[8], but were still highly controlled. They found that MFA using models pre-trained on adult speech produced the best alignments, with 86% accuracy (using midpoint overlap). Again, vowels were the best aligned segments.

Szalay et al. (2022) have also evaluated forced aligners on child speech, comparing the MAUS HTK-based aligner with three custom aligners trained using Kaldi's DNN functionality, rather than using HMMs. Their test data were 153 single words elicited from 11 Australian English (AusE) speaking children (7 boys and 4 girls) aged between 4;10 and 11;11. Their custom aligners differed by training data; one was trained on AusE speaking adults, another was trained on speech by similar aged children speaking a different dialect – American English (AmE) – and the third was trained on a mixture of adult AusE and child AmE speech.

They found that the custom aligners trained on adult AusE training data, and the aligner that combined this with AmE child data, had similar high comparative accuracy – with 65% and 66% boundaries within 20ms of the manual boundary, and mean Overlap Rate of 0.74 and 0.73, respectively – better than the aligner that used AmE child speech alone, with 46% accuracy and 0.71 mean Overlap Rate, and MAUS with 59% accuracy and 0.69 mean Overlap Rate. They conclude that matching dialect is more important than matching age.

Our Data

We have a growing corpus of New Zealand children performing an oral language assessment task at their pre-school. Each child heard a story and was asked to re-tell it.

[7] This may have been an caused by the single-word utterances being more finely chunked than the spontaneous utterances

[8] Mahr et al. did not report the total duration of their recordings.

The initial corpus for forced alignment included 38 children (21 boys, 17 girls) aged 3;6 - 4;11.

The literature on adult and child forced alignment would appear to offer clear guidelines for aligning a corpus of child speech.

- The more similar the training and alignment speech, the better; the speaker's own speech is best (i.e. speaker-specific models) but if not, speaker-independent models trained on similar speech work better.
- The closer the dictionary is to the actual pronunciations, the better; a dictionary for the same language variety (with the same phoneme inventory, rhoticity etc.) should be preferred.
- The more training data, the better.

However our initial attempts to force align the speech using LaBB-CAT's default HTK-based training of speaker-specific models and a non-rhotic dictionary suitable for New Zealand English (NZE) produced poor results. We suspected that this kind of corpus falls within a gap in the forced alignment literature.

Although the literature is clear that speaker-specific models are preferable, it is also necessary to have *enough* training data to produce reliable models. Fromont & Watson (2016) found that, for NZE, at least five minutes of speech is required for each speaker for the Overlap Rate to plateau between 0.5 and 0.6[9]. The most verbose child in our corpus spoke for slightly less than three minutes, and many spoke much less than this; the least amount of speech for a single child was sixteen seconds.

We considered using speaker-independent models, either by grouping children in our corpus together in order to train on more than five minutes of speech; our corpus contains 29 minutes child speech, or 46 minutes including adult examiner speech. Or we could use pre-trained models, which are trained on much more data than our corpus contains. However, most models available for English are pre-trained on adult US English speech, which we suspected would be too different from the speech in our corpus.

Almost all of the data used for evaluation in the child speech forced alignment literature was controlled speech; short predictable sentences, and often single words, elicited in a sound-attenuating laboratory environment. But our corpus is spontaneous speech, and is field data recorded in environments with background noise. In many cases the speech is low volume or the child is whispering. The literature appears to have no recommendation for these circumstances; Mahr et al. (2021) are clear about

[9] Overlap Rate is a value between 0 meaning no overlap at all, and 1 meaning perfect overlap; see the section called Overlap Rate for details.

this: "we are hesitant to extrapolate beyond elicited laboratory speech."[10] Furthermore, in some cases the speech is articulated in a manner that's so divergent from adult norms, that even the correct transcription is debatable.

Faced with many doubts about how to proceed, we performed a number of experiments in order to determine 1) which tool/procedure would yield the most accurate alignments, and 2) how the resulting accuracy measured up against accuracies reported in the literature. We expected some configuration involving a non-rhotic dictionary and training on some mix of the children's own speech to result in the most accurate alignments, but that the best accuracy would still be lower than in other studies, due to the age of the speakers and the spontaneous nature of the utterances.

Methods

We compared three commonly used HMM-based aligners, LaBB-CAT's HTK forced-alignment, P2FA (also built on HTK), and MFA (built on Kaldi), and different alignment procedures using those tools:

- train/align with speaker-specific models
- train/align with speaker-independent models
- pre-trained models using a pronunciation dictionary matching our non-rhotic NZE data
- widely-used pre-trained models using a rhotic pronunciation dictionary

In order to easily and reproducibly automate specific configurations, we used LaBB-CAT, which integrates with all three aligners[11], and includes the `nzilbb.labbcat` R package[12], allowing the implementation of an R script to precisely specify forced alignment configurations, and run forced alignment on different subsets of the corpus.

We used ten different forced alignment configurations, which are all easily configurable options with the chosen forced aligners, requiring the minimum manual intervention. The train/align configurations generally use the default options for the given forced aligner (except where otherwise noted), and the pre-trained model configurations use models and dictionaries that are readily available. They represent options that were not only convenient for us to set up quickly for our own LaBB-CAT-based corpus, but also would be easily configured for other sociophonetic research with similar data, either via LaBB-CAT, or in the case of MFA and P2FA, independently of LaBB-CAT

[10] Mahr et al. (2021), p. 2221.

[11] Although LaBB-CAT integrates with BAS Web Services, we could not try MAUS for forced alignment, because our data cannot be shared with a third party

[12] Fromont (2023)

by using the command line interfaces of those forced aligners. The configurations are compared in Table 1. We describe them in detail now.

LaBB-CAT-HTK configurations

The configurations we refer to as 'LaBB-CAT-HTK' use LaBB-CAT's direct integration with the HTK toolkit, which automates the eight steps for training acoustic models with HTK laid out by Young et al. (2006) in Chapter 3 of 'The HTK Book'.

For all train/align configurations using LaBB-CAT-HTK, the same pronunciation dictionary was used: the CELEX English lexicon (Baayen et al. (1995)), a non-rhotic lexicon based on 'British English', supplemented to include words not present in the original lexicon, including non-standard child wordforms such as "comed", "goed", "runned", etc. Phonemic transcriptions are encoded using CELEX's 'DISC' phoneme symbols[13].

The initial base-line configuration was for speaker-specific models; each child's speech was aligned using models trained only on their own speech (*Speaker specific* in Table 1). We also specified three speaker-independent configurations which grouped speakers together for the training phase in groups of increasing size and decreasing speaker similarity. Firstly, speakers were grouped by gender; each child's speech was aligned using models trained on speech of children of the same gender (*Gender specific* in Table 1). Secondly, one set of speaker-independent models were trained using the speech of all children together (*Child independent* in Table 1). Thirdly, one set of speaker-independent models were trained using the speech of all children and also adults in the corpus (*Speaker independent* in Table 1). All speaker-independent models were trained on more than five minutes of speech.

P2FA

The final HTK-based configuration uses the P2FA pre-trained models (*P2FA* in Table 1) in order to compare accuracy of the LaBB-CAT-HTK train/align configurations above with this commonly-used aligner. These models use ARPAbet phoneme symbols[14] that are different from those used by CELEX, and are trained on rhotic US English adult speech[15]. As a result, this configuration used a supplemented version of the CMU Pronouncing Dictionary (CMUdict).[16]

[13] Appendix A includes a table showing how these symbols relate to other symbol sets, and they are described in section 2.4.1 of the CELEX English manual included with Baayen et al. (1995)

[14] See Appendix A.

[15] The P2FA models were trained on 25.5 hours of speech by adult American English speakers, specifically speech of eight Supreme Court Justices selected from oral arguments in the Supreme Court of the United States (SCOTUS) corpus (Yuan & Liberman (2008)).

[16] See Rudnicky & Weide (2014)

MFA configurations

By default MFA uses a train/align procedure that first trains speaker-independent models using all speech, and then adapts these models to each speaker, so that the final alignments use speaker-specific models. Our first MFA configuration used this procedure, using the same CELEX pronunciation dictionary as used by the LaBB-CAT-HTK configurations (*Speaker adapted* in Table 1).

MFA also supports using a variable number of HMM states; each model uses fewer or more states depending on what type of phoneme is being modelled (e.g. fewer states for certain stops, but more for diphthongs). In order to achieve this, MFA requires the phonemic transcriptions to use a specific set of IPA symbols, so we used a supplemented dictionary based on a non-rhotic 'British English' dictionary supplied by MFA[17] (*Variable state* in Table 1).

MFA provides different sets of pre-trained models, so our final three configurations used pre-trained models and corresponding dictionaries. The first two configurations use 'General American English' models using a rhotic dictionary encoded with the same ARPAbet symbols as used by P2FA[18]. The first configuration uses the models 'as-is', without adapting the models to each speaker before alignment (*GAM Unadapted* in Table 1), and the second includes the speaker adaptation step (*GAM Speaker adapted* in Table 1) in order to be able to determine how much difference the speaker adaptation of the models might make with our child speech data. The last configuration uses models trained on different varieties of English using a non-rhotic 'UK English' dictionary encoded using IPA (*UK Speaker adapted* in Table 1)[19]. This final configuration includes much more training data, including non-rhotic (as well as rhotic) varieties of English, and a non-rhotic dictionary, so we suspected it might provide more accurate alignments for our non-rhotic NZE speech than the GAM-based configurations above. Because the dictionary is non-rhotic, as is much of the training data, it is marked as such in Table 1.

[17] See https://mfa-models.readthedocs.io/en/latest/dictionary/English/English%20%28UK%29%20MFA%20dictionary%20v2_0_0a.html.

[18] The English (US) ARPA acoustic model v2.0.0a (McAuliffe & Sonderegger (2022b)) was trained on speech by 2484 American English speakers from the LibriSpeech English corpus (Panayotov et al. (2015)) - for more information see https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20%28US%29%20ARPA%20acoustic%20model%20v2_0_0a.html

[19] English MFA acoustic model v2.0.0a (McAuliffe & Sonderegger (2022a)) trained on a number of varieties of English from the following corpora: 2479.95 hours from Common Voice English v8.0 (Ardila et al. (2020)), 982.3 hours from Librispeech English (Panayotov et al. (2015)), 124.31 hours from The Corpus of Regional African American Language (Kendall & Farrington (2018)), 5.77 hours from Google Nigerian English (Butryna et al. (2019)), 31.29 hours from the Open-source Multi-speaker Corpora of the English Accents in the British Isles (Demirsahin et al. (2020)), 56.43 hours from The NCHLT speech corpus of the South African languages (Barnard et al. (2014)), and 7.13 hours from the ARU Speech Corpus (University of Liverpool) (Hopkins et al. (2019)) - for more information see https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20MFA%20acoustic%20model%20v2_0_0a.html

Table 1: Comparison of forced alignment configurations

Aligner	Model	Training	Non-rhotic	Training Data
LaBB-CAT-HTK	Speaker specific	Train/Align	✓	0.3 – 2.9 min
LaBB-CAT-HTK	Gender specific	Train/Align	✓	13.1 – 15 min
LaBB-CAT-HTK	Child independent	Train/Align	✓	29.1 min
LaBB-CAT-HTK	Speaker independent	Train/Align	✓	46.1 min
P2FA	P2FA	Pre-trained	×	25.5 hours
MFA	Speaker adapted	Train/Align	✓	29.1 min
MFA	Variable state	Train/Align	✓	29.1 min
MFA	GAM Unadapted	Pre-trained	×	982.3 hours
MFA	GAM Speaker adapted	Pre-trained	×	982.3 hours
MFA	UK Speaker adapted	Pre-trained	✓	3687.0 hours

Evaluation of Alignments

Manual alignments, for comparison purposes, were provided by one of the authors, a graduate student in linguistics doing research specific to this data, using Praat (Boersma & Weenink (2001)). The best pronunciation was selected from all possibilities in CELEX for each word, using the 'DISC' phoneme symbols. 613 utterances were manually aligned, totalling 28:32 duration, and including 8,514 aligned segments. Manual alignment took approximately 40 hours.

In order to compare each manually aligned phone with its corresponding automatic counterpart, it was necessary to create a mapping between the two sets of alignments. This was complicated by two factors: a) each word may have a different phonemic transcription in the two alignments, because different dictionaries might use different phonemes to transcribe the word,[20] and forced alignment systems can select different pronunciations among all possible pronunciations of a word,[21] b) each dictionary employs a different set of symbols for each phoneme,[22] and don't necessarily use the same phoneme inventories.[23]

[20] e.g. the word "for" is transcribed with two phonemes in CELEX (`f$`), but with three in CMUdict (`F AO1 R`)

[21] e.g. CELEX transcribes the word "and" variously as `{nd` (ænd), `@nd` (ənd), `@n` (ən), `Hd` (n̩d), `H` (n̩), `F` (ŋ̩), or `C` (ŋ̩).

[22] e.g. the word "transcription" is transcribed using the 'DISC' symbols in CELEX, `tr{nskrIpS@n`, the ARPAbet symbols in CMUdict, `T R AE2 N S K R IH1 P SH AH0 N`, and using the IPA in the MFA 'UK English' dictionary, t r æ n s c r ɪ p ʃ ə n.

[23] e.g. the CELEX includes diphthongs `7` (NEAR), `8` (SQUARE) and `9` (CURE), but in CMUdict they are transcribed as multiple phonemes: `IY R`, `EH R`, and `UH R` respectively, and are similarly mismatched in MFA's 'UK English' dictionary, ɪ ə, ɛː, and ʊ ə respectively

In order to ensure the best possible mapping between different alignments, we used a common Minimum Edit Distance algorithm (Wagner & Fischer 1974), modified to ensure matching of similar phonemes across phoneme sets. Appendix A provides a table showing direct correspondences assumed between different symbol sets. The arrows in Figure 1 illustrate how these mappings work; despite the presence of inserted/deleted segments (coloured grey), and also despite the difference in encoding of the segment labels (the manual alignments above use CELEX 'DISC' symbols, while the automatic alignments below use ARPAbet symbols), the algorithm correctly maps corresponding phones to each other.

The literature includes a wide array of metrics for comparing alignments. We wanted to be able to compare our child NZE accuracy with the adult NZE accuracy reported by Fromont & Watson (2016)[24], and that reported by Gonzalez et al. (2020)[25], who reported Overlap Rates of 0.569 and 0.646 respectively. We also wanted to compare accuracies with other evaluations that used laboratory-based child speech; Knowles et al. (2018) reported 75%-90% accuracy using what we call 'Midpoint Containment', and Mahr et al. (2021) reported 86% accuracy using the same metric. In addition Szalay et al. (2022, Table 1.) reported Overlap Rates of 0.69-0.74. We report both of these metrics in our results purely to enable comparison with results from these previous experiments.

Both metrics are independent of the units used, and neither involve arbitrary thresholds to be decided.

Overlap Rate

Paulo & Oliveira (2004) devised Overlap Rate (OvR) as a measure of how much two intervals overlap, independent of their absolute durations. OvR is a value between 0, where the two intervals being compared do not overlap at all, and 1, where the two intervals have the same start and end times. OvR is calculated as follows:

$$OvR = \frac{CommonDur}{DurMax} = \frac{CommonDur}{DurRef + DurAuto - CommonDur},$$

where $CommonDur$ is the duration in common between the automatically aligned and manually aligned segments, $DurRef$ is the duration of the manually aligned segment, and $DurAuto$ is the duration of the automatically aligned segment. $DurMax$ is the maximum duration of the sound file covered by the pair of segments.

Figure 1 visualises how this works; the automatic alignment of the first vowel overlaps with only a third of the corresponding manual alignment, so OvR is 0.333. The second

[24]Fromont & Watson (2016), Section 4.1, p418
[25]Gonzalez et al. (2020) p6, Figure 2.

manually aligned vowel only covers half of the duration of the corresponding automatic alignment, so OvR is 0.5. For the final consonant, both alignments completely overlap each other, resulting in an OvR of 1.

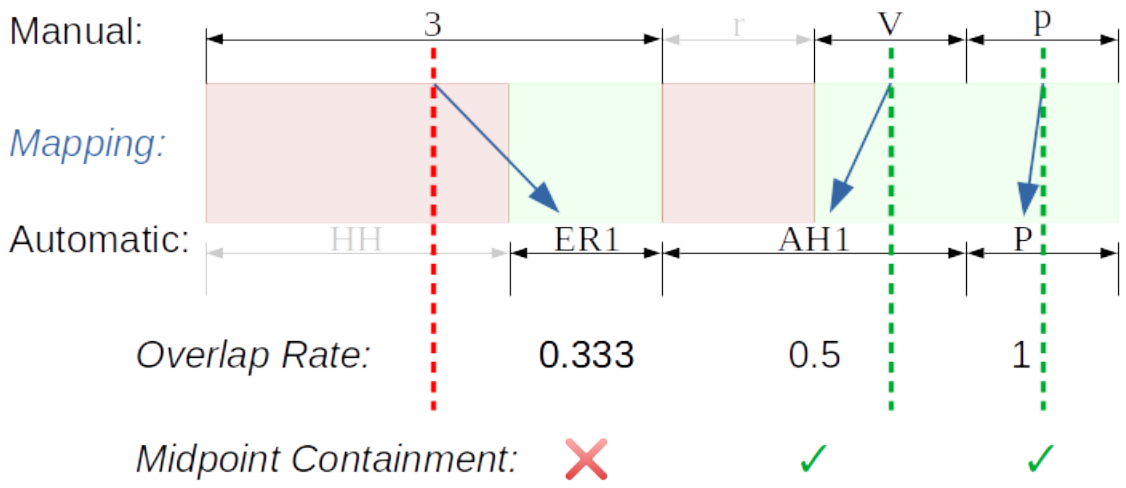

Figure 1. ***Example of mapping manually to automatically aligned phones, and metric computation***

Midpoint Containment

Knowles et al. (2018) devised a measure that calculates the percentage of segments that are 'approximately correct', defined as follows: 'the force-aligned segment overlapped with the midpoint of the corresponding manually aligned phone.'[26] Mahr et al. (2021) use the same metric, calling it a 'gross measure'.[27] Here we prosaically but descriptively call it "midpoint containment".

Figure 1 illustrates how alignments may match or not; the midpoint of the first manually aligned vowel falls outside the bounds of the corresponding automatic interval, so these alignments do not match. For both the overlapping second vowel, and the perfectly aligned final consonant, the manual alignment's midpoint falls within the bounds of its automatic counterpart, so these alignments match.

[26]Knowles et al. (2018) p. 2491, under "Comparisons"
[27]Mahr et al. (2021), p. 4, under "Outcome Variables"

Expectations

Given the general conclusions from the literature our expectations were as follows:

1. Overall performance would be lower than with adult speech, i.e. OvR will be lower than 0.646 (Gonzalez et al. (2020)) and also 0.569 (Fromont & Watson (2016)), because child speech is more varied than adult speech.
2. Overall performance would be lower than with controlled child speech, i.e. Midpoint Containment would be lower than 86% (Mahr et al. (2021)), 75% (Knowles et al. (2018)), and also lower than the 0.69 mean OvR reported by Szalay et al. (2022), because spontaneous speech is more varied than controlled speech.
3. Models trained on child speech would be better than those trained on adult speech, because in general the more similar the training and alignment speech, the better.
4. Non-rhotic dictionaries/models should perform better than rhotic ones; rhotic alignments will include alignments for post vocalic /ɹ/ phones that are not present in our non-rhotic NZE speech, so neighbouring automatic phones will overlap less with their manual counterparts.
5. MFA will perform better than the HTK-based aligners (LaBB-CAT-HTK and P2FA in our case), as found by González et al. (Gonzalez, Grama, et al. (2018), Gonzalez et al. (2020)).
6. Vowels will be the best aligned segments, as previously reported by Knowles et al. (2018) and Mahr et al. (2021).

Results

Table 2 compares both Overlap Rate and Midpoint Containment percentages for each of the forced alignment configurations. All train/align configurations have a mean OvR less than 0.3, and less than 50% Midpoint Containment, with the MFA configurations performing worse than the LaBB-CAT-HTK ones. Conversely, all configurations using models pre-trained on adult speech have a mean OvR greater than 0.3; the P2FA models produce a mean OvR of 0.345, the MFA GAM Unadapted models, 0.429, the MFA UK Speaker adapted models, 0.440, and the MFA GAM Speaker adapted models, the highest mean OvR at 0.458. In terms of Midpoint Containment, 48% of the P2FA alignments contain the midpoint of the corresponding manual alignment, and more than 50% of MFA alignments contain the manual alignment midpoint; 59% for GAM Unadapted models, 62% for UK Speaker adapted models, and 63% for GAM Speaker adapted models.

Figure 2 shows the distributions of Overlap Rates for each configuration. All train/align configurations have a third quartile of less than 0.6, and a first quartile of 0 (along with the P2FA pre-trained models). The *variable state* train/align models perform worst of

Table 2: Mean OvR and percent midpoint-contained, for each forced alignment configuration, with the best performing configuration in bold typeface

Aligner	Model	Training	Non-rhotic	Mean OvR	%
LaBB-CAT-HTK	Speaker specific	Train/Align	✓	0.228	37
LaBB-CAT-HTK	Gender specific	Train/Align	✓	0.261	42
LaBB-CAT-HTK	Child independent	Train/Align	✓	0.298	46
LaBB-CAT-HTK	Speaker independent	Train/Align	✓	0.276	42
P2FA	P2FA	Pre-trained	×	0.345	48
MFA	Speaker adapted	Train/Align	✓	0.239	34
MFA	Variable state	Train/Align	✓	0.155	22
MFA	GAM Unadapted	Pre-trained	×	0.429	59
MFA	**GAM Speaker adapted**	**Pre-trained**	×	**0.458**	**63**
MFA	UK Speaker adapted	Pre-trained	✓	0.440	62

all, with a median of 0, although curiously there are a number of outliers with OvR greater than 0.5. Only the pre-trained MFA models manage a first quartile greater than 0, and all have a third quartile greater than 0.7.

Figure 3 shows the distributions of Overlap Rates for each configuration, broken down by segment category. For the HTK-based tools (the left five configurations), there appears to be little differentiation in accuracy between different segment types. But for MFA (the right five configurations), vowels in particular seem to be very inaccurate for train/align configurations, but quite accurate for pre-trained configurations. Apart from those using MFA pre-trained models, none of the configurations had a first quartile higher than zero for any segment category.

Figure 3 also shows that, although the *GAM Speaker adapted* and *UK Speaker adapted* configurations have similar first and third quartiles for fricatives, the second quartile for *GAM Speaker adapted* is somewhat lower than for *UK Speaker adapted*. The mean fricative OvR for *GAM Speaker adapted* is 0.359 and the corresponding mean for *UK Speaker adapted* is 0.411.

Discussion

The most obvious result is that expectation 3., that 'models trained on child speech would be better than those trained on adult speech', was not borne out by the configurations we tested. All configurations that used only adult data were more accurate than all configurations that used any child data. This surprised us and apparently contradicts Knowles et al. (2018): 'For both corpora, training on adult speech led to poorer

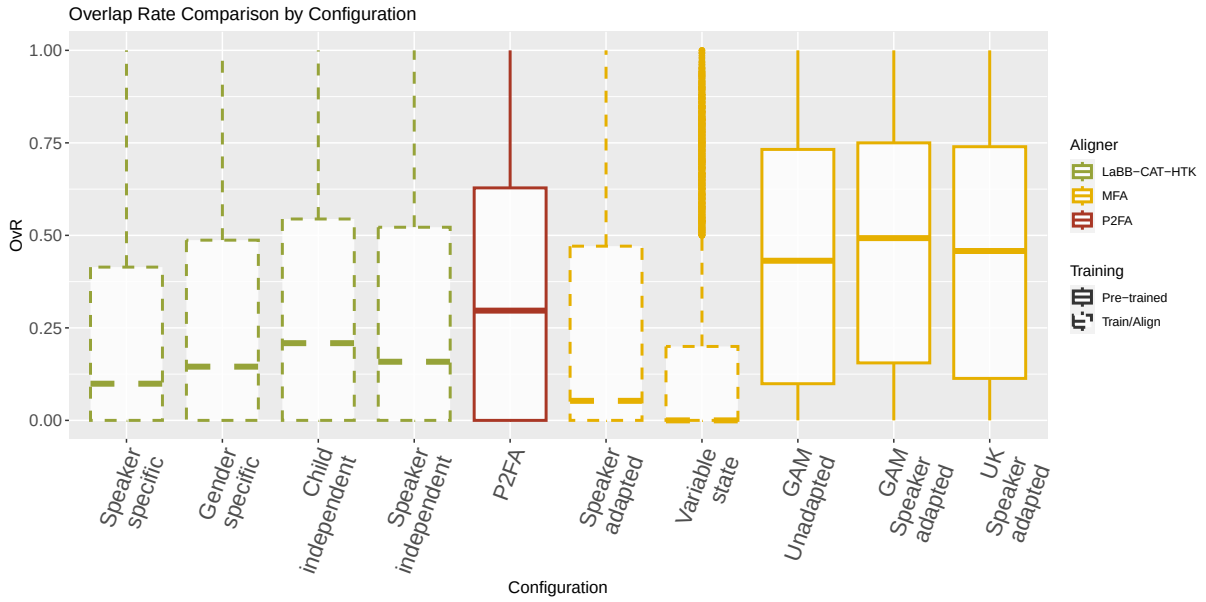

Figure 2. ***Overlap Rate distributions by Configuration. Green lines indicate LaBB-CAT-HTK configurations, dark red lines indicate the P2FA configuration, and yellow lines indicate MFA configurations. Filled lines indicate pretrained configurations, while dashed lines indicate Train/Align configurations.***

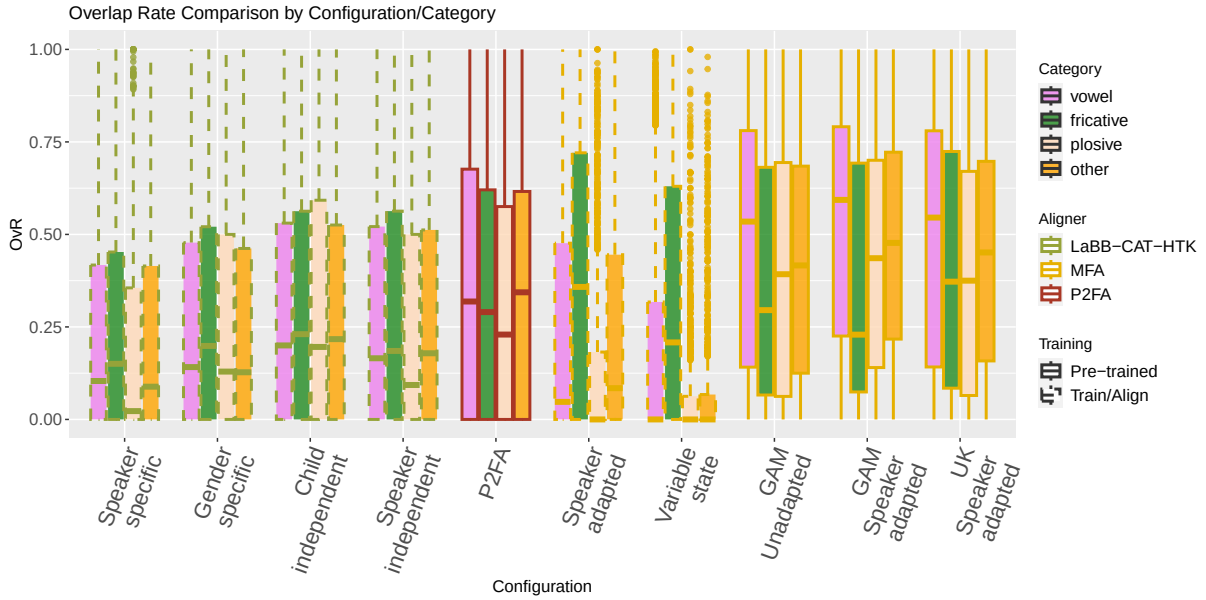

Figure 3. ***Overlap Rate distributions by Configuration, broken down by Segment Category. Line colour indicates the aligner used. Box colour indicates segment category.***

accuracy than training on child speech and can be summarized as follows: Adult-only < Adult–child < Child speech only'[28] although it's in line with Szalay et al. (2022) (section 4.1, p. 39) for whom the best aligners included adult data.

There are various possible explanations for this. Adult speech should be less phonologically varied than child speech, and represents the 'target' forms that children have not yet settled on; perhaps this stability leads to more discerning acoustic models. Or perhaps it's simply because there was more adult speech (25 - 3687 hours) than child speech (29.1 minutes) to train on. Knowles et al. had ten times this amount of child data (5 hours), which yielded alignments that were more accurate than models trained on adult speech (10 hours), so this may indicate that the latter explanation is correct: volume of training data trumps similarity to the speech to be aligned. It's clear that in some cases adult training data leads to higher accuracy for child speech, but further work is required to settle the question of whether this is because of the magnitude of the training data or its qualities.

Another surprise is that the configurations using a rhotic dictionary outperformed those using a non-rhotic dictionary. Using a rhotic dictionary for non-rhotic spontaneous speech inserts tokens of post-vocalic /ɹ/ which do not correspond to the speech. This inevitably decreases alignment accuracy[29], as the extra phone will invade the durations of surrounding phones. This can be seen in Figure 4., which shows an utterance from our corpus, with the correct manual alignment shown above, and the the automatic alignment produced by MFA below. The fourth word, "for", is correctly transcribed with two phonemes, `f $`, but MFA has used the three-phoneme transcription from its rhotic dictionary, `F AO1 R`, the last phone of which is an incorrect insertion taking up most of the duration of the vowel, which has a resulting low OvR of 0.099.

The MFA rhotic dictionary produced marginally better alignments (0.458 mean OvR) than the non-rhotic one (0.440 mean OvR) despite this 'inserted /ɹ/ penalty'. We investigated the incidence of spurious /ɹ/ phones in these alignments, and found that there were only 65 inserted /ɹ/ phones with a mean duration of 74ms, less than 1% of all the phones found in this alignment.[30]

The *English (US) ARPA* models are seemingly so much better than the *English MFA* models that the effect of having extra post-vocalic /ɹ/ tokens is rendered irrelevant. This

[28] Knowles et al. (2018) p. 2492.

[29] If the spurious phones are of zero length, accuracy would not be affected, but there were no zero duration insertions of this type in our data.

[30] Indeed /ɹ/ wasn't even the most common spurious phone; there were more spurious /d/ and /ə/ phones (118 and 68 tokens respectively), mainly representing the final phoneme of the words "and", "the" and "to".

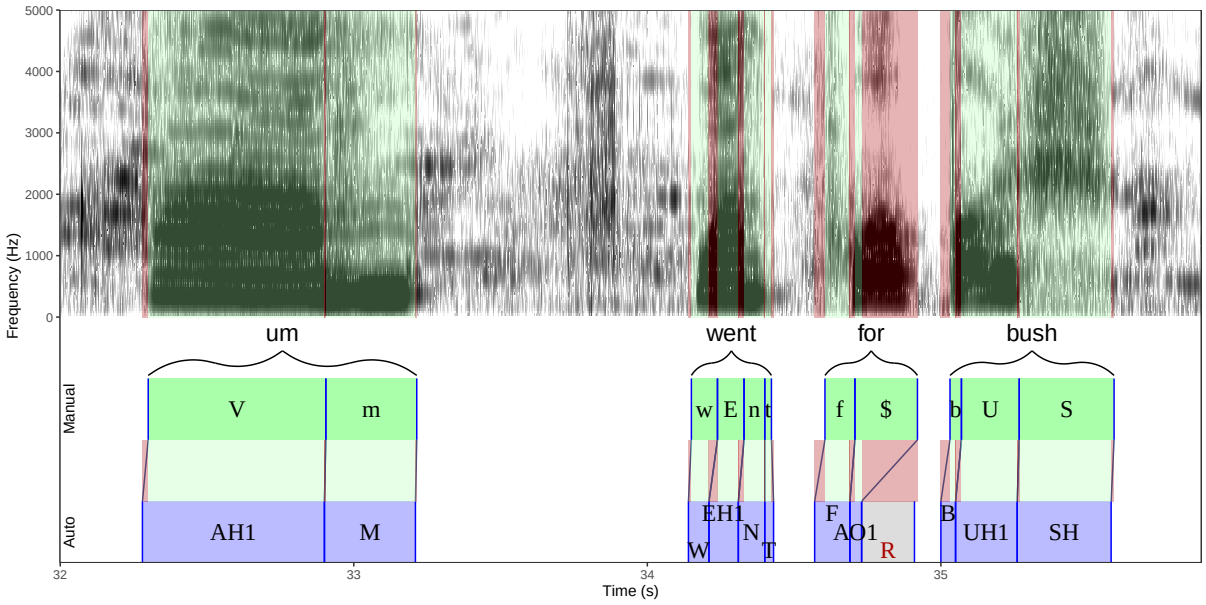

Figure 4. ***Example utterance alignment including inserted rhotic /ɹ/ in the word 'for' - correctly aligned regions are shown in light green, incorrectly aligned regions are shaded in red, and the utterance spectrogram is shown above for reference***

supports and perhaps explains conclusions of Gonzalez et al. (2020)[31] and MacKenzie & Turton (2020)[32] that dictionary/variety don't greatly impact measured performance: the incidence of features that define differences between varieties of the same language (in terms of insertion or deletion of segments) are not frequent enough to make much difference to overall alignment accuracy. However, this contradicts the advice of Szalay et al. (2022) (section 4.1, p. 39): "using a dialect matched, AusE pronunciation dictionary is recommended", and the impact of these discrepancies may indeed be important for downstream research that uses the resulting automatic alignments. For example if sociophonetic research is later conducted on word-final vowels, or on rhoticity itself, the spurious /ɹ/ tokens may significantly interfere with the results.

This better performance cannot be explained by differences in training set size; the models used with the GAM English dictionary were trained on under a thousand hours of speech, but still produced better alignments than the models used with the UK English dictionary, which were trained on over three thousand hours of speech. Apart from the amount of training data and the pronunciations, there are two other differences between these configurations: the latter was trained on numerous varieties of English, and the phoneme sets were differently distributed; The GAM English dic-

[31]Gonzalez et al. (2020) p. 9, section 5.
[32]MacKenzie & Turton (2020) p. 11 section 6.

tionary includes 39 stress-marked vowels and 24 consonants encoded in ARPABET[33], where the UK English dictionary includes 22 vowels and 46 consonants including vocalic and aspirated variants encoded with IPA symbols[34]. Investigating the impact each of these factors has is outside the scope of the current experiment, but it's clear from our results that more training data does not inevitably result in better alignments.

When comparing the performance of HTK-based aligners with MFA, the results are also nuanced. MFA did indeed perform better than HTK-based aligners under the conditions we tested, and as the *GAM Unadapted* accuracy is only slightly below *GAM Speaker adapted*, the difference in accuracy apparently doesn't come down to MFA's speaker-adaptation process. But MFA was more accurate only using pre-trained models. MFA produced the worst alignments among the configurations we tested when using a train/align procedure. The amount of training data is probably the important factor here. Michael McAuliffe, the primary software developer of MFA, notes that 3-5 hours of speech is required for good alignments.[35] It seems that under the conditions of our experiment, LaBB-CAT-HTK works better with scarce data than MFA does. More rigorous comparison between these ASR toolkits may well identify forced alignment methods, or attributes of training data, that yield different results. However, under the conditions we were working with – a relatively small amount of child speech, using the default procedures for LABB-CAT-HTK and MFA – train/align forced alignment was more accurate using LaBB-CAT-HTK, although accuracy was low for both aligners.

When compared with results from other studies, our expectations were borne out. Accuracy was lower than with adult speech, as the best mean OvR of 0.458 was lower than both 0.646 (Gonzalez et al. (2020)) and 0.569 (Fromont & Watson (2016)). Similarly, accuracy was lower with our spontaneous speech than with controlled child speech; our best Midpoint Containment of 63% was lower than 86% (Mahr et al. (2021)) and 75% (Knowles et al. (2018)), and our best mean OvR was lower than 0.69 (Szalay et al. (2022)).

Using MFA pre-trained models, vowels were indeed the best-aligned segments, confirming results from Knowles et al. (2018) and Mahr et al. (2021). However, this was not the case with other configurations. Although the *English (US) ARPA* models are marginally better than the *English MFA* models overall, the latter was better at aligning fricatives. As noted earlier with reference to rhoticity, which models/dictionaries turn out to be best depends somewhat on what types of segment will be analysed downstream.

[33]The GAM English phoneme set is shown in Appendix A

[34]The consonant variants of the UK English phoneme set is shown in Appendix A, Table 4

[35]Michael McAuliffe, "How much data do you need for a good MFA alignment?" (24 August 2021): https://memcauliffe.com/how-much-data-do-you-need-for-a-good-mfa-alignment.html

Conclusion

While there is an established literature on forced alignment methodology for adult speech, accuracy with child speech has only recently received any attention from researchers, and the best approach for dealing with field recordings of children has not been established.

We found that MFA, using acoustic models pre-trained on 'General American English', produced the most accurate alignments of spontaneous NZE child speech in our corpus. These alignments were less accurate than is possible with adult speech of the same variety, and with controlled child speech, and all future alignments in our growing corpus will require manual checking/correction.

Although the results of our experiments resolved a practical problem for us, identifying a clear way forward for the force-alignment of our own corpus, we recognise that they are specific to our speech data and the configurations we tried, using conveniently configurable tools designed specifically for sociophonetic research. More rigorous further work would be required to tease apart the relative importance of the various factors – toolkit, technology, data preparation, amount of and nature of the speech, age and dialect of speakers in the training vs. alignment data, etc.

For example somewhere between the half hour of speech we had available for training, and the two to five hours of speech used by Knowles et al. there may be a threshold where training on child speech alone yields better alignments than those produced by using models pre-trained on adult speech. Furthermore, the recursive method of forced alignment studied by Gonzalez, Travis, et al. (2018) may provide a boost in performance. These are questions to be resolved by future investigation, on a larger corpus of child speech.

In addition the present results compared only HMM-based forced alignment. However, Kaldi also supports the use of DNNs for forced alignment. It would be useful to compare performance of DNN-based alignments with HMM-based ones, using Gentle out of the box, or by training custom aligners as done by Szalay et al. (2022). They point out that their best custom aligner was trained on the same AusE dataset as the HMM-based MAUS aligner, which performed the worst, concluding that the difference in performance can be attributed to using Kaldi and DNNs, rather than HTK and HMMs (Szalay et al. (2022), p. 39, section 4.2). If Kaldi alone were used to discover whether DNNs or HMMs produce more accurate alignments, it could be determined whether it's the toolkit or the technology that makes the difference.

We conclude that alignment procedures that work well with adult data are not guaranteed to produce the best results for children. To maximise accuracy, automated alignment of language acquisition corpora requires special attention, and evaluating differ-

ent options on specific corpora is well worth the effort. Even so, our finding with NZE child speech was the same as that of Szalay et al. (2022, p.38 section 4) with AusE child speech: manual correction is still required. We echo MacKenzie & Turton (2020)[36] who recommend that "these aligners are used in the manner for which they were designed — as tools, and not as the complete replacement of a dedicated researcher".

References

Adell, J., Bonafonte, A., Gómez, J. A., & Castro, M. J. (2005). Comparative study of automatic phone segmentation methods for TTS. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, *1*, I/309–I/312 Vol. 1.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222. https://aclanthology.org/2020.lrec-1.520

Assmann, P. F., & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. *The Journal of the Acoustical Society of America*, *108*(4), 1856–1866. https://doi.org/10.1121/1.1289363

Athanasopoulou, A. A., & Vogel, I. (2016). Acquisition of prosody: The role of variability. *Speech Prosody*, 716–720.

Baayen, H., Piepenbrock, R., & Rijn, H. V. (1995). *The CELEX Lexical Database (Release 2, CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania. https://catalog.ldc.upenn.edu/LDC96L14

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. *Proceedings of the Linguistic Society of America*, *4*(1), 3-1-12. https://doi.org/10.3765/plsa.v4i1.4468

Barnard, E., Davel, M., Heerden, C. van, Wet, F., & Badenhorst, J. (2014). The NCHLT Speech Corpus of the South African languages. *SLTU 2014*, 194–200.

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Brand, J., Hay, J., Clark, L., Watson, K., & Sóskuthy, M. (2021). Systematic co-variation of monophthongs across speakers of New Zealand English. *Journal of Phonetics*, *88*, 101096. https://doi.org/https://doi.org/10.1016/j.wocn.2021.101096

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2012). Automatic Phone Alignment - A Comparison between Speaker-Independent Models and Models Trained on the Corpus to Align. *JapTAL*.

Butryna, A., Chu, S. H. C., Demirsahin, I., Gutkin, A., Ha, L., He, F., Jansche, M., Johny, C. C., Katanova, A., Kjartansson, O., Li, C. F., Merkulova, T., Oo, Y. M., Pipatsrisawat, K., Rivera, C. E., Sarin, S., Silva, P. D., Sodimana, K., Sproat, R., … Wibawa,

[36]MacKenzie & Turton (2020) p. 12 section 6

J. A. E. (2019). Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview. *2019 UNESCO International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide*, 91–94. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.23.pdf

Chen, L., Liu, Y., Harper, M. P., Maia, E., & McRoy, S. (2004). Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus. *LREC*.

Coleman, J. (2005). *Introducing speech and Language Processing*. Cambridge University Press.

Cosi, P., Falavigna, D., & Omologo, M. (1991). A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *2nd European Conference on Speech Communication and Technology (Eurospeech 1991)*.

Coto-Solano, R. (2022). Computational sociophonetics using automatic speech recognition. *Language and Linguistics Compass*, *16*(9), e12474. https://doi.org/https://doi.org/10.1111/lnc3.12474

Demirsahin, I., Kjartansson, O., Gutkin, A., & Rivera, C. (2020). Open-source Multispeaker Corpora of the English Accents in the British Isles. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6532–6541. https://aclanthology.org/2020.lrec-1.804

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., & García, R. C. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, *134*(3), 2235–2246. https://doi.org/10.1121/1.4816491

Fromont, R. (2023). *nzilbb.labbcat R package* (Version 1.2-0). https://github.com/nzilbb/labbcat-R/

Fromont, R., Black, J. W., Clark, L., & Blackwood, M. (2022). Forced alignment of child speech: Comparing HTK and kaldi-based aligners. *Third Workshop on Sociophonetic Variability in the English Varieties of Australia*.

Fromont, R., & Hay, J. (2012). LaBB-CAT: an Annotation Store. *Proceedings of Australasian Language Technology Association Workshop*, 113–117.

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

Gnevsheva, K., Gonzalez, S., & Fromont, R. (2020). Australian English Bilingual Corpus: Automatic forced-alignment accuracy in Russian and English. *Australian Journal of Linguistics*, *40*(2), 182–193. https://doi.org/10.1080/07268602.2020.1737507

Gonzalez, S., Grama, J., & Travis, C. (2018). *FoACL: Forced-Alignment Comparison for Linguistics*. https://cloudstor.aarnet.edu.au/plus/s/gyC6vuX5uvc5soG

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, *6*(1), 20190058. https://doi.org/doi:10.1515/lingvan-2019-0058

Gonzalez, S., Travis, C., Grama, J., Barth, D., & Ananthanarayan, S. (2018). *Recursive forced alignment: A test on a minority language* (pp. 145–148).

Gordon, E., Maclagan, M., & Hay, J. (2007). The ONZE corpus. In J. C. Beal, K. P. Cor-

rigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora: Volume 2: Diachronic databases* (pp. 82–104). Palgrave Macmillan UK. https://doi.org/10.1057/9780230223202_4

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

Hawkins, M., Strob, R. B., & andrakeshshrestha31, D. B. B. (2017). *Gentle*. http://lowerquality.com/gentle/

Hopkins, C., Graetzer, S., & Seiffert, G. (2019). *ARU speech corpus (University of Liverpool)*.

Kendall, T., & Farrington, C. (2018). The Corpus of Regional African American Language. *Version, 6*, 1.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45*, 326–347. https://doi.org/https://doi.org/10.1016/j.csl.2017.01.005

Knowles, T., Clayards, M., & Sonderegger, M. (2018). Examining Factors Influencing the Viability of Automatic Acoustic Analysis of Child Speech. *Journal of Speech, Language, and Hearing Research, 61*(10), 2487–2501. https://doi.org/10.1044/2018/_JSLHR-S-17-0275

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455–1468. https://doi.org/10.1121/1.426686

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations [Journal Article]. *International Journal of Corpus Linguistics, 22*(3), 319–344. https://doi.org/https://doi.org/10.1075/ijcl.22.3.02lov

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard, 6*(s1), 20180061. https://doi.org/doi:10.1515/lingvan-2018-0061

Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., & Hustad, K. C. (2021). Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research, 64*(6S), 2213–2222. https://doi.org/10.1044/2020/_JSLHR-20-00268

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proc. Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McAuliffe, M., & Sonderegger, M. (2022a). *English MFA acoustic model v2.0.0a*.

McAuliffe, M., & Sonderegger, M. (2022b). *English (US) ARPA acoustic model v2.0.0a*.

Meer, P. (2020). Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *The Journal of the Acoustical Society of America, 147*(4), 2283–2294. https://doi.org/10.1121/10.0001069

Moreno, P. J., Joerg, C., Thong, J.-M. V., & Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. *Proc. 5th International Conference*

on Spoken Language Processing (ICSLP 1998), paper 0068. https://doi.org/10.21437/ICSLP.1998-603

Niekerk, D. van, & Barnard, E. (2009). Phonetic alignment for speech synthesis in under-resourced languages. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 880–883. https://doi.org/10.21437/Interspeech.2009-266

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

Paulo, S., & Oliveira, L. C. (2004). Automatic Phonetic Alignment and Its Confidence Measures. In J. L. Vicedo, P. Martínez-Barco, R. Muńoz, & M. Saiz Noeda (Eds.), *Advances in natural language processing* (pp. 36–44). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30228-5_4

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. K., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Veselý, K. (2011). *The Kaldi Speech Recognition Toolkit*.

Rudnicky, A., & Weide, R. (2014). *Carnegie Mellon University Pronouncing Dictionary*. Carnegie Mellon University; http://www.speech.cs.cmu.edu/cgi-bin/cmudict. http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

Schiel, F. (1999). *Automatic Phonetic Transcription of Non-Prompted Speech*.

Schiel, F. (2015). A statistical model for predicting pronunciation. *International Congress of Phonetic Sciences*.

Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., & Steffen, A. (2012). *The Production of Speech Corpora*. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-13693-2

Smith, B. L. (1992). Relationships between Duration and Temporal Variability in Children's Speech. *The Journal of the Acoustical Society of America*, *91*(4 Pt 1), 2165–2174. https://doi.org/10.1121/1.403675

Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., & Thomas, E. R. (2019). Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. *Proceedings of the 19th International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1322.pdf

Szalay, T., Shahin, M., Ballard, K., & Ahmed, B. (2022). Training forced aligners on (mis)matched data: The effect of dialect and age. *Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology*, 36–40.

Tang, K., & Bennett, R. (2019). Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (Mayan). *Proceedings of the 19th International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1768.pdf

Toledano, D. T., & Gómez, L. A. H. (2002, May). HMMs for Automatic Phonetic Segmentation. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. http://www.lrec-conf.org/proceedings/lrec2002/pdf/142.

pdf
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchovaltchev, & Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department; http://htk.eng.cam.ac.uk/docs/docs.shtml.
Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*, 5687–5690.

Data, Code and Materials Availability Statement

The script used for configuring forced alignment conditions and processing data, and also the edit paths mapping manually aligned phones to their automatically aligned counterparts, are available using the following URL:

https://doi.org/10.17605/OSF.IO/8R9FP

Ethics statement

Approval for use of this data was gained from the University of Canterbury Human Ethics Committee (Ref 2020/10/ERHEC).

Authorship and Contributorship Statement

- Robert Fromont: Conceptualization, Methodology, Software, Visualization, Writing, Review & editing
- Lynn Clark: Conceptualization, Funding acquisition, Review & editing
- Joshua Wilson Black: Data curation, Visualization, Review & editing
- Margaret Blackwood: Data curation, Review & editing

Acknowledgements

We gratefully acknowledge the support of the Marsden Fund | Te Pūtea Rangahau a Marsden (Application number: 20-UOC-064).

We would like to thank Tristan Mahr for his valuable feedback on an earlier draft, and our anonymous reviewers for taking the time to thoroughly critique our manuscript, whose comments were invaluable for sharpening our thinking and refining our conclusions.

License

Appendix A - Phoneme Symbol Sets

Different English pronunciation dictionaries use different sets of symbols. In many cases, there are quirks that relate to the provenance and purpose of the dictionary; for example the CMU Pronouncing dictionary (CMU Dict) has no symbol for schwa, because unstressed vowels are instead suffixed with 0. Some of the MFA dictionaries use IPA symbols, but transcribe diphthongs in unfamiliar ways, perhaps because of efforts by the developer to develop multi-lingual models. CELEX's 'DISC' symbols are similar to the SAMPA symbols familiar to many linguists, except they conform to the principle that each phoneme can be represented by exactly one character. Below is a table showing how different symbols sets relate to each other.

Table 3: Vowels

Example	IPA	MFA	DISC	CMU Dict[37]
k**i**t	ɪ	ɪ	I	IH
dr**e**ss	ɛ	ɛ	E	EH
tr**a**p	æ	æ	{	AE
str**u**t	ʌ	ɐ	V	AH
f**oo**t	ʊ	ʊ	U	UH
another	ə	ə	@	
fl**ee**ce	iː	iː/**i**	i	IY
b**a**th	ɑː	ɑː	#	AA
l**o**t	ɒ	ɒ	Q	AO
th**ou**ght	ɔː	ɒː	$	AO
g**oo**se	uː	ʉː/ʉ	u	UW
n**ur**se	ɜː	ɜː/ɜ	3	ER
f**a**ce	eɪ	ej	1	EY
pr**i**ce	aɪ	aj	2	AY
ch**oi**ce	ɔɪ	ɔj	4	OY
g**oa**t	əʊ	əw	5	OW
m**ou**th	aʊ	aw	6	AW
n**ear**	ɪə	ɪ ə	7	IY R
sq**uare**	ɛə	ɛː	8	EH R
c**ure**	ʊə	ʊ ə	9	UH R
t**i**mbre	æ̃		c	
dét**en**te	ɑ̃ː	ɑ	q	
l**in**gerie	æ̃ː		0	
bouill**on**	ɒ̃ː		~	

[37] All vowels in CMU Dict's ARPABET encoding have three variants, each suffixed with a digit: 0 for unstressed, 1 for primary stress, and 2 for secondary stress

Table 4: Consonants

Example	IPA	MFA	DISC	CMU Dict
pat	p	p/pʰ/pʲ	p	P
bad	b	b/bʲ	b	B
tack	t	t/tʰ/tʲ	t	T
dad	d	d/dʲ	d	D
cad	k	k/kʰ/c/cʰ	k	K
game	g	g/ɟ	g	G
ba**ng**	ŋ	ŋ	N	NG
mad	m	m/mʲ/ɱ	m	M
nat	n	n/ɲ	n	N
lad	l	l/ɬ/ʎ	l	L
rat	ɹ	ɹ	r	R
fat	f	f/fʲ	f	F
vat	v	v/vʲ	v	V
thin	θ	θ	T	TH
then	ð	ð	D	DH
sap	s	s	s	S
zap	z	z	z	Z
sheep	ʃ	ʃ	S	SH
mea**s**ure	ʒ	ʒ	Z	ZH
yank	j	j	j	Y
had	h	h/ç	h	HH
wet	w	w	w	W
cheap	tʃ	tʃ	J	CH
jeep	dʒ	dʒ	_	JH
lo**ch**	x		x	
baco**n**	ŋ̍		C	
idealis**m**	m̩	m̩	F	
burde**n**	n̩	n̩	H	
dang**l**e	l̩	ɫ̩	P	
ca**r** alarm	*		R	
uh-oh	ʔ	ʔ		