2019. *Proc Ling Soc Amer* 4. 3:1-12. https://doi.org/10.3765/plsa.v4i1.4468

# A Robin Hood approach to forced alignment: English-trained algorithms and their use on Australian languages

Sarah Babinski, Rikker Dockum, Dolly Goldenberg,
J. Hunter Craft, Anelisa Fergus & Claire Bowern*

**Abstract.** Forced alignment automatically aligns audio recordings of spoken language with transcripts at the segment level, greatly reducing the time required to prepare data for phonetic analysis. However, existing algorithms are mostly trained on a few well-documented languages. We test the performance of three algorithms against manually aligned data. For at least some tasks, unsupervised alignment (either based on English or trained from a small corpus) is sufficiently reliable for it to be used on legacy data for low-resource languages. Descriptive phonetic work on vowel inventories and prosody can be accurately captured by automatic alignment with minimal training data. Consonants provided significantly more challenges for forced alignment.

**Keywords.** Australian languages; phonetics; forced alignment; language documentation; Yidiny

**1. Introduction.** In order to conduct phonetic analysis, the alignment of an audio recording with its transcript at the segment level is necessary. The technology known as forced alignment (FA) is the use of computer algorithms to accomplish this task. Without the use of forced alignment, the manual segmentation and alignment of a transcript with a sound file for phonetic analysis is often prohibitively time and labor intensive. The use of digital recording technologies has made this issue more pronounced by increasing the amount of data available and thus the amount of time needed for manual alignment. In these situations, existing FA algorithms are very helpful, however most are trained on only a small number of well-documented and highly-resourced languages (Lin et al. 2005; Yuan and Liberman 2008). This situation presents a challenge to researchers working on under-resourced and endangered languages, because there are often no existing language models for FA algorithms. Furthermore, data from under-resourced and endangered languages may be legacy data recorded on analog media that is overlooked in favor of digital recordings that are widely available and easier to work with, further exacerbating the digital divide.

Another complicating feature of many FA algorithms is that the amount of data required to train entirely new language models accurately is often prohibitive. For endangered languages with small corpora of legacy data, this is simply not possible. In order to conduct phonetic analyses of such languages, a solution that circumvents the limitations of manual alignment while working with the available FA algorithms is necessary. For this reason, we investigate whether existing (pre-trained) alignment algorithms are in fact usable for languages without large corpora and financial resources.

*Thanks to Jason Shaw and the Phonology and Historical Reading Groups at Yale. This work was funded by NSF Grant BCS-1423711. Authors: Sarah Babinski (sarah.babinski@yale.edu), Rikker Dockum (rikker.dockum@yale.edu), Dolly Goldenberg (dolly.goldenberg@yale.edu), J. Hunter Craft (hunter.craft@yale.edu), Anelisa Fergus (anelisa.fergus@yale.edu), & Claire Bowern (claire.bowern@yale.edu), all of Yale University. Author contributions are as follows: designed project: CB, SB, RD; coded manual data: HC, AF, RD, SB, CB; ran automated alignments: DG, CB; scripting and data analysis: RD, SB, CB; wrote paper: RD, SB, HC, CB.

The language used for our test is Yidiny, a Pama-Nyungan language from the Cairns Rainforest region of Australia's Cape York Peninsula. Yidiny's closest relative is Djabugay (Patz 1991) and to our knowledge currently has no fluent speakers, which limits the possibility of adding to the corpus gathered by R.M.W. Dixon approximately 50 years ago. Researchers such as Dixon (1977b) have done some work on its sound system, but much important analysis remains to be done. Yidiny, as with most Australian languages, certainly qualifies as highly endangered and under-resourced, and as such makes an ideal candidate for the consideration of alternative documentation methods.

For our forced alignment algorithms, we use three models. Two of these are trained on English: namely P2FA (Evanini et al. 2009) and DARLA (Reddy and Stanford 2015). The third algorithm, MFA (the Montreal Forced Aligner), is not trained on English, but allows for training on small corpora such as Yidiny's (Povey et al. 2011).[1] After training MFA on Yidiny, we use all three aligners on the same corpus and then compare the results of these algorithms with manually-corrected data as our gold standard. In doing so, we assess how accurately FA algorithms capture the alignment of Yidiny segments for the purpose of acoustic phonetic description.

## 2. Methods.

2.1 DATA SOURCE. Though there are no longer fluent speakers of Yidiny to our knowledge, a body of available data makes further linguistic analysis possible. The Yidiny materials used in this project come from a group of eight recordings and their associated transcriptions made by Dixon in the late 1960s and deposited at the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). The recordings comprise narratives from two different speakers, Tilly Fuller and Dick Moses, and range in duration from about 4 to 8 minutes. In total, we use about 45 minutes of speech. The speakers were both born at the end of the 1800s and were fluent speakers of the language. It should be noted that Dick Moses spoke the coastal dialect while Tilly Fuller spoke the inland (tablelands) dialect. While this is not the complete extent of audio materials recorded for Yidiny, it is a substantial part of the publicly available narrative corpus.

2.2 DATA PREPARATION. We began by creating a preliminary ARPABET pronunciation dictionary for Yidiny.[2] We then used it to align the transcripts, and then corrected the alignments manually in Praat. We created customized ARPABET pronunciation dictionaries to introduce multiple test conditions, which we then automatically aligned using P2FA and MFA. The manually corrected alignments became the basis for comparison to the various conditions of automatic alignment. Further discussion of the workflow and possible use cases are given in Section 4 below.[3]

[1] We initially also wished to include the Praat Forced Alignment system (Boersma et al. 2002). Since the phonetic analysis is conducted in Praat, it woudl make sense to try to contain the alignment within the same program, simplifying the workflow. The Praat forced alignment algorithm works by matching the audio signal against a text to speech language model. This makes it potentially idea for working with low-resource languages, since it is possible to align without training data. However, none of the pre-defined language models produced results which were usable. We attempted to create our own model, but this required discussion with the eSpeak-edit developers, who unfortunately did not respond to our email. We hope to pursue this in future work.

[2] The ARPABET is a plain text set of conventions for representing phonological transcription using only letters and numbers. It is often used in forced alignment research; see, for example, the CMU Pronouncing Dictionary (of English) for its full conventions. The conventions used in this project are given in Tables 1 and 2.

[3] A member of the audience pointed out that this procedure – aligning from one of the test cases – could bias our results in favor of that algorithm. We acknowledge the possibility, but we note that our results do not, in fact, show

Because Yidiny orthography is surface phonemic, the transcripts correspond fairly closely to the surface pronunciation, including allomorphic variation (e.g. alternations in phonemic vowel length). This fact made the transcription of most segments into ARPABET straightforward. However, because the transcriptions and orthography do not take into consideration allophony, and ARPABET is limited to English phonemes, some segments were more difficult to map than others. This then raises the question, where multiple ARPABET – Yidiny mappings are possible – whether such choices affect the accuracy of automatic alignment. This question provides the basis for our different conditions. Furthermore, the orthography corresponds more closely to Dick Moses's dialect than Tilly Fuller's.[4]

Yidiny's syllable structure is primarily CV(C) with a few consonant clusters occurring word medially (Nash 1979; Dixon 1977a). Yidiny's vowel system distinguishes three vowel qualities: /i/, /a/, and /u/ with phonemic length distinctions for all three (/i:/, /a:/, and /u:/). Because P2FA is primarily concerned with examining consonant–vowel transitions, we were not not concerned with distinguishing in the ARPABET pronunciation dictionary between stressed and unstressed vowels. We coded each vowel as having primary stress (indicated by a 1 following the ARPABET segment). However, we did code for phonemic length distinctions. Table 1 summarizes the segments of Yidiny that readily map to ARPABET segments in the P2FA pronunciation dictionary. Segments that have less clear ARPABET targets were /ɹ/, /r/, /b/, /d/, /ɟ/, /g/, and /ɲ/.

| Orthography | Idealized IPA | ARPABET |
|---|---|---|
| a | /a/ | AH1 |
| aa | /a:/ | AA1 |
| u | /u/ | UH1 |
| uu | /u:/ | UW1 |
| i | /i/ | IH1 |
| ii | /i:/ | IY1 |
| m | /m/ | M |
| n | /n/ | N |
| ng | /ŋ/ | NG |
| l | /l/ | L |
| w | /w/ | W |
| y | /j/ | Y |

Table 1: Yidiny to ARPABET mappings.

Other segments present challenges for the Yidiny pronunciation dictionary. The two rhotics, /ɹ/ and /r/, exhibit allophonic variation and neutralization, with both being realized as taps in certain contexts. Because the trill is more commonly realized as a tap than the approximant, we chose to represent this in some conditions as R and in others as D, while the approximant was represented as R in all conditions. Because Yidiny shows no phonemic distinction between voiced and voiceless stops, we represented the stops /b/, /d/, /ɟ/, and /g/ as voiced B, D, JH, and G in the Voiced condition,

that P2FA is better than MFA, where the alignments were created independently.

[4]We did not alter any transcription, even though there were places in the texts where Tilly Fuller's pronunciation appeared to differ in systematic ways from Dick Moses'.

but voiceless P, T, CH, and K in all others. This decision meant that there was some overlap in the Voiced condition between /r/ and /d/, because both were coded as D.[5] Of particular challenge was the palatal nasal stop. We represented this segment in three different ways: as N, as Y, or as a cluster of N followed by Y. These mappings are summarized in Table 2. We also chose the Voiceless condition as the basis for our manually corrected alignment.

| Yidiny | IPA | Voiced | Voiceless | R Condition | NY Condition | Y Condition | N Condition |
|---|---|---|---|---|---|---|---|
| r | /ɹ/ | R | R | R | R | R | R |
| rr | /r/ | D | D | R | D | D | D |
| d | /d/ | D | T | T | T | T | T |
| b | /b/ | B | P | P | P | P | P |
| dy | /ɟ/ | JH | CH | CH | CH | CH | CH |
| g | /g/ | G | K | K | K | K | K |
| ny | /ɲ/ | N | N | N | N+Y | Y | N |

Table 2: Mappings of difficult segments to ARPABET in different test conditions.

While manually aligning the texts, we encountered several issues. The first problem was the presence of non-speech sounds in the recordings. These sounds included birdsong, laughter, hesitations on the part of the speaker, and strong winds (many recordings were made outside). Where it became obvious that these sounds had interfered significantly with the accuracy of the automatic transcription, we did not use the affected portions of the files.[6] We removed these portions from the transcriptions.

Furthermore, Dixon's transcription did not always map to the audio in the recording. In the case of the narratives from Tilly Fuller, this was a result of differences in dialect. Because we were interested in the underlying phonemic representations, we chose to align according to the transcription in these instances, even where individual segments were not immediately present at a cursory investigation. In several other instances, the transcript did not match the audio as a result of simply being incorrectly transcribed or including hesitations, stuttering, or other sounds (e.g. backchanneling) from the speaker. In these cases the transcripts were manually corrected to match the audio file.

The alignment for DARLA followed a different method. Because DARLA does not use an ARPABET pronunciation dictionary, we did not have different conditions from ARPABET transcription decisions. DARLA allows alignment in two different ways. The first way requires a text grid file that already has utterance boundaries notated. The second uses a a plain text transcript with no boundaries marked. We used the second method to align our transcripts and recordings. Because DARLA extrapolates segmentation from the orthography, the number and nature of segments detected by DARLA did not always correspond to the manually corrected and automatically aligned P2FA alignments.

[5]When analyzing results, we recovered the original phoneme.

[6]We did this because we consider the task here to be how to make accurate analyses (and parts of such recordings would likely be omitted even in manual segmentation and analyses. Note that these recording artifacts are very typical for archival field recordings, which are almost never made under anything like 'lab' conditions, and so having a protocol for what to do with them is important. See further Johnson et al. (2018) for a test of forced alignment which compares 'sanitized' versus original recordings for Tongan.

The methods for the MFA alignment were similar to that used for P2FA. MFA uses an ARPABET-based pronouncing dictionary (we used the same file as for the P2FA alignment).

MFA requires the alignment audio to be in small chunks (of several seconds), while P2FA can align long audio files, working successfully on files of 10 minutes in duration or longer. We ran both P2FA and MFA using files segmented at the utterance level, with a 80 ms buffer before and after each segment. Because Darla works by manually uploading single files, we could not test Darla in the same way. Instead, we uploaded each text as its own audio file and transcript.

2.3 DATA PROCESSING. After generating our automatic alignments and completing the corrections for the Manual condition, we extracted various measures from all of the resulting TextGrids. We adapted a duration logger script by Christian DiCanio to extract measures from each TextGrid interval; we also used his corpus analytics script which takes measures that include segment and word durations, F0, intensity, and vowel formants. We wrote a script in R (R Core Team 2018) for post-processing and analysis of the resulting data.

Linear mixed effects models were used to compare different alignment algorithms and different conditions within P2FA, using the `lmertest` package in R (Kuznetsova et al. 2017). Fixed effects for algorithm comparison were speaker gender and alignment model, with a random effect of word. We used the `vowels` package (Kendall and Thomas 2018) to process data on the vowel spaces.

**3. Results.** How 'good' is forced alignment used in this way? The answer depends on what sort of measurements the researcher needs. Certain types of questions are more robust to the error these automatic aligners introduce than others. Prosodic measurements are highly robust to aligner error, while consonant alignment and durations are less so. This section considers the accuracy of all alignment conditions in detecting prosodic factors (pitch maximum, pitch peak), vowel measurements, and consonant durations.

3.1 PROSODY. Projects looking at prosody and stress are likely to require accurate F0 measurements. Figure 1 shows the peak F0 measurements from each of the alignment conditions. DARLA differed significantly from manual alignment, but P2FA and MFA were not significantly different from the gold standard. Peak F0 was within 1-2Hz of the manual condition.

Figure 1 shows the location of F0 peak measurements in the word, and again little difference is found across alignment condition. DARLA results were significantly different than the manual in this case; other conditions do not differ from manual. Average peak locations were within 0.6%-1% of manual alignment.

Overall, these results suggest that studies of prosody would benefit greatly from the use of forced alignment, with little to no loss of accuracy. Such findings are particularly encouraging for work on Australian languages, where work on prosody is still at an early stage.

3.2 VOWEL SPACE. One way to measure the accuracy of vowel segmentation is to consider how accurate the vowel space is (i.e. first and second formants). Measurements of vowel formants are slightly less robust to aligner error than F0 measurements. Vowel means are shown in Figure 2 for both speakers. DARLA clearly performed most poorly on the data, with all measurements being

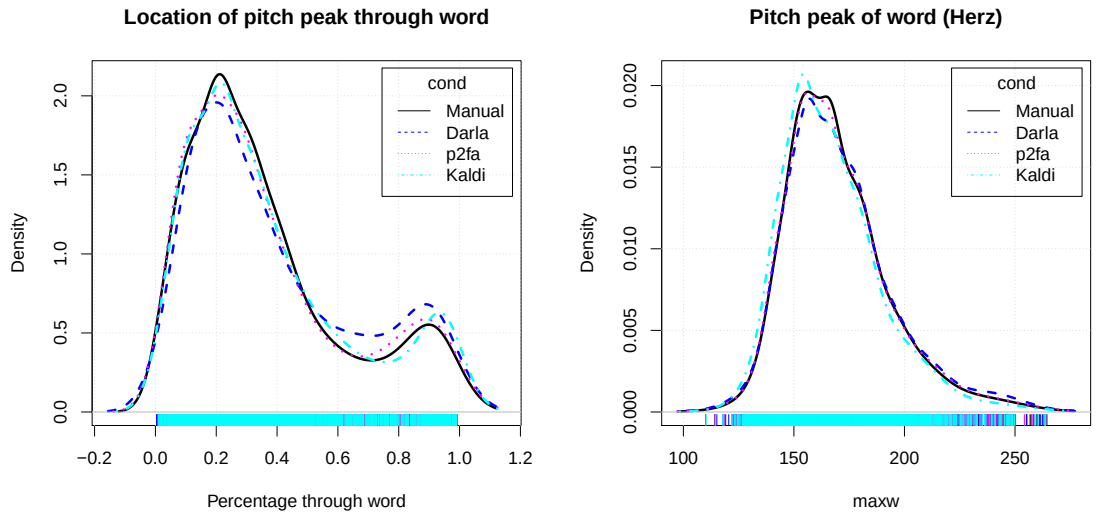


Figure 1: Density plots showing, across all alignment algorithms, the location of pitch peak in the word (left) and the maximum pitch values across all words (right). DARLA distributions are significantly different from manual on both counts, but P2FA and MFA (Kaldi) are not.

quite far off from the manual results.[7] Vowel means from P2FA and MFA are within 6Hz of the manual F1 means, and within 20Hz of manual F2 means. Vowels whose means deviate more from manual alignment are those vowels which have the fewest tokens in the data set.

Research requiring accurate vowel space measurements may need more manual correction than is needed for prosody studies. However, preliminary forced alignment greatly speeds up the task.

3.3 CONSONANT DURATION. Figure 3 compares average duration results for each consonant across all alignment conditions. Automatic algorithms varied with respect to their accuracy in different groups of consonants. For example, the mean durations for the oral stop consonants show that DARLA and P2FA tended to pick out shorter stops than manual, while MFA trended long. For nasal stops, on the other hand, MFA trended short and P2FA long, and overall accuracy for all algorithms was improved. Glides and liquids showed greater variation, with DARLA giving long /l/ segments but being fairly accurate on other segments. MFA and P2FA trended long in their /y/durations.

Projects requiring an analysis of consonants, VOT, lengthening, etc. require accurate and consistent consonant segmentation. Consonant duration measurements were least robust to aligner error, and showed wide variation across different phonemes in Yidiny. This sort of measurement is also likely to be the most variable cross-linguistically, as the accuracy of these automated mod-els on consonant segmentation depends on the similarity of a language's consonant inventory to English. Therefore forced alignment should not be used without manual inspection and correction when this is the object of study.

[7]It should be noted that while DARLA did not work well for this project, we are trying to make it do things which it is not designed to do. These results should not be taken as implying that DARLA is a poor aligner overall. Clearly, from the results achieved on English, it performs excellently.

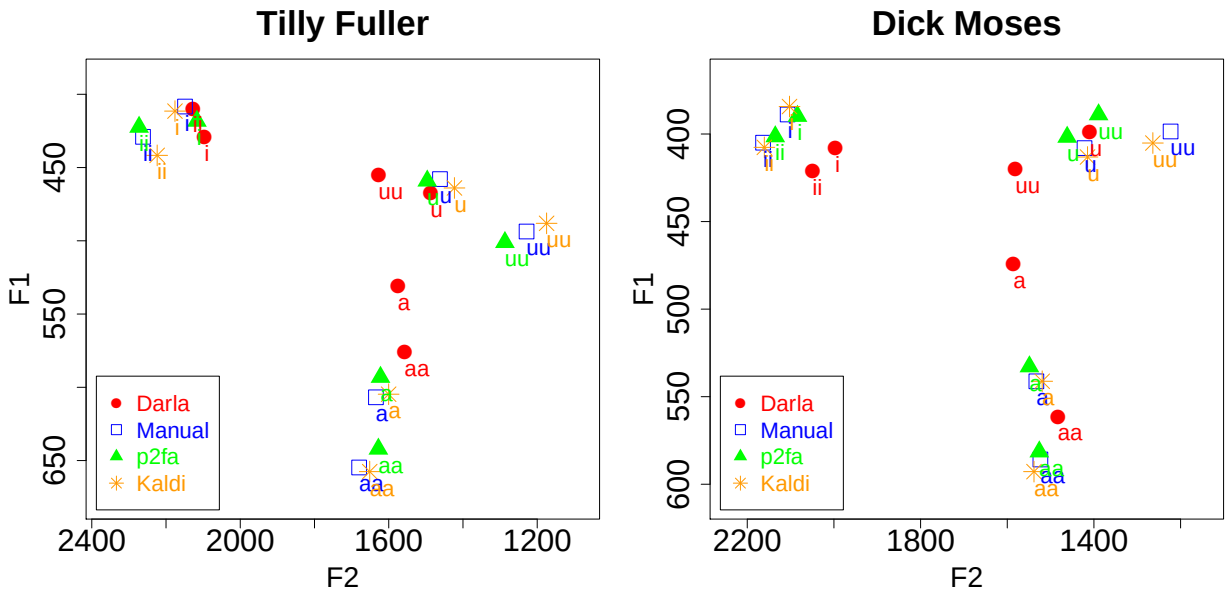


Figure 2: Space of vowel means for Yidiny speakers Tilly Fuller (left) and Dick Moses (right) across all alignment conditions. DARLA results were poor; P2FA and MFA (Kaldi) results within 6Hz for F1 and within 20Hz for F2.

**4. Implications for workflow.** The findings of this paper have implications for workflows, both of primary linguistic fieldwork and for those working with archival materials. In this section, we sketch out the workflow and procedures required to conduct an analysis of the type we did here, noting particular software requirements and analytical bottlenecks.

4.1 MINIMAL REQUIREMENTS. For forced alignment, the minimal requirement is a sound file and its transcript. If the files are not in a digital format, they need to be digitized. For example, the Yidiny sound files for our test were digitized from reel-to-reel cassettes. Some of the transcripts were retyped from scans of typewriter typescripts, while others were reformatted from Microsoft Word documents which contained Yidiny, English, and annotations.

Digital audio files need to have a sampling rate of 16KHz. Some of these forced alignment algorithms will resample in theory, but in practice we found that files that were not already down-sampled threw errors. Text files need to be in plain text format, with just the language in the file. In our case, we extracted the Yidiny line from the Word documents. Because our Word documents already used a practical orthography in common use amongst researchers on Australian languages, we retained that system. We have also made forced alignment workflows that begin with ELAN transcripts. In that case, the transcription tier can be exported to a TextGrid file directly, or optionally presegmented (see below).

Once the files are in the right digital formats, the next stage is to presegment the files to utterance-level chunks. This is required for using the Montreal Forced Aligner, which crashes with files that are longer than about 10 seconds in duration. P2FA can handle very long sound files.[8]

[8] Amalia Skilton raised the point during our presentation of this research that the file length can impact the accuracy

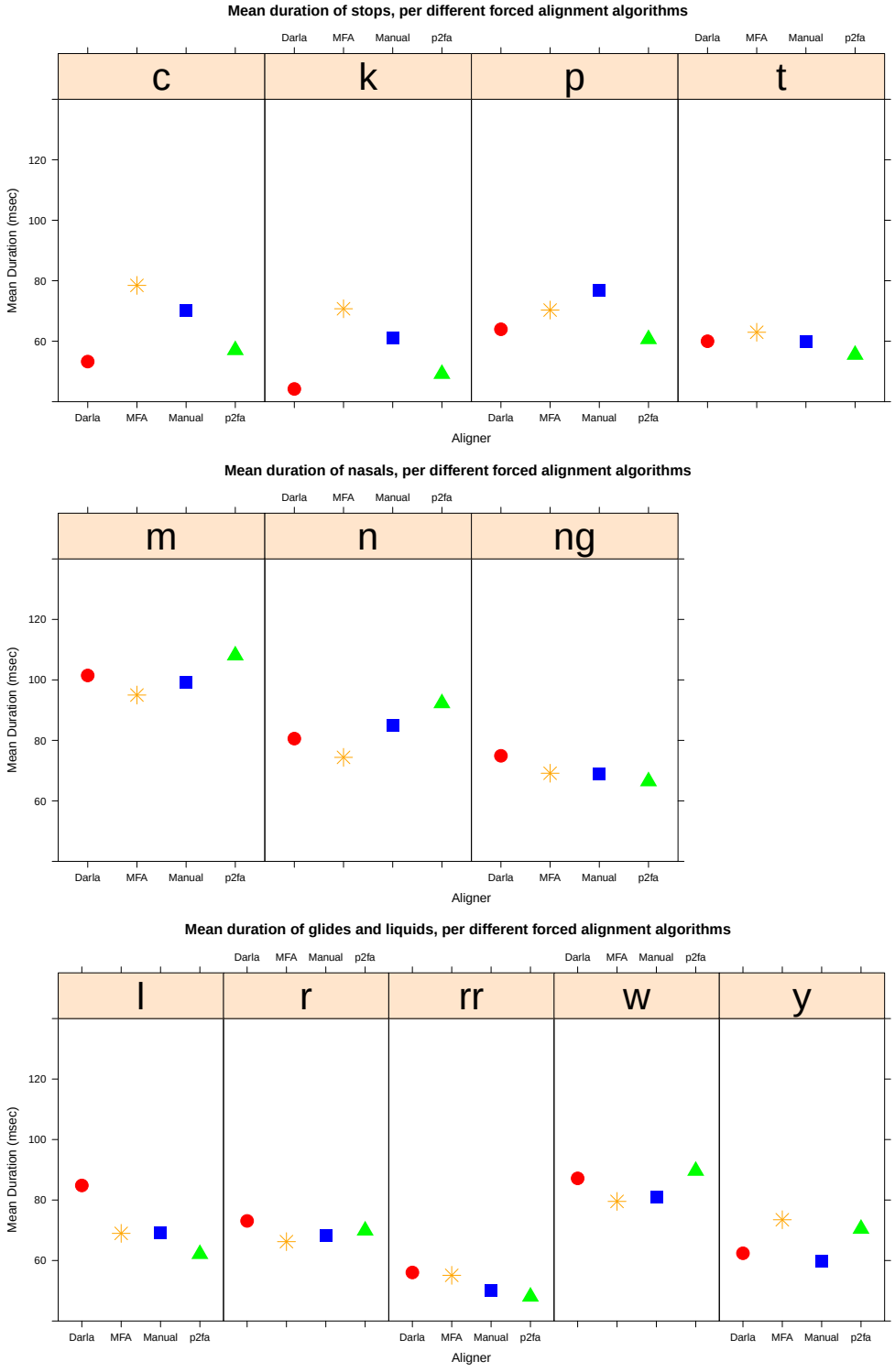


Figure 3: Mean consonant durations across all alignment conditions. Manual means indicated with blue squares. Inventory broken up into oral stops (top), nasals (middle), and glides/liquids (bottom).

Thus, if the input to the forced alignment is an ELAN transcript file which is already segmented at the utterance level, it is straightforward to extract those intervals for use with the forced aligner (either P2FA or MFA). However, if the input to the project is a paper transcript and digitized file, unless the researcher wishes to manually align at the utterance level, P2FA affords an easier choice. Presegmentation is done from utterance-level segmentations (in ELAN or Praat). If in ELAN, the files should be exported to Praat TextGrids. A script[9] is then used to extract the utterances with a boundary buffer of 80ms. It also saves the TextGrid interval as a text file. The files need to have the same filenames as each other, and these filenames cannot contain punctuation, which means that many collections will require additional preprocessing.

In order to run both MFA and P2FA, a pronouncing dictionary is required. We created this by concatenating all text files into a large text file and deleting duplicates, using the free text editor BBEdit. We then created a version of the file which transliterated the orthographic conventions into ARPABET characters (which is read by both MFA and P2FA).

Finally, the models need to be run. We did this with a shell script which called the relevant Python scripts that run the aligner on each sound file. This allowed us to avoid having to enter each filename manually.

4.2 USE CASES. Our project is a particular use case for forced alignment; that is, testing forced alignment algorithms on archival data. In the course of completing this project, we had to make choices for our data that would be different if our aims were different. In this section we document some of the implications of our choices for different use cases.

**Language documentation in progress** Our first use case is a fieldworker who is working on language documentation and needs word-level alignment of transcripts in progress. They are already transcribing manually at the utterance level but wish to use word-level alignment in the documentation. In this case, our procedures for segmenting audio files at the utterance level would not work well (at least, not as we did it), because the segmentation would be subsequently unalignable with the full transcripts. That is, the fieldworker would have to reimport or realign the utterance level segments. Alternatively, they could align the full file using P2FA, or they could write a script for 'stitching' the utterance-based TextGrids back together into a single ELAN file. Given that our methods include the timestamp of the file in the filename, this would be straightforward. However, the fieldworker would probably want to modify the timestamp to add another decimal place or two (so the 'stitching' is more accurate).

Because the fieldworker in this case only needs word-level alignment, not segment-level alignment, the methods here are probably accurate enough for most cases.

Another use case for language documentation in progress is for a fieldworker with a small corpus who wishes to use forced-aligned data to make preliminary observations about the phonetics of the language. For example, perhaps they are unsure about the phonemic status of some transcriptions and would like to plot vowel tokens to see the extent to which they cluster. Perhaps they are unsure whether secondary stress exist in the language, or whether stress ghosting (Fletcher

---

of the alignment. In our work so far, we have found that errors can be compounded, and this can affect multiple phrases. On the other hand, there is no correlation between the position in the file and the accuracy of alignment, in broad terms. Work is currently in progress to further evaluate the effects of file length on accuracy.

[9]This script is available from https://github.com/chirila/ausscripts.

and Butcher 2014) is a factor. In that case, the most important thing to do would be to set up a workflow that automates the process of adding new data. Because the end result is analytical data that feeds back into the documentation project, but does not require direct links between the alignments and the original transcripts, it is possible to use utterance level segmentations as are required for MFA.

This fieldworker should be aware that the results are likely to be accurate enough for impressionistic findings, but the alignments should be manually checked as much as possible. Note also that if the purpose of the phonetic research is to determine differences that crucially rely on segment length (like stress or gemination), alignments done automatically should be checked manually before relying on the results.

**Archival research** Another use case is where the linguist uses forced alignment to create a segment-aligned corpus for phonetic research. Here the methods outlined in this paper will probably be sufficient (with subsequent manual checking if duration results are required). However, for recordings made in the field, outside of the lab, considerable preprocessing may be required. The work of Johnson et al. (2018) makes clear that there is gain from removing extraneous noise that interferes with forced alignment.

**Community research** Another use case is when the fieldworker requires word-level segmentation to create talking dictionaries. This will work for words recorded in isolation, but probably not directly for words extracted from running speech.

4.3 FORCED ALIGNMENT WITH SPEECH TO TEXT. Finally, a note is warranted about speech to text. Forced alignment takes speech data and text data, and aligns the two. Speech to text models take speech data and create transcripts from them. We imagine an ultimate workflow where a preliminary speech to text corpus is trained on manually transcribed data. Persephone (Adams et al. 2018) is such a project. Data could then be automatically transcribed, corrected by the researcher, and then fed to a forced alignment algorithm for segment-level alignment (text to speech transcribes utterances but does not align at the segment level). Such a workflow could provide more data for language projects where untranscribed audio recordings exist.

**5. Conclusions.** For many languages without available forced alignment algorithms, data for phonetic analysis exists but is underutilized due to the prohibitively time-intensive manual alignment process. For endangered languages, such as those of Australia, this situation is further complicated by the presence of small corpora of legacy data in the form of hand-transcribed audio tape recordings (Austin 2013). This creates both a need and opportunity to leverage new technology for the documentation of these languages. One possible way, determining the accuracy of English-trained forced alignment algorithms on non-English language data, has great potential for elevating the quality and rigor of phonetic work on low-resource and under-documented languages, especially those for which there are legacy recordings but no contemporary speakers who would be available to provide training data for entirely new FA models. Our work with Yidiny shows promising results for these languages, implying that for at least some tasks, unsupervised alignment (either based on English or trained from a small corpus) is sufficiently reliable for it to be used on legacy data for low-resource languages, whether endangered or not. In particular, descriptive phonetic work on vowel inventories and prosody can be accurately captured by automatic alignment with

minimal training data. The novel use of this technology on under-resourced languages raises new possibilities for more detailed language documentation and for including more languages and data in comparative work with phonetics, phonology, and sound change.

**References**


Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018*. https://halshs.archives-ouvertes.fr/halshs-01709648v4/document.

Austin, Peter K. 2013. Language documentation and meta-documentation. *Keeping languages alive: Documentation, pedagogy and revitalization* 3–15.

Boersma, Paul et al. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5. http://www.praat.org.

Dixon, Robert M. W. 1977a. *A grammar of yidin*, vol. 19, Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Dixon, Robert M. W. 1977b. Some phonological rules in yidin. *Linguistic Inquiry* 8. 1–34.

Evanini, Keelan, Stephen Isard and Mark Liberman. 2009. Automatic formant extraction for sociolinguistic analysis of large corpora. *Tenth Annual Conference of the International Speech Communication Association.* http://languagelog.ldc.upenn.edu/myl/BayesianFormants.pdf.

Fletcher, Janet and Andrew Butcher. 2014. *Sound patterns of Australian languages*. Berlin: Walter de Gruyter.

Johnson, Lisa M., Marianna Di Paolo and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-aligner with Tongan data. *Language Documentation & Conservation* 12. 80–123. https://scholarspace.manoa.hawaii.edu/bitstream/10125/24763/1/johnson_et_al.pdf.

Kendall, Tyler and Erik R. Thomas. 2018. *vowels: Vowel manipulation, normalization, and plotting*. https://CRAN.R-project.org/package=vowels. R package version 1.2-2.

Kuznetsova, Alexandra, Per B. Brockhoff and Rune Haubo Bojesen Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82(13). https://cran.r-project.org/web/packages/lmerTest/index.html.

Lin, Cheng-Yuan, Jyh-Shing Roger Jang and Kuan-Ting Chen. 2005. Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 2, June 2005: Special Issue on Annotated Speech Corpora* 10(2). 145–166.

Nash, David. 1979. Yidin stress: A metrical account. In E. Battistella (ed.), *Proceedings of the ninth annual meeting of the North East Linguistic Society, CUNY Forum 7-8*, 112–130. New York: Queens College Press.

Patz, Elisabeth. 1991. Djabugay. In R. M. W. Dixon and Barry J. Blake (eds.), *The aboriginal language of Melbourne and other sketches*, vol. 4, Handbook of Australian Languages, 245–348. Melbourne: Oxford University Press.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian and Petr Schwarz. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and under-standing*. IEEE Signal Processing Society. http://kaldi-asr.org/.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Reddy, Sravana and James Stanford. 2015. A web application for automated dialect analysis. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 71–75. http://www.aclweb.org/anthology/N15-3015.

Yuan, Jiahong and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123(5). 3878. https://asa.scitation.org/doi/10.1121/1.2935783.