Gonzalez, Simon, Catherine E. Travis, James Grama, Danielle Barth and Sunkulp Ananthanarayan. 2018.
Recursive forced alignment: A test on a minority language.
Paper submitted to: *17th Speech Science and Technology Conference*,
University of New South Wales, Sydney.

Recursive forced alignment: A test on a minority language

Simon Gonzalez[1], Catherine E. Travis[1], James Grama[1], Danielle Barth[1], Sunkulp Ananthanarayan[2]

[1] ARC Centre of Excellence for the Dynamics of Language, Australian National University
[2] University of Texas at Austin

simon.gonzalez@anu.edu.au, catherine.travis@anu.edu.au, james.grama@anu.edu.au,
danielle.barth@anu.edu.au, sunny.a@utexas.edu

Abstract

We compare recursive and linear approaches to force-aligned data from Matukar Panau, an endangered language of Papua New Guinea. Data were force aligned with the train/align procedure in the Montreal Forced Aligner. Using manual alignments produced by a trained phonetician as a benchmark, the recursive approach was found to outperform the linear approach. The recursive approach produced alignments that overlapped more with those made by human coders, and resulted in fewer fluctuations in both Overlap Rate and Error Rate. We conclude that a recursive approach enhances the quality of automated alignment of languages lacking a pre-existing acoustic model.

Index Terms: forced alignment, accuracy, robustness, recursion, minority language

1. Introduction

Forced alignment is increasingly prevalent in phonetic research, as it dramatically increases the speed of achieving analysable data, and therefore the number of tokens that can be examined phonetically [1, 2]. Forced alignment has primarily been applied to major world languages with fully established acoustic models (in particular, English) [3, 4]. Thus, there exists a significant gap between the forced alignment resources available for minority languages, and those available for major languages. Minority languages have been force aligned using acoustic models from major languages, with varying success [5-8]. This approach is less than ideal because of the reliance it places on matching phonemic inventories and orthographic systems of often completely unrelated languages. We argue that current forced-alignment programs with a train/align procedure offer the means to effectively align data from languages which lack acoustic models, and, by applying a recursive approach, this is so even in the absence of large amounts of speech data with which to work.

A number of factors that affect the accuracy of forced alignment have been presented in the literature to date. Focus has been placed on best practices for addressing transcription errors [9], the impact of long pauses and noisy environments [10], as well as the optimal number of speakers and type of data for successful alignment [3]. Previous studies have also found that alignment accuracy tends to reach a ceiling, after which point additional data does not significantly improve the alignment. One study on spontaneous spoken English found that this ceiling was reached at five minutes of transcribed speech [3]; another, on read French data, exhibited a ceiling effect at two minutes [11]. Instead of increasing alignment quality, in some cases additional data was associated with in poorer alignment accuracy [3].

One of the more powerful tools in forced alignment is the train/align method. While some forced alignment works on the basis of a pre-existing acoustic model, with the train/align method, an acoustic model is created on the basis of the data input to the program, and that model is then applied to the forced alignment of the same input data [11]. This procedure has been successfully used to force align minority languages without established acoustic models [12, 13].

In working with minority, and under-resourced, languages, there may be limited data available, and thus maximal use must be made of the data that is available in order to build an acoustic model from scratch. The standard treatment of force-aligned data follows a linear approach, whereby the data is examined only once prior to creating a model. In contrast, in a recursive approach, the data is examined several times in different stages, and at every new stage, the algorithm learns from the previous stage and adapts accordingly [10]. It is therefore particularly valuable for working with small datasets.

This paper compares the application of a linear vs. recursive approach to force-aligned data from Matukar Panau, a minority language of Papua New Guinea, with no pre-existing acoustic model and with a moderately sized speech corpus [14, 15]. We demonstrate that the recursive approach yields very high quality alignment, and suggest that applying a recursive approach facilitates high quality forced alignments of under-described languages.

2. Methodology

2.1. Montreal Forced Aligner

The aligner chosen for this study was the Montreal Forced Aligner (MFA) [4]. MFA has been demonstrated to be more accurate than FAVE [16], MAUS [17] and Prosody lab-Aligner [18], and marginally more accurate than the train/align procedure in LaBB-CAT [19]. (See [20] for a comparison of these aligners.) One key difference is that these other forced aligners use the HTK toolkit, while MFA uses the Kaldi toolkit, which employs triphone acoustic models to better capture variability in phone realisations. Another is that MFA (like LaBB-CAT) allows for the application of the train/align method, facilitating extension to languages lacking an acoustic model.

2.2. Matukar Panau Speech Data

Matukar Panau is an endangered Oceanic language spoken by around 300 people in Madang Province, Papua New Guinea, in the village Matukar and hamlet Surumarang.

Documentation for this language is ongoing [14, 15]. It is an agglutinating, non-tonal language with 17 consonants, a small vowel inventory, and a fairly transparent orthography. There is no existing acoustic model of the language.

The data for this study come from a corpus of sixty short recordings of monologic narratives produced by 36 native speakers of Matukar Panau. The narratives were transcribed by a trained linguist in conjunction with six trained, semi-speakers of Matukar Panau (native speakers of Tok Pisin who have familiarity with Matukar Panau). Transcription was done at the utterance level, using a phonemically transparent orthography. For this study, we worked with 3.75 hours (or 225 minutes) of transcribed speech. We built a dictionary to map the 2,468 word types that occur in this sub-corpus to their phonemic representations, and force aligned the audio files using the train/align procedure in MFA. Data were then prepared following both a recursive and linear approach, as described below.

2.3. Data Recursion

How much data is required for effective forced alignment when working with a language with no acoustic model? To test the quality of forced alignment with different quantities of data, we needed to create subsets of the data of different lengths. To control for speaker effects, we had to include multiple speakers in each subset. Thus, a script was written to create a TextGrid file in Praat [21] which separated all files into increments of 30 seconds. The starting point was one minute from each of four files representing four speakers. From there, we created five-minute iterations by drawing 30-second increments from each file, until the file's duration was exhausted, at which point we drew a 30-second increment from a new file. We increased the data being aligned by five-minute iterations, until we reached the maximum duration of transcribed speech (225 minutes). The 30-second increments ensure that speakers were equally represented at each five-minute iteration, controlling for speaker effects at each step. The five-minute iterations allow for the quality of the forced alignment with different amounts of data to be compared, to identify the point at which alignment quality is optimised. This process resulted in a mean of 487 word tokens and 2,444 segments per five-minute iteration.

Two datasets were prepared from the forced alignment output: a *linear dataset* and a *recursive dataset*. Figure 1 provides a representation of the difference between the two. Linear processing is the default for forced alignment. To prepare the linear dataset for this study, the force-aligned boundaries were *reset* at each five-minute iteration. In contrast, the force-aligned boundaries for the recursive dataset were *adjusted*. That is, for each iteration, the alignment was recalculated based on the information from the current iteration and from previous iterations, utilising an algorithm that was written for this process. This methodology is adapted from [10], which utilised a recursive algorithm to improve forced alignment in long audio segments. A recursive algorithm works by inspecting the data multiple times; at every new iteration, new information is added, re-evaluated, fed back into the existing information from previous iterations, and then applied at the next iteration. This output then serves as the basis for analysis.

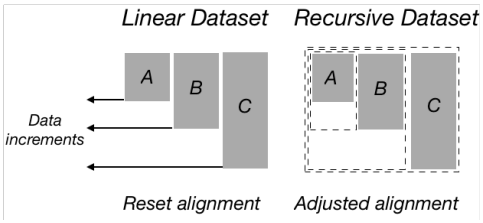

Figure 1 *Alignment process according to linear and recursive datasets*

3. Measures of alignment quality

The standard for determining the quality of forced alignment is comparison with a human benchmark. One method available for doing this is a comparison between boundaries placed by forced alignment vs. boundaries placed by a trained phonetician [3, 22]. Following the same protocol as that applied in previous work, we selected two speakers (one male and one female), and manually corrected the automatic alignment of the first 60 seconds of the file of each speaker. These 549 segments (261 consonants and 288 vowels) serve as the benchmark against which the automatic alignments are compared.

Two quality measurements are employed in this study: *accuracy* and *robustness* [cf., 23]. *Accuracy* was operationalized as the time difference between the placement of a boundary as the result of forced alignment vs. the human benchmark. *Robustness* is the rate of alignment error based on a specified boundary threshold, here set at 20 *ms*, following [3]. Any force-aligned boundary placed greater than 20 *ms* from the human benchmark is classified as an alignment error.

These measurements provide different indications of the quality of the forced alignment. An alignment may have high accuracy but not be robust if there are a large number of alignments that occur just beyond the 20 *ms* threshold, but close to that threshold. On the other hand, an alignment can have robust alignments with low accuracy if there are few alignments beyond 20 *ms*, but those that are beyond 20 *ms* are at a high degree of distance from the benchmark. Together, these measures provide a strong indication of the overall quality of the resulting alignments.

3.1. Accuracy Measurement

For accuracy, we calculated Overlap Rate (OvR) [3, 23], that is, the proportion of overlap between the intervals established by the human coder and the intervals established by the forced aligner. Greater overlap is associated with greater accuracy. The time representation is shown in Figure 2.

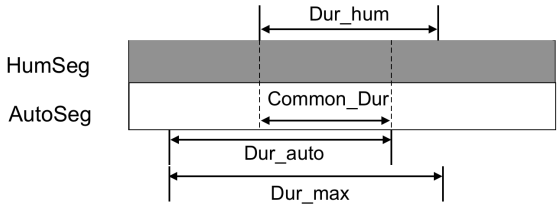

Figure 2: *Representation of Overlap Rate*

Common_Dur is the time shared by the interval created by the human coder (*Dur_hum*), and that created automatically, by the forced aligner (*Dur_auto*). This is measured as a

proportion of the duration from the earliest onset and latest offset boundary of the two intervals (*Dur_max*). The Overlap Rate was calculated for both linear and recursive datasets.

3.2. Robustness Measurement

For robustness, we are interested in the proportion of boundaries that lie beyond a pre-determined threshold, here 20 ms. Following [13], this was calculated on the basis of the difference between the midpoint of the manually created interval and the force-aligned interval. Figure 3 shows a hypothetical midpoint of an interval produced by a human coder, and two hypothetical midpoints from distinct forced alignments. If a force-aligned midpoint falls within 20 ms of a manual midpoint (as for (a)), it is considered a non-error; if it is at a greater distance (as for (b)), it is considered an error. The Error Rate is the ratio of total number of errors to non-errors. As a further measure of robustness, we calculated the mean distance from the manual midpoint of the error tokens. Higher mean distances correspond to less robust tokens, thus less reliable alignments. The two measures of robustness were calculated for both linear and recursive datasets.

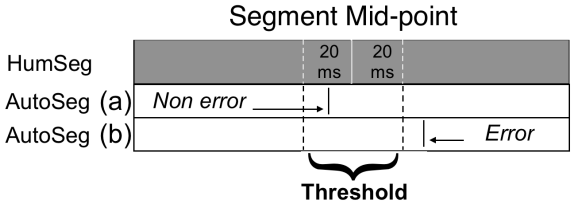

Figure 3: *Representation of errors, using a 20 ms threshold*

4. Results

4.1. Accuracy

Figure 4 compares the Overlap Rate for the linear and recursive datasets across iterations for the two speakers who were manually aligned. The greater accuracy for the recursive dataset, captured with the solid line, can be seen in three ways (all of which hold for each speaker).

First, the overall mean Overlap Rate is higher in the recursive than the linear dataset. Second, the recursive dataset exhibits fewer fluctuations than the linear dataset, suggesting that a recursive approach smooths out major alignment errors. And third, in the recursive dataset, there is a marked increase in Overlap Rate up to 35 minutes of data; beyond this point, the increase is more gradual. In comparison, the linear dataset retains significant fluctuations throughout, though they become less marked from approximately 125 minutes. Thus, the recursive dataset improves more rapidly, follows a steadier trajectory, and overlaps more with alignments placed by a human coder than the linear dataset.

Evidence of the quality of the alignment can be seen by comparing these results with the findings of [3] for English. The Overlap Rate of 0.67 attained at 35 mins in the recursive dataset here is at the upper end of the range reported in [3]; but in [3], this was attained earlier, with just five minutes of data.

4.2. Robustness

The overall Error Rate, that is, the proportion of midpoints determined automatically that occurred at a distance of greater than 20 *ms* from the human benchmark, was very similar across the two approaches (recursive dataset = 21.5%, linear dataset = 23.2%). Overall mean distance from the human benchmark for tokens classified as errors was also similar (recursive dataset = 93.6 *ms*, linear dataset = 94.9 *ms*). Thus, according to this measure of robustness, the recursive dataset produces only marginally better results.

However, mean distances across data iterations differ. Figure 5 shows the mean distance from the benchmark across data iterations for tokens classified as errors. Here we see that, while the recursive dataset exhibits a steadier trajectory throughout, the linear approach is characterized by heavy fluctuation, with pronounced differences between peaks and troughs. As with Overlap Rate, the recursive approach seems to be able to soften the impact of errors more efficiently than the linear approach.

We also see here that both linear and recursive datasets show stabilisation at approximately 35 minutes (a similar point at which improvement in accuracy according to Overlap Rate begins to diminish). And both datasets show a decrease in robustness at the latter stages of data iteration—the linear dataset exhibits striking variability, and the recursive dataset shows a gradual increase in mean distance. One possible explanation is that the greater number of speakers included at these latter stages may result in more variability, and we leave this for future exploration.

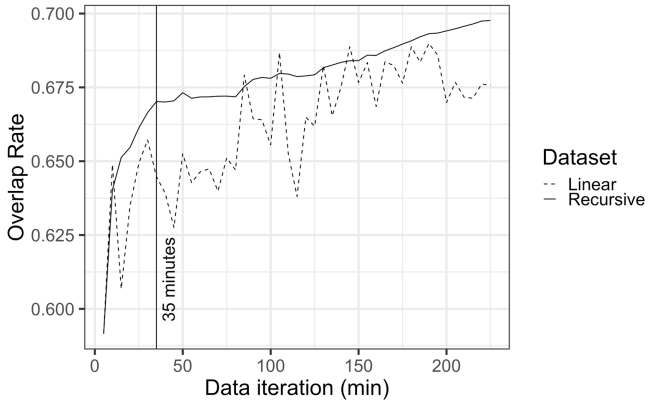

Figure 4: *Overlap Rate across data iterations: Linear and Recursive datasets*

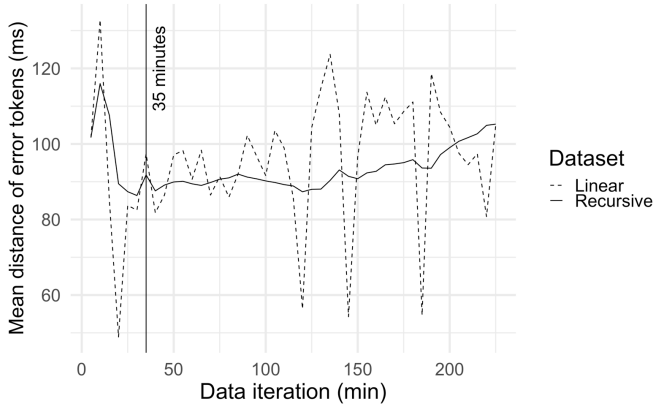

Figure 5: *Mean distance of error tokens across data iterations: Linear and Recursive datasets*

5. Discussion

In this study, we applied a recursive approach to forced aligned data from Matukar Panau, a minority language of Papua New Guinea lacking an acoustic model. To do this, we utilised the train/align procedure in MFA, and aligned the same data at different iterations, from 5 to 225 minutes. We then applied two approaches to prepare data for analysis: a linear approach, where alignment values are reset at each iteration, and a recursive approach, where alignment values are adjusted based on previous iterations. Results indicate that a recursive approach outperforms a traditional linear approach—forced alignments derived via recursion were more accurate, with a higher rate of overlap between manually and automatically placed boundaries. The recursive approach was also more robust than the linear approach in the sense that it was less susceptible to major alignment errors. This suggests that the algorithm learns as more data is processed, in line with observations that adjustments early on in alignment improve alignment in later stages [10]. In this way, the recursive approach may be able to protect alignment in sections of audio files (or whole audio files) that prove challenging to aligners. As the recursive approach employs an algorithm that is self-correcting, these mistakes can be adjusted for as the data is processed; the algorithm learns from these examples and this ultimately improves the alignment later in the data stream.

6. Conclusions

The recursive approach to the force-aligned data outperforms traditional linear implementations, and yields highly accurate alignment, even from a relatively small dataset—here, 35 minutes of transcribed speech was sufficient to achieve high quality alignment. This method expands the potential for large-scale phonetic and sociophonetic studies for under-resourced minority languages, and is a very promising step towards making available to minority languages tools that to date have been primarily utilised for work on major world languages. Future studies on languages of different types will further advance methods and the ability to obtain the best results for automated phonetic alignment.

7. Acknowledgements

We gratefully acknowledge support from an ARC Centre of Excellence for the Dynamics of Language Transdisciplinary & Innovation Grant (TIG952018), as well as the local transcription team (Justin Willie, Rudolf Raward, Amos Sangmei, Alfred Sangmei, Michael Balias and Zebedee Kreno), and expert consultant (Kadagoi Rawad Forepiso).

8. References

[1] Labov, W., Rosenfelder, I., and Fruehwald, J., “One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis,” Language, 89(1):30-65, 2013.

[2] Schiel, F. et al., "The Production of Speech Corpora," 2012.

[3] Fromont, R., and Watson, K., “Factors influencing automatic segmental alignment of sociophonetic corpora,” Corpora, 11(3):401-431, 2016.

[4] McAuliffe, M. et al., “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” Proceedings of the 18th Conference of the International Speech Communication Association, 2017.

[5] Kempton, T., "Cross-language forced alignment to assist community-based linguistics for low resource languages," Paper presented at the 2nd Workshop on Computational Methods for Endangered Languages, ComputEL-2, Honolulu. 2017.

[6] Kempton, T., Moore, R. K., and Hain, T., “Cross-language phone recognition when the target language phoneme inventory is not known,” INTERSPEECH-2011:3165-3168, 2011.

[7] Kurtic, E. et al., “A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English,” Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC’12, Istanbul, Turkey, 1323-1327, 2012.

[8] Strunk, J., Schiel, F., and Seifart, F., “Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS,” Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC’14, Reykjavik, Iceland, 3940–3947, 2014.

[9] Bordel, G., Penagarikano, M., Rodriguez-Fuentes, L. J., and Varona, A., “A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions,” INTERSPEECH-2012:1840-1843, 2012.

[10] Moreno, P. J., Joerg, C. F., Van Thong, J.-M., and Glickman, O., "A recursive algorithm for the forced alignment of very long audio segments," Paper presented at the 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Australia. 1998.

[11] Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R., “Automatic phone alignment: A comparison between speaker-independent models and models trained on the corpus to align,” Proceedings of the 8th International Conference on NLP, JapTAL:300–311, 2012.

[12] DiCanio, C. et al., “Assessing agreement level between forced alignment models with data from endangered language documentation corpora,” INTERSPEECH-2012:130-133, 2012.

[13] Coto-Solano, R., and Flores, S., “Comparison of two forced alignment aystems for aligning Bribri speech,” CLEI Electronic Journal, 20(1):2:1-2:13, 2017.

[14] Anderson, G. D. S., Barth, D., and Rawad Forepiso, K., “The Matukar Panau online talking dictionary: Collective elicitation and collaborative documentation,” Language documentation and cultural practices in the Austronesian world, Papers from 12-ICAL, Canberra, Australia, 111-126, 2015.

[15] Barth, D., “Matukar Panau Language Documentation (DGB1), Digital collection managed by PARADISEC. [Open Access] DOI: 10.4225/72/56E97A2420C64,” 2010.

[16] Rosenfelder, I. et al., "FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2 10.5281/zenodo.22281," 2014.

[17] Schiel, F., “Automatic phonetic transcription of non-prompted speech,” ICPhS-14:607-610, 1999.

[18] Gorman, K., Howell, J., and Wagner, M., “Prosody lab-aligner: A tool for forced alignment of laboratory speech,” Canadian Acoustics, 39(3):192-193, 2011.

[19] Fromont, R., and Hay, J., “LaBB-CAT: An annotation store,” Proceedings of the Australasian Language Technology Workshop:113-117, 2012.

[20] Gonzaléz, S., Grama, J., and Travis, C. E., “Comparing the accuracy of forced-aligners for sociolinguistic research,” Poster presented at CoEDL Fest, University of Melbourne (available at: https://cloudstor.aarnet.edu.au/plus/s/gyC6vuX5uvc5soG - pdfviewer), 2018.

[21] Boersma, F., J., and Weenink, D., Praat: Doing phonetics by computer [Computer Software] Amsterdam: Department of Language and Literature, University of Amsterdam. Retrieved from http://www.praat.org/, 2011.

[22] Cosi, P., Falavigna, D., and Omologo, M., “A preliminary statistical evaluation of manual and automatic segmentation discrepancies,” 1991.

[23] Paulo, S., and Oliveira, L. C., "Automatic phonetic alignment and its confidence measures," Advances in Natural Language Processing: 4th International Conference, EsTAL 2004, Luis Vicedo, J. et al., eds., 36-44, Berlin / Heidelberg: Springer-Verlag 2004.