UNITE AND CONQUER: BOOTSTRAPPING FORCED ALIGNMENT TOOLS FOR CLOSELY-RELATED MINORITY LANGUAGES (MAYAN)

Kevin Tang[a] and Ryan Bennett[b]

Zhejiang University[a] and University of California, Santa Cruz[b]
linguist@kevintang.org, rbennett@ucsc.edu

ABSTRACT

Forced alignment, a technique for aligning segment-level annotations with audio recordings, is a valuable tool for phonetic analysis. While forced alignment has great promise for phonetic fieldwork and language documentation, training a functional, custom forced alignment model requires at least several hours of accurately transcribed audio in the target language—something which is not always available in language documentation contexts. We explore a technique for model training which sidesteps this limitation by pooling smaller quantities of data from genetically-related languages to train a forced aligner. Using data from two Mayan languages, we show that this technique produces an effective forced alignment system even with relatively small amounts of data. We also discuss factors which affect the accuracy of training on mixed data sets of this type, and provide some recommendations about how to balance data from pooled languages.

Keywords: forced alignment, field phonetics, small data, representativeness, Mayan languages

1. INTRODUCTION

The transcription and annotation of field recordings pose major logistical challenges for phonetic fieldwork. The effort required to annotate many hours of acoustic data can be prohibitive, and as a result fieldworkers often collect far more data than they ever fully annotate or transcribe. The recent development of accessible forced alignment tools for linguistics has helped reduce the severity of this problem. Forced alignment is a computational technique which can be used to semi-automatically time-align phonetic transcriptions with associated audio recordings, at both the word and phone levels [11]. Though extremely promising, forced alignment is not always practical for languages which are relatively under-documented [5]. Training an accurate forced alignment model typically requires at least several hours of recorded and transcribed data. Indeed, most forced aligners for majority languages like English have been trained on hundreds or even thousands of hours of transcribed data—resources which will likely never be available for most minority languages.

How, then, can forced alignment be exploited for under-documented languages? One approach is to apply models trained on majority languages (e.g. English) to the target language ('cross-language forced alignment', CLFA). This approach raises several challenges. First, the researchers must decide which existing language model should be used for alignment. Second, researchers must decide how to map the phones of the target language to phones in the majority language which the alignment model was trained on. Given the high researcher degree of freedom involved, it is perhaps unsurprising that the overall performance of CLFA varies considerably across studies and parameters [5, 11–13, 16].

A different approach is to train a model directly on the target language, using existing transcriptions (language-specific forced alignment, LSFA). This approach has been especially facilitated by two forced aligners: the Prosodylab-Aligner (PLA) [8], and the Montreal Forced Aligner (MFA) [14]. PLA in particular has been used to produce time-aligned transcriptions for a number of minority languages [11]. This approach has advantages over CLFA: there are fewer researcher degrees of freedom; language-specific alignment models produce better alignment quality; and models can be further refined as more data is collected and transcribed. Nonetheless, LSFA does not overcome the central problem that large sets of transcribed recordings are simply unavailable for most languages.

In this study, we employed a method which uses LSFA to train alignment models on mixed data from two phonetically-similar and genetically-related languages. This approach combines strengths of both CLFA and LSFA: we harness data from a non-target language (henceforth NTL) to improve alignment, but we train a model for the target language (henceforth TL) specifically, which can be updated

as more data is collected. We illustrate this approach with two languages from the K'ichean branch of the Mayan family, Kaqchikel and Uspanteko [3]. Previous work suggests that alignment models trained on multiple languages improve when the languages in the training set are genetically similar [9]. Using genetically similar languages also simplifies the process of mapping phones between languages [5], as Kaqchikel and Uspanteko have very similar phonological and phonetic inventories.

We focus on two issues here. First, we explore in detail how additional NTL data might benefit the alignment of a TL, and consider the limitations of this approach. Second, we ask whether the *kind* of NTL data used for model training might affect how well the model ultimately aligns TL data. We investigate these questions by manipulating (i) the *amount* of TL and NTL data used to train alignment models (Exp. 1), and (ii) the *representativeness* of NTL data used in model training (Exp. 2).

2. METHOD

We used the MFA forced aligner (v1) for this study because it has some advantages over PLA, e.g. using triphone windows to capture context-dependent acoustic variability for each phone. The PLA aligner should also produce similar results, as it shares an underlying architecture with the MFA aligner.

All of the default MFA parameters were used in this study. Preliminary testing showed that training speaker-specific alignment models with MFA consistently improved alignment quality [15, 18]. To ensure the robustness of our results, for each model specification, five sets of training data were generated by random sampling from our full data set. The results from the five models were then averaged.

2.1. Acoustic and alignment data

Our target language (TL) was Kaqchikel, and our non-target language (NTL) Uspanteko. The Kaqchikel data consists of spontaneous monologues from 16 speakers, recorded in Sololá, Guatemala in 2013. These recordings include ∼32,000 phones. A random subset of these (∼4,500 phones, drawn roughly equally from all 16 speakers) was aligned using the PLA and then hand-corrected. These hand-corrected alignments serve as the gold-standard to which we compare the alignments generated by our alignment models. The Uspanteko data consists of sentences elicited from 10 speakers in 2017. These recordings include ∼13,000 phones.

2.2. Manipulation: Sample size

We manipulated the amount of TL and NTL data in the input to model training by varying the number of phones included from each language (= 0, 500, 1000, 2000, 4000, or 8000 phones).

2.3. Manipulation: Sampling with representativeness

We define the 'representativeness' of an NTL sample in this study as how closely that NTL sample matches the distribution of phones in the TL training data. In addition to random sampling of data at different phone counts (Exp. 1), we considered how alignment is affected when samples of the NTL data are most representative (closest to the TL), and least representative (furthest from the TL).

We determined the representativeness of each NTL sample with a measure known as *tf-idf* [10]. Our NTL and TL data is composed of sets of .WAV files and transcriptions corresponding roughly to utterances. For each phone P_x in an NTL or TL utterance, we computed the normalized frequency of that phone relative to that utterance (tf = raw frequency of P_x in the utterance / total # of phones in the utterance). We then computed how evenly distributed that phone is within the sample ($idf = log_e$(total # of utterances / # of TL utterances containing P_x)). Intuitively, the product $tf \times idf$ then expresses how 'important' a given phone is within a particular sample, balancing frequency with dispersedness.

Each utterance in a sample can then be expressed as a vector of $tf \times idf$ values reflecting the phones it contains. To compute the representativeness of each NTL utterance, we computed the cosine similarity between that utterance and all of the TL utterances included in that experiment. The average cosine similarity of an NTL sample then expresses its representativeness relative to the TL data.

2.4. Pronunciation data

Kaqchikel and Uspanteko have very shallow orthographic systems, which makes it straightforward to convert orthographic transcriptions to phonetic form. Two custom grapheme-to-phone converters were used to create pronunciation dictionaries for each language, taking into account some allophonic detail. In both languages, plain stops are aspirated word finally, and vowel initial words undergo initial glottal stop insertion [1, 4, 6]. Glottalized stops are contrastive: implosive /ɓ/ varies phonetically between [ɓ ʔ w], and /ʔ/ is sometimes deleted. In Kaqchikel, sonorants also devoice in coda position. These patterns of allophony and variability were in-

cluded in the dictionary entries used for alignment.

Phonetically, the two languages differ mainly in their vowel inventories (though Uspanteko also permits a wider range of consonant clusters). Vowels in Uspanteko include long and short /aeiou/, and contrast for tone [2]. Vowels in Kaqchikel include tense /aeiou/ and lax /əɛɪɔʊ/, and do not contrast for tone. In unstressed syllables, only short vowels (Uspanteko) and tense vowels (Kaqchikel) occur. In stressed syllables, there are parallel contrasts which reflect the historical relatedness of these languages [3]: long vowels (Uspanteko) correspond to tense vowels (Kaqchikel), and short vowels correspond to lax vowels (e.g. [ˈt͡ʃáːχ]∼[ˈt͡ʃaχ] 'ash' and [ˈt͡ʃaχ]∼[ˈt͡ʃəχ] 'pine'). Based on these correspondences, we formulated three mapping rules to convert Uspanteko words (NTL) into pseudo-Kaqchikel forms (TL): (i) the tonal contrast is ignored; (ii) in stressed syllables, long and short vowels in Uspanteko are mapped to tense and lax vowels, respectively; (iii) in unstressed syllables, short vowels in Uspanteko are mapped to tense vowels.

3. RESULTS

We computed errors relative to our hand-corrected standard for the onset boundaries of each phone, as predicted by each model. The results were similar when using phone offsets instead, because phone offsets are usually also the onset of the next phone.

3.1. Experiment 1: Sample size

3.1.1. Average error

Table 1 shows the average error size (in ms) with different sizes and proportions of TL and NTL training data. Adding NTL data clearly improves alignment of the TL, under at least some conditions. When there are zero TL phones in the training data (first column), there is a steady reduction in average error from 500 phones (215ms) all the way to 8000 phones (42ms) of NTL data. However, the positive effect of adding NTL phones diminishes after about 3000-4000 total input phones (in any proportion of NTL:TL). Lastly, NTL phones do not contribute as much to alignment quality as TL phones. By comparing equivalent cells along the diagonal, the cells at the top right are consistently lower than those at the bottom left. This demonstrates that adding NTL data leads to a smaller increase in alignment quality than adding TL data to the training input.

Table 1: Average error in onset boundary placement (in ms) relative to gold-standard annotations, for different amounts of NTL+TL training data, averaged across 5 samples in each cell.

NTL \ TL	0	500	1000	2000	4000	8000
0		190	118	39	32	31
500	215	116	60	37	33	31
1000	150	76	44	37	31	29
2000	72	59	37	32	31	28
4000	49	40	37	31	31	30
8000	42	35	35	32	29	27

3.1.2. Accuracy thresholds

Table 2 shows the percentage of alignments accurate within two tolerances (error size <20ms, <30ms) for different combinations of TL and NTL training data. (For reference, inter-annotator agreement on hand-aligned data is about 80% within a 20ms tolerance, [5, 11].) Some of the findings shown in Table 2 differ from what was suggested by average error rates in Table 1. Here, the positive effect of adding NTL phones levels out around 8000 total phones rather than 4000, with accuracies of about 72% for a <20ms error threshold, and 82% at <30ms.

Table 2: % accurate alignment relative to gold-standard annotations at different tolerances (20ms, 30ms), for different amounts of NTL+TL training data, averaged across 5 samples in each cell.

Tolerance < 20ms

NTL \ TL	0	500	1000	2000	4000	8000
0		15.0%	27.5%	63.2%	68.3%	71.7%
500	13.7%	33.0%	49.7%	63.7%	68.2%	72.5%
1000	22.7%	45.5%	60.4%	62.1%	71.4%	73.3%
2000	44.2%	54.6%	62.4%	68.4%	70.8%	73.1%
4000	56.4%	64.4%	65.1%	70.2%	72.3%	73.9%
8000	62.4%	66.9%	67.2%	71.0%	72.4%	73.7%

Tolerance < 30ms

NTL \ TL	0	500	1000	2000	4000	8000
0		21.8%	38.6%	75.8%	79.5%	82.5%
500	18.8%	43.2%	62.2%	76.2%	80.1%	82.6%
1000	31.2%	57.7%	72.5%	76.1%	82.0%	83.8%
2000	57.0%	67.4%	76.2%	79.6%	81.3%	83.3%
4000	68.7%	75.9%	77.4%	81.1%	82.2%	83.2%
8000	73.4%	78.2%	78.7%	80.9%	82.7%	83.7%

3.2. Experiment 2: Representativeness

3.2.1. Average error

Table 3 shows the average alignment error (in ms) for different combinations of TL and NTL data, as the representativeness of NTL data varies. Cell values are differences in mean alignment error between most and least representative samples: positive val-

ues indicate an advantage for models using *less* representative NTL samples, while negative values indicate an advantage for *more* representative samples.

The results suggest that sampling less representative NTL data yields a better model than sampling more representative NTL data. However, this sampling effect diminishes and (again) levels out for models with at least 4000 total phones.

Table 3: Differences (MOST-LEAST representative) for average error in onset boundary placement (in ms) relative to gold-standard annotations, for different amounts of NTL+TL training data, averaged across 5 samples in each cell.

NTL \ TL	500	1000	2000	4000	8000
500	27	5	4	-2	-1
1000	47	17	8	2	-2
2000	18	-4	1	0	2
4000	12	7	0	-1	2
8000	2	3	1	1	0

3.2.2. Accuracy thresholds

Table 4 shows differences for the percentage of alignments accurate within two tolerances (<20ms, <30ms), when comparing models using more representative NTL data to those using less representative NTL data. Here, negative values indicate an advantage for models using less representative NTL samples. Unlike the results based on average error size, the effect of representativeness persists across most combinations of TL and NTL data. Even at 16,000 total phones (8000 TL and 8000 NTL), using less representative NTL data still provides at least a $\sim$3.5% increase in accuracy across tolerances.

4. CONCLUSION

We have shown that combining two genetically-similar languages as the input to model training can improve the quality of forced alignment for at least one of those languages, even with a rather small amount of data. This has implications for work on endangered languages: as long as a suitable amount of NTL data is available, it can be leveraged to improve the alignment of TL data. This could be useful in any scenario where the amount of TL data is limited, e.g. the early stages of data collection, or even cases when the TL is no longer spoken. In future work we intend to explore whether this method can be extended to languages which have similar segmental phonologies, but which are not genetically related (e.g. Kaqchikel and Quechua [17]).

Exp. 1 found diminishing returns for the addition of NTL data during model training. Our results suggest some cut-off points for the usefulness of NTL data, which could provide ballpark figures for the minimum amount of data needed to produce usable forced alignment models (e.g. at least 8000 phones). Crucially, fairly good alignments are produced at these cut-off points, with levelled-off accuracy values of 72% for a tolerance of 20ms, and 82% for 30ms. These values are comparable to models trained on thousands of hours of recordings: for instance, an MFA aligner trained on 1000 hours of English data [14] aligned the *Phonsay* corpus with 72% accuracy within 25ms error. Still, the amount of data needed to train a minimal working aligner may be language-dependent, e.g. the EasyAlign aligner needs different amounts of training data to produce accurate models for French vs. Taiwan Min [7].

Table 4: Differences (MOST-LEAST representative) for % accurate alignment relative to gold-standard annotations at different tolerances (20ms, 30ms), for different amounts of NTL+TL training data, averaged across 5 samples in each cell.

Tolerance < 20ms

NTL \ TL	500	1000	2000	4000	8000
500	0.70%	1.22%	-6.06%	-0.56%	-1.79%
1000	-11.48%	-4.54%	-4.03%	-4.97%	-0.52%
2000	-11.68%	-2.02%	-4.71%	-2.06%	-5.04%
4000	-8.00%	-5.38%	-8.20%	-4.68%	-4.20%
8000	-7.55%	-7.14%	-5.12%	-6.32%	-5.38%

Tolerance < 30ms

NTL \ TL	500	1000	2000	4000	8000
500	-0.24%	1.25%	-5.06%	-0.57%	-1.10%
1000	-13.94%	-5.51%	-4.00%	-3.14%	-0.24%
2000	-12.33%	-0.61%	-3.43%	-2.44%	-2.94%
4000	-7.76%	-5.81%	-5.35%	-3.02%	-2.60%
8000	-6.50%	-5.48%	-4.12%	-4.05%	-3.52%

Exp. 2 found that the representativeness of NTL data affects alignment quality: less representative samples provide greater model improvement. We speculate that less-representative NTL samples increase the diversity of phones and phone sequences in the training data, leading to greater balance in the number of tokens of each phone/sequence, and thus improving alignment. In future work we will investigate whether the alignment of some phones (e.g. rare phones) benefits more from the addition of less representative NTL data. More generally, these results suggest that researchers can improve the quality of alignment by carefully selecting the kind of NTL data included during model training.[1]

[1] This research was partially supported by NSF grant BCS/DEL-1757473 to Bennett.

REFERENCES

[1] Bennett, R. 2016. Mayan phonology. *Language and Linguistics Compass* 10(10), 469–514.

[2] Bennett, R., Henderson, R. 2013. Accent in Uspanteko. *Natural Language & Linguistic Theory* 31(3), 589–645.

[3] Campbell, L. 1977. *Quichean linguistic prehistory* volume 81 of *University of California Publications in Linguistics*. Berkeley, CA: University of California Press.

[4] Can Pixabaj, T. 2007. *Gramática descriptiva Uspanteka*. Antigua, Guatemala: Oxlajuuj Keej Maya' Ajtz'iib' (OKMA).

[5] 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.

[6] García Matzar, P. O., Toj Cotzajay, V., Coc Tuiz, D. 1999. *Gramática Kaqchikel*. Antigua, Guatemala: Proyecto Lingüístico Francisco Marroquín.

[7] Goldman, J.-P. 2011. *EasyAlign: an automatic phonetic alignment tool under Praat*. Interspeech'11, 12th Annual Conference of the International Speech Communication Association. ID: unige:18188.

[8] Gorman, K., Howell, J., Wagner, M. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3), 192–193.

[9] Imseng, D., Bourlard, H., Garner, P. N. 2012. Boosting under-resourced speech recognizers by exploiting out-of-language data-case study on Afrikaans. *Spoken Language Technologies for Under-Resourced Languages* 60–67.

[10] Itoh, N., Sainath, T. N., Jiang, D. N., Zhou, J., Ramabhadran, B. 2012. N-best entropy based data selection for acoustic modeling. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE 4133–4136.

[11] Johnson, L. M., Di Paolo, M., Bell, A. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation & Conservation* 12, 80–123.

[12] Kempton, T. March 6–7 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. Paper presented at ComputEL-2, Honolulu.

[13] Kurtic, E., Wells, B., Brown, G. J., Kempton, T., Aker, A. May 2012. A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English. Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., (eds), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey. European Language Resources Association (ELRA).

[14] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. 2017. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association* 498–502.

[15] Peters, A., Tse, H. 2016. Evaluating the efficacy of prosody-lab aligner for a study of vowel variation in Cantonese. Workshop on Innovations in Cantonese Linguistics (WICL 3), The Ohio State University, Columbus, OH.

[16] Strunk, J., Schiel, F., Seifart, F. May 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* Reykjavik, Iceland. European Language Resources Association (ELRA).

[17] Weber, D. J. 1989. *A Grammar of Huallaga (Huánuco) Quechua*. University of California Press.

[18] Wilbanks, E. 2015. The development of FASE (Forced Alignment System for Español) and implications for sociolinguistic research. New Ways of Analyzing Variation 44, Toronto, Canada.