Corpus Phonetics for Under-Documented Languages: A Vowel Harmony Example

Timothy Kempton and Mary Pearce
SIL International

1 Introduction

1.1 *Corpus phonetics* Corpus phonetics is transforming the analysis of acoustic and articulatory recordings in the same way that corpus linguistics has transformed the analysis of transcriptions. This "semiautomatic analysis of digital speech collections" (Liberman, 2019) is opening up new opportunities in phonetics and phonology and inspiring "a movement from the study of small, mostly artificial datasets to the analysis of published corpora of natural speech that are thousands of times larger" (Yuan et al., 2018). Although many linguists want to use corpus phonetics in their own research, there are still barriers, such as the particular skills needed to use the software and some of the challenges in applying the tools to under-documented languages. In this paper, we describe the development of a language-independent workflow. We used this workflow to analyse the vowel system of a recently documented language, replicating existing results and revealing new perspectives on the speech data.

1.2 *Kera* The reduction of vowels is a popular topic for research, but little has been said about the effects of vowel harmony on vowel reduction. In Kera (Chadic) it has been shown (Pearce, 2012), that not only is phonetic reduction linked to the phonetic duration of the vowel, but also that reduction is blocked in vowel harmony domains.

A pilot study of several other languages suggests this same hypothesis concerning vowel reduction is true for them too. For the study in Kera, the statistical results are compelling, but for the pilot study, which included several under-documented languages, the task of collecting enough data, applying the correct harmony rules and then measuring formants manually meant that the data were insufficient to give a confident result. A corpus phonetics approach would allow us to test the hypothesis in several languages in a more robust manner. To develop this capability, in this study we worked with Kera, and attempted to replicate existing acoustic measurements. This helped to validate and develop the automated workflow which we believe will be useful in analysing other languages.

Table 1: The six vowels of Kera

	- round	- front	- front
	front	- round	round
high	i	ɨ	u
- high	ɛ	a	ɔ

1.3 *Kera vowel harmony* Kera has six oral vowels. The vowels have long and also nasalised counterparts, although the nasalised vowels are fairly rare (Pearce, 2011). The six vowels of Kera are shown in Table 1. The harmony rules taken from Pearce (2012) are as follows (with each pair of parentheses indicating a foot):

[*] We would like to thank Faith Comes By Hearing for providing the original Kera recordings, Danielle Gootjes for producing the gold-standard alignments, Rejoice Lasing for preparing some of the initial data, David Lovell for providing the phonetic feature function in R, and Matthew Kempton for advice with the statistics. This work is partially funded by the SIL International Pike Scholars Program.

Proceedings of *AMP 2019*

1.3.1 *Epenthetic feature-filling* Verb epenthetic vowels copy the root vowel if there is no suffix vowel.

(1) mirk-t-n → (mīr)(kītīn) ‘greets me repeatedly’

But if there is a suffix vowel, that gets copied.

(2) mirk-t-u → (mír)(kútúː) ‘greets him repeatedly’
(3) mirk-t-n-u → (mír)(kút)(núː) ‘greeted him repeatedly’

1.3.2 *Height* If one vowel is high, the other vowels also become high.

(4) mirk-a → (mír)(kɨ́ː) ‘greet her (3sf)’
(5) baːd-u → (bɨ̀ː)(dùː) ‘wash him (3sm)’

1.3.3 *Fronting and rounding* Vowels that are not front or round, but which are high, become a front vowel providing the suffix is a high front vowel, and become a round vowel if the suffix is a high round vowel.

(6) cɨɨri → (cīi)(rīː) ‘your (f) head’
(7) cɨɨru → (cúu)(rúː) ‘his head’
(8) ɨski → (īs)(kīː) ‘hear you (f)’
(9) ɨsku → (ús)(kúː) ‘hear him’
(10) vɨɨgɛ → (vɨ̀ː)(gìː) ‘is emptying’
(11) baadi →(bɨ̀ː)(dìː) ‘wash you’

1.3.4 *Fronting within feet* This harmony targets central vowels and is triggered by any front vowel. Spreading only takes place within the foot, from right-to-left. The iambic foot is shown in parentheses.

(12) bàl-ɛ → (bèlɛ̀ː) ‘love’
(13) bɨ̀ŋ-ɛ → (bìŋìː) ‘open’
(14) ɨ̄s-ɛ → (ísíː) ‘sit down’

1.3.5 *Total* In nouns, all vowels agree in features with the possible exception of the final vowel in final CV syllables.

(15) (dɨ̀bɨ̀ː)(bɨ̀r) ‘lizard’
(16) (káː)(sáw) ‘millet’
(17) (kɛ̄f)(tɛ̄r) ‘book’
(18) (kúpúr)kí ‘male goat’
(19) (gàdàː)mɔ́ ‘horse’
(20) (dògɔ̀y)nà ‘now’

Table 2 summarises the different types of harmony acting over various domains and in various directions. PrWd is the Prosodic Word, which does not include unfooted syllables at the end of nouns. MWd is the morphological word, which includes all roots and affixes.

Table 2: The domains and direction of Kera harmony (Pearce, 2007)

Harmony	Direction	Target	Trigger	Domain
Epenthetic feature-filling	← if possible	epenthetic V	underlying V	PrWd
Height	↔	all V	high V	MWd
Fronting and rounding	←	high central V	high fr/rd (suffix) V	PrWd
Fronting	←	central V	front (suffix) V	Foot
Total	←	all V (not suffix)	last head V (not suffix)	PrWd

2 Automated workflow for corpus phonetics

Our approach to the automated workflow is outlined in Figure 1. The complete workflow is unicode compliant so there is no need to convert the language orthography or IPA characters into a different ASCII machine readable format. Some of the additional resources we have developed as part of this workflow are available online[1].

2.1 *Utterance alignment* Most phone alignment algorithms require short segments of audio, e.g. Montreal Forced Aligner requires speech segments less than 30 seconds in duration (McAuliffe et al., 2017b). Cross-language forced alignment can be used to automatically break up a transcribed recording session into utterances. We used the Aeneas forced aligner (Petterin, 2017). The SIL tool, Scripture App Builder (SIL, 2019) uses Aeneas in a cross-language configuration. The default source language Esperanto works well for mapping across to the alphabet or phoneme inventory of most target languages, including Kera. The fine tune timings HTML output file was converted into a simple tabulated form (Audacity label format) and the names of the speaker for each utterance was manually added (thus creating a text file conforming to the FAVE align format). ELAN (ELAN, 2019) was then used to open this tabulated file, interpreting the speaker column as a tier column. The resulting file of the recording session was then saved as a Praat TextGrid. In this TextGrid each tier corresponds to a speaker and the intervals correspond to utterances. This format is suitable for the next stage.

2.2 *Phone alignment* Phone alignment can be completed by cross-language forced alignment if there is a small amount of data, or more commonly, target-language forced alignment when there is an adequate amount of data to train the model. Montreal forced aligner (MFA) is a state-of-the-art forced aligner (McAuliffe et al., 2017a) which includes triphone modelling and speaker adaptation. MFA requires audio and transcriptions for each utterance, and usually a "pronunciation dictionary" where each line contains the word in the orthography followed by a phone sequence. For languages with fairly transparent orthographies this can be achieved with relatively simple letter-to-phoneme mappings. We adapted an existing script[2] developed by Coto-Solano & Solórzano (2017) which used the above FAVE files to generate the pronunciation dictionary. We also used regular expressions to implement some slightly more complex letter-to-sound rules e.g. for the letter <ə> in Kera. MFA was then trained on this material and produced word-alignments and phone-alignments.

2.3 *Analysis* The EMU Speech Database Management System (EMU-SDMS) (Winkelmann et al., 2017) is used in the analysis stage. From its origins thirty years ago, it has had the strengths of being set in a statistical and visualisation programming environment (Watson, 1989), and allows querying across different levels of the linguistic hierarchy (Harrington et al., 1993).

Since EMU requires separate files for each speaker, we created a simple script that splits up the TextGrids from MFA. Once EMU loads these in alongside the audio files, it constructs a database that can be interrogated using a flexible query language. For example, the query: `phone == ɛ` will return the timing information of all the tokens of [ɛ] which can then be analysed acoustically. Using the ability of EMU to define groups of elements, and making use of a set of binary features defined by Hayes (2009), we have added the capability to easily form groups from these features. For example `phone == non_high_vowel` will return all the vowels tokens in the recording that are not considered to be high. Feature definitions can be redefined for the target language if necessary. EMU also allows both regular expressions and queries combining different levels in the annotation hierarchy. For example `word =~ .* ^ phone == non_high_vowel` returns all the word tokens that contain a non-high vowel. The fact that EMU is used within the R programming language (R Core Team, 2019) means that the functionality of queries can be extended even further. For example some queries for checking vowel harmony are difficult to form [3] (which has been a challenge for other phonetic search tools e.g. Dingemanse (2008)). However, by using R set operations this difficulty can be overcome, for example we use `setdiff(all_words,words_with_non_high_vowel)` to return all the words in the Kera corpus that exhibit vowel height harmony.

[1] Resources such as the feature data and example notebooks can be found at github.com/speechchemistry

[2] `makeArpabet.py`

[3] Raphael Winkelmann (2019) personal communication

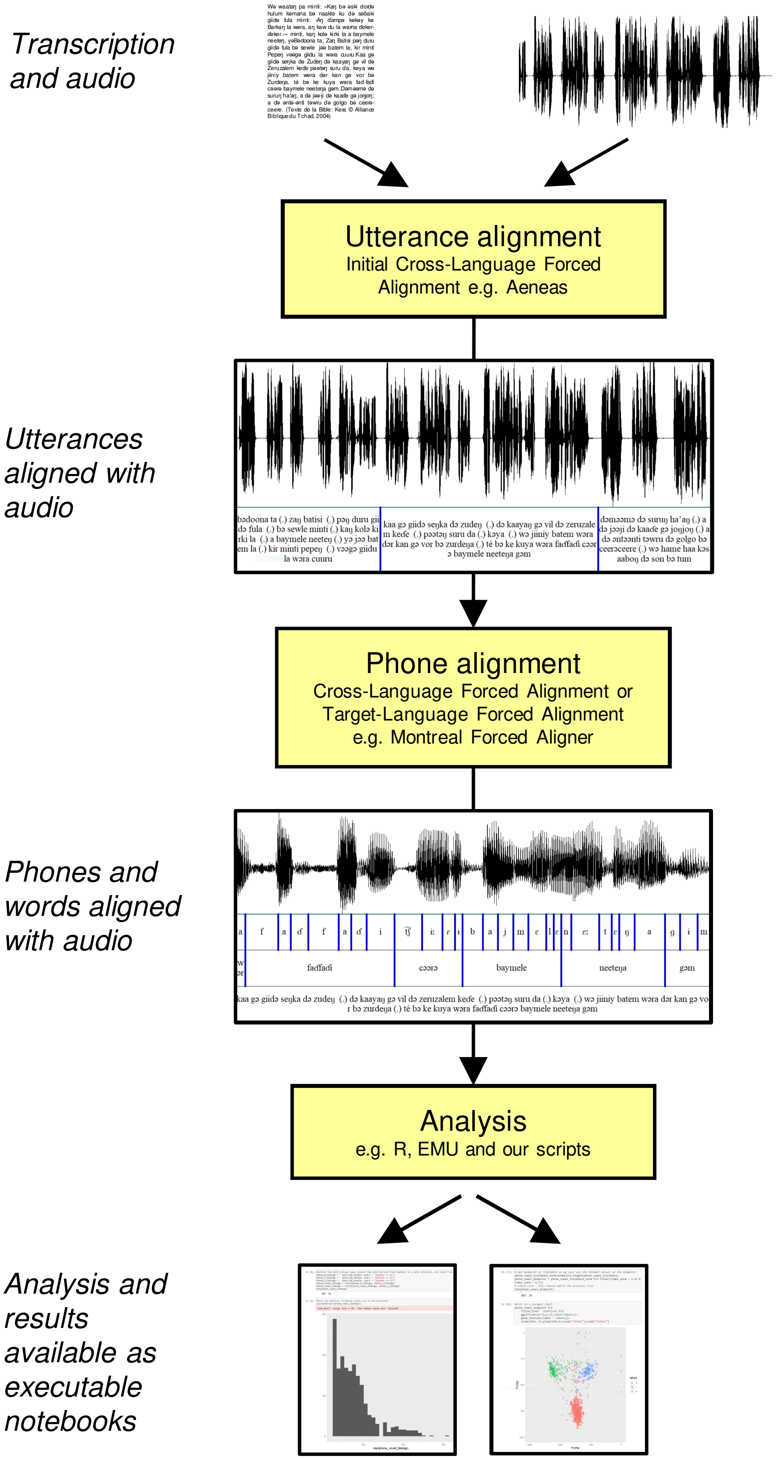
Transcription and audio
Wə waataŋ pa minti: «Kaŋ bə əski dordə hulum kəmana bə naakte ku də səɓaki giidə fula minti: ‹Aŋ ɗampə kəkəy kə Barkaŋ la wəra, aŋ kaw du la warna ɗeker-ɗeker.›» minti, kaŋ kolə kirki la a baymele neeteŋ, yəBədoona ta, Zaŋ Batisi pəŋ duru giidə fula bə sewle jəə batem la, kir minti Pepeŋ vəəgə giidu la wəra cuuru.Kaa gə giidə seŋka də Zudeŋ də kaayaŋ gə vil də Zeruzalem keɗe pəətəŋ suru da, kəya wə jiiniy batem wəra dər kan gə vor bə Zurdeɲa, té bə ke kuya wəra faɗ-faɗi cəərə baymele neeteɲa gəm.Dəməəmə də suruŋ ha'aŋ, a də jəə-ji də kaaɗe gə joŋjoŋ; a də əntə-ənti təwru də golgo bə ceerə-ceere. (Texte de la Bible: Kera © Alliance Biblique du Tchad, 2004)
Utterance alignment
Initial Cross-Language Forced Alignment e.g. Aeneas
Utterances aligned with audio
bədoona ta (.) zaŋ batisi (.) pəŋ duru giidə fula (.) bə sewle minti (.) kaŋ kolə kirki la (.) a baymele neeteŋ (.) yə jəə batem la (.) kir minti pepeŋ (.) vəəgə giidu la wəra cuuru
kaa gə giidə seŋka də zudeŋ (.) də kaayaŋ gə vil də zeruzalem keɗe (.) pəətəŋ suru da (.) kəya (.) wə jiiniy batem wəra dər kan gə vor bə zurdeɲa (.) té bə ke kuya wəra faɗfaɗi cəərə baymele neeteɲa gəm
dəməəmə də suruŋ ha'aŋ (.) a də jəəji də kaaɗe gə joŋjoŋ (.) a də əntəənti təwru də golgo bə ceerəceere (.) wə hame haa kəsaaboŋ də son bə tum
Phone alignment
Cross-Language Forced Alignment or Target-Language Forced Alignment e.g. Montreal Forced Aligner
Phones and words aligned with audio
a f a ɗ f a ɗ i t͡ʃ ɨː r ɨ b a j m ɛ l ɛ n ɛː t ɛ ŋ a g ɨ m
wər faɗfaɗi cəərə baymele neeteɲa gəm
kaa gə giidə seŋka də zudeŋ (.) də kaayaŋ gə vil də zeruzalem keɗe (.) pəətəŋ suru da (.) kəya (.) wə jiiniy batem wəra dər kan gə vor bə zurdeɲa (.) té bə ke kuya wəra faɗfaɗi cəərə baymele neeteɲa gəm
Analysis
e.g. R, EMU and our scripts
Analysis and results available as executable notebooks

Figure 1: Automated workflow

3 Experiments

In an effort to keep our research reproducible, we used a publically available corpus - a recording of the Kera New Testament. Speakers of the Kera language confirm that this contains natural sounding speech produced by fluent readers (Pearce, personal communication).

There is a tension between the technique of corpus phonetics and under-documented languages. The former technique thrives on large datasets but the latter can only offer a limited amount of data. As we wanted to ensure our workflow would be shaped by the challenge of under-documented languages we kept the data down to a realistic size. In our preliminary experiments we used a corpus size of 17 minutes. This was later expanded to 37 minutes. This was made up from the two most common speakers in the recording: Speaker N (23 minutes) and Speaker AF (14 minutes). In the following experiments we test the accuracy of the phone alignments, attempt to replicate some existing acoustic results, and finally use our approach to investigate a Kera vowel harmony question.

3.1 *Accuracy of phone alignment* To evaluate the accuracy of the machine alignments, gold-standard alignments were produced independently rather than using corrected existing machine alignments. The most gold-standard alignments were created from the Speaker N recordings, amounting to just under 200 boundaries. These were used to calculate the results shown in Table 3. The accuracy was similar for the other speaker and clearly increases with corpus size.

Table 3: Accuracy of forced alignment on Kera phone boundaries

Corpus Size	Mean error	Median error
17 minutes	21 ms	15 ms
37 minutes	19 ms	13 ms

The trained model shows surprising accuracy given the size of the corpus (compare with McAuliffe et al. (2017a)). If acoustic measurements are targeted at the midpoint of each phone, they will clearly succeed in falling between the boundaries of long duration phones more often than short duration phones. The shortest duration phones in the gold standard alignments were around 30 ms and this also corresponds to the minimum duration of MFA intervals. The vowel midpoint median error was calculated as 13 ms, so slightly more than half of the measurements targeted at the midpoint should fall within these 30 ms boundaries on average. However this would still have a high proportion of errors. MFA quantizes the durations of each interval in steps of 10 ms. We decided to exclude the 30 ms category from our acoustic measurements and start measuring from 40 ms.

Pearce (2012) divided the phonetic duration into three categories; short, medium and long. The definition of short was any duration less than 50 ms, and the definition of long was any duration greater than 70 ms. We took a similar approach but made a slight change. Given the quantization of the MFA output durations described above and our decision not to use the 30 ms duration, we decided to include the 40 ms and 50 ms durations in the short duration category and the 60 ms and 70 ms durations in the medium category. All other durations were included in the long category.

3.2 *Reduction of non-high vowels* Phonetically shorter vowels in Kera have a different vowel quality to their longer counterparts. This is particularly the case with non-high vowels; as they become shorter the first formant (F1) value significantly decreases. This has lead to phonetically short forms of the vowels [ɛ,a,ɔ] being transcribed as the allophones [e,ə,o] respectively. We used the corpus phonetics workflow to investigate this further. Figure 2 shows an attempt to replicate the results of Figure 1 in Pearce (2011). It shows the relationship between the F1 and the duration of the non-high vowels. The largest change is in the vowel quality of /a/. This large change is one of the reasons that this vowel is written as both <a> and <ə> in the orthography. The individual vowel tokens appear in vertical lines because MFA quantizes the durations of each interval in steps of 10 ms. Estimating the confidence intervals from the curves given in Pearce (2011) suggests that the trends of the three non-high vowels match these previous results.

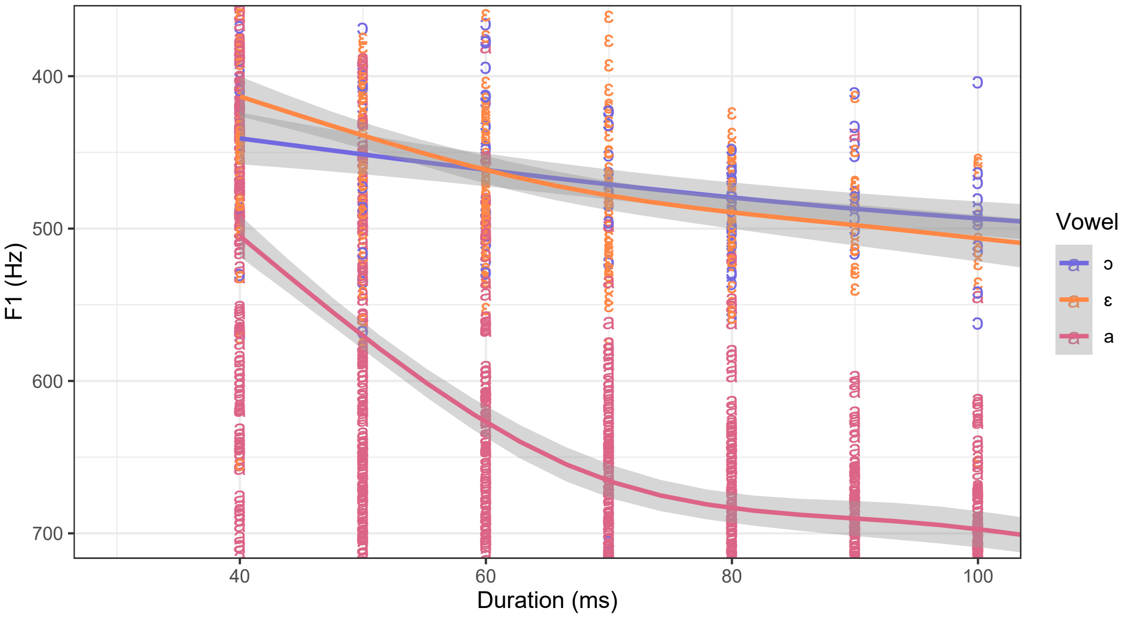

Figure 2: Kera non-high vowels: phonetically shorter vowels have a lower first formant (Speaker AF, grey bands indicate 95% confidence intervals)

3.3 *Vowel height harmony* Pearce (2012) investigates a number of vowel harmonies in Kera. Compelling results are given to demonstrate how reduction is blocked in vowel harmony domains. However in the height harmony domain there is less of an effect:

> "In high vowels, there might be a small effect of height harmony blocking the reduction in F1. The movements in this case would be small, so these results could not provide conclusive evidence either way, but might still add supporting evidence. So the prediction is that the F1 value should stay marginally lower for short high vowels when involved in height harmony. But in general, reduction will occur in F1 whether vowels are in harmony or not." (Pearce, 2012)

Her results confirm that there is little movement: "The /i/ and /u/ vowels have very little F1 reduction in either case". Statistical tests confirmed that there is no proof of an F1 difference between the harmony domain and the non-harmony domain for /i/ and /u/ (Pearce, 2012).

We attempted to use corpus phonetics to study this particular vowel harmony to see if these results are replicated or if we can measure a difference.

As in Pearce (2012), we treat the phonetically long vowels as the canonical vowels. A short version of this vowel may undergo reduction. The amount of reduction can be determined by measuring the distance in F1 between the short and the long version of vowel that are in the same domains. For example, we statistically compare the distribution of short /u/ tokens in the height harmony domain with the long /u/ tokens in the height harmony domain. Short /u/ tokens not within the height harmony domain are compared with long /u/ tokens that are also not within the height harmony domain.

In converting the orthographic form to a phonetic form, we weren't able to precisely model every phonological process. This meant that some vowels that appeared in the phonetic transcript were actually elided in the speech recording. We found that monitoring an acoustic energy measure was a fairly effective way of detecting these elisions. Since MFA was forced to include these missing vowels it usually chose the minimum duration of 30 ms which we automatically excluded anyway.

Loan words were identified and excluded from the acoustic measurements.

The Shapiro-Wilk normality test across the dataset indicated that the majority of samples were not normally distributed in the F1 dimension so we used a non-parametric test (Wilcoxon rank sum test) to test for statistically significant differences.

Results showed that for both speakers there was no statistically significant difference in F1 between short /i/ vowels and long /i/ vowels in the harmony domain. This was also true for /i/ vowels not in the harmony domain. The same set of results were found for the /u/ vowel. We used a threshold of $p < 0.01$ for statistical significance because early experiments indicated that a level of $p < 0.05$ was not robust enough and led to conclusions about the smaller dataset that didn't hold in the larger dataset. Even so, the only difference in our results if we had used $p < 0.05$ was negligible (a small shift in short /u/ tokens that were not in the harmony domain for Speaker AF).

Although we had hoped to detect a small effect, our results end up matching Pearce (2012) where no difference was found between the harmony and the non-harmony condition.

4 Discussion

Our results using corpus phonetics to analyse Kera are in agreement with previous studies on Kera where manual acoustic measurements were used. This gives us more confidence in the automated workflow. For the vowel harmony experiment we were pushing the technology to the limit - investigating short duration phones where accuracy at short duration level is a challenge. Additional data and improvements in forced alignment accuracy may reveal a difference in the future. However it is possible the difference is small or maybe non-existent.

We decided to get an overall picture of what was happening with the vowel reduction. Figure 3 shows the six oral vowels for long and short durations as spoken by speaker AF. The formant plot for speaker N (not shown) is similar. It can be seen that short vowels tend to reduce toward a central point that is quite high in the vowel space. The change in F1 between the long and short high vowels (/i/ and /u/) is particularly small. This makes it difficult to determine the difference between reduction occuring and reduction being blocked.

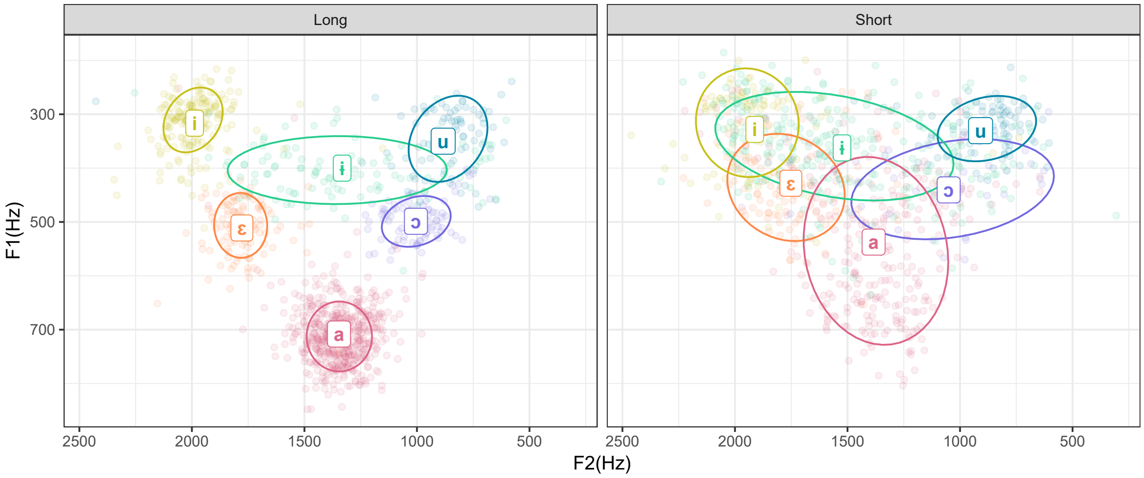

Figure 3: Formant plot of the six Kera oral vowels (the long category includes durations greater than 70 ms, the short category includes durations of 40 ms and 50 ms, vowels shown are from speaker AF, ellipses show one standard deviation of the samples)

Clearly phonetic environment can have an effect on vowel formants. We noticed that in Kera, surrounding nasal sounds as well as proceeding alveolar and velar consonants had a noticeable effect on formant frequencies. Different harmony domains tended to have roughy different consonant environments, so comparing the short vowel to the long vowel within the same domain helped to normalise the effect of these environments. Ideally we would want to explicitly control for these different phonetic environments but that would require substantially more data.

As we have applied this workflow for corpus phonetics we have made a number of additional observations that might be helpful for future analyses and could benefit from being empirically assessed. Firstly, when starting from an orthography we have observed it is important to be aware of the differences between the estimated phonetic (or estimated phonemic) representation and the actual surface level phonetic realisation as expressed in the acoustic signal. A first phase of acoustic analysis can help to reveal any additional transformations and phonological processes that may need to be applied. Secondly, a corpus will have a different balance of data compared to data collected for manual acoustic analysis. Aggregated acoustic measurements can be skewed by high-frequency words. Thirdly, loanwords can lead to acoustic measurements that are significant outliers, so it is important these are labelled and dealt with appropriately.

The forced alignment accuracy results indicate that satisfactory accuracy can still be attained with a small corpus size. It is still an unresolved question when cross-language forced alignment is needed instead of trained alignment. Anecdotally we would suggest that cross-language forced alignment is needed if the dataset is smaller than 10 minutes, but this needs to be investigated more thoroughly.

5 Conclusion

In this study we developed and tested a corpus phonetics automated workflow for the analysis of under-documented languages. An outline of the workflow is shown in Figure 1. Two key components are Montreal Forced Aligner (McAuliffe et al., 2017a) and EMU Speech Database Management System (EMU-SDMS) (Winkelmann et al., 2017). We also added functionality to allow phones to be queried by binary features. Every stage in the process is unicode compliant. We have applied this workflow to Kera, investigating the vowel system. We have replicated previously published results on vowel reduction and vowel harmony that were measured manually as well as gaining new perspectives on the data. Our results are derived from automatic measurements; the only data provided was the recordings, the transcription and any necessary transformation rules to convert the orthography into a phonetic representation.

This study gives us confidence in using this approach to analyse additional languages. We also hope that this workflow or similar approaches to corpus phonetics will be of value to other phonologists.

References

Coto-Solano, Rolando & Sofía Flores Solórzano (2017). Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. *CLEI ELECTRONIC JOURNAL* 20:1, p. 13.

Dingemanse, Mark (2008). Review of Phonology Assistant 3.0.1. *Language Documentation & Conservation* 2:2, 325–331.

ELAN (2019). ELAN v5.7-FX [Computer program]. Nijmegen: Max Planck Institute for Psycholinguistics.

Harrington, Jonathan, Steve Cassidy, Janet Fletcher & Andrew McVeigh (1993). The mu+ system for corpus based speech research. *Computer Speech & Language* 7:4, 305–331.

Hayes, Bruce (2009). *Introductory Phonology*. Wiley-Blackwell.

Liberman, Mark Y. (2019). Corpus Phonetics. *Annual Review of Linguistics* 5, 91–107.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger (2017a). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Interspeech*.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger (2017b). Montreal Forced Aligner v1.0.0 [Computer program].

Pearce, Mary D. (2007). *The Interaction of Tone with Voicing and Foot Structure: Evidence from Kera Phonetics and Phonology*. PhD Thesis.

Pearce, Mary D. (2011). Kera. *Journal of the International Phonetic Association* 41:2, 249–258.

Pearce, Mary D. (2012). Effects of harmony on reduction in Kera. *Linguistic Variation* 12:2, 292–320.

Petterin, Alberto (2017). Aeneas - a Python/C library and a set of tools for forced alignment v1.7.3 [Computer program].

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

SIL (2019). Scripture App Builder v6.0.2 [Computer program]. SIL International.

Watson, Gordon S. (1989). APS: An environment for acoustic phonetic research. *The Journal of the Acoustical Society of America* 85:S1, S56–S56.

Winkelmann, Raphael, Jonathan Harrington & Klaus Jänsch (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45, 392–410.

Yuan, Jiahong, Wei Lai, Chris Cieri & Mark Liberman (2018). Using Forced Alignment for Phonetics Research. *Chinese Language Resources and Processing: Text, Speech and Language Technology. Springer* .