Cross-language forced alignment to assist community-based linguistics for low resource languages

Timothy Kempton
SIL Nigeria, PO Box 953
Jos, Plateau State
Nigeria
tim_kempton@sil.org

Abstract

In community-based linguistics, community members become involved in the analysis of their own language. This insider perspective can radically increase the speed and accuracy of phonological analysis, e.g. providing rapid identification of phonemic contrasts. However, due to the nature of these community-based sessions, much of the phonetic data is left undocumented. Rather than going back to traditional fieldwork, this paper argues that corpus phonetics can be applied to recordings of the community-based analysis sessions. As a first step in this direction, cross-language forced alignment is applied to the type of data generated by a community-based session in the Nikyob language of Nigeria. The alignments are accurate and suggest that corpus phonetics could complement community-based linguistics giving community members a powerful tool to analyse their own language.

1 Background

1.1 Community-based linguistics

Fieldwork is traditionally directed by the linguist. It is the linguist who elicits data from members of a speech community. It is the linguist who phonetically transcribes a wordlist and makes an audio recording. It is the linguist who performs the analysis.

In community-based or participatory-based linguistics, members of the speech community participate in many of these stages (Czaykowska-Higgins, 2009). This includes linguistic analysis, with community members making discoveries and deepening their understanding of the patterns in their own language.

One particular approach to participatory-based phonological analysis is described by Kutsch-Lojenga (1996), Norton (2013), and Stirtz (2015). In this approach, members of the speech community write down words in their language on small cards. A trial orthography is used for the writing since the work is usually part of a language development project to help establish a writing system. The trial orthography may be no more sophisticated than a best-guess spelling using an alphabet of another language. Picking up each card, the language speaker calls out the word aloud and starts to arrange these cards into piles. The choice of pile depends on same/different judgments regarding a specific sound in the word. For example, during a session on the Nikyob[1] language of Nigeria where single syllable nouns were being investigated, the Nikyob speakers placed the words in six different piles representing six different tone patterns. Such piles represent the different contrastive categories of the phonological feature being investigated, e.g. tone might be investigated in one session, and voicing in another session.

The results of participatory-based linguistics are often presented as if they were generated purely from the language speakers' (insider) perspective. This is true most of the time. However, there is an interesting contribution from the (outsider) linguist which can be easily overlooked. Occasionally the linguist who is facilitating a session will hear a consistent difference that the language speaker does not at first notice, sometimes because the distinction is obscured by the trial orthography. For example, during the Nikyob session, speakers were so familiar with writing the five vowels of the

[1]The full name of the language is Nikyob-Nindem (ISO693-3 code kdp) covering two main dialects. The focus of this paper is on the dialect of of Nikyob [nīŋkʲóp̄]. The spelling of Nikyob has varied, both within the community and in the academic literature, due to the fact that the orthography is still developing. The Nikyob speaker recorded for this experiment is from the village of Garas.

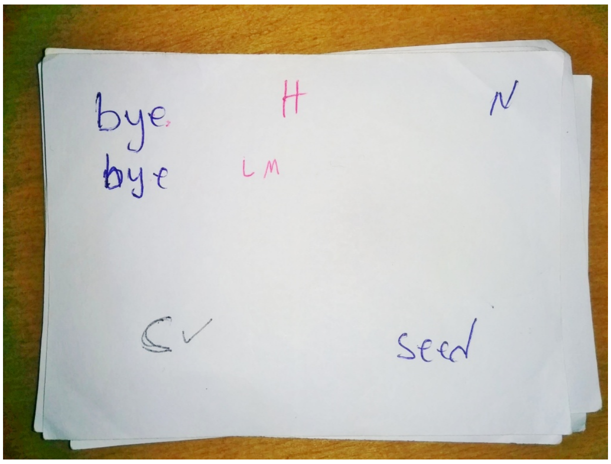

Figure 1: Card for the Nikyob word <bye> “seed”

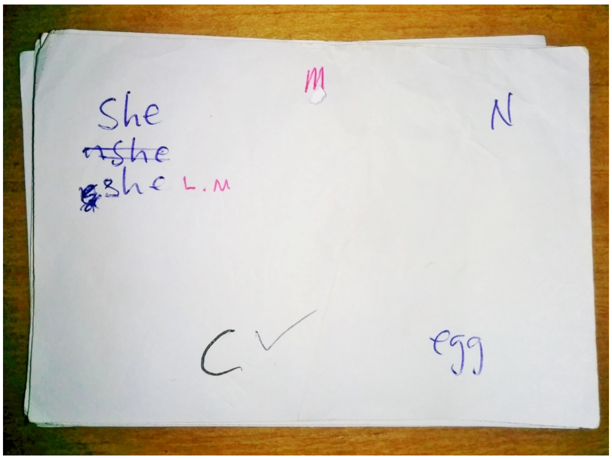

Figure 2: Card for the Nikyob word <she> “egg”

Hausa language /i,e,a,o,u/ they didn’t always notice the extra vowel distinctions in Nikyob, i.e. /o/ versus /ɔ/ and /e/ versus /ɛ/. When the linguist suggested a distinction, the speakers quickly caught on and were soon able to hear their own distinction consistently. The speakers were also quick to recognise which phonological feature was being investigated, e.g. learning to focus on the vowel quality and ignoring the tone.

In these sessions, the primary contribution of the language speakers is their ability to make phonological distinctions, and the primary contribution of the linguist is her broad knowledge of phonetics and phonology. The speakers’ language ability is often unconscious and the collaborative approach raises awareness of that ability. This then gradually accelerates the whole analysis process so that it is quicker than the linguist-only approach. There is also the added advantage that community members have greater motivation to continue in the language development project.

Annotations to the word cards, which are primarily a record of phonemic distinctions, form much of the documentation of these participatory sessions. This is valuable information reached by consensus by a group of speakers. However, the wealth of phonetic data generated in speaking the words is rarely recorded. This lost data limits analysis — not just analysis at the time, but particularly analysis in the future.

Figure 1 shows an example word card that the Nikyob speakers have written. First the singular form is written in the trial orthography <bye>. The “H” indicates the high tone, and “N” indicates a noun. The plural form is then given <bye> and “LM” indicates a low tone rising to mid tone. “C” indicates that the data on the card has been entered on the computer. Finally there is the gloss: “seed”. Note that the phonetic or the phonological representation ([bʲé], /bʲé/ respectively) is not used directly by the Nikyob speakers. Another example word card is shown in Figure 2 for a mid tone word.

1.2 Corpus phonetics

During the development of community-based linguistics in the area of fieldwork, there has been a separate interesting development in the area of phonetics. This is the rise of “corpus phonetics” which involves the “large-scale quantitative analysis of acoustic corpus data” (Yao et al., 2010). In a similar way that corpus linguistics has provided new insights into large collections of texts and transcriptions, corpus phonetics is providing new insights on large sets of acoustic data (Chodroff et al., 2015).

Much of this large-scale analysis is made possible with speech recognition technology and one of the fundamental tools is forced alignment — to automatically align phone transcripts with acoustic data.

2 A first step in combining these two approaches

Combining the participatory-based approach with corpus phonetics should be a fruitful method for analysing and documenting a phonology of the language. For example, corpus phonetics could help describe the phonetic character of the phonemic distinctions suggested from the participatory-based sessions and in turn suggest possible distinctions that may have been missed.

The work described in this paper takes the first step towards combining these two approaches. A fundamental tool of corpus phonetics, forced alignment, is evaluated to see if it can be successfully applied to the type of data generated by the participatory approach.

One characteristic of the data is that it is not adequate to train a forced alignment system. This is because the language has few resources, i.e. no labelled data or a pronunciation dictionary. However, it is still possible to use a forced alignment system trained on a different language. Having its roots in cross-language speech recognition (Schultz and Waibel, 1998), this is called cross-language forced alignment (Kempton et al., 2011), or untrained alignment (DiCanio et al., 2013).

Other characteristics of the data generated by the participatory approach include the lower quality of recording with background noise present and the transcription in a trial orthography.

3 Experimental set-up

An initial pilot corpus was elicited to simulate the data from a participatory-based session, a Swadesh 100 list in the Nikyob language. Each item elicited included an isolated word and the word in a frame sentence. The recording was made in the same room that would be used in a participatory-based session which was a slightly reverberant environment and no special effort was made to mask background noise.

Transcriptions in a trial orthography were taken primarily from a participatory-based tone workshop held in 2015. The trial orthography at the time was adapted from previous work by Kadima (1989), and corresponded with a tentative phoneme inventory derived from Blench (2005).

The cross-language forced alignment system uses a phone recogniser with a 21.5% phone error rate on the TIMIT corpus, so it is still fairly close to state-of-the-art (Schwarz, 2009, p46; Lopes and Perdigao, 2011). The artificial neural network uses a 310 ms window so it is implicitly context dependent (Schwarz, 2009, p39). The neural network produces phone posterior probabilities which are fed into a Viterbi decoder. This means that the system can easily be configured for forced alignment.

Phone set	BFEPP
Czech	0.27
Hungarian	0.49
Russian	0.62

Table 1: Expressing the Nikyob phoneme inventory: phonetic distance

Metric	Value
20 ms error	34%
Mean error	25 ms
Median error	15 ms

Table 2: Cross-language forced alignment on Nikyob Swadesh 100 list

Freely available phone recognisers trained on Czech, Hungarian and Russian were used (Schwarz et al., 2009). A phonetic distance measure, binary feature edits per phone (BFEPP) (Kempton, 2012), was used to predict which phone recogniser would be most suitable for the Nikyob language, and the same phonetic distance measure was used to automatically map the letter labels (reflecting the tentative phoneme inventory) from the Nikyob language to the phone recogniser. For example, the Nikyob <sh> letter represents the Nikyob /ʃ/ phoneme which can be automatically mapped to the Czech /ʃ/ phone recogniser. The Nikyob <w> letter represents the Nikyob /w/ phoneme. However, there is no Czech /w/ phone recogniser so the letter is automatically mapped to the closest recogniser which is the Czech /u/ phone recogniser.

The accuracy of the alignment was evaluated by comparing the boundary timings of the forced aligned labels with gold standard alignments. Gold standard alignments were created by a phonetician for the first 50 words of the Swadesh 100 list along with their frame sentences producing approximately 750 gold standard boundary alignments. The evaluation measure used in forced alignment is the proportion of alignments outside a particular threshold: 20 ms is a common choice. Some recent studies have used mean and median of the absolute timing error instead. In this paper all three evaluation measures are reported.

4 Results

Table 1 shows how close the phone sets of the different phone recognisers were able to express

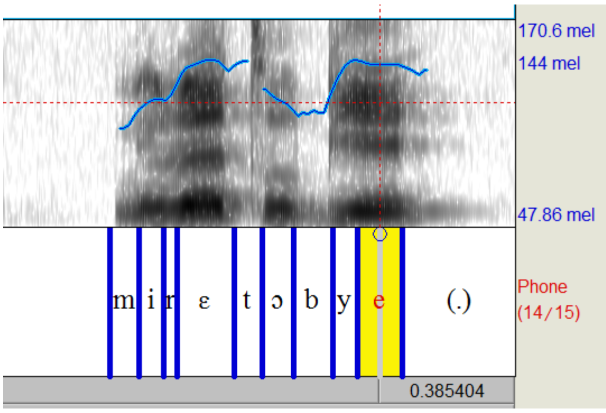

Figure 3: Forced alignment of high tone word <bye> "seed" with its frame sentence displayed in Praat

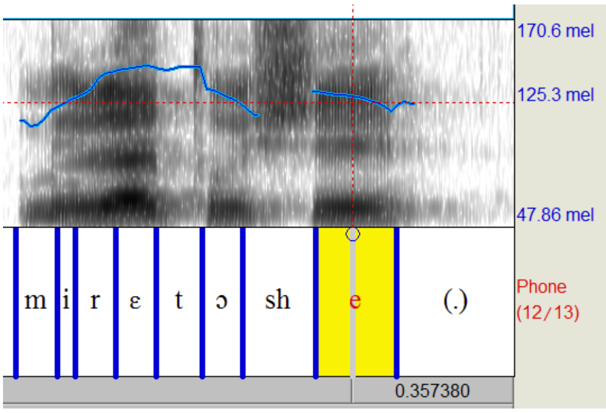

Figure 4: Forced alignment of mid tone word <she> "egg" with its frame sentence displayed in Praat

the Nikyob phoneme inventory. The phone set of the Czech recogniser was closest to the Nikyob phone inventory. So this recogniser was used in the forced alignment of Nikyob.

Results for the first 50 words in the Swadesh 100 list are shown in Table 2. These are the primary results of this paper.

Figure 3 shows an example forced alignment displayed in Praat (Boersma and Weenink, 2014). At the top there is a spectrogram with a pitch tracker and at the bottom there is the alignment of letter labels. Only the second half of the recording is shown where the word is included in the frame sentence: <mi rɛ tɔ bye>, /mī ɾɛ́ tɔ̄ bʲé/, "I say seed". Another forced alignment example is shown in Figure 4.

5 Discussion

The results in Table 2 are encouraging when compared to previous studies on cross-language forced alignment. Previous 20 ms threshold results include a 39% error on isolated words (DiCanio et al., 2013), a 36% error on simple sentences (Kempton et al., 2011), and a 51% error on conversational speech (Kurtic et al., 2012). In a slightly different evaluation of word alignments within long utterances (Strunk et al., 2014), the error averaged across eight corpora revealed a mean error of 187 ms and a median error of 65 ms. There was also a measure of how much disagreement there was between human transcribers. The mean transcriber disagreement was 86 ms and the median transcriber disagreement was 34 ms.

These earlier studies have put forth the argument that such alignments are accurate enough to be usable, either as they are or with a small number of boundaries corrected. In the participatory-based linguistics scenario, there are many repetitions of words recorded and the subsequent aggregation of acoustic measurements would suggest that manual correction to the boundaries would be unnecessary.

The particular alignments reported in this paper are being used to assist with a tone analysis of the Nikyob language. For example, it is a straightforward mechanical process to extract pitch contours from the alignment shown in Figure 3 revealing that the high tone word <bye> "seed" has a pitch contour about 20 mels higher than the known mid tone in the frame sentence. Figure 4 shows that the mid tone word <she> "egg" has a pitch contour much closer to the known mid tone with a difference of about 1 mel. Forced alignment allows many such measurements to be taken. Figure 1 shows the word card for <bye> "seed" is actually part of a pile of word cards that have been judged by Nikyob speakers as high tone words. In the same way, Figure 2 shows a pile of mid tone words. Aggregated acoustic measurements can indicate the extent of phonetic differences within these piles and between these piles, i.e. the phonetic character of these phonemic distinctions can be documented.

Inspecting all the 50 forced aligned utterances indicates that about 8% of the utterances contain alignment errors that would produce erroneous pitch contour measurements. It seems unlikely that this would cause problems in the analysis but further investigation would be needed to confirm this.

6 Conclusion and future work

The results of this paper indicate that cross-language forced alignment can be applied to the data produced in a participatory-based session. With this promising first step, the prospect of combining participatory-based linguistics and corpus phonetics looks viable.

One could imagine a future scenario where the piles of paper cards are simulated on a touchscreen tablet, and as participants select words and speak them, the computer associates a set of recordings with each transcribed word. Phonemic distinctions could be easily tracked along with acoustic data. This would give speech communities a powerful tool to help them discover the phonology of their language.

Acknowledgments

I am grateful to Dushe Haruna who was recorded speaking the Nikyob words and Laura Critoph who produced the gold standard alignments. I appreciate the feedback from Gary Simons, Linda Simons and Matthew Harley on an earlier draft of this paper. This work is partially funded by the SIL International Pike Scholars Program.

References

Roger Blench. 2005. The Ninkyop language of central Nigeria and its affinities (Draft).

Paul Boersma and David Weenink. 2014. Praat: doing phonetics by computer. *Version 5.3.77* [Software].

Eleanor Chodroff, John Godfrey, Sanjeev Khudanpur, and Colin Wilson. 2015. Structured variability in acoustic realization: A corpus study of voice onset time in American English stops. *The Scottish Consortium for ICPhS* .

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation & Conservation* 3(1).

Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García. 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.

Hauwa Kadima. 1989. iByan Rwɛ wa Ninkyob 1 (a first alphabet of the Ninkyob language). Kadima.

Timothy Kempton. 2012. *Machine-assisted phonemic analysis*. Ph.D. thesis, University of Sheffield.

Timothy Kempton, Roger K. Moore, and Thomas Hain. 2011. Cross-language phone recognition when the target language phoneme inventory is not known. *Proc. Interspeech, Florence, Italy* .

Emina Kurtic, Bill Wells, Guy J. Brown, Timothy Kempton, and Ahmet Aker. 2012. A Corpus of Spontaneous Multi-party Conversation in Bosnian Serbo-Croatian and British English. *International Conference on Language Resources and Evaluation, Istanbul, Turkey* .

Constance Kutsch-Lojenga. 1996. Participatory research in linguistics. *Notes on Linguistics* (73):13–27.

Carla Lopes and Fernando Perdigao. 2011. Phone recognition on the TIMIT database. *Speech Technologies/Book* 1:285–302.

Russell Norton. 2013. The Acheron vowel system: A participatory approach. *Nuba Mountain Language Studies. Cologne: Rüdiger Köppe* pages 195–217.

Tanja Schultz and Alexander Waibel. 1998. Multilingual and Crosslingual Speech Recognition. *Proc. DARPA Workshop on Broadcast News Transcription and Understanding* pages 259–262.

Petr Schwarz. 2009. *Phoneme recognition based on long temporal context*. Ph.D. thesis, Brno University of Technology.

Petr Schwarz, Pavel Matějka, Lukáš Burget, and Ondřej Glembek. 2009. Phoneme recognition based on long temporal context. *phnrec v2.21* [Software].

Timothy M Stirtz. 2015. Rapid grammar collection as an approach to language development. *SIL Electronic Working Papers* (2015-004).

Jan Strunk, Florian Schiel, Frank Seifart, et al. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS .

Yao Yao, Sam Tilsen, Ronald L Sprouse, and Keith Johnson. 2010. Automated measurement of vowel formants in the Buckeye corpus. *UC Berkeley Phonology Lab Annual Reports* .