Text-Speech Alignment: A Robin Hood Approach for Endangered Languages

Claire Bowern
Yale University, claire.bowern@yale.edu

Rikker Dockum
Yale University, rikker.dockum@yale.edu

Sarah Babinski
Yale University, sarah.babinski@yale.edu

Hunter Craft
Yale University, hunter.craft@yale.edu

Anelisa Fergus
Yale University, anelisa.fergus@yale.edu

See next page for additional authors

Presenter/Creator Information
Claire Bowern, Rikker Dockum, Sarah Babinski, Hunter Craft, Anelisa Fergus, and Dolly Goldenberg

* = joint first author

CHIRILA
Yale Linguistics

Automated Speech–Text Alignment: A Robin Hood Approach for Endangered Languages

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

Introduction

- Linguists need very fine-grained sound file labeling to study speech.
- Problem: manual speech–text alignment is extremely time-intensive, a severe bottleneck for phonetic research on the under-resourced languages of Australia.
- **Forced alignment (FA)** automatically aligns audio recordings of spoken language with transcripts at the segment level and saves manual alignment time.
- FA models must be trained on existing segmented language data, and the amount required for training is often greater than the amount of data available for many under-documented languages (also the case for much of Australia).
- This project tests the performance of a 'Robin Hood' approach to to FA by using English-trained models to align low-resource (particularly Australian) languages.

Prior work

- Models trained on English are widespread (Evanini et al., 2009; Gorman et al., 2011; Reddy and Stanford, 2015), but work on under-resourced languages is less common.
- **DiCanio et al. (2013)** tested the performance of two English-trained FA models (P2FA and hmalign) on Yoloxóchitl Mixtec data.
- Compared to totally manual labeling, use of FA greatly reduced the time required for processing, but a substantial amount of manual correction was required.
- **Johnson et al. (2018)** tested the accuracy of the Prosodylab aligner (Gorman et al., 2011) on uncleaned Tongan field recordings against both FA-aligned cleaned audio and manual corrections.
- They found that as long as the model is trained on cleaned audio, it is just as good at aligning uncleaned as cleaned audio, and differences between FA and manual alignment were not significantly different from the differences between two manual alignments done by different humans.

Methods

Data

- 45 minute corpus of spontaneous running Yidiny (Pama-Nyungan) speech, recorded & transcribed in the 1970s, from Dick Moses & Tilly Fuller.

Models

- Three different FA models were tested:
  - DARLA (Reddy and Stanford, 2015): Trained on an English model
  - P2FA (Evanini et al., 2009): Trained on English and can only use ARPAbet characters
  - KALDI (Povey et al., 2011) Trained on Yidiny
- P2FA outputs were manually corrected and used as a gold standard against which to compare automatic results.

Statistical methods

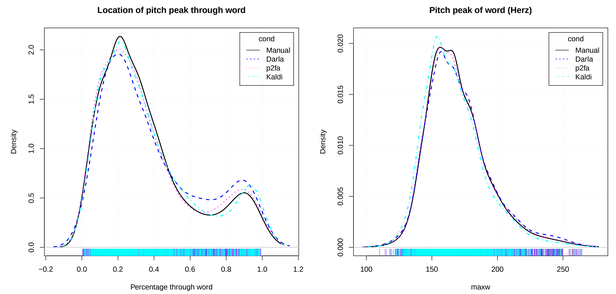
- F0 peak and F0 alignment were extracted from the TextGrids for comparison.
- Models were compared using linear mixed effects models:
  - Fixed effects: speaker gender & alignment model
  - Random effect of word
- Difference in fixed effects means were calculated with the `Lsmeans` test in the R package `lmertest`.
- Vowel spaces were extracted and plotted using the `vowels` package.

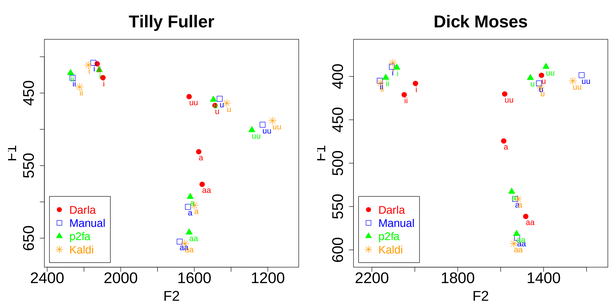
Results: Prosody

We compare F0 peak location (% through the word) and F0 measurement (in Hz).

- P2FA and KALDI are statistically indistinguishable from manual data.
- The P2FA-Manual factor comparison had no significant contribution to the model; a smaller standard error, and higher correlation coefficient with the manual coding.
- The DARLA condition was significantly different on all counts.
- Average peak locations within 0.6%–1% and 1–2 Hz.
- This implies that for word-level prosodic studies, automatic alignment is acceptable.

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	3.587e-01	9.414e-03	3.671e+03	38.101	<2e-16	***
condDarla	2.627e-02	1.327e-02	4.074e+03	1.980	0.0477	*
condp2fa	-1.982e-03	9.614e-03	7.531e+03	-0.206	0.8367	
condKaldi	8.113e-04	1.263e-02	3.713e+03	0.064	0.9488	
speakerTF	-2.922e-02	1.160e-02	9.290e+03	-2.519	0.0118	*



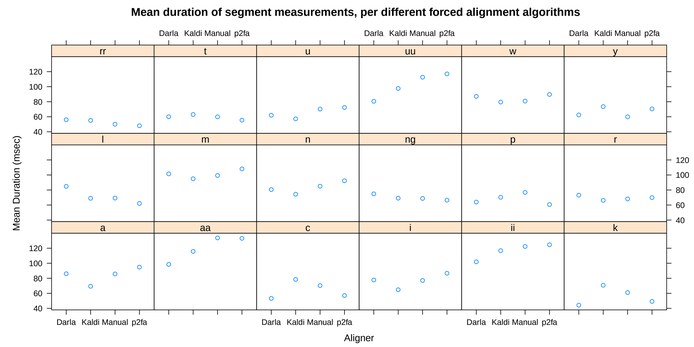


Results: Vowel Space

- Darla performed poorly on this data. Other means are within 6Hz for F1 and 20Hz for F2 (except for uu).
- Vowel measurements were close for both speakers, except uu and ii; in both speakers, P2FA recorded uu as further fronted on average.
- The least accurately captured vowels are also those with the fewest tokens.
- Uncorrected automatic alignment may not be sufficiently accurate to capture accurate renderings of vowel spaces.

Results: Segment Duration

- p2fa and Kaldi algorithms produce results that differ from the gold standard data by an average 7 and 8 ms (respectively); Darla results differ by 10 ms.
- However individual phoneme results vary substantially.
- All algorithms differed significantly from the manual data.
- Sounds least accurately captured are found in both languages, i.e. those with unclear transitions between consonants and vowels.



Impacts

- Australian languages lack research on prosodic questions, an important subject for documenting these endangered languages.
- Archival collections, such as those available for Australian and other endangered language groups, are rarely aligned.
- Manual alignment is prohibitively time consuming. Using unsupervised methods can overcome this bottleneck.

Conclusions

- P2FA and KALDI did not perform significantly differently from 'gold standard' manually aligned data in detecting basic prosodic features.
- Performance on vowels requires more caution, depending on the task and the amount of data available.
- Our results demonstrate the promise in forced alignment for preliminary study of prosody in underdocumented languages.
- Caveat: Results may be language specific, meaning the performance of these methods on other languages is unclear.

Acknowledgements

This research was funded by NSF grant BCS-9423711. Thanks also to participants in the Histling@Yale Lab.