

Evaluating the efficacy of Prosody-lab Aligner for a study of vowel variation in Cantonese

Andrew Peters & Holman Tse

York University & University of Pittsburgh

In this talk, we discuss the effectiveness of using Prosody-lab Aligner (Gorman, Howell, & Wagner, 2011) as a tool for the study of vowel variation and change in Cantonese. Automated (Forced-)aligner programs have recently been introduced as a computational tool for facilitating the process of creating time-aligned transcripts of speech data. The most widely used program, FAVE (Rosenfelder, Fruehwald, Evanini, & Yuan, 2011), however, is designed to work only on English. We therefore use Prosody-lab Aligner (Gorman et al., 2011) as an alternative because of its ability to train models for alignment of any language.

Speech samples used in this project come from sociolinguistic interviews that were collected as part of the Heritage Language Variation and Change in Toronto (HLVC) project (Nagy, 2011). We investigate two questions for evaluating the efficacy of this methodology for use in a larger project on intergenerational change in Heritage Cantonese vowels: 1) Is Prosody-lab aligner effective at producing sufficiently accurate transcript alignment to permit automated measurement of vowel data? 2) What sort of data used to train models for Prosodylab-aligner is most effective at producing results that require minimal manual adjustments?

We address these questions by running Prosodylab on 10 speakers, including four GEN 1 speakers (born and raised in Hong Kong), and six GEN 2 speakers (raised in Toronto). For each speaker, 50% of their transcript data was set aside for model training, and on each speaker the aligner was run using 3 different models: once with data from that speaker alone in the model training, once with data from all speakers in the respective generation used in model training, and a final time with data from all speakers used in model training.

The three types of model training were compared for their efficacy quantitatively by measuring the differential between the automatically-generated boundaries of 468 monophthong vowel tokens, and “gold-standard” manually-aligned vowel boundaries for the same vowels. On this data, the root-meansquare-deviation was calculated for the time-aligned results of each model type (Chen, Liu, Harper, Maia, & McRoy, 2004). The percentage of occurrences in which the center of the automatically-aligned vowel segment lay within the manually-aligned target vowel area was also calculated for each instance.

Our results show that models trained on the individual speakers alone produced the least-deviant data from the ideal manually-aligned vowel targets, and the model trained on data from all speakers produced the most deviant results. However, as requirements on a minimum amount of data to be made available for model training would necessitate up to 50% loss in analyzable vowel tokens if taken from one speaker’s interview alone, an individual-based training model is rejected as impractical. A model trained on data from the respective generational cohort is accepted as the best compromise that produces results requiring the least manual adjustment post-alignment, without sacrificing large amounts of data to model training.

References:

- Chen, L., Liu, Y., Harper, M. P., Maia, E., & McRoy, S. (2004). Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus. In LREC. Retrieved from <https://www-new.comp.nus.edu.sg/~rpnlp/ir/proceedings/lrec-2004/pdf/307.pdf>
- Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3), 192–193.
- Nagy, N. (2011). A Multilingual Corpus to Explore Variation in Language Contact Situations. *Rassegna Italiana Di Linguistica Applicata*, 43(1/2), 65–84.
- Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. Retrieved from <http://fave.ling.upenn.edu>