MEASURING AND MODELLING THE DURATION OF INTERVOCALIC ALVEOLAR TAPS IN PENINSULAR SPANISH

Scott James Perry[1], Matthew C. Kelley[2], Benjamin V. Tucker[3,1]

[1]University of Alberta, [2]University of Washington, [3]Northern Arizona University
sperry1@ualberta.ca, mattck@uw.edu, benjamin.tucker@nau.edu

ABSTRACT

Factors predicting acoustic variation in the production of Spanish taps have yet to be investigated outside of their relationship to the tap-trill contrast. The present study models the duration of alveolar taps with occlusions visible on a spectrogram in spontaneous Spanish and compares durations measured by automated methods to hand-placed boundaries. We model tap duration from the Nijmegen Corpus of Casual Spanish [1] with a combination of lexical, phonetic, and phonological predictors. Results indicate a high degree of uncertainty regarding the relationship between most of our predictors and tap duration. However, we are confident that faster speech rates are associated with decreased duration. Our automated measurements show deviations from hand-measured duration, indicating a need to evaluate the performance of the automated methods in future research.

Keywords: Spanish, phonetic variation, acoustic measurement, speech production

1. INTRODUCTION

Studies analyzing the duration of Spanish alveolar taps (hereafter taps) have generally focused on comparing productions across speaker groups [2], investigating acoustic correlates of the tap-trill contrast [3, 4] or both [5]. The present study's primary goal is to augment our understanding of variation in tap production. We model tap duration with phonetic, phonological, lexical, and predictability-related factors. Our secondary goal is to compare tap durations based on experimenter markup to three automated methods that measure duration with minimal researcher markup and evaluate the methods' impact on model estimates.

In research on other languages, various predictors have been associated with changes in duration at the segmental, syllabic, and word levels. Increased frequency and predictability have been associated with decreased duration [6, 7]. Duration differences by phonetic factors such as phonetic environment and speech rate have been attested [8, 9]. Lexical stress and pitch accents have also been associated with changes in stop closure duration [10].

When measuring the duration of segments without an apparent onset or offset, which includes many Spanish taps [3, 4], it is desirable to have a measurement method that applies to most realizations. One alternative to human markup includes using boundaries placed by an acoustic model through forced alignment, although this method has some known issues [11]. Another option is to use intensity to measure duration, which can be done in various ways [12, 8, 13]. Before applying these methods to Spanish taps, we believe it is important to compare them to hand measurement for tokens with visual cues to onset and offset, where experimenter markup can be consistent.

2. METHOD

The data and the script documenting the analysis are available through the University of Alberta Education and Research Archive here: https://doi.org/10.7939/r3-5k3f-4t63

2.1. Data coding and measurement

We analyzed a random 10% sample of intervocalic taps from the Nijmegen Corpus of Casual Spanish [1], containing 20 conversations between groups of three university students from Madrid, Spain. Of the 2,606 hand-coded taps, 1,312 had occlusions visible in a spectrogram ('True' or 'Approximant' taps following previous studies [3, 4, 2, 5, 14]).

Each tap's duration was measured in four ways using a script in Praat (v 6.1.47; [15]). The first method was the manual placement of boundaries, carried out by the first author. For stop-like taps, the onset was placed at the beginning of the stop closure, and the offset was placed at the onset of periodic energy after the burst release. For approximant taps, onset and offset were placed where the spectrogram abruptly changed intensity. The second method used force-aligned boundaries from the Montreal Forced

Aligner [16] trained on the corpus under analysis. The third method took the tap onset and offset as the midway points between the maximum intensities of the surrounding vowels and the minimum intensity during the tap [8]. The final method placed the onset and offset of the tap at the largest absolute values of the spline-smoothed intensity slope in and out of the tap, using methods from [12, 13].

When using automated measurements, outliers that are likely measurement errors are removed based on domain knowledge. To fairly compare methods, we removed observations that we judged to have impossible values in any of the measurement methods. Spanish taps have average durations below 50 ms [3, 4], and virtually all taps reported in [5] were under 150 ms. Therefore, we removed nine observations with values of 200 ms or more and seven with negative values. This left a total of 1,296 tokens for statistical analysis.

2.2. Statistical analysis

To analyze tap duration, we fit four hierarchical Bayesian models (one for each measurement method) with lognormal likelihoods using `brms` (v 2.18.0; [17]) in `R` (v 4.1.1; [18]). Priors were weakly-informative in the context of Spanish taps, and their assumptions were assessed through prior predictive simulation. Following best practices [19], we checked our priors' influence on the posterior by using wider priors, which resulted in identical posteriors upon visual inspection.

The population-level predictors for our models appear below. Group-level effects were varying intercepts by speaker and correlated varying slopes for all population-level effects by speaker. Unigram and bigram frequencies from the corpus under analysis were added to counts from the Spanish OpenSubtitles corpus [20], and conditional probabilities derived therefrom. We extracted speech rate (syllables/second), surrounding vowels, and lexical stress from force-aligned text grids. Word length and content/function status were derived from the corpus data file.

The following predictors were used in the models:

- **Unigram freq.** Log unigram frequency for word containing tap
- **Bigram freq. prev.** Log bigram frequency for word and previous word
- **Bigram freq. fol.** Log bigram frequency for word and following word
- **Cond. prob. prev.** Conditional probability of the word based on previous word
- **Cond. prob. fol.** Conditional probability of the word based on following word
- **Number of syllables** Word length in number of syllables (log-transformed)
- **Function word** Sum-coded, difference between function (0.5) vs. content words (-0.5)
- **Local speech rate** Log average speech rate of the utterance containing tap (syllables/s)
- **Prev. vowel** Preceding vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Fol. vowel** Following vowel, treatment coded (/i,e,a,o,u/) with /e/ as reference level
- **Prev. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) previous vowel
- **Fol. stress** Sum-coded, unstressed (-0.5) vs stressed (0.5) following vowel

To measure the similarity between posterior distributions from the manual model to those from the automated methods, we calculated overlap in the population-level posteriors using the `overlap()` function from package `bayestestR` (v 0.11.0; [21]). We entered the overlap values as the dependent variable in a hierarchical Beta regression predicting posterior overlap with the manual model by method. We included varying intercepts for predictors with varying slopes by method.

3. RESULTS

For complete model summaries, the reader is directed to the materials hosted in the repository, which include the saved models for convenience.

3.1. Descriptive comparison of methods

Figure 1 contains visualizations that compare manual duration to the three automated methods. In the top row (A, B, & C) are scatter plots between manual measurements on the y-axis and each automated method on the x-axis. Marginal density distributions are placed along the edge of the plots. All automated methods were weakly correlated with manual measurements. In Figure 1 D, we plot the density distributions of the difference between the manual measurement and the automated methods, subtracting the automated duration from the manual measurement for each token. The red dotted line at zero is where the methods had the same value.

3.2. Factors affecting duration

For the model of manual durations, we report in Table 1 the percentages of each posterior that fell below, within, and above a Region of Practical Equivalence (ROPE). A ROPE establishes a minimum effect size the researchers consider practically different from zero. We chose a ROPE

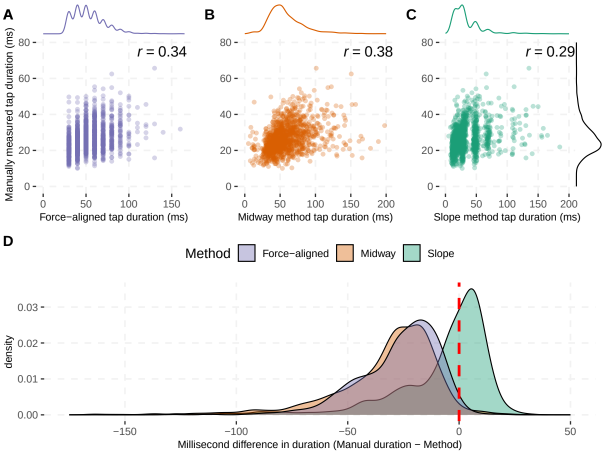

Figure 1: Scatter plots for comparisons between the manually measured durations and the durations taken from forced alignment (A), the midway method (B), and the slope method (C). The density distribution of manual measurements are along the right-hand y-axis of C. Pearson's correlation coefficients are printed on the scatterplot for each comparison. D displays density distributions of by-token differences.

of -0.05 to 0.05 for log-scale duration, which for our model states that if the total effect of a predictor is less than ≈ 1.2ms, then we consider the effect to be negligible. This approach allows us to consider the evidence from our model in terms of the probability of both the existence and direction of an effect. An effect below the ROPE is associated with shorter taps, and an effect above the ROPE signals an association with longer taps. For example, our models suggest we cannot be sure if function words contain shorter taps than content words, as roughly 68% of the posterior is within the ROPE, and 32% is below it. We interpret this as a 2/3 chance that there is no difference in tap duration for function words and a 1/3 chance that function words contain shorter taps. The posterior distribution did not extend above the ROPE, meaning we are confident function words don't contain longer taps.

3.3. Modelling differences & posterior overlap

In Figure 2, we plot the population-level posteriors from the models fit to durations from the four methods. The mean of the posterior is plotted by shape, and the two-stage lines visualize the most probable 80% and 95% of each posterior. The posteriors for the same predictor from the four models show variable levels of overlap. For some predictors (e.g., previous bigram frequency), all methods have similar posterior distributions. For others, there are larger differences between the methods, sometimes in a way that could change model interpretation as compared to hand-measured duration or compared to other automated methods.

Table 1: Percentages of posteriors from the model of manually-measured duration that fall below, within, and above the established Region of Practical Equivalence. The ROPE range of (-0.05, 0.05) was divided by the range of continuous predictors to evaluate the total effect size.

Predictor	% Below/Within/Above ROPE
Unigram freq.	2.8 / 26.1 / 71.0
Bigram freq. prev.	56.6 / 42.5 / 0.9
Bigram freq. fol.	17.3 / 76.2 / 6.5
Cond. prob. prev.	70.1 / 7.8 / 22.2
Cond. prob. fol.	91.8 / 7.2 / 1.0
Number of syllables	56.3 / 39.1 / 4.6
Function word	32.4 / 67.6 / 0.0
Local speech rate	98.9 / 1.1 / 0.0
Prev. vowel /i/	0.0 / 10.0 / 90.0
Prev. vowel /a/	4.7 / 95.3 / 0.0
Prev. vowel /o/	23.8 / 76.1 / 0.1
Prev. vowel /u/	22.4 / 73.3 / 4.3
Fol. vowel /i/	0.0 / 0.0 / 100
Fol. vowel /a/	0.7 / 93.0 / 6.3
Fol. vowel /o/	0.0 / 40.2 / 59.7
Fol. vowel /u/	83.1 / 15.7 / 1.3
Prev. stress	14.4 / 83.9 / 1.7
Fol. stress	5.3 / 93.4 / 1.3

The posteriors from the Beta regression estimating overlap as well as group-level standard deviations among the predictors are plotted by method in Figure 3. We cannot be confident that any automated method overlaps more or less with our manual model, although the Slope method had the highest estimated overlap (Figure 3 A). We are confident the Slope method showed less variability by predictor than the other methods (Figure 3 B).

4. DISCUSSION & CONCLUSIONS

The present study measured tap durations in conversational Spanish and modelled this duration with several predictors. As many taps lack visible occlusions, we wanted to evaluate alternative methods of measuring duration that could be applied to more variable realizations. In predicting manually-measured taps, most predictors are highly uncertain regarding the presence or direction of an effect, although we can rule certain patterns out. Measurements from all automated methods correlated similarly with manual measurements, and the Slope method had the lowest absolute error. The force-aligned durations were multiples of 10ms with a floor at 30ms. The midway

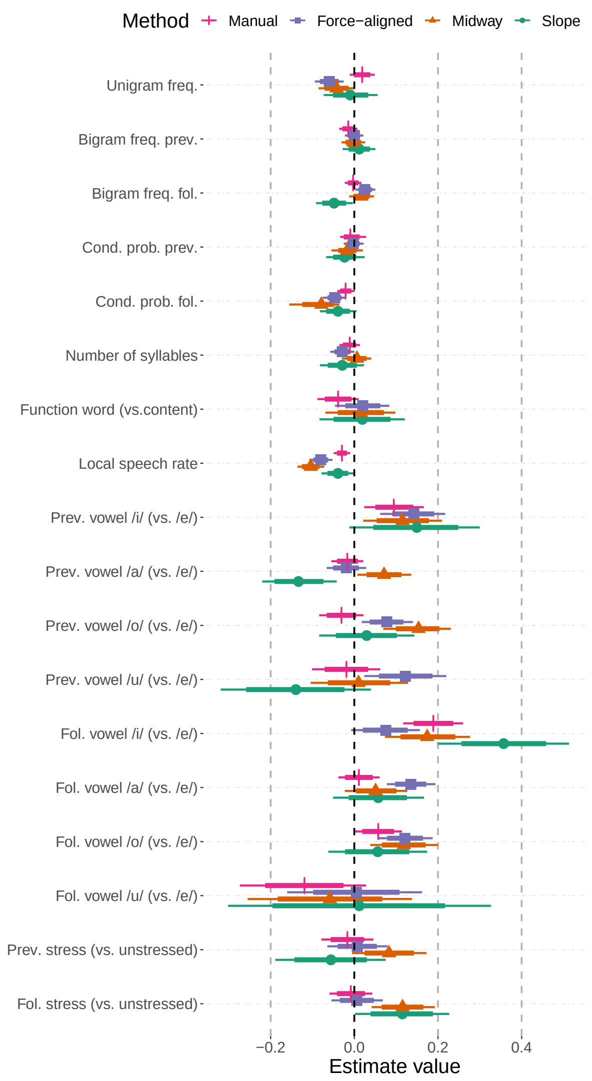

Figure 2: Posteriors for population-level effects models of tap duration calculated by four methods: manually-placed boundaries, force-aligned boundaries, using the midway points of the intensity curve, and using the intensity slope. The thick line corresponds to 80% of the posterior, and the extended, thin line corresponds to 95%.

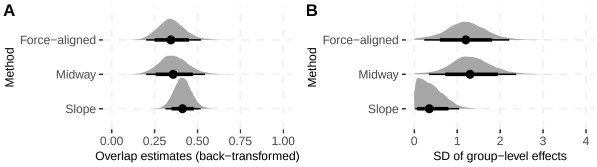

Figure 3: Posterior distributions of estimated overlap with model estimates from manual durations. The thick lines correspond to 80% of the posterior and the extended thin line 95%. ‘A’ plots predicted overlap, zero meaning posteriors do not overlap and one meaning posteriors are identical. ‘B’ plots the standard deviations of group-level effects (variation across predictors).

method had a similarly-shaped distribution to manual measurements but with longer values, while the Slope method had clusters at multiple modes.

In our model of manually-measured tap duration, we can only generalize results to taps that have a visible presence in a spectrogram. We only claim with confidence that increased speech rate is associated with shorter taps, and that taps are longer before /i/ than before /e/. Other specific effects show uncertainty regarding the presence of a meaningful association, but effects in specific directions are incompatible with our model. For some effects, large percentages of the posterior fall within the ROPE, indicating the most likely interpretation should be that these predictors are not associated with changes in tap duration.

When comparing estimates from our four models, the patterns of similarity and difference varied depending on the predictor. For some, like speech rate, the effects from all models are reliably negative, but some methods overestimate the effect’s size, which is likely due to overestimating tap duration overall. If we consider our manual model the gold standard, other methods make both Type S errors (getting the direction wrong) and Type M errors (wrong effect magnitude). This assumption is reasonable, as retrodictive checks showed the manual model fit the data well, while other models did not. Several predictors with posteriors centered around zero in the manual model show reliably negative or positive effects in one or more automated methods (e.g., Bigram freq. fol, Fol. stress). A potential explanation is that some variables are related to intensity changes independent of duration.

From our model of posterior overlap, all automated methods had less than 60% overlap with the manual model for an average predictor. This result is not reassuring, although the slope method, which may have slightly more overlap, also showed less variation across our predictors, possibly due to having wider posteriors than the other methods. Based on these results, we cannot recommend these automated methods be used to measure Spanish taps. We also must question the reliability of measuring segment duration using intensity more generally, and recommend researchers evaluate their measurement methods as standard practice.

Hand-correcting data is costly, but building knowledge on results skewed by measurement error will be more so. Sharing hand-corrected data publicly will allow for data to be used by the wider research community to answer research questions and generate informative priors that allow them to use their data more efficiently.

5. ACKNOWLEDGEMENTS

This research was supported by the Social Sciences and Humanities Research Council of Canada in the form of a CGS-D scholarship awarded to the first author.

6. REFERENCES

[1] F. Torreira and M. Ernestus, "The Nijmegen corpus of casual Spanish," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Paris: European Language Resources Association (ELRA), 2010, pp. 2981–2985.

[2] N. C. Henriksen, "Acoustic analysis of the rhotic contrast in Chicagoland Spanish: An intergenerational study," *Linguistic Approaches to Bilingualism*, vol. 5, no. 3, pp. 285–321, 2015.

[3] E. Willis and T. Bradley, "Contrast maintenance of taps and trills in Dominican Spanish: Data and analysis," in *Selected proceedings of the 3rd conference on laboratory approaches to Spanish phonology*, L. Colantoni and J. Steele, Eds. Somerville, MA: Cascadilla Proceedings Project, 2008, pp. 87–100.

[4] T. G. Bradley and E. W. Willis, "Rhotic variation and contrast in Veracruz Mexican Spanish," *Estudios de fonética experimental*, no. 21, pp. 43–74, 2012.

[5] M. Amengual, "Acoustic correlates of the Spanish tap-trill contrast: Heritage and L2 Spanish speakers," *Heritage Language Journal*, vol. 13, no. 2, pp. 88–112, 2016.

[6] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, no. 1, pp. 31–56, 2004.

[7] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[8] N. Warner and B. V. Tucker, "Phonetic variability of stops and flaps in spontaneous and careful speech," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1606–1617, 2011.

[9] U. Cohen Priva, "Informativity affects consonant duration and deletion rates," *Laboratory Phonology*, vol. 6, no. 2, pp. 243–278, 2015.

[10] T. Cho and J. M. McQueen, "Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress," *Journal of Phonetics*, vol. 33, no. 2, pp. 121–157, 2005.

[11] B. V. Tucker and Y. Mukai, *Spontaneous Speech*, ser. Elements in Phonetics. Cambidge University Press, 2023.

[12] J. Kingston, "Lenition," in *Proceedings of the 3rd conference on laboratory approaches to Spanish phonology*, L. Colantoni and J. Steele, Eds. Somerville, MA: Cascadilla, 2008, pp. 1–31.

[13] J. Katz and G. Pitzanti, "The phonetics and phonology of lenition: A Campidanese Sardinian case study," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 10, no. 1, pp. 1–40, 2019.

[14] J. Y. Kim and G. Repiso-Puigdelliura, "Deconstructing heritage language dominance: Effects of proficiency, use, and input on heritage speakers production of the Spanish alveolar tap," *Phonetica*, vol. 77, no. 1, pp. 55–80, 2020.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." 2022. [Online]. Available: http://www.praat.org/

[16] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of the annual conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[17] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[18] R. C. Team, "R: A language and environment for statistical computing," Vienna, Austria, 2021. [Online]. Available: http://www.R-project.org/

[19] J. K. Kruschke, "Bayesian analysis reporting guidelines," *Nature Human Behaviour*, vol. 5, no. 10, pp. 1282–1291, 2021.

[20] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 923–929.

[21] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework." *Journal of Open Source Software*, vol. 4, no. 40, p. 1541, 2019.