ARTICULATION RATE IN CONSONANTS AND VOWELS: RESULTS AND METHODOLOGICAL CHALLENGES FROM A CROSS-LINGUISTIC CORPUS STUDY

Roger Yu-Hsiang Lo, Márton Sóskuthy

Department of Linguistics, University of British Columbia
roger.lo@ubc.ca, marton.soskuthy@ubc.ca

ABSTRACT

We investigate how articulation rate is implemented by different segment types such as consonants and vowels using corpus data from eight unrelated spoken languages. Our study also sheds light on methodological issues that arise in the analysis of cross-linguistic corpus data. With forced-aligned data from 20-40 speakers for each language, we observe a robust effect of articulation rate on segment duration, such that vowels undergo a significantly greater extent of duration adjustment than consonants. However, the size of this effect varies across languages in ways that may be partially due to alignment quality and cross-linguistic differences in segment distributions. We propose methods for accounting for these factors and observe that bringing them into our model does indeed change some of the language-specific conclusions. We further discuss the cross-linguistic variation in our data in terms of language-specific phonetic implementation, and speculate on its implications for sound change.

Keywords: articulation rate, corpus, cross-linguistic, generalized additive mixed model

1. INTRODUCTION

Speech rate is an important and well-documented dimension of linguistic variation [1, 2, 3, 4, 5, 6]. We focus on *articulation rate*, based solely on segmental material to the exclusion of pauses, as opposed to *speaking rate*, which includes pauses as well. Articulation rate varies across age [7], gender [3], and dialect [8], and even within individual speakers and utterances [5]. However, much remains unknown about the implementation of changes in articulation rate. To this end, we ask the following question: how is variation in articulation rate reflected in consonant and vowel durations? It is reasonable to expect differences between consonants and vowels due to their articulatory and aerodynamic properties. An experimental study reports that the ratio of vocalic to consonantal material increases as articulation rate slows down [4] (though see [6]). In other words, they find that vowels stretch more than do consonants in slower speech. Here, we use speech corpora from multiple languages to address this research question. This allows us to test the universality of any observed effects, and also to look at variation across languages.

Corpus data come with significant challenges. The amount of speech data renders manual processing impractical and requires automated methods. The accuracy of these methods is often affected by recording quality. Therefore, when comparing corpora that were recorded under different conditions, it is crucial to explicitly account for quality differences. This is relevant for the current study, as we are interested in exploring cross-linguistic variation, but our corpora differ substantially in terms of their recording quality (see Section 2.1).

Cross-linguistic differences in linguistic structures and their use may also introduce confounds into statistical models. Of particular relevance to the current study is the fact that the precise make-up of broad categories such as consonants and vowels varies across languages: for instance, a given language may have a higher proportion of stops than another language. There may also be differences in the extent to which these finer-grained segment types respond to articulation rate (e.g., stops may be less stretchy than other consonants). Thus, any observed cross-linguistic differences at a broader level may simply be artifacts of different frequency distributions at a lower level (e.g., language A has more stops, which makes its consonants appear generally less stretchy). It may therefore be important to control for such differences across languages.

The current study attempts to answer the main research question while mitigating the aforementioned issues via statistical modeling. Thus, we look at the differential roles of consonant and vowel durations in carrying articulation rate, and explore variation across languages. At the same time, we also highlight the extent to which recording/alignment quality and varying segmental frequency distributions can distort the results of cross-linguistic corpus studies. This allows us to strengthen our results, and also provides important pointers for future work on cross-linguistic corpus data.

2. METHODOLOGY

2.1. Corpora

Our data come from three databases that contain speech samples from eight genetically unrelated languages: Korean and Mandarin (the Origins of Patterns in Speech or OoPS-Lab corpus, created in-house); Amharic, Georgian, Swahili, Turkish, and Vietnamese (IARPA; [9]); and English (Buckeye Speech Corpus; [10]).

Our own OoPS-Lab corpus consists of read and spontaneous speech from 20 native speakers per language.[1] All speech data was collected remotely on speakers' computers or mobile devices via a dedicated website and transcribed by undergraduate research assistants. Each language contains 2-3 hours of speech data, and the recordings are of reasonably high quality.

The Intelligence Advanced Research Projects Activity (IARPA) corpus was created to develop speech recognition technology for noisy telephone conversations [9]. As a result, the recording quality for these corpora is substantially lower. To mitigate this issue, we only include telephone conversations from a home or office environment from a total of 40 speakers per language, but quality issues remain (more on this in Section 2.2). We analyse a total of 1-2 hours of data per language.

The Buckeye Speech Corpus [10] contains high-quality spontaneous speech recordings from 40 speakers (20 female and 20 male) from Columbus, Ohio. Importantly, given that the phonetic alignments were manually checked, we expect the alignments to be more accurate than the other corpora used in this study. We can therefore use it to rule out the possibility that our results from the other corpora are artifacts of the forced-alignment process.

For the recordings from the OoPS and the IARPA corpora, we used the Montreal Forced Aligner [11] to obtain alignments. All durations were extracted using the PolyglotDB [12] corpus management package. We analyse a total of 1.429 million segments (median of 89,547 per language) representing 35,483 utterances (median of 2,822 per language). Analyses and visualisations were generated using R [13]. All data and code are available at https://bit.ly/3GVjt0J.

2.2. Analyses

Articulation rate is operationalized as average segment duration in seconds in an utterance (a stretch of speech surrounded by at least 150 ms of pauses on each side). Note that average duration is the *inverse* of rate: a 1 s utterance with 5 segments has a segment rate of $5/1 = 5$ and an average segment duration of $1/5 = 0.2$. We define this measure at the level of segments (as opposed to, e.g., syllables) as our dependent variable—the duration of individual segments—is also at this level.

Our research question is operationalized as follows: do the durations of consonants and vowels change at different rates as average segment duration (i.e., inverse articulation rate) changes? For instance, when the average segment duration is short, consonants and vowels may have the same duration; but when the average segment duration is high, vowels may be 1.5 times longer than consonants. This should correspond to diverging consonant *versus* vowel durations when plotted over average segment duration in log-log space.[2]

This question is tested using generalized additive mixed models (GAMMs; [14]) with **segment duration** as the outcome and **C/V** (consonant vs. vowel), **average segment duration** and their interaction as fixed effects. We include random smooths over **average segment duration** by **speaker** × **C/V** (i.e. separate groups for each combination of **speaker** and **C/V**) and by **language** × **C/V**. This allows the **C/V** effect to vary across speakers and languages. The online materials include the full model structure. We fitted one model as described above (baseline model); one that also controls for alignment quality (alignment model); and one that controls for alignment quality and different segment distributions across languages (alignment + segment model).

To quantify the alignment quality of the recordings, we randomly sampled 20 aligned segments for every speaker. A research assistant judged each alignment to be "OK" or "misaligned", where misalignment meant that more than 50% of the aligned interval was placed incorrectly. This measure focuses on gross misalignment, not fine-grained errors, which are more subjective and harder to identify. The proportion of "OK" alignments for individual speakers are shown in Figure 1A. These proportions were passed through a logistic function with a midpoint of 0.6 and a growth rate of 15 (Figure 1B). The resulting weights were then used to run a weighted GAMM, where each data point contributes to the model in accordance with their weight. This means

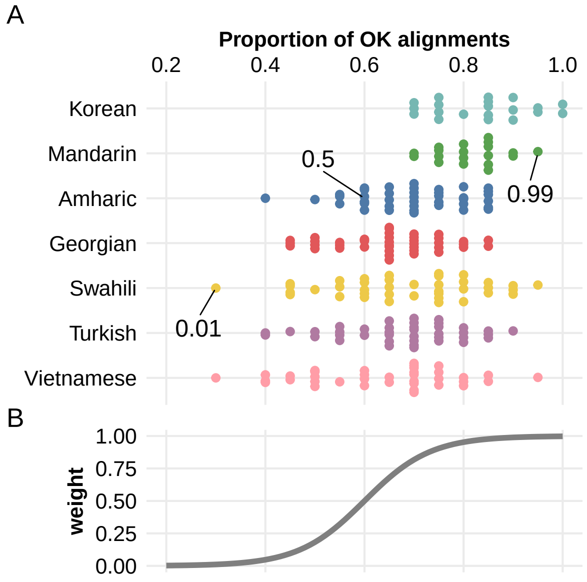

Figure 1: (A) Proportions of alignments judged "OK" across languages. Each point represents a speaker. (B) The mapping between proportions and model weights. The weights of three speakers are also annotated in the top panel.

that data points from speakers with lower alignment quality have less of an influence on model estimates.

Our third model controls for differing segment distributions across languages. While the proportion of vowels is approximately the same across all languages, the proportions of consonant types tend to vary more; for instance, Vietnamese has more nasals than all the other languages, while Amharic has more stop consonants. We bring this variation into the model by including a random smooth over **average segment duration** by **segment type**, where **segment type** has the levels STOP, AFFRICATE, FRICATIVE, LIQUID, NASAL, APPROXIMANT, HIGH VOWEL, and NON-HIGH VOWEL.

3. RESULTS

We first show the results from the alignment + segment model. Figure 2 shows model predictions plotted over data aggregated at the utterance level, with average consonant and vowel durations shown along the y-axis, and average overall segment duration along the x-axis. In all languages, vowels scale more readily with articulation rate than do consonants—that is, vowels are stretchier. These differences are significant across all languages. However, the effect sizes vary: for instance, Turkish shows only a small effect compared to the other languages. To compare the consonant and vowel slopes more holistically, we derive a metric called Δ effect size ($\Delta\beta$), defined as the difference in derivatives for the consonant and vowel prediction curves estimated at the median articulation rate. A bigger $\Delta\beta$ indicates a more pronounced difference in stretchiness. Turkish has the lowest $\Delta\beta$ (see Figure 4).

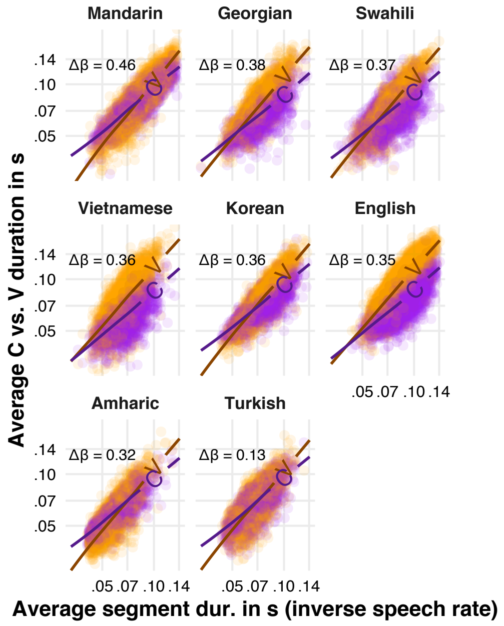

Figure 2: Average vowel and consonant durations as a function of average segment duration for each language. The dots are within-utterance averages. The lines are model predictions from the alignment + segment GAMM. Both axes are on the log scale.

Figure 3 shows the results from the three different models for Amharic. For Amharic, bringing alignment quality into the model only makes a small difference, but controlling for segment distributions leads to a much higher effect size. Figure 4 shows the changes in $\Delta\beta$ across the three models for all the languages. Weighting by alignment quality tends to lead only to modest changes in the size of the effect (with the exception of Georgian). However, some effect sizes shift markedly once we bring variation in segment distribution under control. One noticeable difference between the baseline model and the alignment + segment one is that while the former shows a wide range of variation in effect sizes, the latter shows a tight cluster with only Turkish and Mandarin as outliers. Controlling for non-phonetic sources of variation in the data removes a great degree of apparent cross-linguistic variation.

4. DISCUSSION AND CONCLUSION

Our findings echo the work of [4], showing that vowels are stretchier than consonants: as articulation rate decreases, the proportion of an utterance occupied by vowels tends to increase. This pattern is stable

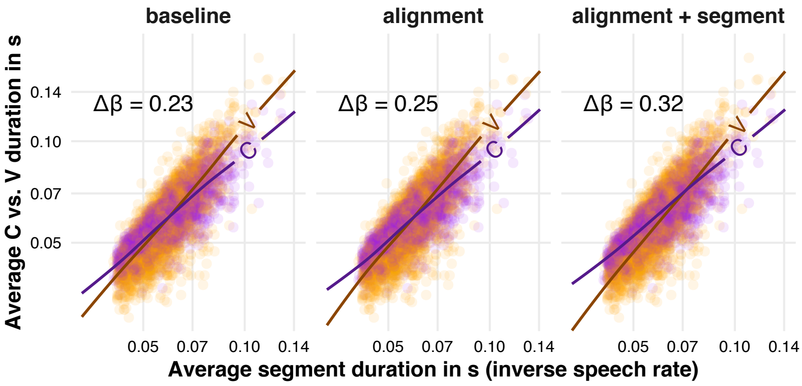

Figure 3: Differences in effect sizes between average consonant and vowel durations across models for Amharic. Each dot in the background represents an original data point. The solid lines show model predictions for the two segment categories. Note that both axes are on the log scale.

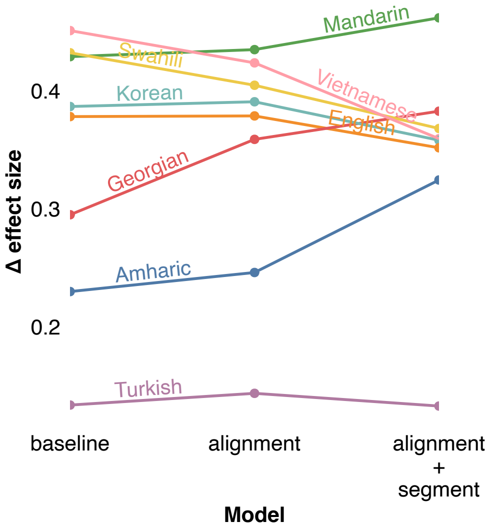

Figure 4: Changes in Δ effect size ($\Delta\beta$) across models for each language.

even after controlling for alignment quality and segment types. While all eight languages show a robust pattern, the size of the effect varies across languages. The relative ranking of different languages also changes across models with different degrees of control, and removing external sources of variation leads to more uniform estimates.

The differences in the way consonants and vowels carry articulation rate can be partly explained by reference to their articulation. Airflow is constricted for consonants, resulting in more aerodynamic and coordinatory complexity. To give an example, stops with closure voicing are impossible to sustain indefinitely due to the pressure build-up behind the constriction. It seems plausible that the added complexity of consonants makes them less responsive to changes in articulation rate.

Even after controlling for differences in segment distributions, languages do differ with respect to our key finding, suggesting that the same segment types (e.g., stops) can be more or less stretchy in different languages. This can potentially be attributed to subtle cross-linguistic differences in the phonetic implementation of the "same" sound (cf. [15]). For instance, [16] finds that English and Japanese /s/ exhibit differences in their spectral-temporal properties. Sundara [17] finds differences between the coronal stops of Canadian English and Canadian French in voice onset time, burst intensity and spectral shape. We speculate that at least some of the differences in our sample may be due to such fine-grained differences in phonetic implementation. This could be tested by taking a more detailed look at duration variation across different segment types in different languages, which, however, is outside the scope of this paper.

How articulation rate modulates the phonetic realization of segments also has implications for sound change. For instance, segments that are particularly inflexible in terms of their duration may be prone to 'catastrophic failure' under high articulation rate, that is, full deletion or reduction. To give another example, length contrasts may also be more susceptible to neutralization at high articulation rates in segments with limited durational maneuverability.

Finally, it is important to emphasise that the study conclusions do change as data quality and non-phonetic differences across languages are brought under statistical control. Cross-linguistic corpus phonetics [18, 19] has seen a marked surge in recent years. Our findings testify to the importance of accounting for systematic differences across data sources in cross-linguistic big data approaches. Statistical techniques such as the differential weighting of data points and random effects provide powerful tools for achieving this goal.

5. REFERENCES

[1] D. Byrd and C. C. Tan, “Saying consonant clusters quickly,” *Journal of Phonetics*, vol. 24, no. 2, pp. 263–282, 1996.

[2] T. H. Crystal and A. S. House, “Articulation rate and the duration of syllables and stress groups in connected speech,” *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 101–112, 1990.

[3] E. Jacewicz, R. A. Fox, and L. Wei, “Between-speaker and within-speaker variation in speech tempo of American English,” *The Journal of the Acoustical Society of America*, vol. 128, no. 2, 2010.

[4] V. A. Kozhevnikov and L. A. Chistovich, *Speech: Articulation and perception*. Washington, D.C.: Joint Publications Research Service, 1965.

[5] J. L. Miller, F. Grosjean, and C. Lomanto, “Articulation rate and its variability in spontaneous speech: A reanalysis and some implications,” *Phonetica*, vol. 41, no. 4, pp. 215–225, 1984.

[6] S. Wood, “What happens to vowels and consonants when we speak faster?” *Working Papers in Linguistics, Lund University*, vol. 9, pp. 8–39, 1973.

[7] H. Quené, “Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo,” *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1104–1113, 2008.

[8] M. P. Robb, M. A. Maclagan, and Y. Chen, “Speaking rates of American and New Zealand varieties of English,” *Clinical Linguistics & Phonetics*, vol. 18, no. 1, pp. 1–15, 2004.

[9] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy, “Developing speech recognition systems for corpus indexing under the IARPA Babel program,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6753–6757.

[10] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier. Buckeye Corpus of Conversational Speech (2nd release). Columbus, OH.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” in *Proceedings of INTERSPEECH 2017*, 2017, pp. 498–502.

[12] M. McAuliffe, E. Stengel-Eskin, M. Socolof, and M. Sonderegger, “Polyglot and Speech Corpus Tools: A system for representing, integrating, and querying speech corpora,” in *Proceedings of INTERSPEECH 2017*, 2017, pp. 3887–3891.

[13] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[14] S. N. Wood, *Generalized additive models: An introduction with R*, 2nd ed. Boca Raton, FL: CRC Press, 2017.

[15] J. B. Pierrehumbert, “What people know about sounds of language,” *Studies in the Linguistic Sciences*, vol. 29, no. 2, pp. 111–120, 1999.

[16] P. F. Reidy, “Spectral dynamics of sibilant fricatives are contrastive and language specific,” *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2518–2529, 2016.

[17] M. Sundara, “Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French,” *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1026–1037, 2005.

[18] M. Y. Liberman, “Corpus phonetics,” *Annual Review of Linguistics*, vol. 5, no. 1, pp. 91–107, 2019.

[19] E. P. Ahn and E. Chodroff, “VoxCommunis: A corpus for cross-linguistic phonetic analysis,” in *Proceedings of the 13th Conference of Language Resources and Evaluation*, 2022, pp. 5286–5294.

[1] One female Taiwan Mandarin speaker was excluded from analysis due to poor recording quality.

[2] We look at the effects in log-log space as it is possible to observe diverging raw durations that maintain a constant ratio.