Journal of Phonetics 118 (2026) 101530

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

A crosslinguistic corpus phonetic analysis of intrinsic vowel duration

Eleanor Chodroff *, Miao Zhang

University of Zurich, Department of Computational Linguistics, Andreasstrasse 15, 8050 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 29 September 2025
Received in revised form 20 July 2026
Accepted 21 July 2026
Available online 7 August 2026

Keywords:
Intrinsic vowel duration
Crosslinguistic phonetics
Typology
Corpus phonetics

ABSTRACT

This study investigated the crosslinguistic distribution of intrinsic vowel duration, the observation that high vowels tend to be shorter than low vowels. Using a large-scale corpus phonetic analysis, we analyzed duration contrasts among /i/, /u/, and /a/ in inter-obstruent contexts from over 60 languages and 16 language families in the VoxCommunis Corpus. Across languages, high vowels were shorter than low vowels in approximately 90% of comparisons, while no systematic durational asymmetry was observed between high front and high back vowels. The directional consistency of the height effect, along with a positive correlation between duration and F1 within low vowels, provides support for a biomechanical basis to the effect. However, we argue that an automatic biomechanical source critically depends on the assumption of uniform phonetic targets associated with the shared vowel feature that, when instantiated across different vowel height configurations, allow biomechanical factors to generate the observed durational pattern. Variation in the magnitude of the effect across languages and speakers, as well as some minor relationships between the height effect and vowel inventory size provide some limited support for contrast enhancement. Overall, these findings support intrinsic vowel duration as a robust statistical phonetic universal grounded primarily in a biomechanical source, while target uniformity makes explicit the phonetic target structure that allows the pattern to emerge.
© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Intrinsic vowel duration refers to the observation that "other things being equal, high vowels are shorter than low vowels" (Maddieson, 1996). The phenomenon is theoretically important as it has often been treated as a potential mechanistic phonetic universal: a crosslinguistically recurring pattern grounded in the physical properties of speech production. Although the effect has been reported in only a handful of languages, it is commonly accounted for by biomechanical explanations. For instance, the greater jaw displacement required for low vowels is assumed to yield correspondingly longer acoustic durations compared to high vowels (Lehiste, 1970; Lindblom, 1967). On this account, intrinsic vowel duration may reflect a general consequence of articulatory implementation rather than a language-specific pattern.

Nevertheless, the empirical foundation for treating intrinsic vowel duration as a crosslinguistically robust pattern remains limited. Existing evidence is largely restricted to Indo-European languages and small speaker samples, which leaves open the question of how consistent and robust the effect is when examined at scale. In addition, previous work has questioned whether intrinsic vowel duration can be attributed to biomechanical factors alone, raising the possibility that speaker-controlled phonetic targets or enhancement may also contribute to the observed pattern (Coretta, 2026; Solé and Ohala, 2010; Toivonen et al., 2015; Westbury and Keating, 1980). These two issues motivate the present study: first, the need for a broader empirical assessment of the crosslinguistic distribution of intrinsic vowel duration, and second, the need to clarify how the observed pattern bears on biomechanical and enhancement-based interpretations.

The present study addresses these issues through a large-scale corpus phonetic analysis spanning more than 60 languages and 16 language families. Our primary goal is to establish the crosslinguistic distribution of intrinsic vowel duration and to evaluate how the direction and variability of the effect bear on existing explanations. We further argue that biomechanical accounts depend on a key assumption: some aspect of phonetic realization must remain relatively uniform across vowels for differences in articulatory configuration to yield systematic differences in acoustic duration.

* Corresponding author.
E-mail addresses: eleanor.chodroff@uzh.ch (E. Chodroff), miao.zhang@uzh.ch (M. Zhang).

https://doi.org/10.1016/j.wocn.2026.101530
0095-4470/© 2026 The Author(s). Published by Elsevier Ltd.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1.1. Previous observations

Intrinsic vowel duration differences have been reported in a number of languages, though existing evidence has been limited in terms of linguistic diversity and sample size. Most observations come from Indo-European languages, including Catalan (Solé and Ohala, 2010), Danish (Bundgaard, 1980; Fischer-Jørgensen, 1955), English (Black, 1949; Crystal and House, 1988; Hillenbrand et al., 1995; House and Fairbanks, 1953; Peterson and Lehiste, 1960; Solé and Ohala, 2010; Tauberer and Evanini, 2009), French (O'Shaughnessy, 1981), German (Maack, 1949; Meyer, 1904), Northwestern Italian (Coretta, 2026), Saterland Frisian (Schoormann et al., 2017, for long vowels), Spanish (Marín Gálvez, 1995; Navarro, 1916; Stone, 2015), and Swedish (Elert, 1964; Lindblom, 1967; Toivonen et al., 2015). Comparatively, few studies have explicitly examined the durational difference between high and low vowels in non-Indo-European languages (Äimä, 1918 for Finnish; Johnson and Martin, 2000 for Muscogee, also known as Creek; Solé and Ohala, 2010 for Japanese; Lapierre et al., 2026 for Panãra), and some frequently cited cases, such as Thai (Abramson, 1974), are based on limited data and do not directly mention the effect. One notable exception to the high-low pattern has been reported for Muscogee (Johnson and Martin, 2001), spoken in North America, although the authors suggest this result may be attributable to prosodic imbalances in the dataset.

In addition to limited language coverage, most studies rely on very small speaker samples. Many analyses include fewer than four speakers, and several are based on a single speaker. Even the largest studies are restricted to a small number of languages, primarily American English, with studies ranging from 16 speakers in Black's (1949) final analysis to 435 speakers in Tauberer and Evanini's (2009) corpus study, alongside intermediate-sized samples in Hillenbrand et al. (1995) and Long et al. (2018). More recent studies include 35 speakers of Saterland Frisian (Schoormann et al., 2017) and, outside Indo-European, 21 speakers of Panãra, a Jê language spoken in Brazil (Lapierre et al., 2026). Apart from these studies, the small samples typical of the literature limit the ability to estimate population-level effects for a language and to characterize variation across speakers within a language.

1.2. Sources of the intrinsic vowel duration effect

Intrinsic vowel duration is often discussed within a broader class of proposed mechanistic universals, defined as phonetic patterns that arise automatically from the biomechanics of speech production (Maddieson, 1996). Other frequently cited examples include intrinsic vowel f0, in which high vowels have a higher fundamental frequency (f0) than low vowels, consonant f0, in which vowel onset f0 is higher following voiceless than voiced consonants, and place-related differences in stop voice onset time (VOT), in which labial stops have shorter VOTs than dorsal ones (see Keating, 1985; Maddieson, 1996 for a review).

Mechanistic universals are traditionally explained as a result of biomechanical factors in articulation. For intrinsic vowel duration specifically, durational differences are assumed to arise as a consequence of the articulatory movements required to achieve different vowel height targets. Lindblom (1967) developed a model in which intrinsic vowel duration arises from the interaction of jaw movement, lip separation, and coarticulatory timing between the two. In this model, the timing and duration of applied jaw force are assumed to be constant across vowels, while the actual jaw force differs. Because the jaw must travel further to reach the low vowel setting, the acoustic interval during which the vocal tract remains open (measured by the midsagittal lip separation) is correspondingly longer (see Fig. 1). In addition, Lehiste (1970) proposed that low vowels are longer than high vowels because the jaw must travel a greater distance to reach the lower articulatory configuration (p. 19).

Implicit in these biomechanical accounts is the assumption that while some aspects of articulation differ across vowel categories, others remain uniform (Chodroff and Wilson, 2017,

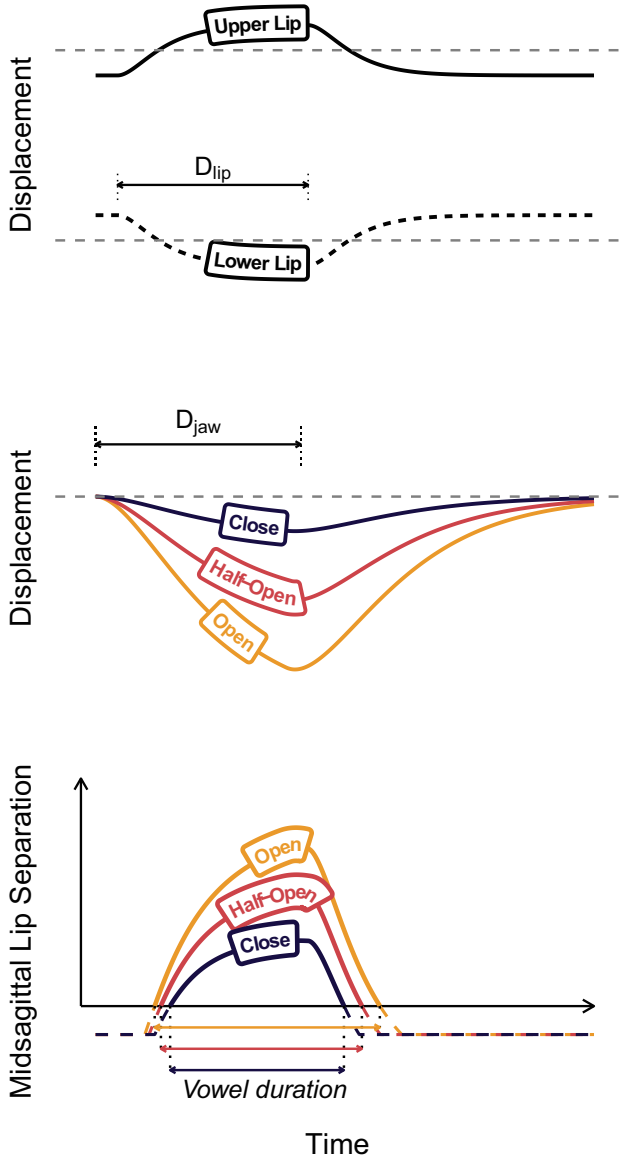


Fig. 1. The model from Lindblom (1967) in which a constant upper and lower lip displacement duration (top panel), constant jaw displacement duration, but different jaw force inputs per vowel height (middle panel), can result in different acoustic vowel durations by height as measured in the midsagittal lip separation. This figure is based on Figures I-A-5 and I-A-6, derived from Equation (3) on p. 5 in Lindblom (1967).

2022; Keating, 2003). Related proposals have also considered this in terms of gestural economy (Maddieson, 1995) or reuse of phonetic targets (Ménard et al., 2008). This assumption of uniformity is important because it concerns constraints on phonetic targets or articulatory parameters that are physically implemented (i.e., the input), rather than the biomechanical consequences of implementation itself (the observed output). In Lindblom's (1967) model, the jaw force duration and lip separation input are explicitly assumed to be the same for all vowels. Under these assumptions, the observed acoustic differences emerge as a mechanical consequence of implementing identical temporal targets across vowels with different articulatory configurations. Similarly, in Lehiste (1970)'s proposal, the jaw displacement differs between high and low vowels, but we note that the velocity of jaw movement would need to remain constant in order for an automatic durational difference to arise. If speakers instead produced low vowels with faster jaw movement or high vowels with slower movement, the durational difference could disappear or even reverse.

An alternative perspective is that speakers may actively control phonetic cues for enhancement of a phonetic or phonological contrast. Enhancement accounts are often framed in terms of speaker control, though we note the term enhancement has been used to refer to a wide range of mechanisms that increase the perceptual distinctiveness between contrasts (Kirby and Harrington, 2026; Ting et al., 2025). In some approaches, enhancement involves perceptual integration among multiple acoustic cues that jointly signal a phonological feature or contrast (Kingston and Diehl, 1994). In others, speakers may recruit or adjust additional cues to support an existing contrast, allowing cues to enter into a compensatory or trading relation (e.g., a "trading relation-based enhancement": Dmitrieva et al., 2015). These and related proposals have been discussed not only for intrinsic vowel duration, but also for other proposed mechanistic universals, including intrinsic vowel f0 and consonant f0 (Kingston and Diehl, 1994; Kingston, 2007; Kirby and Ladd, 2016; Solé and Ohala, 2010; Ting et al., 2025). What these proposals share is the assumption that phonetic realization is not determined solely by biomechanics but may also be shaped by pressures to maintain or increase contrastive distinctiveness.

Evidence supporting such interpretations often comes from perceptual studies demonstrating that listeners rely on multiple cues in the identification of speech sounds. For example, listeners have been shown to use f0 in addition to F1, the primary correlate of vowel height, when identifying vowel height contrasts (Hoemeke and Diehl, 1994; Miller, 1953; Potter and Steinberg, 1950). Similarly, the f0 of a vowel onset contributes to the perception of consonant voicing alongside VOT (Haggard et al., 1981; Massaro and Cohen, 1976; Summerfield and Haggard, 1977). In these cases, speakers might actively manipulate secondary cues, such as f0, to reinforce the phonological contrasts of vowel height or consonant voice.

For intrinsic vowel duration, some studies indicate that listeners do indeed perceive duration differently depending on the vowel height. For instance, Gussenhoven (2007) found that Dutch listeners perceived high vowels as sounding longer than mid or low vowels even when the acoustic durations were equal. Similar results were reported in Stone (2015), in which American English listeners perceived high vowels as longer than low vowels in English and Swedish stimuli. This suggests that a physically shorter high vowel may be perceptually comparable in duration to a physically longer low vowel. Solé and Ohala (2010) further reported that Catalan listeners used vowel duration in distinguishing the relatively higher /e/ from /ɛ/. Such findings raise the possibility that speakers could enhance vowel height contrasts by manipulating vowel duration. Importantly, though, the use of a phonetic dimension in perception does not necessarily mean that the speaker intentionally produced that difference.

Articulatory evidence has also been used to distinguish between automatic biomechanical explanations and accounts invoking speaker control. Westbury and Keating (1980) conducted an electromyographic study of American English vowel production and found that acoustic vowel duration correlated with jaw displacement, with low vowels longer than high vowels. Consistent with Lindblom (1967)'s assumptions, the relevant jaw-lowering muscle exhibited greater activation. However, contrary to the assumption of a constant duration of jaw force input, the duration of muscle activation was also longer for /a/ than for /i/. Westbury and Keating (1980) interpreted this result as evidence that speakers may specify distinct articulatory durations for different vowels rather than relying on a uniform timing target specified across vowels.

Other studies have examined whether intrinsic vowel duration persists across speaking rates. Solé and Ohala (2010) proposed that if vowel duration differences are purely automatic consequences of articulation, then they should remain constant regardless of overall speaking rate. In contrast, if speakers actively enhance the contrast, the durational difference should scale proportionally with overall speaking rate. In a crosslinguistic comparison of American English, Catalan, and Japanese, Solé and Ohala (2010) found that durational differences between high and low vowels remained constant across speaking rates in Japanese but varied proportionally to speaking rate in American English and Catalan. They interpreted this pattern as evidence that intrinsic vowel duration reflects an automatic articulatory consequence in Japanese but active enhancement in American English and Catalan.

Finally, recent studies have questioned whether biomechanical explanations alone can account for the observed patterns of vowel duration. One line of evidence comes from studies of the relationship between vowel duration and F1, an acoustic correlate of vowel height. If intrinsic vowel duration arises automatically from articulatory displacement associated with vowel height, then within-category variation in vowel height might be expected to covary with duration. In American English, Tauberer and Evanini (2009) found categorical duration differences across vowel qualities, with /i/ shorter than /a/, but no systematic covariation between duration and F1 within vowel categories. Toivonen et al. (2015) reported a similar pattern for English and Swedish. Coretta (2026), by contrast, found relationships between duration and F1 in the expected automatic direction for /a/, /e/, and /o/, but not for /i/ or /u/. These findings were interpreted as evidence that the duration of the vocalic gesture may contribute to observed

vowel duration, while category-specific timing specifications are still needed to account for vowel-specific duration patterns.[1] Overall, these studies suggest that vowel duration is not reducible to vowel height alone. Biomechanical accounts provide a coherent explanation for the general direction of the effect, but existing evidence leaves open the possibility that speaker-controlled phonetic targets and enhancement processes may also contribute.

1.3. Current study

In the current study, we investigate the crosslinguistic distribution of durational differences across vowel height and backness using a large-scale corpus phonetic approach. Specifically, we analyze duration contrasts among /i/, /u/, and /a/ in inter-obstruent contexts from over 60 languages and 16 language families in the VoxCommunis Corpus (Ahn and Chodroff, 2022; Zhang et al., 2025a). The analyses are organized around three empirical questions:

1) How stable is the intrinsic vowel duration effect between high and low vowels across languages and individual speakers?
2) Does vowel duration covary with vowel height, indexed acoustically by F1, as expected under an automatic biomechanical relationship between the two dimensions?
3) Does the magnitude of the height effect vary systematically across languages, for instance with vowel inventory size or the phonological status of duration, in ways that suggest modulation beyond the baseline biomechanical effect?

Languages with larger vowel inventories may be more likely to show enhanced intrinsic duration effects, whereas the presence of a phonological length contrast may limit the extent to which duration can be recruited to support additional height-based contrasts. In addition to examining the height effect directly, we compare /i/ and /u/ as a baseline for between-category durational variability. Because these vowels share the same height value, there is no clear biomechanical basis for a consistent durational asymmetry between them. Overall, these analyses characterize the direction and variability of intrinsic vowel duration across languages and provide a basis for evaluating how the observed pattern bears on biomechanical and enhancement-based accounts.

2. Methods

2.1. Speech corpus

2.1.1. Mozilla common voice dataset

Language data was sourced from the Mozilla Common Voice Dataset (Ardila et al., 2020), with phone- and word-level alignments from the VoxCommunis Corpus (Ahn and Chodroff, 2022; Zhang et al., 2025a). The Mozilla Common Voice Corpus is an online, crowdsourced collection of speech recordings in over 120 languages. Although Common Voice now includes efforts to collect spontaneous speech, the present study uses the original read speech datasets, in which speakers record prompted sentences. For each available language, users can contribute to the corpus in three ways: by writing sentences for recording, producing recordings of existing sentences, or validating recordings or sentences contributed by other users. For recordings, users are instructed to read the sentence as faithfully as possible. Users may optionally self-report demographic information. During validation of recordings, users listen to recorded sentences and provide an upvote if the recording faithfully represents the intended sentence. Recordings do not need to be in a standard accent or free of background noise, but each expected linguistic unit must be audible and recognizable. A recording is considered validated if it receives at least two upvotes and more upvotes than downvotes. Each language dataset includes MP3 recordings, sentence transcripts, upvote and downvote counts, sentence and client IDs, and speaker demographic information when available.

2.1.2. VoxCommunis corpus

The VoxCommunis Corpus contains phone- and word-level alignments for the validated portion of 76 language datasets available through Common Voice (Ahn and Chodroff, 2022; Zhang et al., 2025a). The alignments were created using an automated pipeline that converted the orthographic transcript to a canonical phonetic transcription, which was then force-aligned to the audio file. A language was considered for processing if it had at least two hours of validated recordings and if we could obtain a reasonable pronunciation lexicon. A detailed description of the processing pipeline can be found in Appendix A1. The assumed inventories, resulting phone- and word-level time alignments, and updated speaker IDs are publicly available.[2] All original audio files can be downloaded from Mozilla Common Voice.[3]

2.2. Measurement

For each language and length series (short or long), the most frequent high front, high back, and low vowel were selected for measurement. We refer to a contrast between two vowel qualities within a given language and vowel length series as a *comparison*. The high front vowel categories were [i ɪ ɨ y ʏ ə]; the high back vowel categories were [u ʉ ʊ ɯ]; and the low vowel categories were [ɑ a æ ɐ ɒ].[4] For convenience, we will refer to the selected vowel from each region as /i/, /u/, and /a/.

Measurements were extracted from vowel tokens provided they occurred in the initial 50% of the utterance, had a defined midpoint f0, and occurred between two obstruents. The initial 50% of the utterance was selected to avoid major effects of

[1] Note that Tauberer and Evanini (2009) computed correlations between mean F1 and mean duration across dialect regions, whereas Toivonen et al. (2015) examined token-level correlations between F1 and duration within individual speakers. We follow Toivonen et al. (2015), as well as Coretta (2026) in treating within-speaker token-level covariation as the more relevant measure for assessing whether vowel height and duration covary in production; Coretta implemented this approach using mixed-effects regression with by-speaker random effects.

[2] The data are available on the Hugging Face repository https://huggingface.co/datasets/pacscilab/VoxCommunis. The original VoxCommunis repository can also be found at https://osf.io/t957v/, but due to space limitations, we have since migrated the repository to Hugging Face. For posterity, we will continue to keep the OSF repository active, but as of writing, all new additions to the corpus will be on Hugging Face.

[3] As of writing, the Common Voice datasets can be downloaded from https://datacollective.mozillafoundation.org/.

[4] [ə] was only considered a high front vowel for Abkhaz and otherwise excluded from the analysis. Stress was indicated on the token for Catalan only, and both stressed and unstressed vowels for a given vowel quality were combined into one category.

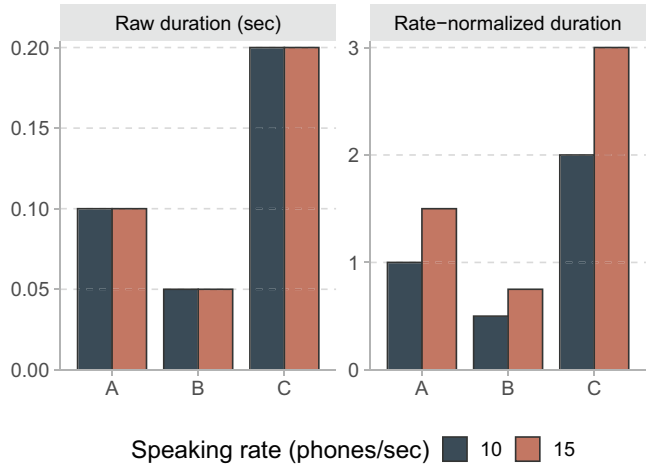


Fig. 2. A demonstration of the applied speaking rate correction to three hypothetical vowel durations (A, B, C) in two different speaking rate contexts (10 phones/sec, 15 phones/sec). The left panel shows the same raw durations (sec) within different speaking rate contexts, while the right panel shows how these values are adjusted with the multiplicative speaking rate correction. The same physical vowel duration should seem longer within a faster speaking rate (15 phones/sec) than a slower one (10 phones/sec).

phrase-final lengthening, and the necessity of a defined f0 was to ensure that the aligned interval was likely vocalic. The obstruents were defined as all stops and fricatives ([p b t d ʈ ɖ c ɟ k g ɡ q ʔ ɸ β f v θ ä s z ʃ ʒ ʂ ʐ ç ʝ ɕ ʑ x ɣ χ ʁ ħ ʕ h ɦ]) and did not need to be within the same word as the target vowel. In addition, the utterance had to contain at least five phones, and the word labels of the preceding and following phones were recognized words in the alignment procedure.[5] The minimum of five phones was chosen to obtain slightly better estimates of speaking rate. The extracted measurements were raw vowel duration, midpoint vowel f0, number of phones within the utterance, utterance duration (measured as the difference between the offset of the final phone and the onset of the initial phone), the preceding phone label, the following phone label, the word label of the target vowel, the word label of the preceding phone, and the word label of the following phone. The phone and word labels were extracted such that we could control for these in the statistical models. We then derived the speaking rate as the number of phones minus the target vowel, divided by the utterance duration minus the target vowel duration in seconds (Tilsen and Tiede, 2023).

The normalized vowel duration was then calculated as the target vowel duration in seconds multiplied by the speaking rate (phones per second). This measurement reflects the target vowel's duration relative to the average phone duration in that utterance, excluding the target phone. In general, as shown in Fig. 2, a 0.1 sec vowel duration produced in an utterance of 15 phones/sec ("fast", 0.1 * 15 = 1.5) should be perceived as longer than a 0.1 sec vowel duration produced in an utterance of 10 phones/sec ("slow", 0.1 * 10 = 1.0). This can also be interpreted as the first vowel being equivalent to 1.5 average phone durations within the utterance and the second vowel as equivalent to 1.0 average phone duration within the utterance.

Additional filtering and primary analyses were then conducted separately for the /i/–/a/, /u/–/a/, and /i/–/u/ contrasts. We first removed speakers with fewer than ten tokens for each of the two categories within the contrast (each of /i/ and /a/, /u/ and /a/, or /i/ and /u/), then calculated the speaker-specific mean and standard deviation for the normalized vowel duration and removed tokens more than 2.5 standard deviations from the speaker-specific mean. Following this, speakers with fewer than ten tokens per category and comparisons with fewer than five speakers were removed from the analysis.[6]

2.3. Final dataset

Following measurement and filtering, we analyzed 67 languages for the /i/–/a/ contrast (82 comparisons across 18 language families), 65 for the /u/–/a/ contrast (77 comparisons across 16 families), and 65 for the /i/–/u/ contrast (78 comparisons across 16 families).[7] The two language families represented only in the /i/–/a/ dataset were Abkhaz-Adyghe (Abkhaz) and Koreanic (Korean). The languages and their families are depicted geographically in Fig. 3, with language family information from Glottolog (Hammarström et al., 2025).

2.4. Manual validation

In addition to the automatic pipeline, we manually aligned a subset of vowel tokens from the final /i/–/a/ analysis. For each comparison, we randomly selected five speakers and ten tokens each of /i/ and /a/ from the automated dataset for a total of 8200 vowels (82 comparisons, five speakers, 20 tokens).[8] During this process, we identified occasional speaker heterogeneity within some original Common Voice client IDs. This issue was addressed in the full corpus using the speaker similarity scores described in Zhang et al. (2025b) (see Appendix A1). A few vowel tokens were subsequently excluded due to low speaker similarity scores. No systematic issues were noticed in the vowel labels. The final validation set included 8044 manually aligned vowels, with each comparison retaining at least 95 of the 100 expected manually aligned vowels. Manual and automatic mean differences were then estimated from the same token set, using the difference between the grand means of the five selected speakers' /i/ and /a/ averages.

2.5. Analyses

2.5.1. Analyses of manually validated data

Manual validation was conducted to assess whether the automatic alignments preserved the expected relative duration

[5] In a few cases, a word was "out of vocabulary" in that it did not have a phonetic transcription in the pronunciation lexicon. These were then skipped in alignment and thus excluded from the analysis.

[6] The Dhivehi – long /u/–/a/, Dhivehi – long /i/–/u/, and Ligurian /i/–/u/ comparisons were also removed. For Dhivehi, the long /u/ was always 10 ms, which was suggestive of an alignment problem and would pose a problem for a model of variance in duration. For Ligurian /i/–/u/, the linear mixed-effects model did not converge even with a minimal random intercept for speaker.

[7] As an example, Czech /i/–/a/, Czech /iː/–/aː/ (Czech – long /i/–/a/) and Yoruba /i/–/a/ would count as three /i/–/a/ comparisons.

[8] Because the manual audit involved verifying every measurement and this was a very time-consuming task, relying on data from just five speakers was tolerated. For the main analysis, all measurements were obtained from the automatic alignments. To mitigate variability due to alignment quality, we wanted more speakers in the automatic analysis, particularly in case data from some speakers was more affected than from others.

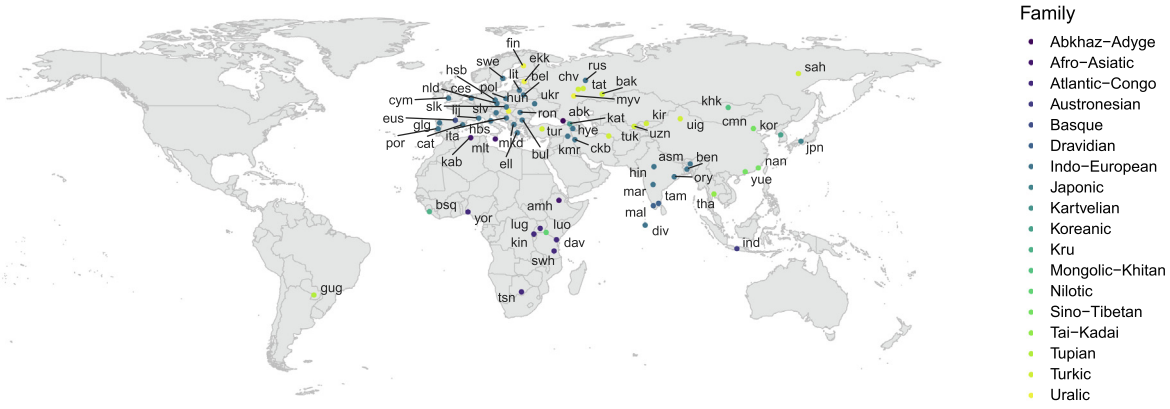


Fig. 3. Geographic distribution of languages represented in the analysis. Language families are represented by color; space-permitting, specific languages are labeled with their ISO-693–3 code. A full inventory of languages is provided in Table A1.

pattern between /i/ and /a/ across comparisons and to evaluate the overall quality of the forced alignment. We first compared the mean /i/–/a/ durational difference estimated from the manual alignments with the corresponding estimate from the automatic alignments for the five manually aligned speakers in each comparison. This provided a descriptive assessment of the agreement between the manual and automatic analyses.

We then fit a linear mixed-effects regression to test whether vowel duration differed by alignment type and whether this difference depended on vowel quality. The model included alignment type (automatic +1, manual –1), vowel height (/i/ +1, /a/ –1), and their interaction, along with by-comparison random intercepts and slopes for vowel quality. Pairwise differences between manual and automatic durations were estimated separately for /i/ and /a/ using the emmeans R package (Lenth, 2025).

A second linear mixed-effects regression assessed absolute deviations between manual and automatic boundaries. This model included boundary position (onset +1, offset –1), vowel height (/i/ +1, /a/ –1), and their interaction, with a random intercept by comparison; additional random slopes did not converge. This analysis tested whether alignment error differed between vowel onsets and offsets and whether alignment precision depended on vowel quality.

Finally, for consistency with previous studies assessing forced alignment performance, we report the percentage of onset and offset boundaries that fell within 20 ms of the corresponding manual boundary.

2.5.2. Analyses of automatically aligned data

The automatically aligned data were used for four sets of analyses: comparison-specific models of vowel quality effects, by-speaker analyses of duration rankings, by-speaker correlations between vowel duration and F1, and cross-language models of variation in effect size.

Comparison-specific vowel quality effects. For the analysis of the relationship between vowel duration and vowel quality,[9] a separate mixed-effects linear regression model was fit to the rate-normalized vowel duration of each comparison within the /i/–/a/, /u/–/a/, and /i/–/u/ subsets. Rate-normalized vowel duration was predicted from vowel quality, all possible interactions between preceding voicing, preceding manner, following voicing, and following manner, as well as the utterance proportion of the target vowel. An additional analysis of raw vowel duration with a further predictor of speaking rate and its interaction with vowel quality is reported in Appendix A2. For /i/–/a/ and /u/–/a/, the effect of vowel height was sum-coded with the high vowel set to +1 and /a/ set to –1; for /i/–/u/, the effect of vowel backness was sum-coded with /i/ set to +1 and /u/ set to –1; all other categorical predictors were sum-coded. A by-speaker random intercept and slope for vowel quality was also included, along with a by-word random intercept.[10] Given the large number of comparisons (82 comparisons for /i/–/a/, 77 for /u/–/a/, and 78 for /i/–/u/), *p*-values were adjusted using the Benjamini–Hochberg method while retaining an alpha level of 0.05 for significance (Benjamini and Hochberg, 1995).

By-speaker analyses. To assess the consistency of intrinsic vowel duration across different individuals in each language, we also conducted a by-speaker analysis. For each speaker, we applied a simple linear regression, with rate-normalized duration as the dependent variable and vowel quality as a predictor. Speaker-specific models were run for each of /i/–/a/, /u/–/a/, and /i/–/u/ with treatment coding of the vowel quality predictor. In the two models comparing high and low vowels, /a/ was used as the reference category, while in the comparison between /i/ and /u/, /i/ served as the reference. The slopes from these models thus indicate the difference between the two vowels. For each language, we calculated the proportion of speakers in one of four categories determined by the direction of the slope (±) and the significance of the *p*-value. The number of speaker-specific models per comparison ranged from 5 to 4306 with a mean of 191 and a median of 44.

[9] We use the term "vowel quality" here instead of "vowel height" as the /i/–u/ analysis assesses the effect of vowel backness as opposed to vowel height.

[10] A few notes on exceptions for individual languages are provided here. For Cantonese, the predictors for context included only preceding manner of articulation; for Mandarin Chinese, the predictors for context included only preceding manner, following manner, and their interaction. For convergence, the random by-speaker slope for vowel quality had to be removed for several models. Removing the by-speaker slope for vowel quality results in the assumption that speakers have the same vowel quality durational difference, which can affect our estimation of the fixed effect for vowel quality. For the /i/–/a/ analysis, this was the case for Amharic, Assamese, Bassa, Taita – long, Finnish – long, Korean, Ligurian, Macedonian, Erzya, and Turkmen; for /u/–/a/, Assamese, Bassa, Taita – long, Hindi – long, Upper Sorbian, Ligurian, Malayalam, Odia, Saha – long, Slovenian, Serbian, and Turkmen; and for /i/–/u/, Bulgarian, Bengali, Finnish, Hindi – long, Upper Sorbian, Malayalam, Maltese, Min Nan, Sakha, Turkmen, Tswana, and Yoruba. For the /i/–/u/ analysis, Ligurian had to be removed as it did not converge with a random intercept for speaker.

The speaker-specific analyses were intended to characterize the overall pattern of results rather than to identify individual statistically significant effects. Accordingly, uncorrected *p*-values are reported only as descriptive summaries.

Duration–F1 correlations. To examine whether lower vowels automatically exhibited longer durations within each vowel quality, we also analyzed the by-speaker correlation between raw vowel duration and vowel F1, which served as an acoustic proxy for tongue or jaw height. We used raw vowel duration rather than rate-normalized duration because we were interested in the intrinsic relationship between actual duration and F1. F1 was calculated in hertz from the midpoint of the vowel using formant parameters estimated based on whether the utterance had an overall "high" (>160 Hz) or "low" ($\leq$160 Hz) f0 value. Formants were estimated using Praat's Formant (Burg) method. For tokens with "high" f0 values, the maximum formant was set to 5500 Hz and five formants were tracked; for tokens with "low" f0 values, the maximum formant was set to 4500 Hz and four formants were tracked. The F1 data was then merged with the filtered /i/–/a/ and /u/–/a/ datasets, the former of which provided the data for /i/ and /a/, and the latter for /u/. Using the WRS2 R package (Mair and Wilcox, 2020), a percentage bend correlation with a bending percentage of 20% was calculated for each speaker within each vowel category (/i/, /u/, and /a/) between vowel duration and F1. A percentage bend correlation first "bends in" outlier points, then applies a Pearson correlation.[11] Correlations were only implemented provided the speaker had at least ten tokens of the vowel category with measurable F1 values.

Cross-language variation. Finally, to assess cross-language variation, a linear fixed-effects model was implemented to assess variation in the estimated comparison-specific vowel height effects from the rate-normalized vowel duration models. The model included fixed effects of vowel inventory size and phonological length-contrast status. Phonological length-contrast status was a three-level predictor distinguishing vowel comparisons from languages without a phonological length contrast, short vowel comparisons from languages with a length contrast, and long vowel comparisons from languages with a length contrast. The interaction between the number of vowels and phonological length-contrast status was also included. Phonological length-contrast status was treatment-coded using vowel comparisons from languages without a length contrast as the baseline; the number of vowels was centered in the data.

3. Results

3.1. Manually validated data

The relationship between the manual and automatic mean /i/–/a/ differences is shown in Fig. 4. Each point corresponds to one comparison, estimated from the five manually aligned speakers. Points on the diagonal indicate equivalent /i/–/a/ differences in the manual and automatic alignments. Most comparisons fell in the lower-left quadrant, indicating that /i/ was shorter than /a/ in both analyses. Three comparisons fell in the upper-right quadrant, with /i/ longer than /a/ in both analyses: Min Nan, Guaraní, and Swedish – long. Several comparisons fell in the upper-left quadrant, where the automatic alignment yielded /i/ > /a/ but the manual alignment yielded /i/ < /a/: Belarusian, Dhivehi, Dhivehi – long, Greek, Basque, Finnish, Galician, Indonesian, Kabyle, Lithuanian – long, Dholuo, Malayalam, Mongolian, Kinyarwanda, Slovenian, Serbian, and Swedish. Overall, based on the estimated means for the five speakers in each comparison, the manual analysis showed that /i/ was shorter than /a/ in 95% of comparisons (78/82).

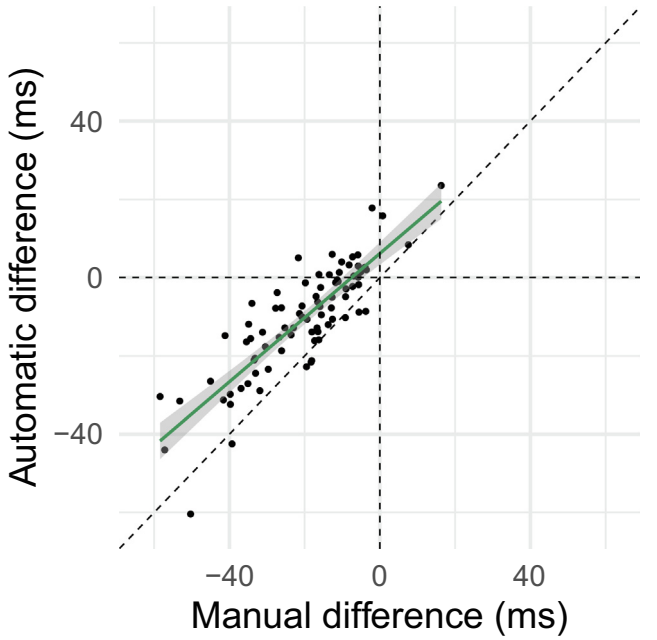


Fig. 4. Relationship between mean /i/–/a/ durational differences in milliseconds (ms) derived from manual and automatic alignments. Each point represents a single comparison, plotting the mean manual difference against the corresponding automatic difference. Negative values indicate that /i/ was shorter than /a/. The solid line shows the best-fit linear regression, and the shaded band represents the 95% confidence interval. The dashed diagonal line indicates the line of identity (i.e., equal durational differences in the manual and automatic alignments). Points above the diagonal reflect larger durational differences in the manual alignment than in the automatic alignment. Points in the lower-left quadrant indicate cases in which /i/ is shorter than /a/ in both alignments, and points in the upper-right quadrant indicate that /i/ is longer than /a/ in both alignments.

Durational differences between the manual and automatic alignments were assessed using a linear mixed-effects regression (Table 1). The automatic alignment tended to underestimate the difference between /i/ and /a/, driven primarily by an overestimation of the /i/ duration compared to the /a/ duration in the automatic alignment. Pairwise tests using emmeans indicated that automatic /i/ durations were significantly longer than manual /i/ durations by approximately 8 ms, whereas

Table 1
Summary of the linear mixed-effects model predicting vowel duration (ms) based on alignment type (automatic +1 or manual –1), vowel height (/i/ +1 or /a/ –1), and their interaction.

Term	β	*p*-value
Intercept	83.76	$p < 0.001$
Alignment type	1.31	$p < 0.001$
Vowel height	–7.74	$p < 0.001$
Alignment type:Vowel height	2.53	$p < 0.001$

[11] If the data distribution strictly follows a normal distribution, the percentage bend correlation should be identical to Pearson's correlation.

Table 2
Summary of the linear mixed-effects model predicting absolute boundary deviation between automatic and manual alignments (ms) from boundary position (onset +1 or offset –1), vowel height (/i/ +1 or /a/ –1), and their interaction.

Term	β	p-value
Intercept	15.65	$p < 0.001$
Boundary position	–1.64	$p < 0.001$
Vowel height	0.32	$p < 0.05$
Boundary position:Vowel height	–2.08	$p < 0.001$

automatic /a/ durations were significantly shorter than manual /a/ durations by approximately 2 ms ($\beta_{/i/}$ = 7.69, $p < 0.001$; $\beta_{/a/}$ = –2.44, $p < 0.01$).

Onset and offset boundary differences between the manual and automatic alignments were assessed using a second linear mixed-effects regression (Table 2). Offset boundaries showed significantly greater absolute deviations from the manual boundaries than onset boundaries, and this difference was especially pronounced for /i/ compared to /a/. Overall, the automatic analysis tended to underestimate both the magnitude of the /i/–/a/ effect and the number of comparisons conforming to the expected direction. Across comparisons, an average of 80% of onset boundaries fell within 20 ms of the manual boundary (/i/: 84%, /a/: 70%), compared to 72% of offset boundaries (/i/: 66%, /a/: 77%). The percentage of onset boundaries within 20 ms ranged from 16% (Min Nan) to 99% (Mongolian), with a median of 83%; for offset boundaries, the range was 26% (Armenian) to 99% (Mandarin Chinese), with a median of 75%.

3.2. Comparison-specific rate-normalized vowel duration

/i/–/a/. The estimated vowel height effect for each /i/–/a/ comparison is shown in Fig. 5a (67 languages, 82 comparisons). A negative value indicates that /i/ was estimated to be shorter than /a/, which was the expected direction (/i/ < /a/). The effect of vowel height was in the expected direction for 87% of comparisons (71/82) and significantly so for 82% of comparisons (68/82). The estimate of the difference ranged from –0.52 to 0.08 with a median difference of –0.07. An opposite and significant effect was observed in Swedish – long, Min Nan, Guaraní, Thai – short, Cantonese – short, Indonesian, and Kabyle vowels. An opposite effect that did not reach significance was observed in Finnish – long, Malayalam, Dhivehi – short, and Lithuanian – long vowels.

/u/–/a/. The estimated vowel height effect for each /u/–/a/ comparison is shown in Fig. 5b (65 languages, 77 comparisons). A negative value indicates that /u/ was estimated to be shorter than /a/, which again was the expected direction of the effect (/u/ < /a/). The effect of vowel height was in the expected direction for 90% of comparisons (69/77) and significantly so for 77% of comparisons (59/77). The estimate of the difference ranged from –0.61 to 0.13 with a median difference of –0.07. An opposite and significant effect was observed in

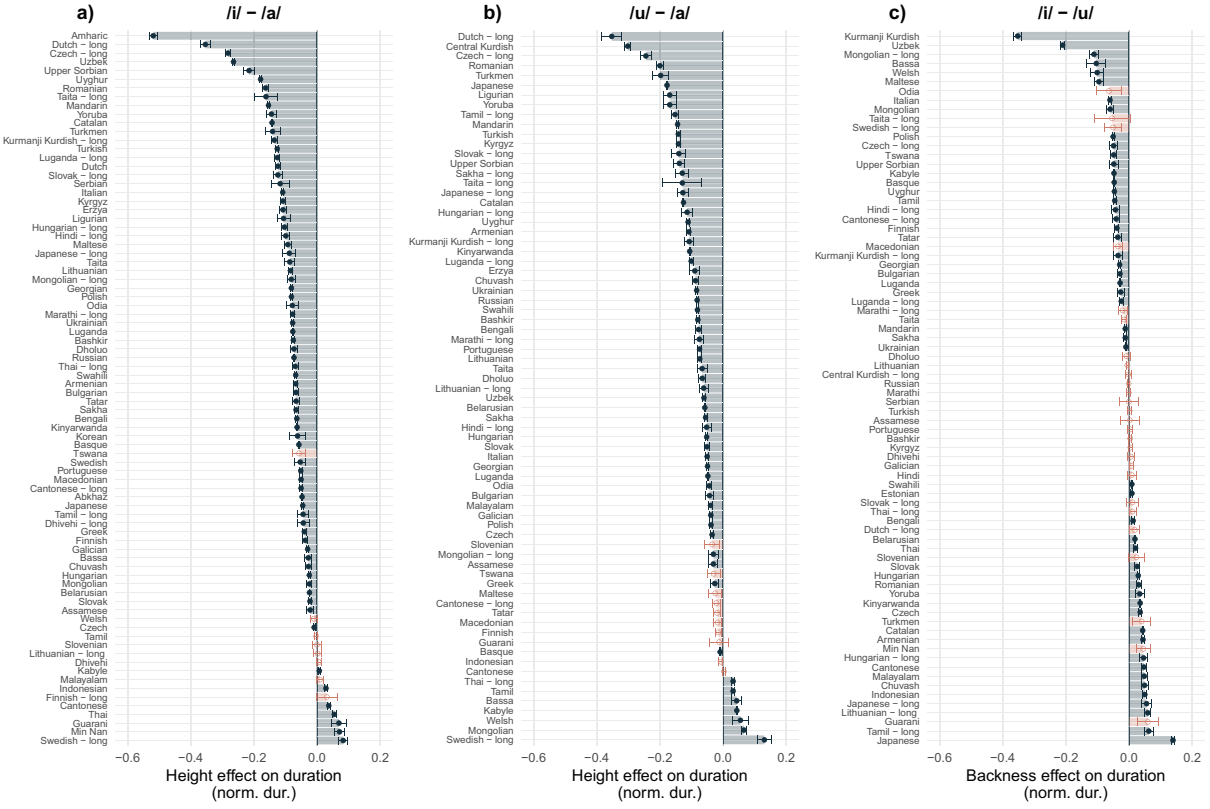


Fig. 5. Estimated effect of vowel quality (height or backness) on the rate-normalized vowel duration from the linear mixed-effects models for each comparison for a) /i/–/a/, b) /u/–/a/, and c) /i/–/u/. A negative estimate indicates that the first listed vowel had a shorter relative duration than the second listed vowel. A filled dark blue circle indicates a significant estimate; an open light orange circle indicates a non-significant estimate. Error bars reflect ± 1 standard error of the estimate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Swedish – long, Mongolian, Welsh, Kabyle, Basaa, Tamil, Thai – long, and Cantonese – short.

/i/–/u/. The estimated vowel backness effect for each /i/–/u/ comparison is shown in Fig. 5c (65 languages, 78 comparisons). A negative value indicates that /i/ was estimated to be shorter than /u/. /i/ was shorter than /u/ for 53% of comparisons (41/78) and significantly so for 37% (29/78). In addition, /u/ was shorter than /i/ for 47% of comparisons (37/78) and significantly so for 28% of comparisons (22/78). Overall, the effect did not differ significantly from zero for 35% of comparisons (27/78). The estimate of the difference ranged from –0.35 to 0.14 with a median of –0.001.

3.3. By-speaker analysis on rate-normalized vowel duration

Fig. 6 illustrates the distribution of by-speaker paired vowel duration rankings between /i/–/a/, /u/–/a/, and /i/–/u/. Across comparisons, most speakers exhibited a numerically shorter duration for high vowels than low vowels, regardless of statistical significance (87% of comparisons for /i/–/a/ and 78% for /u/–/a/). Additionally, 45% for /i/–/a/ comparisons and 44% for /u/–/a/ comparisons had a majority of speakers with a significantly shorter duration for high than low vowels. In contrast, for /i/–/u/, speakers were roughly evenly distributed in directionality across comparisons, though comparison-specific directional tendencies are present. Only 11 /i/–/u/ comparisons had more than 50% of speakers with a statistically significant durational difference (8 comparisons with /i/ > /u/ and 3 comparisons with /u/ > /i/).

3.4. Correlational analysis of vowel duration and F1

Fig. 7 shows the distribution of binned correlational magnitudes between vowel duration and F1 across speakers within each comparison. The biomechanical direction corresponds to a positive correlation, in which vowel duration increases with F1, where a high F1 corresponds to a lower tongue or jaw height. For /i/ and /u/, the by-speaker correlations spanned both positive and negative directions, where only 48% and 54% of speakers, respectively, showed a positive correlation. However, for /a/, 93% of speakers showed a positive correlation between vowel duration and F1. This pattern was also not limited to certain comparisons: all but one comparison had more than 50% of speakers in the positive direction for /a/ (99%, 80/81 comparisons), whereas only 38% and 56% of /i/ and /u/ comparisons, respectively, had more than 50% of speakers in the positive direction.

3.5. Cross-language variation of rate-normalized vowel duration

The relationship between the number of vowels, the presence of a phonological length contrast, and the durational difference between vowel qualities is shown in Fig. 8 with model results presented in Table 3.

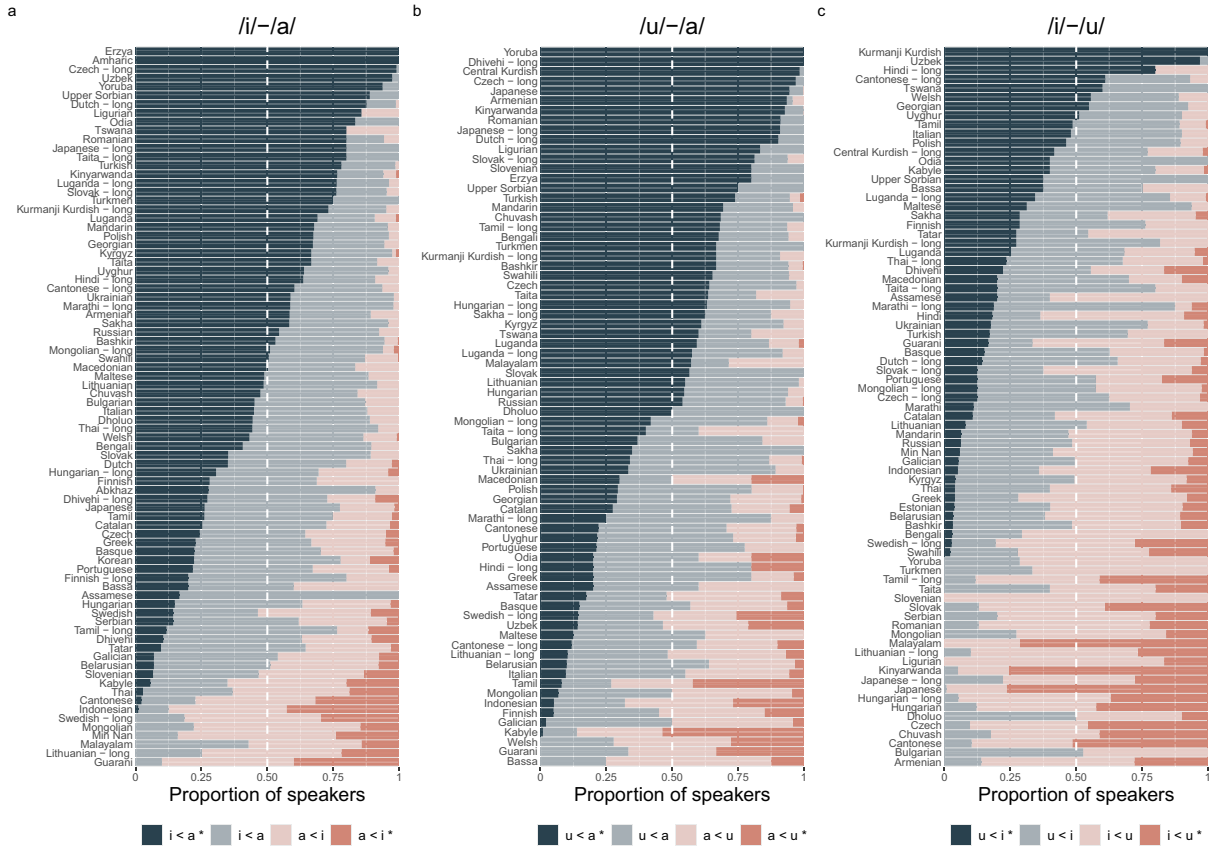


Fig. 6. Distribution of by-speaker vowel duration rankings for each comparison for a) /i/–/a/, b) /u/–/a/, and c) /i/–/u/. The asterisk (*) in the legend indicates a statistically significant difference at $p < 0.05$.

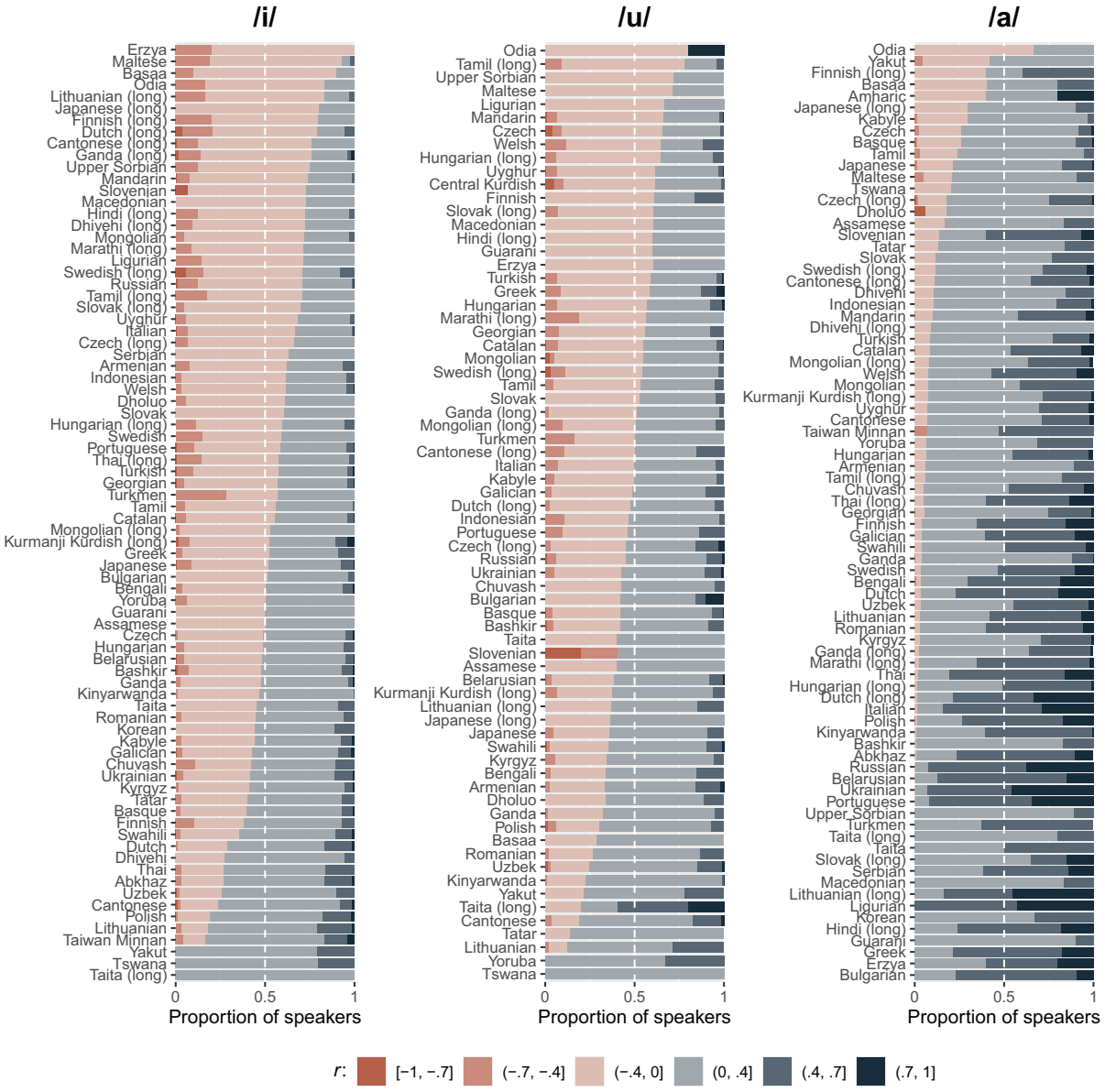


Fig. 7. Distribution of by-speaker correlations between vowel duration (ms) and F1 (Hz) for each comparison for a) /i/, b) /u/, and c) /a/.

/i/–/a/. In assessing cross-language variation for /i/–/a/, the estimated durational difference was significantly different from zero for comparisons from languages without a length contrast (null contrast comparisons). Short vowel comparisons from languages with a length contrast had significantly smaller estimated differences than null contrast comparisons, while long vowel comparisons did not differ from null contrast comparisons. The effect of vowel did not reach significance, where the expected effect would be an increase in the durational difference (in this case, a more negative number) with a greater number of vowels.

/u/–/a/. For /u/–/a/, the estimated durational difference was significantly different from zero for null contrast comparisons, and short and long vowel comparisons from languages with a length contrast did not differ significantly from null contrast comparisons. For this particular contrast, the number of vowels did significantly impact the size of the vowel height effect, but differently so depending on whether the vowel comparisons came from a language with or without a length contrast. Among null contrast comparisons, the estimated durational difference between /u/ and /a/ increased with inventory size, consistent with the predictions of enhancement. Note that a more negative number reflects a greater difference between the relatively short high vowel duration and long low vowel duration. However, among vowel comparisons from languages with a length contrast, this relationship was nullified or even reversed.

/i/–/u/. Finally, for /i/–/u/, the model revealed no significant effects or interactions. The durational difference did not differ significantly from zero, indicating no significant bias across comparisons in the direction of the front–back durational difference, regardless of whether the comparison came from a language with or without a length contrast. In addition, the number of vowels did not significantly impact the durational difference, regardless of whether phonological length was contrastive or not.

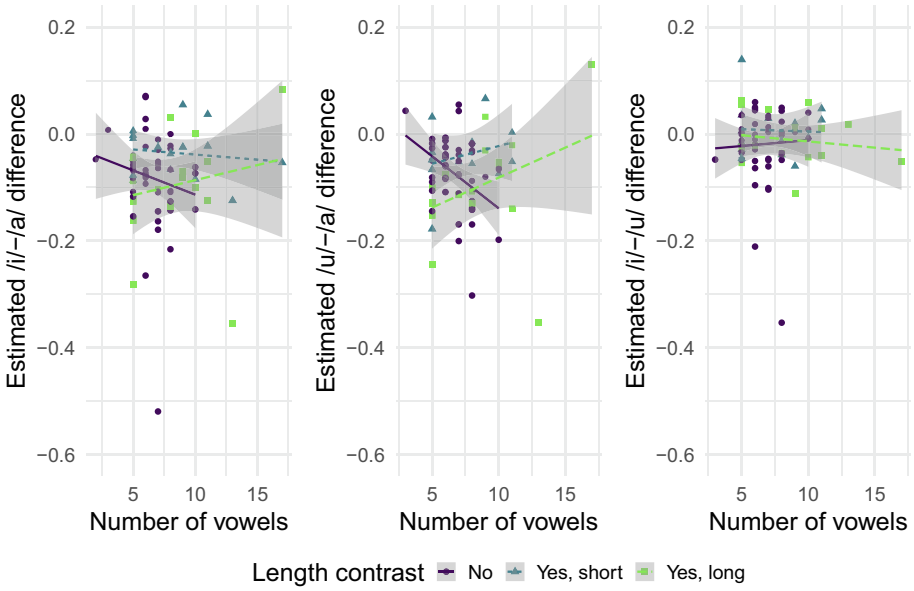


Fig. 8. Relationship between vowel inventory size and the estimated rate-normalized durational difference, by phonological length-contrast status, for a) /i/–/a/, b) /u/–/a/, and c) /i/–/u/. Purple circles indicate comparisons from languages without a phonological length contrast; teal triangles indicate short vowel comparisons from languages with a length contrast; lime squares indicate long vowel comparisons from languages with a length contrast. Each point reflects a single comparison's estimated vowel quality effect from the comparison-specific linear mixed-effects models described in Section 3.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Summary of the linear fixed-effects models assessing cross-language variation in the estimated vowel quality effects for /i/–/a/, /u/–/a/, and /i/–/u/. Estimates for the short and long vowel levels, and their interactions with inventory size, are reported relative to the null (no contrast) baseline. Note that a negative coefficient corresponds to an increased durational difference between the first and second listed vowels (first vowel shorter than second vowel).

Term	/i/–/a/		/u/–/a/		/i/–/u/	
	β	*p*-value	*β*	*p*-value	*β*	*p*-value
Null contrast	**–0.09**	**<0.001**	**–0.08**	**0.001**	–0.02	0.08
Short vowels	**0.06**	**<0.05**	0.04	0.08	0.02	0.22
Long vowels	–0.01	0.65	–0.03	0.22	0.01	0.61
Number of vowels × Null	–0.01	0.62	**–0.02**	**<0.05**	0.002	0.73
Number of vowels × Short	0.01	0.49	**0.03**	**<0.05**	–0.003	0.75
Number of vowels × Long	0.01	0.15	**0.03**	**<0.01**	–0.005	0.57

4. Discussion

The present study investigated the empirical stability of the intrinsic vowel duration effect in a large-scale crosslinguistic corpus analysis. Across more than 60 languages from at least 16 language families, the majority of languages and speakers showed the expected pattern that high vowels were on average shorter than low vowels. At the same time, the magnitude of the height effect varied continuously across languages and speakers, rather than clustering into clearly distinct types. Because this is an observational corpus study, the analyses cannot directly establish the causal mechanisms responsible for the observed patterns. They can, however, provide evidence for evaluating which accounts are consistent or inconsistent with the data.

4.1. Source of intrinsic vowel duration

Overall, the results conform closely to the direction predicted by traditional biomechanical explanations: across languages and speakers, high vowels are on average shorter than low vowels, with no systematic durational asymmetry between high front and high back vowels.[12] The prevalence of this high-low durational difference, present in approximately 90% of comparisons for both /i/–/a/ and /u/–/a/, supports its classification as a statistical phonetic universal, reflecting a robust but non-exceptionless crosslinguistic tendency. At the same time, the magnitude of this effect varies continuously across languages and speakers. The combination of consistent directionality, duration–F1 covariation within /a/, and gradient variability in magnitude is central to the interpretation of intrinsic vowel duration.

The consistent direction of the height effect across languages and speakers is compatible with biomechanical accounts, in which greater articulatory displacement for low vowels yields longer acoustic durations (Lehiste, 1970; Lindblom, 1967). Additional support for this interpretation comes from the relationship between vowel duration and F1, an acoustic correlate of vowel height. Consistent with Coretta (2026) for Northwestern Italian, /a/ showed reasonably strong

[12] Although previous studies have occasionally reported a tendency for back vowels to be slightly longer than front vowels (Fischer-Jørgensen, 1955; Maack, 1954; Malmberg, 1944), this effect was weak and largely non-significant in the present dataset.

positive correlations between duration and F1 across speakers, such that lower vowel realizations tended to be longer (cf., Toivonen et al., 2015). This pattern is consistent with an automatic or biomechanical contribution to intrinsic vowel duration: within the low vowel category, greater jaw and/or tongue displacement may lead to longer acoustic intervals (see also Westbury and Keating, 1980). The weaker or absent relationship for /i/ and /u/ does not necessarily undermine this interpretation, since high vowels may involve a narrower range of relevant jaw-opening or tongue-height variation, limiting the degree to which within-category covariation between duration and F1 could be observed.

These findings suggest that biomechanical factors contribute not only to the categorical difference between high and low vowels, but also to within-category variation in duration, at least for /a/. At the same time, the biomechanical account requires an additional assumption: some aspect of articulation must be held relatively constant across vowel categories. In Lindblom's (1967) model, the relevant constants include the duration of jaw force input and the duration and extent of lip separation. In Lehiste's (1970) account, the greater jaw displacement associated with low vowels yields longer durations but, in our interpretation, only if it is not fully offset by differences in velocity. Thus, even when durational differences arise automatically from articulatory displacement, the effect depends on an assumption that some aspect of phonetic realization remains relatively uniform across vowels.

To see why this assumption is necessary, consider a case in which speakers produced low vowels with faster jaw movement than high vowels, fully compensating for the greater distance the jaw must travel for /a/. Articulatory displacement would still differ between vowel heights, but acoustic duration would not: the observed duration depends not on displacement alone, but on how that displacement is timed. The durational asymmetry observed here therefore depends on the additional assumption that gesture timing, movement velocity, or some comparable set of phonetic targets remains sufficiently uniform across vowel heights, even as the spatial targets for height differ.

This assumption can be stated in terms of target uniformity, a constraint independently proposed to account for structured similarity in phonetic realization across segments that share phonological features; it has previously been applied to stop consonant voicing and sibilant place of articulation (Chodroff and Wilson, 2017, 2022). We extend this idea here to vowels and their observed durations. Target uniformity codifies the assumption of uniform phonetic targets within a framework in which phonological features are associated with phonetic targets, which are subsequently instantiated through the articulatory system. Automatic biomechanical effects may then arise during physical instantiation, but their observable consequences depend in part on how those underlying abstract phonetic targets are specified. In the case of intrinsic vowel duration, this means vowels may have relatively uniform phonetic targets for the shared [+vocalic] feature (e.g., duration of jaw force input, duration of lip separation, jaw movement velocity, etc.), while differing in the phonetic targets for phonological height, such as the tongue and jaw height configurations. The shared phonetic targets for [+vocalic] are therefore instantiated across different height configurations, allowing downstream automatic biomechanical factors to yield different observed acoustic durations. By contrast, high front and high back vowels share the same height value and therefore lack a comparable biomechanical basis for a systematic durational asymmetry; under target uniformity, they are expected to show similar durations.

This formulation separates two issues that are often conflated: while automaticity explains why durational differences arise during physical articulation, target uniformity concerns what is specified prior to physical instantiation in the structure of the abstract phonetic targets. With this interpretation, target uniformity is not an alternative to a biomechanical account, but a way of making explicit one of the assumptions that allows such an account to generate the observed pattern.

Interpreting the findings in terms of target uniformity also clarifies the relationship between automatic explanations and speaker control. Patterns that appear to arise automatically from articulatory mechanisms may nevertheless depend on phonetic targets that are specified similarly across vowel categories, whether that similarity is imposed by the mechanics of articulation or maintained as part of phonetic realization. The present results do not directly determine which of these possibilities is correct. However, once the relevant aspect of uniformity is stated in terms of phonetic targets, it becomes possible to ask how other pressures on phonetic realization might interact with it. In particular, contrast enhancement may modulate a baseline pattern established by biomechanical implementation and target uniformity by recruiting duration as a secondary cue to vowel height. While the present data provide only limited evidence for such systematic modulation, this framework leaves room for limited language- or speaker-specific deviations when contrastive pressures favor greater differentiation.

4.2. Assessing contrast enhancement

Framing automatic accounts in terms of target uniformity also makes it possible to ask whether contrast enhancement contributes to deviations from that baseline. In principle, duration could be recruited or amplified as a secondary cue to vowel height, alongside primary spectral cues associated with tongue height. Enhancement need not necessarily introduce a qualitatively different pattern from the automatic account; it could instead modulate the magnitude of a durational asymmetry already made available by the combination of relatively uniform phonetic targets that underlie acoustic duration and the corresponding automatic biomechanical implementation.

For intrinsic vowel duration, however, the evidence for such enhancement-based modulation is limited. One observation consistent with enhancement is the gradient variability in the size of the height effect across languages (Fig. 5) and individual speakers (Fig. 6). This variability indicates that the effect is not simply categorical or fixed in magnitude, and it leaves open the possibility that language- or speaker-specific pressures can amplify or attenuate the durational difference. At the same time, variability alone does not identify the source of modulation, and it could reflect many factors other than contrast enhancement, including prosodic structure, phonological length, syllable structure, lexical distributions, speaking rate, dialectal variation, or measurement noise.

Other diagnostics provide marginal evidence that enhancement systematically shapes intrinsic vowel duration in the

present data. For example, the relationship between vowel inventory size and the durational difference depended on both the vowel comparison and whether the language had a phonological length contrast (Section 3.4). For /u/–/a/, languages without a phonological length contrast showed a significant positive relationship between vowel inventory size and durational difference. This pattern is consistent with an enhancement-based account, and the same numerical trend was observed for /i/–/a/, though it did not reach significance. One possibility is that languages with a phonological length contrast are less likely to enhance the durational difference because duration is already recruited to maintain the long–short contrast. This would suggest that the available space for durational modulation may be constrained when duration serves an independent phonological function. Overall, however, the weak and inconsistent relationship between inventory size and segmental distinctiveness resembles findings for other segmental phonetic patterns, including sibilant variance (Hauser, 2022), consonant f0 effects (Ting et al., 2025), and vowel intrinsic f0 effects (Whalen and Levitt, 1995). Though some studies have reported modest correlations between vowel inventory size and intrinsic f0 magnitude (Ting et al., 2025; Van Hoof and Verhoeven, 2011; Whalen and Levitt, 1995), these relationships are not robust.

Because speaking rate has been proposed as a possible diagnostic for automaticity and speaker-controlled enhancement in intrinsic vowel duration (Solé and Ohala, 2010), we report a supplemental analysis including rate effects in Appendix A2. We treat this analysis as exploratory, since rate-dependent patterns may arise from multiple articulatory, prosodic, and perceptual sources. The results did not reveal a consistent crosslinguistic pattern that would independently motivate an enhancement account.

Finally, previous studies indicate that listeners may use duration in vowel perception, such that high vowels can sound longer than low vowels of the same physical duration (e.g., Gussenhoven, 2007; Stone, 2015). However, perceptual sensitivity to duration does not by itself imply that speakers actively enhance intrinsic vowel duration; listeners may simply be sensitive to durational differences that arise automatically from biomechanical implementation. Previous studies on just-noticeable differences for vowel and segment duration report thresholds ranging from approximately 5 ms for short segments (Nooteboom and Doodeman, 1980) to 10–25 ms depending on phonetic and sentential context (Fujisaki et al., 1975; Klatt and Cooper, 1975). Sensitivity is further modulated by factors such as prosodic position and listener characteristics (Bergeson et al., 2001; Huggins, 1972). When these thresholds are applied to the present data, only a minority of languages exhibit durational differences that clearly exceed the more conservative proposed thresholds. For /i/–/a/ comparisons, 58% of cases exceed 5 ms, 41% exceed 10 ms, 27% exceed 15 ms, and 17% exceed 20 ms, with comparable proportions for /u/–/a/. These results suggest that, although intrinsic vowel duration is widespread and potentially perceptible in some cases, the present data provide limited evidence that it is systematically enhanced for perception.

Overall, the evidence for a particularly strong role of contrast enhancement in the present data is limited and indirect. The results are more consistent with a framework in which enhancement may contribute to language-specific variation in the magnitude of intrinsic vowel duration, without overriding the baseline pattern consistent with the biomechanical implementation and target uniformity.

4.3. Limitations to the study and future directions

Based on the observed evidence, we argue that the pressure for a uniform specification of the phonetic targets that underlie acoustic duration constitutes a universal constraint across languages and speakers. This gives rise to both an analytic universal and a descriptive statistical universal: across languages, high vowels are overwhelmingly shorter than low vowels. While the data strongly support this tendency, a small number of exceptions do occur. These do not undermine the general pattern, but they do warrant closer examination.

Several limitations of the present study are relevant for interpreting these exceptions. Ideally, stronger controls would be implemented for prosodic structure at multiple levels, as these are well-known for influencing duration. These include prosodic phrasing (e.g., Fougeron and Keating, 1997; Oller, 1973; Wightman et al., 1992), stress (e.g., Fry, 1955), pitch accent (e.g., Gordon, 2008; Sluijter and van Heuven, 1995), syllable structure (Maddieson, 1985), tone (e.g., Burroni and Sukanchanon, 2022; Wu et al., 2023), intricate quantity systems as in Estonian or Finnish (e.g., Lehiste, 1965; Lippus et al., 2013), as well as interactions with speaking rate. Our attempts to control for these effects were modest, such as restricting analyses to the first half of each utterance and including an approximation of speaking rate. Additional factors that should be further considered include dialectal variation or the presence of proper names and loanwords, which could result in potential inaccuracies in the grapheme-to-phoneme (G2P) mappings. Alignment and transcription errors also remain a source of noise, though in the case of Common Voice, transcription inaccuracies are minimized by the platform's validation procedure (Ahn et al., 2023). Further language-specific annotations would be particularly beneficial for refining our estimates of intrinsic vowel duration.

With these considerations in mind, the observed exceptions can be interpreted more cautiously. In some cases, they likely reflect confounds in the data or annotation; in others, they may point to genuine variation in how phonetic targets are implemented across languages. Importantly, the lack of a consistent pattern across the non-conforming cases suggests that these exceptions do not constitute a systematic alternative to the general crosslinguistic tendency.

Among the results for /i/–/a/, /i/ was significantly longer than /a/ in Swedish – long (cf., Elert, 1964; Lindblom 1967), Guaraní, Thai – short, Cantonese – short, Indonesian, and Kabyle, and /i/ was numerically, but not significantly longer than /a/ for Min Nan, Finnish – long, Dhivehi, and Malayalam. The non-conforming results were also confirmed in the manually corrected subset for Swedish – long, Min Nan, and Guaraní, while the other languages were otherwise conforming. For /u/–/a/, /u/ was significantly longer than /a/ in Swedish – long, Mongolian, Welsh, Kabyle, Bassa, Tamil, and Thai – long, and numerically but not significantly so in Cantonese – short.

In some cases, plausible explanations can be identified. For Thai short vowels, the pattern may reflect mismatches

between the G2P system and the vowel quantity system of the recorded speakers. The G2P for Thai assumes a length contrast across vowel qualities (e.g., /iː i uː u aː a/) that is indeed consistent with standard descriptions of the language (Maspong and Burroni, 2024; Tingsabadh and Abramson, 1993). However, previous research has suggested that this contrast may be weakening in the high and mid vowels, with long vowels in particular undergoing shortening (Abramson, 2001; Brown, 1979). Under this interpretation, the observed pattern would align with expectations for long vowels (with the intermediate duration of /i u/ shorter than long /aː/) but appear as a counterexample for short vowels (with the intermediate duration of /i u/ longer than short /a/). This is precisely the pattern observed in the Thai vowel contrasts.

For Min Nan, the unexpected pattern is likely attributable to a confound with syllable structure, where vowel duration is often longer in open than closed syllables (Maddieson, 1985). High vowels were almost twice as likely to occur in open syllables, while /a/ was almost twice as likely to occur in closed syllables. When syllable structure is included as a predictor in the Min Nan linear regression, the vowel height effect goes in the expected direction, though it remains small and non-significant for both the /i/–/a/ and /u/–/a/ rate-normalized analyses. Min Nan has a predictable syllable structure where the onset or coda can only be occupied by a single consonant as opposed to a cluster, which allowed us to estimate whether the syllable was open or closed for this language. However, controlling for syllable structure openness in languages with more complex phonotactic structure is something we leave to future research.

For the remaining exceptions, no single explanatory factor was clearly apparent. It may be worth noting that Swedish long vowels only occur before short consonants, whereas short vowels only occur before long consonants (Elert, 1964). In addition, Swedish vowel duration is influenced by factors including stress and pitch accent (e.g., Fant et al., 1991; Fant and Kruckenberg, 1994), which may have affected results. In some cases (Min Nan, Finnish – long, and Malayalam), small sample sizes may have contributed to unstable estimates. Each of these comparisons had ten or fewer speakers. We also explored potential influences of word length, syllable structure where possible, and observed lexical frequency on raw vowel duration. Using the number of orthographic characters as an approximation for word length, we found a very small numeric bias for /a/ to occur in words with more orthographic segments compared to /i/ for each of the additional significant languages (Swedish – long, Guaraní, Indonesian, and Kabyle). Previous research suggests that segment duration can be reduced in prosodic units with more segments (Beckman et al., 1990; Gibson and Summers, 2018; Lahiri and Fikkert, 1999; White and Turk, 2010), such that for these languages, /a/ might be reduced disproportionately to /i/. This could potentially be ameliorated with an improved speaking rate estimate or controlling for number of syllables and syllable type. Beyond the reported result for Min Nan and the confounding syllable structure, no clear patterns emerged in the effects of syllable structure or observed lexical frequency for the other exceptions.

Among the exceptions in the /u/–/a/ data, a minor bias was observed in word length such that /a/ tended to occur in longer words than /u/ for Swedish – long, Kabyle, Bassa, and Min Nan, but not for Mongolian, Welsh, or Tamil. Otherwise, syllable structure type did not correspond to any likely durational influence. For lexical frequency, /u/ tended to become shorter than /a/ among high frequency words in Mongolian, and the opposite trend was observed in Tamil; no other major bias was observed.

Future research should more closely examine the limitations of the present dataset, particularly with respect to languages that did not conform to the general pattern. Improved prosodic and dialectal control, along with increased geographic and phylogenetic coverage would also be beneficial. Importantly, the perceptual role of duration in vowel height contrasts could be further clarified: even if speakers do not actively enhance duration as a cue, listeners may nevertheless make use of durational differences to facilitate perceptual processing of these vowel qualities. Experimental work examining whether and under what conditions listeners exploit these differences would help determine whether intrinsic vowel duration functions as a perceptually relevant cue or remains largely incidental.

5. Conclusion

Across more than 60 languages, high vowels were reliably shorter than low vowels, while high front and back vowels showed no comparable asymmetry. Overall, these results support intrinsic vowel duration as a statistical phonetic universal. The observed pattern is most consistent with a biomechanical source for intrinsic vowel duration, but one which can only arise under the assumption that the phonetic targets underlying acoustic duration are near-uniform in their abstract specification. We make this assumption explicit through the notion of target uniformity: phonetic targets that are near-uniform across the natural class of vowels, when instantiated with different vowel heights can give rise to durational differences in the acoustics between high and low vowels. This can account for the overwhelming directional bias of the effect, as well as the positive correlation between vowel duration and F1 within /a/, which points to a direct relationship between jaw or tongue height and vowel duration.

Nevertheless, the considerable variation in effect magnitude across languages and individual speakers suggests that other factors, such as contrast enhancement, may also be at play. Enhancement could likewise explain why the durational difference between high and low vowels grew with vowel inventory size, but only in languages lacking a phonological length contrast. On the whole, however, the evidence for enhancement was limited.

Future work should aim to provide further refinements and additions to the corpus data. It will also be helpful to evaluate the perceptual relevance of these durational differences and to determine the conditions under which enhancement may play a more substantial role, perhaps via experimental evidence. Extending this framework to other proposed mechanistic universals may also provide a more unified account of how phonetic patterns emerge from the interaction of constraints on speech production and perception.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT for editing the manuscript and Claude Code for generating code to create figures. The authors reviewed and edited the output and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Eleanor Chodroff: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Miao Zhang:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

The derivative data and code for the measurement and analyses in this study are available at https://osf.io/wafut/. Phone- and word-level forced alignments are available at https://huggingface.co/datasets/pacscilab/VoxCommunis. Corresponding audio files can be downloaded from Mozilla Common Voice at https://datacollective.mozillafoundation.org/.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank three anonymous reviewers, as well as audiences at LabPhon 19, McGill University, University of British Columbia, New York University, and LMU Munich for helpful feedback. This research was funded by SNF Grant PR00P1_208460.

Appendix A

A.1. VoxCommunis processing pipeline

To obtain the forced alignments for the audio files, a custom pronunciation dictionary and acoustic model were developed for each language. The pronunciation dictionaries were obtained by converting the orthographically transcribed sentences to a canonical phonetic transcription using grapheme-to-phoneme (G2P) conversion or a pre-existing pronunciation lexicon. The pronunciation dictionaries were created using the rule-based G2P systems of Epitran (Mortensen et al., 2018), XPF (Cohen Priva et al., 2021), CommonVoice Utils (Tyers, 2021), or a custom rule-based G2P system based on orthographic and phonetic descriptions of the language. We additionally employed Charsiu, a neural network-based G2P system (Zhu et al., 2022) or developed custom G2P systems using pre-existing pronunciation lexicons and the Montreal Forced Aligner G2P training algorithm (McAuliffe et al., 2017). For languages requiring tokenization—that is, the separation of orthographic sequences into smaller white-space delimited chunks—we used various language-specific systems. These were pycantonese for Cantonese (Lee et al., 2022), PKUSEG for Mandarin (Luo et al., 2019), fugashi for Japanese (McCann, 2020), and PyThaiNLP for Thai (Phatthiyaphaibun et al., 2023). To obtain a more accurate post-lexical pronunciation of Korean words, the Korean texts were first preprocessed by using g2pK (Park, 2019). Our assumptions regarding the language-specific phoneme inventory were based on the original assumptions used in these G2P systems.

Following the lexicon development, we employed the Montreal Forced Aligner to train an acoustic model using the validated MP3 files, their corresponding transcripts, and the pronunciation dictionary that provided the mapping between the orthographic form of the language and the canonical phone sequence (McAuliffe et al., 2017). The acoustic model used context-dependent phone models (triphones) in a GMM–HMM architecture with speaker adaptation. Phone- and word-level alignments with the validated audio files were then obtained using the acoustic model and pronunciation dictionary. For both acoustic model training and alignment, the default settings were employed. A full list of each language, language family, and corresponding processing information is shown in Table A1.

One additional processing step within the VoxCommunis development was the consideration of the speaker ID (Zhang et al., 2025b). Speaker IDs were approximated from the client ID provided in the Common Voice dataset following refinement. The client ID corresponds to “an anonymized ID for the speaker” (Ardila et al., 2020), but more likely reflects a unique recording device that can theoretically have more than one speaker. For instance, a few cases might have involved a family computer that was used by multiple household members with distinct voices (e.g., perceived male and female voices) that had the same client ID.

Based on this observation, we applied a speaker verification system to the corpus with the final recording associated with each client ID as an enrollment file and the remaining recordings as test files. We employed the ResNet-293 system trained on the VoxBlink2 Corpus and used the cosine similarity between the speaker embeddings of the enrollment and test file to obtain a speaker similarity score (Lin et al., 2024). If the similarity score was below 0.4, it was assumed to be either a different speaker or a noisy recording and discarded from further analysis. The 0.4 threshold was chosen based on evidence from a perceptual audit of similarity scores for the Common Voice Corpus and the optimal threshold for the speaker verification system on the English VoxCeleb1-H test set (Nagrani et al., 2020). For additional details on the speaker verification and on auditing process results, see Zhang et al. (2025b).

Table A1
A list of the languages and contrasts included in the analyses, organized by language family (Glottolog; Hammarström et al., 2025). The Common Voice code (CV) and ISO-639–3 code are provided for each language. TG refers to the Common Voice dataset version used for the TextGrids, AM to the Common Voice dataset version used for training the acoustic model, and G2P to the grapheme-to-phoneme system. The G2P systems were Epitran (epi; Mortensen et al., 2018), the XPF Corpus (xpf; Cohen Priva et al., 2021), CommonVoice Utils (cvu; Tyers, 2021), Charsiu (chr; Zhu et al., 2022), the Montreal Forced Aligner system (mfa; McAuliffe et al., 2017), or a custom G2P either obtained via a rule-based script or an existing pronunciation lexicon (vxc). More details regarding the custom G2P systems can be found in the Hugging Face repository at https://huggingface.co/datasets/pacscilab/VoxCommunis. The final columns indicate whether the comparison was included in the /i/–/a/, /u/–/a/, or /i/–/u/ analyses. The number of unique languages per language family is included in parentheses for each language family with more than one language.

Family	Language	CV	ISO-639-3	TG	AM	G2P	/i/–/a/	/u/–/a/	/i/–/u/
Abkhaz-Adyge	Abkhaz	ab	abk	17	16	xpf	✓		
Afro-Asiatic	Amharic	am	amh	20	20	epi	✓		
(3)	Kabyle	kab	kab	20	20	vxc	✓	✓	✓
	Maltese	mt	mlt	17	17	epi	✓	✓	✓
Atlantic-Congo	Taita	dav	dav	20	20	vxc	✓	✓	✓
(6)	Taita – long	dav_long	dav	20	20	vxc	✓	✓	✓
	Luganda	lg	lug	20	20	vxc	✓	✓	✓
	Luganda – long	lg_long	lug	20	20	vxc	✓	✓	✓
	Kinyarwanda	rw	kin	17	17	epi	✓	✓	✓
	Swahili	sw	swh	17	17	epi	✓	✓	✓
	Tswana	tn	tsn	20	20	cvu	✓	✓	✓
	Yoruba	yo	yor	17	16	epi	✓	✓	✓
Austronesian	Indonesian	id	ind	17	17	epi	✓	✓	✓
Basque	Basque	eu	eus	17	17	xpf	✓	✓	✓
Dravidian	Malayalam	ml	mal	16	16	epi	✓	✓	✓
(2)	Tamil	ta	tam	17	17	epi	✓	✓	✓
	Tamil – long	ta_long	tam	17	17	epi	✓	✓	✓
Indo-European	Assamese	as	asm	20	20	cvu	✓	✓	✓
(31)	Belarusian	be	bel	17	17	xpf	✓	✓	✓
	Bulgarian	bg	bul	16	16	xpf	✓	✓	✓
	Bengali	bn	ben	17	17	epi	✓	✓	✓
	Catalan	ca	cat	17	16	vxc	✓	✓	✓
	Central Kurdish	ckb	ckb	17	17	epi		✓	
	Central Kurdish – long	ckb_long	ckb	17	17	epi			✓
	Czech	cs	ces	17	17	xpf	✓	✓	✓
	Czech – long	cs_long	ces	17	17	xpf	✓	✓	✓
	Welsh	cy	cym	20	20	cvu	✓	✓	✓
	Dhivehi	dv	div	17	16	xpf	✓	✓	✓
	Dhivehi – long	dv_long	div	17	16	xpf	✓		
	Greek	el	ell	16	16	xpf	✓	✓	✓
	Galician	gl	glg	20	20	vxc	✓	✓	✓
	Hindi	hi	hin	17	17	epi			✓
	Hindi – long	hi_long	hin	17	17	epi	✓	✓	✓
	Upper Sorbian	hsb	hsb	17	17	xpf	✓	✓	✓
	Armenian	hy-AM	hye	17	17	xpf	✓	✓	✓
	Italian	it	ita	17	17	vxc	✓	✓	✓
	Kurmanji Kurdish	kmr	kmr	17	17	epi			✓
	Kurmanji Kurdish – long	kmr_long	kmr	17	17	epi	✓	✓	✓
	Ligurian	lij	lij	17	17	epi	✓	✓	
	Lithuanian	lt	lit	17	17	chr	✓	✓	✓
	Lithuanian – long	lt_long	lit	17	17	chr	✓	✓	✓
	Macedonian	mk	mkd	19	19	xpf	✓	✓	✓
	Marathi	mr	mar	17	17	epi			✓
	Marathi – long	mr_long	mar	17	17	epi	✓	✓	✓
	Dutch	nl	nld	17	16	epi	✓		
	Dutch – long	nl_long	nld	17	16	epi	✓	✓	✓
	Odia	or	ory	19	19	epi	✓	✓	✓
	Polish	pl	pol	17	17	epi	✓	✓	✓
	Portuguese	pt	por	17	17	epi	✓	✓	✓

Table A1 (*continued*)

Family	Language	CV	ISO-639-3	TG	AM	G2P	/i/–/a/	/u/–/a/	/i/–/u/
	Romanian	ro	ron	17	17	xpf	✓	✓	✓
	Russian	ru	rus	17	17	epi	✓	✓	✓
	Slovak	sk	slk	17	17	chr	✓	✓	✓
	Slovak – long	sk_long	slk	17	17	chr	✓	✓	✓
	Slovenian	sl	slv	17	17	chr	✓	✓	✓
	Serbian	sr	hbs	17	17	chr	✓	✓	✓
	Swedish	sv-SE	swe	17	17	epi	✓		
	Swedish – long	sv-SE_long	swe	17	17	epi	✓	✓	✓
	Ukrainian	uk	ukr	17	17	xpf	✓	✓	✓
Japonic	Japanese	ja	jpn	17	17	epi	✓	✓	✓
	Japanese – long	ja_long	jpn	17	17	epi	✓	✓	✓
Kartvelian	Georgian	ka	kat	17	16	xpf	✓	✓	✓
Koreanic	Korean	ko	kor	20	20	mfa	✓		
Kru	Basaa	bas	bsq	20	20	cvu	✓	✓	✓
Mongolic-Khitan	Mongolian	mn	khk	19	19	epi	✓	✓	✓
	Mongolian – long	mn_long	khk	19	19	epi	✓	✓	✓
Nilotic	Dholuo	luo	luo	20	20	cvu	✓	✓	✓
Sino-Tibetan	Min Nan	nan-tw	nan	21	21	vxc	✓	✓	✓
(3)	Cantonese	yue	yue	21	21	epi	✓	✓	✓
	Cantonese – long	yue_long	yue	21	21	epi	✓	✓	✓
	Mandarin	zh-CN	cmn	17	17	vxc	✓	✓	✓
Tai-Kadai	Thai	th	tha	21	21	epi	✓	✓	✓
	Thai – long	th_long	tha	21	21	epi	✓	✓	✓
Tupian	Guarani	gn	gug	17	17	xpf	✓	✓	✓
Turkic	Bashkir	ba	bak	17	16	xpf	✓	✓	✓
(9)	Chuvash	cv	chv	17	17	xpf	✓	✓	✓
	Kyrgyz	ky	kir	17	17	epi	✓	✓	✓
	Sakha	sah	sah	19	19	xpf	✓	✓	✓
	Sakha – long	sah_long	sah	19	19	xpf		✓	
	Turkmen	tk	tuk	17	17	epi	✓	✓	✓
	Turkish	tr	tur	17	17	xpf	✓	✓	✓
	Tatar	tt	tat	20	20	xpf	✓	✓	✓
	Uyghur	ug	uig	17	17	xpf	✓	✓	✓
	Uzbek	uz	uzn	17	17	epi	✓	✓	✓
Uralic	Estonian	et	ekk	20	20	cvu			✓
(4)	Finnish	fi	fin	20	20	cvu	✓	✓	✓
	Finnish – long	fi_long	fin	20	20	cvu	✓		
	Hungarian	hu	hun	17	17	xpf	✓	✓	✓
	Hungarian – long	hu_long	hun	17	17	xpf	✓	✓	✓
	Erzya	myv	myv	20	20	xpf	✓	✓	

A.2. Speaking rate analysis

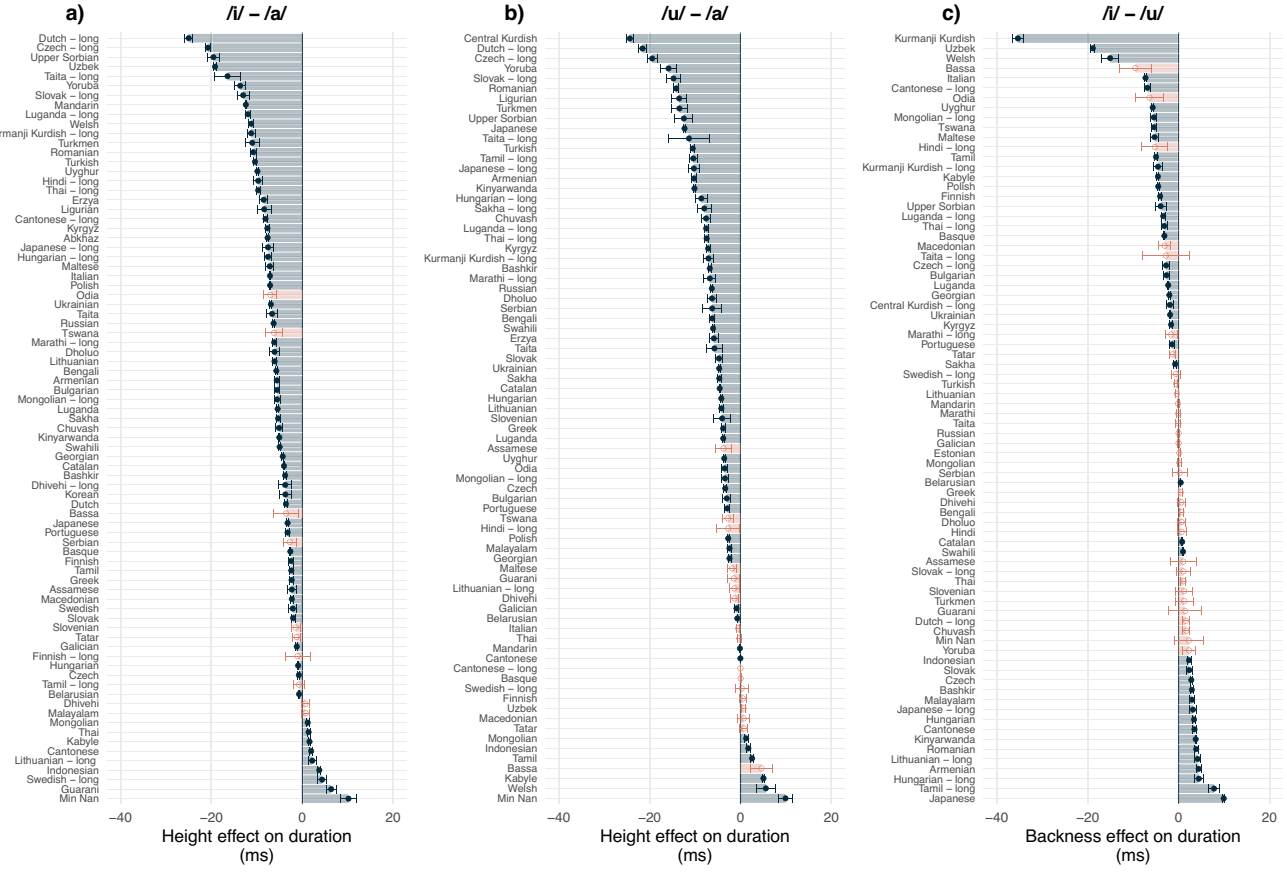
A.2.1. Methods: Language-specific raw vowel duration

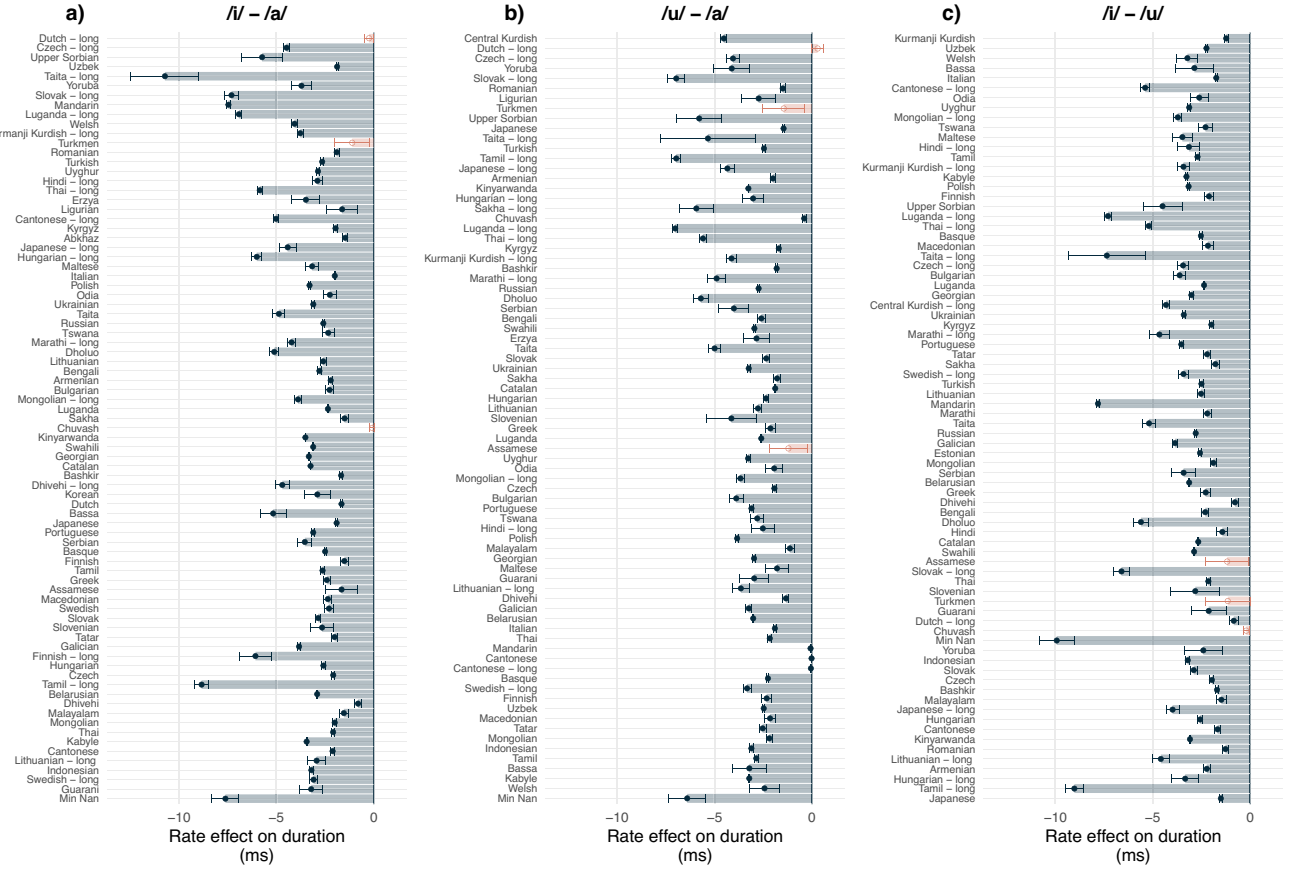
To gain insight into the relationship between speaking rate, vowel quality, and vowel duration, separate linear mixed-effects regression models were also fit to the raw vowel duration of each comparison in the /i/–/a/, /u/–/a/, and /i/–/u/ subsets. These models mirrored the rate-normalized model in all respects, with the critical additional predictor of speaking rate, calculated as number of phones per second, and its interaction with vowel quality. Speaking rate was centered on the language-specific mean prior to model fitting. Raw vowel duration (ms) was thus predicted from vowel quality, speaking rate (centered), the interaction between quality and rate, all possible interactions between preceding voicing, preceding manner, following voicing, and following manner, as well as the utterance proportion of the target vowel. A random intercept and slope for vowel quality was fit per speaker. All other model aspects were kept the same between the rate-normalized and raw duration model fits.[13] *P*-values were corrected using the Benjamini–Hochberg method, and alpha was set to 0.05 for the assessment of significance (Benjamini and Hochberg, 1995).

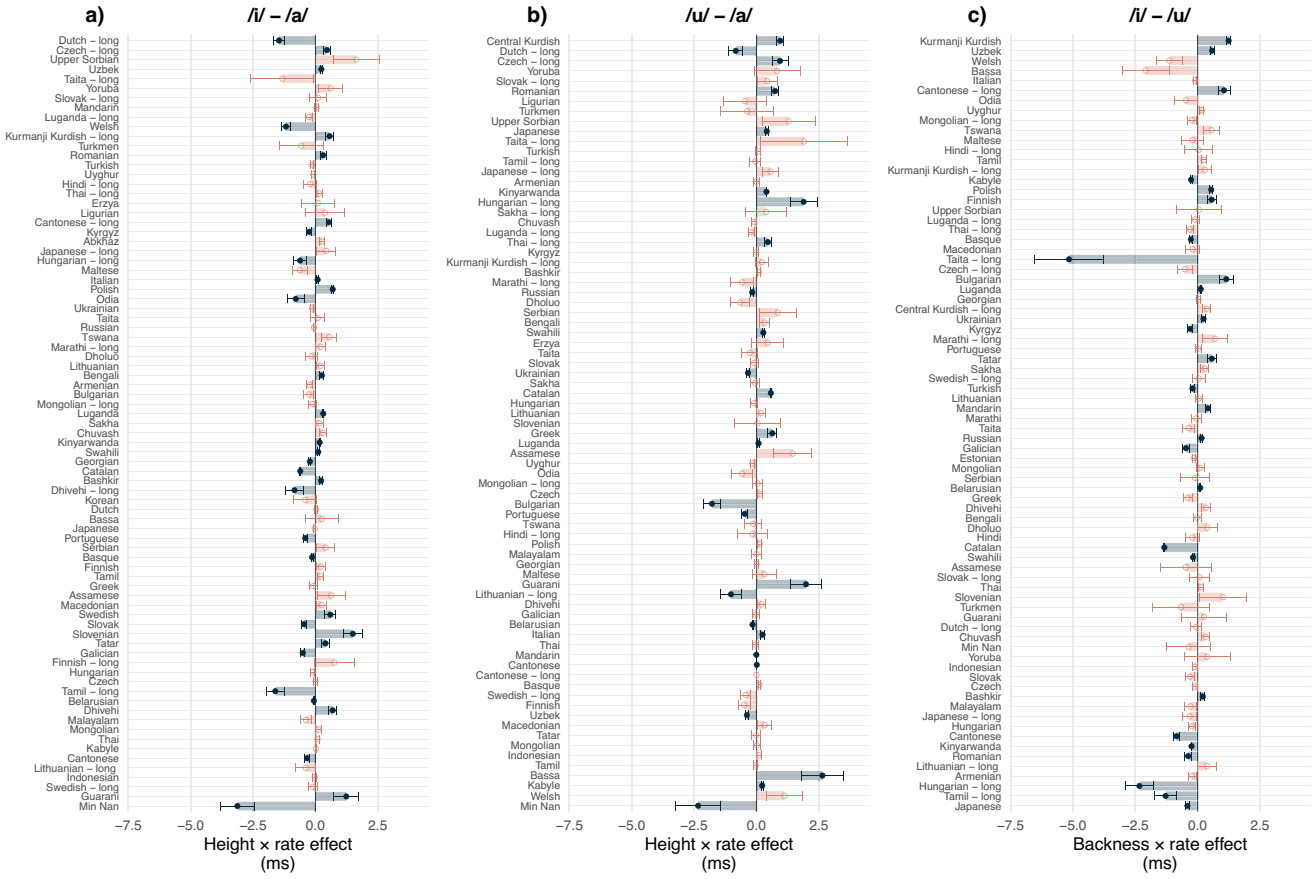
A.2.2. Results: Language-specific raw vowel duration

For the raw vowel duration models, the estimated effects of vowel height, speaking rate, and their interaction for each comparison is shown in Fig. A2. The left panel shows the effect of vowel quality (which, in the cases of /i/–/a/ and /u/–/a/, corresponds to vowel height), the middle panel the effect of speak-

[13] For the raw vowel duration, Amharic /i/–/a/ had to be removed due to lack of convergence, even without the random by-speaker slope for vowel quality. The random by-speaker slope for vowel quality had to be removed from the following languages: for /i/–/a/, Assamese, Taita – long, Guaraní, Upper Sorbian, Korean, Ligurian, Macedonian, Erzya, Turkmen, and Mandarin Chinese; for /u/–/a/, Taita – long, Greek, Guaraní, Ligurian, Malayalam, Maltese, Eryza, Min Nan, Odia, Sakha – long, Slovenian, Serbian, and Turkmen; and for /i/–/u/, Bengali, Taita, Greek, Finnish, Upper Sorbian, Ligurian, Malayalam, Maltese, Min Nan, Sakha, Serbian, Turkmen, Tswana, and Yoruba.


a)
/i/ – /a/
Height effect on duration (ms)
b)
/u/ – /a/
Height effect on duration (ms)
c)
/i/ – /u/
Backness effect on duration (ms)
−40
−20
0
20


a)
/i/ – /a/
Rate effect on duration (ms)
b)
/u/ – /a/
Rate effect on duration (ms)
c)
/i/ – /u/
Rate effect on duration (ms)
−10
−5
0



ing rate, and the right panel, the interaction between vowel quality and speaking rate, each on the estimated raw vowel duration in milliseconds.

/i/–/a/. The effect of vowel height on duration was in the expected direction with /i/ shorter than /a/ (a negative estimate) for 86% of the comparisons (70/81; 87% for rate-normalized) and significant for 77% of the comparisons (62/81; 82% for rate-normalized). The estimate of the difference ranged from –25.0 ms to 10.2 ms with a median difference of –5.3 ms (interquartile range: –7.7 ms to –2.1 ms). For all comparisons, vowel duration was, as expected, shorter at faster speaking rates; this effect was significant for 96% of comparisons (78/81). Finally, the interaction between vowel height and speaking rate on vowel duration did not reach significance in either direction for the majority of comparisons (59%; 48/81). The interaction revealed a significantly bigger difference at faster rates for 20% of comparisons (16/81) and a significantly smaller difference at faster rates for another 21% of languages (17/81).

/u/–/a/. The effect of vowel height on duration was in the expected direction with /u/ shorter than /a/ for 83% of comparisons (64/77; 90% for rate-normalized) and significantly so for 70% of comparisons (54/77; 77% for rate-normalized). The estimate of the difference ranged from –24.4 ms to 9.9 ms with a median difference of –4.0 ms (interquartile range: –7.4 ms to –0.9 ms). As before, vowel duration was shorter at faster speaking rates for 99% of comparisons (76/77) and significantly so for 96% (74/77). In addition, the interaction between vowel height and speaking rate on vowel duration did not reach significance in either direction for the majority of comparisons (66%; 51/77). The interaction was, however, significant with a bigger difference at faster rates, for 21% of comparisons (16/77), and significant with a smaller difference at faster rates for 13% of languages (10/77).

/i/–/u/. /i/ was shorter than /u/ for 54% of comparisons (42/78; 53% for rate-normalized) and significantly so for 35% of comparisons (27/78; 37% for rate-normalized). In addition, /u/ was shorter than /i/ for 46% of comparisons (36/78; 47%

Fig. A1. The effects of vowel quality (height or backness; top panel), rate (middle panel), and their interaction (bottom panel) across languages for a) /i/–/a/, b), /u/–/a/ and c) /i/–/u/ in milliseconds (ms). A negative effect of vowel quality indicates that the first listed vowel was shorter than the second listed vowel. A negative effect of speaking rate indicates that vowel duration was shorter with faster speaking rates. A negative interaction between vowel height and speaking rate indicates that the difference between the listed vowels became smaller at faster speaking rates. A filled dark blue circle indicates a significant estimate, while an open light orange circle indicates that the estimate did not significantly differ from zero. Error bars reflect ±1 standard error of the estimate. Within each figure, comparisons are ordered by the respective estimated effect of vowel quality (height or backness).

for rate-normalized) and significantly so for 23% of comparisons (18/78; 28% for rate-normalized). Overall, the effect of vowel height on duration did not differ significantly from zero for 42% of comparisons (33/78; 35% for rate-normalized). The estimate of the difference ranged from –35.4 ms to 9.9 ms with a median difference of –0.07 ms (interquartile range: –3.2 ms to 1.4 ms). The effect of speaking rate on vowel duration was in the expected direction, such that faster rates corresponded to shorter vowel durations, for all comparisons (78/78) and significantly so for 96% of comparisons (75/78). Finally, the interaction between vowel height and speaking rate on vowel duration did not reach significance in either direction for the majority of comparisons (65%; 51/78). The interaction was, however, significant with a larger durational difference at faster rates for 17% of comparisons (13/78) and significant with a smaller difference at faster rates for 18% of comparisons (14/78).

Table A2
Types of interactions and possible interpretations, along with the observed frequency in the present data for /i/–/a/, /u/–/a/, and /i/–/u/. For /i/–/a/ and /u/–/a/, the quality effect reflects vowel height; for /i/–/u/, it reflects vowel backness. Because /i/ and /u/ share the same height value, the automaticity and enhancement interpretations that presuppose a height-based contrast do not apply to the /i/–/u/ comparison; the /i/–/u/ frequency is reported for comparison only.

Interaction	Possible interpretations	Observed frequency		
		/i/–/a/	/u/–/a/	/i/–/u/
Null (no change in durational difference with rate)	Active maintenance of contrast despite undershoot	59%	66%	65%
Smaller durational difference at faster rates	Perceptual rescaling or articulatory undershoot	21%	21%	17%
Larger durational difference at faster rates	Trading relation in production: duration compensates for degraded spectral cues	20%	13%	18%

A.2.3. Discussion: Language-specific raw vowel duration

Overall, the data reveal consistent effects of vowel height and speaking rate on raw vowel duration, and no consistent crosslinguistic pattern in the interaction. Interactions between vowel height and speaking rate have previously been used to distinguish automatic and controlled accounts of intrinsic vowel duration (Solé and Ohala, 2010). The absence of a height-by-rate interaction has been interpreted as evidence for automaticity, while a negative interaction, with faster rates associated with smaller durational differences, has been interpreted as evidence for controlled enhancement. We argue, however, that this diagnostic may not be reliable, because all three possible interaction patterns may be compatible with a controlled account (Table A2).

Specifically, a null interaction could reflect controlled maintenance of the contrast, such that the durational difference is preserved across speaking rates despite the possibility of articulatory undershoot at faster rates (Flege, 1988; Gay, 1968; Recasens, 2015; Tsao et al., 2006). An interaction with smaller durational differences at faster rates could reflect perceptual rescaling in which durational differences are enhanced at slower rates, or simply articulatory undershoot in which durational differences reduce to greater degrees at faster rates. (The articulatory undershoot explanation could plausibly be consistent with automaticity, but as argued in the main text, would nevertheless only be possible with underlyingly uniform phonetic targets across high and low vowels.) An interaction with greater durational differences at faster rates could reflect a trading relation in production, in which duration becomes more informative if spectral cues are reduced via undershoot. Because each outcome allows for a plausible controlled-account explanation, we argue that the sign of the height-by-rate interaction does not on its own discriminate between automatic and controlled sources of intrinsic vowel duration.

References

Abramson, A. (1974). Experimental phonetics in phonology: Vowel duration in Thai. *Pasaa, 4*(1), 71–90.

Abramson, A. S. (2001). The stability of distinctive vowel length in Thai. *Essays in Tai linguistics*, 13–26.

Ahn, E. P., & Chodroff, E. (2022). VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5286–5294). European Language Resources Association.

Ahn, E. P., Levow, G.-A., Wright, R. A., & Chodroff, E. (2023). An outlier analysis of vowel formants from a corpus phonetics pipeline. *Proceedings of Interspeech, 2023*, 2573–2577. https://doi.org/10.21437/Interspeech.2023-1052.

Äimä, F. G. (1918). *Phonetik und Lautlehre des Inarilappischen* (Vol. 1). Société finno-ougrienne.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4218–4222).

Beckman, M. E., Swora, M. G., Rauschenberg, J., & de Jong, K. (1990). Stress shift, stress clash, and polysyllabic shortening in a prosodically annotated discourse. *First International Conference on Spoken Language Processing (ICSLP 1990)*, 5–8. https://doi.org/10.21437/ICSLP.1990-2.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: S B (Methodological), 57*(1), 289–300.

Bergeson, T. R., Schneider, B. A., & Hamstra, S. J. (2001). Duration discrimination in younger and older adults. *Canadian Acoustics, 29*(4), 3–9.

Black, J. W. (1949). Natural frequency, duration, and intensity of vowels in reading. *Journal of Speech and Hearing Disorders, 14*(3), 216–221. https://doi.org/10.1044/jshd.1403.216.

Brown, J. M. (1979). Vowel length in Thai. *Studies in Tai and Mon-Khmer Phonetics and Phonology In Honour of Eugénie JA Henderson*, 10–25.

Bundgaard, M. (1980). An acoustic investigation of intrinsic vowel duration in Danish. *Annual Report of the Institute of Phonetics University of Copenhagen, 14*, 99–119.

Burroni, F., & Sukanchanon, T. (2022). *Tonal effects on vowel duration in Bangkok Thai*. Canberra, Australia: Science and Technology.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001.

Chodroff, E., & Wilson, C. (2022). Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English. *Language, 98*(2), 250–289. https://doi.org/10.1353/lan.2022.0007.

Cohen Priva, U., Strand, E., Yang, S., Mizgerd, W., Creighton, A., Bai, J., Mathew, R., Shao, A., Schuster, J., & Wiepert, D. (2021). The Cross-linguistic Phonological Frequencies (XPF) Corpus manual. Accessible online, https://cohenpr-xpf.github.io/XPF/manual/xpf_manual.pdf.

Coretta, S. (2026). Is "intrinsic vowel duration" bio-mechanical or more? Preliminary results from Northwestern Italian. *Language and Speech* 00238309251399657.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America, 83*(4), 1553–1573.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics, 49*, 77–95.

Elert, C.-C. (1964). *Phonologic studies of quantity in Swedish: Based on material from stockholm speakers* (PhD Thesis). Stockholm University.

Fant, G., & Kruckenberg, A. (1994). Notes on stress and word accent in Swedish. In *Proceedings of the international symposium on prosody* (pp. 2–3).

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics, 19*(3–4), 351–365.

Fischer-Jørgensen, E. (1955). Om vokallængde i dansk rigsmål. *Nordisk Tidsskrift for Tale Og Stemme, 15*, 33–56.

Flege, J. E. (1988). Effects of speaking rate on tongue position and velocity of movement in vowel production. *The Journal of the Acoustical Society of America, 84*(3), 901–916.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America, 101*(6), 3728–3740.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*(4), 765–768.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. *Auditory Analysis and Perception of Speech*, 197–219.

Gay, T. (1968). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America, 44*(6), 1570–1573. https://doi.org/10.1121/1.381717.

Gibson, T. A., & Summers, C. (2018). Polysyllabic shortening in speakers exposed to two languages. *Bilingualism: Language and Cognition, 21*(3), 471–478.

Gordon, M. (2008). Pitch accent timing and scaling in Chickasaw. *Journal of Phonetics, 36*(3), 521–535.

Gussenhoven, C. (2007). A vowel height split explained: Compensatory listening and speaker control. In J. Cole & J. I. Hualde (Eds.). *Laboratory phonology* (Vol. 9, pp. 145–172). Mouton de Gruyter.

Haggard, M., Summerfield, Q., & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F0 cues in the voiced–voiceless distinction. *Journal of Phonetics, 9*(1), 49–62.

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2025). Glottolog 5.2. doi: 10.5281/zenodo.15525265.

Hauser, I. (2022). Speech sounds in larger inventories are not (necessarily) less variable. *The Journal of the Acoustical Society of America, 152*(5), 2664–2674.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111.

Hoemeke, K. A., & Diehl, R. L. (1994). Perception of vowel height: The role of F_1–F_0 distance. *The Journal of the Acoustical Society of America, 96*(2), 661–674.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America, 25*(1), 105–113. https://doi.org/10.1121/1.1906982.

Huggins, A. W. F. (1972). Just noticeable differences for segment duration in natural speech. *The Journal of the Acoustical Society of America, 51*(4B), 1270–1278.

Johnson, K., & Martin, J. (2001). Acoustic Vowel Reduction in Creek: Effects of Distinctive Length and Position in the Word. *Phonetica, 58*(1–2), 81–102. https://doi.org/10.1159/000028489.

Keating, P. (1985). Universal phonetics and the organization of grammars. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 115–132). Academic Press.

Keating, P. A. (2003). Phonetic and other influences on voicing contrasts. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 20–23).

Kingston, J. (2007). Segmental influences on F0: Automatic or controlled? In T. Riad & C. Gussenhoven (Eds.), *Tones and tunes 2* (pp. 171–210). Mouton de Gruyter. https://doi.org/10.1515/9783110207576.2.171.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language, 70*(3), 419. https://doi.org/10.2307/416481.

Kirby, J., & Harrington, J. (2026). Acoustic-perceptual factors in the actuation of sound change. *The Wiley Blackwell Companion to Diachronic and Historical Linguistics*, 1–37.

Kirby, J. P., & Ladd, D. R. (2016). Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. *The Journal of the Acoustical Society of America, 140*(4), 2400–2411.

Klatt, D. H., & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In *Structure and process in speech perception: Proceedings of the symposium on dynamic aspects of speech perception* (pp. 69–89). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lahiri, A., & Fikkert, P. (1999). Trisyllabic shortening in English: Past and present. *English Language & Linguistics, 3*(2), 229–267.

Lapierre, M., De Falco, E., Tosolini, A., & Steffman, J. (2026). Disentangling prominence strengthening: Evidence for independent stress and length effects in Panãra vowel acoustics. *Journal of Phonetics, 117* 101502.

Lee, J., Chen, L., Lam, C., Lau, C., & Tsui, T.-H. (2022). PyCantonese: Cantonese linguistics and NLP in Python. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6607–6611). Marseille, France: European Language Resources Association.

Lehiste, I. (1965). The function of quantity in Finnish and Estonian. *Language, 41*(3), 447–456.

Lehiste, I. (1970). *Suprasegmentals*. The MIT Press.

Lenth, R. (2025). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.11.2-80003, https://rvlenth.github.io/emmeans/.

Lin, Y., Cheng, M., Zhang, F., Gao, Y., Zhang, S., & Li, M. (2024). VoxBlink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark. *Proceedings of Interspeech, 2024*, 4263–4267.

Lindblom, B. (1967). Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress Status Report, 4*, 1–29.

Lippus, P., Asu, E. L., Teras, P., & Tuisk, T. (2013). Quantity-related variation of duration, pitch and vowel quality in spontaneous Estonian. *Journal of Phonetics, 41*(1), 17–28.

Long, A. Y., Solon, M., & Bongiovanni, S. (2018). Context of learning and second language development of Spanish vowels. *Studies in Hispanic and Lusophone Linguistics, 11*(1), 59–87.

Luo, R., Xu, J., Zhang, Y., Zhang, Z., Ren, X., & Sun, X. (2019). PKUSEG: A toolkit for multi-domain Chinese word segmentation. arXiv Preprint arXiv:1906.11455.

Maack, A. (1949). Die spezifische Lautdauer deutscher Sonanten. *Sprachtypologie und Universalienforschung, 3*(3), 190.

Maack, A. (1954). Die Beeinflussung der Sonantendauer durch die Nachbarkonsonanten. *STUF-Language Typology and Universals, 7*(JG), 104–128.

Maddieson, I. (1985). Phonetic cues to syllabification. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 85–101). Academic Press.

Maddieson, I. (1995). Gestural economy. In *Proceedings of the 13th* (pp. 574–577).

Maddieson, I. (1996). Phonetic universals. *UCLA Working Papers in Phonetics*, 160–178.

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*(2), 464–488.

Malmberg, B. (1944). *Die Quantität als phonetisch-phonologischer Begriff: Eine allgemeinsprachliche Studie*. Lund: C. W. K. Gleerup.

Marín Gálvez, R. (1995). La duración vocálica en español. *ELUA. Estudios de Lingüística, N., 10*, 213–226.

Maspong, S., & Burroni, F. (2024). Are long and short vowels articulatorily different? Spatial and durational effects of vowel length in Thai. In *Proceedings of ISSP 2024–13th international seminar on speech production* (pp. 165–168). https://doi.org/10.21437/issp.2024-42.

Massaro, D. W., & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *The Journal of the Acoustical Society of America, 60*(3), 704–717.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech, 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386.

McCann, P. (2020). Fugashi, a tool for tokenizing Japanese in Python. In *Proceedings of second workshop for NLP open source software (NLP-OSS)* (pp. 44–51). Association for Computational Linguistics. Online.

Ménard, L., Schwartz, J. L., & Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech Communication, 50*(1), 14–28.

Meyer, E. (1904). Zur Vokaldauer im Deutschen. *Nordiska studier tillegnade Adolf Noreen*. K.W. Appelberg.

Miller, R. L. (1953). Auditory tests with synthetic vowels. *The Journal of the Acoustical Society of America, 25*(1), 114–121.

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018),* Miyazaki, Japan. European Language Resources Association (ELRA).

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). VoxCeleb: Large-scale speaker verification in the wild. *Computer Speech & Language, 60* 101027.

Navarro, T. (1916). Cantidad de las vocales acentuadas. *Revista de Filología Española, 3*, 387–408.

Nooteboom, S. G., & Doodeman, G. J. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America, 67*(1), 276–287.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America, 54*(5), 1235–1247.
O'Shaughnessy, D. (1981). A study of French vowel and consonant durations. *Journal of Phonetics, 9*(4), 385–406.
Park, K. (2019). g2pK [Python]. https://github.com/Kyubyong/g2pK (Original work published 2019).
Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America, 32*(6), 693–703. https://doi.org/10.1121/1.1908183.
Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., Chormai, P., Limkonchotiwat, P., Suntorntip, T., & Udomcharoenchaikit, C. (2023). PyThaiNLP: Thai natural language processing in Python. In *Proceedings of the 3rd workshop for natural language processing open source software (NLP-OSS 2023)* (pp. 25–36). Singapore: Association for Computational Linguistics.
Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *The Journal of the Acoustical Society of America, 22*(6), 807–820.
Recasens, D. (2015). The effect of stress and speech rate on vowel coarticulation in Catalan vowel–consonant–vowel sequences. *Journal of Speech, Language, and Hearing Research, 58*(5), 1407–1424. https://doi.org/10.1044/2015_JSLHR-S-14-0196.
Schoormann, H. E., Heeringa, W. J., & Peters, J. (2017). A cross-dialectal acoustic study of Saterland Frisian vowels. *The Journal of the Acoustical Society of America, 141*(4), 2893–2908.
Sluijter, A. M., & van Heuven, V. J. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica, 52*(2), 71–89.
Solé, M.-J., & Ohala, J. J. (2010). What is and what is not under the control of the speaker: Intrinsic vowel duration. *Papers in Laboratory Phonology, 10*, 607–655.
Stone, A. W. (2015). *The production and perception of vowel height and duration* (Master's thesis). Carleton University.
Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *The Journal of the Acoustical Society of America, 62*(2), 435–448.
Tauberer, J., & Evanini, K. (2009). Intrinsic vowel duration and the post-vocalic voicing effect: Some evidence from dialects of north American English. *Proceedings of Interspeech, 2009*, 2211–2214. https://doi.org/10.21437/Interspeech.2009-629.
Tilsen, S., & Tiede, M. (2023). Parameters of unit-based measures of speech rate. *Speech Communication, 150*, 73–97. https://doi.org/10.1016/j.specom.2023.05.006.
Ting, C., Sonderegger, M., Clayards, M., & McAuliffe, M. (2025). The crosslinguistic distribution of vowel and consonant intrinsic F0 effects. *Language, 101*(1), 1–36.
Tingsabadh, M. K., & Abramson, A. S. (1993). Thai. *Journal of the International Phonetic Association, 23*(1), 24–28.
Toivonen, I., Blumenfeld, L., Gormley, A., Hoiting, L., Logan, J., Ramlakhan, N., & Stone, A. (2015). Vowel height and duration. In *Proceedings of the 32nd west coast conference on formal linguistics* (pp. 64–71).
Tsao, Y.-C., Weismer, G., & Iqbal, K. (2006). The effect of intertalker speech rate variation on acoustic vowel space. *The Journal of the Acoustical Society of America, 119*(2), 1074–1082.
Tyers, F. (2021). Commonvoice-utils [Python]. https://github.com/ftyers/commonvoice-utils (Original work published 2021).
Van Hoof, S., & Verhoeven, J. (2011). Intrinsic vowel F0, the size of vowel inventories and second language acquisition. *Journal of Phonetics, 39*(2), 168–177. https://doi.org/10.1016/j.wocn.2011.02.007.
Westbury, J. R., & Keating, P. A. (1980). Central representation of vowel duration. *The Journal of the Acoustical Society of America, 67*(S1), S37–S.
Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics, 23*(3), 349–366.
White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics, 38*(3), 459–471. https://doi.org/10.1016/j.wocn.2010.05.002.
Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America, 91*(3), 1707–1717.
Wu, Y., Adda-Decker, M., & Lamel, L. (2023). Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position. *Speech Communication, 146*, 45–52.
Zhang, M., Ahn, E. P., and Chodroff, E. (2025). VoxCommunis. Accessible online, https://huggingface.co/datasets/pacscilab/VoxCommunis.
Zhang, M., Farhadipour, A., Baker, A., Ma, J., Pricop, B., & Chodroff, E. (2025). Quantifying and reducing speaker heterogeneity within the Common Voice Corpus for phonetic analysis. In *Proceedings of interspeech 2025*, Rotterdam, Netherland.
Zhu, J., Zhang, C., & Jurgens, D. (2022). Phone-to-audio alignment without text: A semi-supervised approach. In *Proceedings of ICASSP 2022* (pp. 8167–8171).