Speaking rate effects on Japanese vowel and consonant length contrasts

Hironori Katsuda*, Yoonjung Kang

*Corresponding author: katsuda1123@gmail.com
University of Toronto Scarborough, 1265 Military Trail, Toronto, ON, M1C 1A4, Canada,

Abstract
This study examines the sensitivity of vowels and consonants to speaking rate variations in both production and perception, using Japanese as a case study. In contrast to prior studies, which suggest that vowels are more responsive to speaking rate changes than consonants in production, our results indicate a more nuanced distinction between vowels and stops versus fricatives and nasals, with the former group exhibiting greater sensitivity to speaking rate changes. Furthermore, this production pattern was also generally reflected, though to a lesser extent, in the perception results. These findings point to the need for further research into factors such as the presence or absence of length distinctions, language-specific prosodic and rhythmic characteristics, and the relationship between the ratios of long to short segments and slow to fast speaking rates.

Keywords
Speaking rate, Segmental duration, Length contrasts, Perception-production link

1. Introduction

Segmental durations are influenced by various factors, including surrounding segments and prosodic structure. Among these, speaking rate has been particularly well-studied. It is generally expected that slower speech results in longer segmental durations. However, there are still questions regarding how changes in speaking rate influence different segments and how listeners adapt to such variation. Specifically, when speaking rate changes, how do different categories of segments, such as vowels and consonants, respond? Are all types of segments equally lengthened or shortened, or are some more sensitive to rate changes than others? Similarly, in perception, how do the boundaries between short and long segments shift with speaking rate, and are all segments equally affected, or do differences exist across segment types?

On the one hand, numerous production studies have investigated factors influencing segmental duration, with several specifically focusing on speaking rate (Gay, 1978; 1981; Kuwabara, 1996; 1997; Port, 1981; Port et al., 1980). A key question raised by this line of research is how changes in speaking rate affect different segmental categories, particularly contrasting vowels and consonants. The prevailing view is that vowels are more sensitive—or more elastic—with respect to speaking rate than consonants. However, as we will discuss in Section 2, two important issues remain unresolved. First, many prior studies have either closely examined a narrow set of segments or made broad vowel-consonant comparisons without further distinguishing segment types (e.g., stops vs. fricatives), with the notable exception of Tilsen and Tiede (2023). This limits the generalizability of their findings, raising the possibility that the greater elasticity observed for vowels may reflect properties of the specific stimuli or language under investigation. Second, there is a lack of perception studies to determine whether such rate effects are perceptually relevant. This gap may stem from the fact that these production studies often did not target phonemic contrasts, which are crucial for designing perception experiments.

On the other hand, another line of research examines the impact of speaking rate variations on the temporal acoustic cues that signal phonemic contrasts. Examples of such cues include Voice Onset Time (VOT) for signaling stop voicing contrasts and closure duration for differentiating stop lengths (e.g., Beckman et al., 2011; Kessinger & Blumstein, 1997; Miller & Baer, 1983; Miller, Green, & Reeves, 1986; Miller & Liberman, 1979; Miller & Volaitis, 1989; Mitterer, 2018; Pickett, Blumstein, & Burton, 1999; Summerfield, 1981; Volaitis & Miller, 1992). These studies delve into how speakers produce these crucial cues at different speaking rates and how listeners perceive them, given that any variations in these cues could potentially compromise effective communication. As we will explore further in Section 2, existing evidence indicates that the influence of speaking rate is evident in both production and perception: speakers change the duration of these cues based on speaking rate, and listeners adjust their interpretation of these cues correspondingly. However, because these studies primarily focus on the production and/or perception of individual length contrasts rather than comparing different segmental categories, it remains unclear how closely listeners' adjustments reflect production behaviors and whether there are any differences among various segmental categories in this respect.

The present study aims to fill these gaps in the literature by addressing two primary research questions. First, we investigate whether distinct segmental categories, whose lengths are contrastive, exhibit varying degrees of sensitivity to speaking rate variations in production. Second, we explore whether the observed differences, or lack thereof, among these segmental categories in production are also evident in perception. By examining these questions, we aim to deepen our understanding of how speaking rate influences both the production and perception of length contrasts across different segmental categories.

To examine these research questions, we conduct our study using Tokyo Japanese (henceforth "Japanese") as the test language. Japanese provides a valuable context for investigation due to its length contrasts in various segments, including both vowels and consonants, which are primarily distinguished based on durational differences (Vance, 2008; Kawahara, 2015). We note that structural differences exist between vowels and consonants, which will be discussed in Section 2.3. We conduct two experiments, each consisting of production and perception tasks, to assess the effects of speaking rate on different segments in each modality.

The structure of this paper is as follows. Section 2 delves into the background and literature relevant to this study, providing a comprehensive understanding of existing research. Section 3 presents Experiment 1, and Section 4 covers Experiment 2. The broader implications of the findings from these experiments and their limitations are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Background

2.1. Effects of speaking rate on the production of different segments

A limited number of production studies have examined the effects of speaking rate on different types of segments (Gay, 1978; Kuwabara, 1996; Port, 1981; Port et al., 1980). These studies typically assessed the proportional changes in vowels and consonants due to speaking rate variations. Generally, they found that vowels are more sensitive to rate changes than consonants.

Two experimental studies have focused on English. In one study, Gay (1978) had four participants produce a series of CVC syllables, incorporating nine vowels (/i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ/) within a /p_p/ environment. These syllables were embedded in a carrier sentence and articulated at both slow and fast rates. By calculating the duration ratios of fast to slow rates, Gay found that vowels exhibited a more significant reduction in duration during fast speech compared to consonants. Specifically, the mean duration ratio for vowels was 0.75 (25% reduction), while it was 0.9 (10% reduction) for consonants (a lower ratio indicates a greater difference between fast and slow rates). Similarly, Port (1981) had five participants produce four CVCV words (*deeber*, *dibber*, *deeper*, and *dipper*) within a carrier sentence at fast and slow speaking rates. Analyzing the percentage changes in duration for the initial stressed vowel (/ˈCV̲CV/) and the subsequent stop (/ˈCVC̲V/) during fast speech, Port observed a 26% reduction in vowel duration compared to a 20% reduction in stops, aligning with Gay's (1978) results.

A comparable pattern has also been observed in languages other than English. Port et al. (1980) explored how Arabic speakers produced trisyllabic words (/ˈCV(V)CVCV/) within a carrier phrase at three distinct rates: fast, normal, and slow. The study measured the degree of reduction in fast speech and expansion in slow speech for the initial stressed vowel (/a(a)/) and the subsequent consonant (/t/, /d/, or /ɾ/). The results indicated that vowels underwent more significant changes, reducing by 25% and expanding by 50%, while consonants exhibited lesser variability, changing by 12% in both cases.

In another study focusing on Japanese, Kuwabara (1996) had four participants produce 15 short sentences at fast, normal, and slow speaking rates. These sentences contained a range of segmental categories, including short and long vowels, singleton and geminate consonants, and moraic nasals. When calculating the proportion of consonants and vowels in CV syllables (excluding long vowels, geminate consonants, and moraic nasals) at each rate, Kuwabara observed that the consonant-to-vowel ratios remained relatively stable for normal and fast rates (approximately 35% for consonants and 65% for vowels for both rates). In contrast, at the slow speaking rate, the proportion of vowels increased substantially to 75.7%, with consonants making up the remaining 24.3%. Based on this data, Kuwabara concluded that vowels were more influenced by speaking rate than consonants.

In a more recent study, Lo and Sóskuthy (2023) further explored this issue through a corpus analysis. They examined data from eight unrelated languages: Korean, Mandarin, Amharic, Georgian, Swahili, Turkish, Vietnamese, and English. They described "articulation rate" as the average segment duration within an utterance. For instance, an utterance lasting 1 second with 5 segments would have an articulation rate of 5 segments per second (5/1 = 5). Their goal was to determine whether consonant and vowel durations change differently as articulation rate shifts. Their results largely corroborated earlier studies, showing that vowels undergo a significantly greater degree of duration adjustment compared to consonants. They argued that this difference could partly be attributed to the increased aerodynamic and coordinative complexities involved in constricting airflow for consonants, making them less responsive to changes in articulation rate than vowels.

Tilsen and Tiede (2023) conducted a corpus study to investigate methods for quantifying speaking rate and evaluate different parameter choices involved in doing so. They used the Haskins Production Rate Comparison database (Tiede et al., 2017), which contains acoustic and articulatory recordings from eight English speakers producing 720 phonetically balanced sentences at both normal and relatively fast speaking rates. The study assessed various methods of estimating speaking rate by examining how well different rate measures correlated with target segment durations. The parameters included unit type (i.e., the linguistic units used to calculate rate, such as words, syllables, or phones) and window size (i.e., the temporal window over which units were counted). Crucially, target segment durations were analyzed across phonologically defined categories, including vowels vs. consonants, onset vs. coda consonants, stressed vs. unstressed vowels, and stop vs. non-stop consonants. The findings most relevant to the present study are that rate measures showed stronger correlations with the durations of coda consonants

compared to onset consonants, stressed vowels compared to unstressed vowels, and non-stop (more sonorous) consonants compared to stops. Tilsen and Tiede interpret these patterns in relation to Tilsen's (2022) model, which proposes that speaking rate is modulated by attention to sensory feedback. Segments that show stronger correlations with speaking rate are thought to be those for which speakers rely more heavily on such feedback. However, they also acknowledge the possibility that these asymmetries may reflect limitations in the accuracy of forced alignment.

In summary, most previous studies have concluded that vowels are more responsive to changes in speaking rate than consonants. However, this conclusion is typically based on a limited set of stimuli—often involving only stops as consonants (Gay, 1978; Port, 1981; Port et al., 1980, which also included /ɾ/) and a narrow range of vowels (Port, 1981; Port et al., 1980), with the exception of Gay (1978), which examined nine vowels. Other studies rely on broad comparisons between vowels and consonants, potentially overlooking important distinctions among more specific categories (Kuwabara, 1986; Lo and Sóskuthy, 2023), though Tilsen and Tiede (2023) provide a notable exception. These limitations underscore the need for further research using carefully selected stimuli and more fine-grained analyses.

2.2. Effects of speaking rate on phonemic contrasts

Numerous studies have explored the effects of speaking rate on the production and perception of temporal acoustic cues that signal phonemic contrasts. A non-exhaustive list of previous studies includes Miller and Liberman (1979), Summerfield (1981), Miller and Baer (1983), Miller, Green, and Reeves (1986), Miller and Volaitis (1989), Volaitis and Miller (1992), Kessinger and Blumstein (1997), Pickett, Blumstein, and Burton (1999), Beckman et al. (2011), and Mitterer (2018). Production studies have consistently yielded two primary findings: first, as speaking rate varies, the duration of acoustic cues also changes, potentially shifting category boundary separating contrasting phonemes; second, the influence of speaking rate on these contrasting phonemes is asymmetric in terms of raw duration, with segments having shorter acoustic cues being less impacted by rate changes compared to those with longer acoustic cues, though the reasons for this remain unclear (see Magen and Blumstein, 1993 for possible hypotheses and counterexamples). Regarding perception, studies demonstrate that listeners adjust their use of acoustic cues based on speaking rate, a phenomenon often referred to as "rate normalization".

For instance, Miller, et al. (1986) conducted an experiment to study the effects of speaking rate on VOT for English stops. In their production study, three speakers produced the syllables /bi/ and /pi/ at various speaking rates. The findings revealed that speakers produced longer VOT values for bilabial stops in slow speech compared to fast speech, with a more pronounced effect observed for the long-lag [p^h] (/p/) compared to the short-lag [p] (/b/) in terms of raw duration. Furthermore, their perception study illustrated the effect of speaking rate in perception as well, showing that listeners required longer VOT values to perceive /p/ in slow speech compared to fast speech. Similar results have been observed for other types of phonemic contrasts, including the contrast

between stops and glides (i.e., /ba/ and /wa/) in English (Miller & Liberman, 1979; Miller & Baer, 1983) and the contrast between singleton and geminate stops in Italian (Picket, et al., 1999).

The effects of speaking rate on phonemic contrasts in Japanese have been extensively studied within the framework of acoustic invariance theory, which posits that certain acoustic patterns remain constant across various speaking conditions. (Amano & Hirata 2010; Amano & Hirata, 2015; Hirata, 2004; Hirata & Amano, 2012; Hirata & Lambacher, 2004; Hirata & Whiton, 2005; Idemaru & Guion-Anderson, 2010; Idemaru, Holt, & Seltman, 2012). Although the primary objective of these studies was not to examine the impact of speaking rate specifically, they did observe significant effects of speaking rate on both the vowel and stop length contrasts. Specifically, Hirata (2004) and Hirata and Whiton (2005) revealed that both short and long vowels, as well as singleton and geminate stops, increase in duration with slower speaking rates, with a more notable lengthening for long vowels and geminate stops. Additionally, Hirata and Lambacher (2004) demonstrated that listeners' perception of vowel length is influenced by speaking rate. Amano and Hirata (2010), as well as Idemaru and Guion-Anderson (2010), showed that a longer closure duration is required to perceive geminate stops in slower speech.

Some indirect evidence suggesting potential differences between vowels and consonants can be obtained from recent perception studies focusing on specific aspects of perception. Reinisch (2016) explored how a speaker's typical speaking rate, or "habitual speaking rate", influences the perception of phonemic contrasts, particularly in distinguishing vowel length in German. In the experiment, German listeners first heard a dialogue between a fast and a slow speaker. During the test phase, they identified whether the words spoken by these speakers contained a short /a/ or a long /aa/. The results showed that words with a long /aa/ were more frequently identified when spoken by the fast speaker, demonstrating perceptual compensation based on habitual speaking rates.

Ting and Kang (2023a; 2023b) attempted to replicate these effects of habitual speaking rate on the distinction between English /p/ and /b/. However, they observed these effects only when the speaking rates of the dialogue were manipulated more dramatically than in Reinisch's (2016) study. This led them to speculate that listeners might be more attuned to rate changes in vowels than in consonants. However, this conclusion remains tentative as it is based on an indirect comparison of vowel duration and VOT duration across different languages and experimental designs. Similarly, Heffner et al. (2017) found that in English, while distal speaking rate (i.e., speaking rate of surrounding words) influenced the perception of coda voicing contrasts, signaled by the duration of the preceding vowel, it did not affect the perception of word-initial voicing contrasts, cued by VOT.

Gadanidis and Kang (under revision) investigated the sensitivity of vowels and stops to speaking rate changes in perception. They found that the perception of stop length contrasts was more affected by speaking rate changes than that of vowel length contrasts. Additionally, they noted that vowel perception was consistently influenced by rate changes, regardless of whether the speaking rate of the carrier sentence was altered by manipulating the duration of vowels in the carrier sentence or not. In contrast, stop perception was more impacted when the duration of

consonants in the carrier sentence was manipulated. This led the authors to conclude that consonants are more sensitive to their specific duration, while vowels are more responsive to the global speaking rate.

Finally, Kawahara, Kato and Idemaru (2022) conducted a perception experiment to examine the persistence of the speaking rate effects across different talkers. They aimed to determine whether the influence of speaking rate remains consistent when the precursor and target words are spoken by different speakers, and whether there are differences between vowels and stops in this respect. They found no differences between vowels and stops in terms of how speaking rate was normalized across different speakers. However, they unexpectedly observed that the influence of speaking rate, specifically fast vs. normal, was significantly greater for stops than for vowels. They attributed this asymmetry to the limited duration range for the vowel duration continuum used in their experiment, speculating that it may not have sufficiently captured the natural variability of vowel durations.

In summary, while many studies have separately investigated the effects of speaking rate on length contrasts in various segment types, none have directly compared these effects across different segments in both production and perception. Some comparisons exist, but they are either indirect, drawn from different studies, or limited to perception alone. Additionally, perception studies on Japanese have suggested that stops might be more sensitive to speaking rate changes than vowels, which does not align with the general production tendency, if perception reflects production behavior. These underscores the need for direct comparisons in both production and perception within a single study.

2.3. The present study

As summarized in Section 2.1, previous studies have indicated that vowels generally exhibit greater sensitivity to speaking rate changes compared to consonants during production. However, this conclusion is based on a limited set of studies, which may be influenced by the specific stimuli, language, or methods of analysis used. Section 2.2 highlighted that, while many studies have explored the impact of speaking rate on length contrasts in various segmental categories, none have provided a comprehensive, systematic comparison of the effects on length contrasts across different segmental categories in both production and perception. Our study aims to deepen our understanding of the effects of speaking rate on the production and perception of length contrasts by comparing different segmental categories. We use Japanese as the test language due to its extensive length contrasts across a variety of segments, which are primarily distinguished by durational differences (Kawahara, 2015; Vance, 2008).

This study comprises two experiments, each involving both production and perception tasks. In the production task, participants performed an imitation task in which they produced target words embedded in a carrier phrase, attempting to match the speaking rate of auditory prompts. This method was chosen instead of instructing participants to speak at a specific rate or using visual cues, in order to better align the production data with the perception data. Specifically,

it offers more precise control over speaking rate by reducing variability in how participants interpret rate-related instructions. It also helps better control over prosodic features, which is particularly important when eliciting full sentences rather than isolated words. In the perception task, participants completed a two-alternative forced-choice task, categorizing target words presented at fast and slow speaking rates as containing either phonemically short or long segments.

Experiment 1 focused on comparing two segments, /o(o)/ and /k(k)/, using real-word stimuli. Experiment 2 expanded the scope to a broader set of segments, including five vowels (/i(i)/, /e(e)/, /a(a)/, /o(o)/, /u(u)/) and five consonants—two stops (/t(t)/, /k(k)/), one fricative (/s(s)/), and two nasals (/n(n)/, /m(m)/)—using nonce-word stimuli. All obstruents were voiceless, as voiced obstruent geminates are absent from native Japanese words and are clearly dispreferred, as evidenced by several phonological processes (see Kawahara, 2005 for detailed discussion).

It is important to acknowledge that the comparison between vowels and consonants inherently involves structural differences in terms of prosodic properties, making it challenging to compare them purely on phonetic grounds. The most fundamental difference lies in their positions within syllable structure: vowels occur as nuclei while consonants usually serve as onsets or codas. Additionally, in Japanese phonology, it is generally assumed that short and long vowels correspond to one and two moras, respectively, while singleton (short) and geminate (long) consonants are considered mora-less and one mora, respectively (McCawley, 1965; Poser 1984). Although these structural differences present potential confounds, they are unavoidable when comparing vowels and consonants. We acknowledge them as potential sources of variation in how segments respond to changes in speaking rate. We revisit these issues in Section 5.1.3.

3. Experiment 1

The purposes of this experiment were twofold: (i) to determine whether the influence of speaking rate differs between two types of segments, namely /o(o)/ and /k(k)/, in production, and (ii) to assess whether the differences, or lack thereof, observed in production are reflected in perception. According to the consensus in the literature, vowels tend to be more sensitive to speaking rate changes than consonants in production. Therefore, /o(o)/ would be expected to exhibit greater sensitivity than /k(k)/. If perception closely mirrors production, the perception results are expected to be consistent with the production patterns. Specifically, if there is a difference between /o(o)/ and /k(k)/ in production, a similar difference should be observed in perception. Conversely, if no difference is found in production, no difference is expected in perception either. It is also noted that some previous studies indicated that stops are more sensitive to speaking rate changes than vowels in perception (Gadanidis & Kang, under revision; Kawahara et al. 2022).

3.1.Methods

3.1.1. Participants

Participants were recruited through Crowdworks.jp, a crowdsourcing platform based in Japan. A total of 69 self-identified Tokyo Japanese speakers completed the experiment and were compensated for their participation (49 females, mean age: 38, age range: 20-64). Due to various issues observed during the production task, as discussed later, data for this task include only a subset of participants. In contrast, data for the perception task include all participants.

3.1.2. Stimuli

The target words selected for the study were the minimal triplet, /sjokan/ 'letter', /sjookan/ 'summon', and /sjokkan/ 'texture'. The first and second words formed a minimal pair contrasting in vowel length (i.e., /sj**o**kan/-/sj**oo**kan/), while the first and third words formed a minimal pair contrasting in stop (closure) length (i.e., /sjo**k**an/-/sjo**kk**an/) (although there is also a structural difference, as discussed in Section 2.3). All target words were unaccented, which as we will discuss in the next paragraph, simplifies control over their fundamental frequency (f0) patterns since unaccented words carry only phrasal tones, unlike accented words. The lexical frequencies of the three target words, based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ), differ slightly, with raw counts as follows: /sjokan/: 545, /sjookan/: 49, and /sjokkan/: 522, making /sjookan/ less frequency than the other two. Although lexical frequency differences may influence baseline perception responses, we argue that matching frequencies is not essential for assessing the effect of speaking rate (note that Experiment 2 uses nonce words, eliminating this concern altogether).

These target words were embedded in the carrier sentence shown in (1). We deliberately avoided using long segments in the carrier sentence to eliminate cues that listeners could directly reference to determine the length of long segments. Instead, listeners are required to infer the category boundary based on the duration of short segments and the speaking rate.

(1) /takéutisan-wa odájakani sono **[target]** to hatuon-sita/
Mr. Takeuchi-TOP gently that **[target]** COMP pronounce-past
"Mr. Takeuchi gently pronounced that **[target]**"

In Japanese, the left edge of an accentual phrase (AP), the smallest unit of prosodic phrasing, is marked by an f0 rise. This rise is analyzed as a combination of a low boundary tone (%L) and a high phrasal tone (H-) associated with the second mora in the Autosegmental-Metrical model of Japanese intonational phonology (Beckman & Pierrehumbert, 1986; Venditti, 1997). This presents a challenge when creating a perceptually natural duration continuum, particularly for the /o/~/oo/ contrast, as the words /sjokan/ and /sjookan/ exhibit distinct f0 patterns in AP-initial positions. Specifically, /sjokan/ displays low and high f0 values on the first (/sjo/) and

second syllables (/kan/), respectively, while /sjookan/ carries an f0 rise on the first syllable (/sjoo/). Such pitch differences have been shown to affect Japanese listeners' perception of vowel length (Kinoshita, Behne, & Arai, 2002; Kozasa, 2005; Takiguchi, Takeyasu, & Giriko, 2010). To prevent the f0 contour from providing a cue to vowel length, the target words were positioned AP-medially following the demonstrative determiner /sono/ 'that'. In this way, the f0 rise occurs on the demonstrative determiner, while the target words themselves are realized with a gradual f0 fall, allowing for better control over f0 variation in the /o/~/oo/ continuum. However, in the production task, not all participants adhered to this intended phrasing, resulting in the exclusion of data from some participants, as we will discuss this later.

Having introduced the target words and carrier sentence, we will now explain the details of stimulus creation for both the production and perception tasks. Since the stimuli for the production task are derived from those used in the perception task, we will first describe the perception task stimuli, followed by an explanation of the production task stimuli.

Perception

The speech materials for the experiment were recorded by a male native speaker of Tokyo Japanese in his 30s. The speaker produced the sentence described in (1) with each target word repeated five times at a normal speaking rate. Manual annotations were performed on all utterances to measure the durations of individual segments and enable subsequent manipulations using PSOLA in Praat (Boersma & Weenink, 2021). Figure 1 illustrates the segmentation of the carrier sentence into three parts: the pre-target carrier phrase (labeled as carrier 1), the demonstrative determiner /sono/ (dem), and the post-target carrier phrase (carrier 2), as shown at the top of the figure. The bottom part of Figure 1 depicts the segmentation of the target words into five parts, as shown in (2).

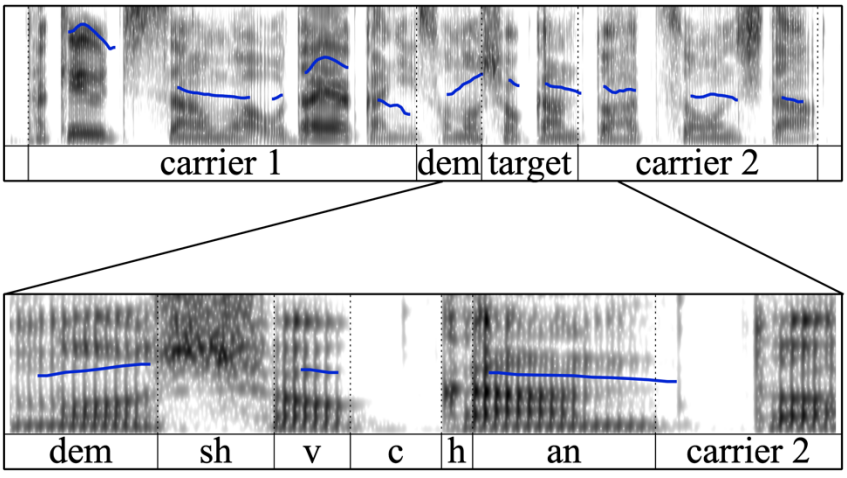


Figure 1. Sample spectrograms showing the segmentation of utterances.

(2) Segmentation of the target words

- sh: /sj/
- v: /o(o)/
- c: stop closure of /k(k)/
- h: aspiration of /k(k)/
- an: /an/

From the recorded utterances of /sjokan/, one token was selected as the baseline for the target word, and another was chosen as the baseline for the carrier sentence components (carrier 1, demo, carrier 2). These two selected utterances were then spliced together to create the baseline utterance. The baseline utterance served as the foundation for both the perception stimuli and production prompts, which will be described later.

The spliced baseline utterance was manipulated to create 12-step duration continua for both /o/~/oo/ length and /k/~/kk/ length contrasts. The duration continuum for the former (v) ranged from the average duration of /o/ in /sjokan/ (58ms) to the average duration of /oo/ in /sjookan/ (156ms), as produced by the model speaker. Similarly, the duration continuum for the latter (c) ranged from the average closure duration of /k/ in /sjokan/ (71ms) to the average closure duration of /kk/ in /sjokkan/ (179ms). The intervals between each step were maintained as equidistant.

For the /sjokan/~/sjookan/ continuum, the closure duration was adjusted to the average value of /k/ in both /sjokan/ and /sjookan/ (based on 10 tokens). Similarly, the duration of /o/ for the /sjokan/~/sjokkan/ continuum was adjusted to the average duration of /o/ in both /sjokan/ and /sjokkan/ (based on 10 tokens). Other parts of the target words (i.e., sh, h, an) and subparts of the carrier sentence (carrier 1, dem, carrier 2) were also adjusted to match the average duration of all productions by the model speaker (based on 15 tokens). This was done to prevent the possibility that the duration of any specific part might influence the responses. The intensity of the stimuli was scaled to 70dB.

To establish the ranges and steps for the duration continua used in the main experiment, a pretest was conducted. Fifteen native speakers of Tokyo Japanese, who did not take part in the main experiment, completed an identification task remotely and were compensated for their participation. They were instructed to listen to the stimuli using headphones and categorize the target word as containing either phonemically short or long segment, depending on the contrast (i.e., /sj**o**kan/~/sj**oo**kan/, /sjo**k**an/~/sjo**kk**an/, presented in Chinese characters). The trials were organized into blocks based on the type of contrast, with the order of the blocks counterbalanced across participants. Within each block, the trials were completely randomized. Each unique stimulus was categorized twice, resulting in a total of 48 trials (12 steps × 2 types of contrasts × 2 repetitions).

Figure 2 presents the results of the pretest, shown as solid lines with circles. The left panel (a) illustrates the categorization pattern for the /o(o)/ contrast, while the right panel (b) shows the categorization pattern for the /k(k)/ contrast. As anticipated, segment duration (horizontal axis)

significantly influences categorization (vertical axis) in both contrasts: longer segment durations elicit more responses indicating a long segment.

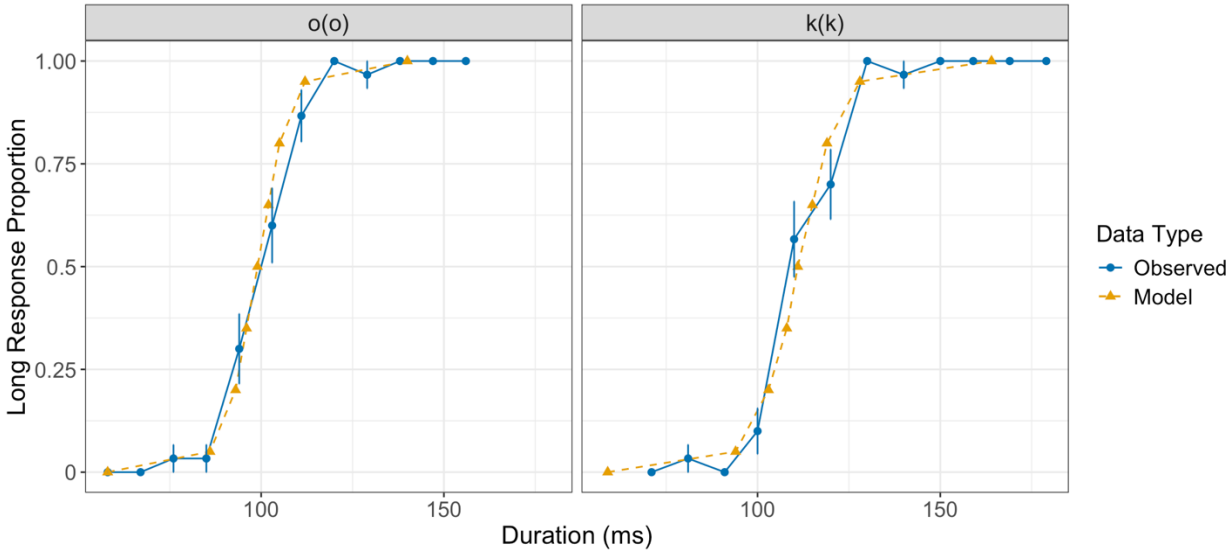


Figure 2. Proportion of long responses at each duration step for the /o(o)/ contrast (left) and the /k(k)/ contrast (right). Error bars indicate the standard error. Solid lines with circles represent the raw data, while dashed lines with triangles show duration steps estimated by the mixed-effects logistic regression models.

To account for the inherent duration differences between /o(o)/ and /k(k)/, and to make the two continua as comparable as possible, the ranges and steps of the duration continua in the main experiment were determined based on model-estimated probabilities of long responses. A logistic mixed-effects regression model[1] was fitted separately for the /o(o)/ and /k(k)/ contrasts. The model predicted listeners' responses (long or short) using centered duration as the predictor, with random intercepts for participants and by-participant random slopes for centered duration. Based on the model, we estimated the durations corresponding to nine long-response probability levels, as shown in Table 1. These estimated durations, represented by dashed lines with triangles in Figure 2, formed the duration steps for each length contrast continuum.

[1] The model was specified in R as: RESPONSE~DURATION.CENTERED+(DURATION.CENTERED|PARTICIPANT).

Probability of perceiving long	Duration (ms)	
	/o(o)/	/k(k)/
0.0001	58	59
0.05	86	94
0.2	93	103
0.35	96	108
0.5	99	111
0.65	102	115
0.8	105	119
0.95	112	128
0.9999	140	164

Table 1. Duration steps for the /o(o)/ contrast and the /k(k)/ contrast estimated by the mixed-effects logistic regression models.

The baseline utterance was adjusted to incorporate the new duration steps for each condition, and then manipulated to create two speaking rate conditions: fast and slow. These conditions were created by modifying the duration of the carrier sentence components (i.e., carrier 1, dem, carrier 2). In the fast condition, the duration of the carrier sentence was reduced to 80% of its original duration, while in the slow condition, it was extended to 120% of the original duration.[2]

It is important to note that the manipulation of speaking rate involved only altering the portions of the utterance that were not directly adjacent to the target segment. The durations of the non-target parts in the target word (i.e., sh, c, h, an in Figure 1, for the /o(o)/ contrast and sh, v, h, an in Figure 1, for the /k(k)/ contrast) were kept constant across both speaking rate conditions. This approach aimed to prevent potential confounds where the duration of an adjacent segment, rather than the speaking rate itself, could serve as a cue for length contrasts (Miller & Liberman, 1979; Summerfield, 1981; Toscano & McMurray, 2012; Toscano & McMurray, 2015). A possible drawback of this approach in the context of our study is discussed in Section 5.2.3.

Production

The production prompts were derived from the baseline stimuli, which were initially created during the process of generating perception stimuli. To prevent the model speaker's production of the target word from influencing participants' productions, the demonstrative determiner /sono/

[2] A reviewer raised concerns about whether the rate manipulation, implemented using PSOLA, resulted in unnatural stimuli, especially given our conclusion that different types of segments respond differently to speaking rate changes. However, the degree of manipulation used in our study was considered entirely natural by the first author of the paper, a native speaker of Japanese. We believe that the rate-induced variation among segments is subtle enough that listeners are unlikely to notice it, particularly when the manipulation is not extreme. In fact, the degree of manipulation was relatively modest compared to other studies, such as Maslowski et al. (2019), where the slow and fast conditions are set at 160% and 62.5%, respectively, and Bosker (2017), where the conditions are 133% and 75%.

and the target word were beeped out.[3] The duration of the beep for the demonstrative determiner was adjusted to match the model speaker's average duration (based on 15 tokens), and the duration of the beep for each target word was adjusted to match the model speaker's average duration for that word (based on 5 tokens for each target word). This process resulted in the final production prompt for each target word.

Similar to the perception stimuli, the durations of the other parts of the carrier sentence (i.e., carrier 1 and carrier 2) was adjusted to match the average duration of the model speaker's productions. Finally, the duration of each stimulus was adjusted to 80% of the original duration for the fast condition and 120% for the slow condition, consistent with the perception stimuli.

3.1.3. Procedure

The participants completed the experiment remotely and were instructed to perform the tasks in a quiet room. The perception task was administered first, followed by the production task, to prevent any potential influence of the production task on the perception task. This order was chosen because perception can be more sensitive to subtle factors than production. For example, previous research has shown that even the speaking rate of a participant's own production can influence their perception of subsequent speech signals (Bosker, 2017).[4]

The procedure for the perception task was identical to that of the pretest outlined in Section 3.1.2, with three exceptions. First, the duration continua comprised 9 steps instead of 12. Second, the speaking rate of the stimuli was either fast (at 80% of the original duration) or slow (at 120% of the original duration), deviating from the normal rate. Lastly, each unique stimulus was presented four times instead of twice. Consequently, the perception task involved a total of 144 trials (9 steps × 2 segment types × 2 speaking rates × 4 repetitions).

The production task consisted of an imitation task, following Kang et al. (2018). In each trial of the production task, participants first heard a production prompt at either a fast or slow rate while simultaneously viewing the target sentence on the screen without the demonstrative determiner and the target word (as shown in Figure 3a). Subsequently, the full target sentence was displayed on the screen (Figure 3b). Participants were instructed to repeat the full sentence twice, aiming to closely match their speaking rate to the prompt. Repeating the sentence twice ensured that at least one full utterance was captured, even if the recording was stopped too early (as noted in Section 3.1.4 below, only the first utterance in each trial was analyzed, unless it was unusable due to errors or noise—in which case the second utterance was used instead). Trials were organized into blocks based on the target words (/sjokan/, /sjookan/, /sjokkan/), with the order of the blocks counterbalanced across participants. Within each block, fast and slow trials were randomized. Each

[3] Initially, we beeped out the target word only. However, a pilot study revealed that this approach often let to participants inserting an AP boundary between the demonstrative determiner /sono/ and the target word. This may have been due to participants feeling compelled to emphasize the masked part in the prompt in their productions.

[4] We acknowledge that the perception task could influence the production task. However, we decided to administer the perception task first, based on the belief that any potential influence from this order would be smaller than the influence that might arise from the opposite order.

participant produced each unique stimulus eight times, resulting in a total of 48 tokens (3 target words × 2 rates × 8 repetitions).

At the beginning of each block in the production task, participants heard a model sentence spoken at the normal speaking rate and were instructed to mimic its intonation pattern. This measure aimed to prevent instances where participants might insert an AP boundary between the demonstrative determiner /sono/ and the target word, as discussed in Section 3.1.2. However, it proved insufficient to completely eliminate such cases, as discussed in 3.1.4.

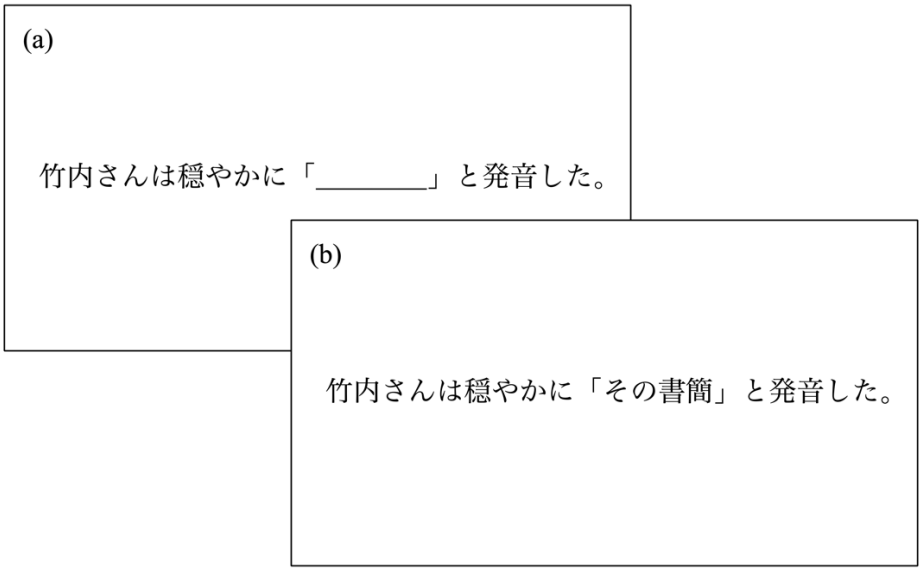


Figure 3. Presentation of production stimuli in Experiment 1.

Before starting the perception task, participants completed a language background questionnaire. The entire experiment, including the questionnaire, and the perception and production tasks, took approximately 30 minutes to complete.

3.1.4. Acoustic measurement and exclusion

The utterances produced by the participants were manually segmented by the first author for carrier 1, carrier 2, and the target segments (i.e., /o/ and /k/ for /sjokan/, /oo/ for /sjookan/, and /kk/ for /sjokkan/). Only the first utterance in each trial was annotated and included in the data. However, in rare instances where the first utterance was unusable due to errors or noise, the second utterance was annotated instead. The durations of these components were then extracted using a Praat script. A sample segmentation of the target word /sjokan/ is shown in Figure 4.

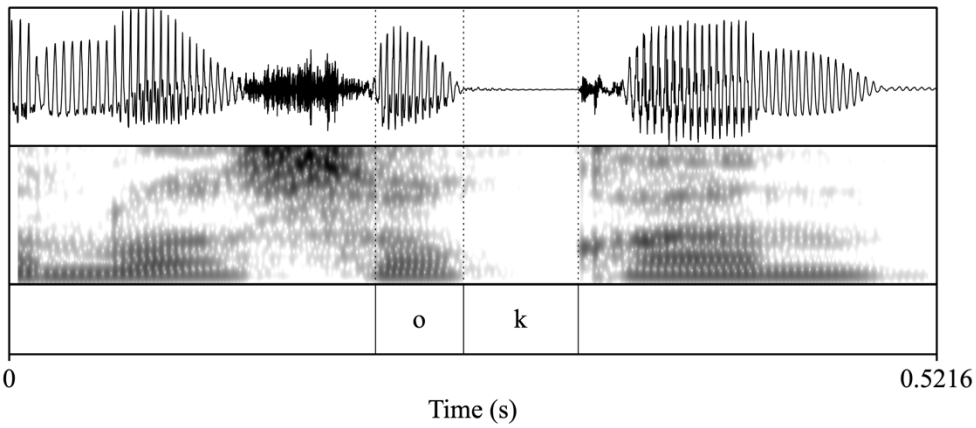


Figure 4. Sample segmentation of the target word /sjokan/.

Out of 69 participants, 23 were excluded from the production data for the following reasons: 16 participants inserted an AP boundary between the demonstrative determiner /sono/ and the target word for at least all eight tokens in a uniquely defined condition by segment target (/o(o)/ and /k(k)/), length (short and long), and speaking rate (fast and slow).[5] The presence of AP boundaries was identified by comparing the average f0 value of the first vowel /o(o)/ and the second vowel /a/ in the target word: if the f0 of the first vowel was lower than the second, it was considered to be lowered by the presence of an AP boundary. Additionally, seven participants produced the target sentences with significantly different intonation patterns compared to the model production. Although these prosodic variations were subtle and their potential influence on the results is unclear, we opted to be conservative and excluded these participants to avoid any potential influence on the production of length contrasts.

Furthermore, eight participants were excluded for other issues: four produced only the demonstrative determiner and the target word, two had poor sound quality, one mispronounced the carrier sentence, and one produced the wrong target word in the target sentence. Consequently, a total of 38 participants (27 females, mean age: 37, age range: 20-62) remained for data analysis for production.

Within the productions of the included participants, 53 utterances had to be excluded for the following reasons: inserting an AP boundary between the demonstrative determiner and the target word (42 cases), failure to record (9 cases), producing only the demonstrative determiner

[5] A reviewer inquired about within-participant variation and why so many participants appeared not to follow the task instruction to mimic the prosodic pattern of the stimuli. Indeed, there was both inter- and within-participant variation. Some participants consistently inserted an AP boundary, while others did so for only a subset of stimuli. Although there may be a systematic pattern regarding when participants tend to insert an AP boundary, we refrain from making speculative comments as this is not the focus of our study. We suggest that one reason for the unexpected insertion of an AP boundary by many participants is that the distinction between the presence and absence of an AP boundary is both phonetically and semantically subtle. Speakers do not appear to be sensitive to this difference, especially in experimental settings where the presence or absence of an AP boundary does not lead to miscommunication.

and the target word (1 case), and background noise affecting the recording (1 case). After excluding these cases, the dataset was reduced to 1771 utterances out of the original 1824 for data analysis. This translates to 2363 tokens out of 2432 tokens (note that utterances containing /sjokan/ involve two target tokens, i.e., /o/ and /k/, while those containing /sjookan/ or /sjokkan/ involve only one, i.e., /oo/ for the former and /kk/ for the latter).

No participants or responses were excluded from the perception data. Consequently, a total of 9936 responses from 69 participants were analyzed.

3.2.Results

In this section, we first analyze the production data (Section 3.2.1) and then the perception data (Section 3.2.2). For the production data, we begin by analyzing the two target segments separately to demonstrate the validity of the data and enable direct comparisons with previous studies. We then compare the two target segments to examine potential differences in their sensitivity to speaking rate, addressing the central question of this study. In contrast, the perception data are analyzed in a single step that incorporates both of these goals. The data and analysis scripts are available on our OSF page.

It is important to note that we do not directly compare sensitivity across production and perception because our perception results may not fully reflect the effects of speaking rate as observed in our production results. This is because we did not manipulate the speaking rate of non-target segments in the target word (/sj/ and /kan/ for the /o(o)/ contrast and /sjo/ and /an/ for the /k(k)/ contrast) in the perception stimuli. This was necessary to avoid participants basing their responses on the duration of surrounding segments. If we had manipulated the speaking rate of those segments, it would be unclear whether participants' responses were influenced by the surrounding segments specifically or the overall speaking rate of the sentence. We consider this a limitation in cross-modality comparisons and therefore approach this question by comparing the results of independent analyses rather than directly comparing across modalities.

3.2.1. Production

We first assess whether our production prompts successfully elicited the productions at the intended speaking rate. Figure 5a displays the average duration of the carrier sentence by participant in both the fast and slow speech conditions. We present average durations rather than durations of individual tokens to account for inter-speaker variation and differences in the number of tokens provided by participants, due to some tokens being excluded. The duration of the carrier sentence includes "carrier 1" (/takéutisan-wa odájakani/) and "carrier 2" (/to hatuon-sita/). As depicted in the figure, the duration of the carrier sentence was shorter in the fast speech condition (mean = 2012ms) compared to the slow condition (mean = 2671ms). These durations are comparable to those of the corresponding parts in the production prompts (fast = 1960ms, slow = 2920ms), which are indicated by cross marks in the figure. While the values indicate that the

speaking rates employed by participants were slightly less extreme than those of the production prompts, this is expected in this type of imitation task (Kang et al., 2018).

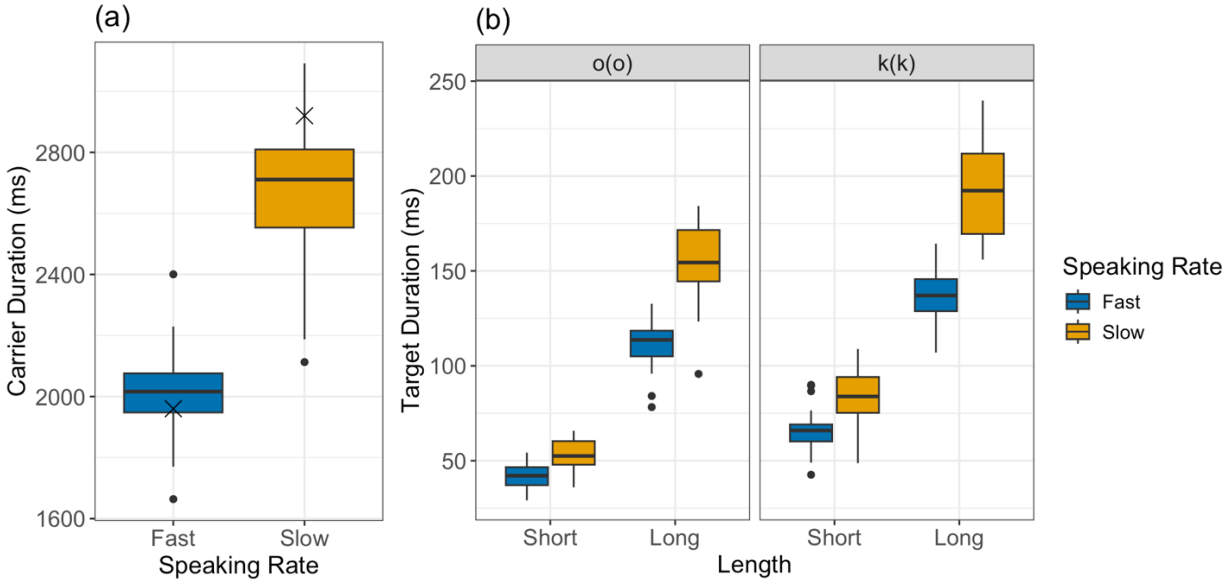


Figure 5. (a) Carrier duration by speaking rate, with cross marks indicating the durations of the corresponding parts in the production prompts. (b) Target duration by length, target segment, and speaking rate.

Figure 5b displays the average durations of target segments by participant in both fast and slow speech conditions. The left panel shows the durations of short /o/ and long /oo/, while the right panel shows the durations of singleton /k/ and geminate /kk/. To analyze the raw production data, we employed a linear mixed-effects regression model[6] for each contrast. The model was fitted to predict the duration of the target segment based on the factors of length (short or long), rate (fast or slow), and their interaction. The categorical predictors were contrast-coded, with short mapped to -0.5 and long mapped to 0.5, while fast mapped to -0.5 and slow to 0.5. Maximal random effects were specified with random intercepts for participants and by-participant random slopes for length, rate, and their interaction.

We observed significant effects for all fixed-effects predictors in both /o(o)/ and /k(k)/ productions. Specifically, phonemically long segments were significantly longer than short segments (/o(o)/: $\beta = 85.89$, $t = 42.17$, $p < 0.001$; /k(k)/: $\beta = 89.86$, $t = 31.36$, $p < 0.001$). Moreover, segments in the slow condition were significantly longer than those in the fast condition (/o(o)/: $\beta = 26.94$, $t = 22.44$, $p < 0.001$; /k(k)/: $\beta = 37.49$, $t = 17.79$, $p < 0.001$). Additionally, phonemically long segments lengthened more than short segments in the slow condition in terms of absolute duration, as expected (/o(o)/: $\beta = 32.45$, $t = 17.60$, $p < 0.001$; /k(k)/: $\beta = 37.65$, $t = 14.64$, $p < 0.001$). Post-hoc analyses using *emmeans* (Lenth, 2020) revealed that the effect of rate was significant for both short and long segments, with long segments exhibiting greater estimates of rate effect for

[6] The model was specified in R as: DURATION~LENGTH*RATE+(LENGTH*RATE|PARTICIPANT).

both /o(o)/ (short: $\beta = 10.7$, $t = 12.50$, $p < 0.001$ vs. long: $\beta = 43.2$, $t = 22.00$, $p < 0.001$) and /k(k)/ (short: $\beta = 18.7$, $t = 12.816$, $p < 0.001$ vs. long: $\beta = 56.3$, $t = 17.735$, $p < 0.001$). Detailed results are provided in Appendix A. We also fit an alternative model using scaled continuous sentence duration in place of the binary rate factor; the results were comparable and are included in the appendix.

Having confirmed the basic effects of speaking rate on /o(o)/ and /k(k)/ separately in production, we now examine whether these segments differ in their sensitivity to speaking rate. To investigate this, we estimated individual speakers' category boundaries between short and long for each speaking rate condition (fast and slow) and each target segment (/o(o)/ and /k(k)/), modeled after the methods used in Nagao & de Jong (2007) and Kang, et al. (2018). We applied a Bayesian logistic regression model[7], to each participant's data for each combination of speaking rate and target segment, using the *bayesglm* function in the *arm* package (Gelman et al., 2016). The model predicts length (short or long) based on target duration. From the model outcomes, we estimated the 50% category boundary (between short and long) for each of the four conditions (2 target segments × 2 speaking rates) for each participant.

One issue in estimating category boundaries is the variation in how closely speakers imitate the model speaker's speaking rate. Specifically, one speaker's "slow" might not be as slow as another's, and the same holds for "fast" speech, even in an imitation task. For example, if one speaker's boundary shifts from 100ms to 150ms, while another's shifts from 100ms to 200ms, it may appear that the second speaker shows a stronger rate effect. However, if the first speaker's carrier duration (a proxy for speaking rate) changes from 2000ms at fast to 2500ms at slow, while the second speaker's variation is from 2000ms to 3000ms, the rate effect for the first speaker might be underestimated. To account for this variability in achieving the target speech rates, we adjusted our category boundary estimates to reflect the expected values if the carrier duration matched the duration of the production prompts.

Figure 6a displays the distribution of estimated category boundaries for individual participants after adjusting for individual speech rate variation, separated by target segment and speaking rate in production.

[7] The model was specified in R as: LENGTH~DURATION.

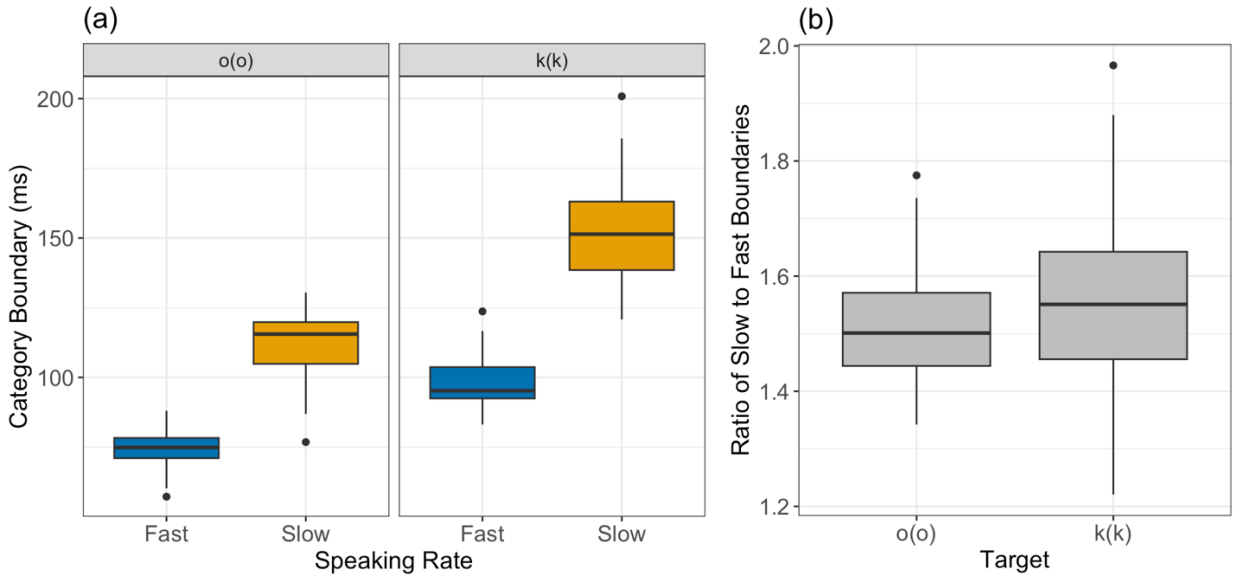


Figure 6. (a) Individual participants' estimated category boundaries by target segment and speaking rate. (b) Slow-to-fast boundary ratios by target segment.

A question arises when comparing the elasticity of segments with differing inherent durations: should we focus on changes in absolute duration or proportion? Earlier studies, such as Peterson and Lehiste (1960) and Klatt (1973), noted that segments with shorter durations exhibit smaller changes in absolute duration compared to those with longer durations. This suggests that comparing changes in absolute duration might overestimate the elasticity of inherently longer segments (e.g., /k(k)/ in this case). It is thus customary to focus on proportional changes when comparing elasticity across different segments (Gay, 1978; Kuwabara, 1996; Port, 1981; Port et al., 1980). Therefore, our analysis focused on proportion, calculating the ratio of category boundaries in fast versus slow speaking rates for each target segment per participant. This approach yields two ratios per participant, reflecting the sensitivity of /o(o)/ and /k(k)/ to speaking rate changes. The larger the ratio, the greater the sensitivity to speaking rate variations. Figure 6b displays these ratio distributions. As shown, the ratios for /o(o)/ are centered around the expected ratio of 1.5, based on the fast (80%) and slow (120%) production prompts (assuming speaking rate affects all segments equally). In contrast, the ratios for /k(k)/ are slightly higher.

We employed a mixed-effects linear regression model[8] to examine whether these target segments differ in their sensitivity to speaking rate (i.e., ratios). The model aimed to predict the ratio between the fast and slow boundaries as a function of target segment (/o(o)/ and /k(k)/). The categorical predictor was contrast-coded, with /o(o)/ mapped to -0.5, /k(k)/ mapped to 0.5. Random effects were specified as random intercepts for participants.

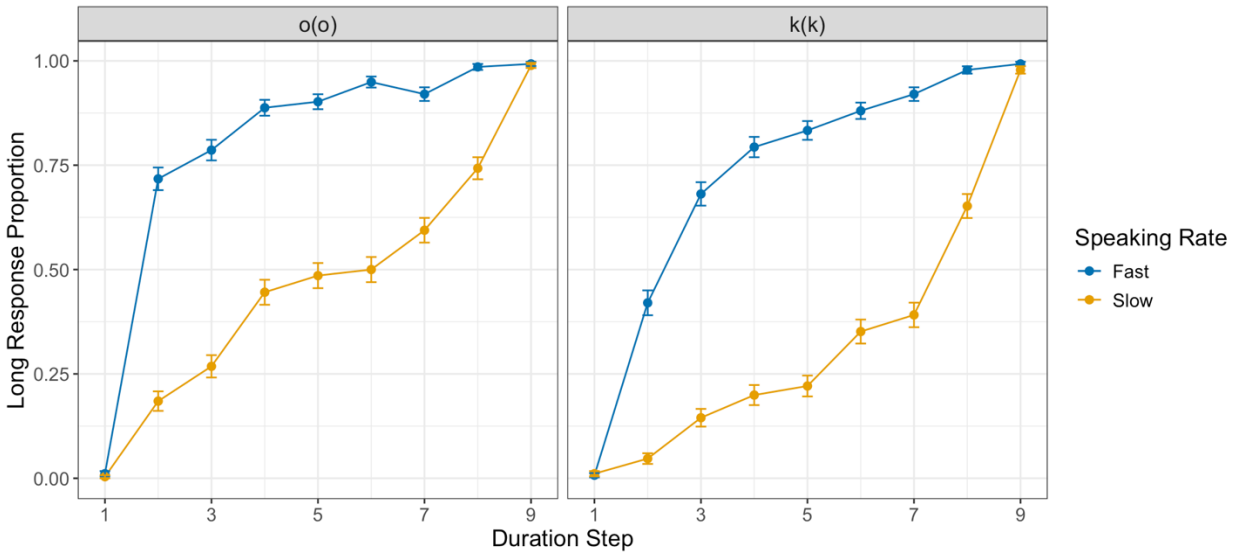
The results indicate that the effect of target was not significant ($\beta = 0.03$, $t = 1.09$, $p = 0.28$), suggesting no evidence of a difference in sensitivity to speaking rate changes between the two

[8] The model was specified in R as: RATIO~TARGET+(1|PARTICIPANT).

segments, /o(o)/ and /k(k)/, in production. Although there was a non-significant tendency for the ratios for /k(k)/ to be slightly larger than those for /o(o)/, as indicated by the positive coefficient, this difference was not statistically significant.

3.2.2. Perception

Figure 7 presents proportion of listeners' long phoneme responses across nine duration steps, separated by target segment (/o(o)/ on the left, /k(k)/ on the right). These results were analyzed using a logistic mixed-effects regression model[9] implemented in R with the *lme4* package (Bates et al, 2015). The model predicted listeners' responses (short or long) from duration step (ranging from 1 to 9), speaking rate (fast or slow), target segment (/o(o)/ and /k(k)/), including the interactions between duration step and rate, duration step and target, and rate and target. The rate × target interaction was included to compare the sensitivity of the two target segments to speaking rate. This approach, unlike in the production analysis, is feasible here because inherent duration differences between the segments were controlled: the duration continua were constructed based on pretest results. The dependent variable was coded such that a short response was assigned a value of 0, while a long response was assigned a value of 1. The duration step was centered, with -4 representing the shortest step and 4 representing the longest step. Speaking rate was contrast-coded, with fast coded as -0.5 and slow coded as 0.5. The target was contrast-coded, with /o(o)/ coded as -0.5 and /k(k)/ coded as 0.5. The random effects were maximally specified including random intercepts for participants, with by-participant random slopes for all fixed effects.



[9] The model was specified in R as: RESPONSE~DURATION.STEP*RATE+DURATION.STEP*TARGET+RATE*TARGET (DURATION.STEP*RATE+DURATION.STEP*TARGET+RATE*TARGET|PARTICIPANT).

Figure 7. Proportion of long responses at each duration step split by speaking rate for the /o(o)/ length contrast (left) and /k(k)/ length contrast (right), with error bars representing the standard error.

The results revealed significant effects for all fixed-effects predictors. Duration step, rate, and target each had robust effects (DURATION.STEP: $\beta = 1.57$, $z = 15.30$, $p < 0.001$; RATE: $\beta = -7.11$, $z = -11.37$, $p < 0.001$; Target: $\beta = -1.37$, $z = -5.30$, $p < 0.001$), indicating that longer durations led to more "long" responses, slow speech reduced "long" responses compared to fast speech, and /k(k)/ elicited fewer "long" responses than /o(o)/. The interaction between duration step and rate was also significant ($\beta = -1.03$, $z = -5.32$, $p < 0.001$), suggesting that the effect of duration step was attenuated in slow speech. The interaction between duration step and target ($\beta = 0.12$, $z = 1.97$, $p < 0.05$) indicated a slightly stronger duration effect for /k(k)/ than for /o(o)/. Crucially, the interaction between rate and target was significant ($\beta = -0.81$, $z = -2.48$, $p < 0.05$), showing that the rate effect was more pronounced for /k(k)/. Post-hoc analyses using *emmeans* confirmed that the rate effect was significant for both segments (/o(o)/: $\beta = -6.70$, $z = -10.35$, $p < 0.001$ vs. /k(k)/: $\beta = -7.51$, $z = -11.66$, $p < 0.001$). Full statistical results are provided in Appendix B.

3.3. Discussion

Our results from Experiment 1 confirmed the basic effects of speaking rate on the segments /o(o)/ and /k(k)/ in both production and perception. In production, both target segments are produced with longer durations in slow speech compared to fast speech. Furthermore, in slow speech, long segments were lengthened more than short segments in terms of absolute duration. This finding aligns with earlier observations that segments with shorter durations exhibit smaller changes in absolute duration compared to those with longer durations (Peterson & Lehiste, 1960; Klatt, 1973) and the asymmetrical effects observed in short versus long acoustic cues to phonemic contrasts (e.g., Hirata, 2004; Hirata and Whiton, 2005; Miller & Baer, 1983, Miller, et al., 1986; Picket, et al., 1999). In perception, listeners provided fewer long phoneme responses in slow speech than in fast speech for both target segments, demonstrating a clear case of rate normalization.

Regarding the comparison between /o(o)/ and /k(k)/, our production results indicated no significant difference in their sensitivity to speaking rate. This result is somewhat surprising given previous production studies indicating that vowels are more sensitive to speaking rate changes than consonants (Gay, 1978; Kuwabara, 1996; Lo & Sóskuthy, 2023; Port, 1981; Port et al., 1980). Interestingly, our results exhibited a non-significant tendency for /k(k)/ to be slightly more sensitive than /o(o)/. While it is important not to draw conclusions from null results, one possible explanation for the discrepancy between our results and the literature is that the observed tendency in the literature may not be as general as previously assumed. It could result from the use of specific stimuli in certain languages or the aggregation of a variety of target segments into a broad contrast between vowels and consonants. Expanding the range of target segments in the stimuli may lead to different, more nuanced results.

Our perception results revealed a significant interaction between speaking rate and target, indicating that the effect of speaking rate is greater for /k(k)/ than for /o(o)/. While this aligns with previous findings on Japanese (Gadanidis & Kang, under revision; Kawahara, et al. 2022), it diverges from our production results, which showed only a non-significant trend. This discrepancy suggests that listeners may overestimate the effect of speaking rate on /k(k)/. However, strong conclusions should be avoided based on a single experiment with just one contrast pair. One possibility is that the production task lacked power due to a smaller sample size, as many participants were excluded, whereas the perception task included the full sample. Alternatively, the perception findings may have been shaped by subtle acoustic factors in the stimuli. For example, an anonymous reviewer raised the possibility of a palatalization effect of /sj/ on the following vowel. Whether such an effect exists—and if so, how it impacts the results—remains unclear. Further converging evidence is needed to better understand these findings.

Building on the findings and limitations of Experiment 1, we will extend our stimuli in Experiment 2 to cover a broader range of vowels and consonants. This will help us determine whether more detailed subcategories of segments exhibit variation in sensitivity to speaking rate and whether perception patterns mimic production patterns or are orthogonal to them.

4. Experiment 2

In this experiment, we revised the stimuli of Experiment 1 to include a wider range of target segments, specifically five vowels (/i(i)/, /e(e)/, /a(a)/, /o(o)/, /u(u)/) and five consonants, including two stops (/t(t)/, /k(k)/), one fricative (/s(s)/), and two nasals (/m(m)/, /n(n)/). We investigated whether there are any differences among these segments in their sensitivity to speaking rate changes in production. Additionally, we examined whether the differences, or lack thereof, observed in production are reflected in perception. To also address the methodological challenge of Experiment 1—specifically, the frequent insertion of an AP boundary that led to the exclusion of many participants from the production data—we used nonce words as targets and created a new carrier phrase.

4.1.Methods

4.1.1. Participants

As in Experiment 1, participants were recruited from Crowdworks.jp. A total of 39 self-identified Tokyo Japanese speakers completed the experiment and were compensated for their participation (32 females, mean age: 40, age range: 20-63).

4.1.2. Stimuli

The target words consisted of 10 minimal pairs of nonce words contrasting in segment length (e.g., /kempina/ and /kempiina/) shown in Table 2. These target words were designed to meet the

following five criteria: (i) have at least two moras preceding the target segments (i.e., /kem/) to avoid an initial f0 rise on the target segments, (ii) include a voiceless stop (i.e., /p/) before the target vowels to ensure a clear boundary between the stop and the following vowel, (iii) avoid same place of articulation for consonants separated by a vowel for ease of pronunciation as much as possible (with only /kem**p**o**m**(m)a/ violating this), (iv) avoid vowel devoicing (i.e., a high vowel between two voiceless consonants), and (v) avoid palatalization of target coronals (i.e., /ti/ [tɕi] and /si/ [ɕi]).

Target segment	Target word	Target segment	Target word
/i(i)/	/kempi(i)na/	/t(t)/	/kempot(t)a/
/e(e)/	/kempe(e)na/	/k(k)/	/kempok(k)a/
/a(a)/	/kempa(a)na/	/s(s)/	/kempos(s)a/
/o(o)/	/kempo(o)na/	/n(n)/	/kempon(n)a/
/u(u)/	/kempu(u)na/	/m(m)/	/kempom(m)a/

Table 2. Target stimuli for Experiment 2.

A potential issue with using novel words is the uncertainty regarding the lexical prosody that speakers might apply to them. In the context of Japanese, it is unclear whether speakers would produce these words as accented or unaccented, and if accented, where they place the accent—though some statistical tendencies have been reported in the lexicon (Kubozono, 2006). To control for this variability, we introduced the target words as names of novel bacteria by attaching the suffix /-kin/ 'bacteria', which categorically renders the entire word unaccented.

The target words (i.e., novel words, specifically names of bacteria, with the suffix /-kin/ (e.g., /kempina-kin/)) were embedded in the carrier sentence shown in (3). The length of the carrier sentence was slightly shortened compared to that in Experiment 1 to accommodate the increased number of target words, resulting in a greater number of trials.

(3) /tanakasan-wa kitínto [**target-kin**]-o siɾábe-ta/
Mr. Tanaka-TOP properly [**target-kin**]-OBJ examine-past
"Mr. Tanaka properly examined [**target-bacteria**]"

Perception

The stimulus creation process was similar to Experiment 1, with minor adjustments for the more complex stimuli in Experiment 2. The speech materials were recorded by the same male speaker of Tokyo Japanese as in Experiment 1, with each target word produced five times at a normal speaking rate. Manual annotations were performed on all utterances to segment each sentence into three parts: the pre-target carrier phrase (/tanakasan-wa kitínto/) (carrier 1), the target word (/target-kin/), and the post-target carrier phrase (/wo siɾábe-ta/) (carrier 2). The target word was further divided into subparts as shown in (4).

(4) Subparts of target words

- For all target words
  - /kem/
  - Stop closure of /p/
  - Aspiration of /p/
  - /a/
  - /kin/
- Additional subparts for vowel targets (e.g., /kemp**i**na-kin/)
  - Target vowel
  - /n/
- Additional subparts for consonant targets (e.g., /kempo**t**a-kin/)
  - /o/
  - Target consonant
  - Aspiration (for target /t/ or /k/ only)

From the recorded utterances with the target /kempossa/, one token was selected as the baseline for the surrounding parts (carrier 1, /kem/, stop closure of /p/, /a/, /kin/, carrier 2). Additionally, an utterance for each short segment target was chosen as the baseline for the middle part. These were then spliced together to create the baseline utterance for each of the 10 minimal pairs. As in Experiment 1, these baseline utterances served as the foundation for both the perception stimuli and production prompts.

Each of the spliced baseline utterances was manipulated to create 12-step duration continua for each length contrasts. The duration continuum for each length contrast ranged from 80% of the average duration of the short target to 120% of the average duration of the long target, as produced by the model speaker.[10] The intervals between each step were maintained as equidistant. Subparts of the middle part were adjusted to match the average duration of all productions for the specific length contrast by the model speaker (based on 10 tokens). Subparts of the surrounding part (carrier 1, /kem/, stop closure of /p/, /a/, /kin/, carrier 2) were adjusted to match the average duration of all productions by the model speaker (based on 95 tokens; note that /kempona/ serves as targets for /o/ and /n/). The intensity of the stimuli was scaled to 70dB.

As in Experiment 1, a pretest was conducted to determine the ranges and steps for the duration continua used in the main experiment. Fifteen native speakers of Tokyo Japanese, who did not participate in the main experiment, completed an identification task remotely and were compensated for their participation. Unlike Experiment 1, the trials from different length contrasts were randomized rather than blocked, with the constraint that all unique tokens were presented once before any second repetitions. Each unique stimulus was categorized twice, resulting in a total of 240 trials (12 steps × 10 contrasts × 2 repetitions).

[10] A pilot study revealed the average durations were not sufficiently short or long enough to cover the range for some length contrasts.

Figure 8 presents the results of the pretest, shown as solid lines with circles. Segment duration (horizontal axis) significantly influences categorization responses (vertical axis) across all contrasts.

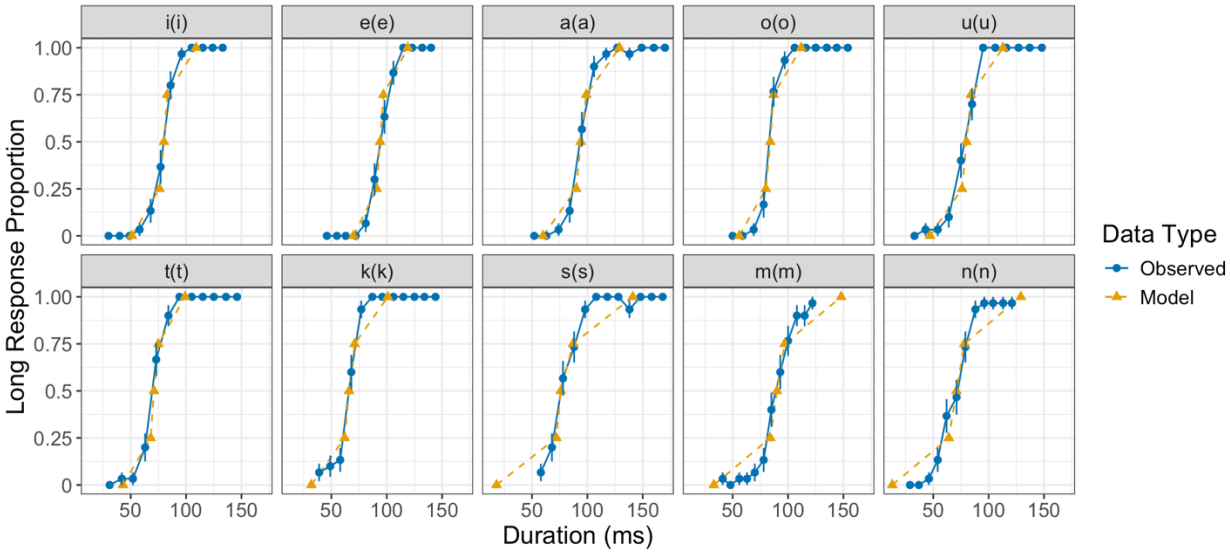


Figure 8. Proportion of long responses at each duration step for all 10 contrasts. Error bars indicate the standard error. Solid lines with circles represent the raw data, while dashed lines with triangles show duration steps estimated by the Bayesian mixed-effects logistic regression models.

As in Experiment 1, the ranges and steps of the duration continua in the main experiment were determined based on model-estimated probabilities of long responses. For each of the 10 contrasts, we fitted the same logistic mixed-effects regression model[11] used in the Experiment 1 pretest using the *brm* function in the *brms* package (Bürkner, 2016).[12] From the model, we derived estimated durations corresponding to five levels of long-response probability, as shown in Table 3. These estimates, shown as dashed lines with triangles in Figure 8, formed the duration steps in each length contrast continuum.

[11] The model was specified in R as: RESPONSE~DURATION.CENTERED+(DURATION.CENTERED|PARTICIPANT).
[12] The Bayesian model was employed because the original frequentist model failed to converge for the /n(n)/ contrast.

Probability of perceiving long	Duration (ms)									
	/i(i)/	/e(e)/	/a(a)/	/o(o)/	/u(u)/	/t(t)/	/k(k)/	/s(s)/	/m(m)/	/n(n)/
0.0001	51	70	60	56	47	43	32	18	33	13
0.25	76	91	90	80	76	68	62	72	84	64
0.5	80	94	94	84	80	71	66	79	90	71
0.75	83	97	99	87	84	75	71	87	97	78
0.9999	109	119	129	112	113	99	101	141	148	129

Table 3. Duration steps estimated by the Bayesian mixed-effects logistic regression models.

The baseline utterance was adjusted to incorporate the new duration steps for each contrast and then manipulated to create two speaking rate conditions by modifying the duration of the carrier sentence components (i.e., carrier 1 and carrier 2) to 80% of their original duration for the fast condition and 120% for the slow condition. As in Experiment 1, the durations of the non-target parts in the target word (from /kem/ to /kin/) were kept constant across both speaking rate conditions.

Production

The production prompts were derived from the baseline stimuli for /kempoka/ created during the process of developing the perception stimuli. As in Experiment 1, the target word (from /kem/ to /kin/) was beeped out. The duration of the beep was adjusted to match the model speaker's average duration of the corresponding part for the short targets (based on 45 tokens) in the prompt for short targets, and the model speaker's average duration of the corresponding part for the long targets (based on 50 tokens) in the prompt for long targets. The duration of the carrier sentence components (i.e., carrier 1 and carrier 2) was adjusted to match the average duration of the model speaker's productions (based on 95 tokens). Additionally, the duration of each stimulus was modified to 80% of the original duration for the fast condition and 120% for the slow condition.

4.1.3. Procedure

The general procedure was identical to that of Experiment 1, with only one minor difference: participants completed the perception and production tasks as separate experiments rather than as two tasks in a single experiment, due to the increased trials of each task. As in Experiment 1, they always completed the perception task first, followed by the production task either immediately after or within a few days.

The procedure for the perception task was identical to that of Experiment 1, except that the duration continua comprised five steps instead of nine, and each unique stimulus was presented three times instead of four. Consequently, the perception task involved a total of 300 trials (5 steps × 10 contrasts × 2 speaking rates × 3 repetitions).

The procedure for the production task was also similar to that of Experiment 1, with three exceptions. First, participants were instructed to produce each sentence once instead of twice due

to the increased number of trials. They were explicitly instructed to take a breath after pressing the recording button and before starting the production, as well as after finishing the production and before pressing the stop button, to ensure the entire production of each sentence is recorded. Additionally, they were instructed to repeat the sentence if they make a mistake. Second, each participant produced each unique stimulus four times, resulting in a total of 152 tokens (19 target words × 2 rates × 4 repetitions). Finally, unlike Experiment 1, participants did not listen to a model sentence at a normal speaking rate because the insertion of an AP boundary was not an issue in Experiment 2.

Both the perception task, including the questionnaire, and the production task took approximately 30 minutes to complete.

4.1.4. Acoustic measurement and exclusion

The acoustic measurement process was mostly identical to that of Experiment 1. However, in Experiment 2, each sentence was usually produced only once, not twice. If participants followed the instructions to repeat the sentence if they made a mistake, the second production was annotated and included in the data.

Out of 39 participants, nine participants were excluded from the production data for the following reasons: seven had poor sound quality, one produced target word with unnatural intonation, and one had a recording issue. Consequently, a total of 30 participants (24 females, mean age: 41, age range: 20-63) remained for data analysis for production.

Within the productions of the included participants, 28 utterances had to be excluded for the following reasons: failure to record (9 cases), producing the target segment too weak to be measured (6 cases), making mistakes or disfluency either in the target or carrier sentence (5 cases), stop recording before finishing the sentence (4 cases), background noise affecting the recording (3 cases), and having a recording issue (1 case). After excluding these cases, the dataset was reduced to 4,532 utterances out of the original 4,560 for data analysis. This translates to 4,771 tokens out of 4,800 tokens (note that utterances containing /kempona/ involve two target tokens, i.e., /o/ and /n/, while those containing the other target segments involve only one).

No participants or responses were excluded from the perception data. Consequently, a total of 11,700 responses from 39 participants were analyzed.

4.2.Results

As in Experiment 1, we begin by analyzing the production data (Section 4.2.1), followed by the perception data (Section 4.2.2). For the production analysis, we first examine each target segment individually, then compare across segments to assess differences in their sensitivity to changes in speaking rate. The perception data are analyzed in a single step that addresses both of these goals.

4.2.1. Production

Figure 9a displays the duration of the carrier sentence in both the fast and slow speech conditions, averaged by speaker. The duration of the carrier sentence includes carrier 1 (/tanakasan-wa kitínto/) and carrier 2 (/wo siɾábe-ta//). As depicted in the figure, the carrier sentence was shorter in the fast speech condition (mean = 1570ms) compared to the slow condition (mean = 2137ms). These durations are comparable to the durations of the corresponding parts in the production prompts (fast = 1500ms, slow = 2250ms), indicated by cross marks. One speaker (F29) produced sentences excessively slowly in the slow condition (mean = 3352ms). However, since this variation is controlled for in the calculation of category boundaries, data from this participant was not excluded.

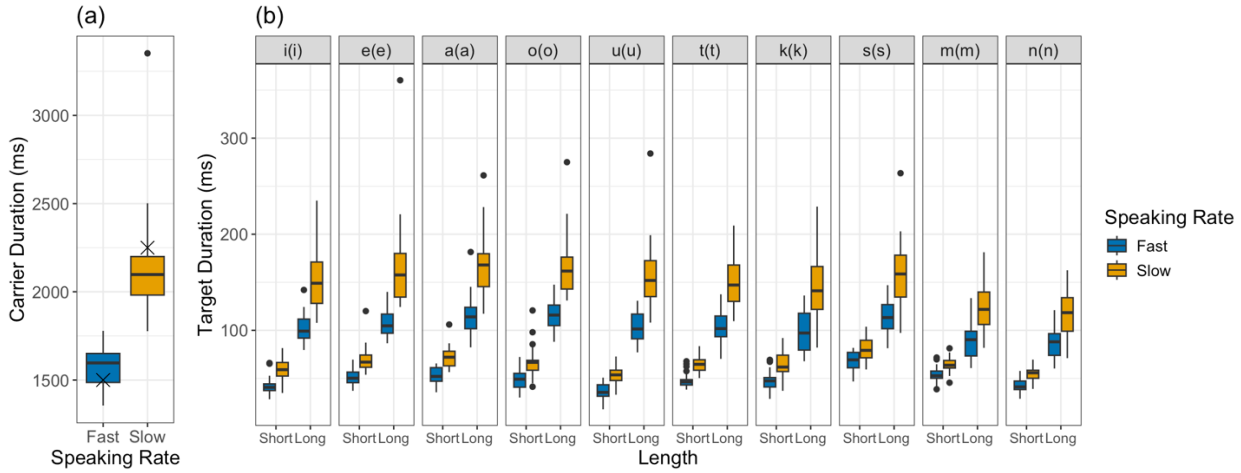


Figure 9. (a) By-speaker average carrier duration split by speaking rate. (b) by-speaker average target duration split by length, segment type, and speaking rate.

Figure 9b displays the average durations of target segments by participant in both fast and slow speech conditions. Note that the average durations longer than 250ms were all produced by F29. To analyze the raw production data, we utilized the same linear mixed-effects regression model[13] as in Experiment 1 for each contrast separately, with the same variable coding schemes (LENGTH: short = -0.5, long = 0.5; RATE: fast = -0.5, slow = 0.5).

We observed significant effects for all fixed-effects predictors across all contrasts. Specifically, phonemically long segments were significantly longer than short segments, segments in the slow condition were significantly longer than those in the fast condition, and phonemically long segments lengthened more than short segments in the slow condition in terms of absolute duration. Post-hoc analyses using *emmeans* confirmed that the effect of rate was significant for both short and long segments, with long segments exhibiting greater estimates of the rate effect for all contrasts. As in Experiment 1, we also fit a model using scaled continuous sentence duration

[13] The model was specified in R as: DURATION~LENGTH*RATE+(LENGTH*RATE|PARTICIPANT).

instead of the binary rate factor, and the results remained consistent. Detailed statistical results are provided in Appendix C.

As in Experiment 1, we examined whether these segments differ in their sensitivity to speaking rate changes by estimating individual speakers' category boundaries between short and long for each speaking rate and each target segment. We applied the same Bayesian logistic regression model[14] as in Experiment 1, and from the model outcomes, we estimated the 50% category boundary, adjusted to account for speaking rate variation, for each of the 20 conditions (10 target segments × 2 speaking rates) for each participant.

Figure 10a displays the distribution of estimated category boundaries for individual participants, separated by target segment and speaking rate, while Figure 10b presents the ratio distributions. The descriptive pattern indicates that the median values for the vowels (/i(i), e(e), a(a), o(o), u(u)/) and stops (/t(t), k(k)/) are higher than the expected ratio of 1.5, while those for the fricative (/s(s)/) and nasals (/m(m), n(n)/) are smaller than 1.5. To examine whether these target segments differ in their sensitivity to speaking rate, we used the same mixed-effects linear regression model[15] as in Experiment 1. We then conducted planned pairwise comparisons of the TARGET levels using *emmeans*, without applying p-value corrections. Detailed results are presented in Appendix D.

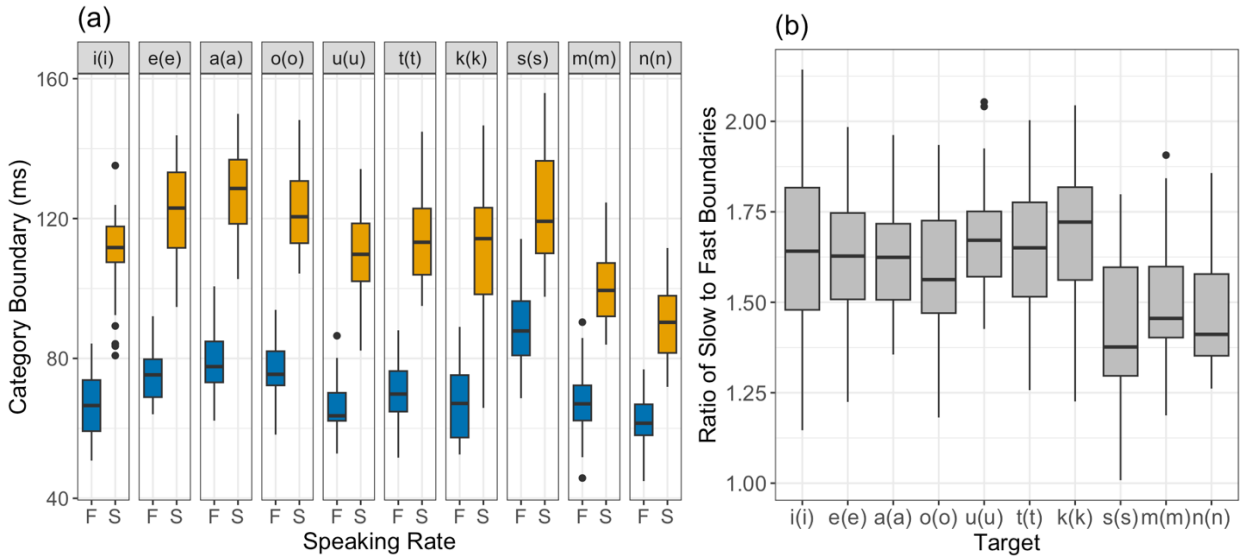


Figure 10. (a) Individual participants' estimated category boundaries by target segment and speaking rate. (b) Slow-to-fast boundary ratios by target segment.

The results of the pairwise comparisons indicate significant differences between /i(i), e(e), a(a), o(o), u(u), t(t), k(k)/ and /s(s), m(m), n(n)/, showing that the effect of speaking rate is greater for the former group, except for the marginally significant difference between /o(o)/ and /m(m)/. Additionally, significant differences were found between /u(u), k(k)/ and /o(o)/, indicating that the

[14] The model was specified in R as: LENGTH~DURATION.
[15] The model was specified in R as: RATIO~TARGET+(1|PARTICIPANT).

effect of speaking rate is greater for /u(u)/ and /k(k)/ than for /o(o)/. Overall, the evidence suggests a distinction between vowels (/i(i), e(e), a(a), o(o), u(u)/) and stops (/t(t), k(k)/) versus the fricative (/s(s)/) and nasals (/m(m), n(n)/), with vowels and stops being more influenced by changes in speaking rate, though there may be additional subtle distinctions within these categories.

4.2.2. Perception

Figure 11 presents the proportion of listeners' long phoneme responses at each duration step for each contrast. The data was analyzed using a logistic mixed-effects regression model[16] predicting the likelihood of "long" response. Fixed effects included duration step, speaking rate, and target segments, along with their two-way interactions: duration step × rate, duration step × target, and target × rate. Coding schemes were as follows: RESPONSE was coded as short = 0 and long = 1; DURATION.STEP was centered from -2 to 2; RATE was coded as fast = -0.5 and slow = 0.5; TARGET was dummy coded with /a/ as the reference level. Due to convergence issues, the random-effects structure was simplified to include by-participant intercepts and by-participant slopes for duration step, rate, and target.

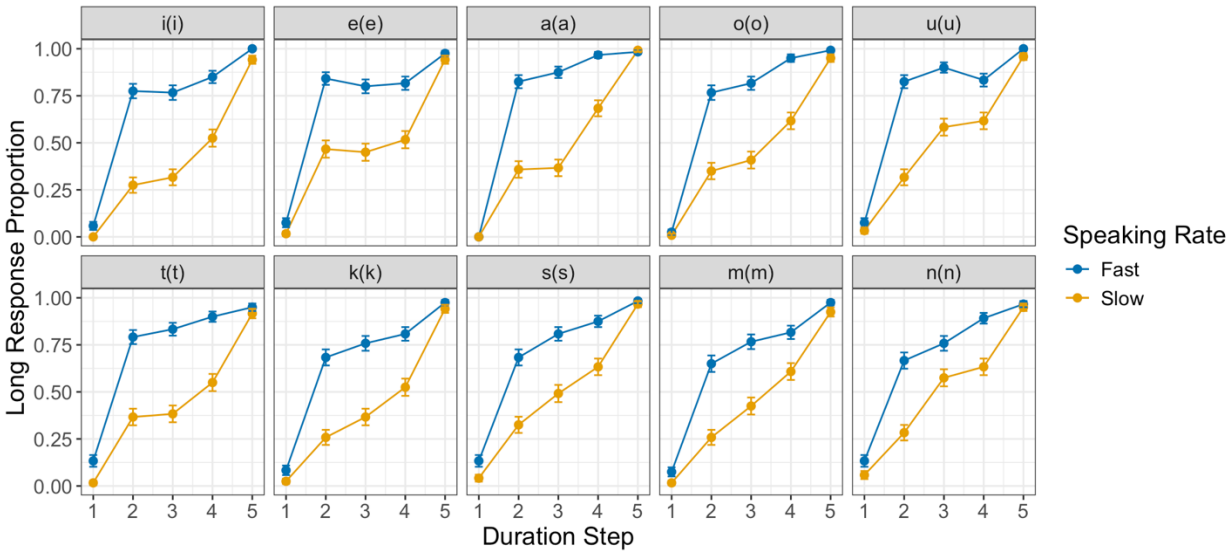


Figure 11. Proportion of long responses at each duration step split by speaking rate for the 10 length contrasts.

Wald chi-square tests implemented using the *Annova()* function in the *car* package (Fox & Weisberg, 2018) revealed significant effects for all fixed-effects predictors: DURATION.STEP: χ^2 = 715.86, df=1, $p < 0.001$; RATE: $\chi^2 = 203.91$, df=1, $p < 0.001$; TARGET: χ^2 =29.64, df=9, $p < 0.001$; DURATION.STEP * RATE: $\chi^2 = 21.64$, df=1, $p < 0.001$; DURATION.STEP * TARGET: $\chi^2 = 38.14$, df=9,

[16] The model was specified in R as: RESPONSE~DURATION.STEP*RATE+DURATION.STEP*TARGET+RATE*TARGET (DURATION.STEP+RATE+TARGET|PARTICIPANT).

$p < 0.001$; RATE * TARGET: $\chi^2 = 27.12$, df=9, $p < 0.005$). The significant interaction of RATE and TARGET was followed up by post-hoc tests using *emmeans*, which revealed a significant effect of RATE ($p < 0.001$) for each individual target segment. Full results are provided in Appendix E.

As in the production analysis, we conducted planned pairwise comparisons of RATE effect between different TARGET levels using *emmeans*, without applying p-value corrections. The comparisons revealed significant differences between /i(i)/, /a(a)/, /o(o)/ and /s(s)/, /m(m)/, /n(n)/, with the former group showing a greater effect of speaking rate, except the differences between /i(i)/ and /m(m)/ and /o(o) and /m(m)/ were only marginally significant. Additionally, /t(t)/ showed a significantly greater rate effect than /n(n)/, and a marginally greater effect than /s(s)/. Among vowels and stops, /a(a)/ was significantly more influenced by speaking rate than /e(e)/, /u(u)/, and /k(k)/. Furthermore, /i(i)/ and /o(o)/ showed significantly greater effects than /e(e)/, and /i(i)/ was also more affected than /k(k)/. While these results were somewhat less conclusive than the production findings, they still suggest that /s(s)/, /m(m)/, and especially /n(n)/ were less influenced by speaking rate, and that there is no clear distinction among vowels and stops—broadly aligning with the rate sensitivity patterns in production. Detailed statistical results are presented in Appendix E.

4.3. Discussion

Our results from Experiment 2 confirmed the fundamental effects of speaking rate on all 10 target segments tested, in both production and perception. In production, all target segments were produced with longer durations in slow speech compared to fast speech, and long segments were lengthened more than short segments in terms of absolute duration. In perception, listeners provided fewer long phoneme responses in slow speech than in fast speech for all target segments. These findings are consistent with Experiment 1 and align with previous studies.

Crucially, the comparisons among the target segments in production revealed a general distinction between vowels (/i(i), e(e), a(a), o(o), u(u)/) and stops (/t(t), k(k)/) on one hand, and the fricative (/s(s)/) and nasals (/m(m), n(n)/) on the other, with the former group generally exhibiting greater sensitivity to speaking rate changes than the latter. These results suggest that the broad distinction between vowels and consonants does not fully capture the variations within each group or the lack of differences across groups regarding sensitivity to speaking rate changes. Specifically, we found differences between stops versus fricatives and nasals, but no consistent differences between vowels and stops (while /a(a)/ and /i(i)/ differed significantly from /k(k)/, significant differences also emerged within the vowel category, such as between /a(a)/ and /e(e)/ or /u(u)/). Overall, these findings indicate that the general tendency observed in the literature is not universal and may be influenced by specific stimuli in particular languages or broad comparisons between vowels and consonants that overlook finer details.

The perception data revealed a similar, though statistically weaker, distinction between vowels and stops versus fricatives and nasals. Pairwise comparisons of the rate effect showed that /i(i)/, /a(a)/, and /o(o)/ exhibited significantly greater sensitivity to speaking rate than /s(s)/, /m(m)/,

and /n(n)/, though the differences between /i(i)/ and /m(m)/ and between /o(o) and /m(m)/ were only marginally significant. Additionally, /t(t)/ showed a significantly greater rate effect than /n(n)/, and a marginally greater effect than /s(s)/. Therefore, it can be concluded that the perception results in Experiment 2 broadly reflect the production patterns.

5. General discussion

5.1. Production

The production results from our experiments provided converging evidence that vowels and stops exhibit similar sensitivity to changes in speaking rate (with the exception that /k(k)/ and /u(u)/ were more affected than /o(o)/ in Experiment 2), while stops were more sensitive than fricatives and nasals. These findings challenge the commonly held view that vowels are more responsive to speaking rate changes than consonants, suggesting instead that this pattern may stem from limited stimuli in specific languages or from overlooking differences among consonant subcategories.

5.1.1. Why do vowels and stops differ from fricatives and nasals?

What accounts for the difference between vowels and stops, on the one hand, and fricatives and nasals, on the other? Lo and Sóskuthy (2023) suggest that the increased aerodynamic and coordinative complexities involved in constricting airflow for consonants partly explains why consonants are less responsive to changes in articulation rate than vowels in their study. We follow this reasoning but also propose that different degrees of articulatory complexity among different consonants may lead to differing responses to speaking rate changes. Specifically, fricatives and nasals may involve more articulatory complexity than stops, especially voiceless stops[17], limiting their flexibility in responding to rate changes.

Given that the limited stretchability of certain consonants under speech variation likely reflects general articulatory challenges in lengthening or shortening them, we might expect a similar asymmetry to arise in the durational distinction between singleton and geminate consonants. Many studies have explored the durational properties of singleton and geminate consonants across languages, typically reporting geminate-to-singleton (SG) ratios (the mean duration of geminates divided by the mean duration of singletons). Indeed, studies on Japanese singletons and geminates have reported consistent findings. Kawahara (2015) notes that SG ratios are higher for stops than for fricatives, suggesting this may be because singleton fricatives tend to be longer than singleton stops in Japanese (Beckman, 1982; Port, Dalby, & O'Dell, 1987), as in other languages (Lehiste, 1970). Note that this trend is replicated in our data as well: in the fast condition of Experiment 2, the average duration of /s/ was 68.38ms, compared to 47.71ms for /t/ and 47.84ms for /k/.

[17] Voiced obstruent geminates are considered difficult to produce due to aerodynamic reasons (Hayes & Steriade, 2004). Indeed, they are absent from native Japanese words and are phonologically marked (see Kawahara, 2005 for a detailed discussion).

Furthermore, Sano (2018) examined spontaneous speech in a corpus and found that the SG ratio varies according to the sonority hierarchy: it is highest for stops, followed by affricates and fricatives, and lowest for nasals.[18] Sano argues that this ordering of SG ratios may be explained by the fact that more sonorous consonants are more marked for gemination, both in production and perception. Sonorous consonants are less effective at signaling length contrasts because their boundaries are harder to perceive precisely (Kawahara & Pangilinan, 2017), and they are also more challenging to geminate due to articulatory challenges. Sano also suggests some functional reasons, which will be discussed in Section 5.1.2. While we did not observe any differences between fricatives and nasals, our results are generally consistent with existing findings on SG ratios.

Additionally, the observed phonetic patterns have phonological consequences. The marked status of fricative and nasal gemination is reflected in the phenomenon of emphatic gemination in mimetic words in Japanese. Typically, the second consonant is geminated in these words to indicate emphasis (e.g., pika-pika → pikka-pika 'shiny'). However, when the second consonant is a voiced stop, the third consonant may be geminated instead, due to a dispreference for voiced geminates (e.g., sube-sube → subes-sube 'smooth') (Nasu, 1999). Crucially, Kawahara (2013), in a nonce word experiment, shows that Japanese speakers most prefer voiceless stop geminates, followed by fricative geminates, and least of all, nasal geminates.

We also note that the limited stretchability of the fricative /s/ and the nasals /m/ and /n/ may stem from different underlying factors. As Kawahara (2013; 2015) observed in his analysis of SG ratios, the singleton /s/ in our study was also longer than other singleton segments (see Figure 9b), which may account for the smaller slow-to-fast ratio observed for /s/. In contrast, the reduced slow-to-fast ratios for nasals appear to result primarily from the relatively shorter durations of their geminate forms. In Experiment 2, the average durations of /mm/ and /nn/ in the slow condition were 125.79ms and 118.17ms, respectively, compared to 150.91ms for /tt/, 147.39ms for /kk/, and 157.92ms for /ss/. These findings suggest that the articulatory complexity of fricatives makes them harder to shorten, while the complexity of nasals limits their ability to lengthen.

5.1.2. Why do our results differ from previous findings?

Our findings differ from earlier studies comparing vowels and stops. Specifically, Gay (1978) and Port (1981) showed that vowels are more sensitive to speaking rate changes than stops in English, and Port et al. (1980) found similar results in Arabic. In contrast, our study found little differences between vowels and stops in Japanese. Notably, we observed a stronger effect of speaking rate for a stop than for a vowel in at least one pair (/k/ vs. /o/) in Experiment 2. We propose four potential explanations for this discrepancy. (We do not consider the findings of Kuwabara (1996) and Lo and Sóskuthy (2023), as their conclusions are based on analyses that collapse across subcategories).

[18] A similar pattern is observed in Kelantan Malay (Hamzah et al., 2016).

First, it is possible that the presence or absence of phonemic length influences how segments respond to changes in speaking rate. Specifically, in languages without length distinctions, vowels might be more influenced by speaking rate changes than stops (for reasons yet to be fully understood), as shown in the previous studies on English (Gay, 1978; Port, 1981). However, in languages like Japanese, where both vowels and stops have length distinctions, stops may become as responsive to speaking rate changes as vowels in order to avoid confusion between singleton and geminate stops.[19] Note that Port et al. (1980)'s finding that vowels stretch more than consonants by speaking rate in Arabic may seem like a counterexample to this hypothesis, given that Arabic contrasts length in both vowels and consonants. However, their stimuli included short and long vowels (/a/ and /aa/) and singleton consonants (/t/, /d/, and /r/), but no geminate consonants, making it unclear whether the observed difference stems from segment type or an imbalance in the stimulus design.

Somewhat relatedly, it has been observed that the duration differences between length distinctions are more consistently preserved when the distinction relies more heavily on duration among a range of phonetic correlates. Specifically, Engstrand & Krull (1994) examined conversational speech in Finnish, Estonian, and Swedish, and found that Finnish and Estonian, whose quantity (length) contrasts are primarily correlated with duration cues, maintain these duration differences more consistently than Swedish, whose quantity contrasts are also correlated with vowel quality or diphthongization. While this finding is not directly related to speaking rate effects, it highlights that even subtle differences in the phonetic correlates of length distinctions can influence how segments respond to speech variation. Given this, it is plausible a more salient factor—the presence or absence of length distinction—may influence how segments respond to variation in speaking rate. It would be valuable to compare the effects of speaking rate on segments across languages that contrast length in both vowels and consonants, in only vowels or consonants, and in languages without such contrasts.

Second, it is also possible that the degree of stretchability is influenced by communication-related factors. Relevant findings can be seen in Sano's (2018) study on the SG ratio, discussed in Section 5.1.1. Sano examined the SG ratio in Japanese spontaneous speech and found that differing ratios across consonants can be partially explained by the informativity of the contrast, as quantified by entropy values, as well as by functional load determined by the presence or absence of minimal pairs. More specifically, a greater phonetic difference between singleton and geminate, leading to a higher SG ratio, is observed when the singleton/geminate contrast is less predictable (i.e., both are equally likely to occur) in a given context, and when a minimal pair for the length

[19] Note that while this hypothesis suggests that length distinctions *increase* the degree of stretchability in response to speaking rate changes, such distinctions can also *limit* the stretchability of otherwise systematic duration alternations. A potentially relevant example is Nakai et al. (2009), who studied Northern Finnish and revealed complex interactions between vowel length contrast and utterance-final lengthening. They found that the second vowel in CVCV words, traditionally considered "half-long", is restricted in terms of utterance-final lengthening to preserve its distinction with double (long) vowels. If utterance-final lengthening can be viewed as a "slowing-down" effect (Cho, 2016), similar to the effect of speaking rate, this would represent a case where the length contrast *limits* contextually conditioned duration variation.

contrast exists. If there is indeed a link between the SG ratio and the slow-to-fast ratio, as discussed in Section 5.1.1, we would expect the same functional mechanism to play a role in the speaking rate-based variation as well.

Third, another factor potentially contributing to cross-linguistic differences is the rhythmic properties of languages. Languages can be classified into three types based on their rhythmic properties: stress-timed, syllable-timed, and mora-timed (Nespor, Shukla, & Mehler, 2011, for a review). Japanese is traditionally considered a mora-timed language (Warner & Arai, 2001 for a review of experimental work validating this classification), while English is considered stress-timed. Although the exact nature of this classification remains unclear (e.g., whether it refers to the domain of isochrony in production, the perception of speech timing, or a consequence of different phonological properties), it is reasonable to assume that these rhythmic differences could affect the stretchability of segments.

A key distinction is that in Japanese, certain consonants, a nasal or the first half of a geminate obstruent, can be moraic (e.g., [ge**n**dai] 'modern times', [ki**t**te] 'stamp'), having a duration roughly equivalent to vowels, which are always moraic, or contributing to a similar perceived duration. In contrast, English rhythm is primarily defined by isochrony within the domain of feet, where vowels, typically serving as syllabic nuclei, may play a more prominent role in the rhythmic structure. As a result, vowels in stress-timed languages like English may be more sensitive to changes in speaking rate than stops.[20] To deepen our understanding of potential interactions between speaking rate and linguistic factors, it is essential to conduct studies on diverse languages with varying rhythmic properties.

Finally, another relevant factor may be the prosodic characteristics of languages studied. Unlike Japanese, the languages examined in previous experimental studies—English (Gay, 1978; Port, 1981) and Arabic (Port et al. (1980)—use stress to mark word-level prominence. Recall that Tilsen and Tiede (2023) found stronger correlations between rate measures and the durations of stressed vowels compared to unstressed ones. Indeed, the earlier studies on English and Arabic focused exclusively on stressed vowels. This raises the possibility that the observed rate effects were overestimated due to the prosodic profiles of these languages and the omission of unstressed vowels. Further research is needed to investigate how stress and broader prosodic properties influence sensitivity to changes in speaking rate.

Before concluding this section, we note that our finding that stops were more sensitive to speaking rate changes than fricatives and nasals differ from Tilsen and Tiede's (2023) result, which showed stronger correlations between rate measures and stops than non-stops. While we do not have a definitive explanation for this discrepancy, it may stem from cross-linguistic differences—Tilsen and Tiede examined English—and structural differences between the languages, as discussed in this section. Additionally, Tilsen and Tiede caution that their findings should be interpret carefully, as potential inaccuracies in forced alignment cannot be ruled out. Further

[20] Ham (2001) noted the possibility that SG ratios are potentially larger for mora-timed languages than for syllable-timed languages, although the potential relationship with SG ratios between slow-to-fast rates remain unclear.

experimental research is needed to determine whether a similar asymmetry between stops and non-stops is observed in English.

5.1.3. Limitations of the production study

Our production study has several limitations. In this section, we focus on two that we consider especially important.

First, the comparison between vowels and consonants inherently involves structural differences in terms of prosodic properties, making it difficult to compare them purely on the phonetic grounds. The most fundamental difference lies in their positions within syllable structure: vowels occur as nuclei while consonants usually serve as onsets or codas. Additionally, as discussed in the final paragraph of Section 5.1.2, Japanese is a mora-timed language, where short and long vowels are generally assumed to correspond to one and two moras, respectively, while singleton and geminate consonants are considered mora-less and one mora, respectively (McCawley, 1965; Poser 1984). Since it remains unclear how these structural differences influence how segments respond to speaking rate variation, our results comparing the stretchability of vowels and consonants should be interpreted with caution. In contrast, the observed differences among consonants are more robust. Future research should investigate how structural differences influence the stretchability of segments.

Second, we employed an imitation task rather than instructing participants to speak at a specific rate or using a visual cue to indicate the intended rate. This decision was intended to better align production and perception data by minimizing variability in how participants interpret rate-related instructions, such as "speak slowly" (even visual cues can introduce considerable variability (Bosker 2017)) and by standardizing sentence prosody, as discussed in Section 2.3. However, as a reviewer pointed out, it remains unclear how hearing a production prompt may influence participants' subsequent productions. Although we masked the target words with a beep to avoid any direct influence of the model speaker's production, the potential influence of the beep itself remains unknown. For these reasons, we believe it is important to investigate segmental stretchability using a range of task designs. Providing verbal instructions may be entirely appropriate, especially for studies focused solely on production, and future research should test whether similar effects are observed under such conditions.

5.2. Perception

5.2.1. Differences between Experiment 1 and Experiment 2

Experiment 1 showed a greater effect of speaking rate for /k(k)/ than for /o(o)/, consistent with findings by Gadanidis and Kang (under revision) and Kawahara et al. (2022). In contrast, Experiment 2 revealed a general alignment between perception and production patterns, although the effects of segment type were weaker, and no clear difference emerged between /k(k)/ than for /o(o)/—if anything, /o(o)/ was marginally more sensitive to speaking rate than /k(k)/. The reasons

for this discrepancy remain unclear. However, we note that perception studies are often more susceptible to subtle influences than production studies, and our two experiments differed in several important ways, including the use of real versus nonce words, the segmental makeup of the target items, and the structure of the carrier sentences. Since Experiment 2 was designed to be more comprehensive and interpretable, the following discussion is based primarily on its results. That said, additional perception studies following this line of research are needed to assess the replicability of our findings and to help clarify the source of the observed differences.

5.2.2. Perception-production link

While our perception results were somewhat mixed across the two experiments, they generally mirror the production findings. In particular, Experiment 2 revealed that the asymmetry between vowels and stops on the one hand, and fricatives and nasals on the other, observed in production was also evident in perception, although the effect was statistically weaker. This pattern aligns with the commonly held view that perception patterns tend to reflect production patterns to some extent (Clayards et al., 2008; Newman, 2003).

To better understand the empirical contribution of this study, it is helpful to distinguish between two types of acoustic cues. Schertz and Clare (2019) differentiate "independently informative" cues and "contexualizing" cues, also referred to as "context effects" (Repp, 1982). Independently informative cues directly signal phonemic contrasts, such VOT and f0 in English stop voicing contrasts, whereas contexualizing cues, like speaking rate and coarticulation, are not intrinsically contrastive but still influence phoneme perception. Viewed through this lens, our results suggest that listeners are sensitive to how strongly contexualizing cues (i.e., speaking rate) affect different types of length contrasts (i.e., vowels and stops vs. fricatives and nasals).

A conceptually similar finding was reported by Katsuda and Steffman (2021), who conducted a perception experiment in Japanese. Building on production data showing that final lengthening is greater for unaccented words than for accented words (Seo et al., 2019), they used a forced-choice identification task to test whether this pattern is likewise reflected in perception. In this case, final lengthening serves as a contextualizing cue whose impact varies depending on the contrast (i.e., accented vs. unaccented). Their results showed that listeners required longer vowel durations to perceive a long vowel in unaccented words than in accented ones, suggesting sensitivity to systematic variation in production. The present study contributes to this line of research by providing further evidence that listeners track how contextualizing cues, such as speaking rate, affect different types of phonemic contrasts.

Building on our findings, we may expect that the factors influencing a segment's sensitivity to speaking rate in production also shape its sensitivity in perception. Areas for further investigation include the functional load of length contrasts (i.e., the number of minimal pairs) and the salience of duration as a cue for distinguishing target contrasts. Kang, et al. (to appear) offer a relevant example of the latter. Their study showed that listeners modulate their speaking rate adjustments based on assumed reliability of duration as a cue in the speaker's speech. In Daejeon

Korean, the duration (VOT) contrast between aspirated and lenis stops has been merging among younger and/or female speakers, but remains robust in the speech of older male speakers. Accordingly, listeners exhibited rate normalization only when the speaker was an older male, but not when the speaker was younger or female, despite the acoustic stimuli being otherwise comparable. Future research would benefit from extending this line of investigation to test whether and how listeners track production variability shaped by social factors.

5.2.3. Limitations of the perception study

As in the production study, our perception study also faces limitations in comparing vowels and consonants due to inherent structural differences, as discussed in Section 5.1.3. Specifically, vowels and consonants cannot occur in exactly the same linear position. In our experiments, vowel contrasts always precede consonantal contrasts (e.g., sj[o~oo]kan and sjo[k~kk]an), which presents a general challenge for comparing these segment types in both production and perception. However, a perception-specific issue arises from the inclusion of buffer segments between the rate-manipulated portion of the carrier sentence and the target segment, as illustrated in (5). While the buffer segments were necessary to avoid potential confounds—such as adjacent segment duration serving as a cue for length contrasts—they also introduce the possibility that vowels and consonants are differentially affected by speaking rate, due to their relative distance from the rate-altered context: vowels are closer to the preceding portion, while consonants are closer to the following portion. Since the precise impact of this positional difference is unclear, any interpretation of vowel-consonant differences in the perception results should be made caution, as in the case of the production data.

(5) Perception stimuli in Experiments 1 and 2 (carrier phrases are shown in *italics*, non-rate manipulated parts of the target word are shown in **bold**, target segments are shown in [brackets])
- Experiment 1
  - Vowel target: *takéutisan-wa odájakani sono* **sj**[o~oo]**kan** *to hatuon-sita*
  - Consonant target: *takéutisan-wa odájakani sono* **sjo**[k~kk]**an** *to hatuon-sita*
- Experiment 2
  - Vowel target: *tanakasan-wa kitínto* **kemp**[V~VV]**nakin**-*o sirábe-ta*
  - Consonant target: *tanakasan-wa kitínto* **kempo**[C~CC]**akin**-*o sirábe-ta*

6. Conclusion

This study explored whether and how sensitivity to speaking rate variations differs between vowels and consonants in both production and perception, using Japanese as a test language. Unlike earlier studies, our results indicate a more nuanced distinction between vowels and stops versus fricatives and nasals, with the former group exhibiting greater sensitivity to speaking rate changes. This

challenges the generality of previous findings and suggests that earlier trend could be influenced by the specific stimuli used in particular languages or by overlooking subcategories. Therefore, our findings underscore the importance of further research using carefully designed and controlled stimuli, especially across languages with diverse linguistic and prosodic properties. It is also crucial to examine the relationship between the ratio of long to short segments (SG ratios) and that of slow to fast rates. Additionally, while the production pattern was generally reflected in the perception results, there were mixed results between two experiments, indicating a need for further exploration to understand the generality and mechanism of this phenomenon.

CRediT authorship contribution statement

Hironori Katsuda: Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing.

Yoonjung Kang: Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing - review & editing

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.

Funding

This project was funded by the Social Sciences and Humanities Research Council (SSHRC) (435-2020-0209).

Acknowledgements

We would like to thank the audiences at the Acoustics Week in Canada 2023, the 187th Meeting of Acoustical Society of America, and members of the Phonetics/Phonology Reading Group at the University of Toronto for their helpful comments and suggestions.

References

Amano, S., & Hirata, Y. (2010). Perception and production boundaries between single and geminate stops in Japanese. *The Journal of the Acoustical Society of America*, *128*(4), 2049-2058.

Amano, S., & Hirata, Y. (2015). Perception and production of singleton and geminate stops in Japanese: Implications for the theory of acoustic invariance. *Phonetica*, *72*(1), 43-60.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beckman, J., Helgason, P., McMurray, B., & Ringen, C. (2011). Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of phonetics*, *39*(1), 39-49.

Beckman, M. (1982). Segment duration and the 'mora' in Japanese. *Phonetica*, *39*(2-3), 113-135.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255-309.

Beddor, P. S., Coetzee, A. W., Styler, W., McGowan, K. B., & Boland, J. E. (2018). The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change. *Language*, *94*(4), 931-968.

Boersma, P., and Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.56 http://www.praat.org

Bosker, H. R. (2017). How our own speech rate influences our perception of others. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1225.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*, 1-28.

Cho, T. (2016). Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass*, *10*(3), 120-141.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804-809.

Engstrand, O., & Krull, D. (1994). Durational correlates of quantity in Swedish, Finnish and Estonian: Cross-language evidence for a theory of adaptive dispersion. *Phonetica*, *51*(1-3), 80-91.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 Speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The journal of the Acoustical society of America*, *63*(1), 223-230.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*: Sage publications.

Gopal, H. S. (1990). Effects of speaking rate on the behavior of tense and lax vowel durations. *Journal of Phonetics*, *18*(4), 497-518.

Gadanidis, T. & Kang, Y. (under revision). Speech rate and the perception of consonant and vowel length in Japanese: neural entrainment and episodic memory. https://www.yoonjungkang.com/uploads/1/1/6/2/11625099/gadanidis_kang_under_revision.pdf

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., Zheng, T., & Dorie, V. (2016). arm package, version 1.9-3 [Comupter software]. https://cran.r-project.org/web/packages/arm/

Hamzah, M. H., Fletcher, J., & Hajek, J. (2016). Closure duration as an acoustic correlate of the word-initial singleton/geminate consonant contrast in Kelantan Malay. *Journal of Phonetics*, *58*, 135-151.

Ham, W. (2001). *Phonetic and phonological aspects of geminate timing*. New York: Routledge.

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America*, *96*(1), 73-82.

Hayes, B. & Steriade, D. (2004). Introduction: The phonetic bases of phonological markedness. In Bruce Hayes, Robert Kirchner and Donca Steriade (eds.), *Phonetically based phonology*, 1–33. Cambridge: Cambridge University Press.

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*, 964-988.

Hirata, Y. (2004). Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics*, *32*(4), 565-589.

Hirata, Y., & Amano, S. (2012). Production of single and geminate stops in Japanese three-and four-mora words. *The Journal of the Acoustical Society of America*, *132*(3), 1614-1625.

Hirata, Y., & Lambacher, S. G. (2004). Role of word-external contexts in native speakers' identification of vowel length in Japanese. *Phonetica*, *61*(4), 177-200.

Hirata, Y., & Whiton, J. (2005). Effects of speaking rate on the single/geminate stop distinction in Japanese. *The Journal of the Acoustical Society of America*, *118*(3), 1647-1660.

Homma, Y. (1981). Durational relationship between Japanese stops and vowels. *Journal of Phonetics*, *9*(3), 273-281.

Idemaru, K., & Guion-Anderson, S. (2010). Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica*, *67*(1-2), 25-46.

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, *132*(6), 3950-3964.

Kang, Y., Kung, K., Li, J., Ting, C., & Yeung, J. (2018). Speaking rate variation in English stop production and perception. *Toronto Working Papers in Linguistics*, *40*.

Kang, Y., Yun, S., Ryu, N.-Y. (to appear). merger in progress and speech rate normalization in perception: a case study of Daejeon Korean. *U. Penn Working Papers in Linguistics, Volume 30.2, 2024: Selected Papers from the New Ways of Analyzing Variation (NWAV) 51 conference.*

Katsuda, H., & Steffman, J. (2021). Prominence-boundary interactions in speech perception: Evidence from Japanese vowel length. In *Proceedings of the 1st International Conference on Tone and Intonation* (pp. 200-204). ISCA.

Kawahara, S. (2005). Voicing and geminacy in Japanese: An acoustic and perceptual study. *University of Massachusetts occasional papers in linguistics*, *31*, 87-120.

Kawahara, S. (2013). Emphatic gemination in Japanese mimetic words: A wug-test with auditory stimuli. *Language sciences*, *40*, 24-35.

Kawahara, S. (2015). The phonetics of sokuon, or geminate obstruents. In H. Kubozono (ed.), *Handbook of Japanese phonetics and phonology*, 79-119. Berlin: De Gruyter Mouton.

Kawahara, S., Kato, M., & Idemaru, K. (2022). Speaking rate normalization across different talkers in the perception of Japanese stop and vowel length contrasts. *JASA Express Letters*, *2*(3).

Kawahara, S., & Pangilinan, M. (2017). Spectral continuity, amplitude changes, and perception of length contrasts. In H. Kubozono (ed.), *The phonetics and phonology of geminate consonants*, 13-33. Oxford: Oxford University Press.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of phonetics*, *25*(2), 143-168.

Kinoshita, K., Behne, D.M., Arai, T. 2002. Duration and F0 as perceptual cues to Japanese vowel quantity. In *Proceeding of Seventh International Conference on Spoken Language Processing.*

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *The Journal of the Acoustical Society of America*, *54*(4), 1102-1104.

Kozasa, T. (2005). *An acoustic and perceptual investigation of vowel length in Japanese and Pohnpeian*. University of Hawai'i at Manoa.

Kubozono, H. (2006). Where does loanword prosody come from?: A case study of Japanese loanword accent. *Lingua*, *116*(7), 1140-1170.

Kuwabara, H. (1996). Acoustic properties of phonemes in continuous speech for different speaking rate. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 4, pp. 2435-2438). IEEE.

Kuwabara, H. (1997). Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. In *Fifth European Conference on Speech Communication and Technology*.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge: MIT Press.

Lenth, R. (2020). emmeans: Estimated marginal means, aka least-squares means. R package version 1.5.3. https://CRAN.R-project.org/package=emmeans

Lo, R. Y.-H. & Sóskuthy, M. (2023). Articulation rate in consonants and vowels: results and methodological challenges from a cross-linguistic corpus study. In: Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences (pp. 3206–3210). Guarant International.

Magen, H. S., & Blumstein, S. E. (1993). Effects of speaking rate on the vowel length distinction in Korean. *Journal of Phonetics*, *21*(4), 387-409.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 128.

McCawley, J. D. (1965). *The accentual system of standard Japanese* (Doctoral dissertation, Massachusetts Institute of Technology).

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *The Journal of the Acoustical Society of America*, *73*(5), 1751–1755.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1-3), 106-115.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*, 457-465.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*, 505-512.

Mitterer, H. (2018). The singleton-geminate distinction can be rate dependent: Evidence from Maltese. *Laboratory Phonology*, *9*(1).

Nakai, S., Kunnari, S., Turk, A., Suomi, K., & Ylitalo, R. (2009). Utterance-final lengthening and quantity in Northern Finnish. *Journal of phonetics*, *37*(1), 29-45.

Nagao, K., & de Jong, K. (2007). Perceptual rate normalization in naturally produced rate-varied speech. *The Journal of the Acoustical Society of America*, *121*(5), 2882-2898.

Nasu, A. (1999). Chouhukukei onomatope no kyouchou keitai to yuuhyousei [Emphatic forms of reduplicative mimetics and markedness]. Nihongo/Nihon Bunka Kenkyuu [Japan/Japanese Culture] 9, 13–25.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio*, *2*(2), 203-230.

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. syllable-timed languages. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.). *The Blackwell companion to phonology*. West Sussex, UK: John Wiley & Sons Ltd.

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *The Journal of the Acoustical Society of America*, *113*(5), 2850-2860.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693-703.

Pickett, E. R., Blumstein, S. E., & Burton, M. W. (1999). Effects of speaking rate on the singleton/geminate consonant contrast in Italian. *Phonetica*, *56*(3-4), 135-157.

Port, R. F. (1981). Linguistic timing factors in combination. *The Journal of the Acoustical Society of America*, *69*(1), 262-274.

Port, R. F., Al-Ani, S., & Maeda, S. (1980). Temporal compensation and universal phonetics. *Phonetica*, *37*(4), 235-252.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, *81*(5), 1574-1585.

Poser, W. J. (1984). *The phonetics and phonology of tone and intonation in Japanese* (Doctoral dissertation, Massachusetts Institute of Technology).

Reinisch, E. V. A. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, *37*(6), 1397-1415.

Repp, B. H. (1982). Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychological bulletin*, *92*(1), 81.

Sano, S. I. (2018). Durational contrast in gemination and informativity. *Linguistics Vanguard*, *4*(s2), 20170011.

Schertz, J. L. (2014). *The structure and plasticity of phonetic categories across languages and modalities*. The University of Arizona.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of phonetics*, *52*, 183-204.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(2), e1521.

Seo, J., Kim, S., Kubozono, H., & Cho, T. (2019). Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter?. *The Journal of the Acoustical Society of America*, *146*(3), 1817-1823.

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95-EL101.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074-1095.

Takiguchi, I., Takeyasu, H., & Giriko, M. (2010). Effects of a dynamic F0 on the perceived vowel duration in Japanese. In *Speech Prosody 2010-Fifth International Conference*.

Tiede, M., Espy-Wilson, C. Y., Goldenberg, D., Mitra, V., Nam, H., & Sivaraman, G. (2017). Quantifying kinematic aspects of reduction in a contrasting rate production task. *The Journal of the Acoustical Society of America*, *141*(5_Supplement), 3580-3580.

Tilsen, S. (2022). An informal logic of feedback-based temporal control. *Frontiers in Human Neuroscience*, *16*, 851991.

Tilsen, S., & Tiede, M. (2023). Parameters of unit-based measures of speech rate. *Speech Communication*, *150*, 73-97.

Ting, C., & Kang, Y. (2023a). The effect of habitual speaking rate on speaker-specific processing in English stop voicing perception. *Language and Speech*, 00238309231188078.

Ting, C., & Kang, Y. (2023b). Tracking speaker-specific speaking rate: Habitual vs. local influences on English stop voicing. In: Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences (pp. 545–549). Guarant International.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*, 1284-1301.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, cognition and neuroscience*, *30*(5), 529-543.

Vance, T. J. (2008). *The sounds of Japanese with audio CD*. Cambridge University Press.

Venditti, J. J. (1997). Japanese ToBI labelling guidelines. *Working Papers in Linguistics-Ohio State University Department Of Linguistics*, 127-162.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America*, *92*(2), 723-735.

Warner, N., & Arai, T. (2001). Japanese mora-timing: A review. *Phonetica*, *58*(1-2), 1-25.

Yu, A. C. L. (2019). On the nature of the perception-production link: Individual variability in English sibilant-vowel coarticulation. *Laboratory Phonology*, *10*(1).

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13-29.

Appendix A1. Fixed effects from the TARGET-specific models for the production data in Experiment 1.

TARGET	Predictor	Estimate (*β*)	Std. error	*t*-value	*p*-value
/o(o)/	(Intercept)	90.47	1.56	57.84	< 0.001
	LENGTH	85.89	2.04	42.17	< 0.001
	RATE	26.94	1.20	22.44	< 0.001
	LENGTH:RATE	32.45	1.84	17.60	< 0.001
/k(k)/	(Intercept)	119.81	1.77	67.72	< 0.001
	LENGTH	89.86	2.87	31.36	< 0.001
	RATE	37.49	2.11	17.79	< 0.001
	LENGTH:RATE	37.65	2.57	14.64	< 0.001

Appendix A2. Results of post-hoc tests of the effect of RATE on duration, separated by TARGET and LENGTH using *Emmeans* (production data, Experiment 1).

TARGET	LENGTH	Estimate *(β)*	Std. error	*t*-value	*p*-value
/o(o)/	Short	10.7	0.86	12.50	< 0.001
	Long	43.2	1.96	22.00	< 0.001
/k(k)/	Short	18.7	1.46	12.82	< 0.001
	Long	56.3	3.18	17.74	< 0.001

Appendix A3: Fixed effects from TARGET-specific models for the production data in Experiment 1, using continuous scaled sentence duration (SEN_DUR) instead of the binary RATE factor.

Target	Predictor	Estimate *(β)*	Std. error	*t*-value	*p*-value
/o(o)/	(Intercept)	89.95	1.32	68.37	< 0.001
	LENGTH	85.21	1.82	46.88	< 0.001
	SEN_DUR	14.86	0.51	29.22	< 0.001
	LENGTH:SEN_DUR	17.76	0.86	20.62	< 0.001
/k(k)/	(Intercept)	120.17	1.81	66.27	< 0.001
	LENGTH	91.40	3.20	28.60	< 0.001
	SEN_DUR	21.26	0.85	25.04	< 0.001
	LENGTH:SEN_DUR	21.07	1.27	16.54	< 0.001

Appendix B1: Fixed effects from the model predicting the "long" responses in the perception data from Experiment 1.

Predictor	Estimate *(β)*	Std. error	*z*-value	*p*-value
(Intercept)	0.98	0.20	4.82	< 0.001
DURATION.STEP	1.16	0.05	24.55	< 0.001
RATE	-4.08	0.29	-14.25	< 0.001
TARGET	-1.24	0.23	-5.30	< 0.001
RATE:TARGET	-0.70	0.25	-2.77	< 0.01

Appendix B2: Results of post-hoc tests examining the effect of RATE on "long" responses by TARGET using *Emmeans* (perception data, Experiment 1).

Target	Estimate *(β)*	Std. error	*z*-value	*p*-value
/o(o)/	-3.73	0.320	-11.664	< 0.001
/k(k)/	-4.43	0.306	-14.484	< 0.001

Appendix C1: Fixed effects from the TARGET-specific models for the production data in Experiment 2.

TARGET	Predictor	Estimate *(β)*	Std. error	*t*-value	*p*-value
/i(i)/	(Intercept)	88.571	2.257	39.241	< 0.001
	LENGTH	76.378	3.413	22.376	< 0.001
	RATE	32.517	2.724	11.939	< 0.001
	LENGTH:RATE	31.195	3.291	9.479	< 0.001
/e(e)/	(Intercept)	98.869	2.983	33.14	< 0.001
	LENGTH	76.787	3.848	19.954	< 0.001
	RATE	37.412	4.845	7.722	< 0.001
	LENGTH:RATE	38.125	6.409	5.949	< 0.001
/a(a)/	(Intercept)	102.425	2.524	40.57	< 0.001
	LENGTH	80.358	3.420	23.49	< 0.001
	RATE	37.391	3.167	11.80	< 0.001
	LENGTH:RATE	36.835	4.696	7.84	< 0.001
/o(o)/	(Intercept)	99.851	2.620	38.117	< 0.001
	LENGTH	84.13	2.955	28.475	< 0.001
	RATE	35.14	3.788	9.278	< 0.001
	LENGTH:RATE	33.902	4.052	8.366	< 0.001
/u(u)/	(Intercept)	86.927	2.358	36.87	< 0.001
	LENGTH	83.989	3.722	22.57	< 0.001
	RATE	34.512	3.422	10.09	< 0.001
	LENGTH:RATE	35.621	4.788	7.44	< 0.001
/t(t)/	(Intercept)	91.953	2.188	42.021	< 0.001
	LENGTH	71.109	3.290	21.612	< 0.001

	RATE	32.067	2.180	14.712	< 0.001
	LENGTH:RATE	29.467	3.246	9.079	< 0.001
/k(k)/	(Intercept)	89.457	3.089	28.959	< 0.001
	LENGTH	66.439	4.050	16.405	< 0.001
	RATE	33.257	2.995	11.104	< 0.001
	LENGTH:RATE	32.345	4.155	7.786	< 0.001
/s(s)/	(Intercept)	105.311	2.562	41.10	< 0.001
	LENGTH	60.948	3.506	17.38	< 0.001
	RATE	28.348	3.278	8.64	< 0.001
	LENGTH:RATE	30.846	4.639	6.64	< 0.001
/m(m)/	(Intercept)	83.03	1.973	42.085	< 0.001
	LENGTH	48.122	3.372	14.27	< 0.001
	RATE	24.27	1.969	12.324	< 0.001
	LENGTH:RATE	26.043	3.697	7.044	< 0.001
/n(n)/	(Intercept)	75.534	1.892	39.925	< 0.001
	LENGTH	53.14	3.085	17.226	< 0.001
	RATE	22.26	2.129	10.457	< 0.001
	LENGTH:RATE	19.731	2.800	7.046	< 0.001

Appendix C2: Results of post-hoc tests of the effect of RATE on duration, separated by TARGET and LENGTH using *Emmeans* (production data, Experiment 2).

TARGET	LENGTH	Estimate *(β)*	Std. error	*t*-value	*p*-value
/i(i)/	Short	16.9	1.83	9.249	< 0.001
	Long	48.1	4.11	11.702	< 0.001
/e(e)/	Short	18.4	2.12	8.649	< 0.001
	Long	56.5	7.94	7.116	< 0.001
/a(a)/	Short	19.0	1.59	11.938	< 0.001
	Long	55.8	5.34	10.443	< 0.001
/o(o)/	Short	18.2	2.41	7.557	< 0.001
	Long	52.1	5.58	9.339	< 0.001
/u(u)/	Short	16.7	1.83	9.128	< 0.001
	Long	52.3	5.61	9.319	< 0.001
/t(t)/	Short	17.3	1.62	10.691	< 0.001
	Long	46.8	3.48	13.431	< 0.001
/k(k)/	Short	17.1	1.96	8.733	< 0.001
	Long	49.4	4.77	10.364	< 0.001
/s(s)/	Short	12.9	1.88	6.858	< 0.001
	Long	43.8	5.36	8.170	< 0.001
/m(m)/	Short	11.2	1.23	9.115	< 0.001
	Long	37.3	3.62	10.315	< 0.001
/n(n)/	Short	12.4	1.27	9.751	< 0.001

	Long	32.1	3.37	9.528	< 0.001

Appendix C3: Fixed effects from TARGET-specific models for the production data in Experiment 2, using continuous scaled sentence duration (SEN_DUR) instead of the binary RATE factor.

Target	Predictor	Estimate *(β)*	Std. error	*t*-value	*p*-value
/i(i)/	(Intercept)	88.50	1.63	54.25	< 0.001
	LENGTH	76.32	2.78	27.44	< 0.001
	SEN_DUR	20.41	0.98	20.91	< 0.001
	LENGTH:RATE	19.29	1.58	12.19	< 0.001
/e(e)/	(Intercept)	98.60	1.72	57.18	< 0.001
	LENGTH	76.07	2.56	29.71	< 0.001
	SEN_DUR	22.63	1.02	22.16	< 0.001
	LENGTH:RATE	22.50	1.71	13.13	< 0.001
/a(a)/	(Intercept)	102.85	1.61	63.70	< 0.001
	LENGTH	80.51	2.98	27.05	< 0.001
	SEN_DUR	23.87	0.90	26.54	< 0.001
	LENGTH:RATE	23.17	2.04	11.37	< 0.001
/o(o)/	(Intercept)	99.29	1.56	63.57	< 0.001
	LENGTH	83.64	2.57	32.51	< 0.001
	RATE	22.16	1.22	18.19	< 0.001
	LENGTH:RATE	21.60	1.90	11.37	< 0.001
/u(u)/	(Intercept)	87.38	1.75	49.87	< 0.001
	LENGTH	84.86	2.98	28.45	< 0.001
	SEN_DUR	21.65	0.97	23.64	< 0.001
	LENGTH:RATE	23.10	1.72	13.47	< 0.001
/t(t)/	(Intercept)	92.96	1.71	54.25	< 0.001
	LENGTH	73.54	3.11	23.68	< 0.001
	SEN_DUR	21.79	1.17	18.58	< 0.001
	LENGTH:RATE	21.28	1.81	11.77	< 0.001
/k(k)/	(Intercept)	90.11	2.55	35.34	< 0.001
	LENGTH	68.92	3.85	17.88	< 0.001
	SEN_DUR	21.84	1.23	17.70	< 0.001
	LENGTH:RATE	23.33	1.78	13.08	< 0.001
/s(s)/	(Intercept)	105.52	2.064	51.12	< 0.001
	LENGTH	61.73	3.249	19.00	< 0.001
	SEN_DUR	18.23	1.19	15.37	< 0.001
	LENGTH:RATE	20.12	1.83	10.99	< 0.001
/m(m)/	(Intercept)	83.42	1.53	54.58	< 0.001
	LENGTH	49.07	3.04	16.13	< 0.001
	SEN_DUR	15.71	0.89	17.55	< 0.001
	LENGTH:RATE	17.55	1.76	9.95	< 0.001
/n(n)/	(Intercept)	75.63	1.52	49.88	< 0.001
	LENGTH	53.61	2.96	18.09	< 0.001
	SEN_DUR	13.85	0.79	17.54	< 0.001

LENGTH:RATE	12.86	1.29	9.99	< 0.001

Appendix D: Results of planned pairwise comparisons of category boundary by TARGET using *Emmeans* for the production data in Experiment 2.

Target_pairwise	Estimate *(β)*	Std. error	*t*-value	*p*-value
n - m	-0.034	0.042	-0.798	0.426
n - s	0.044	0.042	1.041	0.299
n - k	-0.214	0.042	-5.087	<.0001
n - t	-0.173	0.042	-4.107	<.0001
n - u	-0.216	0.042	-5.138	<.0001
n - o	-0.113	0.042	-2.682	0.008
n - a	-0.165	0.042	-3.929	<.0001
n - e	-0.161	0.042	-3.822	<.0001
n - i	-0.192	0.042	-4.569	<.0001
m - s	0.077	0.042	1.839	0.067
m - k	-0.180	0.042	-4.289	<.0001
m - t	-0.139	0.042	-3.309	0.001
m - u	-0.182	0.042	-4.340	<.0001
m - o	-0.079	0.042	-1.884	0.061
m - a	-0.132	0.042	-3.131	0.002
m - e	-0.127	0.042	-3.024	0.003
m - i	-0.158	0.042	-3.771	<.0001
s - k	-0.258	0.042	-6.128	<.0001
s - t	-0.216	0.042	-5.148	<.0001
s - u	-0.260	0.042	-6.179	<.0001
s - o	-0.156	0.042	-3.723	<.0001
s - a	-0.209	0.042	-4.970	<.0001
s - e	-0.204	0.042	-4.863	<.0001
s - i	-0.236	0.042	-5.610	<.0001
k - t	0.041	0.042	0.980	0.328
k - u	-0.002	0.042	-0.051	0.959
k - o	0.101	0.042	2.405	0.017
k - a	0.049	0.042	1.158	0.248
k - e	0.053	0.042	1.265	0.207
k - i	0.022	0.042	0.518	0.605
t - u	-0.043	0.042	-1.031	0.303
t - o	0.060	0.042	1.424	0.156
t - a	0.007	0.042	0.178	0.859
t - e	0.012	0.042	0.285	0.776

t - i	-0.019	0.042	-0.462	0.644
u - o	0.103	0.042	2.456	0.015
u - a	0.051	0.042	1.209	0.228
u - e	0.055	0.042	1.316	0.189
u - i	0.024	0.042	0.569	0.570
o - a	-0.052	0.042	-1.247	0.214
o - e	-0.048	0.042	-1.140	0.256
o - i	-0.079	0.042	-1.887	0.060
a - e	0.005	0.042	0.107	0.915
a - i	-0.027	0.042	-0.640	0.523
e - i	-0.031	0.042	-0.747	0.456

Appendix E1: Results of post-hoc tests examining the effect of RATE on "long" responses by TARGET using *Emmeans* (perception data, Experiment 2).

Target	Estimate *(β)*	Std. error	*z*-value	*p*-value
/i(i)/	-2.59	0.25	-10.52	< 0.001
/e(e)/	-1.91	0.23	-8.50	< 0.001
/a(a)/	-2.88	0.27	-10.52	< 0.001
/o(o)/	-2.55	0.26	-10.02	< 0.001
/u(u)/	-2.18	0.24	-9.13	< 0.001
/t(t)/	-2.33	0.23	-10.07	< 0.001
/k(k)/	-2.07	0.23	-8.95	< 0.001
/s(s)/	-1.88	0.23	-8.21	< 0.001
/m(m)/	-2.04	0.24	-8.63	< 0.001
/n(n)/	-1.74	0.23	-7.65	< 0.001

Appendix E2: Results of planned pairwise comparisons of RATE effect by TARGET using *Emmeans* for the perception data in Experiment 2.

Target_pairwise	*β*	Standard error	*t*	*p*
a - e	0.97	0.30	3.23	0.0012
a – i	0.29	0.32	0.93	0.3542
a – k	0.81	0.31	2.65	0.0080
a - m	0.84	0.31	2.72	0.0065
a - n	1.14	0.30	3.76	0.0002
a - o	0.33	0.32	1.02	0.3066
a - s	1.00	0.30	3.31	0.0009
a - t	0.56	0.31	1.82	0.0686
a - u	0.70	0.31	2.25	0.0243

e - i	-0.68	0.28	-2.45	0.0142
e - k	-0.16	0.26	-0.60	0.5457
e - m	-0.13	0.27	-0.48	0.6341
e - n	0.17	0.26	0.64	0.5202
e - o	-0.64	0.28	-2.26	0.0236
e - s	0.03	0.26	0.12	0.9029
e - t	-0.42	0.26	-1.58	0.1136
e - u	-0.27	0.27	-1.02	0.3099
i - k	0.52	0.28	1.84	0.0663
i - m	0.55	0.29	1.92	0.0549
i - n	0.85	0.28	3.03	0.0024
i - o	0.04	0.30	0.12	0.9035
i - s	0.71	0.28	2.54	0.0111
i - t	0.26	0.28	0.93	0.3519
i - u	0.41	0.29	1.41	0.1589
k - m	0.03	0.27	0.12	0.9081
k - n	0.33	0.27	1.23	0.2203
k - o	-0.48	0.29	-1.67	0.0958
k - s	0.19	0.27	0.72	0.4739
k - t	-0.26	0.27	-0.95	0.3403
k - u	-0.11	0.28	-0.41	0.6798
m - n	0.30	0.27	1.09	0.2763
m - o	-0.61	0.29	-1.75	0.0802
m - s	0.16	0.27	0.59	0.5569
m - t	-0.29	0.27	-1.05	0.2921
m - u	-0.15	0.28	-0.52	0.6034
n - o	-0.81	0.29	-2.83	0.0047
n - s	-0.14	0.26	-0.51	0.6072
n - t	-0.58	0.27	-2.20	0.0281
n - u	-0.44	0.27	-1.62	0.1051
o - s	0.67	0.29	2.35	0.0188
o - t	0.23	0.29	0.78	0.4341
o - u	0.37	0.29	1.25	0.2099
s - t	-0.45	0.27	-1.68	0.0923
s - u	-0.31	0.27	-1.12	0.2625
t - u	0.14	0.27	0.52	0.6025