Investigation of the acoustic effects of orthographic consistency on Japanese homophones

Yoichi Mukai (Vancouver Island University), Benjamin V. Tucker (Northern Arizona University)
yoichi.mukai@viu.ca, benjamin.tucker@nau.edu

1 Introduction

Research has claimed that the phonetic realization of word duration in homophone pairs, such as *time* and *thyme*, systematically differs based on their lexical characteristics, including part of speech, frequency, and orthographic consistency. This research has shown that higher orthographic consistency is associated with shorter homophone duration in English (Gahl, 2008). The present study builds on this line of research in two ways. First, we extend it to Japanese, examining whether orthographic consistency effects on homophone duration generalize beyond alphabetic languages (e.g., English). Unlike alphabetic languages, Japanese employs logographic (or morphographic) orthography, in which logographic characters do not reliably correspond to any sub-lexical phonological unit (Wydell, 1998). Second, we investigate two directions of orthographic consistency: spelling-to-sound (feedforward) and sound-to-spelling (feedback) (Ziegler et al., 1997). While previous research has demonstrated effects of sound-to-spelling consistency on word production (e.g., Ziegler et al., 2003), studies on acoustic duration have focused exclusively on spelling-to-sound consistency (Brewer, 2008; Muschalik and Kunter, 2024), overlooking possible effects of sound-to-spelling consistency on word duration.

1.1 The phonetic realization of word duration

The phonetic realization of words is variable, especially in spontaneous, casual speech (Tucker and Mukai, 2023). This phonetic variability is governed by communicative efficiency and probability. Lindblom's Hypo- and Hyper-articulation theory (1990) states that speakers adjust their speech according to situational demands to maximize communicative efficiency. That is, speakers seek the most efficient way to convey a message while ensuring that the listener can understand it. When the message is easily understood, such as when talking to a friend in a quiet room, speakers expend less articulatory effort, producing shorter word and segment durations. When the message is harder to understand, such as when talking to a stranger in a noisy room, they expend more articulatory effort, producing longer word and segment durations.

Relatedly, the Smooth Signal Redundancy Hypothesis (Aylett and Turk, 2004) proposes that less probable elements are articulated more carefully (e.g., longer duration) so that they are more likely to be understood. Crucially, phonetic variability emerges from the effort to keep the flow of information constant across elements. In a similar vein, the Probabilistic Reduction Hypothesis (Jurafsky et al., 2001) proposes that words are shortened when they are more probable: high relative frequency and high conditional probability lead to shorter word duration. This hypothesis fits well with Lindblom's account in that highly probable words are easier for listeners to understand. Gahl (2008) and Lohmann (2018) likewise demonstrate that for homophone pairs such as *time* and *thyme*, the higher-frequency member *time* is produced with shorter duration than the lower-frequency one *thyme*.

1.2 Orthographic effects on spoken word processing

Communication also relies heavily on written texts. As a result, there has been an increasing body of research on orthography and its interactions with speech. Much of the evidence for orthographic effects on spoken word processing comes from studies on orthographic consistency either in spelling-to-sound (feedforward) and sound-to-spelling (feedback). In particular, feedback consistency has been shown to affect the ease with which spoken words are processed. For instance, auditory lexical decisions are slower and less accurate for inconsistent words, whose rhyme can be spelled in multiple ways (e.g.,/-ip/ in *leap* or *keep*) than for consistent words, whose rhyme can be spelled in only one way (e.g., /-ʌk/ in *luck* and *duck*) (Ziegler and Ferrand, 1998; Ziegler et al., 2008).

Importantly, when attempting to pinpoint the locus of such orthographic effects, two hypotheses have been suggested: (1) representations of orthography are co-activated with phonological representations and

(2) orthographic information "contaminates" phonological representations with exposure to written language. The first account hypothesizes that orthographic representations are co-activated with phonological representations online, and these representations are bidirectionally connected, influencing each other incrementally (e.g., Dijkstra et al., 1995). The second account hypothesizes that orthographic information is used to restructure phonological representations, creating 'phonographic' representations: they are more specified representations. They emerge from combining phonological representations with their corresponding orthography during the course of learning to read and write (Muneaux and Ziegler, 2004).

1.3 Orthographic effects on spoken word duration

Research on speech production has measured orthographic effects using varying tasks (e.g., Ziegler et al., 2003; Mukai et al., 2023). However, few studies have extended the investigation to spoken word duration in relation to their corresponding spellings. For example, Warner et al. (2006) reported durational differences for consonants and vowels in Dutch heterographic homophones, such as /heːtən/ being orthographically realized as both <heten> 'to be called' and <heetten> 'were called'. These segments were produced longer in words spelled with a double letter. Similarly, Gahl (2008), as well as Lohmann (2018), found that in heterographic homophone pairs, the orthographically inconsistent member is produced with a longer duration than the consistent one.

However, these studies were not designed specifically to test the influence of orthography on speech production; orthographic effects were covariate. Brewer (2008) focused on the effect of orthography and found that when more letters are used to spell a sound, the sound's duration is longer. Muschalik and Kunter (2024) attempted to replicate Brewer's result with more careful considerations of items using more sophisticated statistical methods and found that orthographic consistency affects segment duration rather than the number of letters: inconsistent mapping leads to longer segment duration. Grippando (2021) showed that Japanese homophone pairs exhibit similar durational differences according to the number of logographic characters: homophonous members represented by two characters, such as 海苔 /noɾi/ 'seaweed' were produced with longer duration than those represented by a single character, such as 糊 /noɾi/ 'glue'. Grippando (2021) also examined the effect of orthographic complexity in terms of the number of pen strokes, but found no comparable effect on word duration.

Importantly, however, these duration studies only looked at the feedforward consistency although previous research has demonstrated orthographic effect in both directions (e.g., Ziegler et al., 2003), overlooking possible effects of feedback consistency on word duration. The present study therefore examines the duration of Japanese homophones to address two questions: First, does orthographic consistency influence the duration of Japanese homophones, and if so, is the direction of the effect comparable to that observed in alphabetic languages? Second, is the orthographic consistency effect bidirectional; that is, do both types of consistency influence the duration of Japanese homophones? If the orthographic consistency effect on homophone duration is a language-general phenomenon, it should also occur in a logographic language, as it does in alphabetic languages. In addition, if orthographic consistency affects the acoustic duration of homophones, then it is necessary to examine consistency in both directions because the degree of consistency may differ between feedforward and feedback mappings, and effects may emerge from both directions.

2 Method

2.1 Corpus data

To obtain Japanese homophonous words, we used the Core component of the Corpus of Spontaneous Japanese (Maekawa, 2003), which consists of approximately 500,000 words and 44 hours of speech by 155 individuals. It is segment-labeled, mostly phonemically. The phonemic labels were generated and time-aligned using a Hidden Markov Model-based algorithm and then adjusted by human labelers. Details of the procedure and evaluation of the segmental labeling can be found at `https://clrd.ninjal.ac.jp/csj/misc/preliminary/labeling.html`. The corpus comprises of four speech styles: The Academic Presentation includes live recordings of various academic talks, whereas the Simulated Public Speech consists of studio recordings on everyday topics, delivered in front of a small friendly audience. The Dialogue component is composed of interviews, task-oriented dialogue, and free conversation. Follow-

ing the Academic Presentation, speakers were asked to read transcriptions of their presentations, producing the Read Speech component.

For the present study, we restricted the data set to: (1) four-mora, two-logograph words to control for the length of morae and logographic characters, (2) nouns to control for part of speech, and (3) spontaneous monologue (i.e., Academic Presentation and Simulated Public Speech) to control for style of speech. As a result, we extracted 176 homophone sets (148 pairs, 22 triples, and 6 quadruples), comprising 385 word types and 3,166 word tokens. We also used another corpus, the Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2014), to obtain lexical information of the homophonous words (e.g., lexical frequency and the number and frequency of their phonological and orthographic neighbours), as it is one of the largest Japanese corpora and covers a wider range of genres.

Following previous research (Hino et al., 2017), we calculated both Phonological-Orthographic (P-O: sound-to-spelling) and Orthographic-Phonological (O-P: spelling-to-sound) consistencies based on the number and frequency of their phonological and orthographic neighbours. As exemplified in Table 1, P-O consistency was calculated as the frequency of the target word plus the summed frequency of its orthographic friends, divided by the frequency of the target word plus the summed frequency of its orthographic friends and enemies. Orthographic friends are defined as phonological neighbours that are also orthographic neighbours, whereas orthographic enemies are defined as phonological neighbours that are not orthographic neighbours of the target word. Similarly, O-P consistency was calculated as the frequency of the target word plus the summed frequency of its phonological friends, divided by the frequency of the target word plus the summed frequency of its phonological friends and enemies. Phonological friends are defined as orthographic neighbours that are also phonological neighbours, whereas phonological enemies are defined as orthographic neighbours that are not phonological neighbours of the target word.

Table 1: Example Japanese P-O consistency index calculation

	Orthography	Phonological Form	Meaning	Frequency	Types of Neighbor
Target word	現在	/ge.N.za.i/	'now'	24	—
Phonological Neighbours	**現**代	/**ge**.**N**.da.**i**/	'modern times'	38	OF (PN and ON)
	存**在**	/so.**N**.**za**.**i**/	'existence'	28	OF (PN and ON)
	犯罪	/ha.**N**.**za**.**i**/	'crime'	35	OE (PN not ON)
	限界	/**ge**.**N**.ka.**i**/	'limit'	31	OE (PN not ON)
Consistency Ind.			0.58		

Note. OF = Orthographic Friend, OE = Orthographic Enemy, PN = Phonological Neighbour, ON = Orthographic Neighbour. Bold text indicates orthographic or phonological elements shared with the target word. The consistency index is calculated as the sum of the target and OF frequencies divided by the total frequency of the target, OFs, and OEs: $(24+38+28)/(24+38+28+35+31)$.

2.2 Analysis

We used linear mixed-effects modeling in the statistical environment R (R Core Team, 2026), using the function *lmer* in the *lme4* package (Bates et al., 2015). P-values were calculated using Satterthwaite's method with the *lmerTest* package (Kuznetsova et al., 2017).

To separate the effect of consistency differences within a homophone set from that of differences between homophone sets, we adapted the approach of Lohmann (2018) and created four new consistency variables: *P-O-between*, *O-P-between*, *P-O-within*, and *O-P-within*. P-O-between and O-P-between are the mean P-O and O-P consistency values of the word types in each homophone set. They are constant across the members of each set and capture consistency effects between sets. P-O-within and O-P-within are each member's own consistency value minus its set mean. They are unique to each homophone and capture consistency effects among the members within each set.

Accordingly, our variables of interest are as follows. The dependent variable is Log-transformed Word Duration. The independent variables are P-O-within and O-P-within. As covariates, we added P-O-between and O-P-between to account for consistency effects between homophone sets. Additionally, we included

variables that previous research has shown to influence word duration: Speech Rate, Log-transformed Frequency, Log-transformed Bigram Probability, and Speech Style. Speech Rate was calculated as the number of morae in the inter-pausal unit (IPU) containing the target homophone (excluding the morae of the target homophone), divided by the duration of that IPU (excluding the duration of the target homophone). IPUs are defined as segments bounded by pauses of 0.2 seconds or longer. There were 15 tokens in which the target homophone (four morae) constituted the entire IPU. These were excluded from the analysis, leaving 2 homophone sets without a pair. These unpaired homophones (5 tokens) were also excluded. The final dataset comprised 3149 tokens, 381 word types, and 174 homophone sets (146 pairs, 22 triples, 6 quadruples). We employed a backward selection procedure for fixed effects and a forward fitting procedure for random effects to fit the optimal model (Matuschek et al., 2017).

3 Results

Our final model includes the following variables: Speech Rate, Log-transformed Frequency, and Speech Style as fixed effects, along with Speaker, Word, and Homophone Set as random effects. Neither of the consistency measures for within homophone set comparisons remained in the final model: P-O-within ($t = 1.234$, $p = 0.219$) and O-P-within ($t = -0.223$, $p = 0.824$). The consistency measures for between homophone set comparisons did not remain in the final model either: P-O-between ($t = -0.368$, $p = 0.713$) and O-P-between ($t = 1.036$, $p = 0.301$). As a result, we found no effect of orthographic consistency, in either the spelling-to-sound or sound-to-spelling direction, or whether measured within or between homophone sets. For the control variables, we observed a large effect of Speech Rate, where faster speech is associated with shorter duration ($t = -17.525$, $p < 0.001$), and a modest effect of Log-transformed Frequency, where higher frequency is associated with shorter duration ($t = -3.078$, $p < 0.01$). For Speech Style, Academic Presentation is associated with shorter duration compared to Simulated Public Speech ($t = 2.356$, $p < 0.05$).

4 Discussions and Conclusions

The present study investigated the effects of orthographic consistency on the acoustic duration of Japanese homophones. Unlike in previous studies, our results revealed no effect of orthographic consistency, in either the spelling-to-sound or sound-to-spelling direction, and whether measured within or between homophone sets. This suggests that such effects may be harder to detect in logographic orthographies, possibly because logographic characters do not reliably correspond to any sub-lexical phonological unit.

One possible explanation for the absence of an effect lies in the distribution and range of consistency in our data set. Figure 1 shows the distribution of P-O and O-P consistency against their phonological and orthographic neighbourhood sizes. The figure reveals the limited range of orthographic neighbourhood sizes. It also shows that consistency values cluster at the lower end of the scale. As a result, the consistency contrast between words is minimal. The contrast is even narrower for within-set comparisons: the median within-set difference is only 0.11 for O-P consistency and 0.07 for P-O consistency, on an index ranging from 0 (low consistency) to 1 (high consistency). A contrast this small may be too subtle for an effect on duration to be detected. A post hoc exploratory analysis restricted to homophone sets whose within-set consistency difference exceeded 0.4 (26 sets, 686 tokens) revealed an effect of O-P consistency ($t = 2.182$, $p < 0.05$), with higher consistency associated with longer duration as shown in Figure 2. For P-O consistency, only 9 sets met this criterion, which is too few to fit the model reliably. This effect of O-P consistency is in line with Ziegler et al. (2003), in which dense phonological neighbourhoods inhibit word production while dense orthographic neighbourhoods facilitate it, and high orthographic density tends to correspond to high P-O consistency. As can be seen in Figure 1, higher O-P consistency tends to be associated with dense phonological neighbourhoods (strong inhibition) and sparse orthographic neighbourhoods (weak facilitation), resulting in overall inhibition. In contrast to our results, Gahl (2008) shows that higher O-P consistency is associated with shorter duration. This reversal may stem from the way consistency is calculated. While Ziegler et al. (2003) and the present study use the same calculation, based on the number and frequency of phonological and orthographic neighbours, Gahl (2008) used the M-score, which averages the probabilities of a word's graphemes, normalized by the probability of the most probable pronunciation of each grapheme.

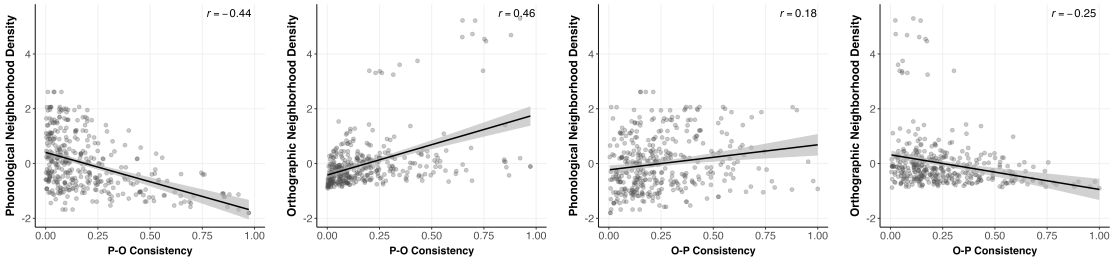

Figure 1: Correlations between P-O and O-P consistency and neighbourhood size (n = 381 word types). Lines show linear fits with 95% confidence intervals; r values are Pearson correlations.

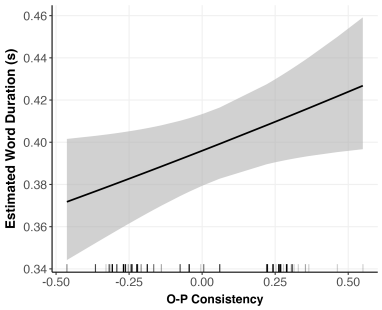

Figure 2: Partial effect of O-P consistency on word duration (within-set comparison) estimated by the post hoc subset analysis.

In conclusion, because we did not find effects of orthographic consistency in the main analysis, the present results do not speak directly to the discussion of how orthography affects speech production, or the locus of those effects. Nevertheless, a substantial body of research shows that speech production is influenced by orthography and its mapping to sound even though orthography is secondary to speech. As discussed in the introduction, our communication relies increasingly on written text, through texting and online chatting, which often include speech-like spellings. At the same time, the ways in which we interact with written text have been changing across generations. These changes are partly driven by technological advances such as text-to-speech and speech-to-text. Such technologies also blur the boundary between the spoken and written modalities: written language is increasingly produced by speaking and consumed by listening. The relationship between orthography and speech may therefore be shifting, in ways that could either strengthen or weaken that influence. This raises the question of whether the effect of orthography on speech processing varies across generations. Investigating this question would shed additional light on how spoken and written language interact.

References

Aylett, M. and Turk, A. (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech". *Language and Speech*, 47:1, 31–56.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). "Fitting linear mixed-effects models using **lme4**". *Journal of Statistical Software*, 67:1.

Brewer, J. B. (2008). *Phonetic reflexes of orthographic characteristics in lexical representation*. PhD thesis, The University of Arizona.

Dijkstra, T., Roelofs, A., and Fieuws, S. (1995). "Orthographic effects on phoneme monitoring". *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 49:2, 264–271.

Gahl, S. (2008). "*Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech". *Language*, 84:3, 474–496.

Grippando, S. (2021). "Japanese orthographic complexity and speech duration in a reading task". *Phonetica*, 78:4, 317–344.

Hino, Y., Kusunose, Y., and Lupker, S. J. (2017). "Phonological-orthographic consistency for Japanese words and its impact on visual and auditory word recognition". *Journal of Experimental Psychology: Human Perception and Performance Human perception and performance*, 43:1, 126–146.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (2001). "Probabilistic relations between words: Evidence from reduction in lexical production". In Bybee, J. and Hopper, P.(eds.), *Frequency and the emergence of linguistic structure*, (pp. 229–254). Amsterdam: John Benjamins.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). "lmerTest package: Tests in linear mixed effects models". *Journal of Statistical Software*, 82, 1–26.

Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H&H theory". In Hardcastle, W. J. and Marchal, A.(eds.), *Speech production and speech modeling*, (pp. 403–440). Dordrecht: Kluwer.

Lohmann, A. (2018). "Time and thyme are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis". *Language*, 94:2, e180–e190.

Maekawa, K. (2003). "Corpus of spontaneous Japanese: Its design and evaluation". In *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), Tokyo*.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). "Balanced corpus of contemporary written Japanese". *Language Resources and Evaluation*, 48:2, 345–371.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). "Balancing Type I error and power in linear mixed models". *Journal of Memory and Language*, 94, 305–315.

Mukai, Y., Järvikivi, J., and Tucker, B. V. (2023). "The role of phonology-to-orthography consistency in predicting the degree of pupil dilation induced in processing reduced and unreduced speech". *Applied Psycholinguistics*, 44:5, 784–815.

Muneaux, M. and Ziegler, J. C. (2004). "Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task". *Language and Cognitive Processes*, 19:5, 641–660.

Muschalik, J. and Kunter, G. (2024). "Do letters matter? The influence of spelling on acoustic duration". *Phonetica*, 81:2, 221–264.

R Core Team (2026). "R: A language and environment for statistical computing".

Tucker, B. V. and Mukai, Y. (2023). *Spontaneous speech*. Elements in Phonetics. Cambridge: Cambidge University Press. Publisher: Cambridge University Press.

Warner, N., Good, E., Jongman, A., and Sereno, J. (2006). "Orthographic vs. morphological incomplete neutralization effects". *Journal of Phonetics*, 34:2, 285–293.

Wydell, T. N. (1998). "What matters in kanji word naming: Consistency, regularity, or On/Kun-reading difference?". *Reading and Writing*, 10:3, 359–373.

Ziegler, J. C. and Ferrand, L. (1998). "Orthography shapes the perception of speech: The consistency effect in auditory word recognition". *Psychonomic Bulletin & Review*, 5:4, 683–689.

Ziegler, J. C., Muneaux, M., and Grainger, J. (2003). "Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation". *Journal of Memory and Language*, 48:4, 779–793.

Ziegler, J. C., Petrova, A., and Ferrand, L. (2008). "Feedback consistency effects in visual and auditory word recognition: Where do we stand after more than a decade?". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34:3, 643–661.

Ziegler, J. C., Stone, G. O., and Jacobs, A. M. (1997). "What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English". *Behavior Research Methods, Instruments, & Computers*, 29:4, 600–618.