Both intrinsic and allophonic vowel duration matter in textsetting

Nicole Gilroy (nikki.gilroy@carleton.ca)
Faculty of Arts and Social Sciences, 1125 Colonel By Drive
Ottawa, ON K1S 5B6 CAN

Lev Blumenfeld (lev.blumenfeld@carleton.ca)
Department of Linguistics and Language Studies, 1125 Colonel By Drive
Ottawa, ON K1S 5B6 CAN

Ida Toivonen (ida.toivonen@carleton.ca)
Department of Cognitive Science and Department of Linguistics and Language Studies
1125 Colonel By Drive
Ottawa, ON K1S 5B6 CAN

Abstract

In studies of song corpora, longer vowels have been shown to be preferentially aligned with longer notes in textsetting. Here we test this alignment preference in English in an experimental setting and replicate the finding for duration in a task where participants constructed a textsetting by placing target words in appropriate slots. We test two types of vowel duration: intrinsic duration and vowel duration that is contextually determined by the voicing of the following consonant. We show that both of these types of duration have an effect on textsetting preferences.

Introduction

Background

Music and language have a lot in common. Structures in both domains display prominence, rhythm, and hierarchical constituency (Lerdahl & Jackendoff, 1983). At some level of abstraction, musical well-formedness principles resemble phonological well-formedness; proficient performers and music listeners have intuitions about musical structures similar to grammaticality judgments of native speakers. For all these reasons music is of interest to linguists, as it can shine a novel light on linguistic structure and grammar (e.g., Katz & Pesetsky, 2011).

The importance of music to linguistics is apparent where linguistic representations are aligned with music in songs and chants. Such behavior is called *textsetting*. A textsetting structure contains both a linguistic form (the phonological and phonetic structure of a text), and a musical form (the melody, rhythm, phrasing and other aspects of musical organization) as seen in Figure 1.

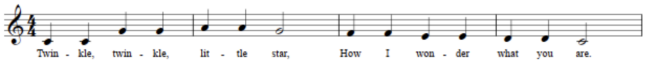

Figure 1: Twinkle Twinkle Little Star

Given a line of text and a musical phrase, there are many ways to align them—i.e. to sing the line to the notes of the music—but only some are actually used and accepted by native speakers. Characterizing the set of possible alignments between a given text and a given musical structure is the task of a *textsetting grammar*. The research program of studying textsetting, inspired by Halle and Lerdahl (1993) and continued by Halle (1999); Kiparsky (2006); Dell and Halle (2009); Hayes (2009a, 2009b), among others, aims to describe specific textsetting systems and to circumscribe the possible ways textsetting can operate, in particular the kinds of information it may access on both the musical and linguistic sides.

The question of what information from one module or domain (e.g., language) is available to another module or domain (e.g., music) is a common one in the study of interfaces. In parallel to the issues pursued below, the question also arises in the study of metrics, which shares with textsetting the alignment of two structures, a line of text and a metrical template (Kiparsky, 1977; Hanson & Kiparsky, 1996; Blumenfeld, 2015). In this process of alignment, does the metrical grammar access the surface structure, the underlying form, or some intermediate representation? These issues have been of concern to metrists since Kiparsky (1968, 1972).

The question of abstractness is relevant in our study from another point of view. It offers a novel take on an old question: to what extend does predictable, non-contrastive information play a role in perception? This question takes center stage in exemplar models which attempt to account for grammatical behavior by assuming storage of finely detailed perceived tokens, or exemplars (Goldinger, 1996; Palmeri, Goldinger, & Pisoni, 1993; Johnson, 1997b, 1997a; Coleman, 2002; Hawkins, 2003; Pierrehumbert, 2003, 2016, a.o.). While some authors have suggested that exemplars are stored as fully detailed acoustic representations (Johnson, 1997b), experimental results in recent years have accumulated that show at least partial abstraction, or stripping of predictable information in perception or exemplar storage.

For example, the "stress deafness" literature (Dupoux, Pallier, Sebastian, & Mehler, 1997; Peperkamp & Dupoux, 2002) demonstrates that stress is harder to perceive in languages where it is predictable. In segmental phonology, Boomershine, Currie Hall, Hume, and Johnson (2008)

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

showed that pairs of sounds such as [d] and [ð] are perceived as more distinct in languages where they contrast (English) than in languages where they are allophones (Spanish). Harnsberger (2001) reports analogous result for nasal allophones in Malayalam. In the context of exemplar theory, Manker (2020) shows that expected coarticulation, such as f0 perturbation by voicing of consonants, undergoes abstraction in exemplar storage. Results such as these prompted the development of hybrid models accommodating some degree of abstraction in addition to storage of phonetic detail (Pierrehumbert, 2002, 2016).

Our paper reports on an experiment addressing the question of how duration of notes and duration of vowels interacts in English textsetting, in particular focusing on the nature of information that is available to the textsetting grammar. In our experiment, we examine the role of two kinds of duration in textsetting: intrinsic duration, and allophonically induced duration. Our results are in line with the literature cited in the preceding paragraph: while allophonic duration plays a role, its contribution to the textsetting grammar appears to be weaker than the contribution of intrinsic duration.

Duration and textsetting

Unsurprisingly, longer vowels are preferentially aligned with longer notes in textsetting. Hayes and Kaun (1996, 260) proposed the Syllable Duration Rule in (1):

(1) Syllable Duration:
Reflect the natural phonetic durations of syllables in the number of metrical beats they receive.

However, duration is a complex property. Phonetic duration of a vowel can be determined by many factors: its phonemic status as "long" or "short", its features such as tenseness or height, its status as stressed or unstressed, its position within the word, and segmental contextual factors. Each of these effects contributes to the phonetic duration of a vowel, and it is an open question whether the textsetting grammar can access that surface duration directly, or interfaces in a more abstract way with the phonological grammar (see, e.g., Hayes & Kaun, 1996, 260–261). Phonological quantity, while more coarse-grained than raw phonetic duration, also displays gradient or at least multivalued behavior (Ryan, 2011, 2014, 2019).

The general longer-note-to-longer-vowel principle has been demonstrated for various aspects of duration. In Finnish, vowel length is phonemic: short vowels differ contrastively from long vowels. For example, the word *muta* with a short [u] means 'mud' and *muuta* with a long [uː] means 'other'. Arjava and Kentner (2022) examined the role of prosodic weight in textsetting in 27 well-known Finnish songs. One of the factors that determines prosodic weight in Finnish is vowel length, and Arjava and Kentner's study shows that phonemically long vowels are preferred on longer notes, and short vowels are preferred on shorter notes. More generally, they found that syllable weight aligns with musical length.

A few studies have considered the role of intrinsic vowel duration in textsetting. Certain vowels are inherently longer than others: low vowels vowels are longer than high vowels, for example. Based on different data sets, Ryan (2022) and Fenk-Oczlon (2022) both find that syllables with low vowels were more likely to occur on longer notes, and syllables with high vowels were more likely to occur on shorter notes. Ryan's data set consists of 2371 English pop songs, and Fenk-Oczlon's data set consists of 20 traditional Alpine yodels, which are nonsense syllables pronounced by native speakers of German.

Hayes and Kaun (1996) find further systematic interactions between prosodic positions, syllable weight and duration. They are mainly interested in prosodic structure and phrase-final lengthening, and they do not specifically examine the role of intrinsic duration depending on segment identity, but they do suggest that it is likely to play a role (300).

Less work has been devoted to contextually determined duration, and most existing work is corpus-based, not experimental. We aim to fill this research gap in the present study.

Current study: research questions

We explore experimentally the effect of two aspects of duration on the textsetting grammar in English: intrinsic duration of vowels due to their status as tense vs. lax, and duration determined contextually by the voicing of the following consonant.

Tense vowels are generally longer than lax vowels; for example, the vowel in *beat* is longer than the vowel in *bit*. We explore whether speakers prefer to align *beat* to longer notes than *bit*. Likewise for the contextual factor: the vowel in *bead* is longer than the vowel in *beat*, and we explore the alignment preferences here as well. In the context of the general syllable duration rule (1), we expect the following results.

(2) a. Monosyllabic words with tense vowels are preferred on long notes, and words with lax vowels are preferred on short notes.
b. Monosyllabic words with voiced final consonants are preferred on long notes, and words with voiceless final consonants are preferred on short notes.

These questions are explored by an experimental task where participants are asked to generate a textsetting by placing target words into appropriate slots of a song. The study reported here is the second component of a larger study that consisted of two experiments. In the first experiment, listeners are asked about their preferences between alternative textsettings. The results were not significant. That component is described in detail in Gilroy (2021).

Our study thus addresses both the basic question of whether duration influences alignment in textsetting, and the more fine-grained question of what durational information is accessible to the textsetting grammar: intrinsic duration, allophonic contextually-determined duration, or both.

Methodology

Participants

The analysis included the responses of 87 participants. Originally, 170 participants were recruited, however the responses of 83 participants were immediately excluded because their responses did not meet the minimal possible time to complete the survey.[1] Participants were recruited using Carleton University Cognitive Science Department's recruitment system (SONA). As compensation for their participation, participants received 0.75% of a credit towards a Cognitive Science course. The eligibility requirements stated that any student with an active SONA account could participate.

Experimental design

The experiment made use of a pre-recording of the beginning of the song *Twinkle Twinkle Little Star*, depicted in Figure 1. The participants were asked to fill in blanks in the song. The song was recorded specifically for this experiment by the first author of this paper.

Instead of singing the lyrics, the singer repeated the syllable "la" and left the fourth quarter note and the second half note silent. These notes were selected because a quarter note is half the length of a half note. Thus, only two notes of the melody were tested in this experiment. There was a play button beside the top line that the participants could press to listen to the song. All participants were familiar with the complete song from before. *Twinkle Twinkle* is a well-known song, and the participants had also heard it as part of the first part of the study, which we do not report on here.

Each stimulus consisted of a minimal pair of English monosyllabic words that differed in vowel length. Vowel length differed as a result of vowel tenseness or allophonic lengthening due to the voicing of the final consonant. Participants were asked to drag one member of the minimal pair into one of two boxes, which correspond to the missing notes labeled *Blank 1* for the short note and *Blank 2* for the long note, as shown in Figure 2.

The participants repeated this task 29 times with 29 different minimal pairs. 16 minimal pairs differed in whether the vowel was tense or lax and and 13 minimal pairs differed in whether the final consonant was voiced or voiceless. The members of the minimal pairs can be described as "long" and "short" words, since tense vowels are longer than lax vowels, and vowels are longer before voiced than voiceless consonants. The minimal pair in Figure 2, *hit hid*, illustrates the voicing distinction. An example minimal pair from the experiment that displays the tenseness distinction is *hit heat*. Appendix A contains the full list of minimal pairs.

[1] We caluclated that participants would need a minimum of 15 minutes to consider all experimental items. Listening to the audio files alone would take seven minutes, and the audiofiles were crucial for the first part of the experiment (see Gilroy, 2021).

Figure 2: Example stimulus: blanks in the musical phrase

After the participants selected a word for one of the empty slots, the second member of the minimal pair was automatically placed into the remaining slot. The order in which the word pairs were presented to the participants was randomized. In addition, the order in which the words in each pair were presented to the participants was randomized. Once a participant had made their choice for one pair and both words were assigned to notes, they would click on an arrow to bring them to the next word pair. Participants were not able to backtrack to change previous responses. There was no time limit.

Results

The following table shows the 'match' and 'mismatch' responses by pair type, and the match/mismatch ratios, illustrating that there was a stronger effect for intrinsic duration than for allophonic duration. The first two rows illustrate the intrinsic vowel duration in voiceless and voiced context (exemplified by *bit/beat*, *bid/bead*). The last two rows illustrate allophonic duration for lax and tense vowels (exemplified by *bit/bid*, *beat/bead*).

PAIR	MATCH	MISMATCH	RATIO	
bit/beat	449	224	2.00	intrinsic
bid/bead	464	228	2.04	
bit/bid	307	215	1.43	allophonic
beat/bead	371	237	1.57	

Table 1: Intrinsic vs. allophonic effects

A Chi-squared Goodness of Fit test was performed to test whether there was a difference between the matching and non-matching condition. A paring between the long note and a long word (i.e., a word with a tense vowel or final voiced consonant) was considered a match, and a pairing between the long note and a short word (i.e., a word with a lax vowel or final voiceless consonant) was considered a mismatch. The participants matched vowel length to the length of the note 1591 times and mismatched 924 times. There was a significant preference for matches over mismatches ($N = 2515, \chi^2(1) = 176.9,\ p < .01$).

In order to examine if vowel tenseness or consonant voicing had a stronger effect, a logistic regression was conducted. The independent variable was the type of duration difference (voicing or tenseness), and the dependent variable was the participants' choice (match or mismatch). The effect was significant, $\chi^2(1) = 9.372,\ p < .01$. Both groups were significant predictors of the match/mismatch preferences, but the effect of vowel tenseness was 1.29 times stronger than the effect of allophonic lengthening. In other words, when participants were presented with the tense-lax vowel distinction they were 29% more likely to match than when they were presented with the allophonic lengthening condition. See Table 1 and Figure 3.

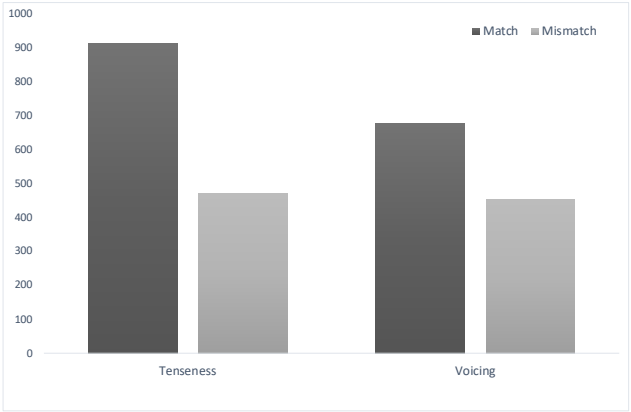

Figure 3: Matching between notes and vowel duration in tense/lax vowel word pairs and word pairs with voiced/voiceless final consonant word pairs

Conclusion

The study described in this paper supports the conclusion that duration influences alignment in textsetting: there is a preference for "matching" the duration of vowels and notes in the sense that longer vowels align with longer notes and shorter vowels with shorter notes. The present study adds experimental evidence to the corpus evidence provided in previous studies. Our experiment found a long-to-long matching preference for intrinsic vowel duration determined by tenseness, and also for allophonic vowel duration reflecting the voicing of the postvocalic consonant. Allophonic contextually-determined duration has not been a focus of previous studies, and the fact that it influences textsetting is thus a novel finding.

Previous studies have argued that textsetting is sensitive to different kinds and levels of phonological representation (see Hayes & Kaun, 1996, and also McPherson, 2019 for discussion and further references). Our results can be interpreted as additional support for this claim, since we found that the effect of intrinsic vowel duration (*beat-bit*) was stronger than the effect of contextually-determined duration (*bead-beat*), in line with general results that predictable information can be stripped away in perception.

As suggested by a reviewer, future studies could test less well-known melodies, as well as more target notes in the melody to examine the role of other potential sources of bias, such as ordering.

References

Arjava, H., & Kentner, G. (2022). Alignment of prosodic weight and musical length in Finnish vocal music textsetting. In M. Scharinger & R. Wiese (Eds.), *How language speaks to music: Prosody from a cross-domain perspective* (pp. 161–189). Berlin: Walter de Gruyter. doi: 10.1515/9783110770186-007

Blumenfeld, L. (2015). Meter as faithfulness. *Natural Language and Linguistic Theory*, *33*, 79–125. doi: 10.1007/S11049-014-9254-8

Boomershine, A., Currie Hall, K., Hume, E., & Johnson, K. (2008). The impact of allophony versus contrast on speech perception. In P. Avery, B. E. Dresher, & K. Rice (Eds.), *Contrast in phonology: theory, perception, acquisition* (pp. 145–171). Berlin: Mouton de Gruyter.

Coleman, J. (2002). Phonetic representations in the mental lexicon. In J. Durand & B. Laks (Eds.), *Phonetics, phonology, and cognition* (pp. 96–130). Oxford University Press.

Dell, F., & Halle, J. (2009). Comparing musical textsetting in French and English songs. In J.-L. Aroui & A. Arleo (Eds.), *Towards a typology of poetic forms: from language to metrics and beyond* (pp. 63–78). Amsterdam: John Benjamins. doi: 10.1075/lfab.2.03del

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A destressing 'deafness' in French? *Journal of Memory and Language*, *36*, 406–421.

Fenk-Oczlon, G. (2022). Iconic associations between vowel acoustics and musical patterns, and the Musical Protolanguage Hypothesis. *Frontiers in Communication*, *7*, 887739. doi: 10.13140/RG.2.2.17540.48007

Gilroy, N. (2021). *English vowel duration in textsetting*. Unpublished master's thesis, Carleton University.

Goldinger, S. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Halle, J. (1999). *A grammar of improvised textsetting*. Unpublished doctoral dissertation, Columbia University.

Halle, J., & Lerdahl, F. (1993). A generative textsetting model. *Current Musicology*, *55*, 3–23.

Hanson, K., & Kiparsky, P. (1996). A parametric theory of poetic meter. *Language*, *72*, 287–335. doi: 10.2307/416652

Harnsberger, J. (2001). The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis. *Journal of Phonetics*, *29*, 303–327.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3–4), 373–405.

Hayes, B. (2009a). Faithfulness and componentiality in metrics. In K. Hanson & S. Inkelas (Eds.), *The nature of the word: Essays in honor of Paul Kiparsky* (pp. 113–148). Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/7894.003.0009

Hayes, B. (2009b). Textsetting as constraint conflict. In J.-L. Aroui & A. Arleo (Eds.), *Towards a typology of poetic forms: from language to metrics and beyond* (pp. 43–62). Amsterdam: John Benjamins. doi: 10.1075/lfab.2.02hay

Hayes, B., & Kaun, A. (1996). The role of phonological phrasing in sung and chanted verse. *The Linguistic Review*, *13*, 243–303. doi: 10.1515/tlir.1996.13.3-4.243

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics*, *50*, 101–113.

Johnson, K. (1997b). Speech perception without speaker normalization. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Katz, J., & Pesetsky, D. (2011). *The identity thesis for language and music.* (Ms., MIT)

Kiparsky, P. (1968). Metrics and morphophonemics in the Kalevala. In C. Gribble (Ed.), *Studies presented to Roman Jakobson by his students.* Cambridge, MA: Slavica. (Reprinted in D. C. Freeman, ed., *Linguistics and literary style*. New York: Holt, Rinehart and Winston. 1971)

Kiparsky, P. (1972). Metrics and morphophonemics in the Rigveda. In M. Brame (Ed.), *Contributions to generative phonology* (pp. 171–200). Austin, TX: University of Texas Press.

Kiparsky, P. (1977). The rhythmic structure of English verse. *Linguistic Inquiry*, *8*(2), 189–247.

Kiparsky, P. (2006). A modular metrics for folk verse. In B. E. Dresher & N. Friedberg (Eds.), *Formal approaches to poetry* (pp. 7–49). Berlin: Mouton de Gruyter. doi: 10.1515/9783110197624.1.7

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: The MIT Press.

Manker, J. (2020). The perceptual filtering of predictable coarticulation in exemplar memory. *Laboratory Phonology*, *11*(1), 1–17. doi: 10.5334/labphon.240

McPherson, L. (2019). Musical adaptation as phonological evidence: Case studies from textsetting, rhyme, and musical surrogates. *Language and Linguistic Compass*, *13*(12), e12359. doi: 10.1111/lnc3.12359

Palmeri, T., Goldinger, S., & Pisoni, D. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 203–240). Berlin: Mouton de Gruyter.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–139). Berlin: Mouton de Gruyter. doi: 10.1515/9783110197105.101

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*, 115–154.

Pierrehumbert, J. (2016). Phonological representation: beyond abstract versus episodic. *Annual review of linguistics*, *2*, 33–52. doi: 10.1146/annurev-linguist-030514-125050

Ryan, K. (2011). Gradient syllable weight and weight universals in quantitative metrics. *Phonology*, *38*(3), 413–454. doi: 10.1017/S0952675711000212

Ryan, K. (2014). Onsets contribute to syllable weight: statistical evidence from stress and meter. *Language*, *90*, 309–341. doi: 10.1353/lan.2014.0029

Ryan, K. (2019). *Prosodic weight: categories and continua*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198817949.001.0001

Ryan, K. (2022). Syllable weight and natural duration in textsetting popular music in English. *English Language and Linguistics*, *26*(3), 559–582. doi: 10.1017/S1360674322000156

Appendix A

VOICING	TENSENESS
hit hid	hit heat
sit Sid	sit seat
bit bid	fit feet
fit fid	bit beat
grit grid	grit greet
mitt mid	mitt meet
meet mead	hid heed
greet greed	mid mead
seat seed	fid feed
heat heed	Sid seed
beat bead	bid bead
feet feed	grid greed
boot booed	took toque
	look Luke
	pull pool
	full fool

Figure 4: Stimuli