Perception of vowel quantity in sung Estonian

Kaili Vesik
University of British Columbia
kaili.vesik@ubc.ca

February 12, 2021

Abstract

This paper investigates representation and perception of contrastive vowel length in Estonian songs. In speech, vowels in Quantity 1 (short; Q1) and Quantity 2 (long; Q2) are correlated with $\sigma 1/\sigma 2$ duration ratios of 2:3 and 3:2, respectively (Lehiste, 1960). However, in atypical timing contexts such as song, adhering to speech ratios may not always be feasible. I aim to determine whether native Estonian speakers perceive vowel quantity in song adhering to the same criteria that they are shown to use in speech. A small corpus exploration suggests that there is a general trend for composers to assign Q1 words to shorter:longer note pairs and Q2 words to longer:shorter pairs, without specific adherence to the ratios typical of speech. Given that text setting in musical composition does not necessarily follow expected speech ratios, a perception study was implemented with the purpose of determining whether listeners nevertheless perceive vowel quantity in song according to the speech ratios. Native Estonian listeners were asked to identify sung bisyllabic nonce words of varying ratios as either Q1 or Q2, and results showed that participants identify sung tokens according to the same ratios as they do in speech. This suggests that in the absence of clues from lexical information, native listeners use the same perceptual tools from speech to identify vowel length in song, even though words in composed music are not necessarily always presented in the same way.

***Keywords*—** phonology, Estonian, vowel length, perception, music

1 Introduction

In many of the world's languages (at least 11% and up to 32% based on data from the UPSID (Maddieson, 1984) and PHOIBLE (Moran & McCloy, 2019) databases), the length of a vowel sound holds lexical significance. That is, phonological length is contrastive: changing the duration of one sound can change the meaning of a word. For example,

the Estonian words *kalu* [kɑlu] (plural dative"fish"), *kaalu* [kɑɑlu] (singular imperative "weigh"), and *kaalu* [kɑɑːlu] (singular partitive "scale") differ only in the length of the first vowel, where the [ɑ] sound in [kɑlu] is short, in [kɑɑlu] is long, and in [kɑɑːlu] is overlong. Thus, the acoustic duration of the vowel is critical to a listener's correct interpretation of the underlying phonological length, and therefore the speaker's intended meaning.

Estonian speakers produce and perceive these differences in length as a natural part of spoken language. In some settings, however, typical temporal patterns are disrupted; for example, in music. Hence clear communication in Estonian could be impeded in singing, since the relative speed at which neighbouring syllables are sung could change the meaning. This is particularly relevant to Estonian in that joint singing (e.g. choral music) is a large part of national identity and cultural transmission in Estonia (Raudsepp & Vikat, 2009, 2011). This paper investigates whether Estonian listeners are able to adjust to a greater degree of length variation in sung language than is typical in spoken language. More specifically, via a corpus exploration as well as a perception experiment, I show that (a) in compositions of Estonian choral music, in which there exist lexical clues to vowel quantity, the measures that determine Estonian vowel quantity contrasts in speech are not entirely preserved, and (b) in a context where those lexical clues are not applicable, native listeners perceive vowel quantity contrasts in sung Estonian similarly to the way they do according to the speech literature.

2 Background

2.1 Estonian quantity

Estonian is in the Finnic branch of the Uralic language family, and contrasts length in both vowels and consonants. In particular, it has been described as having a three-way length contrast (Ross & Lehiste, 1998; Eek, 1983). The three-way contrast is available in the primary stressed syllable in a disyllabic foot; that is, the initial syllable of a polysyllabic word (Asu & Teras, 2009; Lippus, Pajusalu, & Allik, 2009; Lehiste, 1960). These contrasting values are referred to as Quantity 1 (associated with short phonological length), Quantity 2 (long), and Quantity 3 (overlong). This study focuses solely on vowel quantity; see examples in Table 1.

Estonian quantities, assigned to word-initial syllables, are correlated perceptually with particular values for the ratios of first-syllable to second-syllable durations, rather than the raw durations of the stressed syllables themselves (Eek & Meister, 1997).

In a study investigating the relative durations associated with Estonian quantity, Lehiste (1960) recorded a series of minimal pairs differing only in quantity. These bisyllabic stimuli were recorded with varying relative durations of their first and second

	Q1 (short)	Q2 (long)	Q3 (overlong)
(1)	*vina*	*viina*	*viina*
	[vi.nɑ]	[vii.nɑ]	[vii:.nɑ]
	vapour.NOM	vodka.GEN	vodka.PART
(2)	*koli*	*kooli*	*kooli*
	[ko.li]	[koo.li]	[koo:.li]
	trash.NOM	school.SG.GEN	school.SG.PART
(3)	*krabid*	*kraabid*	-
	[krɑ.pit]	[krɑɑ.pit]	-
	crab.PL.NOM	scrape.2SG.PRES	-

Table 1: Length contrast in Estonian. (1) - Spahr (2014); (2) - Ehala (2003); (3) - author's; notation from Prillop (2013).

syllables. Twenty native Estonian listeners were asked to identify whether the word they heard would fit into either of two suggested sentences (or neither), thus categorizing the quantity of the stimulus based on relative syllable durations or rejecting it entirely. Any stimulus whose quantity was agreed upon by at least 75% of listeners was included in the analysis; see Table 2 for sample data.

Results showed that the average $\sigma 1/\sigma 2$ duration ratio of Q1 words is 2:3, that of Q2 words is 3:2, and of Q3 words is 2:1 (see Figure 1); for the first of two syllables to sound "short" to a native listener, $\sigma 1$ must in fact have a shorter duration than $\sigma 2$.

Fox and Lehiste (1989) showed that native Estonian speakers are able to distinguish between two pairs of sounds (bursts of speech noise) when one pair's sounds had a duration ratio less than one (1:2 or 2:3) and the other pair's sounds had a duration ratio greater than one (3:2 or 2:1). However, they were not able to distinguish between pairs of sounds

Words of quantity 2

Agreement	Number of words	$\sigma 1$ duration (cs)	$\sigma 2$ duration (cs)	$\sigma 1/\sigma 2$ ratio
100%	10	29.8	19.6	$\approx 30/20$
75-95%	25	29.2	18.1	$29/18$
75-100%	35	29.4	18.5	$29/19$
Below 75%	165			

Table 2: Sample data adapted from Lehiste (1960, p. 57)

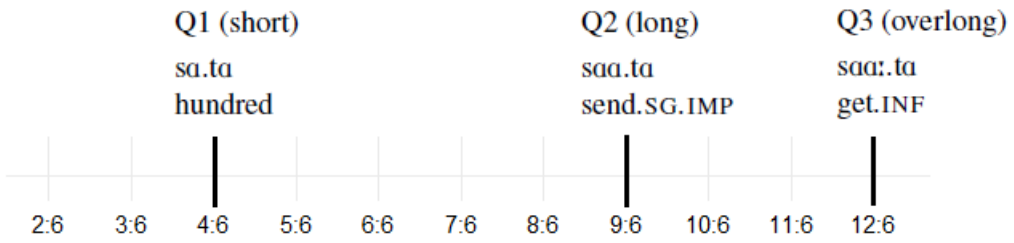


Figure 1: Syllable duration ratios associated with Estonian quantity (Lehiste, 1960; Lippus et al., 2009).

when both pairs had duration ratios either less than one or greater than one. In an investigation of quantity cues amongst ratios greater than one, Lippus et al. (2009) showed that using bisyllabic tokens as stimuli, native Estonian speakers required the addition of falling pitch on the first syllable in order to successfully identify targets as Q3 rather than Q2 at levels better than chance. Thus only the distinction between Q1 and Q2 is available independent of pitch.

2.2 Phonological contrast in music

Due to the necessity of f0 cues for distinguishing Q2 from Q3, the focus of this study is simply on the perception of Q1 vs Q2 (i.e., short vs long). In sung Western music, text is often set within a metrical framework in which $\sigma 1/\sigma 2$ ratios of 1:1, 2:1, or 3:1 (or their reciprocals) are much more likely than those of 2:3 or 3:2. Although Native Estonian listeners' discrimination of duration ratios for these contrasts have been studied as they occur in typical speech contexts, there is currently no work that investigates these percepts in singing. This study addresses this gap: investigating representation and perception of Q1 vs Q2 in sung speech facilitates the analysis of speakers' ability to make crucial distinctions with respect to the length of speech sounds, in contexts with atypical speech timing.

Given the lack of research on how vowel length is represented and/or perceived in singing, similar studies involving lexical tone in singing will be used to inform the approaches used in this study. The techniques described in these studies, for categorizing suprasegmental properties of speech that are also manipulated in music, can be applied to analysis of vowel length as well.

Kirby and Ladd (2016) analyzed the correspondence of between-syllable tone vs musical transitions in Vietnamese songs, developing three categories to describe the alignment between the melodic pitch change from one syllable in a song to the next, compared to the lexical tone change. They used the musical concepts of similar, oblique, and con-

trary motion to compare linguistic vs musical transitions, where similar motion means that two melodies simultaneously increase, decrease, or remain level in pitch, oblique means that one melody increases or decreases in pitch while the other remains level, and contrary means that one melody increases in pitch while the other decreases. Kirby and Ladd found that the degree of correspondence between linguistic and musical structure can be represented by similarity of linguistic vs musical transitions from one unit (syllable) to the next, with levels of linguistic phenomena grouped and ordered such that maximum similar motion and minimum contrary motion are achieved.

Wong and Diehl (2002), in a two-part study of contemporary Cantonese music, investigated how songwriters balance the competing pressures of melodic expression vs lexical tone structure when setting text to music. In an analysis of existing songs, they found that composers of contemporary Cantonese music differentiate lexical tones via general information about whether a syllable's pitch is higher or lower than another's, rather than being concerned with precisely how much higher or lower (as is typical in speech). In a subsequent perception experiment, they found similarly that the *size* of a musical interval between a sequence of sung syllables was irrelevant to listeners' perception; only the *direction* had an effect on lexical tone perception. Thus Wong and Diehl showed that Cantonese songwriters and listeners use an ordinal system (relative pitch) to both set as well as perceive lexical tones, which is less rigid than that seen in speech. This allows for musical freedom without completely ignoring lexical intelligibility. I use a similar approach to Wong and Diehl, considering how Estonian vowel quantity is both set by composers and perceived by listeners.

3 Research questions

3.1 Questions for corpus exploration

Given a small collection of Estonian choral music, I use a 1:1 note-length ratio as a reference point and ask:

1. Are Q1 and Q2 syllable pairs set to note pairs with ratios less than and greater than one, respectively (i.e., as similar as possible to 2:3 and 3:2 given the metrical structure of these songs)?

2. If not, how can we describe the way that Q1 and Q2 syllable pairs are instead assigned to note pairs?

I predict that Western musical conventions prevent composers from setting text in a way that faithfully follows Lehiste's ratios. Further, I predict that Q1 syllables tend to be assigned to note-length ratios of 1:1 (i.e., text is set according to phonological segment

length rather than speech-based syllable ratios) while Q2 syllable pairs tend to be assigned to note-length ratios of greater than 1:1.

3.2 Questions for perception experiment

In this experiment, informed by the corpus exploration, I analyze native listeners' perception of vowel quantity in sung Estonian, investigating how native listeners perceive sung Estonian vowel quantities over a continuum of $\sigma 1/\sigma 2$ duration ratios. Given bisyllabic stimuli whose $\sigma 1/\sigma 2$ ratios range over a continuum from 1:3 to 2:1, I ask:

3. Whether Lehiste's (1960) ratios for Q1 (2:3) and Q2 (3:2) in speech are identified the same way in sung Estonian (with ratios $\leq$ 2:3 perceived as Q1 and $\geq$ 3:2 as Q2).

I predict that sung syllable duration ratios of 2:3 and below are perceived as being in Q1 (having a short V1) and those with duration ratios of 3:2 and above are perceived as Q2 (long V1). Further, I predict that the crossover point of a predicted categorization curve will be located at approximately 1:1.

4 Corpus exploration

4.1 Corpus

The small collection of music explored in this study includes eight traditional and contemporary Estonian choral music compositions in notated form (Ernesaks & Koidula, 2018; Kaskmann, 1989; Härma & Karlson, 2018; Kõrvits, 2018; Vabarna, 2016; Uusberg, 2013; Pehk, 2013; Tamberg, 2013). These works were selected from various collections of traditional songs as well as the author's personal collection of repertoire performed at national and regional Estonian choral events, with the aim to cover a range of styles (including traditional runic songs, patriotic songs, and contemporary choral music purpose-composed for performance) and composition dates (from 1897 to 2017).

The pieces in the corpus comprise 722 tokens, 462 of which have at least two syllables. In order to focus on vowel length to the exclusion of any potential interference from coda consonants, only those words with open first syllables (314) were used. From this set, pairs of first and second syllables were analyzed for their vowel quantities as compared to their comparative note lengths set by the composers.

Notes from the various compositions in the corpus were coded for length using the fractions corresponding to their value (a quarter note would be recorded as having length $1/4$). Based on these fractional lengths, note1-note2 pairs had their ratios computed as the quotient of the two fractions, multiplied by 6 in order to maintain a common second

term (denominator) of 6. For example, if the second note of the pair is three times as long as the first, these notes are in the ratio 1:3. Since the $\sigma 1/\sigma 2$ duration ratios for speech include second terms (denominators) of 1, 2, and 3, a common second term of 6 was used for all ratios described in this paper. Hence if a note1-note2 ratio is 1:3, it is recorded as 2:6. Sample data are shown in Table 3.

song	word	syl1	longv1	note1len	syl2	note2len	numover6
Isamaa	uneta	u	0	0.083333	ne	0.083333	6
Isamaa	öödel	öö	1	0.250000	del	0.250000	6
Isamaa	kõigil	kõi	1	0.083333	gil	0.083333	6
Isamaa	vaevastel	vae	1	0.083333	vas	0.083333	6
Isamaa	töödel	töö	1	0.500000	del	0.375000	8
Isamaa	ütelda	ü	0	0.125000	tel	0.125000	6
Isamaa	oled	o	0	0.166667	led	0.083333	12
Isamaa	seesama	see	1	0.083333	sa	0.083333	6
Isamaa	elu	e	0	0.083333	lu	0.083333	6
Isamaa	sinult	si	0	0.083333	nult	0.166667	3

Table 3: Sample data from corpus exploration

It is possible that, due to lexical competition, quantity-based minimal pairs present in this dataset may have been more likely to align with their expected speech-based syllable ratios in order to preserve contrast (Baese-Berk & Goldrick, 2009; Wedel et al., 2018). Given the exploratory nature of the study, the existence of minimal pairs is not taken into account here; however, there is some discussion of minimal pairs present in the perception experiment (see Section 5.3.3).

4.2 Note-length ratio vs vowel quantity

To address the question of whether Q1 and Q2 syllable pairs are set to note pairs with ratios less than and greater than one, respectively, note-length contrasts are compared to vowel-quantity contrasts. The data from this study are described using a system similar to Kirby and Ladd (2016), who refer to *similar*, *oblique*, and *contrary* motion. Similar settings are those whose ratios are both equal to, less than, or greater than 1:1; oblique settings are those where one ratio is equal to 1:1 and the other is either less than or greater than 1:1; contrary settings are those where one ratio is less than 1:1 and the other is greater.

Considering the raw counts in Table 4, the majority of both Q1 syllable pairs and Q2 syllable pairs are assigned to note pairs of equal length, rather than shorter (for Q1) or longer (for Q2). This confirms my prediction that text would not tend to be set according

to the speech-based ratios. It also confirms my prediction that Q1 syllable pairs tend to be assigned to note pairs of ratio 1:1. However, it disconfirms my prediction that Q2 syllables would tend to be assigned to note pairs of ratio greater than 1:1 (though NOTE that there is only a single Q2 syllable pair assigned to a note pair whose ratio is less than 1:1).

		Note1:Note2 length ratio			
		shorter:longer	equal	longer:shorter	
Vowel quantity	Q1 (short V1)	Oblique 19 6.0% of total	**Similar** **174** **55.1% of total**	Oblique 31 9.8% of total	Total 316
	Q2 (long V1)	*Contrary* *1* *0.3% of total*	Oblique 69 21.8% of total	**Similar** **22** **7.0% of total**	

Table 4: Similar, oblique, and contrary settings in corpus

Further to considering simply the tallies in each cell, if the syllable pairs as analyzed as having similar, oblique, or contrary settings, the majority of the data (62.0%) are seen to have similar settings, 31.6% have oblique settings, but only 1 pair = 0.3% has a contrary setting. So although it is not the case that Q2 syllable pairs tend to be assigned to note-length ratios of greater than 1:1, it is at least the case that composers avoid (whether intuitively or intentionally) completely inverted text-music length relationships.

Figure 2 shows a plot of the corpus data (note that jitter has been added in order to differentiate individual data points). This facilitates a more detailed examination of the assignment of note-length ratios to words of Q1 vs words of Q2, showing how note-length ratios are distributed as a function of quantity. The horizontal axis indicates whether a Q1 or a Q2 word is being considered, and the vertical axis represents the range of note-length ratios attested in the data, from 2:6 to 18:6. As noted above in Table 4, the majority of the data for both quantities is clustered at 6:6; that is, a pair of equal-length notes. However, there are some asymmetries in that the lowest note-length ratios (i.e. 2:6) are only ever assigned to Q1 words and never Q2 words, and also that there is more data in the higher note-length ratios (i.e. 9:6 and above) for Q2 words than for Q1 words. Though it is not dramatic by any means, there does appear to be a subtle trend of assigning greater note-length ratios to Q2 words than Q1 words.

In order to determine whether quantity has a significant positive effect on note-length ratio as suggested above, a mixed effects linear regression model was fitted to the data in R (R Core Team, 2020), using the lme4 package (Bates et al., 2015). The outcome variable is `ratio_num` (first term of the note-length ratio of the target word's two syllables, with second term normalized as 6), with `quantity` (Q1 or Q2) as predictor.

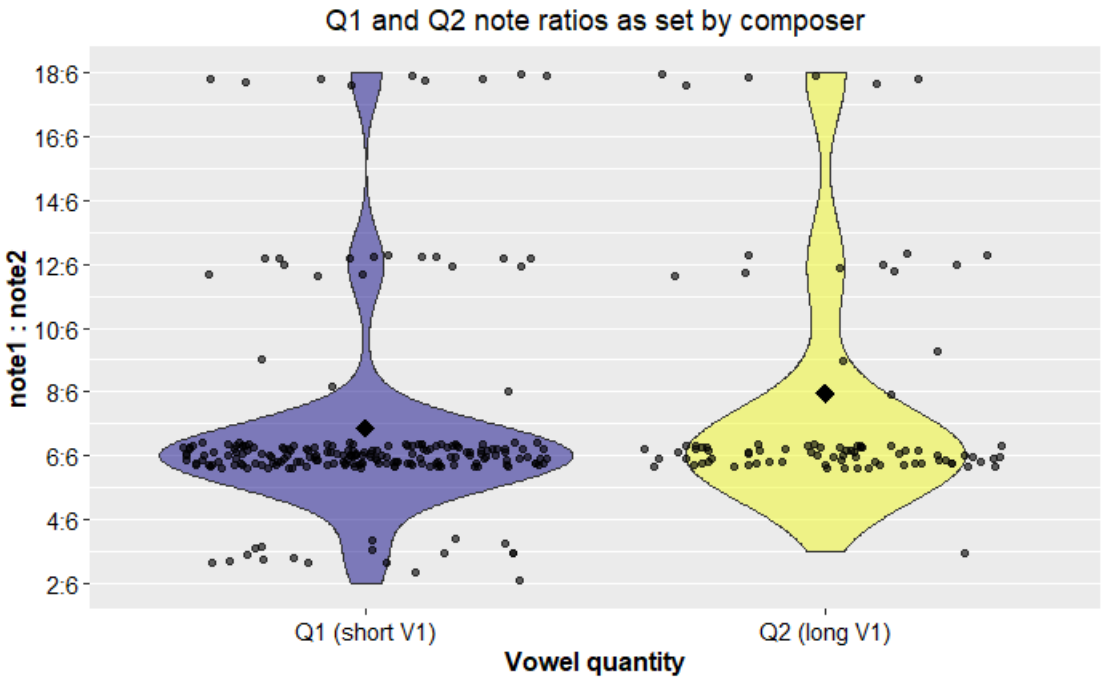


Figure 2: Assignment of note-length ratio based on vowel quantity. Black diamonds represent the mean ratio for each quantity.

Random intercepts by `song` and `word`, as well as random slopes over `quantity` by `song`, were included. According to the model,[1] there is no significant effect of quantity on note-length ratio ($\beta = 1.18$, $SE = 0.59$, $t = 2.01$, $p = 0.08$, with p in this and following models estimated using the afex package (Singmann et al., 2021)).

4.3 Q2 assignment vs note-length ratios

As mentioned in Section 4.2, there is a great degree of variation in this small corpus, in terms of vowel quantity vs note-length ratio assignments. Figure 3 serves to visualize this variation, showing for each song the proportion of note pairs assigned syllable pairs in which the first vowel is phonologically long (Q2). Although the ratios less than or equal to 1:1 tend to be assigned Q2 syllables less than 50% of the time, it is not that case that note-length ratio of 2:3 (4:6) or less are consistently assigned Q1 syllable pairs. Looking toward the higher ratios, there is even more variation in the way that ratios greater than or equal to 3:2 (9:6) are assigned text. In terms of the ratios between 2:3 and 3:2, there is no

[1] $\texttt{ratio\_num} \sim \texttt{quantity} + (1 + \texttt{quantity}|\texttt{song}) + (1|\texttt{word})$

clear relationship between the proportion of items assigned Q2 syllable pairs as compared to note-length ratio.

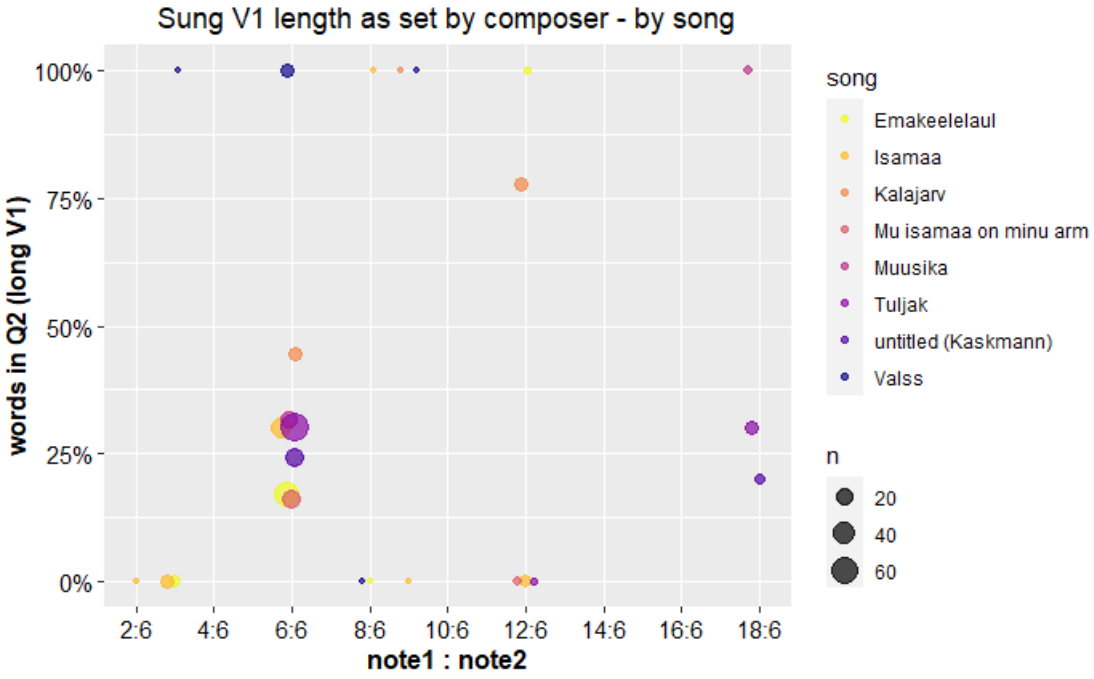


Figure 3: Percentage of words in Q2 vs $\sigma 1/\sigma 2$ note-length ratio

Figure 3 does suggest a general increasing trend in terms of percentage of note pairs assigned a Q2 syllable pair as the note-length ratios increase; however, there is also a large amount of variation within the data. This is due to several potential factors, including particularly restricted metrical structure in some songs (Valss, untitled (Kaskmann)), words at ends of songs being slowed to a more even pace for dramatic effect (Mu isamaa on minu arm), and uneven distribution of data (few data points at a specific note-length ratio mean that the effects of each observation are larger).

Overall, the results shown in Figure 3 align with the lack of evidence for text being set to music according to the same ratios found in speech (though the small size of the corpus may well have contributed to this null result).

5 Perception experiment

While the corpus exploration found no evidence that composers set text to rhythms that parallel quantity ratios in speech, it remains possible that listeners use rhythms in per-

formed music to identify or categorize vowel quantity in music, particularly in the absence of lexical information (i.e., in the case of nonce words).

5.1 Methods

5.1.1 Participants

Study participants were sixteen native speakers of Estonian, recruited via Facebook and word of mouth. They did not receive any compensation, financial or otherwise, for completing the task. Participants ranged in age from 22 to 78 years old (median 55.5 years) and included residents of Canada, the USA, and Estonia.

5.1.2 Materials

Stimuli consisted of sung bisyllabic targets set phrase-medially into sung carrier phrases. The targets comprised twenty $C_1V_1(V_1)C_{12}V_2$ nonce words, for each of which neither the Q1 ($C_1V_1C_{12}V_2$) nor the Q2 ($C_1V_1V_1C_{12}V_2$) version was an actual Estonian wordform. 240 candidate nonce words were generated by crossing $C_1\in\{$k,m,p,s,t$\}$ x $V_1\in\{$e,ø,o,ɤ$\}$ x $C_2\in\{$k,p,t$\}$ x $V_2\in\{$e,ø,o,ɤ$\}$. Only mid vowels were used as this facilitated the largest set of vowels while maintaining consistent vowel height, in an attempt to avoid inherent vowel duration effects associated with height. Candidates were excluded if (a) the Q1 or the Q2 version was (or could be misinterpreted as a legitimate mispronunciation of) an existing Estonian word, (b) C_1=C_2 (these words sounded like babble or baby-talk), (c) and/or the the word sounded like it belonged to either a specific Estonian dialect or Finnish (this resulted in many of the $V_2\in\{$ø,o,ɤ$\}$ candidates being excluded). The complete list of twenty nonce-word pairs (for example, [sepe]/[seepe]) is provided in Appendix A.1.

The carrier phrases comprised six different Estonian sentences, syntactically constructed such that two presented the target word in noun position, two in verb position, and two in adjective position. Only words consisting of open syllables with simple (or no) onsets were included in carrier phrases, in order to avoid distracting participants from focusing on the variation in the targets. Each carrier phrase was assigned its own melody. The complete list of six carrier phrases (for example, [tule ___ minukɑ]), along with their associated melodies, is provided in Appendix A.2.

Targets and carrier phrases were recorded by the author using a Blue Yeti USB microphone at a sampling rate of 44.1 kHz. Typical laboratory facilities were unavailable due to COVID-19 related restrictions; therefore the recording environment was a room with hardwood floors as well as a single-pane window facing a main road. In an attempt to attenuate room reverberation, the recordings were made under a blanket fort.

Target words were sung at a rate of one syllable per second (two beats per syllable at 120bpm), with both syllables of all target words sung at the same pitch. Carrier phrases

were recorded at a rate of two syllables per second (one beat per syllable at 120 bpm), with varied melodies as shown in Appendix A.2.

In order to expose listeners to a continuum of $\sigma 1/\sigma 2$ duration ratios, syllable durations in target recordings were scaled using the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma & Weenink, 2020). Each vowel's duration was iteratively scaled and resynthesized to produce thirteen versions of the target word, each with a total duration of one second but with $\sigma 1/\sigma 2$ duration ratios ranging from 2:6 to 12:6 with step size 1:6. These endpoints were chosen such that both the low end and the high end have a natural-sounding ratio with respect to a western musical context, and so that the continuum includes Q1, Q2, and Q3 speech ratios as well as 1:1 (6:6), the most commonly-used $\sigma 1/\sigma 2$ note length ratio observed in the corpus exploration. This process resulted in a total of $20 \times 11 = 220$ unique targets. Targets and carrier phrases were truncated at the increasing zero-crossing closest to each word boundary involved in the insertion of target words into carrier phrases. All possible combinations of target continua and carrier phrases ($220 \times 6 = 1320$) were concatenated using Praat. The concatenated stimuli ranged in duration from 4.3 to 6.7 ms.

5.1.3 Procedure

The experiment was developed using the jsPsych (de Leeuw, 2015) library, accessed online, and delivered in Estonian. Upon accessing the experiment and completing a consent form and an audio test designed to succeed only if the listener is wearing headphones, each participant answered a brief demographic questionnaire (including questions about age, language background and current use, and musical training) and was given instructions for the task. The experiment consisted of two practice trials followed by 180 experimental trials.

From a linguistic perspective, the ratios of primary interest to include in the set of stimuli were in the range [4:6,9:6], corresponding to Q1 and Q2 ratios in speech. As mentioned above, this interval was extended to include musically natural endpoints of 2:6 and 12:6. However, for the sake of keeping the experiment at a maximum duration of approximately twenty minutes, each participant heard targets through the continuous range of continuum steps in either the interval [2:6,10:6] or the interval [4:6,12:6]. For each participant, the twenty targets were randomly distributed as evenly as possible across the six carrier phrases; those associations varied between but not within participants. Each participant heard a crossing of 20 targets (in their assigned carrier phrases) x 9 continuum steps = 180 stimuli. These 180 trials were split into five blocks, between which participants had the opportunity to take a short break.

For each trial, the participant was asked to listen to the sung phrase while two (disabled) buttons were displayed on screen, one showing the Q1 version of the target word (orthographically presented) and the other showing the Q2 version (see Figure 4). The

Figure 4: Participants select which version of the target nonce word they heard. Translated instructions: *Click your selection once audio has finished playing.*

arrangement of these buttons (left vs right) was randomized per participant and remained the same throughout each participant's experience. One the audio ended, the buttons were enabled and the participant clicked the one indicating which version of the target word they heard. The trials did not have a set time limit; the experiment did not move forward to the next trial until the participant clicked their selection.

5.2 Results

Of the sixteen speakers who participated, three were omitted due to completing less than 80% of the full experiment. Each of the thirteen remaining participants completed all 180 trials. Of the remaining observations, those with reaction times greater than three standard deviations above the mean (11.2 seconds, measured from the start of audio play) were excluded; this resulted in 34 of 2340 observations being discarded.

In order to determine whether note-length ratio has a significant positive effect on perceived quantity, a mixed effects logistic regression model was fitted to the data in R (R Core Team, 2020), using the lme4 package (Bates et al., 2015). The outcome variable is long_id (whether the participant identified the target word as having a long V1; i.e., being in Q2), with predictor variable `ratio_num` (first term of the note-length ratio of the target word's two syllables, with second term normalized as 6). Random intercepts, as well as random slopes over `ratio_num`, by each of `nonce`, `carrier`, and `participant` were included. In addition, the optimx optimizer was used, with method L-BFGS-B. According to this model,[2] there is a significant positive effect of `ratio_num` on `long_id` ($\beta = 1.30$, $SE = 0.11$, $z = 11.56$, $p < 0.001$); for every increase of one unit in the numerator of the duration ratio, the odds of a listener identifying a target as being in Q2 increase by a factor of $e^1.33 \approx 3.78$.

Figure 5 (produced using the package ggplot2 (Wickham, 2016)) shows the effect plot of `ratio_num` on `long_id`, with the categorization curve in the context of both the

[2] `long_id` $\sim$ `ratio_num` + (1 + `ratio_num`|`nonce`) + (1 + `ratio_num`|`carrier`) + (1 + `ratio_num`|`participant`)

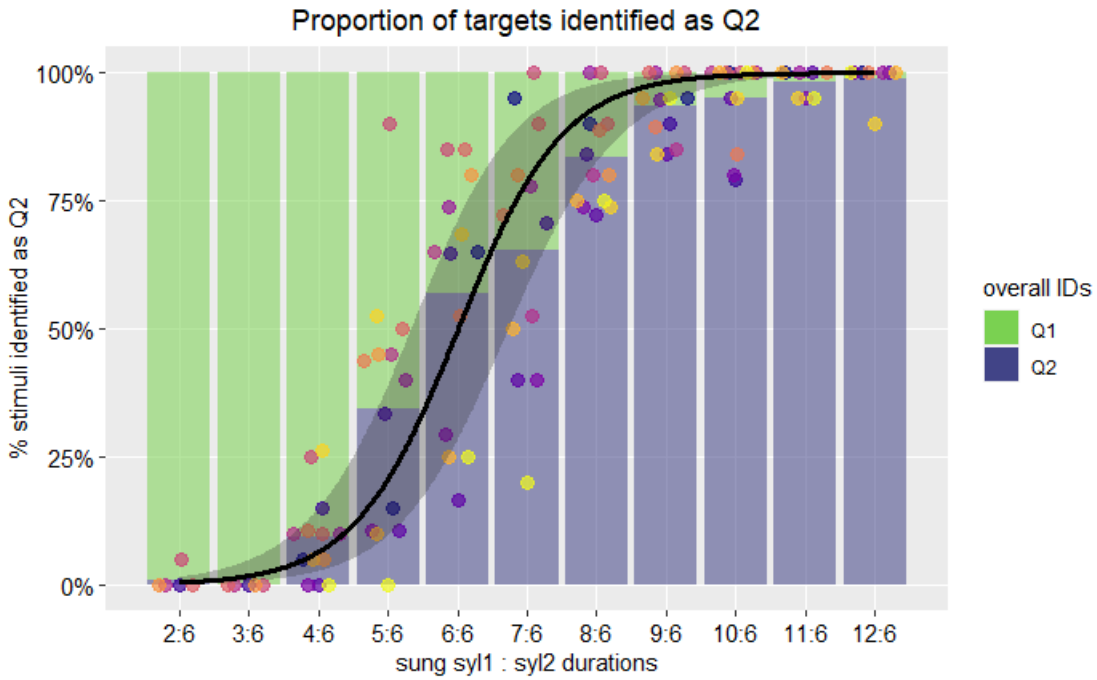


Figure 5: Effect of note-length ratio on vowel quantity identification. Bars show overall proportions of Q1/Q2 identifications; points show proportions by participant.

raw binary responses for `long_id` as well as the by-participant means (with the proportion of Q2 identifications calculated for each participant at each ratio).

A more precise calculation of model predictions shows that the crossover point is at about 6.01:6 (that is, almost exactly at 1:1). Thus a 1:1 ratio is just about as likely to be identified as Q1 as it is to be identified as Q2. In addition, ratios less than or equal to 3:6 and greater than or equal to 9:6 are within 5% of 0% and 100%, respectively, reflecting the consistent Q1/Q2 identifications that would be expected at the extremes of the continuum.

5.3 Post-hoc explorations

5.3.1 Lower- and higher-interval subgroups

As mentioned in Section 5.1.3, participants heard targets with duration ratios in one of two intervals. In order to determine whether the set of targets that the participants heard had an effect on their perception of those targets, the data were partitioned into two subsets: those whose participants heard targets in the interval [2:6,10:6], and those whose participants

heard targets in the interval [4:6,12:6]. The same regression model[3] from Section 5.1.3 was fitted to both subsets of data, and a similar significant positive effect of `ratio_num` on `long_id` was observed for both groups; see Table 5.

Participants	β	SE	z	p
All	1.30	0.11	11.56	< 0.001
[2:6,10:6]	1.32	0.11	11.74	< 0.001
[4:6,12:6]	1.23	0.11	10.97	< 0.001

Table 5: Model summaries for all participants vs two subsets grouped by ratio continua heard.

The effect plots are shown in Figure 6; they are extremely similar not only to each other, but also to the predictions for the main model in Figure 5. The maximum difference between any of the three predictions (where defined) is 1.73%, well within the range of the standard error for any of the three prediction functions. This shows that the effect of which interval of targets was heard is negligible.

5.3.2 Previous musical training

One possible interacting factor in this task might be a listener's musical training. In particular, previous musical experience might lead to better rhythmic discrimination, resulting in a steeper slope at the category boundary. Five of sixteen participants reported having had some musical training (ranging from 2 to 8 years). The data, including a binary-valued predictor for music background, were fitted to a logistic regression model[4] adapted from the one described in Section 5.1.3. Random intercepts, as well as random slopes over `ratio_num`, by each of `nonce`, `carrier`, and `participant` were included. Random slopes over `music_bg` by both `nonce` and `carrier` were included as well. The optimx optimizer was used, with method L-BFGS-B. An ANOVA comparison showed that including the additional predictor led to significantly improved fit over the originally-defined model ($\chi^2 = 21.52$, $Df = 7$, $p = 0.003$). However, although note-length ratio was a significant predictor of Q2 in this expanded model, musical experience was not; see Table 6 for model summary.

[3] `long_id` $\sim$ `ratio_num` + (1 + `ratio_num`|`nonce`) + (1 + `ratio_num`|`carrier`) + (1 + `ratio_num`|`participant`)

[4] `long_id` $\sim$ `ratio_num` + `music_bg` + (1 + `ratio_num` + `music_bg`|`nonce`) + (1 + `ratio_num` + `music_bg`|`carrier`) + (1 + `ratio_num`|`participant`)

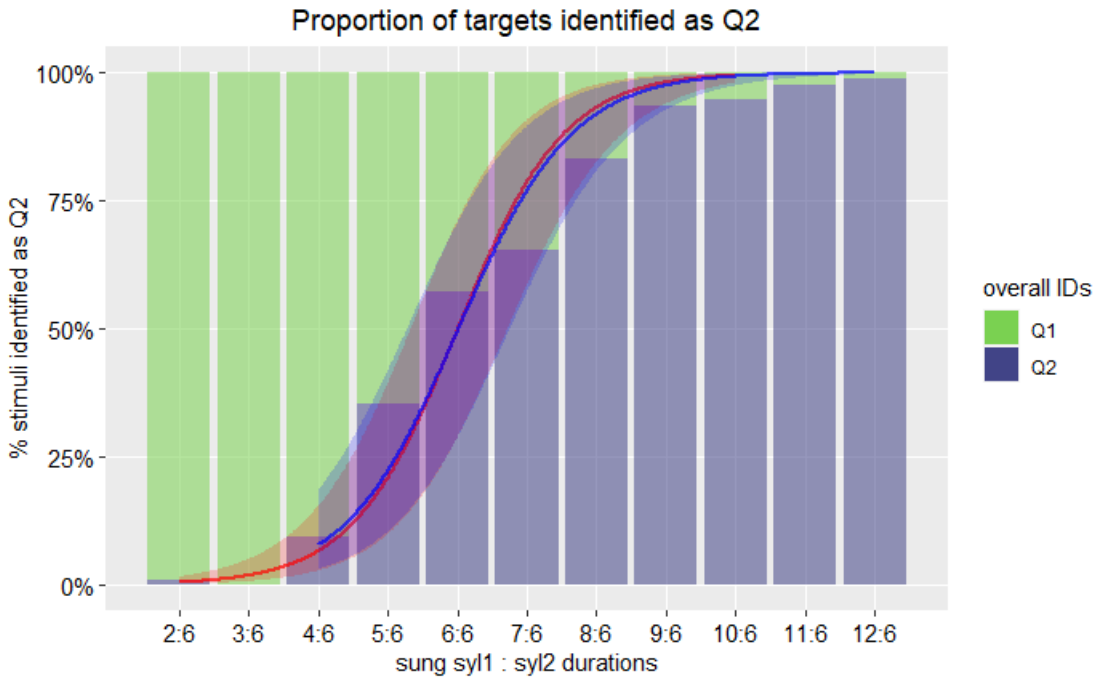


Figure 6: Effect of note-length ratio on vowel quantity identification for two subgroups of participants partitioned by duration ratios heard.

5.3.3 Minimal pairs in carrier phrases

In two of the carrier phrases (phrases 2 and 3; see Appendix A.2) the preceding context for the target word contains a member of a Q1/Q2 minimal pair: in phrase 2, "tema" (3SG.NOM) has the pair "teema" (theme.NOM), and in phrase 3, "tule" (come.SG.IMP) has the pair "tuule" (wind.SG.GEN). The design of the carrier phrases did not intentionally include these words with minimal pairs; however, the presence of such words, potentially ambiguous in their vowel quantity, might have influenced listeners' category boundaries. Curves for the relevant carrier phrases may have been shifted further right than the others, since the preceding context could prime the listener to associate one beat with either a short or a long V1 rather than an unambiguously short V1.

The data, including a binary-valued predictor for the presence of an element of a minimal pair, were fitted to a logistic regression model[5] adapted from the one described in Section 5.1.3. Random intercepts, as well as random slopes over `ratio_num`, by each of `nonce`, `carrier`, and `participant` were included. Random slopes over `minpair` by

[5] `long_id` $\sim$ `ratio_num` + `minpair` + (1 + `ratio_num` + `minpair`|`nonce`) + (1 + `ratio_num`|`carrier`) + (1 + `ratio_num` + `minpair`|`participant`)

Predictor	β	*SE*	z	p
`ratio_num`	1.37	0.12	10.99	< 0.001
`music_bg`	-0.63	0.82	-0.77	0.44

Table 6: Model estimates for predictors `ratio_num` and `music_bg`.

both `nonce` and `participant` were included as well. The optimx optimizer was used, with method L-BFGS-B. An ANOVA comparison showed that including the additional predictor did not lead to significantly improved fit over the originally-defined model ($\chi^2 = 7.01$, $Df = 7$, $p = 0.428$).

6 Discussion

6.1 Corpus exploration

The findings from the exploratory corpus study suggest that preserving speech-based durational ratios is not of primary importance to composers of Estonian music, which could be interpreted to mean that they instead choose text settings that allow for a greater degree of musical expression. There is no significant effect of quantity on assigned note-length ratio in this small sample, though there is a slight positive trend. It is fairly clear, however, that composers do avoid contrary settings such as assigning a Q2 word to a short-long note pair. The ambiguity in how individual words are set rhythmically could potentially be accommodated by listeners due to the availability of the lexical information in the surrounding text. In the presence of lexical clues, duration need not be a strong cue for vowel quantity. The current study comprises a relatively small dataset, especially considering the degree of variation evident for some subsets of the data. A future expansion of this corpus could better illuminate the relationship of note-length ratio to vowel quantity.

6.2 Perception experiment and Estonian speech

The significant effect of note-length ratio on listeners' perception of vowel quantity shows that without the advantage of lexical clues from lyrics, listeners use durational cues to identify sung words as being in Q1 vs Q2. The curve predicted by the model, while similar to the results from work done on perception of quantity in speech, is not entirely identical. Consider the duration ratios determined by Lehiste (1960); that is, 2:3 for Q1 and 3:2 for Q2. This model predicts that a ratio of 2:3 (4:6) will be identified as Q2 about 6% of the time and that a ratio of 3:2 (9:6) will be identified as Q2 about 98% of the time, with ratios more extreme than these having correspondingly more extreme predictions. These results are certainly aligned with Lehiste's 75% agreement rate for Q1 and Q2 identifications.

However, in the interval between these canonical speech-based ratios (that is, from 5:6 to 8:6 inclusive), the predictions are slightly different than would be expected in speech. For instance, 8:6 is also predicted to be identified as Q2 greater than 75% of the time (7:6 is as well, though its lower error limit drops to 64%), suggesting that participants in this experiment were more consistently certain of their identification of a 4:3 duration ratio than participants in Lehiste's speech-based study. Overall, however, the results at the endpoints of the note-ratio continuum are quite well-aligned with Lehiste's results for speech.

6.3 Perception experiment and speech-rate normalization

As discussed in Section 5.2, the crossover point predicted for the categorization curve in the perception experiment is almost exactly at 1:1. Given that we do not have any detailed information about the category boundary in speech from Lehiste's (1960) work, a best guess for where it might be located is at the midpoint between 2:3 and 3:2; that is, at 13:12 (or, in the normalized terms used in this study, 6.5:6), which is very similar to the crossover point predicted by the model described above. Under this assumption, it would be reasonable to predict that a 1:1 duration ratio in speech would, similarly to song, be approximately equally likely to be identified as Q1 or Q2. This is surprising considering the literature on speech-rate normalization and the difference in context of speech vs singing.

In the perception experiment described in this paper, carrier phrases were recorded at an even pace of one syllable per beat, with all syllables being open and having a phonologically short vowel. Words used in the carrier phrases comprised anywhere from one to four syllables; therefore while listening to the phrases - especially the preceding contexts - listeners were consistently being exposed to duration ratios of 1:1 for Q1 words. Based on results from Brown et al. (2015) and Theodore et al. (2009) related to online normalization of speech information, it would be expected that being exposed to this kind of priming would cause listeners to perceive the targets with 6:6 duration ratios as more likely to be in Q1 than Q2.

In a production experiment investigating the effect of speech rate on voice onset time (VOT) for voiceless stops, Theodore et al. (2009) found that talkers differed not only in their characteristic VOTs for utterances produced at the same speaking rate, but also in the rate at which their VOTs increased with decreased speaking rate. Hence listeners must have the flexibility to adjust online to speech rate changes for any given talker.

Brown et al. (2015) showed in an eye-tracking experiment involving English stress patterns, that listeners fixated a target word alternative that would continue the stress pattern they heard in the preceding part of the carrier phrase. Again, listeners were flexibly adjusting their expectations online for suprasegmental information with respect to the target word, primed by the initial portion of the carrier phrase.

Given (a) the speech-based estimate of 6.5:6 for the Q1/Q2 category boundary, (b) the online ratio information being processed as listeners heard the carrier phrases, and (c) the findings of Brown et al. and Theodore et al. with respect to listeners flexibly adjusting to new information and incorporating it into their perception, I would have expected listeners to perceive the targets with 6:6 duration ratios as Q1 with a higher probability than Q2. However, the model's prediction of the category boundary quite near to 6:6 shows that in fact, listeners' perception of sung targets in this interval parallels composers' text-setting habits. In both studies, 1:1 ratios appear to be ambiguous as to whether they are associated with vowels in Q1 or Q2. This suggests that listeners' perception is affected by the fact that they are within a musical context, possibly indicating that the same associations are not made when listening to a musical phrase as compared to a spoken one. It is also possible that perception of Estonian quantity, a ratio-based measure, is less likely to be affected by rate or other cues in context, as shown by Pind (1986) for Icelandic quantity contrasts.

6.4 Further work

The perception experiment described in this paper takes the first steps toward investigating how Estonian listeners perceive vowel quantity in song. From the results above, it is clear that in the absence of any lexical clues, native Estonian listeners perceive quantity in sung syllable pairs with duration ratios in the neighbourhoods of 2:3 and 3:2 the same way they do as in speech. Continuing the investigation into listeners' perception of sung Estonian, in particular assessing to what extent perception through the intervening interval is categorical or gradient, requires discrimination data as well as identification data. Subsequent perception studies, involving both identification of and discrimination between target words presented in isolation, will facilitate a deeper understanding of the Q1-Q2 boundary implied by the results of the identification task discussed herein.

7 Conclusion

No significant relationship was found between vowel quantity and note-length ratio in a small corpus of existing natively-composed Estonian music. Composers, while appearing to avoid setting text to rhythmic patterns that are contrary to expected ratios from speech, do not adhere to any more precise timing relationships in their text setting, using primarily 1:1 note-length ratios for the majority of $\sigma 1/\sigma 2$ pairs. In contrast, there is a significant relationship between duration ratio and identified vowel quantity in Estonian listeners' perception of sung phrases. Greater note-length ratios predicted higher proportions of Q2 identifications, but even with consistent contextual timing information, listeners perceive 1:1 duration ratios in sung Estonian as Q1 with equal probability as Q2. This finding

draws a link between composers' approach to text setting and listeners' perception, and shows that effects demonstrated in speech-rate normalization experiments do not extend to all speech-like contexts.

References

Asu, E. L., & Teras, P. (2009). Estonian. *Journal of the International Phonetic Association*, *39*(3), 367–372. doi: 10.1017/S002510030999017X

Baese-Berk, M., & Goldrick, M. (2009, 05). Mechanisms of interaction in speech production. *Language and cognitive processes*, *24*, 527-554. doi: 10.1080/01690960802299378

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2020). *Praat: doing phonetics by computer [computer program].* Retrieved February 15, 2020, from `http://www.praat.org/` (Version 6.1.09)

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of experimental psychology: Human perception and performance*, *41*(2), 306-323.

de Leeuw, J. R. (2015). jspsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1-12. doi: 10.3758/s13428-014-0458-y

Eek, A. (1983). Kvantiteet ja rõhk eesti keeles (I). *Keel ja kirjandus*, *9*, 481-489.

Eek, A., & Meister, E. (1997, 01). Simple perception experiments on estonian word prosody. In *Estonian prosody: Papers from a symposium* (p. 71-99).

Ehala, M. (2003). Estonian quantity: Implications for moraic theory. In D. Nelson & S. Manninen (Eds.), *Generative approaches to Finnic and Saami linguistics* (p. 51-80). Stanford: CSLI.

Ernesaks, G., & Koidula, L. (2018). Mu isamaa on minu arm [musical score]. In A. Sopp (Ed.), *XXVII laulupeo segakooride laulik* (p. 94-95). Ellington Printing OÜ.

Fox, R. A., & Lehiste, I. (1989). Discrimination of duration ratios in bisyllabic tokens by native English and Estonian listeners. *Journal of Phonetics*, *17*(3), 167-174.

Härma, M., & Karlson, K. F. (2018). Tuljak [musical score]. In A. Sopp (Ed.), *XXVII laulupeo segakooride laulik* (p. 55-60). Ellington Printing OÜ.

Kaskmann, L. (1989). Item 60i (untitled) [musical score]. In E. Laugaste (Ed.), *Vana kannel* (Vol. 6, p. 359-160). Tallinn: Eesti Raamat.

Kirby, J., & Ladd, D. (2016, 01). Tone-melody correspondence in Vietnamese popular song..

Kõrvits, T. (2018). Emakeelelaul [musical score]. In A. Sopp (Ed.), *XXVII laulupeo segakooride laulik* (p. 20-23). Ellington Printing OÜ.

Lehiste, I. (1960). Segmental and syllabic quantity in estonian. In *American studies in uralic linguistics* (p. 21-82). Bloomington: Indiana University.

Lehiste, I. (1989). Current debates concerning Estonian quantity. In J. A. Nevis (Ed.), *FUSAC '88: Proceedings of the sixth annual meeting of the finno-ugric studies association of canada.*

Lehiste, I., & Fox, R. A. (1992). Perception of prominence by estonian and english listeners. *Language and Speech*, *35*(4), 419-434. Retrieved from `https://doi.org/10.1177/002383099203500403` doi: 10.1177/002383099203500403

Lippus, P., Pajusalu, K., & Allik, J. (2007, 08). The tonal component in perception of the Estonian quantity. In *Proceedings of the 16th international congress of phonetic sciences* (p. 1049-1052).

Lippus, P., Pajusalu, K., & Allik, J. (2009). The tonal component of Estonian quantity in native and non-native perception. *Journal of Phonetics*, *37*(4), 388-396.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.

Moran, S., & McCloy, D. (Eds.). (2019). *Phoible 2.0*. Jena: Max Planck Institute for the Science of Human History. Retrieved from `https://phoible.org/`

Pehk, J. (2013). Valss [musical score]. In A. Sopp (Ed.), *XXVI laulupeo segakooride laulik* (p. 32-36). Chaser Print Agency.

Pind, J. (1986). The perception of quantity in icelandic. *Phonetica*, *43*, 116 - 139.

Prillop, K. (2013). Feet, syllables, moras and the Estonian quantity system. *Linguistica Uralica*, *49*(1), 1-29.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raudsepp, I., & Vikat, M. (2009). Joint singing as a phenomenon of Estonian cultural transmission. *Problems of Education in the 21st Century*, *13*, 103-109.

Raudsepp, I., & Vikat, M. (2011). The role of the phenomenon of joint singing in the development of national identity in Estonia. *Procedia Social and Behavioral Sciences*, *29*, 1312-1319.

Ross, J., & Lehiste, I. (1998). Timing in Estonian folk songs as interaction between speech prosody, meter, and musical rhythm. *Music Perception: An Interdisciplinary Journal*, *15*(4), 319-333.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=afex` (R package version 0.28-1)

Spahr, C. (2014, 01). Prosodic epenthesis and floating vowels in estonian quantity [conference handout]. In *Eleventh old world conference on phonology.* Retrieved from `http://individual.utoronto.ca/spahr/spahr_ocp11_handout.pdf`

Starr, R. L., & Shih, S. S. (2017). The syllable as a prosodic unit in Japanese lexical strata: Evidence from text-setting. *Glossa: A Journal of General Linguistics*, *2*(1), 1-34.

Tamberg, E. (2013). Isamaale [musical score]. In A. Sopp (Ed.), *XXVI laulupeo segakooride laulik* (p. 28-31). Chaser Print Agency.

Theodore, R., Miller, J., & DeSteno, D. (2009, 07). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*, 3974-82. doi: 10.1121/1.3106131

Uusberg, P. (2013). Muusika [musical score]. In A. Sopp (Ed.), *XXVI laulupeo segakooride laulik* (p. 17-19). Chaser Print Agency.

Vabarna, A. (2016). Kalajärv [musical score]. In J. Oras & K. Sarv (Eds.), *Eesti rahvamuusika antoloogia.* Eesti Kirjandusmuuseumi Teaduskirjastus. Retrieved from `http://www.folklore.ee/pubte/eraamat/rahvamuusika/ee/075-Kalajarv`

Wedel, A. B., Nelson, N. R., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, *100*, 61-88.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from `https://ggplot2.tidyverse.org`

Wong, P., & Diehl, R. (2002, 10). How can the lyrics of a song in a tone language be understood? *Psychology of Music*, *30*, 202-209. doi: 10.1177/0305735602302006

Appendices

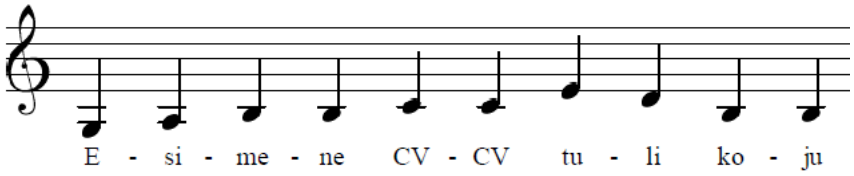
A Perception experiment materials

A.1 Nonce words with corresponding orthographic representations

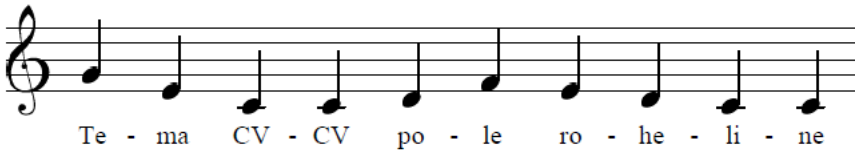
IPA	Orthography (Q1)	Orthography (Q2)
[ke(e)pe]	kebe	keebe
[kø(ø)pe]	köbe	kööbe
[kø(ø)te]	köde	kööde
[ko(o)te]	kode	koode
[kɤ(ɤ)te]	kõde	kõõde
[me(e)ke]	mege	meege
[mø(ø)ke]	möge	mööge
[mø(ø)te]	möde	mööde
[mo(o)ke]	moge	mooge
[mo(o)pe]	mobe	moobe
[mɤ(ɤ)pe]	mõbe	mõõbe
[pø(ø)ke]	pöge	pööge
[pø(ø)te]	pöde	pööde
[pɤ(ɤ)te]	põde	põõde
[se(e)ke]	sege	seege
[se(e)pe]	sebe	seebe
[te(e)pe]	tebe	teebe
[tø(ø)ke]	töge	tööge
[tø(ø)pe]	töbe	tööbe
[tɤ(ɤ)ke]	tõge	tõõge

A.2 Carrier phrases with associated melodies

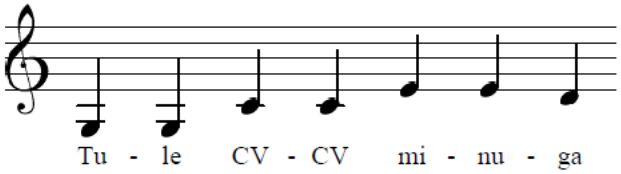
(1) Esimene ___ tuli koju.
esimene ___ tuli koju
"The first ___ came home."



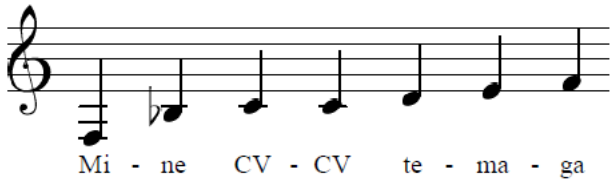
(2) Tema ___ pole roheline.
temɑ ___ pole roheline
“His/her ___ isn’t green.”



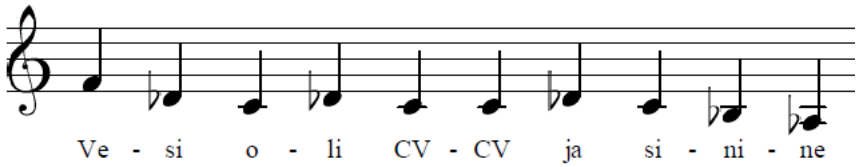
(3) Tule ___ minuga.
tule ___ minukɑ
“Come ___ with me.”



(4) Mine ___ temaga.
mine ___ temɑkɑ
“Go ___ with him/her.”



(5) Vesi oli ___ ja sinine.
vesi oli ___ jɑ sinine
“The water was ___ and blue.”



(6) Tähe ___ sära oli imeline.
tæhe ___ særɑ oli imeline
“The star’s ___ shine was wonderful.”