the acoustic correlates of vowel prominence in concord/conflict at song/syllable levels: Estonian *regilaul* folksong meter.

sally ransom

University of Texas at Austin

May 12, 2022

I don't use folksongs,

rather,

folksongs use me

-Veljo Tormis, Estonian composer and *regilaul* revitalist

Abstract

This corpus phonetics study of Estonian folk songs examines vowel space and duration in conflict and concordance at the song and the syllable-levels of prosodic prominence. Songs are useful for examining acoustic correlates of stress because they are restricted in the domain of time, having less temporal variation than natural "running" speech. By examining the interaction of temporal and spatial acoustic cues in various positions, we can see how a singer accomodates prosodic conflicts to preserve contrasts in the language. The paper first introduces prosodic prominence in general, then delves into the metrical specifics of Estonian and its folksong tradition. After outlining the research questions, the methods section describes the corpus of *regilaul*: its construction, annotation, and processing. Results show that for vowel duration, * is the best predictor of vowel length. In the spatial domain, however, * is the best predictor of vowel length.

Introduction

At the interface of language and music is the lyrical song, with natural language participating in the musical domain as the (adjective) resonator for the vocal instrument. In this domain we can examine the effects of the musical context on the language: while bpm varies naturally in human performances, phrases in musical contexts are more consistent than in running speech . Both music and language share the notion of rhythmic prominence, with certain beats (or syllables) carrying extra weight or "stress."

A primary focus of research in metrical prosody is the acoustic properties of stress or prominence at different levels of the prosodic heirarchy and its bounds. In a language, syllables have a prominence status at every level: every foot has a strongest syllable, each word has a strongest foot, each clause a strongest word, and on and up. When lyrics are put to a melody, an additional level is added to the prosodic hierarchy, this one with heavy restrictions in the temporal domain, subordinating all lower prosodic levels to that of the song.

This paper takes the interface of music and language as an opportunity for an exploratory corpus phonetics study of syllable prominence using the Estonian language, which has three syllable quantities, a predictable stress pattern at the word level and a robust tradition of folksongs *regilaul* which follow a strict metrical pattern.

Acoustic Prominence

At present, typological data shows that across languages, syllable prominence is indicated by a convergence of several cues: i.e., duration, spectral balance, vowel space, and onset consonant length Sluijter and van Heuven (1996); Gordon (1997); Liberman and Prince (1977); Berinstein (1979); de Jong (2004). Syllable prominence can also be thought of as localized hyperarticulation de Jong (1995). It is well documented that speakers

spontaneously adjust the articulatory and acoustic dimensions of their speech to a given context, i.e., when communicating to someone who is hard of hearing, to an L2 speaker, or in a noisy restaurant. **?**. In these situations, certain acoustic dimensions are exaggerated: expanding the vowel space, lengthening certain duration cues, increasing the intensity of certain frequencies . Conversely, in facilatory situations such as speaking to someone very familiar, the opposite (hypo articulation) happens, where certain acoustic elements are diminished. Lindblom (1990). Decades since H & H theory, an abundance of research has shown gradient acoustic effects of speech modification. Rather than a manner of speaking in [plusminus] difficult context, we see a convergence of cue weighting and adjustments along a several continua of contexts.

We can think of strong syllables as hyperarticulated compared to their weaker (hypoarticulated) peers, with the strongest syllable at the highest level the most hyperarticulated compared to its strong peers.

When a listener hears a syllable in some sort of conflict, such as word-stressed but song-unstressed, they may or may not perceive it as conflicting or concordant. However, we might assume the singer's production to contain acoustic evidence of the conflict if they are a native speaker of the language they are singing. The null hypothesis is that song and word-level prominent syllables are no different from each other: that is, the highest prosodic hierarchy in this case the song) will dominate prominence in the

temporal domain, confirming the earlier Ross & Lehiste studies where a syllable's position in the song's metrical pattern is the best predictor for syllable duration.

Temporal and spectral correlates of prominence:

Stress in Estonian

Syllables in Estonian have three quantities, resulting in three types of durationally contrastive feet, so that the same bisyllabic sequence can mean three different things depending on the quantity. Ross and Lehiste (1998).

Word type	Q1	Q2	Q3
Vcv(cv...)	/sata/	/saata/	/saa:ta/
	'hundred'	'send' 2^{nd} p. sg. imper.	'to get',
vCv(cv...)	/lake/	/lakke/	/lak:ke/
	'bare'	'thin gruel' nom. sg,.	'ceiling' ill. sg.

Figure 1: Three Way Quantity Contrast (*Krull, 1999)*

There is a documented tendency for foot isochrony, a consistency in duration of feet in Estonian which results in an inverse relationship between the durations of the two syllables in a foot: the longer the first syllable, the shorter the second. This is realized acoustically by the duration ratio of first and second syllable in a foot, i.e., second syllables that follow an overlong syllable will be realized shorter than a second syllable that follows a long syllable.

- Q1 – short ratio 2/3 (short-long(er))

- Q2 – long ratio 3/2 (long(er)-short(er))
- Q3 – overlong ratio 2/1 (long(er)-short(er))

In list reading speech, the duration ratio is greater between short and long syllables, but in conversational speech this is only seen in situations where the bearer of quantity contrast is a vowel, and f0 is available as a secondary cue for quantity. In conversational speech when the contrast bearer is a consonant, the duration ratio is greater between long and overlong syllables.

With few exceptions, such as in borrowed words, Estonian stress is predictable from the following rules:

- if the syllable is in the third quantity, stress falls on its nucleus.
- otherwise, primary word stress is on the first syllable.

Estonian is mostly trochaic, with main stress falling on the first syllable and secondary stress falling iteratively on odd-numbered syllables Lehiste (1965). In other words, a left-edge, quantity sensitive iterative stress pattern.

Kalevala Meter

The Kalevala meter is part of the musical tradition of both Estonia and Finland. The main element in the structure of the songs is the verse line, consisting of eight positions divided between four (usually trochaic) feet. From the point of view of musical rhythm, the majority of old folksong melodies are roughly isochronous, i.e. consisting of notes of about the same

duration. In most cases, each of the eight positions holds one syllable for one melody note. As an exception, two syllables may fit one note, or a syllable (usually a diphthong) may be divided between two notes. This is readily evident in the musical score, which is divided into four or eight. Ictus refers to a position that is stressed in a song, and off-ictus is an unstressed position in a song. In *regilaul* songs, which make use of the Kalevala meter, the pattern is trochaic: ictus starts on the first beat and is applied to every other syllable of a phrase.

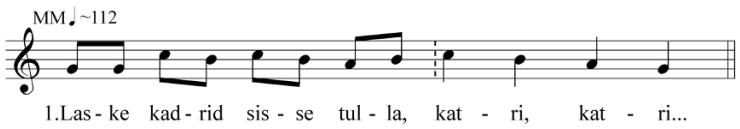

Figure 2: one *regilaul* phrase in music notation

[...]

Previous Work

A previous study in Estonian Ross and Lehiste (1996) found that duration was a better predictor of musical verse position than of word stress: stressed syllables in off-ictus lost their durational contrast with unstressed syllables. When syllable duration is subordinated to the song meter, what acoustic modifications does the singer make to preserve word stress? A previous study Ross and Lehiste (1996) examined the relationship of the conflict between word stress and metrical ictus (stress position in the song), finding that word-level duration cues were subordinated to the musical prominence

pattern.

Another study, Ross and Lehiste (1994) compared S1/S2 ratios for all three quantity degrees in three Estonian funeral laments, finding that duration was best predicted by metrical position and not syllable quantity. Ross and Lehiste (1996) Syllables in ictus position were longer than syllables in off-ictus position measured syllable duration ratios (S1/S2) for two categories of Q1 (short) words: those with initial syllable(which would be stressed in speech) falling in ictus position and with initial syllable falling in off-ctus position (resulting in a conflict between word stress and metrical ictus). They found that the duration ratios of Q1 words that started in ictus was greater than that which started in off-ictus. Q1 words starting in ictus position tended to have notes of approximately the same duration, while those that started in off-ictus position tended to shorten the initial syllable.

More recent work has found that proportional duration increases **?** between adjacent syllables is a more robust metric than simple duration ratio. This paper aims to extend and increase the n of the aforementioned studies of these songs' prosodic hierarchies by annotating a larger corpus of *regilaul*, to compare measurements of syllable duration ratios and proportional duration differences of syllable quantity and prominence in songs.

The extension of the findings arises after confirming the duration results. If duration is not the contrastive cue for syllable stress or quantity in songs, what (if any) are the acoustic cues for syllable prominence and quantity contrast? Therefore in addition to duration, vowel space and spectral tilt will be measured in syllable vowel nuclei.

If duration is subordinated to the prosody of the music, we can predict that another cue to linguistic stress will be retained in the singer's production: for this paper, we will look at vowel quality.

Research Questions and Hypotheses

The hypothesis is that stressed syllables that fall in off-ictus positions will be have raised intensity in higher frequency bands Sluijter and van Heuven (1996) due to constricted vocal folds, and will be hyperarticulated Lindblom (1990) compared to unstressed syllables in those same positions.

- Stressed syllables that fall in off-ictus positions will be hyperarticulated (higher high vowels, lower low vowels) compared to unstressed syllables in those same positions.
- Duration will be subordinated to metrical structure, so vowel duration will be predicted more accurately by song position than word stress.
- Stressed syllables in off-ictus positions will have an increase in intensity at higher frequencies than unstressed syllables in those same positions.

[...]

Methods

This section is outlined as follows: first, the corpus of *regilaul* songs is described in detail, including information about the archival source and collectors, the performers in the recordings, text transcriptions of the lyrics, and the digital audio signal. Then, the annotation methods for phrase, beat, word, syllable, and segment are demonstrated, proceeding finally to the acoustic measurements.

Materials

The Anthology of Estonian Traditional Music (Tampere, n.d.)provides an overview of the earlier folk music tradition of Estonia, providing a sample of lyrics, English translations, and archival audio (.ogg) recordings of 98 *regilaul* songs and 17 instrumental tunes collected and compiled across Estonia's many parishes by Herbert Tampere, Erna Tampere, Ottilie Kõiva between 1912 and 1966 for the Estonian Folklore Archives in Tartu, Estonia.

For this project, a sample of 9 *regilaul* songs were chosen by the following criteria: all were recorded between 1960 and 1965 (likely with same or similar equipment) in the same region (Parnumaa, which according to an informant is a dialect with the aforementioned stress pattern). The singers were three women aged 67 and 92 (avg 75) years, and long-time residents of Parnumaa county at the time of the recordings.

Preparing Audio

After downloading from the archive, the files were converted from .ogg to .wav files using sox, as PRAAT**?** (Boersna, n.d.) cannot read the null bytes contained in .ogg files.

0.0.1 TEMPO MAPPING

First, I used the flex tempo feature of Logic Pro X to find the best fit of beats and tempo, adjusting the map to fit the downbeat of each measure and fitting the variable notes in time. Using this map enabled me to make a metronome of sorts that followed the true tempo of the performance. This midi metronome was programmed to strike on the down beat and at every other syllable division (given the melody of the song in question). Song data is available in the appendix. This midi file was then trimmed to match the song file exactly, so that PRAAT could automatically annotate sounds and silences for each song based on its variable tempo metronome. This had the effect of marking every ictus syllable with a simple script. The textgrids from this script were then the basis for the analysis tiers: first, every eight beats is taken as a large phrase interval, its appropriate lyric line inserted. Using the built-in eSpeak forced-aligner for Estonian, phrases were annotated to words, checked for accuracy, and then annotated to segments. Ictus and off-ictus boundaries were adjusted during the segmentation process to match onsets and offsets. First and second syllable vowels were annotated to new tiers, and then the intersection of ictus and first syllable, ictus and second syllable, off-ictus and first syllable, and off-ictus and second syllable were

taken for analysis. From this intersection, midpoint formant measurements were extracted, as well as spectral tilt.

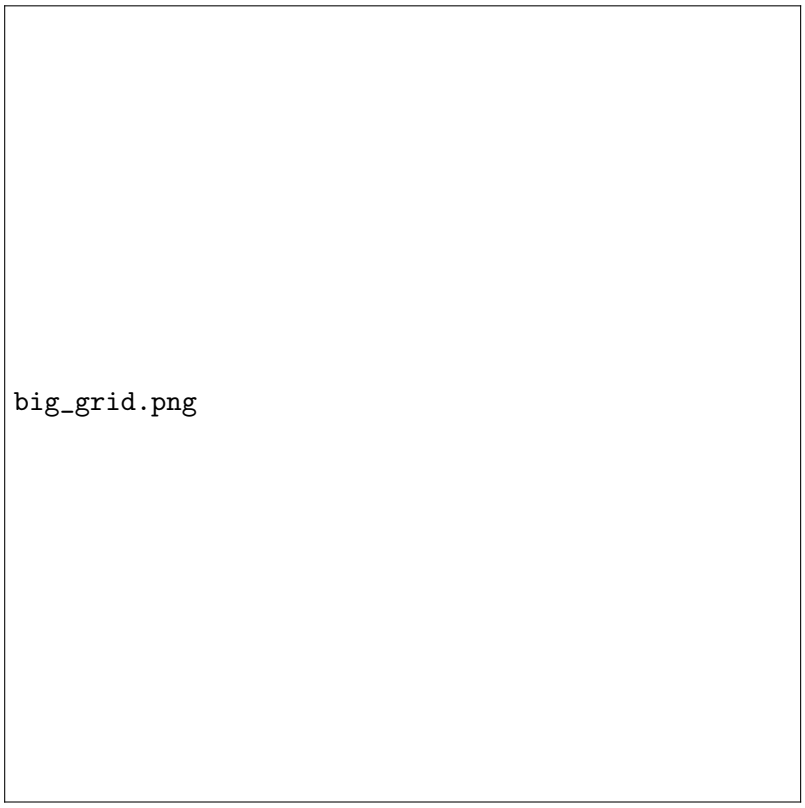

Abbildung 3: whole song

Syllable Segmentation, Temporal and Spectral Measurement Criteria

The following criteria were used to adjust the results of the forced aligner when the segment boundaries were verifiably off.

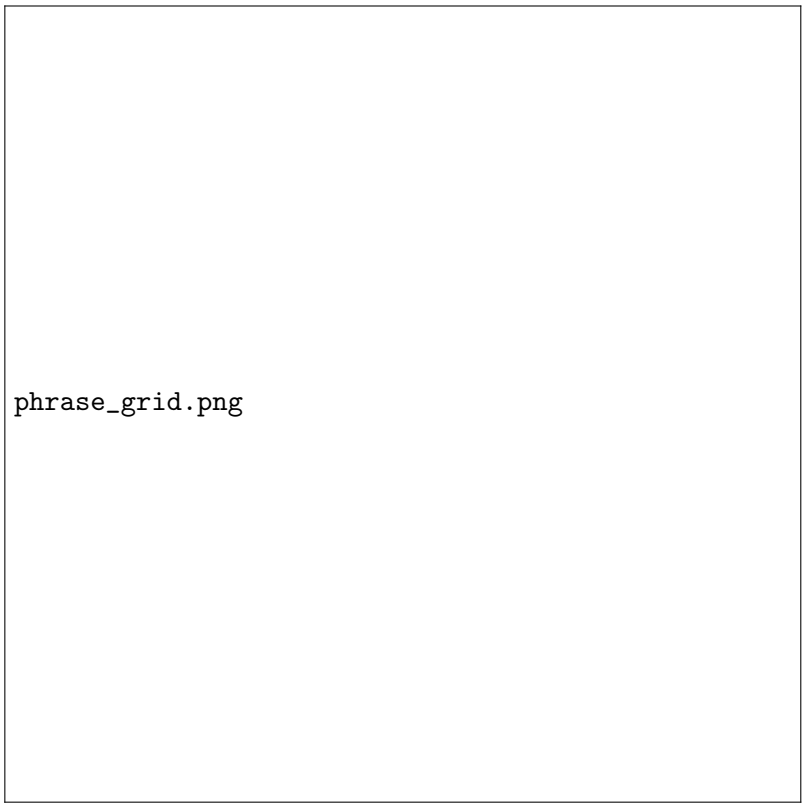

Abbildung 4: single phrase

Segmental/Temporal

onsets:

Oral stops at phrase boundaries are excluded. Due to the continuous nature of singing, most words with oral stop onsets have enough preceding acoustic energy to determine the boundary, but care is still taken to exclude stop consonants at word boundaries adjacent to pauses.

vowels:

The criteria for a vowel onset following stop consonants is at the beginning of the release burst, and at the clear increase in amplitude in the waveform and visible formants in spectrogram following other obstruents. Following nasals, vowel onset is defined at the boundary of detectible anti-formants. The boundaries of vowel onsets following liquids and glides are disregarded and instead these sequences are analyzed as a diphthong.

The criteria for the offset of a vowel preceding stop consonants is defined at the drop in acoustic energy upon closure preceding the small tail of f0 dropoff. Vowel offset preceding other obstruents is at the boundary between visible vowel formants and noisier acoustic energy, and for nasals at the boundary between vowel formants and antiformants. Preceding liquids and glides, no boundary is determined between vowel and coda, these are segmented with the vowel and analyzed as a diphthong. Long vowels are annotated as geminate, with the boundary between the two identical vowels being considered arbitrary, and the midpoint of the full double sequence

used as the location of spectral measurements. coda:

liquid codas are included in total syllable duration, analyzed as a part of the vowel

Spectral Measurement Parameters

tilt, cog [..]

Preparing Text

The lyrics of each song are downloaded in a text file and aggregated into a corpus of songs together with annotated and transcribed audio. Using estnltk's Estonian toolkit, text data is used to filter and sort the acoustic data by quantity, syllable boundary, syllable-word index (i.e. stressed or not in speech) and syllable-phrase index (ictus or not at this metrical position in the song).

syllabification

At this point, the corpus is annotated for metrical position, phrase, word, and segment. However, in order to examine duration, a syllable quantity and stress annotation are necessary. Using an estonian version of nltk. https://github.com/estnltk, which has a useful automatic "varba-morfßyllabifier library. The output of the syllabification in this module includes syllable quantity and prominence data as well as the phoneme segments in each respective syllable. With the quantity annotated, duration measurements between syllable quantities can be compared with each other, to see if results are similar to Ross and Lehiste (1996).

Results

0.1 Vowel Duration

Vowel Space (F1 and F2)

Discussion

0.2 Goin' Fishin' *some exploratory analysis I'd like to do with the data I've got*

Stress in Estonian is acoustically most often realized by onset consonant lengthening (Gordon, 1997; Hint, 1973). Gordon 1997 also found evidence for acoustic correlates of prosodic domains by measuring peak nasal flow, nasal amplitude, and duration in initial positions of four prosodic domains. The syllable, word, phrase, and utterance, were each also cross-classified for stressed and unstressed positions.

functional load hypothesis (Berinstein, 1979),
but another cross-linguistic study specifically analyzed the acoustic correlates of stress in languages that have a duration contrast, and found no support for the functional load hypothesis (Lunden et al., 2017). Thus this study has another potential theoretical bearing to weigh in on.
.The lowered amplitude of initial nasal segments (also found by (Fougeron 1996)), Gordon argues, is to maximize consonantness compared to the sonorance of the following vowel. Nasal data provided evidence for progressively larger domains of prosodic constituency: syllable, word, and phrase.

Previous research has shown lengthening of domain-final segments (Oller 1973, Beckman et al 1992, Wightman et al 1992), wherein the higher the prosodic level, the greater the lengthening effect in general. Will this remain the case at different levels of the prosodic domain of the song?

0.3 Syllable index within phrase compared to beats detected

- rate of beats/syllables in each song
- occurence of beats in odd (ictus) syllables
- beats detected outside of ictus
- overall robustness of the beat detection

0.4 Duration Ratios:

metrical analysis of text database of *regilaul*

a text database of *regilaul* songs: https://www.folklore.ee/regilaul/andmebaas/?ln=en. With this, I will analyze the frequency of stress and ictus conflicts in the larger pool of songs. For example, are there more occurrences of syllables strong at the word level ending up in weak metrical positions, or more weak word-level syllables in strong metrical positions? How often do overlong (Q3, heaviest possible) syllables end up in metrically weak positions, if at all?

At the very least, this database can provide some useful data, such as lexical frequency within the broader set of regilaul songs to provide a gradient picture of the present corpus in the context of *regilaul* songs on the whole.

0.5 Growing the Corpus

Available on the archive are eight more songs from the same region and time period, which would bring the total songs to 17 *regilaul* songs and seven singers. Once these annotations are aligned and verified, the audio corpus will total thirty-two minutes and twelve seconds of recorded audio.

Conclusion

[...]

Literatur

Berinstein, A. E. (1979). WPP, No. 47: A Cross-linguistic Study on the Perception and Production of Stress.

de Jong, K. (2004). Stress, Lexical Focus, and Segmental Focus in English: Patterns of Variation in Vowel Duration. *Journal of Phonetics*, 32(4):493–516.

de Jong, K. J. (1995). The Supraglottal Articulation of Prominence in English: Linguistic Stress as Localized Hyperarticulation. *The Journal of the Acoustical Society of America*, 97(1):491–504.

Gordon, M. (1997). PHONETIC CORRELATES OF STRESS AND THE PROSODIC HIERARCHY IN ESTONIAN. *Estonian prosody: Papers from a symposium*, pages (pp. 100–124).

Lehiste, I. (1965). The Function of Quantity in Finnish and Estonian. *Language*, 41(3):447.

Liberman, M. and Prince, A. (1977). On Stress and Linguistic Rhythm. *Linguistic Inquiry*, 8(2):249–336.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, NATO ASI Series, pages 403–439. Springer Netherlands, Dordrecht.

Ross, J. and Lehiste, I. (1994). Lost Prosodic Oppositions: A Study of Contrastive Duration in Estonian Funeral Laments. *Language and Speech*, 37(4):407–424.

Ross, J. and Lehiste, I. (1996). Trade-off between quantity and stress in Estonian folksong performance? *Folklore: Electronic Journal of Folklore*, 02:116–123.

Ross, J. and Lehiste, I. (1998). Timing in Estonian Folk Songs as Interaction between Speech Prosody, Meter, and Musical Rhythm. *Music Perception*, 15(4):319–333.

Sluijter, A. M. C. and van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, 100(4):2471–2485.