Phonetic tone signals phonological quantity and word structure

Martti Vainio[a)]
Institute of Behavioural Sciences (Phonetics and Speech Synthesis Research Group), University of Helsinki, P.O. Box 9, Helsinki FI-00014, Finland

Juhani Järvikivi
Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Daniel Aalto
Institute of Mathematics, Aalto University School of Technology (Helsinki University of Technology), P.O. Box 1100, FI-02015, Finland

Antti Suni
Department of Speech Sciences, University of Helsinki, P.O. Box 9, Helsinki FI-00014, Finland

(Received 8 August 2009; revised 2 July 2010; accepted 2 July 2010)

Many languages exploit suprasegmental devices in signaling word meaning. Tone languages exploit fundamental frequency whereas quantity languages rely on segmental durations to distinguish otherwise similar words. Traditionally, duration and tone have been taken as mutually exclusive. However, some evidence suggests that, in addition to durational cues, phonological quantity is associated with and co-signaled by changes in fundamental frequency in quantity languages such as Finnish, Estonian, and Serbo-Croat. The results from the present experiment show that the structure of disyllabic word stems in Finnish are indeed signaled tonally and that the phonological length of the stressed syllable is further tonally distinguished within the disyllabic sequence. The results further indicate that the observed association of tone and duration in perception is systematically exploited in speech production in Finnish.
 [DOI: 10.1121/1.3467767]

PACS number(s): 43.70.Fq [AL] Pages: 1313–1321

I. INTRODUCTION

In addition to contrastive lexical features like segmental quality (pig, big, peg, beg), some languages exploit suprasegmental means, such as duration and fundamental frequency, in signaling word meaning. Traditionally, duration and tonality (or pitch) have been taken as mutually exclusive: whereas the latter is prototypically associated with tone languages like Chinese, segmental duration has been taken as the main vehicle to signal lexical contrasts in so-called quantity languages like Finnish. Consequently, the presence or absence of phonological-lexical tone forms one of the fundamental divisions in the (phonological) typology of languages.[1,2] Furthermore, it is assumed that in the latter type of non-tonal languages fundamental frequency is used only post-lexically to signal various discourse and information structure functions, such as e.g., emphasis or focus.[2] Moreover, some research suggests that listeners are specifically tuned to the phonological categories of their native languages, even to such extent that they may have problems identifying contrasts that are alien to their native language system.[3]

In contrast, however, some evidence indicates that, in addition to durational cues, phonological quantity is associated with and co-signaled by changes in fundamental frequency, as suggested by earlier evidence for Finnish, Estonian, and Serbo-Croat.[4,5] Recently, Järvikivi *et al.*[6] investigated the influence of tone on the perception of quantity using a two-alternative forced-choice categorization task. They manipulated both the tonal structure and segmental durations within pairs of Finnish disyllabic words, e.g., *sika* 'pig' vs. *siika* 'whitefish'. The results showed a clear influence of tone such that a falling contour tone was likely to induce more "long" syllable categorizations than a static high tone in a disyllabic sequence where the second syllable always had a low horizontal f_0. Furthermore, the results indicated that there is quite a large durationally ambiguous space where the tonal structure may serve as the only cue for phonological length. The results suggest that both the durational and tonal differences responsible for the perception of quantity within a word are distributed within the disyllabic stem of the word. Thus, despite duration being a salient feature for e.g., speakers of Finnish, perceptual sensitivity to the length opposition in, e.g., (short) *tuli* 'fire'-(long) *tuuli* 'wind', seems to be modulated by a difference in melody.[6] Furthermore, even though it cannot be conclusively ruled out that this is due to a more general, perhaps psycho-physically based, propensity to "hear" contour tones as longer than static ones in general independent of the language in question[7,8] recent evidence suggests that this association may even be unique to speech sounds and perhaps especially systematic in languages with phonological quantity.[6,9,10]

Importantly, however, whether the observed association of tone and duration in perception is fundamental to certain

[a)]Author to whom correspondence should be addressed. Electronic mail: martti.vainio@helsinki.fi

types of phonological systems or a general psycho-physical trait characteristic of human perceptual mechanisms, this association must be systematically exploited in speech production in order for people to be able to benefit from it in perception at all. Yet this question has not been studied systematically as far as we know. In what follows we will investigate these issues in Finnish.

Finnish is a full-fledged quantity language with two degrees of length for practically all sounds in the language and applicable in almost all positions within a word. The majority of Finnish non-derived roots are disyllabic, most of which have a heavy (two or more morae in the syllable rime) first syllable.[11] In phonological terms the heavy syllables contain two or more segments after the onset, which is never counted as a mora in Finnish. Thus the most common word patterns in Finnish are CVC.CV, CVV.CV and CV.CV (where the full stop stands for a syllable boundary). Lexical, or rather, word stress is on the first syllable of the word by rule; i.e., the first syllable invariantly receives utterance or sentence level stress. Due to the fixed place on the first syllable there are no stress based minimal pairs. Otherwise Finnish word stress behaves similarly as in Germanic languages, like English. The standard assumption regarding quantity is that the length difference can be explained by the durations of different segments and their mutual relationships.[12] In this view no tonal differences should exist between the vowels of the two quantity categories.

In a recent series of studies, Suomi[13–15] argues that Finnish disyllabic words follow a uniform tonal pattern that is timed according to their moraic structure. According to Suomi, there is a pitch rise during the first mora of the word and a fall during the second mora. Therefore, both the rise and the fall then either occur during a single (heavy–two or more morae, e.g., CVV, CVC) first syllable, or the fall during the second syllable, if the first syllable consists of only one mora (e.g., CV). Thus, not only would all Finnish word types be uniform with respect to their tonal form, but segmental durations of the stressed syllables would in fact "conspire" to guarantee the tonal uniformity across different word structures within a disyllabic foot. Moreover, in this view the tonal uniformity afforded by the durational conspiracies would serve only accentuation and subsequently the prominence relations within an utterance. Thus, rather than seeing the tonal realization of the accent varying as a function of the segmental structure of the stressed syllable, this account takes on the view that it is the segmental durations that vary in order for the accent to be realized temporally. However, as we noted above, the tonal differences within the nucleus of the stressed syllable have been shown to be relevant for the perception of quantity in both Finnish[5,6] and Estonian.[4,16] It is not clear how the above approach would account for these perceptual effects. Second, there is evidence indicating that the tonal means to achieve prominence in the form of a contrastive accent/prosodic focus vary depending on the position of the contrasted words in an utterance.[17] Namely, utterance initial narrowly focused words behave differently from utterance final focused words; the initial focus is characterized by a rise, whereas the final focus is characterized by a fall. As Suomi has mainly studied contrastively accented words, it is not entirely clear whether the results would generalize to non-contrastive words or the position of the word in an utterance.

We would like to pursue an alternative view; that tonal shape may not be relative to the word alone, but would instead be a systematic feature related to marking quantity. Thus, it could be that even in a quantity language like Finnish long stressed vowels (and possibly bimoraic syllables in general) are co-signaled by a falling tone as opposed to a static level tone in short (one mora) stressed syllables in production. It is, therefore, important to see whether these tonal differences are due to similar articulations or if there are, in fact, different underlying structures that are being signaled tonally during the stressed syllable. Moreover, for the tonal differences to be communicatively useful, they should be robust with respect to external factors: namely, accentuation and speech rate.

In summary, we expect the tonal structure of a disyllabic Finnish word stem to phonetically mirror a similar sequence of tones as in e.g., Mandaring Chinese. Thus we can define *phonetic tone* operationally as a tonal pattern that co-signals a lexical contrast.

II. EXPERIMENT

In order to investigate the above issues, we carried out a production experiment in Finnish. Participants produced pairs of nouns and adjectives (CV.CV vs. CVV.CV or CV.CV vs. CVC.CV) embedded in carrier sentences. In addition, the impact of accentuation was studied by using verbs, because–unlike nouns and adjectives–verbs tend to be only weakly accented in an unmarked context. That is, verbs in Finnish tend to have similar tonal shapes as nouns and adjectives, but with smaller tonal movements. In addition, the effect of speech rate was studied by repeating the material using slow speech.

A. Materials and procedure

Four sets of nouns and adjectives were selected from the Turun Sanomat lexical database[18] as follows: Twenty-four word pairs were selected for the CV-CVV (conditions CV_a-CVV_a; e.g., /pu.ro/-/puu.ro/; 'stream'-'porridge'), and twenty-one pairs for the CV-CVC (conditions CV_b-CVC_b; e.g., /ka.ma/-/kam.pa/; 'stuff'-'comb').

In the CV_a-CVV_a condition the words differed with respect to the length of the vowel in the first syllable. In the second condition (CV_b-CVC_b) the words in each pair were identical except for the extra syllable-final consonant in the latter words. The onset consonant of the second syllable [(l),(r),(m) or (n)] was always the same for both members of a pair.

Using liquids and nasals at the end of the bimoraic syllable ensured that the f_0 could be detected throughout the syllable. A further set of 21 word-pairs consisting of verbs was also selected. The verb pairs consisted of CV and CVV words comparable to the CV_a-CVV_a conditions above. Some of the verbs in the CVV condition had diphthongs rather than long vowels. In a moraic analysis they are, however, equally heavy or long. In the subsequent analyses the conditions are

marked as CV_c and CVV_c. In summary, the sets marked CV_a, CV_b, and CV_3 were all CV words, but matched against different bimoraic sets; CVV_a, CVC_b, and CVV_c, respectively.

The words in the sets were matched for lemma frequency (CV_a and CVV_a: mean frequency, 34 and 26 per million, respectively [*t—test*, $t<1$]; CV_b and CVC_b: mean frequency, 40 and 42 per million, respectively). Each word was embedded in a sentence frame keeping the surrounding context identical or as similar as possible; in all contexts at least the words that immediately preceded and followed the target were always identical within a single pair. The target words were, thus, always in an utterance medial position. There were altogether 66 sentence frames with a different sentence frame for each experimental pair. The sentences were counterbalanced between two experimental lists with each list containing only one of the target words per word pair: *tuli* 'fire' vs. *tuuli* 'wind', *kama* 'stuff vs. *kampa* "comb," *nukkua* 'to sleep' vs. *nuokkua* 'to nod'. Both lists had an equal number of target sentences from each condition. For example, one participant encountered the following sentence only once with either the CV.CV or CVV.CV version of the targe word: *Parasta mitä tiedän, on paistettu* ***sika/siika*** *tarjoiltuna perunoilla* 'The best thing I know is fried **pig/whitefish** served with potatoes'.

Twenty-five native speakers of Finnish (24 female and one male) were randomly assigned to one of the lists (13 and 12 per list). The speakers were all students at the University of Helsinki and none reported any hearing problems. Only one of the speakers was familiar with speech prosody research in general, but not with the aim of the present study. Each speaker was paid €7 for their participation.

Each speaker was first asked to read the sentences aloud one by one at a normal speech rate. The reading was then repeated using a slow speech rate: The speakers were instructed to read the sentences as slowly as possible without breaking the utterances into separate words.

The recording was done in a sound proof room at the Department of Speech Sciences of the University of Helsinki using a high quality condenser microphone placed approximately 20 cm from the subjects' mouth. The sound was stored on a computer hard drive using a high quality AD converter.

B. Data analysis

The recordings were split into utterance sized chunks and manually segmented and labeled. A phonetically trained labeler who was not familiar with the study marked either the nucleus of the syllable (CV and CVV cases) or the nucleus and the coda (CVC or closed syllable cases). At this point the labeler discarded 165 tokens due to creaky voice and/or missing f_0 data. For the rest of the data, the raw pitch values using vocal pulse marking and manual fixing in Praat[19] were extracted for further analysis.

Before statistical analyses both the 165 items marked as problematic by the transcribers as well as trials more than 2.5 standard deviations above and below the mean were discarded. Because in some target trials the second syllable vowel was too short or the f_0 values fluctuated too much for a reliable assessment, these trials were also removed. Thus, altogether 6.8% and 10.1% of the first and second syllable data were removed, respectively. The larger number for the second syllable data was due to pre-pausal creaky voice, in a subset of the cases where the target word was immediately followed by a relative clause in the sentence frame. The percentage of accepted items by speaker ranged from 73% to 97%.

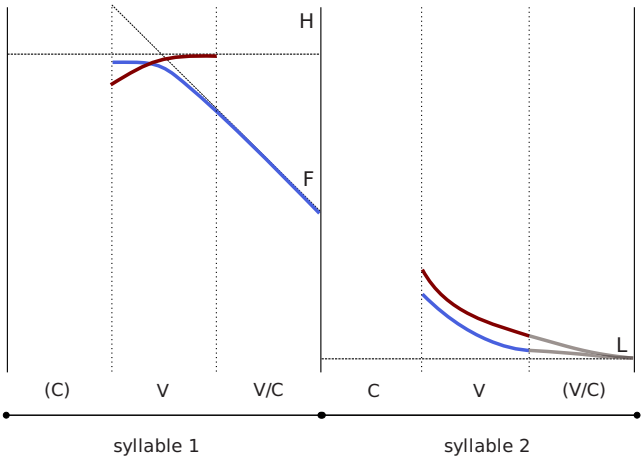

FIG. 1. (Color online) A schematic view of two pitch contours (time vs. frequency) of a disyllabic sequence in Finnish. The dark grey line depicts the CV.CV sequence and the light grey line the CVV.CV or CVC.CV sequences. The Target Approximation model targets (H,F, and L) are also shown (dotted diagonal line and the dotted horizontal lines). The last optional vowel or consonant segment in the second syllable is also shown; the gray lines show the possible continuation of the f_0 contour at that point.

We used the Target Approximation (TA) model[20,21] to analyze our data. The model views speech melody as a series of consecutive syllable synchronic pitch targets modulated by the articulatory system; the tonal target is best seen on the surface f_0 contour toward the end of the syllable, where the approximation is maximal with respect to the underlying target. The model is therefore designed to handle several consecutive syllables and provides a means to investigate the disyllabic structure of Finnish words.

The TA model has four parameters which determine the pitch contour at any point: local pitch target (static or dynamic), duration, pitch range, and the articulatory effort. Figure 1 shows the schematic representation of a disyllabic sequence with two different tone sequences and the corresponding targets. The red line shows a pitch contour during a high static (H) and low static (L) target sequence; the light grey line shows a dynamic falling (F) and low static sequence. In all cases the f_0 is seen to approximate the targets (dashed lines). In both H and F cases the f_0 onset is at the dotted vertical line depicting the end of the syllable onset. In the F target case the syllable is longer in duration than in the H case. In both cases the contour continues to the second syllable, which has a slightly reduced articulatory strength (effort) which can be seen in that the contours do not reach the target.

Establishing possible tonal targets for languages that are by definition non-tonal is naturally a non-trivial task. Any *a priori* determinants have to be based on purely

phonological–rather than phonetic–factors. In Finnish the natural determinants are syllable weight and word structure, i.e., quantity and disyllabicity, respectively. Consequently, any disyllabic sequence of Finnish word root would be analyzed as a sequence of two tonal targets as depicted in Fig. 1. Since the second (unstressed) syllable is almost always lower than the first (stressed) one, it most likely corresponds to either a low or a mid tone. In turn, the tonal differences between the long (heavy), e.g., CVV, and short (light), e.g., CV, stressed syllables (two vs. one moraic, respectively) could be analyzed as stemming from two distinct tonal targets, or tones, in an articulatory phonetic sense.

We used different distances (20 to 50 ms) before the syllable offset in our analyses depending on the duration of the syllable nucleus in order to avoid the effects of segmental perturbation on the results. The temporal range is further justified by findings of Xu and Liu,[22] who show that the actual syllable offset is 20–50 ms earlier than the traditional one, which is used as the basis of segmentation in the current study, as well.

Each f_0-signal from Praat was read in Matlab (version 7.1.0.183, R14) and was then smoothed by trimming and averaging. The trimming algorithm runs through every pitch mark and if the preceding and the following pitch marks are both considerably higher (or lower) the value is replaced by linear interpolation. The criterion for replacement is such that the change rate to and from the deviant point would be more than 120 semitones per second. This value is based on the estimate of maximum excursion speed of f_0 change,[23,24] although we did not take into account that—according to[24]—the actual maximum velocity varies directly with the size of the pitch movement in question. Since the pitch period based f_0 data is not temporally evenly distributed, it had to be sampled with 200 Hz sample frequency where the samples where linear interpolants of the neighboring data points. The signal was further smoothed by standard 7-point triangular smoothing. The resulting signal was continuous with no gaps for unvoiced parts. To approximate the underlying tonal target, the difference between two successive f_0 points at the given time before syllable boundary was calculated. This yielded the slope of the f_0 curve in semitones per second.

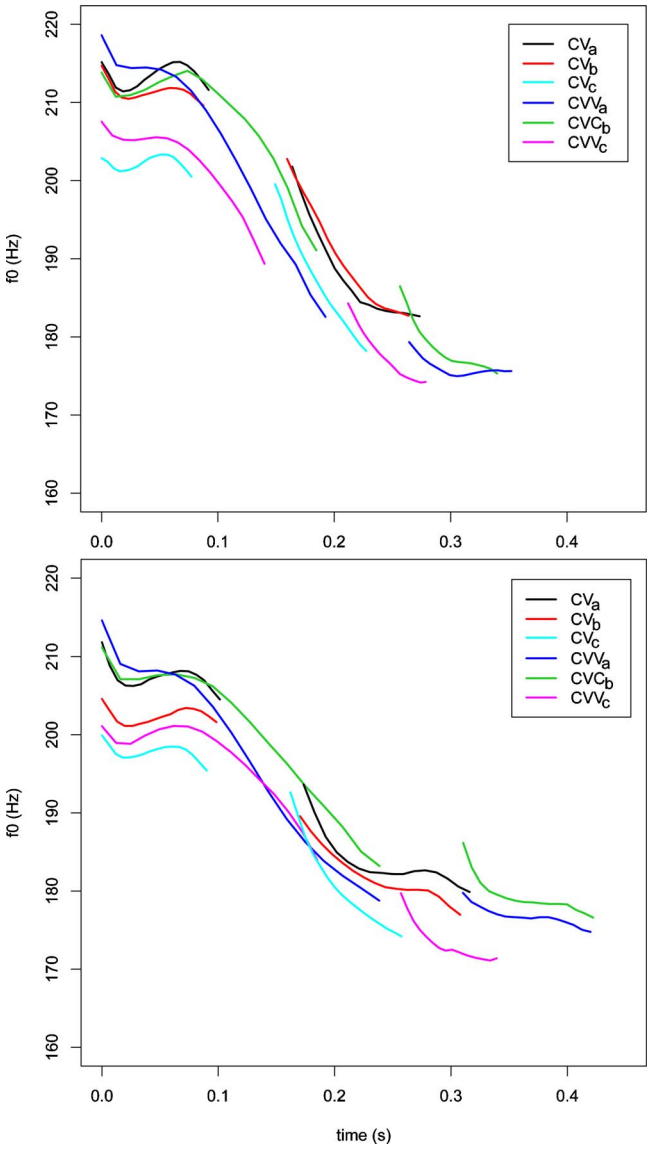

FIG. 2. (Color online) Average f_0 contours for the different syllable structure conditions on the experiment. Upper panel shows the contours for normal speech rate; the lower panel shows slow speech rate. The gaps in the contours are due to the second syllable consonant onsets. Time zero stands for the pitch onset of the first syllable nucleus. The computation of the averages were done in semitones.

C. Results

Figure 2 shows average contours for all data in the study. The different colors depict the different syllable structure conditions. Normal speech rate is shown in the upper panel and the slow speech rate in the lower one. The contours have been aligned to start at the first syllable voicing onset. The gaps in the contours stand for the intervening second syllable consonant onsets. It should be noted that each subscript stands for a comparable sets of word pairs sharing the same set of sentence frames.

The same contours in time normalized form in semitones are shown in Fig. 3. Two contours have been drawn for each syllable structure separately; one for normal speech rate (black lines) and another for slow speech rate (grey lines; red lines online). The upper panel depicts the contours for the unimoraic first syllable (CV), the lower panel depicts the contours for bimoraic syllables (CVV and CVC). Solid lines in the lower panel stand for CVV words and the dashed line the CVC words. Dotted lines in both panels depict the contours for the verb items. Time normalization was done using 16 equidistant points and linear interpolation for each vocalic segment of the syllable. It should be noted that in the case of CVC-syllables the contours depict the VC-segment of the syllable. Only the vocalic segments of the second syllables are depicted in both panels.

The overall patterns that emerge from both Figs. 2 and 3 seem to support the hypotheses stated before; namely, that the disyllabic sequence itself has a word level pattern, and that the stressed syllable has two distinct patterns depending on the weight of the syllable.

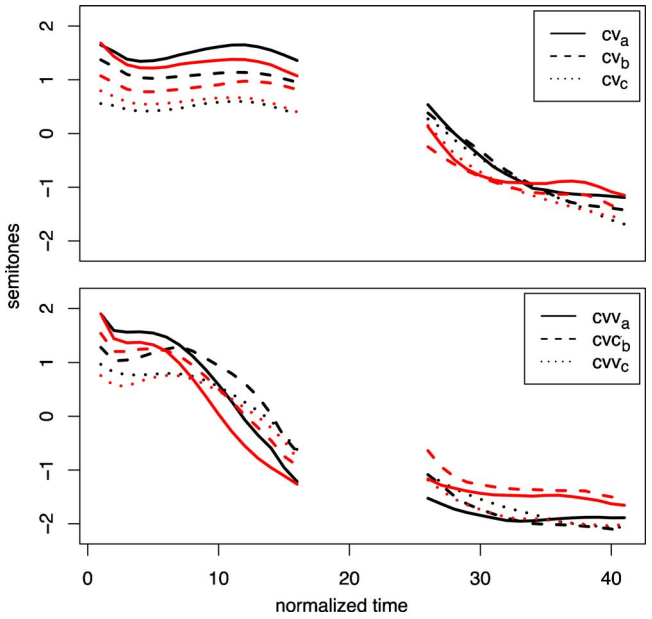

FIG. 3. (Color online) Time normalized average f_0 contours for the different syllable structures. Black lines depict normal speech rate and grey lines slow speech rate. The upper panel depicts the contours for the unimoraic first syllable (CV), the lower panel depicts the contours for bimoraic syllables (CVV and CVC). Solid lines in the lower panel stand for CVV words and the dashed line the CVC words. Dotted lines in both panels depict the contours for the verb items. The index numbers in the legend depict the experimental conditions used for statistical analyses.

1. Tonal differences in the stressed syllable

Table I shows the means and standard deviations of the slopes 50 ms before the offset of the first syllables broken down by tonal conditions and speech rate. Table II shows the corresponding rhyme durations. To assess the statistical significance of the slope values we fitted linear mixed-effects models with participants and items as a crossed-random effect to the first syllable f_0 slope measured at 50 and 25 ms from the end of the syllable.[25,26] With respect to the target change rate in semitones per second as the dependent measure, the predictors were Speech Rate (normal, slow), and the tonal categories (CV_a-CV_c, CVV_a, CVV_c, and CVC_b). Markov Chain Monte Carlo (MCMC) sampling was used to obtain the p-values. Model selection was done using likelihood ratio tests. The models were estimated without intercepts and the significance tests were done comparing the estimates against zero; i.e., the expected target value of a static (H) tone. The results are summarized in Tables III and IV.

TABLE I. Means and standard deviations of slopes (ST/s) for the first syllables 50 ms before syllable offset. Both speech rates (normal and slow) are shown.

	Mean		SD	
Syllable type	Normal	Slow	Normal	Slow
CV_a	3.18	2.08	20.95	17.65
CV_b	0.27	0.87	19.96	17.89
CVC_b	−12.2	−12.23	20.75	19.66
CVV_a	−17.14	−13.44	17.65	16.86
CV_3	1.84	0.84	19.16	17.69
CVV_c	−13.7	−13.48	16.87	15.79

TABLE II. Means and standard deviations of rhyme durations (milliseconds) for the first syllables. Both speech rates (normal and slow) are shown.

	Mean		SD	
Syllable type	Normal	Slow	Normal	Slow
CV_a	92.02	101.47	22.13	24.48
CV_b	87.61	98.45	19.85	22.81
CVC_b	184.59	238.60	30.00	45.74
CVV_a	192.45	238.39	34.42	47.96
CV_c	77.26	90.15	17.90	28.30
CVV_c	140.04	185.09	25.45	36.72

We found a significant main effect of syllable weight (conditions CVV_a, CVV_c, and CVC_b) indicating that the heavy and light syllables are tonally different. The verb conditions being similar to their counterparts indicates that the tonal categories do not change regardless of the degree of accentuation; in other words, the weakly accented verbs (CV_c and CVV_c) did not differ from the moderately accented nouns.

There was no main effect of speech rate. However, there was a significant RATE:CVV_a interaction (p=0.023). That is, the slopes are significantly less steep in slow speech. By inspecting Fig. 1, one can see that the interaction is most probably due to either segmental perturbation or–as pointed out above–to the fact that the actual syllable offset is earlier and it is the syllable boundary which is actually visible in the figure.[27] A model fitted closer to the syllable end (25 ms) showed similar results with an additional RATE:CVC_b interaction (Table IV) indicating further that the previous interaction was most likely due to segmental or syllable boundary effects rather than speech rate per se.

All in all, the results are in line with our hypothesis that the f_0 slopes at the end of the syllable differ according to the syllable structure defined in terms of its quantity and its weight. In summary, bimoraic stressed syllables–whether they contain a long vowel or vowel and a coda consonant–show a fairly negative slope at the syllable end, whereas other syllables have slopes that are in effect horizontal. Moreover, speech rate has no effect on the slopes.

The verb materials (conditions CV_c and CVV_c) were included in order to elicit weakly accented disyllabic forms. Visual inspection of the averaged contours in Fig. 2 shows that the verbs (dotted lines in both panels) indeed had a narrower pitch range than the nouns. With regard to slopes, however, they did not differ significantly from the other conditions, indicating further that the underlying tonal form of the words was not based on accentuation but on word and syllable structure.

2. Tonal differences in the second syllable

With regard to the way low tones behave in e.g., Mandarin,[21,28] we did not expect the f_0 of the second syllable to be horizontal at the end of the syllable. We did, however, expect them to behave fairly uniformly and show no effect by the structure of the previous syllable with respect to the

TABLE III. Mixed effects model results for the first syllable slopes 50 ms from syllable end.

Condition	Estimate	MCMC mean	HPD 95 lower	HPD 95 upper	pMCMC	Pr(>|t|)
CV_a	3.1837	3.1910	−0.1917	6.631	0.0716	0.0677
CV_b	0.2739	0.2636	−3.3435	3.726	0.8888	0.8810
CVC_b	−12.3918	−12.3834	−15.9165	−8.876	0.0001	0.0000
CVV_a	−17.0836	−17.0632	−20.4780	−13.686	0.0001	0.0000
CV_c	2.5462	2.5608	−1.1802	6.224	0.1812	0.1796
CVV_c	−14.4998	−14.4681	−18.1824	−11.000	0.0001	0.0000
RATE	−1.1309	−1.1263	−3.9194	1.867	0.4308	0.4357
CV_b:RATE	1.6398	1.6402	−2.4908	5.911	0.4560	0.4419
CVC_b:RATE	1.1425	1.1205	−2.8915	5.514	0.6066	0.5905
CVV_a:RATE	4.6093	4.6116	0.3783	8.406	0.0234	0.0242
CV_c:RATE	0.1271	0.1346	−4.1580	4.415	0.9412	0.9534
CVV_c:RATE	1.3248	1.3066	−2.6616	5.503	0.5326	0.5263

underlying targets. Since we did not expect the tonal shape of the second syllable to vary as a function of its syllable structure, most of the verbs in the experiment (19 out of 21) had a third syllable (a morpheme signifying person) as opposed to the noun materials (35 out of 45 disyllabic). Also, as some of the second syllables were followed by phrase break, we expected that they would show more variability than the initial syllables.

Since the durations of the second syllable are fairly short after a heavy first syllable, we did the analyses at 20 ms before the syllable offset. The second syllable slopes are summarized in Table V and the corresponding vowel durations in Table VI. As Table V shows the slopes at the end of the second syllable are negative rather than horizontal throughout the data, and the added time provided by the slower speech rate does not result in the removal of these slopes. However, we were mainly interested in whether speech rate or the structure of the preceding syllable had an effect on the slopes, the hypothesis being that no effect should be seen.

Results from the mixed effects model fitted on the slopes 20 ms before the syllable offset are summarized in Table VII. CV_a condition was used as the intercept and significance was tested on the differences from this condition. As can be seen in the table, the only significant main effect was in the CV_c (verb) condition, where the slopes are significantly steeper. Similarly, the two significant interactions with speech rate were with the verb materials. The fact that the slopes were steeper at the end of verbs beginning with a CV syllable goes against expectations based on reduced pitch ranges and the interaction effect of the speech rate seems to further neutralize the differences. That is, the differences in the CV_c category were probably due to segmental perturbation and do not invalidate our original hypothesis that the slopes are not influenced by the structure of the preceding syllable, speech rate, or degree of accentuation. Another possibility–as pointed out by one of the reviewers–is that there is a lowering of all the targets in verbs relative to nouns and adjectives and the more negative slope is actually due to speakers' continued effort to lower f_0 even near the end of the syllable.

3. Tonal differences within the disyllabic sequence

One of the main assumptions of this study concerned the prosodic shape of the word as a disyllabic tonal sequence where the first syllable is relatively higher than the second: that is, a static high tone as opposed to a static low one; or a (high) falling tone followed by a static low one. In either case we expected there to be a significant difference in the average f_0 values between the syllables. In order to assess the differences we calculated the mean pitch values for the syllables by averaging the semitone transformed f_0 values within the center half of the voiced sections (calculated from

TABLE IV. Mixed effects model results for the first syllable slopes 25 ms from syllable end.

Condition	Estimate	MCMC mean	HPD 95 lower	HPD 95 upper	pMCMC	Pr(>|t|)
CV_a	−1.5750	−1.5554	−4.917	2.0431	0.3804	0.3871
CV_b	−2.0015	−1.9957	−5.613	1.6494	0.2816	0.2875
CVC_b	−21.9486	−21.9510	−25.839	−18.3760	0.0001	0.0000
CVV_a	−18.8184	−18.8157	−22.266	−15.2793	0.0001	0.0000
CV_c	−0.0314	−0.0362	−3.893	3.6715	0.9850	0.9870
CVV_c	−18.1505	−18.1398	−21.875	−14.5206	0.0001	0.0000
RATE	−1.7931	−1.8161	−4.571	0.9439	0.1952	0.1970
CV_b:RATE	2.1337	2.1632	−1.890	5.9403	0.2762	0.2939
CVC_b:RATE	8.9513	8.9792	5.017	13.1658	0.0001	0.0000
CVV_a:RATE	8.9817	9.0094	5.084	12.8382	0.0002	0.0000
CV_c:RATE	0.5116	0.5562	−3.365	4.7396	0.7870	0.8031
CVV_c:RATE	0.9287	0.9550	−3.019	4.8439	0.6310	0.6458

TABLE V. Means and standard deviations of slopes (ST/s) for the second syllables 20 ms before syllable offset. Both speech rates (normal and slow) are shown.

	Mean		SD	
Syllable type	Normal	Slow	Normal	Slow
CV_a	−8.12	−10.95	27.78	26.21
CV_b	−11.44	−12.09	28.91	26.71
CVC_b	−10.32	−12.39	29.57	28.88
CVV_a	−5.91	−10.69	25.24	29.63
CV_c	−21.11	−11.25	21.67	18.19
CVV_c	−13.65	−9.64	19.24	24.78

the time normalized data). A paired t-test, calculated for both participant means over items and item means over participants indeed, revealed that the means do differ from each other significantly [$t(24)=8.339$, $p<0.0001$ and $t(65)=20.422$, $p<0.0001$]. The difference between the means was 2.2 semitones (yielding −10.8 ST/s) which is much greater than would be expected if the differences were due to general downtrends alone. For instance, Swerts, Strangert, and Heldner have reported a value of −0.88 ST/s for Swedish.[29]

III. DISCUSSION

The present research studied the tonal variation of Finnish words. First, we investigated the tonal differences within the stressed syllable of the word: namely, the contrasting, paradigmatic differences between light and heavy syllables, with an emphasis on the quantity differences between the nuclear vowels. Second, we investigated syntagmatic tonal differences between the stressed and unstressed syllables in a disyllabic sequence determining the basic root structure of a Finnish word. The results of this study clearly showed that phonological quantity, and syllable weight in general, is signaled tonally: syllables with more weight (bimoraic syllables) are associated with a dynamic (falling) tonal target as opposed to a static (high) target for a light syllable with only one mora (or a short nuclear vowel). With regard to the mechanism and articulation, there probably are no differences between the way a Finnish speaker produces these tones as opposed to e.g., a Mandarin speaker producing lexical tones.

Because the second syllable–always being unstressed in Finnish–would usually be discarded as unimportant by models/theories that are not explicitly syllable based (one exception being Suomi's mora based model which does take it into account when it contains the word's second mora), we opted for an approach that views the f_0 contour as a series of syllable-synchronic pitch targets, namely Target Approximation. As shown in this study, the unstressed second syllable in Finnish is produced very consistently regardless of either the tonal context, amount of accentuation, or rate of speech. Its stability can be seen to provide an anchor for the tonal structure of the whole word; especially when contrasted with the more variable stressed syllable preceding it.

TABLE VI. Means and standard deviations of second syllable vowel durations (ms). Both speech rates (normal and slow) are shown.

	Mean		SD	
Syllable type	Normal	Slow	Normal	Slow
CV_a	109.50	142.69	40.85	51.40
CV_b	104.37	137.62	39.85	56.63
CVC_b	83.51	111.64	36.34	57.14
CVV_a	87.83	109.65	39.26	49.23
CV_c	78.36	95.34	28.73	39.91
CVV_c	66.58	82.33	30.92	44.83

The results further suggest that the tonal structure of Finnish words stems from different articulations determined by the segmental makeup of the words' first syllable as well as the syllabic structure of the word stem. As opposed to the stable second syllable, which can be characterized as having a low static target, the first syllable falls in two distinct tonal categories determined by its segmental structure and/or the quantity of the nuclear vowel. This indicates that there probably are no conspiracies regarding one prosodic parameter over another (cf. Refs. 13–15), for both segmental durations and f_0 are used together in achieving the same communicative goals, which in this case have to do with word structure and syllable weight (and consequently phonological quantity) rather than accentuation. Accentuation and the consequent prominence relations in Finnish, it seems, are thereby not related to pitch peaks themselves (which are, in fact, often absent), but rather to the dynamic aspects of the f_0 contour as a whole.[17,30]

We have not taken into account any regional varieties of Finnish in this study; we simply regard it as representing standard spoken Finnish. Both Suomi[13–15] and Ylitalo[31] have attested to regional differences in both segmental durations and tonal structure. Their methods differ considerably from the ones presented in this study, and the results are not directly comparable. However, the tonal differences reported by both Suomi and Ylitalo do not seem to be in disagreement with the current results; especially if one were to compare velocity profiles instead of raw f_0 contours.

As to the widely researched issue of tonal (or segmental) alignment, our results can be quite easily interpretated within the Target Approximation model. Seen in this light, the production of the pitch contour in Finnish would be syllable based and the disyllabic word root/stem would act as an "anchor" for the tonal targets of the two consecutive syllables. Seen in this way, the results would be far less natural to subsume within frameworks that attempt to correlate pitch peaks and their temporal positions with the segmental material. Even though these types of approaches have been successfully applied to (especially) Indo-European languages, cross-linguistic differences, albeit a possibility, may not explain this difference completely. For example, Schepman *et al.* found that vowel length (i.e., quantity) was the only factor explaining the alignment of f_0 landmarks in Dutch stressed vowels.[32] Even though the authors assume–with reservations that this effect was independent of syllable structure, seen against the background of the present results, this may have been a hasty conclusion. Rather, their results suggest that underlyingly Dutch and Finnish systems might not be that different after all.

TABLE VII. Mixed effects model results for the first syllable slopes 20 ms before syllable offset.

Condition	Estimate	MCMC mean	HPD 95 lower	HPD 95 upper	pMCMC	Pr(>|t|)
CV_a (intercept)	−8.0487	−7.998	−12.556	−3.4176	0.0012	0.0007
CV_b	−3.2722	−3.289	−9.133	2.1676	0.2556	0.2745
CVC_b	−2.8540	−2.864	−8.847	2.7170	0.3374	0.3495
CVV_a	2.0636	2.042	−2.033	6.1599	0.3208	0.3265
CV_c	−16.1122	−16.083	−21.814	−10.2089	0.0001	0.0000
CVV_c	−3.9406	−3.931	−10.169	1.7860	0.2038	0.2074
RATE	−2.8757	−2.903	−7.124	0.8834	0.1544	0.1639
CV_b:RATE	1.8624	1.914	−4.243	7.5302	0.5166	0.5365
CVC_b:RATE	0.9809	1.022	−4.961	7.0245	0.7376	0.7512
CVV_a:RATE	−2.0250	−1.999	−7.991	3.5405	0.4932	0.4922
CV_c:RATE	12.7020	12.711	6.787	18.5753	0.0001	0.0000
CVV_c:RATE	7.1287	7.139	1.265	13.2231	0.0176	0.0225

The tonal structure of the word can be seen to serve at least two distinct communicative purposes. First, the static/dynamic distinction probably works against tendencies which lengthen the duration of the stressed syllable (making the durational distributions between long and short segments overlap) and thus providing the listeners with a disambiguating cue when the durational one is insufficient or totally absent. This idea has, in fact, been shown to work in off-line categorization[6] indicating that tonal structure probably needs to be specified at least on the sub-lexical–if not lexical-level. What exactly is the locus of the effect that the tonal structure exerts in Finnish word production and recognition, compared to e.g., the specification of tones in Mandarin Chinese, is an important question that needs to be pursued further.

Second, as a consequence of having a structurally specified tonal default pattern, the signaling of post-lexical, i.e., utterance level phenomena, such as for instance continuation, would be straight-forward. For instance, producing a sequence of two high level tones rather than the default high-low sequence for a CV.CV word, would result in a typical non-falling pitch contour at the end of a phrase to signal continuation. That is, an intonation contour can be produced largely just by executing a sequence of words with pre-specified tonal targets for each syllable of the utterance; any post-lexical alterations can be achieved by manipulating the default target patterns and choosing non-default patterns at the stage when the utterance is still being planned. Whether this is the case, calls for further investigation where syllable structure, in addition to word structure, and factors responsible for post-lexical intonation are manipulated in a controlled manner. A syllable based model could also be used to describe the terminal rises and other intonation phenomena that have been attested in Finnish.[33] In general, syllable based models of prosody would be easily consolidated with speech production models, such as the one proposed by Levelt,[34,35] which use the syllable as a central production unit.

Based on the results shown here the question arises as to how widespread the use of what we might call the phonetic tone is among the languages of the world. Is it only constrained to tone and quantity languages or is it, perhaps, even more universal a feature? In fact–as pointed out by one of the reviewers–there is already some evidence that English also uses syllable-sized pitch targets in signaling word stress contrast,[36,37] except that there are probably more post-lexical shifts of the targets in English than in Finnish.[38]

In conclusion, our results suggest that the phonological quantity distinction in Finnish is co-signaled by a systematic difference in tonal structure. We would like to argue that taking the results reported here into account would have consequences for a more thorough and cognitively motivated theory of the (mental) lexicon. Furthermore, it seems to us that a taxonomy based purely on contrasting tone and quantity may not be sufficient for the linguistic typology of prosody. Instead, a more refined language typology, taking into account the underlying similarities of tone and quantity languages, would perhaps shed further light on attempts to relate speaker populations with external factors such as, for instance, the geographical distribution of certain genes.[39]

ACKNOWLEDGMENTS

We would like to thank Leena Wahlberg for her help in recording and preparing the data. The present study was supported by grants no. 107606, 125940, and 128204 from the Academy of Finland to the first author. A subset of the current data with preliminary analyses was reported in Ref. 40.

[1]M. Haspelmath, M. Dryer, D. Gil, and B. Comrie, *The World Atlas of Language Structures* (Oxford University Press, Oxford, UK, 2005), pp. 58–61.
[2]A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," Lang Speech **40**, 141–202 (1997).
[3]E. Dupoux and S. Peperkamp, "Fossil markers of language development: Phonological 'deafnesses' in adult speech processing," in *Phonetics, Phonology, and Cognition*, edited by J. Durand and B. Laks, (Oxford University Press, Oxford, UK, 2002), pp. 168–190.
[4]I. Lehiste, "Influence of fundamental frequency pattern on the perception of duration," J. Phonetics **4**, 113–117 (1976).
[5]M. O'Dell, "Intrinsic timing and quantity in Finnish," Ph.D. thesis, University of Tampere, Tampere, Finland (2003).
[6]J. Järvikivi, D. Aalto, R. Aulanko, and M. Vainio, "Perception of vowel length: Tonality cues categorization even in a quantity language," in Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI) (2007), edited by J. Trouvain and W. J. Barry, pp. 693–696, http://www.icphs2007.de/conference/Papers/1509/1509.pdf (Last viewed 03/17/2010).
[7]S. Brown, "Time, change, and motion: The effects of stimulus movement on temporal perception," Percept. Psychophys. **57**, 105–116 (1995).
[8]L. Leboe and T. Mondor, "The role of a change heuristic in judgments of

sound duration," Psychon. Bull. Rev. **15**, 1122 (2008).
[9]W. van Dommelen, "Interactions of fundamental frequency contour and perceived duration in Norwegian," Phonetica **52**, 180–187 (1995).
[10]H. Lehnert-LeHouillier, "The influence of dynamic f0 on the perception of vowel duration: Cross-linguistic evidence," in Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI) (2007), edited by J. Trouvain and W. J. Barry, pp. 757–760.
[11]F. Karlsson, "Phonotactic complexity of Finnish nouns," in *Inquiries into Words, Constraints and Contexts*, edited by A. Arppe, L. Carlson, K. Linden, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, and A. Yli-Jyra (CSLI, Stanford, 2005), pp. 65–70.
[12]J. Lehtonen, *Aspects of Quantity in Standard Finnish*, in Studia Philologica Jyväskyläensia (University of Jyväskylä, Jyväskylä, 1970), pp. 123–134.
[13]K. Suomi, "Temporal conspiracies for a tonal end: Segmental durations and accentual f0 movement in a quantity language," J. Phonetics **33**, 291–309 (2005).
[14]K. Suomi, "On the tonal and temporal domains of accent in Finnish," J. Phonetics **35**, 40–55 (2007).
[15]K. Suomi, "Durational elasticity for accentual purposes in Northern Finnish," J. Phonetics **37**, 397–416 (2009).
[16]P. Lippus, K. Pajusalu, and J. Allik, "The tonal component in perception of the Estonian quantity," in Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI) (2007), edited by J. Trouvain and W. J. Barry.
[17]M. Vainio and J. Järvikivi, "Focus in production: Tonal shape, intensity and word order," J. Acoust. Soc. Am. **121**, EL55–EL61 (2007).
[18]M. Laine and P. Virtanen, Turun Sanomat computerised lexical data base (database and program) (1996).
[19]P. Boersma, "Praat, a system for doing phonetics by computer," Glot International **10**, 341–345 (2001).
[20]Y. Xu, "Speech melody as articulatorily implemented communicative functions," Speech Commun. **46**, 220–251 (2005).
[21]S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," J. Acoust. Soc. Am. **125**, 405–424 (2009).
[22]Y. Xu and F. Liu, "Determining the temporal interval of segments with the help of F0 contours," J. Phonetics **35**, 398–420 (2007).
[23]J. Sundberg, "Maximum speed of pitch changes in singers and untrained subjects," J. Phonetics **7**, 71–79 (1979).
[24]Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," J. Acoust. Soc. Am. **111**, 1399 (2002).
[25]R Development Core Team, *R: A language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2005), http://www.R-project.org (Last viewed: 03/17/2010).
[26]R. H. Baayen, *Analyzing Linguistic Data. A Practical Introduction to Statistics* (Cambridge University Press, Cambridge, England, 2007), pp. 241–259.
[27]We thank one of the reviewers for pointing out this and other facts regarding the TA model.
[28]Y. Xu, "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83 (1997).
[29]M. Swerts, E. Strangert, and M. Heldner, "F0 declination in spontaneous and read-aloud speech," in Proceedings of ICSLP, Philadelphia (1996), pp. 1501–1504.
[30]M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," J. Phonetics **34**, 319–342 (2006).
[31]R. Ylitalo, "The realisation of prominence in three varieties of standard spoken Finnish," Acta Univ. Oulu, B **88**, 98–121 (2009).
[32]A. Schepman, R. Lickley, and D. Ladd, "Effects of vowel length and 'right context' on the alignment of Dutch nuclear accents," J. Phonetics **34**, 1–28 (2006).
[33]R. Ogden and S. Routarinne, "The communicative functions of final rises in Finnish intonation," Phonetica **62**, 160–175 (2005).
[34]W. Levelt, A. Roelofs, and A. Meyer, "A theory of lexical access in speech production," Behav. Brain Sci. **22**, 1–38 (1999).
[35]J. Cholin, W. Levelt, and N. Schiller, "Effects of syllable frequency in speech production," Cognition **99**, 205–235 (2006).
[36]D. Fry, "Experiments in the perception of stress," Lang Speech **1**, 126–152 (1958).
[37]Y. Xu and C. Xu, "Phonetic realization of focus in English declarative intonation," J. Phonetics **33**, 159–197 (2005).
[38]F. Liu and Y. Xu, "Question intonation as affected by word stress and focus in English," in Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany (2007), pp. 1189–1192.
[39]D. Dediu and D. Ladd, "Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin," in Proceedings of the National Academy of Sciences (2007), Vol. **104**.
[40]M. Vainio, D. Aalto, J. Järvikivi, and A. Suni, "Quantity and tone in Finnish lexically stressed syllables," in Proceedings of the Second International Symposium on Tonal Aspects of Langugages-TAL2006 (2006), edited by A. Belotel-Grenié and M. Grenié, pp. 121–124.