Hierarchical Levels of Rhythm in Conversational Speech

Michael L. O'Dell, Mietta Lennes and Tommi Nieminen

Univ. of Tampere, Univ. of Helsinki, Univ. of Jyväskylä

michael.odell@uta.fi, mietta.lennes@helsinki.fi, tommi.nieminen@campus.jyu.fi

Abstract

This study continues our previous investigation into the rhythms contributing to the temporal structure of speech. The relative significance of different hierarchical levels of rhythm was evaluated using Bayesian inference on a linear regression model based on coupled oscillators. Results strengthen our previous conclusions that stress, mora and possibly foot timing are all simultaneously present as rhythmic factors in Finnish conversational speech.

1. Introduction

Languages have been classified as *stress timed* or *syllable timed*; more recently other types of timing have been proposed, such as *mora timing* [14] and *foot timing* [18]. Rather than assuming a strict classification, we view speech rhythm as resulting from synchronization of these or other hierarchical levels of cyclical behavior. We consider the possible relevance of these rhythmic components in conversational speech, using results from coupled oscillator theory to evaluate their influence on timing.

It has been suggested for example that Finnish may exhibit signs of foot timing [18] or mora timing [1]. In our previous research [11], we found evidence that at least for one speaker, Finnish conversational speech appears to have a strong component of rhythm at about the level of phrasal stress in addition to mora timing, along with a weaker third component which could be labeled a foot rhythm.

In the present study, we examine an additional speaker from the same data base, covering almost twice as much total speech time as before. We also extend our statistical model in two ways: an additional category of syllable boundary is included to allow for different probabilities of phrasal stress realization, and a parameter is also added to take into consideration alternative definitions of the mora.

2. Theoretical background

In recent years several researchers have utilized the mathematical apparatus of coupled oscillators to model speech rhythm [2, 4, 12, 16]. One result of a general model of hierarchically coupled oscillators is that the period of the slowest rhythm tends toward a value which can be expressed as a linear function of the number of lower level units it includes [13]. In order to evaluate the effect that hypothetical rhythmic levels have on timing we utilize a linear regression model for pause group duration (T_1) with five (possible) levels:

$$T_1 = c_1 + c_2 n_2 + c_3 n_3 + c_4 n_4 + c_5 n_5 \qquad (1)$$

1. Pause group (stretch of speech between physical pauses; coefficient c_1)
2. Number of stress groups in each pause group n_2 (determined stochastically subject to the restriction that each pause group contains at least one and that a stress group boundary does not fall within a word; coefficient c_2)
3. Number of feet in each pause group n_3 (determined stochastically subject to the restriction that every stress group boundary is also a foot boundary, that every lexical stem begins a new foot and that a foot boundary does not fall within a syllable; coefficient c_3)
4. Number of syllables in each pause group n_4 (coefficient c_4)
5. Number of morae in each pause group n_5 (coefficient c_5)

3. Corpus of conversational Finnish speech

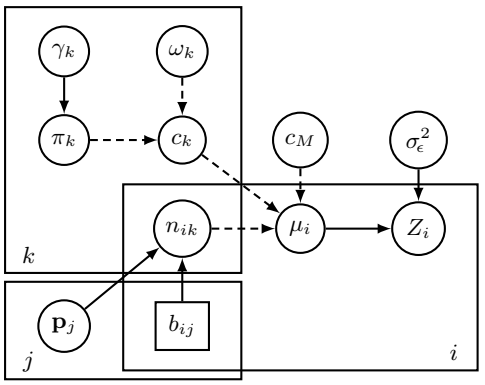
Informal unscripted dialogues were recorded from young Finnish adults in an anechoic room. The participants in each dialogue were close friends and they were allowed to chat freely and unmonitored for a total of 40 to 60 minutes on either given or self-selected topics. The speakers were sitting a few meters apart and facing opposite directions. Each speaker's speech was recorded to a separate channel of a DAT recorder using high-quality headset microphones. The recorded material was then transferred to a computer and sampled at 22050 Hz. The two channels of the stereo files were separated, resulting in one audio file per speaker. Each speaker's utterances were delineated and orthographically transcribed using the Praat program [3]. Parts of the material were phonetically segmented and transcribed, and pause (including short hesitations), word, syllable and mora boundaries were marked. Previously we analyzed 1200 seconds of speech from one female speaker (age 24 years) [11]. For the present study, we analyzed 2360 seconds of conversational speech from a second female Finnish speaker (age 21 years). As before, unclear cases, e.g. hesitation noises, were excluded from analysis.

4. Statistical treatment

We used the WinBUGS program [17] to perform Bayesian inference using the regression model described above. The directed acyclical graph (DAG) in Fig. 1 shows the structure of the total statistical model employed. Arrows with broken lines indicate logical (deterministic) links and those with solid lines indicate stochastic links.

The measured duration of pause group i is represented in this figure by Z_i. As in equation (1), n_{ik} is the number of level k cycles in pause group i and c_k is the coefficient for level k. Together with μ_i (expected duration) and σ_ϵ^2 (error variance) these form the main regression model. Error variance does not appear to increase with duration in our data, so we assume a normal distribution of the Z_i with mean μ_i.

Figure 1: Directed acyclical graph (DAG) of stochastic model.



In our previous work [11] the c_k were given a noninformative half-normal prior. Because of the nature of the oscillator model, estimates for the c_k are expected to be highly (negatively) correlated, causing slow convergence and mixing in the WinBUGS simulation. This problem became even more acute in the present case with an increased number of cases. The problem was alleviated by adding redundant parameters to the model. We take advantage of the relation $\sum c_k \omega_k = 1$ holding in the coupled oscillator model, where ω_k is the natural frequency of the k-level oscillator. Defining auxiliary weight parameters $\pi_k = c_k \omega_k$ which sum to unity, the original coefficients can be calculated as $c_k = \pi_k / \omega_k$. Although the parameters π_k and ω_k would be nonidentifiable in a classical regression analysis, they can be used with informative priors and additional constraints in the Bayesian setting. We give the π_k a joint Dirichlet prior (with all parameters equal to 1) to ensure they sum to unity. The ω_k are given a reasonable lognormal prior with the added restriction that a higher level oscillator is assumed to be intrinsically slower, ie. $\omega_k < \omega_{k'}$ for $k < k'$.

An added advantage of this reparameterization is that π_k can be interpreted as the importance of level k in the coupled oscillator model. To see this, we express equation (1) as

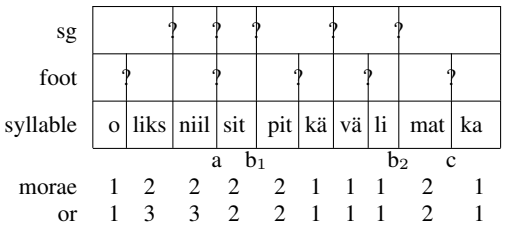
$$T_1 = \pi_1 \frac{1}{\omega_1} + \pi_2 \frac{n_2}{\omega_2} + \pi_3 \frac{n_3}{\omega_3} + \pi_4 \frac{n_4}{\omega_4} + \pi_5 \frac{n_5}{\omega_5} \qquad (2)$$

and note that n_k/ω_k is the duration we would get if the level k oscillator were the only one in the model. Thus equation (2) expresses T_1 as a weighted average of durations. Unlike the raw coefficients c_k, the weights π_k are expressed on the same dimensionless scale independent of the natural period $1/\omega_k$ of the corresponding oscillator. Equation (2) also shows a connection with (nonhierarchical) intrasyllabic gesture coupling currently being investigated in the task dynamic literature [9]. In both cases the resulting composite oscillator has a period which can be expressed as the weighted average of the periods of its components.

The indicator variable $\gamma_k = 0, 1$ in Fig. 1 serves to exclude or include level k in the regression with prior probability 0.5 (so called Gibbs Variable Selection, GVS, cf. [10]).

Since some of the n_{ik} are not known exactly (specifically number of stress groups $n_{i,2}$ and number of feet $n_{i,3}$), a prior distribution must be set up for them as well. This is the purpose of b_{ij} and $\mathbf{p}_j$ in Fig. 1. For each syllable boundary type j, b_{ij} gives the number of such boundaries in pause group i. The vector $\mathbf{p}_j$ expresses the prior probabilities for all possible boundary realizations given boundary type j. These probabilities are given a noninformative hyperprior distribution (Dirichlet with all parameters equal to one).

Figure 2: Four types of syllable boundary, illustrated with the phrase *Oliks niil sit pitkä välimatka?* 'Then did they have a long distance between them?' (sg = stress group).



Given the restrictions outlined above in Section 2, a minimum of three boundary types must be distinguished (see Fig. 2 for an example pause group from our data):

Type a (before a function word) could be a stress group boundary and a foot boundary or only a foot boundary or neither.

Type b (before a lexical word) is at least a foot boundary, but could also be a stress group boundary.

Type c (word internal) cannot be a stress group boundary but might be a foot boundary.

In our previous investigation we restricted the number of prior boundary types to these three. In the present analysis the second group (type b) was further divided into two types:

Type $\mathbf{b}_1$ the boundary is not inside a compound word (ie. corresponds to a space in written Finnish)

Type $\mathbf{b}_2$ the boundary is inside a compound word (no space in written Finnish)

This division was motivated by the widespread consensus among Finnish scholars that the orthographic rules for writing Finnish words together without an intervening space are justified by the facts of Finnish stress. In that case we would not expect boundary type b_2 to begin a stress group, or at least less often than type b_1.

Another difference in the present analysis concerns the method of counting morae. Traditional accounts of the mora in Finnish assume that each additional segment in a syllable after a possible initial consonant adds one mora to the total making it possible to have one, two, three or four mora syllables. An alternative traditional description divides Finnish syllables into light syllables (ending in a short vowel) and heavy syllables (all other types), which corresponds to the restriction of syllables to monomoraic or bimoraic type. To illustrate, in the case of the pause group illustrated in Fig. 2, we get a total mora count of either $n_5 = 17$ or $n_5 = 15$. With this distinction in mind, we wanted to allow the stochastic model to choose between these two ways of counting morae ("bimoraic" *vs.* "multimoraic" hypothesis). We first included a categorical choice in the model with a prior probability of 0.5, but this lead to very poor mixing of the simulation. An alternative approach was therefore taken to allow assessment of the importance (in durational terms) of the "extra segments/morae" in a syllable (ie. those after the first

two morae). These extra segments were counted separately and given a coefficient of their own in the regression model (c_M in Fig. 1), with a noninformative normal prior. This arrangement allowed us to test the two hypotheses statistically: in the bimoraic case we expect $c_M \approx 0$ while in the multimoraic case we expect $c_M \approx c_5$.

5. Results and discussion

Both mora and stress group counts had a very significant effect on pause group duration. Table 1 shows the significance of each level considered. It is quite unlikely that there is any effect at the top (pause group) level, or at the syllable level. The foot level on the other hand showed a likely effect, although it did not reach the significance of the stress group and mora levels. All these results replicate our earlier results for a different speaker.

Table 1: Significance of terms in the regression model (posterior marginal probability that the term is not in model).

rhythmic level	term	Pr($\gamma_k = 0$)
pause group	c_1	0.98654
stress group	$c_2 n_2$	< 0.00001
foot	$c_3 n_3$	0.24100
syllable	$c_4 n_4$	0.87360
mora	$c_5 n_5$	< 0.00001

Table 2 shows the posterior probabilities of the most probable combinations of terms included in the regression model. Mora and stress group terms are included in all models. The most likely model of all ($p = 0.6778$) is the one including stress group, foot and mora (model 1 in Table 2). This is the same model which was most likely in our earlier study.

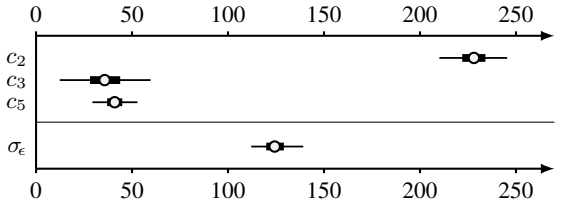
Table 2: Posterior probabilities of the most likely models (only those models for which $p > 0.01$ are shown).

	model	p
1	$c_2 n_2 + c_3 n_3 + c_5 n_5$	0.6778
2	$c_2 n_2 + c_5 n_5$	0.1846
3	$c_2 n_2 + c_3 n_3 + c_4 n_4 + c_5 n_5$	0.07002
4	$c_2 n_2 + c_4 n_4 + c_5 n_5$	0.05406

We now turn to the bimoraic and multimoraic hypotheses. The posterior probability that $c_M > 0$ is $p = 0.7512$. While this is greater than 50%, the 95% credible interval (CI) is $(-14.03, 27.63)$, which includes zero, so the bimoraic hypothesis cannot be rejected at this level of confidence. On the other hand the posterior probability that $c_M < c_5$ is $p = 0.98913$ so we can reject the multimoraic hypothesis at more than a 95% confidence level. Outside of these two hypotheses there remains the possibility of a more complex effect, but in what follows we retain the simplest assumption consistent with the present data, namely that Finnish syllables are either monomoraic (C)V or bimoraic (all other types).

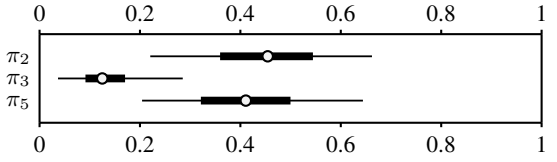
We next consider the other parameters of the model, conditional on the choice of the most likely model (model 1 in Table 2). Fig. 3 shows the posterior distributions for coefficients c_2, c_3 and c_5. In this and the following diagrams the dot indicates the estimate (median of the posterior distribution), the thick line indicates the 50% CI, while the thin line indicates the 95% CI. The value of the coefficient for stress group is quite large (median 228.1 ms). This is surprising since Finnish has not traditionally been described as stress timed, but this is not likely to be just an idiosyncratic feature of a single speaker, since it exactly parallels the result we obtained with our previous speaker (median 251.7 ms) [11].

Figure 3: Credible intervals (in ms) for stress group, foot and mora coefficients and SD of error.



One problem with interpreting the raw coefficients is that in a sense they represent different scales, the scales of their respective oscillator periods ($1/\omega_k$) or speech rhythms. Fig. 4 on the other hand shows the posterior distributions of the π_k from equation (2) and Fig. 1, which as mentioned earlier can be interpreted as indicating the relative weighting or importance of each level in the model. Judging from Fig. 4 it would appear that stress group and mora levels are weighted approximately equally, but the so called foot rhythm is much weaker. These conclusions should be considered only tentative owing to the nonidentifiability of these parameters in the model (dependence on the ω_k).

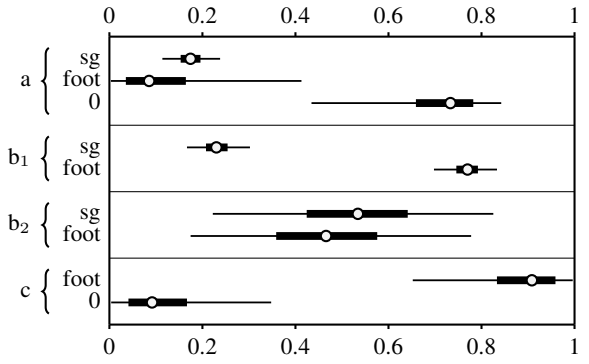
Figure 4: Credible intervals for π_k.



Posterior distributions for the boundary probability vectors $\mathbf{p}_j$ are shown schematically in Fig. 5 (sg means stress group (and foot); 0 means neither stress group nor foot). The distributions for types a, b_1 and c are remarkably similar to the distributions for types a, b, and c for our previous speaker. The posterior distribution for type b_2 is obviously quite different in being quite wide with a median close to 50%. The explanation for this could be that there are too few cases in the data to determine stress probability more accurately for this boundary category. Another possible explanation is that b_2, defined as it is by orthography, does not represent a well defined boundary type for speech. In any case, we do not find evidence that lexical words inside an orthographic compound are less likely to start a stress group; on the contrary the median probability for type b_2 is *greater* than for type b_1.

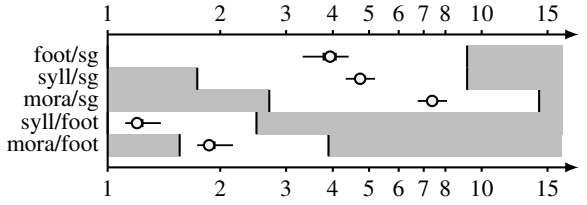
We now turn to the relative sizes of the units uncovered by the analysis. This is summarized in Fig. 6 in terms of the posterior distributions (CI) for average number of feet, syllables

Figure 5: Credible intervals for boundary probabilities, types a, b_1, b_2 and c.



and morae per stress group, and the average number of syllables and morae per foot, plotted on a logarithmic scale. The grey areas indicate impossible values given the present data and the restrictions outlined above in Section 2.

Figure 6: Credible intervals for average stress group and foot length in number of feet, syllables and morae.



Average foot length is only slightly above one syllable (median 1.198), which does not coincide with traditional descriptions of the foot in Finnish (Finnish *tahti*, most usually two syllables in length [15]). It is however almost identical to the value obtained for our previous speaker (median 1.234 syllables). Average foot length counted in morae is almost 2 (median 1.867), again very close to the value for our previous speaker (median 2.017 morae, with multimoraic syllables allowed). Does Finnish perhaps have a foot rhythm which is predominantly bimoraic rather than bisyllabic? Such units have been proposed for other languages, eg. Japanese [5], as well as (in a limited way) for Finnish [6, p. 151]. While the present data are suggestive, they are far from conclusive. Obviously this possibility deserves more attention.

6. Conclusions

In this study, we extended our investigation of the main components contributing to speech rhythm in Finnish. Pause group durations in conversational Finnish were modeled as resulting from an interaction between several hypothetical oscillators with different natural frequencies. Since the present results from a second Finnish speaker accord well with our previous findings, we are more confident than before that Finnish conversational speech has a strong component of phrasal stress rhythm in addition to strong mora timing. A third, weaker component between these two might be labeled a foot rhythm, although it appears to be dominatly bimoraic rather than the traditional bisyllabic foot.

In addition to analyzing data from more speakers, our future plans include investigating oscillator behavior within cycles. In this way we hope to clarify how synchronization is achieved between hierarchical levels. Put another way, we would like to investigate how the deviations in angular velocity required for synchronization are alloted throughout the cycle. To this end two tools will likely prove valuable. It is now possible to model stochastic differential equations using a specially developed module for WinBUGS [7]. Another possibility is task dynamic modeling of Finnish including so called prosodic gestures using TADA [8].

7. References

[1] K. Aoyama. *A Psycholinguistic Perspective on Finnish and Japanese Prosody: Perception, Production and Child Acquisition of Consonantal Quantity Distinctions*. Boston: Kluwer Academic Publishers, 2001.

[2] P. A. Barbosa. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In *Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence*, pages 163–166, 2002.

[3] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 4.6.02) [Computer program], 2007. Last retrieved May 18, 2007, from http://www.praat.org/.

[4] F. Cummins. Speech rhythm and rhythmic taxonomy. In *Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence*, pages 121–126, 2002.

[5] Y. Kondo. Within-word prosodic constraint on coarticulation in Japanese. *Language and Speech*, 49(3):393–416, 2006.

[6] J. Lehtonen. *Aspects of Quantity in Standard Finnish*. Number VI in Studia Philologica Jyväskyläensia. University of Jyväskylä, 1970.

[7] D. Lunn. WinBUGS differential interface—worked examples, 2004. URL http://www.winbugs-development.org.uk/wbdiff.html.

[8] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. TADA: An enhanced, portable Task Dynamics model in MATLAB. *Journal of the Acoustical Society of America*, 115:2430, 2004.

[9] H. Nam and E. Saltzman. A competitive, coupled oscillator model of syllable structure. In M. J. Solé, D. Recasens, and J. Romero, editors, *Proceedings of the 15th International Congress of Phonetic Sciences*, volume 3, pages 2253–2256. Universitat Autònoma de Barcelona, Spain, 2003.

[10] I. Ntzoufras. Gibbs variable selection using BUGS. *Journal of Statistical Software*, 7(7), 2002.

[11] M. O'Dell, M. Lennes, S. Werner, and T. Nieminen. Looking for rhythms in conversational speech. In J. Trouvain and W. J. Barry, editors, *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 1201–1204. Universität des Saarlandes, Saarbrücken, Germany, 2007.

[12] M. O'Dell and T. Nieminen. Coupled oscillator model of speech rhythm. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. Bailey, editors, *Proceedings of the XIVth International Congress of Phonetic Sciences*, volume 2, pages 1075–1078. University of California, Berkeley, 1999.

[13] M. O'Dell and T. Nieminen. Speech rhythms as cyclical activity. In S. Ojala and J. Tuomainen, editors, *21. Fonetiikan päivät Turku 4.–5.1.2001*, number 67 in Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja / Publications of the Department of Finnish and General Linguistics of the University of Turku, pages 159–168, 2001.

[14] R. F. Port, J. Dalby, and M. O'Dell. Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, 81(5):1574–1585, 1987.

[15] M. Sadeniemi. *Metriikkamme perusteet*. Number 236 in SKS:n toimituksia. Helsinki: Suomalaisen Kirjallisuuden Seura, 1949.

[16] E. Saltzman and D. Byrd. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19:499–526, 2000.

[17] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual, Version 2.10*. Cambridge: Medical Research Council Biostatistics Unit, 2005.

[18] K. Wiik. On a third type of speech rhythm: Foot timing. In M. Rossi, A. Rival, A. di Cristo, et al., editors, *Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France August 19–24, 1991*, volume 3, pages 298–301, 1991.