LOOKING FOR RHYTHMS IN CONVERSATIONAL SPEECH

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

Univ. of Tampere, Univ. of Helsinki, Univ. of Joensuu, Univ. of Jyväskylä
michael.odell@uta.fi, mietta.lennes@helsinki.fi,
stefan.werner@joensuu.fi, tommi.nieminen@legisign.org

ABSTRACT

Our exploratory study looks for units of temporal structure in conversational Finnish 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 suggest that stress, mora and possibly foot timing as rhythmic factors in Finnish are more relevant than traditionally assumed.

Keywords: rhythm, unscripted speech, Finnish, coupled oscillators

1. INTRODUCTION

Languages have been traditionally classified as *stress timed* or *syllable timed*, but more recently there have been proposals that other types of timing exist, such as *mora timing* and *foot timing*. It has been suggested that Finnish might be a language with foot timing [12]. On the other hand, while the term mora timing has generally been associated with Japanese [8], it has also been suggested that Finnish might be even more mora timed than Japanese [1].

There are several ways in which such questions could be approached. However, it is likely that different types of rhythmical units affect the temporal organization of speech to varying degrees. In the present study we use results from coupled oscillator theory to evaluate the influence of various levels in a supposed rhythmical hierarchy. Also we turn attention to conversational (unscripted) speech rather than trying to elicit various rhythmic patterns from speakers in an experimental setting.

It is relatively straightforward in Finnish, even in conversational speech, to determine the number of syllables as well as the moraic structure. We were not as sure about our ability to locate phrasal stresses or determine foot structure in conversational speech. (Position of stress within the word is not distinctive in Finnish.) Our initial solution was to use a panel of listeners to judge phrasal stresses or prominent syllables, but it was soon evident that the listeners were far from being unanimous in their judgments. We decided therefore to leave the exact determination of both foot and stress group open and ask a slightly different question: Is there evidence in conversational speech for hierarchical rhythmic units larger than the syllable, based on their possible effects on timing (durations)? Because the number of "stress groups" and "feet" are determined stochastically based on durational effects, these terms should be interpreted with caution in what follows. However, for clarity we continue to use these terms without quotes.

2. THEORETICAL BACKGROUND

In recent years several researchers have utilized the mathematical apparatus of coupled oscillators to model speech rhythm [2, 4, 6, 10]. 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 [7]. For example, assuming a five oscillator model gives the expression

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

for the expected duration T_1 of a top level cycle, where n_k indicates the number of level k oscillator cycles synchronized within it. While the coefficients c_k can be further expressed in terms of the eigenfrequencies of the individual oscillators and their mutual coupling relations, for our purposes the result expressed in equation (1) is sufficient. It provides a justification for using a linear regression model to investigate whether a hypothetical rhythmic unit has an effect on duration.

In what follows we consider a regression model with five (possible) levels:

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. For the present study, we analyzed 1200 seconds of conversational speech from one female Finnish speaker (age 24 years). Unclear cases, e.g. hesitation noises, were excluded from analysis.

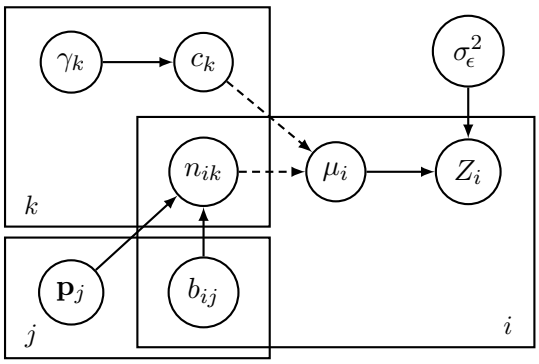
4. STATISTICAL TREATMENT

We used the WinBUGS program [11] 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.

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 Stochastic Search Variable Selection, SSVS, cf. [5]). The c_k themselves are given a noninformative (very large variance) half-normal prior (positive half of a normal

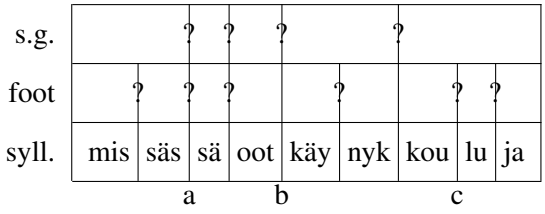
Figure 1: Directed acyclical graph (DAG) of stochastic model.



distribution with mean zero) when $\gamma_k = 1$. (The *a priori* restriction of the coefficients to positive values is a consequence of the coupled oscillator model.)

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 *prior* 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 configurations given boundary type j. These probabilities could have been given fixed values, but we chose instead to give them a noninformative hyperprior distribution (Dirichlet with all parameters equal to one).

Figure 2: Three types of syllable boundary, illustrated with the phrase *Missäs sä oot käynyk kouluja?* 'Where did you go to school?'.



The concept of prior boundary type requires some discussion. The idea is to encapsulate what we want to assume *a priori* about the probabilities of the different level boundaries occurring in various places in the speech stream. Given the restrictions outlined above in Section 2, we need to distinguish a minimum of three types (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 it might be a foot boundary.

Utilizing other information such as syntactic or discourse structure it would be possible to distinguish more prior boundary types but in our present investigation we restrict the number to these three.

5. RESULTS

The main result of the statistical analysis is that at least 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.

Table 1: Significance of terms in the regression model (posterior marginal probability that the term is not in model, ie. $\gamma_k = 0$).

term	p
c_1	0.9545
$c_2 n_2$	< 0.0001
$c_3 n_3$	0.2491
$c_4 n_4$	0.7291
$c_5 n_5$	< 0.0001

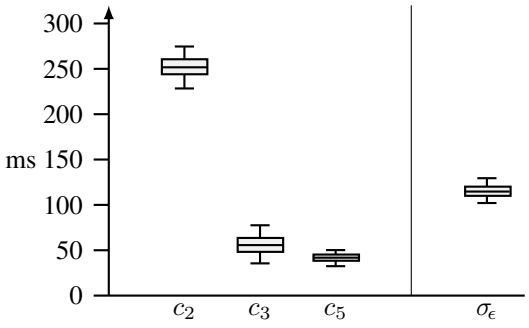
Table 2 shows the posterior probabilities of the most probable combinations of terms included in the regression model. Mora and stress group terms are included in all models. All models also include either foot or syllable terms. The most likely model of all ($p = 0.7007$) is the one including stress group, foot and mora (model 1 in Table 2).

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

	model	p
1	$c_2 n_2 + c_3 n_3 + c_5 n_5$	0.7007
2	$c_2 n_2 + c_4 n_4 + c_5 n_5$	0.2330
3	$c_1 + c_2 n_2 + c_3 n_3 + c_5 n_5$	0.02838
4	$c_2 n_2 + c_3 n_3 + c_4 n_4 + c_5 n_5$	0.02076
5	$c_1 + c_2 n_2 + c_4 n_4 + c_5 n_5$	0.01613

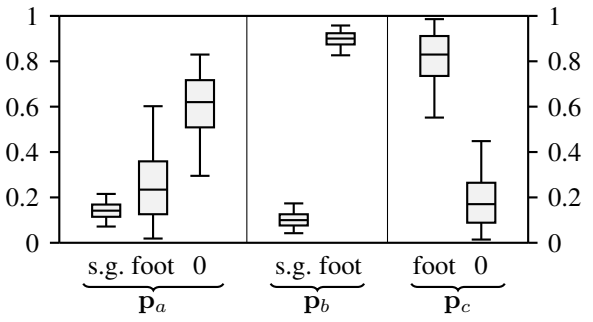
Fig. 3 shows the posterior distributions for coefficients c_2, c_3 and c_5, given the choice of model 1. In this and the following diagrams the gray box indicates the 50% credible interval (CI) with median value at the cross-bar, while the whiskers show the 95% CI. The value of the coefficient for stress group is quite large, on the order of 250 ms, a result which is very striking since Finnish has not traditionally been described as stress timed. Whether this reflects individual or stylistic variation or merely the use of unscripted speech rather than carefully elicited sentences cannot be determined with the present data.

Figure 3: Credible intervals for coefficients and SD of error, given model 1.



Posterior distributions for the boundary probability vectors $\mathbf{p}_j$ are shown schematically in Fig. 4 (s.g. means stress group (and foot); 0 means neither stress group nor foot). We note that the probability of a new stress group at a word boundary is quite well determined and very similar for both types a and b, $p \approx 0.1$. Foot boundary probability is not so well determined by the present data, but it appears to vary considerably depending on prior boundary type.

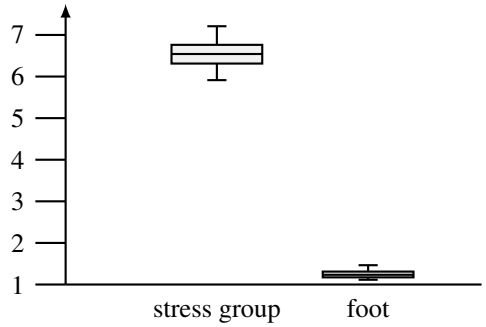
Figure 4: Credible intervals for boundary probabilities, types a, b, c, given model 1.



It remains to examine the size of the units uncovered by our Bayesian analysis. We summarize this in Fig. 5 in terms of average length in number of syllables, again assuming model 1. The average foot length is only slightly above one syllable, which is

very interesting since in traditional descriptions of Finnish the foot (Finnish *tahti*, [9]) is assumed to cover usually two syllables, sometimes three, least often one. This is no doubt due in part to the high incidence of single syllable utterances (such as *joo* 'yeah' or *nii* 'right') in conversation, but it may be a sign that our so called foot is actually a syllable, with occasional fusion of adjacent syllables into one.

Figure 5: Credible intervals for average stress group and foot length in number of syllables, given model 1.



Average stress group length is about 6.5 syllables. To see whether this is a plausible value for a stress group unit, we compare it with the results of our panel of judges. Prominent or accented syllables were marked for the first 345 seconds of the speech material by three trained phoneticians. By listening to one utterance at a time, each word-initial syllable was judged to be either prominent or not. Values for average stress group length in syllables based on these judgments were as follows: 3.75, 5.42 and 7.08. Counting a syllable as stressed only when a majority of listeners judged it prominent resulted in the same total number of stress groups (though divided differently) as for the second judge, giving the same average length, 5.42 syllables per stress group. It is obvious that there was considerable dispersion among the judges, a fact which in part goes to show that phrasal stress in Finnish is not always very prominent. On the other hand, the value from the statistical analysis is well within this range of candidates for phrasal stress rhythm.

6. CONCLUSIONS

It is tempting on the basis of these results to suggest that Finnish may have both stress timed and mora timed rhythms, and possibly a foot timed rhythm as well. Of course there are several reasons to be cautious about such a pronouncement. First of all our data so far encompass only one speaker. Also, the so called stress groups and feet determined here stochastically on the basis of durational effects may not correspond well to traditional stress or feet. It would be safer to say 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. A third component could be labeled a foot rhythm, although it may represent an (augmented) syllable instead.

The statistical method developed here is very general and could easily be extended to the analysis of other cases with varying degrees of uncertainty as to individual rhythmic cycles and the levels of rhythm which are durationally relevant in a language.

7. REFERENCES

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

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

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

[4] Cummins, F. 2002. Speech rhythm and rhythmic taxonomy. *Proc. Speech Prosody 2002, Aix-en-Provence* 121–126.

[5] George, E. I., McCulloch, R. E. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Society* 88(423), 881–889.

[6] O'Dell, M., Nieminen, T. 1999. Coupled oscillator model of speech rhythm. *Proc. XIVth ICPhS, San Francisco* volume 2 1075–1078.

[7] O'Dell, M., Nieminen, T. 2001. Speech rhythms as cyclical activity. Ojala, S., Tuomainen, J., (eds), *Papers from the 21st Meeting of Finnish Phoneticians, Turku 4.–5.1.2001* 159–168.

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

[9] Sadeniemi, M. 1949. *Metriikkamme perusteet*. Helsinki: Suomalaisen Kirjallisuuden Seura.

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

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

[12] Wiik, K. 1991. On a third type of speech rhythm: Foot timing. *Proc. XIIth ICPhS, Aix-en-Provence* volume 3 298–301.