SYLLABLE RATE, SYLLABLE COMPLEXITY AND SPEECH TEMPO PERCEPTION IN FINNISH

Michael O'Dell[1], Tommi Nieminen[2]

[1]Tampere University, [2]no affiliation
michael.odell@tuni.fi, tommi.nieminen@legisign.org

ABSTRACT

There is evidence that perceived speech tempo in English corresponds closely to syllable rate, while segment rate is largely irrelevant when syllable rate is controlled for. However, in a full fledged quantity language such as Finnish, tempo can remain constant even though syllables of different complexity (short or long, one or two morae) vary widely in duration. We therefore hypothesized that tempo would also be perceived differently.

Recorded versions of 60 Finnish phrases, manipulated to control for syllable and mora rate, were played in pairs to 57 native speakers who judged which phrase sounded faster. Our results show that, contrary to English, syllable complexity (mora count) does have a clear effect on tempo perception. On the other hand, the effect of increasing mora rate was weaker than decreasing phrase duration by the same amount, suggesting that perceived tempo is based on several rhythmic factors instead of mora rate alone.

Keywords: speech tempo, Finnish, syllable rate, mora

1. INTRODUCTION

Responding to earlier studies on speech tempo but criticizing their use of syllable and segment rates as a simple proxy for speech tempo, Plug & Smith [10] have shown with manipulated samples that when subjects judge which of two English utterances is faster, syllable rate matters but segment rate doesn't.

We extend this research to Finnish. Like English, Finnish allows a wide variety of syllable complexity, but it is also a full fledged quantity language and hence there are bound to be differences. While segment rate is not expected to be very important, there are reasons to expect that mora rate will be [1, 9], and of course adding morae entails adding segments as well. For instance one would expect a 4 syllable, 8 mora word such as *huo.les.tut.taa* 'disconcerts' to be longer than the 4 syllable, 4 mora word such as *hy.ti.se.vä* 'shivering' when spoken at the same tempo, and therefore it might sound faster if manipulated to be the same duration (contrary to the English results).

To study this, we recorded and manipulated the durations of a number of phrases and then conducted a perception experiment with native speakers. Although it is known that many factors correlate with tempo (e.g. [5, 7, 14]), here we restricted attention to syllable complexity (mora structure). Our results confirmed the influence of mora count on speech tempo but also indicated that tempo perception involves other factors as well.

2. EXPERIMENT

2.1. Stimuli

Phrases were chosen from a corpus of Finnish newspaper texts [13]. They consisted of two words, a 4 syllable adjective followed by a 3 syllable noun, forming a grammatical noun phrase agreeing in case and number. Phrases were chosen so that the total mora count ranged from 7 (e.g. *sinisenä savuna* 'as blue smoke') to 12 morae (e.g. *seuraavista ryhmistä* 'from the following groups'). The final syllable of the adjective and the noun were always short (one mora) for all phrases. Ten phrases were chosen for each mora count, giving a total of $6 \times 10 = 60$ phrases.

Recordings were made of all phrases spoken by one native speaker (the second author) at a comfortable rate. Fig. 1 shows the average syllable rate of the original recorded versions for each mora count. Although no specific procedure was used to insure that all tokens were indeed spoken at the same rate, we may assume this was roughly the case. This figure suggests that, in production at least, syllable rate alone is not the best indicator of tempo—as the number of morae increases in these phrases containing seven syllables the total duration also increases on average so that syllable rate declines.

Each recorded phrase was used to make two stimuli by uniformly manipulating the duration in Praat [4]. One stimulus was produced with 5 syllables per second, the other with 6 syllables per second (com-

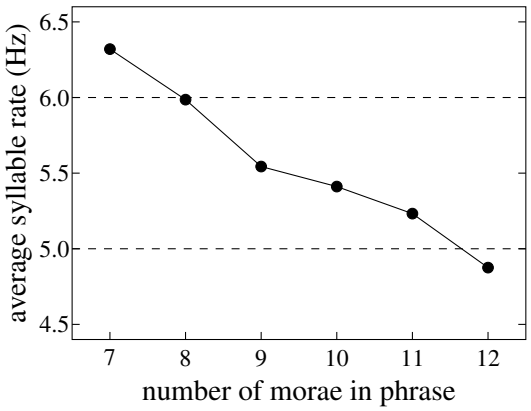

Figure 1: Average syllable rates for original recorded phrases

pare the dashed lines in Fig. 1). This difference (ratio $6/5 = 1.2$) was expected to be well above the perceptual threshold for tempo changes (cf. [11]).

Following the procedure of [12] (see also references therein), C/V boundaries at the beginning (t_b) and end of the phrase (t_e) were used as alignment points for rate manipulation. In the case of voiceless stops the alignment point was set to the onset of voicing after release, which for these tokens was close to the value obtained by the algorithm of [12]. The interval from t_b to t_e thus encompassed 6 syllables and a single Duration point was set in a Praat Manipulation object equal to $(6/5)/(t_e - t_b)$ to achieve 5 syll/s, and equal to $(6/6)/(t_e - t_b)$ to achieve 6 syll/s. $F0$ was set for all stimuli to 135 Hz at t_b with linear interpolation to 100 Hz at t_e.

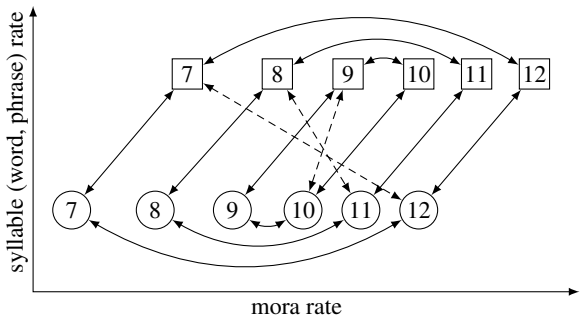

Figure 2: Stimulus pair structure

2.2. Perception test

Pairings used in the experiment (both orders in all cases) are indicated by the arrows in Fig. 2. Numbers in this diagram refer to the number of morae (7–12) in the phrase, circles indicate a rate of 5 syllables/s whereas squares indicate 6 syllables/s. Thus syllable rate increases upward, while mora rate increases to the right. Straight solid arrows indicate control pairs differing in duration, but with identical numbers of morae. These are expected to show a clear difference in responses. Curved arrows indicate pairings with constant duration but varying numbers of morae. Because all stimuli have seven syllables, if syllable rate is the sole determinant of perceived tempo, responses to these pairs should be close to guessing (50%). Dashed arrows indicate pairings which differ in both duration (syllable rate) and in number of morae. Responses to these pairs should shed light on the relative influence of the two rates on perceived tempo.

Each run of the perception experiment used all phrases once (no repetitions) in pairs, giving $60/2 = 30$ judgments of whether the first or second phrase was spoken faster. Order of presentation of the pairs was completely randomized for each run and individual stimuli were randomly allocated to appropriate pairs according to stimulus type (duration and mora count). Each test pair in a run was preceded by a 3 second silence followed by a short piano like sound alerting the listener to the upcoming pair, followed by 1 second of silence. The two phrases in each pair were separated by 600 ms of silence (cf. [11]).

2.3. Subjects

Subjects were 57 native speakers of Finnish residing in the vicinity of Tampere at the time of testing, but coming originally from several areas of Finland. Their ages ranged from 18 to 61 years (median 23).

2.4. Statistical Analysis

Responses were tabulated and analyzed using a Bayesian logistic regression model, in particular, a form of the so called *ideal point* or *item response* model [6, pp. 314–320], [2].

In a general ideal point model, probability of each binary response is modeled on the logit scale as the difference between two opposing sets of coefficients (latent *ideal points*), for instance political inclination of politicians vs. political issues, or individual abilities vs. difficulty of test items. In the present case the responses are judgments that the second phrase was faster, and probability is modeled on the logit scale as the difference in latent tempo (β_i) of the second and the first phrase. Thus this a restricted form of the general model (also known as a Bradley-Terry model) in which the "ideal points" for differencing come from the same continuum. Because some subjects may be more sensitive than others to any given

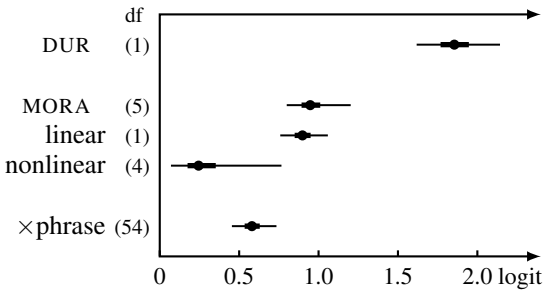

Figure 3: Estimated standard deviations of effects (posterior median, 50% CI & 95% CI)

tempo difference, the model also includes a sensitivity parameter γ_j allowed to vary by subject. In addition, because there may be a bias for perceiving the second phrase faster (or slower) regardless of the phrases presented, the model includes a subject specific bias parameter δ_j (see eq. 1).

$$\text{(1)} \quad \text{logit}(p_{j,a,b}) = \gamma_j(\beta_b - \beta_a) + \delta_j$$

$$\text{(2)} \quad \beta_k = \beta_k^{\text{DUR}} + \beta_k^{\text{MORA}} + \beta_k^{\text{MORA}\times\text{phrase}}$$

Latent tempo itself is modeled as a linear combination of factors including total duration (syllable rate, DUR), mora count (MORA), and the effect of different phrases with the same mora count (MORA×phrase). Note that all parameters in this model are not uniquely determined [6]. In particular, adding any constant to all β_k will leave results unaltered (differences will remain the same), as will dividing all γ_j by any constant and multiplying all β_k by the same constant. Additional restrictions are therefore imposed: $\sum \beta_i = 0$, $\prod \gamma_j = 1$, $\gamma_j > 0$.

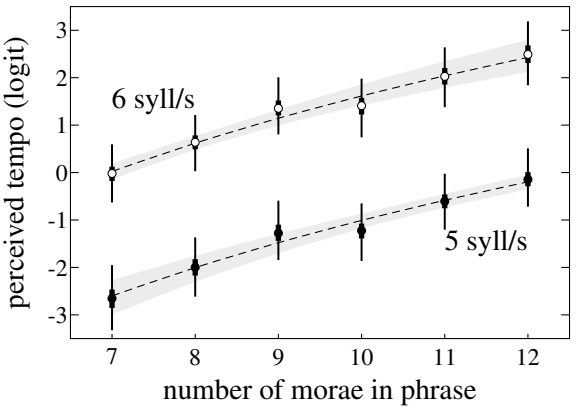

Figure 4: Estimated average latent perceived tempo (95% CI and median for linear effect alone indicated by gray areas and dashed lines)

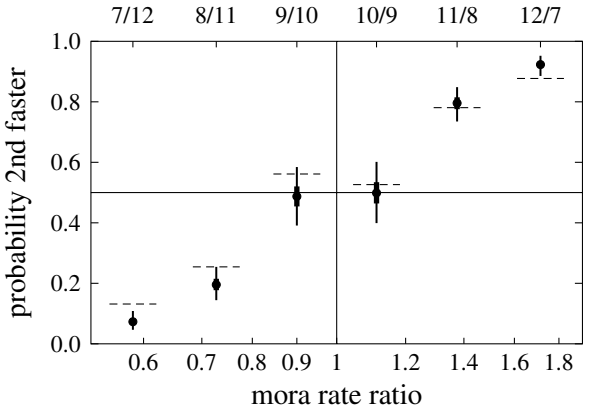

Figure 5: Posterior probabilities of judgments for constant duration pairs based on mora rate ratio (raw pooled proportions shown with dashed lines)

3. RESULTS

3.1. General sensitivity to tempo change

As expected, our subjects were quite sensitive to the durational change of 5 to 6 syllables per second with mora count held constant (cf. straight solid arrows in Fig. 2): the DUR effect is clearly nonzero in Fig. 3. The size of this effect is approximately ± 1.3 on the logit scale, (SD approx. 1.86 in Fig. 3; difference approx. 2.62 between the 2 series in Fig. 4), corresponding to a change from about 6.8 % to 93.2 % "faster" judgments.

There was clear indication of variation in tempo sensitivity among subjects. Several subjects performed much better than the overall average on the control stimulus pairs. Five subjects out of the 57 tested were relatively insensitive to the duration change (posterior median $\log(\gamma_j) < -1$).

3.2. Mora count

As hypothesized, mora count had a very clear effect on perceived tempo: the MORA effect is clearly nonzero in Fig. 3. The effect is also clear in Fig. 5 showing response probabilities for pairs with constant duration (curved arrows in Fig. 2; cf. Plug & Smith's Fig. 4 for English). To get a better idea of the form of this effect, the overall variance due to mora count was divided into a linear trend and a residual nonlinearity (cf. Fig. 3). The linear trend of the mora count effect is also clearly nonzero. The posterior probability that this linear trend is positive (i.e. that mora count *increases* perceived tempo) is $p > 0.9999$. While the magnitude of the nonlinearity remains unclear, it would also appear to be nonzero. This may be due primarily to phrase pairs with 9

vs. 10 morae being difficult for listeners to distinguish. An experimental design including a wider variety of pairs, such as 8 vs. 9 or 10 vs. 11 mora phrases might shed more light on this question.

Another way of judging the MORA effect is to compare opposite ends of the stimulus series. The posterior probability that phrases with 12 morae are perceived as faster than phrases with 7 morae (duration and syllable rate held constant) is $p > 0.9999$. The magnitude of this difference is approximately 2.51 on the logit scale (cf. the difference between 7 mora and 12 mora phrases in Fig. 4), corresponding to a change from about 7.5 % to 92.5 % "faster" judgments (cf. Fig. 5).

3.3. Other factors

Order bias (δ_j in eq. 1) was so small that it was not clear whether it favored the first or the second phrase in comparisons. The posterior probability that the second phrase was favored overall was $p = 0.3827$, and subject based variation in bias was also very small. This contrasts with the results of [10], which showed a slight bias in favor of hearing the second phrase faster.

There was a clear indication that phrases with the same mora count varied in their perceived tempo (standard deviation of the MORA×phrase effect was nonzero, cf. Fig. 3). These differences in perception may be related to several differences in phrases that were not controlled for, such as intrinsic durations, type and location of "extra" morae (long vowel, long consonant, diphthong, consonant cluster). For instance, mora count of the second word may have had a greater influence than mora count of the first word.

4. DISCUSSION

The results indicate that mora count has a clear effect on tempo perception, but also that tempo is not just a function of mora rate *per se*. One type of model for speech rhythm which allows this type of "rhythmic compromise" is the coupled oscillator model (COM, [3, 8]).

In general the COM predicts that natural (equilibrium) duration for some unit should be a linear function of the number of units (or oscillator cycles, with n_i cycles for oscillators $i = 1$ to K) included within it: $T = c_1 n_1 + \cdots + c_K n_K$ [8]. Probability of phrase b sounding faster than phrase a should be a (logistic) function of the ratios (Q) of the phrase durations to their reference (natural, expected) durations.

$$\text{(3)} \quad \text{logit}(p_{j,a,b}) = G_j \log(Q_b/Q_a)$$
$$Q_k = T_k/T'_k, \quad k = a,b$$

If there is no coupled mora oscillator in the COM, natural reference durations would be constant for the phrases used here (since all other counts are constant, 1 phrase, 2 words, 7 syllables, etc.), so that stimulus duration would have an effect but mora count would not (all $\beta_k^{\text{MORA}} = 0$). On the other hand if a mora oscillator is completely dominant, reference duration would be just a multiple of mora count, so the difference between the two duration series (ratio 5/6) should be equal to the difference between phrases of equal duration with a 5/6 mora ratio, e.g. 10 morae vs. 12 morae. More generally, the COM postulates that several rhythms may synchronize, in which case a compromise is predicted (see [8] for overview). If mora rhythm and other rhythms coexist, the COM model predicts that the perceived tempo difference from 10 to 12 morae at constant duration should be positive, but less than the tempo difference for 5 to 6 syll/s with constant mora count. This is indeed the case, as can be seen in Fig. 4.

5. CONCLUSIONS

In general, the effect of increasing mora count from 7 to 12 morae in a 7 syllable phrase of constant duration is about the same as reducing phrase duration by a factor of 1.2 (from 1400 ms to 1167 ms) keeping mora count constant. Put another way, a 7 mora phrase at 6 syllables per second is perceived as roughly equal in tempo to a 12 mora phrase at 5 syllables per second ($\boxed{7}$ and $\textcircled{12}$ in Figure 2).

It is clear that mora count (or segment count, which is confounded with mora count in the present experiment) *does* influence tempo perception in Finnish. This contrasts sharply with the results cited above for English. It is equally clear that mora rate on its own cannot completely account for Finnish tempo perception: there would appear to be other additional rhythms which also come into play. The present experiment was not designed to indicate whether this includes, for example, syllable rate, word rate, phrase rate, or indeed some combination of these. Previous studies have pointed to Finnish having strong rhythmic components at least at the mora and phrase levels [9, 8].

Of course, for purposes of cross-linguistic comparison, using syllable rate as a proxy for speech tempo may still be a good strategy, even if some other rhythm such as mora rhythm dominates in some language(s), because other rhythms will on average have a strong correlation with syllable rate. However, it should be kept in mind, that this procedure may produce results that are less accurate for some languages than for others.

6. REFERENCES

[1] Aoyama, K. 2001. *A Psycholinguistic Perspective on Finnish and Japanese Prosody*. Boston: Kluwer Academic Publishers.

[2] Bafumi, J., Gelman, A., Park, D. K., Kaplan, N. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 13, 171–187.

[3] 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.

[4] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer (version 6.0.39) [computer program].

[5] Cumming, R. 2011. The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics* 39, 375–387.

[6] Gelman, A., Hill, J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[7] Koreman, J. 2006. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America* 119(1), 582–596.

[8] O'Dell, M., Nieminen, T. 2009. Coupled oscillator model for speech timing: Overview and examples. Vainio, M., Aulanko, R., Aaltonen, O., (eds), *Nordic Prosody: Proc. Xth Conference 2008* Helsinki. Peter Lang 179–189.

[9] O'Dell, M. L., Lennes, M., Werner, S., Nieminen, T. 2007. Looking for rhythms in conversational speech. Trouvain, J., Barry, W. J., (eds), *Proc. 16th ICPhS* Saarbrücken. Universität des Saarlandes 1201–1204.

[10] Plug, L., Smith, R. 2018. Segments, syllables and speech tempo perception. *Proc. Speech Prosody 2018* Poznań. 279–283.

[11] Quené, H. 2007. On the just noticeable difference for tempo in speech. *Journal of Phonetics* 35(3), 353–362.

[12] Quené, H., Port, R. F. 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* 62(1), 1–13.

[13] University of Helsinki, 2017. 1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä -korpus [text corpus]. Kielipankki, http://urn.fi/urn:nbn:fi:lb-2017091902.

[14] Weirich, M., Simpson, A. P. 2014. Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics* 43(1), 1–10.