Some lengthening factors in English speech combine additively at most rates

Fred Cummins

Department of Linguistics

2016 Sheridan Road

Northwestern University

Evanston, IL 60208

Suggested running title: Lengthening factors in combination

Received:

Abstract

The known lengthening effects of phrase final position and of contrastive emphasis have been predicted by Klatt to combine super-additively. In a new experiment texts elicited at a wide range of speaking rates were measured and the separate and combined effects of these lengthening factors were found to combine approximately additively at all rates studied. The proportion of lengthening attributable to each factor was found to be relatively invariant except at the fastest speaking rates, where lengthening was eventually eliminated. The results support the interpretation of absolute speaking rate as an inessential variable for characterizing speech at a range of moderate rates.

PACS numbers: 43.70.Fq, 43.70.Bk

I. Introduction

This paper deals with the individual and combined effects of three of the best known factors which influence macroscopic speech timing: speech rate, phrase final lengthening and contrastive emphasis.

Phrase final lengthening (hereafter, PFL) refers to the relatively longer durations observed within a syllable which lies at the right edge of a major prosodic constituent such as an intonational phrase. PFL effects are largely restricted to the rhyme of the final syllable (Klatt, 1976). This distinguishes PFL from utterance final lengthening, which is characterized by global deceleration and reduction in articulatory effort distributed over several syllables. Estimates of lengthening due to PFL vary from 30% (Klatt, 1975) to as much as 120% for long stressed vowels in V# position (Crystal and House, 1988a).

Contrastive emphasis (CE) is a form of accenting used to highlight a particular syllable or word. An accented syllable will bear the nuclear accent of a phrase, often with an exaggerated pitch excursion. Lengthening is frequently associated with CE, and unlike PFL, all parts of the syllable are found to be affected (Beckman, Edwards, and Fletcher, 1992; Turk and Sawusch, 1997). In English, if a syllable bearing CE is followed by an unstressed syllable, the latter may also show some lengthening. Accenting due to CE may be considered an extreme form of phrasal, or nuclear, accenting, especially when the elicitation form suggests a previous misunderstanding (e.g. "I said *BEEF arm*, not *REEF arm*", from Turk and Sawusch, 1997). No reliable estimates of the degree of lengthening due to CE are available.

In an early study (Klatt, 1973), Klatt examined the combined effect of two factors each of which has a shortening effect on vowels in stressed syllables: voiceless coda consonants (relative to voiced) and the addition of an unstressed syllable after the stressed syllable but within in the same word. He found that their combined effect was considerably less than a simple additive model would predict. Thus he proposed that vowel duration be computed based on an inherent (relatively long) duration D_i, which is analyzable as an incompressible part D_{min} and a compressible part which is multiplied by a constant:

$$D_o = k(D_i - D_{min}) + D_{min} \tag{1}$$

The serial application of shortening rules of this sort will produce a sub-additive modification of the overall duration. In his well-known review, Klatt (1976) lists many factors which can influence segmental durations. Each, he suggests, can be associated with a different constant k, with $0 < k < 1$ for shortening rules and $k > 1$ for lengthening rules. This approach to combining factors has found practical application in speech synthesis algorithms. In the present context, Klatt's model predicts that two lengthening factors, such as PFL and CE, will combine super-additively.

The combination of utterance final lengthening and CE was examined in a study by Cooper, Eady and Mueller (1985). They had subjects read isolated sentences with contrastive emphasis on specific key words, as induced by a preceding question (e.g. "Did Chuck like the letter or the present that Shirley sent to her sister?"). They found that the placement of CE on sentence final words resulted in much less durational increase than an additive model would predict. This suggests that there might be an expandibility constraint, analogous to the compressibility constraint. A similar expandibility constraint has been independently proposed by Berkovits (1991).

The present study examines the separate and combined durational effects of PFL and CE at a range of rates. Rate is seen as a possible lengthening or shortening factor which can take on values over a continuous range, and which combines with the two factors of PFL and CE to influence the final duration of segments and syllables. Careful indexing of rate should allow separation of the relative contributions of these three factors to observed acoustic durations. This approach should also reveal any possible interaction between rate of speech and lengthening due to CE or PFL.

II. Method

A. Sentence materials

Four short texts consisting of three sentences were devised. Each was of the form *Didn't he say X? The message was Y. Surely that's not what he said.*, where X and Y differ minimally in one syllable, so that contrastive emphasis is placed on that syllable in Y. All measurements are taken from the second sentence, where Y is phrase final but not utterance final. The X/Y pairs used were chosen so that a target syllable /peɪ̯n/ was or was not phrase final (±PFL), and received or did not receive contrastive emphasis (±CE). Table 1 lists X/Y pairs for each condition.

Insert Table I about here

B. Subjects and recording conditions

Four subjects (3 female, one, JC, male) were paid a flat rate for their participation. Three were undergraduates at Northwestern University, WG was a full-time mother. Subjects JC, GC and CB had lived exclusively in the American upper Midwest (Wisconsin, Michigan, Northern Illinois), WG was a native of Connecticut who had lived 6 years in Southern Indiana. All were monolingual native speakers of American English. None had any known speech or hearing defects.

Subjects were seated at a computer screen and used a mouse to control the succession of trials. On each trial they were presented with one of the four three-sentence texts. They were also given a nominal speech rate which was one of "slow", "comfortable", "medium", "fast" and "very fast", together with a graphic which had an arrow pointing to the appropriate point on a five-point scale from "slow" to "very fast". To further ensure that they would attempt to vary rate across trials, they were asked to repeat the nominal rate aloud before reading the text. Once they were ready to repeat the text, they initiated recording and read the text at their best estimate of the nominal rate. Subjects were given several practice runs, and they were instructed to place contrastive emphasis on the capitalized syllable (see Table 1). No subject had overt difficulty with the required task or with producing the required prosody. Trials were self-paced, and after every block of 20 trials a computer message encouraged subjects to take a break.

The four conditions were crossed with five nominal rates to yield blocks of 20 trials which were randomized within blocks. Each subject completed 12 such blocks in a single session, providing 240 trials in all. Recordings were done in a quiet but not soundproofed room using a Shure SM10A head mounted microphone. Speech was captured directly onto disk via a ProPort D/A unit which digitized at 11025 Hz, with 16 bit linear resolution.

C. Acoustical measurement

Segmentation was done by hand using Entropics Xwaves software. Of the points measured from the waveform and with simultaneous spectrographic control, the following are relevant here: the onset of a nasal formant pattern for /m/ in "message was", the offset of frication in "was", which coincided with the abrupt drop in energy at the /p/ or /k/ closure of the following word, the onset of the syllable *pane/pain*, and the offset of the syllable.

A randomly generated subset of 10% of the utterances was selected and remeasured. Means and standard deviations for the differences between the two measurements showed that all points were reliably measured (mean discrepancy < 5 ms) with the exception of the offset of the final /n/ in "cancer PAIN" and "COUNTerpane", which was occasionally uncertain from the acoustic record due to phrase-final weakening (mean discrepancy: 11 ms, s.d. 30 ms). All measurements were done by the author.

In addition to the remeasurement described above, interval distributions were examined and all obvious outliers were remeasured. A few measurement errors were easily detected in this manner.

Insert Figure 1 about here

FIG. 1.

D. Indexing speech rate

Several recent kinematic studies of rate have opted to treat speech rate as a continuous variable, rather than the categorical division into two or three self selected rates which has been more usually employed (Byrd and Tan, 1996; Shaiman, Adams, and Kimelman, 1995). In the present study, the duration of a common stretch of speech provided a starting point for computing a rate measure. All texts contain the words *message was* in the second sentence, i.e. within the same intonational phrase as the target syllable. These three syllables constitute a single prosodic foot. The reciprocal of the duration of this foot was used as a continuous measure of rate, which thus has the units feet/sec. This measure served to make the variance across nominal rate conditions approximately equal and there was an approximately linear increase in median rate across nominal values. The measure does not take into account any phonological reorganization which may underlie production at fast rates.

III. Results

A. Rate variation

From Figure 1 it can be seen that subject WG produced a much greater variety of rates than the remaining subjects. Given that there are three syllables in the reference foot *message was*, her

Insert Figure 2 about here

FIG. 2.

fastest rates correspond to about 12 syll/sec which is very fast indeed. Subjects CB and JC each produce a wide variety of rates with clear separation across nominal rate classes, while subject GC produces very little rate variation from the slowest to the fastest.

B. Syllable durations

Figure 2 shows syllable durations as a function of speech rate for subject WG only. Note that rate is a continuous variable (feet/sec), and not merely nominal. The relationship between syllable duration and rate appears to be non-linear. As WG speaks more rapidly, there comes a point at which the non-final syllable /peɪ̯n/ does not compress further while the trisyllabic foot *message was* is still getting shorter. This effect is less obvious when the syllable is in final position (right hand panels).

While the non-linearity is not as obvious in the data for the other subjects, nothing in the following analysis depends on linearity in this relationship between rate and syllable duration.

For statistical analysis of the data from all four subjects, syllable durations were rank ordered by rate and then binned into four bins with 15 tokens per bin. Inhomogeneity of variance, as evidenced by Levine's test (Snedecor and Cochran, 1989) was corrected for by taking a log transform of the duration data. A $2*2*4$ (PFL*CE*RATE) factorial analysis with repeated measures on all factors was then carried out. The main effects associated with PFL [F(1,3) = 89.24, $p < 0.01$], CE [F(1,3)

Insert Figure 3 about here

FIG. 3.

= 916.6, $p < 0.001$] and rate [F(1,3) = 12.26, $p < 0.05$] were all significant at $\alpha = 0.05$. The rate effect was evaluated after the conservative Geisser-Greenhouse adjustment to the degrees of freedom to allow for non-circularity (Hays, 1988, 525). All 2-way interactions and the 3-way interaction were not significant at $\alpha = 0.05$.

C. Estimating the degree of lengthening

In order to estimate the degree of lengthening of an individual token y_i which is directly attributable to the factor PFL and/or CE, a prediction of the duration of that token in the baseline ([-PFL,-CE]) condition, $\hat{y}_0$, based on its measured rate, is required. As shown in Fig 2, the relationship between rate and target syllable duration is nonlinear, precluding a simple linear model of the form $\hat{y}_0 = mx + c$. However, a more general additive model of the form $\hat{y}_0 = f(x)$ is possible, where the function $f(x)$ is estimated using a locally linear smooth fit to the data. For a given token y_i in the [+PFL,-CE] condition, the proportion of its duration attributable to the factor [+PFL] is $y_i - \hat{y}_0$.

The Splus function "lo", which fits a locally weighted least squares linear regression was used for all smoothing (Statistical Sciences, 1995). This fit was used as a predictor in a generalized additive model. The dashed lines in Figure 2 illustrate the local fits computed for each condition for Subject WG.

Figure 3 shows the proportion of total syllable duration attributable to condition-specific factors for the [+PFL], [+CE] and [+PFL,+CE] conditions. From this figure, it is evident that PFL and CE contribute approximately the same amount to total syllable lengthening when each is present alone. Only subject JC shows a consistent difference, with [+CE] occasioning greater lengthening than [+PFL].

Because the estimates of lengthening are based on a smooth fit to the data, rather than on the raw data directly, lengthening estimates are not independent from token to token, and rate effects cannot be estimated using classical methods. However one main effect of rate stands out: the dissapearance of any lengthening for either prosodic effect at the fastest rates for Subject WG. No other subject approaches these extremely fast production rates. Within the range of rates produced by all speakers (approximately 1.2–2.5 feet/sec), there is no obvious systematic effect of rate on the proportion of duration attributable to lengthening. It is also apparent that the proportion of lengthening which is due to each of the two factors changes in similar fashion as rate changes. The lengthening due to [+PFL] thus seems to be directly comparable to that due to [+CE], irrespective of speaking rate.

D. Comparison with model-based predictions

Klatt's model given in Equation 1 was originally intended to account for the lengthening or shortening of segments, and in fact, Klatt applies it to compute shortened durations only. The model extends directly to the prediction of lengthening ($k > 1$) and for estimating durational changes in units larger than the segment. Given two factors, each of which adds $l_i, i \in \{1, 2\}$ to a baseline duration, a simple additive model predicts that the combined effect will add $l_1 + l_2$ to the baseline. Klatt's model predicts a superadditive effect equal to $l_1 + l_2 + \frac{1}{(D_0 - D_{min})} l_1 l_2$. The magnitude of the superadditive term $\frac{1}{(D_0 - D_{min})} l_1 l_2$ depends on the size of the hypothesized incompressible portion

Insert Figure 4 about here

FIG. 4.

(estimated by Klatt to be about 0.45 for vowels. See Klatt 1975), and is minimized for $D_{min} = 0$[1].

Figure 4 shows the predicted length of the target syllable in the [+PFL,+CE] condition as generated by a simple additive model and by Klatt's model, where the incompressible portion D_{min} has been fitted for each subject separately. The actual data have been included in the plot. For WG, CB and GC the simple additive model provides a better fit as estimated from the sum of the squared residuals, and in each case, the best fit using Klatt's model requires $D_{min} = 0$. For JC, Klatt's model is the better fit and $D_{min} = 0.61$.

IV. Discussion

This study examined the combination of lengthening effects over a range of (continuously measured) speaking rates. Although previous studies of durational modifiers in combination have suggested that there are limits to both expansion and compression of syllable duration, a simple additive model was found to provide a good fit over a wide range of rates. An algorithmic approach to computing durations could conceivably accommodate limits on both expandibility and compressibility by treating the segment (or syllable) as a hard spring with a neutral or preferred duration. Over some medial range, factors which influence duration combine in simple additive fashion, while

[1] Although naïve as a production model, Klatt's model still finds application in synthesis algorithms and it has the inestimal virtue of making numerically testable predictions.

beyond that range, little compression or stretching is effected.

This simple interpretation is premised on the assumption that timing factors all behave in approximately the same manner. Port (1981) raised the possibility that factors which instantiate phonological features may combine by constant ratios, while other factors (tempo, number of syllables in a word) may be more likely to exhibit subadditive combination. This accords well with the present data, where both PFL and CE can be seen as phonologically specified, and both are manifested by lengthening of constant proportion. The influence of tempo on durations was more complex, but for all but the fastest rates, relative proportions were essentially unaffected by tempo. Models of constant proportion of duration have not fared well as predictors of timing in speech movement (Löfqvist, 1991). They may, however, be appropriate in the acoustic domain for specific factors over a wide range of rates.

Much work remains to be done in examining the edges of the range within which simple timing effects are found. Fast speech studies need to take into account the existence of a continuum of rates, with the likelihood of a discontinuity or reorganization in articulation at some fast rate, beyond which timing is likely to be heavily influenced by the biomechanical limitations of the production system and less obviously dictated by linguistic factors. Slow speech (as opposed to clear speech, see e.g. Uchanski, Choi, Braida, Reed and Durlach, 1996) has attracted less attention to date, but merits closer scrutiny not least because it serves to demarcate a range of “normal” operation of speech production.

V. Acknowledgments

The present work has benefited from discussions with Janet Pierrehumbert, Art Kovitz, Bob Port and Ken de Jong. Funding came from a postdoctoral fellowship from the Andrew W. Mellon Foundation and SNF grant 21-49144.96 (Long Short-Term Memory).

References

Beckman, M. E., Edwards, J., and Fletcher, J. (1992). "Prosodic structure and tempo in a sonority model of articulatory dynamics," in *Gesture, Segment, Prosody: Papers in Laboratory Phonology II*, edited by G. J. Docherty and D. R. Ladd (CUP, Cambridge), pp. 68–86.

Berkovits, R. (1991). "The effect of speaking rate on evidence for utterance-final lengthening," Phonetica **48**, 57–66.

Byrd, D. and Tan, C. C. (1996). "Saying consonant clusters quickly," Journal of Phonetics **24**, 263–282.

Cooper, W. E., Eady, S. J., and Mueller, P. R. (1985). "Acoustical aspects of contrastive stress in question-answer contexts," Journal of the Acoustical Society of America **77**, 2142–2156.

Crystal, T. H. and House, A. S. (1988a). "The duration of American-English vowels: an overview," Journal of Phonetics **16**, 263–284.

Crystal, T. H. and House, A. S. (1988b). "Segmental durations in connected-speech signals: Current results," Journal of the Acoustical Society of America **83**, 1553–1573.

Hays, W. L. (1988). *Statistics* (Harcourt Brace, Orlando, FA), fourth ed.

Klatt, D. H. (1973). "Interaction between two factors that influence vowel duration," Journal of the Acoustical Society of America **54**, 1102–1104.

Klatt, D. H. (1975). "Vowel lengthening is syntactically determined in a connected discourse," Journal of Phonetics **3**, 129–140.

Klatt, D. H. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," Journal of the Acoustical Society of America **59**, 1208–21.

Löfqvist, A. (1991). "Proportional timing in speech motor control," Journal of Phonetics **19**, 343–350.

Port, R. F. (1981). "Linguistic timing factors in combination," Journal of the Acoustical Society of America **69**, 262–274.

Shaiman, S., Adams, S. G., and Kimelman, M. D. Z. (1995). "Timing relationships of the upper lip and jaw across changes in speaking rate," Journal of Phonetics **23**, 119–128.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods* (Iowa State University Press, Ames, IA).

Statistical Sciences (1995). *S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.3* (StatSci, a division of MathSoft, Inc., Seattle).

Turk, A. E. and Sawusch, J. R. (1997). "The domain of accentual lengthening in American English," Journal of Phonetics **25**, 25–41.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. (1996). "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," Journal of Speech and Hearing Reaearch **39**, 494–509.

Table I

	-PFL	+PFL
-CE	X: painful SHOT Y: painful BLOW	X: WINdowpane Y: COUNTerpane
+CE	X: ARTfully Y: PAINfully	X: cancer PILL Y: cancer PAIN

Words used in the X and Y slots of the first and second sentences for each condition. Capitalization was used to highlight the required contrastive emphasis. The labels PFL and CE refer to the target syllable /peɪ̯n/.

Figure Legends

Figure 1 Distribution of the reciprocal of the duration of the words *message was*, shown as a function of nominal rate. White horizontal bars show the median, filled bars delimit the interquartile range, and the whiskers mark the range of values lying within 1.5*interquartile range. Points outside this range are shown individually.

Figure 2 Syllable durations as a function of (continuously valued) speech rate, fast rates to the right. Only subject WG is shown here. Each dataset has been fitted with a smooth curve based on locally weighted least squares linear regression.

Figure 3 Lengthening ascribable to [+PFL] (open diamonds), [+CE] (closed triangles) and both (open circles). The proportion of lengthening is based on an estimate of the unlengthened [-PFL,-CE] condition.

Figure 4 Comparison of predictions from a simple additive model (crosses), Klatt's 1975 model with best fit of D_{min} (open circles) , and actual data (filled triangles) in the [+PFL,+CE] condition.

Figure 1

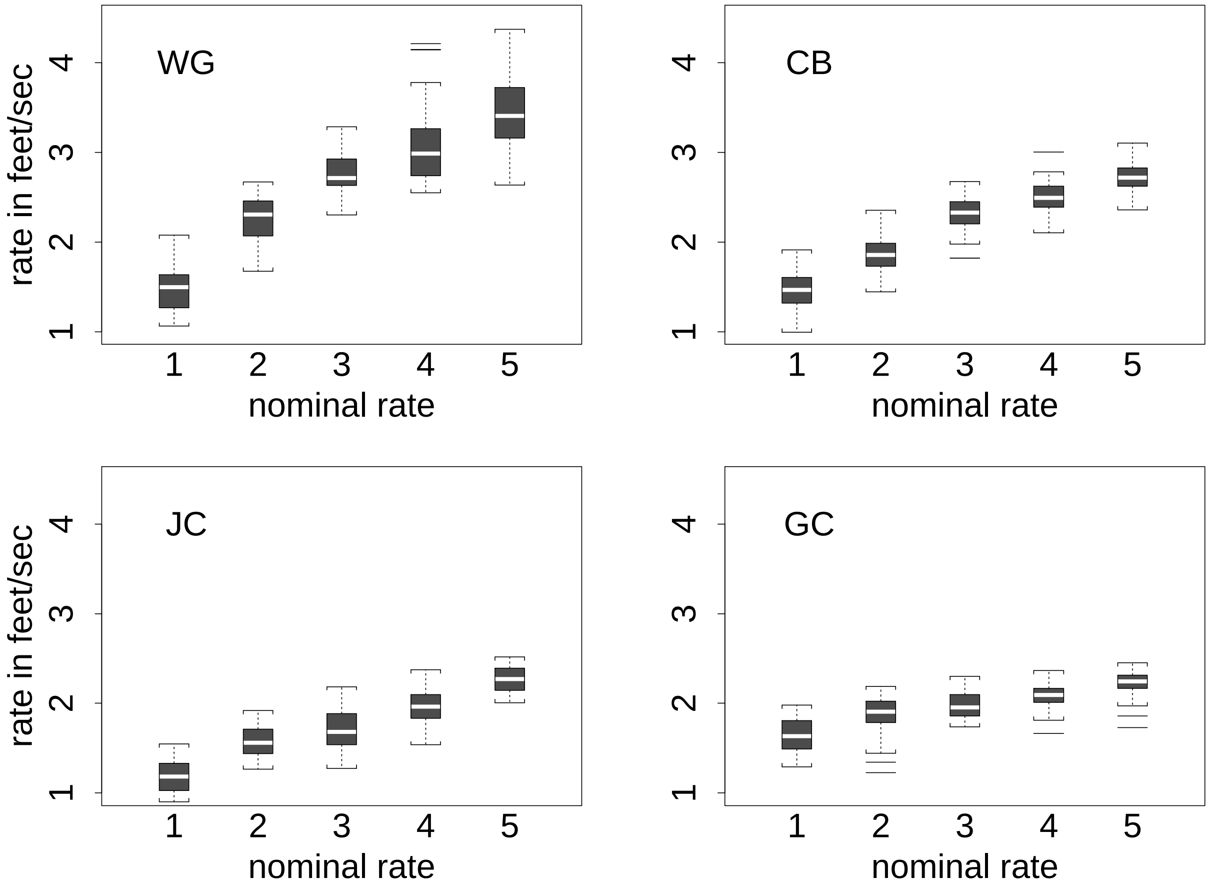

WG
CB
JC
GC
rate in feet/sec
rate in feet/sec
1
2
3
4
1
2
3
4
5
nominal rate
nominal rate
nominal rate
nominal rate

Figure 2

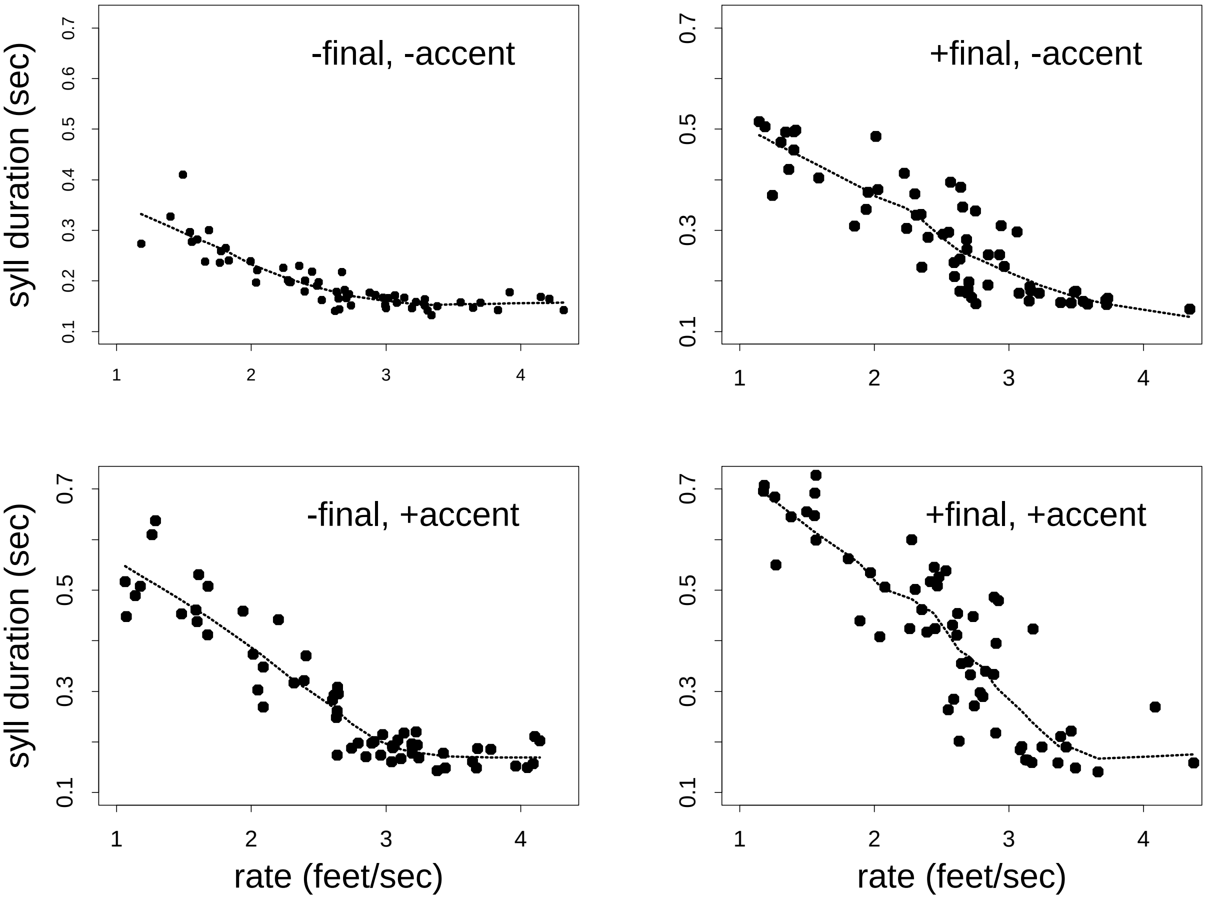
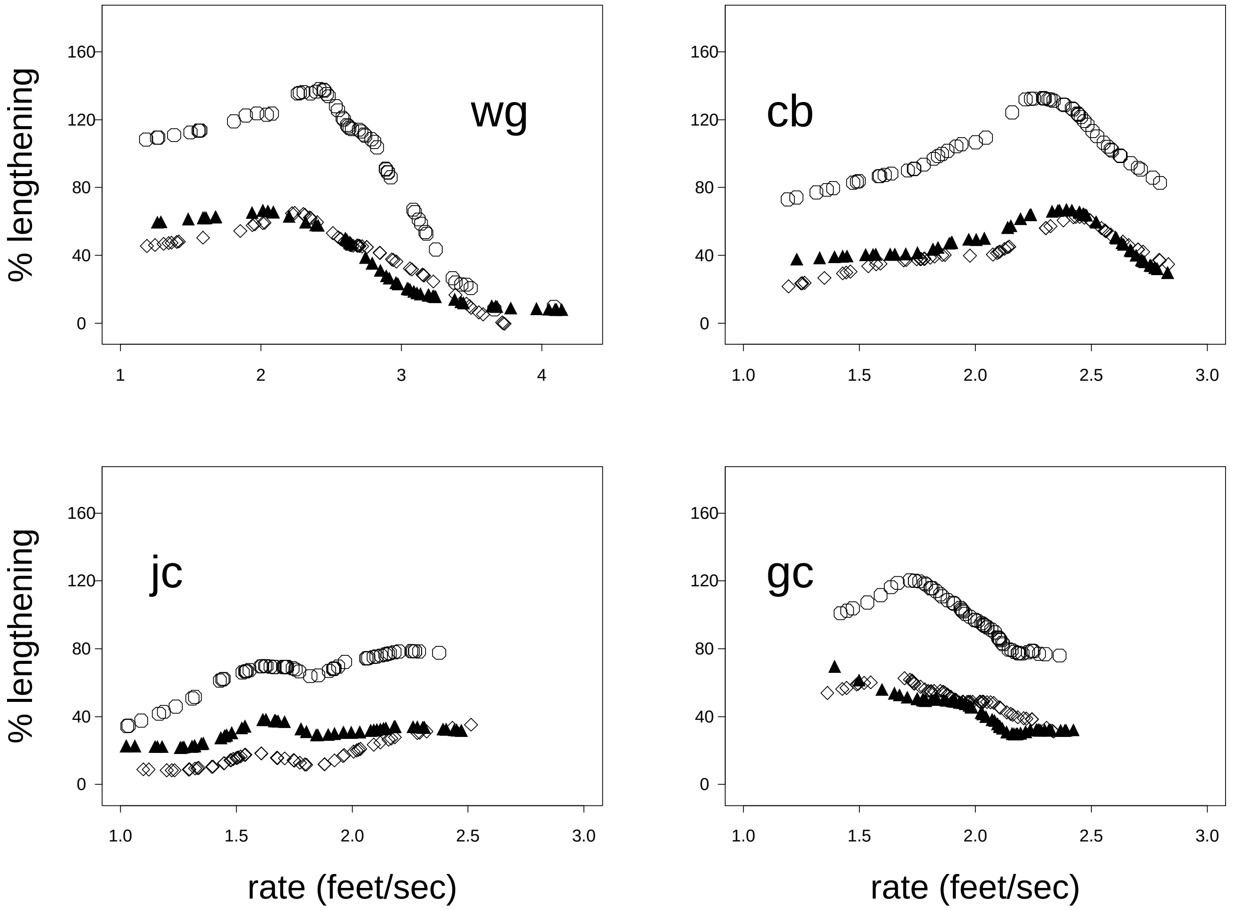
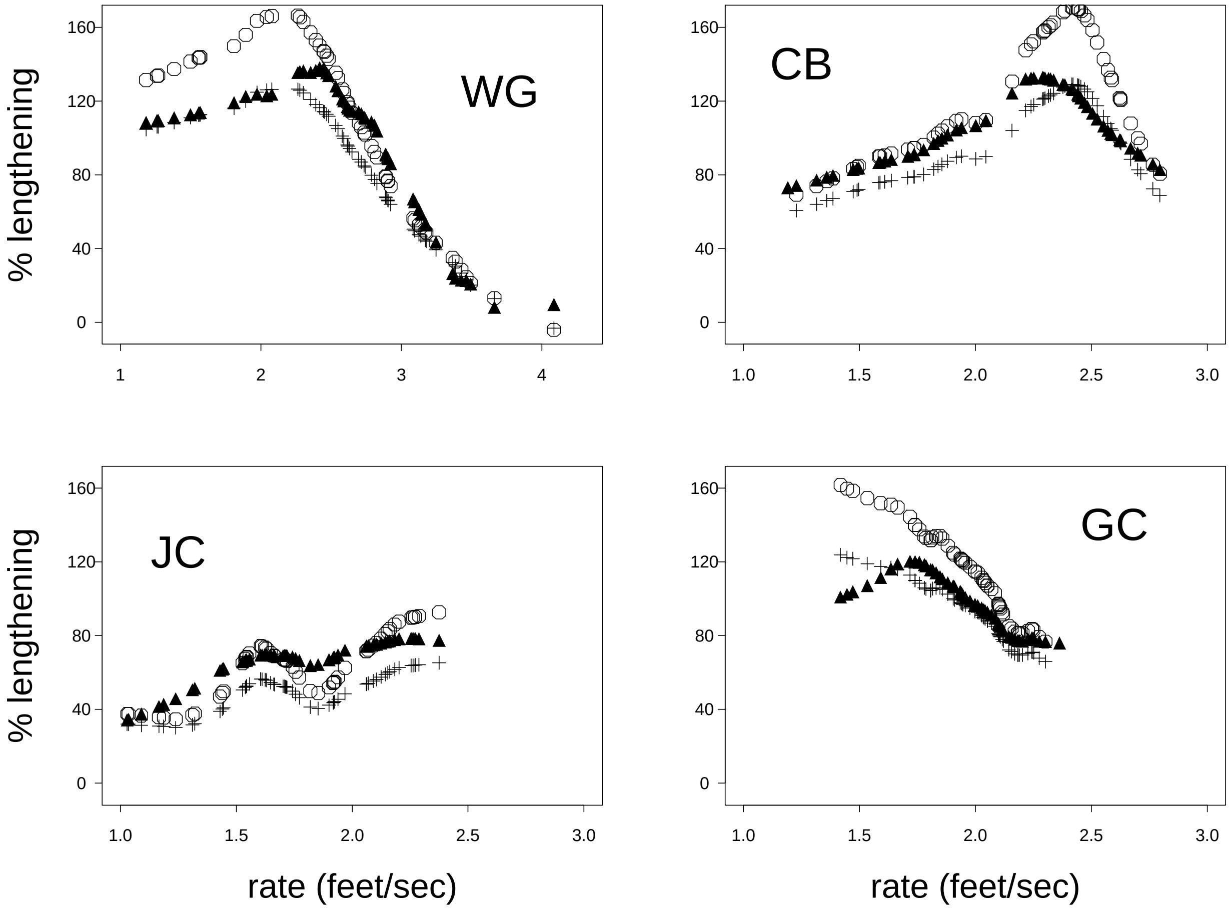

-final, -accent
+final, -accent
-final, +accent
+final, +accent
syll duration (sec)
syll duration (sec)
rate (feet/sec)
rate (feet/sec)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.1
0.3
0.5
0.7
1
2
3
4

Figure 3



% lengthening
wg
cb
jc
gc
0
40
80
120
160
1
2
3
4
1.0
1.5
2.0
2.5
3.0
rate (feet/sec)

Figure 4



WG
CB
JC
GC
% lengthening
% lengthening
0
40
80
120
160
1
2
3
4
1.0
1.5
2.0
2.5
3.0
rate (feet/sec)
rate (feet/sec)