10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Listeners' sensitivity to syllable complexity in speech tempo perception

Leendert Plug[1], *Robert Lennon*[1], *Rachel Smith*[2]

[1]University of Leeds, UK
[2]University of Glasgow, UK
l.plug@leeds.ac.uk

Abstract

Studies of speech tempo commonly use syllable or segment rate as a proxy measure for perceived tempo. While listeners' sensitivity to syllable rate is well-established [1-4], clear evidence for listeners' additional sensitivity to segment rate—that is, to syllable complexity alongside syllable rate—is as yet lacking. In [5, 6] we reported on experiments that showed no evidence for listeners' orientation to segment rate differences between stimuli that have the same syllable rate. In these experiments, we kept syllable rate constant by working with a single carrier phrase and equalizing phrase durations. Given that phrase duration is a separate temporal parameter from syllable rate, it is important to complement this work with experiments using less homogeneous stimulus sets, in which syllable rate is controlled without equalizing stimulus durations. In this paper we report on an experiment that uses stimuli selected from a corpus of unscripted British English speech. Within crucial subsets there was minimal variation in one out of syllable and segment rate, and substantial variation in the other. Stimulus duration varied independently. Listeners ranked stimuli for perceived tempo. Results suggest that faced with these more variable stimuli, listeners do orient to segment rate in ranking stimuli that have near-identical syllable rates—presumably reflecting the influence of syllable complexity. Moreover, stimulus duration emerges as a separate factor influencing listeners' rankings.

Index Terms: speech perception, tempo, syllable structure, unscripted speech

1. Introduction

Syllable and segment (or phone) rate are often used as proxy measures for perceived tempo. As we have pointed out previously [5, 6], these measures can yield quite divergent results in languages whose phonologies allow substantial variation in syllable complexity. For example, English allows a wide range in syllable shapes, such that one syllable can contain between one and seven segments. Increases in syllable complexity are not associated with uniform increases in syllable duration: increased onset complexity in particular is accompanied by a relative shortening of consonants, such that the midpoint of the onset is in a stable timing relation with that of the vowel [7, 8]. Thus, as syllable complexity increases, segment rate (the number of individual sound segments, or phones, per second) tends to go up, but syllable rate tends to go down [9]. In other words, syllable and segment rate can make different predictions as to the ranking of utterances according to perceived tempo. In this paper, we assess the impact of syllable and segment rate variation on listeners' impressions of speech tempo, extending previous work with highly controlled stimuli to unscripted speech.

Many tempo perception studies report that syllable rate measurements correlate with listeners' tempo judgements in the region of r=0.80 [1, 2]. Research into rhythm perception also highlights listeners' attention to syllable rate when judging whether utterances are rhythmically alike or distinct [10, 11]. Evidence that segment rate is a separate influence on tempo perception, however, is scarce if not lacking altogether. In [3], listeners ranked short utterances from a corpus of German spontaneous speech for tempo; tempo rankings were then correlated with rate measurements including both syllable and segment rate. The correlation between the two rates was not controlled. Both yielded pairwise correlations with tempo rankings in the region of r=0.80, and a regression analysis suggested both had independent explanatory value in modelling the rankings. This can be taken as evidence for listeners' sensitivity to syllable complexity alongside syllable rate, but it is indirect at best.

In [5, 6] we systematically varied the segment rate of English phrases on a constant syllable rate by embedding monosyllabic nouns of varying degrees of complexity (CVC, CCVC, CCVCC etc.) in the phrase structure *this* N_1 *or that* N_2: for example *this kit or that pack, this trust or that stock, this prank or that stunt*. Embedding the nouns in two positions in the utterance frame yielded a range of segment numbers across the phrases, while all had five syllables. We equalized phrase durations so that all phrases had the same syllable rate while varying in segment rate. Listeners compared phrases for tempo in a pairwise discrimination task. We found no evidence that listeners heard phrases with higher segment rates as faster.

In [5, 6] we equalized syllable rates by equalizing phrase durations. As phrase duration is an independent temporal parameter from syllable rate, it might in principle be independently relevant for tempo perception. Thus, our design may have created too much temporal uniformity within phrase pairs for listeners to orient to phrase-internal variation. A follow-up experiment [12] kept syllable rate constant across three template phrases with different syllable numbers and thus different overall durations, and here we did see a significant relationship between segment rate and perceived tempo. These observations warrant further research using less homogeneous stimulus sets, i.e. stimuli in which syllable rate is controlled independently from stimulus duration. In this paper we report on an experiment along these lines. The experiment has a similar overall design to that of [3], but incorporates a more systematic approach to stimulus selection, derived from [13], which yielded subsets of stimuli within which there was minimal variation in one of the two rates but substantial variation in the other.

10.21437/SpeechProsody.2020-2

2. Method

2.1. Participants

55 monolingual native English speakers (40 female; mean age 23; age range 18–36) participated in the experiment. All reported normal hearing, and all received payment. Since listeners' tempo perceptions might be informed by their own production tendencies [14], participants completed three short speech production tasks before commencing the tempo rating task described below. We describe these tasks in [15]; as the measures extracted from them proved uninformative in modelling listeners' tempo ratings we leave them aside here.

2.2. General design

The experiment as a whole was designed to allow three analyses, each of which compared two rate parameters in terms of their mappings to tempo judgements: (1) canonical vs surface syllable rate, (2) canonical vs surface segment rate, and (3) surface syllable vs segment rate. For each, we constructed a set of 60 stimuli. Analyses (1) and (2) are described in [15]; here we focus exclusively on analysis (3).

2.3. Stimulus selection

We selected stimuli from a corpus of 920 'memory stretches' extracted from the DyVIS database [16] by [17], produced by 30 male Standard Southern British English speakers aged 18–25. The data comprise stretches of unscripted, although guided speech: the speakers were given a scenario in which they were accomplices in a crime settling on a narrative to report in subsequent police interviews. Mean stretch duration is 1.5 sec (range 0.5–2.7). We used WebMAUS [18] for phone-level segmentation, with a protocol for correcting substantive misparsings. We derived canonical and surface syllable and segment rates from the output segmentations. As the four rates were highly inter-correlated (r=0.84–0.91) across the corpus, it was challenging to select stretches that would allow for meaningful pairwise comparisons of rates' mappings to tempo judgements. The stimulus set needed to comprise subsets within which syllable rate was close to constant but segment rate varied considerably, and vice versa. To this end, we selected a set of 60 stimuli, using a method along the lines of that of [13]. Our starting point was a scatterplot of the two relevant (log) rates in all 920 stretches: see Figure 1. We identified the 10–20%, 45–55% and 80–90% quantile ranges for both rates to represent slow, medium and fast rates respectively. Within each of these narrow ranges on each axis, we selected 10 data points that were as widely dispersed on the other axis—that is, in the comparison rate's range—as allowed by the shape of the overall scatter. For Figure 1, this yields three sets of 10 stimuli that are very similar in syllable rate but vary substantially in segment rate (dots) and three sets of ten stimuli that are very similar in segment rate but vary substantially in syllable rate (triangles).

2.4. Acoustic analysis

While the use of stimuli sampled from a corpus of unscripted speech maximizes the ecological validity of a tempo judgement task, it also introduces variables that may have an impact on participants' judgements. Multiple studies have shown that utterances with a relatively high overall f0 level, a relatively high magnitude of f0 movement and a relatively high overall intensity are perceived as relatively fast [19-21]. Therefore, we extracted f0 and intensity measures for all of the stimuli using Praat [22]. We used Mausmooth [23] to extract editable F0 contours (time step 0.05s, range 15–400Hz). We manually corrected clearly erroneous points before calculating the mean f0 for each corrected contour as a measure of f0 level and the f0 distribution's kurtosis as a measure of f0 span—acknowledging that the perceptual relevance of these and related f0 measures remains a matter of investigation [24, 25]. We also took a mean intensity measure for each stretch. We used these acoustic measures (which were not significantly inter-correlated) as control variables in our quantitative analyses.

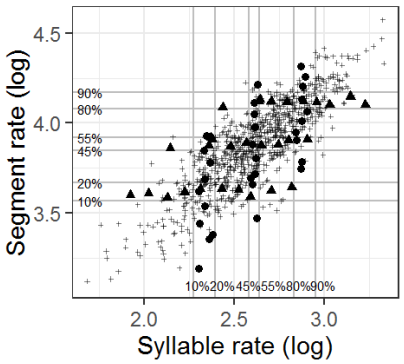

Figure 1: Scatterplot illustrating the stimulus selection procedure: see text for details. Black dots and triangles represent selected stimuli.

2.5. Tempo rating task

We elicited perceptual tempo ratings using an on-screen interface similar to that of [3], implemented in PsychoPy2 [26]. The stimuli in each set of 60 were presented together on one screen in the form of a vertical line of colored dots in the centre of the screen. When the participant clicked on a dot, an orthographic transcription of the stimulus appeared on the screen, and the corresponding audio played over headphones. The participant's task was to move each dot along a horizontal reference line to reflect its perceived tempo. Vertical gridlines and the labels 'Slowest, Slower, Average, Faster, Fastest' aided orientation. Stimuli appeared in the same randomised order for all participants. Participants could listen to stimuli repeatedly and revise their ratings until they were happy with the overall ranking of the 60 stimuli on each screen.

2.6. Quantitative analysis method

Dot placements were extracted as ratings on a scale between 0 and 1000, with 500 corresponding to the dot's original position and a perception of 'average speed', 0 meaning maximally slow and 1000 meaning maximally fast. We analyzed the ratings through fitting linear mixed effects models using the lme4 package [27] in R [28]. Participant and speaker identities were treated as random intercepts. Stimulus duration, f0 mean, f0 kurtosis and intensity mean were assessed as fixed effects. Durations and rates were log-transformed prior to modelling.

We took the set of 60 stimuli illustrated in Figure 1 as a starting point in constructing two focused data sets: one in which segment rate was the 'stable' rate and syllable rate the 'variable' one (Set A), and one in which syllable rate was the 'stable' rate and segment rate 'variable' (Set B). We maximized the size of these data sets by adding any other stimuli from the experiment as a whole that fell within the appropriate quantile ranges, although we had selected them for analyses (1) and (2)

described in 2.2 above. We narrowed the quantile ranges where relevant to ensure that correlations between syllable and segment rate within the quantile range subsets were all below *r*=0.3. The resulting data sets are shown in Figure 2. (Since these data sets comprise stimuli that appeared on different screens in the experimental setup, we assessed whether ratings varied systematically by screen—and found no evidence that they did.)

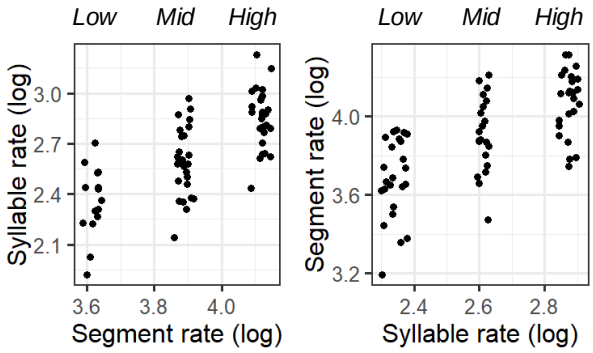

Figure 2: Scatterplots for Sets A (left) and B (right), each with the 'stable' rate on the x-axis and the 'variable' rate on the y-axis. Each data point represents one stimulus. Low, Mid and High subsets are labelled.

While we could in principle model participants' tempo ratings for each of the smallest subsets of stimuli (Low, Mid, High in each of Sets A and B) separately, resulting in six models, we deemed it preferable to fit fewer models over larger sets of stimuli. We therefore fitted one model for each of Sets A (71 stimuli, 3905 ratings) and B (70 stimuli, 3850 ratings). For each set, we modelled ratings across the Low, Mid and High subsets. This raised a methodological issue in that the 'stable' rate is only close to stable *within* these subsets, and the 'variable' rate varies systematically across them. To ensure that our 'stable' and 'variable' rate measures remained independent even when modelling ratings across the Low, Mid and High subsets, we centred and standardized the 'variable' rate measures within the Low, Mid and High subsets. This removed all variation in the 'variable' rate that correlates with the observed variation in the 'stable' rate between subsets.

In modelling tempo ratings for each of Sets A and B, we first fitted a control model with random intercepts for listener and speaker identities and a three-level fixed factor for stimulus subset (Low, Mid, High). We predicted that this factor would yield significant effects, such that stimuli in the Low subset would be rated lower (i.e. slower) than stimuli in the Mid subset, and stimuli in the High subset would be rated higher (i.e. faster) than stimuli in the Mid subset. We then assessed the relevance of our acoustic factors—stimulus duration, f0 mean, f0 kurtosis, intensity mean—before turning to our rate variables. We predicted that the 'stable' rate variable would lack predictive power with stimulus subset already accounted for. Adding the z-scored 'variable' rate measure allowed us to establish whether participants' ratings were systematic in relation to rate variation captured only by the 'variable' rate measure. Critically, we could test whether segment rate variation affected tempo perception across sets of relatively spontaneously-spoken stimuli with variable duration but close to stable syllable rates.

3. Results

3.1. Set A stimuli

In Set A, segment rate was the 'stable' rate and syllable rate the 'variable' one. Given the available evidence that listeners orient to syllable rate in estimating speech tempo, we predicted a significant positive relationship between listeners' tempo ratings and our z-scored syllable rate variable. As explained above, we also predicted that once the stimulus subsets Low, Mid and High were distinguished by a categorical variable, a segment rate variable would not add predictive power.

The modelling method outlined above resulted in an optimal model with the fixed effects summarized in Table 1. For the stimulus subset variable, Mid was treated as the reference level. The predicted effect of this variable was indeed observed: stimuli in the Low subset were rated as having a lower tempo compared with stimuli in the Mid subset, and stimuli in the High subset were rated as having a higher tempo. A segment rate variable showed no additional effect. By contrast, our z-scored syllable rate variable showed a positive effect: across the Set A stimuli, those with relatively high syllable rates were rated as relatively fast. Figure 3 illustrates the effect, which is observed in each of the three stimulus subsets. The model also shows a negative effect for stimulus duration—relatively long stimuli were rated as relatively slow—and positive effects for f0 mean and intensity mean.

Table 1: Fixed effects in an optimal model for Set A tempo ratings (est=estimate, se=standard error, df=degrees of freedom, t=t-statistic, p=probability).

	est	se	df	t	p
(intercept)	–119.2	73.6	802	–1.6	0.106
Low	–64.1	6.2	3588	–10.2	<0.001
High	14.5	5.0	3741	2.8	0.004
log duration	–22.1	5.2	3450	–4.2	<0.001
f0 mean	1.2	0.2	576	5.2	<0.001
intensity mean	7.9	1.2	571	6.5	<0.001
syllable rate (z)	10.9	2.2	3518	4.7	<0.001

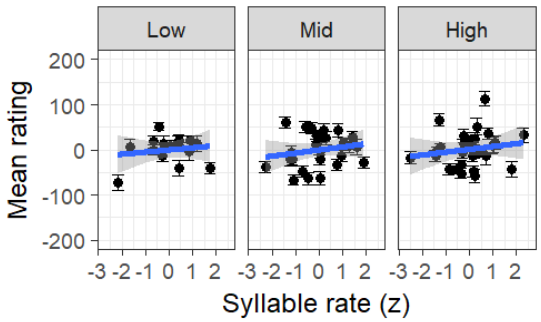

Figure 3: Scatterplots of (z-scored) syllable rate (x-axis) against tempo ratings (y-axis; raw values have been replaced by the control model residuals) within Subset levels for Set A, with linear fit lines. Data points are average ratings associated with all unique x-axis values. Whiskers are standard errors.

3.2. Set B stimuli

In Set B, syllable rate was the 'stable' rate and segment rate the 'variable' one. Again we predicted that once the stimulus subsets Low, Mid and High were distinguished by a categorical variable, a syllable rate variable would not add predictive power. Our crucial question was whether our z-scored segment

rate variable *would* significantly improve model fit. As shown in Table 2 and Figure 4, this was indeed the case: among stimuli with very similar syllable rates, those with relatively high segment rates—that is, relatively complex syllables—were rated as relatively fast. Like the model for Set A, that for Set B also shows a negative effect for stimulus duration—relatively long stimuli were rated as relatively slow—and positive effects for f0 mean and intensity mean. For the stimulus subset variable, the predicted effect was not entirely observed: stimuli in the Low subset were rated as having a lower tempo compared with stimuli in the Mid subset, but stimuli in the High subset were not rated as having a significantly higher tempo. As expected, with this variable in the model, entering a syllable rate variable did not improve fit.

Table 2: Fixed effects in an optimal model for Set B tempo ratings; see Table 1 for further details.

	est	se	df	t	p
(intercept)	–102.7	63.1	772	–1.6	0.1
Low	–32.2	5.4	3635	–5.8	<0.001
High	–3.4	5.9	3341	–0.5	0.6
log duration	–35.5	5.9	2724	–5.9	<0.001
f0 mean	2.2	0.3	441	7.4	<0.001
intensity mean	6.4	1.1	450	5.7	<0.001
segment rate (z)	23.2	2.4	3192	9.6	<0.001

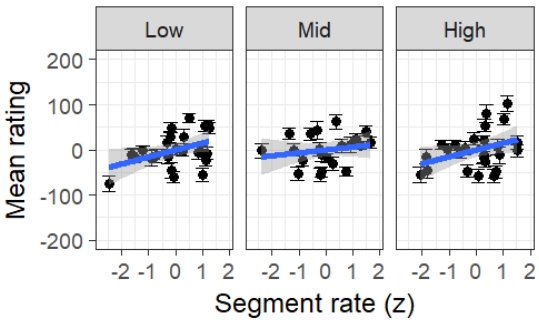

Figure 4: Scatterplots of (z-scored) segment rate (x-axis) against tempo ratings; see Figure 3 for further details.

4. Discussion

Our earlier research [5, 6] found that English listeners hear little difference in tempo among stimuli that vary in segment rate but are constant in syllable rate. Our aim in this study was to assess whether this finding generalizes to experimental designs with less homogeneous stimulus sets: in particular, designs in which stimulus duration and syllable rate are independent parameters. The results suggest that stimulus duration is indeed treated as a separate parameter by listeners, such that stretches of speech that take a speaker longer to complete are judged as slower than stretches whose production takes less time. Moreover, with this effect accounted for, segment rate variation yields a significant positive effect: among stretches of speech with various durations but very similar syllable rates, stretches with higher segment rates—in other words, more complex syllable structures—are judged as faster than stretches with lower segment rates. The results confirm those from [12], so we have converging evidence from experiments using both highly controlled and unscripted materials.

Several other aspects of the results deserve comment. The results from the stimulus subset where segment rate was stable and syllable rate was variable confirm that syllable rate predicts perceived tempo well. The results for the control variables confirm that high mean f0 and mean intensity both support the perception of fast tempo. This is broadly in line with previous findings [19-21], although our measure of f0 span was not a significant predictor of tempo ratings.

The results suggest that when judging tempo, listeners attend to the rates of production—and by implication the durations—of linguistic units at multiple levels: segments, syllables, and phrases. Taken together with previous findings [5, 6, 12] they show that the interplay among these different units is not simple. In [5], keeping both syllable rate and phrase rate constant suppressed any influence of segment rate, yielding results that seemed at odds with the idea that any manipulation of relative spectral complexity should trigger systematic variation in perceived tempo [29]. In [12] and the present experiment, a stimulus set that was less homogeneous in phrase rate allowed the influence of segment rate to emerge. Together, our findings in relation to both durational and non-durational parameters underscore the multi-dimensional nature of tempo perception.

5. Acknowledgements

This research was supported by a Leverhulme Trust Research Grant. The authors would like to thank Erica Gold for sharing her corpus.

6. References

1. Den Os, E., Perception of speech rate of Dutch and Italian utterances. *Phonetica*, 1985. **42**: 124–134.
2. Vaane, E., Subjective estimation of speech rate. *Phonetica*, 1982. **39**(2-3): 136–149.
3. Pfitzinger, H. Local speech rate perception in German speech. *Proceedings of the International Congress of Phonetic Sciences 1999*. 1999. San Francisco.
4. Gibbon, D., K. Klessa, and J. Bachan, Duration and speed of speech events: A selection of methods. *Lingua Posnaniensis*, 2015. **56**(1): 59–83.
5. Plug, L. and R. Smith, Phonological complexity, segment rate and speech tempo perception. *Proceedings of Interspeech 2017*. 2017. Stockholm.
6. Plug, L. and R. Smith. Segments, syllables and speech tempo perception. *Proceedings of Speech Prosody 2018*. 2018. Poznan.
7. Marin, S. and M. Pouplier, Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control*, 2010. **14**(3): 380–407.
8. Byrd, D., C-centers revisited. *Phonetica*, 1995. **52**(4): 285–306.
9. Greenberg, S., et al., Temporal properties of spontaneous speech: A syllable-centric perspective. *Journal of Phonetics*, 2003. **31**(3–4): 465–485.
10. White, L., S.L. Mattys, and L. Wiget, Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, 2012. **66**(4): 665–679.
11. Arvaniti, A. and T. Rodriquez, The role of rhythm class, speaking rate, and F-0 in language discrimination. *Laboratory Phonology*, 2013. **4**(1): 7–38.
12. Plug, L. and R. Smith, Segments, syllables and speech tempo perception by English listeners. Under review, 2019.

13. Koreman, J., Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America*, 2006. **119**(1): 582–596.
14. Schwab, S., Relationship between speech rate perceived and produced by the listener. *Phonetica*, 2011. **68**(4): 243–255.
15. Plug, L., R. Lennon, and R. Smith, Measured and perceived speech tempo: Canonical vs surface syllable and phone rates. *Proceedings of the International Congress of Phonetic Sciences 2019*. 2019. Melbourne.
16. Nolan, F., et al., The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, 2009. **16**(1): 31–57.
17. Gold, E., Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters. PhD dissertation, 2014, University of York.
18. Kisler, T., U.D. Reichel, and F. Schiel, Multilingual processing of speech via web services. *Computer Speech & Language*, 2017. **45**: 326–347.
19. Feldstein, S. and R.N. Bond, Perception of speech rate as a function of vocal intensity and frequency. *Language and Speech*, 1981. **24**: 387–394.
20. Kohler, K.J., Parameters of speech rate perception in German words and sentences: Duration, f0 movement, and f0 level. *Language and Speech*, 1986. **29**: 115–139.
21. Rietveld, A.C.M. and C. Gussenhoven, Perceived speech rate and intonation. *Journal of Phonetics*, 1987. **15**(3): 273–285.
22. Boersma, P. and D. Weenink, *Praat: Doing phonetics by computer*. 2017, www.praat.org.
23. Cangemi, F., *Mausmooth*. 2015, http://phonetik.phil-fak.uni-koeln.de/fcangemi.html.
24. Mennen, I., F. Schaeffler, and G. Docherty, Cross-language differences in fundamental frequency range: A comparison of English and German. *Journal of the Acoustical Society of America*, 2012. **131**(3): 2249–2260.
25. Niebuhr, O. and R. Skarnitzl, Measuring a speaker's acoustic correlates of pitch – but which? A contrastive analysis based on perceived speaker charisma. *Proceedings of the International Congress of Phonetic Sciences 2019*. 2019. Melbourne.
26. Peirce, J., Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2009. **2**(10).
27. Bates, D., et al., Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 2015. **67**(1): 1–48.
28. R Development Core Team, *R: A language and environment for statistical computing*. 2008.
29. Weirich, M. and A.P. Simpson, Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics*, 2014. **43**: 1–10.