INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Phonological complexity, segment rate and speech tempo perception

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

[1]University of Leeds, United Kingdom
[2]University of Glasgow, United Kingdom
l.plug@leeds.ac.uk, rachel.smith@glasgow.ac.uk

Abstract

Studies of speech tempo commonly use syllable or segment rate as a proxy measure for perceived tempo. In languages whose phonologies allow substantial syllable complexity these measures can produce figures on quite different scales; however, little is known about the correlation between syllable and segment rate measurements on the one hand and naïve listeners' tempo judgements on the other.

We follow up on the findings of one relevant study on German [1], which suggest that listeners attend to both syllable and segment rates in making tempo estimates, through a weighted average of the rates in which syllable rate carries more weight. We report on an experiment in which we manipulate phonological complexity in English utterance pairs that are constant in syllable rate. Listeners decide for each pair which utterance sounds faster. Our results suggest that differences in segment rate that do not correspond to differences in syllable rate have little impact on perceived speech tempo in English.

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

1. Introduction

Studies of speech tempo commonly use syllable or segment rate as a proxy measure for tempo. However, in languages whose phonologies allow substantial syllable complexity these measures can produce figures on quite different scales. English is a case in point. Its phonology allows a wide range in syllable shapes, such that one syllable can correspond to two to 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 [2, 3]. As a result of this 'C-center effect', increases in syllable complexity tend to mean increases in segment rate but decreases in syllable rate: in the corpus of [4], the mean duration of a stressed CVC syllable is 310ms, and that of a stressed CCVC syllable 382ms. The former yields a segment rate of 9.7 and a syllable rate of 3.2; the latter a segment rate of 10.5 (up 8%) and a syllable rate of 2.6 (down 19%).

Surprisingly, there has been little research on the validity of using syllable and segment rates as proxy measures of tempo. More generally, it remains largely an open question how well acoustically-based measures reflect mechanisms by which naïve listeners estimate speech tempo. This question is far from trivial, as for various purposes, including speech synthesis and fluency assessment, it is more important to have reliable estimates of *perceived* tempo than it is to have precise measurements of *produced* articulation rate. In the study reported here, we address this general question by assessing the impact of segment rate manipulations on English listeners' impressions of speech tempo.

Several studies report correlations of listeners' tempo judgements and syllable rate measurements in the region of r=0.80 [5, 6], and [7] reports stronger correlations with measurements of 'consonant-vowel interval' rate. To date only one study has included segment rate measurements in the comparison [1]. In this research, listeners ranked a series of short utterances taken from a corpus of German spontaneous speech according to their perceived tempo; tempo rankings were then correlated with syllable and segment rate measurements. Both measurement methods yielded correlations in the region of r=0.80. While the correlation coefficients for segment and syllable rate were not significantly different, [1] reports that combining both measurements in a single equation, with syllable rate weighted higher, yields a significantly closer correlation with listeners' tempo judgements (r=0.91).

The findings of [1] are consistent with a model in which listeners attend to both syllable and segment rates in making tempo estimates. While the two rates yield similar tempo estimates, there will be occasions where they diverge. On these occasions listeners must refer either to one or the other rate, or to some weighted average of the two, in estimating speech tempo. The finding that in an optimal combination of the two rates for German, syllable rate is weighted higher than segment rate suggests that on occasions of rate 'mismatch', German listeners either default to estimating tempo on the basis of their impression of syllable rate, or use a weighted average of their impressions of both rates balanced in favour of syllable rate.

In the study reported here, we systematically create rate 'mismatches' by manipulating phonological complexity in English utterance pairs that are constant in syllable rate. The manipulations yield a range of segment rate differences across the utterance pairs. If listeners consistently default to using syllable rate as the basis for their tempo estimations when syllable and segment rate diverge in their implications, these utterance pairs should yield a majority of 'same tempo' responses. If listeners consistently work with a weighted average of the two rates, we should find evidence of a 'consequential difference' threshold for segment rate: small differences are likely not to have an impact on listeners' tempo judgements, but large differences should do.

2. Experimental design

We used a pairwise discrimination paradigm commonly used in research on speech tempo perception [8, 9]: subjects were asked to judge tempo differences in pairs of stimuli. We kept syllable rate constant both within and across pairs, while varying the syllable complexity of stimuli within pairs.

http://dx.doi.org/10.21437/Interspeech.2017-987

2.1. Speech materials

To create the stimuli, we used a short phrase containing two nouns of varying phonological shapes (*this* N1 *or that* N2). The phonological shapes were chosen to include no complexity (CVC), onset complexity only (CCVC), coda complexity only (CVCC), and both onset and coda complexity (CCVCC). We minimized segmental variation by allowing only voiceless obstruents in initial and final position, and allowing only short vowels in the nucleus: e.g. *pack*, *clock*, *tact*, *stunt*. Embedding the nouns in two positions in the utterance frame gives a range of segment numbers across the utterances, from 13 to 17; moreover, it allows us to assess whether the utterance position of complexity has an impact on tempo judgements. We created two sets of 16 stimuli with all logical combinations of N1 and N2 shapes. Nouns were not reused across N1 and N2 or across sets, and an attempt was made as much as possible to construct utterances with semantically compatible nouns that do not share onsets or vowel nuclei: e.g. *this kit or that pack*, *this trust or that stock*, *this pump or that plank*, *this prank or that stunt*. Stimuli were paired exhaustively across the two sets, resulting in 256 stimulus pairs. Stimuli were separated by a 0.5s silence.

We used the same phrase to produce 134 filler pairs. Each filler included one bisyllabic or trisyllabic noun, and one semantically related monosyllabic noun with no restrictions on segmental make-up: e.g. *this kestrel or that kite*, *this bean or that potato*, *this adventure or that tour*.

2.2. Recording and manipulation

Stimuli and fillers were recorded in a sound-proof studio by a female native speaker of Southern Standard British English. In order to minimize rhythmical and prosodic variation prior to manipulation, the speaker recorded one utterance as a model, and listened to it before producing each of the remaining utterances. Syllable rate varied between 3.19 and 3.87 sylls/sec, and the syllable rate distributions of the utterance constituents (*this*, N1, *or that*, N2) were normal, with no outliers.

All stimuli were manipulated using *PSOLA* in Praat [10]. The manipulation equalized durations, F0 contours and mean amplitudes across stimuli. Stimulus duration was set to yield a syllable rate of 4; this was confirmed as sounding neither noticeably fast nor slow in a small-scale survey with four listeners. The F0 contour was set to a stylized version of one of the natural tokens. Mean amplitude was set to 62 dB.

Fillers were manipulated for duration and amplitude only, so that across the experiment as a whole, participants heard some degree of pitch variation. Durations were set differently for subsets of fillers, so that across the filler set, syllable rate varied between 4 and 4.75, 60 out of 134 filler pairs had a measured syllable rate difference of 0.5 and smaller subsets had a measured difference of 0.25 or no difference. This was done to ensure that across the experiment as a whole, participants were regularly faced with 'easy' tempo judgements. Mean amplitude was again set to 62 dB.

2.3. Participants and procedure

The experiment was run at the University of Leeds in accordance with all institutional ethics regulations. 50 native British English listeners between the ages of 18 and 35 participated. None reported known hearing problems. All were reimbursed for their time.

The experimental procedure was similar to that described in [8, 9]. We used the *ExperimentMFC* facility in Praat [10] to run the experiment. Listeners were introduced to the task on-screen. Experimental and filler utterance pairs were presented in random order. For each pair of utterances, listeners were asked to indicate whether the second utterance was faster, slower or the same in tempo as the first utterance using a 7-point response scale ranging from -3 (slower) through 0 (same) to 3 (faster). The next pair played 1.5s after each judgement was recorded. Listeners could replay each pair once. The main experiment followed a familiarization run on five utterance pairs.

2.4. Analysis variables

Our analysis focused on the relationship between the listeners' responses (henceforth *Response*), and the difference in syllable complexity, and therefore segment rate, between the two utterances in the experimental pairs.

To capture complexity, we defined several variables. An overall measure, *Segment rate ratio*, was the segment rate of the second utterance divided by that of the first utterance. Values above 1 identify pairs with a more complex second member, values below 1 pairs with a more complex first member. The result of our selection of syllable shapes is that the smallest non-null segment rate difference within pairs is 6%, which is around the general Just Noticeable Difference for speech tempo [8], and the largest 31%.

Two further complexity measures were sensitive to the location of complexity within syllables. *Onset complexity ratio* and *Coda complexity ratio* were the number of onset/coda consonants in the second utterance minus the number of onset/coda consonants in the first utterance. Higher values meant that the second utterance had more complex onsets/codas than the first. A final measure, *Final noun complexity,* was designed to capture any recency effect in tempo estimation, i.e. the extent to which complexity was located at the end of the stimulus: it was the difference in number of segments between N2 and N1 in the second utterance.

We also included several control variables in our analysis. First, while the syllable rate distributions of *this*, N1, *or that*, and N2 were normal, suggesting a fairly consistent rhythm across stimuli, temporal analysis confirmed an expected significant correlation between the number of segments in N1 and N2 and the duration of these constituents, which survived the temporal manipulation of the stimuli (for N1, r=0.404, t(30)=2.42, p=0.022; for N2, r=0.419, t(30)=2.53, p=0.017). This means that the higher an utterance's segment numbers, the more of the utterance duration is taken up by N1 and/or N2, and the less by *this* and *or that.* As this systematic utterance-internal temporal variation may provide listeners with tempo cues, we incorporated constituent durations in our statistical analysis of responses. For each constituent (*this,* N1, *or that,* and N2, and for the sum of the two nouns) we defined two variables, *{Constituent} duration* (in milliseconds), and *{Constituent} duration ratio* (the duration of a given constituent in the second utterance, divided by that of the same constituent in the first).

Second, since our complex onsets and codas contain a mix of voiceless and voiced consonants, and we know that the ratio of voiced and voiceless portions of speech is relevant for rhythm perception [11], we used the *fraction of locally unvoiced frames* within Praat's *voice report* function to obtain the proportion of voiceless material for each utterance. For each utterance pair we divided the proportion for the second utterance by that for the first (*Voicelessness ratio*).

For filler pairs, we defined *Syllable rate ratio* as the syllable rate of the second utterance divided by that of the first.

2.5. Quantitative analysis

We used the *lme4* and *lmerTest* packages in R [12] to fit linear mixed-effects models for *Response*. We used the *step()* function to eliminate non-significant predictors.

3. Results

3.1. Responses to filler pairs

Before we turn to the listeners' responses to the experimental stimuli, we consider their responses to the filler utterance pairs. Figure 1 shows that listeners were sensitive to the syllable rate manipulations: responses (here averaged across listeners) were generally negative ('second utterance slower') when the second filler utterance had a lower syllable rate than the first, and positive ('second utterance faster') when it had a higher syllable rate. In a mixed-effects model with *Listener, Utterance1* and *Utterance2* as random effects, *Syllable rate ratio* had a strongly significant effect on *Listener response* (t=5.5, p<0.0001). This tells us that listeners were paying appropriate attention to the tempo of the utterance pairs they were exposed to.

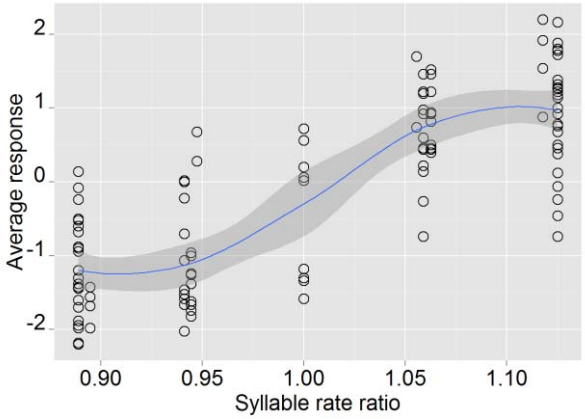

Figure 1: *Relationship in filler pairs between* Syllable rate ratio *and* Response *(averaged across listeners).*

3.2. Relationships between factors

As the *{Constituent} duration* and *{Constituent} duration ratio* variables were highly inter-correlated, we used exploratory modelling to select among them. We fit models containing each constituent duration variable plus *Voicelessness ratio* and *Segment rate ratio,* and selected the model with the lowest AIC value as the best fit. This procedure showed that *N2 duration ratio* was the most informative variable to capture duration, so it was entered into the main analysis.

Likewise, *Segment rate ratio* was highly correlated with *Onset complexity ratio,* with *Coda complexity ratio,* and with *Final noun complexity*. These latter three variables were not correlated with each other. The exploratory modelling procedure revealed that *Onset complexity ratio* and *Final noun complexity* were the most informative complexity variables, so they were entered into the main analysis.

Correlations between *Onset complexity, Final noun complexity, N2 duration ratio* and *Voicelessness ratio* were all low enough that collinearity was not a cause for concern (r<0.5 in all cases and r<0.2 in all but two cases). In other words, *N2 duration ratio* and *Voicelessness ratio* should not capture substantial proportions of any effect of our syllable complexity variables on *Response*.

3.3. Modelling *Response*

Listeners' responses to the experimental stimuli clustered close to zero, as seen in Figure 2: i.e. listeners perceived very little variation in tempo across the utterance pairs. We fitted a linear mixed-effects model with *Onset complexity ratio, Final noun complexity, N2 duration ratio* and *Voicelessness ratio* as fixed effects, and *Listener, Utterance1* and *Utterance2* as random effects. *Onset complexity ratio* and *Final noun complexity* were eliminated as non-significant. Table 1 shows the final model.

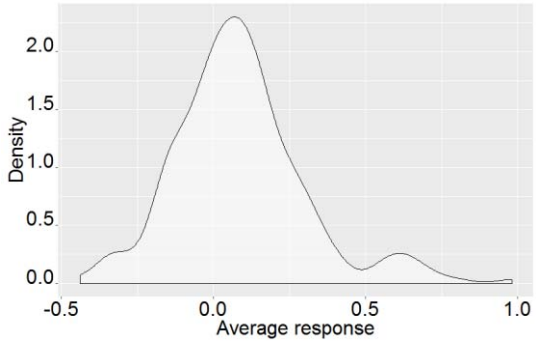

Figure 2: *Distribution of* Response *(averaged across listeners).*

Table 1: *Linear mixed-effects model for* Response.

Predictor	Estimate	SE	df	*t*	*p*
(Intercept)	-1.19	0.38	25	-3.1	<0.005
Voicelessness ratio	0.38	0.16	33	2.4	<0.025
N2 duration ratio	0.93	0.43	25	2.2	<0.05

We did not observe a significant effect of any complexity measure on tempo perception, i.e. there was no evidence that listeners judged utterances as faster when they had greater syllable complexity, and therefore a higher segment rate. Further inspection of the data revealed no evidence of utterance pairs with large complexity differences being judged more consistently than pairs with smaller differences: i.e. no evidence of a 'consequential difference' threshold for segment rate.

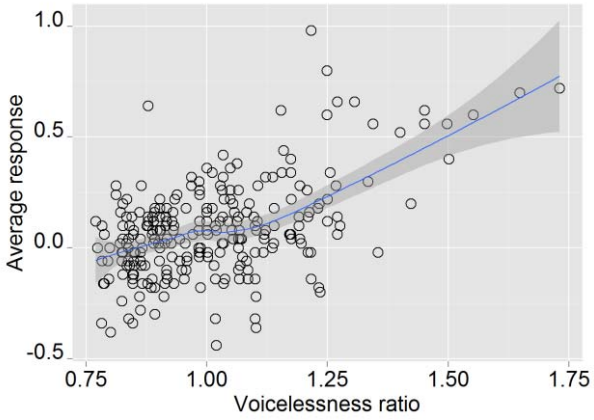

Figure 3: *Relationship between* Voicelessness ratio *and* Response *(averaged across listeners).*

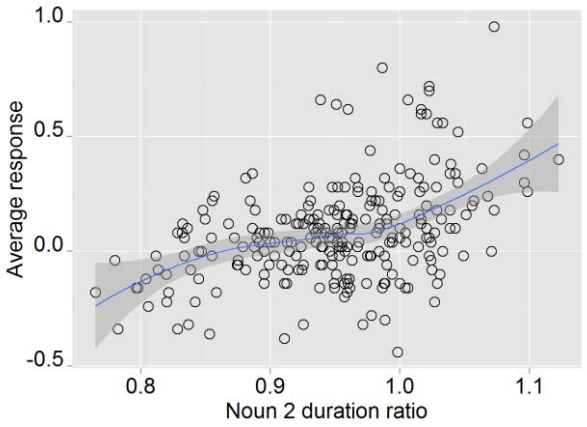

Figure 4: *Relationship between* N2 duration ratio *and* Response *(averaged across listeners)*.

Figure 3 illustrates the effect of *Voicelessness ratio*: the higher the proportion of voicelessness in the second utterance relative to the first, the more likely the second utterance was to be perceived as faster. Figure 4 illustrates the effect of *N2 duration ratio*: the longer the duration of N2 in the second utterance relative to the first, the more likely listeners were to perceive the second utterance as faster.

4. Discussion

As indicated above, our experimental design manipulated phonological complexity, and therefore segment rate in English utterance pairs that are constant in syllable rate. Given that previous research has shown that listeners attend to both syllable and segment rates in making tempo estimates, although syllable rate carries more weight in the calculation [1], we hypothesized that we might find evidence of a 'consequential difference' threshold for segment rate differences. We found no such evidence. Our findings suggest that even if English listeners track segment rate in judging speech tempo, when faced with a task in which the two rate calculations give rise to substantially different tempo estimates, listeners default to the syllable rate calculation.

It is possible that our design created too artificial a task, and that the recurrence of clause constituents within and across pairs (*this* N *or that* N) together with the identical utterance durations in the experimental pairs made listeners insensitive to segment rate differences. Still, our design was similar to that of [9], who found that listeners are sensitive to the peripherality vs centrality of vowels in pairs of otherwise near-identical utterances in judging their tempo. If this means, as [9] suggest, that listeners generally estimate complex spectral events as faster than less complex events that take the same amount of time to complete, then a lack of impact of syllable complexity is somewhat unexpected – especially since we measured it in multiple ways to allow for differential weighting of complexity in onsets vs codas and in recently heard words.

However, our listeners' tempo judgements do reveal sensitivity to other aspects of the segmental structure of the utterances. We found small but robust effects of two control variables. First, utterances with more acoustically voiceless material relative to their paired utterance were more likely to be judged as faster. For example, *this test or that step* might be judged faster than *this trump or that stomp,* despite having fewer segments. One interpretation is that rather than exclusively counting phonological segments, listeners assess the temporal distribution of consonantal and vocalic intervals in speech, with sonorant consonants behaving as vocalic. This tallies well with the findings that the proportions and standard deviations of voiced and voiceless intervals are a good proxy for the V and C measures that are known to correlate with rhythm class affiliation [11], and that acoustic 'rhythmicity parameters' yield tempo estimates that closely correlate with segment-based rate measurements [13]. Arguably, the presence of more voiceless material leads to a more spectrally differentiated signal, which is more complex in [9]'s sense.

Second, utterances with more internal variation in constituent durations (a longer N2, and therefore shorter unstressed syllables and/or a shorter N1) were likely to be judged as faster. This appears to be a rhythmic effect, i.e. an effect on tempo perception of local *alternation* in duration, and a sensitivity to the rate of the fastest parts of the utterance. Further work is needed to establish whether controlling for such alternation in the experimental design, by manipulating constituent durations as opposed to utterance durations alone can make an effect of syllable complexity appear.

5. Conclusion

Our findings suggest that differences in segment rate that do not correspond to differences in syllable rate have little impact on perceived speech tempo in English. We take our findings regarding our control variables to point to an interesting relationship between tempo perception and rhythm perception, which warrants further investigation.

6. Acknowledgements

This research was supported by a British Academy and Leverhulme Trust Small Research Grant. The authors would like to thank Nathan Clarke for his contribution to this study.

7. References

[1] H. Pfitzinger. "Local speech rate perception in German speech," *Proceedings of ICPhS*, 1999.
[2] S. Marin and M. Pouplier, "Temporal organization of complex onsets and codas in American English," *Motor Control*, 14, pp. 380–407, 2010.
[3] D. Byrd, "C-centers revisited". *Phonetica*, 52, pp. 285–306, 1995.
[4] S. Greenberg, et al., "Temporal properties of spontaneous speech – a syllable-centric perspective," *Journal of Phonetics*, 31, pp. 465–485, 2003.
[5] E. Vaane, "Subjective estimation of speech rate," *Phonetica*, 39, pp. 136–149, 1982.
[6] E. Den Os, "Perception of speech rate of Dutch and Italian utterances," *Phonetica*, 42, p. 124–134, 1985.
[7] V. Dellwo, et al. "The perception of intended speech rate in English, French, and German by French speakers," *Proceedings of ICSP*, 2006.
[8] H. Quené, "On the just noticeable difference for tempo in speech," *Journal of Phonetics*, 35, pp. 353–362, 2007.
[9] M. Weirich, and A.P. Simpson, "Differences in acoustic vowel space and the perception of speech tempo," *Journal of Phonetics* 43, pp. 1–10, 2014.
[10] P. Boersma, and D. Weenink, *Praat: Doing phonetics by computer*, 2017: http://www.praat.org/.
[11] V. Dellwo, et al. "Rhythmical classification of languages based on voice parameters," *Proceedings of ICPhS*, 2007.
[12] R Development Core Team, *R: A language and environment for statistical computing*, 2008: http://www.R-project.org/.
[13] C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters," *Proceedings of Interspeech*, 2010.