9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Segments, syllables 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. Listeners' sensitivity to syllable rate has been demonstrated in multiple studies in which listeners judge the rhythm or tempo of spoken utterances, although these studies do not control for segment rate. Evidence for listeners' sensitivity to segment rate is much rarer. We report two experiments aimed at clarifying the contributions of syllable and segment rate to English listeners' tempo judgements. In the first experiment, we manipulate syllable rate in utterance pairs that are constant in segment rate; in the second, we keep syllable rate constant and manipulate segment rate. Listeners decide for each pair which utterance sounds faster. Our results suggest that syllable rate differences are perceived as tempo differences even if segment rate is constant, while 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, rhythm

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 quite divergent results. English is a case in point. Its phonology allows a wide range in syllable shapes, such that one syllable can correspond to one 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 [1, 2]. Thus, increases in syllable complexity tend to mean increases in segment rate but decreases in syllable rate [3].

Surprisingly, there has been little research on the validity of using syllable and, in particular, 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. In this study, we address this general question by assessing the impact of syllable and segment rate manipulations on English listeners' impressions of speech tempo.

Several tempo perception studies report correlations of listeners' judgements and syllable rate measurements in the region of r=0.80 [4, 5]. Research into rhythm perception highlights listeners' attention to syllable rate when judging whether utterances are rhythmically alike or distinct [6, 7]. To date only one study has included segment rate measurements in the comparison [8]. 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 rate measurements. Both syllable and segment rate measurements yielded correlations in the region of r=0.80, but combining both measurements in a single equation, with syllable rate weighted higher, yielded a closer correlation with listeners' tempo judgements (r=0.91).

The findings of [8] are consistent with a model in which listeners attend to both syllable and segment rates in making tempo estimates. One hypothesis we can formulate is that for utterances that are very similar in one of the rates but different in the other, the difference must exceed a certain threshold to be consequential for listeners' tempo estimates: assuming a heavier weighting of syllable rate [8], we could expect that utterances with a similar syllable rate must be considerably different in segment rate to be perceived as different in tempo, while utterances that are very similar in segment rate must show syllable rate differences above a certain minimum to be perceived as different. In this study, we assessed whether experimental evidence for such 'consequential difference thresholds' can be found. We did this through two complementary listening experiments.

2. General method

In both experiments, we used a pairwise discrimination paradigm [9, 10]: subjects judged tempo differences in pairs of stimuli. In Experiment 1, we created a range of syllable rate differences within the pairs, while keeping segment rate constant. In Experiment 2, we did the opposite: pair members varied in segment rate on a constant syllable rate.

All stimuli were produced specifically for these experiments, and manipulated using *PSOLA* in Praat [11] as specified below. Both experiments were run at the University of Leeds with standard ethics approval, using *ExperimentMFC* in Praat. In each, listeners were introduced to the task on-screen. Utterance pairs were presented in random order, with 0.5s separating the two utterances. For each pair, listeners indicated whether the second utterance was faster, slower or the same in tempo as the first using a 7-point response scale (–3 'much slower', 0 'the same', 3 'much faster'). The next pair played 1.5s after each judgement. Listeners could replay pairs once.

Quantitative analysis of listeners' responses was done through linear mixed effects modeling, using the *lme4* and *lmerTest* packages in R [12].

10.21437/SpeechProsody.2018-57

3. Experiment 1

The aim of Experiment 1 was to assess listeners' sensitivity to syllable rate variation in a tempo judgement elicitation task in which segment rate is not a potential confound. We hypothesized that utterances that are very similar in segment rate must show syllable rate differences above a certain minimum to be perceived as different in tempo.

3.1. Materials and participants

To create the stimuli, we embedded monosyllabic nouns and verbs of varying phonological shapes (VC, CVC, CVCC, CCVCC, CCCVCC, CCCVCCC) in nine five-syllable carrier utterances, yielding a set of 24 six-syllable utterances. The utterances with a noun all started with *the* and ended in a prepositional phrase; the utterances with a verb started with *I* or *he* and ended in a noun phrase: e.g. *the* ***springs*** *on the car door* (CCCVCC), *I* ***add*** *it the wrong way* (VC), *he* ***sprints*** *to the low bough* (CCCVCCC). Because of the variation in phonological complexity in the crucial nouns and verbs, utterance segment numbers varied between 12 and 17. Pairing was done so that carrier phrases were not repeated within utterance pairs. This resulted in 120 stimulus pairs. We used similar carrier phrases to produce 120 filler pairs.

Stimuli and fillers were recorded by a female speaker of Southern Standard British English. Syllable rate in the unmanipulated recordings varied between 3.72 and 4.78 syll/sec. The durations of the utterance chunks *Word1* (*I*, *he*, *the*), *Words2-3* (*springs on, twist it* etc.) and *Words 4-6* (*the car door* etc.) were normally distributed.

PSOLA manipulations equalized F0 and amplitude, as in [10]. F0 contours were set to a stylized version of one of the natural tokens. Mean amplitude was set to 62 dB. Stimulus duration was set to yield a constant segment rate of 10 seg/sec across stimuli; this was confirmed as neither noticeably fast nor slow in a small-scale survey with four listeners. This generated syllable rate variation between 3.5 syll/sec (17 segments, 6 syllables, 1.7 sec) and 5 syll/sec (12 segments, 6 syllables, 1.2 sec). Fillers were manipulated for overall duration and amplitude only. This meant 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 5.77 syll/sec.

50 native British English listeners aged 18–35 participated in the experiment. None reported known hearing problems.

3.2. Analysis methods

Analysis focused on the relationship between listeners' responses (*Response*) and the difference in syllable rate between the two utterances in the experimental pairs. *Syllable rate ratio* was the rate of the second utterance divided by that of the utterance. The result of our selection of crucial nouns and verbs is that the smallest non-null syllable rate difference within pairs is 6%, which is around the general Just Noticeable Difference for speech tempo [9], and the largest 42%.

We also included several control variables in our analysis. First, we included control variables based on lexical frequency counts for the crucial nouns and verbs taken from the 19-billion-word web corpus EnTenTen [13], on the assumption that tempo judgements may be affected by processing speed. 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 [14], we used the *fraction of locally unvoiced frames* within Praat's *voice report* function to obtain the proportion of voiceless material for each utterance and calculated the ratio for each utterance pair (*Voicelessness ratio*). Third, since our complexity and duration manipulations leave some variation across stimuli in the durations of successive consonant and vowel intervals, we quantified commonly used rhythm metrics designed to capture this variation. These included %V, nPVI for consonants, vowels and syllables, rPVI for consonants and Varco for consonants, vowels and syllables [15, 16]. We again calculated corresponding differences (Δ) for all utterance pairs (henceforth *Δ %V*, etc.).

3.3. Results

In judging the experimental utterance pairs, listeners recorded hearing a difference between the pair members more often than they recorded hearing no difference: Figure 1 shows more non-'0' than '0' responses. Figure 1 also suggests that listeners had a small bias towards hearing the second pair member as relatively fast.

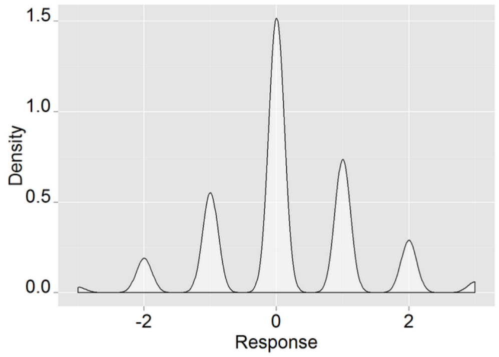

Figure 1: *Distribution of* Response.

Modelling *Response* (with significant random effects for listener and utterance identities) revealed a strong effect of *Syllable rate ratio*, in the expected direction: utterance pairs were judged to have faster second members as the syllable rate ratio of the second to the first member increased. We found additional effects of the rhythm metrics *Δ %V* and *Δ Varco-Syll*. The optimal model is in Table 1. As can be seen in Figure 2, the relationship between *Syllable rate ratio* and *Response* is a straightforward linear one (*r*=0.41): differences in *Syllable rate ratio* as small as 6% were interpreted by our listeners as differences in relative tempo.

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

Predictor	**Estimate**	**SE**	**df**	***t***	***p***
(Intercept)	–1.60	0.26	9	–6.1	<0.001
Syllable rate ratio	1.72	0.24	7	7.1	<0.001
Δ Varco-Syll	–3.62	0.67	21	–5.4	<0.001
Δ %V	0.01	0.01	8	2.4	0.04

3.4. Discussion

Our results confirm that measured syllable rate is a reliable proxy for perceived tempo. The observed correlation between listeners' responses and syllable rate differences would appear to be weaker than that observed in [4, 5] (r=0.41 vs r≈0.80), and it is possible that this is due to our control of segment rate. However, it could be due to other methodological differences between these studies too, and there is no evidence to support our hypothesis that that utterances that are very similar in segment rate must show syllable rate differences above a certain minimum to be perceived as different in tempo. Our results also suggest that temporal differences captured by commonly used rhythm metrics but not by syllable rate measures have some impact on tempo perception.

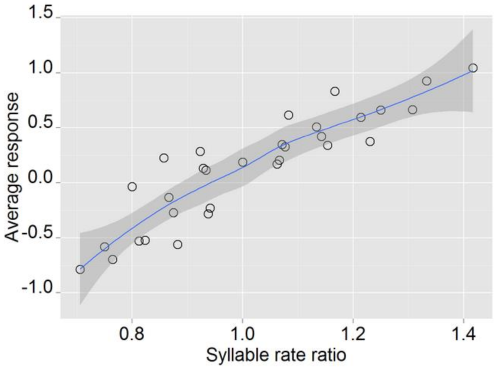

Figure 2: *Relationship between* Syllable rate ratio *and* Response *(averaged across listeners).*

4. Experiment 2

The aim of Experiment 2 was to assess listeners' sensitivity to segment rate variation in a tempo judgement elicitation task in which syllable rate is kept constant. Based on the findings of [8], we hypothesized that utterances with a similar syllable rate must be considerably different in segment rate to be perceived as different in tempo. Experiment 2 builds on the experiment reported in [17], in which listeners were insensitive to segment rate variation. [17] had a confounding factor of utterance-internal temporal variation, which Experiment 2 removes.

4.1. 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. We created two sets of 16 stimuli with all logical combinations of N1 and N2 shapes: 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. We used the same phrase to produce 134 filler pairs.

Stimuli and fillers were recorded by a female speaker of Southern Standard British English. Syllable rate varied between 3.19 and 3.87 syll/sec in the unmanipulated stimuli, and the rate distributions of the utterance chunks (*this*, N1, *or that*, N2) were normal, with a significant correlation between the number of segments in N1 and N2 and the durations of these chunks: as might be expected, more complex nouns were longer than less complex ones (e.g. *stunt* > *pack*).

PSOLA manipulation equalized F0, amplitude and utterance duration. As in Experiment 1, the F0 contour was set to a stylized version of one of the natural tokens. Mean amplitude was set to 62 dB. Stimulus duration was set to 1.25s to yield a syllable rate of 4 syll/sec; this was confirmed as neither noticeably fast nor slow in a small-scale survey with four listeners. In order to remove systematic utterance-internal temporal variation resulting from the significant correlation between N1 and N2 complexity and duration, we then set the duration of each chunk (*this*, N1, *or that*, N2) to its mean across the stimulus set. Thus, all stimuli have both the same overall syllable rate and the same syllable rate 'contour'. Moreover, one might expect this manipulation to highlight segment rate variation in N1 and N2, as more complex nouns (e.g. *stunt*) are slightly compressed relative to their natural productions, while less complex nouns (e.g. *pack*) are slightly stretched. As in Experiment 1, fillers were manipulated for overall duration and amplitude only. Durations were set differently for subsets of fillers, so that syllable rate varied between 4 and 4.75 syll/sec across the experiment.

34 native British English listeners aged 18–35 participated in the experiment. None reported known hearing problems.

4.2. Analysis methods

Analysis focused on the relationship between listeners' responses (*Response*), and the difference in segment rate between the two utterances in the experimental pairs. *Segment rate ratio* was the segment rate of the second utterance divided by that of the first. 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 [9], and the largest 31%.

As in Experiment 1, we included control variables based on lexical frequency counts for the nouns taken from the 19-billion-word web corpus EnTenTen [13], as well as *Voicelessness ratio* and difference measures for %V, nPVI and Varco for consonants and vowels, and rPVI for consonants. Syllable-based rhythm metrics are uninformative for this experiment as syllable durations are fully controlled.

4.3. Results

Listeners' responses to the experimental stimuli clustered close to zero, as seen in Figure 3: listeners perceived very little variation in tempo across the utterance pairs. Again we can see evidence of a small bias towards hearing the second pair member as relatively fast. Modelling *Response* (as in Experiment 1 with significant random effects for listener and utterance identities) revealed no significant effects of *Segment rate ratio* or any of our control variables. As seen in Figure 4, there is also no clear evidence of the more extreme *Segment rate ratio* values being associated with systematic responses in the expected direction (that is, *Segment rate ratio* values above 1 being associated with larger *Response* values than *Segment rate ratio* values below 1): this should have resulted in visible clusters of points in the bottom left and top right sections of the scatter.

4.4. Discussion

Our results offer no evidence that listeners are using segment rate to influence their tempo judgements when syllable rate is constant. In addition to finding no general correlation between *Segment rate ratio* and *Response*, we find no evidence for large differences in segment rate – that is, differences that are several multiples of the general JND for temporal patterns in speech – having an impact on listeners' tempo judgements. This replicates the finding reported in [17], even though controlling for internal temporal variation should, if anything, have highlighted segment rate variation in the crucial nouns.

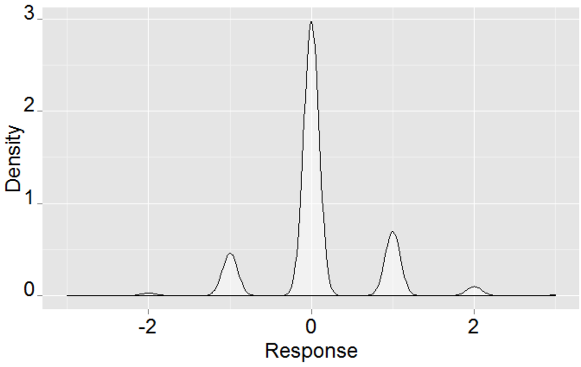

Figure 3: *Distribution of* Response*.*

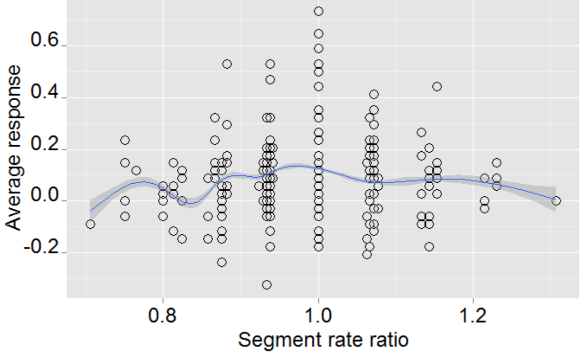

Figure 4: *Relationship between* Segment rate ratio *and* Response *(averaged across listeners).*

5. General discussion

The aim of Experiment 1 was to assess listeners' sensitivity to syllable rate variation in a tempo judgement elicitation task in which segment rate is not a potential confound. We found clear evidence to support the finding of multiple previous studies that syllable rate measurements are reliable proxies of perceived tempo, and no evidence for the existence of a 'consequential difference threshold': even the smallest differences in syllable rate that we included in our design were interpreted by listeners as tempo differences. In addition, we found evidence that our manipulations of complexity and duration affected the internal rhythm of our experimental utterance in ways that were not captured well by our syllable rate measurements, but were consequential for listeners' tempo judgements. At present, we cannot offer a comprehensive account for the observed effects of %V and syllable-based Varco metrics on listeners' tempo judgements, but we take our findings to confirm that the relationship between rhythm and tempo perception may be closer than previously assumed, and is worthy of focused investigation [6, 15].

The aim of Experiment 2 was to assess listeners' sensitivity to segment rate variation in a tempo judgement elicitation task in which syllable rate is kept constant. This experiment yielded no evidence of such sensitivity, even to large differences in segment rate. Our findings suggest that even if English listeners track segment rate in judging speech tempo, as suggested for German by [8], 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, but as pointed out in [17], this seems unlikely: our task 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.

However, it is worth considering that the absence of an effect of segment rate on listeners' tempo judgements could be due to the fact that we manipulated segment rate through varying the phonological complexity of selected syllables while keeping utterance durations constant. As suggested above, in normal speech, increases in phonological complexity are associated both with the relative compression of segmental units *and* with temporal expansion of relevant syllables. Our results suggest that when estimating speech tempo, listeners are not sensitive to the compression of segmental units in the absence of temporal expansion (which would give rise to syllable rate differences). It is possible that English listeners *expect* average segment durations – and thus segment rate – to vary with phonological complexity, and therefore ignore, for the purpose of tempo estimation, any variation that is consistent with lower segment rates for less complex syllables and higher segment rates for more complex ones.

That is, as utterance durations were constant in the experimental pairs in Experiment 2, it may have been obvious to listeners that the variation in phonological complexity they observed was associated with syllable-internal compression and expansion of segment durations only. For the future, it would seem useful to test whether the same results are obtained in an experimental design that varies segment rate on a constant syllable rate, but with variable utterance durations.

6. Conclusion

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

7. 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.

8. References

1. Marin, S. and M. Pouplier 2010. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control* **14**: 380–407.
2. Byrd, D. 1995. C-centers revisited. *Phonetica* **52**: 285–306.
3. Greenberg, S., et al. 2003. Temporal properties of spontaneous speech — A syllable-centric perspective. *Journal of Phonetics* **31**: 465–485.
4. Vaane, E. 1982. Subjective estimation of speech rate. *Phonetica* **39**: 136–149.
5. Den Os, E. 1985. Perception of speech rate of Dutch and Italian utterances. *Phonetica* **42**: 124–134.
6. White, L., S.L. Mattys, and L. Wiget 2012. Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language* **66**: 665–679.
7. Arvaniti, A. and T. Rodriquez 2013. The role of rhythm class, speaking rate, and F-0 in language discrimination. *Laboratory Phonology* **4**: 7–38.
8. Pfitzinger, H. 1999. Local speech rate perception in German speech. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco.
9. Quené, H. 2007. On the just noticeable difference for tempo in speech. *Journal of Phonetics* **35**: 353–362.
10. Weirich, M. and A.P. Simpson 2014. Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics* **43**: 1–10.
11. Boersma, P. and D. Weenink 2017. *Praat: Doing phonetics by computer*. http://www.praat.org/
12. R Core Development Team 2008. *R: A language and environment for statistical computing*. http://www.R-project.org
13. Jakubíček, M., et al. 2013 The TenTen corpus family. *Proceedings of the 7th International Corpus Linguistics Conference*, Lancaster.
14. Dellwo, V., A. Fourcin, and E. Abberton 2007. Rhythmical classification based on voice parameters. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken.
15. Arvaniti, A. 2012. The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* **40**: 351–373.
16. Rathcke, T.V. and R.H. Smith 2015. Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies. *Journal of the Acoustical Society of America* **137**: 2834–2845.
17. Plug, L. and R.H. Smith 2017. Phonological complexity, segment rate and speech tempo perception. *Proceedings of Interspeech 2017*, Stockholm.