The role of segment rate in speech tempo perception by English listeners

Leendert Plug[1], Rachel Smith[2]

[1] University of Leeds, United Kingdom

[2] University of Glasgow, United Kingdom

Corresponding author details

Dr Leendert Plug, Linguistics and Phonetics, University of Leeds, Leeds LS2 9JT, l.plug@leeds.ac.uk

Abstract

Studies in which speech tempo is quantified commonly use either syllable or segment rate as a proxy measure of tempo. Perception studies have shown that syllable rate measurements correlate closely with elicited tempo judgements across languages. Some research suggests that segment rate is an additional, independent predictor of perceived tempo — in other words, that both syllable *rate* and syllable *complexity* matter for tempo perception. However, direct empirical evidence for this is as yet lacking. This paper reports on three experiments that test the hypothesis that when segment rate is varied on a constant syllable rate, listeners estimate utterances with higher segment rates as faster. Our results provide evidence for listeners' orientation to syllable rate in estimating tempo, and evidence for listeners' additional orientation to segment rate — that is, to syllable complexity. However, the latter orientation is only observable when stimuli are variable in duration: when presented with stimuli that are identical both in syllable rate and duration, listeners do not appear to hear stimuli with more complex syllables as faster.

Keywords

speech tempo, articulation rate, perception, English

The role of segment rate in speech tempo perception by English listeners

1 Introduction

Studies of speech tempo, or studies in which speech tempo is controlled, commonly use syllable or segment rate as a proxy measure for tempo. Syllable rate, or its inverse average syllable duration, is a popular measure in a wide range of linguistic and phonetic studies (e.g. Crystal & House, 1990; Jacewicz, Fox, & Wei, 2010; Quené, 2013). Segment rate (or 'phone rate') is used frequently in corpus-based research and in phonetic studies which present measurements over very short stretches of speech (e.g. Byrd & Tan, 1996; Plug & Carter, 2014; Seifart et al., 2018).

In this paper we are concerned with the extent to which syllable and, in particular, segment rate correlate with measures of perceived tempo. This is a pertinent question especially for languages whose phonologies allow substantial syllable complexity: in these languages, the two 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 between one (V) and seven segments (CCCVCCC). Moreover, the temporal organisation of syllables is such that increases in syllable complexity are not associated with uniform increases in syllable duration. Increases in onset complexity in particular are accompanied by a relative shortening of consonants, such that the midpoint of the onset is in a stable timing relation with the midpoint of the vowel (Browman & Goldstein, 1988; Byrd, 1995; Marin & Pouplier, 2010). The result of this organisation is that more complex syllables tend to be spoken with higher segment rates but lower syllable rates relative to less complex ones. For example, in the 45-minute corpus of American English telephone speech of Greenberg, Carvey, Hitchcock, and Chang (2003), the mean duration of a stressed CVC syllable is 310ms, and that of a stressed CCVC syllable is 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%). This means that it is not difficult to find utterance pairs for which a syllable rate measure identifies one member as faster while a segment rate measure suggests the opposite. By extension, drawing reasonable conclusions about perceived tempo from comparisons of articulation rate analyses across language varieties with different typical timing patterns, or across languages with different syllable structure phonologies is far from straightforward.

The relationship between measured syllable rate and perceived tempo has been assessed in a range of languages including Dutch (Den Os, 1985; Vaane, 1982), German (Pfitzinger, 1999), French (Dellwo, Ferrange, & Pellegrino, 2006), Italian (Den Os, 1985), Polish (Gibbon, Klessa, & Bachan, 2015) and Japanese (Pfitzinger & Tamashima, 2006). All of these studies focus partly or exclusively on L1 speech perception; we consider L2 listeners' tempo perception outside the scope of this paper. These studies all point to a strong correlation between syllable rate and perceived tempo as elicited through rating or

comparison tasks: correlation coefficients between 0.7 and 0.9 are typical. (Data reported by Moore, Adams, Dagenais, and Caffee (2007) point to a similarly strong correlation between perceived tempo and a measure of words per minute in English.) Studies of rhythm perception (e.g. Arvaniti & Rodriquez, 2013; White, Mattys, & Wiget, 2012) and rhythmical entrainment (e.g. Lidji, Palmer, Peretz, & Morningstar, 2011; Schultz et al., 2016; Wilson & Wilson, 2005) provide further support for the relevance of syllables, or syllable-sized units in temporal perception generally. Wilson and Wilson (2005) suggest that the universality of the syllable as a phonological unit, the existence of fairly clear indicators of syllable beginnings, middles and ends, and the frequency range of syllable cycles all speak in favour of the idea that the syllable is a crucial timing unit in speech processing.

The relationship between measured segment rate and perceived tempo has been less widely assessed. Koreman (2006) summarizes the results of a tempo rating task with utterances sampled from a corpus of German spontaneous speech; unfortunately, while he presents mean segment rates for utterance groups, he does not directly address the correlation between segment rate measures and listeners' tempo ratings. Pfitzinger (1999) and Gibbon et al. (2015) both report that segment rate is a weaker predictor of listeners' tempo ratings than syllable rate, although the correlation coefficients they cite for segment rate are still in the range between 0.7 and 0.85. In Pfitzinger's study, 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 in a regression analysis. Pfitzinger (1999) shows that syllable rate is a marginally better predictor of the tempo rankings, and that when syllable rate and segment rate are combined in a single regression model, German listeners' tempo ratings can be approximated very closely ($r=0.91$). Mixdorff and Pfitzinger (2005) use the resulting equation, in which syllable rate is weighted considerably more heavily than segment rate, as an alternative to syllable or segment rate in an analysis of speech tempo in German map-task dialogue.

An issue with these previous studies is that the correlation between syllable rate and segment rate is not explicitly addressed. Pfitzinger's (1999) regression analysis suggests that much of the explanatory power of segment rate in predicting perceived tempo is due to its correlation with syllable rate, although segment rate still adds explanatory power once the syllable rate effect is 'partialled out'. We can hypothesize on this basis that while listeners get most of their impression of speech tempo from the durations of intervals between syllable nuclei, they may judge intervals with similar durations differently when they differ in the number of segments they contain. To our knowledge, the available empirical evidence does not allow us to confirm or refute this hypothesis. In this study, we address it through experiments in which listeners judge the relative tempo of phrases in which syllable rate is kept constant and segment rate is varied systematically.

Several previous studies have included experiments in which listeners judge the tempo of utterances in which syllable rate is kept constant and some other parameter is systematically varied. These studies have shown that higher and more dynamic pitch and intensity contours make utterances sound faster (Cumming, 2011; Feldstein & Bond, 1981; Kohler, 1986), and more peripheral vowel qualities make utterances sound faster (Weirich &

Simpson, 2014). As Weirich and Simpson (2014) indicate, these results warrant the generalisation that listeners estimate complex spectral events as faster than less complex events that take the same amount of time to complete — because 'if more (in terms of spectral information) happens, the listener reasons it has to take longer compared to a less complex event' (p.2). Given this, we can formulate the hypothesis that when segment rate is varied on a constant syllable rate, listeners will estimate utterances with higher segment rates as faster. It may be that the variation in segment rate needs to be considerable to have an impact on listeners' tempo judgements: if listeners get most of their impression of speech tempo from the durations of intervals between syllable nuclei, a stable syllable rate will prompt an impression of no difference in tempo. We therefore further hypothesise that we can identify a 'consequential difference threshold' for segment rate variation which may be substantially higher than the general Just Noticeable Difference (JND) for temporal variation, estimated at around 5% by Quené (2007). In this study, we assess the perceptual relevance of multiple degrees of segment rate variation among stimuli that do not vary in syllable rate.

2 Experiment 1

2.1 Aims

In Experiment 1, we wanted to assess listeners' sensitivity to segment rate variation in a tempo judgement elicitation task in which syllable rate is not a potential confound. We did this using a pairwise discrimination paradigm (Quené, 2007; Weirich & Simpson, 2014): subjects were asked to judge tempo differences in pairs of phrases in which we varied segment rates but kept syllable rates constant.

2.2 Method

2.2.1 Participants

The experiment was run at the University of Leeds in accordance with all institutional ethics regulations. 50 monolingual British English listeners (34 female) between the ages of 18 and 35 (mean=22) participated.[1] None reported known hearing problems. All were paid a small fee for their time.

2.2.2 Phrase design and recording

To create our phrases, we embedded two nouns of varying phonological shapes in the carrier phrase *this* ***Noun$_1$*** *or that* ***Noun$_2$***. First, we created two sets of 16 *experimental* phrases which each included all possible configurations of two monosyllabic nouns with the phonological shapes CVC, CCVC, CVCC and/or CCVCC. The 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

[1] In this and subsequent experiments we asked participants to indicate what accent of English they deemed themselves to speak. Unfortunately their responses varied too widely in specificity for us to have confidence that we could derive an informative control variable.

voiceless obstruents in initial and final position and allowing only short vowels in the nucleus: e.g. *kit*, *pack* (CVC), *clock*, *spot* (CCVC), *tact*, *cost* (CVCC), *prank*, *stunt* (CCVCC).[2] Embedding the nouns in two positions in the (7-segment) carrier gave us a range of segment numbers across the phrases, from 13 (*this* $N_{1(CVC)}$ *or that* $N_{2(CVC)}$) to 17 (*this* $N_{1(CCVCC)}$ *or that* $N_{2(CCVCC)}$). Each noun occurred in only one position (N_1 or N_2) and in only one set. We tried as much as possible to construct phrases with semantically compatible nouns that do not share onsets or vowel nuclei — *this kit or that pack*, *this trust or that stock*, *this pump or that plank*, *this prank or that stunt* — in order to minimise variation in information density and spectral complexity across phrases, which may have an impact on tempo perception (Bosker & Reinisch, 2017; Weirich & Simpson, 2014). We took lexical frequency counts for the nouns from SUBTLEX-UK, a corpus of 201.3 million words from the subtitles of 45,099 BBC broadcasts (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The two sets of phrases, which we used to create 256 experimental phrase pairs, are given in Appendix A.

Second, we created two sets of 20 *distractor* phrases which each included one monosyllabic noun and one disyllabic or trisyllabic noun and had no restrictions on segmental make-up. We varied the position of the multi-syllabic noun to create some variation in phrase-internal rhythm. Again, we tried to combine nouns with some semantic connection: for example *this kestrel or that kite*, *this bean or that potato*, *this adventure or that tour*. The two sets of phrases, which we used to create 116 *distractor* phrase pairs, are given in Appendix A.

All phrases were produced by the second author, who is a female speaker of Southern Standard British English. (As the second author is based in Glasgow and the experiment was run in Leeds, the likelihood of any of our participants being familiar with the speaker's voice was minimal.) The recordings were made in a sound-proof studio at the University of Glasgow using a cardioid condenser microphone (Sennheiser MKH40), a pre-amplifier (Rolls MX34c LiveMix) and the recording software Audacity running on a Windows PC. The recordings were produced in mono at a sampling rate of 44100 Hz with 16-bit resolution. The speaker produced the phrases at what felt like a comfortable pace. In order to minimise rhythmical and prosodic variation, the speaker recorded one phrase as a model, and listened to it before producing each of the remaining phrases. The speaker also took care to be consistent in segmental realisations, in particular that of /t/ in *that*: this was glottalised and without audible release [$\widetilde{\text{ʔt}}$] on most productions. The speaker recorded the entire series of phrases twice, and corrected any disfluent tokens immediately as they occurred.

2.2.3 Pre-manipulation analysis

Using *Praat* (Boersma & Weenink, 2017), we segmented all experimental phrase productions into the constituents *this*, N_1, *or that* and N_2, using standard criteria (Turk, Nakai, & Sugahara, 2006). The segmentation is illustrated in Figure 1. Note that the speaker realised /ð/ in *this* with a complete voiceless closure on all productions; we therefore consistently placed the

[2] Given our speaker's accent, we can follow Roach (2004) and Collins and Mees (2013) in transcribing the vowels as /ɪ/ (e.g. *kit*), /e/ (e.g. *tent*), /æ/ (e.g. *pack*), /ʌ/ (e.g. *stunt*) and /ɒ/ (e.g. *clock*).

start boundary for /ð/ at the burst, as seen in Figure 1. We extracted durations and f0 means for each of these constituents (pitch settings: time step 0.01, pitch floor 95 Hz, pitch ceiling 250 Hz). Inspection of the corresponding distributions allowed us to select the production of each phrase (out of the two recorded productions) whose duration and f0 values were closest to the corresponding grand means. In the resulting set of experimental phrases, the durations of *this*, N_1, *or that* and N_2 were all normally distributed (Shapiro-Wilks tests: *this* W=0.955, p=0.203; N_1 W=0.986, p=0.940; *or that* W=0.958, p=0.243; N_2 W=0.972, p=0.551) without outliers. The speaker's articulation rate averaged 3.6 syllables per second (range 3.2–3.9) and 10.7 segments per second (range 8.9–12.0) across the phrases.[3] There was evidence of phrase-final lengthening: N_2 durations were between 20 and 212 ms longer than N_1 durations (cf. Turk & Shattuck-Hufnagel, 2007; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). Noun duration was weakly correlated with phonological complexity (N_1: Pearson's r=0.32, p=0.068; N_2: r=0.34, p=0.055).

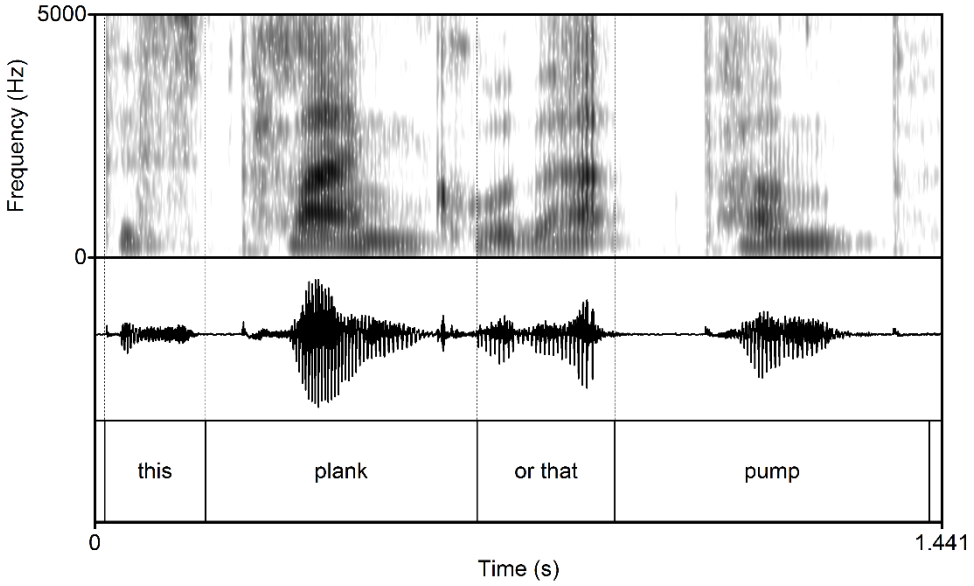


Figure 1. Segmented experimental phrase before manipulation

2.2.4 Manipulation and pairing

We manipulated the experimental phrases to equalize their durations, f0 contours and mean intensities, using PSOLA in *Praat* (Boersma & Weenink, 2017). We controlled f0 contours and mean intensities because both parameters have been shown to inform listeners' tempo judgements (Cumming, 2011; Feldstein & Bond, 1981; Kohler, 1986). We set all phrases' durations to 1.25s to yield a constant syllable rate of 4 sylls/s; as the speaker's natural syllable rate varied between 3.2 and 3.9 sylls/s, this entailed temporal compression in all cases. We chose this syllable rate on the basis of a survey with four listeners (who did not participate in the main experiment). They were presented with 16 phrase productions, each

[3] We will henceforth abbreviate 'syllables per second' to 'sylls/s' and 'segments per second' to 'segs/s'.

manipulated to yield one of eight syllable rates between 3 and 5 sylls/s (0.20 steps). They were asked to rate phrase tempo on a 5-point Likert scale ('very slow', 'quite slow', 'normal', 'quite fast', 'very fast'). The rates 4, 4.2 and 4.4 sylls/s were near-unanimously rated as 'normal'; among these, the rate 4 represents the smallest divergence from the original productions. Along with the temporal compression, we stylised the f0 contour of one of the phrase productions and resynthesized all phrases with it: f0 rose from 195 to 200 Hz during *this*, fell from 215 to 130 Hz during N_1, rose during *or that* to reach 185 Hz at the start of N_2, falling to 115 Hz at the end of N_2. We set mean intensity at 62 dB. We recalculated phrase-internal constituent durations so we could take these into consideration in the response analysis; as illustrated in Figure 2, in this experiment individual constituent durations varied as they did in the original productions, within a fixed phrase duration of 1.25s.

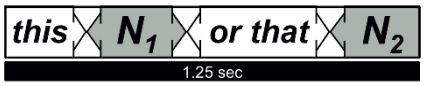


Figure 2. Temporal organisation of experimental phrases; the shading of the noun blocks reflects that our complexity manipulation affected the nouns, while the arrow heads reflect that constituent durations varied between individual phrases, within a fixed phrase duration.

We used the manipulated experimental phrases to create 256 phrase pairs. We did this by pairing exhaustively across the two sets of phrases we created, i.e. every phrase from the first set (set 1 in Appendix A) was paired with every phrase from the second set (Set 2 in Appendix A), with 0.5s silence separating the pair members (see Figure 3). Each phrase pair featured in one order only.[4] Within each phrase pair, phrase durations and syllable rates were equal. Segment rates, however, varied as a result of our phrase design and manipulations. In absolute terms, the difference in segment rate between the first and second member of a pair ranged from –3.2 segs/s to 3.2 segs/s. In relative terms, the *proportional* segment rate of the second pair member relative to the first ranged in 17 steps between 0.76 and 1.31. The smallest divergence from equality was ±0.06, which is just above the JND for speech tempo (around 5%) according to Quené (2007). Therefore, we could expect that the smallest divergences from segment rate equality should be *noticeable* to our listeners, and therefore at least potentially consequential for their tempo judgements.

[4] Since the set membership of each individual phrase (Set 1 or Set 2, see Appendix A) was a matter of random assignment, we did not deem it necessary to counterbalance phrase orders within pairs.

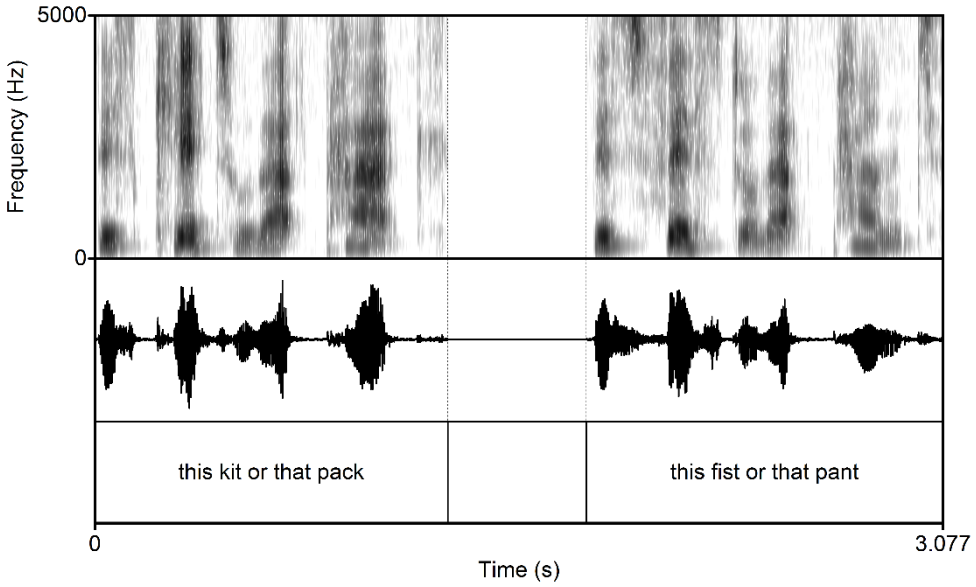


Figure 3. Experimental phrase pair (*this kit or that pack* 1.25s, 10.4 segs/s; silence 0.5s; *this fist or that pant* 1.25s, 12 segs/s)

For the distractor phrases, we manipulated duration and mean intensity only. We retained their original f0 contours to ensure that across the experiment, listeners heard a certain amount of natural pitch variation. We set phrase durations differently for subsets of distractor phrases, so that across the phrase set, articulation rate varied between 4 sylls/s and 4.75 sylls/s (and between 9.3 segs/s and 12.9 segs/s). We used the manipulated phrases to create 116 phrase pairs.

Before proceeding with the experiment, we played all phrase pairs to two colleagues with experience of running listening experiments, but no specific knowledge of our manipulations. They listened to the pairs in random order and deemed their productions to be sufficiently natural for a tempo judgement task.

2.2.5 Procedure

The experimental procedure was similar to that described by Weirich and Simpson (2014). We used the *ExperimentMFC* facility in *Praat* (Boersma & Weenink, 2017) to run the experiment. Participants were introduced to the task on-screen. For each pair of phrases, participants were asked to indicate whether the second phrase was faster, slower or the same in tempo as the first phrase using a 7-point response scale ranging from –3 ('much slower') through 0 ('same') to 3 ('much faster'). The next pair played 1.5s after each judgement was recorded. Participants could replay each pair once.[5] Participants did a practice run consisting of five phrase pairs, followed by the main experiment. Experimental and distractor phrase pairs were presented in random order by participant.

[5] *Praat* did not record usage of this function, so we could not enter it into our quantitative analysis.

2.2.6 Analysis method

We treated participants' judgements as an ordinal variable and assessed the impact of our manipulations and control variables through fitting cumulative link mixed models (Orme & Combs-Orme, 2009) using the *ordinal* package (Christensen, 2018) in R (R Core Team, 2016). We should note that in line with much previous research, we had initially treated judgements as a numerical variable and fitted linear mixed effects models using *lme4* (Bates, Maechler, Bolker, & Walker, 2015). Among other things, this assumes that judgement categories are equidistant and responses are normally distributed — neither of which is necessarily the case (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018). As it happens, in the case of our data the two modelling methods led to optimal models with the same fixed effects. In what follows we report on the more statistically appropriate modelling method using cumulative link mixed models. (We still treat *Response* as a numerical variable for illustration purposes, as mean *Response* values provide a more intuitive visual summary than *Response* value proportions.) Our starting point in modelling was a base model with random intercepts for the listener's identity (*Participant*) and phrase identity ($Phrase_1$, $Phrase_2$); the latter is appropriate as individual phrases were used in multiple pairs, but pairs were not repeated in the procedure above. We then followed a stepwise procedure to arrive at an optimal model; we describe this as we present the results below.[6] We conducted log-likelihood tests using the *anova* function for pairwise model fit comparison in this procedure.

The variables we entered into the analysis are listed in Table 1. We should note that as our stimuli are phrase pairs, we can in principle quantify any continuous parameters in a number of ways. Since we were primarily interested in the effect of the temporal relationship between the two phrases in each pair on listeners' perceptions of tempo, and the perceptions we elicited were themselves relative, we opted for using only relative implementations of our independent variables — that is, measures yielding one value per phrase pair which quantifies the relationship between the two pair members on the relevant parameter.

For segment rate we used the proportional measure cited above (*Proportional segment rate*), which allows us to relate differences within phrase pairs straightforwardly to the reported JND of about 5%.[7] For lexical frequency we subtracted the mean (log) frequency across the two nouns in the first phrase pair member from the mean (log) frequency across the two nouns in the second. Higher values on the resulting measure (*Lexical frequency difference*) represent phrase pairs in which the second phrase contains higher-frequency nouns than the first phrase. Finally, we calculated duration ratio measures by constituent: for each phrase pair, we divided the durations of *this*, N_1, *or that* and N_2 in the second phrase by the corresponding constituent durations in the first. These measures are only weakly correlated with *Proportional segment rate* (Pearson's $r<|0.5|$) since we manipulated

[6] The random effects estimates of the optimal model reported below are: *Participant* variance=0.400, SD=0.632; $Phrase_1$ variance=0.070, SD=0.264; $Phrase_2$ variance=1.149, SD=1.386.

[7] We should note that as we varied consonant numbers in the nouns only, segment rate by phrase is quantitatively equivalent to a numerical measure of syllable complexity across the two nouns. We did not pursue an alternative analysis with binary coding for complexity in each noun's onset and coda, as the resulting proliferation of categorical variable levels would make output models hard to interpret.

complexity in two within-phrase locations and equalized only phrase durations, not individual constituent ones (see Figure 2 above).

random variables	*Participant, Phrase$_1$, Phrase$_2$*	
control variables	*Proportional 'this' duration* *Proportional N_1 duration* *Proportional 'or that' duration* *Proportional N_2 duration* *Lexical frequency difference*	$dur(this_{Phrase2})/dur(this_{Phrase1})$ $dur(N_{1Phrase2})/dur(N_{1Phrase1})$ $dur(or\_that_{Phrase2})/dur(or\_that_{Phrase1})$ $dur(N_{2Phrase2})/dur(N_{2Phrase1})$ $\overline{LogFreq}_{N1+N2}^{Phrase2} - \overline{LogFreq}_{N1+N2}^{Phrase1}$
crucial predictor	*Proportional segment rate*	$segrate_{Phrase2}/segrate_{Phrase1}$

Table 1. Experiment 1 analysis variables

2.3 Results

Figure 4 shows the distribution of *Response* for phrase pairs with a segment rate difference and phrase pairs without. It is clear that listeners heard no difference in tempo in most phrase pairs, whether they had a measurable difference in segment rate (6026 zero responses to 9300 stimuli, or 65%) or not (2269 zero responses to 3500 stimuli, or 65%). The slight asymmetry in the distributions suggests listeners may have been moderately biased towards hearing the second phrase pair member as faster (cf. Lehiste, 1976). Zooming in on particular subranges of segment rate difference does not reveal any further patterns: Table 2 shows that listeners consistently heard no tempo difference in around two thirds of phrase pairs, whether these pairs had measurable segment rate differences exceeding 10%, measurable segment rate differences between 6% and 10%, or no measurable difference at all. In other words, we see no obvious evidence here that listeners responded to our manipulations — at least not as reflected in measures of segment rate difference.[8]

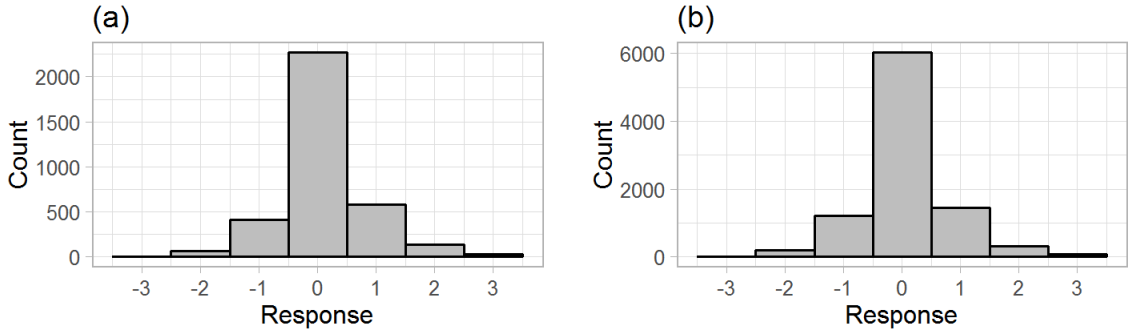


Figure 4. Distribution of *Response* for (a) phrase pairs with no segment rate difference (*Proportional segment rate* = 1) and (b) phrase pairs with a segment rate difference (*Proportional segment rate* ≠ 1)

Proportional segment rate	<0.9	>0.9, <1	1	>1, <1.1	>1.1
N responses	1850	2800	6026	2800	1850

[8] We can also report that response latencies were not significantly correlated with *Proportional segment rate* (r=0.01, p=0.14) and female and male participants produced near-identical *Response* distributions. In modelling, we checked whether *Gender* improved the fit of our base model; it did not.

Response = 0	1174 (63%)	1821 (65%)	9300 (65%)	1857 (66%)	1174 (63%)

Table 2. Numbers and proportions of zero ('no difference') responses by *Proportional segment rate* subrange

In modelling *Response*, we first assessed the relevance of our control variables *Lexical frequency difference*, *Proportional 'this' duration*, *Proportional N_1 duration*, *Proportional 'or that' duration* and *Proportional N_2 duration* by adding them to a base model with random intercepts for *Participant*, *Phrase$_1$* and *Phrase$_2$*. We kept any significant predictors, and then assessed whether *Proportional segment rate* further improved model fit. This revealed a significant fixed effect of *Proportional N_2 duration* only (est=3.665, se=1.249, z=2.933, p=0.003). *Proportional segment rate* did not improve model fit further, whether added as an independent predictor or as an interaction term alongside *Proportional N_2 duration*. It did not yield a significant effect in a model without *Proportional N_2 duration* either (est=–0.800, se=0.918, z=–0.872, p=0.383).

The effect of *Proportional N2 duration* can be seen in Figure 5, in which we have treated *Response* as a numerical variable for convenience of illustration, and calculated a mean across listeners corresponding to every value of *Proportional N2 duration*. As the duration of the second noun in the second phrase went up, listeners' greater tendency to hear the second phrase as faster is reflected in a significantly positive correlation (r=0.46, p<0.001).

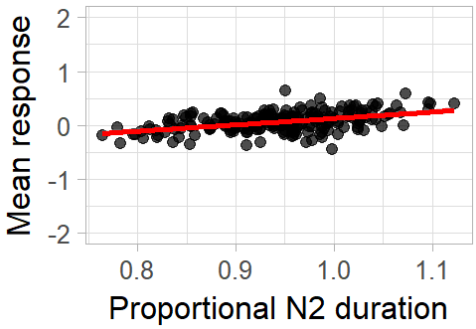


Figure 5. Relationship between *Proportional N_2 duration* (*x*-axis) and *Mean response* (*y*-axis), with linear fit line; each data point is the mean *Response* value for the corresponding *x*-axis value.

While we do not present detailed analysis of listeners' responses to the distractor phrase pairs, we can confirm that these do show clear effects of both syllable rate and segment rate ratios (whose inter-correlation we did not control, unlike in the experimental phrase pairs). For comparability with previous studies, we treat *Response* as a numerical variable here. Averaging it for each unique *Proportional syllable rate* value yields a correlation between *Mean response* and *Proportional syllable rate* of (Pearson's) r=0.88, p<0.001. Averaging by *Proportional segment rate* yields a very similar correlation between *Mean response* and *Proportional segment rate* (r=0.85, p<0.001). These correlations seem in line with those reported by Vaane (1982), Den Os (1985), Pfitzinger (1999) and Gibbon et al.

(2015) — and notably, they show that it is *not* the case that our listeners were insensitive to rate variation throughout the experiment: they were insensitive to segment rate variation in the experimental phrase pairs only.

2.4 Discussion

Experiment 1 showed that listeners' tempo judgements were largely insensitive to the segment rate differences that arose from our manipulation of phonological complexity. Listeners consistently heard no difference in tempo in about two-thirds of phrase pairs, whether these pairs had measurable differences in segment rate or not, and *proportional segment rate* did not feature in the optimal model of *Response*. We did find an effect of the relative duration of the second noun: the longer it was, the faster the phrase was perceived to be. As indicated above, the positive direction of this effect may appear counter-intuitive: one might expect that a relatively *fast* production of the final noun should contribute to listeners' sense that the phrase is relatively fast. Note, however, that as we kept phrase duration constant, the duration of the final noun is inversely proportional to the duration of the remainder of the phrase, and thus directly proportional to the syllable rate in that remainder: Utterances with longer final nouns have a higher syllable rate across the rest of the phrase. Therefore, the effect of *Proportional N_2 duration* could reflect sensitivity to the syllable rate of the second phrase up to the final noun: the long duration of the noun itself might be assumed by listeners to result from final lengthening, rather than reflecting phrase tempo.

More generally, listeners' sensitivity to the duration of the final noun highlights that our design did not control within-phrase temporal variation. The utterances that the speaker originally recorded will have varied in their internal temporal make-up: some phrases will have had relatively long nouns and short function words; others relatively short nouns and long function words; others perhaps just one lengthened word. Since we equalised the overall utterance duration, but not the durations of individual words, this utterance-internal temporal variation was preserved in the manipulated stimuli: see Figure 2 above. In consequence the stimulus set will have contained uncontrolled variation in *local* (word-by-word) syllable rate and segment rate (even though overall syllable rate was constant and overall segment rate followed pre-specified patterns). This may have attenuated the effect of our complexity manipulations. Therefore, we conducted Experiment 2, which controlled constituent durations as well as overall utterance duration.

3 Experiment 2

3.1 Aims

As suggested above, the variation in constituent durations in the Experiment 1 stimuli — which translates to variation in syllable rate 'contour' across phrases — constitutes a potential confound for any effect that segment rate might have on perceived tempo. We therefore reran the experiment, this time controlling individual constituent durations in the

experimental phrases, with the aim of establishing whether this additional control might give rise to a detectable effect of segment rate. We hypothesized that participants would perceive more differences in segment rate than in Experiment 1, as fixing the constituent durations meant concentrating all observable segment variation in the two nouns in each phrase.

3.2 Method

This experiment was an exact replication of Experiment 1 apart from the additional manipulation of the experimental items described below, and the entailed redundancy of control variables derived from constituent durations in the quantitative analysis. Distractor phrase pairs and procedure were identical to those for Experiment 1.

3.2.1 Participants

Like Experiment 1, the experiment was run at the University of Leeds in accordance with all institutional ethics regulations. 34 monolingual British English listeners (25 female) between the ages of 18 and 35 (mean=26) participated. None reported known hearing problems. All were paid a small fee for their time. Six participants had also participated in Experiment 1. The experimental sessions for Experiments 1 and 2 were separated by more than two months.

3.2.2 Additional temporal manipulation

In order to remove all syllable rate variation between experimental phrases, we calculated the mean duration of each phrase constituent (*this*, N_1, *or that* and N_2) across all experimental phrases after manipulation for Experiment 1 — that is, with the overall phrase durations set at 1.25s. We then set the constituent durations in all phrases to these means. Table 3 lists the range of durations observed for each constituent before this additional manipulation, and the mean that became the stable constituent duration as a result of it. In the resulting phrase set, all phrases have the same internal temporal organization apart from the variation in noun complexity, as illustrated in Figure 6. While in the Experiment 1 phrases, segment rate differences were distributed across the phrases, in Experiment 2 they were concentrated in the nouns only: *this* and *or that* had the same duration across phrases, all N_1 nouns had the same duration whether their phonological shape was CVC, CCVC, CVCC or CCVCC, and likewise all N_2s had the same duration regardless of phonological shape. Consistent with the original productions and with phrase-final lengthening, N_2 duration was greater than N_1 duration.

	this	N_1	*or that*	N_2
duration range (ms)	138–188	346–435	201–271	410–536
duration mean (ms)	153	389	224	484

Table 3. Duration ranges and means for phrase constituents in Experiment 1, the latter of which were the fixed constituent durations in Experiment 2

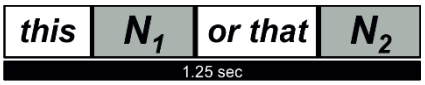


Figure 6. Temporal organisation of experimental phrases; the shading of the noun blocks reflects that our complexity manipulation affected the nouns, while the use of blocks for all constituents reflects that constituent durations were fixed.

As in Experiment 1, we played all phrase pairs to two colleagues with experience of running listening experiments, but no specific knowledge of our manipulations. They listened to the pairs in random order and deemed their productions to be sufficiently natural for a tempo judgement task, despite the inevitable divergences from the original temporal make-up of the phrases introduced by our additional manipulation.

3.2.3 Analysis method

As indicated above, we applied the same analysis method as we did in Experiment 1, except that there was no need to enter control variables derived from constituent durations. Table 4 lists the analysis variables.[9]

random variables	*Participant*, *Phrase*$_1$, *Phrase*$_2$
control variable	*Lexical frequency difference*
crucial predictor	*Proportional segment rate*

Table 4. Experiment 2 analysis variables

3.3 Results

Figure 7 shows the distribution of *Response* for phrase pairs with a segment rate difference and phrase pairs without. It is clear that as in Experiment 1, listeners heard no difference in tempo in most phrase pairs, whether they had a measurable difference in segment rate (4415 zero responses to 6324 stimuli, or 70%) or not (1647 zero responses to 2380 stimuli, or 69%). Again, the asymmetry in the distributions suggests listeners may have been biased towards hearing the second phrase pair member as faster (cf. Lehiste, 1976). Zooming in on particular subranges of segment rate difference does not reveal any further patterns: Table 5 shows that listeners consistently heard no tempo difference in just over two thirds of phrase pairs, whether these pairs had measurable segment rate differences exceeding 10%, measurable segment rate differences between 6% and 10%, or no measurable difference at all. As in Experiment 1, therefore, we see no obvious evidence here that listeners responded to our manipulations of phonological complexity and segment rate difference.[10]

[9] The random effects estimates of the optimal model reported below are: *Participant* variance=0.340, SD=0.584; *Phrase*$_1$ variance=0.035, SD=0.187; *Phrase*$_2$ variance=1.125, SD=1.354.

[10] We can also report that response latencies were not significantly correlated with *Proportional segment rate* (r=0.02, p=0.029) and female and male participants produced near-identical *Response* distributions. In modelling, we checked whether *Gender* improved the fit of our base model; it did not.

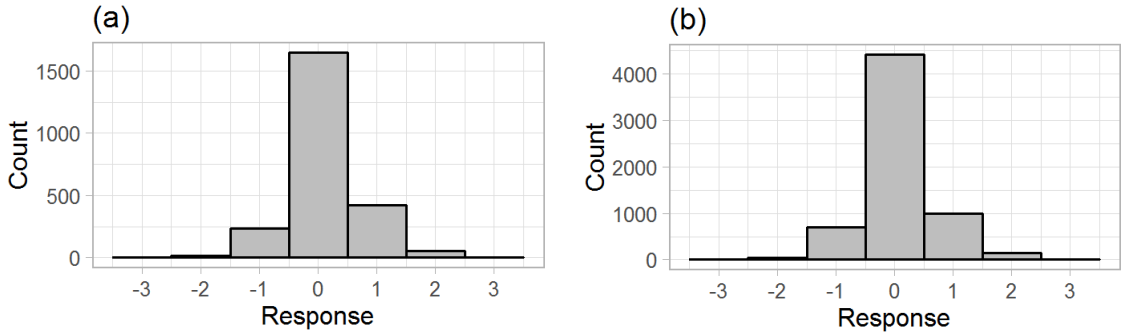


Figure 7. Distribution of *Response* for (a) phrase pairs with no segment rate difference (*Proportional segment rate* = 1) and (b) phrase pairs with a segment rate difference (*Proportional segment rate* ≠ 1)

Proportional segment rate	<0.9	>0.9, <1	1	>1, <1.1	>1.1
N responses	1258	1904	2380	1904	1258
Response = 0	868 (69%)	1344 (71%)	1647 (69%)	1349 (71%)	854 (68%)

Table 5. Numbers and proportions of zero ('no difference') responses by *Proportional segment rate* subrange

Modelling with cumulative link mixed models revealed no significant effect for *Lexical frequency difference*, as in Experiment 1, and no significant effect for our crucial predictor *Proportional segment rate* (addition to model with random effects only: est=–0.222, se=0.672, z=–0.33, p=0.742). Inspection of participants' responses to the distractor pairs again confirms that it is *not* the case that they were insensitive to rate variation throughout the experiment: the correlation between *Mean response* (averaged by *Proportional syllable rate*) and *Proportional syllable rate* is r=0.90, p<0.001 (cf. 0.88 in Experiment 1), and the correlation between *Mean response* (averaged by *Proportional segment rate*) and *Proportional segment rate* is r=0.87, p<0.001 (cf. 0.85 in Experiment 1).

3.4 Discussion

The results of Experiment 2 are very similar to those of Experiment 1. Our additional control of constituent durations removed related variables from the analysis. We hypothesized that differences in segment rate would be more noticeable than in Experiment 1, as fixing the constituent durations meant concentrating all observable segment variation in the two nouns in each phrase. Again, however, we found no evidence that complexity differences in phrase pairs with more extreme *Proportional segment rate* values gave rise to perceived differences in tempo. Overall, participants heard no tempo difference in a greater proportion of phrase pairs than in Experiment 1: even when segment rate differences were well above 10%, listeners heard no tempo difference in around 70% of phrase pairs.

One potential weakness of our design of Experiment 1, and by extension Experiment 2, is that we fixed syllable rate by fixing phrase durations across the board. While listeners were exposed to phrases of variable duration in the distractor phrase pairs, in the experimental phrase pairs both pair members invariably had the same duration. In principle, syllable rate and phrase duration are independent temporal parameters, and in principle, both may independently affect listeners' tempo perception. If so, equalizing *both* in our experimental design may have biased listeners towards a 'no difference' response more strongly than an equalization of syllable rate alone. In Experiment 3, therefore, we fixed syllable rate while allowing phrase durations to vary within and across pairs.[11]

4 Experiment 3

4.1 Aims

As indicated above, we wondered whether the substantial proportion of 'no difference' responses in Experiments 1 and 2 might be due to the fact that the way we chose to equalize the syllable rates of our experimental phrases was by equalizing their durations. It is of course possible to create phrases that are equal in syllable rate but not in duration, by varying the number of syllables within phrases. Our aim in Experiment 3 was to assess whether listeners would show sensitivity to our phonological complexity manipulations if we kept syllable rate constant *without* keeping phrase durations constant. We hypothesized that varying phrase durations and syllable numbers would bring down the overall proportion of 'no difference' responses. Our question was whether this would reveal an effect on perceived tempo of segment rate variation due to phonological complexity, which we failed to find evidence for in Experiments 1 and 2.

4.2 Method

4.2.1 Participants

Experiment 3 was run at the University of Leeds in accordance with all institutional ethics regulations. It was run as part of a longer experimental session, along with another listening experiment that is not reported here. 45 monolingual British English listeners (34 female) between the ages of 18 and 35 (mean=23) participated. None reported known hearing problems. All were paid a small fee for their time. Four participants had participated in Experiment 1, Experiment 2 or both. The experimental sessions for Experiments 2 and 3 were separated by more than six months.

[11] An anonymous reviewer also queried whether listeners might be more sensitive to our manipulations if the experimental phrase pairs had been presented without filler pairs, as tempo differences were relatively salient in the latter. We therefore ran a part-replication of Experiment 2. We included phrases with the noun shapes $N_{1(CVC)}$~$N_{2(CVC)}$, $N_{1(CCVCC)}$~$N_{2(CCVCC)}$) and $N_{1(CCVC)}$~ $N_{2(CCVC)}$ to make 81 experimental pairs with the same overall range of *Proportional segment rate* values as in Experiment 2. These were presented to 14 listeners who had not participated in Experiment 2. The experiment was conducted online using *SoSci Survey* (https://www.soscisurvey.de/). Listeners were given the same response options as in Experiment 2. Applying the quantitative analysis methods described above revealed no effect of *Proportional segment rate*.

4.2.2 Phrase design and recording

To keep the stimuli for Experiment 3 similar to those used in Experiments 1 and 2, we again embedded nouns of varying phonological shapes in the (five-syllable) carrier phrase *this* ***Noun₁*** *or that* ***Noun₂***; however, we added two additional carrier phrases: the four-syllable variant *this* ***Noun₁*** *or* ***Noun₂*** and the six-syllable variant *this* ***Noun₁*** *or that* ***Noun₂*** *then*. In what follows, we will refer to these as carrier phrases A to C, as in Table 6.

	Carrier phrase	N sylls
A	*this* ***Noun₁*** *or* ***Noun₂***	4
B	*this* ***Noun₁*** *or that* ***Noun₂***	5
C	*this* ***Noun₁*** *or that* ***Noun₂*** *then*	6

Table 6. Experiment 3 phrase design

In each, we embedded nouns also used in Experiments 1 and 2 to create five degrees of complexity: $N_{1(CVC)}$~$N_{2(CVC)}$, $N_{1(CVC)}$~$N_{2(CCVC)}$, $N_{1(CCVC)}$~$N_{2(CCVC)}$, $N_{1(CCVC)}$~$N_{2(CCVCC)}$ and $N_{1(CCVCC)}$~$N_{2(CCVCC)}$. We used the same nouns in the three carrier phrases: for example, for $N_{1(CVC)}$~$N_{2(CVC)}$ we created *this kit or pack*, *this kit or that pack* and *this kit or that pack then*. As in Experiments 1 and 2, we created two separate sets of 15 phrases (in each, 5 degrees of complexity times 3 carrier phrases). We paired phrases from the two sets exhaustively to create 225 *experimental* phrase pairs, although after piloting we ran the experiment with a subset of 81. We explain this below. The complete sets of phrases are given in Appendix B.

To create *distractor* phrase pairs, we embedded nouns used in the Experiment 1 and 2 distractor phrase pairs in the three carrier phrases to create two sets of 20 phrases. As in Experiments 1 and 2, each distractor phrase contained one monosyllabic noun and one di- or trisyllabic one with some semantic connection: for example *this church or chapel*, *this disco or dance*, *this cause or that effect*, *this bereavement or that loss*, *this shower or that bath then*, *this bike or that scooter then*. We created 69 phrase pairs using these phrases, which are given in Appendix B.

All phrases were produced by the second author, and recordings were made in a sound-proof studio at the University of Glasgow with the same equipment as for Experiment 1. Again, the speaker produced the phrases at a comfortable pace, correcting any disfluencies immediately as they occurred. The speaker took care to be consistent in prosodic and segmental realisations, and recorded the entire series of phrases twice.

4.2.3 Pre-manipulation analysis

Using *Praat* (Boersma & Weenink, 2017), we segmented all experimental phrase productions into their main constituents: *this*, N_1, *or* and N_2 for phrases with carrier A; *this*, N_1, *or that* and N_2 for phrases with carrier B; and *this*, N_1, *or that*, N_2 and *then* for phrases with carrier C. We extracted durations and f0 values for each of these constituents (pitch settings: time step 0.1, pitch floor 95 Hz, pitch ceiling 250 Hz). The segmentation procedure was the same as that in Experiment 1. Again, the speaker realised /ð/ in *this* with a complete voiceless closure on all

productions; we therefore consistently placed the start boundary for /ð/ at the burst. Inspection of the corresponding distributions allowed us to narrow the productions down to one per phrase, having identified the member of each phrase production pair whose duration and f0 values were closest to the corresponding grand means.

In the resulting set of experimental phrases, constituent durations were mostly normally distributed. For *this* (W=0.955, p=0.228), N_1 (W=0.974, p=0.651) and N_2 (W=0.929, p=0.047) we could assess this across all phrases. For *or*, we assessed it for phrases with carrier A only (W=0.934, p=0.487); for *or that* we assessed it across phrases with carriers B and C (W=0.959, p=0.5311); and for *then* (W=0.907, p=0.260) we assessed it for phrases with carrier C only. The non-normality of the distribution of N_2 durations related to the fact that N_2 duration varied systematically by carrier phrase (A: mean=574 ms, B: mean=607 ms, C: mean=530 ms; F(2, 27)=16.52, p<0.001). Post hoc comparison revealed that it was the mean for carrier C that stood out: while this was significantly lower than those for carrier A (Tukey HSD: p=0.001) and carrier B (p<0.001), the means for carrier A and B were not significantly different from each other (p=0.281). It is of course not surprising that carrier C was associated with the lowest N_2 duration mean: unlike in carriers A and B, N_2 was not final here, so not subject to phrase-final lengthening (Turk & Shattuck-Hufnagel, 2007; Wightman et al., 1992). In support of this interpretation, N_1 duration did not vary systematically by carrier phrase (F(2, 27)=0.05, p=0.951). As in the Experiment 1 materials, N_1 was generally shorter than N_2 (overall mean=421 ms).

The speaker's articulation rate averaged 3.3 sylls/s (range 3.0–3.8 sylls/s) and 10.0 segs/s (range 7.7–11.7 segs/s). The variation in articulation rate was systematic by carrier (syllable rate: F(2, 27)=24.39, p<0.001; segment rate: F(2, 27)=5.84, p=0.008): as the number of syllables went up, articulation rate went up too (for syllable rate, A: mean=3.1 sylls/s, B: mean=3.4 sylls/s, C: mean=3.5 sylls/s; for segment rate, A: mean=9.3 segs/s, B: mean=10.0 segs/s, C: mean=10.5 segs/s). This is presumably in part because the additional syllables in carriers B and C are unstressed, and therefore relatively short. Our complexity manipulation had a systematic effect on N_1 duration: the higher the number of segments in the noun, the longer it was (r=0.72, p<0.001). For N_2, this was not the case (r=0.17, p=0.379). These patterns were stable across carrier phrases.

4.2.4 Manipulation

As in Experiment 1, we manipulated the experimental phrases to control their temporal make-up, f0 contours and mean intensities, using PSOLA in *Praat* (Boersma & Weenink, 2017). In order to control both the phrases' overall syllable rates and internal syllable rate 'contours' — the latter as motivated in Experiment 2 — we did the temporal manipulation in two steps. First, we set all phrase durations to yield a constant syllable rate of 4 sylls/s across phrases: this meant 1s for (4-syllable) carrier A phrases, 1.25s for (5-syllable) carrier B phrases and 1.5s for (6-syllable) carrier C phrases. As in Experiment 1, this involved temporal compression in all cases.

We then calculated the mean durations for the phrase constituents *this*, N_1, *or* (carrier A only), *or that* (carriers B and C only), N_2 and *then* (carrier C only) across all phrases in

which they occurred and set all individual constituents' durations to these means: see Table 7. Because we calculated means across carrier phrases where feasible, fixing constituent durations accordingly sacrificed exact equality of the phrases' overall syllable rates: these now varied between 3.8 sylls/s and 4.1 sylls/s. Our method also meant that in carrier C phrases, N_2 was likely to sound noticeably long: it was produced substantially faster in carrier C than in A and B in the original productions, and thus, setting its duration to the mean of all carriers entailed a relatively greater lengthening of it in C. Both of these points implied the possibility of confounding factors in the analysis of the relationship between segment rate variation and listeners' tempo judgements. However, we deemed this manipulation preferable to one in which the phrases' overall syllable rates were exactly identical, but phrase-internal syllable rate 'contours' varied as in the original productions — as had been the case in Experiment 1. In the current design, all phrases were temporally identical — aside from differences in segment rate — up to and including N_1; phrases with carriers B and C were further identical up to and including N_2. This is illustrated in Figure 8.

<table>
<tr><th></th><th>this</th><th>N₁</th><th>or</th><th>or that</th><th>N₂</th><th>then</th><th>total</th><th>syllable rate</th></tr>
<tr><td>A</td><td rowspan="3">143</td><td rowspan="3">350</td><td>85</td><td></td><td rowspan="3">461</td><td></td><td>1039</td><td>3.85</td></tr>
<tr><td>B</td><td></td><td rowspan="2">251</td><td></td><td>1205</td><td>4.15</td></tr>
<tr><td>C</td><td></td><td>300</td><td>1505</td><td>3.97</td></tr>
</table>

Table 7. Durations (ms) and syllable rates (sylls/s) in Experiment 3 experimental phrases, after manipulation

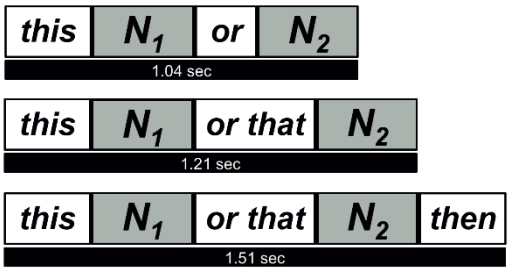


Figure 8. Temporal organisation of experimental phrases (see Figure 6 above): Set A (top), Set B (middle) and Set C (bottom).

Along with the temporal compression, we manipulated the phrases' f0 contours to minimize variation that might have an impact on listeners' tempo judgements. We took f0 measurements at selected points throughout the phrases and used the means across phrases as a basis for a stylised contour; we then resynthesized all individual phrases with this contour. Mean f0 points were calculated separately for each carrier phrase to maintain naturalness. This introduced some degree of variation in f0 means and ranges between phrases with different carriers (A: mean=216 Hz, range=112 Hz; B: mean=209 Hz, range=134 Hz; C: mean=199 Hz, range=103 Hz) which might affect listeners' tempo judgements; we will return to this below. We set mean intensity of all phrases to 65 dB.

For the distractor phrases, we manipulated duration and mean intensity only, as in Experiment 1. To expose participants to a degree of syllable rate variation in the experiment as a whole, we set phrase durations differently for subsets of distractor phrases, so that articulation rate was either 4 sylls/s, 4.25 sylls/s or 4.5 sylls/s. We set mean intensity to 62 dB.

4.2.5 Pairing and stimulus set reduction

We initially used the same pairing method as in Experiments 1 and 2, pairing experimental phrases exhaustively across the two phrase sets with 0.5s silence and each phrase pair featuring in one order only; and pairing distractor phrases within subsets of the phrases we created to produce a range of syllable rate differences across pairs.

After applying this method to the data set as described so far, we decided to proceed with a reduced version of the initial design. This was primarily for practical reasons: while Experiments 1 and 2 were run as stand-alone experiments, Experiment 3 needed to be run following another experiment that is not reported here. Based on pilot runs of both experiments, we estimated that a design of approximately 150 stimuli, presented in three blocks of 50, would minimize the risk of participants losing concentration through the experiment. We therefore reduced the stimulus set systematically by including three instead of five levels of complexity in the nouns. We kept the phrases with the simplest and most complex noun shapes ($N_{1(CVC)}$~$N_{2(CVC)}$ and $N_{1(CCVCC)}$~$N_{2(CCVCC)}$) — that is, with the lowest and highest segment rates — and one instead of three 'intermediate' levels of complexity ($N_{1(CCVC)}$~ $N_{2(CCVC)}$). This reduced design yields two sets of 3 (levels of complexity) x 3 (carriers) = 9 phrases, so exhaustive pairing across the two sets yields 81 experimental phrase pairs. We deemed it unnecessary to recalculate constituent duration means in the corresponding smaller set of original productions and adjust constituent durations accordingly, so the temporal make-up of the stimuli was that summarized in Table 7.

The result of our phrase design and manipulations was that in 9 out of the 81 experimental phrase pairs (11%), segment rate was equal too; in the remaining pairs, the difference in segment rate between pair members varied between –4.4 segs/s and 4.4 segs/s. Because of the greater temporal variability in the design for this experiment, compared with Experiments 1 and 2, the segment rate difference distribution is more fine-grained: the proportional segment rate of the second pair member (relative to the first) took 73 values ranging between 0.68 and 1.46. The smallest divergence from equality was by 0.01. A total of 10 phrase pairs had a proportional segment rate that was not 1, but diverged from 1 by less than 0.06. Since the JND for speech tempo according to Quené (2007) is around 5%, we cannot assume that the measured difference in segment rate for these 10 pairs was noticeable to listeners, let alone that it affected their tempo judgements.

We used the manipulated distractor phrases to create 69 phrase pairs to add to the 81 experimental pairs. For the distractors, we paired 'within' carriers only: that is, in all phrase pairs, the two phrases had the same carrier phrase. We did this so that the carrier phrase variation in the experimental phrase pairs would be more noticeable. We did the distractor phrase pairing so that approximately a third of the resulting pairs had a second pair member

with a lower syllable rate (Δ –0.25 or –0.5 sylls/s), a third had a second pair member with a higher syllable rate (Δ 0.25 or 0.5 sylls/s), and a third had no syllable rate difference between pair members.

As in Experiments 1 and 2, we played all phrase pairs to two colleagues with experience of running listening experiments, but no specific knowledge of our manipulations. They listened to the pairs in random order and deemed their productions to be sufficiently natural for a tempo judgement task

4.2.6 Procedure and analysis method

The experimental procedure was identical to that of Experiments 1 and 2. The analysis method was identical too, except again for the variables entered into our models; these are listed in Table 8.[12] We added measures of proportional phrase duration and proportional syllable rate to the variable set, as well as a categorical variable *Pair type*. *Proportional phrase duration* (the duration of the second phrase expressed as a proportion of the first) and *Proportional syllable rate* (the syllable rate of the second phrase expressed as proportion of the first) result from the duration and syllable rate differences between carriers and are weakly correlated with each other ($r=0.57$, $p<0.001$). Note that the three carriers also have closely comparable, but non-identical f0 contours. A measure quantifying the difference in f0 mean between phrase pair members was so closely correlated with *Proportional phrase duration* ($r=–0.92$, $p<0.001$) that the two measures could be considered quantitatively equivalent. While we took *Proportional phrase duration* forward in the analysis, we should therefore exercise caution in interpreting any effects of this variable. Finally, the categorical variable *Pair type* generalizes across the phonetic differences resulting from our use of three carriers; it has the nine levels ‘AA’, ‘AB’, ‘AC’, ‘BA’, ‘BB’, ‘BC’, ‘CA’, ‘CB’ and ‘CC’.

random variables	*Participant*, *Phrase*$_1$, *Phrase*$_2$	
control variables	*Lexical frequency difference* *Proportional phrase duration* *Proportional syllable rate* *Pair type*	as in Experiment 1 dur$_{Phrase2}$/dur$_{Phrase1}$ syllrate$_{Phrase2}$/syllrate$_{Phrase1}$ ‘AA’, ‘AB’, ‘AC’ etc.
crucial predictor	*Proportional segment rate*	as in Experiment 1

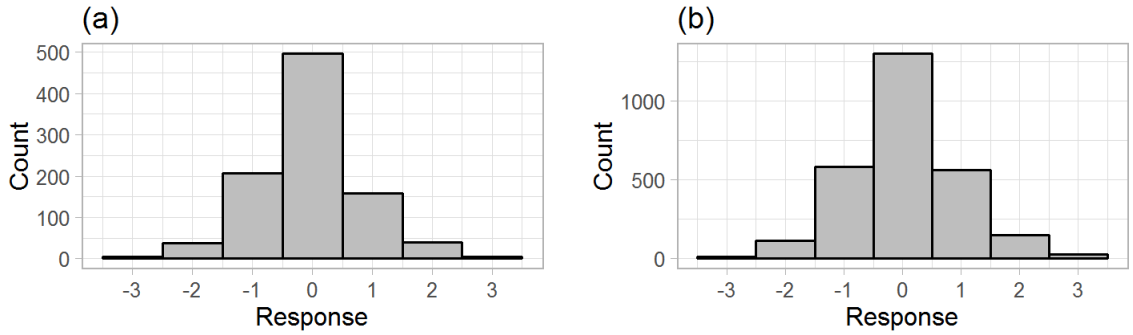
Table 8. Experiment 3 analysis variables

4.3 Results

Figure 9 shows the distribution of *Response* for phrase pairs with a segment rate difference and phrase pairs without. Given the fine-grained distribution of segment rate ratios across the pairs, and the low proportion of pairs with a segment rate ratio of exactly 1, we present *Response* distributions for phrase pairs with segment rate ratios above 0.94 and below 1.06 — whose difference is below or equivalent to the JND estimate for speech tempo of Quené (2007) — and phrase pairs with segment rate ratios outside of this range, thus most likely

[12] The random effects estimates of the optimal model reported below are: *Participant* variance=0.116, SD=0.340; *Phrase*$_1$ variance=0.000, SD=0.000; *Phrase*$_2$ variance=0.098, SD=0.313.

above JND. Figure 9 shows that the distributions both look close to symmetrical, providing no further evidence for a bias towards hearing the second phrase pair member as faster. Table 10 shows that listeners heard no difference in around half of the phrase pairs but slightly less often if they had a difference in segment rate that was above JND (1298 zero responses to 2745 stimuli, or 47%) than if it was not above JND (476 zero responses to 900 stimuli, or 53%). Zooming in on particular subranges of segment rate difference does not reveal any additional patterns: Table 9 shows that listeners consistently heard no tempo difference in about half of phrase pairs, whether these pairs had measurable segment rate differences exceeding 10% or measurable segment rate differences between 6% and 10%.[13]



10

Figure 9. Distribution of *Response* for (a) phrase pairs with a segment rate ratio between 0.94 and 1.06 and (b) phrase pairs with segment rate ratios outside this range

Proportional segment rate	<0.9	0.9–0.94	>0.94, <1.06	1.06–1.1	>1.1
N responses	990	315	900	270	1125
Response = 0	473 (48%)	133 (42%)	476 (53%)	122 (45%)	536 (48%)

Table 9. Numbers and proportions of zero responses by *Proportional segment rate* subrange

In modelling *Response*, using cumulative link mixed models as in Experiments 1 and 2, we first assessed the relevance of our control variable *Lexical frequency difference*, by adding it to a base model with random intercepts for *Participant*, *Phrase*$_1$ and *Phrase*$_2$. *Lexical frequency difference* yielded no significant effect, so was omitted from the model. Next, we considered the three control variables *Proportional phrase duration, Proportional syllable rate* and *Pair type.* Since these all capture partially overlapping information, we added each separately to the base model, in order to identify the one that yielded the greatest improvement in fit. This revealed significant effects for each variable. *Proportional phrase duration* (est=–2.856, se=0.503, z= –5.684, p<0.001) had a negative effect, i.e. the shorter the second phrase was relative to the first, the more likely participants were to judge it as faster. *Proportional syllable rate* (est=9.564, se=3.257, z=2.936, p=0.003) likewise had a negative effect. However, model fit was better with *Pair type* (AIC=9266) than with either

[13]We can also report that response latencies were not significantly correlated with *Proportional segment rate* (r=–0.006, p=0.7) and female and male participants produced near-identical *Response* distributions. In modelling, we checked whether *Gender* improved the fit of our base model; it did not.

Proportional phrase duration (AIC=9282) or *Proportional syllable rate* (AIC=9302). Therefore, we retained *Pair type* in the model. Adding interactions between *Pair type* and either *Proportional phrase duration* and *Proportional syllable rate* yielded no further improvement in fit, so we proceeded with an expanded base model consisting of random intercepts for *Participant*, $Phrase_1$ and $Phrase_2$ and a main effect for *Pair type.*

Next, we added *Proportional segment rate*, our crucial predictor, to the expanded model, along with its interaction with *Pair type*. This yielded a significant effect of *Proportional segment rate* and a significant interaction with *Pair type*. Table 10 shows the statistical results, and Figure 10 displays the interaction, again treating *Response* as a numerical variable for purpose of illustration. Within each row of sub-plots, we see a general decrease in *Mean response* from left to right. This likely reflects the durations of the phrases: Carrier A is the shortest carrier; carrier C is the longest. When an A phrase is paired with a B phrase ('AB'), the B phrase is more often judged to sound slower than another A pair member ('AA'), and when it is paired with a C phrase ('AC') the C phrase is judged yet more often to sound slower. When a B phrase is paired with an A phrase ('BA'), the latter is more often judged to sound faster than another B pair member ('BB'); when it is paired with a C phrase ('BC'), the latter is judged more often to sound slower. When a C phrase is paired with a B phrase ('CB'), the latter is more often judged to sound faster than another C pair member ('CC'), and when it is paired with an A phrase ('CA') the latter is judged yet more often to sound faster. These patterns are reflected in Table 10 in the significant difference between 'BC' and the reference level 'AA', and the marginal difference between 'AC' and 'AA'.

The significant positive effect of *Proportional segment rate* is visible in the linear fit lines, which have a clearly positive slope in some sub-plots and a weakly positive or level slope in most. The slope lines for 'AA', 'BB' and 'CC' are close to level. These phrase pairs were maximally similar to those used in Experiments 1 and 2 — in fact, 'BB' pairs were identical. We therefore see again that when comparing phrases with identical durations and syllable rates, listeners show no obvious orientation to segment rate differences. The slope lines for 'AB', 'AC' and 'CB' look similarly close to level; it is those for 'BA' and 'CA' that are most suggestive of a positive effect of *Proportional segment rate,* and indeed, Table 10 shows that these both differ significantly from the slope for 'AA'. We can only speculate as to why this might be. It may be worth noting that both pair types end in an A phrase; as the listeners' task focused their attention crucially on the second phrase pair member, it may be that a segment rate effect is most likely to surface when the crucial stimuli are kept short and variation in phonological complexity has the greatest proportional impact on segment rate. In general terms, both our descriptive statistics and modelling suggest that the effect of *Proportional segment rate*, while significantly contributing to model fit, in the expected direction, is relatively weak.

	Estimate	Standard error	*z* value	*p* (>|z|)
Pair type 'AB'	–1.4059	0.8664	–1.623	0.1047
Pair type 'AC'	–1.4625	0.8811	–1.66	0.0969 .
Pair type 'BA'	–1.0258	0.7274	–1.41	0.1584
Pair type 'BB'	–0.2729	0.9017	–0.303	0.7621
Pair type 'BC'	–2.415	0.9515	–2.538	0.0111 *
Pair type 'CA'	–0.3016	0.7626	–0.395	0.6925
Pair type 'CB'	–0.3367	0.9848	–0.342	0.7324
Pair type 'CC'	–1.5353	0.996	–1.542	0.1232
Proportional segment rate	1.2558	0.5303	2.368	0.0179 *
Pair type 'AB' : *Proportional segment rate*	0.7345	0.7691	0.955	0.3396
Pair type 'AC' : *Proportional segment rate*	–0.1253	0.7989	–0.157	0.8754
Pair type 'BA' : *Proportional segment rate*	1.7433	0.7365	2.367	0.0179 *
Pair type 'BB' : *Proportional segment rate*	0.188	0.8426	0.223	0.8234
Pair type 'BC' : *Proportional segment rate*	1.6013	0.9164	1.747	0.0806 .
Pair type 'CA' : *Proportional segment rate*	1.5039	0.7582	1.983	0.0473 *
Pair type 'CB' : *Proportional segment rate*	0.2654	0.9076	0.292	0.7699
Pair type 'CC' : *Proportional segment rate*	1.0995	0.942	1.167	0.2431

Table 10. Summary statistics for the optimal model of *Response* for Experiment 3.

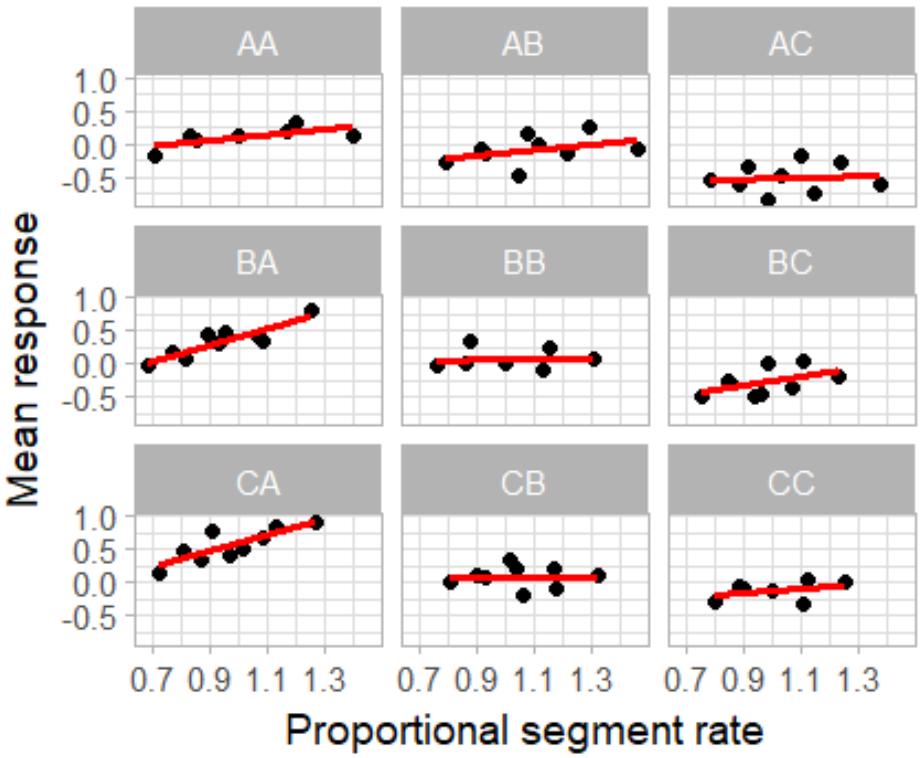


Figure 10. Relationship between *Proportional segment rate* (*x*-axis) and *Mean response* (*y*-axis) split by *Phrase pair*; plots have a linear fit line, and in each, each data point is the mean *Response* value for the corresponding *x*-axis value. A higher *Mean response* value reflects judgements that the second phrase in the pair sounds faster.

4.4 Discussion

Experiment 3 was motivated by the possibility that the substantial proportion of 'no difference' responses in Experiments 1 and 2 might be due to the fact that we equalized the syllable rates of our experimental phrases by equalizing their durations. We predicted that varying phrase durations and syllable numbers would reduce the overall proportion of 'no difference' responses, and that this in turn could reveal sensitivity on our listeners' part to our complexity manipulation, reflected in *Proportional segment rate* values.

These predictions were supported to an extent. Listeners reported 'no difference' for around 50% of phrase pairs, compared with around 65% or 70% for Experiments 1 and 2. Results revealed a significant, though weak, effect of *Proportional segment rate*: at least in some phrase pairs, phrases that had higher segment rates, because they contained more phonologically complex nouns, were perceived as faster. *Phrase pair* emerged as a stronger predictor of responses, and Figure 10 suggests that *Proportional phrase duration* makes a substantial contribution to its effect: phrase pairs ending in the shortest ('A') phrases elicited the most 'faster' responses and phrase pairs ending the longest ('C') phrases elicited the least. Of course, our three carriers are also different in terms of their syntax and prosody; however, we believe that none of our other measures can straightforwardly explain this response

pattern ‘by second phrase’. *Proportional phrase duration* can: stimuli whose production took relatively little time were rated as relatively fast.

5 General discussion

Across three experiments, we expected to find that, perhaps above a certain threshold, segment rate variation due to variation in syllable complexity would affect listeners’ speech tempo judgements. We found no evidence for this in Experiment 1, where phrase duration and syllable rate were held constant, nor in Experiment 2 where the duration of individual phrase constituents was held constant as well. In contrast, Experiment 3 allowed phrase duration to vary while controlling syllable rate. Here, phrases containing more complex syllables *were* more often perceived as relatively fast than phrases containing less complex syllables — at least in some phrase pairs. None of the experiments yielded evidence for a specific threshold above which segment rate variation mattered, and below which it did not. Finally, in all three experiments, responses to distractors revealed robust effects of syllable rate on tempo perception, and some unpredicted factors emerged as consequential, too.

On one level, the results can be taken to support Pfitzinger (1999) and Mixdorff & Pfitzinger (2005)’s conclusions that of the two rate predictors, segment rate is weighted less heavily by listeners than syllable rate. Of course, what is intriguing is why we found a segment rate effect where we found it, and not where we did not. We do not have clear answers to these questions, but can offer several observations. First, to explain why we did not observe an effect of segment rate variation in Experiments 1 and 2, and found no evidence to suggest that ‘AA’, ‘BB’ and ‘CC’ phrase pairs contributed to the observed effect in Experiment 3, it may be helpful to consider how complexity affects duration in natural speech. As reviewed in the introduction, increases in syllable complexity can be manifested by increases in syllable duration and temporal compression at the segmental level (Klatt, 1976; Browman & Goldstein, 1988; Byrd, 1995; Greenberg et al., 2003; Marin & Pouplier, 2010). The same seems to apply at higher levels ― for example, as syllables are added to feet (Eriksson, 1991). If listeners implicitly know that compression of sub-units within a complex structure can occur, and to variable degrees, they may factor this in to their estimation of tempo. When the overall structure of the materials makes it clear that segment-level compression is occurring (as in Experiment 2 in particular, as the most and the least complex nouns had exactly the same duration), listeners might conclude that syllable complexity is *not* a strong influence on duration, and that they should therefore privilege syllable rate in their estimation of tempo. This would mean that causing segment rate and syllable rate to diverge in as extreme a way as we did in Experiments 1 and 2 may *not* have created the ideal conditions for an effect of segment rate to emerge. Rather, listeners may have made the reasonable inference that segments were not the major organising principle for the timing of these particular stretches of speech, and may therefore have based their judgements more upon other factors ― such as the sameness of overall duration and syllable rate.

Some support for this logic comes from O’Dell and Nieminen (2019), who tested the role of sub-syllabic complexity on tempo perception for Finnish. Finnish is a quantity

language where syllables vary considerably in duration according to their number of morae. They found that both mora rate and syllable rate influenced perceived tempo, contrary to our results for English but in accordance with the temporal organisation of Finnish, where morae cannot be compressed too far without compromising the distinctiveness of long and short vowels and consonants. We could further test this logic by looking at higher levels in the prosodic hierarchy for English: we could conduct a parallel to Experiment 2, and examine pairs of utterances that have the same duration, foot-rate, and constituent durations, but where feet vary in complexity by virtue of containing different numbers of unstressed syllables (e.g. *this pattern or that habit* vs. *this splat or that splodge*). If syllable rate is always key to tempo perception, the utterance with more syllables should be perceived as faster than one with fewer, but if listeners compensate perceptually for compression of sub-units, their tempi would be perceived as similar.

Second, regarding the question of why we observed the effect of segment rate variation where we did in Experiment 3, the effect seemed mostly due to listeners' responses to phrase pairs which had in common that their second pair member was of the shortest phrase type (*this N or N*) and their first pair member was of a longer one (*this N or that N* or *this N or that N then*). We pointed out in our presentation of the Experiment 3 results that the task we gave our listeners focused their attention in particular on the second pair member, and in shorter phrases, increases and decreases in syllable complexity have a greater proportional impact on segment rate than in longer phrases. It seems worth considering that listeners' attention to segment-level temporal patterns in making phrase-level tempo judgements varies with phrase duration, with longer phrases promoting less granular judgements. This could be tested through further experiments along the lines of Experiment 3. Interestingly, Pfitzinger (1999) elicited tempo judgements on stimuli that were substantially shorter than those in our experiments (and most of the other experiments we have cited), at around 0.6s.

Moving beyond the findings for segment rate variation, our results support and extend findings in the literature that tempo perception is affected by properties other than those captured by articulation rate measurements. Previous studies identified influences of f0 height and dynamic variation of f0 and intensity contours, and of the peripherality of vowels (Cumming, 2011; Feldstein & Bond, 1981; Kohler, 1986; Weirich & Simpson, 2014). To this list of influences, we can — tentatively — add utterance duration and utterance-internal temporal variability. The Experiment 3 results suggest that other things being equal, a longer stretch of speech sounds slower than a shorter stretch of speech. This warrants further exploration, as it is not obvious that the effect of phrase duration would necessarily generalise beyond the structurally homogeneous phrase set investigated here. If it does, and if our reasoning about a possible limiting effect of phrase length on listeners' attention to segmental detail is on the right lines, stimulus duration is a parameter that must be carefully considered in future experimental work on speech tempo perception.

As for phrase-internal variability, the Experiment 1 results suggest that other things being equal, a stretch of speech with more internal temporal variability (faster "fast portions" and slower "slow portions") sounds faster than one with less variability. In unpublished work using further manipulations of the Experiment 1 stimuli, we confirmed this latter finding.

Lengthening the stressed syllables (N_1 and N_2) by 15% (and correspondingly shortening the unstressed syllables so that phrase duration remained constant) increased perceived tempo. Lengthening the unstressed syllables by 15% (and correspondingly shortening stressed syllables) caused no change. When only N_2 was lengthened, perceived tempo increased, but not as much so as when both N_1 and N_2 were lengthened. These findings strongly suggest that greater rhythmic variation, like greater f0 or intensity variation, can correlate with higher perceived tempo. Again, more work is needed to explore this finding. Nonetheless, our results support the idea that like perceived rhythm (Arvaniti, 2009) tempo perception is multidimensional, in ways that do not only relate to units per unit time.

6 Conclusion

We set out to discover how well segment rate correlates with listeners' perceptions of tempo when it is not itself strongly correlated with syllable rate, whose relevance for speech tempo perception is well-established. Teasing the two apart matters because in a language like English, the two rates can diverge as a function of variation in syllable complexity. In a series of pairwise discrimination tasks, we held syllable rate constant while allowing complexity — and therefore segment rate — to vary. We hypothesized that this variation would have an observable effect on listeners' tempo judgements. Our results suggest that our decision to use an invariant phrase duration as well as invariant syllable rate may have been responsible for the lack of an effect of segment rate in Experiments 1 and 2. When syllable rate and overall duration of frame are held constant, it seems that listeners have a powerful perceptual impression of sameness of tempo, and they do not interpret segment rate variation resulting from phonological complexity as reflecting tempo difference. Once there is some variation in these properties, it appears that segment rate becomes a relevant parameter — although further work is clearly needed to delimit the domain and extent of its relevance.

Acknowledgements

This research was supported by a British Academy and Leverhulme Trust Small Research Grant (SG151790) and a Leverhulme Trust Research Grant (RPG-2017-060). We are grateful to Nathan Clarke and Robert Lennon for their contributions to this research, and to four anonymous reviewers and the associate editor for constructive criticism.

References

Arvaniti, A., & Rodriquez, T. (2013). The role of rhythm class, speaking rate, and F-0 in language discrimination. *Laboratory Phonology, 4*(1), 7-38.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer. www.praat.org.

Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology, 8*.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in Articulatory Phonology. *Phonetica, 45*(2-4), 140-155.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918823199.

Byrd, D. (1995). C-centers revisited. *Phonetica, 52*(4), 285-306.

Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics, 24*(2), 263-282.

Christensen, R. H. B. (2018). Ordinal -- Regression models for ordinal data (R package version 2018.8-25)

http://www.cran.r-project.org/package=ordinal/.

Collins, B., & Mees, I. (2013). *Practical phonetics and phonology* (3 ed.). London: Routledge.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America, 88*(1), 101-112.

Cumming, R. E. (2011). The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics, 39*(3), 375-387.

Dellwo, V., Ferrange, E., & Pellegrino, F. (2006). *The perception of intended speech rate in English, French, and German by French speakers.* Paper presented at the Third International Conference on Speech Prosody, Dresden.

Den Os, E. (1985). Perception of speech rate of Dutch and Italian utterances. *Phonetica, 42*, 124-134.

Feldstein, S., & Bond, R. N. (1981). Perception of speech rate as a function of vocal intensity and frequency. *Language and Speech, 24*(Oct-), 387-394.

Gibbon, D., Klessa, K., & Bachan, J. (2015). Duration and speed of speech events: A selection of methods. *Lingua Posnaniensis, 56*(1), 59-83.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. Y. (2003). Temporal properties of spontaneous speech: A syllable-centric perspective. *Journal of Phonetics, 31*(3-4), 465-485.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America, 128*(2), 839-850.

Klatt, D. H. (1976). Linguistic uses of segmental durations in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*(5), 1208-1221.

Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: Duration, f0 movement, and f0 level. *Language and Speech, 29*, 115-139.

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

Lehiste, I. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics, 4*(2), 113-117.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328-348.

Lidji, P., Palmer, C., Peretz, I., & Morningstar, M. (2011). Listeners feel the beat: Entrainment to English and French speech rhythms. *Psychonomic Bulletin & Review, 18*(6), 1035-1041.

Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control, 14*(3), 380-407.

Mixdorff, H., & Pfitzinger, H. R. (2005). Analysing fundamental frequency contours and local speech rate in map task dialogs. *Speech Communication, 46*(3-4), 310-325.

Moore, R. E., Adams, E. M., Dagenais, P. A., & Caffee, C. (2007). Effects of reverberation and filtering on speech rate judgment. *International Journal of Audiology, 46*(3), 154-160.

Orme, J. G., & Combs-Orme, T. (2009). *Multiple regression with discrete dependent variables*. Oxford: Oxford University Press.

Pfitzinger, H. (1999). *Local speech rate perception in German speech.* Paper presented at the Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco.

Pfitzinger, H., & Tamashima, M. (2006). *Comparing perceptual local speech rate of German and Japanese speech.* Paper presented at the Proceedings of the 3rd International Conference on Speech Prosody, Dresden.

Plug, L., & Carter, P. (2014). Timing and tempo in spontaneous phonological error repair. *Journal of Phonetics, 45*, 52-63.

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

Quené, H. (2013). Longitudinal trends in speech tempo: The case of Queen Beatrix. *Journal of the Acoustical Society of America, 133*(6), El452-El457.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association, 34*(2), 239-245.

Schultz, B. G., O'Brien, I., Phillips, N., Mcfarland, D. H., Titone, D., & Palmer, C. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics, 37*(5), 1201-1220.

Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., Witzlack-Makarevich, A., de Jong, N. H., & Bickel, B. (2018). Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences, 115*(22), 5720-5725.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: a practical guide. In S. Sudhoff & D. Lenertova & R. Meyer & S. Pappert & P. Augurzky & I. Mleinek & N. Richter & J. Schliesser (Eds.), *Methods in Empirical Prosody Research* (pp. 1-28): Mouton de Gruyter.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*(4), 445-472.

Vaane, E. (1982). Subjective estimation of speech rate. *Phonetica, 39*(2-3), 136-149.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

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

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language, 66*(4), 665-679.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707-1717.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*(6), 957-968.

Appendix A: Experiment 1–2 phrases

Experimental phrases

Set 1	Set 2
this kit or that pack	*this tick or that pet*
this clock or that pot	*this trip or that kick*
this temp or that cop	*this camp or that pit*
this print or that cut	*this clamp or that cup*
this pip or that clot	*this tuck or that clip*
this top or that kilt	*this cat or that pest*
this tip or that crest	*this pop or that cramp*
this click or that spot	*this trek or that stop*
this trick or that tact	*this trap or that kink*
this speck or that crust	*this track or that stump*
this tusk or that stick	*this test or that step*
this tank or that tent	*this fist or that pant*
this cost or that slump	*this tuft or that scalp*
this trust or that stock	*this tramp or that skip*
this plank or that pump	*this trunk or that pelt*
this prank or that stunt	*this stomp or that trump*

Distractor phrases

Set 1	Set 2
this church or that chapel	*this mister or that miss*
this stream or that river	*this alto or that bass*
this bank or that bistro	*this panto or that play*
this loft or that attic	*this disco or that dance*
this pill or that powder	*this bottle or that mug*
this snake or that rattle	*this pastor or that priest*
this cause or that effect	*this record or that disk*
this method or that way	*this niece or that cousin*
this ramble or that route	*this verse or that chorus*
this kestrel or that kite	*this bow or that ribbon*
this shower or that bath	*this force or that power*
this sprocket or that chain	*this bike or that scooter*
this steeple or that chase	*this pram or that buggy*
this leopard or that lynx	*this nun or that bishop*
this end or that beginning	*this adventure or that tour*
this bean or that potato	*this disaster or that verve*
this soup or that spaghetti	*this decision or that call*
this bereavement or that loss	*this peach or that banana*
this falafel or that wrap	*this tea or that espresso*
this intestine or that gut	*this flat or that apartment*

Appendix B: Experiment 3 phrases

Experimental phrases

Set 1	Set 2
this kit or pack *(this pip or clot)* *this click or spot* *(this speck or crust)* *this prank or stunt*	*this tick or pet* *(this tuck or clip)* *this trek or stop* *(this track or stump)* *this stomp or trump*
this kit or that pack *(this pip or that clot)* *this click or that spot* *(this speck or that crust)* *this prank or that stunt*	*this tick or that pet* *(this tuck or that clip)* *this trek or that stop* *(this track or that stump)* *this stomp or that trump*
this kit or that pack then *(this pip or that clot then)* *this click or that spot then* *(this speck or that crust then)* *this prank or that stunt then*	*this tick or that pet then* *(this tuck or that clip then)* *this trek or that stop then* *(this track or that stump then)* *this stomp or that trump then*

Phrases between brackets were not used in the final design.

Distractor phrases

Set 1	Set 2
this church or chapel *this stream or river* *this bank or bistro* *this loft or attic* *this pill or powder* *this end or beginning* *this bean or potato*	*this mister or miss* *this alto or bass* *this panto or play* *this disco or dance* *this bottle or mug* *this adventure or tour* *this disaster or verve*
this snake or that rattle *this cause or that effect* *this method or that way* *this ramble or that route* *this kestrel or that kite* *this soup or that spaghetti* *this bereavement or that loss*	*this pastor or that priest* *this record or that disk* *this niece or that cousin* *this verse or that chorus* *this bow or that ribbon* *this decision or that call* *this peach or that banana*
this shower or that bath then *this sprocket or that chain then* *this steeple or that chase then* *this leopard or that lynx then* *this falafel or that wrap then* *this intestine or that gut then*	*this force or that power then* *this bike or that scooter then* *this pram or that buggy then* *this nun or that bishop then* *this tea or that espresso then* *this flat or that apartment then*