Measured and perceived speech tempo: Comparing canonical and surface articulation rates

Leendert Plug[1], Robert Lennon[2], Rachel Smith[3]

[1] University of Leeds, United Kingdom
[2] University of Leeds, United Kingdom
[3] University of Glasgow, United Kingdom

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

Abstract

Studies that quantify speech tempo tend to use one of various available rate measures. The relationship between these measures and perceived tempo as elicited through listening experiments remains poorly understood. This study furthers our understanding of the relationship between measured articulation rates and perceived speech tempo, and the impact of syllable and phone deletions on speech tempo perception. We follow previous work in using stimuli from a corpus of unscripted speech, and in sampling stimuli in distinct 'global tempo' ranges. Within our stimulus sets, the differences between canonical and surface rate measurements are directly due to syllable or phone deletions. Our results for syllable rates suggest that listeners use both canonical and surface rates to estimate speech tempo: that is, deletions do not have a consistent effect on perceived tempo. Our results for phone rates suggest that surface phone rate also influences judgements, but canonical phone rate does not. Our results also confirm previously-reported effects of *f*0 and intensity on speech tempo perception, plus an effect of stimulus duration, but no effect of listeners' own tempo production tendencies.

Keywords

speech tempo, articulation rate, deletion, perception, English

Funding and competing interests

This research was made possible by a Leverhulme Trust Research Grant (RPG-2017-060: *Speech tempo perception and missing sounds*). The authors have no competing interests to declare.

Prior publication

We presented an analysis of part of the data set described in this paper at the Nineteenth International Congress of Phonetic Sciences, Melbourne (Plug, Lennon, & Smith, 2019).

Measured and perceived speech tempo: Comparing canonical and surface articulation rates

Introduction

In quantifying speech tempo through rate measurements, researchers must choose what to count and what temporal domain to count in (e.g. Dankovičová, 1997; Jessen, 2007). When counting syllables or phones, researchers can count units as expected in canonical pronunciations ('canonical rates'), or as observed in their data ('surface rates') (e.g. Koreman, 2006). The correlations between these alternative measures vary across datasets: the mapping between canonical and surface rates depends on the prevalence of deletions, which varies within and across languages (Barry & Andreeva, 2001; Johnson, 2004; Kohler, 2000). In English, differences between canonical and surface rates can be substantial. A phrase like *I suppose this terrain is hard* produced in 1.6s yields a canonical rate of 5 syll/s; when produced with schwa deletion in both *suppose* and *terrain* the surface rate would be 3.75 syll/s because schwa deletion entails syllable deletion in these words as the consonants surrounding the schwas constitute well-formed syllable onsets (/sp/ and /tr/, respectively). The measured difference is well above the JND for temporal variation in speech (Quené, 2007). Our research is motivated by the question of how closely common rate measures map onto listeners' tempo ratings, and what this tells us about the respective roles in perception of canonical and surface forms. Here we assess how closely listeners' tempo judgements are correlated with canonical and surface syllable and phone rates, using stimuli sampled from a corpus of unscripted British English speech.

Assessing how closely listeners' tempo judgements are correlated with canonical and surface rates is, effectively, assessing the impact of deletions on speech tempo perception. Few studies have attempted this: typically, studies of tempo perception present syllable or phone rates calculated using *either* canonical *or* surface unit counts (e.g. Gibbon, Klessa, & Bachan, 2015; Pfitzinger, 1999; Vaane, 1982). Den Os (1985) presents a flawed attempt in an investigation of Dutch and Italian listeners' perceptions of speech tempo across the two languages. Den Os mapped listeners' ratings to canonical and surface syllable rates, predicting that listeners can only orient to canonical syllable rate if they know the language well. However, her design did not control the correlations among the rates. As these were very strong, Den Os concluded that she could not reliably establish which mapped more closely to listeners' tempo ratings.

Only two studies have explicitly investigated the impact of deletions on tempo perception (Koreman, 2006; Reinisch, 2016). Koreman (2006) maps canonical and surface phone rates to listeners' tempo judgements of spontaneously-produced German intonation phrases. Listeners completed two tasks: first they were presented with paired phrases and asked which is faster; then they made scalar judgements on the same phrases presented individually. Koreman selected phrases with reference to their measured phone rates to make up six groups: *fast~clear* (high rate, similar canonical and surface rates), *fast~sloppy* (high rate, divergence between canonical and surface rates), *normal~clear for comparison with fast phrases* (average rate, similar canonical and surface rates, surface rates similar to those of

fast~sloppy phrases), *normal~clear* (average rate, similar canonical and surface rates), *normal~sloppy* (normal rate, divergence between canonical and surface rates), and *slow~clear* (low rate, similar canonical and surface rates, surface rates similar to those of *normal~sloppy* phrases). Koreman selected phrases in three rate ranges to test the hypothesis that listeners respond differently to deletions depending on the overall rate: Koreman predicted that in slow speech, listeners would associate greater numbers of phone deletions with very slow, 'slurred' speech; when rate is in the average range or above, they would associate greater numbers of phone deletions with faster, 'hypo-articulated' speech (and smaller numbers of deletions with slower 'hyper-articulation').

Koreman's results show that listeners perceived tempo differences between utterances with similar surface but different canonical rates (*fast~sloppy* utterances were perceived as faster than *normal~clear* ones). However, listeners also perceived differences between utterances with similar canonical but different surface rates: for example, *fast~sloppy vs fast~clear*, where the higher surface rates in *clear* utterances appeared to be 'taken at face value and interpreted as an indication of faster speech' (Koreman 2006: 592). Koreman concludes that listeners orient to *both* canonical and surface rates and are able to compare utterances along both parameters. Koreman also notes that contrary to prediction, the base rate did not affect tempo judgements: 'clarity' and 'sloppiness' appeared to have similar effects among slower and faster utterances.

Reinisch (2016) reports two experiments in which listeners judged the tempo of naturally-produced normal and fast speech, and speech that results from linear rate manipulations. A German utterance was produced at normal rate, with few deletions, and at fast rate with more deletions. Both were manipulated to create an additional 'normal rate' version with the fast-rate deletions and a 'fast rate' one without. The four versions were first used as context sentences in an implicit tempo perception task, in which listeners were asked to judge the identity of a subsequent ambiguous word; this lexical judgement hinged on whether the word's (acoustically-identical) first vowel was perceived as long or short, so that rate normalisation, and accordingly a tempo judgement of the context sentence, could be inferred (Bosker, 2017; Mitterer, 2018; Newman & Sawusch, 1996, 2009; Reinisch, Jesse, & McQueen, 2011; Sawusch & Newman, 2000). The four utterance versions were then used in an explicit tempo perception task involving paired comparison, like Koreman (2006). In the implicit task, the naturally fast utterance version was perceived as faster than the linearly compressed version. This is neither consistent with orientation to surface phone rate nor consistent with orientation to canonical rate; rather, listeners may have drawn on their knowledge that phone deletions tend to occur in fast speech, and the association between high tempo and high deletion rates may have informed their judgements. However, in the explicit task no consistent difference was perceived between naturally fast and linearly compressed versions.

The results of Reinisch (2016) are intriguing, but based on judgements of just one sentence; this makes it difficult to assess how they relate to those of Koreman (2006). More research into the impact of syllable and phone deletions on tempo perception is warranted. In the current study we followed Koreman (2006) in sampling stimuli from a corpus of unscripted speech; a companion study takes the opposite approach of using highly controlled scripted speech (Plug, Lennon & Smith, in preparation). We elicited tempo ratings through a

ranking task (Pfitzinger, 1999; Pfitzinger & Tamashima, 2006), which combines elements of a paired comparison and scalar judgement task. We compared canonical and surface rates for both phones and syllables, to assess the impact of both phone deletions and syllable deletions on tempo perception. Given that canonical and surface rates tend to be highly correlated, we carefully designed an approach to stimulus selection that allowed us to tease apart their influences as far as possible. If listeners attend to multiple temporal parameters when making tempo judgements, then we expect to see effects of both canonical and syllable rates, in line with Koreman (2006). If instead they attend primarily to a single parameter, then we will see effects either of canonical rate, or surface rate, but not both, at the level of syllables and/or phones.

Method

Participants

The experiment was run at the University of Leeds in accordance with institutional ethics regulations. 55 monolingual British English listeners (40 female) aged 18–35 (mean 23) participated. None reported hearing problems. All were paid a small fee.

Materials

Corpus Our corpus was a set of 'memory stretches' extracted by Gold (2014) from the larger *DyViS* (*Dynamic Variability in Speech*) database (Nolan, McDougall, De Jong, & Hudson, 2009). *DyVIS* comprises studio-quality recordings of 100 male speakers of Standard Southern British English (SSBE) aged 18–25 undertaking reading tasks and role-play tasks relevant in a forensic context (a simulated police interview and a telephone call with a supposed accomplice). The homogeneity of the speaker sample makes the database an excellent source for stimuli: it limits variation due to age, gender and other sociolinguistic variables that are known to influence tempo perception in complex ways(e.g. Buller, Lepoire, Aune, & Eloy, 1992; Feldstein, Dohm, & Crown, 2001; Harnsberger, Shrivastav, Brown, Rothman, & Hollien, 2008; Street & Brady, 1982; Weirich & Simpson, 2014). Gold (2014) used the *DyViS* recordings of role-played telephone calls with a supposed accomplice to derive population statistics for articulation rate in SSBE. In these calls, the speakers relayed to the accomplice (the experimenter) the account they had given of their actions during the previous task, a simulated police interview. Speaker and experimenter both had access to a map, and the experimenter used prompts to ensure that key lexical items featured on the map were produced. Following Jessen (2007), Gold segmented the recordings for each participant into 26–32 'memory stretches'. In this procedure, 'the phonetic expert goes through the speech signal and selects portions of fluent speech containing a number of syllables that can easily be retained in short-term memory' (Jessen, 2007, p. 54). According to Jessen, this method is more efficient in casework practice than delimiting inter-pause stretches or intonation phrases. Our corpus comprises the memory stretches that Gold (2014) selected for 30 speakers (N=920).

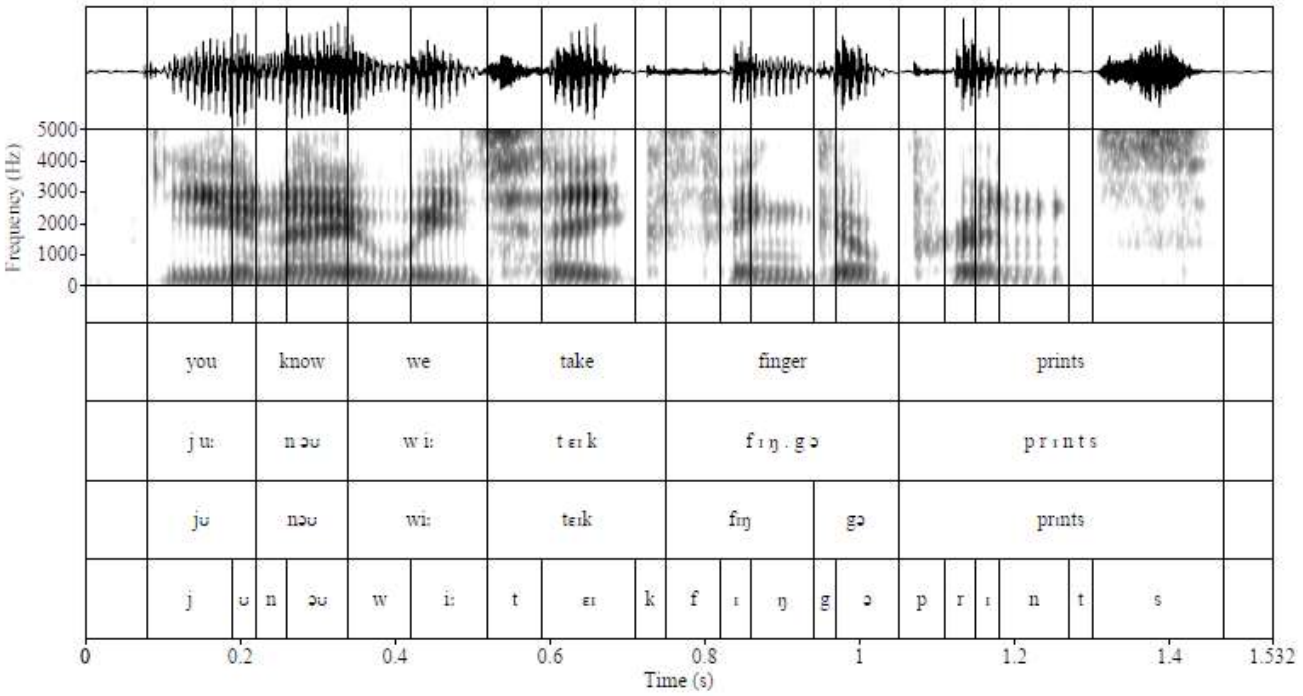
Segmentation We used the BAS web services tools *G2P*, *WebMAUS* and *Phon2Syl* (Kisler, Reichel, & Schiel, 2017) for segmentation, using the 'English (GB)' language model and the stretches' orthographic transcriptions prepared by Gold (2014). These tools produce a

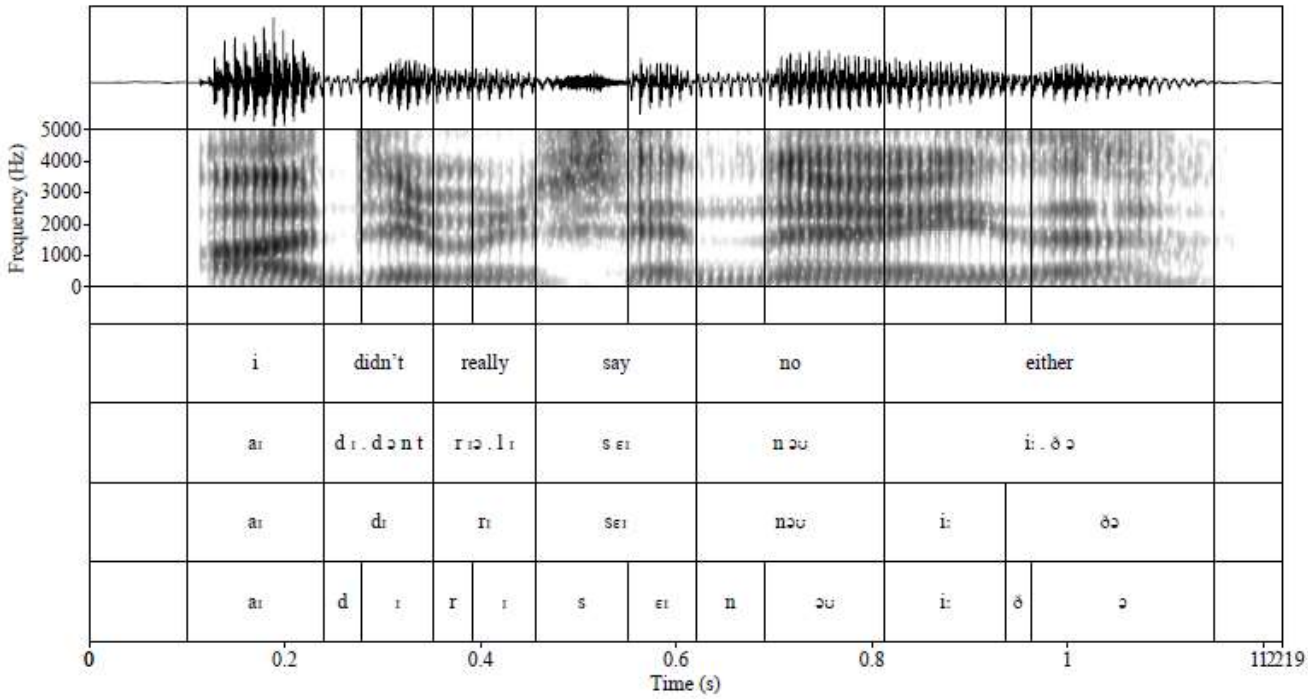
canonical transcription and syllabification (*G2P*), a surface transcription and segmentation (*WebMAUS*), and a syllabification of the surface form (*Phon2Syl*). Syllabification was done within word boundaries.

Plug, Lennon & Gold (2021) have shown for a larger version of the DyViS memory stretch corpus that researcher decisions on canonical forms and those of the BAS tools result in minimally different deletion distributions. Nevertheless we took a cautious approach to checking and correcting the tools' output. The second author checked the output segmentations using *Praat* (Boersma & Weenink, 2017). This revealed three types of inaccuracy, affecting: (1) boundary placements, (2) deletion judgements and (3) syllabifications. We handled these as follows. (1) The precise location of boundaries was not a major concern as we were interested in syllable and phone *rates*, so we manually corrected the alignment only when two or more successive segments with clearly visible acoustic correlates were not aligned with those correlates. Approximately 7% of memory stretches underwent this kind of correction. (2) As phone and syllable deletions are our focus, we applied a more elaborate protocol to correct inaccurate deletion judgements. Our principle was to treat a phoneme as deleted if it was not possible to identify an acoustically-segmentable chunk that corresponded primarily to that phoneme. First, the second author identified a set of frequent lexical items whose heavily reduced productions *WebMAUS* recurrently segmented inaccurately. These items included *actually*, *probably*, *occasionally*, *remember*, and *didn't*. All productions of this set of items were transcribed independently by the other authors, and segmentations were corrected to match consensus transcriptions. Second, the second author listed all other instances of erroneous deletion (where *WebMAUS* treated a phone as deleted when a segmental acoustic correlate could be found; about 10% of stretches) and erroneous non-deletion (where *WebMAUS* treated a phone as present when no segmental acoustic correlate could be delimited; about 30% of stretches). The phone in question was most commonly schwa. All were manually corrected. (3) *WebMAUS* consistently treated surface forms with syllabic consonants as monosyllabic (e.g. *bottle* [ˈbɒtl̩]); these syllabification errors were corrected.

We excluded a number of memory stretches from further consideration. 37 stretches contained an internal silent pause (defined as a silence exceeding 50ms that was not a stop closure). 18 stretches were excluded because accurate segmentation was impossible due to missing initial or final phones, excessive creak or signal disturbances. The resulting corpus had 865 memory stretches. We extracted canonical and surface syllable and phone rates from the corrected segmentations, alongside syllable and phone deletions per memory stretch.

Figure 1 shows segmented waveforms and spectrograms of three example stretches that formed part of our final stimulus set. The top panel shows an utterance with no deletions; the middle panel an utterance with relatively many syllable and phone deletions; and the bottom panel an utterance with no syllable deletions, but relatively many phone deletions.


Frequency (Hz)
5000
4000
3000
2000
1000
0
you
know
we
take
finger
prints
j u:
n əʊ
w i:
t eɪ k
f ɪ ŋ . g ə
p r ɪ n t s
ju
nəʊ
wi:
teɪk
fɪŋ
gə
prɪnts
j
ʊ
n
əʊ
w
i:
t
eɪ
k
f
ɪ
ŋ
g
ə
p
r
ɪ
n
t
s
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.532
Time (s)


Frequency (Hz)
5000
4000
3000
2000
1000
0
i
didn't
really
say
no
either
aɪ
d ɪ . d ə n t
r ɪə . l ɪ
s eɪ
n əʊ
iː . ð ə
aɪ
dɪ
rɪ
seɪ
nəʊ
iː
ðə
aɪ
d
ɪ
r
ɪ
s
eɪ
n
əʊ
iː
ð
ə
0
0.2
0.4
0.6
0.8
1
1.2219
Time (s)

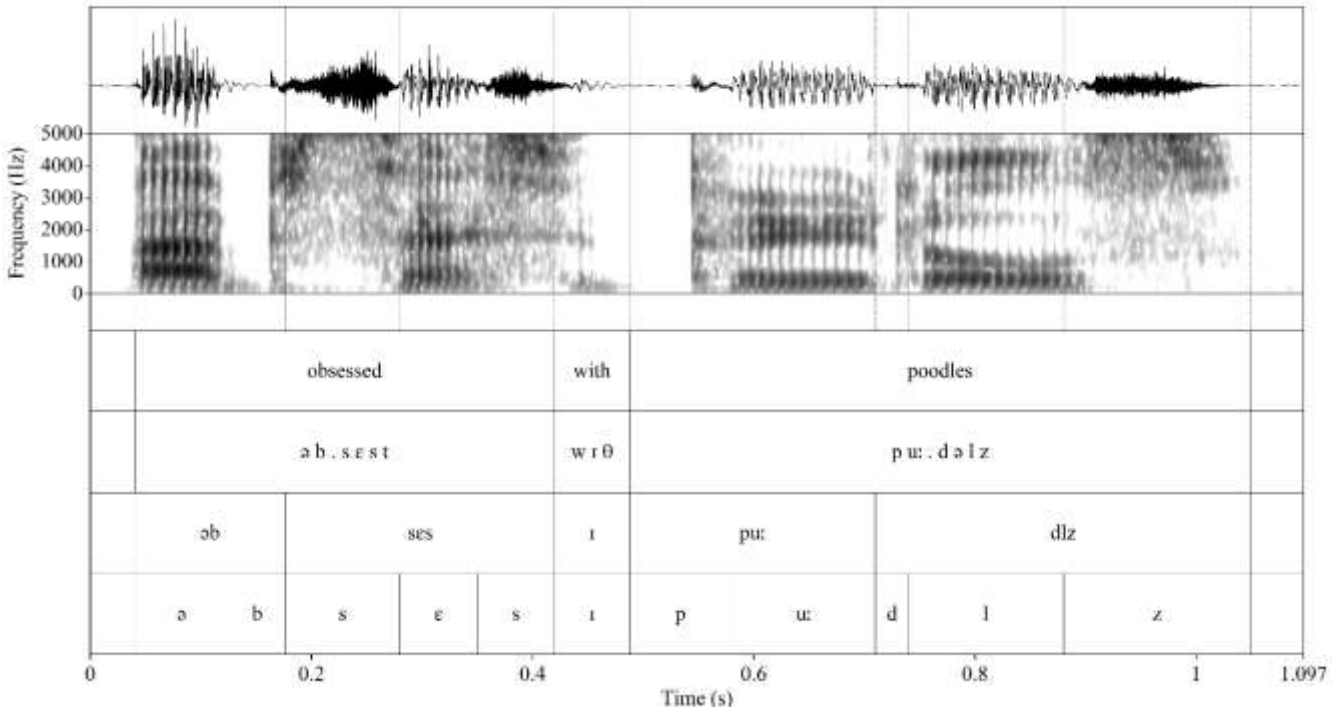


Figure 1: Examples of segmented utterances. Top panel: *you know we take fingerprints,* with no deletions. Middle panel: *I didn't really say no either,* with 2/9 deleted syllables and 6/18 deleted segments. Bottom panel: *obsessed with poodles,* with 0/5 deleted syllables and 4/15 deleted segments.

Corpus statistics To test the relative strength of influences of canonical and surface rates, we first needed to establish the relationship between the two types of rate in the corpus, and then select stimuli in such a way as to decorrelate the two rates as far as possible. The relationship between canonical and surface rates in our corpus is determined by the prevalence of syllable and phone deletion. We identified 314 syllable deletions and 1598 phone deletions: 4% of canonical syllables and 8% of canonical phones in the corpus lack a surface realisation. Syllable deletion occurs in 26% of memory stretches. The maximum number of deleted syllables is seven; the most common number just one. Stretches with four or more syllable deletions are all long (>15 syllables); stretches with less than four cover the full range of stretch lengths. Phone deletion occurs in 73% of memory stretches. The maximum number of deleted phones is 12, but most stretches have between one and four missing phones. Zero deletion is observed in stretches of up to 45 canonical phones. The relationship between syllable and phone deletions is reasonably linear (r=0.69), but each observed number of syllable deletions maps to a considerable range of phone deletions. Overall, these deletion rates seem in line with those reported in corpus-based studies of English (Greenberg, 1999; Johnson, 2004; Robb, Maclagan, & Chen, 2004; Shattuck-Hufnagel & Veilleux, 2007; Tauroza & Allison, 1990).

The correlations between canonical and surface rates are strong. For syllable rates, the correlation between canonical and surface rate is r=0.91; excluding zero-deletion stretches predictably lowers the correlation, but not by much (r=0.89). For phone rates, the correlation

between canonical and surface rate is r=0.90 both including and excluding zero-deletion stretches. For both syllable and phone rates, variability in surface rate increases as canonical rate goes up, reflecting that the likelihood of 'massive reduction' (Johnson, 2004) increases with increasing rate, although speakers do not invariably delete syllables or substantial numbers of phones at higher rates.

Stimulus selection The pairwise correlations just cited pose a challenge: how do we select stimuli that allow us to assess whether canonical or surface articulation rate maps more closely to perceived tempo, when the two rates correlate so strongly? Clearly, random sampling might lead to methodological failure, as in Den Os (1985). Koreman (2006) addressed this challenge by selecting sets of stimuli with distinct combined ranges of canonical and surface phone rate values. For example, Koreman's *fast~sloppy* phrases have canonical phone rates between one and two standard deviations above the mean calculated across his whole corpus, and surface rates within one standard deviation from the mean. His *normal~clear* phrases for comparison have similar surface phone rates to the *fast~sloppy* phrases, but few phone deletions, so little difference between surface and canonical rates. Unfortunately, Koreman does not report how closely canonical and surface rates remained correlated within his six stimulus sets.

We implemented a variant of Koreman's method aimed at ensuring that correlations between comparison rates remained low enough to make the comparisons viable. We selected three sets of 60 stimuli, each optimized to compare two specific rate measures in their mapping to listeners' tempo ratings. As we describe in detail below, set 1 was optimized for comparing canonical and surface *syllable* rates: correlations between these two rates were kept as low as possible, to tease their effects apart to the greatest possible extent and thereby test our hypotheses. Set 2 was similarly optimized for comparing canonical and surface *phone* rates. Set 3 was optimized for comparing surface syllable and surface phone rates. While listeners judged set 3 stimuli, and we incorporated some of these judgements into the analysis we present below, we do not report analysis of the comparison between surface syllable and surface phone rates: our focus is on the comparisons between canonical and surface rates.

To select stimuli for each set from the overall corpus of 865 memory stretches, our approach was to identify a group of stimuli for which one rate remained as stable as possible, while the other rate varied as much as possible: for example, we sought to identify a group of stimuli where surface syllable rate was as stable as the dataset allowed, while canonical syllable rate was as variable as the dataset allowed. The question then arose what the base tempo of the 'stable' rate should be (e.g. fast, medium, or slow). Although base tempo did not materially alter the effects of rate in Koreman (2006), we chose to control base tempo by creating slow, medium and fast stimulus subsets. To this end, we first identified the 10–20%, 45–55% and 80–90% quantile ranges for each of the two (log-transformed) rates being compared (in scatterplot terms, the *x*-axis rate and the *y*-axis rate: see Figure 2) to represent slow, medium and fast rates respectively. Within each of the six quantile ranges (*x*-axis 10–20%, 45–55% and 80–90%; *y*-axis 10–20%, 45–55% and 80–90%) we then selected 10 data points that were as widely dispersed in the *other* (comparison) rate's range as possible (i.e., for *x*-axis quantile ranges, we found the 10 points that had maximum dispersal on the *y*-axis; for *y*-axis quantile ranges, we found the 10 points that had maximum *x*-axis dispersal). In sets

1 and 2 we included one point with identical values for the two rates (that is, no deletion) in each of the six quantile range subsets.

For each rate comparison (i.e., each of sets 1, 2 and 3), this procedure yielded 30 stimuli in three subsets of 10 within which the *x*-axis rate varied little (one subset low in the range, one in the middle, and one high) and the *y*-axis rate varied considerably more widely; and 30 stimuli in three subsets of 10 (again low, medium and high) within which the *y*-axis rate varied little and the *x*-axis rate varied considerably more widely. Figure 2 illustrates the selection method for set 1. The black dots are the three subsets of 10 stimuli within which the *x*-axis rate varies little, and the black triangles are the three subsets of 10 stimuli within which the *y*-axis rate varies little. The grey crosses are the remaining 805 data points in the corpus. Figure 2 illustrates that the ranges of the variable rates in the quantile subsets vary: as noted above, across all memory stretches, surface rate variation increases as canonical rate goes up.

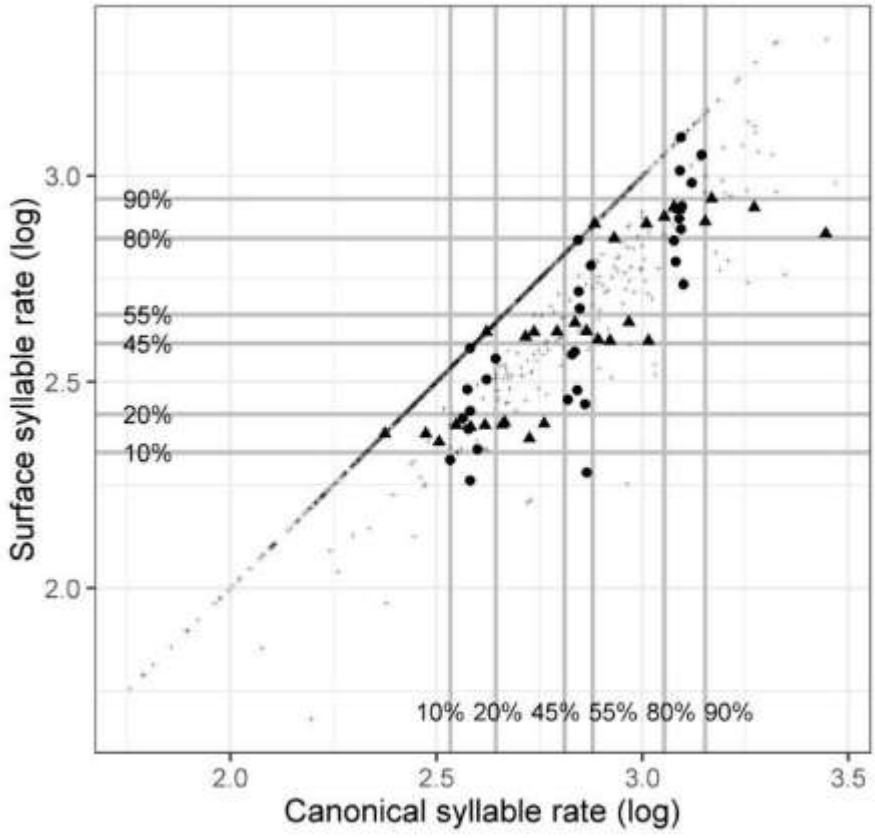


Figure 2: Scatterplot for canonical *vs* surface syllable rates (log syll/s) with quantile range boundaries; black dots and triangles are selected stimuli

Acoustic analysis Experiments show that utterances with a relatively high *f*0 level, a relatively high magnitude of *f*0 movement and relatively high overall intensity are perceived as relatively fast (Feldstein & Bond, 1981; Kohler, 1986; Rietveld & Gussenhoven, 1987) Using *mausmooth* (Cangemi, 2015) in *Praat* (Boersma & Weenink, 2017), we extracted editable *f*0 contours for all of the memory stretches in our stimulus sets (time step of 0.05s, analysis range 15–400Hz). We manually removed clearly

erroneous points. We calculated the mean *f*0 for each corrected contour as a measure of *f*0 level and the kurtosis of the *f*0 distribution as a measure of span (Mennen, Schaeffler, & Docherty, 2012; Niebuhr & Skarnitzl, 2019). We also took a mean intensity measure for each stretch.

Procedure

Tempo rating task We elicited perceptual tempo ratings on a continuous scale using a visual interface similar to that of Pfitzinger and Tamashima (2006), implemented in *PsychoPy2* (Peirce, 2009) and illustrated in Figure 3. The stimuli in each set of 60 were presented together on one (wide-screen, rotated) computer screen in the form of a vertical line of coloured dots in the centre of the screen. When the participant clicked on a dot, an orthographic transcription of the stimulus appeared at the top of the screen and the corresponding audio played over headphones (JVC HA-RX500-E). Like Koreman (2006) we displayed an orthographic transcription to ensure that participants understood lexical content, particularly when deletion rates were high. After listening to the audio, the participant's task was to left-click on the dot and move it along a horizontal guide line to a position that reflected its tempo. The position was recorded through an invisible 1000-point grid. Vertical lines and the labels 'slowest, slower, average, faster, fastest' aided orientation. We did not present selected stimuli as 'anchor points' (Pfitzinger & Tamashima, 2006): Dellwo, Ferrange, and Pellegrino (2006) show that 'listeners have a fairly good idea of what a normal, fast or slow speech rate is' without such guidance. Stimuli were arranged in the same random order for all participants. The randomization left no notable correlations between our crucial rates and screen position ($r<|0.35|$ across rates and screens). Participants could listen to stimuli repeatedly as they worked through each set of 60 stimuli and were encouraged to take a short break after the first and second sets. Participants were tested individually in a quiet laboratory room.

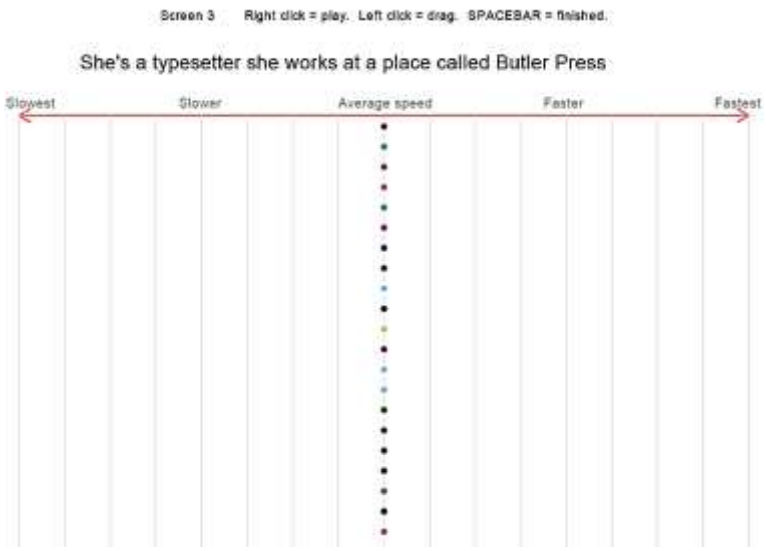


Figure 3: Partial visual interface for eliciting perceptual tempo ratings (see text for description)

Production tasks Like Koreman (2006), we also elicited production data from our participants. Koreman did not find support for his hypothesis that listeners' tempo

judgements are systematically related to their own production habits, but other studies have revealed some evidence for a systematic link (Gósy, 1992; Schwab, 2011), such that relatively slow speakers rate others' speech as faster than relatively fast speakers do. Participants completed several production tasks prior to the tempo rating task, all audio-recorded into a PC at a sampling rate of 44.1kHz using an AudioTechnica AT2020 microphone. First (cf. Alexandrou, Saarinen, Kujala, & Salmelin, 2016; Ruspantini, et al., 2012), participants were instructed to repeat /pa/ at what seemed to them a normal, comfortable rate, for 10 seconds between visual start and stop signals. We counted the number of /pa/ realizations in the 10-second window to yield a '/pa/ rate' per participant. Second (cf. Jacewicz, Fox, & Wei, 2010; Jungers, Palmer, & Speer, 2002; Schultz, et al., 2016), participants saw five sentences from the *Rainbow passage* (see Cartwright & Lass, 1975), one at a time. They had to memorize each sentence, then tap the space bar to reveal a blank screen and produce the sentence (see Dilley & Pitt, 2010). They were encouraged to focus on repeating the sentence verbatim and to correct any erroneous or disfluent productions. No instruction as to tempo or clarity was given. We delimited the starts and ends of the sentence productions using *Praat*, calculated canonical syllable rates, and averaged these across the five sentences to yield a 'production rate' by participant. Third (cf. Collyer, Broadbent, & Church, 1994; Palmer, Lidji, & Peretz, 2014), participants tapped the index finger of their dominant hand on a laptop touchpad at what seemed to them a normal, comfortable rate, for 20 seconds between visual start and stop signs. The laptop recorded the timestamps of the taps using a Python script implemented in *PsychoPy2* (Peirce, 2009). We counted the taps within the 20-second window to yield a 'tap rate' by participant.

Quantitative analysis

We fitted linear mixed effects models using *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and *emmeans* (Lenth, 2022) in *R* (R Development Core Team, 2008). We first modelled ratings across the full stimulus set to assess the general impact of our independent variables on listeners' tempo perceptions. We then addressed the question of which of our alternative articulation rate measures best predicted listeners' ratings in four tailored data subsets.

Dependent variable Tempo ratings were recorded on a numerical scale from 0 to 1000, with 500 representing the initial central placement of the stimuli in the visual interface. As participants were encouraged by the initial placement to work 'from the centre', ratings were reasonably symmetrically distributed around a median of 521. We therefore decided not to log-transform ratings, although listeners' tempo ratings in magnitude estimation tasks have been shown to follow Stevens' (1975) power function law (Cartwright & Lass, 1975; Grosjean, 1977; Grosjean & Lass, 1977; Schwab, 2011).

Predictor variables The crucial predictor variables were our four articulation rate measures: canonical syllable rate, surface syllable rate, canonical phone rate and surface phone rate. We log-transformed these prior to modelling. While we carefully controlled the relationships among these rates in stimulus subsets, across all stimuli the measures remain strongly inter-correlated (r=0.70–0.84 across pairwise comparisons). This confirms that modelling across all stimuli will not allow us to draw firm conclusions about which of these measures maps most closely to tempo ratings: modelling ratings of smaller subsets of stimuli is necessary for this (correlations reported below).

Random variables All of the models presented below contain random intercepts for participant and speaker identities (see Baayen, 2008, p. 241). There were 55 participants, each contributing 180 responses. The 30 speakers selected from the *DyViS* corpus contributed on average 6 stimuli each (SD=2.4, range=2–11); thus there were on average 330 responses per speaker (range=110–605). We report random intercept models only, as most models with random slopes failed to converge. We did not include stimulus identity as a randomeffect as our participants judged each stimulus only once.

Additional variables Before modelling tempo ratings we checked for collinearity (Tomaschek, Hendrix, & Baayen, 2018) in the relationships among the additional variables derived from our production tasks, acoustic analysis, and articulation rate variables. Participants' canonical syllable rates ranged from 4.0 to 6.8 syll/s, in line with the ranges observed by Tauroza and Allison (1990), Robb, et al. (2004) and Jacewicz, et al. (2010). The /pa/ and finger tapping rate ranges were lower, averaging 1.7 and 2.3 per second, respectively, in line with previous studies (Collyer, et al., 1994; Lidji, Palmer, Peretz, & Morningstar, 2011; Palmer, et al., 2014; Ruspantini, et al., 2012). Syllable rates were not correlated with /pa/ rates (r=0.22) or tapping rates (r=0.02), and /pa/ and tapping rates were only moderately correlated with each other (r=0.53). We therefore treated the three measures as independent in our analysis and remain agnostic as to how close any of them comes to reflecting participants' 'normal' speech tempo.

The memory stretches in our stimulus set ranged in duration between 0.5s and 2.7s (mean 1.5s). Their distribution was left-skewed; only 24 stretches (13%) had a duration above 2s. 77 stretches (43%) were in the duration range 1–1.5s used by Koreman (2006). The length of the stretches was on average 7.5 words (SD=2.7), 9.8 canonical syllables (SD=3.3), 9.0 surface syllables (SD=3.1). Stretch duration was strongly correlated with the number of canonical syllables in the stretch (r=0.86), but not notably correlated with any of our four articulation rate measures (r=|0.18|–|0.27|), or with the proportion of deleted syllables in the stretch (r=–0.065). We therefore included (log-transformed) stretch duration among our additional variables. We found no notable correlation between *f*0 mean and kurtosis values (r<0.20), between either of these and mean intensity (r<|0.10|), or among the *f*0, intensity and articulation rate measures (r<|0.25|).

Analysis data sets We took stimulus sets 1, 2 and 3 described above (see Figure 2 for set 1) as a starting point in constructing four tailored data sets, each of which allows us to compare canonical and surface rates in terms of their mapping to tempo ratings. Table 1 lists the main characteristics of the data sets (Sets A to D), with summary statistics.

Sets A and B allow us to assess how listeners respond to syllable deletions, which yield divergence between canonical and surface syllable rates. Set A stimuli have narrow canonical syllable rate ranges and wider surface syllable rate ranges: variation in syllable deletion occurrence is associated with variable lowering of surface syllable rate relative to the canonical rate. Set B stimuli have narrow surface syllable rate ranges and substantially wider canonical syllable rate; here, variation in syllable deletion occurrence produces similar surface syllable rates. Sets C and D similarly allow us to assess how listeners respond to phone deletions. Set C stimuli have narrow canonical phone rate ranges and wider surface phone rate ranges; Set D stimuli have narrow surface phone rate ranges and wider canonical phone rate ranges.

Our stimulus selection procedure resulted in some overlap between the smallest stimulus subsets within and across sets 1, 2 and 3. Some of this overlap is visible in Figure 2: for example, several stimuli that we had included to populate the 10%–20% quantile canonical syllable rate subset (circles) also fit in the 10%–20% quantile surface syllable rate subset (triangles). Likewise, some of the set 3 stimuli, which we had included to facilitate a comparison on which we do not report in this paper, fit in the quantile ranges that we distinguished in sets 1 and 2. To maximise statistical power, we constructed the data sets in Table 1 using as many as possible of the stimuli in our total stimulus set (N=180) that fell within the boundaries of the 10%–20%, 45%–55% and 80%–90% quantile ranges for each 'stable' rate. This amounted to expanding the size of each smallest stimulus subset ('Low', 'Mid' and 'High'), where possible, beyond the 10 stimuli that we had included in our design to populate that subset.

Set A

'stable' rate	canonical syllable rate		
'variable' rate	surface syllable rate		
quantile range subset	'Low'	'Mid'	'High'
stable rate range (raw)	0.20 syll/s	0.28 syll/s	0.26 syll/s
variable rate range (raw)	1.20 syll/s	2.33 syll/s	1.87 syll/s
N stimuli	10	14	11
N ratings	550	770	605

Set B

'stable' rate	surface syllable rate		
'variable' rate	canonical syllable rate		
quantile range subset	'Low'	'Mid'	'High'
stable rate range (raw)	0.15 syll/s	0.19 syll/s	0.50 syll/s
variable rate range (raw)	1.43 syll/s	1.93 syll/s	3.51 syll/s
N stimuli	11	12	15
N ratings	605	660	825

Set C

'stable' rate	canonical phone rate		
'variable' rate	surface phone rate		
quantile range subset	'Low'	'Mid'	'High'
stable rate range (raw)	0.72 phon/s	0.65 phon/s	0.71 phon/s
variable rate range (raw)	2.60 phon/s	3.22 phon/s	4.87 phon/s
N stimuli	15	16	13
N ratings	825	880	715

Set D

'stable' rate	surface phone rate		
'variable' rate	canonical phone rate		
quantile range subset	'Low'	'Mid'	'High'
stable rate range (raw)	0.59 phon/s	0.47 phon/s	0.66 phon/s
variable rate range (raw)	3.36 phon/s	4.84 phon/s	4.21 phon/s
N stimuli	22	21	23
N ratings	1210	1155	1265

Table 1. Summary characteristics of analysis data sets A, B, C and D

Table 1 shows that within each of the 'Low', 'Mid' and 'High' subsets, the 'variable' rate range is always at least 3.6 times that of the 'stable' rate, and on average 7.1 times. Within each subset, the 'stable' and 'variable' rates are correlated at less than r=0.30; this means that we can safely treat the two rate variables as independent. Figure 4 shows the stimuli on scatter plots. In each plot, the canonical rate is on the *x*-axis, and the surface rate on the *y*-axis; the 'stable' rate is the canonical rate for panels A and C, and the surface rate for panels B and D. The dotted diagonal represents equivalence between the 'stable' and 'variable' rates. Data points clearly fall into three subsets ('Low', 'Mid', 'High') with respect to the 'stable rate'. The positioning of the data points with respect to the diagonal reflects that where canonical rate is the 'stable' rate (Sets A and C), deletion is associated with a relative decrease in surface rate; where surface rate is the 'stable' rate, (Sets B and D), deletion is associated with a relative increase in canonical rate.

We fitted a model for each of Sets A (35 stimuli, 1925 ratings), B (38 stimuli, 2090 ratings), C (44 stimuli, 2420 ratings) and D (66 stimuli, 3630 ratings). For each set, we included stimuli from all three subsets: 'Low', 'Mid' and 'High'. Of course the 'stable' rate is only close to stable *within* these subsets, and the 'variable' rate varies systematically across them. To ensure that our 'stable' and 'variable' rate measures remained independent even when using all three subsets, we centred and standardized the 'variable' rate measures within the 'Low', 'Mid' and 'High' subsets. The z-score transformation removes all variation in the 'variable' rate that correlates with the observed variation in the 'stable' rate between subsets, leaving only the variation that is the result of the variable occurrence of deletion.

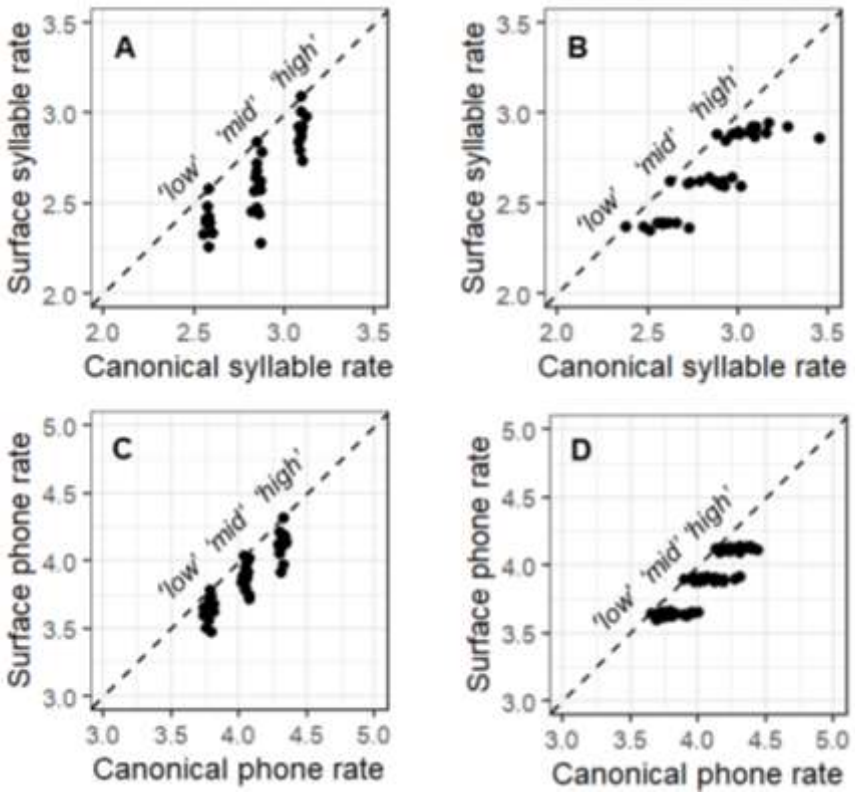


Figure 4. Scatterplots for Sets A, B, C and D, each with the canonical rate (log values) on the *x*-axis and the surface rate (log values) on the *y*-axis; each data point represents one stimulus.

The 'stable' rate is the canonical rate (*x*-axis) for Sets A and C, and the surface rate (*y*-axis) for Sets B and D. See text for details.

In modelling tempo ratings for each of Sets A, B, C and D, we first fitted a control model which contained a three-level factor for stimulus subset ('Low', 'Mid', 'High'). We predicted that 'Low' stimuli should sound slower than 'Mid' ones, and 'High' stimuli should sound faster. Adding the z-scored 'variable' rate measure then allowed us to assess whether ratings were also systematic in relation to rate variation captured only by the 'variable' rate measure, and attributable directly to syllable or phone deletions. Inspection of the relationship between the 'variable rate' and participants' ratings within the 'Low', 'Mid' and 'High' subsets further allowed us to assess the evidence for listeners responding differently to deletions in slow and fast speech.

Results

Modelling all ratings

We started with a base model containing random intercepts for participant and speaker identities, then added each of the independent variables in Table 2 (centered to its mean) in turn, using the *anova* function to assess whether the addition resulted in a significant improvement of model fit. The single variable which yielded the greatest significant improvement of fit, as reflected in AIC values, was added to the model. We repeated this procedure with the remaining independent variables, until model fit could not be significantly improved. Variables that yielded non-significant fixed effects in the model were removed, and the model refitted. Finally, we assessed whether interactions between the independent variables improved the fit of the model. This was not the case.

Production measures (by participant)	*Syllable rate, /pa/ rate, Tap rate*
Stimulus placement	*Screen position*
Stimulus duration	*Duration*
f0 and intensity measures	*f0 mean, f0 kurtosis, Intensity mean*
Articulation rates	*Canonical syllable rate, Surface syllable rate, Canonical phone rate, Surface phone rate*

Table 2. Independent variables used in modelling tempo ratings across the complete stimulus set

Table 3 summarizes and Figure 5 visualizes the fixed effects of the optimal model of ratings. None of the production measures predicted participants' ratings, nor did *Screen position*. However, our phonetic parameters—stimulus duration, *f*0, intensity and articulation rate—are all informative. Longer stimuli were rated as slower (*Duration*). Stimuli with higher *Intensity mean* and *f0 mean* were rated as faster, consistent with Feldstein and Bond (1981), Kohler (1986) and Rietveld and Gussenhoven (1987). The positive effect of articulation rate confirms that participants were sensitive to our crucial experimental manipulation, and the model shows that *Surface phone rate* resulted in the best overall model fit.

	Estimate	SE	df	t	p
(Intercept)	522.380	9.722	39.643	53.730	<0.001
Duration	–43.505	2.856	9789.293	–15.229	<0.001
Intensity mean	7.071	0.650	3215.656	10.867	<0.001
f0 mean	0.486	0.150	3258.982	3.226	<0.001
Surface phone rate	204.925	8.724	9805.217	23.490	<0.001

Table 3. Summary of fixed effects in the optimal model of tempo ratings across the complete stimulus set

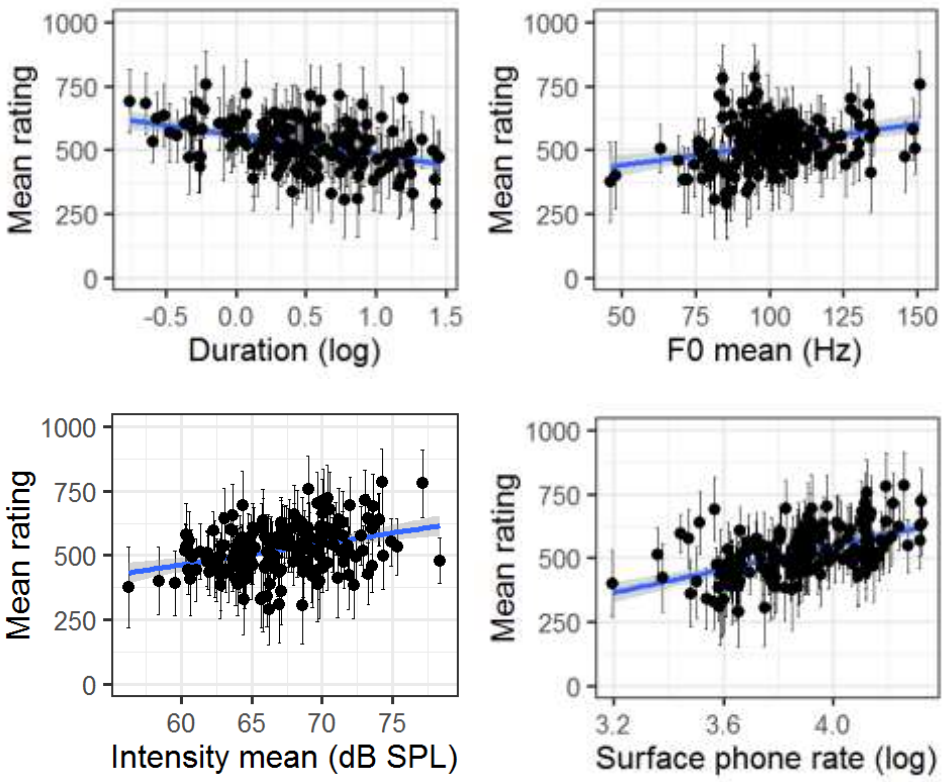


Figure 5. Scatterplots showing the significant independent variables in the model in Table 3 (*x*-axes) against tempo ratings (*y*-axes, with 500 representing 'average speed'), with linear fit lines. In each plot, data points represent mean ratings, with whiskers showing standard deviations.

Modelling ratings in stimulus sets A, B, C and D

For each of sets A, B, C and D, we first added the factor *Subset* with levels 'Low', 'Mid' and 'High' to the base model; as predicted, this improved model fit in all cases. We then added the independent variables that featured in the model Table 3 and verified that the independent variables listed in Table 2 which did not feature in the model in Table 3—the production measures by participant, *Screen position* and *f0 kurtosis*—did not improve model fit. We considered the resulting model our 'control model'. We then added the relevant z-scored 'variable' rate measure, and its interaction with *Subset*, to assess the relationship between the

'variable' rate and listeners' tempo ratings within the three stimulus subsets. The variables are summarized in Table 4.

Stimulus duration	*Duration*
*f*0 and intensity measures	*f0 mean*, *Intensity mean*
'Stable' rate quantile range	*Subset* ('Low', 'Mid', 'High')
Articulation rates (z-scored by *Subset* level)	*Canonical syllable rate*, *Surface syllable rate*, *Canonical phone rate*, *Surface phone rate*

Table 4. Independent variables used in modelling tempo ratings in stimulus sets A, B, C and D

Set A Here the 'stable' rate is canonical syllable rate; the 'variable' rate is surface syllable rate. More syllable deletions yield lower surface syllable rate values. Our control model contains significant fixed effects for *Duration*, *f0 mean* and *Subset*. Adding (z-scored) *Surface syllable rate* and its interaction with *Subset* improves fit; the resulting model is summarised in Table 5. Table 6 and Figure 6 show estimates of the effect of *Surface syllable rate* for each level of *Subset.* Table 6 also shows the pairwise differences in the effect of *Surface syllable rate* across the three levels of *Subset* (obtained using the *emtrends* function). *Surface syllable rate* has a positive effect in the 'High' subset, a negative effect in the 'Low' subset, and no effect in the 'Mid' subset. The effect of *Surface syllable rate* for each subset differs significantly from that for the other two. As Figure 6 shows, among the relatively fast stretches of the 'High' subset, those with fewer deletions (and therefore higher surface syllable rates) are rated as faster than stretches with more deletions (and lower surface syllable rates), whereas for the relatively slow stretches of the 'Low' subset, those with more deletions are rated as faster than those with fewer deletions.

	Estimate	SE	df	t	p
(Intercept)	529.048	20.338	18.106	26.012	<0.0001
Duration	–58.724	9.435	1460.151	–6.224	<0.0001
f0 mean	1.951	0.404	948.137	4.826	<0.0001
Surface syllable rate	10.219	3.738	1549.098	2.734	0.006
Subset ('Low') *vs* mean	38.123	6.361	1247.862	5.994	<0.0001
Subset ('High') *vs* mean	–3.168	5.327	1391.945	–0.595	0.552
Surface syllable rate: Subset ('Low') *vs* mean	–40.051	7.651	1166.224	–5.235	<0.0001
Surface syllable rate: Subset ('High') *vs* mean	45.236	6.902	1406.351	6.555	<0.0001

Table 5. Model of tempo ratings across Set A stimuli; *Subset* was sum coded with contrasts shown for 'Low' *vs* mean and 'High' *vs* mean.

(a)	Estimate	SE	df	Lower limit of CI	Upper limit of CI
'Low'	–29.83	9.79	1109	–49.05	–10.6
'Mid'	5.03	4.38	1863	–3.56	13.6
'High'	55.45	7.34	1706	41.07	69.8
(b)	Estimate	SE	df	t	p
'Low' *vs* 'High'	–85.3	14.03	1247	–6.078	<0.0001
'Low' *vs* 'Mid'	–34.9	10.40	1283	–3.352	0.0024
'High' *vs* 'Mid'	50.4	8.62	1751	5.846	<0.0001

Table 6. Estimated slope for *Surface syllable rate* for each level of *Subset* for Set A stimuli (a), and pairwise differences in slope for *Surface syllable rate* between levels of *Subset* (b).

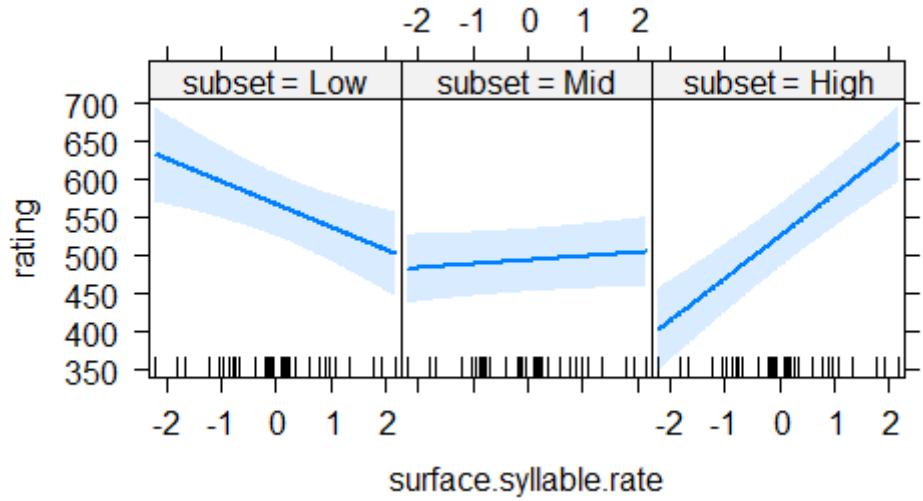


Figure 6. Estimated effect of (z-scored) *Surface syllable rate* (*x*-axis) on tempo ratings (*y*-axis) for the three levels of *Subset* for Set A stimuli.

Set B stimuli Here the 'stable' rate is surface syllable rate; the 'variable' rate is canonical syllable rate. More syllable deletions lead to higher canonical syllable rate values. Our control model contains significant fixed effects for *Duration*, *f0 mean* and *Subset*. Adding (z-scored) *Canonical syllable rate* and its interaction with *Subset* improves fit, but removes the effect of *f0 mean*. The resulting model is summarised in Table 7 and the effect of *Canonical syllable rate* across the three levels of *Subset* is shown in Table 8. *Canonical syllable rate* has a positive effect in the 'High' subset, a negative effect in the Mid subset, and no effect in the 'Low' subset. The effect of *Canonical syllable rate* for each subset differs significantly from that for the other two. As Figure 7 shows, among the relatively fast stretches of the 'High' subset, those with more deletions (and therefore higher canonical syllable rates) are rated as faster than stretches with fewer deletions, whereas for the slower stretches of the 'Mid' subset, those with fewer deletions (and therefore lower surface syllable rates) are rated as faster than stretches with more deletions. This would seem the opposite pattern of that observed in the Set A stimuli; we will return to this below.

	Estimate	SE	df	t	p
(Intercept)	548.893	22.537	18.101	24.355	<0.0001
Duration	–50.499	9.056	1967.077	–5.577	<0.0001
Canonical syllable rate	–6.344	3.957	1692.605	–1.603	0.109
Subset ('Low') *vs* mean	–47.562	5.107	2014.015	–9.479	<0.0001
Subset ('High') *vs* mean	27.421	5.913	1568.349	4.637	<0.0001
Canonical syllable rate: Subset ('Low') *vs* mean	1.472	4.781	1877.186	0.308	0.758
Canonical syllable rate: Subset ('High') *vs* mean	41.944	4.981	1649.234	8.421	<0.0001

Table 7. Model of tempo ratings across Set B stimuli; *Subset* was sum coded with contrasts shown for 'Low' *vs* mean and 'High' *vs* mean.

(a)	Estimate	SE	df	Lower limit of CI	Upper limit of CI
'Low'	–4.87	5.30	2024	–15.3	5.52
'Mid'	–49.76	9.62	1174	–68.6	–30.89
'High'	35.60	4.72	2029	26.3	44.85
(b)	Estimate	SE	df	t	p
'Low' *vs* 'High'	–40.5	6.99	2027	–5.787	<0.0001
'Low' *vs* 'Mid'	44.9	10.75	1403	4.177	0.0001
'High' *vs* 'Mid'	85.4	11.03	1275	7.739	<0.0001

Table 8. Estimated slope for *Canonical syllable rate* for each level of *Subset* for Set B stimuli (a), and pairwise differences in slope for *Canonical syllable rate* between levels of *Subset* (b).

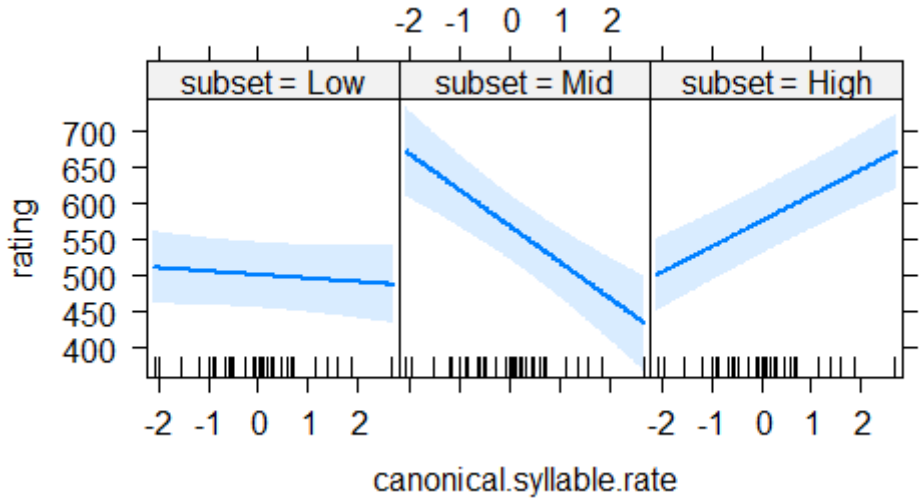


Figure 7. Estimated effect of (z-scored) *Canonical syllable rate* (*x*-axis) on tempo ratings (*y*-axis) for the three levels of *Subset* for Set B stimuli.

Set C stimuli Here the 'stable' rate is canonical phone rate and the 'variable' rate is surface phone rate. More phone deletion yields lower surface phone rate values. Our control model contains significant effects for *Duration*, *Intensity mean* and *Subset*. Adding (z-scored) *Surface phone rate* and its interaction with *Subset* improves fit. The resulting model is summarised in Table 9 and the effect of *Surface phone rate* across the three levels of *Subset* is shown in Table 10. *Surface phone rate* has a positive effect in the 'Low' and 'High' subsets and a negative effect in the 'Mid' subset. The effect for 'Mid' differs significantly from the other two, which do not differ from each other. As Figure 8 shows, among the relatively slow and fast stretches of the 'Low' and 'High' subsets, those with fewer deletions (and therefore higher surface phone rates) are rated as faster than stretches with more deletions. Among the mid-tempo stretches of the 'Mid' subset, those with more deletions (and therefore lower surface rates) are rated as faster than stretches with fewer deletions.

	Estimate	SE	df	t	p
(Intercept)	510.562	16.104	24.487	31.705	<0.0001
Duration	–116.926	11.021	1492.875	–10.609	<0.0001
Intensity	2.873	1.347	1032.409	2.132	0.033
Surface phone rate	9.701	3.467	2094.654	2.798	0.005
Subset ('Low') *vs* mean	–17.650	5.583	1989.635	–3.161	0.002
Subset ('High') *vs* mean	37.395	5.177	2126.537	7.223	<0.0001
Surface phone rate: Subset ('Low') *vs* mean	18.829	4.705	2057.606	4.002	<0.0001
Surface phone rate: Subset ('High') *vs* mean	5.554	4.852	2057.806	1.145	0.252

Table 9. Model of tempo ratings across Set C stimuli; *Subset* was sum coded with contrasts shown for 'Low' *vs* mean, and 'High' *vs* mean.

(a)	Estimate	SE	df	Lower limit of CI	Upper limit of CI
'Low'	28.5	5.33	2330	18.08	38.98
'Mid'	–14.7	7.28	1202	–28.97	–0.39
'High'	15.3	5.79	2338	3.89	26.62
(b)	Estimate	SE	df	t	p
'Low' *vs* 'High'	13.3	7.70	2337	1.723	0.197
'Low' *vs* 'Mid'	43.2	9.29	1507	4.653	<0.0001
'High' *vs* 'Mid'	29.9	9.51	1534	3.147	0.005

Table 10. Estimated slope for *Surface phone rate* for each level of *Subset* for Set C stimuli (a), and pairwise differences in slope for *Surface phone rate* between levels of *Subset* (b).

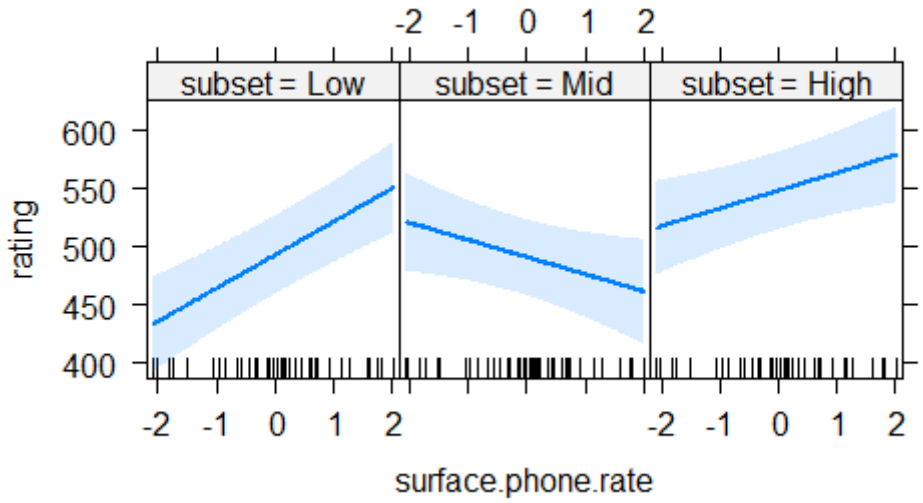


Figure 8. Estimated effect of (z-scored) *Surface phone rate* (*x*-axis) on tempo ratings (*y*-axis) for the three levels of *Subset* for Set C stimuli.

Set D stimuli Here the 'stable' rate is surface phone rate and the 'variable' rate is canonical phone rate. More deletion means higher canonical phone rate values. Our control model contains significant main effects for *Duration*, *f0 mean* and *Subset*. Adding (z-scored) *Canonical phone rate* does not significantly improve fit, either as a main effect or in interaction with *Subset*. Therefore, the optimal model is our control model, summarized in Table 11. *Subset* has the expected effect, and *Canonical phone rate* variation within these subsets does not predict ratings further.

	Estimate	SE	df	t	p
(Intercept)	515.704	11.868	32.906	43.454	<0.0001
Duration	–24.770	5.756	3281.431	–4.303	<0.0001
f0 mean	1.020	0.213	2339.925	4.787	<0.0001
Subset ('Low') *vs* mean	–46.225	4.740	2369.508	–9.752	<0.0001
Subset ('High') *vs* mean	37.210	3.790	3027.344	9.817	<0.0001

Table 11. Model of tempo ratings across Set D stimuli; *Subset* was sum coded with contrasts shown for 'Low' *vs* mean, and 'High' *vs* mean.

Summary

Analysis of the full dataset revealed a negative effect of *Duration* and positive effects of *f0 mean*, *Intensity* and articulation rate, in line with previous studies. Modelling within Sets A, B, C and D revealed a more complex picture of the roles of canonical and surface rates. For syllable rates, the participants' ratings of Set A stimuli were partly predicted by the surface rate variation in these stimuli. At the same time, ratings of Set B stimuli were partly predicted by the canonical rate variation. Clearly, therefore, participants were not consistently orienting only to canonical syllable rate or only to surface syllable rate: they oriented to both. Moreover, in both sets of stimuli, the direction of the effect varies with general speaking tempo: different effects were observed across the 'Low', 'Mid' and 'High' stimulus subsets.

The picture was different for phone rates. While ratings of Set C stimuli were partly predicted by surface phone rate variation, ratings of Set D stimuli were not systematically related to canonical rate variation. The predictive power of surface phone rate variation was observed for stimuli at all tempi. Thus there is evidence for listeners' orientation to surface phone rates in making tempo judgements, and no evidence for their orientation to canonical phone rates. Again, different effects of surface phone rate were observed across the 'Low', 'Mid' and 'High' stimulus subsets.

Discussion

The purpose of this study was to further our understanding of the relationship between measured articulation rates and perceived speech tempo, and the impact of syllable and phone deletions on speech tempo perception. We followed the work of Koreman (2006) in using stimuli from a corpus of unscripted speech, and in sampling stimuli in distinct 'global tempo' ranges. Within our stimulus sets, the difference between canonical and surface rate measurements was directly due to deletions. Results showed that listeners used both canonical and surface syllable rates to rate tempo: the effects in the models for Sets A and B were very similar, even though in Set A stimuli, surface syllable rate varied while in Set B stimuli, canonical syllable rate varied. Surface phone rate also influenced judgements (Set C), but canonical phone rate did not appear to (Set D). Our data also confirmed previously-reported influences of *f*0 and intensity, plus a new influence of duration—but no detectable influence of participants' own production tendencies.

Findings relating to our crucial variables

Our results resemble those of Koreman (2006) in that they suggest that in judging tempo, listeners do not consistently attend to some particular temporal parameter best captured by one rate measurement method. They also suggest that listeners' observation of phone and syllable deletions does not have a consistent effect on their tempo judgements. This is in line with Koreman (2006) but not Reinisch (2016), whose results from an implicit tempo judgement task suggested that deletions consistently make speech sound faster, and whose results from an explicit task suggested listeners did not attend to deletions. Like Koreman (2006) we found multiple effects of our manipulations—but different effects in different subsets of stimuli. It is worth emphasizing that these subsets—Sets A, B, C and D, and the 'Low', 'Mid' and 'High' subsets within them—were not presented as such to participants, so there was no *a priori* motivation for participants to respond differently between them. In what follows, we draw the findings summarized above together, and suggest directions for future work.

Our findings for syllable rate mirror those of Koreman (2006) quite closely, though he did not distinguish syllable from phone rate. He suggested that listeners orient to both canonical and surface rates and are good at identifying similarities and differences between utterances along both parameters. Where multiple utterances are similar along one parameter but different along the other, listeners' judgements are guided by the difference. Koreman hypothesized that in mid-tempo and fast speech, deletions would raise perceived tempo, as increased deletion rates are consistent with increased hypo-articulation; in slow speech, by contrast, deletions might make speech sound 'slurred', lowering perceived tempo. Our results

for Set A and Set B stimuli do indeed suggest that listeners respond differently to deletions—or to the relationship between canonical and surface articulation rates—in different general tempo ranges. The results for Set B stimuli seem consistent with Koreman's hypothesis in that for stimuli in the 'High' subset, those with more deletions are associated with higher tempo ratings. However, no significant effect is observed in the 'Low' subset, while more deletions are associated with lower tempo ratings in the 'Mid' subset. Moreover, the ratings for the Set A stimuli are inconsistent with Koreman's hypothesis: here relatively fast stretches with fewer syllable deletions are rated as faster than stretches with more deletions, whereas relatively slow stretches with more deletions are rated as faster than those with fewer deletions.

The data patterns observed for Set A and Set B stimuli is hard to interpret as compatible. It seems plausible that at higher speaking rates, the *absence* of deletions might raise perceived tempo: compared with fast hypo-articulation, fast hyper-articulation requires greater average articulatory velocity and results in a spectrally more complex signal. This reasoning is supported by the finding that more peripheral vowel productions make speech sound faster when articulation rate is controlled (Weirich & Simpson, 2014). However, our results appear to provide equal support for two opposing hypotheses regarding the effect of deletions on the perception of tempo in relatively fast speech. One observation we can make on our dataset is that as a result of our sampling method, the canonical syllable rates for the Set B 'High' subset include, at the top end, rates that are considerably higher than those of the Set A 'High' subset (see Figure 4). It is possible that listeners' interpretations of what constitutes relatively slow, mid-tempo and relatively fast speech is more complex than we assume, and this complexity is one source of the complexity we observe in the effects of the 'variable rates'.

Our findings for phone rate diverge somewhat from those of Koreman (2006). Set C and Set D provide clear evidence that listeners attend to surface phone rate variation—but no direct evidence for listeners attending to variation in canonical phone rate. The effect of surface phone rate variation was observed for relatively fast as well as relatively slow stimuli—and it also yielded the best mapping to tempo ratings across the entire stimulus set. The different pictures produced by our syllable rate and phone rate comparisons may be reconcilable. While the relationship between syllable and phone deletions in our corpus is linear, each observed number of syllable deletions maps to a considerable range of phone deletions. This is to be expected: while syllable deletions entail phone deletions, phone deletions do not necessarily contribute to syllable deletions—as illustrated by zero syllable deletion mapping to up to 6 phone deletions. Perhaps listeners orient to a canonical syllable string when estimating tempo, based on their understanding of the lexical content of the incoming speech signal—but not necessarily a fully elaborated canonical phone string. This would mean that listeners effectively ignore phone deletions, taking the surface phone string 'at face value', when phone deletions do not result in the deletion of entire syllables. Testing this hypothesis requires stimuli in which numbers of phone and syllable deletions are carefully controlled, so that multiple stimuli have the same numbers of phone deletions but different numbers of syllable deletions. Unfortunately our stimulus set does not lend itself to rigorous testing of this kind.

Findings relating to our additional variables

Moving on to the findings gleaned from our additional variables, like Koreman (2006) we found no evidence for our participants' perceptions of speech tempo being affected by their own production tendencies. As noted above, Schwab (2011, p. 253) does report such effects, on the basis of an experiment in which 28 participants each read the same passage at slow, normal and fast rates, and then rated the total set of passages for tempo through a magnitude estimation task. Schwab (2011, p. 253) reports some variation across participants and tempi, and a relatively weak overall effect. Perhaps that a similar effect failed to emerge in our design because our listeners, unlike Schwab's, judged different phrases from the ones they had produced. Interestingly, Schwab (2011, p. 253) notes that the perception of intended-as-fast passages seemed less constrained by listeners' production tendencies t intended-as-slow and intended-as-normal passages—and wonders whether other factors, including deletions, might be more salient in fast speech. Fast speech is exactly where we found most impact of deletions on our participants' tempo ratings. Finally, it is probably wise not to assume that participants' performance in a short speech production task closely reflects their 'habitual' speaking tempo. Our finding that rates for unguided production, /pa/ repetition and finger tapping were not closely correlated with each other warrants further research into the validity and reliability of these control tasks.

With reference to our temporal and acoustic variables, the positive effects of *Intensity mean* and *f0 mean* are consistent with Feldstein and Bond (1981), Kohler (1986) and Rietveld and Gussenhoven (1987). What are the reasons for rating a louder or higher-pitched speaker as speaking at a higher tempo? Production studies show covariations between speech rate, intensity and f0 (overall level, and also details of the intonation contour; Black 1961; Kohler 1983). One account therefore is simply that listeners know about these covariations, expect them and infer their presence. An alternative possibility, put forward by Feldstein & Bond (1981) is methodological, i.e. that listeners attribute variation on irrelevant stimulus dimensions to the only characteristic the experiment allows them to judge, in this case tempo. Kohler (1986) showed that phonetic expertise reduces listeners' tendency to use f0 in tempo judgements, but does not eliminate it altogether: this suggests that elements of both accounts are in play.

A new finding was a negative effect of stimulus duration on perceived tempo, such that longer stimuli tended to be rated slower. This was significant in all subsets of the data. Plug and Smith (2021) observed a similar effect in a tempo discrimination experiment where participants compared phrases with the same syllable rates but different phone rates and phrase durations. It is arguably not surprising that stimuli that are completed relatively quickly are perceived as relatively fast, while stimuli that take longer to complete are perceived as slower. However, further research is needed to establish the robustness and precise nature of this effect. We noted above that in the current experiment, stimulus durations were strongly correlated with word and canonical syllable numbers. Therefore, the observed effect might in principle be due to differences between stimuli in speed of processing—although previous research suggests that increasing cognitive load has the effect of making speech sound faster (Bosker, Reinisch, & Sjerps, 2017). In any case, our finding suggests that stimulus duration is an informative parameter in tempo perception, and must therefore be controlled in the experimental design.

Conclusion

We have shown that both canonical and surface syllable rates, and surface phone rate, combine with a number of other temporal and non-temporal acoustic parameters to influence English listeners' perception of tempo in short stretches of unscripted speech. Interestingly, Pfitzinger (1999) has proposed that perceived speech tempo in German is best approximated by an equation combining measured syllable and phone rates (see Mixdorff & Pfitzinger, 2005; Pfitzinger & Tamashima, 2006). His articulation rate measures were surface ones only; our results, which ongoing work seeks to disentangle using highly controlled lab speech (Plug et al., in preparation), suggest that the best approximation might be derived from a combination of canonical syllable rate and surface phone rate.

References

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2016). A multimodal spectral approach to characterize rhythm in natural speech. *Journal of the Acoustical Society of America, 139*, 215-226.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barry, W., & Andreeva, B. (2001). Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association, 31*, 51-66.

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

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

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention Perception & Psychophysics, 79*, 333-343.

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language, 94*, 166-176.

Buller, D. B., Lepoire, B. A., Aune, R. K., & Eloy, S. V. (1992). Social Perceptions as Mediators of the Effect of Speech Rate Similarity on Compliance. *Human Communication Research, 19*, 286-311.

Cangemi, F. (2015). Mausmooth. In: http://phonetik.phil-fak.uni-koeln.de/fcangemi.html.

Cartwright, L. R., & Lass, N. J. (1975). Psychophysical Study of Rate of Continuous Speech Stimuli by Means of Direct Magnitude Estimation Scaling. *Language and Speech, 18*, 358-365.

Collyer, C. E., Broadbent, H. A., & Church, R. M. (1994). Preferred Rates of Repetitive Tapping and Categorical Time Production. *Perception & Psychophysics, 55*, 443-453.

Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics, 25*, 287-312.

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

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

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychol Sci, 21*, 1664-1670.
Feldstein, S., & Bond, R. N. (1981). Perception of speech rate as a function of vocal intensity and frequency. *Language and Speech, 24*, 387-394.
Feldstein, S., Dohm, F. A., & Crown, C. L. (2001). Gender and speech rate in the perception of competence and social attractiveness. *Journal of Social Psychology, 141*, 785-806.
Gibbon, D., Klessa, K., & Bachan, J. (2015). Duration and speed of speech events: A selection of methods. *Lingua Posnaniensis, 56*, 59-83.
Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters.* University of York, York.
Gósy, M. (1992). *Speech perception*. Frankfurt am Main: Hector.
Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication, 29*, 159-176.
Grosjean, F. (1977). Perception of rate in spoken and sign languages. *Perception & Psychophysics, 22*, 408-413.
Grosjean, F., & Lass, N. J. (1977). Some factors affecting listeners' perception of reading rate in English and French. *Language and Speech, 20*, 198-208.
Harnsberger, J. D., Shrivastav, R., Brown, W. S., Jr., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *J Voice, 22*, 58-69.
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*, 839-850.
Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science & Justice, 47*, 50-67.
Johnson, K. (2004). Massive reduction in conversational American English. In *Tenth International Symposium on Spontaneous Speech: Data and Analysis.* (pp. 29-54): Citeseer.
Jungers, M. K., Palmer, C., & Speer, S. R. (2002). Time after time: The coordinating influence of tempo in music and speech. *Cognitive Processing, 1-2*, 21-35.
Kisler, T., Reichel, U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45*, 326–347.
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.
Kohler, K. J. (2000). Investigating unscripted speech: Implications for phonetics and phonology. *Phonetica, 57*, 85-94.
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*, 582-596.
Lenth, R. V. (2022). Emmeans: Estimated marginal means, aka least-squares means. In (Vol. R package version 1.7.3.). https://CRAN.R-project.org/package=emmeans.
Lidji, P., Palmer, C., Peretz, I., & Morningstar, M. (2011). Listeners feel the beat: Entrainment to English and French speech rhythms. *Psychonomic Bulletin & Review, 18*, 1035-1041.
Mitterer, H. (2018). The singleton-geminate distinction can be rate dependent: Evidence from Maltese. *Laboratory Phonology, 9*, 1-16.
Mixdorff, H., & Pfitzinger, H. R. (2005). Analysing fundamental frequency contours and local speech rate in map task dialogs. *Speech Communication, 46*, 310-325.
Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics, 58*, 540-560.

Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics, 37*, 46-65.
Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law, 16*, 31-57.
Palmer, C., Lidji, P., & Peretz, I. (2014). Losing the beat: Deficits in temporal coordination. *Philosophical Transactions of the Royal Society B-Biological Sciences, 369*.
Peirce, J. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*.
Pfitzinger, H. (1999). Local speech rate perception in German speech. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 893-896). San Francisco.
Pfitzinger, H., & Tamashima, M. (2006). Comparing perceptual local speech rate of German and Japanese speech. In *Proceedings of the 3rd International Conference on Speech Prosody* (pp. 105-108). Dresden.
Plug, L., Lennon, R., & Smith, R. (2019). Measured and perceived speech tempo: Canonical vs surface syllable and phone rates. In *Nineteenth International Congress of Phonetic Sciences*. Melbourne.
Plug, L., & Smith, R. (2021). The role of segment rate in speech tempo perception by English listeners. *Journal of Phonetics, 86*, 101040.
Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics, 35*, 353-362.
R Development Core Team. (2008). R: A language and environment for statistical computing. In.
Reinisch, E. (2016). Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception,& Psychophysics, 9*, 9.
Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech, 54*, 147-165.
Rietveld, A. C. M., & Gussenhoven, C. (1987). Perceived speech rate and intonation. *Journal of Phonetics, 15*, 273-285.
Robb, M. P., Maclagan, M. A., & Chen, Y. (2004). Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics, 18*, 1-15.
Ruspantini, I., Saarinen, T., Belardinelli, P., Jalava, A., Parviainen, T., Kujala, J., & Salmelin, R. (2012). Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2-3 Hz. *Journal of Neuroscience, 32*, 3786-3790.
Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception & Psychophysics, 62*, 285-300.
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*, 1201-1220.
Schwab, S. (2011). Relationship between speech rate perceived and produced by the listener. *Phonetica, 68*, 243-255.
Shattuck-Hufnagel, S., & Veilleux, N. (2007). Robustness of acoustic landmarks in spontaneously-spoken American English. In *Sixteenth International Congress of Phonetic Sciences* (pp. 925-928): Saarland University Saarbrücken.
Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.
Street, R. L., & Brady, R. M. (1982). Speech Rate Acceptance Ranges as a Function of Evaluative Domain, Listener Speech Rate, and Communication Context. *Communication Monographs, 49*, 290-308.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics, 11*, 90-105.
Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71*, 249-267.
Vaane, E. (1982). Subjective estimation of speech rate. *Phonetica, 39,* 136-149.
Weirich, M., & Simpson, A. P. (2014). Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics, 43*, 1-10.