Speech Communication 132 (2021) 40–54

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Articulation rates' inter-correlations and discriminating powers in an English speech corpus

Leendert Plug [a,*], Robert Lennon [a], Erica Gold [b]

[a] University of Leeds, UK
[b] University of Huddersfield, UK

ARTICLE INFO

Keywords:
Articulation rate
Speaker comparison
Correlations

ABSTRACT

Studies that quantify speech tempo on acoustic grounds typically use one of various rate measures. The availability of multiple measurement techniques yields 'researcher degrees of freedom' which call the robustness of generalisations across studies into question. However, explicit assessments of the possible impact of researchers' choices amongst the available measures are rare. In this study we attempt such an assessment by comparing the distributions of five common rate measures–canonical and surface syllable and phone rates, and CV segment rate–calculated over fluent stretches of unscripted speech produced by 100 English speakers. We assess the measures' inter-correlations across the corpus as a whole as well as in relevant data samples to simulate multiple analysis scenarios. We also report on deletion rates in our corpus, as they determine the relationship between canonical and surface rates; we assess the impact on rate figures of variable assumptions as to what constitutes deletion; and we compare the measures' discriminating powers in a forensic analysis context using Bayesian likelihood ratios. Our results suggest that in a sizeable English corpus with normal deletion rates, the five rates are closely inter-correlated and have similar discriminating powers; decisions as to the segmental make-up of canonical forms also have limited impact on distributions. Therefore, for common analytical purposes and forensic applications the choice between these measures is unlikely to substantially affect outcomes.

1. Introduction

Studies that quantify speech tempo through signal-based measurements tend to use one of various available measurement techniques. Researchers choose what to count—words, syllables, phones, or derived units such as C and V segments (e.g. Dellwo et al., 2006; Pfitzinger, 1996) —and what temporal domains to count in—total speaking time including or excluding pauses, or stretches of speech such as clauses, intonation phrases, inter-pause stretches or memory stretches (e.g. Dankovičová, 1997; Jessen, 2007). When counting syllables or phones, researchers can count units as expected in canonical pronunciations, or as actually observed in their data (e.g. Koreman, 2006). These methodological choices are one example of 'researcher degrees of freedom' in phonetic and related research (Roettger, 2019; Simmons et al., 2011): the availability of multiple alternative methods for operationalizing a phonetic parameter brings with it the risk that researchers—intentionally or unintentionally—select the method that produces the clearest analysis results. In any case, the availability of multiple methods means that comparing findings across studies is not always straightforward (Jessen, 2007). Individual studies typically present the outputs of one technique only, and studies that do refer to multiple techniques do not necessarily report on correlations between their outputs. We are generally left to wonder, therefore, whether this particular methodological choice has been consequential for the reported data patterns.

One might argue that differences between the distributions of values yielded by alternative tempo quantification techniques are likely to be small and therefore do not need our attention. Still, empirical confirmation of this likelihood is better than the absence thereof, especially since speech tempo is quantified in various applied contexts, such as those of forensic casework (Gold and French, 2011; Jessen, 2007), speech therapy (Martens et al., 2015; Pellowski, 2010), mental health diagnosis (Cummins et al., 2015; Mundt et al., 2007) and language learner assessment (Bosker et al., 2013; Wang et al., 2018). In this paper we quantify speech tempo using four commonly used measures: syllable rate based on canonical unit counts ('canonical syllable rate') and counts of observed units ('surface syllable rate'), and phone rate based on canonical unit counts ('canonical phone rate') and counts of observed units ('surface phone rate'). We also include a fifth measure ('CV rate') to

* Corresponding author.
E-mail address: l.plug@leeds.ac.uk (L. Plug).

https://doi.org/10.1016/j.specom.2021.05.006
Received 14 May 2020; Received in revised form 9 March 2021; Accepted 10 May 2021
Available online 17 May 2021

which we will return below.

In a language like English, the four common measures just mentioned may yield rather different figures on selected stretches of speech. In English, one syllable can correspond to between one (V) and seven phonemes (CCCVCCC). The temporal organisation of syllables is such that increases in syllable complexity are not associated with uniform increases in syllable duration (Browman and Goldstein, 1988; Byrd, 1995; Marin and Pouplier, 2010): therefore, increases in complexity tend to correspond to decreases in measured syllable rate but increases in phone rate. In the corpus of American English telephone speech of Greenberg et al. (2003), the mean duration of a stressed CVC syllable is 310 ms, and that of a stressed CCVC syllable is 382 ms. The former yields a phone rate of 9.7 and a syllable rate of 3.2; the latter a phone 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 and a phone rate measure identifies the other. Similarly, Jessen (2007) describes the 'curious artefacts when a speaker in speaking rapidly deletes canonical syllables, whereas another speaker might reduce or delete perhaps the same number of canonical sounds but still preserves the number of underlying syllables'. In these cases, deciding between the four measures is non-trivial.

With specific reference to 'canonical' and 'surface' rates, Den Os (1985) reports correlations of up to $r = 0.98$ for her experimental stimuli. However, these comprise two sets of nine read sentences (in Dutch and Italian). As the relationship between canonical and surface rates is largely determined by the prevalence of (syllable or phone) deletion, it is no surprise that in carefully produced speech, canonical and surface are close to equivalent. In spontaneous, or at least unscripted speech, by contrast, differences between 'canonical' and 'surface' syllable rates may be substantial. A phrase like I suppose this terrain is hard produced in 1.6 s yields a canonical syllable rate of 5; when produced with schwa deletion in both suppose and terrain the surface rate would be 3.75. The difference between the two figures is well above the 'just noticeable difference' for temporal variation in speech of around 5% (Quené, 2007). A pertinent question is how such differences translate to measurements taken over collections of utterances used in actual studies: for example, stimulus sets used in listening experiments, language learner speech samples, or larger corpora.

One context in which explicit comparison of tempo measures has taken place is that of forensic analysis, in which tempo is generally considered a relevant parameter for voice comparison (Gold and French, 2011).[1] Here, the aim is to establish the relative discriminating power of available measurement techniques. For German, Künzel (1997) reports that 'speech rate' calculated over stretches of speech including pauses and hesitations shows more speaker-internal variation than 'articulation rate' calculated over fluent stretches of speech only; therefore, articulation rate has greater speaker-discriminating power. Jessen (2007) reports articulation rate distributions for 100 German speakers, and compares population statistics reported across studies of speech tempo in German. Gold (2014) presents population statistics for 100 speakers of Southern Standard British English (SSBE), comparing articulation rates calculated over inter-pause and memory stretches with variable minimum length requirements. Gold (2014) quantifies the discriminating power of the alternative measures using Bayesian likelihood ratio (LR) calculations, which provide an assessment of 'strength of evidence' given competing hypotheses concerning the relationship between samples of speech (Gold and Hughes, 2014). Gold (2014) reports that syllable rates calculated over inter-pause and memory stretches yield near-equivalent discriminating powers: in other words, this particular methodological decision—the choice of domain over which to quantify articulation rates—has no substantial impact on analysis outcomes. Gold also investigated the effect of imposing different minimum length requirements when identifying stretches to calculate articulation rates over. She reports that as the length requirement moves up, within-speaker variation goes down; however, this does not have a substantial effect on discriminating power. So again, this methodological decision—the choice between possible minimum stretch lengths—appears to have little impact on analysis outcomes.

In this study we build on this previous work by comparing further alternative tempo measures, calculated on the memory stretch corpus of Gold (2014), in terms of their inter-correlations and relative discriminating powers. By reporting inter-correlations, we hope to inform any future studies in which the analyst is faced with a choice between the alternative tempo measures, and may wonder whether the choice is likely to be consequential for analysis outcomes. The higher the inter-correlations, the lower the likelihood that the methodological choice is consequential. We report correlations calculated across a sizeable corpus of Standard Southern British English unscripted speech, as well as correlations calculated over various data subsets, to simulate analysis scenarios that involve sampling stretches of speech from a larger corpus. By reporting relative discriminating powers, quantified using Bayesian LR calculations, we provide an example of one particular type of analysis in which a choice between alternative measures might in theory be consequential. High inter-correlations across speakers should translate to small differences in this analysis.

We should emphasize at the outset that it is impossible to concretely quantify the likelihood that a choice between two alternative measures will affect analysis outcomes. Consequently, there is no simple criterion for deciding whether two variables are correlated closely enough to be considered 'effectively equivalent'. Still, some general guidance can be gleaned from the literature on collinearity in multivariate analysis (e.g. Dormann et al., 2013; Tomaschek et al., 2018; Tu et al., 2005). Collinearity is observed when two predictor variables are linearly related to each other; when both are entered into a model predicting a third variable, the resulting model parameters can be difficult to interpret and unstable across alternative stepwise model building procedures. A common remedy is to reduce the number of predictor variables prior to modelling, either by conflating linearly related variables or by simply not entering individual predictors that are linearly related to others (Dormann et al., 2013). The latter approach is justified when it is not clear to what extent the information provided by an individual collinear predictor adds to that provided by the predictor with which it is collinear (Tu et al., 2005). In relation to pairwise correlations between predictors, Dormann et al. (2013) refer to a 'folk lore' threshold' of r>0.70 for potentially removing predictors. They conclude on the basis of regression modelling simulations that deselecting predictors correlated with each other at r>0.70 is indeed an effective 'rule of thumb'. This suggests that for a range of analytical purposes, variables which are correlated with each other at r>0.70 overlap sufficiently in the information they can provide to be considered 'effectively equivalent'. We will interpret observed inter-correlations among our tempo measures with reference to this criterion.

We compare articulation rates derived from syllable, phone and CV segment counts; for syllable and phone rates, we compare rates based on canonical and surface unit counts. To elucidate the latter comparison, we also report syllable and phone deletion rates for our corpus. Canonical and surface syllable and phone rates are reported in a wide range of studies. We additionally included the less widely used 'CV rate'. This is an available measure in the multilingual BonnTempo corpus (Dellwo et al., 2005), and Dellwo et al. (2006) assert that it 'has the advantage [over syllable rate] that labelling can be performed more objectively especially for the faster speech rates because acoustically phonetic categories such vowel and consonants are easier identifiable on an acoustic

[1] In relation to automatic speech recognition, Fosler-Lussier and Morgan (1999) summarize some inconsistent findings as to whether word rates are more predictive of system error than syllable and phone rates. In the clinical domain, comparison between rate measures has focused on finding automated rate extracting methods whose output figures map closely to syllable counts by human practitioners (e.g. Bakker, Brutten, & McQuain, 1995).

level than the phonological category ‘syllable’’. We were therefore particularly interested in the correlation between syllable and CV rates.

2. Method

2.1. Corpus

Our corpus comprises 2786 ‘memory stretches’ extracted from the Dynamic Variability in Speech Corpus (DyViS) (Nolan et al., 2009) by Gold (2014). The DyViS database was designed primarily to yield reliable population statistics for forensic phonetic work on British English (see Hughes et al., 2016; McDougall and Duckworth, 2017). It comprises recordings of 100 male speakers of Standard Southern British English (SSBE) in the age range 18–25 undertaking two role-play tasks relevant in a forensic context (a simulated police interview and a telephone call with a supposed accomplice) and two reading tasks, all recorded in studio conditions. Gold (2014) used the recordings of telephone calls with a supposed accomplice to derive population statistics for articulation rate. In this task, participants were engaged in a conversation with an interlocutor who wanted to compare accounts of a fictional crime. Participants were given visual stimuli such as pictures of people and places to allow them to construct their accounts.

Gold (2014) adopted the general methodology of Jessen (2007), segmenting 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). According to Jessen, this method is more efficient in casework practice than delimiting inter-pause stretches or intonation phrases, which are commonly used as phrasal domains in non-forensic studies (e.g. Dankovičová, 1997; Jacewicz et al., 2010; Mixdorff and Pfitzinger, 2005; Quené, 2008). Gold extracted a total of 2993 memory stretches, of which we excluded 207 from our analysis for reasons given below. Like Jessen, Gold (2014) stipulated a minimum stretch length of four (canonical) syllables, and left the maximum stretch length up to her own judgement of ease of recollection. Gold counted syllables through close listening and native speaker intuition as to the canonical syllabic make-up of each memory stretch; articulation rate values were derived from these counts.

Gold (2014) produced a verbatim transcription for each memory stretch. For the purpose of automatic alignment we had to edit some of these transcriptions, as explained below. Table 1 provides summary statistics derived from the edited transcriptions. For numbers of words, contracted forms (e.g. do not) were counted as single words. Abbreviations had to be written out as multi-word phrases, as explained below, and were therefore counted as multiple words. Street names (e.g. Harper Avenue) were also counted as multiple words. For numbers of canonical phones, the verbatim transcriptions were translated into SAMPA symbols (Wells, 1997). Long vowels, diphthongs and affricates were all counted as single phones on phonological grounds. The distributions summarised in Table 1 all show positive skew, exemplified in Fig. 1 for duration: they approximate a normal shape up to about 2 s duration, 10 words, 12 syllables and 35 phones, covering approximately 80% of the corpus—and the remainder of memory stretches have values up to the maxima in Table 1.

Table 1
Summary statistics for stretch length, calculated across our corpus of 2786 memory stretches.

	mean	minimum	maximum
duration (sec)	1.43	0.43	3.45
N words	7	2	24
N syllables (canonical)	9	4	26
N phones (canonical)	25	9	72

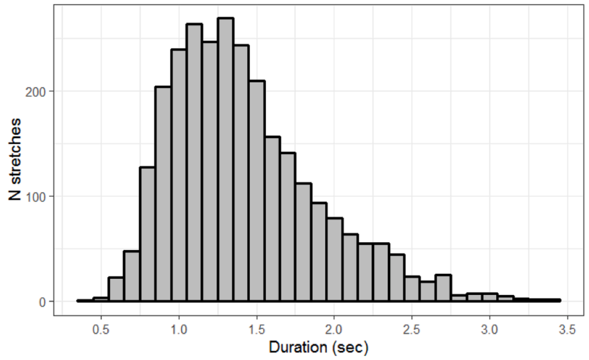

Fig. 1. Distribution of memory stretch durations across our corpus of 2786 memory stretches.

2.2. Segmentation and rate calculation

We used WebMAUS (Kisler et al., 2017) for segmentation, using a ‘pipeline’ of G2P (which converts input graphemes to SAMPA phones), MAUS (which does the automatic alignment) and PHO2SYL (which adds syllable boundaries to output segmentations). We used the ‘English (GB)’ language model, which was trained on the phonetic transcriptions of the Aix-MARSEC database (Auran et al., 2004). We initially worked with Gold’s total corpus of 2993 memory stretches. The input orthographic transcriptions were those provided by Gold (2014) with some edits to prevent recurrent errors identified in G2P trials: for example, abbreviations had to be written out ‘phonetically’ (e.g. ‘vee double you’ for ‘VW’) for the corresponding phones to be identified, while ‘no’ had to be rewritten ‘know’ to avoid it being treated as an alternative spelling of ‘number’. Syllabification was done within word boundaries. The second author manually checked the output phone and syllable segmentations using Praat (Boersma and Weenink, 2017). An example TextGrid is shown in Fig. 2. As we were interested in syllable and phone rates, the precise location of boundaries was not a major concern. The second author therefore applied a relatively light correction protocol to deal with misalignments: when two or more successive segments with clear acoustic correlates in the signal missed their targets (i.e. the segment clearly did not accurately delimit the acoustic correlates), the segments’ boundaries were manually moved to a more accurate position. Approximately 7% of memory stretches underwent this kind of correction.

As we were interested in phone and syllable deletions, the second author applied a more elaborate correction protocol to deal with WebMAUS’ inaccuracies in judging whether phones were delimitable as segments. First, the second author identified a set of frequent lexical items whose productions included heavily reduced ones which 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 first author and another phonetician, and segmentations were corrected to match consensus transcriptions. Second, leaving these frequent lexical items aside, WebMAUS sometimes treated a phone (most commonly schwa) as deleted when a segmental acoustic correlate could be delimited relatively easily; this happened in approximately 10% of memory stretches. More commonly, in approximately 30% of stretches, WebMAUS treated a phone (again most commonly schwa) as present in the surface form when no segmental acoustic correlate could be delimited. All of these cases were manually corrected by the second author in consultation with the first author. Third, WebMAUS often correctly identified the presence or absence of a schwa in ‘syllabic consonant’ contexts such as *bottle* [ˈbɒtəl]∼[ˈbɒtl̩]—but inaccurately treated the surface forms as monosyllabic when schwa was absent. These syllabification errors (which occurred in approximately 12% of memory stretches) were also corrected. Several additional issues were identified while checking the

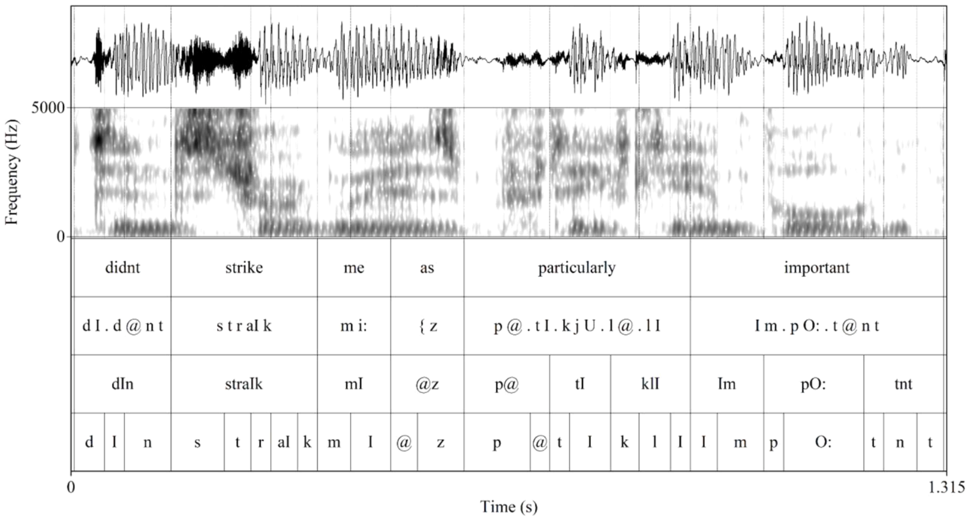

Fig. 2. Segmentation of one memory stretch (did not strike me as particularly important, with phone and syllable deletions in did not, particularly and important). Tier 1 shows the simplified orthographic transcription; Tier 2 the canonical form transcription with syllable boundaries marked by full stops; Tier 3 the surface form transcription by syllable; and Tier 4 the surface form transcription by phone. The word boundaries overlaying Tiers 1 and 2 and the syllable boundaries overlaying Tier 3 are derived from the phone-level segmentation shown on Tier 4.

segmentations, including a small number of stretch-internal silent pauses with a duration above 50 ms (although short enough for Gold not to have identified them), missing initial or final phones, excessive creak and signal disturbances making accurate segmentation impossible. We excluded the relevant memory stretches (N = 207, or 7%) from our analysis, taking the corpus size down to 2786 memory stretches.

We extracted canonical and surface syllable and phone rates (in syllables and phones per second) from the corrected segmentations, alongside syllable and phone deletions per memory stretch. As indicated above, we also calculated CV segment rates, following Dellwo et al. (2006). For this, we grouped any immediately consecutive consonantal phones (within and across word boundaries) into a combined C segment, and any immediately consecutive vocalic phones into a V segment. Following Arvaniti (2012) we treated /w/ and /j/ as consonantal if there was evidence of frication in their production; we treated /l/ and /ɹ/ as consonantal throughout. We then divided the total number of C and V segments in each memory stretch by the stretch's duration.

2.3. Deriving deletion rates

We coded all phone deletions for the phones involved, and mapped both syllable and phone deletions to part-of-speech tags for the words they occurred in. The latter allowed us to complement our primary analysis of articulation rates by memory stretch with analysis of syllable and phone deletions at the word level, as well as analysis of phone deletions at the phone level. In particular, it allowed us to quantify deletions within and across function and content words, for comparison with Johnson (2004) and others, and to rank canonical phones according to their individual deletion rates. For the word-level analysis, we derived part-of-speech tags from the KELLY English lexicon (Kilgarriff et al., 2014), which comprises approximately 7500 lemmas with part-of-speech tags. As the KELLY lexicon does not list individual word forms, we manually tagged plural, possessive, past tense, and other inflected forms (e.g. *friends, sister's, got, drives*). We also manually tagged the small number of words in our corpus whose lemmas are not included in the KELLY lexicon; many of these are proper nouns.

We then followed Bell et al. (2009) in coding nouns, verbs, adverbs and adjectives as content words and all other words as function words. Like Bell et al., we excluded from analysis at the word level the high-frequency sequences *you know* ($N = 58$) and *I mean* ($N = 14$), the discourse markers *yeah* ($N = 41$) and eh ($N = 24$), acronyms parsed as individual words ($N = 71$) and utterances of three words or fewer (108 utterances). We also excluded contracted forms such as *didn't* and *I'm*, which WebMAUS parsed as single word forms ($N = 618$). Moreover, we excluded words that could be classed as either function or content depending on context ($N = 2693$), such as *about, in, on, over* (preposition or adverb) and homonyms like *can* (modal verb and noun). Finally, we excluded a small number of short function words such as *a* and *are* that were transcribed by Gold (2014) because grammatically licensed, but not associated with a segmental realisation ($N = 76$).[2] These exclusions took our corpus size for the word- and phone-level analyses of deletion down to 16,041 words comprising 20,596 canonical syllables and 51, 741 canonical phones.

2.4. Quantitative analysis

We carried out all quantitative analysis in R (R Development Core Team, 2008). We used basic tidyverse functionality (Wickham et al., 2019) for exploring distributions, running Pearson's correlation tests and data visualisation, along with two tailored scripts for implementing the 'moving window' and random sampling methods we describe below.

We used the package fvclrr (Lo, 2018) for calculating the discriminant power for each of the five tempo measures in terms of Bayesian likelihood ratios (LRs). This package is based on the MATLAB implementation of Aitken and Lucy's multivariate kernel-density (MVKD) formula (Aitken and Lucy, 2004) by Morrison (2007). We used the package to run cross-validated, univariate same speaker (SS) and different speaker (DS) LR calculations. Same-speaker LR tests were run for each tempo measure individually, such that each of the 100 speakers acted as a same-speaker comparison (whereby their tokens were split in half for comparison purposes) and the remaining 99 speakers acted as the reference population. Different-speaker LR tests were run for each tempo measure as well, such that each speaker was compared to every other of the 100 speakers, while the remaining 98 speakers in each test would serve as the reference population. In total we ran 100 same-speaker tests and 9900 different-speaker tests.

We evaluated the discriminant performances of the tempo measures in terms of equal error rate (EER), which provides a 'hard' accept–reject measure of validity, and log-LR cost (Cllr), which provides a more 'gradient' measure of performance (Morrison, 2009). EER is the point at which the percentages of false hits and false misses are equal (Brümmer and Du Preez, 2006). Cllr is a Bayesian error metric that quantifies the ability of a system to align correctly with the expected outcome of

[2] Greenberg (1999) acknowledges that 'Under extreme conditions words of high frequency (and hence predictability) may be entirely deleted from the utterance, but without the listener's conscious awareness' and reports that this affects less than 1% of words in the Switchboard Corpus (Godfrey and Holliman, 1993). This is accurate for our corpus too: 76/(16,041+76) =0.47%.

whether speech samples are produced by the same or different speakers. As this paper is largely concerned with the relative performance of the five speaking tempo measures against one another, we did not calibrate the results in order to maximize the number of same speaker and different speaker tests that we could run.

3. Results

3.1. Syllable and phone deletions

We first consider the extent of syllable and phone deletion in our corpus, as this determines the relationship between canonical and surface rates. Our method identified 1305 syllable deletions and 9835 phone deletions. This means that 5% of canonical syllables and 14% of canonical phones in the corpus as a whole lack a surface realisation. Deletion of at least one syllable is observed in 33% of memory stretches. As shown in Fig. 3, the maximum number of deleted syllables per stretch is 7, but most stretches with deletions have just one missing syllable. Stretches with 4 or more syllable deletions are all long, at 15 canonical syllables or above (total range 4–22); stretches with no deletion or deletion of up to 3 syllables cover the full range of stretch lengths. Deletion of at least one phone is observed in 90% of memory stretches. As shown in Fig. 3, the maximum number of deleted phones per stretch is 17, but most stretches with deletions have between 1 and 4. Memory stretches with 10 or more phone deletions all have more than 20 canonical phones; still, zero deletion is observed in stretches of up to 45 canonical phones (16 words). The relationship between syllable and phone deletions is reasonably linear ($r = 0.59$), but each observed number of syllable deletions maps to a considerable range of phone deletions. For example, a stretch without any syllable deletions may still have up to 40% of its canonical phones missing a surface realisation.

While in-depth analysis of word- and phone-level deletion patterns is outside of the scope of this paper, we report summary statistics for comparison with those observed in other English corpora. The corpus as a whole has 6889 function words and 9152 content words. 5% of all words ($N = 809$) are subject to at least one syllable deletion and 18% ($N = 2844$) are subject to at least one phone deletion. Table 2 breaks these deletion rates down by word type. Table 2 shows that function words are on average shorter than content words and function words are slightly less likely to exhibit deletion.

At the phone level, the two consonants most prone to deletion (across content and function words) are /d/ (32% of canonical phones deleted) and /h/ (15% deleted); the two vowels most prone to deletion are /ʊ/ (32% of canonical phones deleted) and /ə/ (25% deleted). Of course word-specific pronunciation patterns influence these numbers: for example, more than half of /d/-deletions occur in the function word *and*; /h/-deletion is the norm in *him, her, his* and so on; and more than half of /ʊ/-deletions occur in the content word *actually*. These observations arguably call the validity of our canonical forms into question. While we consider this an interesting theoretical question (Ernestus, 2014; Kohler, 2000; Pierrehumbert, 2002), we must leave it aside here: for the purpose of our articulation rate analyses, we follow common practice in defining relevant linguistic units, and this entails referring to a standard lexicon of canonical word forms.

The deletion rates reported above are roughly comparable to those reported in previous corpus-based studies of English—although admittedly the pool for comparison is small and focused on American English. In comparison with our gross syllable deletion rate of 5%, Fosler-Lussier and Morgan (1999) cite a rate of 3% in the Switchboard Corpus (Godfrey and Holliman, 1993). In comparison with our gross phone deletion rate of 14%, Greenberg (1999) and Fosler-Lussier and Morgan (1999) both report a rate of 13% in Switchboard, while Shattuck-Hufnagel and Veilleux (2007) report deletion of 8% of all 'acoustic landmarks' in a small corpus of American English dialogue. In comparison with the percentages of words with at least one unit deletion reported above, Johnson (2004) reports that 'over 20%' of words in the Buckeye Corpus of Conversational Speech (Kiesling et al., 2006) have at least one phone deletion; Dilts (2013) reports 22% for the same corpus. Johnson reports identical percentages for syllable deletion to those in Table 2 at least one syllable deletion in 5% of function words and 6% of content words.

This means that collectively, the speakers in our corpus do not appear to be unusually careful articulators or speakers of a variety of English with relatively little deletion. Therefore, we can be reasonably confident that correlations between canonical and surface rates observed in our corpus generalize to other English corpora.

They may indeed generalize beyond English: for example, for Dutch, Van Bael, Baayen, and Strik (2007) report a gross syllable deletion rate of 6% and a gross phone deletion rate of 8% in speech selected from the Spoken Dutch Corpus (Oostdijk, 2002); Schuppler et al. (2011) report a gross syllable deletion rate of 9% in the Ernestus Corpus of Spontaneous Dutch (Ernestus, 2000). Van Bael et al. (2007) also report that 7% of words have at least one syllable deletion and 20% of words have at least one phone deletion. Strik et al. (2008) report that 15% of words in a different selection of speech from the Spoken Dutch Corpus (Oostdijk, 2002) have at least one phone deletion. For German, Zimmerer (2009) reports that in the Kiel Corpus of Spontaneous Speech (IPDS, 1994), 16% of all segments lack a surface realisation. For French, Adda-Decker et al. (2005) report an overall syllable deletion rate of 6%, as well as deletion rates of 13% for consonants and 15% for vowels in a corpus of spontaneous speech from radio interviews,. This means that these corpora are likely to yield similar correlations between canonical and surface articulation rates to those reported below, too.

We should note that in our corpus there is considerable inter-speaker variation in the prevalence of deletion: gross syllable deletion rates vary from close to zero to 14% between speakers (cf. 5% across speakers), and gross phone deletion rates vary between 4% and 16% (cf. 14% across speakers). For this reason we consider correlations between canonical and surface articulation rates by speaker below.

3.2. Canonical syllable rate: Gold (2014) vs WebMAUS

As indicated above, we derived our five rate measures from a semi-automatic syllabification and forced alignment workflow in WebMAUS. However, we also had access to the syllable rate figures of Gold (2014), which were derived from Gold's own estimations of canonical syllable counts. Of course decisions as to the composition of canonical forms have an impact on canonical syllable rate measures, as well as on the relationship between canonical and surface rate measures—that is, on what constitutes deletion. As these decisions are open to debate (Cangemi and Niebuhr, 2018; Ernestus, 2014; Kohler, 2000; Pierrehumbert, 2002), they can be considered another layer of 'researcher degrees of freedom'. To explore this layer, we compared the syllable counts of Gold (2014) with those generated by PHO2SYL in our WebMAUS workflow. We found that these are identical for 2500 out of 2786 memory stretches (90%). Over half of the discrepancies (167 out of 286, 58%) are related to lexical items whose syllabification is indeed debatable, such as *actually* (3~4 syllables), *particularly* (4~5), *secondary* (3~4) *camera* (2~3) and *tour* (1~2) ;[3] most of the remainder contain contracted forms, which Gold (2014) appears to have 'reconstructed' (e.g. *he's* 2 syllables) and PHO2SYL has not (e.g. *he's* 1 syllable); and in a small number of cases Gold's counts simply seem erroneous.

Unsurprisingly, given the high level of agreement on syllable counts, the canonical syllable rates of Gold (2014) and WebMAUS are very strongly correlated, at $r = 0.97$ (CI 0.972–0.975). This is illustrated in Fig. 4. In what follows, we will leave the syllable rate values of Gold (2014) aside and quantify the relationships among the five rate

[3] The complete list of lexical items is *actually, Barbara, camera, evening, eventually, every, general, geography, interest(ed), memory, occasionally, particularly, really, regularly, secondary, several, specifically, theatre, tour, travelling, usually.*

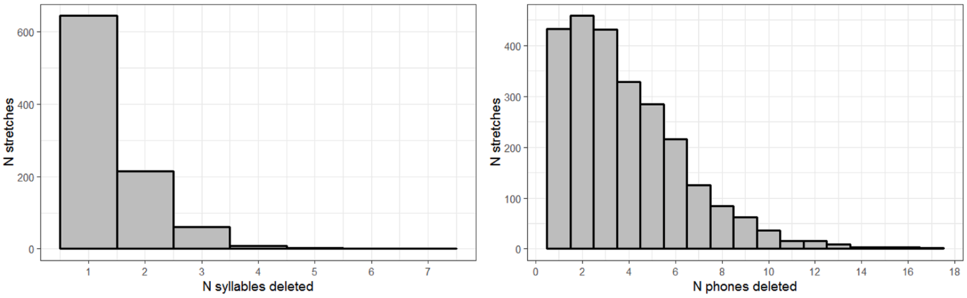

Fig. 3. Histograms for N syllable and phone deletions, excluding stretches with zero deletion.

Table 2
Summary statistics for the correlations between canonical syllable and phone rates (left) and surface syllable and phone rates (right); *r* is the Pearson's correlation coefficient and CI the corresponding 95% confidence interval. In the linear model equations, CPR = canonical phone rate, CSR = canonical syllable rate, SPR = surface phone rate, SSR = surface syllable rate.

	Canonical rates	Surface rates
r	0.85	0.82
CI	0.84–0.86	0.81–0.84
linear model	CPR = 2.9 + 2.2 × CSR	SPR = 3.9 + 1.8 × SSR
r range (by speaker)	0.63–0.94	0.50–0.95

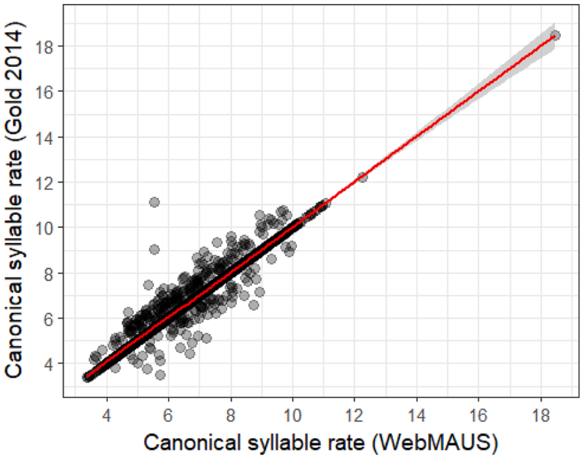

Fig. 4. Scatterplot illustrating the correlation between the canonical syllable rates derived from WebMAUS syllabification (x-axis) and calculated by Gold (2014) (y-axis), with LOESS fit lines.

measures derived from our semi-automatic syllabification and segmentation using WebMAUS.

3.3. Rate distributions

Before turning to our five rate measures, we can note one distant outlier in Fig. 4a single memory stretch produced at an extremely high tempo (e.g. canonical syllable rate > 18 sylls/sec, surface syllable rate > 13 sylls/sec). While the memory stretch in question (*What do you do for a living?*) appears to have been segmented accurately, and Fig. 4 confirms that Gold's syllable count was the same as that of WebMAUS, we deemed it appropriate to exclude it from the data set for the purpose of our correlation analysis. Fig. 5 shows that after the removal of this one data point, the distributions of the five rate measures are reasonably symmetrical. For both syllable and phone rates, the surface rate distribution appears somewhat closer to normal than the canonical rate distribution: canonical syllable rate shows some right skew which is less obvious for surface syllable rate, while canonical phone rate has several positive outliers that are absent in the surface phone rate distribution. The data points making the difference here are memory stretches produced at high tempo with multiple instances of syllable or phone deletion: these have high canonical rates but substantially lower surface rates.

With reference to the observed rate ranges and means (see the vertical dashed lines in Fig. 5), comparison with figures reported in other studies of speech tempo in British English is not straightforward due to methodological variation (cf. Jessen, 2007): for example, Tauroza and Allison (1990) report a mean canonical syllable rate of 4.3 sylls/sec (to be precise, 260 sylls/minute) for their sample of conversational speech, but appear to have quantified speaking rate (including pauses) as opposed to articulation rate. Moreover, it is clear from variationist research that there is considerable variation in speech tempo within language varieties delimited as broadly as 'British English' (see Clopper and Smiljanic, 2015; Coats, 2019; Jacewicz et al., 2010; Kendall, 2013; Kowal et al., 1983; Quené, 2008). The corpus of Lee and Doherty (2017) seems comparable in design to the DyViS database, although its speakers are simply described as 'Irish English'.

In very general terms, we can note that the mean syllable rates (6.55 canonical sylls/sec, 6.21 surface sylls/sec) seem high compared with similarly quantified rates reported for samples of American English spontaneous speech by Jacewicz et al. (2010), Kendall (2013) and Clopper and Smiljanic (2015). The latter tend to be closer to 5 for males. Note that Gold (2014) calculated a mean canonical syllable rate of 6.59 sylls/sec on our data set (see Fig. 4 above). Lee and Doherty (2017), who motivate their investigation of speech tempo in Irish English with the impressionistic observation that Irish English speakers speak faster than speakers of other varieties of English, report a mean articulation rate of 5.88 surface sylls/sec for male speakers' spontaneous speech. Unlike in Kendall's data, but as in Jacewicz et al.'s, measured rates are negatively correlated with stretch duration: shorter memory stretches are (weakly) associated with higher articulation rates (canonical syllable rate: r=–0.17, surface syllable rate: r=–0.16, canonical phone rate: r=–0.20, surface phone rate: r=–0.18, CV rate: r=–0.22).[4]

In relation to inter-speaker variation, Fig. 6 plots means calculated by speaker (that is, across the approximately 30 memory stretches produced by each speaker) against corresponding coefficients of variance. The coefficient of variance, or 'relative standard deviation' is calculated here by dividing the standard deviation by the mean for each individual speaker's rate distribution. It is a conservative measure of variance which corrects for the general finding in the analysis of temporal events that higher means are associated with higher standard deviations (see Jessen, 2007; Shaw et al., 2009, 2011). As shown in Fig. 5, speaker means vary substantially around the across-speaker

[4] Quené (2008) and Schwab and Avanzi (2015) observe the same effect of duration as Kendall does, in analyses of Dutch and French corpora, respectively—that is, longer phrases are associated with higher articulation rates. Quené (2008) accounts for this effect in terms of 'anticipatory shortening'. Note that this account makes sense if the stretches of speech under study constitute planning units in speech production. This is arguably less obvious for memory stretches than for inter-pause stretches or intonation phrases.

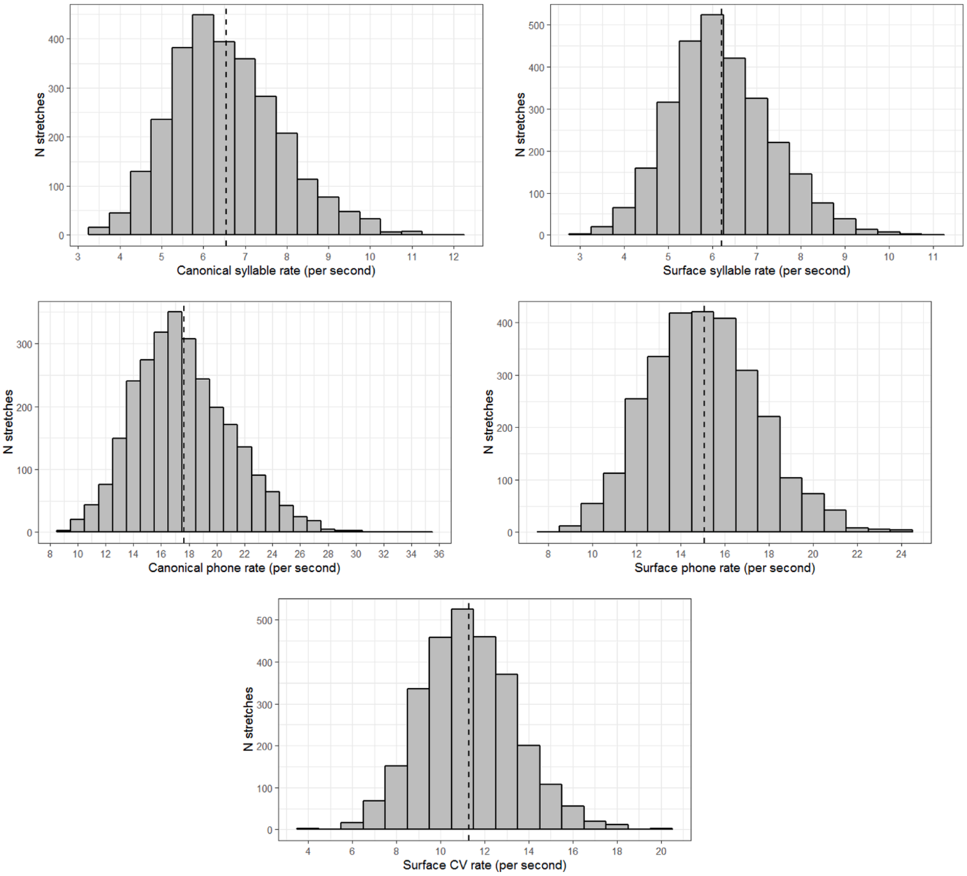

Fig. 5. Histograms for articulation rates: canonical and surface syllable rates (top, bin size 0.5 sylls/sec), canonical and surface phone rates (middle, bin size 1 phone/sec) and CV rate (bottom, bin size 1 segment/sec); in each graph, the dotted vertical line locates the mean.

means (shown again by the vertical dashed lines). We also see considerable variation in within-speaker variance: for example, for all rates the most highly variable speakers have a coefficient of variance that is at least twice that of the least variable speakers. There is no obvious linear relationship between means and coefficients of variance: that is, faster speakers are not necessarily more variable relative to their means.

3.4. Inter-correlations

We now turn to the inter-correlations among the five rate measures. In what follows, we compare syllable and phone rates, canonical and surface rates, and syllable and CV rates in turn. For each individual comparison, we provide a correlation plot and associated statistics, and we examine how the relationship between the two measures in question varies by speaker.

3.4.1. Syllable ~ phone rates

Fig. 7 shows scatterplots for two comparisons of syllable and phone rate: one in which both are calculated on the basis of canonical unit counts (left), and one in which both are calculated on the basis of observed unit counts (right). It is clear that in both cases, the syllable and phone rate are strongly correlated. Note that the straight lines of data points visible in both plots represent stretches for which the phone rate is exactly twice or three times the syllable rate. The strength of the correlations is confirmed by Pearson's correlation tests, results of which are shown in Table 2. The correlation for canonical rates is stronger by a small margin.

Table 2 also shows that the correlation coefficients vary considerably by speaker. This is of course to be expected: correlation coefficients calculated over about 30 data points are far more likely to be influenced by small sets of data points, or even individual ones, compared with coefficients calculated over hundreds of data points. It also makes sense that surface rates show more variation by speaker: for canonical rates, the relationship between syllable and phone rate is primarily determined by the phonotactics of the speakers' memory stretches, while for surface rates, the relationship depends on phonotactics and on speakers' production tendencies. As shown in Fig. 8, for canonical rates only 5 out of 100 correlation coefficients are below 0.7; for surface rates 15 out of 100 are below 0.7. Note that the speaker with the lowest correlation coefficient for surface rates seems somewhat of an outlier on that measure, but closer inspection suggests that the coefficient is heavily influenced by the speaker's production of a single memory stretch. In this stretch, *friends from secondary school* [fɹɛnz fəm sɛkdɹɪ skuːl], the syllable deletion in secondary (assumed canonical form /sɛkəndɹɪ/) yields an untypically low surface syllable rate given the surface phone rate. Excluding this single stretch from the speaker's sample increases the speaker's correlation coefficient to $r = 0.63$, which is no longer separated from the rest of the distribution.

3.4.2. Canonical ~ surface rates

Fig. 9 shows scatterplots for two comparisons of canonical and surface rate: one for syllables (left) and one for phones (right). Again it is clear that in both cases, the syllable and phone rate are strongly correlated. Note that the prominent straight lines of data points in both plots represent stretches with no deletions: for these stretches, canonical and surface rates are equivalent. Most other data points predictably fall below these lines: for these stretches, deletions lower the surface rate relative to the canonical rate. Just a few data points fall above the lines; these have no deletions and one syllable or phone insertion each. There were 83 such instances, 74 of which reflect insertion of a phone due to

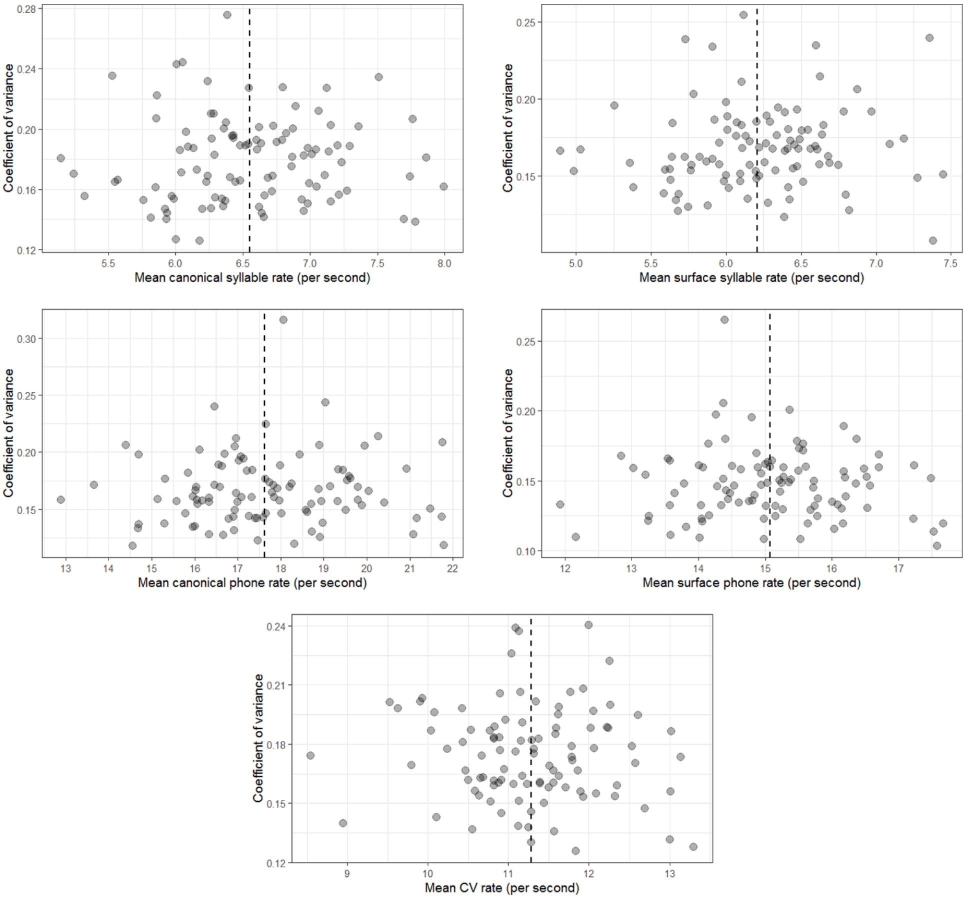

Fig. 6. Scatterplots illustrating variation in measured rate by speaker, plotting means (x-axis) against coefficients of variance (y-axis). Each data point represents one speaker, and in each graph, the dotted vertical line locates the overall mean.

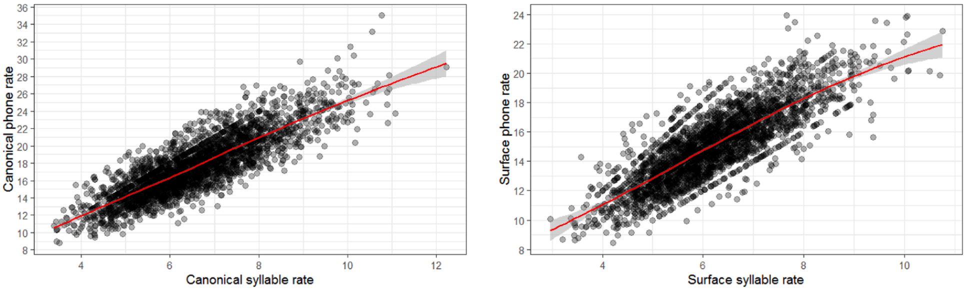

Fig. 7. Scatterplots illustrating the correlations between canonical syllable and phone rates (left) and surface syllable and phone rates (right), with LOESS fit lines.

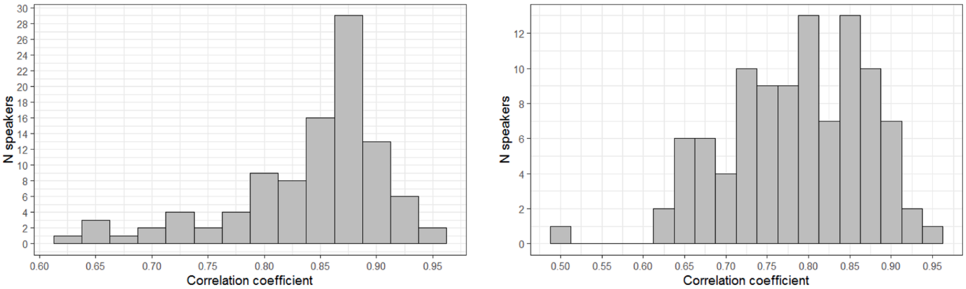

Fig. 8. Histograms for correlation coefficients (Pearson's *r*) calculated by speaker: correlations between canonical syllable and phone rates (left) and surface syllable and phone rates (right).

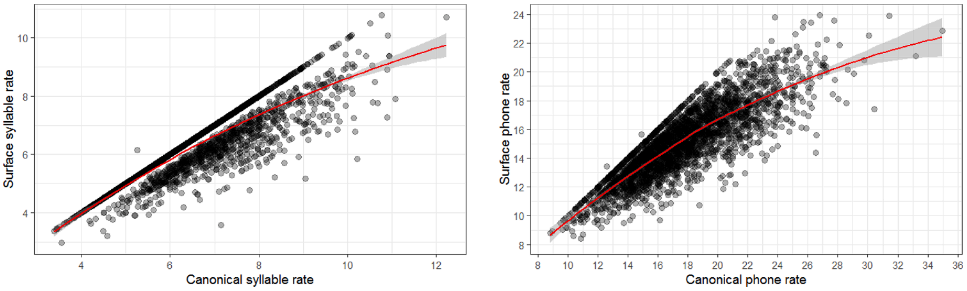

Fig. 9. Scatterplots illustrating the correlations between canonical and surface syllable rates (left) and canonical and surface phone rates (right), with LOESS fit lines.

linking /r/, for example in the sequences for a, or whether I. As it was trained on SSBE, a non-rhotic variety, the MAUS forced aligner did not parse postvocalic /r/ as a canonical realisation. Nevertheless, since these insertions only occur in 0.5% of words it is clear that their impact is minimal. The strength of the correlations is confirmed by Pearson's correlation tests, results of which are shown in Table 3. Excluding the memory stretches with no deletions has almost no impact on the strength of the correlations (r = 0.90 for syllable rates, r = 0.83 for phone rates).

However, it should be clear from the shape of the data scatters in Fig. 9 that the correlations vary systematically with overall tempo. Some of the memory stretches produced at the higher ends of the rate ranges are produced with very few deletions; at the same time, 'massive deletion' is more likely than at the lower ends of the rate ranges. As a result, the range of differences between canonical and surface rates increases as rates increase, and corresponding correlations weaken. This is shown by the LOESS fit lines in Fig. 9. To get closer to capturing this pattern with linear correlations, we implemented a 'moving window' approach to sampling: for both syllable and phone rate, we sampled 5 subsets of memory stretches, each making up a 60-percentile portion of the canonical rate distribution, with a step size of 10 percentiles—0–60, 10–70, 20–80, 30–90 and 40–100—and ran correlation tests on these subsets. We chose these window and step sizes to have reasonably sizeable samples to calculate correlations over (>1500 data points). Fig. 10 plots the correlation coefficients against the central canonical rates in the percentile ranges. The figure suggests that for both syllable and phone rates, the overall correlation reported in Table 3 is strongly determined by the correlation observed at lower rates. Still, for syllable rate, correlation coefficients stay above r = 0.7 throughout, while for phone rate, correlation coefficients stay above r = 0.6.[5]

Table 3
Summary statistics for the correlations between canonical and surface syllable rates (left) and canonical and surface phone rates (right); r is the Pearson's correlation coefficient and CI the corresponding 95% confidence interval. In the linear model equations, CSR = canonical syllable rate, SSR = surface syllable rate, CPR = canonical phone rate, SPR = surface phone rate.

	Syllable rates	Phone rates
r	0.90	0.84
CI	0.89–0.91	0.83–0.85
linear model	CSR = 0.2 + 1.0 × SSR	CPR = –0.1 + 1.2 × SPR
r range (by speaker)	0.73–0.99	0.57–0.98

Again, the correlation coefficients vary by speaker; however, the range of this variation is narrower than in the comparison of syllable and phone rates. For syllable rates no individual speaker sample has a correlation between canonical and surface rates with a coefficient below r = 0.7. As shown in Table 3 and Fig. 11, for phone rates only one speaker's sample has a correlation coefficient below r = 0.7. Inspection of the sample for this speaker suggests that the low correlation coefficient (r = 0.57) is not due to individual memory stretches with atypical rate values: the sample is simply characterised by a wide range of phone deletion numbers per stretch, yielding a wide range of differences between canonical and surface rate values. The speaker produces only one memory stretch with no phone deletion; by comparison, he produces 16 stretches with no syllable deletions, which means the correlation coefficient for canonical and surface syllable rates in the same sample is considerably higher (r = 0.80).

3.4.3. *CV rate*

As CV rate is a surface rate, we compare it with surface syllable and phone rates. Fig. 12 and Table 4 show that the two correlations are almost identical, at r = 0.78. Note that the straight lines of data points visible in the left plot in Fig. 12 correspond to memory stretches for which the CV rate is exactly 1.5 times or twice the syllable rate; these are fairly frequent. Again there is considerable variation between speakers, and again there are some apparent 'outlier' speaker samples, as seen in Fig. 13. We leave the details of these samples aside here.

3.4.4. *Correlations in random data samples*

So far we have reported correlations calculated across our entire corpus and correlations calculated by level or percentile range for relevant variables, such as speaker and stretch duration. As a final step in our examination of inter-correlations among our five rate measures, we also implemented a random sampling procedure. This was to simulate a range of research or applied scenarios in which analysis might be done on smaller versions of a corpus such as ours—for example, on a language learner production corpus with fewer speakers and utterances per speaker, or on a small set of phrases sampled from a corpus for use as stimuli in a listening experiment—and articulation rates are to be quantified for analysis. Even if correlations calculated across our whole

[5] We also examined how the correlations vary with stretch duration, particularly as syllable and phone deletions have a proportionally greater impact on articulation rates in shorter stretches of speech. We sampled 5 subsets of memory stretches, each making up a 60-percentile portion of the stretch duration distribution, with a step size of 10 percentiles and ran correlation tests on these subsets. This revealed minimal variation around the correlation coefficient calculated across the entire distribution, so we do not report the resulting figures. Comparison runs with alternative sizes yielded very similar results, so we are confident in concluding that stretch duration does not substantially affect the relationships among the rate measures we report.

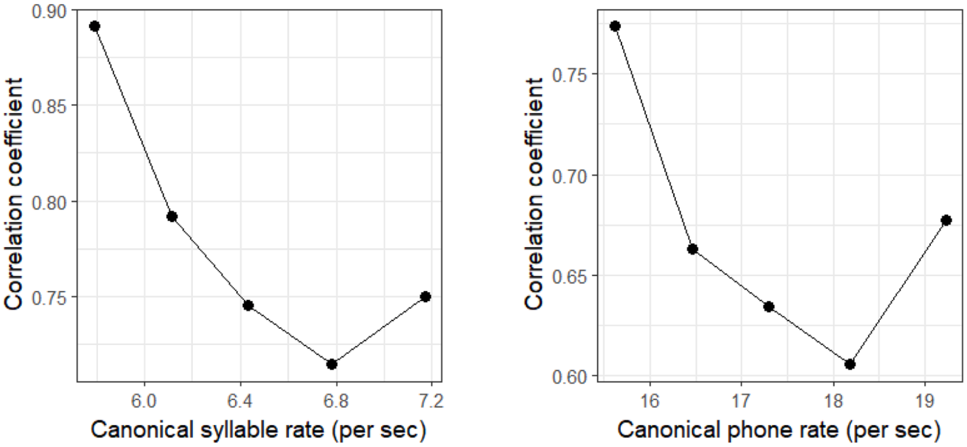

Fig. 10. Correlation coefficients (Pearson's *r*) calculated by percentile range in the canonical rate distribution: correlations between canonical and surface syllable rates (left) and canonical and surface phone rates (right). The x-axis value of each data point is the central value in a 60% portion of the canonical rate distribution.

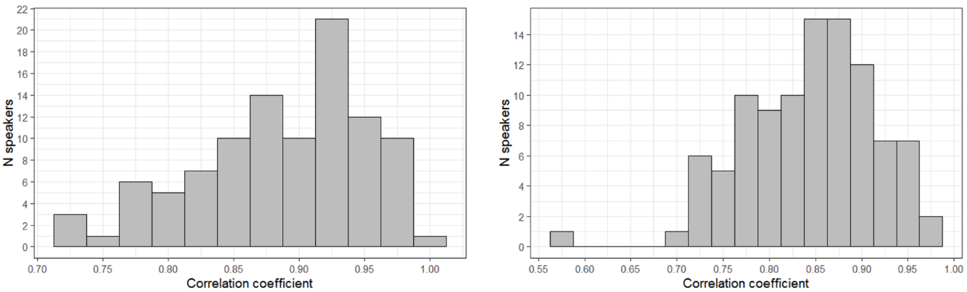

Fig. 11. Histograms for correlation coefficients (Pearson's *r*) calculated by speaker: correlations between canonical and surface syllable rates (left) and canonical and surface phone rates (right).

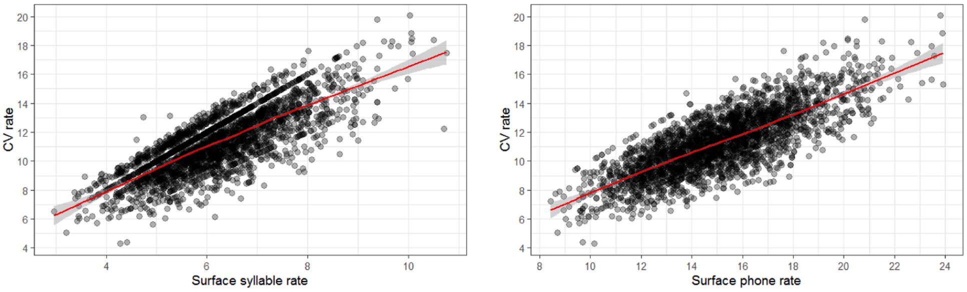

Fig. 12. Scatterplots illustrating the correlations between surface syllable and CV rates (left) and surface phone and CV rates (right), with LOESS fit lines.

Table 4
Summary statistics for the correlations between surface syllable and CV rates (left) and surface phone and CV rates (right); r is the Pearson's correlation coefficient and CI the corresponding 95% confidence interval. In the linear model equations, CVR = CV rate, SSR = surface syllable rate, SPR = surface phone rate.

	Syllable rate	Phone rate
r	0.78	0.78
CI	0.77–0.80	0.77–0.80
linear model	CVR = 2.2 + 1.5 × SSR	CVR = 1.1 + 0.7 × SPR
r range (by speaker)	0.43–0.91	0.42–0.90

corpus are very strong, smaller subsets may show less predictable relationships between rates: we have already seen this in the results of the analysis by speaker. We were interested in how low correlation coefficients might go in random data samples of varying sizes. To check this, we created random samples of memory stretches with the sizes $N = 10$, $N = 25$, $N = 50$, $N = 100$, $N = 150$, $N = 200$, $N = 250$, $N = 500$ and $N = 1000$. For each size, we created 100 random samples. We ran a Pearson's correlation test in each of the 100 samples at each of the nine sample sizes: this way, each sample size generated a distribution of 100 correlation coefficients. We ran this procedure separately for each pairwise rate comparison. We were particularly interested in how many of the coefficients were below $r = 0.7$.

Fig. 14 visualises the output of this procedure. For each pairwise rate comparison and each sample size, it shows a boxplot summarizing the distribution of correlation coefficients ($N = 100$ random samples of the specified size).[6] Looking first at the plot for the correlation between canonical syllable and phone rates (top row, left), we see that the correlation coefficient distributions are centred close to the coefficient

[6] Note that because sampling is random, the procedure produces different output each time it is run. We therefore ran it multiple times for each pairwise comparison, and were satisfied that the general shapes of the output distributions are stable across runs.

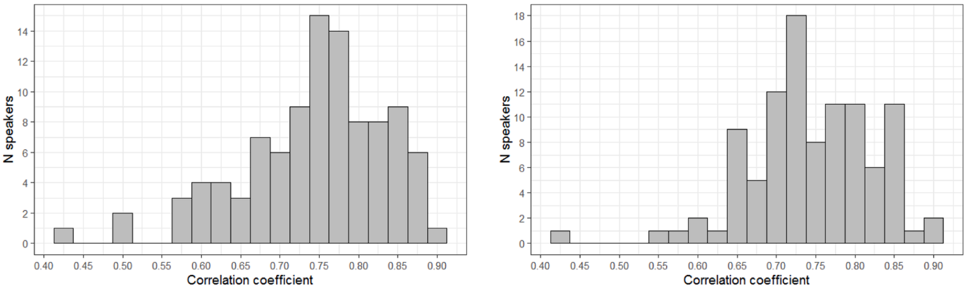

Fig. 13. Histograms for correlation coefficients (Pearson's *r*) calculated by speaker: correlations between CV and surface syllable rates (left) and CV and surface phone rates (right).

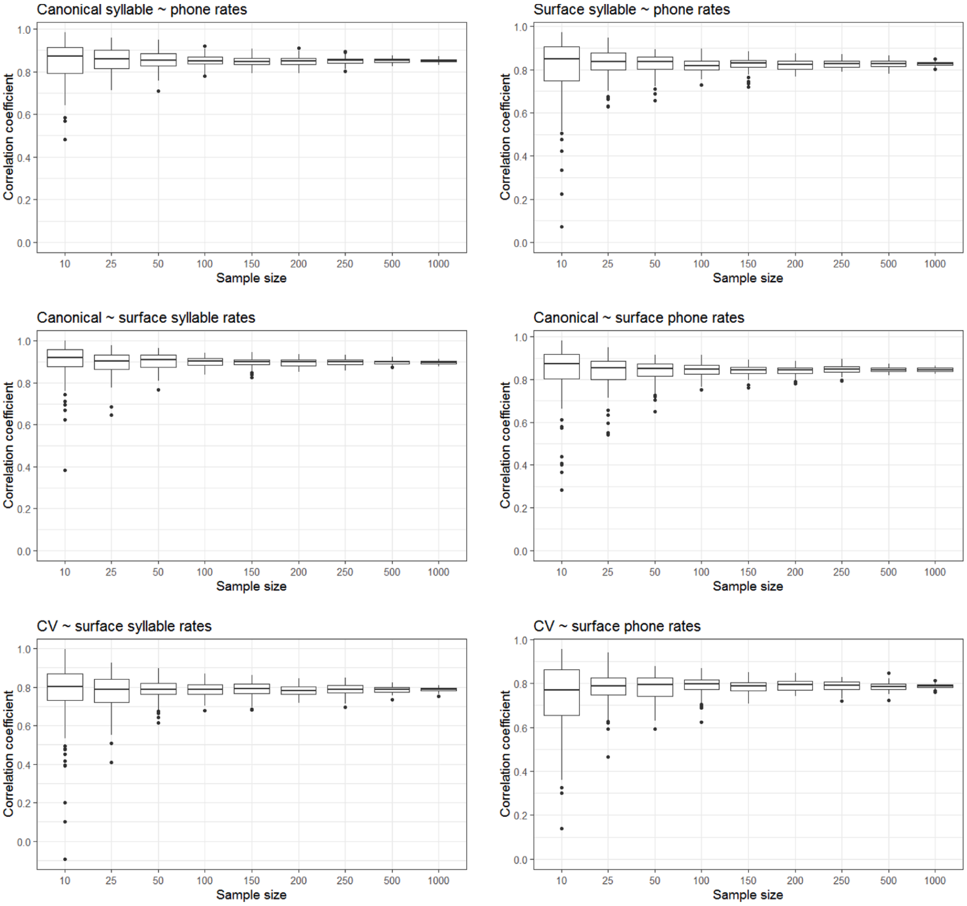

Fig. 14. Boxplots showing correlation coefficient distributions in random data samples of varying sizes; for each size, 100 samples were taken.

calculated across the whole corpus ($r = 0.85$, see Table 2) for all sample sizes, and the distributions for sample sizes of 150 and up are very narrow. This means that the correlation between these two measures comes out in the region of $r = 0.85$ pretty much whatever random data sample we consider even when working with sample sizes as small as 150 (about 5% of the total corpus). Distributions are predictably widest for the smallest sample sizes, but fewer than 10 individual samples yield a correlation coefficient below $r = 0.7$, at sample sizes 10 and 25.

Interpreting the other plots in the same way, we see that the correlation between surface syllable and phone rates is somewhat weaker than that between the canonical rates, as also reflected in the correlation coefficient calculated across the corpus ($r = 0.82$, see Table 2). Here fewer than 30 correlation coefficients fall below $r = 0.7$, at sample sizes 10, 25 and 50. The correlation between canonical and surface syllable rates is the strongest overall ($r = 0.90$, see Table 3), and fewer than 5 individual samples yield a correlation coefficient below $r = 0.7$. The correlation between canonical and surface phone rates ($r = 0.84$, see Table 3) is very similar to that between surface syllable and phone rates; here correlation coefficients fall below $r = 0.7$ at sample sizes 10 and 25. The correlations between CV rate and surface syllable and phone rates are the weakest overall ($r = 0.78$ for both comparisons, see Table 4), and this is reflected in comparatively large proportions of coefficients below $r = 0.7$ (maximum around 30 out of 100, for CV and surface phone rates at sample size 10) spread across a comparatively wide range of sample sizes.

3.5. Discriminating power

As indicated above, we decided to compare our five rate measures in one particular type of analysis in which the choice between alternative measures might in theory be consequential. Given the results of our correlation analysis, we can predict that the five measures are highly

Table 5
Summary results of a Bayesian likelihood ratio analysis comparing the five rate measures: equal error rate (EER) and likelihood ratio cost (Cllr). The rate measures are presented in descending order by EER.

	Equal error rate (EER)	Likelihood ratio cost (Cllr)
Canonical phone rate	28%	0.80
Surface phone rate	29%	0.84
Surface syllable rate	32%	0.87
Canonical syllable rate (WebMAUS)	34%	0.86
CV rate	37%	0.88
Canonical syllable rate (Gold 2014)	40%	0.94

similar in discriminant power as quantified through Bayesian likelihood ratio calculations. Table 5 confirms that this is the case. As indicated above, the equal error rate (EER) provides a 'hard' accept–reject measure of validity while the log-likelihood ratio cost (Cllr) provides a more 'gradient' measure of discriminant performance. For both, higher values are interpreted as poorer performance. For Cllr, values close to zero indicate a good system performance; values above 1 a poor one. The figures in Table 5 suggest that tempo is a relatively poor speaker discriminant parameter on its own, regardless of methodology, as it is characterised by rather high EER values and Cllr values close to 1, and that the differences among the five rate measures are small. Still, it is interesting to note that canonical syllable rate as calculated by Gold (2014) appears to be the weakest measure: therefore, relying on a semi-automated syllabification and segmentation workflow certainly does not have a negative effect on discriminant performance.

Fig. 15 plots the EER and Cllr figures alongside those for other parameters quantified by Gold (2014) on the same corpus. All figures were derived using the same method used in this paper, without calibration, using the MATLAB or R implementation of Aitken and Lucy's multivariate kernel-density (MVKD) formula (Aitken and Lucy, 2004). In this kind of plot, measures closest to the bottom left have the greatest discriminating power. Our five rate measures occupy a narrow region of the plot close to several individual formant measurements, but at considerable distance from fundamental frequency. Moreover, several studies have shown that measures generalising over multiple formants have considerably greater discriminating powers than individual formant measures (see Gold et al., 2013; Hughes et al., 2016). The figure confirms that the differences between canonical and surface variants of phone and syllable rates are particularly small. The two phone rate measures are somewhat stronger discriminant parameters than the syllable rate ones, particularly for EER. CV rate is weaker for EER than its closest comparison measure, surface syllable rate.

4. Discussion

In this study we investigated the extent to which articulation rates derived from syllable, phone and CV segment counts are correlated, and how they compare in terms of Bayesian likelihood ratios. For syllable and phone rates, we included canonical and surface rate calculations; for canonical syllable rate, we included figures calculated 'manually' by Gold (2014) and figures derived from the syllabifications generated in our WebMAUS workflow. As indicated at the outset of this paper, explicit comparisons of the distributions generated by such alternative measures are rare, although they elucidate whether the methodological choice between these measures is likely to be consequential for analysis outcomes. As such, they also inform comparisons of findings across studies in which different methodological choices were made.

In comparing the relevant distributions, we have proposed to adopt the 'rule of thumb' that variables which are correlated with each other at $r>0.70$ overlap sufficiently to be considered 'effectively equivalent' (Dormann et al., 2013). If we do, our general conclusion can be that canonical syllable rate, surface syllable rate, canonical phone rate, surface phone rate and CV rate are 'effectively equivalent' in our corpus, and alternative decisions as to the constitution of canonical forms do not jeopardise this effective equivalence. We examined correlation coefficients across the corpus as well as within a series of subsets—by speaker, by stretch duration, by quantile range and by random sampling—and in the majority, correlations stay at $r>0.70$. Some individual speakers show weaker correlations, and when randomly sampling very small sets of utterances from the corpus, such as sets of 10, it seems likely that different measures place individual utterances quite differently on a rate scale. It is debatable, of course, whether correlation coefficients calculated over 30 or fewer data points are reliable; in any case, most comparisons confirmed the strong inter-correlations among the five rate measures.

We also noted that the measures yield very similarly shaped

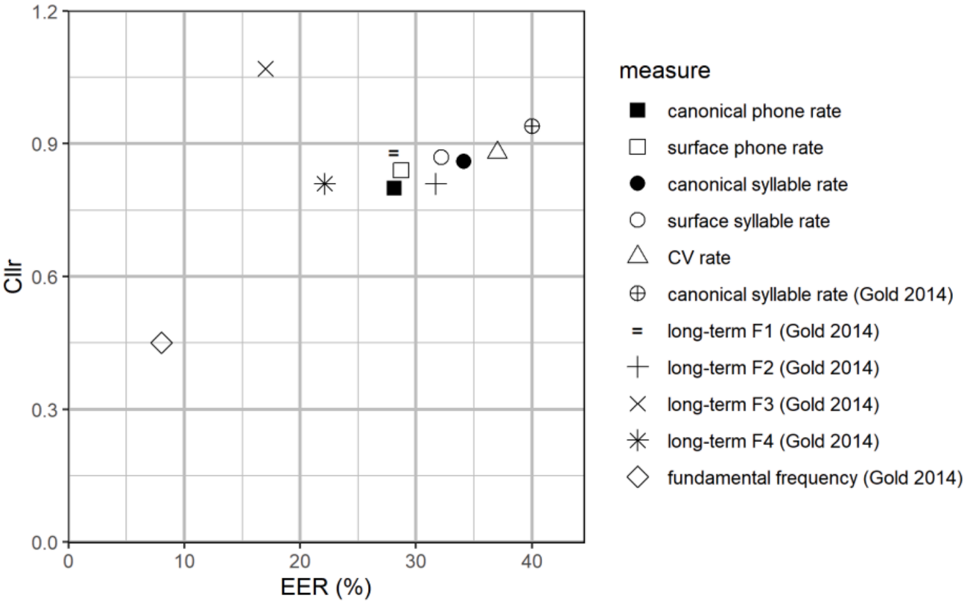

Fig. 15. Scatterplot showing equal error rates (x-axis) and log-likelihood ratio costs (y-axis) for our five rate measures and selected comparison measures from Gold (2014).

distributions with similar patterns of inter-speaker variation. Unsurprisingly, our Bayesian log-likelihood analysis revealed very similar performance across the five measures. This confirms that in a forensic context, it is unlikely that a choice between these five measures of speech tempo will be consequential for analysis outcomes. More generally, our findings suggest that we can assume with some confidence that past studies which used one of the five measures to quantify speech tempo would not have yielded substantially different results had they used another of the five measures: since the measures produce very similarly-shaped distributions, so that any analyses involving group comparisons or using speech tempo as a control variable should produce very similar results whichever measure is implemented.

Of course the measures do produce figures on different scales, so absolute values cannot be directly compared across studies. However, the regression equations we have reported above can be used as conversion formulas, and our comparisons of rate distributions provide an insight into the quantitative effects of several of the methodological decisions analysts are faced with when measuring tempo. In particular, Jessen (2007) suggests that Künzel (1997) may have reported a mean syllable rate of about 0.7 sylls/sec above that of his own observed mean (5.89 vs 5.21) because Künzel (1997) measured canonical syllable rate, not surface syllable rate. This seems reasonable, although in our study the difference in mean between the two measures is considerably smaller, at around 0.3 sylls/sec (6.55 vs 6.21). Moreover, our comparison between the canonical syllable rates calculated by Gold (2014) and WebMAUS suggests that making different decisions on the canonical syllable make-up of contentious lexical items has a small effect on canonical syllable rate figures: WebMAUS' syllabification yielded a mean of only 0.05 sylls/sec below Gold's. Gold (2014) reports a similarly small difference between means calculated over sets of memory stretches and inter-pause stretches for the same 25 speakers (5.96 vs 5.98 respectively).

It seems unlikely, therefore, that the observed differences between Gold's and our means on the one hand and those of previous studies on other English corpora on the other—which are in the region of 1 syll/sec—can be attributed to differing decisions of the type just described. As Jessen (2007) suggests, another methodological decision that is likely to have an impact on rate figures is whether stretches with filled pauses or other markers of hesitation, in particular noticeable segmental or syllabic lengthenings, are included or excluded. They were excluded in our study and, insofar as we can make out, in the studies of Jacewicz et al. (2010), Kendall (2013), Clopper and Smiljanic (2015) and Lee and Doherty (2017). As Jessen (2007) points out, studies in which they are included are likely to report lower mean rates. Unfortunately our analysis has not allowed us to quantify the impact of this decision.

Our results have at least two further practical implications. First, CV rate was described by Dellwo et al. (2006) as an efficient alternative to syllable rate, as its calculation does not involve making phonological decisions as to where syllable boundaries may be, how to treat 'syllabic' consonants and so on. In our corpus, CV rate is correlated with surface syllable rate below $r = 0.80$ and its by-speaker correlation coefficients include some of the lowest that we have observed across comparisons. The measure is not particularly strong in terms of discriminating power, so offers no obvious advantage over phone or syllable rates in a forensic context. Furthermore, we can question its efficiency as an alternative to syllable rate given that its calculation requires phone-level segmentation and phone rates are more closely correlated with syllable rates than CV rate; moreover, phone rates appear to be the strongest rates in terms of discriminating power. In a workflow like ours, therefore, there would seem to be no practical advantage to calculating CV rates compared with calculating syllable or phone rates.

Second, our findings suggest that at least for English and languages with similar phonotactics, speech rate estimators that depend on the automatic identification of acoustic correlates of syllables, (Bakker et al., 1995; de Jong and Wempe, 2009; Heinrich and Schiel, 2010; Martens et al., 2015) may well yield very similar output distributions to 'rough-and-ready' phone rate calculations using a general-purpose forced alignment system, or even canonical syllable rate calculations based on the output of a text-to-syllables conversion. This is because these estimators are typically evaluated against 'manual' syllable rate calculations. As we have seen, in our data set the 'manual' syllable rate calculations of Gold (2014) are very strongly correlated with the syllable rate figures derived from the syllabifications generated in our WebMAUS workflow. Therefore, while acoustically-based estimators are valuable models of speech tempo perception, for practical purposes much more basic tools may well produce 'effectively equivalent' figures.

5. Concluding remarks

The quantification of speech tempo is just one example of a procedure in phonetic analysis that can be operationalised in multiple ways, yielding 'researcher degrees of freedom' (Roettger, 2019; Simmons et al., 2011) which call the robustness of generalizations across studies into question. In this study we have attempted an explicit assessment of the impact of researchers' choices among some of the available measures. Our results suggests that in a sizeable English corpus with normal deletion rates, five common articulation rates are closely inter-correlated and have similar discriminating powers; decisions as to the segmental make-up of canonical forms also have limited impact on distributions. Therefore, for common analytical purposes and forensic applications the choice between the measures considered is unlikely to substantially affect outcomes.

Of course, the confidence with which we can make methodological recommendations on the basis of our results depends on how representative our corpus of Standard Southern British English memory stretches is in terms of the measures under consideration. We established that our corpus shows similar syllable and phone deletion frequencies to other English corpora and at least one Dutch corpus. This suggests that the relationships between canonical and surface rates should also generalize beyond this study. The relationships among syllable, phone and CV rates mostly depend on the phonotactics of the utterances under consideration—so while these are language-specific, they should in principle generalize reasonably well across varieties of English.

We hope that this study makes a valuable addition to available population data for articulation rate variation, as well as syllable and phone deletion, in British English—although our results confirm Gold (2014) conclusion that speech tempo is a relatively weak discriminant parameter. We also hope that our study provides a straightforwardly replicable model for research into the 'researcher degrees of freedom' (Roettger, 2019; Simmons et al., 2011) that alternative operationalizations of major phonetic parameters generate.

CRediT authorship contribution statement

Leendert Plug: Conceptualization, Methodology, Supervision, Formal analysis, Visualization, Writing - original draft, Writing – review & editing. **Robert Lennon:** Data curtion, Methodology, Formal analysis, Visualization, Writing – review & editing. **Erica Gold:** Data curtion, Formal analysis, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was made possible by a Leverhulme Trust Research Grant (RPG-2017-060: Speech tempo perception and missing sounds). We are grateful to Rachel Smith for her involvement with this research, Chris Norton for Praat and R scripting help, and Justin Lo for assistance

with the FVCLRR package.

References

Adda-Decker, M., Boula de Mareuil, P., Adda, G., Lamel, L., 2005. Investigating syllabic structures and their variation in spontaneous French. Speech Commun. 46 (2), 119–139.

Aitken, C.G., Lucy, D., 2004. Evaluation of trace evidence in the form of multivariate data. J. R. Stat. Soc. Ser. C Appl. Stat. 53 (1), 109–122.

Arvaniti, A., 2012. The usefulness of metrics in the quantification of speech rhythm. J. Phon. 40 (3), 351–373.

Auran, C., Bouzon, C., Hirst, D., 2004. The Aix-MARSEC project: an evolutive database of spoken English. In: Proceedings of the Paper presented at the Second International Conference on Speech Prosody. Nara.

Bakker, K., Brutten, G.J., McQuain, J., 1995. A preliminary assessment of the validity of three instrument-based measures for speech rate determination. J. Fluency Disord. 20 (1), 63–75.

Bell, A., Brenier, J.M., Gregory, M., Girand, C., Jurafsky, D., 2009. Predictability effects on durations of content and function words in conversational English. J. Mem. Lang. 60 (1), 92–111.

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

Bosker, H.R., Pinget, A.F., Quené, H., Sanders, T., de Jong, N.H., 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. Lang. Test. 30 (2), 159–175.

Browman, C.P., Goldstein, L., 1988. Some notes on syllable structure in articulatory phonology. Phonetica 45 (2–4), 140–155.

Brümmer, N., Du Preez, J., 2006. Application-independent evaluation of speaker detection. Comput. Speech Lang. 20 (2–3), 230–275.

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

Cangemi, F., & Niebuhr, O. (2018). Rethinking reduction and canonical forms. In F. Cangemi & M. Clayards & O. Niebuhr & B. Schuppler & M. Zellers (Eds.), Rethinking reduction (pp. 277-302): De Gruyter Mouton.

Clopper, C.G., Smiljanic, R., 2015. Regional variation in temporal organization in American English. J. Phon. 49, 1–15.

Coats, S., 2019. Articulation rate in American English in a corpus of YouTube videos. Lang. Speech, 0023830919894720.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F., 2015. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 71, 10–49.

Dankovičová, J., 1997. The domain of articulation rate variation in Czech. J. Phon. 25 (3), 287–312.

de Jong, N.H., Wempe, T., 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behav. Res. Methods 41 (2), 385–390.

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

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J., Wagner, P., 2005. The Bonntempo-corpus and Bonntempo-tools: a database for the study of speech rhythm and rate. In: Proceedings of the Paper presented at the Ninth Annual Conference of the International Speech Communication Association (Interspeech 2005). Lisbon.

Den Os, E., 1985. Perception of speech rate of Dutch and Italian utterances. Phonetica 42, 124–134.

Dilts, P., 2013. Modelling Phonetic Reduction in a Corpus of Spoken English using Random Forests and Mixed-Effects Regression. University of Alberta, Edmonton, Alberta.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R.G., Gruber, B., Lafourcade, B., Leitao, P.J., Munkemuller, T., McClean, C., Osborne, P.E., Reineking, B., Schroder, B., Skidmore, A.K., Zurell, D., Lautenbach, S, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.

Ernestus, M., 2000. Voice Assimilation and Segment Reduction in Casual Dutch. A corpus-Based Study of the Phonology–Phonetics Interface. Netherlands Graduate School of Linguistics, Utrecht.

Ernestus, M., 2014. Acoustic reduction and the roles of abstractions and exemplars in speech processing. Lingua 142, 27–41.

Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word frequency on pronunciations in conventional speech. Speech Commun. 29 (2–4), 137–158.

Godfrey, J.J., Holliman, E., 1993. Switchboard-1 Release 2 LDC97S62. Philadelphia: Linguistic Data Consortium.

Gold, E., 2014. Calculating Likelihood Ratios For Forensic Speaker Comparisons Using Phonetic and Linguistic Parameters. University of York, York.

Gold, E., French, P., 2011. International practices in forensic speaker comparison. Int. J. Speech Lang. Law 18 (2), 293–307.

Gold, E., French, P., Harrison, P., 2013. Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. J. Acoust. Soc. Am. Proc. Meet. Acoust. 19, 1–8.

Gold, E., Hughes, V., 2014. Issues and opportunities: the application of the numerical likelihood ratio framework to forensic speaker comparison. Sci. Justice 54 (4), 292–299.

Greenberg, S., 1999. Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation. Speech Commun. 29 (2–4), 159–176.

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S.Y., 2003. Temporal properties of spontaneous speech: a syllable-centric perspective. J. Phon. 31 (3–4), 465–485.

Heinrich, C., Schiel, F., 2010. Estimating speaking rate by means of rhythmicity parameters. In: Proceedings of the Paper presented at the Twelth Annual Conference of the International Speech Communication Association. Florence. Interspeech 2011.

Hughes, V., Wood, S., Foulkes, P., 2016. Strength of forensic voice comparison evidence from the acoustics of filled pauses. Int. J. Speech Lang. Law 23 (1).

IPDS. (1994). The Kiel Corpus of Spontaneous Speech. Kiel: Für Phonetik Und Digitale Sprachverarbeitung.

Jacewicz, E., Fox, R.A., Wei, L., 2010. Between-speaker and within-speaker variation in speech tempo of American English. J. Acoust. Soc. Am. 128 (2), 839–850.

Jessen, M., 2007. Forensic reference data on articulation rate in German. Sci. Justice 47 (2), 50–67.

Johnson, K., 2004. Massive reduction in conversational American English. In: Proceedings of the Paper presented at the Tenth International Symposium on Spontaneous Speech: Data and Analysis.

Kendall, T., 2013. Speech rate, pause, and Sociolinguistic variation: Studies in Corpus Sociophonetics. Palgrave Macmillan, London.

Kiesling, S., Dilley, L., Raymond, W., 2006. The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of Conversational Speech. Department of Psychology, Ohio State University, Columbus, Ohio. www.buckeyecorpus.osu.edu.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J.B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., Volodina, E., 2014. Corpus-based vocabulary lists for language learners for nine languages. Lang. Resour. Eval. 48 (1), 121–163.

Kisler, T., Reichel, U.D., Schiel, F., 2017. Multilingual processing of speech via web services. Comput. Speech Lang. 45, 326–347.

Kohler, K.J., 2000. Investigating unscripted speech: implications for phonetics and phonology. Phonetica 57 (2–4), 85–94.

Koreman, J., 2006. Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech. J. Acoust. Soc. Am. 119 (1), 582–596.

Kowal, S., Wiese, R., O'Connell, D.C, 1983. The use of time in storytelling. Lang. Speech 26 (4), 377–392.

Künzel, H.J., 1997. Some general phonetic and forensic aspects of speaking tempo. Forensic Linguist. 4 (1), 48–83.

Lee, A., Doherty, R., 2017. Speaking rate and articulation rate of native speakers of Irish English. Speech Lang. Hear. 20 (4), 206–211.

Lo, J. (2018). Fvclrr: Likelihood ratio calculation and testing in forensic voice comparison. https://github.com/justinjhlo/fvclrr.

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.

Martens, H., Dekens, T., Van Nuffelen, G., Latacz, L., Verhelst, W., De Bodt, M., 2015. Automated speech rate measurement in dysarthria. J. Speech Lang. Hear. Res. 58 (3), 698–712.

McDougall, K., Duckworth, M., 2017. Profiling fluency: an analysis of individual variation in disfluencies in adult males. Speech Commun. 95, 16–27.

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

Morrison, G.S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. http://geoff-morrison.net/#MVKD.

Morrison, G.S., 2009. The place of forensic voice comparison in the ongoing paradigm shift. In: Proceedings of the Paper presented at the Second International Conference on Evidence Law and Forensic Science.

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S., 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguist. 20 (1), 50–64.

Nolan, F., McDougall, K., De Jong, G., Hudson, T., 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. Int. J. Speech Lang. Law 16 (1), 31–57.

Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In P. Peters & P. Collins & A. Smith (Eds.), New frontiers of corpus research (pp. 105-112). Amsterdam: Rodopi.

Pellowski, M.W., 2010. Speech-language pathologists' knowledge of speaking rate and its relationship to stuttering. Contemp. Issues Commun. Sci. Disord. 37, 50–57.

Pfitzinger, H., 1996. Two approaches to speech rate estimation. In: Proceedings of the Paper presented at the Sixth Australian International Conference on Speech Science and Technology. Canberra.

Pierrehumbert, J.B., 2002. Word-specific phonetics. Lab. Phonol. 7 (4–1), 101–139.

Quené, H., 2007. On the just noticeable difference for tempo in speech. J. Phon. 35 (3), 353–362.

Quené, H., 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. J. Acoust. Soc. Am. 123 (2), 1104–1113.

R. Development Core Team. (2008). R: A language and environment for statistical computing.

Roettger, T.B., 2019. Researcher degrees of freedom in phonetic research. Lab. Phonol. J. Ass. Lab. Phonol. 10 (1), 1. https://doi.org/10.5334/labphon.147.

Schuppler, B., Ernestus, M., Scharenborg, O., Boves, L., 2011. Acoustic reduction in conversational Dutch: a quantitative analysis based on automatically generated segmental transcriptions. J. Phon. 39 (1), 96–109.

Schwab, S., Avanzi, M., 2015. Regional variation and articulation rate in French. J. Phon. 48, 96–105.

Shattuck-Hufnagel, S., & Veilleux, N. (2007). Robustness of acoustic landmarks in spontaneously-spoken American English. Paper presented at the Sixteenth International Congress of Phonetic Sciences.

Shaw, J.A., Gafos, A.I., Hoole, P., Zeroual, C., 2009. Syllabification in Moroccan Arabic : evidence from patterns of temporal stability in articulation. Phonology 26 (1), 187–215.

Shaw, J.A., Gafos, A.I., Hoole, P., Zeroual, C., 2011. Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. Phonology 28 (3), 455–490.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 22 (11), 1359–1366.

Strik, H., van Doremalen, J., Cucchiarini, C., 2008. Pronunciation reduction: how it relates to speech style, gender, and age. In: Proceedings of the Paper presented at the Ninth Annual Conference of the International Speech Communication Association. Brisbane. Interspeech 2008.

Tauroza, S., Allison, D., 1990. Speech rates in British English. Appl. Linguist. 11 (1), 90–105.

Tomaschek, F., Hendrix, P., Baayen, R.H., 2018. Strategies for addressing collinearity in multivariate linguistic data. J. Phon. 71, 249–267.

Tu, Y.K., Kellett, M., Clerehugh, V., Gilthorpe, M.S., 2005. Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. Br. Dent. J. 199 (7), 457–461.

Van Bael, C., Baayen, R.H., Strik, H., 2007. Segment deletion in spontaneous speech: a corpus study using mixed effects models with crossed random effects. In: Proceedings of the Paper presented at the Eighth Annual Conference of the International Speech Communication Association. Interspeech 2007.

Wang, Y., Gales, M.J.F., Knill, K.M., Kyriakopoulos, K., Malinin, A., van Dalen, R.C., Rashid, M, 2018. Towards automatic assessment of spontaneous spoken English. Speech Commun. 104, 47–56.

Wells, J.C. (1997). SAMPA computer readable phonetic alphabet. https://www.phon.ucl.ac.uk/home/sampa/.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., Francois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Milton Bache, S., Mueller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutanill, H., 2019. Welcome to the Tidyverse. J. Open Source Softw. 4 (43), 1686.

Zimmerer, F., 2009. Reduction in Natural Speech. Goethe Universität, Frankfurt.