A COMPARISON OF MULTIPLE SPEECH TEMPO MEASURES: INTER-CORRELATIONS AND DISCRIMINATING POWER

Robert Lennon[1], Leendert Plug[1], Erica Gold[2]

[1]University of Leeds, United Kingdom, [2]University of Huddersfield, United Kingdom
r.w.lennon@leeds.ac.uk

ABSTRACT

Studies that quantify speech tempo on acoustic grounds typically use one of various rate measures. Explicit comparisons of the distributions generated by these measures are rare, although they help assess the robustness of generalisations across studies; moreover, for forensic purposes it is valuable to compare measures in terms of their discriminating power. We compare five common rate measures ―canonical and surface syllable and phone rates, and CV segment rate―calculated over fluent stretches of spontaneous speech produced by 30 English speakers. We report deletion rates and correlations between the five measures and assess discriminating powers using likelihood ratios. Results suggest that in a sizeable English corpus with normal deletion rates, these five rates are closely inter-correlated and have similar discriminating powers; therefore, for common analytical purposes the choice between these measures is unlikely to substantially affect outcomes.

Keywords: phonetics, forensic speaker comparison, speech tempo, correlations, likelihood ratios.

1. INTRODUCTION

Studies that quantify speech tempo through signal-based measurements tend to use one of many available measurement techniques. Researchers choose what to count—words, syllables, phones, or derived units such as C and V segments [8, 25]—and what temporal domains to count in—total speaking time including or excluding pauses, or stretches of speech such as clauses, intonation phrases, interpause stretches or memory stretches [6, 15]. When counting syllables or phones, researchers can count units as expected in canonical pronunciations, or as actually observed in their data [18].

These different measurement techniques can yield different figures for subsets of instances, depending on the phonology of the language and on individual speaker characteristics. However, direct comparisons are rare: typically, studies present the outputs of one technique only, and studies that do refer to multiple techniques do not necessarily report on correlations between their outputs. For one thing, this can make it difficult to compare tempo figures across studies [15]; for another, it can leave in doubt the robustness of reported data patterns.

One context in which comparison of measures has taken place is that of forensic analysis, in which tempo is a candidate parameter for voice comparison [10]. Here, the aim is to establish the relative discriminating power of available measurement techniques. [19] reports (for German) 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. [15] reports articulation rate distributions for 100 German speakers, and compares population statistics reported across studies of speech tempo in German. [9] presents population statistics for 100 speakers of Southern Standard British English (SSBE), comparing articulation rates calculated over interpause and memory stretches with variable minimum length requirements. [9] quantifies the discriminating power of the alternative measures using Bayesian likelihood ratio (LR) calculations, which provide a gradient assessment of 'strength of evidence' given competing hypotheses concerning the relationship between samples of speech [11]. [9] concludes that articulation rates calculated over interpause and memory stretches yield similar discriminating power, and that articulation rate is a relatively poor discriminant parameter due to considerable speaker-internal variation.

In this study we extend this previous work by comparing further alternative tempo measures, calculated on a subset of the corpus of [9], in terms of their inter-correlations and relative discriminating powers. We compare articulation rates derived from syllable, phone and CV segment counts; for syllable and phone rates, we compare rates based on canonical and 'actual' unit counts. [15] speculates that some of the differences in reported articulation rate means across studies can be attributed to different choices among these alternatives. One aim of this study was to provide empirical data that may inform such attributions; another was to assess whether the choices between these alternatives is consequential for discriminant power.

2. METHOD

2.1. Corpus

Our corpus comprises 865 'memory stretches' extracted from non-contemporaneous speech in the Dynamic Variability in Speech Corpus (*DyViS*) [24] by [9]. Memory stretches are short samples of speech over which measurements can be taken; these are commonly used in forensic practice as alternatives to interpause stretches or intonation phrases. The stretches were produced by 30 SSBE speakers (males, aged 18–25).

2.2. Segmentation and rate calculation

We used WebMAUS [17] for forced alignment. Audio and orthographic transcriptions served as input; a pipeline of Grapheme2Phoneme, MAUS and Pho2Syl generated a Praat TextGrid [4] for each audio file. Table 1 shows the contents of one of the output TextGrids. Tier 1 contains the orthography, with boundaries between words. Tier 2 contains the canonical phonemic transcription rendered in the SAMPA alphabet. Tier 3 adds syllable boundaries (dots). Tier 4 contains the surface segmentation. Tier 5 contains the surface syllables, following the principle of word boundary syllabification. To generate the segmentation, MAUS balances the likelihood that each phone will appear according to its phonetic model for British English with the acoustic 'landmarks' present in the audio. In this example, the /t/ in *that* and the /l/ in *all* were deemed absent. All of the other tiers take their alignment from the surface segmentation.

Table 1: Contents of TextGrid for ms016_03

Tier	Segmentation
1	| *And* | *I* | *said* | *that* | *she* | *cycles* | *to* | *work* | *all* | *the* | *time* |
2	| @nd | aI | sed | D{t | Si | saIk@lz | t@ | w3k | Ol | D@ | taIm |
3	| @nd | aI | sed | D{t | Si | saI.k@lz | t@ | w3k | Ol | D@ | taIm |
4	| @|n|aI|s|e|d|D|@|S|i|s|aI|k|l|z|t|@|w|3|k|O|D|@|t|aI|m |
5	| @n | aI | sed | D@ | Si | saI | klz | t@ | w3k | O | D@ | taIm |

We re-processed instances where G2P had failed to recognize specific spellings. We manually corrected stretches where MAUS had missed two or more successive phones with clear acoustic correlates, and identified a set of frequent lexical items for which phone deletions were regularly missed; this included *actually* and *didn't*. Two phoneticians agreed broad transcriptions for all instances of these items, and we corrected TextGrids accordingly. This protocol ensured consistency and reasonable accuracy in identifying phone deletions, while maximizing potential for replication by keeping manual correction to a minimum.

We then extracted phone and syllable counts for all stretches and calculated canonical and surface syllable rates (CSR, SSR), and canonical and surface phone rates (CPR, SPR). We also calculated CV rate (CVR) [8]. For this, we grouped any immediately consecutive consonantal phones into a combined C segment, and any immediately consecutive vocalic phones into a V segment. We then divided the total number of C and V segments in each memory stretch by the stretch's duration. Alongside these articulation rates, we derived phone and syllable deletion counts to compare with previous studies.

2.3. Likelihood ratio analysis

The discriminant power for each of the five tempo measures was calculated using Bayesian LRs. LRs were calculated in MatLab using an implementation of [2]'s Multivariate Kernel-Density (MVKD) formula [21]. A MatLab script [13] was used to run multiple same speaker (SS) and different speaker (DS) LR calculations. Two LR tests were run for each tempo measure individually, such that the first iteration used the first 15 speakers as SS comparisons and the second 15 speakers as DS comparisons. The second iteration reversed the group roles. The outputs for the two tests were then combined, which resulted in 30 SS comparisons and 420 DS comparisons per tempo measure.

The discriminant performance of the measures is evaluated in terms of equal error rate (EER) and log-LR cost (C_{llr}). EER provides a 'hard' accept-reject measure of validity. This is based on the point at which the percentage of false hits and the percentage of false misses are equal [5]. C_{llr} is a Bayesian error metric that quantifies the ability of a system to align correctly with the expected outcome of whether speech samples are produced by the same or different speakers [22]. C_{llr} provides a more 'gradient' measure of performance, unlike EER [23]. The EER and C_{llrs} for the five tempo measures were computed within Bio-Metrics [1], which requires only the SS and DS log LR scores as input. Calibration was not used in this preliminary analysis of relative discriminant power. Further work on a larger set of data is forthcoming and will look at calibration and discriminant power in greater detail.

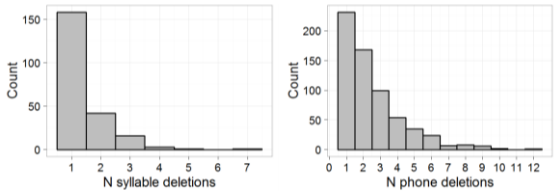
3. RESULTS

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 314 syllable deletions and 1598 phone deletions. This means that 4% of

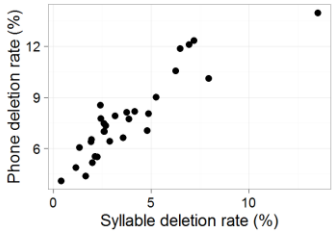
canonical syllables and 8% of canonical phones in the corpus lack a surface realisation. Syllable deletion occurs in 5% of words and 26% of memory stretches. As shown in Figure 1, the maximum number of deleted syllables per stretch is 7, but most stretches with deletions have just one missing syllable. 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. On average, 0.4 syllables are deleted per stretch. Phone deletion occurs in 25% of words and 73% of memory stretches.

Figure 1: Histograms for N syllable and phone deletions, excluding stretches with zero deletion



As shown in Figure 1, the maximum number of deleted phones per stretch is 12, but most stretches with deletions have between 1 and 4. The relationship between canonical (total range 9–53) and deleted phones is reasonably linear (r=0.53), although zero deletion is observed in stretches of up to 45 canonical phones. On average, 1.8 phones are deleted per memory stretch. 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: for example, zero syllable deletion maps to up to 6 phone deletions.

Figure 2: Deletion rates by speaker

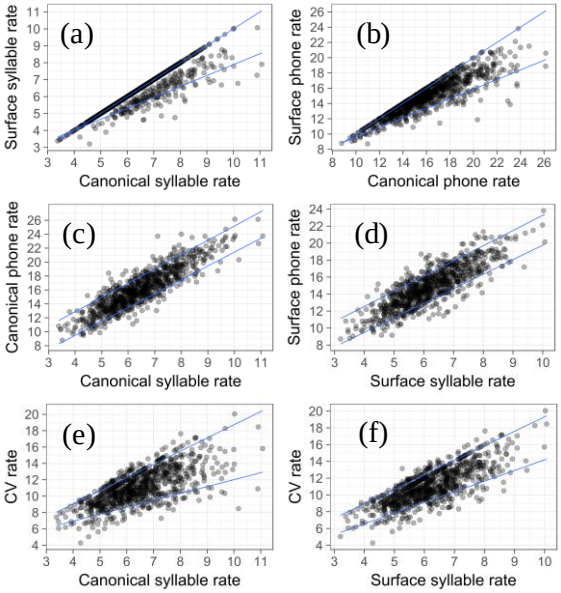


These deletion rates are comparable to those reported in previous corpus-based studies of English: [12] reports a 12.5% phone deletion rate; [26] reports deletion of 8% of 'acoustic landmarks'; [16] reports syllable deletion in 4.5–6% of words and phone deletion in 25%. This means that collectively, the speakers in our corpus do not appear to be unusually careful articulators, or speakers of a variety of English with little deletion. Therefore, correlations between canonical and surface rates in our corpus are likely to generalize to other English corpora, and they may indeed generalize beyond English: for example, [27] report deletion rates of 5% (syllables) and 8% (phones) for Dutch. We should note that there is considerable inter-speaker variation in deletion rates: as shown in Figure 2, syllable deletion rates vary from close to zero to 14% between speakers, and phone rates vary between 4% and 14%.

3.2. Correlations

Turning now to the correlations between our rate measures, Figure 3 shows scatter plots for the crucial comparisons. In plots (a) and (b), for canonical vs surface rates, we see a diagonal line of data points associated with stretches with no deletion; points below the diagonal reflect variable deletion. Overall the rates are highly correlated (CSR~SSR: r=0.91, CPR~SPR: r=0.90), even when zero-deletion stretches are excluded (CSR~SSR: r=0.89, CPR~SPR: r=0.90). In both plots we see evidence of the variability in surface rate increasing as canonical rate goes up. This is confirmed by quantile regression: the 0.9 and 0.1 quantile fit lines have a very similar intercept, but different slopes. This suggests that as speech tempo increases, the likelihood of 'massive reduction' [16] increases, although speakers do not invariably delete syllables and substantial numbers of phones.

Figure 3: Scatter plots for (a) CSR~SSR (b) CPR~SPR (c) CSR~CPR (d) SSR~SPR (e) CSR~CVR and (f) SSR~CVR with 0.1 and 0.9 quantile regression lines



In plots (c) and (d), for syllable vs phone rates, we see consistent relationships through the tempo range. Canonical rates are slightly more closely correlated than surface rates (CSR~CPR: r=0.89, SSR~SPR:

r=0.84); this is due to the relative weakness of correlation between phone and syllable deletions mentioned above. Plots (e) and (f) show that CV rate and syllable rates are less strongly correlated, at r=0.68 (CSR) and r=0.73 (SSR), with the quantile fit lines showing that CV rate variability increases as tempo increases, particularly against CSR.

3.3. Discriminating power

The results comparing the discriminant power of the five speaking tempo measures (CSR, SSR, CPR, SPR, and CVR) are summarised in Table 2. For C_{llr}, values close to zero indicate a good system performance; values above 1 a poor performance. The results are very similar to those found in [9]. They suggest that tempo is a relatively poor speaker discriminant regardless of methodology, as it is characterized by rather high EERs and C_{llrs} close to 1. Comparing the results of the five systems against each other, we can see that CSR, SSR, CPR, and SPR perform relatively similarly for both EER and C_{llr}; CVR, however, has a much higher EER than the other four systems as well as the highest C_{llr} (along with SSR). The best performing system in terms of C_{llr} is CPR, while CSR performs the best in terms of having the lowest EER.

Table 2: Performance of speaking tempo measures

Measure	EER	C_{llr}
CSR	28.1%	0.88
SSR	31.1%	0.89
CPR	33.5%	0.86
SPR	32.5%	0.87
CVR	37.5%	0.89

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.

We found very close relationships among the four syllable and phone rates. As might be expected, surface rates are on average lower than canonical rates, due to syllable and phone deletions. Surface rates become more widely dispersed as tempo goes up; we attribute this to the likelihood of substantial deletion increasing with overall tempo, while substantial deletion does not become the norm. Canonical and surface measures are very strongly correlated, in the region of r=0.90, even if zero deletion stretches are excluded. Corresponding syllable and phone rates are also strongly correlated, above r=0.80. The four measures yield very similar log likelihood values, with canonical rates marginally outperforming surface rates.

CV rate was introduced by [8] 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 in the region of r=0.70, and it is the poorest measure in terms of discriminating power.

The confidence with which we can make methodological recommendations of course depends on how representative our relatively small corpus of SSBE 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 relationship between canonical and surface rates should also generalise beyond this study. Assuming that it does, we can suggest that in analyses of at least English data in which speech tempo functions as an independent variable, quantifying tempo through surface or canonical syllable or phone rates is not likely to be consequential for analysis outcomes. Similarly, our results suggest that for forensic purposes there is very little difference between the four rate measures. In analyses where speaker comparison is central, the *relationship* between surface and canonical rate may well be informative: in our corpus, deletion rates vary considerably by speaker, with one speaker standing out as particularly prone to 'massive reduction'. It seems likely that speaker differences become clearer as tempo goes up. This warrants more detailed analysis of inter- and intra-speaker variation in syllable and phone deletion.

With reference to CV rate, we can question its efficiency as an alternative to syllable rate given that its calculation requires phone-level segmentation, and in our corpus, phone rates are more closely correlated with syllable rates than CV rate is. One possible implication of the latter finding is that speech rate estimators that depend on the automatic identification of acoustic correlates of syllables, and are generally evaluated against manual syllable rate calculation [3, 7, 14, 20], may in fact yield very similar output distributions to 'rough-and-ready' phone rate calculation using a general-purpose forced alignment system such as WebMAUS.

5. ACKNOWLEDGEMENTS

This research was supported by Leverhulme Trust Research Grant RPG-2017-060. We would like to thank Phil Harrison for troubleshooting on MatLab code.

6. REFERENCES

[1] Bio-metrics 1.5 performance metrics software, Oxford Wave Research ltd.

[2] Aitken, C. G., Lucy, D. 2004. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* vol. 53, pp. 109-122.

[3] Bakker, K., Brutten, G. J., McQuain, J. 1995. A preliminary assessment of the validity of three instrument-based measures for speech rate determination. *Journal of Fluency Disorders* vol. 20, pp. 63-75.

[4] Boersma, P., Weenink, D. 2017. Praat: Doing phonetics by computer. www.praat.org.

[5] Brümmer, N., Du Preez, J. 2006. Application-independent evaluation of speaker detection. *Computer Speech & Language* vol. 20, pp. 230-275.

[6] Dankovičová, J. 1997. The domain of articulation rate variation in Czech. *Journal of Phonetics* vol. 25, pp. 287-312.

[7] de Jong, N. H., Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* vol. 41, pp. 385-390.

[8] Dellwo, V., Ferrange, E., Pellegrino, F. 2006. The perception of intended speech rate in English, French, and German by French speakers. *Proceedings of the 3rd International Conference on Speech Prosody*, Dresden.

[9] Gold, E. 2014. *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*, PhD thesis, University of York.

[10] Gold, E., French, P. 2011. International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* vol. 18, pp. 293-307.

[11] Gold, E., Hughes, V. 2014. Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice* vol. 54, pp. 292-299.

[12] Greenberg, S. 1999. Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* vol. 29, pp. 159-176.

[13] Harrison, P. 2012. Matlab script, ss_ds_lrs.M, downloaded: May 2012.

[14] Heinrich, C., Schiel, F. 2010. Estimating speaking rate by means of rhythmicity parameters. *Proceedings of Interspeech 2010* Florence.

[15] Jessen, M. 2007. Forensic reference data on articulation rate in German. *Science & Justice* vol. 47, pp. 50-67.

[16] Johnson, K. 2004. Massive reduction in conversational American English. *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, Tokyo.

[17] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* vol. 45, pp. 326–347.

[18] Koreman, J. 2006. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America* vol. 119, pp. 582-596.

[19] Künzel, H. J. 1997. Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics* vol. 4, pp. 48-83.

[20] Martens, H., Dekens, T., Van Nuffelen, G., Latacz, L., Verhelst, W., De Bodt, M. 2015. Automated speech rate measurement in dysarthria. *Journal of Speech, Language, and Hearing Research* vol. 58, pp. 698-712.

[21] Morrison, G. S. 2007. Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using Multivariate-Kernel-Density Estimation, downloaded: December 2011.

[22] Morrison, G. S. 2009. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America* vol. 125, pp. 2387-2397.

[23] Morrison, G. S. 2009. The place of forensic voice comparison in the ongoing paradigm shift. *Written version of an invited presentation given at the 2nd International Conference on Evidence Law and Forensic Science* Beijing.

[24] 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* vol. 16, pp. 31-57.

[25] Pfitzinger, H. 1996. Two approaches to speech rate estimation. *Proceedings of the 6th Australian International Conference on Speech Science and Technology* Canberra.

[26] Shattuck-Hufnagel, S., Veilleux, N. 2007. Robustness of acoustic landmarks in spontaneously-spoken American English. *Proceedings of the 16th International Congress of Phonetic Sciences* Saarbrücken.

[27] 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. *Proceedings of Interspeech 2007* Antwerp.