MEASURED AND PERCEIVED SPEECH TEMPO: CANONICAL VS SURFACE SYLLABLE AND PHONE RATES

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

[1]University of Leeds, United Kingdom, [2]University of Glasgow, United Kingdom
l.plug@leeds.ac.uk

ABSTRACT

Studies that quantify speech tempo tend to use one of various available rate measures. The relationship between these measures and perceived tempo as elicited through listening experiments remains poorly understood. We assess how canonical and surface syllable and phone rates compare in terms of their mapping to listeners' tempo ratings. Native speakers of English rated short stretches of spontaneous speech for tempo; we modelled ratings for stimulus samples in which correlations between canonical and surface rates were low. Our findings suggest that listeners' ratings map most straightforwardly to *canonical* rate for syllables, but to *surface* rates for phones. We find little evidence of global tempo affecting the mappings, and consistent effects of stimulus duration. We discuss implications for the role of phoneme restoration in temporal processing.

Keywords: phonetics, speech perception, tempo, phoneme restoration, English

1. INTRODUCTION

Studies that quantify speech tempo through signal-based measurements tend to use one of various available measures. Researchers choose what to count [1, 2], what domains to count in [3, 4], and whether to count units as observed in their data, or as expected in canonical pronunciations [5, 6]. The corresponding measures may yield different figures for subsets of instances; however, few studies have investigated how closely the outputs of available measures are correlated, and how closely they map onto perceived tempo ratings elicited through listening experiments [7]. In this paper we focus on the relationship between syllable and phone rates on the one hand and listeners' tempo ratings on the other, implementing both rates in two ways: counting canonical units ('canonical rate'), and surface units ('surface rate').

Few studies have directly compared canonical and surface rates: for example, in [8-10], syllable and phone rates were calculated on the basis of either canonical or surface unit counts. [6] includes both canonical and surface rates, but it was 'impossible to decide for the best-fitting measure'.

However, assessing whether listeners' tempo judgements are most closely correlated with canonical or surface rates is of both practical and theoretical interest. While evidence for 'phoneme restoration'—listeners thinking they heard sounds that are masked or absent altogether in the signal—is robust [e.g. 11, 12, 13], it remains unclear whether this has an impact beyond word recognition. Assessing the impact of deletions on tempo perception [14] is a way of addressing this.

To date, two studies have explicitly attempted this assessment [5, 15]. In [15], a German utterance was produced at normal tempo, with few deletions, and at fast tempo with deletions. Both productions were manipulated to create a 'normal rate' version with deletions and a 'fast rate' one without. Listeners heard little difference between utterance versions with the same surface rate. In [5], German utterances were binned on the basis of phone rate measurements. Bins included 'fast-clear' (high rate, similar canonical and surface rates), 'normal-sloppy' (average rate, divergence between canonical and surface rates), and so on. Listeners judged pairs of utterances, selected across bins: 'fast-clear'~'fast-sloppy', 'fast-clear'~'normal-clear' and so on. Results suggested that listeners do perceive tempo differences between utterances with similar surface but different canonical rates: consistent with phoneme restoration, some 'sloppy' utterances were perceived as faster than 'clear' ones despite similar surface rates. However, global tempo modulated this result: listeners were more consistent in perceiving difference when the utterances were both relatively fast. In the study we report on here, we aimed to build on [5, 15], focusing on English.

2. METHOD

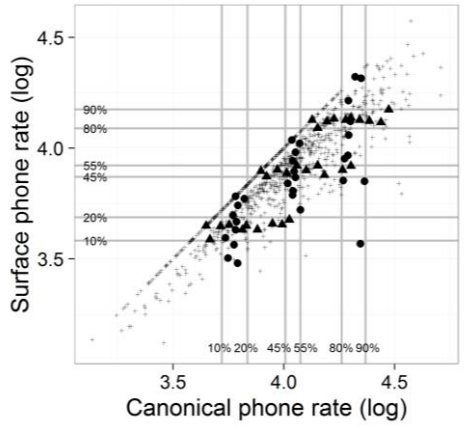
2.1. Stimuli

We selected stimuli from a set of 920 'memory stretches' extracted from the DyVIS corpus [16] by [17], produced by 30 male Standard Southern British English speakers aged 18–25. Mean stretch duration is 1.5 sec (range 0.5–2.7). We used WebMAUS [18] for segmentation, with a protocol for correcting substantive misparsings and under-identifications of phone deletion in frequent words. We derived

canonical and surface syllable rates (CSR, SSR) and phone rates (CPR, SPR) from the output segmentations.

As the four rates were very highly inter-correlated (r=0.84–0.91), selecting stretches that would allow for a meaningful comparison of the rates' mappings to perceptual tempo ratings was a methodological challenge. We selected three sets of 60 stimuli, each optimized for comparing two specific rate measures: (1) CSR~SSR; (2) CPR~SPR; (3) SSR~SPR. To create each set, we first inspected a scatterplot of the two (log) rates in all 920 stretches, as in Figure 1 for CPR~SPR. Here, points on the diagonal line have identical canonical and surface values; points below have varying amounts of deletion. For each rate, we identified the 10–20%, 45–55% and 80–90% quantile ranges to represent slow, medium and fast rates respectively. Within each of these, we selected 10 data points that were as widely dispersed in the 'comparison' rate's range as possible, and including one point with identical values for the two rates (i.e. no deletion). For Figure 1, this procedure yields 30 stimuli that are minimally variable in CPR but maximally variable in SPR (10 low CPR, 10 mid, 10 high: black dots) and 30 stimuli that are minimally variable in SPR but maximally variable in CPR (10 low, 10 mid, 10 high: triangles). We followed these steps for comparisons (1), (2) and (3) in turn. We anticipated that the subsets of stimuli within which variation was minimized on one rate but maximized on another would allow for meaningful comparisons of mappings to perceptual ratings. Moreover, sampling at low, mid and high rates might allow us to assess the impact of global tempo on the relationship between the alternative measured rates.

Figure 1: Scatterplot for CPR~SPR, with quantile ranges; black dots and triangles are selected stimuli



2.2. Tempo rating task

We elicited perceptual tempo ratings using an on-screen interface similar to that of [8], implemented in PsychoPy2 [19]. The stimuli in each set of 60 were presented together on one screen in the form of a vertical line of coloured dots in the centre of the screen. When the participant clicked on a dot, an orthographic transcription of the stimulus appeared on the screen, and the corresponding audio played (over headphones). The participant's task was to move each dot along a horizontal reference line to reflect its perceived tempo. Vertical gridlines and the labels 'Slowest, Slower, Average, Faster, Fastest' aided orientation. Stimuli appeared in the same order for all participants. Participants could listen to stimuli repeatedly.

2.3. Participants and production tasks

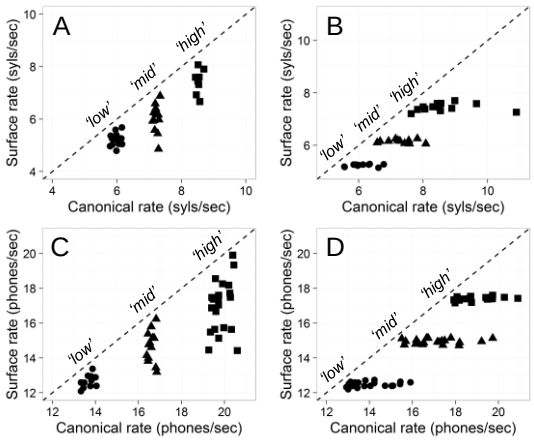
36 monolingual native English speakers (31 female; aged 18–36) were recruited at Leeds. All reported normal hearing, and all received payment.

As tempo perceptions might be informed by listeners' production tendencies [20], participants completed three tasks adopted from prior studies. In the first (e.g. [21]), participants repeated /pa/ at a 'comfortable rate' for 10 seconds. In the second (e.g. [20]), participants were presented with five sentences (from the *Rainbow passage* [22]) in turn. They memorized each sentence, then tapped the space bar to reveal a blank screen and produced the sentence (see [23]). In the third task (e.g. [24]), participants tapped the index finger of their dominant hand on a laptop touchpad for 20 seconds at a 'comfortable rate'. We extracted /pa/ rates, canonical syllable rates and tap rates per second.

2.4. Analysis methods

Dot placements were extracted as ratings on a scale between –500 and 500, with 0 corresponding to the dot's original position and a perception of 'average speed', –500 meaning maximally slow and 500 meaning maximally fast. We analysed the ratings through fitting linear mixed effects models using the lme4 package [25] in R [26]. Participant identities were treated as random intercepts. We report models with raw rate values; log rates revealed the same patterns. We focus on the canonical vs surface comparisons (CSR~SSR, CPR~SPR). To make the relevant analysis samples as large as possible, we pooled stimuli from the total stimulus set (N=180) that fell within relevant quantile ranges. We excluded stimuli with identical canonical and surface rate values and narrowed quantile ranges where relevant to keep correlations in the smallest relevant subsets below r=0.30 to ensure we could treat canonical and surface rates as orthogonal. The samples are shown in Figure 2; smallest relevant subsets are labelled 'low', 'mid' and 'high'.

Figure 2: Analysis samples for CSR~SSR (A: variable SSR, B: variable CSR) and CPR~SPR (C: variable SPR, D: variable CPR); dashed lines mark equivalence of the two rates



3. RESULTS

We found the overall distribution of ratings (N=36×180=6480) to be close to symmetrical with a large majority (85%) between –200 and 200. Participants varied in how widely they dispersed their ratings. Below we present results for analysis samples A, B, C and D in turn. Our approach was to fit a base model with a random intercept for participant; then assess the relevance of control variables (production measures, screen, screen position, stimulus duration) as fixed effects, keeping only those that significantly contributed to model fit; then assess whether the relevant rate variables improved model fit further. In the last step, we first checked whether the rate that varied most widely in the stimulus sample improved model fit, and then compared the fit of the resulting model to that of a model containing the more stable rate instead. In what follows we list coefficients for significant duration and rate effects only ($p<0.05$).

3.1. Sample A (CSR~SSR)

In sample A, SSR is variable; CSR is relatively stable. Lower SSR values reflect more syllable deletions.

We modelled ratings across the sample (N=1296), including CSR quantile range ('low', 'mid', 'high') as a factor to minimize the potential effect of CSR. The optimal model has effects for quantile range, position, log duration (β=–70.33, se=7.87, |t|=8.9: longer stimuli are rated slower) and CSR (β=136.85, se=38.16, |t|=3.59: stimuli with higher CSR are rated faster even when quantile range is accounted for). Including SSR instead of CSR results in poorer model fit, and the effect of SSR is negative (β=–29.37, se=9.39, |t|=3.13): stimuli with more syllable deletions were rated faster. We modelled ratings within the quantile ranges following the same procedure. The optimal model for the 'low' subset (N=540) has effects for screen, screen position and log duration (β=–97.26, se=13.85, |t|=7.02); neither rate improves model fit. The optimal model for the 'mid' subset (N=432) has effects for screen, log duration (β=–33.62, se=13.28, |t|=2.53) and SSR (β=–34.03, se=10.16, |t|=3.34); note that the effect of SSR is negative. The optimal model for the 'high' subset (N=324) has effects for log duration (β=–72.46, se=17.74, |t|=4.08) and CSR (β=337.28, se=98.62, |t|=3.42); SSR is non-significant added instead of CSR.

In sum, sample A provides little evidence of SSR being informative in modelling ratings; the evidence we do find points towards stimuli with more deletions sounding faster—in effect, orientation to *canonical* rate. We also find evidence for CSR being informative, despite its low variability.

3.2. Sample B (CSR~SSR)

In sample B, CSR is variable; SSR is relatively stable. Higher CSR values reflect more syllable deletions.

As above, we modelled ratings across the sample (N=1296), including SSR quantile range as a factor. The optimal model has effects for quantile range, screen, position, log duration (β=–43.12, se=8.42, |t|=5.12) and CSR (β=38.52, se=6.72, |t|=5.73). Including SSR instead of CSR results in significantly poorer fit. The optimal model for the 'low' subset (N=396) has fixed effects for log duration (β=–101.91, se=16.96, |t|=6.00) and SSR (β=494.00, se=156.76, |t|=3.151). CSR is also significant instead of SSR (β=39.20, se=19.13, |t|=2.05), but the resulting model has poorer fit. The optimal model for the 'mid' subset (N=396) has effects for screen and CSR (β=73.65, se=13.48, |t|=5.46). SSR is non-significant when added instead of CSR. The optimal model for the 'high' subset (N=504) has effects for screen, position and CSR (β=69.08, se=10.24, |t|=6.75). Adding SSR instead of CSR results in significantly poorer fit.

In sum, sample B provides clear evidence of CSR being informative in modelling ratings, although SSR shows some significance too. The effects of CSR are all positive, consistent with listeners hearing stimuli with more syllable deletions as faster.

3.3. Sample C (CPR~SPR)

In sample C, SPR is variable; CPR is relatively stable. Lower SPR values reflect more phone deletions.

As above, we modelled ratings across the sample (N=1908) with CPR quantile range as a factor. The optimal model has effects for quantile range, screen, log duration (β=–78.24, se=6.05, |t|=12.93) and SPR (β=11.01, se=2.26, |t|=4.87). CPR is non-significant when added instead of SPR. The optimal model for the 'low' subset (N=504) has effects for screen, log

duration (β=–43.85, se=15.13, |t|=2.90) and SPR (β=41.27, se=9.10, |t|=4.54). CPR is significant added instead of SPR (β=53.54, se=23.88, |t|=2.24), but the resulting model has poorer fit. The optimal model for the 'mid' subset (N=504) has effects for screen, log duration (β=–88.95, se=13.57, |t|=6.56) and CPR (β=148.70, se=41.82, |t|=3.56). SPR is non-significant added instead of CPR. The optimal model for the 'high' subset (N=900) has effects for screen, log duration (β=–63.63, se=8.19, |t|=7.77) and SSR (β=10.13, se=2.65, |t|=3.82). CPR is non-significant added instead of SPR.

In sum, sample C provides clear evidence of SPR being informative in modelling ratings, although CPR shows some significance too. The effects of SPR are all positive, consistent with listeners hearing stimuli with fewer phone deletions as faster.

3.4. Sample D (CPR~SPR)

In sample D, CPR is variable; SPR is relatively stable. Higher CPR values reflect more phone deletions.

As above, we modelled ratings across the sample (N=2160) with SPR quantile range as a factor. The optimal model has effects for quantile range, screen, position, log duration (β=–69.51, se=6.28, |t|=11.07) and SPR (β=142.32, se=22.43, |t|=6.35). CPR is non-significant added instead of SPR. The optimal model for the 'low' subset (N=756) has effects for screen, log duration (β=–59.71, se=13.19, |t|=4.53) and SPR (β=78.86, se=39.39, |t|=2.00). CPR is non-significant added instead of SPR. The optimal model for the 'mid' subset (N=720) has effects for screen, log duration (β=–93.29, se=9.70, |t|=9.61) and CPR (β=268.62, se=30.57, |t|=8.78). CPR is significant added instead of CPR (β=–9.75, se=4.92, |t|=1.98), but its effect is both very weak and negative, suggesting that stimuli with more phone deletions were rated as slower. The optimal model for the 'high' subset (N=684) has effects for position, log duration (β=46.27, se=9.94, |t|=4.65) and CPR (β=–19.60, se=4.31, |t|=4.54), again suggesting that stimuli with more phone deletions were rated as slower. SSR is non-significant added instead of CPR.

In sum, the evidence that we find of CPR being informative in modelling ratings for this sample points towards stimuli with more deletions sounding faster—in effect, orientation to *surface* rate. We also find evidence for SPR being informative, despite its low variability. The latter is consistent with the effects of SPR in modelling sample C ratings.

3.5. Further modelling

The models above suggest that CSR outperforms SSR (samples A, B), while SPR outperforms CPR (C, D). Given this, it would seem reasonable to compare CPR and SPR in modelling A responses, and CSR and SSR in modelling D responses. Unfortunately, our design does not allow for these comparisons, as in sample A, CPR and SPR are correlated at r=0.84, and in sample D, CSR and SSR are correlated at r=0.90.

4. DISCUSSION

In this study we set out to assess whether listeners' tempo judgements are most closely correlated with canonical or surface rates, for both syllables and phones. Our sampling and analysis methods have revealed a complex picture. For syllables, canonical rate maps most closely to listeners' tempo ratings. Stimuli with syllable deletions were rated faster than their surface syllable rate predicted. This can be taken as evidence for listeners restoring missing syllables in making tempo judgements, in line with [5]. For phones, however, surface rate maps most closely to listeners' tempo ratings: stimuli with phone deletions were rated slower than their canonical segment rate predicted. This provides no evidence for listeners restoring missing phonemes, in line with [15]. Of course, syllable deletions entail phone deletions, while phone deletions do not entail syllable deletions. Assuming our results are robust, listeners might ignore phone deletions in assessing the tempo of an utterance with all canonical syllables realized, while counting any missing syllables. This would mean that phone deletions become consequential for tempo perception when they contribute to syllable deletions. This hypothesis is worth testing in future work.

Our modelling within low, mid and high rate ranges has revealed that the patterns summarized above are mostly consistent across subsamples. In sample A, SSR outperforming CSR in the 'low' range might suggest, in line with [5], that listeners restore missing syllables more when processing speech at rates that are normally associated with regular syllable deletion. In sample C, CPR outperforming SPR in the 'mid' subset is difficult to account for.

Finally, in line with [5] our analysis revealed no evidence of participants' performance in production tasks co-varying with tempo ratings. We did observe a negative effect of stimulus duration in most analysis samples: listeners heard longer stimuli as slower independent of measured rates. This raises interesting questions about window size in listeners' online temporal processing [8], and warrants studies in which stimulus duration is varied while rates are controlled.

5. ACKNOWLEDGEMENTS

This work was supported by Leverhulme Trust Research Grant RPG-2017-060.

6. REFERENCES

[1] 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.

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

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

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

[5] 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.

[6] Den Os, E. 1985. Perception of speech rate of Dutch and Italian utterances. *Phonetica* vol. 42, pp. 124-134.

[7] Plug, L., Smith, R. 2017. Phonological complexity, segment rate and speech tempo perception. *Proc. Interspeech 2017*, Stockholm.

[8] Pfitzinger, H. 1999. Local speech rate perception in German speech. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco.

[9] Vaane, E. 1982. Subjective estimation of speech rate. *Phonetica* vol. 39, pp. 136-149.

[10] Gibbon, D., Klessa, K., Bachan, J. 2015. Duration and speed of speech events: A selection of methods. *Lingua Posnaniensis* vol. 56, pp. 59-83.

[11] Warren, R. M. 1971. Phonemic restoration in speech perception. *Journal of the Acoustical Society of America* vol. 49, pp. 85-&.

[12] Mitterer, H., Yoneyama, K., Ernestus, M. 2008. How we hear what is hardly there: Mechanisms underlying compensation for /t/-reduction in speech comprehension. *Journal of Memory and Language* vol. 59, pp. 133-152.

[13] Kemps, R., Ernestus, M., Schreuder, R., Baayen, H. 2004. Processing reduced word forms: The suffix restoration effect. *Brain and Language* vol. 90, pp. 117-127.

[14] Baayen, R. H., Shaoul, C., Willits, J., Ramscar, M. 2016. Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language Cognition and Neuroscience* vol. 31, pp. 106-128.

[15] Reinisch, E. 2016. Natural fast speech is perceived as faster than linearly time-compressed speech. *Atten Percept Psychophys* vol. 9, p. 9.

[16] 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.

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

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

[19] Peirce, J. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* vol. 2.

[20] Schwab, S. 2011. Relationship between speech rate perceived and produced by the listener. *Phonetica* vol. 68, pp. 243-255.

[21] Ruspantini, I., Saarinen, T., Belardinelli, P., Jalava, A., Parviainen, T., Kujala, J., Salmelin, R. 2012. Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2-3 Hz. *Journal of Neuroscience* vol. 32, pp. 3786-3790.

[22] Cartwright, L. R., Lass, N. J. 1975. Psychophysical study of rate of continuous speech stimuli by means of direct magnitude estimation scaling. *Language and Speech* vol. 18, pp. 358-365.

[23] Dilley, L. C., Pitt, M. A. 2010. Altering context speech rate can cause words to appear or disappear. *Psychol Sci* vol. 21, pp. 1664-70.

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

[25] Bates, D., Maechler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* vol. 67, pp. 1-48.

[26] R Development Core Team 2008. R: A language and environment for statistical computing.