Schwa deletion and perceived tempo in English

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

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

Abstract

We report on an experiment aimed to test the hypothesis that listeners orient to canonical forms when judging the tempo of reduced speech. Orientation to canonical forms should yield higher tempo estimates than orientation to surface phone strings when canonical phones are deleted. We tested the hypothesis for English, capitalizing on the fact that the non-realization of schwa in an unstressed syllable (e.g. *support*) may result in a surface phone string associated with a different word than the intended one (*sport*). We presented listeners with sentences containing ambiguous surface realizations, along with orthographic representations which convinced some that they were listening to disyllabic words (*support* etc.) and others that they were listening to monosyllabic ones (*sport* etc.). Asking listeners to judge the tempo of the sentences allowed us to assess whether the difference in imposed lexical interpretation had an impact on perceived tempo. Our results reveal the predicted effect of the imposed interpretation: sentences with a 'disyllabic' interpretation for the ambiguous word form were judged faster than (the same) sentences with a 'monosyllabic' interpretation.

Index Terms: speech perception, tempo, phonetic reduction, deletion, canonical forms, English

1. Introduction

Phonetic reduction phenomena, including phone and syllable deletion, are ubiquitous in normal speech [1, 2]. There is considerable evidence to support the notion that full pronunciation forms—'canonical forms'—have a different status from reduced forms in speech perception. For example, listeners' phonotactic generalizations appear to be based on full pronunciation forms [3], and listeners may report hearing phonemes that are absent from the signal due to assimilation or other reduction processes [4-6]. The latter suggests that the perception of reduced forms involves the activation of their corresponding canonical forms, and that this activation may override bottom-up information regarding the presence or absence of acoustic cues [5].

In this paper we address the impact of deletions on speech tempo perception. Given that listeners' tempo ratings tend to correlate closely with articulation rate measures in phones or syllables per time unit [7-10], it seems reasonable to assume that something akin to phone or syllable counting is involved in speech tempo estimation. Given the findings above, we might expect that listeners show an orientation to full pronunciation forms in doing the relevant counts. To date only a few experimental studies have attempted to test this expectation: most notably [11] and [12].

[11] maps 'intended' and 'observed' phone rates—that is, rates based on phone counts in full pronunciation forms and rates based on phone counts in surface forms, respectively—to listeners' tempo judgements of spontaneously-produced German intonation phrases. Phone deletions yield divergence between the two rates: stretches with deletions have a higher intended rate than observed rate. The results show some evidence for listeners' orientation to canonical forms in that listeners perceive tempo differences between utterances with similar observed rates but different intended rates. However, they also perceive tempo differences between utterances with similar intended rates but different observed rates, suggesting that *both* rates are oriented to. [12] reports two experiments in which listeners judged the tempo of naturally-produced normal and fast speech, and speech that results from linear tempo manipulations—including speech produced at a normal tempo, with few deletions, sped up. Neither experiment shows straightforward evidence for orientation to canonical forms.

Both [11] and [12] propose that listeners may perceive utterances with more deletions as faster not because they are counting deleted phones or syllables in estimating tempo, but because they know that phonetic reduction is more common at higher speaking rates. In effect, listeners may be making a stylistic judgement, linking both high speech tempo and high phonetic reduction to 'casual speech, the register in which reduction is most common' [1].

In the experiment we report on here we aimed to test the hypothesis that English listeners orient to canonical forms in making speech tempo judgements—in a way that made it unlikely that listeners were making a stylistic judgement. We achieved this by using the speech of one speaker only, speaking in one style. We capitalized on the fact that in English, the non-realization of schwa in an unstressed syllable may create ambiguity as to the intended word form. For example, schwa deletion in *support* results in a surface realization that is highly similar to that of *sport*. We presented listeners with utterances containing such surface realizations, using orthographic primes to convince some that they were listening to disyllabic words (*support* etc.) and others that they were listening to monosyllabic ones (*sport* etc.). Asking listeners to judge the tempo of the utterances allowed us to test whether the difference in interpretation had an impact on perceived tempo.

2. Method

2.1. Participants

70 native speakers of British English (56 female) in the age range 18–35 (mean=22) participated in this experiment. All self-reported as having grown up in a monolingual household and having no known hearing problems. All provided informed consent in line with institutional ethics guidance.

2.2. Stimuli

We identified the word pairs in Table 1 as potential loci of lexical ambiguity due to schwa deletion. In all of these, the disyllabic pair member contains a pre-stress schwa. In this position, schwa deletion has been shown to be a gradient process [13, 14]; we therefore assume that the disyllabic pair member has one canonical form which includes schwa [15].

Table 1: *Lexical pairs used in the experiment.*

lexical pair	[–schwa]	[+schwa]
blow~below	/ˈbləʊ/	/bəˈləʊ/
claps~collapse	/ˈklaps/	/kəˈlaps/
clean~Colleen	/ˈkliːn/	/kəˈliːn/
clone~cologne	/ˈkləʊn/	/kəˈləʊn/
clued~collude	/ˈklu:d/	/kəˈlu:d/
cream~Kareem	/ˈkɹiːm/	/kəˈɹiːm/
cress~caress	/ˈkɹɛs/	/kəˈɹɛs/
crowed~corrode	/ˈkɹəʊd/	/kəˈɹəʊd/
dried~deride	/ˈdɹaɪd/	/dəˈɹaɪd/
drive~derive	/ˈdɹaɪv/	/dəˈɹaɪv/
griller~gorilla	/ˈgɹɪlə/	/gəˈɹɪlə/
Kroner~corona	/ˈkɹəʊnə/	/kəˈɹəʊnə/
plight~polite	/ˈplaɪt/	/pəˈlaɪt/
prayed~parade	/ˈpɹeɪd/	/pəˈɹeɪd/
sport~support	/ˈspɔːt/	/səˈpɔːt/
train~terrain	/ˈtɹeɪn/	/təˈɹeɪn/

We also included morphologically related pairs such as *sports~supports*, *sported~supported* and *sporting~supporting* for *sport~support*. This resulted in a list size of N=26 lexical pairs. For each, we constructed a carrier sentence in which either pair member was similarly semantically fitted. This was verified through an online semantic acceptability survey (25 participants). Examples are given in (1). The same carriers were used to construct comparison sentences that contained no ambiguity (N=52), illustrated in (2).

(1) a. He wanted to *sport~support* it.
b. He spotted the *Kroner~corona.*
c. He predicted the *claps~collapse.*

(2) a. He wanted to *start~restart* it.
b. He spotted the *killer~instiller.*
c. He predicted the *terms~concerns.*

All sentences were produced by a female speaker of British English (age 27) who grew up in the South East of England. Recordings were done in a soundproof room with a cardioid condenser microphone, at a sampling rate of 44100 Hz. The speaker produced each sentence once at a normal pace and once at a fast pace with as little variation in pitch and loudness across sentences as feasible. Following an exploratory listening survey (10 participants), we spliced the speaker's fast productions of [+schwa] words (*support* etc.) into their corresponding normal-pace carriers as a starting point for creating ambiguous sentence forms. We then manipulated schwa duration and, where relevant, plosive VOT in the crucial word forms to create multiple candidate ambiguous forms. The manipulations were done manually in *Praat* [16], reducing durations in 25% steps.

In a further listening survey (37 participants), listeners heard resulting candidate forms while reading their orthographic transcriptions, and judged for each how well the audio and the written form matched. The participants were divided into two groups such that for each candidate ambiguous form, one group of participants judged how well the audio matched the [–schwa] orthography (*sport* etc.) while the other group judged how well it matched the [+schwa] orthography (*support* etc.). For each lexical pair in Table 1, the sentence form with the smallest difference in goodness rating between the two groups was selected for inclusion in the main experiment. For example, the most ambiguous sentence form containing *corrode* was found to be the token manipulated to have 100% VOT for [k] and 25% duration for [ə]; it was judged acceptable at an average of 83/100 by participants who thought they were hearing *they crowed on the roof* and at an average of 86/100 by participants who thought they were hearing *they corrode on the roof*.

2.3. Task design

We used a gradient implementation of an *abx* task in which participants rated the tempo of an *x* sentence form relative to two realizations of a different sentence: a slow realization (*a*) and a fast realization (*b*). Participants were presented with a horizontal scale; the slow comparison realization (*a*) was at the left end and the fast comparison realization (*b*) at the right end. The participants' task was to place the *x* sentence form on the scale according to its tempo relative to the two comparison realizations. The 26 sentences containing an ambiguous word form (*sport~support* etc.) formed the crucial set of *x* sentences; the 52 similar sentences without lexical ambiguity were used as comparison sentence forms, as well as to construct filler trials.

Figure 1 shows the interface, coded in *PsychoPy2* [17], for one trial. The *x* sentence is *He wanted to sport~support it*; the comparison sentence *It was that dream again*. The leftmost square represents a slow realization of *It was that dream again* (*a*); the rightmost one a fast realization (*b*). The arrow underneath the middle square aided the participant in dragging the square to position it on the spectrum.

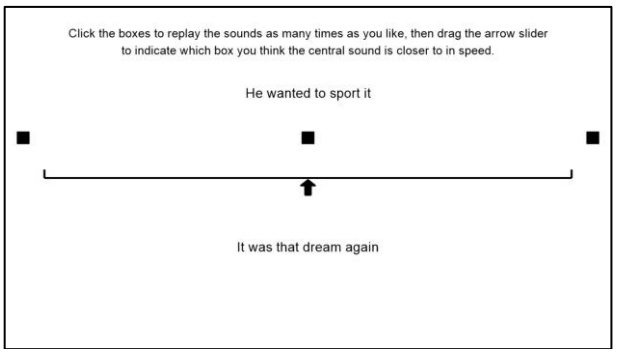

Figure 1: *Visual interface for one experimental trial.*

We predicted that listeners who interpreted an *x* sentence as containing a [–schwa] word (e.g. *sport*, as in Figure 1) would rate the sentence form as slower than those who interpreted the same sentence as containing the corresponding [+schwa] word (*support*). To test this prediction, we presented participants with the orthography of both the *x* sentence and the *a/b* sentence on each trial, and we divided participants into two groups. Group A participants were given a [–schwa] orthography for half of the 26 sentences and a [+schwa]

orthography for the remainder; for group B participants the halves were reversed. Participants were randomly assigned to Group A or Group B. Crucially, Group A and B participants heard the same audio stimuli: what differed was only the orthography of the *x* sentence.

We presented participants with each *x* sentence twice, in two experimental blocks. In Block 1, the *a/b* sentence consistently had the same number of syllables as the [–schwa] interpretation of the *x* sentence, while in Block 2, the *a/b* sentence consistently had the same number of syllables as the [+schwa] interpretation.

To create the slow and fast realizations of the *a/b* sentence forms, we calculated the syllable rate for each *x* sentence form and then resynthesized the *a/b* sentence form (using PSOLA in *Praat*) to create versions at 0.9 (*a*) and 1.1 (*b*) times this syllable rate. This process was done separately for each block, with the *x* rate calculated assuming a [+schwa] interpretation for Block 1, and a [+schwa] interpretation for Block 2. Informal piloting suggested that the 10% rate adjustments were easily perceivable without resulting in extremely slow or fast realizations.

We created 14 filler trials containing *x* and *a*/*b* sentences without any ambiguity, so that each block consisted of (26+14=)40 trials.

2.4. Quantitative analysis

Our central prediction was that participants who interpreted an *x* sentence as containing a [–schwa] word (*sport*) would rate the sentence form as slower than those who interpreted the same sentence as containing the corresponding [+schwa] word (*support*). To test this, we fitted linear mixed effects models using *lme4* [18] and *lmerTest* [19] in *R* [20], through a stepwise model fitting procedure [21]. The dependent variable was the participants' tempo ratings (*Rating*, N=3640), recorded on a 0–1000 scale with 500 representing the initial central placement of the stimuli in the visual interface. The crucial predictor variable was the *x* sentence orthography (*Orthography*): [–schwa] or [+schwa].

We included random intercepts for participant identity (*Participant*) and item identity (*Item*). The latter distinguishes the *x* sentence forms irrespective of the imposed interpretation. Each level of *Item* is repeated within subjects (as each *x* was presented in Blocks 1 and 2) and across subjects (as each *x* was presented with a [–schwa] or [+schwa] interpretation depending on the participant). As the experiment consisted of two blocks of trials and we varied the presentation order of the *a* and *b* sentence forms, we coded for *Block* (1 or 2), *Trial* (within blocks) and *Order* (*a* first or *b* first).

We included a number of control predictors. First, we derived two variables from each of the surveys described above (the semantic acceptability survey and the listening survey): a measure of the goodness of each sentence or crucial word form given its [–schwa] or [+schwa] interpretation, and a measure of the difference between the two for each sentence and audio stimulus (irrespective of its imposed interpretation). Second, we took acoustic measures of intensity and *f*0 level and span for each sentence, given the relevance of these parameters for tempo perception [22]. As none of these variables significantly predicted tempo ratings, we do not describe them in detail here.

3. Results

As seen in Figure 2, the distribution of *Rating* is skewed towards participants judging *x* as closer in tempo to *b*. We fitted our models on the raw values of *Rating* and on the results of a square root transformation. As the outcomes were the same, we present the modelling procedure using the raw values.

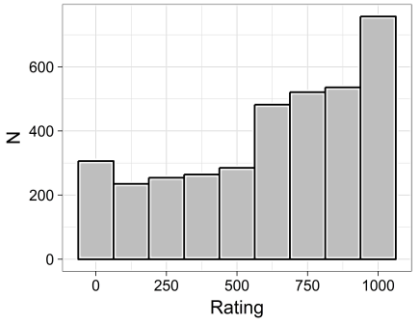

Figure 2: *Distribution of* Rating.

In modelling *Rating*, we started with a base model with random intercepts for *Participant* and *Item* and first assessed the predictive value of *Block*, *Trial* and *Order*. This revealed that adding *Block* to the model improved fit; adding *Trial* or *Order* did not. We then added our crucial predictor, *Orthography* and established that an interaction between *Block* and *Orthography* significantly improved fit. The final model is shown in Table 2 and its estimated effects are illustrated in Figure 3.

Table 2: *Summary of effects in the optimal model of* Rating. *For* Block*, '1' is the reference level; for* Orthography*, '[–schwa]' is the reference level.*

	Estimate	SE	df	t	p
(Intercept)	537.55	40.36	27.77	13.32	<0.001
Block '2'	119.82	11.48	3542	10.44	<0.001
Orthography '[+schwa]'	64.15	11.48	3542	5.59	<0.001
Block '2' * *Orthography* '[+schwa]'	–42.21	16.23	3542	–2.60	0.009

The model confirms that listeners' tempo ratings were significantly higher in Block 2 than in Block 1. They were also significantly higher for [+schwa] sentences than for [–schwa] sentences, in line with our prediction. The significant interaction between *Block* and *Orthography* reflects that the effect of *Orthography* is considerably greater in Block 1 than in Block 2.

4. Discussion

The results of this experiment confirm our prediction that listeners who were primed by orthography to interpret an ambiguous sentence form as containing a [–schwa] word (e.g. *sport*) would rate it as slower than those who were primed to interpret it as containing the corresponding [+schwa] word, which has an additional syllable (*support*): *Orthography* was a significant predictor of *Rating*. This can be taken as evidence for listeners doing more in estimating speech tempo than

mapping the content of a surface signal to time: activation of corresponding full pronunciation forms, or more abstract phonological representations based on them [5], explains our findings. We should note that it might also be possible that listeners judged *x* sentences with [+schwa] words as more 'casual' overall compared with *x* sentences with [–schwa] words, and associated 'casualness' with relatively high tempo. However, we believe that this interpretation is less convincing for our experimental design than for [11]'s design, which used stylistically highly variable stimuli sampled from spontaneous speech: in our experiment, all listeners were exposed to the same stimuli and the carrier sentence productions were minimally variable in terms of speech style. Our result is more likely to reflect that listeners' tempo judgements involve mapping signal content to time, and full pronunciation forms play a role in listeners' interpretation of signal content.

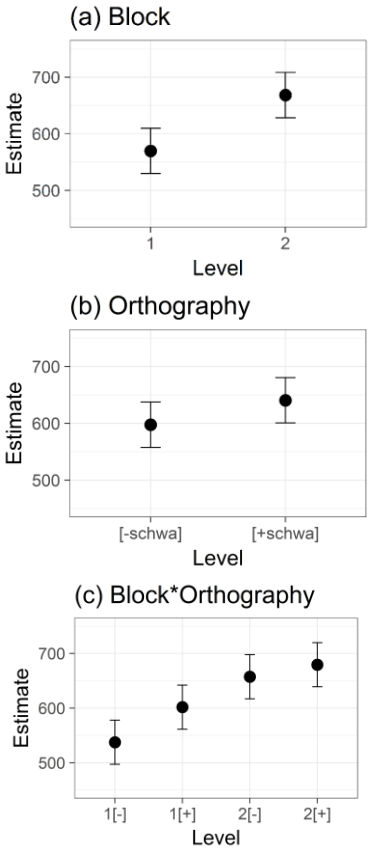

Figure 3: *Effect estimates in the optimal model of* Rating *(see Table 2).*

In relation to the observed effect of *Block* and its interaction with *Orthography*, our results suggest that the *Orthography* effect was considerably more robust in Block 1 than in Block 2. From one perspective, evidence consistent with orientation to the canonical form in Block 1 can be taken as particularly strong, as in Block 1 the *a/b* sentence did not contain a word with an unstressed schwa. This means that in Block 1, the *a/b* sentences could not be said to facilitate 'schwa restoration' through priming. In Block 2, the *a/b* sentences could be said to have this facilitating effect, as they contained words like *concerns, preceded, contrite* and so on. We do not have a straightforward explanation for the observed interaction, but it is possible that it is related to the general effect of *Block*: in Block 2, the *x* sentences were generally rated as closer to *b*, and perhaps participants made less fine distinctions towards the outer edges of the scale. The effect of *Block* may be due to our decision to use the syllable rate of the [–schwa] interpretation of the *x* sentence as reference for deriving the *a* and *b* syllable rates in Block 1, and the syllable rate of the [+schwa] interpretation in Block 2. This means that the *a* and *b* syllable rates were higher in Block 2 than in Block 1, perhaps prompting participants to generally estimate the *x* tempo as higher in Block 2. These considerations warrant further experiments using the same or similar stimuli in alternative designs.

5. Conclusion

We aimed to test the hypothesis that English listeners orient to canonical forms in making speech tempo judgements, using ambiguous audio stimuli for which two groups of participants were given different orthographical representations. Our results reveal the predicted effect of the imposed interpretation. While the results are complicated by effects associated with our experimental design, they are consistent with the idea that when phones or syllables are absent in surface forms due to deletion, but listeners are able to interpret the intended word forms, the latter's full pronunciation forms inform listeners' perceptions of speech tempo.

6. Acknowledgements

This research was supported by a Leverhulme Trust Research Grant (RPG-2017-060).

7. References

1. Ernestus, M., *Acoustic reduction and the roles of abstractions and exemplars in speech processing.* Lingua, 2014. **142**: p. 27-41.
2. Warner, N., *Reduced speech: All is variability.* WIREs Cognitive Science, 2019. **10**(4).
3. Bürki, A., *Variation in the speech signal as a window into the cognitive architecture of language production.* Psychonomic Bulletin & Review, 2018. **25**(6): p. 1973-2004.
4. Mitterer, H., K. Yoneyama, and M. Ernestus, *How we hear what is hardly there: Mechanisms underlying compensation for /t/-reduction in speech comprehension.* Journal of Memory and Language, 2008. **59**(1): p. 133-152.
5. Kemps, R., et al., *Processing reduced word forms: The suffix restoration effect.* Brain and Language, 2004. **90**(1-3): p. 117-127.
6. Mitterer, H. and L. Blomert, *Coping with phonological assimilation in speech perception: Evidence for early compensation.* Perception & Psychophysics, 2003. **65**(6): p. 956-969.
7. Pfitzinger, H. *Local speech rate perception in German speech*. in *Proceedings of the 14th International Congress of Phonetic Sciences*. 1999. San Francisco.
8. Vaane, E., *Subjective estimation of speech rate.* Phonetica, 1982. **39**(2-3): p. 136-149.
9. Den Os, E., *Perception of speech rate of Dutch and Italian utterances.* Phonetica, 1985. **42**: p. 124-134.

10. Gibbon, D., K. Klessa, and J. Bachan, *Duration and speed of speech events: A selection of methods*. Lingua Posnaniensis, 2015. **56**(1): p. 59-83.
11. Koreman, J., *Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech*. Journal of the Acoustical Society of America, 2006. **119**(1): p. 582-596.
12. Reinisch, E., *Natural fast speech is perceived as faster than linearly time-compressed speech*. Atten Percept Psychophys, 2016. **9**: p. 9.
13. Davidson, L., *Schwa elision in fast speech: Segmental deletion or gestural overlap?* Phonetica, 2006. **63**(2-3): p. 79-112.
14. Patterson, D., P.C. LoCasto, and C.M. Connine, *Corpora analyses of frequency of schwa deletion in conversational American English*. Phonetica, 2003. **60**(1): p. 45-69.
15. Bürki, A. and M.G. Gaskell, *Lexical representation of schwa words: Two mackerels, but only one salami*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2012. **38**(3): p. 617-631.
16. Boersma, P. and D. Weenink, *Praat: Doing phonetics by computer*. 2017: www.praat.org.
17. Peirce, J., *Generating stimuli for neuroscience using PsychoPy*. Frontiers in Neuroinformatics, 2009. **2**(10).
18. Bates, D., et al., *Fitting linear mixed-effects models using lme4*. Journal of Statistical Software, 2015. **67**(1): p. 1-48.
19. Kuznetsova, A., P.B. Brockhoff, and R.H.B. Christensen, *LmerTest package: Tests in linear mixed effects models*. Journal of Statistical Software, 2017. **82**(13): p. 1-26.
20. R Development Core Team, *R: A language and environment for statistical computing*. 2008.
21. Baayen, R.H., *Analyzing linguistic data: A practical introduction to statistics using R*. 2008, Cambridge: Cambridge University Press.
22. Feldstein, S. and R.N. Bond, *Perception of speech rate as a function of vocal intensity and frequency*. Language and Speech, 1981. **24**(Oct-): p. 387-394.