CLEAR SPEECH SOUNDS FAST: THE IMPACT OF SPEAKING MODE VARIATION ON PERCEIVED TEMPO

Leendert Plug[1], Yue Zheng[1], Rachel Smith[2]

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

ABSTRACT

Research has identified multiple acoustic cues for perceived speech tempo variation. Together, these warrant the proposal that speech which conveys more complex spectral information is heard as taking more time than speech conveying less complex information; if timing is controlled, the former sounds faster than the latter. We therefore hypothesize that hyper-articulated speech sounds faster than normal speech when articulation rates are controlled. We report on two listening experiments which address this hypothesis using clear and normal sentence productions. Experiment 1 assesses listeners' ability to separate tempo and speaking mode judgements; Experiment 2 uses the same stimuli to probe the direction of any effect of speaking mode variation on perceived tempo. The results confirm that compressed clear sentence productions sound faster than normal productions with the same duration, although the precise mechanism underlying this effect remains to be established.

Keywords: speech perception, tempo, clear speech, English

1. INTRODUCTION

Research on speech tempo perception has identified multiple non-temporal acoustic cues for perceived tempo variation: increases in *f*0 span and level, intensity level and vowel space size have all been shown to raise perceived tempo [1–4]. Generalizing across these findings, Weirich and Simpson [4] propose that 'there is a link between the perception of spectral information and the perception of time', such that '[a]n event that conveys more information in terms of acoustic parameters (a greater range in *f*0, a greater vowel space traversed) is perceived as taking more time than an event conveying less information'.

The acoustic parameters listed above are all implicated in variation along the 'H&H continuum' [5, 6], including shifts between normal and clear speaking modes [7]. According to [5, 6], speakers adjust their articulatory effort and precision on a continuum between loosely and firmly controlled ('hypo-' and 'hyper-') articulation on the basis of the estimated need for speech clarity given situational constraints. Many studies have investigated speakers' adjustments when asked to speak *clearly*—as if communicating in noise, or with a hearing-impaired or second-language listener [8–14]. These studies have shown that recurrent correlates of clear speech include increased *f*0 and intensity levels and *f*0 span, and greater dispersion of vowels in the F1–F2 space [7–9, 14]. Clear speech is also associated with a decrease in coarticulation, resulting in more easily delimitable, 'canonical' articulations [10, 12, 13, 15].

These findings warrant the hypothesis that when articulation rate is controlled, the non-temporal features of clear speech will make it sound fast relative to normal speech, as clear speech conveys more spectral information in the same time window [4]. A complicating factor is that clear speech is generally articulated slowly [7–9, 14], and listeners may draw on their knowledge of associations among non-temporal and temporal parameters in making tempo judgements [16, 17]. This would predict that when listeners recognize speech as clear, they are biased towards hearing it as relatively slow. Either way, we can hypothesize that speech mode variation between normal and clear has an impact on speech tempo perception when articulation rate is controlled.

We report on two experiments that address these related hypotheses. Both were pairwise comparison tasks with pairs of normal and clear sentence productions, controlled for articulation rate, serving as stimuli. In Experiment 1, participants judged whether the productions in each pair differed in tempo, speaking mode, both, or neither. This allowed us to assess the extent to which listeners can separate these two perceptual parameters and test the general hypothesis that speaking mode variation has an impact on tempo perception. In Experiment 2, participants heard the same pairs and judged how the productions differed in tempo; here their attention was not drawn to, nor were they asked to make judgements about, speaking mode. This allowed us to test the more specific hypothesis that clear speech sounds faster than normal speech when articulation rate is controlled.

2. EXPERIMENT 1

2.1. Participants

82 native British English speakers aged 18–35 years were recruited through *Prolific* (www.prolific.co). All reported normal hearing and passed a screening task in which they identified sentence production pairs with identical members. All were paid.

2.2. Stimuli

We used the LUCID Corpus [14], available via *SpeechBox* [18], to create our stimuli. The corpus includes a set of 144 sentences which were read by 40 Southern Standard British English speakers in both normal and clear speaking modes. For the normal mode, speakers were instructed to speak 'casually, as if talking to a friend'; for the clear mode, 'clearly as if talking to someone who is hearing impaired' [14]. We selected the ten sentences in Table 1 as produced by six speakers—four female and two male—who did not make large loudness and voice quality adjustments in speaking clearly, but whose 'clear' productions still had distinct articulatory and prosodic characteristics.

Table 1: Stimulus sentences with summary statistics for articulation rate (mean and SD, N=6) split by speaking mode.

Sentence	Articulation rate (sylls/sec)	
	normal	clear
The bear belongs to the children.	4.52 (0.69)	3.13 (0.67)
I've lost my box of pins.	5.36 (1.33)	3.47 (0.65)
The dog barked at the sheep.	4.41 (0.65)	3.32 (0.61)
The old lady ate the peach.	5.12 (0.73)	3.50 (0.26)
The music blared from the shack.	4.26 (0.35)	3.07 (0.26)
The pear belongs to the teacher.	5.12 (0.41)	3.30 (0.60)
The seat came with the car.	4.03 (0.20)	2.83 (0.54)
She's going to sue the firm.	4.52 (0.39)	3.09 (0.26)
The suit was full of holes.	4.49 (0.82)	3.08 (0.26)
Jonathan gave his wife a bush.	5.22 (0.79)	3.28 (0.59)

Acoustic analysis following [14] confirmed that across the 120 sentence productions, all clear ones are longer and more slowly articulated than their corresponding normal production (Table 1); most also have a higher Long-Term Average Spectrum and greater *f*0 dispersion. Segmentations produced by the BAS tools *G2P* and *WebMAUS* [19] and additional auditory analysis highlighted systematic differences in phone-level articulation between the normal and clear sentence productions consistent with descriptions of SSBE full forms and connected speech phonetics. For example, in the normal productions [t, k, g] are mostly unreleased when followed by another consonant: *dog barked* [dɒgˀbɑːkˀt]. In the clear sentence productions, these plosives are predominantly released.

For each speaker's production of each sentence in Table 1, we created four sentence pairs, as outlined in Table 2. Pair members were separated by a 1-second within-pair silence. Compressions were done through PSOLA in Praat [20]. The target of the compression of a clear production was always the duration of the matching normal production. In SPEED pairs, members differ in (utterance-level) articulation rate but not speaking mode. In PRECISION pairs, they differ in speaking mode but not articulation rate. In BOTH pairs, they differ in speaking mode and articulation rate. In NEITHER pairs, members are identical. To manage the number of trials per participant, we created four lists (N=120), counterbalancing for sentence and within-pair order.

Table 2: Stimulus pair types; '~' indicates that pairs were included in both possible orders.

Type	Pair members
SPEED	clear ~ $\text{clear}_{\text{compressed}}$
PRECISION	normal ~ $\text{clear}_{\text{compressed}}$
BOTH	normal ~ clear
NEITHER	normal – normal clear – clear

2.3. Task and procedure

We used *Gorilla* to run the experiment online [21]. Participants were assigned randomly to one of the four lists. In each trial, participants were informed on-screen that in the sentence production pair they were going to hear, the productions might sound different in the speed of the articulation, the precision of the articulation, both speed and precision, or neither (i.e. the same). They were asked to decide which of these four descriptions seemed most appropriate to them given the following audio. The production pair then played twice with a 2-second between-pair silence, before the four response options 'neither', 'speed', 'precision' and 'both' appeared on screen. The next trial began once the participant had submitted a response, with a 0.5-second between-trial silence. Trial order was randomised by participant.

2.4. Predictions

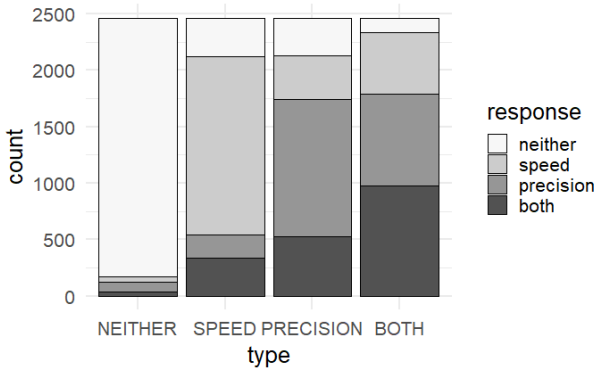
The hypothesis that speech mode variation has an impact on speech tempo perception when articulation rate is controlled yields several specific predictions. We expected participants to be less accurate at identifying PRECISION pairs ('precision') than at identifying SPEED pairs ('speed'). More specifically, we expected PRECISION pairs to attract more 'speed' and 'both' responses than SPEED pairs attract 'precision' and 'both' responses—indicating that

speaking mode differences trigger tempo percepts more than the other way around. We also expected PRECISION pairs to attract more 'speed' and 'both' responses than NEITHER pairs—indicating that participants hear precision differences as tempo differences, rather than simply producing occasional false alarms when no rate difference is present.

2.5. Results

Figure 1 shows the response proportions by stimulus type. Listeners are very good at identifying NEITHER pairs (93% 'neither'). For SPEED pairs, 'speed' is also the majority response (64% 'speed'). For PRECISION pairs, 'precision' is the majority response (50% 'precision'), but this majority is significantly smaller than that for SPEED pairs (χ^2= 108.64, df=1, p<0.001).

Figure 1: Cumulative bar chart of the Experiment 1 response proportions by stimulus type.



As predicted, PRECISION pairs attract significantly more 'speed' and 'both' responses than SPEED pairs attract 'precision' and 'both' responses (37% vs. 22%; χ^2=131.2, df=1, p<0.001). This is partly because the number of 'speed' responses for PRECISION pairs is almost twice that of 'precision' responses to SPEED (16%~8%). PRECISION pairs also attract significantly more 'speed' and 'both' responses than NEITHER pairs (37% vs. 4%; χ^2=843.60, df=1, p<0.001). Participants are least accurate at identifying BOTH pairs (40%): a majority of responses suggests participants were hearing difference in one parameter only (33% 'precision', 22% 'speed').

Having shown that speaking mode variation is heard as tempo variation more often than vice versa, we tested the direction of the effect in Experiment 2: is clear speech heard as faster, or slower, than normal speech with the same articulation rate?

3. EXPERIMENT 2

3.1. Participants

26 native British English speakers aged 18–35 years were recruited primarily from student cohorts. None had participated in Experiment 1. All reported normal hearing and passed the same screening task as in Experiment 1. Most were paid.

3.2. Stimuli, task and procedure

Experiment 2 was identical to Experiment 1 except for the on-screen instructions and response options. In each trial (N=120), participants were informed that the pair of productions might sound the same or different in tempo. They were asked to identify the faster pair member ('first', 'second', 'neither').

3.3. Predictions

Our key prediction concerned PRECISION pairs, in which articulation rate was identical but speaking mode differed. We expected participants to perceive a tempo difference in at least the same percentage of PRECISION trials as observed in Experiment 1 (37%), and we expected that when a tempo difference was heard, the (compressed) clear production would be heard as faster. Beyond this, we expected participants to veridically report articulation rate differences where these were present. That is, in SPEED pairs we expected the member with the higher articulation rate to be heard as faster, and in NEITHER pairs we expected the members to be heard as the same. In BOTH pairs, where uncompressed clear and normal productions were presented, we expected participants to accurately identify the clear pair member as slower.

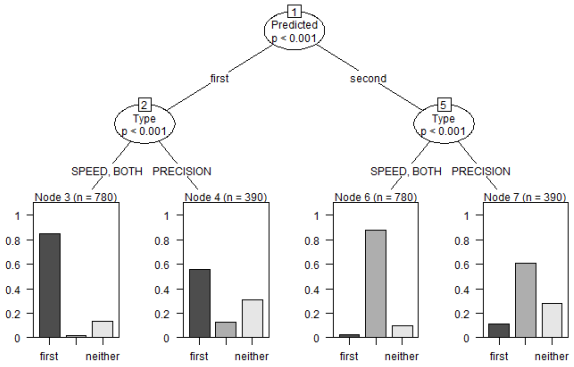
3.4. Results

Consistent with our prediction for PRECISION pairs, 71% of responses reflect a tempo difference; this is almost twice as many as in Experiment 1 (37%). The responses are split near-evenly as to which pair member ('first' or 'second') was heard as faster. While in these pairs, both productions have the same duration, 'neither' is the minority response (29%). For SPEED and BOTH pairs, 88% and 90% of responses identify a tempo difference, again split evenly between 'first' and 'second'. As expected, listeners accurately identified NEITHER pairs (96% 'neither').

To examine the responses to PRECISION, SPEED and BOTH pairs further we created a variable to reflect our predictions as to which pair member should be heard as faster: the clear production for PRECISION pairs and the higher-rate production for SPEED and BOTH pairs. We fitted a mixed ordinal regression model (random intercept for participant) [22] and a conditional inference regression tree model [23] to the responses with this variable (*Predicted*) as well as the pair type (*Type*: 'PRECISION', 'SPEED' or 'BOTH') as predictors. Both models revealed the same effects; we visualize them using the tree model in Figure 2.

The tree algorithm iteratively implements binary data splits according to the strongest predictor for the relevant data subset. The first split is at the top of the tree (node 1). In Figure 2, this split is for *Predicted*, such that 'predicted first' maps to a clear majority of 'first' responses and 'predicted second' to a clear majority of 'second' responses across PRECISION, SPEED and BOTH pairs. Within each of the two subsets of responses, *Type* gives rise to a further split (nodes 2, 5) which shows that responses to PRECISION pairs are significantly different from those to SPEED and BOTH pairs. The bar plots suggest that this is partly due to the greater proportion of 'neither' responses. Moreover, while for SPEED and BOTH pairs (nodes 3, 6) nearly all 'first' and 'second' responses are in the predicted direction, for PRECISION pairs (nodes 4, 7) the majority is smaller: about 10% of responses are in the opposite direction. Still, the modelling confirms that responses to PRECISION pairs show a significant listener preference for hearing the compressed clear sentence productions as relatively fast.

Figure 2: Regression tree for the Experiment 2 responses to PRECISION, SPEED and BOTH pairs; predictors explained in the text.



4. DISCUSSION

We have reported on two experiments that address the hypothesis that clear, hyper-articulated speech sounds faster than normal speech when articulation rates are controlled. The Experiment 1 results show that listeners are able to separate the parameters of 'speed' and 'precision' reasonably well when encouraged to do so. However, the fact that responses to PRECISION and BOTH pairs were more variable (and less accurate) than responses to SPEED pairs is consistent with the notion that speaking mode variation triggers tempo variation percepts. The Experiment 2 results show that when listeners assess tempo only, speaking mode clearly influences judgements: as predicted, clear speech sounds faster than normal speech when articulation rates are matched.

Our results show no support for the notion that listeners are biased by their knowledge of typical production patterns towards hearing clear speech as relatively slow [16, 17]. Participants in Experiment 2 showed remarkably robust sensitivity to the duration differences between PRECISION and BOTH pairs: when clear and normal productions differed naturally in duration, clear ones were heard as slower; when the duration difference was absent, clear productions were mostly heard as faster. This is not to dispute the possible perceptual relevance of listeners' knowledge of typical production patterns: listeners are acute at drawing on *any* relevant dimension of variation in making tempo judgements and may weight cue parameters depending on the task at hand [24]. The large difference between Experiments 1 and 2 in the proportions of responses to PRECISION pairs which reflect a perceived tempo difference demonstrates this: participants responded differently to the same stimuli as the details of their tasks were different.

Of course we cannot tell exactly what acoustic parameter(s) participants were weighting highly in Experiment 2. The clear productions were spectrally more complex in the sense of [4], but it remains an open question as to whether this is the best way to characterize the crucial perceptual dimension(s). Note that the increased spectral complexity in clear speech enhances intelligibility [7, 25]. Studies have shown that listeners judge speech as fast in conditions of high cognitive load [26] and low intelligibility [27]. In these contexts, 'complexity' is of a different type, posing a perceptual challenge. While a unified account for all of these effects on perceived tempo is clearly desirable, we cannot propose one at present.

Interestingly, one Experiment 2 participant observed after debriefing that in some of the PRECISION pairs, some specific segment transitions sounded fast. We assume this relates to the low degree of coarticulation in the compressed clear productions; we also cannot rule out some impact of our linear compression method on perceived segment-to-segment timing. Another participant noted that some sentences sounded 'urgent', which made them sound fast. It would appear, then, that both detailed utterance-internal timing patterns and more holistic interpretations of speaker intentions may feed into listeners' tempo judgements. These observations should inform further experimental work.

5. ACKNOWLEDGEMENTS

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

REFERENCES

[1] Kohler, K. J. 1986. Parameters of speech rate perception in German words and sentences:

Duration, f0 movement, and f0 level. *Language and Speech* 29, 115–139.
[2] Rietveld, A. C. M., Gussenhoven, C. 1987. Perceived speech rate and intonation. *Journal of Phonetics* 15, 273–285.
[3] Feldstein, S., Bond, R. N. 1981. Perception of speech rate as a function of vocal intensity and frequency. *Language and Speech* 24, 387–394.
[4] Weirich, M., Simpson, A. P. 2014. Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics* 43, 1–10.
[5] Lindblom, B., 1990, Explaining phonetic variation: A sketch of the H & H theory, in *Speech production and speech modelling*, ed. W. J. Hardcastle and A. Marchal, Amsterdam: Kluwer Academic, 403–439.
[6] Lindblom, B. 1996. Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America* 99, 1683–1692.
[7] Smiljanić, R., Bradlow, A. R. 1999. Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass* 3, 236–264.
[8] Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., Brungart, D. 2018. Clear speech adaptations in spontaneous speech produced by young and older adults. *Journal of the Acoustical Society of America* 144, 1331–1346.
[9] Krause, J. C., Braida, L. D. 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America* 115, 362–378.
[10] Searl, J., Evitts, P. M. 2013. Tongue-palate contact pressure, oral air pressure, and acoustics of clear speech. *Journal of Speech, Language, and Hearing Research* 56, 826–839.
[11] Bradlow, A. R., Torretta, G. M., Pisoni, D. B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20, 255–272.
[12] Smiljanić, R., Bradlow, A. R. 2008. Temporal organization of English clear and conversational speech. *Journal of the Acoustical Society of America* 124, 3171–3182.
[13] Picheny, M. A., Durlach, N. I., Braida, L. D. 1986. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research* 29, 434–446.
[14] Hazan, V., Baker, R. 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America* 130, 2139–2152.
[15] Matthies, M., Perrier, P., Perkell, J. S., Zandipour, M. 2001. Variation in anticipatory coarticulation with changes in clarity and rate. *Journal of Speech, Language, and Hearing Research* 44, 340-353.
[16] Koreman, J. 2006. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America* 119, 582–596.
[17] Reinisch, E. 2016. Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception and Psychophysics* 9, 9–23.
[18] Bradlow, A. R. n.d. SpeechBox. Retrieved from https://speechbox.linguistics.northwestern.edu.
[19] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.
[20] Boersma, P., Weenink, D. 2017. Praat: Doing phonetics by computer. Version 6.0.25. Retrieved from www.praat.org.
[21] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52, 388–407.
[22] Bürkner, P.-C., Vuorre, M. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science* 2, 77–101.
[23] Tagliamonte, S.A., Baayen, R.H. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24**,** 135–178.
[24] Plug, L., Lennon, R., Smith, R. 2022. Measured and perceived speech tempo: Comparing canonical and surface articulation rates. *Journal of Phonetics* 95, 101–193.
[25] Smiljanic, R., Gilbert, R. C. 2017. Intelligibility of noise-adapted and clear speech in child, young adult, and older adult talkers. *Journal of Speech, Language, and Hearing Research* 60, 3069–3080.
[26] Bosker, H. R., Reinisch, E., Sjerps, M. J. 2017. Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language* 94, 166–176.
[27] Bosker, H. R., Reinisch, E. 2017. Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology* 8.