Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 119-122

The influence of articulation rate, and the disfluency of others, on one's own speech

Ian R. Finlayson[1], *Robin J. Lickley*[1], *Martin Corley*[2]

[1]Speech Science Research Centre, Queen Margaret University, UK
[2]Department of Psychology, University of Edinburgh, UK
ifinlayson@qmu.ac.uk

Abstract

Disfluencies are a regular feature of spontaneous speech, and much has been learnt about the effects of various linguistic factors on their production. Speech usually occurs within dialogue, yet little is known about the influence of an interlocutor's speech on a speaker's own fluency. It has been shown that speakers tend to align on various levels, converging, for example, on lexical, and syntactic levels. But we know little about convergence in rate of speech or disfluency. Little is also known about the effects of speech rate on fluency in a speaker's own speech. In this paper, we examine these effects through analysis of speech rate, hesitation and error correction in a corpus of task-oriented dialogues (the HCRC Map Task Corpus). Our findings demonstrate that different types of disfluencies can be influenced in different ways by speech rate. Furthermore, the probability of an interlocutor being disfluent appears to affect the speaker's own likelihood, raising the possibility that interlocutors may "align" on disfluent, as well as fluent, speech.
Index Terms: articulation rate, alignment, accommodation theory, dialogue

1. Introduction

Spontaneous speech is rarely fully fluent. Natural dialogue is peppered with disfluencies such as repairs, repetitions and fillers (e.g. *uh* and *um*). While much has been learned about various linguistic factors which influence the production of disfluencies (e.g. [1, 2]), relatively little attention has been paid to how, on the one hand, rate of speech and the speech and, on the other, the speech of the interlocutor may affect fluency. This study explores these factors and their interaction.

Common in many theories of language production (e.g. [3]) is that the process involves a series of stages. In conceptualisation, a general plan for the utterance is formed; in formulation, the lexical forms are selected and ordered, in articulation, motor commands are sent to the articulatory mechanism. A delay at any of these stages, whether through indecision or error and repair, may result in a need for the speaker to pause, as the rate of motor execution exceeds the rate of planning. Various strategies are possible: The speaker could simply elect to remain silent until the next chunk of speech is ready (though this seems to be the least preferred option); they could maintain some form of vocalization, either by prolonging the previous sound (if the pause is mid-utterance), or by producing a filler (e.g., um/uh); or they may elect to restart the current phrase, thus producing a repetition (cf. [4]).

If restarts are the consequence of the articulator executing one plan before it is able to formulate the next then in the case of fast speech, where we may expect to see a greater incidence of the articulator "out-pacing" the components upon which it draws from, speakers may be more likely to produce repetitions. Repetitions may not, however, be the only strategy available, and we may expect to see other forms of disfluency produced as a consequence of the articulator stalling for time while it awaits new input. A possible alternative would be to produce a filler, a superficially meaningless sound which could occupy the delay.

Where articulation rate has been indirectly manipulated it has been shown that faster speech is more likely to contain repetitions [5], but not likely to contain more fillers. Participants were asked to describe the path taken by a dot around a network which contained images of everyday objects. In half of these trials, the speed of the dot's movement was increased and analysis showed a corresponding significant increase in the articulation rate. This faster condition also saw a higher frequency of errors and error repairs.

It should perhaps come as no surprise that when forced to accelerate our speech, which the dot's faster movements require, participants made more errors. However within this paradigm it is not clear where exactly causality lies. Do participants make more repairs, and repetitions, because they are speaking faster? Or is this increase in errors a consequence of the time pressures caused by having to keep up with the speed of the dot's movements?

In the absence of pressure we may expect there to be no difference in the volume of disfluencies produced by naturally faster and slower speakers. Why, after all, would a person consistently speak at a rate which impaired their ability to maintain fluency? The map task corpus [6] affords us the chance to explore the effects of a naturally occuring range of articulation rates on the production of disfluencies without the confound of having to place pressure upon speakers to produce differences.

The conversational nature of the map task also allows us to explore what effects, if any, an interlocutor may have on both articulation rate and disfluencies. Possible interactions between the speech rates of interlocutors have already received considerable theoretical attention [7]. Accommodation theory suggests that speakers' speech rates may be sensitive to those of their interlocutors: when speaking to an interlocutor who speaks slower, a speaker may tend to reduce their own speech rate in order to converge with their partner.

Recent years have seen a change in our ideas about communication, with the gap between production and comprehension continuing to narrow [8]. While it has been suggested that production and comprehension rely upon one another in order to ensure the alignment of representations between interlocutors in dialogue necessary for successful communication, little consideration has been given to the effects upon production of the disfluent speech that speakers regularly hear from their inter-

locutors.

In the present study we will investigate the influence of speaker's articulation rate upon their likelihood of producing substitutions, repetitions and fillers, before exploring if interlocutors show a tendency to align on articulation rate and their production of disfluencies.

2. Method

2.1. Corpus

Materials came from the HCRC Map Task corpus [6]; transcribed and annotated dialogues recorded between 64 University of Glasgow students (32 male, 32 female) taking part in a cooperative task. Participants took turns to direct their partner along a path which used labelled images of objects as landmarks. Each participant had their own map which their partner could not see. Participants were split up into quads, each consisting of two pairs of friends (although each pair was unfamiliar to the other). In addition, half of all quads performed the task with a screen preventing them from seeing each other at all. All coding has been converted to XML which can be queried using the NITE XML Toolkit [9].

2.2. Unit of analysis

The markup of the corpus is segmented into individual units, known as timed units, which correspond to individual words or silences. Each timed unit has a corresponding unit for its part of speech, one of which being fillers, which include: *eh*, *ehm*, *er*, *erm*, *uh* and *um*.

For disfluencies which did not correspond to individual units we used Lickley's [10] taxonomy of editing disfluencies, as this forms the basis for the disfluency coding which appears in the corpus. Editing disfluencies come in four forms but we will only focus on two of these, with examples of both given in (1): (1a) substitutions, which correspond to what are commonly called repairs; and (1b) repetitions, where one or more words are repeated immediately following their first mention. The structure of each disfluency follows that set out by Levelt [11], with each containing a reparandum, an interruption point and a repair.

(1a) I don't suppose you've got [the balloons] the baboons

(1b) Right [there's a] there's a line about quarter of the way down

All timed units were extracted from the corpus, with the exception of those silences where the participant was listening to their partner, and non-vocal noises, creating a data set of 174,049 units. Each of these units were then coded for whether or not they were disfluent (either a filler, or a reparandum word). Variables were subsequently created for individual types of disfluency (fillers, substitutions, repetitions, insertions and deletions) and each unit was coded appropriately.

2.3. Analysis

As our dependent variables of interest in the present study were binomial (whether or not a unit was disfluent), likelihood was modelled using logit mixed-effects models [12, 13]. Logit mixed-effects models provide an advantage over ordinary logistic regression in that they allow us to include random effects in models, removing the need for separate by-participants and by-items analyses. All analyses were performed in R [14], using the lme4 package [15].

Prior to analysing the effect of articulation rate and interlocutor disfluency on our dependant variables, "control models" were constructed containing participants and the map being described as random effects. Both linguistic and non-linguistic factors were tested for inclusion in these models. In order to control for the finding that disfluencies are associated with the production of longer utterances, a measure of utterance length was required. As it is difficult to get a measure of utterance length in a unstructured dialogue (see [16] for a discussion of these difficulties), the length of each move was used as an approximate analog.

As each individual word of a reparandum was considered, it was possible that speakers who tended to produce longer reparanda may produce spurious results. To control for this, the mean lengths of each speaker's reparanda within each conversation was included. In order to control for the fact that the more a participant said the greater opportunity they had to produce disfluencies, the word count for each conversation was first added. This figure included reparandum words and fillers, in addition to fluent words, however word fragments were excluded. Both word count and mean reparanda length were converted to z-scores.

The following non-linguistic factors were subsequently tested for inclusion in the model: speaker's role (giver, or follower, of instructions), interlocutors' ability to make eye-contact, interlocutors' familiarity, and finally speaker's gender.

These control models were then compared with "full models" containing potential predictors of interest: speaker's articulation rate, partner's articulation rate and the probability of the partner producing each type of disfluency. Articulation rate was measured in syllables per second. The number of syllables in the 100 most used words in the corpus (accounting for 75% of all speech) were counted. Considering only these 100 words, the total number of syllables produced by each speaker in each conversation, was divided by the summed duration of each of these words. Finally, probabilities were obtained by dividing the total number of each disfluency a speaker produced by the total number of words they produced in each conversation.

Once all predictors had been tested, the Wald statistic was used to assess if the estimated slope for each predictor was significantly different from zero. Where the addition of a predictor was found to improve the fit of a model its coefficient will be reported alongside the odds ratio (OR; e^{β}).

3. Results

3.1. Articulation Rate and Disfluency

All control models constructed include utterance length, mean reparandum length and word count where they significantly improve model fit. Where other predictors were included they are reported (see [16] for a full account of the construction of these models).

3.1.1. Substitutions

In order to examine the likelihood of words appearing in the reparanda of substitutions a control model was first constructed, which included the speakers role, giving log-likelihood -8390.8. The addition of the speaker's articulation rate did not improve upon the fit of the model, with a log-likelihood of -8390.8 ($\chi^2(1) = 0.012, p < 1$), suggesting that articulation rate has no effect on speaker's substitutions.

3.1.2. Repetitions

A control model was constructed containing interlocutors' ability to make eye-contact and an interaction between gender and role, with a log-likelihood of -13484. This model was found to be improved by the addition of articulation rate, giving log-likelihood -13481 ($\chi^2(1) = 4.8296, p < .05$), suggesting that faster speakers were more likely to produce repetitions ($\beta = 0.130, p < .05; \mathrm{OR} = 1.138$).

3.1.3. Fillers

In order to model the probability of speakers producing a filler, a control model was constructed which included speaker's role, and provided a log-likelihood of -9578.9. The addition of articulation rate significantly improved the fit of the model, giving a log-likelihood of -9575.0 ($\chi^2(1) = 7.778, p < .01$). This finding suggests that faster speakers are less likely to produce fillers ($\beta = -0.180, p < .01; \mathrm{OR} = 0.835$).

3.2. Alignment of Articulation Rate

To assess the effect of interlocutor's articulation rate we modelled speaker's articulation rate with linear mixed-effects models, with Markov chain Monte Carlo sampling over 10,000 simulations to estimate coefficients. A control model was constructed containing word count, speaker's role, familiarity, speaker and partner's genders; with log-likelihood of -215.83. The addition of the interlocutor's articulation rate was found to improve the fit of the model, with log-likelihood -213.22 ($\chi^2(1) = 5.216, p < .05$), suggesting that talking to faster speakers increases one's own articulation rate. However, we must be cautious in interpreting this finding as role may be confounded with utterance length.

Additionally, our model was further improved by the inclusion of an interaction between speaker's role and the articulation rate of their partner, with log-likelihood -211.02 ($\chi^2(1) = 4.420, p < .05$). This suggests that while all speakers are sensitive to their interlocutor's articulation rate, the effect is increased for instruction followers.

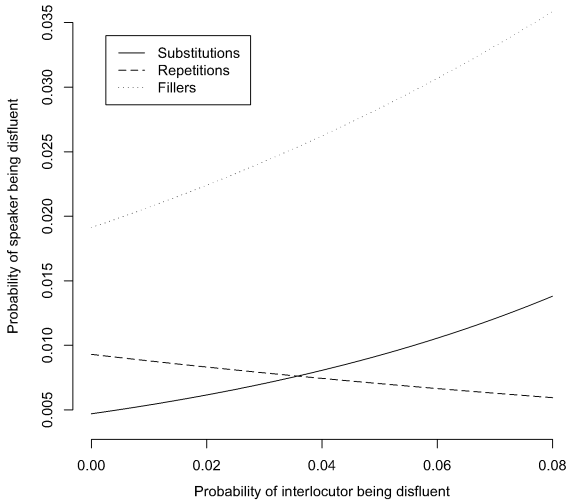
3.3. Alignment of Disfluency

The alignment of articulation rates suggest that aspects of a speaker's production appears sensitive to corresponding aspects of their interlocutors' speech, and we may rightly question if disfluencies behave similarly. However, as articulation rate may influence certain forms of disfluency, we will test the effect of interlocutor's articulation rate on our models before considering their own likelihood of being disfluent. For all types of disfluency, the best fitting models constructed in section 3.1 were used as controls. Estimated probabilities for each type of disfluency are shown in Figure 1.

3.3.1. Substitutions

The addition of interlocutor's articulation rate was not found to improve model fit, giving a log-likelihood of -8390.5 ($\chi^2(1) = 0.579, p < 1$); however including the probability of an interlocutor producing a substitution did significantly improve the model, with log-likelihood -8386.2 ($\chi^2(1) = 9.230, p < .1$). This suggests that speakers are more likely to produce substitutions when speaking to people who are themselves more likely to produce substitutions ($\beta = 13.597, p < .01; \mathrm{OR} = 803437$).

Figure 1: Relationship between speaker's and interlocutor's likelihood of production substitutions, repetitions and fillers. Where models include word count, mean reparandum length or articulation rates, mean values were used.



3.3.2. Repetitions

Similarly to substitutions, interlocutor's articulation rate was not found to improve our model of likelihood of producing repetitions, log-likelihood -13481 ($\chi^2(1) = 0.996, p < 1$). However, speakers were found to be less likely to repeat words when their interlocutor had a greater tendency to repeat ($\beta = -5.631, p < .05; \mathrm{OR} = 0.004$), with the addition of interlocutor probability improving model fit, giving a log-likelihood of -13478 ($\chi^2(1) = 5.809, p < .05$).

3.3.3. Fillers

Including interlocutor's articulation rate was found to improve upon the fit of our control model, giving a log-likelihood of -9571.3 ($\chi^2(1) = 7.280, p < .01$), suggesting that participants were less likely to produce fillers when talking to faster speakers ($\beta = -0.123, p < .01; \mathrm{OR} = 0.884$). Furthermore, the probability of a participant producing a filler increased when their partner was more likely to ($\beta = 8.073, p < .05; \mathrm{OR} = 3207.484$), log-likelihood -9568.7 ($\chi^2(1) = 5.294, p < .05$).

4. Discussion

Our results suggest that different types of disfluencies may be influenced in different ways by articulation rate, and the likelihood of one's interlocutor being disfluent:

- Faster speakers were more likely to produce repetitions; however they were less likely to produce fillers, and no effect was found with substitutions.
- Participants' articulation rate increased with faster speaking partners, particularly when the participant had the role of follower.
- Participants tended to be more likely to produce substitutions and fillers with interlocutors who were themselves

more likely to produce these two types of disfluency, while they produced fewer repetitions when speaking to a partners who were more prone to repeating.

Our finding that faster speakers were more likely to produce repetitions appears consistent with Blackmer & Mitton's [4] idea of an articulator possessing an autonomous restart capability. Increasing the rate of articulation may increase the likelihood of the articulator becoming "out of sync" with the components which precede it, and the prediction that in these situations speakers will tend to repeat the most recent plan appears to be borne out. While speakers may choose to stall through repetition, our finding that faster speakers are, in fact, less likely to produce fillers suggest that they are not used as a similar stalling mechanism.

While it appears that participants' articulation rate is variable, particularly in response to the articulation rate of their partner, it does not seem that they allow themselves to speak at such an accelerated rate that they are more prone to make the sorts of errors which require substitutions to resolve. This suggests that Oomen & Postma's [5] finding that errors occurred more frequently when a speaker was under time pressure may be due solely to the greater demand of having to keep up with a faster moving dot in order to describe its path, rather than being a consequence of speaking too quickly.

The alignment of articulation rates between partners provides evidence for the convergence of linguistic factors suggested by accommodation theory. However deeper understanding of this phenomena requires closer examination of variations in articulation rate across not only conversations, but also across conversational turns, which is beyond the scope of the present study.

It is not clear how the disfluencies of one's interlocutor may influence our own likelihood of being disfluent, with converging upon disfluencies seeming unlikely to bring a benefit. One possibility may be that hearing a conversational partner produce a greater number of errors makes one more comfortable in being error prone, reducing the care one takes in being fluent. This account would hold despite the same pattern not emerging with repetitions, as repetitions may reflect a different type of problem to those solved with substitutions. However it would offer no explanation for why the opposite pattern was found when we observed the effects of repetitions.

Similarly, an explanation is lacking for why one is less likely to produce fillers when speaking to a faster partner, even after one's own articulation rate has been taken into account. If fillers are used as signals for turn-taking, as has been suggested (e.g. [17, 18]), then we may speculate that engaging in a dialogue with someone who often uses fillers may lead one to employ them more often themselves in order to manage their side of the conversation, which may explain the link observed between our participant's likelihood of using fillers and their interlocutor's.

Future research may explore further how both linguistic and non-linguistic features of dialogues may influence speaker's likelihood to be disfluent, and how these factors may interact.

5. Acknowledgements

We thank Jean Carletta and Jonathan Kilgour for technical help.

6. References

[1] G. W. Beattie and B. Butterworth, "Contextual probability and word frequency as determinants of pauses in spontaneous speech," *Language and Speech*, vol. 22, pp. 201–211, 1979.

[2] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and Speech*, vol. 44, no. 2, p. 123, 2001.

[3] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and brain sciences*, vol. 22, no. 01, pp. 1–38, 1999.

[4] E. R. Blackmer and J. L. Mitton, "Theories of monitoring and the timing of repairs in spontaneous speech," *Cognition*, vol. 39, no. 3, pp. 173–194, 1991.

[5] C. C. E. Oomen and A. Postma, "Effects of time pressure on mechanisms of speech production and self-monitoring," *Journal of Psycholinguistic Research*, vol. 30, pp. 163–184, 2001.

[6] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991. [Online]. Available: http://www.hcrc.ed.ac.uk/maptask/

[7] H. Giles and P. Smith, *Language and Social Psychology*. Oxford: Basil Blackwell, 1979, ch. Accommodation Theory: Optimal Levels of Convergence.

[8] M. J. Pickering and S. Garrod, "Do people use language production to make predictions during comprehension?" *Trends in Cognitive Sciences*, vol. 11, no. 3, pp. 105–110, 2007.

[9] J. Carletta, S. Evert, U. Heid, and J. Kilgour, "The NITE XML Toolkit: Data model and query language," *Language Resources and Evaluation*, vol. 39, pp. 313–334, 2006. [Online]. Available: http://groups.inf.ed.ac.uk/nxt/index.shtml

[10] R. J. Lickley, "HCRC Disfluency Coding Manual," HCRC, Tech. Rep. 100, 1998. [Online]. Available: http://www.ling.ed.ac.uk/ robin/maptask/disfluency-coding.html

[11] W. J. M. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[12] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Society*, vol. 88, pp. 9–25, 1993.

[13] S. DebRoy and D. M. Bates, "Linear mixed models and penalized least squares," *Journal of Multivariate Analysis*, vol. 91, pp. 1–17, 2004.

[14] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. [Online]. Available: http://www.R-project.org

[15] D. Bates, M. Maechler, and B. Dai, *lme4: Linear mixed-effects models using S4 classes*, 2009, r package version 0.999375-32. [Online]. Available: http://CRAN.R-project.org/package=lme4

[16] I. R. Finlayson, M. Corley, and R. J. Lickley, "Why use disfluency? The effect of situational factors on the production of disfluent speech in dialogue," Submitted.

[17] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[18] H. Maclay and C. E. Osgood, "Hesitation phenomena in spontaneous speech," *Word*, vol. 15, pp. 19–44, 1959.