2017. *Proc Ling Soc Amer* 2, 35:1-10. https://doi.org/10.3765/plsa.v2i0.4088.

How automatic is phonetic convergence? Evidence from working memory

Jevon Heath[*]

Abstract. Phonetic convergence has been argued to be automatic as well as socially motivated. Previous studies have shown effects of socially-driven factors on working memory (Beilock et al. 2007), suggesting that working memory might mediate the socially-driven and automatic components of convergence. We used Amazon's Mechanical Turk to test convergence toward a voice with lengthened VOT, with short-term and working memory (modified digit span) conditions as well as a control condition. Overall, participants converged in VOT while shadowing, but converged less when working memory was occupied. These results suggest that social factors affect phonetic convergence indirectly, through their effect on working memory load.

Keywords. Sociophonetics, accommodation, imitation, Amazon Mechanical Turk, working memory.

1. Introduction. Phonetic convergence, also called phonetic imitation, is the phenomenon where an individual's speech becomes more similar to the speech that they hear over the course of an interaction. Studies have demonstrated an automatic component to imitation in speech (e.g., Delvaux & Soquet 2007), indicating that imitation is default behavior for humans. However, studies have also repeatedly shown effects of social factors on the degree of imitation that speakers evince in various situations (Bourhis & Giles 1977; Abrego-Collier et al. 2011; Babel 2010, 2012; *inter alia*) — and some studies seem to indicate that social factors affect whether speakers evince phonetic imitation at all (e.g., Pardo et al. 2010). If imitation is indeed automatic, then social factors must affect imitation indirectly, rather than being the direct cause of imitation behavior.

One way in which social factors may indirectly affect imitation is by affecting the processing of input before it can be imitated. This is possible either by affecting the perception of the incoming phonetic signal, or by affecting the cognitive processes that recognize the received phonetic signal as input for convergence. Social factors have been shown to affect speech perception (Peterson & Barney 1952; Strand & Johnson 1996). One candidate cognitive process for affecting the processing of phonetic input is working memory, which has been shown to have an effect on socially driven factors in other contexts, such as the activation of stereotypes (Schmader & Johns 2003; Beilock et al. 2007). The current study examines the effects of working memory load on phonetic convergence, measuring VOT in English voiceless stops in a modified digit span shadowing task.

2. Background. Delvaux and Soquet (2007) describe phonetic imitation as "automatic": "Our goal is to provide evidence for a general tendency for spontaneous and automatic

[*] This research has benefited greatly from input and support from Melinda Fricke, Susanne Gahl, Keith Johnson, Susan Lin, and Ronald Sprouse, among others. Author: Jevon Heath, University of Pittsburgh (jsh82@pitt.edu).

imitation of the way of speaking of the ambient language, regardless of the complex history and nature of the social relationships that may exist between 2 interacting speakers" (2007:148). However, *automatic* is ambiguous in this context, such that it will be useful to distinguish two potentially pertinent senses of the word before proceeding with a discussion of automaticity in convergence. A process may be called automatic when it always takes place without conditions; such a process is *systematic*. Alternatively, a process may be called automatic when it always takes place given appropriate trigger conditions; such a process is *reflexive*. Delvaux and Soquet later make clear that they use *automatic* to mean *reflexive* rather than *systematic*: "[U]nless hindered by higher-order sociopsychological factors (e.g. deliberate will to dissociate from a particular social group, or to distance oneself from a specific individual), speakers automatically tend to adjust their phonetic realisations to ambient speech" (2007:146).

Along this line, studies of phonetic imitation have found certain circumstances in which no measurable convergence takes place. Pardo et al. (2010) used a dyadic map task to look at the effects of conversational role on imitation. They found increased convergence in dyads in which the information giver was explicitly instructed to imitate their interlocutor, but divergence when the information receiver was instructed to imitate.

If phonetic convergence is reflexive – always taking place given the appropriate trigger conditions – discrepancies in the occurrence of convergence across individuals and contexts must relate either to the presence of appropriate trigger conditions, or to their recognition as such. The executive functions of cognition that are potentially germane to the presence and/or recognition of trigger conditions for phonetic convergence are inhibitory control and working memory.

Inhibitory control governs the suppression of information (Radvansky et al. 1996). Working memory involves actively using information that is held in memory, i.e. that is not currently perceptible (Baddeley & Hitch 1994). Working memory and inhibitory control generally work in concert: working memory is used to keep track of what to inhibit, and inhibitory control is used to focus on the correct manipulations to perform on the information being held in working memory (Diamond 2013). Because of this, inhibitory control is sometimes grouped in with working memory (cf. Hasher & Zacks 1988), or considered a component or consequent function of working memory (cf. Baddeley & Hitch 1994; Munakata et al. 2011). Additionally, working memory has been shown to constrain linguistic processing in other ways, including resolution of syntactic ambiguity (MacDonald et al. 1992).

Convergence in speech can be conceptualized in terms of the incorporation of unnecessary or unspecified detail into one's speech patterns. If such detail is actively attended to, its incorporation will take up working memory capacity. If such detail is instead actively filtered out, its nonincorporation will take up working memory capacity. Either way, we have reason to expect working memory to have an effect on convergence. However, depending on whether phonetic detail is attended to or blocked by working memory, we are led to two hypotheses regarding its influence on convergence in speech. According to the first hypothesis (H1), interlocutors with an increased working memory load will exhibit *less* convergence, as they will not be able to retain the perception of fine phonetic detail in order to converge toward it. According to the second hypothesis (H2), interlocutors with an increased working memory load will exhibit *greater* phonetic convergence, as they will not be able to inhibit the attendance to fine phonetic detail that triggers convergence. The null hypothesis (H0) is that working memory load will not affect the rate of phonetic convergence.

In order to determine which of these hypotheses has more empirical support, I designed a shadowing task intended to manipulate working memory load. In three separate conditions, participants repeated words after a model talker. In the control condition, there was no other component to this task. In the working memory condition, participants conducted a simple mathematical operation on a string of digits in their head while shadowing the model talker. In the third condition, participants memorized the string of digits but did not do anything else with them; this third condition is intended to disambiguate working memory load from short-term memory (Engle et al. 1999).

I measured convergence to VOT of voiceless stops in English, which has been repeatedly demonstrated to show imitation effects (Shockley et al. 2004; Abrego-Collier et al. 2011). Because the data in question were durational in nature, a standardized recording fidelity was not crucial. As such, I was able to conduct this experiment using Amazon Mechanical Turk (Buhrmester et al. 2011) in order to recruit a sufficient number of participants quickly for each of the three conditions.

3. Experiment. Participants completed a Human Intelligence Task (HIT) on Amazon Mechanical Turk in which they requested qualification for the working memory task. In order to receive the necessary qualifications, participants had to report their age as between 18-35; that they had an external microphone and headphones; that they had no hearing loss; that they were native speakers of English; and that they consented to have their voices recorded. Participants who self-reported all of these qualifications (*n* = 76) were enabled to complete a second HIT which was the experiment.

In the second HIT, participants navigated in their web browser to a webpage with an interface allowing them to start and stop recordings of their voice. The webpage began with a page of text instructions for using the interface. When participants clicked the "start recording" button, a word appeared on the screen for them to say; when they subsequently pressed "stop recording", the word disappeared. Participants proceeded at their own pace through 120 English words in this manner. The 120 words included 100 words whose stressed syllable had a voiceless stop (one of /p t k/) as the onset, and 20 filler words. Upon finishing this block, participants saw another page of text instructions for the next section. For the second section, instead of words appearing visually when "start recording" was pressed, recordings of the words were played. The recordings in this condition were modified speech from a female undergraduate college student with a California Bay Area accent; the VOT of her voiceless stops was doubled via a Praat script (modified from Pacilly 2008). This section included 80 words, after which another page of text instructions appeared for the next section. The third section was the same as the first except that the words presented were in a different order.

There were three conditions for this task. The control condition (*n* = 20 after exclusions) was as described in the preceding paragraph. In the short-term memory condition (*n* = 19 after exclusions), participants were given a series of six digits every eight trials during the shadowing task. They were to hold these digits in mind and type them into a text field eight trials later. (Digit series were originally planned to be eight digits in length, but pilot subjects indicated that the task was too difficult.) The working memory condition (*n* = 25) was the same as the short-term memory condition except that participants were instructed to add 1 to each digit before recalling them.

Mechanical Turk workers are paid by the task, and so they have incentive to finish as quickly as possible. This was reflected in the recordings of two participants, who clicked "stop recording" before they finished speaking for a large percentage of trials. While this truncation often occurred after the focal stressed syllable, it more often occurred during the stressed syllable. The two participants whose recordings were truncated in this manner were excluded from analysis. Similarly, some participants in the short-term memory condition copied and pasted the strings of digits they were given instead of memorizing them (as evinced by 100% accuracy in recall and the presence of spaces between the digits, which they were instructed not to include). Participants who had no errors in their digit recall in this condition were likewise excluded from analysis. Participants in the working memory condition were not excluded for having perfect responses, as it was assumed that individuals who intended to edit the string of digits were occupying their working memory with this intent. (The inclusion of these participants' data did not affect the models, lending credence to this assumption.)

Participants' recordings were aligned using the Penn Forced Aligner (Yuan & Liberman 2008), and VOTs were measured with a Python script (Johnson 2015). To quantify convergence in VOT, I used the difference-in-distance measure (*following* Babel 2009). Difference-in-distance is calculated by subtracting the difference between the speaker and the model's values before the exposure phase from the difference between their values after exposure. As such, a negative difference-in-distance is analyzed as convergence – the difference between speaker and model shrank after exposure.

3.1 ANALYSIS. Participants showed global convergence across all task conditions, as shown in Figure 1. Overall, participants' voiceless stops had the longest VOT during the shadowing block, and the shortest VOT during the pre-exposure block ($p < 0.01$ for all three block pairs).

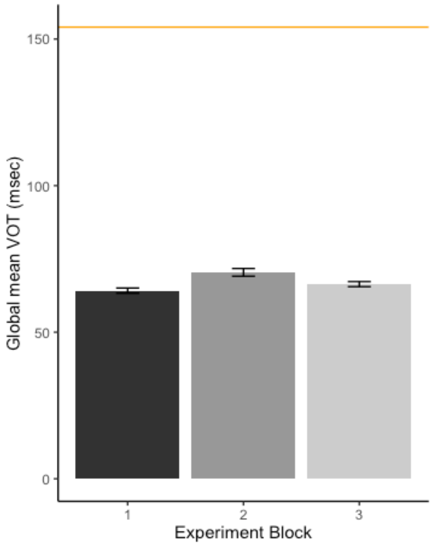

Figure 1: Global VOT convergence between conditions (model mean VOT in gold)

As shown in Figure 2, there was no effect of experimental condition on difference-in-distance between VOT in blocks 1 and 3. However, a weak interaction effect was observed between task condition and consonant, such that participants in the working memory condition diverged from the model talker in their VOT for /p/; this interaction is shown in Figure 3.

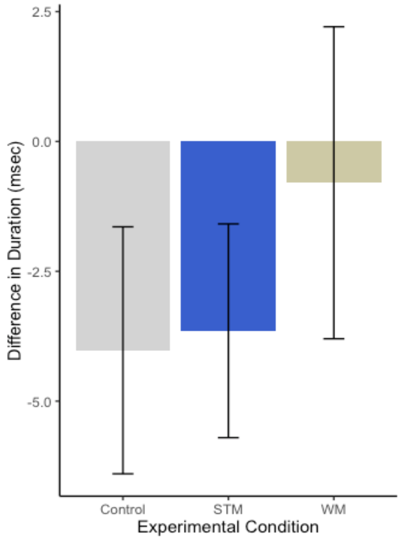

Figure 2: VOT convergence by condition (*n.s.* for all pairs)

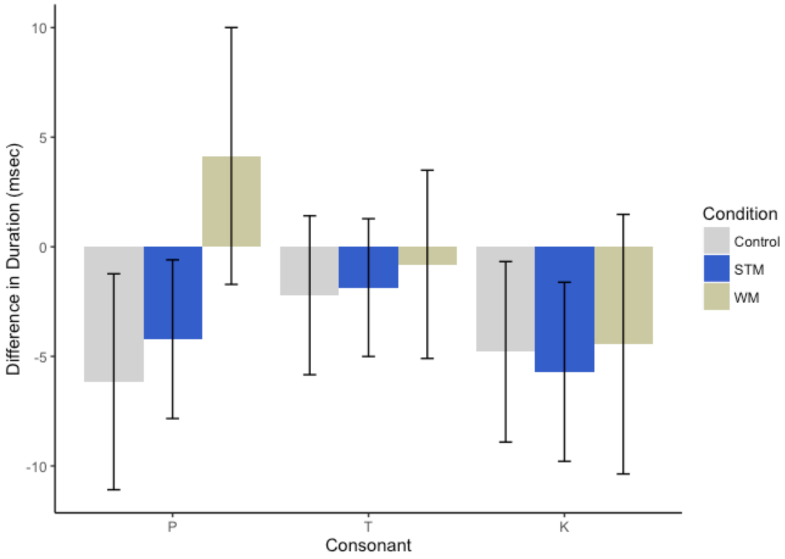

Figure 3: VOT convergence by condition and consonant

Figure 4 shows histograms of individuals' mean difference-in-distance on a by-word basis, broken down by condition. Compared to the control condition, participants in the two task conditions have mean difference-in-distance values that cluster around 0, indicating no change between pre-exposure and post-exposure recordings, although this pattern did not prove statistically significant.

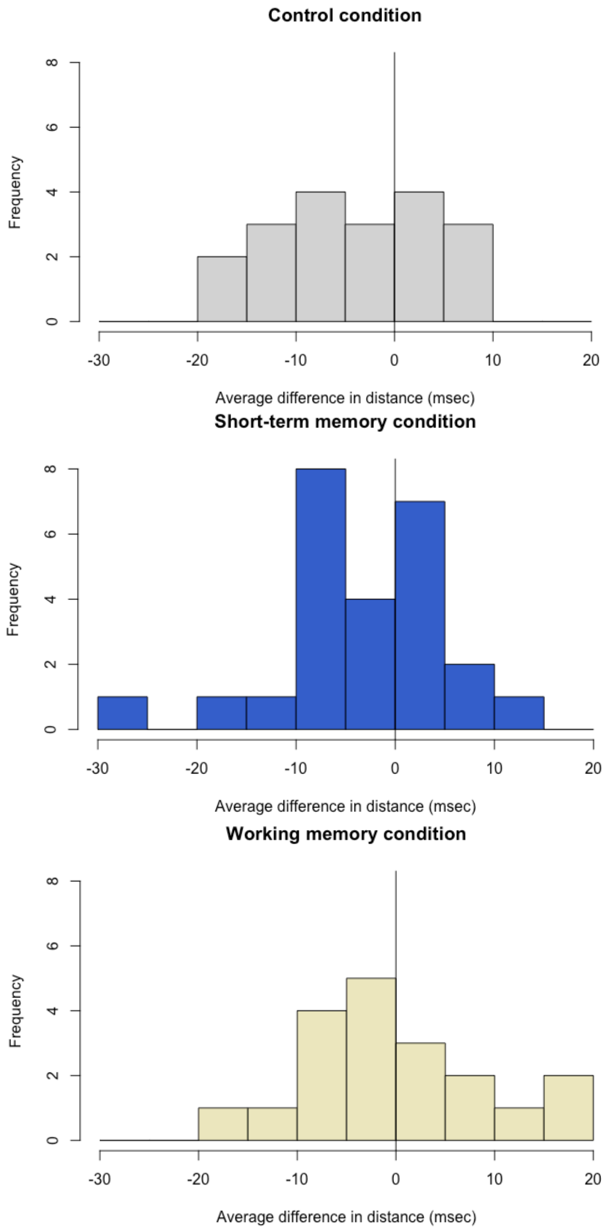

Figure 4: Histograms of difference-in-distance by participant for each condition

3.2 MODELING. I ran a linear mixed-effects regression model using the lme4 and lmertest packages in R (Bates et al. 2016; Kuznetsova et al. 2016; R Core Team 2016) with VOT as the outcome variable; main effects of consonant, experimental condition, test block, and an interaction effect between condition and block; and random intercepts by subject and word. The output of the resulting model is shown in Table 1. (/k/ is the reference level for the consonant effect; the control condition is the reference level for the condition effect.) There was no main effect of condition; however, there was a significant interaction effect of condition with the shadowing task block (*M* = 4.2 msec, *SE* = 1.8 msec, *t* = –2.36), indicating that during the shadowing block, the participants with a working memory load-occupying distractor task had VOTs significantly shorter than participants with no such distractor task.

```
Random effects:
 Groups   Name        Variance  Std.Dev.
 Word     (Intercept) 0.0001007 0.01003
 Subject  (Intercept) 0.0001216 0.01103
 Residual             0.0007249 0.02692
Number of obs: 11305, groups:  Word, 104; Subject, 63

Fixed effects:
                      Estimate Std. Error         df t value Pr(>|t|)
(Intercept)          7.619e-02  3.160e-03  1.260e+02  24.108  < 2e-16 ***
Block2               5.870e-03  1.229e-03  1.123e+04   4.777 1.80e-06 ***
Block3               2.120e-03  1.010e-03  1.114e+04   2.098   0.0359 *
ConditionSTM        -4.424e-03  3.492e-03  6.900e+01  -1.267   0.2095
ConditionWM         -5.915e-03  3.731e-03  7.000e+01  -1.585   0.1174
ConsonantP          -1.249e-02  2.552e-03  1.050e+02  -4.896 3.56e-06 ***
ConsonantT          -1.973e-03  2.450e-03  1.030e+02  -0.805   0.4225
Block2:ConditionSTM  2.305e-03  1.628e-03  1.115e+04   1.416   0.1568
Block3:ConditionSTM -5.188e-05  1.349e-03  1.114e+04  -0.038   0.9693
Block2:ConditionWM  -4.248e-03  1.797e-03  1.115e+04  -2.364   0.0181 *
Block3:ConditionWM  -8.880e-04  1.483e-03  1.115e+04  -0.599   0.5493
```

Table 1: Random and fixed effects for model predicting VOT by task condition

```
Correlation of Fixed Effects:
            (Intr) Block2 Block3    STM     WM     P      T   B2:STM  B3:STM B2:WM
Block2      -0.143
Block3      -0.162  0.412
STM         -0.627  0.119  0.145
WM          -0.587  0.111  0.136  0.531
ConsonantP  -0.388  0.013  0.002  0.000  0.001
ConsonantT  -0.404  0.017  0.002  0.000  0.001  0.499
Block2:STM   0.102 -0.720 -0.311 -0.161 -0.084 -0.004 -0.004
Block3:STM   0.121 -0.309 -0.749 -0.196 -0.102  0.000  0.000  0.419
Block2:WM    0.093 -0.652 -0.282 -0.082 -0.165 -0.004 -0.005  0.495  0.211
Block3:WM    0.110 -0.281 -0.682 -0.099 -0.200  0.000 -0.002  0.212  0.511  0.415
```

Table 2: Correlation of fixed effects for model predicting VOT by task condition

4. Discussion. *People converge less when they are distracted.* In a digit span task designed to tax working memory, participants converged less to a model talker while handling an active working memory load. This finding supports the hypothesis (H1) in which phonetic convergence relies on working memory resources. As such, phonetic convergence is not systematic; it only occurs when resources are allocated to it. Once "turned on" by the allocation of sufficient working memory, convergence is a reflexive response to the processing of speech.

Accommodation may be moderated by attention, such that people converge more toward speech that they are attending: If a listener is concentrating on something other than the received speech signal, their working memory is occupied by that other thing. Working memory needs to be occupied by the received speech in order for details of that speech to be processed. This finding may also point to an explanation as to why particular social factors such as attractiveness (Babel 2012) affect phonetic convergence: people pay more attention to people they are attracted to.

Only accommodation to VOT was measured for this experiment, so it might be argued that some phonetic features are moderated by working memory whereas others bypass working memory. If this were the case, would we expect more or less salient features to be the ones to bypass working memory? One might predict that salient features do not need to be overtly noticed; perception systems will pick up anything salient without direct attention being paid. However, given the phonological importance of VOT in English, this account would seem to predict that English VOT would bypass working memory, which is not the pattern observed here. This suggests that phonetic features are moderated by working memory regardless of their salience.

While the feature measured in this experiment is phonetic, it is reasonable to suppose that similar results would hold for convergence toward linguistic features at other levels of structure. This finding is wholly in line with Interactive Alignment Theory (Pickering & Garrod 2004), in which convergence is a mostly unconscious, resource-free process that cannot be switched off: "The activation of a representation in one interlocutor leads to the activation of the matching representation in the other interlocutor directly" (2004:9). In Pickering and Garrod's model, convergence results from "implicit common ground", which is built up between interlocutors over the course of a dialogue.

The findings here also suggest that we may expect a difference between convergence behavior in laboratory speech and convergence behavior in the "real world". Laboratory speech tends not to involve distractions that are not specifically included in the design of an experiment, which presumably means that working memory limits are not taxed. In contrast, natural speech often involves all sorts of distractions, ranging from the environmental to the metadiscursive. If working memory affects the magnitude of accommodation, we would expect a difference in accommodative behavior inside the lab and outside of it.

According to Kunda and Spencer (2003), one of the main purposes of stereotypes is to facilitate comprehension of a situation by simplifying it. Gilbert and Hixon (1991) found that subjects aided by an Asian lab assistant showed evidence of stereotype activation, but subjects who were rehearsing an 8-digit number during their exposure to the same assistant showed no evidence of such activation. Under an episodic model of convergence (Goldinger 1998, Pierrehumbert 2001), it may be useful to think of phonetic features in terms of stereotypes. In this approach, people form expectations of what they will hear, and if they hear speech that meets those expectations, their stereotypes are reinforced, and they rely more heavily on those stereotyped features, resulting in convergence.

References

Abrego-Collier, Carissa, Julian Grove, Morgan Sonderegger, & Alan C. L. Yu. 2011. Effects of speaker evaluation on phonetic convergence. In *Proceedings of the 17th International Congress of the Phonetic Sciences*. 192–195.

Babel, Molly E. 2009. *Phonetic and social selectivity in speech accommodation.* Berkeley, CA: University of California dissertation.

Babel, Molly. 2010. Dialect divergence and convergence in New Zealand English. *Language in Society* 39(4). 437–456. https://doi.org/10.1017/S0047404510000400.

Babel, Molly. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40(1). 177–189.

https://doi.org/10.1016/j.wocn.2011.09.001.

Baddeley, Alan D., & Graham J. Hitch. 1994. Developments in the concept of working memory. *Neuropsychology* 8(4). 485–493. https://doi.org/10.1037/0894-4105.8.4.485.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2016. "lme4: Linear mixed-effects models using Eigen and S4." R package.

Beilock, Sian L., Robert J. Rydell, & Allen R. McConnell. 2007. Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General* 136(2). 256–276. https://doi.org/10.1037/0096-3445.136.2.256.

Bourhis, Richard Y., & Howard Giles. 1977. The language of intergroup distinctiveness. In Howard Giles (ed.), *Language, Ethnicity, and Intergroup Relations*. 119–135. London: Academic Press.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1). 3–5. https://doi.org/10.1177/1745691610393980.

Delvaux, Véronique, & Alain Soquet. 2007. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica* 64(2-3). 145–173. https://doi.org/10.1159/000107914.

Diamond, Adele. 2013. Executive functions. *Annual Review of Psychology* 64. 135–168. https://doi.org/10.1146/annurev-psych-113011-143750.

Engle, Randall W., Stephen W. Tuholski, James E. Laughlin, & Andrew R. A. Conway. 1999. Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General* 128(3). 309–331. https://doi.org/10.1037/0096-3445.128.3.309.

Gilbert, Daniel T. & J. Gregory Hixon. 1991. The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Psychology and Social Psychology* 60. 509–517. https://doi.org/10.1037/0022-3514.60.4.509.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105(2). 251–279. https://doi.org/10.1037/0033-295X.105.2.251.

Hasher, Lynn, & Rose T. Zacks. 1988. Working memory, comprehension, and aging: A review and a new view. *Psychology of learning and motivation* 22. 193–225. https://doi.org/10.1016/S0079-7421(08)60041-9.

Johnson, Keith. 2015. VOT.py – measure VOT in all of the stops. Python script.

Kunda, Ziva, & Steven J. Spencer. 2003. When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin* 129(4). 522–544. https://doi.org/10.1037/0033-2909.129.4.522.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2016. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0–32. https://CRAN.R-project.org/package=lmerTest.

MacDonald, Maryellen C., Marcel Adam Just, and Patricia A. Carpenter. 1992. Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology* 24. 56–98. https://doi.org/10.1016/0010-0285(92)90003-K.

Munakata, Yuko, Seth A. Herd, Christopher H. Chatham, Brendan E. Depue, Marie T. Banich, & Randall C. O'Reilly. 2011. A unified framework for inhibitory control. *Trends in cognitive sciences* 15(10). 453– 459. https://doi.org/10.1016/j.tics.2011.07.011.

Pacilly, Jos J. A. 2008. "Modify duration of 2nd interval in 16 steps from 5..80 ms." Praat script.

Pardo, Jennifer S., Isabel Cajori Jay, & Robert M. Krauss. 2010. Conversational role influences speech imitation. *Attention, Perception, & Psychophysics* 72(8). 2254–2264. https://doi.org/10.3758/BF03196699.

Peterson, Gordon E., & Harold L. Barney. 1952. Control methods used in the study of vowels. *Journal of the Acoustical Society of America* 24. 175–184. https://doi.org/10.1121/1.1906875.

Pickering, Martin J., & Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2). 169–189. https://doi.org/10.1017/S0140525X04000056.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*. 137–157. John Benjamins Publishing.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Radvansky, Gabriel A., Rose T. Zacks, & Lynn Hasher. 1996. Fact retrieval in younger and older adults: The role of mental models. *Psychology and Aging* 11(2). 258–271. https://doi.org/10.1037/0882-7974.11.2.258.

Schmader, Toni, & Michael Johns. 2003. Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology* 85(3). 440–452. https://doi.org/10.1037/0022-3514.85.3.440.

Strand, Elizabeth A. & Johnson, Keith. 1996. Gradient and visual speaker normalization in the perception of fricatives. In Dafydd Gibbon (ed.), *Natural language processing and speech technology: results of the 3rd KONVENS conference, Bielefeld*. 14–26. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110821895-003.

Yuan, Jiahong, and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123(5). 3878. https://doi.org/10.1121/1.2935783.