Journal of Memory and Language 89 (2016) 68–86

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

# Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations

Esteban Buz [a,*], Michael K. Tanenhaus [a,b], T. Florian Jaeger [a,b,c]

[a] Department of Brain and Cognitive Sciences, University of Rochester, United States
[b] Department of Linguistics, University of Rochester, United States
[c] Department of Computer Science, University of Rochester, United States

ARTICLE INFO

*Article history:*
Received 4 November 2014
revision received 17 December 2015
Available online 2 February 2016

*Keywords:*
Language production
Hyper-articulation
Communication
Interlocutor feedback
Perceptual confusability

ABSTRACT

We ask whether speakers can adapt their productions when feedback from their interlocutors suggests that previous productions were perceptually confusable. To address this question, we use a novel web-based task-oriented paradigm for speech recording, in which participants produce instructions towards a (simulated) partner with naturalistic response times. We manipulate (1) whether a target word with a voiceless plosive (e.g., *pill*) occurs in the presence of a voiced competitor (*bill*) or an unrelated word (*food*) and (2) whether or not the simulated partner occasionally misunderstands the target word. Speakers hyper-articulated the target word when a voiced competitor was present. Moreover, the size of the hyper-articulation effect was nearly doubled when partners occasionally misunderstood the instruction. A novel type of distributional analysis further suggests that hyper-articulation did not change the *target* of production, but rather reduced the probability of perceptually ambiguous or confusable productions. These results were obtained in the absence of explicit clarification requests, and persisted across words and over trials. Our findings suggest that speakers adapt their pronunciations based on the perceived communicative success of their previous productions in the current environment. We discuss why speakers make adaptive changes to their speech and what mechanisms might underlie speakers' ability to do so.



## Introduction

Speech production is context sensitive. This is most obvious and best understood with regard to linguistic context. For example, how a sound is articulated and pronounced depends on the surrounding sounds (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) and its position in the larger linguistic structure (e.g., due to stress assignment and other prosodic factors, Klatt, 1976). Speech production is also sensitive to the broader non-linguistic context. This includes, for example, adjustments in how we talk due to the levels of acoustic noise in the local environment—speakers tend to talk louder when in a noisy environment (known as the Lombard effect, Lombard, 1911; Van Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988). It also includes adjustments based on whom we are talking to. For example, we sometimes revert to our home dialect when talking to friends or family from that

* Corresponding author.
*E-mail addresses:* ebuz@bcs.rochester.edu (E. Buz), mtan@bcs.rochester.edu (M.K. Tanenhaus), fjaeger@bcs.rochester.edu (T.F. Jaeger).

http://dx.doi.org/10.1016/j.jml.2015.12.009
0749-596X/

region, or switch to less formal registers when talking to people we know, resulting in changes to speech rate and clarity of articulation, among other things (e.g., Bell, 1984; Finegan & Biber, 2001). Sensitivity to the socio-indexical context in which speech takes place goes beyond adjustments to interlocutors we know. Speakers also can adjust their pronunciations based on *types* of interlocutors. For example, speech directed at adults differs systematically from speech directed at infants (e.g., Kuhl et al., 1997; Pate & Goldwater, 2015) or pets (e.g., Burnham, Kitamura, & Vollmer-Conna, 2002). Similarly, speech directed at typical adult native interlocutors differs from speech directed at non-native interlocutors (e.g., Uther, Knoll, & Burnham, 2007) or audiences with impaired comprehension (e.g., "clear speech", speech directed at the hard of hearing, Picheny, Durlach, & Braida, 1986).

Examples like these illustrate that non-linguistic context can affect pronunciation. They also suggest that speech production is to some extent adaptive, allowing speakers to adjust their productions depending on their audience. These examples leave open, however, how *dynamic* such adjustments are. The present study begins to address this question. This question is both under-explored and of central importance to our understanding of the architecture underlying language production. In the longer-term, understanding adaptive processes holds the potential to shed light on the origin of socio-indexically conditioned registers, such as infant- and foreigner-directed speech. Adaptive processes may also be key to reconciling seemingly conflicting results in research on audience design (as proposed in Jaeger & Ferreira, 2013; for discussion see Jaeger & Buz, in press). Beyond contributing to these longer-term goals, our more immediate goal is to gain a better understanding of the nature of adaptation in speech. Specifically, we investigate whether and if so how speakers adapt their pronunciations—by *hyper-articulating* certain sounds—based on feedback from interlocutors about the communicative success of the speaker's previous productions. This then leads us to investigate hyper-articulation in such situations more closely. Precisely how do speakers adapt their articulations in response to feedback that suggests that their previous utterance was perceptually confusing? And what is the likely function or goal of this adaptive behavior?

We approach these questions in a novel web-based task-oriented *simulated partner* paradigm for speech recording. Participants provide instructions to a partner, who—unbeknownst to the participant—is simulated by a computer program. This allows us to control the timing and type of feedback that speakers received from their interlocutors, while maintaining ecological validity (as indexed by ratings reported below). Specifically, we manipulate what feedback participants receive on individual trials about whether their partner understood them. We then assess the degree to which participants hyper-articulate as a function of the perceived communicative success of their previous productions.

Before we describe our study in more detail, we briefly summarize previous work on the effect of interlocutor feedback on speakers' articulations and highlight how the present experiment contributes to this literature.

*Previous work and how the present study contributes to it*

Only a few studies have directly investigated the role of interlocutor feedback on subsequent productions. One line of research that is particularly relevant to the current goals has investigated the articulations of *corrections* following explicit clarification requests (Maniwa, Jongman, & Wade, 2009; Ohala, 1994; Oviatt, Levow, Moreton, & MacEachern, 1998; Oviatt, MacEachern, & Levow, 1998; Schertz, 2013; Stent, Huffman, & Brennan, 2008). For example, Schertz (2013, Study 1) recorded participants as they produced speech directed at what they believed to be an automatic speech recognition system. Target words either had voiced or voiceless plosive onsets (e.g., *pit*). On critical trials, the (simulated) automatic speech recognition system displayed a recognition error and requested clarification. Participants then had to repeat the same word. Schertz found that corrections were hyper-articulated (see also Maniwa et al., 2009; Oviatt, Levow, et al., 1998; Oviatt, MacEachern, et al., 1998; Schertz, 2013; Stent et al., 2008; for similar findings in response to a simulated human partner see Ohala, 1994).

Interestingly, the hyper-articulation observed in these studies was often targeted to the specific part of the production that seemed to have caused the misrecognition. For example, the (simulated) automatic speech recognition system in Schertz (2013) used more or less specific clarification prompts to indicate which part of participants' productions had likely caused the misrecognition. Sometimes clarification prompts were general ("???"). Other times, prompts contained specific guesses that deviated from the target (e.g., *pit*) in either voicing ("bit?"), place ("kit?"), or manner (e.g., "sit?"). Schertz found that voice onset times (VOT)—the primary cue to the English voicing distinction (e.g., *pit* vs. *bit*)—were hyper-articulated only following voicing-contrastive word prompts (e.g. "bit?") but not when participants saw general prompts ("???") or manner/place-contrastive prompts (e.g. "kit?" or "sit?"). Additionally, hyper-articulation after voicing-contrastive prompts was limited to VOTs: neither the overall amplitude nor overall word duration was hyper-articulated (see also de Jong, 2004; Maniwa et al., 2009; though see Ohala, 1994).

These results suggest that speakers can adapt productions of the same word immediately following an explicit request for clarification, and that they can do so in a targeted manner. Here we seek to contribute to this literature and to extend it in several ways. First, the majority of previous studies had participants produce words towards (simulated) automatic speech recognition systems (but see Ohala, 1994). There is evidence that speech directed at automatic speech recognition systems differs qualitatively from speech directed at human interlocutors (Oviatt, Levow, et al., 1998; Oviatt, MacEachern, et al., 1998; see also the discussion in Stent et al., 2008, p. 166). For this reason, the present paradigm employs a (simulated) human interlocutor.

Second, the studies summarized above employed specific requests for clarifications to elicit hyper-articulated productions. In those paradigms, participants typically are asked to produce the same word again immediately