Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 163–178
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Adapting speaking after evidence of misrecognition: Local and global hyperarticulation

Amanda J. Stent [a,c,*], Marie K. Huffman [b], Susan E. Brennan [a,c]

[a] Department of Computer Science, State University of New York at Stony Brook, Stony Brook, NY 11794, USA
[b] Department of Linguistics, State University of New York at Stony Brook, Stony Brook, NY 11794, USA
[c] Department of Psychology, State University of New York at Stony Brook, Stony Brook, NY 11794, USA

Received 14 November 2006; received in revised form 27 July 2007; accepted 28 July 2007

Abstract

In this paper we examine the two-way relationship between hyperarticulation and evidence of misrecognition of computer-directed speech. We report the results of an experiment in which speakers spoke to a simulated speech recognizer and received text feedback about what had been "recognized". At pre-determined points in the dialog, recognition errors were staged, and speakers made repairs. Each repair utterance was paired with the utterance preceding the staged recognition error and coded for adaptations associated with hyperarticulate speech: speaking rate and phonetically clear speech. Our results demonstrate that *hyperarticulation is a targeted and flexible adaptation rather than a generalized and stable mode of speaking*. Hyperarticulation increases after evidence of misrecognition and then decays gradually over several turns in the absence of further misrecognitions. When repairing misrecognized speech, speakers are more likely to clearly articulate constituents that were apparently misrecognized than those either before or after the troublesome constituents, and more likely to clearly articulate content words than function words. Finally, we found no negative impact of hyperarticulation on speech recognition performance.
Published by Elsevier B.V.

Keywords: Hyperarticulation; Clear speech; Speaking rate; Adaptation in speaking; Speech recognition; Spoken dialog

1. Introduction

Speech recognition technology has made its way into many telephone and information applications in wide use by the general public; people routinely encounter the option of speaking to a machine when they request phone numbers, make collect calls, and seek information about schedules, events, or accounts. Most speech applications used by the public achieve acceptable performance by strongly constraining what users can say—for instance by asking users questions with yes or no answers or by presenting menus containing just a few items with short labels that users are invited to repeat. By seizing most or all of the initiative, spoken dialog systems increase the likelihood that input utterances will be predictable and recognizable (Schmandt and Arons, 1984; Schmandt and Hulteen, 1982). In contrast, applications that recognize spontaneous, unconstrained utterances, such as dictation programs, have many fewer users, who need to be motivated enough to co-train with a particular application over time.

A long-standing goal of the speech and dialog research communities has been to enable less constrained, more flexible, mixed-initiative interaction with spoken dialog systems (e.g., Allen et al., 2001; Gorin et al., 2002); this goal has yet to be realized. The problem is that speech is highly variable. In addition to those variations characteristic of

* Corresponding author. Address: Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400, USA. Tel.: +1 631 335 2849; fax: +1 631 632 8334.
E-mail addresses: amanda.stent@stonybrook.edu, amanda.stent@gmail.com (A.J. Stent), marie.huffman@stonybrook.edu (M.K. Huffman), susan.brennan@stonybrook.edu (S.E. Brennan).

0167-6393/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.specom.2007.07.005

individual speakers (e.g., voice quality, dialect, and idiosyncratic pronunciation), there is variation in lexical choice and choice of syntactic structures, as well as prosodic or articulatory variability (due, e.g., to emphasis, affect, fluency, or even the speaker having a cold). Generally speaking, variability is associated with error: larger vocabularies and greater syntactic flexibility are associated with higher perplexity and, correspondingly, with higher word error rates (Huang et al., 2001), and disfluent or fragmented utterances, with recognition errors (Core and Schubert, 1999). To the extent that a source of variability is systematic, it can be described and modeled, which (in theory at least) should lead to ways in which to handle it successfully.

Through the experiment presented in this paper, we examine the causes and consequences of a kind of adaptive variation in speaking that has been loosely labeled *hyperarticulation*. When speakers believe that their addressees cannot understand them, they adapt in a variety of ways, such as by speaking more slowly, more loudly, and more clearly. Speakers have been found to adapt their speech to babies (Fernald and Simon, 1984), to foreigners (Ferguson, 1975; Sikveland, 2006), in noisy rooms (Summers et al., 1988) or on cell phones, as well as to computer-based speech recognizers. Each of these situations inspires a set of distinct but overlapping adaptations (see Oviatt et al., 1998a,b for discussion). For example, utterances directed to young children as well as those directed to speech recognizers tend to be shorter than those to adults; at the same time, child-directed speech typically has expanded pitch contours (Fernald and Simon, 1984) while machine-directed speech does not. Although hyperarticulation can improve intelligibility in speech directed at people (Cutler and Butterfield, 1990; Picheny et al., 1985), especially in the listener's native language (Bradlow and Bent, 2002), it can also result in increased error rates in automated speech recognizers (Shriberg et al., 1992; Soltau and Waibel, 1998; Wade et al., 1992).

The relationship between hyperarticulation in speaking and misrecognition by computers is thought to be bi-directional. This relationship has been described by some as a spiral in which evidence of misrecognition causes speakers to hyperarticulate, in turn causing even more recognition errors (e.g., Hirschberg et al., 1999; Levow, 1998; Oviatt et al., 1998a; Soltau and Waibel, 2000b). For example, in one study of machine speech recognition, an utterance produced right after a misrecognized utterance was itself misrecognized 44% of the time, compared to only 16% when produced after a correctly recognized utterance (Levow, 1998). Because of such observations, it has been widely presumed that increased error rates in automatic speech recognition are due to hyperarticulation. However there is a shortage of systematic data documenting the effects of specific features of hyperarticulation on speech recognition performance, as well as the persistence or actual time course of this kind of adaptation over the course of a human–machine dialog.

1.1. Elements of hyperarticulation

Hyperarticulation is really an umbrella term for many different adaptations in speaking, including prosodic adaptations due to speaking more slowly, pausing more often, and speaking more loudly, as well as segmental adaptations due to replacing reduced or assimilated forms of vowels and consonants with more canonical forms. As used in the literature, the term *hyperarticulation* is sometimes equated with *clear speech*, and often contrasted with *casual speech* (e.g., Moon and Lindblom, 1994) or *conversational speech* (e.g., Picheny et al., 1986; Levow, 1998; Krause and Braida, 2004). But the distinction is not a simple binary one. Hyperarticulate speech is a gradient phenomenon (e.g., Moon and Lindblom, 1994; Oviatt et al., 1998b); the properties of speech that vary during hyperarticulation do not all vary at the same rates or under the same conditions.

Perhaps the most detailed analyses of both prosodic and segmental aspects of hyperarticulate speech have been provided by Oviatt and colleagues (Oviatt et al., 1998a,b). These studies examined the duration of utterances, segments and pauses; pause frequency; F_0 minimum, maximum, range and average; amplitude; intonation contour; and the incidence of these segmental features: stop consonant release, /t/ flapping, vowel quality, and segment deletion. These studies used a simulated ("Wizard of Oz") multimodal spoken dialog system and a form-filling task. Users were given staged error messages at random points in the dialog; this elicited matched pairs of short utterances with the same wording by the same speaker, produced before and after evidence of speech recognition error. In a corpus of 250 paired utterances, speakers spoke more slowly (by about 49 ms/syllable) and paused longer and more often after evidence of recognition failure than before, whether they experienced high (20%) or low (6.5%) error rates; this hyperarticulation was not accompanied by much variation in amplitude and pitch (Oviatt et al., 1998b). Only the speakers who experienced the higher error rate produced clearer phonetic segments (e.g., released stop consonants) after error messages than before (Oviatt et al., 1998b).

The second study in this series by Oviatt and colleagues provided acoustic evidence that hyperarticulation in speech to machines is targeted to the perceived problem within an utterance, rather than produced as a persistent, non-specific adaptation in speaking style. A somewhat larger corpus of 638 pairs of utterances produced by 20 speakers (and elicited using the same task, the same simulated-error technique, and a 15% error rate, with errors distributed randomly during the dialog) yielded consistent increases in features of hyperarticulation across paired utterances (Oviatt et al., 1998b). These included prosodic adaptations such as increased duration and pausing as well as segmentally clearer forms on 6% of repetitions. In a further analysis of 96 paired utterances, speakers hyperarticulated most during the part of the repaired utterance perceived to have been problematic (Oviatt et al., 1998b). That is, speech at