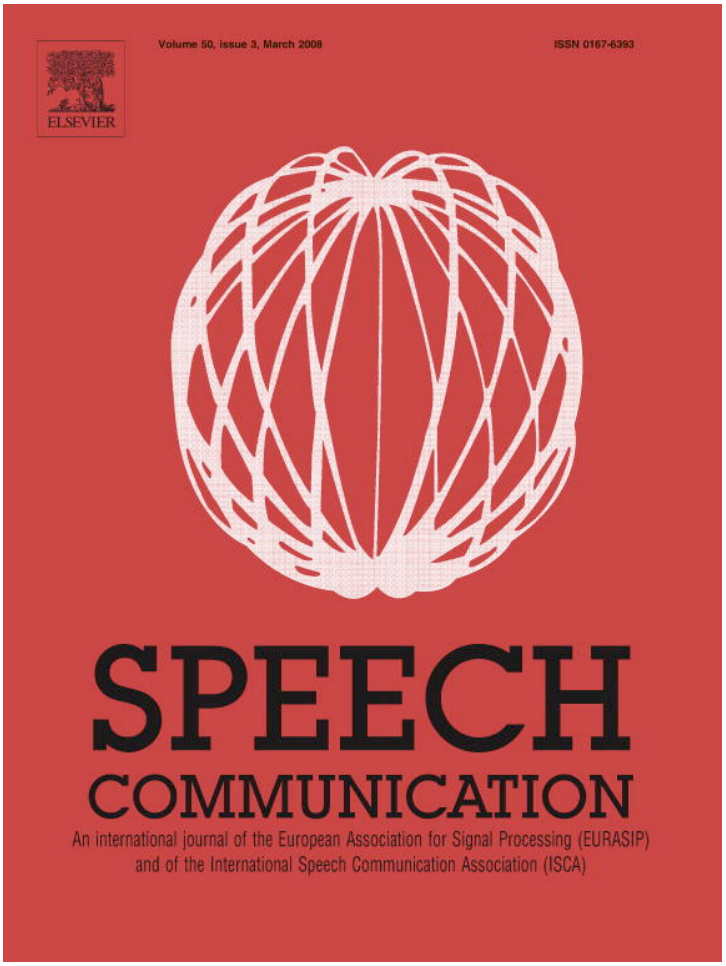

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 163–178
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

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

the focal area of the repair was greater in pitch range (11%), amplitude (1%), pausing (149%), and duration (11%) than adjacent segments before or after.

Another study (Levow, 1999) analyzed spontaneous commands directed at an interactive, working system (the Sun Microsystems SpeechActs system). Utterances produced following two types of system error were analyzed: those after a general failure (where the system simply indicated that it could not process the input) and those after a misrecognition (in which part of the utterance was correctly recognized and part was not, as evident from the feedback provided to the user). Both types of error resulted in more pausing and longer word durations, particularly in utterance final position, but the effect was stronger after misrecognition errors. In addition to these prosodic changes, there were also segmental changes during repairs, in the form of higher incidence of full vowels and released stop consonants. These segmental changes apparently did not depend on the type of the preceding error (misrecognition or general failure).

1.2. Effects of hyperarticulation on automatic speech recognition

Although hyperarticulation has been widely blamed for speech recognition errors (e.g., Levow, 1999; Oviatt et al., 1998a,b), the effect is by no means large, determinate, or well understood. Relatively few studies have systematically examined the effects of hyperarticulate speech on automated speech recognition (ASR). One set of studies (Shriberg et al., 1992; Wade et al., 1992) looked at dialogs with a working spoken dialog system, DECIPHER™, with which speakers were able to take substantial initiative and produce spontaneous, relatively long utterances. Speakers experienced a higher word error rate in their first session with DECIPHER™ than their second (20.4% vs. 16.1%), suggesting that they successfully adapted their speech to the system over time. In this experiment, the speakers' recorded utterances were subjectively categorized by human raters on a three-point scale as natural-sounding, hyperarticulated in portions, or completely hyperarticulated and then re-processed through the speech recognizer with a bigram and a 'nogram' language model. Overall, reduction in word error rate for utterances from the first to the second session was about 4% regardless of language model, suggesting that the reduction in word error rate was due to speakers' prosodic and segmental adaptation rather than any adaptation to the system's grammar.

Most speakers in that experiment actually reduced their use of hyperarticulation from the first to the second session. However, improved performance by DECIPHER™ was due not only to reduced frequency of hyperarticulation, but also to adaptation in the *nature* of hyperarticulation. While utterances rated as strongly hyperarticulated yielded higher word error rates than ones not so rated, even the strongly hyperarticulated utterances from the second session were better recognized than the strongly hyperarticulated ones from the first session (Wade et al., 1992). Wade et al. documented that over time, the hyperarticulated utterances actually became more acoustically similar to the data on which the speech recognizer had originally been trained (whereas the natural-sounding utterances did not). This set of findings highlights the need to better understand just what about the broad category of "hyperarticulation" is detrimental to speech recognizer performance; the mapping of speaking style to word error rate is not a simple one.

Another study that used corpora of utterances directed to a working spoken dialog system (the TOOT and W9 corpora, Hirschberg et al., 1999, 2000, 2004) analyzed the acoustic–prosodic characteristics of recognized vs. misrecognized utterances. Significant differences were found in loudness, pitch excursion, utterance length, pausing, and speaking rate. As in the studies by Shriberg et al. (1992) and Wade et al. (1992), utterances were rated subjectively on a three-point scale as to whether they sounded hyperarticulated. Utterances rated as sounding hyperarticulated were more likely to have been misrecognized, and misrecognized utterances had higher hyperarticulation ratings; moreover, utterances rated as *not* hyperarticulated were more likely to have been misrecognized when they were higher on objective loudness, pitch, and durational measures (Hirschberg et al., 1999). In follow-on work, Hirschberg et al. identified and labeled corrections in these corpora; compared to non-corrections, corrections were significantly longer and louder and had a slower speaking rate, longer prior pause, higher pitch and less silence. 52% of corrections vs. 12% of non-corrections were subjectively rated as sounding hyperarticulated. Corrections were more likely to be misrecognized than non-corrections, and hyperarticulated corrections than non-hyperarticulated ones. However, the number of misrecognized corrections varied by type, with corrections including additional information and paraphrases being misrecognized at higher rates than repetitions and corrections omitting information (Litman et al., 2006).

A third set of studies confirmed that hyperarticulation lowers word accuracy in ASR (Soltau and Waibel, 1998, 2000a,b). These studies elicited a corpus of highly confusable word pairs in either German or English as baseline pronunciations, for comparison with pronunciations after simulated evidence of error in a dictation task. These ASR studies measured not only adaptation in speaking rate but also hyperarticulation of phonetic segments. In English, phone duration increased by 28% on average (44% for voiced plosives but only 16% for vowels; Soltau and Waibel, 2000a) and in German, 20% (with the greatest increases for voiced consonants and schwa sounds, Soltau and Waibel, 2000b). Recognition of before and after error tokens of isolated words was compared using the JANUS-II speech recognition toolkit (with a 60K vocabulary for German and a 30K vocabulary for English). These studies report on the order of 30% more errors in hyperarticulate than casual speech (Soltau and Waibel, 2000a).

1.3. Strategies for avoiding hyperarticulation

Although users of spoken dialog systems are often explicitly instructed to speak naturally, it is questionable whether this strategy works for minimizing misrecognition. For example, when speakers in one study were told not to "overenunciate", they produced utterances that yielded lower subjective ratings of hyperarticulation, and yet this adjustment did not result in reliably lower ASR error rates (Shriberg et al., 1992).

Oviatt and colleagues (Oviatt et al., 1998b) also looked for prosodic and segmental differences in speech in response to three different kinds of error messages: those for which users saw only the message "????", those for which the system apparently substituted a related (semantically plausible) word in the utterance, and those for which it substituted an unrelated (semantically implausible) word. These situations (experienced by all the speakers) led to no differences in prosodic or segmental measures of hyperarticulation.

To summarize, previous research on hyperarticulation in human–computer interaction has shown that when speakers experience misrecognition, they adapt by exaggerating their speech: speaking more loudly and more slowly, with greater variety in pitch, and with greater attention paid to the articulation of certain phonemes. Speakers focus their hyperarticulation on the part of the utterance that was misrecognized. The impact on speech recognition performance is unclear: Misrecognized utterances exhibit features of hyperarticulation, and on isolated word tasks hyperarticulate tokens are more likely to be misrecognized. On the other hand, in spoken dialog to a computer where users can produce continuous speech, reduced word error rates over time are partly due to adapation in the *nature* of hyperarticulate speech and to syntactic and lexical adaptation, as well as to reduction in the *amount* of hyperarticulate speech.

1.4. Rationale and predictions

Our goal for the current project was to investigate:

(1) How speakers adapt spontaneous speech directed at spoken dialog systems after they receive evidence of misrecognition. When speakers encounter evidence that an utterance was misrecognized, they should repair by repeating the utterance more slowly; and forms that had been relaxed in the "before" utterance should tend to be replaced by clear forms in the "after" version.
(2) How long adaptations in response to evidence of misrecognition persist during a dialog. We expected that segmental adaptations would be targeted to troublesome parts of the utterance (local adaptation); we were interested in whether segmental and prosodic adaptations would persist over turns (global adaptation). We were particularly interested in whether hyperarticulation to a computer is like a "switch": an adaptation that, once turned on, persists mostly independent of later system behavior (as suggested by the notion of "spiraling errors"); or whether it is like a "dial" that is adjusted gradually during the interaction.
(3) When (or whether) adaptations in response to evidence of misrecognition cause problems for speech recognition. We investigated the effects of hyperarticulation on ASR systems trained on broadcast speech and conversational speech and configured with different statistical language models (word list, unigram, bigram, and trigram), as well as for a grammar-based ASR. We were not primarily interested in staging a competition between ASR systems, but in establishing whether the features of hyperarticulate speech are really as severe a problem as has been assumed, and which features of hyperarticulation (prosodic or segmental) are problematic.

Because speech read aloud has different prosodic, segmental, and fluency characteristics than spontaneous speech, and because we wanted to examine speech generated by speakers who were trying to repair errors, we did not have speakers read sentences aloud, as in most other controlled studies of hyperarticulation (e.g., Harnsberger and Goshert, 2000; Johnson et al., 1993). We used a Wizard-of-Oz procedure (adapted from Brennan, 1991, 1996; Oviatt et al., 1998a,b) to collect a corpus of spontaneous utterances from naive volunteer speakers who were led to believe that they were interacting with an ASR in order to enter information into a computerized database. In fact, the system's responses were simulated by a human operator behind the scenes. To elicit paired tokens with identical lexical and syntactic content from each speaker that could be compared for hyperarticulation, we adapted Oviatt and colleagues' (Oviatt et al., 1998a,b) and Soltau and Waibel's (1998) method of simulating errors by providing spurious error messages so that speakers would spontaneously repeat utterances.

We wished to extend Oviatt and colleagues' findings by looking not only at focal prosodic adaptations within repairs, but also at segmental adjustments before, during, and after the problematic word(s). Unlike Oviatt et al. (1998a,b), our errors appeared at pre-planned locations in the dialog for *all* the speakers, so that we could examine designated target words for hyperarticulation. This was an important property of the corpus we collected, as it enabled us to systematically conduct both local and global analyses of the persistence of hyperarticulation by multiple speakers, over multiple utterances, and across parts of the dialog that had higher and lower incidence of errors. We also wished to extend previous research on the impact of hyperarticulation in spoken dialog to a computer by looking at the impact of hyperarticulate speech on automatic speech recognition.

We used a task that enabled us to elicit spontaneous speech in the form of complete sentences containing multi-

ple tokens of words with specific phonetic segments, articulated within controlled contexts. This is difficult to do, but not impossible (e.g., Brennan, 1996; Kraljic and Brennan, 2005). Fortunately, what speakers choose to say can be constrained implicitly by the dialog context to some degree. Previous studies of lexical and syntactic entrainment have demonstrated that speakers are strongly influenced by a dialog partner's words and syntax and tend to re-use these elements (Brennan, 1991, 1996; Brennan and Clark, 1996). In fact, the tendency to entrain on a partner's wording and syntax occurs not only with human partners, but also with computer partners (Brennan, 1991), and this is true whether the currency of communication is speech or text (Brennan, 1996).

We aimed to collect a speech corpus that met the following criteria: it should contain (1) *spontaneous* speech, (2) in the form of *sentences*, (3) by *multiple* speakers, (4) who produced *target words with particular phonetic segments*, (5) in relatively *controlled phonetic environments*, (6) in a *dialog context* in which they received responses contingent upon their utterances, (7) enabling us to collect *paired tokens* of the same utterance, before and after the speaker received evidence that the utterance was misrecognized.

2. Method

2.1. Task and setup

We designed an information-entry task to elicit spontaneously planned yet predictable utterances. Participating speakers were supplied with a one-page spreadsheet depicting a database of a hypothetical children's softball team containing the children's names, positions on the team, parents' occupations, and what the children would bring to sell at two fund-raising events (a food and kitchen items sale and a garage sale). Speakers were to use this spreadsheet to look up the answers to questions they would be asked and present their answers by speaking (following the procedure in Brennan, 1996). They were told to answer in complete sentences. Feedback from the "dialog system" was provided as text messages. When the speaker made a speaking error (for example, using an incomplete sentence or abandoning an utterance), the system produced an *unplanned error message* (e.g., "Complete sentences are necessary for the database – please repeat"). In other cases, the system displayed a message in the form "You said:", followed by a transcription of the participant's utterance. Sometimes, when the utterance was the site of a *planned error*, the transcription would contain a "misrecognition" 1–6 words long. This was done to localize the site of the misrecognition within the utterance and the interaction. By analyzing speech before, during, and after these misrecognition sites, we hoped to discern the time course of hyperarticulation adaptations. Table 1 shows example planned and unplanned errors.

The planned errors were distributed throughout the dialog as follows: of the 66 total answer turns to be provided by speakers, #1–24 were associated with only two planned errors (an 8.3% error rate), utterances #25–47 with eight planned errors (a 33% error rate), and utterances #48–66 with nine planned errors (a 50% error rate that began with a highly salient run of five errors in a row, followed by a run of five non-errors and then some alternation). To make the recognition errors as realistic as possible, we had a research assistant produce the answers to the questions while talking to computer-based dictation software; the planned error messages in the experiment were based on the resulting misrecognitions.

We elicited spontaneous spoken sentences by having speakers answer questions heard over a headset, prerecorded in a female voice (e.g., "Who is the catcher of the softball team?"). We made this modality distinction between the questions (speech) and system responses (text) because we wanted speakers to consider what they heard over the headphones to be prompts that did not necessarily originate from the system. This way, the recognition system would not be assumed to be a partner with prior knowledge of the discourse, and speakers would be less likely to use pronouns and ellipses.

2.2. Materials and stimulus words

We identified categories of sounds that can be spoken with both clear and relaxed forms and that would enable us to quantify hyperarticulation. The first is mid-word /t/ before an unstressed vowel, as in *water*. In relaxed speech,

Table 1
Examples of planned and unplanned error/repair sequences (Q = Question; U = Uset's answer; S = System's feedback response)

Planned error	Unplanned error
Q. What is Kate Tolstoy bringing to the food sale?	Q. What is Ryan Dade bringing to the food sale?
U. Kate Tolstoy is bringing some cookie dough and a picnic table to the food sale	U. Ryan Dade is bringing cat collars, and a basket, and pet toys to the food, to the garage sale, oops
S. You said: Kate Tolstoy is bringing some cooking label in a pickle to the food sale	S. Please repeat
U. Kate Tolstoy is bringing some cookie dough and a picnic table to the food sale	U. Ryan Dade is bringing cat collars, a basket, and pet toys to the garage sale
S. You said: Kate Tolstoy is bringing some cookie dough and a picnic table to the food sale	

a /t/ in this context is said as a "flap", which is short and /d/-like, while in clear speech this would be a strong voiceless [t]. The next is word-final /t/ (as in *cat*), which in relaxed speech is produced without audible noise at the end of the oral closure for the sound; a clear speech form would have audible noise as oral pressure is released after oral closure. The third /t/ variant is mid-word /t/ after n, which may be absent in relaxed forms (as when *winter* sounds much like *winner*), and clearly voiceless and released in clear forms. As noted earlier, Levow (1999) and Oviatt et al. (1998a,b) found consonant release and unflapped /t/ to occur more frequently in corrections after system misrecognitions. These same features have also been reported in nonsense sentences when subjects are told to speak more clearly, as for a hearing impaired or non-native speaking partner (Picheny et al., 1986; Krause and Braida, 2004). The latter studies also report a higher occurrence of full vowels in function words, as opposed to the reduced vowel *schwa*. In our materials the indefinite article *a* occurred very frequently, and could thus be examined for changes in vowel quality. Finally, the *d* in the word *and* may be unpronounced in a relaxed form of the word, or audibly produced in a clear form of the word. Since a highly frequent variant of this word is the relaxed form with no /d/ (e.g., Bell et al., 2003 report that in the Switchboard corpus, /d/ was articulated in *and* only 14% of the time), presence of the /d/ was taken as a sign of clearer, or hyperarticulated, speech.

We chose a set of target words containing several tokens for each of these sound categories. The database provided to experiment participants contained these target words (see Table 2). One of the experimenters recorded a set of 66 prompting questions about this database; the questions were worded to evoke answers in the form of target sentences with phonetic environments roughly matched for coarticulatory environment and stress (see Table 3).

Table 2
Database with target words

Child's name	Team position	Mother's job	Father's job	Garage sale	Food and kitchen items sale
Dawn Lepko	catcher	party planner	bail bondsman	a toy whale a suitcase	30 cupcakes a set of coasters
Peter Lipton	pitcher	reporter	scientist	a lab coat arrowheads and spear points	tea cups a pear tart
Joy Wade	first base	tutor	dog catcher	a gold key chain & a teardrop charm a Bad Boyz tape	peach pies soup bowls
Sammy Dale	second base	teacher	dentist	20 lotto tickets a chalk board	40 oatmeal cookies beer mugs
Kate Tolstoy	third base	maid service owner	car mechanic	a talking doll a hope chest	some cookie dough a picnic table
Linda Adams	shortstop	potter	ship captain	pipe cleaners and beads boat poster	a sixpack of Tab a fat gram chart
Donna Esposito	left field	sales clerk	truck driver	a soapdish a peat spreader	a chocolate cake apple strudel
Mary Deed	right field	florist	mail carrier trainer	matching backpack and tote bag a golf bag	soy burgers an angel food cake
Lisa Evans	center field	ghost writer	dance instructor	a Santa outfit and boot polish a hat box and a beard trimmer	soup napkin rings
Ted Smith	relief pitcher	manager	landscaper	a dart game & a deer statue a goose decoy & a beebee gun	kale baby cups
Ryan Dade	outfielder	vetrinarian		cat collars a basket and pet toys	cake dish pet food bowl
Martin McCoy	umpire	lawyer	head coach	starter skis & roller skates a toy space suit	cutting board coffee maker
Bob Carter	infielder	doctor	ad man	some kite string float tubes and a red beach ball	sugar cookies planter's nuts
Marcella Asner	designated hitter	engineer	plumber	a hoop skirt tap shoes and a night light	hot dogs a spatula
Mark Pitney	umpire	talk show host	life insurance salesman	a large Yankees cap an umbrella	chocolate chip cookies
Deb Kanter	substitute catcher	flute teacher	computer programmer	a bait box and a spade a bike pump	a cooler a knife block
Joe Peck	water and bat boy	artist	book publisher	a laptop a pencil and pad set	

Speakers used a printed copy of this database to find information that they then used to spontaneously plan and produce sentences directed to the simulated speech recognizer.
The database contained information about members of a children's softball team, their parents' occupations, and two upcoming fund raising events for the softball team: a garage sale and a sale of food and kitchen items. The speakers were prompted by pre-recorded questions played through a headset. The target words in each cell in the table are underlined.

Table 3
Pre-recorded question prompts

Question	Question text
1	Who are Kate Tolstoy's parents?
2	What is Mary Deed bringing to the food sale?
3	Who is the scientist's son?
4	What is Deb Kanter bringing to the kitchen item's sale?
5	What is Donna Esposito bringing to the garage sale?
6	Who are Joy Wade's parents?
7	What is Ryan Dade bringing to the food sale?
8	Who is Deb Kanter's mother?
9	What is Lisa Evans bringing to the food sale?
10	Who are Dawn Lepko's parents?
11	What is Ted Smith bringing to the garage sale?
12	What is Sammy Dale bringing to the food sale?
13	Who is Lisa Evans' mother?
14	What is Martin McCoy bringing to the food sale?
15	Who is Donna Esposito's father?
16	What is Bob Carter bringing to the food sale?
17	What position does Joe Peck play?
18	What is Linda Adams bringing to the garage sale?
19	Who is Ted Smith's mother?
20	Who are Joe Peck's parents?
21	What is Donna Esposito bringing to the food sale?
22	What is Dawn Lepko bringing to the garage sale?
23	Who are Sammy Dale's parents?
24	What is Dawn Lepko bringing to the food sale?
25	Who is Martin McCoy's father?
26	What is Joy Wade bringing to the food sale?
27	Who plays right field?
28	What is Lisa Evans bringing to the garage sale?
29	Who is the relief pitcher?
30	What is Marcella Asner bringing to the food sale?
31	Who plays second base?
32	What is Linda Adams bringing to the food sale?
33	What is Ryan Dade bringing to the garage sale?
34	Who is the infielder?
35	What is Mark Pitney bringing to the food sale?
36	Who are Mary Deed's parents?
37	What is Ted Smith bringing to the food sale?
38	What position does Lisa Evans play?
39	Who plays third base?
40	Who is Peter Lipton's mother?
41	What is Sammy Dale bringing to the garage sale?
42	Who plays outfield?
43	What is Kate Tolstoy bringing to the food sale?
44	Who is the softball team's umpire?
45	What is Deb Kanter bringing to the garage sale?
46	Whose father is a dance instructor?
47	What is Martin McCoy bringing to the garage sale?
48	What is Mary Deed bringing to the garage sale?
49	Who plays left field?
50	What is Peter Lipton bringing to the food sale?
51	What is Kate Tolstoy bringing to the garage sale?
52	Who are Linda Adams' parents?
53	What is Bob Carter bringing to the garage sale?
54	Who are Marcella Asner's parents?
55	What is Mark Pitney bringing to the garage sale?
56	Who plays shortstop?
57	Who plays first base?
58	What is Peter Lipton bringing to the garage sale?
59	Who is Mark Pitney's father?
60	What is Joy Wade bringing to the garage sale?
61	What is Marcella Asner bringing to the garage sale?
62	What is Joe Peck bringing to the garage sale?
63	Who is not bringing anything to the food sale?
64	Who is the softball team's catcher?
65	Who is the substitute catcher?
66	Who are Bob Carter's parents?

2.3. Subjects

Sixteen undergraduate students (nine women and seven men, mean age approximately 22 years) from the State University of New York at Stony Brook volunteered to participate in the experiment and received either research participation credit in one of their psychology courses or $7. All were computer users with minimal or no experience with speech recognizers. They were told that the purpose of the experiment was to study speech recognition by computers. Two additional students experienced equipment failure during their experimental sessions and their data were discarded. All 16 speakers identified themselves as native speakers of English; 10 were monolingual and the remaining 6 were bilingual but English-dominant. In interviews after the experiment, all speakers reported that they believed that they had been speaking to a computer. They were then debriefed about the need for simulation and told that a research assistant had been providing the responses.

2.4. Procedure

After consenting to participate in the experiment, speakers were fitted with a headphone microphone set and seated before a computer display. The experimenter checked the voice level from the microphone and explained that they would be performing a data entry task. The experimenter then had the speaker listen to the first question over the headset, helped the speaker locate the answer on the paper spreadsheet, and instructed the speaker to press a *Speak* key and speak the answer in a complete sentence. When the key was released, the screen displayed the status message "Processing...", and a few seconds later, a text feedback message corresponding to what the system had recognized appeared on the screen (the first utterance was always recognized exactly as spoken). Saying "let's try another one", the experimenter had the speaker hit the *Ready* key and then listen to and answer the next question. On this second trial, the system displayed a message with evidence of misrecognition. The experimenter reacted by saying "oops, I guess it's not perfect" and then demonstrated how to repair the system's error by having the speaker press the *Speak* key and say the correct answer again. After this, the experimenter asked if the speaker had any questions; if there were any, the experimenter provided clarification and stayed in the room for a third trial. After the speaker indicated that the task was understood, the experimenter mentioned that there would be 66 questions in all, and left the room.

In order to enable the 'system' operator to provide feedback responses that were tailored to the speakers' answers in a realistic way (within a few seconds), we used a configuration in which a research assistant in another room monitored the speaker's utterance (e.g., "The team's catcher is Dawn Lepko") and operated a control panel to select a feedback response from a menu of pre-stored sentences for the prompt (e.g., "You said: 'The team's catcher is

Dawn Lepko"'). The pre-stored feedback messages included a set of variants on the expected answer to each question in the task so that the operator would be able to respond as quickly as possible (we obtained these variants by piloting and adding new variants as they were produced); for a sample set, see Table 4. Whenever speakers departed from this expected set, the operator would select the closest match and edit it to reflect exactly what the speaker had said, apart from any stuttering or disfluencies such as "um". If the speaker restarted an utterance, stuttered more than once or produced a major disfluency, the operator responded with "Please repeat." As in the second practice trial, at the planned error points the operator sent feedback with evidence of misrecognition to the speaker's screen. The operator's control panel is shown in Fig. 1.

Table 4
Sample variants on the expected spontaneous answer to a question prompt, *Who is the catcher of the softball team?* Speakers were instructed to answer in complete sentences

The catcher is Dawn Lepko
The team's catcher is Dawn Lepko
The softball team's catcher is Dawn Lepko
Dawn Lepko is the catcher
Dawn Lepko is the team's catcher
Dawn Lepko is the softball team's catcher
Dawn Lepko is the catcher of the softball team

The control panel (see Fig. 1) enabled the operator to select the text that exactly matched the speaker's utterance, or else to rapidly edit one variant to make it match. A text message was then generated and sent to the speaker's screen (e.g., "You said: *Dawn Lepko is the catcher*.").

During the experimental session, the operator of the Wizard-of-Oz program sent text messages back to the speaker rapidly after each utterance (average response time: 3.6 s; range of response times: .02–40.5 s with all but 100 being under 10 s). The Wizard-of-Oz system logged the filenames for recorded utterances and the feedback messages. After each session, the feedback messages were hand-corrected by a research assistant to produce a true transcript of the speaker's utterances. The transcripts enabled us to identify cut off utterances, to identify when the speaker re-worded the previous utterance as opposed to repeating it verbatim, and later, to compare speech recognition output to what the speaker had originally said.

2.5. The corpus

The experiment was designed to provide a corpus of at least 1360 planned utterances (16 speakers × 66 answers + 19 planned repairs). Speakers varied in the numbers of unplanned error messages they received (from 1 to 25, with a mean of around 6); due to the ensuing repairs, the total corpus comprised 1512 utterances. Utterances containing a major disfluency (other than minor stuttering, pausing, *um* or *uh*) were removed from the corpus, as were those cut off due to speakers depressing the press-to-speak key late or releasing it too early, and those whose speaking rates represented extreme outliers (outliers represented fewer than 2%). This resulted in a corpus of 1202 analyzable utterances. Of these, 373 were planned repairs, each paired with a preceding utterance. The analyzable utterances averaged 3.9 s in length.

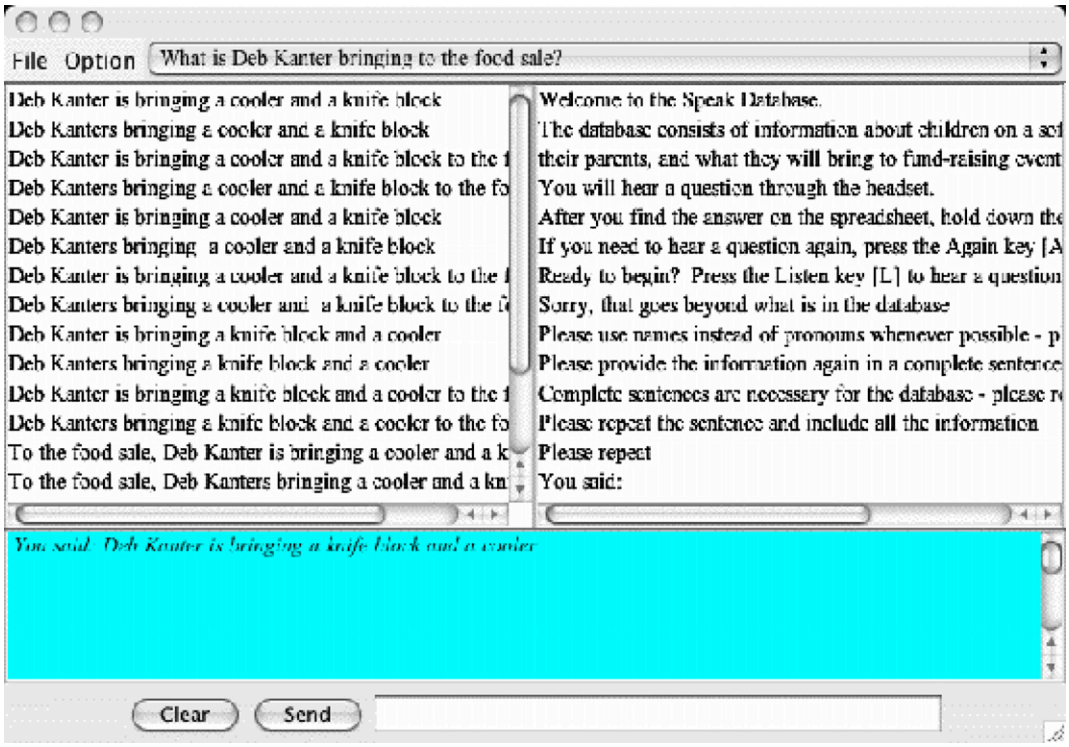

Fig. 1. Screen shot of the Wizard-of-Oz (simulation) control panel.

2.6. Coding and data analyses

Two main sets of measures were made on this corpus to characterize speakers' adaptations to evidence of misrecognition. First, utterance lengths were computed from the recorded audio for each utterance. The number of words in each utterance was automatically converted to a number of syllables using a list of the words in the corpus and their lengths in syllables, and then rate of speaking in syllables per second was calculated for each utterance. Then, a second set of analyses was done on the "before" and "after" tokens of the utterances that were associated with planned errors. These utterances had been designed to include specific target words that contained one or more of the five phonetic segments that we expected would be produced in hyperarticulate form during repairs (/t/ tapping, word-final /t/ release, /t/ release after /n/, the full vowel in indefinite articles, and /d/ in *and*).

Phonetically trained transcribers evaluated the target words in the utterances before and after the planned error messages to determine whether word-medial and word-final segments were either flapped or released and whether indefinite articles contained full vowels. On this basis, the target words were classified as either relaxed or clear e.g., if the word *lotto* contained a medial flap, then it was classified as *relaxed*; if its medial [t] was released, it was classified as *clear*.

For each utterance in the corpus, we then computed the distance in number of utterances from the last error (planned or unplanned). So that we could examine whether hyperarticulation within an utterance would be global (affecting the entire utterance) or focal (focused on the 'misrecognized' part of the utterance), the segmental features of interest were also coded as to whether they appeared before, during, or after the 'misrecognized' part of the utterance.

We conducted both *local* analyses of hyperarticulation (within the utterance) and *global* analyses (across a series of utterances). For local analyses, we used ANOVAs to compare speaking rates (in syllables per second) and percentages of target words that were pronounced in clear (vs. relaxed) form for paired utterances just before or after an error message. We conducted two kinds of global analyses using all the utterances (including those that did not evoke error messages). First, we examined the effects of the recency of the last error message on speaking rate, to chart the decay of hyperarticulation. Second, because the overall rates of planned errors differed across the thirds of the dialog, we looked for an impact of higher vs. lower simulated misrecognition rates on hyperarticulation. Whenever possible, two ANOVAs are presented for each analysis, as we wish to generalize both to the behavior of the average speaker and to the effects on the average utterance or the average word. F_1 collapses the data by-subjects (speakers) and F_2, by-items (items are utterances, unless otherwise noted). To examine relationships among paired variables of interest such as speaking rate, clear speech, and speech recognizer word error rates, correlation coefficients were computed for each of the 16 speakers and transformed into Fischer's Z_r scores so that they could be averaged and tested for significance. The Z_r scores were then transformed back into correlations for reporting purposes.

2.7. Speech recognition testing

Although our primary focus was how hyperarticulation is adapted in response to cues of misrecognition (both within the utterance and across utterance sequences), we also looked at the impact of hyperarticulation on speech recognition performance. This problem has been previously studied (e.g., Wade et al., 1992; Hirschberg et al., 1999; Soltau and Waibel, 1998, 2000a,b), but usually in the context of speech recognizers using statistical language models rather than grammar-based speech recognizers. Also, previous researchers did not have access to the detailed segmental annotations created for this corpus, so could not examine the impact on speech recognition of segmental adaptations. Accordingly, after the experimental sessions were completed, the resulting corpus was processed off-line.

Our corpus is a corpus of conversational speech directed to the computer, but most data used for training speech recognizers is either read speech (e.g., the HUB4 corpus of broadcast news speech) or conversational speech directed to human partners (e.g., the HUB5 corpus, which includes parts of the Switchboard and CALLHOME corpora[1]). There is one large corpus of conversational speech directed at computers that has been used to train acoustic models: the CMU Communicator corpus (Bennett and Rudnicky, 2002). In addition, commercial speech recognizers may train on proprietary corpora of read or conversational speech. We processed our speech through five speech recognition systems, incorporating two approaches to language modeling and three different acoustic models:

(1) **Sphinx-word list**: A version of the Sphinx 3 statistical speech recognizer (available from http://cmu-sphinx.sourceforge.net/) with a language model consisting of a simple word list language model from our corpus, similar to the 'nogram' model of (Wade et al., 1992). We processed our corpus through this speech recognizer using two acoustic models: one trained on HUB4, and one on CMU Communicator.
(2) **Sphinx-unigram**: A version of Sphinx 3, with a unigram language model trained over our corpus. Again, we processed our corpus through this speech recognizer using acoustic models trained on HUB4 and CMU Communicator.

[1] These corpora are available from http://www.ldc.upenn.edu/.

(3) **Sphinx-bigram**: A version of Sphinx 3, with a bigram language model trained over our corpus and acoustic models trained on HUB4 and CMU Communicator.
(4) **Sphinx-trigram**: A version of Sphinx 3, with a trigram language model trained over our corpus and acoustic models trained on HUB4 and CMU Communicator.
(5) **Grammar-based**: A state-of-the-art speaker-independent commercial speech recognizer, with an acoustic model trained on a proprietary corpus of conversational and read speech. For a language model, we created a recognition grammar tailored to our corpus. Grammar-based language modeling has been shown to give a lower word error rate and comparable semantic error rate to statistical *n*-gram language modeling on unconstrained, in-domain utterances for spoken dialog (Knight et al., 2001).

Because our corpus is relatively small, we did no tuning of the acoustic models to our speakers or data. The recognition data from these systems were used to examine the extent to which hyperarticulation affects speech recognition. These data should not be interpreted as measures of the performance of the speech recognizers we used: our only goal was to compare measures of recognition errors for hyperarticulate and non-hyperarticulate speech.

3. Results and discussion

3.1. Relationship between main hyperarticulation measures

Our two main sets of measures of hyperarticulation, speaking rate and phonetically clear speech, were moderately but reliably (and negatively) correlated; that is, the more slowly an utterance was spoken, the higher proportion of clear segmental forms it contained. The average correlation of speaking rate and proportion of clear speech for the 16 speakers was $r_Z = -.239$, $p < .001$. Correlations of speaking rate and clear speech for individual speakers ranged from $r = +.08$ (no relationship) to $-.640$ (a large effect size, according to Cohen, 1988), demonstrating substantial variability among individuals. Details and discussion of individual differences are provided in Section 3.4.

3.2. Local effects of evidence of misrecognition on speaking

3.2.1. Local effects of error messages on hyperarticulation

As we expected, when speakers saw an error message indicating the system had misrecognized an utterance, they adapted speaking during the subsequent repair. This was true for both speaking rate and pronunciation. Speakers produced more segmentally clear forms in the utterance just after an error message than in the utterance before, 38% compared to 30%, $F_1(1, 15) = 17.42$, $p = .001$; F_2[2](1, 27) = 35.46, $p < .001$. They also spoke more slowly after receiving a planned error message than before, 3.62 syllables/s compared to 4.12 syllables/s, $F_1(1,15) = 49.23$, $p = <.001$; $F_2(1, 38) = 14.14$, $p = .001$. The same pattern held when matched pairs of utterances before and after *un*planned errors were added to the analysis (such messages included requests for the speaker to repeat due to disfluency, pronouns, ellipsis, or otherwise incomplete sentences); for all matched pairs of utterances taken together, repairs (*after* versions) were slower than *before* versions, 3.67 compared to 4.17 syllables/s, $F1(1,15) = 30.64$, $p < .001$.[3]

Of the five kinds of segmental variation coded in the paired utterances, three were more likely to be pronounced in their clear forms in the after version than the before version: mid-word /t/ vs. flap /D/, word-final /t/ release vs. non-release, and presence of /t/ release after /n/ vs. absence. The vowel quality of the indefinite article *a* and the presence of /d/ in *and* did not show this pattern of clearer speech after errors (see Table 5, last column). There were not sufficient numbers of observations distributed evenly enough across speakers for a significance test to be done for each of these five segmental features. However, the pattern suggests that hyperarticulation may work differently for content words than for function words. Note that the first three clear speech features occurred only in content words within our data set (mid-word /t/, word-final /t/ release, and /t/ release after /n/), whereas the last two occurred only in function words (/d/ release in *and*, and the full vowel in the indefinite article). When we compared the clear speech features occurring in content words to those in function words (the first three vs. the last two in Table 5), we found that content words were produced in their clear forms 13% more often in a repair than in the preceding utterance, while function words were produced in their clear forms only 4% more often, different at $F_1(1, 15) = 13.32$, $p = .002$; F_2[4](1,13) = 15.66, $p = .002$.

We offer three (non-competing) explanations for this finding. The first, a communication account, holds that, to the extent that the point of clear speech is to repair a particular troublesome message, speakers should be more likely to use the clear phonological form while repairing a content word than a function word, because content words are usually more critical for understanding the message (see, e.g., Ferreira et al., 2002).[5] The second, an ease-of-production account, suggests that it may be somewhat taxing to articulate several consecutive words in their clear forms (such as a determiner or coordinating word followed immediately by a noun) and so *a* or *and* produced in the

[2] Here, F_2 is collapsed by words rather than by utterances, as clear and relaxed forms vary across words.

[3] F_2, the ANOVA collapsed across items cannot be calculated here, as many of the unplanned items generated few or no error messages.

[4] Here, F_2 is collapsed by utterances because a given word can be either a content or a function word, not both (so an analysis collapsed by words is not useful). The second degree of freedom for F_2 is low due to missing data for some of the planned error utterances.

[5] We expect that in a situation in which speakers need to make a meaningful contrast between function words, these words would be hyperarticulated (and contrastively stressed) as well.

Table 5
Proportions of clear forms for the phonetic segments of interest, before and after evidence of misrecognition

Segment	Number of tokens in corpus	Proportion of clear forms before error	Proportion of clear forms after error (during repair)	Difference (after − before)
Word-final released /t/	238	0.2343	0.3697	0.1345
Mid-word /t/ vs. flap /d/	76	0.0026	0.1316	0.1053
/t/ after /n/	43	0.6977	0.8605	0.1628
Full vowel in def. article a	306	0.3954	0.4314	0.0036
/d/ in and	232	0.2716	0.3147	0.0043
Total/mean	897	0.3032	0.3795	0.0076

The positive difference in the last column indicates the proportion of times that segment was pronounced more clearly during a repair than before the repair.

vicinity of a hyperarticulated content word should tend to relax. The third account is that clear forms should be more likely to be associated with contrastive stress or a pitch accent, and since this is rare for function words, such words are less likely to be hyperarticulated. Future corpora designed with these possibilities in mind may be able to distinguish among these accounts.

Although target utterances in our study were not designed to elicit clear speech in vowels other than the indefinite article *a*, each speaker produced enough tokens (1–3) of *ee* (as in *deed*), *eh* (as in *peck*) and *ay* (as in *Dade*) to support a post hoc look at the acoustic properties of vowels in before and after error matched pairs of target words. For these items, we measured the frequencies of the lowest two resonances of the vowels (formants), which roughly represent how front (formant 2) and how low (formant 1) the vowels are produced in the mouth. Figs. 2a and 2b plot frequencies for these two formants at vowel midpoint for two speakers who demonstrated the most common patterns observed.[6] Consistent with previous work (Johnson et al., 1993; Bradlow, 2002), the vowel *ee* as in *deed* was fronter (higher formant two) in repaired forms (for 75% of speakers). Repaired forms of *deed* were also usually lower (higher formant one) in the vowel space (44%). The vowel *eh* as in *peck* was usually fronted (67% of the time), also consistent with (Johnson et al., 1993) and was most often somewhat raised in the vowel space (50% of the time). We could not systematically summarize changes in the *ay* phoneme (as in *Dade*) across all speakers, since *ay* is monophthongal for some speakers and diphthongal for others; however, for the two speakers represented in Figs. 2a and 2b, *ay* was monophthongal, and turned out to be slightly fronter and higher in repaired forms. The overall pattern, then, was for front vowels in repaired target words to become even more fronted (i.e., more peripheral) than they were in their pre-repair forms.

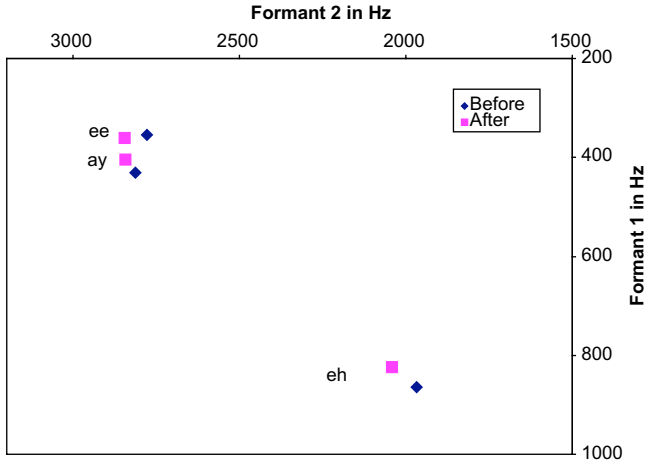

Fig. 2a. Formant values for ee, ay and eh before and after an error, first subject.

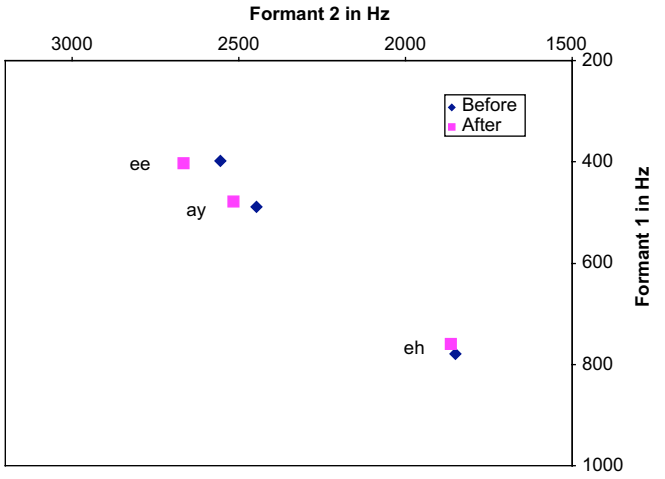

Fig. 2b. Formant values for ee, ay, and eh before and after an error, second subject.

3.2.2. Local hyperarticulation within utterances

How targeted is hyperarticulation? Is it aimed at repairing the most troublesome portion of an utterance? Since our experimental design elicited utterances that consisted of relatively long sentences (up to 26 words long, median length 11 words), we were able to examine the time course of hyperarticulation within an utterance. Recall that our error messages included both correct and incorrect words from the speaker's prior utterance (e.g., *You said "Kate*

[6] Sample utterances from these two speakers can be found at http://www.cs.sunysb.edu/~adaptation/hyperarticulation.html.

Tolstoy is bringing a tackle dog and hopscotch'' when the speaker had said "*Kate Tolstoy is bringing a talking doll and a hope chest*"). Each target word coded for segmental form was categorized as to whether it occurred before, during, or after the part indicated as misrecognized in the error message. We then analyzed the relative percentages of hyperarticulation (the percentage increase in relaxed vs. clear forms from the original utterance to its repair) across these three locations in the utterance.

Consistent with Oviatt et al.'s findings (1998a,b), we found that hyperarticulation has a locally targeted component; that is, during the repair speakers are more likely to modify the part of the utterance that was apparently misunderstood than the other parts. The percentage of clear forms increased 12.6% for the misunderstood portion of the repaired utterance over the "before" version, significantly greater than the increase of only 4.3% for words preceding the misunderstood portion of the repaired utterance ($F_1(1, 15) = 8.41$, $p = .01$; $F_2(1, 2) = 68.97$, $p = .01$) and marginally so for the increase of only 4.7% for words following the misunderstood portion ($F_1(1, 15) = 6.72$, $p = .02$; $F2(1, 2) = 2.65$, *ns*). This pattern emerged as a quadratic trend, $F_1(1, 15) = 8.48$, $p = .01$; $F2(1, 2) = 6.52$, $p = .125$,[7] and is illustrated in Fig. 3.[8]

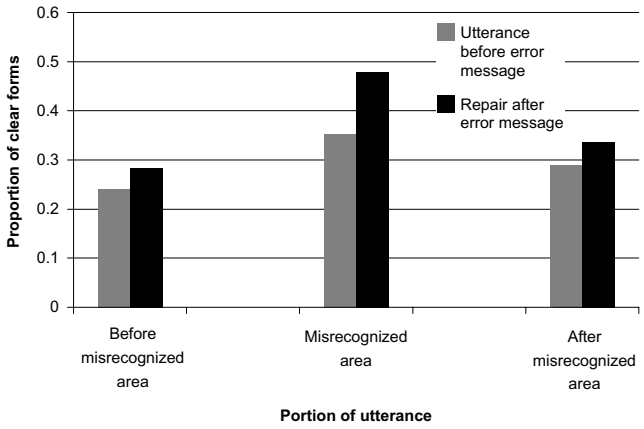

Fig. 3. Local hyperarticulation. Proportion of phonetically clear speech forms before, during, and after the apparently misunderstood portion of utterances produced before error messages and afterward (as repairs).

3.3. Global effects of misrecognition upon speaking

3.3.1. Decay of hyperarticulation

As a style of speaking, does hyperarticulation persist, or end abruptly once there is evidence that an utterance has been understood (like a switch), or decay gradually (like a dial)? Since the error messages in our experiment were staged in advance and occurred at the same points in the dialog for all speakers (as opposed to occurring randomly as in Oviatt et al., 1998b and Soltau and Waibel, 1998), we were able to systematically analyze the effect of error messages upon hyperarticulation *across* utterances. For each utterance in our corpus, we calculated how recently the speaker had experienced an error message (ranging from the immediately preceding turn in the case of a repair, to a maximum of 16 turns previously). This analysis included not only planned error messages, but also unplanned ones. We found that the closer (in turns) an utterance was to the most recent previous error message, the more it was hyperarticulated – that is, the slower its rate of speech, $r_Z = .224$, $p < .001$ and the more likely any target words were to contain clear segmental forms, $r_Z = -.147$, $p < .005$. This demonstrates that hyperarticulation is not only a focal phenomenon, but also global, in that it decays gradually after a repair. Fig. 4 illustrates that speakers returned to relaxed speech gradually after experiencing evidence of misrecognition (note: utterances one utterance away from an error message are repairs).

3.3.2. Effect of error rates

Finally, we looked at the global effect of error rates on hyperarticulation. Oviatt et al. (1998b) found significant slowing of speaking rate (by about 49 ms/syllable) when speakers experienced more (20%) vs. fewer (6.5%) errors. Recall that the design of our study split the dialog sessions into three parts, each with distinct rates of staged error messages, as follows: of the 66 total utterances, the first 24 were associated with only two planned errors (an 8.3% error rate), utterances #25–48 were associated with eight planned errors (a 33% error rate), and utterances #49–66 were associated with nine planned errors (a 52.9% error rate that included a highly salient run of five errors in a row). As expected, mean speaking rate slowed as a function of higher error rate; speakers produced 4.53 syllables/s. during the part of the dialog with an 8.3% error rate vs. 4.08 during the part with a 33.3% error rate, $F(1, 14) = 13.35$, $p = .003$. This amounted to a slowing of about 24 ms/syllable. There was no further slowing from the middle part to the last part of the dialog, during which the error rate was extremely high, $F(1, 14) = .06$, n.s.

3.4. Individual differences

Individual speakers displayed substantial variability in average speaking rate, ranging from 2.43 to 5.27 syllables/s. All speakers slowed their speaking rate during repairs, relative to matched utterances before repairs; the extent to which they did so ranged from .04 to 1.33 syllables/s. Variability in individual rates of speaking may have increased due to a few speakers adopting a globally hyperarticulate style of speaking throughout the experiment; those who experienced the highest error rates (due to

[7] The second ANOVA (by-items) was marginal; power was limited because only three utterances contributed to the analysis (i.e., contained target words that occurred before, during, and after the trouble area).

[8] This finding also suggests a possible fourth explanation for the difference in clear speech between content and function words; in the sample of utterances that resulted from our design, content words were more likely to be indicated as misrecognized in the error messages and function words were more likely to precede or follow them.

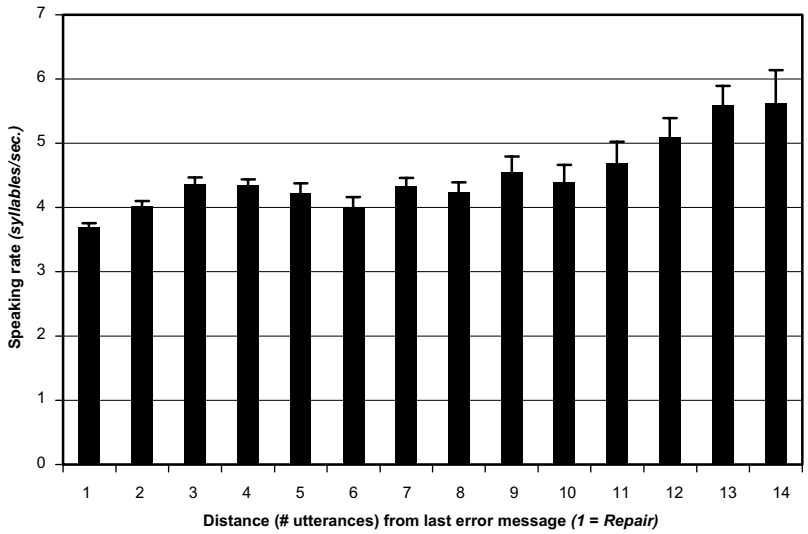

Fig. 4. Global hyperarticulation. Speaking rates during and after repairs, as a function of number of utterances since the last error message (the utterances that are 1 utterance away from an error message are repairs).

receiving an abundance of unplanned error messages in response to producing pronouns, ellipsis, disfluencies, etc.) spoke the most slowly during utterances that were not repairs (marginal at $r = -.443$, $p < .09$).

All but three speakers produced more clear speech during repairs than before repairs. Speaking rate and phonological form were correlated not only within utterances by individual speakers (Section 3.1) but also across speakers; that is, speakers who spoke characteristically fast (averaged over the whole dialog) tended to produce more relaxed forms and those who spoke characteristically slowly tended to produce more clear forms, $r = -.760$, $p < .001$.

Recall that our speakers comprised 10 monolingual native speakers of American English and six bilinguals. Both monolingual and bilingual speakers slowed their speaking rate equally during repairs $F_1(1,15) = .313$, n.s., and there was no difference in average speaking rates of monolinguals vs. bilinguals, $F1(1,14) = 357$, n.s. However, during repairs, monolinguals increased their proportion of segmentally clear forms marginally more than did bilinguals, $F_1(1, 14) = 4.32$, $p = .057$. It has been hypothesized that speakers represent phonetic targets in a hyperarticulate form (see Johnson et al.'s, 1993 proposal and Whalen et al., 2004 for the controversy surrounding this proposal). If this is true, it may be that monolingual native speakers represent phonetic targets as more hyperarticulate prototypes than do bilingual speakers of English, or else monolinguals may manage to approach these targets more closely during hyperarticulation.

3.5. Speech recognition performance

The corpus of utterances was processed through the Sphinx 3 speech recognizer, with two acoustic models, trained on (a) broadcast speech (HUB4) and (b) conversational speech (CMU Communicator), each configured with four different language models (word list, unigram, bigram, and trigram). We computed by-subjects and by-items ANOVAs with Training × Language Model as factors, both for overall word error rate (WER) and for word error rate with a before and after error message factor. Mean word error rates are reported in Table 6.

There was a strong and consistent pattern, with higher word error rates by every recognizer for "before" utterances than for "after" (repair) utterances, $F_1(1, 15) = 12.81$, $p = .003$; $F_2(1, 47) = 14.89$, $p < .001$. This pattern flies in the face of the expectation that hyperarticulate speech always harms automatic recognition performance; for our corpus, the repairs (with their slower speaking rates and clearer pronunciations) were associated with *improved* performance for all the Sphinx-based recognizers. In contrast, Wade et al. (1992) found in their study that hyperarticulation did adversely affect the performance of a speech recognizer whose model was trained on in-domain conversational speech.

The acoustic model training data made a difference as well. Over the entire corpus of utterances, recognizers trained on HUB4 performed better than those trained on CMU Communicator, $F_1(1, 15) = 115.75$, $p < .001$; $F_2(1, 65) = 143.40$, $p < .001$. The difference in word error rates for HUB4- vs. Communicator-trained recognizers was 19.18 for Word list models, 21.2 for unigram, 13.87 for bigram, but only .12 for trigram, yielding a Training Set × Language Model interaction, $F_1(1, 15) = 196.36$, $p < .001$; $F_2(1, 65) = 25.09$, $p < .001$. It is not so surprising that the recognizer trained on HUB4 out-performed the one trained on CMU Communicator, as the latter corpus was smaller. Moreover, the HUB4 corpus is a corpus of broadcast speech, which typically contains an overabundance of pitch accents and is actually sometimes segmentally similar to hyperarticulate speech.

The pattern of ASR mean errors in Table 6 (deletions, insertions, and substitutions) shows that before and after

Table 6
Recognition performance (word error rate) by Sphinx-based speech recognizers trained on broadcast speech (HUB4) and conversational speech (CMU Communicator) for all utterances, "before" utterances only, and "after" utterances only

	HUB4				Comm			
	Word list	Unigram	Bigram	Trigram	Word list	Unigram	Bigram	Trigram
All utterances (N = 1405)								
WER	0.5745	0.3558	0.2527	0.0625	0.7663	0.567	0.3914	0.0637
Deleted	2.95	1.42	0.62	0.13	4.65	3.16	0.81	0.14
Inserted	0.06	0.26	0.48	0.2	0.01	0.08	0.55	0.16
Substituted	3.37	2.12	1.46	0.26	3.93	3.14	2.79	0.36
"Before" utterances (matched to repairs; N = 387)								
WER	0.5702	0.3591	0.2727	0.0929	0.7716	0.5533	0.4214	0.1009
Deleted	3.03	1.49	0.7	0.19	4.82	3.17	0.92	0.18
Inserted	0.05	0.23	0.47	0.21	0.01	0.1	0.5	0.24
Substituted	3.49	2.18	1.57	0.38	4.15	3.15	3.03	0.52
"After" Utterances (repairs; N = 387)								
WER	0.5483	0.3316	0.2448	0.0739	0.745	0.5195	0.3723	0.0775
Deleted	2.9	1.32	0.66	0.17	4.6	3.01	0.84	0.16
Inserted	0.05	0.24	0.52	0.25	0.02	0.12	0.6	0.18
Substituted	3.64	2.15	1.31	0.25	4.31	3.15	2.64	0.38

error differences in recognition rates were due to primarily to fewer words being deleted from repair utterances than from "before" utterances. This is a consistent pattern across all of the recognizers. For all the Sphinx-based recognizers, deletions and substitutions were more common than insertions.

For the grammar-based recognizer, the mean word error rate was 11.17%, with no difference in word error rate for utterances before and after error messages, $t_1(15) = .72$, *ns*; $t_2(16) = .53$, *ns*. However, better speech recognition performance (proportion of words correctly recognized) was correlated with slower speech, $r_Z = -.179$. And better recognition was weakly correlated with higher proportions of clear speech, $r_Z = .092$ (for individual speakers, this ranged from $r = -.316$ to $r = +.615$). So some features of hyperarticulation appear to help recognition performance for this grammar-based recognizer as well. The only other studies of hyperarticulation to use a grammar-based speech recognizer were Hirschberg et al. (2004) and Litman et al. (2006); those studies did not find a significant correlation between speaking rate and speech recognizer performance. However, those data may not be easy to compare to ours, as they involved telephone speech recorded from actual human–computer dialog (rather than a Wizard-of-Oz simulation), and so included more cascading errors (and neither study examined any effects of segmentally clear speech).

4. Conclusions

In this paper, we described the results of an experiment designed to investigate the impact of different components of hyperarticulation in computer-directed speech. Our design enabled us to collect a corpus of spontaneous utterances (a) of varied length, (b) with comparable utterances from multiple speakers, (c) some of which were repetitions in response to planned error messages, yielding lexically identical pairs for within utterance comparison, (d) with locations of errors parameterized for both within target utterances and across a span of utterances, and (e) with multiple tokens of target words to support comparisons of segmental adaptation. With these characteristics combined in a single design, our findings extend those of previous hyperarticulation studies.

Our findings support two major conclusions. First, hyperarticulation is a somewhat targeted adjustment involving both slowed speaking rates and clearer phonetic segments; this adjustment varies both locally (within an repair utterance) and globally (decaying across the span of a dialog). Second, hyperarticulation in speaking to computers is not as maladaptive as previously thought.

We report these new results:

- Hyperarticulation is a 'dial', rather than a 'switch':
  - Speakers return to their pre-error speaking style by 4–7 utterances after evidence of misrecognition.
  - Speakers hyperarticulate more in high-error parts of a dialog than in low-error parts, as well as in high-error parts of an utterance.
- After misrecognition, content words are more likely to be pronounced in their clear forms than function words.
- Monolingual and bilingual speakers alike adjust their speaking rate after evidence of misrecognition. However, monolingual speakers produce marginally more clear forms during repairs than do bilingual speakers.

We also replicate several previous results:

- Speakers speak more slowly and produce more segmentally clear forms after evidence of misrecognition (Levow, 1999; Oviatt et al., 1998a,b; Shriberg et al., 1992; Wade et al., 1992).

- In repairs, speakers are more likely to clearly pronounce the words that appeared to have been misrecognized than other words before or after these trouble spots in the same utterance (Oviatt et al., 1998b).
- There is considerable variation in individual speaking style to a (simulated) speech recognizer (Shriberg et al., 1992; Hirschberg et al., 2004). A minority of speakers produce segmentally clear speech in all interaction with a computer system; the majority, however, use relaxed forms in the absence of evidence of misrecognition. Speaking rate, a prosodic feature, is a more universal feature of hyperarticulation.

Although our corpus was not collected from interactions with an actual dialog system, we replicated significant aspects of typical dialog system interaction (the nature of system feedback, the distribution of misrecognition errors, the need for the user to correct errors by repeating the entire original utterance) while ensuring that our corpus was collected under controlled conditions so that we were able to perform a quantitative analysis of changes in hyperarticulation over successive turns. In an actual interaction with a dialog system, the system behavior might vary. However, the behavior we studied in this experiment was that of the human user in conversation with a computer. What we have learned about the nature of hyperarticulation, then, is expected to apply in any in any situation where users face patterns of misrecognition errors.

Our results do, however, provide some insights for designers of dialog system behaviors. When users of a spoken dialog system experience misrecognition, they alter their behavior in two ways: they may start to hyperarticulate, and they may rephrase, sometimes to out-of-grammar utterances (Wade et al., 1992; Kirchhoff, 2001; Batliner et al., 2003; Choularton and Dale, 2004; Bohus and Rudnicky, 2005; Bulyko et al., 2005; Gieselmann, 2006; Litman et al., 2006). Our results, taken in light of the previous literature, suggest that approaches to preventing or handling maladaptive rephrasing (e.g., as suggested by Hockey et al., 2003; Litman et al., 2006) may have more impact on dialog outcomes than simply encouraging users to "speak naturally".

These results serve the broader goals of characterizing how speakers adapt to their addresses, and of handling variation in input to spoken dialog systems (http://www.cs.sunysb.edu/~adaptation/). An improved understanding of the nature and causes of variation in language use in human–human and human–computer dialog will, we predict, lead to more natural and more powerful interaction with spoken dialog systems.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0325188. We would like to thank Anny Cheng, Alex Christodoulou, Jonathan MacDonald, Yajaida Merced, Randy Stein, and Anthony Weaver for their work on this study. Please address correspondence to Amanda Stent at amanda.stent@gmail.com.

References

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A., 2001. Toward conversational human–computer interaction. AI Magazine 22 (4), 27–38.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., 2003. How to find trouble in communication. Speech Comm. 40, 117–143.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. J. Acoust. Soc. Amer. 113, 1001–1024.

Bennett, C., Rudnicky, A., 2002. The Carnegie Mellon communicator corpus. Proc. Internat. Conf. Spoken Language Process. International Speech Communication Association, Denver, Colorado, pp. 341–344.

Bohus, D., Rudnicky, A., 2005. Sorry, I didn't catch that!–An investigation of non-understanding errors and recovery strategies. Proc. SIGDial Workshop Discourse and Dialogue. Lisbon, Portugal, ACL, pp. 128–143.

Bradlow, A.R., Bent, T., 2002. The clear speech effect for non-native listeners. J. Acoust. Soc. Amer. 112, 272–284.

Bradlow, A.R., 2002. Confluent talker- and listener-oriented forces in clear speech production. In: Gussenhoven, C., Rietveld, T., Warner, N. (Eds.), Papers in Laboratory Phonology VII. Mouton de Gruyter, Berlin/New York, pp. 241–273.

Brennan, S.E., 1991. Conversation with and through computers. User Modeling and User-Adapted Interaction 1, 67–86.

Brennan, S.E., 1996. Lexical entrainment in spontaneous dialog. Proc. 1996 Internat. Sympos. Spoken Dialogue (ISSD-96). Acoustical Society of Japan, Philadelphia, PA, pp. 41–44.

Brennan, S.E., Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation. J. Exp. Psychol.: Learn. Memory Cognit. 6, 1482–1493.

Bulyko, I., Kirchhoff, K., Ostendorf, M., Goldberg, J., 2005. Error-correction detection and response generation in a spoken dialogue system. Speech Comm. 45, 271–288.

Choularton, S., Dale, R., 2004. User responses to speech recognition errors: consistency of behaviour across domains. Proc. 10th Austr. Internat. Conf. Speech Sci. Technol.. The Australian Speech Science and Technology Association, Sydney, Australia, pp. 457–462.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Academic Press, New York.

Core, M.G., Schubert, L.K., 1999. A model of speech repairs and other disruptions. Proc. AAAI Fall Sympos. Psychol. Models Commun. Collabor. Syst.. American Association for Artificial Intelligence, North Falmouth, MA, pp. 48–53.

Cutler, A., Butterfield, S., 1990. Durational cues to word boundaries in clear speech. Speech Comm. 9, 485–495.

Ferguson, C., 1975. Towards a characterization of English foreigner talk. Anthropol. Linguist. 17, 1–14.

Fernald, A., Simon, T., 1984. Expanded intonation contours in mothers' speech to newborns. Develop. Psychol. 20, 104–113.

Ferreira, F., Ferraro, V., Bailey, K.G.D., 2002. Good-enough representations in language comprehension. Curr. Direct. Psychol. Sci. 11, 11–15.

Gieselmann, P., 2006. Comparing error-handling strategies in human–human and human–robot dialogues. Proc. 8th Conf. Nat. Language Process. (KONVENS). Konstanz, Germany, pp. 24–31.

Gorin, A., Abella, A., Riccardi, G., Wright, J., 2002. Automated natural spoken dialog. Computer 35 (4), 51–56.

Harnsberger, J.D., Goshert, L.A., 2000. Reduced, citation, and hyperarticulated speech in the laboratory: an acoustic analysis. Research on Spoken Language Processing Report No. 24. Bloomington, IN, Speech Research Laboratory, Indiana University.

Hirschberg, J., Litman, D., Swerts, M., 1999. Prosodic cues to recognition errors. Proc. Internat. Workshop Automat. Speech Recognit. Understand. (ASRU'99). IEEE, Keystone, CO, pp. 349–352.

Hirschberg, J., Litman, D., Swerts, M., 2000. Generalizing prosodic prediction of speech recognition errors. In: Proc. Internat. Conf. Spoken Language Process. (ICSLP), Vol. 1. International Speech Communication Association, Beijing, China, pp. 254–257.

Hirschberg, J., Litman, D., Swerts, M., 2004. Prosodic and other cues to speech recognition failures. Speech Comm. 43, 155–175.

Hockey, B., Lemon, O., Campana, E., Hiatt, L., Aist, G., Hieronymus, J., et al., 2003. Targeted help for spoken dialogue systems: intelligent feedback improves naïve users' performance. Proc. Euro. Chapter Meet. Associat. Comput. Linguist.. Association for Computational Linguistics, Budapest, Hungary, pp. 147–154.

Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, Upper Saddle River, NJ.

Johnson, K., Flemming, E., Wright, R., 1993. The hyperspace effect: phonetic targets are hyperarticulated. Language 69, 505–528.

Kirchhoff, K., 2001. A comparison of classification techniques for the automatic detection of error corrections in human–computer dialogues. Proc. NAACL Workshop Adaptat. Dialogue Syst. Pittsburgh, PA, Association for Computational Linguistics.

Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R., Lewin, I., 2001. Comparing grammar-based and robust approaches to speech understanding: a case study. Proc. Eurospeech 2001. International Speech Communication Association, Aalborg, Denmark, pp. 1779–1782.

Kraljic, T., Brennan, S.E., 2005. Using prosody and optional words to disambiguate utterances: for the speaker or for the addressee? Cognit. Psychol. 50, 194–231.

Krause, J., Braida, L., 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. J. Acoust. Soc. Amer. 115, 362–378.

Levow, G.A., 1998. Characterizing and recognizing spoken corrections in human–computer dialogue. Proc. COLING-ACL 1998. Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 736–742.

Levow, G.A., 1999. Understanding recognition failurers in spoken corrections in human–computer dialogue. Proc. ESCA Workshop on Dialogue and Prosody, Eindhoven, Netherlands, ESCA.

Litman, D., Hirschberg, J., Swerts, M., 2006. Characterizing and predicting corrections in spoken dialogue systems. Comput. Linguist. 32 (3), 417–438.

Moon, S.-J., Lindblom, B., 1994. Interaction between duration, context, and speaking style in English stressed vowels. J. Acoust. Soc. Amer. 96, 40–55.

Oviatt, S.L., Levow, G., Moreton, E., MacEachern, M., 1998a. Modeling global and focal hyperarticulation during human–computer error resolution. J. Acoust. Soc. Amer. 104, 3080–3098.

Oviatt, S.L., MacEachern, M., Levow, G., 1998b. Predicting hyperarticulate speech during human–computer error resolution. Speech Comm. 24, 1–23.

Picheny, M.A., Durlach, N.I., Braida, L.D., 1985. Speaking clearly for the hard of hearing. I. Intelligibility differences between clear and conversational speech. J. Speech Hear. Res. 28, 96–103.

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. J. Speech Hear. Res. 29, 434–446.

Schmandt, C., Arons, B., 1984. A conversational telephone messaging system. IEEE Trans. Consum. Electron. 30, xxi–xxiv.

Schmandt, C.M., Hulteen, E.A., 1982. The intelligent voice-interactive interface. Proc. Conf. Human Factors Comput. Syst.. ACM Press, Gaithersburg, MD, pp. 363–366.

Shriberg, E., Wade, E., Price, P., 1992. Human–machine problem solving using spoken language systems (SLS): factors affecting performance and user satisfaction. Proc. Workshop Speech and Nat. Language at the Human Language Technol. Conf. Harriman, NY, Association for Computational Linguistics, pp. 49–54.

Sikveland, R.O., 2006. How do we speak to foreigners?—phonetic analyses of speech communication between L1 and L2 speakers of Norwegian. Proc. Fonetik 2006. Centre for Language and Literature, Lund University, Lund, Sweden, pp. 109–112.

Soltau, H., Waibel, A., 1998. On the influence of hyperarticulated speech on recognition performance. Proc. Internat. Conf. Spoken Language Process.. International Speech Communication Association, Beijing, China, pp. 229–232.

Soltau, H., Waibel, A., 2000a. Phone dependent modeling of hyperarticulated effects. In: Proc. Internat. Conf. Spoken Language Process. (ICSLP), Vol. 4. International Speech Communication Association, Beijing, China, pp. 105–108.

Soltau, H., Waibel, A., 2000b. Specialized acoustic models for hyperarticulated speech. In: Proc. IEEE Internat. Conf. Appl. Speech and Signal Process., Vol. 3. IEEE, Istanbul, Turkey, pp. 1779–1782.

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: acoustic and perceptual analysis. J. Acoust. Soc. Amer. 84, 917–928.

Wade, E., Shriberg, E.E., Price, P.J., 1992. User behaviors affecting speech recognition. Proc. 2nd Internat. Conf. Spoken Language Process.. International Speech Communication Association, Banff, Alberta, Canada, pp. 995–998.

Whalen, D.H., Magen, H.S., Pouplier, M., Kang, A.M., Iskarous, K., 2004. Vowel production and perception: hyperarticulation without a hyperspace effect. Language Speech 47, 155–174.