PREDICTING SPOKEN DISFLUENCIES DURING HUMAN-COMPUTER INTERACTION*

Sharon Oviatt

Department of Computer Science and Engineering
Oregon Graduate Institute of Science & Technology
P.O. Box 91000, Portland, Oregon 97291
oviatt@cse.ogi.edu

September 26, 1995

*This research was supported by Grant No. IRI-9213472 from the National Science Foundation, as well as grants, contracts, and equipment donations from Apple Computer, ATR International, AT&T, ETRI, Sun Microsystems, USWest, and Wacom Inc.

Abstract

This research characterizes the spontaneous spoken disfluencies typical of human-computer interaction, and presents a predictive model accounting for their occurrence. Data were collected during three empirical studies in which people spoke or wrote to a highly interactive simulated system as they completed service transactions. The studies involved within-subject factorial designs in which the input modality and presentation format were varied. Spoken disfluency rates during human-computer interaction were documented to be substantially lower than rates typically observed during comparable human-human speech. Two separate factors, both associated with increased planning demands, were statistically related to higher disfluency rates: (1) length of utterance, and (2) lack of structure in the presentation format. Regression techniques demonstrated that a linear model based simply on utterance length accounted for over 77% of the variability in spoken disfluencies. Therefore, design methods capable of guiding users' speech into briefer sentences have the potential to eliminate the majority of spoken disfluencies. In this research, for example, a structured presentation format successfully eliminated 60-70% of all disfluent speech. The long-term goal of this research is to provide empirical guidance for the design of robust spoken language technology.

1 INTRODUCTION

Speech disfluencies of all kinds, from blends to spoonerisms, have long been the subject of descriptive analysis by those interested in the structures and processes underlying linguistic performance (Fromkin, 1971; Hockett, 1967; Meringer & Mayer, 1895; Nooteboom, 1969). Historically, this literature has focused on observing and classifying different types of disfluencies, and then using this information as a basis for drawing inferences relevant to linguistic theory. Empirically-oriented analyses of spoken disfluencies also have been conducted in order to build and refine theories on topics such as speech production and neurolinguistic functioning (Caramazza & Hillis, 1991; Dell, 1986 & 1993; Fay & Cutler, 1977; Garrett, 1982; Goldman-Eisler, 1968; Levelt, 1983 & 1989; MacKay, 1971; Shattuck-Hufnagel & Klatt, 1979; van den Broecke & Goldstein, 1980). Some of this work, through hypothesis testing, modeling, and other analytical tools, has begun to test and compare the adequacy of alternative theoretical models, and to identify cognitive factors that precipitate spoken disfluencies. Although more rare, a few studies also have begun exploring comparisons between spoken disfluencies and those occurring in other modalities, such as handwriting and American Sign Language (Hotopf, 1983; also see Fromkin's 1980 volume).

From a computational viewpoint, researchers interested in spoken language processing recently have begun searching for reliable methods to detect and correct disfluent input automatically during interactions with spoken language systems (Hindle, 1983; Nakatani & Hirschberg, in press; O'Shaughnessy, 1992; Shriberg, Bear & Dowding, 1992). In general, this research has focused on identifying acoustic/prosodic cues for detecting self-repairs, either alone or in combination with syntactic, semantic, and pattern matching information. To date, however, no attempt has been made simply to reduce or eliminate disfluent input through proactive interface design. The development of interface techniques for successfully avoiding or managing spoken disfluencies could benefit from a clearer understanding of the underlying cognitive factors that cause them.

One underdeveloped cognitive theme in disfluency research is the relation between spoken disfluencies and planning demands. Although it is frequently claimed that speech disfluencies rise with increased planning demands of different kinds, the nature of this relation remains poorly understood. Generally speaking, a distinction has been made between "macroplanning" and "microplanning" during speech (Levelt, 1989). During macroplanning, the speaker engages in information retrieval and inference as he or she decides on the basics of what information to express and how to order it. During microplanning, the speaker shifts attention to finalizing the message to be expressed. As macroplanning becomes more effortful, it is assumed that speech disfluencies should increase.

In fact, this general principle has been confirmed in empirical research focusing on intersentential silence as an index of disfluency. That is, tasks involving more macroplanning or cognitive load, such as interpretation versus simple description of a cartoon (Goldman-Eisler, 1968), or description of an unfamiliar versus familiar route (Good & Butterworth, 1980), have been associated with greater sentential disfluency in the form of increased silent hesitations. However, it is not known to what extent macroplanning drives actual speech errors and self-corrections. Furthermore, the major factors contributing to macroplanning have yet to be identified and defined in any comprehensive manner, or linked to disfluency phenomena such as errors and self-repairs. From a design viewpoint, any information on the dynamics of what produces disfluencies, and how best to structure interfaces to minimize them, potentially could improve the robust performance of spoken language systems.

A related question is to what extent qualitatively different types of speech differ in their disflu-

ency rates and, for example, whether the disfluencies differ systematically between human-human and human-computer speech. In particular, if rates are widely variable, can interface design be used to accomplish substantial reductions? Although insufficient data currently exists to answer these questions, a few studies have suggested that disfluency rates may be lower during human-computer exchanges. For example, Levelt (1983) reported that 34% of people's utterances contained self-repairs when they spoke instructions describing a perceptual array into an audiotape machine. In contrast, only 6 to 10% of people's spontaneous utterances to a computer system contained self-repairs when they queried it about travel planning (Shriberg et al., 1992; O'Shaughnessy, 1992; see MADCOW, 1992, for a summary of ATIS domain corpus characteristics). However, to make a meaningful comparison of disfluency rates between human-human and human-computer speech requires: (1) a disfluency rate-per-word dependent measure (i.e., rather than the gross percentage of utterances containing disfluencies, without specifying average utterance length), (2) a comparable definition of disfluencies, (3) parallel transcription conventions, and so forth.

In the present research, past studies by the author and colleagues (Cohen, 1984; Oviatt & Cohen, 1991, 1992) were reanalyzed: (1) to yield data on the rate of disfluencies for four different types of human-human speech, and (2) to compare whether the disfluency rates in these human-human interactions differ systematically from human-computer ones. It was predicted that human-human disfluency rates would be elevated to a greater degree. Due to concern about the quality of previous speech disfluency data (Cutler, 1982), one aim of the present analyses was to achieve methodological uniformity in the application of definitions, transcription and coding conventions, dependent measures, task-oriented context of data collection, and so forth, so that comparison among different types of human-human and human-computer speech would be relatively precise and meaningful. For example, variation in mean length of utterance for different data sets was controlled by calculating a standard disfluency rate per 100 words.

In addition, three simulation studies of human-computer interaction were conducted, which generated data on spoken and handwritten disfluencies. Apart from comparing disfluencies in different communication modalities, two separate factors associated with planning demands were examined. First, presentation format was varied to investigate whether degree of structuring in the format might be associated with disfluencies. It was predicted that a relatively unconstrained format, which requires the speaker to self-structure and plan to a greater degree, would lead to a higher rate of speech disfluencies. Second, the rate of disfluencies was examined in sentences of varying length. Spoken utterances graduated in length were compared to determine whether longer sentences actually have an elevated rate of disfluencies per word, compared to shorter ones, or whether any increase in raw disfluencies simply is proportional to overall length. It was predicted that the disfluency rate would rise in longer utterances, since their production theoretically requires an increase in both macroplanning and microplanning. Finally, implications of this research are discussed for the design of future interfaces capable of reducing disfluent input, thereby enhancing the robustness of spoken language technology.

2 SIMULATION EXPERIMENTS ON HUMAN-COMPUTER INTERACTION

This section outlines three experiments on spoken and handwritten input to a simulated system. A comprehensive report on the linguistic characteristics of these two communication modalities is

provided elsewhere (Oviatt, Cohen & Wang, 1994). For the present research purpose, however, spoken disfluencies constitute the primary analytical focus.

2.1 Method

2.1.1 Subjects, Tasks, and Procedure-

Forty-four subjects participated in this research as paid volunteers. Participants represented a broad spectrum of white-collar professionals, excluding computer scientists, and all were native speakers of English. Participants also represented a broad age range (i.e., mid-20s to 60+ years), and were balanced on gender.

A "Service Transaction System" was simulated that could assist users with tasks that were either (1) verbal-temporal (e.g., conference registration or car rental exchanges, in which most of the content was proper names and scheduling information), or (2) computational-numeric (e.g., personal banking or scientific calculations, in which digits and symbol/sign content predominated). Tasks were selected to be realistic, common, and familiar to people, and sufficiently concrete to establish clear performance criteria. For example, during the verbal-temporal tasks, participants were asked to arrange a car rental for a colleague. They were provided with complete information about the person's needs, along with a copy of his or her business and credit card to confirm the transaction. During the computational-numeric tasks, people paid bills and completed other familiar banking tasks, with the support of actual hardcopy bills.

During the study, subjects first received a general orientation to the Service Transaction System, and then were given practice using it to complete tasks. They received instructions on how to enter information on the LCD tablet when writing, speaking, and combining both modalities. When writing, they were free to use cursive handwriting or printing, and were told to write information with the cordless electronic stylus directly onto highlighted areas on the LCD tablet. When speaking, subjects were instructed to tap and hold the stylus on active tablet areas as they spoke into a table-mounted Crown microphone. During free choice, people were completely free to use either modality in any way they wished. In all cases, they were encouraged to speak and write naturally, and to work at their own pace.

People also were instructed on completing tasks in two different presentation formats. During structured interactions, labeled fields were used to elicit task information (e.g., **Car pickup location** ☐). With this format, linguistic and graphical cues directly guided the content and order of people's input as they worked. In the unconstrained format, people took the initiative to ask questions or state needs that they expressed in an open workspace. In this case, the content and order of their input was self-structured, with no specific system prompting, although a transaction record at the bottom of their screen served as an indirect reminder of items still requiring completion. In both the structured and unconstrained formats, people continued providing information as the system responded interactively with confirmations. The system gradually filled out a transaction record that correctly reflected their requests. Figure 1 (top panel) illustrates one subject's spoken input during a structured interaction involving conference registration. The subject's transcribed speech is recorded in the right column, parallel to the prompt on the left that elicited it. The bottom panel illustrates another subject's spoken input during an unconstrained interaction to rent a car. The unconstrained speech also illustrates several typical disfluencies, including a content self-correction and filled pauses.

Other than specifying the input modality and format, an effort was made not to influence

the manner in which people expressed themselves. People's input was received by an informed assistant, who performed the role of interpreting and responding as a fully functional system would. Essentially, the assistant tracked the subject's written or spoken input, and clicked on predefined fields at a Sun SPARCstation to send confirmations back to the subject. The assistant's task was sufficiently automated that he or she was free to focus attention on monitoring the accuracy of incoming information, and on maintaining sufficient vigilance to respond promptly with correct confirmations.

A conversational model was adopted for providing both backchannel and propositional-level confirmations as people spoke or wrote to the system. These confirmations were designed to function similarly for all input modalities and presentation formats. At the backchannel level, subjects received asterisks (i.e., ***) immediately after their spoken input, and a residual electronic ink trace remained after handwritten input. People were told that this feedback meant that their input had been audible or legible, that it had been processed by the system, and that they should continue. In addition, the task-critical content of people's requests was confirmed by the system textually in their transaction receipt, as illustrated in Figure 1. This receipt remained visible throughout the interaction, and was completed as information was supplied. People were told to verify that their requests were being met successfully by checking the receipt contents during the interaction.

After the session, a post-experimental interview was conducted in which subjects were asked about their preferences to use the two presentation formats, as well as evaluative questions about other aspects of the system and its features. All subjects reported believing that the "system" was a fully functional one, indicating that the simulation was successfully credible for present research purposes. At the end, participants were debriefed about the nature of the simulation, as well as the rationale for using this kind of procedure. Finally, everyone was aware that their interactions with the system were being recorded for research purposes while they participated.

2.1.2 Semi-Automatic Simulation Technique-

In developing this simulation, an emphasis was placed on providing automated support for streamlining the simulation to the extent needed to create facile, subject-paced interactions with clear feedback, and to have comparable specifications for the different input modalities. In the present simulation environment, response delays averaged 0.4 second, with less than a 1-second delay in all conditions. This response speed was achieved in part by using scenarios for which correct solutions could be preloaded (i.e., for task-critical receipt information). Speed also was achieved by automating many functions, such as the delivery of randomly-distributed simulated errors, so that the assistant could focus his or her attention on their task. In addition, the techniques described earlier for providing conversational feedback contributed to the simulation's clarity and speed. In general, semi-automation contributed to a low rate of technical errors, as well as to the fast pace of the simulation. Technical details of the simulation method, capabilities, environment, and performance characteristics have been provided elsewhere (Oviatt et al., 1992, 1993).

2.1.3 Research Design and Data Capture-

Three studies were completed in which the research design was a completely crossed factorial with repeated measures. In all studies, the main factors of interest included: 1) communication

modality – speech-only, pen-only, combined pen/voice, and 2) presentation format – structured, unconstrained. The first two studies examined disfluencies during communication of verbal-temporal content. To test the generality of disfluency findings, a third study was conducted that compared disfluencies in computational-numeric content. In all studies, the order of presenting conditions and tasks was counterbalanced across subjects.

In total, data were available from 528 tasks for analysis of spoken and written disfluencies. All human-computer interactions were videotaped, with split-screen recordings that included an image of the subject as he or she worked at the tablet, a real-time record of all spoken and written input, and responses from the simulated system. From the videotape record, hardcopy transcripts also were created, with the subject's handwritten input captured automatically, and spoken input transcribed onto the printouts.

2.1.4 Transcript Preparation and Coding-

Each subject's speech was transcribed from the videotapes by a native speaker of English, and was second scored for reliability. Transcribed speech was recorded in the right-hand margin next to the tablet image that corresponded in time. All handwritten input was captured on-line, and embedded directly in the printed tablet image. The precise sequencing of all spoken and written input was preserved, as well as its relation to system feedback. For speech, attention was paid to transcribing verbatim input, without "cleaning it up" in any way. This included recording spoken language phenomena such as nonword sounds, repetitions, disfluencies and self-repairs, confirmations, the precise expression used to convey digits, and so forth.

Coding was conducted for the following dependent measures:

(1) Total Words– The total number of spoken and written words was tabulated for each condition and subject in all three studies. For speech, total words were counted as those spoken "on-line," or speech that a subject directed to the system while tapping the stylus to highlight the screen. Data on total number of words provided a baseline for converting disfluencies to a rate per 100 words.

(2) Mean Length of Utterance– The average number of words per utterance was tabulated for each condition and subject in studies 1 and 2. For both modalities, utterance boundary judgments were assisted by the location of sentential constituents and strong cues indicating subject disengagement, such as lifting the stylus off the tablet or gazing away from the tablet. For speech, pausing and sentence-final intonation provided additional cues to utterance boundaries. For writing, spatial displacement and pausing assisted in distinguishing boundaries. Mean length of utterance primarily was used for examining the relation between utterance length and disfluency rate.

(3) Semantic Integration per Utterance– The average number of task-critical items of information expressed in one utterance was summarized for each condition and subject in studies 1 and 2. Items defined as task-critical each were associated with a separate transaction receipt field. This index was used to compare the amount of semantic information that people integrated in a single utterance when interacting with different formats and modalities. Since deciding on what and how much to include in a given sentence is one major aspect of macroplanning described by Levelt (1989), this measure was used as an index of cognitive load during structured versus unconstrained spoken input.

(4) Order of Information Presentation– The predictability with which subjects ordered their information during the task also was summarized for each subject and condition during studies 1 and 2. During structured interactions, labelled fields specifically guided the order of information,

although people could vary their input order within a given page. The extent to which people conformed to this sequence of serial prompts was summarized as the percentage of form slots out of the total that they completed in the order presented. In the unconstrained format, people could present information in any order. However, the order of information in the receipt field, which matched that in the task description, was visible to refer to and model. The extent to which this indirect source of guidance actually predicted people's input order was summarized as the percentage of all task-critical information that was delivered in the order represented on reference materials. Deciding on how to order information within a discourse is a second major aspect of macroplanning identified by Levelt (1989), so this measure was used as a second and independent index of cognitive load during structured versus unconstrained spoken input.

(5) Disfluencies and Self-Corrections– Spontaneously occurring disfluencies and self-corrections were totaled for each subject and condition (i.e., communication modality x presentation format combination) in all 3 studies. The total number of disfluencies per condition then was converted to a rate per 100 words, and average disfluency rates were summarized as a function of condition and utterance length. These data were summarized separately for the verbal-temporal content in studies 1 and 2 and the computational-numeric content in study 3, so that the generality of any findings could be examined across content domains.

Disfluencies were classified into the following types: (1) content self-corrections— errors in task content that were spontaneously corrected as the subject spoke or wrote (e.g., "VISA... AMEX"; "139... 13*39*"), (2) false starts— alterations to the grammatical structure of an utterance that occurred spontaneously as the subject spoke or wrote (e.g., "I do not want to take uh have a ticket for the tour of the marine biology station"; "93 760 equals... 93 *plus* 760 equals"), (3) verbatim repetitions— retracings or repetitions of a digit, letter, phoneme, syllable, word, or phrase that occurred spontaneously as the subject spoke or wrote (e.g., "of the... of the"; "12...12"), (4) filled pauses— spontaneous nonlexical sounds that fill pauses in running speech (e.g., "uh," "um,"), often signaling the start of a new phrase or self-correction, and that have no analogue in writing, (5) self-corrected spellings and abbreviations— spontaneously corrected misspelled words or further specification of abbreviations, which occur in writing but have no analogue in speech (e.g., "Toyata... Toyota," "WDC... Washington D.C."). For the purpose of this study, no attempt was made to code relatively minor spoken mispronunciations involving phenomena such as elongated or slurred sounds, interjection or omission of individual vowel and consonant sounds, and so forth, which also are more difficult to identify reliably during scoring.

2.1.5 Reliability-

A second scorer independently coded 17% of the data for each reported dependent measure, with equal sampling from all conditions. Interrater reliability was calculated as the percentage of agreements out of the total number of codings per category. All dependent measures reported in this paper had reliabilities of 0.83 or above, and 90% of the measures had reliabilities above 0.88.

2.2 Results

2.2.1 Mean Length of Utterance-

In studies 1 and 2 involving verbal-temporal content, utterances ranged from 1 to 26 words, although the mean length of utterance (MLU) was relatively brief in all conditions — 2 to 5 words. In study

3 involving computational-numeric content, MLU again was brief in all conditions — 1.5 to 2 words. However, the range of utterance lengths narrowed to 1 to 12 words.

A repeated measures Anova on MLU in studies 1 and 2 confirmed a significant main effect of modality, $F = 10.46$ (1, 17), $p < .005$, and of format, $F = 19.54$ (1, 17), $p < .0001$, as well as a mode x format interaction, $F = 9.02$ (1, 17), $p < .008$. Whereas unconstrained writing generated approximately 70% longer utterances than writing to a form, unconstrained spoken sentences averaged 100% longer than speech to a form. In addition, unconstrained spoken utterances averaged approximately 50% longer than unconstrained writing.

2.2.2 Semantic Integration per Utterance-

During studies 1 and 2, people integrated more task-critical information into an utterance when they were unconstrained while speaking or writing. Using Wilcoxon Signed Ranks tests, the degree of semantic integration was revealed to be significantly greater in unconstrained spoken utterances than structured ones, $T+ = 120$ ($N = 15$), $p < .0001$, one-tailed, and also in unconstrained written utterances compared to structured ones, $T+ = 99$ ($N = 14$), $p < .001$, one-tailed. However, semantic integration did not differ due to modality per se.

Furthermore, when samples of unconstrained and structured utterances were matched on length (MLU = 11) and compared, a Sign test confirmed that unconstrained spoken utterances still demonstrated a significantly greater degree of semantic integration than did structured ones, $N = 12$, $x = 0$, $p < .001$, one-tailed. That is, independent of the difference in average sentence length between formats, speakers still integrated more information into unconstrained utterances than they did structured ones of equivalent length.

2.2.3 Order of Information Presentation-

It also was confirmed that people's order of presenting information departed significantly from available models when the format was unconstrained, but not when it was highly structured. During unconstrained speech, people's information deviated from the order that would have been predicted based on reference materials 29% of the time, although it never deviated when speaking in the structured format, Wilcoxon Signed ranks, $z = 3.64$ ($N = 17$), $p < .001$. During unconstrained writing, the order of their information departed from that predicted 30% of the time, but it only deviated 1% of the time when writing in the structured format, Wilcoxon Signed ranks, $z = 3.41$ ($N = 15$), $p < .001$. As expected, people did take more liberty in their ordering of task information when no explicit structure was provided.

2.2.4 Disfluencies and Self-Corrections-

The data yielded a corpus totaling over 19,000 words, including over 200 disfluencies for analysis. Figure 2 summarizes the percentage of all spoken and written disfluencies representing different disfluency categories when people communicated verbal-temporal content (left panel) and computational-numeric content (right panel). Figure 2 shows that written disfluencies during both kinds of task predominantly involved content self-corrections. However, spoken disfluencies during the verbal-temporal tasks mainly were filled pauses. This high percentage of spoken filled pauses dropped, however, during tasks in which people spoke digits — perhaps in part due to fewer lengthy

utterances in these tasks. Clearly, the relative distribution of different types of disfluency fluctuates with the modality and content of information being communicated.

The overall baseline rate of spontaneous disfluencies and self-corrections averaged 1.33 per 100 words in the verbal-temporal data, or a total of 1.51 disfluencies per task set. In the computational-numeric data, the overall baseline rate of disfluencies averaged a similar 1.32 per 100 words[1] For the verbal-temporal data, the rate per condition ranged from an average of 0.78 per 100 words when speaking to a form, 1.17 when writing to a form, 1.61 during unconstrained writing, and a high of 1.74 during unconstrained speech. This pattern replicated for the computational-numeric content, with the rate per condition averaging 0.87 when speaking to a form, 1.10 when writing to a form, 1.42 during unconstrained writing, and a high of 1.87 during unconstrained speech. Figure 3 illustrates these disfluency rates, and highlights the increasing rate of spoken disfluencies in the unconstrained format compared with the structured one for both verbal content (left panel) and numeric content (right panel).

Wilcoxon Signed Ranks tests revealed no significant modality difference in the overall rate of disfluent input during either the verbal-temporal or computational-numeric tasks. The verbal-temporal tasks averaged 1.26 disfluencies per 100 words for speech and 1.39 for writing, $T+ = 75$ ($N = 17$), $z < 1$, and the computational-numeric tasks averaged 1.26 disfluencies for writing and 1.37 for speech, $z < 1$. However, the average rate of spoken disfluencies was significantly elevated in the unconstrained format for both types of task. Although analyses revealed no significant difference in the written disfluency rates for either kind of task as a function of format, $T+ = 64.5$ ($N = 14$), $p > .20$ for verbal tasks, and $T+ = 36.5$ ($N = 11$), $p > .35$ for numeric tasks, spoken disfluency rates for both types of task clearly increased significantly in the unconstrained format compared to the structured one, $T+ = 88$ ($N = 14$), $p < .015$, one-tailed for verbal tasks, and $T+ = 77$ ($N = 13$), $p < .015$, one-tailed for numeric tasks. Finally, analyses of multimodal pen/voice input replicated this significant elevation of spoken disfluencies in the unconstrained format, $T+ = 87$ ($N = 14$), $p < .015$, one-tailed.

Since results presented in section 2.2.2 indicated that unconstrained utterances averaged significantly longer than structured ones, disfluency rates were examined further for specific utterances graduated in length between 1 and 18 words.[2] First, these analyses indicated that the average rate of disfluencies per 100 words increased as a function of utterance length for spoken disfluencies, although not for written ones. When the rate of spoken disfluencies was compared for short (1-6 words), medium (7-12 words), and long utterances (13-18 words), it increased from 0.66, to 2.14, to 3.80 disfluencies per 100 words, respectively. Statistical comparisons confirmed that these rates represented significant increases between short and medium sentences, $t = 3.09$ ($df = 10$), $p < .006$, one-tailed, and also between medium and long ones, $t = 2.06$ ($df = 8$), $p < .04$, one-tailed.

A regression analysis indicated that the strength of predictive association between utterance length and disfluency rate was $\rho^2_{XY} = .77$ ($N = 16$). That is, 77% of the variance in the rate of spoken disfluencies was predictable simply by knowing an utterance's specific length. The following simple linear model, illustrated in the scatterplot in Figure 4, summarizes this relation: $Y_{ij} = \mu Y + \beta_{Y*X}(X_j - \mu X) + e_{ij}$, with a Y-axis constant coefficient of -0.32, and an X-axis beta coefficient

[1] Calculation of spoken disfluency rates for the verbal and numeric content included a correction for the average number of syllables articulated per word.

[2] Data from studies 1 and 2 were used to examine the relation between utterance length and disfluency rates. Since the range of utterance lengths observed in the computational-numeric tasks was relatively constricted (i.e., 1 to 12 words), these data did not permit a parallel analysis.

representing utterance length of +0.26. These data indicate that the demands associated with planning and generating longer constructions lead to substantial elevations in the rate of disfluent speech.

To assess whether presentation format had an additional influence on spoken disfluency rates beyond that of utterance length, comparisons were made of disfluency rates occurring in unconstrained and structured utterances matched for length. These analyses revealed that the rate of spoken disfluencies also was significantly higher in the unconstrained format than during structured speech, even with utterance length controlled, t (paired) = 2.42 (df = 5), $p < .03$, one-tailed. That is, independent of utterance length, lack of structure in the presentation format was associated with elevated disfluency rates.

Only 5% of all spoken disfluencies in the present studies were marked with a lexical edit signal, such as "oops," "no," "correction," or "I'm repeating." All of these marked disfluencies except one involved content self-corrections occurring during unconstrained speech. The click-to-speak implementation effectively reduced the rate of lexical signaling, since cases were observed in which speakers lifted their pen off the tablet following a disfluency and then spoke an edit signal off-line before re-engaging the system to speak a correction. These data suggest that lexical edit signals are neither sufficiently frequent, nor reliable enough to aid in automatically detecting disfluency locations.

From a pragmatic viewpoint, it also is informative to compare the total number of disfluencies that would require processing during an application. Different design alternatives can be compared with respect to effective reduction of total disfluencies, which would result in less need for processing of repairs. In studies 1 and 2, a comparison of the total number of spoken disfluencies revealed that people averaged 3.33 per task set when using the unconstrained format, which was reduced to an average of 1.00 per task set when speaking to a form. That is, 70% of all disfluencies were eliminated simply by using a more structured format. Likewise, in study 3, the average number of disfluencies per subject per task set dropped from 1.75 in the unconstrained format to 0.72 in the highly structured one. Again, the more structured presentation format successfully eliminated most of people's spoken disfluencies, or 59% — compared with the same people completing the same tasks when unconstrained.

2.2.5 Self-reported Format Preference

During post-experimental interviews, people reported their preference to interact using the two different presentation formats. Results from studies 1 and 2 indicated that 67% of subjects reported a preference for the structured format rather than the unconstrained one. This preference for the structured format replicated in study 3, with 69% of participants reporting a preference for more structure and guidance.

3 EXPERIMENTS ON HUMAN-HUMAN SPEECH

This section reports on data that were analyzed to explore the degree of variability in disfluency rates among qualitatively different kinds of spoken interaction. In particular, the following data compare whether disfluency rates for human-human and human-computer spoken interactions differ systematically.

3.1 Method

Data originally collected by the author and colleagues during two previous studies were reanalyzed to provide comparative information on human-human disfluency rates for the present research (see Cohen, 1984; Oviatt & Cohen, 1991, 1992).

One of these previous studies focused on telephone speech, providing data on both: (1) prototypical two-person telephone conversations, and (2) three-person interpreted telephone conversations, with a professional telephone interpreter intermediating. Methodological details of this study are provided elsewhere (Oviatt & Cohen, 1992), and only are summarized briefly here. In this study, within-subject data were collected from 12 native speakers during each of the two conditions. Telephone conversations were task-oriented dialogues concerning conference registration and travel arrangements, in which the task content and structure were very similar to the simulation studies just reported. The transcription conventions for recording telephone speech, which were developed by the author, also were comparable to those used in the simulation studies. Finally, the two types of telephone speech were analyzed for the same categories of disfluencies as those coded during the simulation studies (see section 2.1.4). The rate of spoken disfluencies per 100 words was calculated for both types of telephone speech, producing a comparable measure to that used in the simulation studies. In short, an effort was made to achieve methodological uniformity to facilitate precise and meaningful comparisons between the human-computer and human-human data.

In another study, for which methodological details are outlined elsewhere (Cohen, 1984; Oviatt & Cohen, 1991), speech data were collected during task-oriented dialogues conducted in five different communication modalities. For the present comparison, data from two of these modalities were reanalyzed: (1) two-party face-to-face dialogues, and (2) single-party monologues into an audiotape machine. A between-subject design was used, in which 10 subjects provided spoken instructions on how to assemble a water pump. Since the original study's transcription conventions differed from those used in the simulation studies, speech from the original audiotapes was retranscribed. These transcriptions then were coded for the same categories of disfluencies outlined in section 2.1.4. Finally, these data also were converted to a disfluency rate per 100 words.

3.2 Comparative Results

Table 1 summarizes the average speech disfluency rates for the four types of human-human and two types of human-computer interaction that were studied. Disfluency rates for each of the two types of human-computer speech are listed in Table 1 for verbal-temporal and computational-numeric content, respectively, and are corrected for average number of syllables per word. All samples of human-human speech had a substantially higher disfluency rate than the human-computer samples, and these two general categories were confirmed to be significantly different, $t = 5.59$ ($df = 38$), $p < .0001$, one-tailed. Comparison of the average disfluency rate for human-computer speech with that for monologues, the least discrepant of the human-human categories, also confirmed a difference, $t = 2.65$ ($df = 21$), $p < .008$, one-tailed. The magnitude of this disparity ranged from 2-to-11 times higher disfluency rates for human-human as opposed to human-computer speech, depending on the categories compared.

Further analyses uncovered the fact that average disfluency rates were significantly higher for telephone speech than other categories of human-human speech, $t = 2.12$ ($df = 20$), $p < .05$, two-tailed. In addition, the disfluency rate was significantly higher when a person conversed in a typical two-party telephone call than when the same person participated in a three-person interpreted call,

Type of Spoken Interaction	**Disfluency Rate**
Human-human speech:	
Two-person telephone call	**8.83**
Three-person interpreted telephone call	**6.25**
Two-person face-to-face dialogue	**5.50**
One-person noninteractive monologue	**3.60**
Human-computer speech:	
Unconstrained computer interaction	**1.74 / 1.87**
Structured computer interaction	**0.78 / 0.87**

Table 1: Spoken disfluency rates per 100 words for different types of human-human and simulated human-computer interaction. Human-computer disfluency rates on the left represent verbal-temporal content, and the right computational-numeric content.

t (paired) = 3.22 (df = 11), p < .008, two-tailed. Reduced disfluencies had not been anticipated during interpreted calls, although it is known that speakers adopt a particularly conservative, cautious linguistic style during such exchanges (Oviatt & Cohen, 1992).

4 DISCUSSION

Spoken disfluencies are strikingly sensitive to the increased planning demands of generating progressively longer utterances. Of all the variance in spoken disfluencies in the present data, 77% was predictable simply by knowing an utterance's specific length. A linear model was provided, $Y = -0.32 + 0.26X$, to summarize the predicted rate of spoken disfluencies (Y) as a function of utterance length (X). Knowledge of utterance length alone, therefore, is a powerful predictor of speech disfluencies in human-computer interaction. Future work should investigate other contributing factors, as well as incorporating them into a more comprehensive and refined predictive model.

Spoken disfluencies also are influenced substantially by the presentation format used during human-computer interaction. An unconstrained format, which required the speaker to self-structure and plan to a greater degree, led speakers to produce over twice the rate of disfluencies as a highly structured interaction. Furthermore, this format effect was replicated when spoken input was multimodal as well as unimodal, and across qualitatively different types of spoken content. Since this difference between formats occurred in samples matched for length, it is clear that presentation format and utterance length each exert an independent influence on disfluency levels.

Two different sources of planning demand may have accounted for increased disfluencies in the unconstrained format. First, analyses indicated that speakers integrated more task information into individual sentences in this format. Secondly, when the format was unconstrained, speakers imposed a greater degree of self-structuring with respect to the order of information they expressed. These organizational features of unconstrained speech are indices of the main dimensions of macroplanning

that Levelt (1989) articulated— deciding what information to include in a sentence, and how to order it. The present results provide evidence that both types of macroplanning did indeed play a more active role during unconstrained interactions, perhaps contributing to its higher disfluency rate.

In these three simulation studies, a substantial 60 to 70% of all spoken disfluencies were eliminated simply by using a more structured format. That is, selection of presentation format was remarkably effective at guiding a speaker's language to be less disfluent. This apparently was accomplished in part by reducing sentential planning demands during use of the structured format — by reducing the need for people to plan the content and order of information delivered. However, it also was attributable in part to the relative brevity of people's sentences in the structured format. More specifically, the percentage of moderate to long sentences (i.e., 7 words or longer) increased from 5% during structured interactions to 20% during unconstrained speech — a 4-fold or 300% increase. In addition, the average disfluency rate increased from 0.66 for short sentences to 2.81 for moderate or lengthy ones — a 326% increase. In brief, these cumulative effects influenced the reduced rate of disfluencies observed in the structured format, which also was the preferred format for completing these tasks by a factor of 2-to-1. Further research is needed to investigate specific techniques for structuring information that can effectively reduce disfluencies in spontaneous language, including textual, graphical, auditory, and multimodal.

Wide variability can be expected in the disfluency rates typical of qualitatively different types of spoken language. Based on the six categories compared here, rates were found to vary by a magnitude of 2-to-11-fold, with the highest rates occurring in telephone speech and the lowest in human-computer interaction. This variability suggests that further categories of spoken language should be studied individually to evaluate how prone they may be to disfluencies, rather than assuming that the phenomenon is stable throughout spoken language. In particular, given the high rate of disfluencies observed in human-human telephone speech, human-computer telephone speech likewise may be relatively elevated among the class of human-computer interactions. If so, then efforts to effectively minimize disfluencies in human-computer telephone speech may have special payoff, and the development of tailored interface techniques for the next generation of screen-based multimodal phones may be a fertile avenue for accomplishing this goal. Future work exploring disfluency patterns during more complex multimodal exchanges will be of special interest.

With respect to methodological considerations, this research underscores the importance of converting disfluencies to a rate, so that fluctuations in observed disfluencies can be compared in a direct and meaningful manner. Statistical comparison of mean length of utterance in these data revealed considerable variability among different modalities and presentation formats. Given that utterance length is strongly associated with spoken disfluencies, these data emphasize the importance of reporting and controlling for MLU in future disfluency research.

In comparison with speech, analyses of handwritten disfluencies have been rare, informal, and limited to evaluating conventional noninteractive handwriting on paper (Hotopf, 1983). This research provides an initial comparison of spoken disfluencies with written ones occurring during highly interactive human-computer exchanges, and in a manner designed to facilitate comparison between these modalities. Results confirmed that disfluencies in these two modes differ in the categories represented, their frequency distributions, and predictive factors — all of which support the view that modality-specific differences likely exist in their underlying production mechanisms (Caramazza & Hillis, 1991; Hotopf, 1983). Nonetheless, no overall difference was found in the baseline disfluency rates between interactive speech and writing. This latter finding contrasts with

Hotopf's general claim that written disfluencies are more common than spoken ones, based on his informal observations of noninteractive handwriting and speech samples. Although the present data indicate that written disfluencies are not influenced by variations in sentence length and format in the way that spoken ones are, further research nonetheless should probe whether such effects might emerge in written tasks involving longer sentences. In addition, future exploratory work needs to identify the most influential factors that generate written disfluencies, so that predictive modeling of this modality can be developed for the benefit of handwriting recognition and pen-based technology.

Finally, future work needs to explore other major human-computer interface features that may serve to decrease planning load on users, and to estimate the magnitude of their impact on reducing disfluencies. Such information would permit proactive system design aimed at supporting more robust spoken language processing. For future applications in which an unconstrained format is preferred, or disfluencies and self-repairs otherwise are unavoidable, methods for correctly detecting and processing the ones that occur also will be required. To the extent that promising work on this topic can incorporate probabilistic information on the relative likelihood of a disfluency for a particular utterance (e.g., of length N), based on the outlined predictive model or future refinements of it, correct detection and judicious repair of actual disfluencies may become feasible.

5 ACKNOWLEDGMENTS

Sincere thanks to the generous people who volunteered to participate in this research as subjects. Thanks also to Michael Frank and Martin Fong for programming the simulation environment, to Martin Fong and Dan Wilk for playing the role of the simulation assistant during testing, to Jeremy Gaston and Zak Zaidman for careful preparation of transcripts, and to Jeremy Gaston, Zak Zaidman, and Michelle Wang for assistance with data analysis. Finally, special thanks to Phil Cohen and Gary Dell for helpful discussions and manuscript comments.

References

[1] A. Caramazza and A. E. Hillis. (1991). Lexical organization of nouns and verbs in the brain. *Nature, 349*, 788–790.

[2] P. R. Cohen. (1984). The pragmatics of referring and the modality of communication. *Computational Linguistics, 10(2)*, 97–146.

[3] A. Cutler. (1982). The reliability of speech error data. In *Slips of the Tongue and Language Production.* (A. Cutler, ed.), Mouton: Berlin, Germany, 7–28.

[4] G. S. Dell. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

[5] G. S. Dell, C. Juliano, and A. Govindjee. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149—195.

[6] D. Fay and A. Cutler. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry, 8*, 505–520.

[7] V. A. Fromkin. (1971). The nonanomalous nature of anomalous utterances. *Language, 47(1)*, 27–52.

[8] V. A. Fromkin (ed.). (1980). *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand.* Academic Press: New York.

[9] M. F. Garrett. (1982). Production of speech: Observations from normal and pathological language use. In *Normality and Pathology in Cognitive Functions.* (A. W. Ellis, ed.), Academic Press: London, U. K.

[10] F. Goldman-Eisler. (1968). *Psycholinguistics: Experiments in Spontaneous Speech.* Academic Press: New York.

[11] D. A. Good and B. Butterworth. (1980). Hesitancy as a conversational resource: Some methodological implications. In *Temporal Variables in Speech.* (H. W. Dechert and M. Raupach, eds.), Mouton: The Hague.

[12] D. Hindle. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, ACL Press: Cambridge, Mass., 123–128.

[13] C. F. Hockett. (1967). Where the tongue slips, there slip I. In *To Honor Roman Jakobson.* Mouton: The Hague.

[14] W. H. N. Hotopf. (1983). Lexical slips of the pen and tongue: What they tell us about language production. In *Language Production, Volume 2: Development, Writing and Other Language Processes.* (B. Butterworth, ed.), Academic Press: London, U. K., ch. 4, 147–199.

[15] W. J. M. Levelt. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

[16] W. J. M. Levelt. (1989). *Speaking: From Intention to Articulation.* ACL/M.I.T. Press: Cambridge, Mass.

[17] D. G. MacKay. (1971). Stress pre-entry in motor systems. *American Journal of Psychology, 84(1)*, 35–51.

[18] MADCOW Working Group. (1992). Multi-site data collection for a spoken language corpus. In *DARPA Proceedings of the Speech and Natural Language Workshop*, Morgan Kaufmann: San Mateo, California, 7–14.

[19] R. Meringer and K. Mayer. (1895). *Versprechen und Verlesen.* Goschensche Verlag: Stuttgart, Germany. (Re-issued with introductory essay by A. Cutler and D. A. Fay (1978), John Benjamins: Amsterdam.)

[20] C. Nakatani and J. Hirschberg. (in press). A corpus-based study of repair cues in spontaneous speech. In *Journal of the Acoustical Society of America.*

[21] S. G. Nooteboom. (1969). The tongue slips into patterns. In *Leyden Studies in Linguistics and Phonetics.* Mouton: The Hague.

[22] D. O'Shaughnessy. (1992). Analysis and automatic recognition of false starts in spontaneous speech. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, (J. Ohala, ed.), University of Alberta: Banff, Alberta, 931–934.

[23] S. L. Oviatt and P. R. Cohen. (1991). Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language, 5(4)*, 297–326.

[24] S. L. Oviatt and P. R. Cohen. (1992). Spoken language in interpreted telephone dialogues. *Computer Speech and Language, 6*, 277–302.

[25] S. L. Oviatt, P. R. Cohen, M. W. Fong, and M. P. Frank. (1992). A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, (J. Ohala, ed.), University of Alberta: Banff, Alberta, 1351–1354.

[26] S. L. Oviatt, P. R. Cohen, M. Wang, and J. Gaston. (1993). A simulation-based research strategy for designing complex NL systems. In *ARPA Human Language Technology Workshop*, Morgan Kaufmann: San Mateo, Calif.

[27] S. L. Oviatt, P. R. Cohen, and M. Wang. (1994). Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. In *Speech Communication* (European Speech Communication Association), *15, 3-4*.

[28] S. Shattuck-Hufnagel and D. Klatt. (1979). The limited use of distinctive features and markedness in in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18*, 41–55.

[29] E. Shriberg, J. Bear, and J. Dowding. (1992). Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann: San Mateo, Calif., 23–26.

[30] M. P. R. van den Broecke and L. Goldstein. (1980). Consonant features in speech errors. In *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand.* (V. A. Fromkin, ed.), Academic Press: London, U. K.