THE CONTRIBUTING INFLUENCE OF SPEECH AND INTERACTION ON HUMAN DISCOURSE PATTERNS

Technical Note 452

November 18, 1988

By: Sharon L. Oviatt
Philip R. Cohen
Artificial Intelligence Center
Computer and Information Sciences Division

APPROVED FOR PUBLIC RELEASE:
DISTRIBUTION UNLIMITED

This research was supported primarily by the National Institute of Education under contract US-NIE-C-400-76-0116 to the Center for the Study of Reading at the University of Illinois and Bolt Beranek and Newman, Inc., and in part by a contract from ATR International to SRI International. This paper will appear in *Architectures for Intelligent Interfaces: Elements and Prototype*, edited by J. W. Sullivan and S. W. Tyler, Addison-Wesley Publishing Co., Menlo Park, California, 1989.

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

The Contributing Influence of Speech and Interaction on Human Discourse Patterns*

Sharon L. Oviatt
Philip R. Cohen
Artificial Intelligence Center
SRI International

November 18, 1988

*This research was supported primarily by the National Institute of Education under contract US-NIE-C-400-76-0116 to the Center for the Study of Reading at the University of Illinois and Bolt Beranek and Newman, Inc., and in part by a contract from ATR International to SRI International. Authors' current address: Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, Ca. 94025

1 Introduction

Communication channels physically constrain the flow and shape of human language just as irresistably as a river bed directs the river's current. Escarpments speed the current, sculpting jetties and whirlpools. Meadows encourage evenness, a certain recumbency. The flow becomes a deafening cascade as it passes over granite bolders, and is arrested abruptly behind man-made dams. In short, the river is molded, rendered navigable or not, through the physical medium of its own bed. Although communication modalities may be less visually compelling than the terrain surrounding a river, it is a mistake to assume that they are any less influential in shaping the language transmitted within them.

Understanding the influence of communication modalities begins with an identification of their landmark features, and of the observable impact of these features on language. The present chapter focuses on two fundamental and potentially orthogonal landmarks that shape the nature of a communication modality: transmission through speech, and interaction between conversants. One goal of this chapter is to provide a comparison of the discourse and performance characteristics of instructions presented in three different modalities, each of which was classified according to the presence or absence of: 1) speech and 2) interaction. A second goal is to begin constructing an analytical framework from which predictions can be made about the separate impact of speech and interaction on specific aspects of discourse and performance. If we can predict in advance what the desirable and undesirable qualities will be of proposed language interfaces and technologies that use different communication modalities, then we will be in a better position to guide the selection and design of optimal alternatives for future applications.

In the present research, telephone, audiotape, and keyboard instructions were compared for teams in which an expert instructed a novice on how to assemble a hydraulic water pump. The spoken modalities were telephone and audiotape, which contrasted with typed input via keyboard. The two modalities with interactive capability included telephone and linked keyboard, whereas the audiotape instructions were transmitted noninteractively. A final partitioning of the modalities was based on the presence of *interactive speech.* For this contrast, the telephone modality, which was unique in its natural combination of both the speech and interaction components, was distinguished from audiotape and keyboard. A comprehensive set of discourse and performance characteristics was examined, including referential, organizational, and efficiency measures. For each measure, the differences in magnitude were compared among all three modalities, and these data were analyzed to determine whether the characteristic predominated when speech was present, when interaction was present, or only when these factors occurred together in interactive speech. Table 1 illustrates the logical partitioning and predicted data pattern among modalities for those characteristics considered to be influenced by speech, interaction, or interactive speech.

Previous research on communication modalities has focused primarily on establishing descriptive characterizations of individual modalities. Although this research occasionally has viewed modalities in terms of their underlying factors, it has not followed through any given theoretical perspective with the empirical comparisons needed to link particular discourse and performance patterns with the proposed underlying factors. As a result of this limitation, as well as a lack of research on how the interaction factor influences communication, research to date has not considered the relative influence of speech and interaction on the characteristics of a modality. Nonetheless, general descriptions of spoken and written communication provide some basis for predicting which characteristics may be associated with the presence of speech. Among other things, speech has been described as less concise, more repetitious, more replete with pronouns, more rapidly delivered, and associated with faster task performance (Blass & Siegman, 1975; Chafe, 1982; Chapanis, Parrish, Ochsman, & Weeks, 1977; Stoll, Hoecker, Krueger, & Chapanis, 1976). On this basis, the present research predicts that the two spoken modalities, by comparison with the keyboard one, are more likely to display a greater number of words, more personal pronouns, and faster novice assembly time. Since interactive speech has been reported to be more efficient than noninteractive speech (Oviatt & Cohen, 1988), this further distinction is incorporated in the present prediction that interactive speech should display the fastest novice assembly time, followed by audiotape, and then keyboard. In addition, recent research has indicated that frequent repetitions are primarily a feature of noninteractive speech, rather than interactive speech or speech in general (Oviatt & Cohen, 1988). On the basis of this qualification, the prediction is made that repetitions are more likely to vary with the presence or absence of interaction, not speech. Finally, since temporally oriented organizational markers such as "Okay, next..." have been found to introduce almost all assembly segments in both interactive and noninteractive spoken modalities (Oviatt & Cohen, 1988), the prediction is made that such temporal markers are more likely to be characteristic of speech than interaction. If this is true, then the present comparison should find that they occur at a lower rate in the keyboard modality.

Although noninteractive human speech is the required input for a variety of innovations in progress, such as voice mail and automatic dictation devices (Gould, 1982; Gould & Boeis, 1983; Gould, Conti, & Hovanyecz, 1983; Jelinek, 1985; Nicholson, 1985), the organization of human discourse and performance under conditions of restricted interactivity is still poorly understood. In our study contrasting the transmission of instructions in interactive and noninteractive spoken modalities (Oviatt & Cohen, 1988), we identified differences in discourse organization, referential characteristics, and performance efficiency. For example, descriptive elaborations and repetitions were much more prevalent in the noninteractive audiotape mode, as were introductions of upcoming actions and summary descriptions. Furthermore, the absence of interaction in spoken discourse corresponded with reduced performance efficiency. Referential differences between interactive and noninteractive speech also have been highlighted by earlier studies demonstrating the role of "concurrent feedback" in progressively reducing the length of referring expressions that are repeated during dialogue (Krauss & Weinheimer, 1964 & 1966). The present research predicts that the rate of elaborations, repetitions, action introductions, and summaries all will be influenced primarily by the presence or absence of interactive feedback within a modality. In particular, the interactive telephone and keyboard exchanges that include confirmation feedback are predicted to contain fewer of these features than noninteractive audiotape.

With respect to the introduction of new objects, telephone dialogue has been characterized as more indirect and fine-grained than either interactive keyboard (Cohen, 1984) or noninteractive audiotape (Oviatt & Cohen, 1988). Telephone dialogue also has been described as more indefinite than audiotape in its use of determiners to introduce new objects (Oviatt & Cohen, 1988), although this comparison has yet to be made with keyboard. For purposes of the present research, then, it was clear in advance that the habitual use of separate indirect requests for identification of new objects depends on the presence of interactive speech. Based on this background information, as well as on the observed concordance between indirection and indefiniteness of description (Oviatt & Cohen, 1988), it is predicted that experts will habitually use more indefinite determiners to introduce new objects in telephone than in keyboard interactions. That is, it is predicted that indefinite reference to new objects likewise will predominate during interactive speech.

2 Overview of Empirical Study

2.1 Method

For the present research purposes, data from thirty subjects, fifteen experts and fifteen novices, was examined and compared. These subjects represented a subset of the total of fifty who participated in a larger study, partial results of which have been reported elsewhere (Cohen, 1984; Oviatt & Cohen, 1988). The fifteen novices had been randomly assigned to experts to form a total of fifteen expert-novice pairs. All subjects were paid student volunteers. For five of the pairs, the expert related instructions by telephone, and an interactive telephone dialogue ensued as the pump was assembled. For another five pairs, the expert's spoken instructions were recorded by audiotape, and later the novice assembled the pump as he or she listened to the expert's taped monologue. For the last five pairs, the expert typed instructions on a keyboard, and a typed interactive exchange then took place between the participants on linked CRTs. The fifteen pairs of participants were randomly assigned to the telephone, audiotape, and keyboard conditions.

Each expert participated in the experiment on two consecutive days, the first for training and the second for instructing the novice. During training, experts were informed that the purpose of the experiment was to investigate modality differences in the communication of instructions. They were given a set of assembly directions for the hydraulic pump kit, which were written as a list of imperatives, along with a diagram of the pump's labeled parts. Approximately twenty minutes was permitted to practice putting the pump together using these materials, after which the subject practiced administering the instructions to a research assistant. If a subject was doubtful or experienced difficulty during practice, training continued for an additional ten to fifteen minutes.

During the second session, the expert was informed of a modality assignment. Then the expert was asked to explain the task to the novice partner, and to make sure that the partner built the pump so that it would function correctly when completed. The expert was allowed to view the water pump parts for reference while giving instructions, although touching the pieces was prohibited. The novice received similar instructions to the expert regarding the purpose of the experiment, and was supplied with all the water pump parts and a tray of water for testing.

In the telephone condition, the expert spoke through a standard telephone receiver, and the novice listened through a speaker-phone so that his or her hands would be free for assembly. The expert and novice were located in adjacent rooms. In the audiotape condition, a cassette recorder was used for recording and playback of the experts' instructions. The novice was tested after recording, and was at liberty to rewind and review sections of the tape. Keyboard teams typed their instructions on Elite Datamedia 1500 CRT terminals connected by the Telnet computer network to a computer at Bolt Beranek and Newman, Inc. The terminals were linked so that whatever was typed on one would appear on the other. Simultaneous typing was possible and did occur. Participants were informed that their typing would not appear instantaneously on their partner's terminal and, in fact, response times averaged 1 or 2 seconds with occasional longer delays due to system load. For all three modalities, assembly of the pump was videotaped. Written transcriptions were available as a hard copy of the keyboard exchanges, and were composed from audio-cassette recordings of the monologues and coordinated dialogues, the latter of which had been synchronized onto one audio channel beforehand. Signal distortion was not measured in either spoken modality, although no subjects reported difficulty with inaudible or unintelligible instructions and < 0.2%, or 1 in 500 recorded words, were undecipherable to the transcriber and experimenter. In all cases, the participants were aware that their behavior would be recorded for later study by the researchers.

2.2 Sample Transcripts

Discourse fragments are provided below to illustrate all three modalities. Each sample includes instructions on how to assemble two parts.

Telephone dialogue segment:

Expert: "Now, do you see a little pink plastic piece?"
Novice: "Yeah, yeah."
Expert: "With two holes?"
Novice: "Yeah."
Expert: "Okay. You have your blue cap in front of you?"
Novice: "Yeah."
Expert: "Setting down with the two little prongs sticking up?"
Novice: "Yeah."
Expert: "Okay, take that little pink plastic piece,
and the two holes in the plastic piece–"
Novice: "Mm–hm."
Expert: "–go over the two little notches."
Novice: "Does it matter whether the shiny side or
the dull side of the pink thing's up?"
Expert: "No, it doesn't matter."
Novice: "Okay."
Expert: "And put it so that it's covering the hole
in the bottom of that little cap.
Kinda fits hard, doesn't it?"
Novice: "Little bit tight, yeah. Okay."

Audiotape monologue segment:

Expert: "So the first thing to do is to take the metal
with the red thing on one end and the green cap on the other end.
Take that and then look in other parts–
there are three small red pieces. Take the smallest one.
It looks like a nail-
a little red nail–
and put that into the hole in the end of the green cap.
There's a green cap on the end of the silver thing.
Take the little red nail and put it in the hole in the end
of the green cap."

Keyboard interaction segment:

Expert: "Now take the blue cap with the two prongs sticking out
and fit the little piece of pink plastic on it. Okay?"
Novice: "Okay."

2.3 Results

Within the research framework described, the data from all three conditions were collected, and then scored and analyzed for their discourse and performance characteristics. Detailed methods for coding, second scoring, and analyzing the dependent measures are summarized elsewhere (Oviatt & Cohen, 1988). All dependent measures reported in this chapter had reliabilities ranging above .86. The basic statistical comparisons were based on *apriori* t or Fisher's exact probability tests (Siegel, 1956).

2.3.1 Aspects of Discourse and Performance Influenced by Speech

Among the discourse characteristics that were influenced solely by the presence of speech were overall wordiness, use of personal pronouns, and the habitual introduction of individual discourse segments with temporal markers such as "Okay, next..." That is, evaluation of these dependent measures revealed that the interactive and noninteractive speech modes did not differ significantly and that, after collapsing data from the speech modes, the contrast between the speech and nonspeech modes was a significant one. With respect to discourse length, the audiotape and telephone experts spoke an average of 875 and 845 words, respectively, compared with only 303 words in keyboard ($t = 5.87$, $df = 12.19$, $p < .0001$, one-tailed, with separate variance estimates). Likewise, personal pronouns were uttered at the average rate of 4.14 and 4.41 per 100 words in the audiotape and telephone transcripts, by comparison with only 1.64 in keyboard ($t = 4.03$, $df = 13$, $p < .001$, one-tailed). Finally, whereas almost all discourse segments were introduced with temporal markers in the audiotape and telephone modes, or an average of 96.3% and 98.6%, respectively, only 43.0% received such initial marking in the keyboard mode. All five experts in the two speech modalities met the criterion of producing nine or more introductory markers, while only one of five keyboard experts did so — a significant departure from chance based on Fisher's exact probability test ($p = .02$).

With respect to performance, efficiency was enhanced quite substantially by the presence of speech, although interaction contributed further to the efficiency of the spoken modalities. Whereas the average assembly time for the audiotape and telephone novices was 530 and 417 seconds, respectively, novices using interactive keyboard required an average of 1,485 seconds to construct the same water pump ($t = 3.71$, $df = 4.13$, $p < .01$, one-tailed, with separate variance estimates).

2.3.2 Aspects of Discourse and Performance Influenced by Interaction

Discourse characteristics that were influenced solely by the presence of interaction included descriptive elaboration and repetition, as well as the prevalence of action introductions and summaries. For this collection of dependent measures, analyses revealed that the two interactive modes did not differ significantly, although the combined data from the interactive modes was significantly different from that of the noninteractive mode. With respect to referential characteristics, experts operating in the noninteractive audiotape modality elaborated their descriptions at the rate of 3.94 per 100 words, which was a significantly higher rate than the 2.09 and 1.36 elaboration rates produced by telephone and keyboard experts, respectively ($t = 4.74$, $df =13$, $p < .0005$, one-tailed). Perseverative elaborations[1] occurred at the average rate of .86 per 100 words in noninteractive audiotape, whereas they were nonexistent in keyboard and nearly nonexistent in telephone (.075 per 100 words), this latter contrast a significant one ($t = 3.27$, $df = 4.18$, $p < .02$, one-tailed, with separate variance estimates). Furthermore, descriptive reversions[2] were found exclusively in the noninteractive audiotape modality. Expert repetitions also occurred at a significantly higher rate in noninteractive audiotape, which had

[1] The definition and coding of elaborative phenomena such as perseverations is outlined in detail in Oviatt & Cohen (1988). Descriptive perseveration refers to continued elaboration of a piece description *after* the expert has explained how to assemble the piece, but within the same discourse assembly segment (see Sample Transcripts section for context of the following audiotape perseveration: "There's a green cap on the end of the silver thing."). A discourse assembly segment refers to a segment of discourse in which the expert provides instructions for attaching two parts or subassemblies as part of an individual assembly step.

[2] Descriptive reversion refers to an elaborative pattern in which the expert describes a new piece in a direct and definite manner, but then downshifts to an indirect and indefinite elaboration about the same piece (e.g., Audiotape expert: "...you take *the* L-shaped clear plastic tube, *another* tube, there's *an* L-shaped one with a big base...").

.65 repetitions per 100 words, compared with .15 and .13 in the interactive telephone and keyboard modes ($t = 3.19$, $df = 13$, $p < .005$, one-tailed).

With respect to organizational features, both action introductions (e.g., "Now we are going to assemble the base of the pump") and descriptive summaries (e.g., "Okay, so at the moment you are going to have this body of the pump with the plunger in it, and the red cap at the top with the base on it, and standing on this pedestal, this plastic pedestal") were significantly more prevalent in the noninteractive than the interactive modalities. In noninteractive audiotape, experts introduced upcoming actions an average of 3 times per transcript, compared with once in interactive telephone and none in interactive keyboard. The contrast between action introductions in audiotape and telephone was significant ($t = 1.91$, $df = 8$, $p < .05$, one-tailed). Likewise, descriptive summaries were issued at the rate of .47 per 100 words in the noninteractive audiotapes, compared with .17 in interactive telephone and none in the interactive keyboard. The difference between audiotape and telephone summaries again was significant ($t = 2.03$, $df = 8$, $p < .04$, one-tailed.)[3]

The opportunity for speaker interaction further strengthened performance efficiency beyond the level afforded by speech alone. This is reflected in the fact that, while novices in the two spoken modalities performed much faster than keyboard novices, those in the interactive telephone modality also worked significantly faster than audiotape novices, with average assembly times of 417 and 530 seconds, respectively, ($t = 1.87$, $df = 8$, $p < .05$, one-tailed).

2.3.3 Unique Discourse Characteristics of Interactive Speech

Some aspects of referential and illocutionary style appeared only when the combined resources of speech and speaker interaction were both present, as represented in the relatively natural telephone modality. In the introduction of new piece descriptions, four of the five telephone experts met the criterion of introducing nine or more new pieces with an indefinite determiner, whereas none of the five audiotape or keyboard experts did so, a significant difference based on Fisher's exact test ($p = .02$). Four of the five telephone experts also used a separate indirect request for piece identification when introducing nine or more of the pieces, although none of the five audiotape or keyboard experts made a separate request of this type. Again, this was a significant difference based on Fisher's ($p = .02$). Instead, audiotape and keyboard experts adopted a more definite and direct style, one in which they immediately requested that the novice assemble the newly introduced piece.

Table 2 summarizes the discourse and performance characteristics that were influenced primarily by speech, by interaction, or by the confluence of both of these factors.

[3] Although interactive keyboard clearly provided an even more pronounced contrast with respect to both introductions and summaries, parametric analyses were precluded due to nonoccurrence. However, if experts in audiotape and keyboard were classified according to whether or not they used introductions, and again according to whether they used summaries, both classifications would reveal 5 of 5 audiotape experts and 0 of 5 keyboard experts as having qualified as producers of at least one introduction or summary, which constitutes a significant departure from chance based on Fisher's, $p < .01$.

3 Discussion

One of the more remarkable features of spoken language modalities was confirmed to be the sheer copiousness of their output. By comparison, typed interaction was very abbreviated, almost telegraphese. In spite of their verbosity, the speech modalities were characterized by substantially faster novice assembly times. This efficiency advantage for speech ranged approximately three-fold, with task completion in the spoken modalities averaging seven to nine minutes, whereas keyboard novices required over twenty-four minutes for the same task. Previous research has reported an efficiency advantage for spoken modalities of approximately two-fold over written and typed exchanges (Chapanis, Parrish, Ochsman, & Weeks, 1977), based on two-person cooperative assembly and geographical location tasks. These data substantiate what has been the general conjecture that speech may be a particularly apt selection for use with hands-on tasks in which typing or writing otherwise would detract from overall performance time. Furthermore, they establish the margin of advantage for hands-on tasks as falling within the two- to three-fold range. Future research on tasks of practical interest other than assembly tasks could contribute further specifics on the natural advantages that make speech a powerful modality. A clearer perspective on these issues will be vital as we strive to harness speech fully for technological purposes.

Apart from their verbosity and efficiency, spoken language modalities also elicit frequent use of personal pronouns, irrespective of whether direct speaker interaction and feedback is present or not. In the past, the very high rate of pronouns in interactive speech has been construed as an index of personal involvement between the participants (Chafe, 1982). If this were true, then modalities permitting more direct speaker interaction, including clarification subdialogues and confirmation exchanges, should generally be characterized by a profusion of personal pronouns. However, in the present research, the rate of personal pronouns produced by experts was substantially lower during keyboard interactions, by a factor of 2.6. Perhaps equally disconcerting to the "direct involvement" viewpoint, personal pronouns were as frequently used by experts in the noninteractive audiotape modality as they were by those engaged in interactive telephone dialogues. In short, the present data provide evidence implicating speech, not interaction, as the common modality factor underlying frequent use of personal pronouns.

One possible explanation for this finding is that there is a strong tendency in spoken language modalities for speakers to create a subjective sense of direct interaction and involvement with a partner, even when the modality itself actually precludes any such opportunity. For example, it has been argued that audiotape experts fabricate a mute listener (Oviatt & Cohen, 1988), perhaps as an aid in composing instructions[4]. As they interact with this fictitious partner, evidently suspending the realities of the known time delay in audiotape communication, they have been found to engage in residual requests for confirmation, followed by comments typical of interactive responses (e.g., Audiotape expert: "Okay, you got that? Good.") These

[4] Goffman (1981) has aptly described people's subjective sense that engaging in solitary monologues feels distinctly like a failure of decorum, especially if observed publicly. In this sense, audiotape talk can be viewed as a sort of technologically induced social impropriety, which results from the delamination of speech from its usual interactional framework.

data on the characteristics of audiotape experts' speech highlight the possibility that there exists subjective gravitation toward fully interactive speech, or an attempt on the speaker's part to recreate a more natural and familiar form of collaborative interaction. Evidently, this is one strategy that audiotape experts use to cope with the strain induced by noninteractive performance requirements. The extent to which people create interactional placeholders when using limited interaction modalities needs to be investigated if we desire accurate user models that are capable of predicting the discourse and performance patterns of future language technology. To date, the user modelling literature has not addressed the relation between modality constraints and the performance models of users.

By contrast with speech, the presence or absence of interaction was associated mainly with referential features of the discourse. By comparison with telephone dialogues and keyboard interactions, noninteractive monologues were distinguished principally by the extensiveness of their elaborative and repetitive description. These profusely elaborative descriptions focussed on the pieces and actions that formed the essence of the present assembly instructions. They also contained more unique elaborative patterns, such as perseverated and reverted elaborations of pieces. These latter elaborative phenomena gave the impression of being out-of-sequence parenthetical additions that disrupted the smooth continuity of the audiotape discourse. For a variety of reasons outlined by Oviatt & Cohen (1988), this collection of referential characteristics rendered the noninteractive audiotape modality far less integrated and predictably sequenced than either of the interactive modalities and, as such, it created more inferential strain for listeners in this modality. Finally, performance efficiency clearly was eroded by lack of speaker interaction and feedback, although the magnitude of the interaction effect on efficiency was relatively small in comparison with the influence of speech.

These basic differences in discourse organization and performance efficiency have implications for the successful design of various applications using interactive and noninteractive input. For example, technology based on noninteractive dictation will be prone to excessive elaboration and repetitiveness that slow down and undermine the coherence of messages, contributing to inefficient processing by the recipient or system. Such communications also will tend toward poorer integration and less predictable structuring, requiring more effort by the recipient to resolve their meaning. When written output is desired, noninteractive speech will lead to poor copy that requires labor-intensive editing, a disadvantage that must be weighed with respect to the initial advantages of spoken input. In fact, the major current impediment to the acceptability of such systems is the unavailability of technically-adequate editing facilities (Ades & Swinehart, 1986) . To reduce the impact of these outlined difficulties on professional communications such as voice mail and other automatic dictation devices, it may be strategic to limit such technology to brief and informal tasks, and to ones that do not emphasize planning or reviewing during transmission. In fact, the recommendation for brevity accords with the usage patterns and preferences reported in recent voice mail and filing studies (Nicholson, 1985; Gould & Boeis, 1983), which indicated a tendency for users to limit noninteractive messages to under one minute in length.

Above a certain threshold, language systems slower than real-time will elicit user input that has similar characteristics to noninteractive language. When system responding is slow and, in addition, prompt confirmations to support the user—system "dialogue" are not forthcoming, then users will elaborate and repeat themselves (van Katwijk, van Nes, Bunt, Muller, & Leopold, 1979). For practical purposes, that is, users are unable to distinguish between a slow response and no response at all, so their strategy for coping with both situations is similar. Ultimately, to improve both dictation technology and delayed-response language systems, the most direct solution may be to design methods of confirmation feedback that effectively inhibit speaker elaborations, along with the discourse convolutions and inefficiency that they precipitate (see Oviatt & Cohen, 1988).

Speech and interaction each were associated with different patterns of discourse organization. Audiotape experts provided more organizational enhancements in the form of introductions of upcoming actions and summary descriptions than did experts in either interactive modality. These organizational reinforcements may have assisted in offsetting the relative lack of integration and predictability in audioptape discourse by focusing the more rambling audiotape descriptions. By contrast, experts in the spoken language modalities habitually provided local temporal markers, such as "Okay, then...," before describing each individual assembly action, whereas keyboard experts did not. In short, organizational enhancements that occurred due to the absence of speaker interaction operated at the propositional level, through explicit advance introduction or reviewing of main points. Instead, the organizational fortification in spoken modalities was temporal in nature, perhaps simply because visual sequencing of the assembly instructions is not possible in speech as it is in written or typed modes.

The discourse characteristics that surfaced exclusively during interactive telephone speech, but not in modalities in which either speech or interaction were absent, centered on the referential and illocutionary style used to introduce new pieces. Basically, experts engaging in a telephone dialogue habitually introduced new water pump pieces in an indefinite and indirect manner, using a more fine-grained series of illocutionary steps. Telephone experts typically decomposed new piece descriptions into two parts: identify and act. As step one, they indirectly requested identification of a piece, which the novice then confirmed, before they progressed to step two – more detailed instructions for picking up, orienting or acting on the piece. In contrast, both audiotape and keyboard experts maintained a more emphatic and direct illocutionary style. These experts were more presumptuous of the novice's initial recognition of new pieces, as reflected in their definite descriptions of them, and they were more assertive about immediately instructing the novice to act on new pieces in a particular way. These data suggest that the use of an indirect and indefinite illocutionary style, one that imparts instructions through a more fine-grained series of illocutionary steps that are each confirmable, is a relatively fragile discourse pattern that relies on the confluence of both speech and interaction.

One implication of the habitually directive style of both noninteractive and nonspeech modalities, which is at odds with the indirection of dialogue, is that users risk creating misleading impressions of their attitude toward the task and the message recipient. This may be especially problematic for comunication in professional settings, in which the need to observe status distinctions is generally emphasized. For example, potential users may be reluctant to send messages to higher status individuals such as managers using voice mail or other noninteractive modalities that encourage a directive, brusk, or impatient tone. Future research in this area could begin with careful observation of natural usage patterns for technology prone to such problems, and could determine whether the outlined stylistic phenomena are amenable to training.

This chapter has described several basic ways in which speech and interaction each mold the current of language flowing within a communication modality. The long-term goal of the present approach is to construct a model capable of predicting the advantages and disadvantages of different communication modalities for future language interfaces and technology. Such a model can provide one source of guidance in the selection and design of proposed systems. However, unique discourse features always can be expected to emerge that are not derivable in any simple way from underlying modality factors, which was evident from the patterns of interactive speech. As a result, research with prototype systems will be required in order to refine or correct anticipated outcomes. Information generated by empirical models, and supplemented with performance results collected during iterative design, will speed the process of crafting future language systems that are habitable, high quality, and enduring.

References

[1] S. Ades and D. C. Swinehart. Voice annotation and editing in a workstation environment. In *Proceedings of AVIOS '86: Voice I/O Systems Applications Conference*, pp. 13–28, American Voice I/O Society, Alexandria, Virginia, September 1986.

[2] T. Blass and A. W. Siegman. A psycholinguistic comparison of speech, dictation and writing. *Language and Speech*, 18:20–34, 1975.

[3] W. L. Chafe. Integration and involvement in speaking, writing, and oral literature. In D. Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, chapter 3, pp. 35–53, Ablex Publishing Corporation, Norwood, N. J., 1982.

[4] A. Chapanis, R. N. Parrish, R. B. Ochsman, and G. D. Weeks. Studies in interactive communication: II. The effects of four communication modes on the linguistic performance of teams during cooperative problem solving. *Human Factors*, 19(2):101–125, April 1977.

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

[6] E. Goffman. *Forms of Talk.* University of Pennsylvania Press, Philadelphia, Pennsylvania, 1981.

[7] J. D. Gould. Writing and speaking letters and messages. *International Journal of Man-Machine Studies*, 16(1):147–171, 1982.

[8] J. D. Gould and S. J. Boeis. Human factors challenges in creating a principal support office system — the speech filing system approach. *ACM Transactions on Office Information Systems*, 1(4):273–298, October 1983.

[9] J. D. Gould, J. Conti, and T. Hovanyecz. Composing letters with a simulated listening typewriter. *Communications of the ACM*, 26(4):295–308, April 1983.

[10] F. Jelinek. The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, 73(11):1616–1624, November 1985.

[11] R. M. Krauss and S. Weinheimer. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1):113–114, 1964.

[12] R. M. Krauss and S. Weinheimer. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–346, 1966.

[13] R. T. Nicholson. Usage patterns in an integrated voice and data communications system. *ACM Transactions on Office Information Systems*, 3(3):307–314, July 1985.

[14] S. L. Oviatt and P. R. Cohen. *Discourse Structure and Performance Efficiency in Interactive and Noninteractive Spoken Modalities.* Technical Report, Artificial Intelligence Center, SRI International, Menlo Park, California, forthcoming.

[15] S. Siegel. *Nonparametric Methods for the Behavioral Sciences.* McGraw-Hill Publishing Co., New York, New York, 1956.

[16] F. C. Stoll, D. G. Hoecker, G. P. Kruger, and A. Chapanis. The effects of four communication modes on the structure of language used during cooperative problem-solving. *Journal of Psychology*, 94(1):13–26, 1976.

[17] A. F. VanKatwijk, F. L. VanNes, H. C. Bunt, H. F. Muller, and F. F. Leopold. Naive subjects interacting with a conversing information system. *IPO Annual Progress Report*, 14:105–112, 1979.

Table 1

Predicted Patterns Among Modalities for Characteristics Influenced by Speech, Interaction, and Interactive Speech

	Modalities:		
	Audiotape	Telephone	Keyboard
Characteristics Influenced by:			
Speech	+*	+	—
Interaction	—	+	+
Interactive Speech	—	+	—

*The *directionolity* of predicted magnitude differences among modalities for speech, interaction, and interactive speech characteristics is inconsequential. For example, either a — + + or a + — — pattern among modes for a given feature indicates the association of that feature with interaction.

Table 2

Discourse and Performance Characteristics Influenced by Speech, Interaction, and Interactive Speech

	Spoken Modalities		Interactive Modalities
	Audiotape	Telephone	Keyboard
Speech:			
Speed of Assembly Time	+	++	–
Number of Words	+	+	–
Personal Pronouns	+	+	–
Initial Temporal Markers	+	+	–
Interaction:			
Number of Elaborations	+	–	–
Perseverations	+	–	0
Reversions	*	0	0
Repetitions	+	–	–
Introduction of Actions	+	–	0
Summary Descriptions	+	–	0
Interactive Speech:			
Separate Requests for ID of New Pieces	–	+	–
Indefinite Reference to New Pieces	–	+	–

++ and + and – specify greatest, greater, or lesser amounts of a feature, respectively, each of these distinctions statistically significant

0 and * designate a nonexistent or sometimes present feature