SRI International

Technote 454 • November 1988

DISCOURSE STRUCTURE AND PERFORMANCE EFFICIENCY IN INTERACIVE AND NONINTERACTIVE SPOKEN MODALITIES

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

The 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 SR1 International.

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

1 Abstract

Speaker interaction is a central feature of human dialogue, one with a powerful influence on its discourse structure and performance efficiency. The present study examined two speech modalities that represent opposites on the spectrum of speaker interaction — the telephone dialogue and audiotape monologue. Experts provided instructions by either telephone or audiotape as their novice partner completed an assembly task. Within this task framework, a comprehensive analysis is provided of the basic differences in discourse organization, referential characteristics, and performance efficiency for these two spoken modalities. The outlined distinctions are interpreted with special reference to the role of confirmation feedback in promoting dialogue efficiency. Implications are discussed for the development of prospective speech systems designed to be habitable, high quality, and relatively enduring. Finally, a theoretical model of collaborative dialogue is proposed from which several features of interactive and noninteractive speech can be derived.

2 Introduction

In the near future, technological advances incorporating speech will surface in myriad forms. To harness speech fully for such purposes, the natural advantages that make it a powerful modality need to be elucidated. Behaviorally speaking, natural speech is both fundamentally social and highly skilled. People experience it as a very rapid, direct, and tightly interactive modality, one that is governed by an array of conversational rules and is rewarding in its social effectiveness. As with all skilled performance, the regulation of speech also is relatively automatized, and this is reflected in its actual and expected usage patterns. However, the implications of these characteristics for the development of successful speech applications remain to be explored. The present research examines the distinct discourse and performance characteristics of two types of spoken discourse, telephone dialogue and audiotape monologue, which differ in the opportunity for speaker interaction and feedback.

Clearly, many basic issues need to be addressed before technology can leverage from the natural advantages of speech. One issue involves the structuring of spoken interfaces to reflect the realities of speech instead of text. Historically, language norms have been based on written modalities, even though research has established that spoken and written communication differ in major ways (Blass & Siegman, 1975; Chafe, 1982; Chapanis, Parrish, Ochsman, & Weeks, 1977; Hindle, 1983; Stoll, Hoecker, Krueger, & Chapanis, 1976). Within the realm of technologically oriented modalities, telephone speech also has very different qualities from interactive keyboard (Cohen, 1984; Oviatt & Cohen, 1989). This literature indicates that spoken communication, by comparison with written, tends to be delivered much more rapidly, to be less planned, less concise, less complex and well integrated syntactically, with shorter and less varied vocabulary, more dysfluencies, hedges, quantifiers, and function words, more pronouns, more requests for confirmation and listener confirmations, more repetition, more noun phrase reductions with repeated reference, more indirection, a more fine-grained decomposition of requests, and more metacomments about the content and discourse itself. Among other things, this extensive set of differences establishes that the spoken/written distinction is hardly a unidimensional one. Furthermore, exploratory research has begun to highlight the difficulty that a spoken language system can have with such features as the indirection, confirmations and reaffirmations, quantifiers, nonword fillers and overall wordiness of human speech (van Katwijk, van Nes, Bunt, Muller & Leopold, 1979). Many of the speech characteristics cited in the literature will require basic design accommodations if spoken language systems are to function successfully and efficiently.

In addition to these different characteristics of speech, users appear to have more inflated expectations of natural language systems that incorporate speech than they do of keyboard systems. It is well known that users interacting with a natural language question-answering system tend to assume greater linguistic and conceptual coverage than is actually possessed by the system, a mismatch that can seriously hinder performance (Hendrix & Walter, 1987; Small & Weldon, 1983; Turner, Jarke, Stohr, Vassiliou & White, 1984). The fact that people associate speech with highly coordinated interaction and speed can further inflate their expectations of the *interactional* coverage of spoken language systems. Users generally are

intolerant of speech systems that are perceived to have unnatural qualities, or that are demanding to learn, socially inappropriate, slow, or otherwise unresponsive (van Katwijk et. al., 1979). For example, system pauses in speech output that were considered long by human standards have resulted in more repetitions, interruptions, and eventual user aversion to the system (Potosnak & van Nes, 1984; van Katwijk et. al., 1979). In this respect, a speech interface compounds the burden of orienting the user to actual machine capabilities.

The introduction of speech also can lead to unpredictable preferences and use of a system's capabilities, with resulting inefficiency or uninhabitability of the system. For example, people have been found to abandon pointing with a mouse when a potentially redundant speech capability was added (Wulfman, Isaacs, Webber & Fagan, 1989). System integration discontinuities of this type are especially likely to be incurred when speech is added as a "front end" afterthought, rather than being incorporated more fully during iterative system design. In short, designing new speech interfaces will pose many unanticipated hurdles. To the extent that system designers can collect advance speech data that resolves major difficulties and uncertainties, they will be in a better position to craft systems that support optimal dialogue and task performance.

Another major issue concerns the fact that spoken language technology, by comparison with natural face-to-face dialogue, invariably either restricts information channels or reduces interactive feedback. For example, telephone transmission limits communication through removal of the visual dimension of face-to-face contact, and audiotape totally eliminates the interactive exchange and feedback typical of conversation. Essentially, these applications constitute a kind of technologically induced delamination of speech from the context of its original development and constant skilled use. 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.

Unfortunately, there is no solid theoretical foundation available for interactive dialogue from which to build. Recently, a theory of dialogue has begun to generate interest among researchers who view collaborative action as the foundation for a theory of communicative interaction (Clark & Wilkes-Gibbs, 1986; Cohen & Levesque, forthcoming; Grosz & Sidner, 1989; Searle, 1989). In this view, dialogue participants are able to support their collaboration through the formation and maintenance of mutual beliefs, joint commitments to particular communicative goals, and joint intentions to continue cooperating until those goals are achieved. However, this literature is just beginning to introduce the principal issues and terminology necessary for a workable theory of collaborative communication. It has yet to address the fundamental constraints imposed by different communication modalities, particularly the degree and structure of interaction permitted within a given modality, even though collaborative communication most certainly is shaped by the interactions possible between dialogue partners.

In a separate empirical literature on the characteristics of interactive dialogue, it is well established that when an object is referred to repeatedly during a task, the referring noun

phrase will reduce in length (Krauss & Weinheimer, 1964a, 1964b, 1966, & 1967). The original task in this research involves a speaker describing a series of abstract forms to a listener who, without visual access to the speaker, must select a card with the correct matching arrangement. The task is iterative, and results in each figure being described at least 15 times. In a related study, Krauss and Weinheimer (1964b) also observed generally longer name lengths in an audiotaped monologue condition, by comparison with dialogue, although they neither replicated this difference subsequently (Krauss & Weinheimer, 1967), nor did they pursue the issue of modality further.

In a classic study aiming to elucidate the dynamics of progressive noun phrase reduction, Krauss and Weinheimer (1966) found that the phenomenon was substantially influenced by the presence or absence of "concurrent feedback," or continual confirmations of attention and comprehension as the listener received instructions. They concluded that such feedback plays a direct role in enabling the speaker to converge on a brief encoding of the object. More recently, other researchers have asserted that a central feature of collaborative dialogue is the participants' implicit striving for progressive accommodation of one another and for communicative efficiency, which is evident in the gradual tailoring and streamlining of their dialogue descriptions (Clark & Wilkes-Gibbs, 1986; Isaacs & Clark, 1987). Within the framework of collaborative communication, confirmations are a primary vehicle for establishing the minimal level of descriptive detail needed between a particular pair of speakers, which in turn facilitates increased communicative efficiency. Together, these studies clearly identify confirmation feedback as influential in the shaping of economical referring expressions whenever the same objects are designated repeatedly.

Given the limited state of research on how the potential for interactive feedback affects different spoken language modalities, as well as the prospect for limited interactivity in future speech technology, the present study aimed for a characterization of the main distinctions between interactive and noninteractive speech. More specifically, telephone dialogues and audiotape monologues were compared for teams in which an expert instructed a novice on how to assemble a hydraulic water pump. In the audiotape monologues, there was no opportunity for the speakers and listeners to confirm their understanding as the task progressed, or to engage in clarification subdialogues with one another. Both modalities involved spatial displacement of the participants, but participation in the audiotape mode also was disjoint temporally. Unlike the paradigm for studies on referential reduction, the present task and procedure were not designed with an interest in collecting or analyzing repeated references to the same pump pieces or features. The primary purpose of the study was to provide a comprehensive analysis of differences in discourse organization, referential characteristics, and performance efficiency for dialogues and monologues during a task-oriented exchange. In this connection, the literature on referential reductions provided some predictive basis for beginning to hypothesize likely discourse and performance patterns in the present study. A further goal of the study was to examine the implications of observed contrasts between these two spoken modalities for the development of future speech systems that are habitable, high quality, and relatively enduring. Since the design of future speech systems will depend in part on adequate models of spoken discourse, a final goal was to construct a theoretical model capable of accounting

for several principal features of interactive and noninteractive speech.

3 Method

Twenty subjects, ten experts and ten novices, participated in the study. The ten novices were randomly assigned to experts to form a total of ten expert-novice pairs. These ten pairs of participants then were randomly assigned to the telephone and audiotape conditions. All subjects were paid student volunteers. The distribution of male and female participants was comparable for the two conditions, with 50% of the team members in each condition male and 50% female. For five of the pairs, each expert related instructions by telephone, and an interactive telephone dialogue ensued as the pump was assembled. For the other five pairs, each 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.

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. These materials are presented in Figure 1. The experts were allotted approximately twenty minutes to practice putting the pump together using these materials. Then they practiced administering the instructions to a research assistant. If an expert 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 about 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, the novice was tested the day after the expert's instructions were recorded. The novice was completely free to rewind and review sections of the tape. A cassette recorder was used for recording and playback of the experts' instructions. For both modalities, assembly of the pump was videotaped. Written transcriptions 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 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.

4 Sample Transcripts

Discourse fragments are provided below to illustrate the telephone and audiotape 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 rod
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."

5 Transcript Coding

Within the research framework described, the telephone and audiotape data were collected, and then scored and analyzed for their discourse and performance characteristics. Coding methods pertaining to different dependent measures are summarized below.

Discourse Assembly Segments. Each transcript's assembly instructions were divided into separate *discourse assembly segments* that described how to attach new water pump parts. Individual discourse assembly segments constituted one natural unit of analysis for evaluating discourse phenomena in the current seriated assembly task. Within each segment, one part or subassembly was described, then a second, followed by instructions on how to attach them. Examples of telephone and audiotape discourse assembly segments are presented in the Sample Transcripts section. In transcripts containing instructional backtracks (see upcoming definition), any segment that described the correct reassembly of one or more parts also was considered a separate assembly segment. The total number of segments per transcript ranged from 11 to 16. Any segment was operationally defined as "strained" if it was a backtrack segment, a segment in which the initial organizational marker was absent, a transitional segment that introduced or concluded the assembly instructions, or a segment containing three or more elaborations.

Organizational Markers. The total number of *organizational markers* or cue phrases (Grosz & Sidner, 1986; Reichman, 1981) that were produced by experts, which included temporal phrases such as "Okay, the first thing," "Okay, now," "Next," "Then," was tabulated for each transcript. A cluster of several markers together was counted as one. Each discourse assembly segment also was scored for the presence or absence of an initial marker, and the percentage of assembly segments introduced with a marker was calculated for each transcript.

Action Introductions. A tabulation was performed for each transcript of the frequency with which experts spontaneously introduced an upcoming action, or series of actions entailed in a subassembly, before initiating the discourse assembly segment(s) and describing the relevant pieces and action(s) (e.g., "Now we are going to assemble the base of the pump"). Any introductions of either the general assembly task or pump testing, which occurred in the initial and final phases of the discourse, were included in this tabulation of action introductions (e.g., "What we are going to attempt to do this morning is to put together a water pump" ; "Now that you've made it you have to test it and make sure it works").

Summary Descriptions. A tabulation was made of the total number of summaries of a water pump part, subassembly, or action that were described in a preceding discourse assembly segment. Only spontaneously produced summary descriptions were counted in the telephone modality, not expert reviews elicited by the novice during clarifications. Sometimes several summary descriptions were grouped together in a *summary checkpoint,* which was initiated with its own organizational marker (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."). Each of the summary descriptions within such grouped checkpoints were counted individually. In the above example, summaries of the three different subassembly features were counted separately. The total number of these spontaneous summaries per transcript was then converted to a rate per 100 words.

Elaborations. The total number of elaborated descriptions of pieces and actions that were spontaneously produced by each expert was tabulated. Any supplemental descriptions that added to, refined, or cast a new perspective on some earlier description in the discourse segment were coded as elaborations (e.g., "— there are three small red pieces. Take the smallest one. *It looks like a nail — a little red nail* —"). Elaborative phrases that were distinct grammatically or prosodically were scored as separate elaborations, even when they occurred in sequence and in reference to the same piece or action. For example, the two italicized noun phrase elaborations above were coded as separate. However, false starts were not counted as separate elaborations. Likewise, the first modifying relative clause or prepositional phrase in an original description was not considered elaborative. In the telephone modality, elaborations directly prompted by the novice during clarification subdialogues were not counted as spontaneous elaborations.

Elaborations were divided into those that primarily extended piece and assembly action descriptions. Piece elaborations (e.g., "It looks like a nail —") tended to follow the initial description of the relevant pump piece and to precede specification of the main assembly action in the segment. Piece elaborations were considered separate phrases or sentences that further specified a pump piece to be identified, while excluding reference to the assembly action through verb phrases, prepositional phrases, and so forth. By comparison, elaborations of assembly actions (e.g., "and then you push it all the way to the bottom...") always followed the basic assembly action description. These elaborations focused on clarifying the action, although they frequently contained embedded piece descriptions.

Three different dependent measures of discourse elaboration were prepared for analysis. First, the total number of spontaneous piece and action elaborations was converted to a rate per 100 words for each transcript. A second measure based on these codings was the average length of expert elaborations, computed for each transcript. The third measure concerned the number of times per transcript that an expert continued to elaborate a piece description *after* having explained how to assemble the piece, but within the same assembly segment. For this measure, piece elaborations were totaled before and after description of the main assembly action within each segment, and these totals were converted to a rate per 100 words of *initial* versus *perseverative* piece elaborations.

Alternative Piece Descriptions. Experts typically described pump pieces in terms of simple perceptual characteristics (e.g., "there's a little tiny red piece"). Alternative types of description were classifiable as either "holistic" or "functional." A holistic piece description related a piece to a common object (e.g., "looks like a fat thumbtack"), and a functional piece description specified the piece in relation to the functioning water pump (e.g., "that serves as the washer"). The frequency of alternative descriptors was totaled, and the rate per 100 words of these descriptors was calculated for each transcript.

Repetitions. Experts' spontaneous repetitions of noun, verb, and prepositional phrases were tabulated for each transcript. Only phrases that were longer than three words and that directly repeated a phrase presented earlier in the same segment were scored as repetitions. Noun and verb phrases that were embedded both within a prepositional phrase and within an unembedded context were included as repetitions. The following italicized noun phrases are both repetitions from an audiotape transcript: "The first thing that you do to that blue cap is fit the black circle ring into the base of *that blue cap...*"; "Over that you will fit the clear glass tube, *clear glass tube...*"). Repetition of information directly prompted by the novice during clarification subdialogues was not included in this count of spontaneous repetitions. The total number of repetitions was converted to a rate per 100 words.

Illocutionary Acts. Seven different kinds of illocutionary acts were coded that classified the expert's intention during first reference to each of the pump parts. These illocutionary acts included expert requests that the novice identify, pick up, orient, or attach a particular part, other requests, labeling, and listener confirmations. The methods used to construct this coding scheme have been outlined in detail in Cohen (1984), who presents an analytical approach derived from a plan-based theory of communicative action. Based on these illocutionary act codings, the number of times that each expert began by requesting identification of a new pump part, rather than immediately requesting that it be picked up, oriented, or attached, was totaled. Experts within each modality were classified as habitual users of the *identification request strategy* for introducing part descriptions if they first requested identification for nine or more of all the parts.

Determiners. The number of times that experts first referred to pump parts using definite and indefinite determiners (e.g., "the," "this," "that" versus "a," "an," "another") was totaled for each transcript. Each expert was classified as a habitual user of definite first reference if nine or more first references to a pump part were presented with a definite determiner.

Personal Pronouns. The number of personal pronouns, such as "I," "you," and "we," that experts used while instructing the novice was totaled for each transcript and converted to a rate per 100 words.

Instructional Backtracking. The total number of discourse *backtracks* per transcript was tabulated. Backtracking occurred when experts erred in their ordering of instructions on how to assemble the parts and, consequently, had to give instructions for disassembling and reassembling one or more parts in a different order. In addition to the unique content that identified backtracks, the beginning of backtrack segments always was signaled with a meta-comment about the presence of an error, and with a distinguishing organizational marker.

Speech False Starts. An incomplete noun or verb phrase produced by the expert was counted as a false start if it was followed by a self-correction that was not influenced by a novice interjection. The following is an example of two false starts produced by a telephone expert while attempting to identify a part: "Uh - keep the - Let's start with a piece that has - it's a metal rod." No attempt was made to score hesitations, nonword fillers such as "Uh," single word repetitions, or other forms of speech dysfluency. In no case did a scored false start ever qualify as either a simple repetition or an elaboration. The total number of false starts in experts' speech was recorded for each transcript and converted to a rate per 100 words.

Total Expert Words. For each transcript, the total number of words spoken by experts while transmitting their instructions was scored to the nearest 10 words. In the telephone modality, the total number of words that excluded clarification subdialogues also was computed.

Novice Assembly Time. The number of seconds that each novice spent assembling the water pump, from the beginning of instructions until the conclusion of successful testing, was scored to the nearest second.

Reliability. Second scoring was completed for 20-33% of the data for each reported dependent measure within each mode. 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 ranging over .86.

6 Results

The pattern of characteristics that distinguish telephone and audiotaped speech is presented in the following three sections on discourse macrostructure, referential descriptions, and performance and discourse efficiency. The predominant differences between these two spoken modalities are summarized in Figure 2.

6.1 Discourse Macrostructure

Experts' telephone and audiotape discourses shared many organizational features. Typically, they began with a brief introduction to the goals and nature of the task. This was followed by the main body of the discourse, in which a series of instructions were given for assembling the individual parts of the pump. These seriated assembly instructions were highly redundant in their organization. Within each discourse assembly segment, it was typical for two parts to be described, followed by instructions on how to attach them. Almost without exception, the beginning of a new discourse assembly segment was signaled with an organizational marker, such as "Okay, now." These markers occurred at the beginning of new assembly segments 96.3% of the time in the audiotapes and 98.6% in the telephone transcripts. Occasionally, experts erred in their ordering of how to assemble the parts and had to backtrack in their instructions. Once all twelve parts were assembled correctly, the conclusion of the main task was announced, and a transition was made to providing instructions for testing the water pump. In addition to these general similarities in discourse structure between telephone and audiotape, the average length of experts' instructions was comparable in the two modalities.

6.1.1 Organization of Telephone Dialogues

The interactivity of telephone dialogues created a unique discourse structure with many clarification subdialogues between expert and novice. Novices tended to request clarification when they needed a more refined description of an assembly action, such as the exact location for attaching a piece (e.g., "Which end am I supposed to put it on over?"), the orientation of the pump or its pieces during attachment (e.g., "Which way should I point the spout?"), or a calibration of amount, distance, pressure, and so forth (e.g., "Oh, a little bit. How much you want?"). Clarification subdialogues focused on these action refinements 69% of the time, whereas piece identification only required clarification 21% of the time, and task functions and goals 10% of the time. In this study, telephone interactions averaged 5.4 novice-initiated clarification subdialogues per transcript, 2.8 of which could be characterized as brief (i.e., $\leq$ 5 speaker turns) and 2.6 as lengthy (i.e., $>$ 5 speaker turns). An average of 16% of experts' verbal instructions in the telephone modality were transmitted during these subdialogues, although the need for clarification between different novice-expert pairs was notably variable. While four of the five teams engaged in clarification subdialogues totalling less than 11% of the expert's total instructions, this rate was 47% for the fifth team.

Another unique and prominent aspect of the interactive telephone dialogues was their confirmation structure. Listeners regularly confirmed that instructions were received and

understood clearly enough for their assembly of the pump to be progressing smoothly. The vast majority of all confirmations observed in this task followed descriptions of pieces to be identified, which averaged 43% of the total, and descriptions of assembly actions, which averaged 46% of the total. The following is an example that illustrates a multi-turn confirmation exchange during assembly of a piece:

Expert:	"Take that and stick that in the end of the green part of the plunger. In the bottom of the green part of the plunger."
Novice:	*"Okay."*
Expert:	*"You see it?"*
Novice:	*"Yeah."*
Expert:	*"You got it in?"*
Novice:	*"Yeah."*
Expert:	*"Okay."*

The majority of all confirmations, or 65%, were brief single-turn interjections, such as "Okay." Another 35% were multi-turn confirmation exchanges, such as the six-turn exchange in the above telephone fragment. In these longer exchanges, experts often prompted novices by explicitly asking if they had located or acted on a piece in some way (e.g., "You see it?"), which functioned as a confirmation request. Experts also frequently reaffirmed their own receipt of the novice's original confirmation, thus establishing a confirmation as mutual knowledge. In the present task, approximately 18% of the total verbal interaction between telephone novices and experts was spent eliciting and issuing confirmations, such as the ones italicized above[1], with an average rate of one novice or expert confirmation every 5.6 seconds.

6.1.2 Organization of Audiotape Monologues

It was hypothesized that audiotape experts might attempt to compensate for the lack of interactive feedback by relying more on organizational strategies to further clarify their instructions. With respect to discourse organization, audiotape experts did make more frequent explicit introductions of upcoming actions before they began giving direct instructions about the relevant pieces and actions. These introductions either referred to construction or testing of the whole pump, or to several steps required for the subassembly of, for example, the pump base or spout. A comparison of the two spoken modalities revealed that action introductions were a significantly more common feature of audiotaped than telephone speech, with means of 3.0 and 1.0 per transcript, respectively, $t = 1.91$, $df = 8$, $p < .05$, one-tailed.

In addition, experts in both spoken modalities often paused to provide a summary description of the water pump's current status. A comparison of the rate per 100 words of experts'

[1] Requests for piece identification (e.g., "the smallest of the red pieces?") were often cast as questions, prosodically if not grammatically. These speaker requests for piece identification also appeared to function as requests for confirmation of piece identification, to which the novice responded with a confirmation. However, they were not included in the present tabulation of confirmation interactions.

reviews of piece or assembly descriptions, ones originally described in an earlier assembly segment, revealed that summaries also were significantly more common in the audiotape mode, with mean rates of .47 and .17 per 100 words, respectively, $t = 2.03$, $df = 8$, $p < .04$, one-tailed.

Audiotape experts often produced parallel introductions and summaries of particular subassemblies, as if to provide structural bracketing of a group of assembly steps. The presence of these demarcations around certain subassemblies implies that the speaker perceived these steps to be related and, furthermore, that a hierarchical level of organization was perceived in the task beyond the simple sequence of individual assembly steps.

6.2 Referential Descriptions

Telephone and audiotape experts differed extensively in the way they conveyed the piece and action descriptions that formed the essence of task instructions.

6.2.1 Spontaneous Elaborations and Repetitions

It was hypothesized that audiotape experts might take the conservative approach of providing more extensive descriptions to ensure successful task completion, since step-by-step confirmations from the novice were unavailable. Experts' referential descriptions in this task focused primarily on 1) part descriptions– the individual parts of the water pump, and the pump in partial states of assembly, and 2) action descriptions– the different actions required to assemble the parts and test the pump's final performance. Audiotape experts were found to produce significantly more spontaneous elaborative descriptions of parts and actions, which averaged over 34 per transcript, even though they did not use more words overall than telephone experts to convey their instructions. A comparison of the rate of experts' spontaneous elaborations per 100 words revealed an average of 3.94 in the audiotape modality compared with 2.45 in the telephone modality, $t = 2.74$, $df = 8$, $p < .02$, one-tailed. In both the audiotape and telephone modalities, elaborations of piece descriptions were more prevalent than action elaborations by a margin of almost two to one. Piece elaborations constituted 57.6% of the total in the telephone and 59.9% in the audiotape transcripts.

To rule out the possibility that experts in the audiotape condition were making more frequent but briefer elaborations, an independent assessment was conducted of the average length of experts' elaborated descriptions. Elaborations in the audiotape mode also were established to be significantly longer than those in the telephone mode, with means of 9.64 and 7.93 words, respectively, $t = 2.21$, $df = 8$, $p < .03$, one-tailed.

In addition to more frequent and longer spontaneous elaborations in the audiotape modality, there was a difference between modes in the pattern of ordering elaborative descriptions of pieces and actions. The most common pattern of presentation, and the one that predominated in the telephone modality, was to describe one piece, a second piece, and then the action required to assemble them. Elaborations of the piece descriptions followed first reference to those pieces 93% of the time, and action elaborations followed description of the main assembly action 100% of the time in the telephone transcripts. In the audiotapes, a significant departure from this sequence often occurred in which a piece description continued

to be elaborated even after the main action to assemble the piece had been described. In fact, 26% of piece elaborations in the audiotapes followed description of the main action, whereas only 7% did so in the telephone transcripts. An example of this phenomenon was presented earlier in the Sample Transcripts section, in which the audiotape expert describes the metal rod, describes and elaborates the red piece, explains how to attach these two pieces, but then continues to elaborate the metal rod: "There's a green cap on the end of the silver thing."

This observed average rate per 100 words of piece elaborations *following* instructions to assemble the piece, although still occurring within the same assembly segment, was .86 for the audiotape modality and .08 for telephone, $t = 3.27$, $df = 4.18$ (with separate variance estimates), $p < .02$, one-tailed. By contrast, although 82% of all piece elaborations occurred *before* description of the main assembly action, their average rate per 100 words was 1.50 for the audiotapes and 1.34 for the telephone transcripts, which did not represent a significant difference. To summarize, piece elaborations in the more prototypical location *preceding* the main action description were similar in the two modes, although *perseverative* piece elaborations were substantially more abundant in the audiotape modality.

Regarding qualitative analysis of referential descriptions, the majority of experts' piece descriptions during this task focused on simple perceptual characteristics, although alternative holistic and functional descriptors also were common. Since audiotape experts elaborated much more extensively than telephone experts, an analysis was conducted to evaluate whether their piece elaborations also represented more diversity in the form of alternative descriptors. However, referential description of these types was equally prevalent in the two modalities, with experts averaging 2.06 alternative descriptors per 100 words.

Spontaneous phrase and sentence repetitions also were significantly more common in the audiotape modality. The rate of spontaneous self-repetitions per 100 words averaged .65 in audiotapes, compared with .15 in telephone dialogues, $t = 2.56$, $df = 8$, $p < .02$, one-tailed. Further examination revealed that repetitions were distributed very differently in the two modalities, and appeared to function differently. Approximately two-thirds of all telephone repetitions occurred during clarification subdialogues with the novice, rather than as spontaneous repetitions. In fact, 27% of all telephone repetitions were prompted by and immediately followed a request for clarification by the novice. In addition, 54% of the time that telephone experts repeated themselves, they were summarizing piece or action descriptions. In contrast, all of the audiotape experts' repetitions were spontaneous self-repetitions, and 71% of the time they repeated themselves during assembly segments that were classified as organizationally strained. After correcting for the difference in overall frequency of strained segments between the two modes, the audiotapes were confirmed to harbor a 26% higher concentration of repetitions during strained segments than the telephone modality.

6.2.2 Directness and Determinateness

To investigate whether noninteractive audiotape instructions might display less communicative indirection and fine-grainedness than instructions presented during an interactive telephone dialogue, the communicative acts of experts in both modalities were compared. It was confirmed that audiotape and telephone experts introduced piece descriptions differently.

More specifically, a comparison of first reference to each of the pump pieces revealed that 71% of the time telephone experts requested novices to identify a new piece before telling them how to assemble it. These identification requests in the telephone modality most often were stated indirectly in the form of existential descriptions (e.g., "Okay, now, there's a black O-ring"), although it also was common to use perceptually based descriptions (e.g., "You'll see three very small red pieces of plastic") and fragmentary descriptions (e.g., "Now, the smallest of the red pieces?"). Together, these three types of identification request accounted for 85% of all those occurring in the telephone modality. By contrast, audiotape experts often first referred to a piece by immediately instructing the novice to act on it in some way (e.g., "Take that plunger with the red part in it"). In the audiotapes, only 25% of first references to a piece included a separate request for identification.

A comparison was made of audiotape and telephone experts' tendency to use a separate request for identification when referring to new pump pieces before specifying an action. Based on a criterion for each expert of demonstrating ≥ 9 separate requests for identification during first reference to all the pieces, 0 out of 5 audiotape experts and 4 of 5 telephone experts were classifiable as habitual users of the identification request strategy, a significant departure from chance, $p = .02$, based on Fisher's exact probability test (Siegel, 1956). To summarize, in the telephone modality there was a tendency to decompose the instructions into two parts: identify and act. This was accomplished by emphasizing the identification of pump pieces through separate communicative acts. In this sense, experts engaging in interactive telephone dialogues used a more fine-grained and less direct instructional style than those speaking into an audiotape.

It was assumed that experts would use a mixture of definite and indefinite determiners, since the novelty of the task increased the likelihood of indefinite reference, whereas a few perceptually distinct pump pieces might have been referred to with a definite determiner instead. It also was hypothesized that audiotape experts would use more definite referential descriptions, since they generally speak in a more direct and less fine-grained manner than telephone experts. In fact, audiotape and telephone experts were found to differ in their use of definite and indefinite determiners during first reference to each of the pump pieces. A full analysis of the transcripts confirmed that definite first references occurred 88% of the time in the audiotape instructions, but only 48% of the time in telephone instructions. Based on a criterion of producing ≥ 9 definite determiners during first mention of all the water pump pieces, 5 out of 5 audiotape experts and 1 of 5 telephone experts were classified as habitual users of definite reference. Using Fisher's exact probability test, this modality difference departed significantly from chance, $p = .02$. In short, experts' audiotaped instructions were conveyed using a more definite style — a style that was more assertive in its assumption that the novice would recognize particular pieces in common with the expert.

Although audiotape experts generally introduced new pieces with definite determiners, one of the audiotape experts displayed the odd but habitual style of producing descriptive *reversions*. That is, in his descriptions this expert would initially refer to a piece or salient feature with a definite determiner, but subsequently would revert to an indefinite elaborative excursion about the same piece. These reversions were characterized by downshifting to both

a more indirect illocutionary style as well as a more indefinite description. This pattern was not evident in the telephone transcripts. The following are examples:

Audiotape expert: "...you take *the* L-shaped clear plastic tube,
another tube,
there's *an* L-shaped one with a big base..."

Audiotape expert: "...you are going to insert that into
the long clear tube
with two holes on the side.
Okay.
There's *a* tube about one inch in diameter
and about four inches long.
Two holes on the side."

Personal pronouns are a common referential feature of interactive speech. This high frequency in spoken language has been noted by researchers and construed as an index of involvement (Chafe, 1982; Chapanis et. al., 1977). Although less involvement was assumed to exist in the audiotape mode, examination of the transcripts revealed that the audiotape experts did not produce significantly fewer personal pronouns than the telephone experts in this study. In the telephone transcripts, experts often shifted to using personal pronouns during requests for identification of pieces, during requests for confirmation and confirmations, and during requests for clarification and the ensuing clarification subdialogues — all of which were constructions unique to the interactive telephone modality. Since the audiotape experts' instructional style was more direct and assertive (e.g., "you should fit the peg"), they frequently used personal pronouns throughout the main discourse assembly segments, whereas telephone experts rarely did this. In particular, audiotape experts very often used personal pronouns in the introductory phrase of an assembly segment (e.g., "Okay, now, after you finish that, you will see"), even when pronoun use was suspended later in the segment. In conclusion, there was no evidence that audiotape experts used personal pronouns less often than experts speaking directly and interactively. Nonetheless, the distribution and function of their personal pronouns appeared very different.

6.3 Performance and Discourse Efficiency

All five telephone teams and five of six audiotape teams were able to assemble a working water pump. The data from one audiotape team was discarded due to the novice's inability to decipher the instructions and assemble the pump after 30 minutes had elapsed. Of the five audiotape novices who successfully completed the task, only one person rewound and replayed a couple of brief segments of the audiotape. In short, the task appeared relatively easy to complete, and the discard rate was low.

Dysfluencies occurred in both the sequencing of experts' instructions and in their speech production, although no modality difference was evident in either type of dysfluency. The overall incidence of backtracks was relatively low. Sixty percent of the transcripts contained either one or two incidents of instructional backtracking, with an average of .80 backtracks per transcript. All experts produced false starts, with an average rate of 1 per 73 words. Again, no significant difference was evident in the rate of false starts between telephone and audiotape experts.

Analysis of task performance, as defined by the time required for novices to assemble the water pump, revealed that audiotape novices were significantly less efficient than telephone novices. Novices listening to audiotape instructions took an average of 8 minutes, 50 seconds to assemble the pump, whereas telephone novices required only 6 minutes, 57 seconds on average, $t = 1.87$, $df = 8$, $p < .05$, one-tailed. This difference in novice assembly time occurred in spite of the fact that there was no overall difference in the total number of words that telephone and audiotape experts used to transmit their spoken instructions, which averaged 875 and 845 words, respectively. Even with clarification subdialogues excluded from the telephone transcripts, no significant modality difference was evident in the total number of expert words.

For exploratory purposes, an attempt was made to uncover the discourse factors that might have been associated most closely with the length of novice assembly time. Correlations were evaluated between the amount of time required for audiotape and telephone novices to assemble the water pump and the discourse measures assessed in this study. Significant positive correlations were revealed between novice assembly time and 1) experts' elaborations of piece and action descriptions ($r = +.69$, $df = 8$, $p < .02$, one-tailed), 2) experts' introduction of upcoming actions ($r = +.68$, $df = 8$, $p < .02$, one-tailed), and 3) experts' use of personal pronouns ($r = +.66$, $df = 8$, $p < .04$, two-tailed). That is, in both spoken modalities, experts who elaborated their descriptions most extensively, and whose discourse style entailed frequent use of personal pronouns and advance introductions of upcoming actions, were the ones most likely to be part of a team in which the novice assembly time was lengthy.

7 Discussion

In the present communication task, the main feature that differentiated the two modalities was the opportunity for team members to interact directly. In the telephone mode, participants experienced natural and direct dialogue with one another, including clarification exchanges and immediate access to confirmation feedback. By contrast, audiotape experts related instructions to novices without the benefit of interaction or feedback of any kind. However, other important aspects of the communication task were designed to be comparable for all teams. For instance, all experts transmitted their instructions solely through speech, since there was no visual contact between partners. The instructions were devised to provide participants with similar background knowledge of both their partner and the task. Furthermore, the evaluation of task performance provided a global index of comparability, since it ensured that teams in both modalities were functioning at a level sufficient for comparison

purposes.

The following sections discuss the discourse and performance patterns that were distinctly different in interactive and noninteractive speech and, in a more predictive vein, how human language and performance are likely to be altered during communication with future speech systems that will inevitably be limited in their interactive capabilities. The final section presents a theoretical model of collaborative dialogue that accounts for some of the principal features of interactive and noninteractive spoken discourse. A model of the type proposed eventually will be required to design high quality interactive speech systems.

7.1 Interactive and Noninteractive Spoken Discourse Patterns

The principal distinguishing characteristic of the audiotape experts' monologues was profuse elaborative description. Their descriptive elaborations of pieces and assembly actions, which formed the essence of these task instructions, were both more abundant and longer than those in telephone, even though the total verbal output in the audiotapes was not any greater. In addition, unique elaborative patterns were present in the audiotape transcripts. For example, audiotape experts often persisted in describing a particular piece beyond the basic specification of how to assemble the piece. In fact, they only produced more piece elaborations than telephone experts *after* description of the main assembly action in the discourse segment, not before this point.

By contrast, descriptions in the telephone modality were relatively brief and, in addition, a different illocutionary style emerged for specifying pieces. Telephone experts habitually 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 addition to greater communicative indirection, new pieces were initially described in a more tentative, indefinite manner until recognition was confirmed by the novice. After confirmation, the description usually progressed to an anaphoric definite reference to the same piece. By contrast, audiotape experts were more presumptuous of the novice's recognition of new pieces, as reflected in their initial definite descriptions of them, as well as more assertive about immediately instructing the novice to act on new pieces in a particular way. Another unique pattern in the audiotape modality entailed the description of a new piece in a definite manner, but then reverting to an indefinite elaborative excursion about the same piece. The descriptive progression in these reversions was precisely the opposite of the more prototypical indirect/indefinite $\Longrightarrow$ direct/definite one observed in telephone dialogues since, in the second half of reversions, the audiotape expert also downshifted to an indirect request for piece identification in the form of an existential statement (e.g., "stick the green thing into the non-threaded end of the plastic tube. *There's one end that's threaded and one end that's not...*")

Although the indefinite $\Longrightarrow$ definite descriptive progression in the telephone modality is more common, the audiotape experts' first-mention use of the definite article in reference to new pieces was not inappropriate in the current task context, because reference was made via singular count nouns to unique objects. Furthermore, the listener was being informed

of the location of new objects, which were potentially visible in the immediate environment. Both participants knew that the novice was unfamiliar with the objects, needed to rely on the experts' instructions to identify them, and that the purpose of the interaction was for the expert to instruct the novice on the correct identity of each object to be assembled (see Hawkins, 1974). However, recent theories of definite description do not account for systematic modality differences of the type found in this study, nor do they acknowledge the presence of phenomena like audiotape reversions[2] (Clark & Marshall, 1981; Hawkins, 1974). Most likely, the observed modality difference in definiteness simply corresponds with the directness in experts' style at a pragmatic level. A correspondence between illocutionary directness and the definiteness of determiners also has been noted in the keyboard modality (Oviatt & Cohen, 1989). Both the audiotape and keyboard modalities, which eliminated or reduced the interaction possible between participants, produced emphatic and direct illocutionary styles. In both cases, this style was reflected in the less common first-mention use of the definite article.

In many respects, the audiotape instructions were simply less integrated and predictably sequenced than telephone, which may have reflected the strain of operating without feedback. Of course, the high rate of audiotape elaborations introduced more information for the novice to integrate about a piece. Perseverative piece descriptions also required the novice to integrate information from two separate locations in the discourse. This introduced unpredictability with respect to where piece information was located, as well as violating expectations for the prototypical placement of piece information. In addition, piece descriptions that were perseverative or reverted appeared to be out-of-sequence parenthetical additions that disrupted the smooth continuity of the audiotape discourse. The novice had to disambiguate whether such elaborations referred to a new piece, as implied by indefinite introduction or discontinuity, or whether they were indeed anaphoric. Once established as anaphoric, the novice had to integrate the continued or reverted description with the appropriate earlier one. In addition to these poorly integrated elaborative features, simple repetitions also occurred at a higher rate in the audiotapes, and more often were located in discourse segments classified as strained. Finally, audiotape experts never established the predictable illocutionary framework for introducing new piece descriptions that telephone and keyboard experts did, who habitually relied on indirect identification requests and direct requests for the specific assembly action, respectively (Cohen, 1984). None of the audiotape experts provided a clear illocutionary mold for instructing the novice, as would have been reflected in the establishment of a criterial pattern for piece introduction. All of these characteristics produced more inferential strain in the audiotape modality.

The reinforcement of discourse macrostructure in the audiotape modality, which was evident in the audiotape experts' more frequent introduction of actions and summaries of the water pump's status, may have assisted in offsetting this relative lack of integration and predictability in the audiotapes. At least, the effect of these organizational enhancements was to

[2]The unity and subordination of these indefinite elaborations to the preceding main definite description might well be signaled prosodically. If this is true, then prosodic analysis might serve as a primary vehicle for determining how to integrate these elaborations with the preceding discourse.

focus the more rambling audiotape descriptions, which digressed into elaborative detail and otherwise risked losing the thread of their coherence. In this sense, the lack of interactive feedback in the audiotape modality might have been compensated for somewhat by preparing the novice in advance for certain main points, and by reiterating others to emphasize their importance.

During the assembly task, all telephone teams engaged in frequent confirmations, with a substantial 18% of telephone experts' total verbal output allocated during confirmation exchanges and with a confirmation forthcoming from one of the participants every 5.6 seconds. In a task-oriented dialogue, this continual stream of confirmations is the major vehicle available for the listener to signal to the expert that the expert's communicative goals have been achieved. Furthermore, access to concurrent feedback has been associated with increased dialogue efficiency, at least in the form of reduced noun phrases with repeated reference (Clark & Wilkes-Gibbs, 1986; Isaacs & Clark, 1987; Krauss & Weinheimer, 1964 & 1966). Since audiotape experts had to operate without confirmations from the novice, they essentially had no metric for gauging when to inhibit elaborations, and had to make this decision in an arbitrary manner. Therefore, it was impossible for audiotape experts to tailor a description to meet the information needs of their particular partner most efficiently. In this sense, their extensive and perseverative elaborating was a rather understandable conservative strategy.

By comparison with audiotape, telephone experts' piece and action descriptions have been described as a series of more fine-grained steps. Each step entailed a different communicative goal which, after being satisfied, was discharged by a confirmation from the novice. For example, telephone novices confirmed after a piece description to notify the expert that the piece had been located and, therefore, that the expert's goal of making the piece mutually identifiable could be discharged. This signaling and mutual accommodation of telephone novices and experts during confirmation exchanges most likely contributed to their successful curtailment of elaborative description, which in turn promoted the superior efficiency of the telephone modality. The efficiency advantage of the telephone modality was reflected most clearly in the faster average assembly time of their teams. The strong positive correlation between frequent elaborations and lengthy novice assembly time also highlighted the relative inefficiency of excessive elaboration as a discourse strategy and, by implication, of the audiotape mode as well.

7.2 Implications for Future Speech Technology

The differences between interactive and noninteractive speech have implications for the successful design of future speech technology. This section considers how users' language and performance are likely to be influenced as they communicate with systems designed for dictated spoken input and interactive speech exchanges.

7.2.1 Noninteractive Speech Technology

The augmentation of general computing technology with digital signal processing is yielding a proliferation of noninteractive voice systems, primarily for use in professional settings. One

example of this new class of technology is voice mail (Nicholson, 1985), a spoken analogue of electronic mail systems in which users record a message, transmit it to selected recipients, reply to received messages, and so forth. Other examples include voice annotation of text, and personal workstations that provide several speech capabilities in a hypertext system (Ades & Swinehart, 1986).

The results of the present study suggest that noninteractive spoken input has characteristic disadvantages that may limit the practical uses and efficiency of this class of communication technology. For illustration, compare the following assembly instruction expressed by an audiotape expert with the original training instructions:

Audiotape expert: "Then on the other hole, the one that looks
complicated, you take the last little,
the second little red piece that also looks like
a nail with some prongs on the top of that.
There's another red one slightly bigger than
the first one that you used
and that goes, lies right in the other hole
that you have on the long regular tube..."

Training instructions: "Fit the slide valve loosely into the
lower outlet of the main tube."

The present research suggests that technology based on noninteractive speech, exemplified by the above audiotape instruction, will be prone to excessive elaboration and repetition that slows down and undermines the coherence of messages, contributing to inefficient processing by the recipient. The results also indicate that such communications will tend to be more poorly integrated and less predictably structured, requiring more effort by the recipient or system to resolve their meaning. These adverse characteristics are most certain to surface during hands-on assembly tasks such as the one examined during the present study. This type of task and those requiring visual attention to a display (e.g., geographical location tasks) should be of special interest to future speech system designers, since speech is known to reap a two- to three-fold efficiency advantage over nonspeech modalities in these domains[3] (Chapanis, Parrish, Ochsman, & Weeks, 1977; Oviatt & Cohen, 1989).

One major technical impediment that hinders widespread commercialization of voice systems is the lack of adequate editing facilities beyond rerecording the message or editing the signal itself (Ades & Swinehart, 1986; Milne, 1986). Since techniques do not yet exist for

[3]Future research needs to calibrate the relative efficiency of speech for task domains other than those requiring simultaneous visual attention to a display or manual manipulation of objects.

transforming speech input to a textual display, users have no way to easily identify the contents of an extended speech signal so that segments can be located for editing. Instead, these systems represent speech graphically, typically either as a profile of sound energy plotted against time, or as a series of capillary tubes extended over a time line and shaded to distinguish speech from silence. Markers often are used in an effort to highlight the more significant speech, either in terms of content or recency of editing. The user attempts to locate and replay segments and, if correctly identified, then erases and rerecords them as desired. Short of actually playing back the message itself, however, no graphical representation can convey or enable the user to track the actual content of an extended voice signal.

A breakthrough in speech recognition technology that would enable coupling systems such as voice mail and speech annotation with automatic dictation machines (Jelinek, 1985) is one innovation that could substantially improve the feasibility of these systems by enabling the user to edit written transcriptions. However, for applications like voice mail in which the ultimate goal is spoken presentation, the availability of text editing then creates a problem re-synthesizing the edited transcript into the user's voice with appropriate intonation. When written output is desired, either for automatic dictation machines such as IBM's "talkwriter" (Gould, Conti, & Hovanyecz, 1983; Jelinek, 1985) or as a transitional format for editing dictation, noninteractive speech can be expected to produce the kind of poor copy described earlier. As such, it will require labor-intensive editing, a disadvantage that must be weighed with respect to the initial advantages of spoken input. Although the availability of speech-to-text technology could facilitate editing, it will not lessen the total amount of editing required due to original input through noninteractive speech. In addition, since editing by voice requires noninteractive spoken input just as input of the original material does, the quality of voice edited discourse may itself tend to be elaborative and poorly integrated by comparison with material that has received conventional written editing.

To reduce the described disadvantages of noninteractive speech on professional communications, it may be strategic to limit this voice technology to brief and informal tasks, and to ones that do not emphasize planning, reviewing, or editing during transmission. In fact, this recommendation for brevity accords with the usage patterns and preferences reported in a recent voice mail study (Nicholson, 1985), which indicated a natural tendency for users to limit noninteractive spoken messages to less than one minute in length.[4] Theoretically, a more direct and less restrictive solution may be to design methods for providing confirmation feedback that can effectively inhibit speaker elaborations, along with the resulting discourse convolutions and inefficiency. As argued in the previous discussion section, listener confirmations naturally play a role in reducing referential descriptions, thereby contributing to conversational efficiency. However, pursuing this alternative would require that noninteractive systems be reconceptualized as partially interactive ones, which would need to be capable of some speech and language recognition. Furthermore, research would need to explore how a system could provide confirmations that are effective in guiding the speaker's dialogue, and perceived by the speaker to be sincere and acceptable, while also resulting in improved

[4] In two different samples, Nicholson (1985) found voice mail messages averaging approximately 30 seconds in duration.

descriptive detail and structuring for the average recipient.

So far, primary consideration has been given to implications for the structure and integrity of information transmitted by noninteractive speech systems. However, the social-interactional impact of this technology also must be addressed, especially since its goal is to promote professional communication in office settings. For example, one implication of the habitually directive style of noninteractive speech, which is at odds with the indirection of dialogue, is that users risk creating misleading impressions of their attitude toward the task and message recipient. Some potential users may be reluctant to send dictated voice messages to higher status managers if they have concerns or experience indicating that these recipients are likely to be offended by the more directive, brusk, or impatient tone. Furthermore, users may be uncomfortable leaving content messages for more senior colleagues when they cannot receive confirmation feedback to reassure them that the listener is in agreement, or that he or she wishes to comply with a proposed agenda. In short, when the balance of power is asymetrical between two professionals, the more junior individual is especially likely to find dictation without confirmation guidance undesirable and impractical, both because of its greater potential for conveying damaging misimpressions, and due to the higher probability that the senior professional would expect to initiate and exercise control over decision-making.

In addition to the issue of directiveness, potential users who become aware that they have difficulty presenting well integrated monologues may seek ways to avoid creating the impression that they are disorganized. In cases where this proves difficult, one option is to avoid dictated speech altogether. Due to the need to observe status distinctions and to promote a professional image of oneself in professional settings, speaker concerns over the appropriateness and consequences of using dictated speech for professional purposes are likely to be emphasized. Finally, where misimpressions do occur, communication partners who use noninteractive speech devices have the added disadvantage of being unable to resolve problems immediately.

To the extent that certain types of misimpressions are fostered by noninteractive speech technology, they can be expected to lead to broadly counterproductive consequences, including reduced interaction and cooperation among colleagues. Past research has established that perceptions of social attractiveness, such as perceived friendliness, exert an influence on the overall amount of interaction and degree of cooperation between individuals (Gardin, Kaplan, Firestone, & Cowan, 1973; Wichman, 1970; Williams, 1977). Therefore, social perceptions will need to be managed if noninteractive speech technology is to achieve full use, and if it is to avoid adverse consequences for office dynamics. To cultivate the mutual positive images upon which frequent cooperative exchanges are built, individuals need to interact through affect-rich modalities that carry nonverbal and prosodic cues during direct interaction. Noninteractive speech, which eliminates direct interaction and the bidirectional communication of affect through prosody, is simply less well suited than dialogue for constructing a foundation of positive images.[5] Ultimately, the extensive introduction of noninteractive speech systems

[5]Communication through affect-rich channels emphasizes the affective content of a message and is more likely to further polarize the positivity or negativity of conveyed messages (Gardin et. al., 1973). Under circumstances in which a speaker harbors negative feelings toward the content or recipient of a message, he or

may require some restructuring of office communications to compensate for the reduction of more directly interactive, affect-rich channels.

7.2.2 Interactive Speech Systems

A long-term goal for speech systems is to develop dialogic systems with fully interactive capabilities. Several interactive speech systems currently under development are based on task domains like those of the present study, in which instructions are issued to robots for manipulating or assembling a set of physical objects. Among this class of systems are instructable robots that perform household tasks for the handicapped (Crangle & Suppes, 1987) and robots that assist humans during construction of the Space Station (Coler & Edgerton, 1986). In practice, of course, speech applications currently being developed only are capable of limited interaction. For example, system responses typically are more delayed than the average human conversant.

While the natural speed of human dialogue creates an efficiency advantage in tasks, it simultaneously challenges current computing technology to produce more consistently rapid response times. In research on telephone conversations, transmission and access delays[6] of as little as .25 to 1.8 seconds have been found to disrupt the normal temporal pattern of conversation and to reduce referential efficiency (Krauss & Bricker, 1967; Krauss, Garlock, Bricker, & McMahon, 1977). These data reveal that the threshold for an acceptable time lag can be a very brief interval, and that even these minimal delays can alter the organization and efficiency of spoken discourse.

Preliminary research on human–computer dialogue has indicated that, beyond a certain threshold, language systems slower than real-time will elicit user input that has characteristics in common with noninteractive speech. For example, when system response is slow and prompt confirmations to support user–system interaction are not forthcoming, users will interrupt the system to elaborate and repeat themselves, which ultimately results in a negative appraisal of the system (van Katwijk, van Nes, Bunt, Muller, & Leopold, 1979). For practical purposes, then, people typically are unable to distinguish between a slow response and no response at all, so their strategy for coping with both situations is similar. Unfortunately, since system delays typically vary in length, it seems unrealistic to expect users to learn to anticipate and accommodate a slower dialogue pace.

One way to mitigate the adverse effects of time delay on dialogue is to design a confirmation system capable of informing users when input has been received and is continuing to be processed, as well as assuring them that important propositional content has been interpreted correctly. Although a slow system may never be appreciated, at least effective confirmations

she may wish to avoid the use of an affect-rich modality that risks exposing negative feelings and generating friction between the conversants. In these cases, speakers may gravitate naturally to a noninteractive modality like spoken or written messages in an effort to minimize anticipated conflict and to manage their interactions successfully. Future research should evaluate the possibility of this type of natural usage pattern, as well as evidence of its long-term impact.

[6] A *transmission* delay refers to a relatively pure delay of each speaker's utterances for some defined time period. By contrast, an *access* delay prevents simultaneous speech by the listener, and then delays circuit access for a defined time period after the primary speaker ceases talking.

could facilitate users' accurate interpretation and coordination with such a system. In the case of two native speakers engaged in dialogue, Krauss et. al. (1977) demonstrated that the deleterious effects of a one second delay in response time also could be reversed by providing the speakers with video images of one another. Under these circumstances, visual confirmations in the form of headnodding evidently compensated for the absence of verbal confirmations. These data further underscore empirical support for the importance of confirmation feedback in promoting optimal conversational efficiency. Beyond this, they offer an alternative solution for future interactive speech systems in cases where verbal confirmations are either unavoidably blocked or delayed. At least for applications in which it is feasible, cost-effective, and in which two native speakers are involved, the provision of supplementary visual access appears to offer another potential avenue for managing dialogue inefficiencies.

Apart from system delay, another current limitation that will influence future interactive speech systems is the unavailability of full prosodic analysis. Unlike text, speech is fraught with technically ungrammatical constructions and fragments, the meaning of which often are disambiguated prosodically. Therefore, to process speech accurately, systems will be required to analyze significant prosodic cues. Needless to say, users cannot be expected to recognize which aspects of their intended meaning are being conveyed prosodically, or to find alternative literal means for expressing this information.

Since an interactive system must be able to analyze meaning in order to deliver appropriate and timely confirmations of received messages, limited prosodic analysis may make the design of an effective confirmation system more difficult. In spoken interaction, speakers typically convey requests for confirmation prosodically, and such requests occur mid-sentence as well as at sentence end. For example:

Expert: "Put that on the hole on the side of that tube —"
(pause)
Novice: "Yeah."
Expert: "— that is nearest to the top or nearest to the green handle."
Novice: "Okay."

For a system to analyze and respond to requests for confirmation, it would need to detect rising intonation, pausing, and other characteristics of the speech signal which, although elementary in appearance, cannot yet be performed automatically (Pierrehumbert, 1983; Waibel, 1988). A system also would need to derive the contextually appropriate meaning for a given intonation pattern, by mapping the prosodic structure of an utterance onto a representation of the speaker's intentions at a particular moment. Since the pragmatic analysis of prosody barely has begun (Hirschberg & Pierrehumbert, 1986; Waibel, 1988), this important capability is unlikely to be present in initial versions of interactive speech systems. Therefore, the typical prosodic vehicles that speakers use to request confirmation will remain unanalyzed such that confirmations are likely to be omitted. This may be especially true of mid-sentence

confirmation requests that lack redundant grammatical cues to their function. To the extent that confirmation feedback is omitted, speakers' discourse can be expected to become more elaborative, repetitive, and generally similar to monologue as they engage in dialogue with limited-interaction systems.

If supplying apt and precisely timed confirmations for near-term spoken language systems will be difficult, then consideration is in order of the difficulties posed by noninteractive discourse phenomena for the design of preliminary systems. For one thing, the discourse phenomena of noninteractive speech differ substantially from the keyboard discourse upon which current natural language processing algorithms are based. Old algorithms will require extensive alteration, especially with respect to referential features and discourse macrostructure, if designers expect new speech systems to handle human speech input. The treatment of noun phrases, which is currently a poorly understood area of natural language processing, will require considerable reexamination. With respect to reference resolution, the system will have to identify whether a perseverative elaboration refers to a new versus previously mentioned part, whether the initial descriptive expression is being further expanded, qualified, or corrected, and so forth. The potential difficulty of tracking noun phrases throughout a repetitive and elaborative discourse, especially segments that include perseverative descriptions displaced from one another and definite descriptions that revert to indefinite elaborations about the same part, is illustrated in the following brief monologue segment:

> "and then you take *the L-shaped clear plastic tube*, *another tube*, there's *an L-shaped one with a big base*, and that big base happens to fit over the top of this hole that you just put the red piece on. Okay. So there's one hole with a blue piece and one with a red piece and you take the one with the red piece and put *the L-shaped instrument* on top of this, so that..."

For example, a system must distinguish whether "another tube" is a new tube or whether it co-refers with "the L-shaped clear plastic tube" uttered previously, or with the other two italicized phrases. In cases where description of a part persists beyond that of the basic assembly action, the system also must determine whether a new discourse assembly segment has been initiated and whether a new action now is being described. In the above illustration, the system must determine whether "and you take the one with the red piece and put the L-shaped instrument on top of this" refers to a new action, or whether it refers back to the previously described action, "that big base happens to fit over the top of this hole..." The system's ability to resolve such co-reference relations will determine the accuracy with which it interprets the basic assembly actions underway. To optimize the interpretation of spoken monologues, a system needs to continually reexamine whether further descriptive information supports or refutes current beliefs about part identity and action performance. That is, the system's orientation should be geared more toward frequent cross-checking of previous information, rather than automatically positing new entities and actions.

Data on the organizational properties of noninteractive speech could be used to assist with identifying part descriptions that are new since, for example, they are highly likely to occur following the temporal markers, action introductions, and summary checkpoints that generally

divide assembly segments. Prosodic analysis could enhance the resolution of reference further, since Hirschberg and Pierrehumbert (1986) point out that discourse structure often is signalled by changes in pitch. With more specific relevance to this research, it has been noted that speech elaborations tend to be signalled by a shift to faster tempo and to less variable pitch excursions (Goffman, 1981). Further research on the relation of prosody to the pragmatics of spoken discourse could advance heuristics for the accurate interpretation of both interactive and noninteractive speech.

Current natural language systems typically attempt to determine whether a noun phrase refers to a new or old entity based on its determiner, with new entities most often introduced indefinitely, and subsequent anaphoric reference shifting to a definite determiner or pronoun. However, this pattern was not typical of the noninteractive speech observed in this study, which habitually involved the definite introduction of new parts. Future systems would function more efficiently if heuristics more in accord with noninteractive speech patterns were developed. A relatively novel strategy for handling determiners is used in Tacitus (Hobbs & Martin, 1987). This system begins with the assumption that indefinite noun phrases are referential, although less computational effort is expended in the resolution of indefinite than definite noun phrases. When an indefinite reference is not resolved quickly, the assumption of referentiality is relaxed and a new discourse entity is postulated. This approach may operate quite efficiently for noninteractive speech, especially for phenomena like reversions, but it would be inefficient for interactive speech.

As spoken language systems approach the fully interactive ideal, overcoming temporal and prosodic limitations, users' discourse characteristics can be expected to more closely resemble telephone dialogue. When this occurs, keyboard-based standards and heuristics will require restructuring to accommodate the characteristics of interactive speech summarized in Figure 2. Since interactive speech displays more predictable discourse patterns than noninteractive speech, the heuristics developed for fully interactive systems ultimately may yield more accurate intent recognition than is possible for the limited interaction forerunners.

7.2.3 Automatic Telephone Interpretation of Japanese–English Dialogues

In this section, implications of the present research will be discussed for the development of an interactive speech system that interprets Japanese–English telephone dialogues automatically. The design of such a system is complex at many levels, including the discourse and human interface issues entailed. Separate features each predispose such a system to communication breakdown, including Japanese–English linguistic and cultural differences, the process of interpretation, telephone transmission, and current limitations of computer technology. Some of the various difficulties that this collection of factors pose for users have been described (Oviatt, 1988; see pp. 19-21), along with proposed strategies for their management.

Due to the pressures that an automatic telephone interpretation system can be expected to impose on users, including necessary processing delays,[7] one essential design element is user

[7]The expected delays in automatic telephone interpretation of Japanese and English are considerable since, for example, the usual three-party interpretation delay and the delay required to interpret Japanese final verbs will be compounded by system processing and transmission delays, and so forth.

support in the form of a strong confirmation system. Some of the goals and requirements for provision of adequate confirmation support have been discussed in detail elsewhere (Oviatt, 1988; see pp. 44-47). Assuming the standard conduit model,[8] interpretation conducted by telephone results in a blockage of the speakers' natural system of mutual acknowledgment, since the interpreter provides a literal interpretation of what has just been said for the listener without simultaneously relating appropriately-timed confirmations back to the speaker. Furthermore, any confirmations that may be generated by the listener during the receipt of interpreted material are heard exclusively by the interpreter, and do not constitute feedback for the original speaker. The need for a clear and supportive confirmation system also is compounded by the cultural distance between Japanese and English speakers, and by the necessary processing delays that an automatic system would entail. Any interpretation system not designed to support the speakers' sense of mutual comprehension, given that they are confronted with the temporal obstacles described, will undermine their basic goal of conducting a coordinated, intelligible dialogue.

One immediate impact of weak or unavailable confirmations is that speakers' anxiety about being understood is heightened. This is known to be more of a problem for audio-only communication channels (Weston & Kristen, 1973), and during Japanese–English interpretation (Oviatt, 1988). Furthermore, general speaker anxiety over whether the listener is understanding, along with interruptions and turn-taking dysynchronies, have been observed during research on Japanese–English telephone interpretation (Iida, Kogure, Nogaito, & Aizawa, 1987). The constrained interaction between the primary speakers during this type of communication, particularly the blockage of mutual confirmations, leads us to predict that the speakers' discourse is likely to display noninteractive phenomena such as those outlined in the present study. That is, speakers may be prone to repeating and elaborating themselves, thereby producing a poorly integrated discourse, which in turn presents resolution problems for an automatic system like the ones discussed in the last section. At least in the short run, if constructing a confirmation system to minimize these discourse features proves difficult, then optimal system accuracy and efficiency are most likely to be achieved if the heuristics developed are tailored to accommodate noninteractive speech.

In contrast with the conventional conduit model, the brokered approach provides an avenue for alleviating the described confirmation blockage (see Oviatt, 1988, for detailed discussion). The interpreter who brokers engages in a series of alternating subdialogues with the two primary speakers in their own native language, with a full exchange of confirmations possible between each primary speaker and the interpreter. During these subdialogues, the interpreter is both the initiator and recipient of messages, and he or she plays a more active, responsible role than the conduit approach permits. A brokered system also can potentially reduce communication problems betwen Japanese and English speakers generated by differences in confirmation style. Of course, due to the alternation of brokered subdialogues,

[8]The conventional view of interpretation maintains that the interpreter's role is that of a conduit – a semi-automatic, passive, neutral, essentially powerless intermediary through whom communications are transmitted from the speaker of one language to that of another. A standard literal interpretation is the professed outcome or goal of such a model.

confirmation feedback regarding important propositions still would be delayed between the two primary speakers, just as it is in the conduit approach, and an automatic system still would be subject to response time delays. The net effect of these basic differences between the conduit and brokered approaches is that effective support of speaker confirmations should be easier for a brokered system to develop. Especially after the delays currently typical of interactive speech systems are reduced and stabilized, a brokered system would be less subject to disruption from noninteractive speech characteristics. These considerations suggest that, for systems such as automatic telephone interpretation of Japanese–English that require strong confirmation support, a brokered system may be advantageous.

Video transmission of nonverbal confirmation feedback (e.g., headnodding) would be unlikely to provide a workable alternative to verbal confirmations for an automatic interpretation system, since two nonnative speakers are involved. In fact, it has been argued that video transmission of confirmations between Japanese and English speakers may well confuse rather than reassure speakers (Oviatt, 1988), due to the considerable discrepancy between Japanese and English speakers in the frequency, placement and meaning of nonverbal gestures. If it will be necessary to provide a verbal confirmation system, then the important relations between prosody and pragmatics discussed in the last section will need to be worked out for Japanese as well as English. Finally, although we predict that the basic differences between interactive and noninteractive speech patterns found in this study would generalize to the Japanese population, this assumption requires replication to document the exact form and degree of modality differences in Japanese discourse structure and performance.

7.3 Toward a Model of Collaborative Dialogue

One purpose of this discussion is to begin clarifying the influence of speaker interaction, as it occurs within different modalities, on dialogue collaboration. The development and implementation of better spoken language systems depend on improved modeling of discourse characteristics, including major differences resulting from modality constraints. In particular, it has been argued that for noninteractive and interactive systems slower than real time to succeed and function efficiently, they will require confirmation feedback designed to inhibit excessive elaboration and related discourse inefficiencies.

In this section, an analytical model of dialogue as rational, goal-directed activity is presented, based primarily on the work of Cohen & Levesque (1987; 1989; forthcoming), which accounts for several principal discourse characteristics of interactive and noninteractive speech. This logically oriented model attempts to derive the possible ways for speakers to achieve discourse and task goals although it does not, as in a processing or performance theory, attempt to delineate why a speaker would select a particular alternative for achieving a goal. To guide and refine the construction of a more informative working model, predictions based on the logical approach are joined whenever possible with actual empirical patterns and probabilities established in the present experiment. Theoretical emphasis is placed on the collaborative[9]

[9]The term *collaboration* refers to the underlying goals, commitments, and intentions that dialogue participants maintain in common, as the result of an implied or public contract, and which function to drive the physical interaction. The term *interaction* refers to the actual behavioral exchange between the partici-

nature of dialogue, and on the discourse vehicles through which experts and novices achieve both synchronization of their physical exchange and sufficient mutual knowledge to support communication. The model focuses on explaining[10] the following empirical phenomena observed in telephone and audiotape discourse:

1. the presence of confirmations in telephone
2. the presence of organizational markers in both audiotape and telephone
3. the presence of action introductions in both audiotape and telephone
4. more extensive descriptive elaborations in audiotape than telephone (including elaborative perseverations and reversions)
5. more extensive descriptive summaries in audiotape than telephone

Previous research has analyzed individual communication as rational, plan-based activity, and has advanced the idea that communicative action can be treated as a special case of action more generally (Allen & Perrault, 1980; Appelt, 1985; Cohen & Levesque, 1989; Cohen & Perrault, 1979). The present dialogue model departs from these earlier approaches by squarely addressing the fundamental collaborative nature of dialogue. Recently, several researchers have adopted the view that conversation has a collaborative essence in need of explanation (Clark & Wilkes-Gibbs, 1986; Grosz & Sidner, 1989; Searle, 1989), and that people engaging in a dialogue adopt mutual goals, beliefs, and intentions as they plan and interact during their collaboration. Previous approaches differ in their definition of collaboration, particularly in how they relate joint intentions to the individual intentions that lead each speaker to contribute to the overall dialogue. Furthermore, since these approaches have not considered the fundamental differences that communication modalities impose on a collaboration, they do not predict the discourse or performance features that distinguish interactive and noninteractive spoken discourse.

The present model is based on a formal analysis of joint commitment and joint intention developed by Cohen & Levesque (forthcoming), which is parallel to their independently motivated analysis of individual intention and commitment (Cohen & Levesque, 1987). With respect to their earlier work on individual communicative action, a speaker is assumed to have *achievement goals* to secure desired states that he or she believes do not exist at present.

---

pants, which both enables and reflects their mutual collaboration. As it has been used throughout this paper, interaction includes spoken confirmation feedback and clarification subdialogues between the participants.

[10] In the context of an exclusively logical model, explanation refers to the ability to predict or derive results from a set of definitions and principles, such as Cohen & Levesque's definition of the speakers' mutual goals and beliefs, and joint commitments and intentions, combined with their principles of rational interaction (Cohen & Levesque, 1987 & 1989). The listed empirical phenomena are consistent with and should follow logically from their theory of collaborative communication, for reasons to be outlined in this section. Prediction of points 3 and 4 is accomplished, in part, through the acknowledgment of relevant empirical patterns and the incorporation of their likelihood estimates. Any mathematical derivations of these phenomena remain to be anchored in future work.

For example, an expert might have the goal of getting the novice to identify the black ring or attach it to the rest of the pump assembly. A speaker is said to be *committed* to achieving a particular goal if the goal persists until it is either achieved satisfactorally, judged to be impossible to achieve, or believed to be unnecessary due, for example, to changes in supporting beliefs or superordinate intentions. Essentially, a speaker's internal commitments are goals that are maintained over time, resisting capricious abandonment. Based on this concept of commitment, an individual *intention* to perform a particular action is a commitment to entering a future state wherein the individual will believe the action is imminent just before performing it. This latter belief approximates the concept of a present-directed intention (Bratman, 1987; Searle, 1983), i. e., an action one intends to perform now. A future-directed intention to perform an action, then, is a commitment to performing the action intentionally or with a present-directed intention. As these defined features operate within this logical model, intention inherits many of its interesting properties from the persistent nature of internal commitments. One such property is that intentions generally will lead an individual to plan and act again if earlier attempts to achieve a goal are thwarted.

Cohen & Levesque's theoretical approach to collaborative dialogue, which is a special case of their approach to collaborative action more generally, is that speakers engaged in conversation have joint intentions based on joint commitments. The definitions of both terms are parallel to those of individual intentions and commitments, but with mutual belief replacing individual belief. Therefore, they assert that dialogue participants are *jointly committed* to a particular achievement goal, such as attaching the black ring correctly, if they maintain the mutual belief that they share the same achievement goal until they arrive at the mutual belief that the goal has been achieved, is impossible, or is now irrelevant. Likewise, dialogue participants *jointly intend* to perform a particular action if they are jointly committed to entering a future state wherein they will share the mutual belief that their action is imminent just before they perform it. Figure 3 summarizes these distinctions between the principal components defined in Cohen & Levesque's theory of *individual* versus *collaborative* action.

The present model of collaborative action assumes that dialogue partners will engage in a rational progression of steps toward their mutual goals, given the particular task and modality context within which they are operating. As a collaborative task, dialogue entails a complex sequence of steps in which both speakers participate and share responsibility for success. For example, to model the present experimental task as a collaborative activity, the following is proposed:

> For each part of the pump, the participants jointly intend that the expert will request the novice to perform some assembly action on that part, and then the novice will do whatever has been requested.[11]

[11]The term request is construed pragmatically to include a broad range of phenomena such as indirect requests and, in the case of face-to-face interaction, nonverbal actions. Theoretically, the model assumes that the novice can potentially infer the next assembly action, not only from an indirect or partial request, but also in advance of any explicit request at all. For purposes of simplicity, however, the present summary does not present the model's descriptive subtleties on this point. With respect to the present experiment, it actually was very rare for a novice to take the initiative to infer the next assembly action without first being requested to do so.

When two participants, X and Y, jointly intend to act in sequence, as in this case where X requests Y to do assembly action A, and then Y performs A, it follows that X individually intends to request Y to do action A, and Y intends to perform A once that request is made clear. Given the assumption of joint intentions, not only are both participants committed to fulfilling their own roles in the collaborative activity, they also are implicitly committed to each other's actions. In this sense, each individual wants the other to complete his or her prescribed part, and both individuals are disturbed when a breakdown occurs because they have adopted the mutual goals entailed in their joint commitment to achieving the task. This leads the participants to support each other's roles during the ensuing exchange, and to coordinate their individual activities as they act together. Finally, to synchronize their interaction successfully, the participants must mutually recognize when their parts are to be initiated and terminated, which is accomplished through communicative acts.

7.3.1 Collaboration in an Interactive Modality

One of the most frequent and stable characteristics of the interactive telephone modality in this study was the exchange of confirmation feedback between the novice and expert, with one confirmation forthcoming approximately every 5.6 seconds. Behavioral evidence of the role that confirmations play in promoting the efficiency of collaborative dialogues has been discussed earlier in this paper. Within the context of the present model, the definition of joint commitment states that one condition for terminating the participants' joint commitment to a particular task goal is that they come to mutually believe that the goal has, in fact, been achieved satisfactorally. Note that the expert and novice must *mutually* believe that the achievement goal (i.e., in this task, the identification or assembly of a piece) has been satisfied before their commitment to this goal is considered to have been achieved. The present model predicts that the novice will have to inform the expert verbally when an identification or assembly action has been completed successfully, since mutual vision is precluded in both the telephone and audiotape modalities. Of course, the selection of verbal confirmation is only an available strategy in the telephone condition, since the audiotapes were specifically designed to prohibit interaction. As reported in the Results section, verbal confirmations of successful piece identification and assembly actions occur continually throughout the task in the telephone dialogues. In fact, they account for 89% of all confirmations. To summarize, the model predicts that a verbal confirmation from the telephone novice will be used to signal to the expert after each of their mutual goals has been achieved and that, therefore, their joint commitment to the particular goal can be discharged at this point. In this sense, verbal confirmations are a predictable strategy, given the participants' original joint intentions and commitments.

As discussed earlier, two speakers who jointly intend to collaborate on a dialogue must have some means of synchronizing their actions. Within the framework of the model, it is specified that the participants' joint intention to perform an action leads them to be committed to entering a future state in which they will mutually recognize that their collaborative action is about to occur, immediately after which they will perform it. This definition entails both mutual foreknowledge that the collaborative action will occur and that its timing will be

mutually recognized immediately before it occurs. While confirmations function to support the mutual belief that the prior assembly step was completed successfully, the experts' discourse strategy of uttering temporally-oriented organizational markers, such as "Okay, next," signals in a manner that establishes as mutual knowledge the participants' readiness to initiate their next collaborative step. It is in this sense that the participants' joint intention to achieve successful completion of each assembly action also predicts the experts' habitual use of organizational markers to signal the beginning of each new discourse assembly segment. Action introductions would be predicted by the model to function similarly to organizational markers, with the exception that they signal the initiation of a collection of actions required to construct a subassembly. That is, they simply reflect a different and more hierarchically organized series of subtask goals and actions.

7.3.2 Collaboration in a Noninteractive Modality

At first glance, the concept of speaker collaboration in a noninteractive modality may seem inherently odd. From a computer science perspective, it has been demonstrated that in circumstances in which two agents are unable to synchronize their "clocks," a condition true of the noninteractive audiotapes, mutual beliefs cannot be established and, therefore, joint commitment and joint intention become unattainable in a technical sense (Halpern & Moses, 1985). However, in this research, both parties explicitly agreed to perform the task so that, according to the present model of joint intention, their public contract ensured collaboration. All participants agreed in advance to the instructions, including their role and modality assignment. Furthermore, everyone operated under the implied expectation that it was possible to assemble the water pump correctly within the arranged teams and proposed experimental framework. In short, the participants' advance public agreement clarifies the grounds for their collaboration, including the adoption of joint intentions and commitments by the audiotape participants.

To further support their performance, it appears that audiotape experts imagined a mute listener with whom they shared a commitment to accomplish the task.[12] That is, audiotape experts were able to ignore the time delay and act as though their taped discourse was being listened to at present, even though they knew this was not the case. Evidence for audiotape experts' fabrication of a real-time listener, perhaps as an aid in composing instructions, is present in the following excerpts:

> Audiotape expert: Now, you'll put the green part inside the the a–
> hole where there, the hole without the two ridges.
> Okay, *You got that? Good.*

[12] Positing a phantom intermediary or substitute partner also has been used as a technique in artificial intelligence research for implementing mutual belief between two parties. Following this concept, the intermediary's beliefs are only those mutually held by the two agents (Appelt, 1985; Konolige, 1986; McCarthy, Sato, Hayashi & Igarashi, 1977).

Audiotape expert: and you take the red nozzle looking thing with the hole in it– *You know what a nozzle looks like? No problem.*

The audiotape experts' instructions occasionally include such examples of "residual" requests for confirmation, as well as subsequent comments typical of an interactive response, in spite of the absence of any dialogue partner.[13] Once a mute listener is posited for audiotape experts, their use of organizational markers and action introductions no longer seems unusual. The fictitious partner essentially provides them with a "placeholder" for what would have been a more natural and familiar collaborative interaction. It also presents a partner with whom synchronization is required, just as it is for telephone experts. This analysis of the audiotape modality means that it was viewed, at a subjective level, as more equivalent to interactive telephone than it is in actuality.

When a preferred discourse strategy for achieving a goal is not available in one modality, then the model attempts to predict the consequences or rational alternatives for securing the goal. In the audiotapes, unlike telephone dialogue, verbal confirmations were unavailable as a strategy for establishing mutual recognition of goal satisfaction. According to the present model, lack of mutual knowledge of goal satisfaction leads to the prediction that both participants will persist in attempting to achieve the goal, because of the nature of joint intentions. By contrast with audiotape, telephone novices very rarely failed to provide confirmation. However, when a lack of confirmation did occur in the telephone modality, resolution was attempted through one of the three alternative means listed below. Note that only the third alternative was available to audiotape experts:

1. the novice would initiate a clarification subdialogue with the expert, in lieu of providing a confirmation
2. the novice would remain silent, and the expert would *request* confirmation from the novice, either directly or indirectly
3. the novice would remain silent, and the expert would spontaneously provide further description assumed to be needed by the novice

The following combination of a model feature (step 1) with empirical information (steps 2 and 3) converges on the prediction that elaborations should be more prevalent in the audiotape instructions:

[13]The use of an imaginary listener by audiotape experts may account for the equally high rate of personal pronouns observed in the audiotape and telephone modes. That is, actual opportunities for direct interaction within the modality may be less important than the experts' subjective sense of immediacy and involvement with a partner in establishing their frequent use of personal pronouns.

1. given that the participants share joint intentions and commitments, they will persist in their efforts to achieve mutual recognition of goal satisfaction after an initial failure
2. the likelihood of an absent verbal confirmation, which would generate the need for an alternative strategy to signal and discharge goal achievement, rises from a very small percentage (i.e., < 10%) in telephone to 100% in audiotape
3. spontaneous elaborative description is the only alternative available to audiotape experts of the three discourse strategies observed in the relatively unconstrained telephone modality (i.e., alternatives reduced 3:1)

The above steps reveal a much greater need for the audiotape expert to locate alternative strategies (step 2), but with fewer alternatives actually available (step 3), which jointly exert pressure to elaborate in the audiotape modality. Estimating from the combined probabilities expressed in steps 2 and 3, the modality difference in elaborations between audiotape and telephone should be a substantial one. In the context of the present model, perseverations and reversions represent two unique elaborative strategies used by experts during continued efforts to secure an adequate description. As types of elaboration, their increased likelihood in the audiotapes would be predicted to correspond with that of the larger class of elaborations. Finally, the greater prevalence of summary descriptions in the audiotapes, which often terminate hierarchically-structured subassembly units, is predicted in a similar manner to the higher rate of descriptive elaborations. That is, both elaborated original descriptions within an assembly segment and descriptive summaries later in the discourse provide further information about pieces and actions that is assumed to be needed by the novice to achieve discourse goals. With confirmations blocked in audiotape, such that goal confirmation cannot be discharged easily, it is predictable that audiotape experts will persist in spontaneously elaborating information at many levels — including both original and summarized descriptions of information throughout the discourse. Figure 4 summarizes several spoken discourse phenomena that have been outlined during this study, along with their corresponding model explanations.

One purpose of this research has been to begin building a model that can predict the predominant characteristics of different speech modalities, information that will be required to design future speech systems that are habitable, high quality, and relatively enduring. In this pursuit, the proposed logical model will need to be extended and enriched to the point where previously unnoticed discourse phenomena become predictable, and their functions more apparent. Of course, a logical approach in isolation is relatively limited, since competence theories cannot reveal and describe individual processing strategies that are not known, nor can they predict the likelihood that important phenomena will occur under varying circumstances of interest. For this reason, analytical approaches representing both the empirical and logical traditions must begin to be synthesized if we are to achieve a comprehensive and unified treatment of spoken discourse. Through the convergence of these complementary viewpoints, we leverage a deeper understanding of the major factors involved in spoken discourse, their functions, their interplay, and the natural powers of speech that await clever technological application.

8 Acknowledgments

We gratefully acknowledge the valuable assistance of Robert Tierney, Debbie Winograd, Larrie Shirey, and Julie Burke in conducting the experiment, Zoltan Uzhelyi for assistance with video and audiotaping, Scott Fertig and Kathleen Starr for coding transcripts, and Marion Hazen, Elsie Chappell, Joan Hirschkorn, Cindy Hunt, Mike Nivens, Ken Olum, and Norma Peterson for text and transcript preparation. Special thanks also to Hector Levesque for invaluable discussions and manuscript comments.

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*, pages 13–28, American Voice I/O Society, Alexandria, Virginia, September 1986.

[2] J. F. Allen and C. R. Perrault. Analyzing intention in dialogues. *Artificial Intelligence*, 15(3):143–178, 1980.

[3] D. Appelt. *Planning English Sentences.* Cambridge University Press, Cambridge, U. K., 1985.

[4] J. Barwise. Three views of common knowledge. In M. Vardi, editor, *Proceedings of the Second Conference on Reasoning about Knowledge*, Morgan Kaufman Publishers, Inc., Los Altos, California, March 1988.

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

[6] M. Bratman. *Intentions, Plans, and Practical Reason.* Harvard University Press, 1987.

[7] 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, pages 35–53, Ablex Publishing Corporation, Norwood, N. J., 1982.

[8] 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.

[9] H. H. Clark and C. Marshall. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*, Academic Press, New York, 1981.

[10] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.

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

[12] P. R. Cohen and H. J. Levesque. On acting together: Joint intentions for intelligent agents. In preparation.

[13] P. R. Cohen and H. J. Levesque. *Persistence, Intention, and Commitment.* Technical Report 415, Artificial Intelligence Center, SRI International, Menlo Park, California, February 1987. Also appears in *Proceedings of the 1986 Timberline Workshop on Planning and Practical Reasoning*, Morgan Kaufman Publishers, Inc. Los Altos, California.

[14] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, M.I.T. Press, Cambridge, Massachusetts, 1989.

[15] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, M.I.T. Press, Cambridge, Massachusetts, 1989.

[16] P. R. Cohen and C. R. Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979.

[17] H. Gardin, K. J. Kaplan, J. J. Firestone, and G. A. Cowan. Proxemic effects on cooperation, attitude, and approach-avoidance in a prisoner's dilemma game. *Journal of Personality and Social Psychology*, 27(1):13–18, 1973.

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

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

[20] 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.

[21] 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.

[22] B. Grosz and C. Sidner. Plans for discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack (eds.), editors, *Intentions in Communication*, M.I.T. Press, Cambridge, Massachusetts, 1989.

[23] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July-September 1986.

[24] J. Y. Halpern and Y. O. Moses. A guide to the modal logics of knowledge and belief. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, IJCAI, Los Angeles, California, August 1985.

[25] J. A. Hawkins. *Definiteness and Indefiniteness.* PhD thesis, Cambridge University, Cambridge, U. K., 1974.

[26] G. G. Hendrix and B. A. Walter. The intelligent assistant. *Byte*, 251–258, December 1987.

[27] D. Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st. Annual Meeting of the Association for Computational Linguistics*, pages 123–128, Cambridge, Massachusetts, June 1983.

[28] J. Hirschberg and J. Pierrehumbert. The intonational structuring of discourse. In *Proceedings of the 24th Annual meeting of the Association for Computational Linguistics*, New York, June 1986.

[29] J. R. Hobbs and P. Martin. Local pragmatics. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Morgan Kaufman Publishers, Inc., Los Altos, California, August 1987.

[30] H. Iida, K. Kogure, I. Nogaito, and T. Aizawa. *Analysis of telephone conversations through an interpreter.* Technical Report TR-1-002, ATR Interpreting Telephony Laboratories, Twin 21 Bldg. MID Tower, 2-1-61 Shiromi, Higashi-ku, Osaka 540 Japan, May 1987.

[31] E. A. Isaacs and H. H. Clark. References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37, 1987.

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

[33] K. Konolige. *A Deduction Model of Belief.* Pitman Publishing, Ltd., London, U. K., 1986.

[34] R. M. Krauss and P. D. Bricker. Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America*, 41(2):286–292, 1967.

[35] R. M. Krauss, C. M. Garlock, P. D. Bricker, and L. E. McMahon. The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, 35(7):523–529, 1977.

[36] 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.

[37] 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.

[38] R. M. Krauss and S. Weinheimer. The effect of feedback on changes in reference phrases. 1964. Conference presentation, Psychonomic Society, Niagara Falls, Ontario, Canada.

[39] R. M. Krauss and S. Weinheimer. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, 6:359–363, 1967.

[40] J. McCarthy, M. Sato, T. Hayashi, and S. Igarashi. *On The Model Theory of Knowledge.* Report No. STAN-CS-78-657 (AIM-312), Computer Science Department, Stanford University, Stanford, California, April 1978. Also in the *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Cambridge, Massachusetts, August 1977.

[41] S. H. Milne. Integrating voice and text mail. In *Proceedings of AVIOS '86: Voice I/O Systems Applications Conference*, pages 1–12, American Voice I/O Society, Alexandria, Virginia, September 1986.

[42] 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.

[43] S. L. Oviatt. *Management of Miscommunications: Toward a System for Automatic Telephone Interpretation of Japanese-English Dialogues.* Technical Report 438, Artificial Intelligence Center, SRI International, Menlo Park, California, May 1988.

[44] S. L. Oviatt and P. R. Cohen. The contributing influence of speech and interaction on human discourse patterns. In J. W. Sullivan and S. W. Tyler, editors, *Architectures for Intelligent Interfaces: Elements and Prototypes*, Addison-Wesley Publishing Co., Menlo Park, California, 1989.

[45] J. Pierrehumbert. Automatic recognition of intonation patterns. In *Proceedings of the 21st Annual Meeting*, pages 85–90, Association for Computational Linguistics, Cambridge, Massachusetts, June 1983.

[46] K. M. Potosnak and F. L. van Nes. *Effects of replacing text with speech output in an electronic mail application.* Technical Report 19, IPO, N. V. Phillips,, Eindhoven, Netherlands, 1984.

[47] R. Reichman. *Plain-speaking: A theory and grammar of spontaneous discourse.* PhD thesis, Department of Computer Science, Harvard University, Cambridge, Massachusetts, 1981.

[48] J. R. Searle. Collective intentionality. In P. R. Cohen, J. Morgan, and M. E. Pollack (eds.), editors, *Intentions in Communication*, M.I.T. Press, Cambridge, Massachusetts, 1989.

[49] J. R. Searle. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press, New York City, New York, 1983.

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

[51] D. Small and L. Weldon. An experimental comparison of natural and structured query languages. *Human Factors*, 25:253–263, 1983.

[52] 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.

[53] J. A. Turner, M. Jarke, E. A. Stohr, Y. Vassiliou, and N. White. Using restricted natural language for data retrieval: A plan for field evaluation. In Y. Vassiliou, editor, *Human Factors and Interactive Computer systems*, chapter 8, pages 163–190, Ablex Publishing Corp., Norwood, N. J., 1984.

[54] 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.

[55] A. Waibel. *Prosody and Speech Recognition.* Pitman Publishing, Ltd., London, U. K., 1988.

[56] J. R. Weston and C. Kristen. *Teleconferencing: A comparison of Attitudes and Interpersonal Atmosphere in Mediated and Face-to-face Group Interaction.* Technical Report 1, Department of Communications, Ottawa, Canada, 1973.

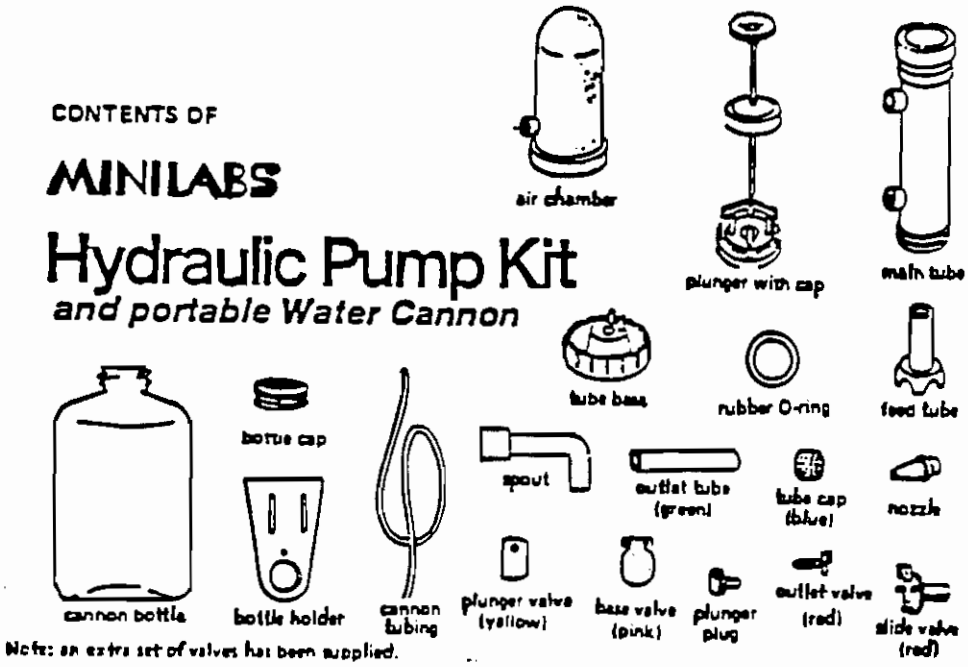
[57] H. Wichman. Effects of isolation and communication on cooperation in a two-person game. *Journal of Personality and Social Psychology*, 16:114–120, 1970.

[58] E. Williams. Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin*, 84(5):963–976, 1977.

[59] C. E. Wulfman, E. A. Isaacs, B. L. Webber, and L. M. Fagan. Integration discontinuity: Interfacing users and systems. In J. W. Sullivan and S. W. Tyler, editors, *Architectures for Intelligent Interfaces: Elements and Prototypes*, Addison-Wesley Publishing Co., Menlo Park, California, 1989.

Figure 1

Parts Diagram of the Water Pump



Instructions for Building a Water Pump

Building a Water Pump

1. Plug the hole in the bottom of the plunger with the plunger plug.

2. Insert the plunger into the main tube. The red handle of the plunger should extend from the non-threaded end of the main tube.

3. Press the blue tube cap down onto the main tube so it fits firmly.

4. Drop the O-ring into the tube base.

5. Fit the pink base valve onto the top of the tube base. The valve should cover the hole in the base.

6. Fit the feed tube onto the bottom of the base.

7. Screw the tube base onto the main tube.

8. Put the tube cap over the upper outlet of the main tube.

9. Fit the slide valve loosely into the lower outlet of the main tube.

10. Put the spout onto the lower outlet of the main tube. The opening of the spout should be pointing away from the tube base.

11. Put the nozzle onto the outlet of the air chamber.

12. Fit the air chamber onto the spout. The nozzle should point away from the main tube.

Using the Water Pump:

1. Place the pump into a tray of water. The pump should be supported by the feed tube.

2. Move the plunger up and down by alternately pushing and pulling on the red handle.

3. Water will be forced out the lower outlet of the main tube, through the spout, through the air chamber, and out through the nozzle.

4. Water will continue to be forced out the nozzle as long as you keep moving the plunger up and down.

5. If nothing happens, check to see that all parts fit tightly and that the valves are properly sealed.

Figure 2

Summary of Major Differences Between Audiotape and Telephone Discourse

Macrostructure	Audiotape	Telephone
Introduction of Actions	+	-
Summary Descriptions	+	-
Clarification Subdialogues		**
Confirmations		**
Reference		
Elaborations		
Number	+	-
Length	+	-
Perseveration on Pieces	+	-
Repetitions	+	-
Separate Requests for ID of New Pieces	-	+
Definite Reference to New Pieces	+	-
Reversions (Definite to Indefinite)	*	
Efficiency		
Speed of Novice Assembly Time	-	+

+ and - specify a greater or lesser amount of a feature.

* and ** designate a sometimes present or characteristic feature.

Figure 3

Principal Definitions
in Cohen /Levesque's Model of
Individual and Collaborative Action

Individual Belief

A belief held by an individual

Mutual Belief

A belief held by two individuals
that they both believe is held by the other,
with their beliefs about each other's beliefs
being infinitely recursive. For more details,
see (Barwise, 1988; Cohen & Levesque, 1989).

Individual Commitment

An individual is committed to achieving
a goal if he or she maintains the goal
until it is either:
- satisfied,
- no longer believed to be achievable,
- or is believed to be irrelevant due to
changes in supporting beliefs,
superordinate intentions, and the like.

Joint Commitment

Two individuals are jointly committed to
achieving a goal if they mutually believe
they share the goal until it is either:
- mutually believed to be satisfied,
- mutually believed to be no longer achievable,
- or is mutually believed to be irrelevant due to
changes in supporting beliefs,
superordinate intentions, and the like.

Individual Intention

An individual's intention to perform
an action is a commitment to entering
a future state wherein the individual
believes the action is imminent
just before performing it.

Joint Intention

Two individuals' joint intention to perform
an action is a joint commitment to entering
a future state wherein the individuals
mutually believe that their action is imminent
just before they perform it.

Figure 4

Theoretical Explanation of Spoken Discourse Characteristics

Discourse Phenomena	*Model Explanation*
1. **Organizational markers** - present in both modalities	Achieves mutual belief that participants are about to initiate a collaborative action, with expert signalling action initiation within an individual assembly segment.
Action introductions - present in both modalities	Same, except expert signals action initiation for a hierarchical grouping of segments
2. **Confirmations** - present in telephone	Achieves mutual belief of goal satisfaction required for discharge of joint commitments and intentions, with novice signalling discharge of mutual goals within individual assembly segments.
Action summaries - present in both modalities	Same, except expert signals discharge of mutual goals for a hierarchical grouping of segments.
3. **Elaborations and summary descriptions** - more extensive in audiotape	Speaker's commitment to dialogue goal persists since there is no route for achieving mutual belief of goal satisfaction