THE EFFECTS OF INTERACTION ON SPOKEN DISCOURSE

Sharon L. Oviatt
Philip R. Cohen
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, California 94025-3493

ABSTRACT

Near-term spoken language systems will likely be limited in their interactive capabilities. To design them, we shall need to model how the presence or absence of speaker interaction influences spoken discourse patterns in different types of tasks. In this research, a comprehensive examination is provided of the discourse structure and performance efficiency of both interactive and noninteractive spontaneous speech in a seriated assembly task. More specifically, telephone dialogues and audiotape monologues are compared, which represent opposites in terms of the opportunity for confirmation feedback and clarification subdialogues. Keyboard communication patterns, upon which most natural language heuristics and algorithms have been based, also are contrasted with patterns observed in the two speech modalities. Finally, implications are discussed for the design of near-term limited-interaction spoken language systems.

INTRODUCTION

Many basic issues need to be addresssed before technology will be able to leverage successfully from the natural advantages of speech. First, spoken interfaces will need to be structured to reflect the realities of speech instead of text. Historically, language norms have been based on written modalities, even though spoken and written communication differ in major ways (Chafe, 1982; Chapanis, Parrish, Ochsman, & Weeks, 1977). Furthermore, it has become clear that the algorithms and heuristics needed to design spoken language systems will be different from those required for keyboard systems (Cohen, 1984; Hindle, 1983; Oviatt & Cohen, 1988 & 1989; Ward, 1989). Among other things, speech understanding systems tend to have considerable difficulty with the indirection, confirmations and reaffirmations, nonword fillers, false starts and overall wordiness of human speech (van Katwijk, van Nes, Bunt, Muller & Leopold, 1979). To date, however, research has not yet provided accurate models of spoken language to serve as a basis for designing future spoken language systems.

People experience speech as a very rapid, direct, and tightly interactive communication modality, one that is governed by an array of conversational rules and is rewarding in its social effectiveness. Although a fully interactive exchange that includes confirmatory feedback and clarification subdialogues is the prototypical or natural form of speech, near-term spoken language systems are likely to provide only limited interactive capabilities. For example, lack of adequate confirmatory feedback, variable delays in interactive processing, and limited prosodic analysis all can be expected to constrain interactions with initial systems. Other speech technology, such as voice mail and automatic dictation devices (Gould, Conti & Hovanyecz, 1983; Jelinek, 1985), is designed specifically for noninteractive speech input. Therefore, to the extent that interactive and noninteractive spoken language differ, future SLSs may require tailoring to handle phenomena typical of noninteractive speech. That is, at least for the near term, the goal of designing SLSs based on models of fully interactive dialogue may be inappropriate. Instead, building accurate speech models for SLSs may depend on an examination of the discourse and performance characteristics of both interactive and noninteractive spoken language in different types of tasks.

Unfortunately, little is known about how the opportunity for interactive feedback actually influences a spoken discourse. To begin examining the influence of speaker interaction, the present research aimed to investigate the main distinctions between interactive and noninteractive speech in a hands-on assembly task. More specifically, it explored the discourse and performance features of telephone dialogues and audiotape monologues, which represent opposites on the spectrum of speaker interaction. Since keyboard is the modality upon which most current natural language heuristics and algorithms are based, the discourse and performance patterns observed in the two speech modalities also were contrasted with those of interactive keyboard. Modality comparisons were performed for teams in which an expert instructed a novice on how to assemble a hydraulic water pump. A hands-on assembly task was selected since it has been conjectured that speech may demonstrate a special efficiency advantage for this type of task.

One purpose of this research was to provide a comprehensive analysis of differences between the interactive and noninteractive speech modalities in discourse structure, referential characteristics, and performance efficiency. Of these, the present paper will focus on the predominant referential differences between the two speech modes. A fuller treatment of modality distinctions is provided elsewhere (Oviatt & Cohen, 1988). Another goal involved outlining patterns in common between the two speech modalities that differed from keyboard. A further objective was to consider the implications of any observed contrasts among these modalities for the design of prospective speech systems that are habitable, high quality, and relatively enduring. Since future SLSs will depend in part on adequate models of spoken discourse, a final goal of this research was to begin constructing a theoretical model from which several principal features of interactive and noninteractive speech could be derived. For a discussion of the theoretical model, which is beyond the scope of the present research summary, see Oviatt & Cohen (1988).

METHOD

The data upon which the present manuscript is based were originally collected as part of a larger study on modality differences in task-oriented communication. This project collected extensive audio and videotape data on the communicative exchanges and task assembly in five different modalities. It has provided the basis for a previous research report (Cohen, 1984) that compared communicative indirection and illocutionary style in the keyboard and telephone conditions. As indicated above, the present research focused on a comprehensive assessment of the discourse and performance features of speech. More specifically, it compares noninteractive audiotape and interactive telephone.

Thirty subjects, fifteen experts and fifteen novices, were included in the analysis for the present study. The fifteen novices were randomly assigned to experts to form a total of fifteen expert-novice pairs. For five of the pairs, the expert related instructions by telephone and an interactive dialogue ensued as the pump was assembled. For another five pairs, the expert's spontaneous spoken instructions were recorded by audiotape, and the novice later assembled the pump as he or she listened to the taped monologue. In this condition, there was no opportunity for the audiotape speakers and listeners to confirm their understanding as the task progressed, or to engage in clarification subdialogues with one another. For the last five pairs, the expert typed instructions on a keyboard, and a typed interactive exchange then took place between the participants on linked CRTs. All three communication modalities involved spatial displacement of the participants, and participation in the noninteractive audiotape mode also was disjoint temporally. The fifteen pairs of participants were randomly assigned to the telephone, audiotape, and keyboard conditions.

Each expert participated in the experiment on two consecutive days, the first for training and the second for instructing the novice partner. 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, along with a diagram of the pump's labeled parts. Approximately twenty minutes was permitted for the expert to practice putting the pump together using these materials, after which the expert practiced administering the instructions to a research assistant. During the second session, the expert was informed of a modality assignment. Then the expert was asked to explain the task to a novice partner, and to make sure that the partner built the pump so that it would function correctly when completed. The novice received similar instructions regarding the purpose of the experiment, and was supplied with all of the pump parts and a tray of water for testing.

Written transcriptions were available as a hard copy of the keyboard exchanges, and were composed from audio-cassette recordings of the monologues and coordinated dialogues, the latter of which had been synchronized onto one audio channel. Signal distortion was not measured for the two speech modalities, although no subjects reported difficulty with inaudible or unintelligible instructions, and < 0.2% or 1 in 500 of the recorded words were undecipherable to the transcriber and experimenter. All dependent measures described in this research had interrater reliabilities ranging above .86, and all discourse and performance differences reported among the modalities were statistically significant based on either *apriori t* or Fisher's exact probability tests (Siegel, 1956).

RESULTS AND DISCUSSION

Compared to interactive telephone dialogues and keyboard exchanges, the principal referential distinction of the noninteractive monologues was profuse elaborative description. Audiotape experts' elaborations of piece and action descriptions, which formed the essence of these task instructions, were significantly more frequent, as well as averaging significantly longer. In addition, repetitions were significantly more common in the audiotape modality, in comparison with interactive telephone and keyboard. Although noninteractive speech was more elaborated and repetitive than interactive speech, these two speech modes did not differ in the total number of words used to convey instructions.

Noninteractive monologues also displayed a number of unusual elaborative patterns. In the telephone modality, the prototypical pattern of presentation involved describing one pump piece, a second piece, and then the action required to assemble them. In contrast, an initial audiotape piece description often continued to be elaborated even after the expert had described the main action for assembling the piece. The following two examples illustrate this audiotape pattern of *perseverative* piece description:

> "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 the 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."*

> "...Now, the curved tube that you just
> put in that should be pointing up still
> Take that, uh — Take the the cylinder that's
> left over — it's the biggest piece that's left over —
> and place that on top of that, fit that into
> that curved tube that you just put on.
> *This piece that I'm talking about is has*
> *a blue base on it and it's a round tube..."*

These piece elaborations that *followed* the main assembly action were significantly more common in the audiotape modality. However, the frequency of piece elaborations in the more prototypical location *preceding* specification of the action did not differ significantly between the audiotape and telephone modes.

Another phenomenon observed in noninteractive audiotape discourse that did not occur at all in interactive speech or keyboard was *elaborative reversion.* Audiotape experts habitually used a direct and definite style when instructing novices on the assembly of pump pieces. For example, they used significantly more definite determiners during first reference to new pump pieces (88% in audiotape, compared with 48% in telephone). However, after initially introducing a piece in a definite and direct manner, in some cases there was downshifting to an indefinite and indirect elaboration of the same piece. All cases of reverted elaborations were presented as existential statements, in which part or all of the same phrase used to describe the piece was presented again with an indefinite determiner. The following are two examples of audiotape reversions:

> "...You take *the* L-shaped clear plastic tube,
> *another* tube, there's *an* L-shaped one
> with a big base..."

> "...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."

These reversions gave the impression of being out-of-sequence parenthetical additions which, together with other audiotape dysfluencies like perseverative piece descriptions, tended to disrupt the flow of noninteractive spoken discourse. Partly due to phenomena such as these, the referential descriptions provided during audiotaped speech simply were less well integrated and predictably sequenced than descriptions in telephone dialogue. To begin with, the high rate of audiotape elaborations introduced more information for the novice to integrate about a piece. In addition, perseverative piece descriptions required the novice to integrate information from two separate locations in the discourse. As such, they created unpredictability with respect to where piece information was located, and violated expectations for the prototypical placement of piece information. In the case of both perseverative and reverted piece elaborations, the novice had to decide whether the reference was anaphoric, or whether a new piece was being referred to, since these elaborations were either discontinuous from the initial piece description or began with an indefinite article. Once established as anaphoric, the novice then had to successfully integrate the continued or reverted description with the appropriate earlier one. For example, did it refine or correct the earlier description? All of these characteristics produced more inferential strain in the audiotape modality.

An evaluation of total assembly time indicated that the audiotape novices functioned significantly less efficiently than telephone novices. Furthermore, the length of novice assembly time demonstrated a strong positive correlation with the frequency of expert elaborations, implicating the inefficiency of this particular discourse feature. Evidently, experts who elaborated their descriptions most extensively were the ones most likely to be part of a team in which novice assembly time was lengthy.

The different patterns observed between interactive and noninteractive speech may be driven by the presence or absence of confirmation feedback. The literature indicates that access to confirmation feedback is associated with increased dialogue efficiency in the form of shorter noun phrases with repeated reference (Krauss & Weinheimer, 1966). During the present hands-on assembly interactions, all interactive telephone teams produced a high and stable rate of confirmations, with 18% of the total verbal interaction spent eliciting and issuing confirmations, and a confirmation forthcoming every 5.6 seconds. Confirmations were clearly a major vehicle available for the telephone listener to signal to the expert that the expert's communicative goals had been achieved and could now be discharged. Since audiotape experts had to operate without confirmation feedback from the novice, they had no metric for gauging when to finish a description and inhibit their elaborations. Therefore, it was not possible 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 an understandably conservative strategy.

In spite of the fact that instructions in the two speech modalities were almost three-fold wordier than keyboard, novices who received spoken instructions nonetheless averaged pump assembly times that were three times faster than keyboard novices (cf. Chapanis, Parrish, Ochsman, & Weeks, 1977). These data confirm that speech interfaces may be a particularly apt choice for use with hands-on assembly tasks, as well as providing some calibration of the overall efficiency advantage. For a more detailed account of the similarities and differences between the keyboard and speech modalities, see Oviatt & Cohen (1989).

IMPLICATIONS FOR INTERACTIVE SPOKEN LANGUAGE SYSTEMS[1]

A long-term goal for many spoken language systems is the development of fully interactive capabilities. In practice, of course, speech applications currently being developed are ill equipped to handle spontaneous human speech, and are only capable of interactive dialogue in a very limited sense. One example of an interactional limitation is the fact that 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[2] 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, their duration is not predictable from the user's viewpoint. Under these circumstances, it seems unrealistic to expect that users will learn to anticipate and accommodate the new dialogue pace as if it had been reduced by some constant amount.

Apart from system delay, another current limitation that will influence future interactive speech systems is the unavailability of full prosodic analysis. Since an interactive system must be able to analyze prosodic 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:

[1] For a discussion of the implications of this research for noninteractive speech technology, see Oviatt & Cohen (1988).

[2] 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.

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 in a reliable manner 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 (Pierrehumbert & Hirschberg, 1989; 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 attempt to 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. Keyboard-based algorithms will require alteration, especially with respect to referential features and discourse macrostructure, if designers expect future systems to handle spontaneous human speech input. With respect to reference resolution, the system will have to identify whether a perseverative elaboration refers to a new part or a previously mentioned one, 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 in "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 will have 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.

In order to see how current algorithms will need to be altered to process noninteractive speech phenomena, we consider how recent dialogue and text processing systems would fare if confronted with such data. The ability to recognize when and how utterances elaborate upon previous discourse is a special case of recognizing how speakers intend discourse segments to be related. The ARGOT dialogue system (Litman & Allen, 1989) takes one important step toward recognizing discourse structures by distinguishing the speaker's *domain* plan, such as for assembling parts, from his or her *discourse* plan, such as to clarify which domain plans are being performed. Although there are technical difficulties, its "identify parameter" discourse plan is designed to process elaborations that further specify the arguments of requested actions during interactive dialogue. However, ARGOT would have to be extended to include a number of new types of discourse plans before it would be able to analyze noninteractive speech phenomena correctly. For one thing, ARGOT does not distinguish different types of elaboration such that information in the two segments of discourse could be integrated correctly. Also, instead of having a discourse plan for self-correction, ARGOT focuses exclusively on a strategy for correcting *other* agents' plans by means of requesting them to perform remedial actions. In addition, ARGOT's current processing scheme is not geared to handle elaborative requests. Briefly, ARGOT performs an action once a sufficiently precise request to perform that action has been recognized. However, since monologue speakers tend to persist in attempting to achieve their goals, they essentially issue multiple requests for the listener to perform a particular action. For example, in the above audiotape fragment, the speaker tried twice to get the listener to put the L-shaped piece over the outlet containing the red valve. Any system unable to recognize that the second request is an elaboration of the first would likely make the fundamental error of positing the existence of two separate actions to be performed.

Although text processing systems are explicitly designed to analyze noninteractive discourse, they fail to provide the needed solutions for analyzing noninteractive speech. These systems currently have no means for identifying basic discourse elaborations and, to date, they have not incorporated discourse structural cues which could be helpful in signaling the relationship of discourse segments (Grosz & Sidner, 1986; Litman & Allen, 1989; Oviatt & Cohen, 1989; Reichman, 1978). In addition, they are restricted to declarative sentences.

One recent text analysis system called Tacitus (Hobbs, Stickel, Martin & Edwards, 1988) appears uniquely capable of handling some of the elaborative phenomena found in our corpus. In selecting the best analysis of a text, Tacitus uses an abductive strategy to search for an interpretation that minimizes the overall cost of the set of assumptions needed to prove that the text is true. The interpretive cost is a weighted function of the individual costs of the assumptions needed to derive that interpretation. Depending on the assignment of costs, it is possible for Tacitus to adopt a non-minimal individual assumption as part of a globally optimal discourse interpretation. Applying this general strategy to noun phrase interpretation, Tacitus' heuristics for referring expressions include a higher cost for assuming that a definite noun phrase refers to a new discourse entity than to a previously introduced one, as well as a higher cost for assuming that an indefinite noun phrase refers to a previously introduced entity than to a new one. These heuristics could handle the prevalent noninteractive speech phenomenon of definite first reference to new pump parts, as well as elaborative reversions, although both would entail higher-cost individual assumptions. That is, if it makes the most global sense, the system could interpret definite first references and reversions as referring to "new" and "old" entities, respectively, contrary to the usual preferences in computational linguistics.

Although such an interpretation strategy may sometimes be sufficient to establish the needed co-reference relations in elaborative discourses, due to the nature of Tacitus' global optimization approach one cannot be certain that any particular case of elaboration will be resolved correctly without first weighing all other local discourse specifics. It is neither clear what percentage of the phenomena would be handled correctly at present, nor whether Tacitus' heuristics could be extended to arrive at consistently correct interpretations. Furthermore, since Tacitus' usual strategy for determining what should be proven is simply to conjoin the meaning representations of two utterances, it would fail to provide correct interpretations for certain types of elaborations, such as corrections in which the latter description supercedes an earlier one. Hobbs (1979) has recognized and attempted to define elaboration as a coherence relation in previous work, and is currently refining Tacitus' computational methods in a manner that may yield improvements in the processing of elaborations.

CONCLUSIONS

In summary, the present results imply that near-term spoken language systems that are unable to provide meaningful and timely confirmations may not be able to curtail speakers' elaborations effectively, or the related discourse convolutions typical of noninteractive speech. Current dialogue and text processing systems are not prepared to handle this type of elaborative discourse. Clearly, new heuristics will need to be developed to accomodate speakers who try more than once to achieve their communicative goals, in the process using multiple utterances and varied speech acts. Under these circumstances, models of noninteractive speech may provide a more appropriate basis for designing near-term spoken language systems than either keyboard models or models of fully interactive dialogue.

To model discourse accurately for interactive SLSs, further research will be needed to establish the generality of these noninteractive speech phenomena across different tasks and applications, and to determine whether speakers can be trained to alter these patterns. In addition, research also will be needed on the extent to which human-computer task-oriented speech differs from that between humans. At present, there is no well developed discourse theory of human-machine communication, and the few studies comparing human-machine with human-human communication have focused on the keyboard modality, with the exception of Hauptmann & Rudnicky (1988). These studies also have relied exclusively on the Wizard of Oz paradigm, although this technique entails unavoidable feedback delays due to the inherent deception, and it was never intended to simulate the interactional coverage of any particular system. Further work ideally would examine human-computer speech patterns as prototypes of interactive SLSs become available.

In short, our present research findings imply that designers of future spoken language systems should be vigilant to the possibility that their selected application may elicit noninteractive speech phenomena, and that these patterns may have adverse consequences for the technology proposed. By anticipating or at least recognizing when they occur, designers will be better prepared to develop speech systems based on accurate discourse models, as well as ones that are viable ergonomically.

ACKNOWLEDGMENTS

This research was supported 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 by a contract from ATR International to SRI International.

References

[1] Chapanis A., 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, 1977.

[2] 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 Corp., Norwood, New Jersey, 1982.

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

[4] 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.

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

[6] A. G. Hauptmann and A. I. Rudnicky. Talking to computers: An empirical investigation. *International Journal of Man-Machine Studies*, 28:583–604, 1988.

[7] 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.

[8] J. Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979.

[9] J. R. Hobbs, M. Stickel, P. Martin, and D. Edwards. Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 95–103, Buffalo, New York, 1988.

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

[11] R. M. Krauss and 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.

[12] 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.

[13] 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.

[14] D. J. Litman and J. F. Allen. Discourse processing and commonsense plans. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* M.I.T. Press, Cambridge, Massachusetts, 1989.

[15] S. L. Oviatt and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. Technical Report 454, Artificial Intelligence Center, SRI International, Menlo Park, California, 1988.

[16] 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.

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

[18] J. Pierrehumbert and J. Hirschberg. The meaning of intonational contours in the interpretatation of discourse. In *Intentions in Communication.* Bradford Books, M.I.T. Press, Cambridge, Massachusetts, 1989.

[19] R. Reichman. Conversational coherency. *Cognitive Science*, 2(4):283–328, 1978.

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

[21] 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*, Eindhoven, Netherlands, 14:105–112, 1979.

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

[23] W. Ward. Understanding spontaneous speech. In *Proceedings of the Darpa Speech and Natural Language Workshop*, February 1989, Morgan Kaufman Publishers, Inc., Los Altos, California.