OSU Working Papers in Linguistics 44 (1-49)

Discourse functions of pitch range in spontaneous and read speech*

Gayle M. Ayers
ayers@ling.ohio-state.edu

Abstract: Functions of intonation and pitch range were compared in matched spontaneous and read speech discourses. Two casual conversations were recorded, and the same speakers read scripts prepared from the original conversations. Sections with one primary speaker were examined. An intonational analysis showed that the locations of accents, phrase boundaries, and pauses differed between the spontaneous and read versions. A discourse segmentation determined that the topic structures were also different, although less so for the second conversation and its read version. Measures of pause and segment durations (as a reflection of speech rate) were made and related to the discourse segmentation units of sentence and paragraph, as well as to turn structure classifications of possible turn, 'rush through', and holding the floor. Since pitch range plays an important role in conveying the hierarchical segmentation of discourse, generally being expanded at the beginning of new topics, corresponding differences in pitch range relationships were expected. Pitch range relationships were represented in phonetic pitch trees based on phrasal peaks. These trees revealed that in addition to signaling topic structure, pitch range was also expanded for corrections and turn taking cues. In spontaneous speech, corrections and turn management disrupted pitch range cues to topic structure. However, the read versions lacked these disruptions, and the pitch range relationships reflected the topic structure more clearly. In a listening test, significantly more read utterances were misperceived as spontaneous in the conversation which had closely matching topic structures in the two versions.

1. Introduction

This study has two main starting points. The first is the distinction between spontaneous speech and read speech, and the second is the role of pitch range in signaling discourse structure. Spontaneous speech and read speech are generally taken to be two quite different modes of speech production and easily distinguishable from one another (Gårding, 1967; Shockey, 1974; Brown et al., 1980; Levin et al., 1982; Remez et al., 1985; Remez et al., 1986; Howell and Kadi-Hanifi, 1991; Blaauw, 1991; Blaauw, 1992). Most detailed prosodic studies of speech have been of speech read in the laboratory or newscast readings. If it is true

*Acknowledgments: Earlier versions of this work were presented at the November 1991 ASA and the January 1992 LSA conferences. The work was supported in part by a National Science Foundation Graduate Fellowship. Thanks to Jennifer Venditti for the discourse segmentation. This paper has benefited from comments made by various people, including Mary Beckman, Julie Boland, Craige Roberts, Julia Hirschberg, Bob Ladd, and participants in the phonetics seminar at the Department of Linguistics and Phonetics, Lund University, Sweden.

that spontaneous and read speech are so easily distinguishable, we can ask ourselves how much of what we learn from studying read speech holds true in spontaneous speech. However, it may be possible for prepared materials to be read in a more spontaneous sounding style and not be so easily distinguishable from true spontaneous speech. Such materials would have the advantage of having controlled content matter typical of read materials, but would be more like natural spontaneous speech than stereotypical read speech. This study matches spontaneous speech materials with read materials based on the spontaneous conversations and read with the aim of sounding spontaneous. I wanted to find out whether the read materials were perceived as spontaneous or read and then compare the two versions, paying particular attention to the role pitch range in signaling discourse structure in the two versions.

There are a few characteristic differences between the traditional classification of spontaneous and read speech. Read speech generally has more complex syntax than spontaneous speech because it is based on written prose. It has fewer hesitations and shorter pauses than spontaneous speech (Gårding, 1967; Brown et al., 1980). Furthermore, the distribution of pauses is different. Pauses in read speech generally align with grammatical phrases and punctuation such as periods and commas (Lehiste, 1975; Brown et al., 1980). Pauses in spontaneous speech may also lie at grammatical boundaries, but they often appear in conjunction with hesitations in the middle of syntactic constituents as the speaker searches for what to say (Gårding, 1967; Butterworth, 1975). In explicit comparisons of matched spontaneous and read speech (where the read text is based on a spontaneous discourse instead of a written text and so is not completely prototypical read speech) the read versions exhibit fewer pauses than the original spontaneous versions, and the pauses are not put in the same locations (Gårding, 1967; Shockey, 1974; Howell and Kadi-Hanifi, 1991). Howell and Kadi-Hanifi also found that readers put stresses and boundaries between tone units in different positions when reading the texts that they had produced spontaneously.

Previous studies have found that listeners are very good at correctly identifying prototypical examples of spontaneous and read speech. Levin et al. (1982) found that listeners could tell an average of 84% of the time whether an utterance was from a spontaneous story or a reading of a story. Even when the speech was low-pass filtered and none of the words were recognizable, they were identified 72% correctly. Informal classification of the differences between the two types of stories were listed as hesitations, long pauses, and non-literary words in the spontaneously told stories. Other studies have found that original spontaneous utterances and matched read productions can be distinguished even if they do not contain hesitations or lexical differences (Remez et al., 1985; Remez et al., 1986; Blaauw, 1991; Blaauw, 1992). Blaauw (1991, 1992) found that listeners could correctly identify a Dutch news reader's spontaneous answer to personal interview questions and his reading from a transcript of the interview 82% of the time when given the full sentence, and even as well as 75% of the time given just the first six syllables of a sentence. She found that the spontaneous versions had lower average F0 and less overall F0 variation, which is in direct contrast to what Remez and his colleagues found for their American English sample. These studies have found that no single acoustic aspect of the signal conveys the spontaneity reliably. If there are not simple acoustic correlates of spontaneous and read speech, perhaps a phonological analysis coupled with a pragmatic analysis will shed more light on the differences between spontaneous and read speech. It is unlikely however that such an analysis will help to explain the high accuracy that listeners have when listening to even small bits of speech or segmentally altered speech.

In spontaneous speech, the principle of turn taking is influential in the production of the talk, even if one person does most of the talking. According to Schegloff (1982:73), speech should be viewed as an interactional achievement.

> The accomplishment or achievement is an interactional one. ... The production of a spate of talk by one speaker is something which involves collaboration with the other parties present, and that collaboration is interactive in character, and interlaced throughout the discourse, that is, it is an ongoing accomplishment, rather than a pact signed at the beginning, after which the discourse is produced entirely as a matter of individual effort.

Put simply, spontaneous speech is not a monologue, even if one speaker does most of the talking. A person speaks with an audience in mind, and interacts with that audience. Schegloff discusses several ways that a single speaker can end up doing most of the talking. One way is that the speaker may actively try to forestall interruption using what he calls 'rush through'. A speaker approaching a possible turn completion speeds up the pace of the talk, withholds a dropping pitch or the intake of breath, and phrases the talk to bridge what would otherwise be the juncture at the end of a unit. The speaker instead breaks in the middle of the next unit. A second way is that the speaker might be forced to continue talking because no coparticipant starts a next turn at an appropriate change of floor. This frequently shows up in the form of a slight gap of silence at the possible turn completion. False starts are also common here since the speaker takes the responsibility to continue talking even though not originally intending to. A third way is that the speaker might continue talking because the conversational partner actively passes an opportunity to produce a full turn of talk, such as by uttering a backchannel continuer (*hmm*, *yeah*, etc.). Spontaneous speech is interactive, and the interaction is both in speaking and listening for all participants. Clark and Schaefer (1989) describe the interactive character of speech in terms of presentations of utterances by speakers and acknowledgments of utterances by listeners. They say that 'a presentation is more than the uttering of a sentence. It is the reaction in real time of a spoken structure from which the partner can identify the words, phrases, and sentences that the contributor intended as final.' False starts, hesitations, and other disfluencies come because the speaker is preparing what to say 'on-line' and reacting to the full situation, including the other conversational participants. In summary, in spontaneous speech there is constant effort put into deciding what to say, making an opportunity to say it, and making sure that it is understood.

Read speech, however, does not involve such a complex communicative process. Prototypical read speech, which comes from a reading of a prepared written text, does not require the reader to figure out what to say next, because the text provides that. In read speech, turns are predefined by the blocks of text on the page and speakers always know when they are to speak. Blocks of text announce the turn structure, so they are not turns in the original sense. In effect (in contrast to what Schegloff says of spontaneous speech), in read speech a pact has been signed and the discourse is produced by individual effort so it is not a true interactional achievement. Even if the read speech is in the form of a multispeaker dialogue based on a spontaneous conversation, it develops more as a series of monologues instead of as a true dialogue like the original. The parts of the telling of the 'story' in the read speech seem to follow one another rather than being sensitive to the spontaneous context in which it was originally produced. The success or failure of recreating the illusion of a spontaneous conversation depends upon the quality of the acting of the readers and their ability to provide a simulation of a natural context. Read speech based on a spontaneous conversation can be seen as having a layer of complexity subtracted away from the original spontaneous conversation because the content is already prepared and the readers do not have to

create it on-line, and furthermore the readers know what order they will speak in so they do not have to negotiate for their turn.

The second large issue explored in this study is the way that discourse structure is signaled. This issue is independent of the distinction between spontaneous and read speech. People organize what they say in terms of relationships of phrases and sentences into larger units, no matter whether they are speaking spontaneously or reading a text. This organization of phrases and sentences into larger units is called the discourse structure. Several different things have been shown to help signal discourse structure. Pauses at the end of sentences and longer pauses at the end of paragraphs have been found in read speech, dividing the speech stream up into units of various sizes and with different groupings (Lehiste, 1979; Brown, 1983; Silverman, 1987; Passenout and Litman, 1993). Speaking rate has also been shown to be related to topic units. Words at the beginning of topics are spoken more slowly and words at the end of topic units are spoken more quickly (Butterworth, 1975; Lehiste, 1980). However, this finding is contradicted by a finding that segment beginnings are faster as compared to segment endings (Grosz and Hirschberg, 1992). Amplitude also relates to topic units. Words at the beginning of topics are louder than words at the end of topics (Brown, 1983).

In addition to these temporal and amplitude cues, pitch range also plays an important role in conveying the hierarchical segmentation of discourse. This is the cue that I will be focusing on primarily, although I will also look at measures of pause durations and speech rate. Pitch range is expanded at the beginning of a new topic (Lehiste, 1975; Butterworth, 1975; Schegloff, 1979; Brazil et al., 1980; Brown, 1983) and compressed to varying degrees at the ends of phrases to reflect the degree of finality of an utterance (Hirschberg and Pierrehumbert, 1986; Silverman, 1987). Cooper and Paccia-Cooper (1980) found that boundary strengths can be reflected by height of F0 targets in the vicinity of the boundaries. Hirschberg and Pierrehumbert (1986) followed up on these observations of the way pitch range cues discourse structure in work with speech synthesis. They found that by systematically varying pitch range of phrases and pause lengths between segments they could signal various hierarchical relationships of topic and subtopic structure. Each discourse segment boundary was marked by a variation in pitch range which correlated with the segment's position in the overall discourse. Grosz and Hirschberg (1992) found in AP news stories that topic segment endings could be identified by relatively long following pauses, and segment beginnings could be identified by larger pitch range, shorter following pause, and by being louder and faster as compared to segment endings. Recall that this tempo cue was in contrast to earlier findings by Lehiste and Butterworth. My data given in Section 4 seemed to support the finding that segment beginnings are faster than segment endings. However, pitch range is not only implicated in signaling topic structure relationships. French and Local (1986) observe that pitch range is used in managing turn taking. They found the prosodic cues of interruptive turn taking to be high pitch and high intensity. Expanded pitch range also marks items as salient, things to pay attention to. Thus, it is also relevant to what Grosz and Sidner (1986) refer to as attentional structure -- what to successively pay attention to over time.

The present study compared how pitch range and intonational structure were used in matched spontaneous speech and read speech discourses. Two different two-speaker conversations were recorded, transcripts of the conversations were prepared, and the scripts were later read by the original speakers. The readers were instructed to read the scripts as if they were involved in a spontaneous conversation. Thus the read speech examined was not prototypical read speech, since it was specifically intended to be a simulation of spontaneous speech. I was interested in finding out how much of the illusion of spontaneity and interactivity

could be created in a read version of a spontaneous conversation. That is, how spontaneous could a read version of a spontaneous conversation sound. The read speech was a reorganization of the spontaneous speech (since it came from the same text), one which was free of the complexity of floor negotiations (since the turns were predefined) and false starts (since those were removed in the preparation of the text). I expected to find that the pitch range cues to topic structure in the read speech versions conformed fairly well to what has been suggested, i.e. that pitch range is expanded at the beginning of new topics and decreases for related subtopics, but that pitch range conveyed the hierarchical discourse structure of the spontaneous speech versions less well. I expected to find that real-time production phenomena such as floor negotiations, corrections, and false starts disrupt clear topic organization. These may complicate the role pitch range plays as a cue to topic structure, because they themselves may have manifestations in the pitch ranges used.

To test these hypotheses, independent discourse segmentations were made of both the spontaneous and read versions for each speaker. The segmentations were not strongly based on a specific theory of discourse, but they were based on the ideas of major topic breaks, turns, and corrections, which were given operational definitions. Pause durations were measured, and a measure was made of speaking rate. These temporal measures were compared with the discourse segmentations and related to previous findings. To see whether the difference in interactivity between a natural spontaneous text and a rehearsed read text could be captured by a symbolic prosodic analysis and an analysis of pitch range, I made a symbolic intonational analysis of the texts, which identified phrases and accented words. Some intonational indications of the interactive character of the spontaneous texts which were not present in the read texts are discussed in Section 5.3. From this prosodic analysis I took a measure of pitch range for each phrase, the peak fundamental frequency occurring on an accented word. Hierarchical pitch trees were constructed from these values, and the segmentations imposed by the pitch trees were compared with the discourse segmentation and events labeled. A perception test with the task of categorizing utterances as spontaneous or read was also carried out to see how "read" the read speech was. The results of the listening test are presented in Section 8.

2. Speech Material

The spontaneous speech used in this study was elicited by recording two separate casual conversations between friends. Each conversation lasted approximately 45 minutes. Both conversations were recorded in a soundproof room with a stereo microphone oriented to concentrate the two speakers' productions on different recording channels. I was a participant in each of the conversations. The first conversation was with FP, and the second conversation was with DW. All three speakers are native speakers of American English. Both FP and DW are male. Even though the conversations took place in a soundproof room, the conversations were very natural. There were no tasks to perform or restricted topic domains; the speakers just spoke about whatever they wished to talk about with each other. The conversations were as close to natural spontaneous speech as they could be, given that the participants knew they were being recorded to provide material for some sort of linguistic investigation. The sections that were chosen for analysis were late in the session and thus past any initial awkwardness or unnaturalness that may have arisen from the studio setting.

The read speech used in the investigation was based on parts of the spontaneous conversations. Approximately seven minutes of each of the original spontaneous conversations were selected to be produced as read speech. I transcribed these selections orthographically, and with the help of a colleague,

edited the transcripts to remove disfluencies such as false starts and pause filling hesitations. These editing decisions were made from the orthographic transcription alone, without direct reference to the audio recording. FP, DW, and I each punctuated our own parts in the edited transcripts. This method of editing and assigning punctuation allowed the maximum opportunity for topic reorganization between the spontaneous and read versions since the groupings of words into phrases, sentences, and paragraphs were determined from the written word and not as a direct simulation of the spontaneous version. Allowing for the possibility of topic reorganization was important because one of the aims of the investigation was to explore the hypothesis that read versions had clearer manifestations of topic structure than the spontaneous versions. The readings were made from clean copies of the scripts which included the punctuation. We studied the scripts and read through them together before the actual recording, so the readings were well rehearsed. We tried to make the readings sound like spontaneous conversations, as if we were acting. None of the readers were trained actors. The read versions were approximately five minutes long.

Most of the decisions of what to remove from the orthographic transcription of the spontaneous speech were quite straightforward, but some of the choices made using this method did not reflect the original intentions of the speakers. Specifically, some false starts were not accurately edited. Consider the examples given in (1). Example (1a) is the orthographic transcription of part of FP's original spontaneous conversation, and example (1b) is the read production. (Pause lengths are also included in these examples, although they were not in the original orthographic transcription. They are shown in milliseconds between angle brackets.)

(1) a. mm but I <64> had I mean the stuff he knows <583> is kind of amazing 'cause <1137> he does a lot of uh environmental impact stuff <694>

b. but I mean the stuff he knows is kind of amazing because he does a lot of environmental impact stuff <454>

The edited orthographic transcription did not reflect that FP stopped after the word *'cause* and started over again with a new sentence *he does a lot of environmental impact stuff.* A more accurate edited and punctuated transcription would have been *The stuff he knows is kind of amazing. He does a lot of environmental impact stuff.* The punctuation in the read version made the phrase after *because* into a subordinate clause, which differs from the structure of the spontaneous production. There were also a few quiet backchannel listening or agreement noises such as *mm-hmm* and *hmm* which were not noted in the original orthographic transcription from which the read script was prepared. This omission of the listener's comments was unintentional and changed the character of the read text. Schegloff (1982:74) comments on the way omissions like these can affect a text. He says that when the behavior of the listeners are separated from the telling of a story, then the parts of the telling seem to follow each other instead of being a response to the behavior of the listeners. Thus the interactivity of the original conversation is destroyed. Since these listener responses (including eye contact, etc., in addition to backchanneling responses) are not present in a subsequent reading of the conversation, the interactive nature of the original conversation is necessarily lost to a certain extent.

The analysis concentrated on sections where FP and DW were the primary speakers in the spontaneous speech and the matching read speech. These sections

were each approximately one minute long. There were two reasons that I chose to concentrate on primarily single speaker sections. The first reason was that when a single speaker talks for a period of time, there is a chance for a topic to develop and be structured by pitch range changes. The second reason was that in sections with one primary speaker, the influence of explicit turn taking is minimized. Short turns and quick interchanges between speakers mix topic structure and turn taking. With these considerations, I expected to find the read speech to be a less complex version of the spontaneous speech, with a clearer topic structure. Then I could try to sort out the contributions of pitch range changes to topic structure from other, more interactive, functions of pitch range changes.

The texts of the conversation excerpts examined can be found in the first six figures. Figures 1 through 4 are of the spontaneous and read versions of two sections of Speaker FP's conversation, and Figures 5 and 6 are of the spontaneous and read versions of Speaker DW's conversation. In each of the figures my utterances are shown in italics and set off in shaded boxes. Silent intervals, a reflection of pauses, are shown in milliseconds between angled brackets (< >). These figures also show the discourse segmentation (see Sections 3 and 4) and intonational phrasing (see Section 5). The symbols **PT, R, H, S, ., __, F,** and **C** are the discourse segmentation codes, and the symbols |, ||, } show the intonational phrasing.

3. Discourse segmentation

There are at least two levels of discourse segmentation which play a strong role in the organization of spontaneous speech. Both interactive turn taking and divisions into major and minor topics are important organizational principles of spontaneous speech. Spontaneous speech also has disruptions to the organized development of topics and turns in the form of on-line production phenomenon such as hesitations, false starts, and corrections. However, neither interactive turn taking phenomena nor hesitation phenomena such as false starts and corrections are particularly crucial to the discourse segmentation of read speech since the scripts provide the explicit turns and exactly what to say.

The principles of turn taking, topic structure, and on-line production phenomena were used as the basis for a qualitative analysis by an independent coder. This analysis then served as the reference for exploring possible acoustic correlates of each sort of phenomena. The independent coder was given an audio recording and a purely orthographic transcription of each of the texts, with no pauses or punctuation marks of any kind. She played the tape as much as she needed to label the data according to the instructions and labels described below. This was a purely auditory-perceptual analysis since she had no instrumental records of the speech. The labels were then related to acoustic measures such as pause lengths, standardized vowel durations (a reflection of speech rate), and pitch range relationships. The pause duration and speech rate results are described in Section 4, and the pitch range relationships are described in Section 7. This analysis was primarily a discourse segmentation and not a strong hierarchical discourse theory version of topic and subtopic relationships. I speculated on the subtopic structure based on the topic segmentations provided by the coder. The full text of the parts of the conversations studied are given in Figures 1 through 6 with the coder's labels. The coding scheme is described in the following paragraphs. The symbols **PT, R, H, S, ., __, F,** and **C** are the discourse segmentation codes, and the symbols |, ||, } show the intonational phrasing (see Section 5). Silent intervals, a reflection of pauses, are also shown in milliseconds between angled brackets (< >). This study looked at the spontaneous and read versions of two different sections of Speaker FP's conversation ('College' shown in Figs. 1 and 2,

and 'Friend' shown in Figs. 3 and 4) and one section of Speaker DW's conversation ('Fernblaster' shown in Figs. 5 and 6).

For turn taking, the coder labeled possible turns (PT), rush through (R), holding the floor (H), and searching for a word (S). A possible turn was described as the possible end of a turn, where the other participant could have started speaking. Rush through was described as a move by the speaker to speak faster and prevent the other speaker from taking a turn. Holding the floor was described as the speaker doing something to keep the floor and indicating that he had more to say. Searching for a word seemed to be a subcase of holding the floor and not reliably distinguishable from holding the floor otherwise. The results in Section 4 treat both holding the floor and searching for a word as holding the floor. The coder remarked that rush through only seemed possible between sentences, and that her percept of possible turn may have been based on the presence of a pause. She said that as a New Yorker (one who tends to trade turns rapidly and tolerate only short pauses at the change of floor) she may have put in more possible turns than the speakers themselves would have perceived, since they are from other parts of the country. Indeed she was correct, because I (one of the speakers) did not perceive as many possible turns as she did. Therefore, I have also marked where I considered the possible turns to be, which I had also done auditorily before I began the instrumental analysis. Those locations are the PTs marked with boxes around them, the ones where we both agreed that there was a possible turn change. I did not perceive any such locations in the read speech, but she did.

For topic structure, the coder labeled ends of sentences (.) and ends of paragraphs (__). Sentences and paragraphs were described loosely. Sentences could be syntactic sentence fragments as well as complete, well-formed syntactic sentences. A paragraph was described as a group of sentences that belonged together, and was possibly divided from the preceding or following paragraph by a change of topic. However, I did not try to impose any strong idea of what a change of topic might be.

For on-line production phenomena, the coder labeled false starts (F) and corrections (C). A false start was described as an incomplete sentence which was abandoned and not completed. A correction was described as a correction of a previous word or phrase -- for example, repeating a word with the correct pronunciation or using a new word or phrase after a false start. All of the corrections marked were self-corrections. The coder remarked that false starts and corrections did not really apply to the read speech data.

Generally the coder's labeling of the phenomena and mine agreed. However, there are a few points where I disagreed with her labels. My labels which disagree with hers use the same coding scheme, but the labels are circled. In FP's spontaneous version of 'College', shown in Fig. 1, I felt that there was a false start and correction between the phrases *Spanish I was uh* <661> and *necessarily had uh* <317>. In FP's spontaneous version of 'Friend' shown in Fig. 3, I strongly disagree with her labeling of the part *the stuff he knows is kind of amazing 'cause he does a lot of uh environmental impact stuff.* She marked an end of sentence after *amazing* and a rush through between *amazing* and *'cause.* I disagree that there is a sentence break there. My judgment is that the break is after *'cause,* at the long pause of 1137 ms, and that that marks the end of a false start and the beginning of a correction to the false start with the phrase *he does a lot of uh.* Otherwise our judgments were basically in agreement. She marked every instance of a repeated word as a correction, while I did not necessarily think of this kind of stuttering as a correction. We perceived hesitations and major paragraph breaks in the same places.

Speaker FP, Spontaneous: 'College'

Why/why wasn't it appealing when it was on computer? <653>
uh because uh <894> || H PT
I F mean |
t- ma- }
to C make a map ||
on a computer would }
C is |
not <47> ||
nearly as much fun as <507> || H
F C to me ||
this seems very obvious <322> || [laugh] . PT
to make it on ||
F C to make it by ||
hand |
is much more fun than to make it on a computer ||. R
but anyway <553> || PT
um <523> || H PT
if you d- C if you can't see that |
then I C I don't know if I can explain it to you <634> ||. PT
[laugh] so I n-}
I C knew I wasn't going to be a cartographer |
but I had no idea what I |
was going to do <323> ||. PT
and <45> || H
I ||
had registered for Spanish ||
simply because I had taken it for
five years in high school <461> ||. PT
and <469> || H
because I was taking |
Spanish I was uh <661> || H PT
FC necessarily had uh <317>) H
F well H the <305>}
the advisor to f- <198>}
C fill out my schedule for the first semester said ||
why don't you take indroduction <141> |
introdu- <130> }
C introductary linguistics ||
which was <53> one ninety ||. R
which is our |
two oh one <1342> ||. PT
and I |
took it |
with |
uh <492> || H PT

F it was taught by |
Thomas ||
Field ||. R
Dr. Thomas Field <623> ||. PT
whom Mary |
knows <352> ||. PT
because he |
also |
went |
to Cornell ||
F graduated from Cornell <302>||. PT
and he does stuff with um <177> || H
Occitan <50> and <855> || H PT
minority |
French languages ||. PT
and speaks them well enough <109> ||
to be <35> mistaken as a native <536> ||. PT
in that part of C of uh France <125> ||. PT
which is amazing <692> ||. PT
and || H
um <918> || H
ever since |
then I knew that <170> ||
linguistics was something I was interested in <530> ||. PT
and <134> || H
I never took any really hard |
core stuff there <620> ||. PT
um <1078> || H
but I knew that <113> }
ling- <378> }
FC being a linguist is what I wanted to do <228> ||. PT
I graduated from college in three years <502> ||. PT
and <275> || H
almost went to graduate school except
that I realized I was con- <325> }
FC had no idea where I was going or <109> | H
what I was going to be doing so <383> ||. R
I ended up teaching ||. PT
but that <663> } H.
F what I did while I was actually there is I was || H
an interdisciplinary studies major <283> ||. PT
y' have any idea what that is <100> ||. PT
Yeah, I've heard.

Fig. 1. Discourse segmentation and coding. Speaker FP, Spontaneous: 'College'.

Coding key:
PT (possible turn), **R** (rush through),
H (hold the floor), **S** (search for word),
. (end of sentence), (end of paragraph),
F (false start), and **C** (correction).

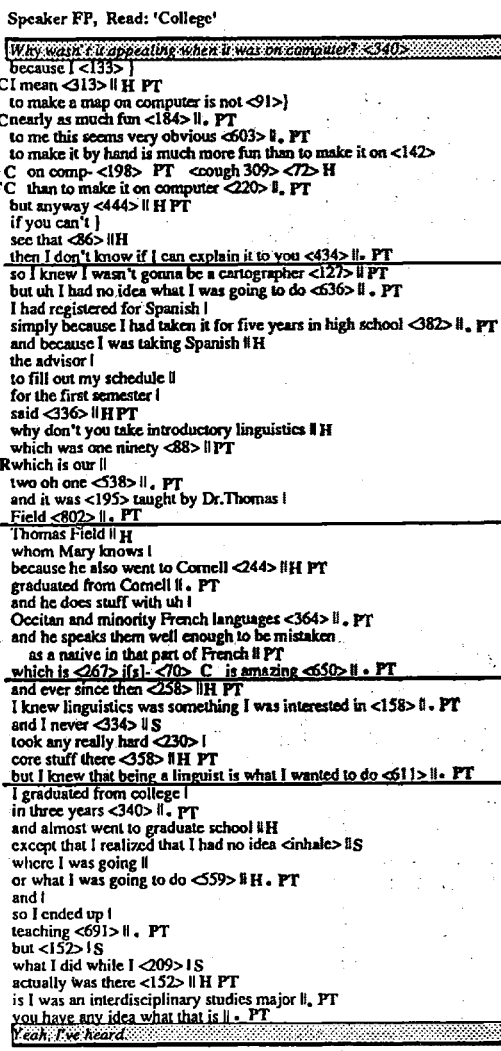


Fig. 2. Discourse segmentation and coding. Speaker FP, Read: 'College'.

Figures 1 and 2 show the spontaneous and read versions of the 'College' part of FP's conversation. Immediately obvious is that the divisions into paragraphs that the coder assigned are not the same for the two versions. The spontaneous version was divided into four paragraphs while the read version is divided into five paragraphs. The end of the first paragraph in the spontaneous version was also perceived as the end of a paragraph in the read version. Then there is a major departure in the paragraph divisions in the two versions. In the spontaneous version the conversation flowed without clearly perceptible breaks from one detail to the next in the second paragraph, from registering for

Speaker FP, Spontaneous: 'Friend'

a friend of mine um <216> ‖ H
works for NASA ‖ . R
he's a physis- <112> |
C physicist <698> ‖ H . PT
and works at NASA <389> ‖ . PT

mm-hmm

R or 'e CF used to work for NASA ‖ . R
he now works for uh <963> | H PT
S uh <98> ‖ H
Federal Energy Regulat- <95> }
no <1597> ‖ H PT
uh S C Department of Energy <197> ‖ . PT
he works for Dept. of Energy <175> ‖ . PT
and he <248> }
visits all the nuclear |
power plants in the country <953> ‖ . PT

hmm

which is |
I suppose |
interesting work <1250> ‖ . PT

mm but I <64> }
had |
I C mean ‖
the stuff he knows <583> ‖ H PT
is kind of amazing . R 'cause <1137> ‖
FC he does a lot of uh | H
environmental impact ‖ H PT
stuff <694> ‖ H . PT
and so <603> ‖ H PT
a lot of things |
that aren't |
necessarily related to physics |
he knows <42> ‖ . R
which is <98> }
FC 'at's <66> really interesting <1196> ‖ . PT
he knows uh | H
geography |
and climate of |
just about every region |
of the United States <391> ‖ . PT

Well that's convenient if ever he wants to move somewhere nice when he retires or gets sick of nuclear energy.

Fig. 3. Discourse segmentation and coding. Speaker FP, Spontaneous: 'Friend'.

Speaker FP, Read: 'Friend'

a friend of mine works for NASA ‖ . PT
he's a physicist ‖
and works at NASA ‖ R
or 'e used to work at }
FC for NASA ‖ . PT
he now works for the Dept. of Energy <300> ‖ . PT
he works for Department of Energy ‖
and he visits all the nuclear |
power plants in the country ‖ R
which is ‖
I suppose ‖
interesting |
work <453> ‖ . PT
but I mean ‖
the stuff he knows is kind of amazing |
because he does a lot of |
environmental impact stuff <454> ‖ . PT
and so <578> ‖ H . PT
a lot of things that aren't necess. related to |
physics ‖
he knows <153> ‖ . PT
which is really interesting <497> ‖ . PT
he knows |
geography and climate of |
just about every region in the United States <67> ‖ . PT

Well, that's convenient if ever he wants to move somewhere nice when he retires or gets sick of nuclear energy.

Fig. 4. Discourse segmentation and coding. Speaker FP, Read: 'Friend'.

introductory linguistics, to the teacher who taught it, to the teacher's research, to his reaction to the course. Only when he said that he graduated from college in three years did the coder say that a new paragraph had begun. In the read version of this same section, the section was divided into three different paragraphs. Essentially the points that flowed from one to the next in the spontaneous version were given stronger emphasis in the read version and were judged to be independent paragraphs. Both versions had a paragraph beginning at *I graduated from college in three years*. Another part of FP's conversation, 'Friend', is shown in Figs. 3 and 4 in both the spontaneous and read versions. Again, as in Figs. 1 and 2, the transcription codes show that the two versions had different divisions into paragraphs. In the spontaneous version there were judged to be two paragraphs, but in the read version there was only judged to be one.

Speaker DW, Spontaneous: 'Fernblaster'

uh H I'm |
exaggerating <1604> ||. PT
but |
I | H
CI <245> like to be spontaneous |
when I teach <677> ||. PT
uh <1628>. PT
well | | C |
kind of invent ||
characters ||
to help me <697> || S H. PT
with my teaching || H R
there's one <183> || PT
What? multiple personalities? <975>
it might be a manifestation of that ||. R
you never know <317> ||. PT
<inhale> uh <548> || H
no | H
uh <754> | S H. PT
for <153> CF to illustrate the idea of ||
pre-scriptive <1492> || S. PT
study of language |
uh <556> ||. PT
I always come from the standpoint of |
everybody's ||
eighth |
grade ||
English teacher || H R
Mrs. |
Edna Fernblaster <513> ||. PT
Fernblaster!
<248> Mrs. Edna |. PT
She doesn't like plants or something then.
<500> | | H
CI don't know ||. PT
there're several || S H
pictures that come to mind || PT
some I'd <448> | S H
CI'd rather not discuss in genteel company bu[t]-
<laugh 420> ||. PT
uh <490> || H.
um no uh <creak 856> || H. PT
no Cno it's just a HC n C it's just a weird sounding name ||. PT
you know ||
whenever I talk about somebody telling you ||
how ||
to speak <342> ||. PT
then |
you know |
I always |
come from CF well um ||
you know Mrs. Fernblaster ||
would tell you to do this || H. R
she would say never say ain't ||
and that type of thing <1391> ||. PT
so uh <459> H PT and and CF and people <creak 316> || H
Cpeople know when she's coming up ||
anymore ||. PT
uh <60> and your eighth grade English teacher || H
and you hear these little titters |
in the back of the room <205> ||. PT
Edna Fernblaster 'n <463> ||. PT
and so ||.
you know |
that type of thing <125> <creak 167> ||. PT
F because <497> || H PT
I li- C like to use characters |
like that |
because <502> || H PT
it's such a basic concept | . R
it's nice to have something <373> || H PT
concrete to hang onto to help <1000> ||. PT
Right.
'n so <228> ||
there's really not been |
too much of a chance for <128> || S H
class participation as yet <642> || . PT

Speaker DW, Read: 'Fernblaster'

no I'm exaggerating <283> ||. PT
but I like to be spontaneous when I teach || . PT
uh <395> ||
C I'll <244> || S
kind of invent |
characters |
to help me with my teaching ||. PT
there's one || . PT
What? multiple personalities? <297>
it might be a manifestation of that || H R
uh you never know <339> ||. PT
uh no <189> || PT
uh to illustrate the idea of |
prescriptive |
study S of language || H
I always come from <256> ||
the standpoint of everybody's ||
eighth grade English |
teacher |
Mrs. Edna Fernblaster || . PT
Fernblaster!
this is Edna <211> ||. PT
She doesn't like plants or something then.
I don't know <263> ||. PT
there are several pictures that come to mind <91> || .PT
eh |
some I'd rather not discuss in genteel company <413> ||. PT
uh <122> ||
no ||
it's just a weird sounding name ||
you know | H. R
whenever you talk about <389> ||
somebody |
telling you how to |
speak ||
then I always say <289> ||
well you know Mrs. |
Fernblaster ||
would tell you to do this <309> ||. PT
she would say ¶
never say ain't ||
and that type of thing <414> ||. PT
so |
and <216> | S H PT
people know when she's coming up anymore <428> || . PT
and your eighth grade English teacher ||. R
and you hear these little |
titters in the back of the room || H. R
Edna Fernblaster <467> ||. PT
and so you know <277> ||
that type of thing <103> ||. PT
I like to use characters | S
like that because it's |
such a basic concept <256> ||. PT
it's nice to have something concrete || S
to hang onto to help ||. PT
there's really not been too much of a chance for |
class participation as yet <497> ||. PT

Fig. 5. Discourse segmentation and coding. Speaker DW, Spontaneous: 'Fernblaster'.

Fig. 6. Discourse segmentation and coding. Speaker DW, Read: 'Fernblaster'.

The spontaneous and read versions of Speaker DW's 'Fernblaster' section shown in Figs. 5 and 6 had similar divisions into paragraphs. The first half was judged to be three paragraphs in the spontaneous version, but only one paragraph in the read version. The first two paragraph divisions in the spontaneous version align with pauses greater than 1.6 seconds. It is as if DW was taking his time making the transition from the previous part of story that he had been telling into a new aspect of the story in the spontaneous version, but that it did not take the same kind of time to make the transition in the read version. The next paragraph division after the matching halfway point was in essentially the same position, either before or after *and so you know that type of thing*, and the final paragraph division was in the same location.

4. Acoustic measures of pause and speech rate

To see whether there were any easily quantifiable correlates of these "paragraphs", "possible turns", and so on, measures were made of pause durations and vowel durations (the latter as a metric of speech rate). Relatively broad phonetic transcriptions of the data were made using both auditory perception and spectrographic analysis of the speech. The spectrograms were made on a DSP Sona-Graph 5500-1, Kay Elemetrics Corporation instrument. Silent pauses and breaths were identified and their durations were measured in milliseconds, based on the spectrograms and waveforms of the speech. The pause durations are reported in the transcriptions as millisecond values shown between angle brackets (< >). Breaths are not distinguished from silent intervals, but rather are included in the pause durations given in the transcriptions. Silent intervals due to segments, such as stop closures, were not counted as pauses. I segmented the vowels guided by spectral changes between consonants and vowels. Vowel durations always included exclusively the voiced portion of a vowel, where there was an obvious voice bar. After stop consonants, the first periodic glottal pulse with both a voice bar and energy in the first formant was taken as the beginning of a vowel. Vowels were segmented from nasal and lateral contexts at the point of spectral change and damping of the first and higher formants. Voiceless vowel durations were not always possible to segment and separate from the surrounding consonants, so they were classified as voiceless and not given any duration in milliseconds.

4.1 Pause measures

Previous investigators have found that read versions of spontaneous texts exhibit fewer pauses than the original versions (Gårding, 1967; Howell and Kadi-Hanifi, 1991). This was also the case for these data, as shown in Table 1. Read speech has also been found to have shorter pauses than spontaneous speech (Gårding, 1967; Butterworth, 1975), and this finding also holds for these data, also shown in Table 1. Both speakers had similar patterns of pause length distributions, with a higher mean and larger standard deviation of pause length in the spontaneous than in the read. Pause durations were significantly different between spontaneous

Table 1
Pause characteristics of the spontaneous and read productions by Speakers FP and DW.

	FP		DW	
	Spon	Read	Spon	Read
total number of pauses	72	46	36	23
mean duration (ms)	439	322	561	293
standard deviation (ms)	358	192	418	112

and read speech for both speakers (FP: t= 2.13, p<.05; DW: t=2.71, p<.01), but the distributions of pauses within the same mode of speech was not significantly different between the speakers (spontaneous: t=-1.34, p>.1; read: t=.51, p>.1). So, pauses were longer and had more variable lengths in the spontaneous speech than in the read speech, for both speakers. On this measure of pause length then these materials are typical of what has been found in spontaneous and read speech in the past, even though this read speech is not prototypical read speech.

Previous investigations have found pauses at the ends of sentences and longer pauses at the ends of paragraphs in read speech (Lehiste, 1979; Brown, 1983; Silverman, 1987; Grosz and Hirschberg, 1992; Passenout and Litman, 1993). Table 2 reports the pause length distributions relative to the discourse coding categories for the current data. The values in the columns of the table represent how many of the data points fall within the range of values. The first two columns are the values of pauses longer than the mean, either greater than one standard deviation above the mean (the first column) or between the mean and one standard deviation above the mean (the second column). The next two columns represent the pause durations less than the mean, either between the mean and one standard deviation below the mean (the third column) or less than one standard deviation below the mean (the fourth column). The fifth column represents those occurrences of each coding category with no following pause, and the final column represents the ones which have *um*, *uh*, or similar filled pauses, which is used mainly as an explanation for the code holding the floor, which is discussed shortly. Values in the "um" column include tokens from the first five columns, since they either did or did not have following pauses. There were not very many paragraphs in the data, but it did not seem necessary for there to be a long following pause (greater than the mean) in order for the coder to mark an end of paragraph. For Speaker FP's read version there seemed to be longer pauses at the ends of paragraphs however. Otherwise there was no compelling evidence for this claim in these data. More sentences ended with pauses than without following pauses, but again, sentences could end without a following pause.

Possible turn locations as marked by the coder correlated very closely with the presence of a following pause but had no clear correspondence with the length of the following pause. There are more pauses than turn labels, but most of her possible turn locations had a following pause. More of the possible turn locations corresponded with a following pause of longer than the mean, but she also marked possible turn locations when there was no pause at all. I marked many fewer possible turn locations than the coder did. Where I marked possible turns, the pauses were generally higher than the mean. Note, however, that as a participant in the conversation I took actual turns in the DW spontaneous conversation at places that I at later listening didn't think were appropriate as possible turn locations. That means that I took interruptive turns, and actively took the floor rather than waiting until it was given to me. Those locations had shorter than the mean pause duration, or no pause at all. I marked no possible turns at all for the read versions, because it didn't seem to me as a listener that there were possible turns in the read version.

Rush through was marked primarily on locations where there was an extremely short pause or no pause at all at the end of a sentence. This would correspond to what Schegloff (1982) called failing to pause for breath and continuing on into the next unit. Holding the floor had no clear relationship to following pause length. Sometimes there were long following pauses and sometimes no following pause at all. However, pause fillers and hesitation words like *um* and *uh* were closely linked with the label of holding the floor. There were more instances of rush through and holding the floor marked in the spontaneous speech than in the read speech, as we would expect if we view these as indications

of interaction with the other conversational participant and of the speaker having to think on-line of what to say.

Table 2
Numbers of each type of discourse coding for each following pause category, by speaker and mode of speech.

	very long (> 1 sd)	long (> m)	short (< m)	very short (< -1 sd)	no pause	"um"
a) FP spontaneous						
topic structure						
sentence end	4	9	13	1	7	0
paragraph end	1	1	3	0	1	0
turn structure						
possible turn (coder)	8	15	12	0	3	4
possible turn (author)	4	3	2	0	0	0
rush through	0	0	0	2	7	0
holding the floor	6	8	8	2	5	13
b) FP read						
topic structure						
sentence end	9	7	5	1	5	0
paragraph end	3	1	0	1	1	0
turn structure						
possible turn (coder)	7	11	11	3	6	0
possible turn (author)	0	0	0	0	0	0
rush through	0	0	0	1	2	0
holding the floor	1	4	5	1	4	0
c) DW spontaneous						
topic structure						
sentence end	5	4	8	0	7	3
paragraph end	2	0	2	0	0	1
turn structure						
possible turn (coder)	4	6	13	0	5	4
possible turn (author)	3	0	2*	0	1*	0
rush through	0	0	0	0	5	0
holding the floor	0	3	7	1	11	6
d) DW read						
topic structure						
sentence end	5	2	5	2	7	0
paragraph end	1	0	0	1	1	0
turn structure						
possible turn (coder)	5	2	6	2	5	0
possible turn (author)	0	0	0	0	0	0
rush through	0	0	0	0	4	0
holding the floor	0	0	1	0	4	0

* I did not mark these as possible turns listening afterwards, but I actually took turns (of the interruptive sort) at these points.

Table 3
Vowel phoneme duration means and standard deviations, by speaker and mode of speech.

a) Speaker FP

vowel phoneme	mean duration (ms)		standard deviation (ms)		number of tokens	
	Spon	Read	Spon	Read	Spon	Read
i	58.2	63.0	24.8	31.1	50	52
I	52.7	54.3	32.5	28.7	140	114
u	106.8	96.5	73.7	55.5	6	4
U	63.5	54.5	10.0	2.1	4	2
ɛ	75.1	78.2	41.5	43.0	31	39
ə	75.5	56.1	82.7	29.6	134	137
o	123.2	97.1	70.2	65.3	22	19
æ	117.3	97.4	69.8	42.5	41	30
a	84.2	83.1	37.6	23.4	24	21
ei	89.0	81.2	31.3	27.8	17	18
ai	115.7	88.2	38.3	31.9	30	30
au	157.5	96.7	58.7	23.0	4	3

b) Speaker DW

vowel phoneme	mean duration (ms)		standard deviation (ms)		number of tokens	
	Spon	Read	Spon	Read	Spon	Read
i	73.4	65.3	31.9	29.1	25	27
I	70.3	57.2	44.7	27.3	61	54
u	66.0	69.8	24.6	25.9	5	5
U	51.7	47.5	17.5	16.3	3	2
ɛ	77.5	71.4	40.5	29.3	23	26
ə	94.1	66.4	64.4	44.5	82	67
o	163.7	135.9	60.8	47.9	15	12
æ	114.0	94.7	43.5	35.9	26	19
a	90.0	106.3	34.1	38.5	10	8
ei	111.4	87.7	34.1	29.9	17	15
oi	111.0	69.0	0.0	15.6	1	2
ai	106.4	101.8	37.7	56.1	20	18
au	134.8	94.5	40.4	16.2	4	4

4.2 Speech rate measure

We might expect that the speech rate varies more in spontaneous speech than in read speech, as a speaker rushes to hold the floor, slows down when thinking of what to say, and the like. There have also been reports in the literature about differences in speech rate at the beginning of a paragraph and the end of a paragraph, although the reports disagree on the direction of the differences. The way I chose to look at speech rate was the durations of vowels. The faster the speech rate, the shorter the vowel duration. Looking at just the raw vowel duration can give a partial answer to the question of whether speech rate varies more in spontaneous speech. Table 3 shows the means of the measured vowel durations

and their standard deviations. Voiceless vowels, which were given a duration of 0 ms by the segmentation criteria, were not included in the values shown here. For both speakers, the majority of the vowels had larger means and larger standard deviations in the spontaneous speech than the read speech. The higher standard deviations around the vowel means in the spontaneous speech indicates that there is more overall variation in rate in the spontaneous speech than the read speech.

However, the raw vowel duration alone can only tell us so much about relative speech rate. Each vowel has an inherent duration (e.g. low vowels are longer than high vowels), and with this sort of information remaining we cannot really know if the vowel at any particular point in the discourse is long or short, unless it is compared to the average for each vowel of its type. A method for factoring out this kind of inherent phoneme duration is to convert the duration of each vowel token to a z-score (i.e. the number of standard deviation units away from the mean for that vowel phoneme -- for a full description of the method see Campbell and Isard, 1991; Campbell, 1992). In this way we can see for each particular token whether it is longer or shorter relative to the others of its class. For these data, the standardized values of each vowel were calculated separately for each speaker and each mode of speech. Segments with the mean duration for their class have z-score values of 0, longer than average segments have positive z-score values, and shorter than average segments have negative z-score values. The z-score vowel durations were used as a measure of local rate of speech. So, for example, if rush through was realized by a local increase in rate, the z-score values of the vowels in those regions would be smaller than the surrounding context. Similarly, if holding the floor was partly accomplished by extending the length of a word while the speaker thought of what else to say, we would see larger z-score values for the vowel(s) of such a word.

Table 4 shows the distribution of the z-score vowel durations for the final vowel before each of the discourse codings for each speaker and mode of speech. The long vowels (i.e. greater than the mean, with positive z-scores) are in the first two columns, and the short vowels (i.e. less than the mean, with negative z-scores) are in the next two columns. The last column is for voiceless vowels, whose durations were 0 by the segmentation criteria used. In these data, sentence initial vowels tended to be shorter than sentence final vowels, and paragraph initial vowels tend to be shorter than paragraph final vowels. This does not agree with the observations that words at the beginning of topic units are spoken more slowly than words at the end of topic units (Butterworth, 1975; Lehiste, 1980). It agrees better with the finding that topic segment beginnings were faster as compared to segment endings (Grosz and Hirschberg, 1992).

The final vowel before possible turns primarily had longer than the mean duration for vowels. So, it seems that a relatively long vowel and a following pause were good cues for the coder to decide that there was a possible turn location. Rush through seemed to have a distribution on the shorter end of the scale than possible turns. Most of the vowels before a rush through were less than one standard deviation unit above the mean (z-score greater than 1), but there were some with longer durations. So, a rush through seemed to correspond to a relatively short vowel and a short following pause. Holding the floor corresponded to long vowels; most were greater than the mean and only one token was shorter than one standard deviation unit below the mean. This seems to be a more reliable correlate of holding the floor than following pause duration, which could be very long or no pause at all.

Table 4
Numbers of each type of discourse coding for each following standardized vowel length category, by speaker and mode of speech.

	very long (> 1 sd)	long (> m)	short (< m)	very short (< -1 sd)	voiceless vowel
a) FP spontaneous					
topic structure					
sentence start	7	8	12	3	4
sentence end	13	6	9	1	5
paragraph start	0	2	4	0	0
paragraph end	1	3	2	0	0
turn structure					
possible turn (coder)	14	13	6	0	5
possible turn (author)	1	6	1	0	1
rush through	1	3	4	0	0
holding the floor	20	6	3	0	0
b) FP read					
topic structure					
sentence start	1	5	10	8	3
sentence end	16	6	2	0	3
paragraph start	0	2	1	2	1
paragraph end	3	1	1	0	1
turn structure					
possible turn (coder)	19	13	3	0	3
possible turn (author)	0	0	0	0	0
rush through	1	0	2	0	0
holding the floor	5	8	2	0	0
c) DW spontaneous					
topic structure					
sentence start	3	7	12	1	1
sentence end	10	2	10	1	1
paragraph start	1	4	1	0	0
paragraph end	2	0	2	1	0
turn structure					
possible turn (coder)	12	6	8	1	1
possible turn (author)	0	2*	1	1	2#
rush through	2	0	2	1	0
holding the floor	9	6	6	1	0
d) DW read					
topic structure					
sentence start	6	2	9	3	1
sentence end	5	9	7	0	0
paragraph start	2	0	2	0	0
paragraph end	0	2	1	0	0
turn structure					
possible turn (coder)	6	8	6	0	0
possible turn (author)	0	0	0	0	0
rush through	1	2	1	0	0
holding the floor	2	2	1	0	0

* both from actual turns and not perceived possible turn
# one from perceived possible turn, one from actual turn

A specific example showing the z-scores for each vowel in a matched spontaneous and read section is given in Fig. 7. The example is from Speaker FP's section 'Friend'. A partial intonational transcription of this excerpt is given in (2), where (2a) is the spontaneous and (2b) is the read version. Accented syllables are underlined, and phrases and pauses are marked in the text as previously. A complete discourse coding of the example can be found in Figs. 3 and 4. The vowel z-scores are plotted against the vowel phonemes occurring in the excerpt. The z-score values of the vowels of the spontaneous version are shown by open circles and the z-score values of the read version by filled circles. These values represent the vowel durations of the words lined up below the graph. The spontaneous and read versions are aligned with each other word-by-word and vowel-by-vowel. The x-axis tick mark labels are the spontaneous version vowels. The symbol ø means that the vowel was voiceless or deleted (e.g. in *he* and *and*), and a dash shows that there was no corresponding word in the other version (e.g. there was no *at* in the spontaneous version). When a vowel was voiceless or deleted in one version relative to the other, that vowel was given a z-score of -2 to graphically show a 'very short' vowel. When a word was missing relative to the other version because of editing or reading differences, the missing vowels were given z-scores of 0 to graphically show no variation in vowel duration. These two kinds of z-score assignments were purely for display purposes and played no role in the calculations.

(2) a. he's a physis- <112> } physicist <698> ||
and works at NASA <389> ||
mm-hmm
or 'e used to work for NASA ||
he now works for uh <963> |
uh <98> || Federal Energy Regulat- <95>}
no <1597> || uh Department of Energy <197> ||

b. he's a physicist ||
and works at NASA ||
or 'e used to work at } for NASA ||
he now works for the Department of Energy <300> ||

Fig. 7 gives the flavor of the rate variation given by this measure of relative vowel duration. The most striking differences between the two versions is where long durations, i.e. relatively slow speaking rates, were used. Very long duration vowels occurred phrase finally, for holding the floor, and for searching for a word. In the read version, the very long duration vowels were phrase final vowels, especially the two tokens of *NASA*. These are clear instances of phrase final lengthening. The phrase *he's a physicist* also shows final lengthening, although not as strikingly. The spontaneous version does not show the same clear tendency for final lengthening as the read version does. In the spontaneous version, the very long vowels were other than phrase final vowels. The final vowel in the aborted word *physic-* was very long, presumably because the speaker was trying to decide if that was what he actually wanted to say. The other two very long duration vowels in the spontaneous version (well over 2) were on the first two occurrences of the pause filler *uh*. The coder and I both marked these as holding the floor and searching for a word. This is a clear example of breaking in the middle of the next unit, both a syntactic and semantic unit, as Schegloff (1982) describes, and it shows lengthening associated with searching for upcoming words. Notice, however, that the next occurrence of holding the floor and searching, on *no uh*, that neither word had a z-score over 1, so very long durations are not necessary

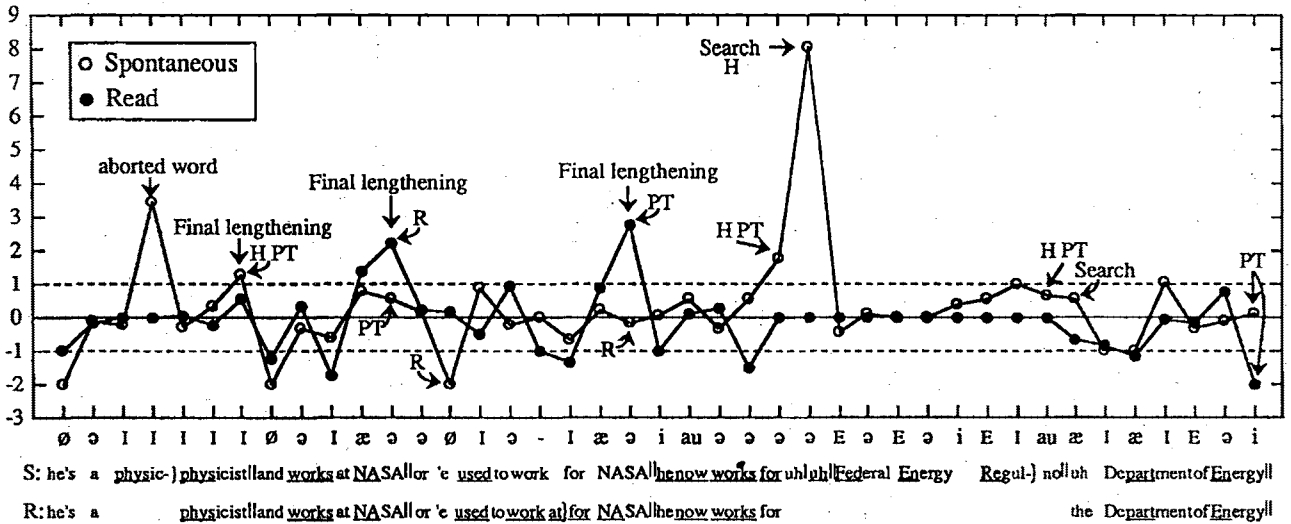


Fig. 7. Vowel duration z-scores for an excerpt from Speaker FP: 'Friend', both spontaneous and read versions. The y-axis shows the z-score value of each vowel, and the x-axis shows the vowel phonemes aligned above the appropriate text for the spontaneous (S) and read (R) versions. Underlined syllables are accented. Discourse codes are as in Figs. 1 to 6.

for holding the floor and searching to be perceived. In this specific case there was a long silence, 1597 ms, between *no* and *uh* after an incorrect mention, *Federal Energy Regulat-*, and the expectation is that he will think of the correct place and continue speaking once he has thought of it. This would be adequate in itself for the perception of holding the floor and searching for a word.

Short vowel durations correspond in some cases to the perception of rush through. In the spontaneous version the speaker uttered the phrase *or 'e used to work for NASA* with just a single accent on *used* and spoke the rest of the words relatively faster than his average rate. There was no pause for breath at the end of the phrase and the vowels were voiceless or right at or below average duration. The vowel for *'e* was a voiceless vowel (hence given a -2 z-score for display purposes), and the beginning of this phrase was perceived as rushed. There was also a rush through marked at the end of this phrase ending with *NASA*, and no possible turn was judged possible there.

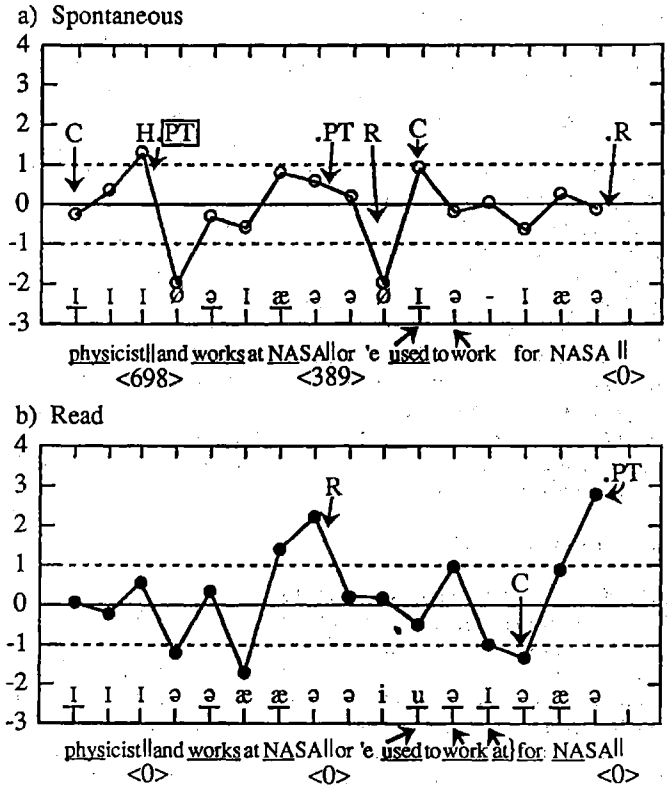


Fig. 8. Vowel duration z-scores for an excerpt from Speaker FP: 'Friend', for (a) spontaneous and (b) read versions. The y-axis shows the z-score value of each vowel, and the x-axis shows the vowel phonemes aligned above the appropriate text. Discourse codes are as in Figs. 1 to 6. Underlined vowels and syllables are accented.

In many of cases discussed above there was a correspondence between short segmental durations and the perception of rushing, and a correspondence between long segmental durations and holding the floor or searching for a word. However, in the spontaneous version *no* was marked as holding the floor where the vowel had a z-score of less than 1, and in the read version a rush through was marked at the end of the phrase *and works at NASA* where the vowel had a z-score of more than 2. Understanding things like rush through and the like from rate information are complicated by noise in the data from accents and phrase boundaries. A vowel can be long because it is an accented syllable, because it is a phrase final syllable, or because the speaker is trying to hold the floor. However, a vowel can be accented and still not be very long. Fig. 8 shows a further expansion of part of this same example. As in Fig. 7, the vowel z-scores are plotted against the vowel phonemes. In (a) the vowels durations plotted and labeled are the spontaneous vowels, and in (b) the vowels plotted and labeled are the read vowels. The underlined vowels were the vowels which were accented, and we can see that in both the spontaneous and read versions there are accented vowels which have z-scores of less than the average of 0. Pause durations are also given.

In the read version the word final vowel of *NASA* is long in both cases showing phrase final lengthening. The [ə] of the first *NASA* is shorter than the second one where the coder marked the end of a sentence and the end of a potential turn, indicating that it had less final lengthening. This potential turn had no following pause, so it must have been partly the extreme final lengthening that contributed to the perception of a potential turn. Notice, however, that the first *NASA* in the read version is marked as a rush through, even though the vowel durations are very long, with the long [æ] from the accent and the long [ə] from the phrase final lengthening. The perception of rush through at that point is probably due to the lack of pause between the phrases. In the spontaneous version, the accented vowel [æ] of *NASA* is longer than the word final vowel. In the spontaneous version the phrase final lengthening is not the primary contribution to length on *NASA*. On the word *physicist* by contrast it is the phrase final vowel which is longest rather than the accented vowel. Listeners and speakers probably have a complex template to compare to for different positions in a sentence, different accent locations, etc. and can tell when something is rushed relative to what it would have been otherwise.

5. Intonational Analysis

5.1 Symbolic transcription framework

One reason for doing an intonational analysis of the materials in this study was to see if a symbolic intonational analysis could express some of the difference in interactivity between natural spontaneous speech and rehearsed read speech. A second reason was to determine the phrasing necessary for a pitch range analysis. The symbolic transcription framework used in this study is based on Pierrehumbert's system for transcribing English (see Pierrehumbert, 1980 for some categories of the system, modified in her later work with Liberman (e.g. Liberman and Pierrehumbert, 1984) and with Beckman (e.g. Beckman and Pierrehumbert, 1986)) and the ToBI standard (Tones and Break Indices) for prosodically labeling data in American English, Australian English, and certain varieties of British English (Silverman et al., 1992). The major components of the intonational transcription system are pitch accents, phrase accents, and boundary tones. The intonational components are listed in Table 5. Only high tones (H) and low tones (L) are assumed in the phonology. Pitch accents are tones associated to certain stressed syllables. The association shows up in the time alignment of F0 to segments. There are single-tone pitch accents and bitonal accents which have two

Table 5
Tonal components of ToBI intonational transcription system.

Pitch accents:	H*, L*, !H*, L+H* (and L+!H*) (L*+H, H+!H* not attested in these data)
Phrase accents:	L-, H-
Boundary tones:	L%, H%

Intonation phrase final sequences: L-L%, L-H%, H-L%, H-H%

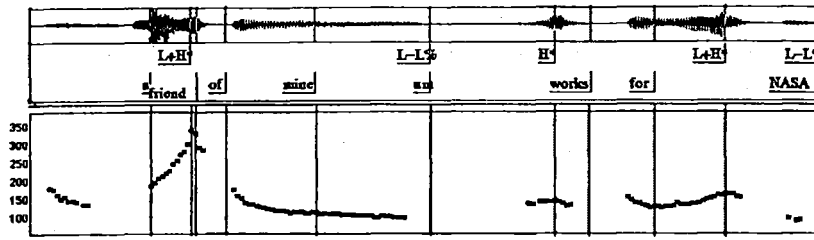
tones, with a tone leading or trailing the explicitly associated tone, the one marked with an asterisk. Words can be grouped into phrases at two levels in this system, intermediate phrases and intonational phrases. Intermediate phrases are marked by phrase accents (L- and H-), and intonational phrases by boundary tones (L% and H%). Intonational phrases can have one or more intermediate phrases. With two kinds of phrase accents and two kinds of boundary tones, there are four possible intonational phrase final sequences (L-L%, L-H%, H-L%, and H-H%). The intonational phrasing is additionally marked in the orthographic transcription of the examples to help the reader see the alignment of tones when they are of interest and to show the phrasing when the specific tones are not of interest. Intermediate phrases are marked by single vertical bars (|), intonational phrases are marked by double vertical bars (||), and intonation contours which are cut off by hesitations or restarts are marked by curly brackets ({}). The symbols for the intermediate and intonational phrase boundaries conform to the IPA guidelines for marking major and minor phrases (International Phonetics Association, 1989).

The examples in Fig. 9 illustrate some of the components of the intonational transcription system. They are examples from Speaker FP 'Friend', the spontaneous (a) and read versions (b) of the sentence *A friend of mine works for NASA*. The intonational transcription can also be found in example (3), with spontaneous (a) and read (b). The figure shows from top to bottom for each version the speech waveform, tonal transcription, word boundaries, and fundamental frequency contour. The time scale is the same for both versions, and shows that the spontaneous utterance is longer than the read version. The ends of words are marked by the labeled lines. The transcriptions are tightly linked to the fundamental frequency contour as well as to the auditory percept. H* signifies a high target F0 on the accented syllable. The accent L+H* is characterized by a rise from a low to a high frequency. This rise for L+H* is seen most clearly in Fig. 9 on the word *friend* in the spontaneous version. Downstepped accents are transcribed explicitly with the downstep diacritic '!' (!H* and L+!H* in these data). We see downstepping in the read version of Fig. 9. The sequence of !H* accents means that each high tone is realized on a lower pitch than it would have been were it not downstepped. There is no specific pitch movement obvious for the accented words in this example, but there is a clear percept of accent on the words *mine*, *works*, and *NASA*.

(3) a. a friend of mine um || works for NASA ||
L+H* L-L% H* L+H* L-L%

b. a friend of mine works for NASA ||
H* !H* !H* L-L%

a) Spontaneous version: *A friend of mine um works for NASA.*



b) Read version: *A friend of mine works for NASA.*

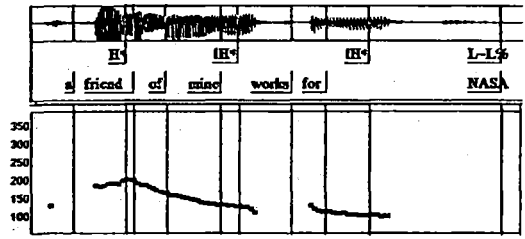


Fig. 9. Spontaneous and read versions of Speaker FP's sentence *'A friend of mine works for NASA.'* Fundamental frequency in Hz.

Four types of differences between utterances can be described given this system of transcription: presence versus absence of pitch accent, type of pitch accent, phrasing, and pitch range of phrases. The examples in Fig. 9 (and example (3)) illustrate all but the first of these differences. The words *friend* and *NASA* show differences in the choice of accent type between the two versions, L+H* in the spontaneous and H* or !H* in the read. There is a difference in phrasing, two intonational phrases in the spontaneous version (both ending with L-L%) in contrast to a single intonational phrase in the read version (also ending with L-L%). The two phrases in the spontaneous version gives two domains for pitch range, in contrast to the read version where there is just one. In addition, the spontaneous version was realized in a much wider pitch range in the first domain -- a peak at 337 Hz versus 200 Hz in the read version. Example (4) shows differences in presence versus absence of pitch accent, as well as phrasing and pitch range of phrases. They are again utterances from Speaker FP 'Friend', spontaneous (a) and read (b). The spontaneous version had only a single accent, on *used*, while the read version had accents on several more words. The read version has a second phrase, after the interrupted phrase correcting *at* with *for*. The pitch range of the spontaneous version, realized by the peak on *used*, was much higher than the read version -- 196 Hz as opposed to 159 Hz.

```
(4)   a.  or 'e  used to work for NASA ||
                 L+H*                  L-L%

      b.  or 'e  used to work at } for NASA ||
                 L+H*       !H* !H*   H* !H* L-L%
```

5.2 Accents

The first part of Table 6 shows the distribution of accent types in the different texts, by speaker and mode of speech. The left-hand half shows the total occurrences of each accent type, and the right-hand half shows the distribution of each accent type as a percentage of the total number of accents. The proportion of words which are accented -- that is, words which have any kind of pitch accent on them -- is similar for the spontaneous and read versions of each speaker, with a slightly smaller percentage of the words accented in the read versions. H* was by far the most common accent type overall. There was a higher percentage of downstepped accents (!H* and L+!H*) in the read speech as compared to the spontaneous speech. There are other differences in the distributions, but these were the most obvious generalizations. The fact that there was a greater percentage of downstepping accents in the read version may be a cue to narrative as opposed to interactive style of communication. Bolinger (1978, p. 490) describes the downstepping intonation as "the only intonation that can be used in starting a story [of the type]: Once there was a bear. His name was Smokey." He characterizes

Table 6
Accents.

a) Overall distribution of accents and accent types, by number of tokens and percentage of the total number of accents.

	number of tokens				percentage			
	FP		DW		FP		DW	
	Spon	Read	Spon	Read	Spon	Read	Spon	Read
total accented	231	203	139	120	54.4	51.8	58.5	57.7
total words	425	392	236	208				
H*	124	90	88	61	53.7	44.3	63.8	50.8
!H*	65	65	19	19	28.1	32.0	13.8	15.8
L+H*	32	34	29	30	13.9	16.7	21.0	25.0
L+!H*	5	12	2	10	2.2	5.9	1.4	8.3
L*	5	2	1	0	2.2	10.0	0.7	0.0

b) Word by word comparison of accent distributions between the spontaneous and read versions (collapsed over downstepped variation), by number of tokens and percentage.

	number of tokens		percentage	
	FP	DW	FP	DW
Words accented in both versions	152	100	54.3	64.5
a. same accents	109	68	38.9	43.9
b. different accents	43	32	15.4	20.6
Words accented in one version	128	55	45.7	35.5
a. spon unaccented	50	19	17.9	12.3
b. read unaccented	78	36	27.8	23.2
Word accent pairs	280	155		

this kind of intonational contour as being used in "self-confident" ... "narratives where a single speaker holds the floor and imposes himself on the audience". Beckman (personal communication) thinks of the downstepping contour in terms of the rather pedantic expectations set up by the act of narration where the narrator is saying something like "This is a story. Each piece flows in a clear rhetorical succession from the last. The story structure of this discourse should give you the connections that I'm evoking by this contour ..." This difference in distribution of accent types in spontaneous and read speech could be interpreted as an essential and important difference between spontaneous and read styles of speech.

However, a numerical tally of the accent types does not reveal the whole picture of the differences in accent type distribution. The distribution of accents on individual words is also important since accent type and accent placement are pragmatic choices for highlighting or downplaying words. The second part of Table 6 shows the comparisons between spontaneous and read word pairs where at least one of the versions had a pitch accent. For Speaker FP, 54% of the words that were accented were accented in both the spontaneous and read versions, and for Speaker DW, the percentage was 65%. However, that leaves 46% and 36% of word pairs, for Speakers FP and DW respectively, where the words were only accented in one of the versions. Of those mismatched accent pairs, more of them were cases where the word was unaccented in the read version than unaccented in the spontaneous version. This goes along with the observation that a slightly smaller percentage of the words were accented in the read versions.

The choice of accent type and accent placement for particular words differed more between the two versions than the pure quantity of accents of a certain type used in a whole text. Since accent type and accent placement are pragmatic choices for highlighting or downplaying words, the two versions differ more by pragmatically determined meanings than a simple count of proportion of accent types used in a whole text can reveal. The differences in choice of accent placement between the two versions reveal differences in attentional structure, what to pay attention to over the unfolding of a discourse (Grosz and Sidner, 1986). Howell and Kadi-Hanifi (1991) also made detailed comparisons of the location of what they called "primary stress" (similar to accented words here, since they mention major pitch obtrusion, loudness and length making the syllables prominent) between the spontaneous and read versions in their data and found that many of the stresses were in different positions. However, they attributed these differences to speech rate differences and made no mention of pragmatic meanings. They said that faster speech tends to have fewer stresses, and were not clear about what that might mean for differences between spontaneous and read speech. I am not aware of any other studies that make detailed comparisons between word-by-word accent locations.

5.3 Phrasing

The first half of Table 7 shows the number of intermediate phrases, intonational phrases, and the mean number of words and accents for each level of phrasing in the different texts, by speaker and mode of speech. The spontaneous texts have more phrases with fewer words and accents than the read texts. The read texts have fewer phrases with more words and more accents than the spontaneous texts. The number of intermediate phrases per intonational phrase is nearly identical for all of the texts. The longer phrases in the read speech may be a reflection of the fact that the words are all there and just have to be read instead of being thought through.

The second half of Table 7 shows the distribution of intonational phrase final tone sequences in the different texts. The left-hand half shows the total occurrences of each boundary tone sequence type, and the right-hand half shows the distribution of each boundary tone sequence type as a percentage of the total

number of intonational phrases. One difference in the distribution of phrase final tone sequences of spontaneous and read speech is fairly easy to interpret. Table 7 shows that while L-L% and L-H% were used heavily in both spontaneous and read speech, H-H% tended not to be used in read speech. While FP had 9 tokens of H-H% in his spontaneous speech, he had only 1 token in his read version. DW had 1 token of H-H% in his spontaneous speech and none in his read version. The contour transcribed as H-H% in this system, a phrase final high rising intonation, is a quite common American contour for inviting listener comments and indicating that the listener is to interpret what was said in terms of what follows (the situational context or the following utterance), and is the standard intonation for a yes/no question. Sacks and Schegloff (1979) calls it a 'try marker'; Clark and Schaefer (1989) calls it a 'trial constituent' when presenting a name or description that the speaker is not sure is factually correct or entirely comprehensible, and Pierrehumbert and Hirschberg (1990) calls it 'forward looking' and 'interpreting in respect to what follows'. The occurrences of H-H% in the spontaneous versions seem to be one reflection of the interaction between speaker and hearer when the speaker is producing an utterance with the hearer in mind. Note that these H-H% all occurred within the sections with one primary speaker which I was examining.

FP's single H-H% in the read speech was an explicit question, and yes/no questions typically have that pattern in American English. The intonation of the question was realized phonologically identically in the two versions. The example is given in (5). (5a) is the spontaneous version, and (5b) is the read version. The

Table 7
Boundary types.

a) Phrasing statistics.

	FP		DW	
	Spon	Read	Spon	Read
intermediate phrases	121	78	69	54
mean words/phrase	3.5	5.0	3.4	3.9
mean accents/phrase	1.9	2.6	2.0	2.2
intonational phrases	73	54	48	37
mean words/phrase	5.8	7.3	4.9	5.6
mean accents/phrase	3.2	3.8	2.9	3.2
mean intermediate phrases per intonational phrase	1.7	1.4	1.4	1.5

b) Overall distribution of intonation boundary tone sequences, by number of tokens and percentage.

	number of tokens				percentage			
	FP		DW		FP		DW	
	Spon	Read	Spon	Read	Spon	Read	Spon	Read
L-L%	36	26	33	30	49.3	48.1	68.8	81.1
L-H%	27	26	11	7	37.0	48.1	22.9	18.9
H-L%	1	1	3	0	1.4	1.9	6.3	0.0
H-H%	9	1	1	0	12.3	1.9	2.1	0.0
total	73	54	48	37				

high rising tone sequence H* H-H% is underlined for ease of comparison. My utterances are shown in italics with sentence punctuation.

(5) a. what I did while I was actually there is I was ||
an interdisciplinary studies major <283> ||
H* H* H* H-H%
you have any idea what that is <100> ||
H* H* H* H-H%
Yeah, I've heard.

b. but <152> | what I did while I <209> |
actually was there <152> ||
is I was an interdisciplinary studies major ||
L+H* !H* !H* L-H%
you have any idea what that is ||
H* H* H* H-H%
Yeah, I've heard.

Example (5) also shows another occurrence of H-H% in the spontaneous version, in the phrase before the explicit question. This H-H% is a reflection of the interaction between speaker and hearer. It is asking a question already, prefiguring the explicit question to come. However, in the read version, the first part is presented as a statement and only the explicit question has final rising intonation. A similar thing happens in example (6), one of DW's spontaneous utterances. The final high rise on *teaching* is as if to say, 'do you follow what I'm saying?', 'do you understand how inventing characters can help with teaching?'. Speaker FP also used the H-H% in example (7) as an indication that he wondered whether he remembered correctly that his friend is a physicist. Examples (8) and (9) seem to be instances of FP using H-H% to invite me to comment on what he has said or make some sort of response. Both the coder and I marked possible turns after the H-H% in examples (8) and (9). All of these examples seem to be implicit questions

(6) I | kind of invent || characters ||
to help me <697> || with my teaching ||
H* !H* H-H%

(7) he's a physis- <112> } physicist <698> ||
H* H L- H* H-H%
and works at NASA <389> ||
H* H* L-L%
Mm-hmm.
or 'e used to work for NASA ||
L+H* L-L%

(8) which is | I suppose | interesting work <1250> ||
H* L- L* L- H* H* H-H%
mm but I <64> } had | I mean ||
the stuff he knows <583> ||
is kind of amazing 'cause <1137> ||

(9) he knows uh | geography | and climate of |
just about every region | of the United States <391> ||
H* H* H-H%
Well that's convenient if ever he wants to move somewhere nice when he retires or gets sick of nuclear energy.

such as 'do you understand what I'm saying?', 'did I say that right?', etc.

All of these examples were realized as L-L% or L-H% in the read productions, except for the explicit question of example (5). These reflections of the dialogue structure of the original conversation were eliminated in the read version when H-H% was replaced by L-L% or L-H%, making the read version more like coordinated monologues rather than a true dialogue. There was no grounding or checking to see that the listener understood by the use of the H-H% high rising contour. Thus there was interaction with the listener in the spontaneous versions which was missing in the read versions. The symbolic intonation transcription captured this reflection of the difference in interactivity of the spontaneous and read texts.

6. Pitch Range

6.1 Peak fundamental frequency as an acoustic measure of pitch range

Sections 6 and 7 specifically address the role of pitch range in structuring discourse. This section describes the measure of pitch range used, and Section 7 discusses how this measure relates to the previously determined discourse structures for the spontaneous and read texts. Evidence from downstep, prominence relations, and asides show that the intermediate phrase is an appropriate domain for local pitch range (Liberman and Pierrehumbert, 1984; Beckman and Pierrehumbert, 1986; Grosz and Hirschberg, 1992; Silverman et al., 1992). Therefore, based on the intonational analysis of each text, I took the peak F0 of each intermediate phrase as an acoustic measure of local pitch range. The peak was taken to lie on an accented word and not on a phrase tone (H- phrase accent or H% boundary tone) if that happened to be the highest point in the pitch contour. Phrase tones were excluded as a measure of the pitch range because the phonological upstep of boundary tones after a H- would artifactually inflate the pitch range estimate by the amount of the upstep. The peaks on *friend* and *NASA* in Fig. 9, spontaneous version, are examples of such peaks. To minimize the effects of segmental perturbation and to provide a consistent measurement criteria, I measured the frequency at the point in time when the vowel of the accented syllable is at its maximum intensity (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992). This measure treats local pitch as a continuous variable, allowing any value of F0 and not just a discrete number of pitch levels.

Fig. 10 plots the F0 peaks of all intermediate phrases for each speaker, in a frequency histogram. The spontaneous tokens are in dark gray. The mean F0 peak is significantly higher for Speaker FP in the read version (spontaneous mean: 131.6 Hz, read mean: 144.3 Hz; $t=-2.97$, $p<.01$), but there is no significant difference in the distribution of peak F0 for Speaker DW (spontaneous mean: 120.8 Hz, read mean: 115.7 Hz; $t=1.01$, $p>.1$). The pure distribution of peak frequency alone then cannot be a reliable characteristic of the difference between spontaneous and read speech because there was no consistent difference in the means. For Speaker FP the read version had a significantly higher mean, although there was an extensive overlap in distribution, and for Speaker DW the spontaneous version had a higher, but not significantly different, mean. This is in keeping with earlier results for average frequency and frequency variation (or range) (Remez et al., 1985; Remez et al., 1986; Blaauw, 1991; Blaauw, 1992). This measure looks at F0 in a global way and neglects the potential organizational principles of pitch range changes over time. Only by looking at the F0 peaks over time can we hope to see how pitch range may be used to signal discourse organization, and perhaps discover differences between uses of pitch range changes in spontaneous and read speech.

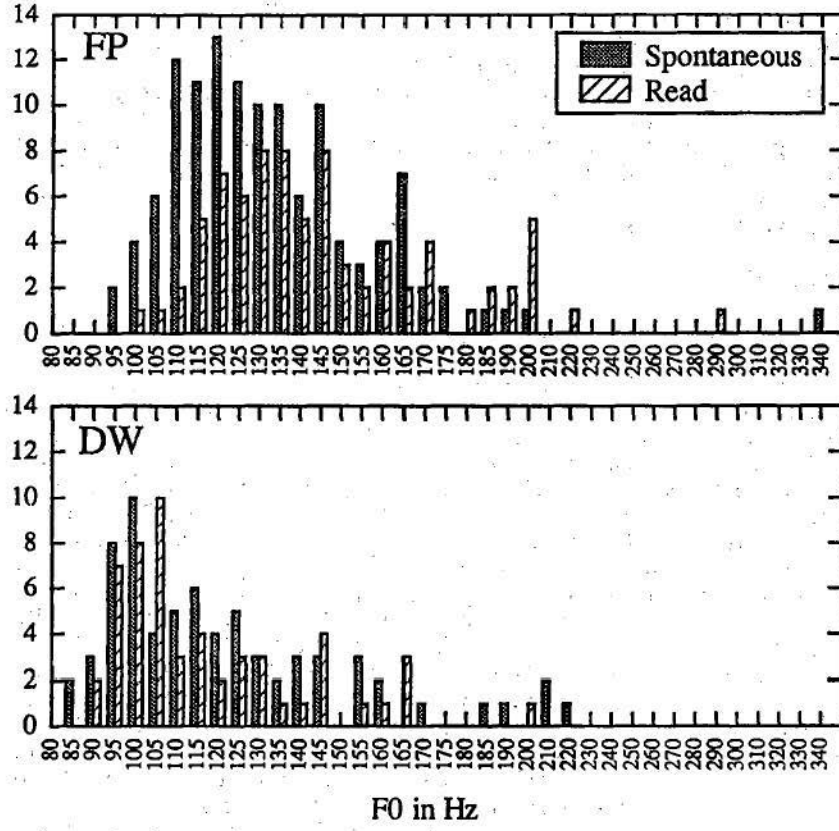


Fig. 10. Histograms of peak F0 of each intermediate phrase for Speakers FP and DW, spontaneous and read versions.

6.2 Local prominence

One thing increasing pitch range is used for is to increase the salience of a phrase or word in a phrase (Pierrehumbert, 1980). In these data of two male speakers, it seemed that accents which have peak F0 values of 150 Hz or more were especially salient. The value 150 Hz was one that I chose based on my impressions; subjectively those words seemed especially salient because of the high pitch. I considered these accents to be realized in an 'expanded' pitch range. No systematic perceptual testing was done in this study, but Ladd (1993 in press) in some preliminary experiments finds a difference in perception of high tones beginning at approximately 150 Hz as well for male speakers. Even though I have been viewing F0 peaks as lying on a continuous scale, it is possible that there are categorical aspects to the distribution as well, such as Ladd's overhigh tone, or uses of 'expanded' pitch range.

Fig. 11 shows an abstract representation of the expanded pitch range in a short discourse segment taken from the FP conversation. The boldface underlined words are the peak accents (i.e. the accents realized with the highest F0 in an

Speaker FP, Spontaneous:

a friend of mine um <216> || works for NASA ||.
he's a physis- <112> } physicist <698> ||.
and works at NASA <389> ||.
[mm-hmm]
or 'e used to work for NASA ||.
he now works for uh <963> | uh <98> || Federal Energy Regulat- <95>}
no <1597> || uh Department of Energy <197> ||.
he works for Department of Energy <175> ||.
and he <248> } visits all the nuclear | power plants in the country <953> ||.
[hmm]
which is | I suppose | interesting work <1250> ||.

Speaker FP, Read:

a friend of mine works for NASA ||.
he's a physicist || and works at NASA || or 'e used to work at } for NASA ||.
he now works for the Department of Energy <300> ||.
he works for Department of Energy || and he visits all the nuclear | power plants in the country || which is || I suppose || interesting | work <453> ||.

Fig. 11. Expanded range for matched spontaneous and read excerpts of Speaker FP's conversation 'Friend'. Boldface underlined words were realized with F0 peaks of 150 Hz or greater.

intermediate phrase) with F0 of more than 150 Hz. These high pitches draw attention to the words or phrases, perhaps as a concrete reflection of something corresponding to Grosz and Sidner's attentional structure (Grosz and Sidner, 1986). The words in expanded pitch range (that is, the boldface underlined ones) were not the same ones in the two versions. If we consider pitch range as one reflection of focus in the local attentional space, the two versions had a different pragmatic or attentional structure, since the spontaneous version focused on place and the read version on time. Thus, even though the sentences in the two versions matched lexically and syntactically, the points that were made salient over the unfolding of the discourse differed. That is true even if it is not words alone but phrases which are made prominent. However, pitch range is implicated in more than just local prominence. It also participates in topic organization and turn structure. I examine these influences in the texts with the help of the hierarchical pitch tree explained in the next section.

6.3 The hierarchical pitch tree

The observations of pitch range and discourse hierarchy and turn taking cues discussed in the introduction suggest that a decrease in pitch between phrases shows some sort of topic subordination and hence groups phrases together, whereas an increase in pitch signals a new unit of some sort, such as a new topic or a new turn. To investigate these predictions and test them against my spontaneous and read speech data, I constructed hierarchical pitch trees. These trees were based on high pitch heads which dominate lower pitch phrases. These trees were specifically designed to capture relationships between phrases in which relationships of increasing pitch between phrases work to divide discourse into different segments and relationships of decreasing pitch between phrases signal coherence between the phrases. That is, if pitch range increases at new topic boundaries and new turns, these boundaries should be captured by a division into separate trees. On the other hand, if an increase in pitch range is used for other purposes besides marking boundaries between discourse segments, we would not expect these trees to capture those relationships clearly. For example, if certain kinds of relationships between phrases are made by increasing pitch from one

phrase to the next instead of by decreasing pitch, the grouping predicted by the trees based on decreasing relationships would be a mismatch with what should be grouped together. The pitch trees impose a segmentation upon the discourses, which I called the pitch tree segmentation.

I considered these trees to be a kind of phonetic structure which captures in a gradient way which phrases are grouped together by decreasing pitch relationships. No *a priori* categories of pitch ranges (e.g. low, mid, high) were assumed. However, these could be assigned later if such a categorization seemed appropriate (see Bruce and Touati, 1992, for work which uses such a categorization). The phonetic structure can be interpreted later, much as a fundamental frequency contour is a phonetic representation which can be interpreted phonologically in terms of accents and phrases. If rising pitch relationships between phrases are uncovered as well, then clearly a richer structure which captures increasing as well as decreasing relationships is called for.

Hierarchical pitch trees were constructed from the peak pitch values of each intermediate phrase in a text (peak measurement criteria as described in Section 6.1). The peak F0 of each intermediate phrase was taken as an acoustic measure of the pitch range for each phrase. This algorithm built hierarchical pitch trees based on the principle that a high pitch dominated all subsequent lower pitches until the next local increase. That is, phrases with subsequently decreasing pitch ranges were grouped together, and phrases where pitch range increased were divided into separate groups. The first higher pitch value in a sequence started a new group. Three levels of groupings were constructed. The first level of trees, Level 1, was based on the measured peak of each phrase. The next two levels, Level 2 and Level 3, were based upon the highest values of each tree in the immediately lower level. The value of the highest daughter became the value used for building the next level of the tree. So, the values for Level 2 were the highest values from the level-1 trees, and the values for Level 3 were the highest values from the level-2 trees. The nested levels of trees captured the large increases in pitch appropriate for changes in topic and the like.

Let us illustrate the step-by-step construction of a hierarchical pitch tree using this algorithm with the example given in Fig. 12, an excerpt which was taken from the read version of FP's conversation. At the left are the intermediate phrases of the text; the underlined words are the accented words which have the peak F0s. The column labeled Peak F0 shows the F0 measured in Hz for the underlined words. The trees at each level begin with a frequency value which is higher than

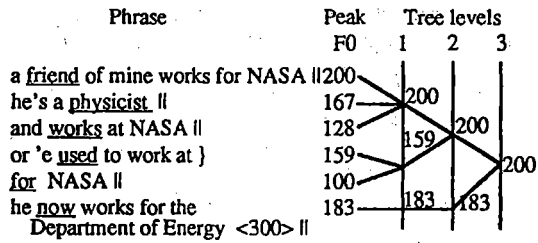


Fig. 12. Building a hierarchical pitch tree. The underlined words are the accented words on which the peak F0 for each phrase was realized. The peak F0 value for the phrase is given in Hz in the column Peak F0. Trees were constructed from these values according to the algorithm described in the text. From Speaker FP, Read: 'Friend'.

the following value. The first level-1 tree begins with the first node, which has a value of 200 Hz. This node dominates the next two nodes, which are realized on progressively lower frequencies. Note that the trees need not be binary branching. A new tree on any level begins at a local increase in frequency values. For example, the second level-1 tree begins with a node of 159 (the peak on *used*), which is greater than the previous node of 128. The 159 is greater than the next node value of 100, so the two nodes are together in a tree. Similarly, the third level-1 tree begins with the node value of 183 (for *now*) because 183 is greater than the previous node of 100. This tree happens to be a tree with only one branch. Levels 2 and 3 proceed similarly, with the value of the highest daughter becoming the value for the next level. So, the first level-2 tree dominates the first two level-1 trees, which have values of 200 and 159 respectively. The second level-2 tree is again the single branch dominating 183. At level 3 there is a single tree which dominates the inherited values of 200 and 183.

7. Comparing discourse and pitch tree segmentations

The trees presented in Figs. 13-18 are schematic but represent the full structure of the pitch segmentation trees for selected parts of Speaker FP's conversation and Speaker DW's conversation presented in Figs. 1-6. Triangles represent selected full trees at the three different levels, neglecting the internal structure of the trees. The pitch values that head the trees are circled in the figures. Heavy lines show the divisions into "paragraphs" that the coder marked. As in Figs. 1-6 my utterances are shown in shaded boxes. The relationships marked by arrows labeled 'C' and 'I' show rising pitch relationships for corrections marked by the coder and what I am calling introductory phrases (see below). We will look at the spontaneous and read versions of two different sections of Speaker FP's conversation ('College' shown in Figs. 13 and 14, and 'Friend' shown in Figs. 15 and 16) and one section of Speaker DW's conversation ('Fernblaster' shown in Figs. 17 and 18).

I will be looking at these data specifically to test my hypotheses that the pitch range cues to topic structure in the read speech versions conform fairly well to what has been suggested, i.e. that pitch range is expanded at the beginning of new topics and decreased for related subtopics, but that pitch range conveys the hierarchical discourse structure of the spontaneous speech versions less well. I expect to find that real-time production phenomena such as floor negotiations, corrections, and false starts disrupt clear topic organization. These may complicate the role pitch range plays as a cue to topic structure, because they themselves may have manifestations in the pitch ranges used. That is, there should be differences between the two versions in how well the pitch range reflects the discourse structure because the read speech versions were reorganizations of the spontaneous speech versions (since they came from the same texts), ones which were free of the complexity of floor negotiations (since the turns were predefined) and false starts (since those were removed in the preparation of the texts).

Recall from Section 3 that the discourse segmentations labeled by the coder differed substantially between the spontaneous and read versions in 'College' for Speaker FP. We will see that the pitch tree segmentations also differed substantially between the two versions, and in fact matched the discourse segmentations quite well. However, for Speaker DW, neither the discourse segmentations nor the pitch tree segmentation for the spontaneous and read versions differed dramatically. The discourse segmentations suggest a substantial reorganization of the topic structure from the spontaneous speech to read speech version in FP's conversation, but a considerably lesser reorganization in DW's conversation, and these differences between the two speakers seem to be reflected in comparable relationships between the pitch trees of the paired versions of text.

7.1 Speaker FP

7.1.1 'College'

Figures 13 and 14 show the spontaneous and read versions of the 'College' part of FP's conversation. It is clear that the major divisions into paragraphs in these two versions are different, and this is a reflection of the fact that different things were emphasized in the two versions. In the spontaneous version many of the subpoints flowed from one to the other (witnessed by the lack of paragraph breaks), while in the read version some of the transitions between points were abrupt enough for the coder to assign them to separate paragraphs. An examination of the pitch trees associated with the read version shows that each of paragraphs 2, 3, and 4 have their own pitch trees. They were all headed by pitch ranges with values of approximately 200 Hz. There were also relatively long pauses, from 434 ms to 802 ms, at these boundaries between the paragraphs. Recall from Fig. 1 that the mean pause duration in the read version was 322 ms, so these are well over the mean. It seems reasonable to assume then that the combination of a regular pause and a large increase of pitch are cues to a strong discourse boundary. This is exactly what Hirschberg and Grosz (1992) found in their AP news reading.

We can also see in the read version the tendency for a fairly hierarchical structure indicated by decreasing pitch range for supporting details of the paragraphs. For example, in paragraph 2, lower level trees headed by local increases seem to correspond nicely to the supporting details. He knew he wasn't going to be a cartographer and registered for Spanish (200 Hz). Because he was taking Spanish, the advisor suggested introductory linguistics (152 Hz) which was a course like our 201 (127 Hz). Then there is another fact, that is, Dr. Thomas Field taught the course (147 Hz). Paragraph 3 has a similarly nice structure with decrease pitch ranges for the subpoints, with the exception that the third level-2 tree peak (156 Hz) has a larger value than the second level-2 tree (133 Hz), but it is still less than the level-3 head (198 Hz).

The most striking difference between the topic structure of the two versions is this part concerning Thomas Field. In the spontaneous version, the speaker mentioned Thomas Field as the instructor of the course without making a strong point of emphasizing who he is, whereas the read version specially highlighted Thomas Field. In the spontaneous version, the pitch tree for the section mentioning him as the instructor of the course had a peak of 127 Hz, and the first mention of Thomas Field had a peak value of 125 Hz. He mentioned Thomas Field a second time along with his title, also in a low pitch range (peak 111 Hz), and then that Mary knows him because he went to Cornell. He elaborated essentially as a parenthetical that Thomas Field graduated from Cornell and not only went there. The only accent on the phrase *graduated from Cornell* was on *graduated*, and this was realized with a peak of 108 Hz. The whole phrase was uttered with low intensity and at a relatively fast pace. All of the vowels which were not devoiced or deleted had z-score durations between 0 and -1. All of these cues together make the parenthetical type meaning of *graduated from Cornell* clear to the listener, and the coder assigned no paragraph breaks separating the discussion of Thomas Field from the previous material. However, the read version specially highlighted Thomas Field, and the discourse segmentation and the pitch tree segmentation both reflect this. The rise in the pitch range (from 139 Hz to 198 Hz) and the pause between the first and second mention of Thomas Field in the read version signaled a clear separation from the previous mention of him as the instructor of the course. A separate paragraph was devoted to him, and one of the points made about him was that he graduated from Cornell. In this version of the phrase *graduated from Cornell* the standardized vowel durations were all over 0, and the final vowel of *Cornell* had a z-score of 2, with the word accent on *Cornell* and phrase final

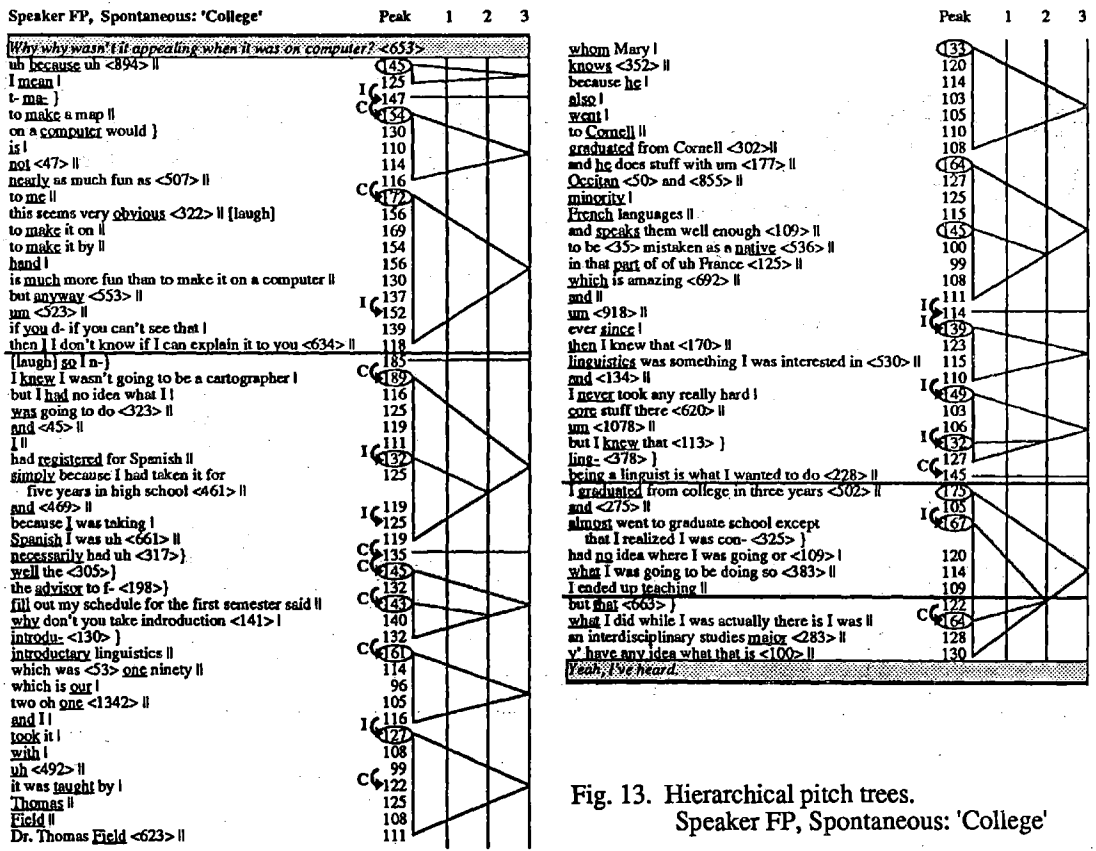


Fig. 13. Hierarchical pitch trees.
Speaker FP, Spontaneous: 'College'

lengthening contributing to the long length. A paraphrase of the meaning of this part is something like the following. The man who Mary knows because she went to Cornell with him, graduated from Cornell. The listener can tell the difference in meaning between the two versions quite easily due to the pitch range relationships, pause structure, and tempo.

These differences in emphasis are not solely due to a difference in spontaneous versus read speech because differences in emphasis might equally be true for different instances of spontaneous speech. However, the fact that in the read version the paragraph boundaries marked by the coder corresponded regularly with long pause durations and extreme pitch rises suggests that the reader produced a clearer indication of the discourse structure in the read version as compared to the

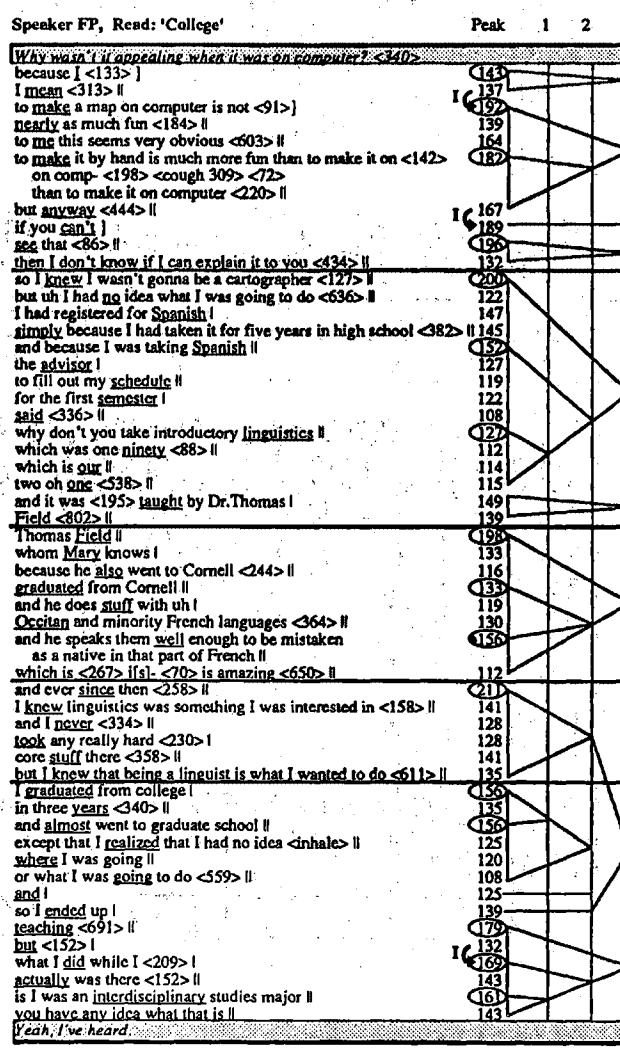


Fig. 14. Hierarchical pitch trees. Speaker FP, Read: 'College'

spontaneous version. I would claim that this is an instance of exactly the kind of reorganization and simplification of discourse structure that I expected to find between the spontaneous and read versions.

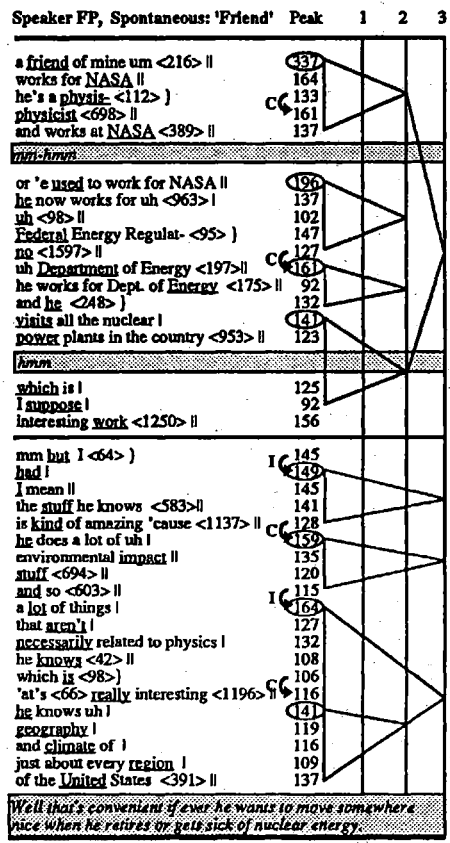
7.1.2 'Friend'

Figures 15 and 16 show the spontaneous and read versions of the 'Friend' part of FP's conversation. The first paragraph of the spontaneous version ends with the sentence which was given as example (8) above, where the high rising intonation was an indication that the speaker expected verbal feedback from me. However, I didn't give it to him, and he continued speaking after an extremely long pause of 1250 ms. He made a few false starts before he started speaking fluently again. The trees in this section after the long pause were headed by increasing pitches: 149, 159, 164. It seemed that FP started out at one pitch range and increased the pitch with each subsequent attempt at topic, a kind of topic reset. The phrases grouped together by the first pitch tree after the pause (headed by 149 Hz) turned out to be a false start which ended with the phrase *the stuff he knows is kind of amazing 'cause*. This phrase had a peak of 128 Hz and a following 1137 ms pause. The next phrase *he does a lot of* was a new attempt after that false start. It had a peak of 159 Hz, which was higher than that of the immediately previous phrase and was also higher than the peak of the whole group which included that phrase. The tree headed by 164, the highest of all the peaks in this section, seems to be his main point, that his friend knows a lot of things. Notice that the tree at level 3 headed by 164 spans a pause gap of 1196 ms after a complete unit, suggesting another possible turn transition point. After the pause, FP raised pitch locally (from 116 to 141 Hz) again as a mark of starting a new topic or a new turn, but not as high or higher than 164 Hz, the peak of the section to which it was topically related. French and Local (1986) note that pitch is raised for competitive turn taking. These data suggest that pitch is also raised (or reset) after a point when a turn could have taken place even when the other speaker did not compete for a turn. Such a turn transition point is an appropriate place to either provide more information as a subtopic or elaboration of the previous topic, and thus make a smaller rise in pitch, or to suggest a new topic, and thus make a larger rise in pitch.

The topic structure of the last part of the read speech, corresponding to the second paragraph in the spontaneous version, seemed to be something like this. The main topic of discussion was the stuff the friend knows. He knows more than physics; specifically he knows geography and climate. *Geography* and *climate* were relatively more prominent than *physics* (realized with a peak of 169 Hz on geography as opposed to 147 Hz on physics), but they were both examples of what he knows. Instead of simply having hierarchical subtopics, this section had levels of parallelism expressed in the pitch ranges. The two versions shared the topic organization that geography and climate are examples of things that he knows. However, in the read version they were given in explicit comparison with physics, whereas in the spontaneous version they seem to have been details added partly because I did not take the floor.

7.1.3 Introductory phrases

The correspondence between the auditory discourse segmentation and the pitch tree segmentation are nearly identical in the read version of the section 'College'. Furthermore, the predicted relationship of decreasing pitch range with subtopic structure seemed to hold fairly well in the read versions. In the spontaneous section 'Friend' an interesting connection with possible turn transition points and pitch trees were shown. However, while the pitch trees grouped phrases together into paragraphs quite well, they did not always group phrases of sentences together correctly. One specific type of situation where the pitch tree



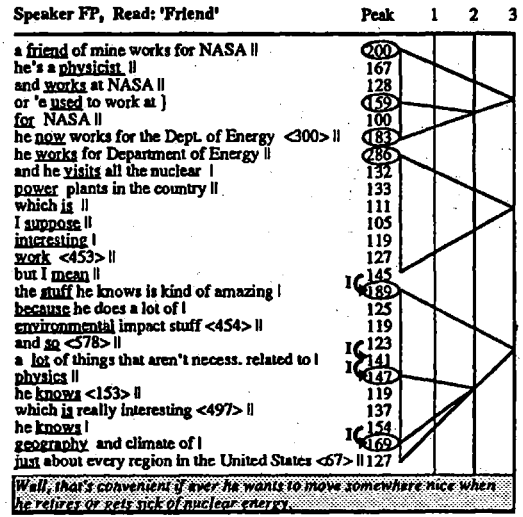


Fig. 15. Hierarchical pitch trees.
Speaker FP, Spontaneous: 'Friend'

Fig. 16. Hierarchical pitch trees.
Speaker FP, Read: 'Friend'

algorithm made wrong predictions about which phrases to group together were sentences which did not start out with a phrase realized in the highest pitch range for the sentence. Introductory phrases such as *and, and so, um, but I mean,* and the like, were generally realized in lower pitch ranges than the more content-containing part of the sentence. These were more common in the spontaneous speech than the read speech, but they occurred in both versions. The tree building algorithm as it was defined did not group such introductory phrases together with the following phrase as it should have, but rather grouped them with a previous phrase which had a higher pitch range. Examples of this sort are marked with arrows labeled 'I' (for 'introductory') in Figures 13 to 16. This sort of relationship between phrases seems to be related to the intonation of cue phrases. Cue phrases introducing a phrase are often realized with a L* accent, and they are therefore realized at a lower frequency than the phrase that they are in relationship to (Hirschberg and Litman, 1987). These introductory phrases were realized with H* accents and not L*, but the increasing pitch relation between an introductory phrase and a following larger discourse unit was similar.

7.1.4 Corrections

The pitch tree segmentations did not correspond as well to the auditory discourse segmentation in the spontaneous versions as they did in the read versions. One thing that is apparent from the spontaneous speech versions is that they were full of false starts, corrections to mispronunciations of words, and irregularly distributed pauses of various lengths. In stark contrast to the read versions, the spontaneous versions were riddled with such reflections of the unprepared nature of the text. The speaker did not know ahead of time what he was going to say, and had to create it on-line as he spoke. Sometimes the speaker made mistakes and had to correct what he said to what he intended. Each of the arrows marked with 'C' (for 'correction') were places where FP aborted a false start and started anew or repeated a word in a new phrase as a correction. Compare the spontaneous versions (Figs. 13 and 15), which had many corrections between phrases with the read versions (Figs. 14 and 16), which did not have such corrections. Recall that the discourse codings discussed in Section 3 listed corrections for the read version. These corrections were of a different nature, however, with the corrections being corrections for the most part being within the same phrase and simple repetitions of a word which was stumbled over in reading. In the spontaneous speech, a correction was almost always uttered with a higher pitch than the word or phrase corrected. That is, there was a local increase in pitch between the phrases. There were a few examples of corrections between phrases being uttered on a lower pitch, such as the *went to Cornell, graduated from Cornell* example, but these were parenthetical additions of information and do not feel like true corrections.

There were several examples of such increases in pitch range for false starts. The spontaneous example described in Section 7.1.2. 'Friend' was such an example. The increase from 122 to 164 Hz in the false start sequence *but <u>that</u> <663> <u>what</u> I did while I was actually there is I was* in the last paragraph of 'College' is another example of such an increasing relationship in false starts. The false start in the first paragraph of 'College' ending with the phrases *is not nearly as much fun as <507>* which was aborted and then corrected by a new approach beginning with the phrase *to me* also had an increasing pitch relationship. The last phrase of the false start had a peak of 116 Hz and the new start beginning had a peak of 172 Hz. A false start reformulation *<u>ling</u>- <378> <u>being a linguist</u>* had an increase from 127 to 145 Hz. The peak pitch for each phrase in the string of false starts *because I was taking <u>Spanish</u> I was uh <661> <u>necessarily</u> had uh <317> <u>well</u> the* increased from 119 to 135 to 145.

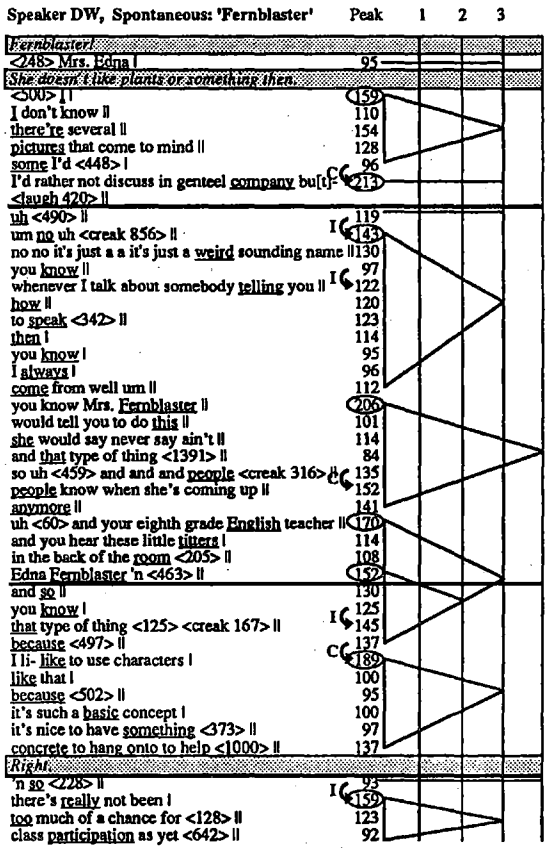
Corrections at the level of the word also exhibit this kind of increasing pitch range relationship. The second mention, the correction, was realized on a higher pitch than the first, incorrect, mention. The correction can be due to incorrect or incomplete pronunciation the first time, such as *introdu- <130> introductory* with an increase from 132 to 162 Hz, *t- ma-- make a map* with an increase from 147 to 154 Hz, and *physic- <112> physicist <678>* with an increase from 133 to 161 Hz (from 'Friend'). Factual corrections also have this sort of pitch relationship, such as the example *Federal Energy Regulat- <95> no <1597> uh Department of Energy <197>* from 'Friend'. There is an increase from 147 to 161 Hz from *Federal Energy* to *Department of Energy*, with *no* at 127 in between. This increase of pitch range seems to be a quite general tendency and a way to mark a new beginning of a correction. These local increases for corrections however disturb the trend for hierarchical topic organization to be marked by decreasing pitch range relationships within a topic group and an increase at the beginning of a new topic group. We might view this kind of pitch increase for a correction as one cue that the listener might take advantage of in recovering the final form of what was intended, as Clark and Schaefer (1989) say listeners can.

7.2 Speaker DW

Figures 17 and 18 show the spontaneous and read versions of part of the 'Fernblaster' part of DW's conversation. Both the discourse segmentations and the pitch tree segmentations were quite similar for the spontaneous and read speech versions. One mismatch between the trees and the discourse segmentation was the division between the second and third paragraphs. The division between the second and the third paragraph did not align neatly with the pitch trees in either version. However, for the spontaneous version there was a 436 ms pause at the end of the second paragraph, and the next few phrases could be taken as introductory phrases to a new point. For the read version there was a short pause of 103 ms at that boundary and a local pitch increase from 110 to 143 Hz.

The rest of the pitch trees were quite similar in the two versions. In the first paragraph the tree was headed by 159 Hz in the spontaneous and 156 Hz in the read. In the second paragraph there were a few trees, which were headed by quite high peaks on the phrases *it's a weird sounding name, Fernblaster*, and *and your eighth grade English teacher* in both versions. Essentially he was role playing and quoting himself and his students, and he used the same sort of changes in pitch range to signal that in both versions. His background comments were uttered with smaller ranges, with peaks of less than 115 Hz, such as the phrases *and you hear these little titters in the back of the room.* The fourth paragraph began with almost exactly the same peak pitch in both versions (neglecting the introductory phrase in the spontaneous version), 159 Hz for the spontaneous and 161 Hz for the read version. So, not only did the pitch tree segmentations essentially match in the two versions of this part of the conversation, the values of the peaks were also nearly identical, signaling parallel emphasis in the two versions. In both the spontaneous and read versions Speaker DW was re-enacting the scene from his class by quoting himself and his students, partly by use of high pitch ranges in the quoted phrases. Using pitch range for quoting in this way disrupted hierarchical topic structure but revealed very similar use of pitch ranges in the two versions.

Just as for Speaker FP, Speaker DW had examples of introductory phrases that were not grouped with the following phrase by the pitch tree as they should have been, but instead with the previous phrases in both the spontaneous and the read versions. Again these are marked in the figures with arrows labeled by 'I' for introductory. The read version had no examples of corrections, but the spontaneous version did, and they are marked with arrows labeled by 'C' for correction in the figure. He repeated the word *people* in the two subsequent



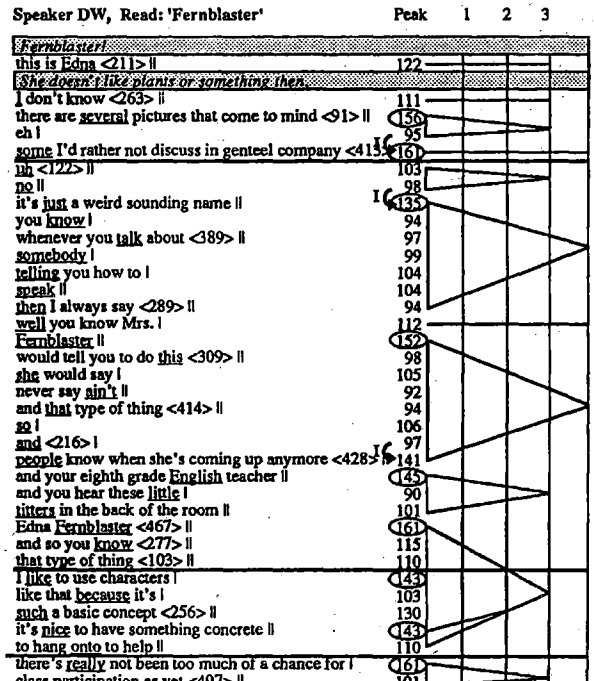


Fig. 17. Hierarchical pitch trees. Speaker DW, Spontaneous: 'Fernblaster'

Fig. 18. Hierarchical pitch trees. Speaker DW, Read: 'Fernblaster'

phrases *so uh <459> and and and people <316> people know when she's coming up* with an increase of pitch from 135 to 152 Hz from the first to the second mention. He corrected a false start beginning with *because* by starting up again after a 502 ms pause with the phrase *I li- like to use characters* with a peak of 189 Hz. The especially high peak could be due to both making a correction after a false start and correcting a mispronunciation of *like* by a second mention.

The pitch trees shown for this section had trees with four levels instead of three levels in order to group together the phrases with the lower peaks that occur between the large peaks. If the pitch tree building algorithm had another criteria of what counted as a 'local increase' (say, for example, that 5 Hz variations are not essentially different values and should not be considered a 'local increase', i.e. they should be considered a tie), three levels of trees would be sufficient to group together what was intended.

7.3 Summary

There were differences between the spontaneous and read speech in how well the discourse segmentations and the pitch tree segmentations matched. They matched better in the read versions than in the spontaneous versions. In Speaker FP's 'College' section, the discourse segmentation and the pitch tree segmentation were almost exactly identical. This section also showed the tendency for hierarchical topic organization to be reflected in the pitch ranges as well. New large topics had increasing pitch ranges at the beginning, and related subtopics had generally smaller pitch ranges.

However, there were also introductory phrases which introduced new topics and sentences, and these were not always grouped together with the appropriate phrases. These phrases were realized with lower pitch range than the following, more content-rich phrases. Because the pitch tree algorithm was designed to treat increasing pitch relationships as division points between units, these introductory phrases were often grouped with previous, higher pitch phrases rather than following, higher pitch phrases as they should have been.

For Speaker DW there were examples of using high pitch ranges for quotes and low pitch ranges for parentheticals. The corresponding phrases in the two versions were treated as quoted and parenthetical material, but again these uses of pitch range did not match with projected hierarchical topic structure. However, since the use of pitch range was similar in the two versions, the pitch trees segmented the discourses very similarly.

The spontaneous versions also had corrections (false starts and word repetitions) that were realized with increasing relationships between phrases. These corrections interrupted the pitch cues to topic subordination, but corrections were expected because the spontaneous versions by their very nature were unplanned and unrehearsed. So, to a certain extent, the topic structure was not as clear to begin with in the spontaneous speech. A further complication to the topic structure in the spontaneous conversations was the possibility of turn taking, since these were two person conversations. It seemed after a point when the other speaker could have spoken but did not, the original speaker also raised the pitch. I propose that the spontaneous conversations were organized both in terms of topic structure and turn taking, with some turns following more easily than others, and the read were organized more in terms of preplanned sections. Speaker FP's read version of section 'College' was a clear example of a reorganization of the contents of the spontaneous version. It was a reading made with knowledge of what was coming up next and how long each turn was to be and without hesitations, false starts, and other corrections.

The discourse segmentations were quite different for Speaker FP between the spontaneous and read versions of the same conversations, even though the

words were nearly identical in the two versions. This means that things were grouped together differently and given different emphasis in the two versions. Pitch range relationships did reflect the differences in discourse structure. The paragraph boundaries in FP's read version corresponded to regular pauses and quite large pitch range expansions, and in FP's spontaneous version again it was at the points of largest pitch range expansions that paragraph divisions were marked. However, the discourse segmentations were nearly identical in the two versions for Speaker DW. We could interpret this as saying that Speaker FP changed the topic relationships more between the spontaneous and read versions than Speaker DW did. Speaker DW seemed to have more or less preserved the organization of the original spontaneous conversation, judging from the discourse segmentations and similar use of pitch range.

8. Perception test

The materials used in this studied differed from the spontaneous versus read speech used in such studies as Remez et al., 1985; Remez et al., 1986; Blaauw, 1991; Blaauw, 1992. Since the read speech was a connected discourse based on spontaneous speech and was deliberately read with the aim of trying to make it sound spontaneous, I wondered how well the readers had succeeded in their task. That is, was the read speech perceived as spontaneous or read? In addition, the two speakers differed dramatically in the extent to which the pitch range patterns reflecting topic organization corresponded between the spontaneous and read versions of the conversations. For Speaker FP the two versions were very different, while for Speaker DW they were very similar. I wondered if these differences between speakers could be partially explained by characteristics of the read speech versions. Specifically, did DW remember or recreate the structure of the spontaneous speech in his read version more so than FP did in his (as the use of pitch range would lead us to believe), and if so, was Speaker DW's read speech more spontaneous sounding than Speaker FP's? If this were true, then we would expect to find excerpts from DW's read version perceived as spontaneous more often than excerpts from FP's read version would be perceived as spontaneous. Finally, I wondered if longer excerpts were more often correctly identified as spontaneous or read than shorter excerpts. Longer excerpts may contain more cues to the spontaneous or read nature of the text than shorter excerpts because the larger amount of material is more likely to contain hesitation phenomena in the spontaneous, etc., and may reveal to the listener differences in patterns of transitions between phrases and topics in the two modes of speech. To address these questions, I designed a perception test to test how well listeners could correctly identify excerpts of these spoken conversations and the reenacted read speech as spontaneous or read. Listeners were presented with utterances from each speaker and different lengths of utterances.

8.1 Method

Subjects. Twenty eight undergraduate linguistics students volunteered to participate in the experiment. All were native speakers of American English and none reported any hearing impairment.

Stimuli. The spontaneous and read conversations of both speakers were segmented into utterances one sentence long, three sentences long, and five sentences long. The three sentence and five sentence long utterances overlapped by one sentence at the beginning and one at the end with other members of the series. Thus each of the single sentences occurred at least once and at most twice in the three sentence utterance set and the five sentence utterance set.

Design and procedure. A stimulus tape consisting of two parts was prepared, and items were presented in blocks of 10, with a three second interstimuli

interval. Listeners could take a break between the two parts. Part I included 110 one sentence long utterances in random order (2 speakers x (29 + 26)), and Part II included 78 three and five sentence long utterances in random order (2 speakers x ((13+13) 3 sentences + (7+6) 5 sentences). There were 28 listeners and 188 items, for a total of 5264 responses.

Listeners sat in a soundproof room listening to the stimulus tape over headphones and for each item circled either 'spoken' or 'read' on an answer sheet. 'Spoken' meant they thought the excerpt they heard could have come from a naturally occurring conversation between two friends, and 'read' meant that they thought that the excerpt they heard came from a reenactment of a conversation, read from a transcript of a naturally occurring conversation. They were told that the readers were trying to make the reading sound as much like a spontaneous conversation as possible, so that it might be difficult to tell whether it was spontaneous or read. They were told they were there to judge how well the readers had done in reading naturally. The task took approximately 40 minutes.

8.2 Results

Chi-squared tests showed that there was a significant effect of speaker on perception of the utterances as spontaneous or read. More of DW's utterances were perceived as spontaneous than were FP's ($\chi^2(1)= 159.14$, $p < .01$). Listeners perceived 68% of DW's utterances as spoken and 51% of FP's as spoken. In actuality, half were spontaneous and half were read for each speaker. Fig. 19 shows these judgments in the columns labeled 'perceived as spontaneous' and 'perceived as read'. The columns labeled 'ss' are the spontaneous utterances which were perceived as spontaneous, and 'rs' are the read utterances which were misperceived as spontaneous. The columns labeled 'rr' are the read utterances which were perceived as read, and 'sr' are the spontaneous utterances which were misperceived as read. The three different shaded columns in each of the categories

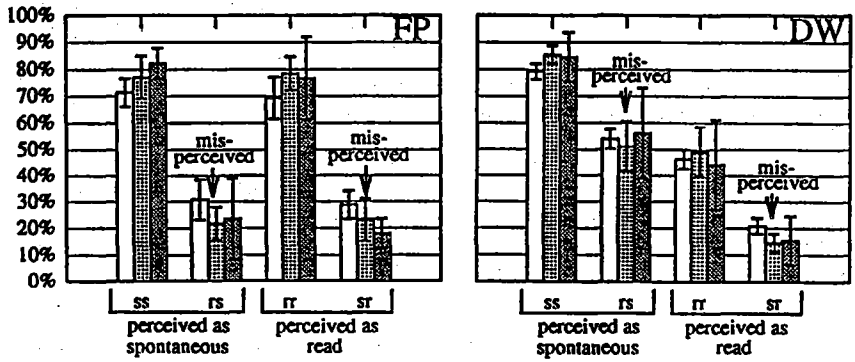


Fig. 19. Perception results in judging material spontaneous or read for the two speakers. (left) Speaker FP, (right) Speaker DW. Many more of DW's read utterances were misperceived as spontaneous utterances than were FP's.

ss: spontaneous perceived as read, rs: read perceived as spontaneous, rr: read perceived as read, sr: spontaneous perceived as read. Three different lengths of material are shown: white fill is 1 sentence long, medium fill is 3 sentences long, darkest fill is 5 sentences long. 95% confidence intervals are marked.

show the three different lengths of utterances. The white fill represents 1 sentence long utterances, the medium fill 3 sentence long utterances, and the darkest fill 5 sentence long utterances. As we can see from the columns labeled 'rs', many more of DW's read utterances were misperceived as spontaneous than FP's. We also see that fewer of DW's spontaneous utterances are misperceived as read as compared to FP. The percentage perceived correctly as spontaneous or read was significantly different for the speakers, 73% for FP and 64% for DW ($\chi^2(1)= 51.9, p < .01$).

For Speaker FP, length of the utterance (i.e. 1 sentence, 3 sentences, or 5 sentences) had a significant overall effect on the number of correct judgments made (i.e. spoken when spoken, read when read), (χ^2 (2) = 27.2, $p < 0.01$). Further analysis revealed that the difference was significant only for the shortest (1) vs. longer (3, 5) utterances (1 vs. 3: $\chi^2 = 17.3$, $p < 0.01$; 3 vs. 5: $\chi^2 = 0.24$, $p > 0.1$). Longer utterances were judged correctly more often. This could be because more turn taking clues are provided in the longer utterances. However, for Speaker DW, length of utterance had no significant overall effect on the number of correct judgments (χ^2 (2)= 3.612, $p > 0.05$). That means that there was no significant difference for the shortest vs. longer utterances for this speaker. Utterances were no more likely to be perceived correctly when they were longer.

8.3 Discussion

The perception test revealed that there were significant differences between the two speakers as to how the speech materials were perceived. Speaker FP's utterances were judged correctly on average about 73% of the time, with more of the longer utterances (three and five sentences) judged correctly than the shortest (one sentence) utterances. However, only 64% of Speaker DW's utterances were judged correctly, and there was no significant effect of length of utterance. More of DW's read utterances were misperceived as spontaneous than were correctly perceived as read. This was not true for Speaker FP. This lends support to the interpretation that DW read more naturally than FP and succeeded in producing a read text that sounded quite spontaneous. The fact that the shorter utterances were more often misjudged than the longer utterances for Speaker FP could mean that longer excerpts from the conversations or readings gave the listeners more clues to the true mode of speech. However, for Speaker DW the listeners did not categorize longer excerpts correctly more often, perhaps meaning that DW succeeded in reenacting spontaneous relationships between phrases in the read speech.

The results of these listening tasks are clearly at odds with the claims that listeners know immediately whether they are listening to natural spontaneous speech or read speech. Perhaps it would be more realistic to say that people do not really know whether an utterance is spontaneous or read, but that they make judgments early. For example, in a gating experiment with Dutch (Blaauw, 1992), listeners were able to classify utterances as spontaneous or read about 82% of the time when given the full sentence, but as well as 63% given the first two syllables and 75% given the first 6 syllables. Since in my experiment listeners were only correct on average 73% or 64% depending on the speaker given full sentences and even several sentences together, we must say that the differences between spontaneous and read speech are not as clear-cut as they might at first seem. It seems more likely that there is a continuum between clearly spontaneous and clearly read speech, with differences in style being quite important. Hesitations, false starts, long pauses and the like are prototypical of spontaneous speech, but spontaneous speech does not have to be disfluent. Read speech is often syntactically distinct since it is based on written texts. Since the read speech in this task was based on spontaneous speech, the syntax was more typical of spontaneous

speech than a read text. The additional instruction to the readers to read the transcript to make it sound like a spontaneous conversation further blurred the edges between the two kinds of speech.

Perhaps the comparison between the two speech styles in this study would benefit from being considered in terms of the scales unprepared versus prepared or unrehearsed versus rehearsed. On those scales, Speaker FP's spontaneous narrative was less prepared than DW's spontaneous narrative, because DW has spoken about his teaching methods and the use of the character Mrs. Fernblaster before. We have talked about teaching experiences together before, and have had conversations discussing pedagogical methods. So, with more rehearsal still, DW's read version is not likely to change its organizational structure as much as FP's read version of a story which he had not told before. There are also individual differences between people and their acting ability, and hence how well they can reenact a conversation and make it seem natural.

9. Discussion

Two different spontaneous conversations were recorded and reenacted as read speech by the original speakers. A listening test involving categorizing excerpts from these conversations as spontaneous or read showed that accurate identification was not entirely straightforward. Many of the read utterances were perceived as spontaneous, and some of the spontaneous utterances were perceived as read. Many more of Speaker DW's read utterances were perceived as spontaneous than Speaker FP's. Apparently skilled readers reading material based on spontaneous conversation can succeed to a certain extent in producing utterances that sound convincingly spontaneous.

The patterns of results for the two speakers were not identical, which is not particularly surprising, given that the listening test determined that the two speakers succeeded to different degrees in producing read speech that sounded like spontaneous speech. Several acoustic measures were made to see if they distinguished the two versions. Pause duration measures revealed that both speakers had similar pause duration distributions, with a higher mean and larger standard deviation of pause duration in the spontaneous than in the read speech. These results match previous findings. A measure of fundamental frequency, the mean F0 peak per phrase, distinguished Speaker FP's spontaneous from read speech, but it did not distinguish Speaker DW's spontaneous and read speech. Measures of average F0 and F0 range have found different relationships depending on language and the specific materials used; this speaker difference is another such result.

A symbolic phonological intonational analysis found some consistent patterns in the differences between spontaneous and read speech. The phrases in the read version were longer on average than the phrases in the spontaneous version. The transcription also showed that there was no use of the H-H% high rising contour as grounding or checking to see that the listener understood. It seems that there was interaction with the listener in the spontaneous version which was missing in the read version. The read speech lacked the hallmarks of interactivity in the spontaneous speech except the ones that are inherent to the text (change of speaker, explicit questions). We could say then that the read version was more like coordinated monologues rather than a true dialogue. The read version was like the spontaneous minus true interaction between the speakers.

The discourse organizations were clearly different between the spontaneous and read versions for Speaker FP, but they were relatively similar for Speaker DW. The pitch tree algorithm (based on measures of the peak pitches of all intermediate phrases) provided a method for comparing the organizational structure of matched spontaneous and read speech discourses. It provided a way of testing the

predictions about how pitch range is used to signal topic structure. The segmentation that the pitch tree algorithm imposed upon the discourses corresponded quite closely to the discourse segmentation that the independent coder assigned to the discourses. The best match between the pitch tree segmentation and the discourse segmentation was in the read version of Speaker FP's section 'College'.

Although the spontaneous and read versions were nearly identical in terms of syntax, different items were marked as salient, and topics were grouped together differently. The read versions were grouped into sections with relatively clear hierarchical topic structures. The spontaneous versions showed some evidence of hierarchical topic structure, but they also had disruptions to these topic organizations due to false starts, corrections, and the influence of possible turns. I hypothesize that the planned production units differ between spontaneous and read speech. I propose that spontaneous conversations are organized both in terms of topic structure and turn taking, with some turns following more easily than others, and read conversations are organized more in terms of preplanned sections. In the read versions, the readers know exactly what is coming up and do not have to negotiate for turns with the conversational partners. This gives them more control over deciding what relationship to give to the various topics. One meaning of pitch increase seems to be a reflection of the start of a new unit, whether it is a new topic or a new turn.

The pitch tree segmentations, together with the discourse annotations, showed that pitch increased in these discourses at the beginning of new topics and at the beginnings of new turns, or potential turns. This matches previous findings. Each such pitch range increase started a new hierarchical tree of descending pitch. The pitch tree algorithm relied on these pitch increases to segment the text, and so could only capture relationships among phrases such as topic subordination and sentence internal declination. However, there were also relationships among phrases based on increasing pitch, for example, false starts, corrections, and introductory phrases. These relationships could only be represented indirectly in the descending pitch trees built by the algorithm. The pitch trees helped to explore the multifunctional use of pitch range changes without first having to posit categories of pitch range and abstract away from the phonetic signal.

The pitch tree algorithm for representing pitch trees could benefit from some fine tuning. As I have defined it now, any local increase in pitch gives rise to a new pitch tree at the appropriate level. Very small differences in frequency, such as 1 to 5 Hz should probably not count as differences in level. Such small differences can be due to measurement errors or inherent fundamental frequencies of different vowels and probably are not even reliably distinguished by listeners. Further work would need to be done to determine how big a difference should be represented as a difference, and if it depends on the absolute location in the frequency range. However, it has been interesting to see how much could be learned by using this extremely simple coding of the conventional wisdom that pitch increases for new topics and that subtopics have pitch ranges less than their main topics. The method showed that this was true to a certain extent in even quite complicated texts, spontaneous as well as read, but that this was not the whole story. It revealed a need to be able to represent connective increasing pitch relationships as well decreasing pitch relationships for such things as the possibility of introducing a new topic and subsequent corrections.

References

Beckman, M.E. & Pierrehumbert, J.B. (1986) Intonational structure in Japanese and English, *Phonology Yearbook*, 3, 255-309.

Blaauw, E. (1991) Phonetic characteristics of spontaneous and read-aloud speech. In *Proceedings, ESCA Workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, pp. 12/1-12/5. Barcelona.

Blaauw, E. (1992) On the perceptual difference between read and spontaneous speech: Two experiments, *OTS Yearbook 1992* (M. Everaert, B. Schouten & W. Zonneveld, editors), pp. 1-16. Utrecht, The Netherlands: LEd.

Bolinger, D. (1978) Intonation across languages. In *Universals of Human Language*, 2 (Phonology) (J. H. Greenberg, editor), pp. 471-524. Stanford University Press.

Brazil, D., Coulthard, M. and Johns, C. (1980) *Discourse intonation and language teaching*. Longman.

Brown, G., Currie, K. L. & Kenworthy, J. (1980) *Questions of Intonation*. London: Croom Helm.

Brown, G. (1983) Prosodic structure and the given/new distinction. In *Prosody: Models and Measurements* (D. R. Ladd and A. Cutler, editors), pp. 67-68. Berlin: Springer-Verlag.

Bruce, G. & Touati, P. (1992) On the analysis of prosody in spontaneous speech with exemplification from Swedish and French, *Speech Communication*, **11**, 453-458.

Butterworth, B. (1975) Hesitation and semantic planning in speech, *Journal of Psycholinguistic Research*, **4**, 75-87.

Campbell, W. N., and Isard, S. D. (1991) Segment durations in a syllable frame, *Journal of Phonetics*, **19**(1), 37-47.

Campbell, W. N. (1992) Prosodic encoding of English speech. In *Proceedings, Second International Conference on Spoken Language Processing*, **1**, pp. 663-666. Banff, Canada.

Clark, H. H., and Schaefer, E. F. (1989) Contributing to discourse, *Cognitive Science*, **13**, 259-294.

Cooper, W. E. & Paccia-Cooper, J. (1980) *Syntax and Speech*. Harvard University Press.

French, P. & Local, J. (1986) Prosodic features and the management of interruptions. In *Intonation in discourse* (C. Johns-Lewis, editor), pp. 157-180. San Diego, CA: College-Hill Press, Inc.

Grosz, B. J. & Sidner, C. L. (1986) Attention, intentions, and the structure of discourse, *Computational Linguistics*, **12**(3), 175-204.

Grosz, B., and Hirschberg, J. (1992) Some intonational characteristics of discourse structure. In *Proceedings, Second International Conference on Spoken Language Processing*, **1**, pp. 429-432. Banff, Canada.

Gårding, E. (1967) Prosodiska drag i spontant och uppläst tal. In *Svenskt talspråk* (G. Holm, editor), pp. 40-85. Uppsala, Sweden: Almqvist & Wiksells Boktryckeri AB.

Hirschberg, J. & Litman, D. (1987) Now let's talk about *now*: Identifying cue phrases intonationally. In *Proceedings, 25th Annual Meeting of the Association for Computational Linguistics*, pp. 163-171. Stanford, CA.

Hirschberg, J. & Grosz, B. (1992) Intonational features of local and global discourse structure. In *Proceedings, Fifth DARPA Workshop on Speech and Natural Language*, pp. 441-446. Harriman, NY: Morgan Kaufmann.

Hirschberg, J. & Pierrehumbert, J. (1986) The intonational structuring of discourse. In *Proceedings, 24th Annual Meeting of the Association for Computational Linguistics*, pp. 136-144. New York.

Howell, P. & Kadi-Hanifi, K. (1991) Comparison of prosodic properties between read and spontaneous speech material, *Speech Communication*, **10**, 163-169.

International Phonetics Association (1989) Report on the 1989 Kiel Convention, *Journal of the International Phonetics Association,* **19**(2), 67-80.

Ladd, D. R. (1993) Constraints on the gradient variability of pitch range. In *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (Patricia Keating, editor), pp. 43-63. Cambridge University Press.

Lehiste, I. (1975) The phonetic structure of paragraphs. In *Structure and Process in Speech Perception* (A. Cohen and S. G. Nooteboom, editors), pp. 195-203. Springer-Verlag.

Lehiste, I. (1979) Perception of sentence and paragraph boundaries. In *Frontiers of Speech Research* (B. Lindblom and S. Öhman, editors), pp. 191-201. London: Academic Press.

Lehiste, I. (1980) Phonetic characteristics of discourse. Presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan.

Levin, H., Schaffer, C. A. & Snow, C. (1982) The prosodic and paralinguistic features of reading and telling stories, *Language and Speech,* **25**, 43-54.

Liberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In *Language Sound Structure: Studies in phonology* (M. Aranoff and R. T. Oehrle, editors), pp. 157-233. MIT Press.

Passenout, R. J. & Litman, D. J. (1993) Feasibility of automated discourse segmentation. In *Proceedings, 31st Annual Meeting of the Association for Computational Linguistics,* pp. 148-155. Ohio State University.

Pierrehumbert, J. B. (1980) *The phonology and phonetics of English intonation.* Doctoral dissertation, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.

Pierrehumbert, J. & Hirschberg, J. (1990) The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication* (P. R. Cohen, J. Morgan & M. E. Pollack, editors), pp. 271-311. MIT Press.

Remez, R. E., Rubin, P. E. & Ball, S. (1985) Sentence intonation in spontaneous utterances and fluently spoken text. Presented at the 109th Meeting of the Acoustical Society of America, Austin, TX, April 1985.

Remez, R. E., Rubin, P. E. & Nygaard, L. C. (1986) On spontaneous speech and fluently spoken text: Production differences and perceptual distinctions. Presented at the 111th Meeting of the Acoustical Society of America, Cleveland, OH, May 1986.

Sacks, H. & Schegloff, E. A. (1979) Two preferences in the organization of reference to persons in conversation and their interaction. In *Everyday Language: Studies in ethnomethodology* (G. Psathas, editor), pp. 15-21. New York: Irvington Publishers.

Schegloff, E. A. (1982) Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In *32nd Georgetown University Roundtable on Languages and Linguistics 1981, Analyzing discourse: Text and talk,* (D. Tannen, editor), pp. 71-93. Washington, DC: Georgetown University Press.

Shockey, L. R. (1974) Phonetic and phonological properties of read speech, *Ohio State University Working Papers in Linguistics,* **17**, iv-143.

Silverman, K. E. A. (1987) *The structure and processing of fundamental frequency contours.* Doctoral dissertation, University of Cambridge.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992) TOBI: A standard for labeling English prosody. In *Proceedings, Second International Conference on Spoken Language Processing,* 2, pp. 867-870. Banff, Canada.