Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous and Read Speech

Julia Hirschberg
AT&T Bell Laboratories

INTRODUCTION

Corpus-based approaches to the study of speaking styles seeks to identify observed differences as potentially perceptually salient. Particularly when reliable differences are found across speakers in such corpora, we hypothesize that such differences may in fact cause subjects to judge one style to be "spontaneous", or "conversational" and another to be "formal" or "read" or "laboratory speech". So the question of what distinguishes one style from another is addressed by examining similar corpora for systematic differences in lexical choice, syntactic constructions, and acoustic and prosodic phenomena. In this paper, results from three American English corpus studies of spontaneous and read speech are discussed, to examine the extent to which they support commonly held beliefs about which prosodic and other acoustic phenomena commonly distinguish between these speaking styles.

DIFFERENCES IN RATE AND CONTOUR

Two phenomena widely believed to distinguish spontaneous from read speech are speaking rate and choice of intonational contour. An examination of 395 read and spontaneous utterances in the ARPA *atis0* training corpus, collected at Texas Instruments for use in the DARPA Air Travel Information System spoken language system evaluation task from seventeen speakers (0), supports these hypotheses. These speakers interacted with a simulated voice response system to make air travel plans for several different hypothetical customers. They later returned to read transcriptions of their own speech.

A simple analysis of speaking rate in this corpus indicates considerable differences in rate for the read vs. spontaneously produced utterances. Of the seventeen speakers in this corpus, the average speaking rate (in syllables per second) was faster for read sentences than for spontaneous sentences in sixteen cases. Ratios of read to spontaneous speaking rate averages ranged from .93 (for the seventeenth speaker) to 1.57. Whether this difference is due primarily to the presence of pauses within spontaneous utterances or simply to an overall reduction in rate remains to be examined.

With respect to choice of contour as a differentiator between spontaneous and read speech, there are similar but less clear-cut findings. It has been claimed that yes-no questions are commonly uttered with rising intonation in spontaneous speech, but not so reliably in read speech. In a small study of yes-no questions in read speech (0), inverted yes-no questions and *wh*-questions from the training and test data of the speaker independent DARPA *Resource Management* (RM) database and from the TIMIT database were sampled for purposes of comparison with standard contours for those sentence types in spontaneous speech (0). While only 55% of yes-no questions in the RM database were uttered with rising intonation, over 80% of yes-no questions in the TIMIT1 sample rose. Production of *wh*-questions appears similar in both databases, with only 8% of the TIMIT1 *wh*-questions and 9% of the RM *wh*-questions uttered with final rise. However, this study examined only read speech, inferring the comparison with

spontaneous.

Comparing spontaneous with read utterances in the *atis0* corpus, we do find apparent differences between the two styles for a broader array of contours. There is a greater tendency in spontaneous speech to utter declaratives with some form of rising intonation; while fully 93.1% of read declaratives are uttered with final fall, only 70.5% of spontaneous declaratives are so uttered. Similarly, while 83.8% of read *wh*-questions are uttered with falling contours, only 62.2% of spontaneous *wh*-questions are. Although the numbers for yes-no questions are even smaller than for *wh*-questions, the observed behavior shows distinctions between read and spontaneous speech which one might not have predicted: more spontaneous yes-no questions (43.3%) are uttered with falling intonation than read yes-no questions (30%), for utterances of the same tokens.

CHARACTERIZING DISFLUENCY IN SPEECH

The presence of various phenomena known collectively as DISFLUENCIES is commonly thought to distinguish spontaneous speech from read speech. However, identifying what constitutes disfluency in spontaneous speech has proven a difficult topic in itself. While intuitively we may know disfluency when we hear it, the impression of disfluent speech may be described in various ways, and mapped to a large number of distinct lexical, acoustic, and prosodic realizations.

How important a cue disfluency is in identifying speech as spontaneous or not must depend in part on how frequently disfluencies occur, a question that corpus-based studies are well-suited to addressing. Blackmer and Mitton (0) report a rate of one disfluency per 4.6 seconds for callers to a Canadian radio talk show. Studies of large recorded speech corpora have found that approximately 9–10% of spontaneous utterances contain one particular form of disfluency, SELF REPAIRS (0; 0; 0). Yet even where a particular type of disfluency was the subject of study, the acoustic, prosodic, and lexical phenomena which realized the disfluency varied considerably from token to token.

Several multi-speaker corpus-based studies of disfluent speech (0) sought to identify the range of acoustic and phonetic features associated with one type of disfluency, self-repairs, and compared some of these features with fluent read speech. The corpus for these studies consisted of 6414 utterances from the ARPA *atis2* database (0) collected at AT&T, BBN, CMU, SRI, and TI and produced by 122 speakers. 346 (5.4%) utterances contained at least one repair, defined as the self-correction of one or more phonemes (up to and including sequences of words) in an utterance. The speech was labeled for intonational prominences and phrasing following Pierrehumbert's description of English intonation (0). Disfluencies had been categorized independently of these studies as REPAIR (self-correction of lexical material), HESITATION ("unnatural" interruption of speech flow without any following correction of lexical material), or OTHER DISFLUENCY; only the first category were examined in detail.

To provide a framework for the investigation, we further labeled each repair instance using the REPAIR INTERVAL MODEL (RIM) (0). This model divides the repair event into three temporal intervals and identifies critical time points within the intervals. A full repair comprises three contiguous intervals, the REPARANDUM INTERVAL, the lexical material which is to be repaired), the DISFLUENCY INTERVAL (the DI, extending from the termination of the fluent portion of the utterance to the resumption of fluent speech, and containing any number of silence, filled pauses and CUE PHRASES, such as "I meant"), and the REPAIR INTERVAL(the correcting material, which is intended to 'replace' the reparandum). The end of the reparandum coincides with the termination of the fluent portion of the utterance, which we term the INTERRUPTION SITE (IS).

Characteristics of the Reparandum Interval

The most reliable cue found to the presence of a self-correction in utterances in the corpus was the speech fragment. However, the fragments observed in the corpus did *not* represent a homogenous class. Most fragments did tend to occur in content words (43%) rather than function words (5%).[1] Also, 91% of fragments were one syllable or less in length. But there were distributional differences in the phonetic composition of fragments. Single consonant fragments that are fricatives occurred more than six times as often as those that are stops. However, fricatives and stops occur almost equally as the initial consonant in single syllable fragments. Furthermore, we observed two divergences from the underlying distributions of initial phonemes for all words in the corpus. Vowel-initial words were less likely to occur as fragments and fricative-initial words were more likely to occur as fragments, relative to the underlying distributions for those classes in the corpus as a whole. Both the overall and repair distributions and the single consonant and single syllable distributions differed significantly.

In addition to speech fragments, we found evidence that glottalization and coarticulation may be associated with reparanda offsets, especially those ending in fragments;[2] hence, these too may serve as cues to speaking style. In our corpus, 30.2% of reparanda offsets are marked by what we will term INTERRUPTION GLOTTALIZATION. Although interruption glottalization was usually associated with fragments, not all fragments were glottalized.[3] Interruption glottalization appears acoustically distinct from LARYNGEALIZATION (creaky voice), which often occurs at the end of prosodic phrases; the latter typically extends over several syllables at the end of an intonational phrase and is associated with a decrease in energy and low fundamental frequency (0). Interruption glottalization on the other hand tends to occur only over the interrupted syllable, and does not appear to be associated with a sustained decrease in energy and fundamental frequency.[4]

A final feature characterizing the end of some reparanda intervals is the presence of coarticulatory gestures preceding silence. Sonorant endings of both fragments and non-fragments in our corpus may exhibit coarticulatory effects of an acoustically unrealized subsequent phoneme. A related feature is the lack of phrase-final lengthening on the last few segments in the reparandum for many cases of repairs. More generally, both of these features are cues to disfluency in the rhythmic structure of pre-pausal segments.

To summarize, findings from the study of reparanda in self-corrections indicate that this spontaneous speech phenomenon can be realized via word fragments, interruption glottalization, and articulatory gestures that are not fully realized.

Characteristics of the Disfluency Interval

In the RIM model, the DI includes all cue phrases and all filled and unfilled pauses from the offset of the reparandum to the onset of the repair. These phenomena are commonly seen as indicators of spontaneous speech in general, and repair phenomena in particular (0; 0; 0; 0). In a superset of our corpus, only 2.8% (333/11900) spontaneous utterances contained one or more filled pauses and 4.8% (572/11900) contained fragments, while 7.1% (844/11900) contained at least one example of either. In fact, while fragments were reliable cues to repair phenomena in our study, filled pauses were not. We did however find that the duration of silent pauses as the DI was a reliable indicator of the presence of a self-repair, and in fact served also to distinguish between non-fragment and fragment repairs. Overall, silent DIs are significantly shorter than fluent pauses, and the DI duration for fragment repairs is significantly shorter than for non-fragment repairs. The fragment repair DI duration is also significantly shorter than fluent pause intervals, while there is no significant difference between non-fragment repairs DIs and fluent phrase boundaries. So, DIs in general appear to be distinct from fluent phrase boundaries. The presence of such unusual boundaries may thus also be seen as a way of distinguishing spontaneous from read speech.

Characteristics of the Repair Interval

One final distinction between spontaneous utterances containing corrections and fluent utterances was in intonational phrasing. We tested the hypothesis that repair interval offsets are marked by the presence of intonational phrase boundaries by examining whether phrase boundaries observed at that offset differed in their occurrence from those observed in fluent speech for the TI corpus as a whole (0). Using Wang and Hirschberg's (0) phrase prediction procedure, with prediction trained on 478 sentences of read, fluent speech from the *atis* TI read corpus, we estimated whether the phrasing at the repair offset was predictably distinct from this model of fluent phrasing. To see whether these boundaries were distinct from those in fluent speech, we compared the phrasing of repair utterances with the phrasing predicted for the corresponding corrected version of the utterance as identified by *atis* transcribers.[5] We found that in these 63 utterances the repair offset co-occurs with minor or major phrase boundaries for 49% of repairs. For 40% of all repairs, an observed boundary occurs at the repair offset where one is predicted in fluent speech; and for 33% of all repairs, no boundary is observed where none is predicted. For the remaining 27% of repairs, observed phrasing diverges from that predicted by a fluent phrasing model. In 37% of these latter cases, a boundary occurs where none is predicted, and, in the remainder, no boundary occurrs when one is predicted.

We also found more general differences from predicted phrasing over the entire repair interval. Two strong predictors of prosodic phrasing in fluent speech are thought to be syntactic constituency (0; 0; 0), especially the relative inviolability of noun phrases (0), and the length of prosodic phrases (0). In our repair utterances, we observed phrase boundaries at repair offsets which occurred within larger NPs when only a portion of the NP was being corrected. We also found cases in which intonational phrases observed in repair utterances were much longer than phrases observed in fluent speech. In such cases, the absence of intonational phrase boundaries appeared to identify the entire repair as a substituting unit.

So, differences in the location of intonational phrase boundaries as well as the realization of boundaries themselves, fragmentation, interruption glottalization, and coarticulatory phenomena may all be found in spontaneous speech but are found more rarely (not at all in our corpus) in read speech. However, when we attempt to use these observations for the purpose of actually distinguishing between spontaneous and read speech, we find some limitations. While a sizeable fraction of spontaneous utterances contain disfluencies (about 20%, for example, of the *atis0* corpus utterances), a much larger portion do not. Even those containing a disfluent phenomenon exhibit vastly different acoustic and prosodic manifestations of that phenomenon. While the presence of fragments, filled pauses, cue phrases,

[1] 52% of intended words were not recoverable by the transcribers.

[2] See also (0; 0).

[3] In our database, 62% of fragments are *not* glottalized, and 9% of glottalized reparanda offsets are *not* fragments.

[4] We suspect that this phenomenon is similar to that of HOLDING SILENCES investigated by Local and Kelly (0), which occur on discourse connectives; Local and Kelly speculate that these serve the general communicative function of holding the floor.

[5] Results reported here are for prediction on only the 63 TI repair utterances, since the prediction tree we used had been developed on TI utterances.

and differences in location and realization of intonational phrasing may variously distinguish some spontaneous utterances from non-spontaneous ones, the large majority of spontaneous utterances in our corpora contained no such cues.

INDICATORS OF DISCOURSE STRUCTURE

In a third set of studies, read and spontaneous elicited speech were compared to see whether the way speakers convey discourse structural information differs from one style to the other, as previous studies have suggested. Our corpus comprises elicited monologues produced by multiple non-professional speakers, who are given written instructions to perform a series of increasingly complex direction-giving tasks. Speakers first explain simple routes such as getting from one station to another on the subway, and progress gradually to the most complex task of planning a round-trip journey from Harvard Square to several Boston tourist sights. The speakers are provided with various maps, and may write notes to themselves as well as trace routes on the maps. For the duration of the experiment, the speakers are in face-to-face contact with a silent experimental partner (a confederate) who traces on her map the routes described by the speakers. The speech is subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted; subjects return several weeks after their first recording to read the transcribed speech. Both sets of recordings are then acoustically and prosodically labeled, the latter using the ToBI labeling convention (0; 0). Results are given for the spontaneous and the read speech for one speaker, performing five direction-giving tasks.

Discourse segmentations based on Grosz & Sidner's theory of discourse structure were obtained from three subjects labeling from text alone (group T) and three labeling from speech and text (group S). Consensus labels (all subjects in the group agreeing) were obtained for segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF). The segmentations of group S differ significantly from those of group T. Unlike results of earlier experiments (0; 0), those who listened to speech while segmenting produced more consensus boundaries for both read and spontaneous speech than did those who segmented from text alone. When the read and spontaneous data are pooled, labelers from speech and text agree upon significantly more SBEG boundaries. Spontaneous speech is generally claimed to exhibit less reliable prosodic indicators of discourse structure than read speech (cf. (0)). Yet, in our corpus, spontaneous speech actually produced significantly more SCONT consensus labels than did read speech, for groups S and T combined. The higher overall percentages of consensus labels for spontaneous speech are attributable to this difference in SCONT labelings.

We examined the following acoustic and prosodic correlates of consensus labelings of intermediate phrases labeled as SBEGs and SFs: f0 maximum and average f0; rms maximum and average; speaking rate; and duration of preceding and subsequent pauses. We compared segmentation labels not only for group S versus group T, but also for spontaneous versus read speech. As noted, while intonational correlates for segment boundaries *have* been identified in read speech, they have been observed in spontaneous speech rarely and descriptively.

We found strong correlations for consensus SBEG and SF labels for groups S and T in both spontaneous speech and read speech.[6] Results on consensus SBEG labels were as follows: given group T segmentations, we found significantly higher maximum and average f0, and maximum and average rms, and shorter subsequent pause for both spontaneous and read speech; for read speech we also found significant correlations for preceding pauses. Given group S segmentations, we found significantly higher maximum and average f0, higher maximum rms, longer preceding and shorter succeeding pauses for read and spontaneous speech; we found higher average rms as well for read speech. Results on consensus SF labels were as follows: given group T segmentations, we found significantly lower average f0 and rms maximum for both read and spontaneous speech, and lower rms average and subsequent pause in addition for read speech. Given group S segmentations, we found lower average f0, rms maximum and average, shorter preceding pause, and longer subsequent pause for both read and spontaneous speech, and in addition, lower f0 maximum for read speech.

So, in this exercise, while we found some differences between read and spontaneous speech, what was *not* different was that speakers indeed seemed to use prosodic and acoustic cues to convey discourse structure.

[6] T-tests were used to test for statistical significance of difference in the means of phrases, e.g. beginning and not beginning segments. Results reported are significant at the .025 level or better.

DISCUSSION

Data from several large corpora of American English read and spontaneous speech thus do provide some comparative information on speaking style issues and raise some interesting additional questions for future analysis.

A study of corpora collected for several DARPA tasks supports the hypotheses that read speech is more rapid than spontaneous speech and that choice of intonational contour does indeed appear to be different for at least some sentence types in read vs. spontaneous utterances. However, a precise characterization of how differences observed in speaking rate between the faster read and slower spontaneous ATIS utterances difference remains to be determined. And, while an examination of read speech found that indeed only 55–80% of yes-no questions were read with rising "yes-no question" contours, thus confirming the hypothesis that read speech does not produce "natural" intonational contours, a comparison of read and spontaneous versions of sentences in the *atis* corpus shows, paradoxically, that over 43% of spontaneous yes-no questions were uttered with **falling** intonation, compared with only 30% of read yes-no questions. So, the notion of a "natural" association between sentence type and intonational contour for yes-no questions in spontaneous speech may need to be re-examined. Similarly, in spontaneous speech, about 30% of declarative sentences were uttered with some type of rising intonation, while over 93% of read declaratives were uttered with final fall. And *wh*-questions, commonly believed to be normally uttered with falling intonation in English, were only so uttered about 62% of the time in spontaneous speech, while in read speech about 84% of such questions were uttered with falling contours. One may speculate that the association between contour type and sentence type may be something that speakers adhere to more consistently when they are asked to read, particularly longer texts, such as the *atis* dialogues, rather than when they produce the same dialogues naturally.

Other questions about comparative speaking style are raised by studies of disfluency in spontaneous speech. While hesitations and self-repairs may occur in collections of read speech, they are rarely preserved; and their presence in speech presented to listeners does seem to provide a useful cue to listeners that the speech they have been presented with is spontaneous. However, do these phenomena, which are quite difficult to define precisely, actually occur often enough to provide reliable cues in general as to speaking style? In the studies of disfluency discussed above, only about 20% of spontaneous utterances contained any form of disfluency — from filled pauses to self-repairs to labeler-observed "hesitations". And the auditory cues that marked such disfluencies themselves range widely, from "abnormal" phrasings to speech fragments to glottalization and coarticulatory devi-

ations from fluent speech. Determining which of these cues are primary indicators of disfluent speech is clearly important, but only speech fragments and filled pauses appear to occur with any great frequency, even for the 20% of spontaneous utterances that contain disfluencies. So, it is clearly important to search for other cues to speaking style variation in addition to acoustically observable indicators of disfluency.

Finally, it has been speculated that a difference between read and spontaneous speaking styles might be inferred from the fact that regularities observable in read speech in the use of prosodic cues to discourse structure, such as pitch range, speaking rate, and pausal duration, had not also been observed for spontaneous speech. However, the comparisons discussed above in the Boston Directions Corpus suggest that this speculation may be incorrect. Statistically reliable associations between prosodic variation and subject-labeled discourse structure have indeed been found for spontaneous as well as read productions of speakers giving directions.

However, before we in fact conclude that the use of intonational variation is not a distinguishing characteristic of read speech, it should be noted that the spontaneous speech collected in this study is not entirely unplanned, but has been elicited from subjects given a task and some time to plan it. While the speech itself is thus spontaneous — subjects did not write out their directions — the productions did follow a planning period. This observation also raises the important point that "spontaneous" speech may vary along many dimensions, including the element of prior planning involved and the presence, size, and type of interlocutors involved.

References

DARPA. *Proceedings of the Speech and Natural Language Workshop*, Hidden Valley PA, June 1990. Morgan Kaufmann.

Julia Hirschberg. Distinguishing questions by contour in speech recognition tasks. In *Proceedings of the Speech and Natural Language Workshop*. Morgan Kaufmann, Cape Cod MA, October 1989.

P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett. The DARPA 1000-word Resource Management Database for continuous speech recognition. In *Proceedings*, volume 1, pages 651-654, New York, 1988. ICASSP88.

Elizabeth R. Blackmer and Janet L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173-194, 1991.

Donald Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting*, pages 123-128, Cambridge MA, 1983. Association for Computational Linguistics.

Elizabeth Shriberg, John Bear, and John Dowding. Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the Speech and Natural Language Workshop*, pages 419-424, Harriman NY, 1992. DARPA, Morgan Kaufmann.

C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, March 1994.

MADCOW. Multi-site data collection for a spoken language corpus. In *Proceedings of the Speech and Natural Language Workshop*, pages 7-14, Harriman NY, February 1992. DARPA, Morgan Kaufmann.

Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.

John Bear, John Dowding, and Elizabeth Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting*, pages 56-63, Newark DE, 1992. Association for Computational Linguistics.

Joseph Olive, Alice Greenwood, and John Coleman. *The Acoustics of American English Speech: A Dynamic Approach*. Springer-Verlag, New York, 1993.

John Local and John Kelly. Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 9:185-204, 1986.

Willem Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41-104, 1983.

Douglas O'Shaughnessy. Analysis of false starts in spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 931-934, Banff, October 1992. ICSLP.

Michelle Q. Wang and J. Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175-196, 1992.

W. E. Cooper and J. M. Sorenson. Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, 62(3):683-692, September 1977.

J. P. Gee and F. Grosjean. Performance structure: A psycholinguistic and linguistic apprasial. *Cognitive Psychology*, 15:411-458, 1983.

E. O. Selkirk. Phonology and syntax: The relation between sound and structure. In T. Freyjeim, editor, *Nordic Prosody II: Proceedings of the Second Symposium on Prosody in the Nordic language*, pages 111-140, Trondheim, 1984. TAPIR.

K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg. TOBI: A standard scheme for labeling prosody. In *Proceedings of the Second International Conference on Spoken Language Processing*, Banff, October 1992. ICSLP.

John Pitrelli, Mary Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 2, pages 123-126, Yokohama, 1994. ICSLP.

B. Grosz and J. Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October 1992. ICSLP.

Barbara Grosz, Julia Hirschberg, and Christine Nakatani. A study of intonation and discourse structure in directions. In *Proceedings of the Workshop on the Integration of Natural Language and Speech Processing*. AAAI, August 1994.

Gayle M. Ayers. Discourse functions of pitch range in spontaneous and read speech. Presented at the Linguistic Society of America Annual Meeting, 1992.