Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dialogue

Gina-Anne Levow
University of Chicago
levow@cs.uchicago.edu

Abstract

Theories of discourse structure hypothesize a hierarchical structure of discourse segments, typically tree-structured. While substantial work has been done on identifying and automatically recognizing the textual and prosodic correlates of discourse structure in monologue, comparable cues for dialogue or multi-party conversation, and in particular human-computer dialogue remain relatively less studied. In this paper, we explore prosodic cues to discourse segmentation in human-computer dialogue. Using data drawn from 60 hours of interactions with a voice-only conversational spoken language system, we identify pitch and intensity features that signal segment boundaries. Specifically, based on 473 pairs of segment-final and segment-initiating utterances, we find significant increases for segment-initial utterances in maximum pitch, average pitch, and average intensity, while segment-final utterances show significantly lower minimum pitch. These results suggest that even in the artificial environment of human-computer dialogue, prosodic cues robustly signal discourse segment structure, comparably to the contrastive uses of pitch and amplitude identified in natural monologues.

Keywords Dialogue Systems, Discourse structure, Prosody in understanding

1 Introduction

Contemporary theories of discourse, both computational and descriptive, postulate a tree-structured hierarchical model of discourse. These structures may be viewed as corresponding to"intentional" structure of discourse segment purposes in the view of (Grosz and Sidner, 1986), to plan and subplan structure directly in the view of (Allen and Litman, 1990) , to nuclei and satellite rhetorical relations in the Rhetorical Structure Theory of (Mann and Thompson, 1987), or to information structures as in (Traum and Hinkelman, 1992). Despite this diversity of views on the sources of structural organization, these theories agree on the decomposition of discourse into segments and subsegments in a hierarchical structure.

Discourse segments help to establish the domain of interpretation for referents or anaphors. (Grosz, 1977) Discourse segmentation can also provide guidance for summarization or retrieval by identifying the topical structure of extended text spans. As a result, an understanding of the mechanisms that signal discourse structure is highly desirable.

While substantial work has been done on identifying and automatically recognizing the textual and prosodic correlates of discourse structure in monologue, comparable cues for dialogue or multi-party conversation, and in particular human-computer dialogue remain relatively less studied. In this paper, we explore prosodic cues to discourse segmentation in human-computer dialogue.

Using data from 60 hours of interactions with a voice-only conversational spoken language system, we identify pitch and intensity features that signal segment boundaries. Specifically, based on 473 pairs of segment-final and segment-initiating utterances, we find significant increases for segment-initial utterances in maximum and average pitch and average intensity, with significantly lower minimum pitch for segment-final utterances. These results suggest that even in the artificial environment of human-computer dialogue, prosodic cues robustly signal discourse segment structure, comparably to the contrastive uses of pitch and amplitude identified in natural monologues.

1.1 Overview

We begin with a discussion of related work on discourse segmentation and dialogue act identification in monologue and dialogue, primarily in the human-human case. Then we introduce the system and data collection pro-

cess that produced the human-computer discourse segment change materials for the current analysis. We describe the acoustic analyses performed and the features chosen for comparison. Then we identify the prosodic cues that distinguish discourse segment boundaries and discuss the relation to previously identified cues for other discourse types. Finally we conclude and present some future work.

2 Related Work

Cues for and automatic segmentation of discourse structure have been most extensively studied for written and spoken monologue. For written narrative, discourse segment boundaries have been identified based on textual topic similarity with a variety of approaches based on Hearst's Textiling(Hearst, 1994). More complex rhetorial structure theory trees have also been extracted based heavily on cue phrases and discourse markers by (Marcu, 2000).

In spoken monologue, prosodic cues to discourse structure and segmentation have been explored by (Nakatani et al., 1995; Swerts, 1997). Increases in pitch range, amplitude, and silence duration appear to signal discourse segment boundaries across different domains - voicemail, broadcast news, descriptive narrative - and across different languages, such as English and Dutch. Comparable prosodic cues have been applied to the related task of news story segmentation, in conjunction with textual cues to topicality, by (Tur et al., 2001), where large pitch differences between pre- and post- boundary positions play the most significant role among prosodic cues.

In spoken dialogue, research has focused on the identification of dialogue acts and dialogue games. Integration of textual and prosodic cues, such as particular pitch accent or contour types, have been found useful for identifying act type(Shriberg et al., 1998; Taylor et al., 1998). Specific classes of dialogue act, such as corrections (request repair), have received particular interest in work by (Levow, 1998; Swerts et al., 2000) in the context of human-computer error resolution. Recent work on the ICSI multi-party meeting recorder data has demonstrated some very preliminary results on multi-party segmentation (Galley et al., 2003); prosodic information in this case was limited to silence duration.

With the exception of work on error resolution, most work on dialogue has focused human-human interaction and on identification of particular act or game types. Here we concentrate on the general question of discourse segmentation in voice-only human-computer interaction. We ask whether the cues to segment structure identified for monologue are robust to the change in number and type of conversational participant.

3 Data Collection

3.1 Speech System Description

The utterances used in the analysis of discourse segmentation for human-computer interaction were drawn from approximately 60 hours of user interactions in a field trial of the Sun Microsystems SpeechActs system (Yankelovich et al., 1995). SpeechActs provides a voice-only interface to common desktop applications such as e-mail, calendar, weather reports, stock quotes, and time and currency conversions. The system allows conversational interactions with speech recognition provided by BBN's HARK recognizer and speech synthesis through Centigram's TruVoice system. A sample interaction appears below. [1]

SpeechActs: You have 14 new messages and
47 old messages. The first new
message is from Eric Baatz
regarding "report completed."
User: Let me hear it.
SpeechActs: "The first draft is ready for
your comments. Eric."
User: Reply and include current message.
SpeechActs: Preparing message to Eric Baatz.
Begin recording after the tone.
When finished, pause several seconds.
User: Eric, I'll get those to you
this afternoon.
SpeechActs: Say cancel, send, or review.
User: Send.
SpeechActs: Message sent. What now?
User: Next message.
SpeechActs: New message two from Stuart
Adams, subject "Can we meet at
3:00 today?"
User: Switch to calendar...

The field trial involved a group of nineteen subjects. Four of the participants were members of the system development staff, fourteen were volunteers drawn from Sun Microsystems' staff, and a final class of subjects consisted of one-time guest users. There were three female and fifteen male regular user subjects.

All interactions with the system were recorded and digitized in standard telephone audio quality format at 8kHz sampling in 8-bit mu-law encoding during the conversation. In addition, speech recognition results, parser results, and synthesized responses were logged. A paid assistant then produced a correct verbatim transcript of all user utterances. Overall there were 7752 user utterances recorded.

[1] Designing SpeechActs: Issues in Speech User Interface Design (Yankelovich et al., 1995) p. 2

3.2 Data Coding and Extraction

Consistent discourse segmentation can be difficult even for trained experts (Nakatani et al., 1995; Swerts, 1997; Passoneau and Litman, 1997), and differences in depth of nesting for discourse structure appear to be the most problematic. As a result, we chose to examine utterances whose segment and topic initiating status would be relatively unambiguous. As the SpeechActs system consists of 6 different applications, we chose to focus on changes from application to application as reliable indicators of topic initiation. These commands are either simply the name of the desirable application, as in "Mail" or "Calendar", possibly with an optional politeness term, or a switch command, such as "Switch to" and the name of the application. Approximately 1400 such utterances occured during the field trial data collection.

We performed an automatic forced alignment in order to identify and extract the relevant utterances from the digitized audio. Using the full sequence of synthesized computer utterances and manually transcribed user utterances, we applied the *align* function of the Sonic speech recognizer provided as part of the University of Colorado (CU) Communicator system to a 16-bit linear version of the original audio recording. 473 utterances that were correctly aligned by this automatic process were used for the current analysis.

4 Acoustic Feature Extraction

Based on prior results for monologue, we selected pitch and amplitude features for consideration. Although silence duration is often a good cue to discourse segment boundary position in narrative, we excluded it from consideration in the current study due to the awkward pace of the SpeechActs human-computer interactions. Users had to wait for a tone to speak, and interturn silences were as long as six seconds.

We used the "*To Pitch...*" and "*To intensity*" functions in Praat(Boersma, 2001), a freely available acoustic-phonetic analysis package, to automatically extract the pitch (in Hertz) and amplitude (in decibels) for the interaction. To smooth out local jitter and noise in the pitch and amplitude contours, we applied a 5-point median filter. Finally, in order to provide overall comparability across male and female subjects and across different channel characteristics for different sessions[2], we performed per-speaker, per-session normalization of pitch and amplitude values, computed as $\frac{val-mean}{mean}$. The resulting pitch and amplitude values within the time regions identified for each utterance by forced alignment were used for subsequent analysis.

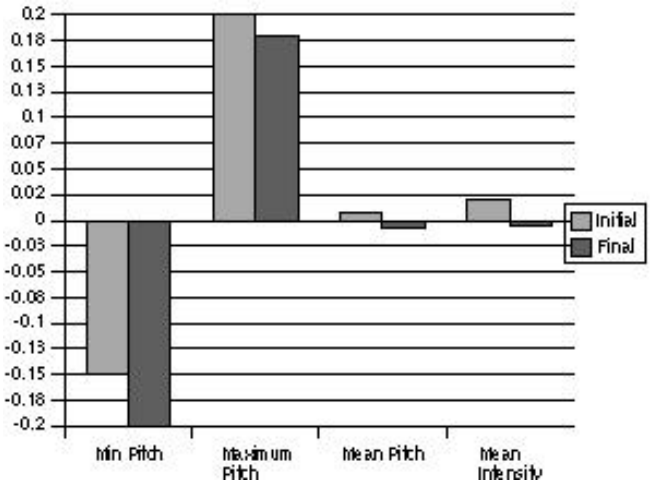

Figure 1: Significant differences in normalized pitch and intensity. Light grey: Segment-initial; Dark grey: segment-final

5 Prosodic Analysis

For both pitch and amplitude we computed summary scalar measures for each utterance. Mean pitch and intensity are intended to capture overall increases or decreases. Maximum and minimum pitch and maximum amplitude served to describe increases in range that might not affect overall average. We compared the segment-initial "application change" utterances with their immediately preceding segment-final utterances.[3] We find significant increases in maximum pitch ($t-test, two-tailed, p < 0.05$), mean pitch ($p < 0.01$), and mean amplitude ($p < 0.001$) of segment-initial utterances relative to segment-final cases. We also find highly significant decreases in minimum pitch ($p < 0.0001$) for segment-final utterances relative to segment-initial utterances. Changes in maximum amplitude did not reach significance. Figure 5 illustrates these changes.

6 Discussion

The significant increases in maximum and mean pitch for segment-initial utterances, coupled with a decrease in pitch minimum for segment-final utterances, suggest a contrastive use of pitch range across the segment boundary. For amplitude, there is a global increase in intensity. These basic features of discourse segment-initial versus discourse segment-final utterances are consistent with the

[2]Since the interface was accessed over a regular analog telephone line from a wide variety of locations - including noisy international airports, the recording quality and level varied widely.

[3]For consistency, we excluded utterances that participated in error spirals, and segment-final utterances which were also segment-initial.

prior findings for monologue. It is interesting to note that in spite of the less than fluent style of interaction imposed on users by the prototype system, cues to discourse segment structure remain robust and consistent. We also observe that the contrasts across discourse segment boundaries are based on the speaker's own baseline prosodic behavior, rather than the conversational partner's, at least in this largely user-initiative system.

7 Conclusions and Future Work

Based on analysis of more than 450 discourse segment boundary pairs, we found significant increases in maximum pitch, average pitch, and average intensity for segment-initial utterances, with a significant decrease in minimum pitch for segment-final utterances. Consistent with prior work on human monologue, new discourse segments in human-computer dialogue are signaled by increases in pitch, contrastive use of pitch range, and loudness, cues which could serve to attract the attention of the other conversational participants.

In future work, we plan to apply these features to automatic extraction of discourse boundaries and global discourse structure. These features could also be used in conjunction with phonetic recognition results to enhance confidence scoring for utterances that would cause a topic shift. In systems such as SpeechActs where topic shift often signals an application change, a somewhat time-consuming activity as a new recognizer is swapped in and new data loaded, it is desirable to have additional implicit confirmation that such an action has in fact been requested. Finally we hope to explore cues to more fine-grained hierarchical discourse structure to distinguish full topic shifts from initiation or completion of subdialogues.

Acknowledgments We thank Nicole Yankelovich and Sun Microsystems for access to the field trial data and their assistance in transcription of these materials.

References

J. F. Allen and D.J. Litman, 1990. *Discourse Processing and Common sense Plans*. MIT Press.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 562–569.

B. Grosz and C. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara Grosz. 1977. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'77)*, page 67=76.

M. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

G.-A. Levow. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of COLING-ACL '98*.

W. C. Mann and S. A. Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In G. Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, pages 85–95. Nijhoff, Dordrecht.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

C. H. Nakatani, J. Hirschberg, and B. J. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 106–112.

Rebecca Passoneau and Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3–4):439–487.

M. Swerts, J. Hirschberg, and D. Litman. 2000. Corrections in spoken dialogue systems. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*, pages 615–619.

Marc Swerts. 1997. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1):514–521.

P. Taylor, S. King, and S. Isard ans H.Wright. 1998. Intonation and dialogue context as constraints for speech recognition. *Language and Speech*, 41(3–4).

D. R. Traum and E. A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599.

G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

N. Yankelovich, G. Levow, and M. Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *CHI '95 Conference on Human Factors in Computing Systems*, Denver, CO, May.