Turn-Taking Cues in a Human Tutoring Corpus

Heather Friedberg
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, 15260, USA
friedberg@cs.pitt.edu

Abstract

Most spoken dialogue systems are still lacking in their ability to accurately model the complex process that is human turn-taking. This research analyzes a human-human tutoring corpus in order to identify prosodic turn-taking cues, with the hopes that they can be used by intelligent tutoring systems to predict student turn boundaries. Results show that while there was variation between subjects, three features were significant turn-yielding cues overall. In addition, a positive relationship between the number of cues present and the probability of a turn yield was demonstrated.

1 Introduction

Human conversation is a seemingly simple, everyday phenomenon that requires a complex mental process of turn-taking, in which participants manage to yield and hold the floor with little pause in-between speaking turns. Most linguists subscribe to the idea that this process is governed by a subconscious internal mechanism, that is, a set of cues or rules that steers humans toward proper turn-taking (Duncan, 1972). These cues may include lexical features such as the words used to end the turn, or prosodic features such as speaking rate, pitch, and intensity (Cutler and Pearson, 1986).

While successful turn-taking is fairly easy for humans to accomplish, it is still difficult for models to be implemented in spoken dialogue systems. Many systems use a set time-out to decide when a user is finished speaking, often resulting in unnaturally long pauses or awkward overlaps (Ward, et. al., 2005). Others detect when a user interrupts the system, known as "barge-in", though this is characteristic of failed turn-taking rather than successful conversation (Glass, 1999).

Improper turn-taking can often be a source of user discomfort and dissatisfaction with a spoken dialogue system. Little work has been done to study turn-taking in tutoring, so we hope to investigate it further while using a human-human (HH) tutoring corpus and language technologies to extract useful information about turn-taking cues. This analysis is particularly interesting in a tutoring domain because of the speculated unequal statuses of participants. The goal is to eventually develop a model for turn-taking based on this analysis which can be implemented in an existent tutoring system, ITSPOKE, an intelligent tutor for college-level Newtonian physics (Litman and Silliman, 2004). ITSPOKE currently uses a time-out to determine the end of a student turn and does not recognize student barge-in. We hypothesize that improving upon the turn-taking model this system uses will help engage students and hopefully lead to increased student learning, a standard performance measure of intelligent tutoring systems (Litman et. al., 2006).

2 Related Work

Turn-taking has been a recent focus in spoken dialogue system work, with research producing many different models and approaches. Raux and Eskenazi (2009) proposed a finite-state turn-taking

Proceedings of the ACL-HLT 2011 Student Session, pages 94–98,
Portland, OR, USA 19-24 June 2011.

model, which is used to predict end-of-turn and performed significantly better than a fixed-threshold baseline in reducing endpointing latency in a spoken dialogue system. Selfridge and Heeman (2010) took a different approach and presented a bidding model for turn-taking, in which dialogue participants compete for the turn based on the importance of what they will say next.

Of considerable inspiration to the research in this paper was Gravano and Hirschberg's (2009) analysis of their games corpus, which showed that it was possible for turn-yielding cues to be identified in an HH corpus. A similar method was used in this analysis, though it was adapted based on the tools and data that were readily available for our corpus. Since these differences may prevent direct comparison between corpora, future work will focus on making our method more analogous.

Since our work is similar to that done by Gravano and Hirschberg (2009), we hypothesize that turn-yielding cues can also be identified in our HH tutoring corpus. However, it is possible that the cues identified will be very different, due to factors specific to a tutoring environment. These include, but are not limited to, status differences between the student and tutor, engagement of the student, and the different goals of the student and tutor.

Our hypothesis is that for certain prosodic features, there will be a significant difference between places where students yield their turn (allow the tutor to speak) and places where they hold it (continue talking). This would designate these features as turn-taking cues, and would allow them to be used as features in a turn-taking model for a spoken dialogue system in the future.

3 Method

The data for this analysis is from an HH tutoring corpus recorded during the 2002-2003 school year. This is an audio corpus of 17 university students, all native Standard English speakers, working with a tutor (the same for all subjects) on physics problems (Litman et. al., 2006). Both the student and the tutor were sitting in front of separate work stations, so they could communicate only through microphones or, in the case of a student-written essay, through the shared computer environment. Any potential turn-taking cues that the tutor received from the student were very comparable to what a spoken dialogue system would have to analyze during a user interaction.

For each participant, student speech was isolated and segmented into breath groups. A breath group is defined as any segment of speech by one dialogue participant bounded by 200 ms of silence or more based on a certain threshold of intensity (Liscombe et. al., 2005). This break-down allowed for feature measurement and comparison at places that were and were not turn boundaries. Although Gravano and Hirschberg (2009) segmented their corpus by 50 ms of silence, we used 200 ms to divide the breath groups, as this data had already been calculated for another experiment done with the HH corpus (Liscombe et. al., 2005). [1]

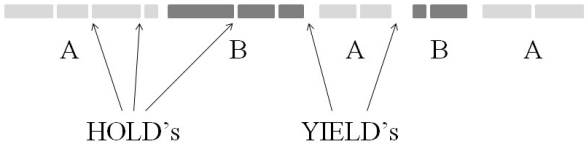

Figure 1. Conversation Segmented into Breath Groups

Each breath group was automatically labeled as one of the following: HOLD, when a breath group was immediately followed by a second breath group from the same person, YIELD, when a breath group was immediately followed by speech from the other participant, or OVERLAP, when speech from another participant started before the current one ended. Figure 1 is a diagram of a hypothetical conversation between two participants, with examples of HOLD's and YIELD's labeled. These groups were determined strictly by time and not by the actually speech being spoken. Speech acts such as backchannels, then, would be included in the YIELD group if they were spoken during clear gaps in the tutor's speech, but would be placed in the OVERLAP group if they occurred during or overlapping with tutor speech. There were 9,169 total HOLD's in the corpus and 4,773 YIELD's; these were used for comparison, while the OVERLAP's were set aside for future work.

Four prosodic features were calculated for each breath group: duration, pitch, RMS, and percent silence. Duration is the length of the breath group in seconds. Pitch is the mean fundamental frequency (f0) of the speech. RMS (the root mean

[1] Many thanks to the researchers at Columbia University for providing the breath group data for this corpus.

	N	duration	percent silence	pitch	RMS
HOLD Group Mean	993	1.07	0.34	102.24	165.27
YIELD Group Mean	480	0.78	0.39	114.87	138.89
Significance		* p = 0.018	* p < 0.001	* p < 0.001	* p < 0.001

Table 1. Individual Results for Subject 111
* denotes a significant p value

	N	duration	percent silence	pitch	RMS
HOLD Group Mean	17	1.49	0.300	140.44	418.00
YIELD Group Mean	17	0.82	0.310	147.58	354.65
Significance		* p = 0.022	p = 0.590	* p = 0.009	* p < 0.001

Table 2. Results from Paired T-Test

squared amplitude) is the energy or loudness. Percent silence was the amount of internal silence within the breath group. For pitch and RMS, the mean was taken over the length of the breath group. These features were used because they are similar to those used by Gravano and Hirschberg (2009), and are already used in the spoken dialogue system we will be using (Forbes-Riley and Litman, 2011). While only a small set of features is examined here, future work will include expanding the feature set.

Mean values for each feature for HOLD's and YIELD's were calculated and compared using the student T-test in SPSS Statistics software. Two separate tests were done, one to compare the means for each student individually, and one to compare the means across all students. $p \leq .05$ is considered significant for all statistical tests. The p-values given are the probability of obtaining the difference between groups by chance.

4 Results

4.1 Individual Cues

First, means for each feature for HOLD's and YIELD's were compared for each subject individually. These individual results indicated that while turn-taking cues could be identified, there was much variation between students. Table 1 displays the results of the analysis for one subject, student 111. For this student, all four prosodic features are turn-taking cues, as there is a significant different between the HOLD and YIELD groups for all of them. However, for all other students, this was not the case. As shown in Table 3, multiple significant cues could be identified for most students, and there was only one which appeared to have no significant turn-yielding cues.

Because there was so much individual variation, a paired T-test was used to compare the means across subjects. In this analysis, duration, pitch, and RMS were all found to be significant cues. Percent silence, however, was not. The results of this test are summarized in Table 2. A more detailed look at each of the three significant cues is done below.

Number of Significant Cues	Number of Students
0	1
1	0
2	6
3	9
4	1

Table 3. Number of Students with Significant Cues

Duration: The mean duration for HOLD's is longer than the mean duration for YIELD's. This suggests that students speak for a longer uninterrupted time when they are trying to hold their turn, and yield their turns with shorter utterances. This is the opposite of Gravano and Hirschberg's (2009) results, which found that YIELD's were longer.

Pitch: The mean pitch for YIELD's is higher than the mean pitch for HOLD's. Gravano and Hirschberg (2009), on the other hand, found that YIELD's were lower pitched than HOLD's. This difference may be accounted for by the difference in tasks. During tutoring, students are possibly

more uncertain, which may raise the mean pitch of the YIELD breath groups.

RMS: The mean RMS, or energy, for HOLD's is higher than the mean energy for YIELD's. This is consistent with student's speaking more softly, i.e., trailing off, at the end of their turn, a usual phenomenon in human speech. This is consistent with the results from the Columbia games corpus (Gravano and Hirshberg, 2009).

4.2 Combining Cues

Gravano and Hirschberg (2009) were able to show using their cues and corpus that there is a positive relationship between the number of turn-yielding cues present and the probability of a turn actually being taken. This suggests that in order to make sure that the other participant is aware whether the turn is going to continue or end, the speaker may subconsciously give them more information through multiple cues.

To see whether this relationship existed in our data, each breath group was marked with a binary value for each significant cue, representing whether the cue was present or not present within that breath group. A cue was considered present if the value for that breath group was strictly closer to the student's mean for YIELD's than HOLD's. The number of cues present for each breath group was totaled. Only the three cues found to be significant cues were used for these calculations. For each number of cues possible x (0 to 3, inclusively), the probability of the turn being taken was calculated by $p(x) = Y / T$, where Y is the number of YIELD's with x cues present, and T is the total number of breath groups with x cues present.

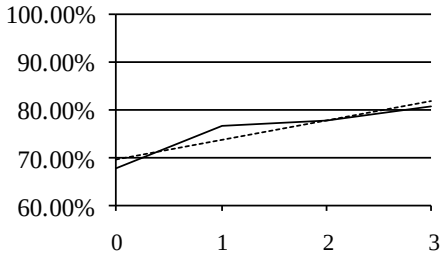

Figure 2. Cues Present v. Probability of YIELD

According to these results, a positive relationship seems to exist for these cues and this corpus. Figure 2 shows the results plotted with a fitted regression. The number of cues present and probability of a turn yield is strongly correlated ($r = .923$, $p=.038$). A regression analysis done using SPSS showed that the adjusted $r^2 = .779$ ($p = .077$).

When no turn-yielding cues are present, there is still a majority chance that the student will yield their turn; however, this is understandable due to the small number of cues being analyzed. Regardless, this gives a very preliminary support for the idea that it is possible to predict when a turn will be taken based on the number of cues present.

5 Conclusions

This paper presented preliminary work in using an HH tutoring corpus to construct a turn-taking model that can later be implemented in a spoken dialogue system. A small set of prosodic features was used to try and identify turn-taking cues by comparing their values at places where students yielded their turn to the tutor and places where they held it. Results show that turn-taking cues such as those investigated can be identified for the corpus, and may hold predictive ability for turn boundaries.

5.1 Future Work

When building on this work, there are two different directions in which we can go. While this work uncovers some interesting results in the tutoring domain, there are some shortcomings in the method that may make it difficult to effectively evaluate the results. As the breath group is different from the segment used in Gravano and Hirschberg's (2009) experiment, and the set of prosodic features is smaller, *direct* comparison becomes quite difficult. The differences between the two methods provide enough doubt for the results to truly be interpreted as contradictory. Thus the first line of future inquiry is to redo this method using a smaller silence boundary (50 ms) and different set of prosodic features so that it is truly comparable to Gravano and Hirschberg's (2009) work with the game corpus. This could yield interesting discoveries in the differences between the two corpora, shedding light on phenomena that are particular to tutoring scenarios.

On the other hand, other researchers have used different segments; for example, Clemens and Diekhaus (2009) divide their corpus by "topic units" that are grammatically and semantically complete. In addition, Litman et. al. (2009) were able to use

word-level units to calculate prosody and classify turn-level uncertainty. Perhaps direct comparison is not entirely necessary, and instead this work should be considered an isolated look at an HH corpus that provides insight into turn-taking, specifically in tutoring and other domains with unequal power levels. Future work in this direction would include growing the set of features by adding more prosodic ones and introducing lexical ones such as bi-grams and uni-grams. Already, work has been done to investigate the features used in the INTERSPEECH 2009 Emotion Challenge using openSMILE (Eyben et. al., 2009). When a large feature bank has been developed, significant cues will be used in conjunction with machine learning techniques to build a model for turn-taking which can be implemented in a spoken dialogue tutoring system. The goal would be to learn more about human turn-taking while seeing if better turn-taking by a computer tutor ultimately leads to increased student learning in an intelligent tutoring system.

Acknowledgments

This work was supported by the NSF (#0631930). I would like to thank Diane Litman, my advisor, Scott Silliman, for software assistance, Joanna Drummond, for many helpful comments on this paper, and the ITSPOKE research group for their feedback on my work.

References

Caroline Clemens and Christoph Diekhaus. 2009. Prosodic turn-yielding Cues with and without optical Feedback. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics.

Anne Cutler and Mark Pearson. 1986. On the analysis of prosodic turn-taking cues. In C. Johns-Lewis, Ed., *Intonation in Discourse,* pp. 139-156. College-Hill.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology,* 24(2):283-292.

Florian Eyben, Martin Wöllmer, Björn Schuller. 2010. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proc. ACM Multimedia (MM), ACM, Florence, Italy. pp. 1459-1462.

James R. Glass. 1999. Challenges for spoken dialogue systems. In *Proceedings of the 1999 IEEE ASRU Workshop.*

Agustín Gravano and Julia Hirschberg. 2009. Turn-yielding cues in task-oriented dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 253--261. Association for Computational Linguistics.

Jackson Liscombe, Julia Hirschberg, and Jennifer J. Venditti. 2005. Detecting certainness in spoken tutorial dialogues. In *Interspeech.*

Diane J. Litman, Carolyn P. Rose, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2006. Spoken Versus Typed Human and Computer Dialogue Tutoring. In *International Journal of Artificial Intelligence in Education*, 26: 145-170.

Diane Litman, Mihai Rotaru, and Greg Nicholas. 2009. Classifying Turn-Level Uncertainty Using Word-Level Prosody. *Proceedings Interspeech*, Brighton, UK, September.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and Challenges of Real-Time Uncertainty Detection and Adaptation in a Spoken Dialogue Computer Tutor. *Speech Communication*, in press.

Diane J. Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *HLT/NAACL*.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proc. NAACL/HLT 2009*, Boulder, CO, USA.

Ethan O. Selfridge and Peter A. Heeman. 2010. Importance-Driven Turn-Bidding for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 177-185.

Nigel Ward, Anais Rivera, Karen Ward, and David G. Novick. 2005. Root causes of lost time and user stress in a simple dialog system. In *Proceedings of Interspeech.*